HAL open science

Independent component analysis by wavelets

Pascal Barbedor

To cite this version:

Pascal Barbedor. Independent component analysis by wavelets. Mathematics [math]. Université Paris-Diderot - Paris VII, 2006. English. NNT : . tel-00119428

HAL Id: tel-00119428

https://theses.hal.science/tel-00119428

Submitted on 9 Dec 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Paris 7 Denis Diderot
UFR de Mathématiques

DOCTORAT

spécialité mathématiques appliquées

PASCAL BARBEDOR

Analyse en composantes indépendantes par ondelettes

Thèse dirigée par Dominique PICARD

soutenue publiquement le 5 décembre 2006 devant le jury composé de

Olivier Bousquet, Pertinence

Jean-François Cardoso, CNRS

Dominique Picard, Université Paris 7

Karine Tribouley, Université Paris 10

Alexandre Tsybakov, Université Paris 6

et au vu des rapports préliminaires de

Jean-François Cardoso, CNRS

Michael I. Jordan, Université de Berkeley

.

J'adresse ma reconnaissance fraternelle et amicale à Dominique Picard qui a dirigé ce travail avec beaucoup de disponibilité et d'ouverture, et mes plus vifs remerciements à Michael I. Jordan et Jean-François Cardoso pour leur appréciation très favorable lors du rapport préliminaire.

Et je remercie tout aussi chaleureusement Olivier Bousquet, Karine Tribouley et Alexandre Tsybakov de me faire l'honneur d'être présents dans le jury. Je regrette que Michael Jordan n'ait pas pu se déplacer, mais c'était un voyage un peu long pour venir jusqu'ici.

Je voudrais aussi remercier le personnel de la bibliothèque du site de Chevaleret que j'ai trouvé toujours charmant et disponible, et l'école doctorale en général.

Une mention spéciale pour Florence Piat qui m'a donné l'idée de me lancer dans cette aventure.

Avant de préciser que la dédicace va à ma famille, amis et collègues ici présents.

P.B.

.

1. Introduction **11**

1. Principales approches du problème ACI 12

2. Motivation 19

3. Résultats théoriques obtenus 24
   Mixage linéaire et appartenance Besov 24
   Contraste en projection 24
   Estimateurs et vitesses 25
   Propriétés de filtrage du gradient et du Hessien 29

4. Résultats pratiques 29
   Utilisation du contraste plug-in 30
   Stabilité numérique 30
   Complexité numérique 31
   Usage d'estimateurs U-statistique 32
   Haar et évaluation directe aux dyadiques 32
   Pouvoir de séparation du contraste 32

5. Éléments de comparaison avec d'autres méthodes 33
   Kernel ICA 34
   Méthodes de la norme de Hilbert-Schmidt 36
   Estimateur RADICAL 38
   Fonctionnelle matricielle 39

6. Perspectives 42
   Prolongements d'ordre pratique 42
   Prise en compte de l'anisotropie 42
   Contraste de type plug-in atteignant une vitesse paramétrique 43
   Test d'indépendance 44
   Adaptation 45
   Gradient exact du contraste L_2 45
   Mixage non linéaire 46
   ACI sans ACP 46

7. Organisation du document 47
   Notations et conventions 47

2. Introduction (english translation) **49**

1. Main approaches in the ICA problem 50

2. Motivation 57

3. Theoretical results 61
   Linear mixing and Besov membership 61
   Contrast in projection 62
   Estimators and rates 62
   Filtering properties of the gradient and the hessian 67

4. Practical results 67
   Plug-in contrast usage 67
   Numerical stability 68
   Numerical complexity 68
   U-statistic estimators usage 69
   Haar and direct evaluation at dyadic rationals 69
   Separating power of the contrast 70

5. Comparison with other methods 71
   Kernel ICA 71
   Hilbert-Schmidt norm method 73
   RADICAL 75
   Matrix functional 76

6. Perspectives 78
   Practical extensions 79
   Taking into account anisotropy 79
   Plug-in type contrast reaching a parametric rate 79
   Test of Independence 81
   Adaptation 82
   Exact gradient of the L_2 contrast 82
   Non linear mixing 82
   ACI without pre-whitening 82

7. Organization of the document 83
   Notations et conventions 84

8. Bibliographie de l'introduction 85

3. ICA by Wavelets : the basics **89**

1. Notations 90

2. Wavelet contrast, Besov membership 94
   Wavelet contrast 94
   Besov membership of marginal distributions 96
   Besov membership of the mixed density 97

3. Risk upper bound 98
   Risk upper bound for $\hat{C}_j$ 99
   Minimizing resolution in the class $B_{s2\infty}$ 99
   minimizing resolution in the class B_{spq} 100

4. Computation of the estimator $\hat{C}_j$ 100
   Sensitivity of the wavelet contrast 101
   Contrast complexity 103

5. Contrast minimization 104
   A visual example in dimension 2 106

6. Appendix 107
   Bias of $\hat{\alpha}_{jk}^2 - 2\hat{\alpha}_{jk}\hat{\lambda}_{jk} + \hat{\lambda}_{jk}^2$ 107
   Decomposition of $\sum F_1(X_{i_1}) \ldots \sum F_m(X_{i_m})$ 109
   rth order moment of Φ_{jk} 110

7. References for ICA by wavelets : the basics 110

4. ICA and estimation of a quadratic functional **113**

Estimation of a quadratic functional 114
Wavelet ICA 115

1. Notations 117

2. Estimating the factorization measure $\int (f_A - f_A^\star)^2$ 119
ICA wavelet contrast 119
Wavelet contrast and quadratic functional 120
Estimators under consideration 120
Notational remark 121
Sample split 122
Bias variance trade-off 123
Minimal risk resolution in the class $B_{s2\infty}$ and convergence rates 123

3. Risk upper bounds in estimating the wavelet contrast 124
Risk upper bound, $d+1$ independent samples — f_A, $f_A^{\star 1}, \ldots, f_A^{\star d}$ 124
Risk upper bound, 2 independent samples — f_A, $f_A^\star$ 124
Full sample $\hat{C}_j^2$ risk upper bound 126
Risk upper bound, full sample — f_A 127

4. Appendix 1 – Propositions 131
2nd moment of $\sum_k \hat{\alpha}_{jk}^2$ about $\sum_k \alpha_{jk}^2$ 131
2nd moments $\sum_k \hat{\lambda}_{jk}^2$ about $\sum_k \lambda_{jk}^2$ and $\sum_k \hat{\lambda}_{jk}\hat{\alpha}_{jk}$ about $\sum_k \lambda_{jk}\alpha_{jk}$ 132
Variance of $\sum_k \hat{B}_j^2$ 135
Variance of multisample $\prod \sum_k \hat{B}_j^2(\tilde{R}^\ell)$ 136
Variance of multi sample $\sum_k \hat{\alpha}_{jk}\hat{\lambda}_{jk}$ 137

5. Appendix 2 – Lemmas 138
Property set 138
Many sets matching indices 139
Two sets matching indices [Corollary and complement] 140
Product of r kernels of degree m 141
Meyer 142
Path of non matching dimension numbers 143
Daubechies wavelet concentration property 143
rth order moment of Φ_{jk} 144

6. References for ICA and estimation of a quadratic functional 145

5. Towards thresholding **147**

1. Estimating $\int (f_A - f_A^\star)^2$ with term-by-term thresholding 148
   ICA wavelet contrast 149
   Wavelet contrast in place of the quadratic functional $\int (f_A - f_A^\star)^2$ 149
   Hard-thresholded estimator 150
   Block threshold 151

2. Risk upper bound 152
   Risk of a thresholded procedure 152

3. Practical issues 159
   Computation of the estimator $\hat{C}_j$ 160
   Choice of the threshold in practice 162

4. Appendix 1 – Propositions 166
   rth moment of $\hat{\beta}_{jk}$ and $\hat{\mu}_{jk}$ 166
   Product of r kernels of degree m 168

5. Appendix 2 – Combinatorial lemmas 169
   Property set 169
   Many sets matching indices 169
   Path of non matching dimension numbers 170

6. Appendix 3 – Thresholding related lemmas and others 171
   wavelet contrast in β_{jk} 171
   Large deviation for term by term thresholding 172
   Number of big coefficients 174
   rth order moment of Ψ_{jk} 175
   Daubechies wavelet concentration property 175
   Bernstein inequality 176

7. References for Towards thresholding 176

6. Stiefel manifold, optimization and implementation issues **178**

1. Direct evaluation of $\varphi_{jk}(x)$ at floating point numbers 178

2. Relocation by affine or linear transform 179

3. Wavelet contrast differential and implementation issues 180
   $\hat{C}_j$ differential in Y 180
   $\hat{C}_j$ differential in W 181
   $\hat{C}_j$ second derivatives 183
   $\hat{C}_j$ hessian in Y 184
   $\hat{C}_j$ hessian in W 184

4. Filter aware formulations for the gradient and the hessian 185
   filter aware gradient formulation 186
   filter aware hessian formulation 187

5. Wavelet contrast implementation 188
   Flat representation routines 188
   DWT in dimension 1 189
   DWT in dimension d 190
   Thresholding in dimension d 192
   Computation of φ values at dyadic rationals 192
   Contrast computation 196
   Haar projection (Daubechies $D2N, N = 1$) 197
   Projection on V_j spanned by a Daubechies $D2N$, general case 198

6. Optimization on the Stiefel manifold 200

1. Introduction

Étant donné un nuage de n points en dimension d, l'analyse en composantes principales (ACP) consiste à trouver la direction de l'espace qui porte la plus grande part de la dispersion totale ; puis la direction orthogonale à la précédente portant la deuxième plus grande part, et ainsi de suite. Du point de vue géométrique, c'est un simple changement de base, qui a pour effet une nouvelle représentation dans des directions correspondant à des variables décorrélées (au sens empirique), et obtenant par là même le statut de facteurs, susceptibles d'expliquer l'information contenue dans le nuage sans la redondance qui pouvait caractériser la représentation de départ.

L'exemple emblématique du cocktail party illustre le point de vue de la séparation de sources : on enregistre d conversations simultanées en plaçant d microphones bien répartis dans la pièce, chacun enregistrant une superposition de toutes les conversations, mais un peu plus nettement celles qui se trouvent à proximité directe. Le problème est d'isoler chacun des discours pour comprendre ce qui s'est dit.

En psychométrie, une des disciplines précurseur de l'analyse factorielle, on utilise la notion de variables latentes, c'est-à-dire inobservables directement mais dont les effets sont mesurables à travers des batteries de test, chaque test révélant, en partie et entre autre, l'effet de telle ou telle variable du modèle latent. Un exemple connu est celui des cinq grandes dimensions de la personnalité (Big Five) identifiées à partir de l'analyse factorielle par des chercheurs en évaluation de la personnalité (Roch, 1995).

D'autres domaines d'applications tels que l'imagerie numérique, le datamining, l'économie, la finance ou encore l'analyse (statistique) de textes, ont recours à des modèles, de type inverse, ayant pour objet de révéler des facteurs explicatifs indépendants à extraire d'une accumulation d'indicateurs éventuellement très divers, et globalement liés à un phénomène.

L'analyse en composante indépendantes (ACI) est un outil de ce type, avec ou sans la notion de réduction de la dimension d, et dont l'ambition est par ailleurs de révéler des facteurs qui soient indépendants au sens plein, et non plus seulement au sens de la non corrélation.

Le modèle se formule a minima de la façon suivante : soit X une variable aléatoire sur $\mathbb{R}^d$, $d \geq 2$, telle que $X = AS$ où A est une matrice carrée inversible et S une variable aléatoire latente dont les composantes sont mutuellement indépendantes. On se propose d'estimer A, pour atteindre $\{S_1, \ldots, S_n\}$, à partir de la donnée d'un échantillon $\{X_1, \ldots, X_n\}$ indépendant, identiquement distribué selon la loi de X, c'est-à-dire tel que X_i est indépendante de X_j pour $j \neq i$, mais tel que les composantes $X_i^1, \ldots, X_i^d$ d'une même observation X_i ne sont a priori pas mutuellement indépendantes.

L'ACI ainsi formulée considère uniquement des superpositions linéaires de signaux indépendants, résultant du mixage par A. C'est souvent une restriction légitime ; par

exemple les systèmes de transmission sont des milieux linéaires où les signaux agissent comme s'ils étaient présents indépendamment les uns des autres, ils n'interagissent pas mais s'additionnent (Pierce, 1980).

D'autres formulation prennent en compte des mélanges dits post nonlinéaire (Taleb, 1999, Achard, 2003) issus de transformations monotones de mélanges linéaires s'exprimant par $X = f(AS)$, f de composantes monotones et A inversible. Dans d'autres cas encore, on ne s'appuie pas sur l'indépendance, et on exploite au contraire une corrélation temporelle des signaux sources. Il existe aussi des modèles convolutifs, où le mixage par A n'est pas instantané, mais de la forme $X(t) = \sum_u A(u)S(t-u)$, $A(u)$ désignant une suite de matrices inversibles. Enfin, on peut aussi considérer des modèles bruités.

Le modèle étudié dans la suite du document est le problème standard défini plus haut. Pour ce problème, en pratique, la transformation ACP (empirique) de la matrice $d \times n$ $M = (\, x_1 \quad \dots \quad x_n \,)$ contenant le signal observé fournit la première partie de la réponse au problème ACI, par décorrélation simple ; et toute la réponse si les signaux mesurés sont purement gaussiens. Dans le cas contraire, pour résoudre le problème entièrement, la procédure usuelle consiste à minimiser une certaine fonction de contraste $C = C(W)$ qui s'annule si et seulement si les composantes de WX sont indépendantes, où W est une matrice $d \times d$ candidate à l'inversion de A.

Le problème ACI standard est toujours paramétrique en A, et est paramétrique ou non paramétrique en S suivant les hypothèses fonctionnelles appliquées à la densité de probabilité de S, qui est le plus souvent supposée admettre une densité par rapport à la mesure de Lebesgue (des modèles de déconvolution discrète sont aussi étudiés, voir notamment Gassiat et Gautherat — 1999).

Le modèle de densité de l'ACI s'écrit de la façon suivante ; soit f la densité de S par rapport à la mesure de Lebesgue, la variable observée $X = AS$ admet la densité f_A, définie par

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

où b_ℓ est la ligne numéro ℓ de la matrice $B = A^{-1}$; cette écriture résulte d'un changement de variable étant donné que f, la densité de S, est le produit de ses marges $f^1 \dots f^d$.

Dans le modèle ACI exprimé ainsi, f et A sont les deux inconnues, et la donnée consiste en un échantillon indépendant et identiquement distribué $\{X_1, \dots, X_n\}$ de f_A. Le cas semi-paramétrique correspond à des hypothèses non paramétriques pour f, dont la forme n'est pas spécifiée, mis à part les critères généraux de régularité nécessaires à l'estimation.

1.1 Principales approches du problème ACI

On trouve au fil de la littérature abondante sur l'ACI les avantages et les inconvénients

des différentes méthodes ayant été proposées depuis les années 1980 et surtout 1990 où les premiers algorithmes performants ont été introduits.

On constate également que les propriétés statistiques (notamment les vitesses de convergence) des modèles classiques sont généralement peu étudiées explicitement, puisque cela ne correspond pas forcément à une priorité pour les chercheurs des domaines concernés, et que d'autre part le critère d'Amari peut tenir lieu de référence dans la validation d'une méthode.

La performance d'un algorithme n'est en effet pas mesurée spécifiquement par le degré de suppression de la dépendance, mais par sa capacité à inverser la matrice A. Le critère utilisé est la distance de Amari (1996) normalisée de 0 à 100 entre la matrice A (connue dans les simulations) et l'estimation de son inverse ACI, W,

$$\frac{100}{2d(d-1)}\left[\sum_{i=1}^{d}(\sum_{j=1}^{d}\frac{|p_{ij}|}{\max_k|p_{ik}|}-1)+\sum_{j=1}^{d}(\sum_{i=1}^{d}\frac{|p_{ij}|}{\max_k|p_{kj}|}-1)\right] \qquad \text{avec } WA=(p_{ij}).$$

Les deux objectifs portent d'ailleurs des noms différents (on parle d'estimation du mélange ou de restitution de source) et ne coïncident pas dans certains types de modèles où A n'est pas carrée inversible.

La plupart des méthodes classiques ne font pas à proprement parler d'hypothèses paramétriques pour f, mais n'entraînent pas pour autant la mise en œuvre de méthodes typiquement non paramétriques.

L'explication tient au fait que ces méthodes ont recours à des contrastes de substitution dont l'annulation n'implique pas exactement l'indépendance mutuelle, et qui, du point de vue théorique, sont plus faciles à estimer qu'un critère exact. Dans beaucoup de cas les contrastes utilisés fournissent de fait un inverse de A très satisfaisant.

Les méthodes mettant en avant leur caractère non paramétrique sont basées sur des contrastes exacts, dont l'annulation implique l'indépendance mutuelle des composantes (parfois seulement par paire), et dont l'évaluation en toute généralité n'est possible qu'à travers les contraintes techniques particulières de l'estimation non paramétrique.

Mis à part le fait que les méthodes classiques sont souvent caractérisées par une plus grande facilité d'implémentation et une complexité numérique peu élevée, la distinction la plus évidente entre les deux groupes est sans doute le recours ou non à un paramètre de régularisation, élément clef dans les techniques d'approximation de fonctions, et élément qui se trouve relié aux performances générales de l'algorithme ACI dans les cas où les propriétés statistiques sont le plus clairement établies.

Méthodes classiques

Des méthodes issues du maximum de vraisemblance et des contrastes basés sur l'information mutuelle ou autre mesures de divergence ont souvent été utilisées. Comon (1994) a défini le concept de l'ICA comme la maximisation du degré d'indépendance statistique entre outputs, en utilisant des fonctions de contraste approchées par un développement d'Edgeworth de la divergence de Kullback-Leibler. Par ailleurs Hyvärinen et Oja (1997) ont proposé l'algorithme FastICA, qui est une méthode de minimisation par point fixe (au lieu de méthode du gradient) applicable à plusieurs types de contrastes ACI.

On identifie habituellement quatre grandes catégories de méthodes (voir Hyvärinen et al. 2001) :

- Maximisation du caractère non gaussien (Hyvärinen, 1999) ;

  On cherche à maximiser le caractère non gaussien d'une combinaison linéaire ${}^t bx$ de l'observation x. Le caractère non gaussien est estimé à partir du kurtosis ou de la negentropie.

  Il s'agit d'une méthode déflationniste, c'est-à-dire où on opère composante par composante, par opposition aux approches simultanées où on cherche directement W permettant d'extraire toutes les composantes en même temps. Cela entraîne généralement que l'erreur dans l'estimation de la première composante s'accumule et augmente l'erreur des composantes suivantes (Hyvärinen et al, 2001, p. 271). C'est également une méthode peu robuste en présence de données aberrantes, en raison d'une croissance cubique en $|y|$ du kurtosis (idem, p. 277).

  La maximisation s'opère par une méthode du gradient simple ou par une méthode de point fixe (FastICA).

- Maximum de vraisemblance et notamment l'algorithme Infomax (Bell, Snejowski, 1995) ;

  La vraisemblance d'un échantillon $\{x_1, ..., x_n\}$ s'écrit

$$\frac{1}{n} \log L(x_1, \ldots, x_n, B) = \log |\det B| + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \log f^j(\ {}^t b_j x_i).$$

  Dans l'ICA par maximisation de la vraisemblance, on cherche les densités f^j dans une famille $\{p_\theta\,;\ \theta \in \Theta\}$ où Θ ne contient que 2 valeurs correspondant l'une à une densité supergaussienne, et l'autre à une densité subgaussienne.

  Cela s'explique par le fait que pour toute fonction G paire et de classe C^2, et soit $z = NAs$ la transformée ACP de $x = As$, les maxima (resp. minima) locaux en w de $EG(\ {}^t wz)$ sous la contrainte $\|w\| = 1$ sont les lignes ℓ de la matrice NA, telles que $E[s^\ell G'(s^\ell) - G''(s^\ell)] > 0$ (resp. > 0), où s^ℓ est la composante ℓ de s. La condition est issue d'un développement de Taylor.

La fonction G divise donc l'espace des densités en deux selon le signe du critère associé à G et s^ℓ, et un représentant de chaque demi-espace est suffisant pour obtenir la convergence de l'algorithme (voir Hyvärinen, 2001, p.201).

Dans le cas du maximum de vraisemblance, la fonction $G = (G^1, \ldots, G^d)$ est donnée par

$$G^\ell = \log f^\ell(\ {}^t wz)\,;$$

dans cette expression, toute densité dans le même demi-espace que f^ℓ convient également ; ainsi dans l'algorithme de Bell et Snejowski (années 1990) on prend $(G^\ell_+)'(y) = -2\tanh(y)$ et $(G^\ell_-)'(y) = \tanh(y) - y$.

L'algorithme Infomax (Bell et Snejowski, 1995), basé sur un réseau de neurones, est assimilable à une méthode de maximum de vraisemblance (Cardoso, 1997).

Il faut noter que les méthodes basées sur la maximisation du caractère non gaussien et les méthodes du maximum de vraisemblance (utilisant le découpage précédent) sont en défaut si on se trompe de demi-espace dans l'initialisation de l'algorithme.

- Méthodes tensorielles, et notamment les méthodes FOBI et JADE (Cardoso, 1990, 1994).

  Pour supprimer les corrélations jusqu'à l'ordre 4, de $x = (x^1, \ldots, x^d)$ on considère les tenseurs de cumulants $\mathrm{cum}(x^i, x^j, x^k, x^l)$ qui généralisent les matrices de covariance $\mathrm{cov}(x^i, x^j)$.

  Soit F le tenseur d'ordre 4, agissant sur des matrices $d \times d$, défini par

  $$M, d \times d \mapsto F_{ij}(M) = \sum_{k,l} M_{kl}\,\mathrm{cum}(x^i, x^j, x^k, x^l)\,;$$

  une matrice propre de F est une matrice M telle que $F(M) = \lambda M$, $\lambda \in \mathbb{R}$.

  On montre que les matrices orthogonales $W = w_m\ {}^t w_m$, $m = 1 \ldots, n$, sont les matrices propres de F. Les valeurs propres associées sont les kurtosis des composantes indépendantes s^ℓ.

  La méthode JADE (joint approximate diagonalization of eigenmatrices) s'appuie sur ce principe (Cardoso, 1993).

- Décorrélation non linéaire, algorithme de Jutten et Hérault (1987), Cichocki et Ubenhauen (1992).

  On cherche à estimer $Ef(y^1)g(y^2)$, idéalement pour toutes fonctions f, g continues à support compact, passant ainsi d'un critère de décorrélation simple à un critère d'orthogonalité, équivalent à une indépendance si au moins une des deux variables est centrée.

  En utilisant un développement de Taylor, le critère devient

  $$\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{1}{j!i!} f^{(i)}(0) g^{(j)}(0) E(y^1)^i (y^2)^j + R = 0$$

et le problème se transforme en une décorrélation des composantes y^1, y^2 à toutes les puissances i, j du développement à l'ordre p. Cela revient donc à exploiter les différents moments des observations, avec ce que cela comporte comme problèmes de robustesse. Jutten et Hérault sont historiquement parmi les premiers à présenter un algorithme ACI. On sait que leur méthode présente des problèmes de convergences dans certaines situations (J.C. Fort, 1991) ; Amari et Cardoso (1997) en ont proposé une extension (fonctions estimables).

Information mutuelle

Hyvärinen et al. (2001, p. 274) font remarquer que le critère théorique de l'information mutuelle entre les composantes de $x = As$ possède un caractère universel, au sens où il peut-être vu comme la limite du principe de maximisation de la vraisemblance auquel s'assimilent beaucoup de méthodes.

Soit un estimateur $\hat{f}_A$ de la densité f_A de $X = AS$; la vraisemblance $L(x_1, \ldots, x_n, \hat{f}_A) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_A(x_i)$ tend vers $E_{f_A} \log \hat{f}_A(x)$, quantité qui s'écrit aussi en utilisant la substitution $\hat{f}_A = \hat{f}_A f_A^{-1} f_A$,

$$\int f_A \log \hat{f}_A = \int f_A \log f_A + \int f_A \log \frac{\hat{f}_A}{f_A} = -K(f_A || \hat{f}_A) - H(f_A),$$

L'entropie différentielle $H(f_A)$ est une constante indépendante de A, puisqu'elle est invariante par transformation inversible (ou plus précisément orthogonale, contrairement à l'entropie de Shannon – Comon 1994, p.293 –, ce qui suppose qu'on s'est ramené au cas A orthogonale par pré-transformation ACP), de même que la divergence de Kullback-Leibler K, et puisque $s = A^{-1}x$,

$$K(f_A(x) || \hat{f}_A(x)) = K(f(s) || \hat{f}(s)) \, ;$$

les composantes de s étant indépendantes, la divergence se décompose aussi en

$$K(f(s) || \hat{f}(s)) = K(f(s) || f^\star(s)) + K(f^\star(s) || \hat{f}(s))$$

où $f^\star(s)$ est la densité produit des composantes de s. Cette quantité est minimum lorsque $\hat{f}_A = f_A^\star$ et vaut $K(f(s) || f^\star(s))$ qui n'est autre que l'information mutuelle des composantes de s.

L'information mutuelle du couple (x, y) est aussi définie par $I(x, y) = H(x) + H(y) - H(x \otimes y)$, où H est l'entropie différentielle et $x \otimes y$ la composée ; Cette quantité est toujours positive et vaut zéro si et seulement si x et y sont indépendantes.

Le critère est généralement utilisé sous la forme d'un minimum d'entropie marginale : si $Y = BS$ est une variable aléatoire de dimension d, $I(Y) = \sum_i H(Y^i) - H(S) - \log|\det B|$, le terme log étant nul si on se restreint aux matrices B orthogonales. On cherche donc dans ce cas $W^* = \operatorname{argmin}_W \left[H(Y^1) + \ldots + H(Y^d)\right]$, où Y^ℓ est la composante ℓ de Y.

On peut noter que cette manière de procéder ne permet pas de savoir si le minimum obtenu est proche de l'indépendance mutuelle, puisque $H(S)$ n'est pas connue (et $H(Y)$, intégrale en dimension d, n'est pas estimée). Le critère du minimum d'entropie marginale élimine donc la possibilité de construire un test statistique permettant de décider que le minimum a été atteint.

Méthodes d'essence non paramétrique

Ces méthodes peuvent être classées en trois groupes : les méthodes à noyau, les méthodes à contraste exact, les méthodes directes.

- Groupe 1 — kernel ICA (Bach et Jordan, 2002) et les méthodes basées sur une norme de Hilbert-Schmidt (Gretton et al., 2003, 2004) qui fournissent des critères exacts d'indépendance des composantes par paire plutôt que mutuelle, à partir de méthodes à noyau dans un espace de Hilbert à noyau reproduisant (RKHS).

  Le contraste exact visé par les méthodes kernel est le critère d'indépendance de deux tribus, c'est-à-dire X et Y sont indépendantes si $Efg = EfEg$ pour toute fonction f de carré intégrable sur les ensembles de la tribu générée par X et toute fonction g de carré intégrable sur les ensembles de la tribu générée par Y. On suppose donc implicitement que le RKHS sur lequel le critère est effectivement vérifié est inclut dans l'espace $L_2(\sigma(X))$ et $L_2(\sigma(Y))$.

  Pour deux variables aléatoires réelles x et y, le contraste de kernel ICA est défini comme la $\mathcal{F}_\sigma$-corrélation $\rho_{\mathcal{F}}$,

$$\rho_{\mathcal{F}_\sigma} = \max_{f,g \in \mathcal{F}_\sigma} \operatorname{corr}(f(x), g(y)).$$

  La $\mathcal{F}_\sigma$-corrélation est nulle pour x et y indépendantes, et, pour un espace de fonction $\mathcal{F}_\sigma$ suffisamment gros, contenant par exemple les fonctions $x \mapsto e^{iwx}$, une $\mathcal{F}_\sigma$-corrélation nulle implique l'indépendance de x et de y. Dans l'ACI à noyau, $\mathcal{F}_\sigma$ est issu du noyau gaussien isotrope $k(x,y) = e^{-\frac{1}{2\sigma^2}\|x-y\|^2}$, avec $\emptyset = \mathcal{F}_0$ et $\mathcal{F}_\sigma$ croît vers $L_2(\mathbb{R}^d)$, quand $\sigma^2 \to +\infty$.

  Dans les méthodes basées sur la norme de Hilbert-Schmidt, on considère deux espaces mesurés $(\mathcal{X}, \Gamma, p_x)$ et $(\mathcal{Y}, \Lambda, p_y)$, et l'espace produit $(\mathcal{X} \times \mathcal{Y}, \Gamma \otimes \Lambda, p_{xy})$.

  On considère également $\mathcal{F}$, un RKHS de fonctions de $\mathcal{X}$ dans $\mathbb{R}$ associé au noyau défini positif $k(.,.): \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, et à la fonction caractéristique ϕ, $\phi(x) = k(x,.)$; et on considère $(\mathcal{G}, l(.,.), \psi)$, un second RKHS de fonctions de $\mathcal{Y}$ dans $\mathbb{R}$.

  Le contraste est défini comme la norme de Hilbert-Schmidt de l'opérateur de covariance croisée

$$C_{xy} = E_{xy}\big[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)\big]$$

  où E_{xy} est l'espérance par rapport à la loi jointe p_{xy}, $\otimes$ est le produit tensoriel défini par $f \otimes g: h \in \mathcal{G} \mapsto f < g, h >_{\mathcal{G}} \in \mathcal{F}$ et μ_x, μ_y sont donnés respectivement par $< \mu_x, f >_{\mathcal{F}} = E_x f(x)$ et $< \mu_y, g >_{\mathcal{G}} = E_y g(y)$.

En dimension supérieure à 2, on opère sur toutes les associations par paires.

- Groupe 2 — information mutuelle, et méthodes basées sur une estimation de densité, implicite ou non ; par exemple la méthode Radical (Miller et Fisher III, 2003) basée sur un estimateur de l'entropie différentielle initialement introduit par Vasicek (1976) ;

$$H_{nm} = n^{-1} \sum_{i=1}^{n} \log \frac{n}{2m} (X_{(i+m)} - X_{(i-m)})$$

où $X_{(i)}$ est la statistique d'ordre associée à un échantillon i.i.d. $X_1, \ldots, X_n$.

Un autre estimateur de l'information mutuelle utilisé dans le contexte ACI par Boscolo et al. (2001) s'écrit sous la forme

$$L(W) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{d} \log f_A^{\star\ell}(w_\ell x_i) + \log |\det W|,$$

où w_ℓ est la ligne ℓ de la matrice candidate W, et les auteurs estiment les densités marginales $f_A^{\star\ell}$ à l'aide d'un estimateur à noyau gaussien classique, qu'ils substituent ensuite dans la formule ci-dessus.

Pham (2004), propose également des algorithmes rapides basés sur l'estimation du critère $C(W) = \sum_{\ell=1}^{d} H(WX^\ell) - \log \det W$, utilisant un estimateur à noyau spline pour les entropies marginales, et des fonctions score liées au gradient de l'entropie pour la minimisation. D'autre part la méthode s'appuie sur une estimation discrétisée de la densité multipliée par son log, $f \log f$.

- Groupe 3 — méthode de la fonctionnelle matricielle (Tsybakov et Samarov, 2001) ; c'est une des méthodes les plus abouties au plan théorique, basée sur un estimateur à noyau d'une fonctionnelle matricielle fournissant une solution algébrique directe au lieu d'un contraste à minimiser numériquement.

Cette méthode donne une estimation consistante de A^{-1} à la vitesse $\sqrt{n}$ sans l'auxiliaire d'une fonction de contraste à minimiser, à partir d'une fonctionnelle utilisant le gradient de f_A. Mis à part certaines densités exclues de fait, la méthode est donc optimale, et donne une consistance de l'estimation directe de A plutôt que de celle d'un contraste qui reste à minimiser. En pratique, la complexité numérique est au minimum en $O(n^2)$.

1.2 Motivation

La différence entre les méthodes classiques et les méthodes non paramétriques (c'est-à-dire ici à paramètre, mais de régularisation) paraît finalement assez mince, puisque on a dans un cas un contraste biaisé, estimé avec un certain degré d'incertitude, et dans l'autre un contraste exact mais dont l'estimation est nécessairement biaisée.

Les critères classiques tendent généralement vers l'information mutuelle, mais sans mention d'aucune vitesse puisqu'on ne s'appuie pas sur une classe fonctionnelle définie par une notion de régularité, par exemple de type module de continuité (on suppose parfois que la densité est "suffisamment dérivable"). D'autre part ces contrastes utilisent souvent des critères basés sur les moments de la distribution, ce qui ne nécessite aucune connaissance de la forme de la densité.

La motivation des approches d'essence non paramétriques est donc de partir de critères d'indépendance exacts, dont on espère qu'ils permettront d'atteindre une inversion de A plus précise, et de s'appuyer pour leur estimation sur le cadre statistique étendu de l'approche non paramétrique. Le critère de l'information mutuelle montre que le problème de l'estimation d'un contraste ACI exact est fondamentalement celui de l'estimation d'une fonctionnelle non linéaire d'une densité.

- Première observation ; dans beaucoup de cas, la précision supplémentaire à laquelle on accède sur le papier est quelque peu écornée par les difficultés de mise en œuvre ou une complexité numérique élevée qui obligent souvent à rétrocéder quelque chose dans le calcul concret.

  Dans la méthode Radical, constatant l'insuffisance de prise en compte de la régularité par le seul paramètre m, les auteurs ajoutent une seconde régularisation par bruitage, en remplaçant chaque observation X_i par $\sum_{j=1}^{R} \epsilon_j$ où ϵ_j suit une loi normale $N(X_i, \sigma^2 I_d)$. Si la consistance de l'estimateur de Vasicek est montrée dans l'article de Vasicek (1976) et dans un article ultérieur de Song (2000), la seconde régularisation change le problème.

  Les méthodes à noyau s'appuient par construction sur une régularisation définissant la largeur du noyau, en général gaussien isotrope. Et là encore c'est insuffisant en pratique. Dans kernel ICA, le problème de valeurs propres généralisé n'a pas de solution sans une régularisation additionnelle, dont l'effet statistique n'est pas totalement clair. La méthode kernel de Hilbert-Schmidt ne demande pas formellement de seconde régularisation, mais le bon fonctionnement du contraste nécessite un seuil de signification γ jouant à peu près le même rôle. D'une manière générale, l'utilisation d'une décomposition de Choleski incomplète dans les méthodes kernel introduit des paramètres supplémentaires dont on ne connaît pas les valeurs optimales.

- Autre remarque d'ordre général, les simulations montrent que les méthodes non paramétriques donnent de bons résultats dans un grand éventail de situations, mais l'une ou l'autre des méthodes classiques n'est jamais très loin en terme d'erreur Amari (voir des exemples y

compris dans les simulations de l'article de Bach et Jordan), et finalement aucune méthode ne se détache des autres dans tous les cas de figure.

Dans la fonctionnelle matricielle, on a pu constater de moins bons résultats que les méthodes classiques dans le cas de densités multimodales (Gómez Herrero, 2004, p. 15).

Dans la méthode Hilbert-Schmidt, les auteurs annoncent d'excellents résultats comparés aux méthodes classiques et à kernel ICA, sauf dans le cas où le nombre d'observations est faible ($n = 250$).

Sur l'ensemble test de Bach et Jordan en dimension 2, les simulations de Radical montrent des résultats en moyenne meilleurs que FastICA, Jade, Infomax et kernel ICA.

D'une manière générale, il semble que les méthodes classiques sont performantes notamment pour des densités unimodales où la notion de supergaussienne et subgaussienne a un sens visuel.

- Troisième observation ; pour avoir un intérêt pratique, une méthode ACI doit posséder une complexité numérique modérée, et sur ce point les méthodes classiques ont l'avantage. Il semble impossible d'obtenir à la fois une optimalité théorique et une complexité numérique faible.

  Les méthodes à noyau réduisent considérablement la complexité du contraste grâce à une décomposition de Choleski incomplète. Mais cela revient à introduire de nouveaux paramètres dont les valeurs optimales ne sont pas connues.

  Les méthodes basées sur l'information mutuelle n'estiment jamais l'information mutuelle de la densité multivariée, mais seulement celle des composantes, pour couper court à une estimation coûteuse. En échange de ce gain dans la complexité numérique, on perd la possibilité de savoir si le minimum atteint est loin du minimum absolu.

  La fonctionnelle matricielle est optimale au plan théorique, mais la forme de l'estimateur implique une complexité numérique élevée, au moins de l'ordre de $O(n^2)$. Il faut également utiliser des noyaux possédant $d + 3$ moments nuls, ce qui n'est pas une situation idéale en grande dimension.

- Enfin, certaines méthodes fonctionnent sous conditions ou bien, de fait, ne fonctionnent pas sans la pré-transformation ACP, qui est une opération non-linéaire dépendant de l'échantillon. Aucune méthode ne prend en compte l'anisotropie.

  Ainsi pour la fonctionnelle matricielle, dans les simulations pratiquées par Gómez Herrero (2004), l'estimation avec une fenêtre h uniforme dans les d dimensions s'avère instable sans une pré-transformation ACP, pourtant non formellement prévue par la méthode.

  Il est important de noter que la pré-transformation ACP dépend entièrement de l'échan-

tillon courant, et que son utilisation introduit dès le départ une rupture non linéaire dans le problème ACI. Cardoso (1994, 1999) a d'ailleurs montré que le pouvoir de séparation des algorithmes ACI adoptant la pré-transformation ACP admettait une borne inférieure ; il souligne aussi le fait que le minimum d'un critère d'indépendance ne correspond pas forcément à une décorrélation totale, puisqu'on n'obtient jamais une indépendance exacte dans la pratique ; autrement dit si la décorrélation est souhaitée, elle doit être effectuée séparément, en plus de la minimisation du contraste.

Dans la plupart des méthodes, au maximum une seule composante gaussienne est autorisée pour le bon fonctionnement des contrastes, et y compris dans le cas de la fonctionnelle matricielle.

Le problème de l'identifiabilité du modèle ACI est discuté dans un article de Comon (1994). Pour une source s à composantes indépendantes et si $x = As$ a des composantes indépendantes, $\operatorname{cov} x$ et $\operatorname{cov} s$ sont diagonales, $\operatorname{cov} x = A \operatorname{cov} s \, {}^tA$, et A peut être identifiée à n'importe quelle matrice de la forme $A = (\operatorname{cov} x)^{\frac{1}{2}} Q (\operatorname{cov} s)^{-\frac{1}{2}}$, où Q est orthogonale. La matrice A de $x = As$ est identifiable à une permutation et normalisation près si et seulement si au plus une seule composante de la source est gaussienne.

Les gaussiennes sont donc exclues du champ de l'ACI, mais en pratique il n'est pas impossible d'en rencontrer, et un contraste qui conserve un sens même en présence de plus d'une gaussienne peut être intéressant, même si une minimisation post-ACP ne pourra pas fonctionner pleinement.

▪

La motivation de ce projet a été d'étudier dans quelle mesure un contraste à base d'ondelettes pouvait apporter quelque chose dans la résolution du problème ACI, étant donné les nombreuses méthodes déjà existantes.

La théorie des ondelettes a été conçue pour pouvoir approcher des fonctions ou des surfaces éventuellement très irrégulières et est un outil efficace utilisé dans les domaines de la compression de données, de l'image et de la théorie du signal. Son utilisation en statistique non paramétrique s'est révélée également très puissante. Les algorithmes computationnels associés aux ondelettes sont rapides ; par exemple la projection sur une base d'ondelettes à une certaine résolution et le changement de résolution sont des opérations numériquement stables et linéaires en n.

A la différence d'une onde sinusoïdale, une ondelette est de durée finie ; le terme renvoie ainsi à une localisation en fréquence et à une localisation temporelle (ou spatiale). Les irrégularités spatiales (multi-modes, discontinuités, mélanges) peuvent être retranscrites efficacement sur une base d'ondelettes et la description de fonctions compliquées tient généralement en un petit nombre de coefficients, souvent plus faible que dans l'analyse de Fourier classique.

La projection linéaire d'une fonction f sur une base d'ondelettes à un niveau de résolution j opère un lissage qui revient à fixer des coefficients de détails à la valeur zéro. Une autre option consiste à ne conserver que les coefficients plus grands que un certain seuil ; le résultat est une projection non linéaire appelée seuillage qui fournit des procédures adaptatives dans le cas où la régularité de f n'est pas connue.

L'autre intérêt des ondelettes est leur articulation avec les espaces de Besov, espaces d'approximation par excellence, qui généralisent des espaces fonctionnels classiques, comme les espaces de Sobolev, Hölder, etc., ayant connu historiquement des développements séparés.

Le cadre fonctionnel des espaces de Besov permet d'obtenir des vitesses de convergence très précisément reliées à la régularité de la densité latente du problème et des conditions d'utilisation éventuellement adaptatives. On espère ainsi pouvoir disposer d'un cadre exact qui ne se dilue pas ensuite dans une succession de paramètres de réglage à fixer selon des principes généralement empiriques.

Une motivation a posteriori de ce projet est d'offrir une alternative à l'information mutuelle, qui reste vue comme un critère difficile à estimer (Bach et Jordan, 2002, p.3).

L'estimateur de l'information mutuelle proposé par Vasicek (1976) s'obtient par le changement de variable $F(x) = p$, $0 \leq p \leq 1$, et en dérivant $F(F^{-1}(p)) = p$, l'entropie différentielle H s'écrit

$$H = -\int_{-\infty}^{+\infty} [\log f(x)] f(x) dx = -\int_0^1 \log f(F^{-1}(p)) dp = \int_0^1 \log \frac{d}{dp} F^{-1}(p) dp.$$

On obtient l'estimateur H_{nm} avec $\hat{F}^{-1}(p) = \inf\{x \colon \hat{F}_n(x) \geq p\}$ où $\hat{F}_n$ est l'estimateur empirique, et en remplaçant la dérivée par une différence première. ; c'est-à-dire $\frac{d}{dp}\hat{F}^{-1}(p) = \frac{n}{2m}(X_{(i+m)} - X_{(i-m)})$ pour $(i-1)/n < p \leq i/n$.

Song (2000) fait remarquer que l'estimation directe de H en fonction de F n'est pas possible en raison de l'opérateur différentiel. H étant fondamentalement une fonctionnelle d'une densité, la technique d'estimation est essentiellement celle d'une densité, d'où la présence du paramètre m jouant le rôle du paramètre de lissage.

Une alternative à l'information mutuelle est donnée par la distance en norme L_2 (au carré) entre la densité et le produit de ses marges.

$$C(f_A) = \int (f_A - f_A^\star)^2,$$

où $f_A^\star$ le produit des marges de f_A.

On peut noter que l'égalité presque sûre de f_A et de $f_A^\star$, conséquence de $C(f_A) = 0$, entraîne bien l'indépendance des composantes puisque les fonctions de répartitions F_A et $F_A^\star$ seront,

elles, égales en tous points. Le contraste $C(f_A)$ est donc un critère d'indépendance mutuelle exact. On remarque aussi la similitude de forme entre les deux quantités $\int f \log f$ et $\int f^2$, et le caractère plus simple de la seconde forme.

Cette mesure de factorisation a été initialement considérée par Rosenblatt (1975) dans le cadre d'un test d'indépendance des composantes d'une fonction en dimension 2 et avec une méthode d'estimation à noyau. On trouve également dans l'article de Rosenblatt une procédure de test permettant de décider de l'indépendance, selon un critère qui tend vers une loi normale. L'ensemble est d'ailleurs directement utilisable dans le cadre ACI, sous une forme éventuellement à moderniser, si on s'en tient à une association des composantes par paire.

La mesure $C(f_A)$ est apparentée au problème classique de l'estimation de $\int f^2$, pour lequel la vitesse minimax est connue pour plusieurs types de régularité. Pour une régularité de type Hölder et pour la perte L_2, Bickel et Ritov (1988) ont montré que la vitesse minimax du problème $\int f^2$ en dimension 1 est de type paramétrique, soit n^{-1}, pour une régularité $s \geq 1/4$; la vitesse tombe à $n^{-8s/1+4s}$ pour $s \leq 1/4$,.

En dimension d, si $s > d/4$, et pour une boule de Besov $\Theta = B_{s2q}(M)$ la constante est identifiée et la vitesse minimax s'exprime par

$$\inf_{\hat{q}} \sup_{f \in \Theta} E_f \left(\hat{q} - \int f^2 \right)^2 = 4A(\Theta) n^{-1}(1 + o(1))$$

où $A(\Theta) = \sup_{f \in \Theta} \int f^2$ est une constante et $4A(\Theta)n^{-1}$ est l'inverse de l'information de Fisher non paramétrique (voir Cai et Low, 2005).

On trouve en introduction d'un article de Kerkyacharian & Picard (1996) un exemple d'estimateur par projection sur une base d'ondelettes atteignant la vitesse minimax dans le cas d'une régularité de type Besov.

La mesure $C(f_A)$ est aussi apparentée au cas $\int (f - g)^2$, pour f et g deux fonctions sans relations entre elles. Ce cas est développé dans un article de Butucea et Tribouley (2006) à propos d'un test d'homogénéité non paramétrique basé sur la perte en norme L_2.

La mesure de factorisation $\int (f - f^\star)^2$ qui nous occupe est d'une forme un peu différente ; la fonctionnelle s'écrit $q(f) = \int \left(f - \int_{\star 1} f ... \int_{\star d} f\right)^2$, où $\int_{\star \ell} f$ est la marge numéro ℓ.

L'autre intérêt a posteriori de ce projet est de pouvoir proposer différents estimateurs du contraste L_2, qui soient numériquement stables, potentiellement optimaux, ou bien sous-optimaux mais de complexité linéaire en n, avec le minimum de paramètres de réglage, et en se plaçant dans un cadre statistique unique, clairement établi.

La procédure peut également distinguer l'indépendance mutuelle de l'indépendance par paire. Cette distinction est en théorie d'un intérêt nul en ACI où par hypothèse au

maximum une seule gaussienne compose la source s ; dans ce cas en effet, d'après un résultat de Comon (1994), l'indépendance par paire est équivalente à l'indépendance mutuelle. Mais elle reprend tout son sens si on réintègre la présence du bruit ou si on admet plusieurs composantes gaussiennes, mis à part que dans ce dernier cas on retombe éventuellement dans des problèmes d'identification de A ; la distinction pourrait néanmoins fournir un angle d'attaque dans le cas de plusieurs composantes gaussiennes.

1.3 Résultats théoriques obtenus

L'étude porte sur le risque de différents estimateurs de la mesure de factorisation L_2 sous une régularité de type Besov, et sur quelques questions liées à l'utilisation du contraste en ondelettes.

Mixage linéaire et appartenance Besov

Dans le cadre de la minimisation du contraste, on s'est assuré que pour une matrice A inversible, toute transformée f_A appartient au même espace Besov que la densité f originale, autrement dit le calcul du risque à une portée générale, valable pour l'ensemble de la procédure.

(Voir propositions 3.2 et 3.3).

Contraste en projection

Soit le contraste en projection

$$C_j(f_A) = \int \left(P_j f_A - P_j f_A^\star\right)^2$$

où P_j est l'opérateur de projection sur l'espace V_j d'une analyse multirésolution de $L^2(\mathbb{R}^d)$.

- Une explicitation de l'écart entre le contraste en ondelettes et le carré de la norme L_2 de $f_A - f_A^\star$ qui représente le contraste idéal est donné par

$$0 \leq \|f_A - f_A^\star\|_2^2 - C_j(f_A - f_A^\star) \leq C2^{-2js} \ ;$$

- Le contraste en projection fournit donc un critère de factorisation approché, mais précisément relié à la régularité s

$$\begin{aligned} f \text{ est factorisable} &\implies C_j(f - f^\star) = 0 \\ C_j(f - f^\star) = 0 &\implies P_j f = P_j f^{\star 1} \dots P_j f^{\star d} \quad p.s. \end{aligned}$$

(voir aussi les propositions 3.1 et 4.6).

- Le contraste plug-in s'exprime également sur les coefficients de détail (voir lemme 5.15),

$$C_{j_1}(f) = C_{j_0}(f) + \sum_{j=j_0}^{j_1} \sum_k (\beta_{jk} - \beta_{jk^1} \dots \beta_{jk^d})^2.$$

où $k = (k^1, \dots, k^d)$ et $\beta_{jk^\ell} = \int f^{\star\ell}(x^\ell)\psi_{jk^\ell}(x^\ell)\,dx^\ell$ et $\beta_{jk} = \int f(x)\Psi_{jk}(x)\,dx$.

Estimateurs et vitesses

Le tableau ci-dessous résume les vitesses de convergence obtenues à la résolution optimale, pour les différents estimateurs détaillés ensuite. La première ligne concerne des estimateurs utilisant des blocs distincts issus d'un découpage de l'échantillon de départ ; l'estimateur figurant à la quatrième ligne est adaptatif.

Vitesses de convergence			
statistic	$2^{jd} < n$	$2^{jd} \geq n$	choix de la résolution j
$\hat{\Delta}_j^2,\ \hat{G}_j^2,\ \hat{F}_j^2$	paramétrique	$n^{\frac{-8s}{4s+d}}$	$2^j = n^{\frac{2}{d+4s}}$, $s \leq d/4$; $2^{jd} \approx n, s \geq d/4$
$\hat{D}_j^2(\tilde{X})$	$n^{-1+\frac{1}{1+4s}}$	$n^{\frac{-8s}{4s+d}}$	idem si $s \leq d/4$; $2^j = n^{\frac{1}{1+4s}}$, $s \geq d/4$
$\hat{C}_j(\tilde{X})$	$n^{\frac{-4s}{4s+d}}$	impraticable	$2^j = n^{\frac{1}{4s+d}}$
$\tilde{\hat{C}}_{j_0,\,j_1}$	$n^{\frac{-2s}{2s+d}} \log n$	impraticable	$2^{j_1 d} = Cn(\log n)^{-1}$, $j_0 = 0$

- Un estimateur de la mesure de factorisation $C(f_A) = \|f_A - f_A^\star\|_2^2$ et plus directement du contraste en ondelette $C_j(f_A - f_A^\star)$ est donné par la statistique

$$\hat{C}_j(X_1, \dots, X_n) = \sum_k (\hat{\alpha}_{jk} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d})^2$$

où $\hat{\alpha}_{jk}$ (resp. $\hat{\alpha}_{jk^\ell}$) est l'estimateur naturel du coefficient α_{jk} de la projection de f_A (resp. marginale numéro ℓ de f_A) sur le sous-espace V_j d'une analyse multirésolution de $L_2(\mathbb{R}^d)$; c'est-à-dire

$$\hat{\alpha}_{jk} = \frac{1}{n}\sum_{i=1}^n \varphi_{jk^1}(X_i^1) \dots \varphi_{jk^d}(X_i^d) \quad \text{et} \quad \hat{\alpha}_{jk^\ell} = \frac{1}{n}\sum_{i=1}^n \varphi_{jk^\ell}(X_i^\ell)$$

où X^ℓ est la coordonnée ℓ de $X \in \mathbb{R}^d$.

On montre que l'erreur en moyenne quadratique de la statistique $\hat{C}_j$ pour une classe Besov B_{spq} est au plus de l'ordre de $n^{\frac{-4s}{4s+d}}$ (voir proposition 4.10), soit mieux que l'estimation de densité $n^{\frac{-2s}{2s+d}}$ mais moins bien que la vitesse optimale de l'estimation d'une fonctionnelle quadratique, n^{-1} pour $s \geq \frac{d}{4}$ et $n^{\frac{-8s}{4s+d}}$ pour $s \leq \frac{d}{4}$, à laquelle se rattache $\int (f_A - f_A^\star)^2$.

- Un estimateur U-statistique du contraste en ondelettes s'écrit

$$\hat{D}_j^2 = \hat{D}_j^2(X_1, \dots, X_n) = \frac{1}{A_n^{2d+2}} \sum_{i \in I_n^{2d+2}} h(X_{i^1}, \dots, X_{i^{2d+2}})$$

avec

$$h = \sum_{k \in \mathbb{Z}^d} \left[\Phi_{jk}(X_{i^1}) - \varphi_{jk^1}(X_{i^2}^1) \dots \varphi_{jk^d}(X_{i^{d+1}}^d)\right] \left[\Phi_{jk}(X_{i^{d+2}}) - \varphi_{jk^1}(X_{i^{d+3}}^1) \dots \varphi_{jk^d}(X_{i^{2d+2}}^d)\right]$$

où φ est une fonction d'échelle, $\Phi_{jk}(x) = \varphi_{jk^1}(x^1) \dots \varphi_{jk^d}(x^d)$, $I_n^m = \left\{(i^1, \dots, i^m) : i^\ell \in \mathbb{N},\, 1 \leq i^\ell \leq n,\, i^{\ell_1} \neq i^{\ell_2} \text{ if } \ell_1 \neq \ell_2\right\}$ et $A_n^m = \frac{m!}{(n-m)!} = |I_n^m|$.

On peut simplifier l'expression, du noyau h en définissant une fonction Λ_{jk} de la façon suivante

$$\Lambda_{jk}(X_{i_1}, \dots, X_{i_d}) = \varphi_{jk^1}(X_{i_1}^1) \dots \varphi_{jk^d}(X_{i_d}^d) \quad \forall i \in I_n^d$$
$$\Lambda_{jk}(X_i) = \Phi_{jk}(X_i) = \varphi_{jk^1}(X_i^1) \dots \varphi_{jk^d}(X_i^d) \quad \forall i \in \{1 \dots, n\}$$

où la seconde ligne est prise comme convention.

On découpe ensuite les paquets de $2d+2$ variables X_i d'indices distincts en 4 sections : Pour $i \in I_n^{2d+2}$, on définit les 4 variables factices $Y_i = X_{i_1}$, $V_i = (X_{i_2}, \dots X_{i_{d+1}})$, $Z_i = X_{i_{d+2}}$, $T_i = (X_{i_{d+3}}, \dots X_{i_{2d+2}})$; c'est-à-dire que Y_i et Z_i admettent la distribution P_{f_A}, V_i et T_i admettent la distribution $P_{f_A}^d$, et Y_i, V_i, Z_i, T_i sont mutuellement indépendantes sous $P_{f_A}^n$.

On peut alors exprimer $\hat{D}_j^2$ sous une forme qui rappelle l'estimateur U-statistique d'ordre 2 de $\int (f-g)^2$ pour f et g deux fonctions quelconques définies sur $\mathbb{R}^d$ (Voir Butucea, Tribouley 2006) :

$$\hat{D}_j^2 = \frac{1}{A_n^{2d+2}} \sum_{I_n^{2d+2}} \sum_k \left[\Lambda_{jk}(Y_i) - \Lambda_{jk}(V_i)\right] \left[\Lambda_{jk}(Z_i) - \Lambda_{jk}(T_i)\right].$$

Ainsi la complexité supplémentaire de $\hat{D}_j^2$ est entièrement encapsulée dans les sections larges V_i et T_i, qui peuvent posséder chacune jusqu'à d coordonnées communes avec la deuxième copie du noyau dans le cadre du calcul de l'EMQ $E_{f_A}^n \left[\hat{D}_j^2\right]^2$.

D'autre part $\hat{D}_j^2$ admet la décomposition $\hat{D}_j^2 = \hat{U}_{j\alpha\alpha} - 2\hat{U}_{j\alpha\mu} + \hat{U}_{j\mu\mu}$, avec

$$\hat{U}_{j\alpha\alpha} = \frac{1}{A_n^2} \sum_{I_n^2} \sum_k \Phi_{jk}(X_{i_1}) \Phi_{jk}(X_{i_2})$$
$$\hat{U}_{j\alpha\mu} = \frac{1}{A_n^{d+1}} \sum_{I_n^{d+1}} \sum_k \Phi_{jk}(X_{i_0}) \Lambda_{jk}(X_{i_1}, \dots, X_{i_d})$$
$$\hat{U}_{j\mu\mu} = \frac{1}{A_n^{2d}} \sum_{I_n^{2d}} \sum_k \Lambda_{jk}(X_{i_1}, \dots, X_{i_d}) \Lambda_{jk}(X_{i_{d+1}}, \dots, X_{i_{2d}}),$$

chacun respectivement étant aussi un estimateur sans biais de $C_{j\alpha\alpha}$, $C_{j\alpha\mu}$ et $C_{j\mu\mu}$, avec

$$C_j = \sum_k \alpha_{jk}^2 - 2 \sum_k \alpha_{jk} \lambda_{jk} + \sum_k \lambda_{jk}^2 \equiv C_{j\alpha\alpha} - 2C_{j\alpha\mu} + C_{j\mu\mu},$$

ce qui démontre que $\hat{C}_j^2$ est la V-statistique associée à $\hat{D}_j^2$.

On peut noter que seul $\hat{U}_{j\alpha\alpha}$ possède un noyau symétrique ; pour cette raison, la plupart des calculs réalisés dans ce projet ne s'appuient pas sur une symétrie du noyau.

Le calcul du risque peut donc se faire sur $\hat{D}_j^2$ ou bien sur chacune des composantes $\hat{U}_{j..}$, mais dans le second cas on perd une propriété du risque, certes relativement accessoire, qui veut que la borne obtenue se compose d'une constante nulle à l'indépendance ; la borne s'exprime en effet par

$$C^\star 2^j n^{-1} + C 2^{jd} n^{-2},$$

où $C^\star = 0$ à l'indépendance (quand $A = I$) (proposition 4.11). On n'obtient pas une telle constante $C^\star$ en majorant le risque sur chacune des composantes prises séparément.

La présence du facteur 2^j dans l'expression ci-dessus matérialise la sous optimalité de $\hat{D}_j^2$ par rapport à un estimateur U-statistique de $\int f^2$.

- On a également utilisé les U-statistiques à noyaux symétriques

$$\hat{B}_j^2(\{X_1, \dots, X_n\}) = \sum_k \frac{1}{A_n^2} \sum_{i \in I_n^2} \Phi_{jk}(X_{i^1}) \Phi_{jk}(X_{i^2})$$

$$\hat{B}_j^2(\{X_1^\ell, \dots, X_n^\ell\}) = \sum_{k^\ell} \frac{1}{A_n^2} \sum_{i \in I_n^2} \varphi_{jk^\ell}(X_{i^1}^\ell) \varphi_{jk^\ell}(X_{i^2}^\ell)$$

qui nous replacent dans le cadre de l'estimation de $\int f^2$ traité en introduction de l'article de Kerkyacharian et Picard (1996), mis à part que dans notre cas l'estimation a lieu en dimension d.

En estimant chaque plug-in $\hat{\alpha}_{jk}(\tilde{R}^0)$ et $\hat{\alpha}_{jk^\ell}(\tilde{R}^\ell)$, et la U-statistique $\hat{B}_j^2(\tilde{R}^0)$, $\hat{B}_j^2(\tilde{R}^\ell)$, $\ell = 1, \dots, d$ sur des tranches indépendantes $\tilde{R}^0, \tilde{R}^1, \dots, \tilde{R}^d$ de l'échantillon de départ, on peut définir l'estimateur mixte plug-in

$$\hat{F}_j^2(\tilde{R}^0, \tilde{R}^1, \dots, \tilde{R}^d) = \hat{B}_j^2(\tilde{R}^0) + \prod_{\ell=1}^d \hat{B}_j^2(\tilde{R}^\ell) - 2 \sum_{k \in \mathbb{Z}^d} \hat{\alpha}_{jk}(\tilde{R}^0) \hat{\alpha}_{jk^1}(\tilde{R}^1) \dots \hat{\alpha}_{jk^d}(\tilde{R}^d),$$

en vue d'estimer la quantité $\sum_k \alpha_{jk}^2 + \prod_{\ell=1}^d \Big(\sum_{k^\ell \in \mathbb{Z}} \alpha_{jk^\ell}^2 \Big) - 2 \sum_k \alpha_{jk} \alpha_{jk^1} \dots \alpha_{jk^d} = C_j^2$ (voir proposition 4.8).

- Deuxième façon de faire : on peut tirer de l'échantillon de départ $\{X_1, \dots, X_n\}$ un échantillon i.i.d. de $f_A^\star$, par exemple $\{X_1^1 \dots X_d^d, X_{d+1}^1 \dots X_{2d}^d, \dots, X_{([n/d]-1)d+1}^1 \dots X_{[n/d]d}^d\}$, à la manière de la décomposition de Hoeffding. Évidemment ce procédé possède l'inconvénient de laisser de côté une grande partie de l'information disponible.

Quoi qu'il en soit on peut supposer qu'on dispose de deux échantillons indépendants, un pour f_A étiqueté $\tilde{R}$ et un pour $f_A^\star$ étiqueté $\tilde{S}$, avec $\tilde{R}$ indépendant de $\tilde{S}$. On définit alors les

estimateurs

$$\hat{G}_j^2(\tilde{R},\tilde{S}) = \hat{B}_j^2(\tilde{R}) + \hat{B}_j^2(\tilde{S}) - 2\sum_{k\in\mathbb{Z}^d} \hat{\alpha}_{jk}(\tilde{R})\hat{\alpha}_{jk}(\tilde{S})$$

et la U-statistique à deux échantillons utilisée dans Butucea et Tribouley (2006)

$$\hat{\Delta}_j^2(\tilde{R},\tilde{S}) = \frac{1}{A_n^2}\sum_{i\in I_n^2}\sum_{k\in\mathbb{Z}^d}\left[\Phi_{jk}(R_{i^1}) - \Phi_{jk}(S_{i^1})\right]\left[\Phi_{jk}(R_{i^2}) - \Phi_{jk}(S_{i^2})\right]$$

en supposant pour simplifier que les deux échantillons ont la même taille n (qui n'est plus le n de départ).

$\hat{\Delta}_j^2(R,S)$ ne recèle plus de sections larges, n'est plus que d'ordre 2 et représente l'exacte réplique de la statistique utilisée dans Butucea and Tribouley (2006) (en dimension d au lieu de 1) pour l'estimation de $\int(f-g)^2$ dans le cas où f et g sont deux fonctions sans relation entre elles.

Cette procédure est optimale et son risque est $C^\star n^{-1} + 2^{jd}n^{-2}$, avec $C^\star = 0$ à l'indépendance (quand $A = I$) (voir proposition 4.9).

- L'estimateur seuillé qui a été étudié s'écrit $\tilde{\hat{C}}_{j_0,j_1} = \hat{C}_{j_0} + \tilde{\hat{T}}_{j_0 j_1}$ avec

$$\tilde{\hat{T}}_{j_0 j_1} = \sum_{j=j_0}^{j_1}\sum_k (\tilde{\hat{\beta}}_{jk} - \tilde{\hat{\beta}}_{jk^1}\dots\tilde{\hat{\beta}}_{jk^d})^2$$

où $\tilde{\hat{\beta}}_{jk}$ est le substitut seuillé de $\hat{\beta}_{jk}$, et $j_0 = 0$ ou quelques unités.

On a étudié le cas d'un seuillage fort

$$\tilde{\hat{\beta}}_{jk} = \hat{\beta}_{jk} I\{\,|\hat{\beta}_{jk}| > t/2\} \text{ et } \tilde{\hat{\beta}}_{jk^\ell} = \hat{\beta}_{jk^\ell} I\{\,|\hat{\beta}_{jk^\ell}| > (t/2)^{\frac{1}{d}}\}, \text{ avec } t \approx \sqrt{\frac{\log n}{n}}.$$

On démarre la procédure à la résolution de l'estimation de densité, soit $2^{j_1 d} = Cn(\log n)^{-1}$.

Le contraste seuillé terme à terme de $\hat{C}_j$ admet une vitesse de convergence de l'ordre de $(\log n)n^{\frac{-2s}{2s+d}}$, soit légèrement moins que la vitesse minimax du problème de l'estimation de densité dans le cas adaptatif qui est $[(\log n)n^{-1}]^{\frac{2s}{2s+d}}$.

Autrement dit le seuillage terme à terme fait perdre plus qu'un log comparé au contraste linéaire dont la vitesse est en $n^{\frac{-4s}{4s+d}}$ (voir proposition 5.18).

Cette perte d'efficacité est due à une forme de seuillage inappropriée : il s'agit d'un seuillage terme à terme de $\hat{C}_j$, tandis que pour le problème $\int f^2$, on sait que le seuillage par bloc fonctionne mieux (voir par ex. Cai et Low, 2005 p.4). Mais dans le cas du contraste plug-in c'était impraticable car le seuillage par bloc démarre typiquement à $2^{j_1 d} \approx n^2$ et s'arrête

à $2^{j_0 d} \approx n$, deux résolutions au-delà du maximum de fonctionnement de $\hat{C}_j$: pour $2^{jd} > n$, l'algorithme est instable, puisque on n'est jamais assuré de la convergence de la borne en $2^{jd} n^{-1}$ lorsque n augmente.

On n'a pas implémenté le seuillage par bloc de l'estimateur U-statistique de $\int (f_A - f_A^\star)^2$, en raison de la complexité élevée du calcul associé, et on n'a pas non plus cherché à démontrer le résultat théorique, puisqu'il devrait découler du cas $\int f^2$ classique, mais c'est théoriquement la configuration tirant le meilleur parti du seuillage.

Pour une boule Besov $\Theta = B_{s2q}(M)$, on sait en effet que le risque minimax de l'estimateur adaptatif de $\int f^2$ est inchangé pour $s > d/4$, c'est-à-dire qu'il reste paramétrique, et est de l'ordre de $n^{\frac{-8s}{d+4s}} (\log n)^{\frac{4s}{d+4s}}$ quand $s \leq d/4$, c'est-à-dire est augmenté d'un terme log (Cai et Low 2005). Voir aussi Gayraud et Tribouley (1999) pour le modèle bruit blanc.

En tant qu'estimateur du contraste en projection lui-même seuillé $\tilde{C}_{j_0,\, j_1}$, défini par $\tilde{C}_{j_0,\, j_1} = C_{j_0} + \tilde{T}_{j_0 j_1}$ avec

$$\tilde{T}_{j_0 j_1} = \sum_{j=j_0}^{j_1} \sum_k (\tilde{\beta}_{jk} - \tilde{\beta}_{jk^1} \dots \tilde{\beta}_{jk^d})^2$$

où $\tilde{\beta}_{jk} = \beta_{jk} I\{\, |\beta_{jk}| > t/2\}$, la statistique $\tilde{\hat{C}}_{j_0,\, j_1}$ admet un risque dont la vitesse ne contient plus de terme log. C'est intéressant en soi puisque l'estimation du contraste est un but intermédiaire, et un contraste seuillé qui omet les petites contributions peut suffire à donner un inverse de A satisfaisant.

Propriétés de filtrage du gradient et du Hessien

Dans le cadre de l'ACI, une fois estimé en un point f_{WA} donné, où W matrice $d \times d$ représente l'estimation courante de A^{-1}, le contraste doit encore être minimisé en W. A ce stade, on rejoint la procédure suivie par n'importe quelle autre fonction de contraste utilisée dans d'autres méthodes de résolution de l'ACI, et notamment la procédure de Bach et Jordan.

Dans le cas d'une ondelette deux fois continûment différentiable, soit par exemple une Daubechies au moins $D4$, il est possible de donner une formulation explicite du gradient et du Hessien de l'estimateur $\hat{C}_j$ en tant que fonction de W ou de $Y = WX$; de plus la formulation permet de passer facilement d'une résolution à une autre par transformation en ondelette discrète (DWT). Voir la partie 6.

1.4 Résultats pratiques

Les résultats pratiques présentés dans cette partie concernent essentiellement l'estimateur plug-in.

Comme pour l'ACI à noyau de Bach et Jordan, l'estimation concrète de la matrice B repose sur une procédure de minimisation du contraste qui peut plus ou moins bien tourner mais se résout en général en quelques itérations.

Utilisation du contraste plug-in

L'estimateur plug-in $\hat{C}_j$ fonctionne seulement à la résolution j telle que $2^{jd} < n$; en effet, le risque étant en $2^{jd}n^{-1}$, on n'a aucune stabilité numérique par augmentation de n sur l'ensemble $\{j, d\colon 2^{jd} >= n\}$ (voir propositions 3.4 et 4.10).

Théoriquement le choix optimal de j dépend de la régularité s et s'établit à j_* tel que $2^{j_*} = n^{\frac{1}{d+4s}}$ (voir proposition 4.7 et voir la proposition 3.5 basée sur une borne du risque sous optimale). Mais en pratique il existe une bande de résolution j fonctionnant tout aussi bien que le j optimal. Si la régularité s n'est pas connue, on peut démarrer par la plus petite résolution technique (tenant compte de la longueur du filtre de l'ondelette Daubechies utilisée), et augmenter le j progressivement si la minimisation ne semble pas s'opérer (voir les simulations dans la partie 3).

L'alternative consiste à utiliser le contraste seuillé, qui s'adapte automatiquement à la bonne résolution ; dans les simulations effectuées on n'a pas observé une plus-value très nette du seuillage comparé à un choix moyen de résolution compris entre 0 et $\log n(d \log 2)^{-1}$ (voir les simulations de la partie 5 p. 164). Il faut préciser que le seuillage optimum dans un problème de type $\int f^2$ est de type global. Dans le seuillage global, on démarre à une résolution 2^{jd} au delà de n, ce qui n'est pas opérant pour le contraste plug-in.

Le contraste n'est pas en défaut en présence de gaussiennes ou si le mixage n'est pas linéaire. Dans tous les cas on a une estimation de la norme L_2 de $f_A - f_A^\star$. Évidemment si le mixage n'est pas linéaire, le risque de la procédure n'est pas celui annoncé puisqu'on n'est plus sûr de l'appartenance Besov de f_A, et la minimisation sur la sous-variété de Stiefel devient probablement inopérante.

Dans ce projet on a utilisé une pré-transformation par ACP et une minimisation sur la sous-variété de Stiefel, pour se rapprocher du modèle utilisé par Bach et Jordan, mais cela n'est pas une nécessité. D'autres méthodes de minimisation sont envisageables et l'absence de pré-transformation ACP n'a pas de conséquence sur la stabilité numérique du contraste.

Dans la pratique, il n'y a qu'un seul paramètre à calibrer, la résolution j.

Stabilité numérique

Basé sur des convolutions et des sommes, le contraste en ondelettes est numériquement stable et trivial du point de vue algorithmique, même si une implémentation efficace indépendante de d est relativement difficile à obtenir. La projection sur l'espace V_j est

l'équivalent d'un histogramme de fenêtre 2^{-j} pour une $D2$ (Haar) ou bien approximativement $2^{-(j \vee L)}$ avec pondération pour une $D2N$ (L est le paramètre d'approximation aux dyadiques), avec une certaine frange de recouvrement due au chevauchement du support des ondelettes ; le calcul du contraste proprement dit est une somme de carrés de différences.

Cette simplicité d'opération est à mettre en perspective avec les approches basées sur des calculs d'algèbre linéaire (valeurs propres généralisées, décomposition de Choleski incomplète) dont la mise en œuvre requiert des paramètres supplémentaires, parfois fixés selon des règles empiriques, et où la sensibilité numérique est nettement plus problématique dans les cas limites (d ou n grands) ou bien si les conditions d'utilisations ne sont pas réunies (mixage non exactement linéaire, présence de gaussiennes, régularisation mal calibrée...).

Dans le cas du contraste en ondelettes, la concrétisation du calcul n'est qu'une question de ressources système. Pour jd grand, on peut lotir l'espace mémoire nécessaire à la projection sur V_j, ainsi que la traversée de la boucle du contraste. Le calcul se prête très bien à une programmation parallèle ou répartie.

Complexité numérique

La statistique $\hat{C}_j$ est un estimateur plug-in ; son évaluation s'appuie en premier lieu sur l'estimation complète de la densité de f_A et de ses marges ; ce qui demande un temps de calcul de l'ordre de $O(n(2N-1)^d)$ où N est l'ordre de l'ondelette de Daubechies, et n le nombre d'observations ; on précise donc ici ce qu'on entend par $O(n)$.

Un n de l'ordre de 10000 ou 100000 ne sature pas la procédure (voir p. 101 et suivantes), on a ainsi le cas échéant la possibilité de démixer des ensembles de données importants, pour des applications de type datamining par exemple (pour diverses applications avancées de l'ACI voir par exemple Bingham, 2003). En comparaison, les méthodes à noyau plafonnent en général à $n = 5000$, tout au moins dans les simulations présentées.

Dans un second temps, le contraste proprement dit est une simple fonction des $2^{jd} + d2^j$ coefficients qui estiment la densité f_A et ses marges ; le temps de calcul additionnel est donc en $O(2^{jd})$.

On voit ici le principal défaut numérique du contraste en ondelettes dans sa formulation totale, celui d'être de complexité exponentielle en la dimension d du problème, mais c'est par définition le coût d'une condition garantissant l'indépendance mutuelle des composantes en toute généralité : d ensembles $B_1, \ldots, B_d$ sont mutuellement indépendants si $P(B_1 \cap \ldots \cap B_d) = PB_1 \ldots PB_d$ pour chacun des 2^d choix d'indices dans $\{1, \ldots, d\}$.

La complexité en jd tombe à $O(d^2 2^{2j})$ si on se concentre sur une indépendance 2 à 2 des composantes, comme dans kernel ICA et la méthode basée sur la norme de Hilbert-Schmidt (voir pp. 71, 73) ou dans la méthode Radical (voir p. 75). L'indépendance par paire est en fait équivalente à l'indépendance mutuelle en l'absence de bruit et dans le cas où au maximum une seule composante est gaussienne (Comon, 1994).

Usage d'estimateurs U-statistique

Les estimateurs U-statistiques de C_j sont de complexité au minimum $O(n^2(2N-1)^{2d})$, c'est-à-dire quadratique en n ; en revanche la complexité en jd est probablement abaissée puisque le contraste se calcule par accumulation, sans qu'il soit nécessaire de conserver toute la projection en mémoire, mais seulement une fenêtre dont la largeur dépend de la longueur du filtre de la Daubechies.

On n'a pas implémenté l'estimateur U-statistique, mais dans une configuration où on minimise les dépendances par paire, puisque on estime au maximum des densités en dimension deux, son utilisation pourrait s'avérer compétitive, sans compter que la U-statistique se prête à un seuillage par bloc.

Haar et évaluation directe aux dyadiques

L'ondelette de Haar ($D2N, N = 1$) ne convient pas pour le calcul d'un gradient, empirique ou théorique ; la variation de l'histogramme (projection sur l'espace V_j engendré par l'ondelette de Haar), en réponse à une faible perturbation de la matrice de démixage W est en général inexistante ou imperceptible (voir notamment les simulations p. 101 et suivantes).

Le problème se pose également au moment de l'initialisation par histogramme à une résolution élevée, souvent utilisée en estimation de densité, avec une série de filtrages vers les résolutions plus basses à suivre.

Pour obtenir un gradient empirique, on utilise donc une ondelette au moins $D4$ et le calcul direct $\frac{1}{n}\sum_i \phi_{jk}(X_i)$ avec approximation des valeurs X_i aux dyadiques d'après l'algorithme expliqué dans Nguyen et Strang (1996). Heureusement, et contrairement peut-être à une idée répandue, ce calcul n'est pas un problème, à partir du moment où les valeurs prises par l'ondelette à une précision donnée sont préchargées en mémoire. Le temps de calcul correspondant n'est en aucun cas un goulet d'étranglement pour l'ACI, d'autant moins qu'on économise des séries de filtrage en dimension d (voir partie 6).

Pouvoir de séparation du contraste

On donne ci-dessous un extrait des simulations figurant dans la partie 3 ; il s'agit du résultat moyen de 100 runs en dimension 2 avec 10000 observations, une Daubechies D4, $j = 3$ et $L = 10$ (précision dyadique) pour différentes densités ; la colonne `start` indique la distance Amari (sur une échelle de 0 à 100) et le contraste en ondelette en entrée ; la colonne `it` est le nombre moyen d'itération de la minimisation Stiefel. Pour quelques densités, après la transformation ACP, on se trouve déjà près du minimum, mais le contraste détecte quand même une dépendance ; la procédure n'est pas exécutée si le contraste ou le gradient empirique sont trop proches de zéro, et cela correspond pratiquement toujours à une erreur Amari inférieure à 1.

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

On peut noter que dans le cas de composantes gaussiennes, la minimisation ne s'opère pas.

Les deux exemples ci-dessous permettent de visualiser la variation du contraste et de l'erreur Amari, à une résolution qui n'est pas forcément la meilleure. On peut noter que le contraste basé sur Haar (D2) donne une courbe visuellement correcte pour la recherche d'une minimum.

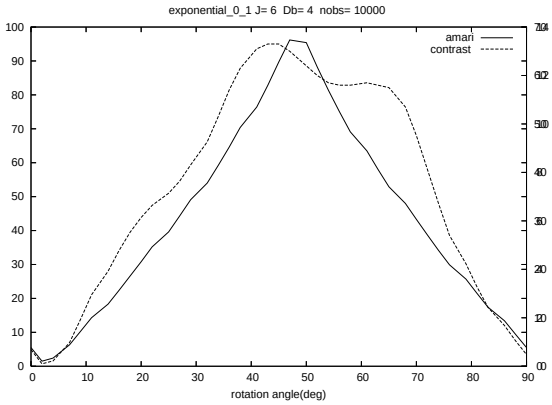

Fig.1. Exponential, D4, j=6, n=10000

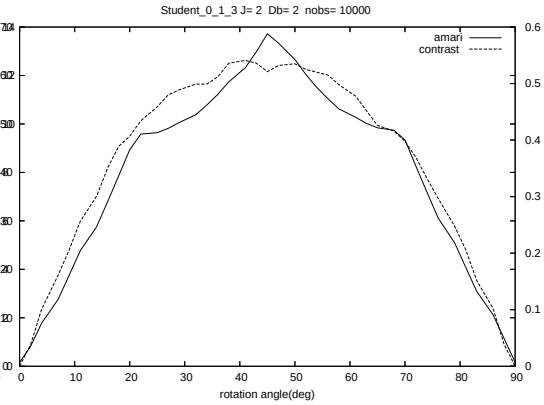

Fig.2. Student, D2, j=2, n=10000

Voir d'autres exemples dans la partie 3.

1.5 Éléments de comparaison avec d'autres méthodes

La méthode que nous proposons fournit un cadre unique dont les propriétés statistiques sont clairement établies, et laisse une assez grande liberté de manœuvre dans la pratique. On a le choix entre des estimateurs plug-in ou U-statistique ; on peut décider de se restreindre à une indépendance des composantes par paires ou par triplets, ou à l'indépendance mutuelle d'une partie seulement des composantes pour couper court à la procédure complète de complexité numérique plus élevée, lorsque son utilisation n'est pas justifiée (c'est-à-dire en utilisation normale : une seule composante gaussienne au maximum et bruit négligé).

Contrairement au critère de l'information mutuelle utilisé sur les marges uniquement, on connaît la valeur du minimum global, qui est zéro, et on a donc au moins en théorie une possibilité de savoir s'il a été atteint.

Dans la suite de cette partie on passe en revue plus précisément les méthodes non paramétriques concurrentes déjà présentées plus haut.

Kernel ICA

Dans l'ACI à noyau de Bach et Jordan (2002), on utilise une forme de décorrélation non linéaire portée par un espace de Hilbert à noyau reproduisant (RKHS) $\mathcal{F}_\sigma$.

L'espace $\mathcal{F}_\sigma$ est issu du noyau gaussien isotrope $k(x,y) = e^{-\frac{1}{2\sigma^2}\|x-y\|^2}$, avec $\emptyset = \mathcal{F}_0$ et $\mathcal{F}_\sigma$ croît vers $L_2(\mathbb{R}^d)$, quand $\sigma^2 \to +\infty$.

Le paramètre σ est donc l'équivalent du paramètre de résolution j dans l'ACI par ondelettes.

Soit W, inverse potentiel de la matrice A solution de $x = As$, soit $y_i = Wz_i$ où z_i est la transformation ACP de x_i, l'observation i.

On utilise l'estimateur $\widehat{\mathrm{cov}}_{\mathcal{F}_\sigma}(Y^1, Y^2) = \frac{1}{n}\, {}^t\alpha^1 K_1 K_2 \alpha^2$, et

$$\hat{\rho_{\mathcal{F}}}(Y^1, Y^2) = \max_{\alpha^1,\alpha^2} \frac{{}^t\alpha^1 K_1 K_2 \alpha^2}{(\,{}^t\alpha^1 K_1^2 \alpha^1)^{\frac{1}{2}} (\,{}^t\alpha^2 K_1^2 \alpha^2)^{\frac{1}{2}}},$$

où K_1 et K_2 sont les matrices $n \times n$ de Gram données par

$$K_\ell = \begin{pmatrix} k(y_1^\ell, y_1^\ell) & \dots & k(y_1^\ell, y_n^\ell) \\ \vdots & \ddots & \vdots \\ k(y_n^\ell, y_1^\ell) & \dots & k(y_n^\ell, y_n^\ell). \end{pmatrix}$$

La solution de ce problème de corrélation canonique est donnée par la solution du problème de valeur propre généralisé

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha^1 \\ \alpha^2 \end{pmatrix} = \rho \begin{pmatrix} K_1^2 & 0 \\ 0 & K_2^2 \end{pmatrix} \begin{pmatrix} \alpha^1 \\ \alpha^2 \end{pmatrix}$$

Dans le cas général on forme une super matrice de Gram des K_ℓ, préalablement centrés, de dimension $(nd \times nd)$ notée $\mathcal{K}$: $\mathcal{K} = (\,K_1 \quad \dots \quad K_d\,)\,(\,K_1 \quad \dots \quad K_d\,)'$.

La complexité du calcul de la solution généralisée est théoriquement de l'ordre de $O(d^3 n^3)$, mais les auteurs ont recours à une décomposition de Choleski incomplète, qui revient à se restreindre à une solution de rang $m = h(\eta/n) \ll nd$, où η est un paramètre de précision et h une fonction dépendant du type de décroissance (inconnu) à l'infini des densités sous-jacentes, soit $h(t) = O(\log t)$ pour une décroissance exponentielle et $h(t) = t^{1/d+\varepsilon}$ pour une

décroissance polynomiale. La complexité est ainsi ramenée à hauteur de $O(d^2n)$, mais il n'existe que des règles empiriques sur le caractère optimal du choix de m en fonction de la précision η souhaitée.

Dans le cas du contraste en ondelettes, la complexité du critère d'indépendance par paire est en $O(d^2n) + O(d^2 2^{2j}) = O(d^2n)$ pour la statistique plug-in et en $O(d^2n^2)$ pour la U-statistique. Il n'y a pas de paramètre supplémentaires à introduire, la connaissance du type de décroissance des marginales à l'infini est inutile.

Un autre point important réside dans la nécessité d'introduire un second paramètre de régularisation $\kappa < 1$ dans la diagonale de $\mathcal{K}$, ainsi K_ℓ^2 est remplacée par $(K_\ell + \kappa I)^2$.

Soit $\mathcal{D}$ la matrice dont la bande diagonale est occupée par le carré des $(K_\ell + \kappa I)$; $\mathcal{D} = \mathrm{diag}((K_1 + \kappa I)^2, \ldots, (K_d + \kappa I)^2)$. On cherche en réalité $\hat{\lambda}$, la plus petite valeur propre de l'équation $\mathcal{K}\alpha = \lambda \mathcal{D}\alpha$ où $\alpha \in R^{dn}$; $\hat{\lambda}$ est dénommée première (kernel) corrélation canonique.

Le paramètre de régularisation a pour effet que le problème devient numériquement stable et doit être choisi de la forme $\kappa = \kappa_0 n$ pour obtenir un critère $\hat{\rho}_{\mathcal{F}}$ indépendant de n, en négligeant le terme d'ordre κ^2. Le critère réellement estimé s'écrit ainsi

$$\hat{\rho_{\mathcal{F}}}(Y^1, Y^2) = \max_{\alpha^1, \alpha^2} \frac{{}^t\alpha^1 K_1 K_2 \alpha^2}{(\mathrm{var}\, f^1(x^1) + 2\kappa n^{-1} \|f^1\|_{\mathcal{F}}^2)^{\frac{1}{2}} (\mathrm{var}\, f^2(x^2) + 2\kappa n^{-1} \|f^2\|_{\mathcal{F}}^2)^{\frac{1}{2}}},$$

puisque ${}^t\alpha(K+\kappa I)^2\alpha = {}^t\alpha K^2\alpha + 2\kappa\, {}^t\alpha K\alpha + \kappa^2\, {}^t\alpha\alpha$, et $n^{-1}\, {}^t\alpha K^2\alpha = \mathrm{var}\, f^1(x^1)$, $n^{-1}\, {}^t\alpha K\alpha = \|f^1\|_{\mathcal{F}}^2$.

Dans le cas général, $d > 2$, $\hat{\lambda}(K_1, \ldots, K_d)$ est une estimation de $\lambda(x_1, \ldots, x_d)$ la $\mathcal{F}$-corrélation entre d variables. On a $0 \leq \lambda \leq 1$, et $\lambda = 1$ si et seulement si les variables $f_1(x_1), \ldots, f_d(x_d)$ sont non corrélées, où $f_i \in \mathcal{F}_\sigma$. Il s'agit donc d'une indépendance par paire, puisque une matrice de corrélation donne une information sur les paires (voir aussi la méthode de Hilbert-Schmidt p. 73 qui généralise kernel ICA) ; dans le cadre ACI usuel, c'est équivalent à une indépendance mutuelle (Comon, 1994).

Dans le cas du contraste en ondelettes on peut distinguer l'indépendance par paire de l'indépendance mutuelle, lorsque il y a une différence, moyennant une augmentation de la complexité numérique.

Au total, kernel ICA demande un choix de quatre paramètres, σ (équivalent au j des ondelettes), m et η pour la décomposition Choleski incomplète, κ pour la stabilité numérique, auxquels s'ajoutent les paramètres de minimisation du contraste.

Le contraste en ondelettes ne demande qu'un seul paramètre (mis à part la minimisation) ; la résolution j.

Dans l'article de Bach et Jordan (2002), les choix opérés sont $\eta = 10^{-3}\kappa$, $\kappa = 10^{-3}n$, $\sigma = 1$ pour $n < 1000$ et $\sigma = 1/2$ pour $n > 1000$. On ne connaît pas de critère théorique permettant de fixer les choix optimaux.

Pour la résolution j du contraste en ondelette, on dispose d'un procédure auto-adaptative, par seuillage ; sans le seuillage, un choix moyen de j avec $2^{jd} < n$ s'avère quasiment auto-adaptatif dans les simulations effectuées avec l'estimateur plug-in. Le j optimal est relié à la régularité de la densité sous-jacente s.

Un ensemble très complet de simulations en dimension 2 avec $n = 256$ et 1024 observations montrent des résultats en moyenne meilleurs pour l'ACI à noyau comparée à Jade, Infomax, et FastICA. Il semble néanmoins que la méthode soit en difficulté avec un nombre d'observation n très élevé, en raison de la taille de la matrice $\mathcal{K}$ dépendant de n. Dans les simulations présentées en grande dimension, $d = 8, 16$, le nombre d'observations maximum s'arrête à 4000 ce qui paraît relativement peu. Pour ces cas, la méthode peut-elle faire mieux en augmentant n ?

Le contraste en ondelettes étant issu d'une méthode de projection classique, un n très élevé (100000) ne compromet pas la réalisation du calcul.

Du point de vue théorique, les propriétés statistiques de Kernel ICA restent à étudier ; par exemple on n'a pas d'ordre de grandeur de la vitesse de convergence de l'algorithme.

Méthodes de la norme de Hilbert-Schmidt

D'autres approches basées sur un espace de Hilbert à noyau reproduisant (RKHS) ont été étudiées par Gretton et al. (Gretton et al. 2003, 2004) qui présentent un critère général de dépendance statistique liée à la norme de Hilbert-Schmidt.

Soit un RKHS $\mathcal{F}$ de fonctions de $\mathcal{X}$ dans $\mathbb{R}$ associé au noyau défini positif $k(.,.)\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, et à la fonction caractéristique ϕ, $\phi(x) = k(x,.)$, c'est-à-dire en résumé

$$\forall x \in \mathcal{X},\ \exists \phi(x) \in \mathcal{F},\ <\phi(x), \phi(x')>_{\mathcal{F}} = k(x, x') \text{ et } <\phi(x), f>_{\mathcal{F}} = \delta_x f = f(x).$$

Soit un second RKHS $(\mathcal{G}, l(.,.), \psi)$ de fonctions de $\mathcal{Y}$ dans $\mathbb{R}$.

On suppose également que $\mathcal{X}$ et $\mathcal{Y}$ sont des espaces mesurés $(\mathcal{X}, \Gamma, p_x)$ et $(\mathcal{Y}, \Lambda, p_y)$, où p_x et p_y sont des mesures probabilités. Et on considère aussi l'entité jointe $(\mathcal{X} \times \mathcal{Y}, \Gamma \otimes \Lambda, p_{xy})$.

Les auteurs introduisent l'opérateur de covariance croisée

$$C_{xy} = E_{xy}\big[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)\big]$$

où E_{xy} est l'espérance par rapport à la loi jointe p_{xy}, $\otimes$ est le produit tensoriel défini par $f \otimes g\colon h \in \mathcal{G} \mapsto f <g, h>_{\mathcal{G}} \in \mathcal{F}$ et μ_x, μ_y sont donnés respectivement par $<\mu_x, f>_{\mathcal{F}} = E_x f(x)$ et $<\mu_y, g>_{\mathcal{G}} = E_y g(y)$.

La mesure de dépendance *HSIC* considérée est le carré de la norme de Hilbert-Schmidt de l'opérateur linéaire C_{xy}

$$HSIC(p_{xy}, \mathcal{F}, \mathcal{G}) = \|C_{xy}\|^2_{HS}$$

avec $\|C_{xy}\|_{HS}^2 = \sum_{i,j} <Cv_i, u_j>_{\mathcal{F}}^2$ et $\{u_j, j \in J\}$ base orthonormée de $\mathcal{F}$, $\{v_i, i \in I\}$ base orthonormée de $\mathcal{G}$.

Les auteurs montrent ensuite que $\|C_{xy}\|_{HS}^2 = 0$ si et seulement si x est indépendant de y, les présupposés étant que $\mathcal{X}$ et $\mathcal{Y}$ sont des domaines compacts et que $\mathcal{F}$ et $\mathcal{G}$ sont des ensembles de fonctions bornées. Pour le cas $d > 2$, on généralise le critère en considérant tous les C_d^2 appariements de 2 variables, ce qui revient donc à se restreindre à une indépendance par paire, moins contraignante que l'indépendance mutuelle, mais suffisante dans le cadre ACI usuel.

Ce qui a été dit dans kernel ICA sur le noyau gaussien isotrope et la décomposition Choleski incomplète reste vrai dans cette méthode. On retrouve donc les trois paramètres σ (équivalent au j des ondelettes), η et m.

Le critère *HSIC* s'exprime en fonction des noyaux par

$$HSIC(p_{xy}, \mathcal{F}, \mathcal{G}) = E_{xx'yy'}\big[k(x,x')l(y,y')\big] + E_{xx'}k(x,x')E_{yy'}l(y,y') - 2E_{xy}\Big[E_{x'}k(x,x')E_{y'}l(y,y')\Big]$$

où $E_{xx'yy'}$ est l'espérance relative à la loi de deux paires indépendantes (x,y), (x',y').

Un estimateur de HSIC est donné par

$$HSIC(Z, \mathcal{F}, \mathcal{G}) = (n-1)^{-2} \text{ trace } KHLH,$$

avec $H, K, L \in \mathbb{R}^{n^2}$, $H = I_n - n^{-1}(1_n {}^t 1_n)$, $K_{ij} = k(x_i, x_j)$, $L_{ij} = l(y_i, y_j)$ et $Z = \{(x_1, y_1), \ldots (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$ est un échantillon indépendant identiquement distribué de p_{xy}.

Les auteurs montrent que $HSIC(Z)$ est biaisé à l'ordre n^{-1} et montrent aussi en utilisant une majoration de grande déviation pour U-statistique que pour tout $\delta > 0$, toute loi p_{xy} et $n > 1$

$$P_Z\left[|HSIC(p_{xy}) - HSIC(Z)| \geq \sqrt{\frac{\log 6/\delta}{\alpha^2 n}} + \frac{C}{n}\right] < \delta$$

où $\alpha > 0,24$.

On a donc une convergence en probabilité P_Z du critère empirique vers le critère exact à une vitesse au mieux $n^{-1/2}$, et on peut donc s'attendre, avec une grande probabilité, à une amélioration de la précision par augmentation de n.

Par comparaison, la consistance du contraste en ondelettes est démontrée en moyenne quadratique. D'autre part la consistance de l'estimateur HSIC ne prend pas en compte le biais d'approximation entre le RKHS et l'espace de Hilbert englobant sur lequel est défini le critère exact d'indépendance des deux tribus $\sigma(X)$ et $\sigma(Y)$.

Pour tester l'indépendance au seuil de signification γ on définit

$$\Delta(Z) = \mathbb{I}\left\{HSIC(Z) > C\sqrt{n^{-1}\log 1/\gamma}\right\}$$

et

$$E_Z[\Delta(Z)=1] = P_Z\left[|HSIC(Z)| > C\sqrt{n^{-1}\log 1/\gamma}\right] < \gamma.$$

La méthode ne demande pas formellement de paramètre de régularisation supplémentaire κ comme pour kernel ICA. En revanche, il faut fixer le seuil de signification γ pouvant éventuellement être vu comme jouant le même rôle, ainsi que la constante C.

La résolution passe par une minimisation du critère HSIC empirique en faisant varier la matrice de démixage selon la même méthode que kernel ICA (minimisation sur la sous variété de Stiefel). Comme pour kernel ICA, le critère HSIC est approché par une décomposition de Choleski incomplète.

L'ACI par ondelettes se prête également à l'utilisation d'un test d'indépendance, mais la minimisation fonctionne même sans un tel test, tandis que dans la méthode de Hilbert-Schmidt, le calcul du critère englobe le test statistique, et ne fonctionne pas sans.

Les auteurs annoncent d'excellents résultats comparés aux méthodes classiques et à kernel ICA, sauf dans le cas où le nombre d'observations est faible ($n = 250$).

Estimateur RADICAL

Miller et Fisher III (2003) ont proposé un contraste ACI basé sur l'information mutuelle à partir d'un estimateur initialement introduit par Vasicek (1976).

Soit $Y_{(1)}, \dots Y_{(n)}$ la statistique d'ordre d'un échantillon i.i.d. de $Y = WX = WAS$. Soit F_A le fonction de répartition de Y.

Comme indiqué plus haut, si $Y = BS$ est une variable aléatoire de dimension d, $I(Y) = \sum_i H(Y^i) - H(S) - \log|\det B|$, le terme log étant nul si on se restreint aux matrices B orthogonales. On cherche donc $W^* = \operatorname{argmin}_W \left[H(Y^1)+\ldots+H(Y^d)\right]$, où Y^ℓ est la composante ℓ de Y.

L'estimateur proposé s'exprime par

$$\hat{H}(X_1, \dots X_n) = \frac{m}{n-1} \sum_{i=0}^{\frac{n-1}{m}-1} \log\left[\frac{n+1}{m}\left(X_{(mi+m+1)} - X_{(mi+1)}\right)\right]$$

où $X_{(i)}$ est la statistique d'ordre associée à l'échantillon.

La consistance de l'estimateur est montrée dans l'article de Vasicek (1976) et dans un article ultérieur de Song (2000).

Le paramètre m joue le rôle du paramètre de régularisation, et on doit s'assurer que $m \to +\infty$ et $m/n \to 0$. Les auteurs ont pris $m = \sqrt{n}$.

Dans la méthode proposée par Miller et Fisher, la minimisation de $\hat{H}(W)$ est réalisée en passant en revue les C_d^2 plans libres de $\mathbb{R}^d$, étant donné que une minimisation en dimension 2 équivaut à faire varier un paramètre θ entre 0 et $\pi/2$. On substitue ainsi à la minimisation sur la sous-variété de Stiefel $S(n,d)$, $d(d-1)/2$ minimisations dans $S(n,2)$. Cela revient à minimiser toutes les dépendances 2 à 2 entre les composantes de X, et concrètement on applique des rotations de Jacobi à la matrice W pour sélectionner le plan dans lequel on veut minimiser la fonction de contraste. Les auteurs proposent un calcul systématique le long d'une grille à $K = 150$ points pour échapper au problème des minimums locaux.

La méthode fournit donc seulement une indépendance par paire, moins difficile à obtenir qu'une indépendance mutuelle, mais suffisante dans le cadre ACI usuel.

On peut utiliser la même procédure de minimisation en substituant le contraste en ondelette au contraste de l'information mutuelle ; ce faisant on obtient en prime un critère qui se prête à la construction d'un test d'indépendance. D'autre part un contraste à base de Haar (de complexité numérique Cn avec $C = 1$) est utilisable dans le cas d'une minimisation sans gradient.

Constatant l'insuffisance de prise en compte de la régularité par le seul paramètre m, les auteurs ajoutent une seconde régularisation par bruitage, en remplaçant chaque observation X_i par $\sum_{j=1}^{R} \epsilon_j$ où ϵ_j suit une loi normale $N(X_i, \sigma^2 I_d)$.

Les choix effectués dans les simulations sont $\sigma = 0,35$ pour $n < 1000$ et $\sigma = 0,175$ pour $n > 1000$, $K = 150$, $R = 30$, $m = \sqrt{n}$. La complexité de l'algorithme est en $O(KRN \log RN)$, le terme log étant dû à la nécessité de trier les observations (typiquement de complexité $n \log n$).

La complexité est donc plus élevée que celle du contraste en ondelette dans sa version plug-in.

Sur l'ensemble-test de Bach et Jordan en dimension 2, les simulations montrent des résultats en moyenne meilleurs que FastICA, Jade, Infomax et kernel ICA.

Fonctionnelle matricielle

Tsybakov et Samarov (2002) ont proposé une méthode d'estimation des directions b_j, où $B = (b_1 \quad \dots \quad b_d)$ basées sur des estimations non paramétriques de fonctionnelles matricielles utilisant le gradient de f_A.

La méthode permet d'estimer la matrice $B = A^{-1}$ à une vitesse paramétrique et permet également d'estimer la densité f, dont les composantes sont indépendantes, à la vitesse de l'estimation d'une densité en dimension 1. La méthode d'estimation est basée sur l'utilisation de noyaux possédant au moins $d+3$ moments nuls, et le choix d'une fenêtre appropriée. Une caractéristique de cette méthode réside dans le fait que l'estimation s'opère directement, sans le recours à une fonction de contraste qui reste à minimiser. Cela revient

à dire qu'on obtient algébriquement des valeurs qui sont à trouver numériquement dans les autres méthodes.

Dans l'ACI par ondelettes, on a également une estimation des densités sous-jacentes du problème à partir de l'estimateur plug-in. On a également une estimation du signal à toutes les étapes du démixage, bien que on n'ait pas donné spécifiquement le risque de cette procédure d'estimation et que le j optimal ne soit pas le même que celui de l'estimation du contraste. L'estimateur U-statistique atteint également la vitesse paramétrique, mais seulement dans le cas $s \geq d/4$. Si on utilise la minimisation par paire, cela revient à dire que l'estimation atteint la vitesse paramétrique dès que $s > 1/2$.

Les auteurs considèrent la fonctionnelle $T(f) = E_f\big[\nabla f(x)\ {}^t\nabla f(x)\big] = \sum_{j=1}^d \sum_{k=1}^d c_{jk} b_j\ {}^t b_k$, où ∇f est le gradient de f, $c_{jk} = (\det B)^2 E\big[\prod_{i\neq j} p_i(\ {}^t x b_i) \prod_{m\neq k} p_m(\ {}^t x b_m) p_j'(\ {}^t x b_j) p_k'(\ {}^t x b_k)\big]$, et $B = (\,b_1 \quad \dots \quad b_d\,)$.

Avec la condition $\int (f^\ell)'(x)(f^\ell)^2 = 0$, $\ell = 1, \dots, d$, la fonctionnelle se simplifie en $T(f) = \sum_{j=1}^d c_{jj} b_j\ {}^t b_j = BC\ {}^t B$, $C = \mathrm{diag}(c_{jj})$.

T est définie positive, $B\ {}^t T^{-1} B = C^{-1}$ et $B\ {}^t \mathrm{var}(X) B = D$; cela implique que $P\ {}^t \Sigma P = \Lambda$ et $P\ {}^t T^{-1} P = I$ avec $\Sigma = \mathrm{var}(X)$, $P = BC^{\frac{1}{2}}$ et $\Lambda = C^{\frac{1}{2}} D C^{\frac{1}{2}}$.

Un résultat d'algèbre matricielle permet d'écrire Λ comme la matrice diagonale des valeurs propres de $T\Sigma$ et les colonnes de P comme les vecteurs propres de $T\Sigma$.

On cherche donc les vecteurs p_j, $j = 1, \dots, d$, solutions de $T\Sigma p_j = \lambda_j p_j$, ce qui est équivalent à chercher les vecteurs q_j solution de $T^{\frac{1}{2}} \Sigma T^{\frac{1}{2}} q_j = \lambda_j q_j$, où $q_j = T^{-\frac{1}{2}} p_j$ et les q_j orthogonaux entre eux.

On estime donc $W = T^{\frac{1}{2}} \Sigma T^{\frac{1}{2}}$, dont on prend ensuite la transformation ACP pour estimer les q_j, puis les $p_j = T^{\frac{1}{2}} q_j$, puis les $b_j = c_{jj}^{-\frac{1}{2}} p_j = p_j \|p_j\|^{-1}$.

L'estimateur de Σ est le classique estimateur empirique de la variance, et l'estimateur de T est donné par

$$\hat{T} = \frac{1}{n} \sum_{i=1}^n \nabla \hat{p}_{-i}(X_i)\ {}^t \nabla \hat{p}_{-i}(X_i)$$

où la composante l de $\nabla \hat{p}_{-i}(X_i)$ est donnée par

$$\frac{d\hat{p}_{-i}(X_i)}{dx^l} = \frac{1}{(n-1)h^{d+1}} \sum_{j=1,\, j\neq i}^n K_1\Big(\frac{X_j^l - X_i^l}{h}\Big) \prod_{k=1,\, k\neq l}^d K\Big(\frac{X_j^l - X_i^l}{h}\Big)$$

Avec des hypothèse de régularité de type Hölder pour les composantes de f, des conditions sur les fenêtres h, $nh^{2d+4} \to \infty$ et $nh^{2b-2} \to 0$, $b > d+3$, et b moments nuls pour les noyaux K, et K_1, et $E\|X\|^4 < \infty$, les auteurs montrent que l'estimation des directions b_j de la matrice B est consistante à la vitesse $\sqrt{n}$.

Dans le cas de l'ACI par ondelettes, le nombre de moments nuls de l'ondelette est indépendant de d, une ondelette D4 s'est avérée suffisante dans les simulations effectuées. Par ailleurs la consistance est démontrée en moyenne quadratique. Il n'y a pas de restriction sur la forme des densités.

La méthode de la fonctionnelle matricielle permet également d'estimer la densité f à partir de l'estimateur

$$\hat{f}(x) = \det \hat{B} \prod_{j=1}^{d} \frac{1}{n\tilde{h}_j} \sum_{i=1}^{n} \tilde{K}\Big(\frac{{}^t X_i \hat{b}_j - {}^t x \hat{b}_j}{\tilde{h}_j} \Big)$$

où $\tilde{h}_j \approx n^{\frac{-1}{2s_j+1}}$ et $\tilde{K}$ admet $s = \min s_j$ moments nuls. Cette estimation est consistante à la vitesse $n^{\frac{-s}{2s+1}}$, soit la vitesse optimale de l'estimation de densité en dimension 1. Mais le nombre de moments nuls du noyau dépend du paramètre s inconnu.

La méthode de Tsybakov et Samarov est optimale du point de vue théorique, possède l'avantage de donner la consistance de l'estimation de la matrice A et a donné de bons résultats dans les simulations (Gómez Herrero, 2004) mais possède aussi quelques inconvénients.

La condition $\int (f^\ell)'(x)(f^\ell)^2 = 0$ vérifiée pour toute densité à support dans $\mathbb{R}$ ou toute densité symétrique à support dans un intervalle, et la condition $E\|X\|^4 < \infty$ excluent de fait les densités ne répondant pas au critère (par exemple béta, Cauchy, χ^2, Pareto, triangulaire,...)

On a pu constater de moins bons résultats que les méthodes classiques dans le cas de densités multimodales (Gómez Herrero, 2004, p. 15).

La forme de l'estimateur $\hat{T}$ implique une complexité numérique élevée, au moins de l'ordre de $O(\kappa d n^2)$, où κ est un facteur multiplicatif constant dépendant du noyau qui doit posséder $d+3$ moments nuls.

Le choix de la fenêtre optimale h dépend de $b \leq \min s_j$, où les s_j sont les régularités des composantes inconnues (comme le j dans l'ACI par ondelettes, mais il y a des estimateurs adaptatifs).

L'estimation avec h uniforme dans les d dimensions s'avère instable sans une pré-transformation ACP, non formellement prévue par la méthode mais de fait largement utilisée dans la pratique de l'ACI.

La méthode donne un résultat relativement indéterminé avec des lois gaussiennes qui introduisent des valeurs propres multiples dans la décomposition spectrale de T. Même si les gaussiennes sont précisément hors du champ de l'ACI, en pratique il n'est pas impossible de tomber sur ces cas là.

En comparaison, en tant que critère d'indépendance très général, le contraste en ondelette garde tout son sens en présence de gaussiennes, même si la minimisation s'en trouve éventuellement compromise. Cela revient à dire que dans le cas de la méthode

par ondelettes, la consistance porte sur l'estimation du contraste exact $C(f_A)$, et pas spécifiquement sur celle de la matrice A.

1.6 Perspectives

Au terme de ce projet de thèse, on a envisagé un certain nombre de prolongements possibles, d'ordre informatique, statistique ou purement pratique.

Prolongements d'ordre pratique

Sur le plan informatique, on peut chercher à améliorer la complexité numérique de la minimisation du contraste.

Des simulations en dimension 2 montrent que l'ondelette de Haar produit une courbe de contraste visuellement elliptique, souvent sans minimum locaux (voir p. 106). Le problème pour Haar vient du calcul du gradient. Des méthodes de minimisation n'impliquant pas de gradient, pourraient rétablir l'utilité de ce contraste. Dans ce cas la complexité de la projection sur V_j passe de $O(n(2N-1)^d)$ à $O(n)$ ($N = 1$ pour Haar).

Par exemple on peut appliquer la méthode testée par Miller et Fisher (2003) et par Gretton et al. (2004), qui consiste à rechercher le minimum dans chacun des C_d^2 plans libres de $\mathbb{R}^d$, en appliquant des rotations de Jacobi pour sélectionner un plan en particulier. La recherche dans chacun des plans est équivalent au cas $d = 2$, où on recherche le minimum en θ d'une fonction réelle, pour $\theta \in [0, \pi/2]$. Pour ce faire le plus simple est encore d'essayer tous les points de 0 à $\pi/2$ selon une grille plus ou moins fine, ou d'employer des méthodes de type bissection si la courbe ne possède pas de minimums locaux (comme c'est le cas pour le contraste en ondelettes dans les exemples p. 106).

On peut envisager d'implémenter l'estimateur U-statistique qui serait au maximum de complexité numérique n^2, sur lequel pourrait être opéré le seuillage par bloc. Cela permettrait avec un nombre d'observation peu important, comme c'est souvent le cas dans les simulations ACI, de donner un critère adaptatif potentiellement plus précis que l'estimateur plug-in, dans sa version linéaire ou adaptative.

Prise en compte de l'anisotropie

Si les composantes de f sont de régularités très différentes, on peut étudier le cas anisotrope.

Dans ce projet, on s'est contenté d'un découpage de l'espace en cube de côté 2^{-j}, mais dans le cas anisotrope on peut modifier la méthode en considérant des rectangles avec des j différents selon les dimensions. Du point de vue de l'implémentation, cela ne constitue pas une extension très difficile. Du point de vue théorique, il faut considérer des espaces de Besov anisotropes.

Contraste de type plug-in atteignant une vitesse paramétrique

Il pourrait être intéressant de prolonger l'étude à des contrastes plus généraux du type $\sum_k |\hat{\alpha}_{jk} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d}|^p$ ou encore $\sup_k |\hat{\alpha}_{jk} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d}|$, ce dernier contraste pouvant éviter le stockage d'un tableau de taille 2^{jd} puisqu'on n'a besoin de conserver que le maximum courant de $|\hat{\alpha}_{jk} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d}|$.

Le cas L^2 peut en effet être prolongé de la façon suivante : à partir du lemme de Meyer (lemme 4.8) on obtient,

$$\int |f|^p \leq 2^{p-1}\big(\|P_j f\|_p^p + 2^{-pjs}\big) \leq C 2^{jd(\frac{p}{2}-1)} \sum_k |\alpha_{jk}|^p + C 2^{-pjs}$$

On définit alors le contraste en ondelettes généralisé

$$C_j^p(f_A - f_A^\star) = \sum_{k \in \mathbb{Z}^d} \left| \int (f_A - f_A^\star)\Phi_{jk} \right|^p$$

et le contraste normalisé $C_j^{p\prime}(f_A - f_A^\star) = 2^{\frac{jd}{2}(p-2)} C_j^p(f_A - f_A^\star)$.

On voit que pour $f \in B_{spq}$, si le contraste généralisé est égal à zéro, $\int |f_A - f_A^\star|^p \leq C 2^{-pjs}$ (et dans tous les cas lorsque la fonctionnelle C_j^p est égale à zéro, le contraste généralisé est aussi égal à zéro).

Pour le problème de l'estimation d'une fonctionnelle non quadratique, Kerkyacharian et Picard (1996) ont utilisé un développement exact de $(P_j f + f - P_j f)^3$, où P_j est l'opérateur de projection associé à l'ondelette de Haar.

Dans notre cas, sauf pour $p = 2$, le contraste généralisé C_j^p défini plus haut n'utilise qu'une inégalité de convexité. Cette approche probablement sous-optimale dans le cas de l'estimation d'une fonctionnelle non quadratique fournit néanmoins une famille de contrastes ACI applicables à n'importe quelle ondelette de Daubechies.

Dans le cas $p = \infty$, on peut définir $C_j^\infty(f_A - f_A^\star) = \sup_{k \in \mathbb{Z}^d} \left| \int (f_A - f_A^\star)\Phi_{jk} \right|$ et on a de le même façon

$$\begin{aligned} \sup_x |f| &\leq \sup_x |P_j f| + 2^{-js} \\ &\leq C 2^{\frac{-jd}{2}} \sup_k |\alpha_{jk}| + C 2^{-js} \end{aligned}$$

c'est-à-dire que si le contraste C_j^∞ est nul, $\sup_x |f| \leq C 2^{-js}$.

Les premières investigations semblent indiquer que un p au-delà de 2, améliore la vitesse du contraste plug-in. On aurait alors un résultat ressemblant à celui-ci :

Soit $p \in \mathbb{N}^*$. Soit $X_1, \dots, X_n$ un échantillon i.i.d. de f, une densité à support compact définie sur $\mathbb{R}^d$. On suppose que φ est une ondelette de Daubechies $D2N$. Soit $\Omega_n^m = \big\{(i^1, \dots, i^m) : i^\ell \in \mathbb{N},\, 1 \leq i^\ell \leq n\big\}$. Soit $I_n^m = \big\{i \in \Omega_n^m : \ell_1 \neq \ell_2 \Rightarrow i^{\ell_1} \neq i^{\ell_2}\big\}$

Pour $i \in \Omega_n^p$, soit $h_i = \sum_k \Phi_{jk}(X_{i^1}) \dots \Phi_{jk}(X_{i^p})$. Soit $\theta = \sum_k \alpha_{jk}^p$. Soit $\hat{B}_j^p = (A_n^p)^{-1} \sum_{i \in I_n^p} h_i$ l'estimateur U-statistique de θ. Soit $\hat{H}_j^p = n^{-p} \sum_{i \in \Omega_n^p} h_i$ l'estimateur plug-in de θ.

Sur l'ensemble $\{2^{jd} < n^2\}$,

$$E_{f_A}^n |\hat{B}_{j\alpha^p} - \theta|^2 \leq C 2^{jd(2-p)} n^{-1} \mathbb{I}\left\{2^{jd} < n\right\} + C 2^{jd} n^{-p} \mathbb{I}\left\{2^{jd} > n\right\}$$
$$E_{f_A}^n |\hat{H}_{j\alpha^p} - \theta|^2 \leq C 2^{jd} 2^{jd(2-p)} n^{-1} \mathbb{I}\left\{2^{jd} < n\right\} + C 2^{jdp} n^{1-2p} \mathbb{I}\left\{2^{jd} > n\right\}$$

Autrement dit le cas $p = 2$ était en fait insuffisant pour le contraste plug-in $(\hat{H}_{j\alpha^p})$, et le choix $p \geq 3$ assurerait une vitesse paramétrique sur l'ensemble $2^{jd} < n$.

Une étude plus complète du cas p général permettrait donc de statuer sur cette hypothèse, et d'envisager une stratégie de choix de p.

Test d'indépendance

On peut envisager la construction d'un test statistique permettant de décider objectivement que le minimum global a été quasiment atteint dans une situation réelle où le critère d'Amari n'est pas disponible.

Butucea et Tribouley (2006) ont déjà proposé un test d'homogénéité basé sur le critère $\int (f-g)^2$, pour f et g deux fonctions réelles. Pour tester l'hypothèse

$$H_0 \colon f = g$$

contre l'alternative

$$H_1 \colon f, g \in V \cap \{f, g \colon \|f - g\|_2^2 \geq C r_{n,m}\}$$

où V est un espace fonctionnel nécessaire à l'estimation, et $r_{n,m}$ une suite qui tend vers zéro quand $n \wedge m$ tend vers l'infini. Les auteurs considèrent la statistique à deux échantillons

$$T_j = [(n \wedge m)(n \wedge m - 1)]^{-1} \sum_{i_1, i_2 = 1}^{n \wedge m} \sum_k \left(\varphi_{jk}(X_{i_1}) - \varphi_{jk}(Y_{i_1})\right)\left(\varphi_{jk}(X_{i_2}) - \varphi_{jk}(Y_{i_2})\right)$$

et une statistique de test D,

$$D = D_j = \begin{cases} 0 & \text{si } |T_j| \leq t_{j,n,m} \\ 1 & \text{si } |T_j| > t_{j,n,m} \end{cases}$$

où j et $t_{j,n,m}$ sont à choisir de façon à obtenir le meilleur test D_j dans une famille $\{D_j, j \in J\}$, ou bien dans le cas adaptatif, $D = \max_{j \in J} D_j$.

Les auteurs obtiennent que la vitesse minimax de séparation des hypothèses H_0/H_1 au niveau γ en dimension 1 est de l'ordre de $n^{\frac{-4s}{4s+1}}$ pour un choix approprié de la résolution j et du seuil du test t_j.

La procédure devrait s'appliquer mutatis mutandis au cas $g = f^\star$, compte-tenu de ce qui a été fait dans ce projet pour le calcul d'une borne du risque (voir notamment la partie p. 26 et suivantes pour les similitudes entre la statistique utilisée par Butucea et Tribouley et la nôtre).

D'autre part, le test d'indépendance qu'on trouve dans l'article de Rosenblatt (1975) devrait également pouvoir s'appliquer.

Considérant en dimension 2 la mesure de factorisation empirique

$$S_n = \int \left[f_n(x) - g_n(x^1)h_n(x^2)\right]^2 dx,$$

où $f_n(x) = \frac{1}{nh^2}\sum_{j=1}^n w\left(\frac{x-X_i}{h}\right)$, $g_n(x^1) = \frac{1}{nh}\sum_{j=1}^n w\left(\frac{x^1-X_i^1}{h}\right)$ (idem pour h) et $w(x) = w^1(x^1)w^2(x^2)$ est un noyau défini positif à 2 moments nuls, et avec un certain nombre de conditions supplémentaires, Rosenblatt aboutit à une convergence de la quantité

$$h^{-1}\Big[-A(h) + nh^2 \int \left[f_n - g_n h_n\right]^2\Big]$$

vers une loi normale de moyenne nulle et de variance $2w^{(4)}\int f^2$ quand $n \to \infty, nh^2 \to \infty$ et $h = o(n^{-\frac{1}{5}})$, où $A(h)$ est une quantité en $O(1)$.

Dans son article, Rosenblatt s'interroge sur les avantages d'un test basé sur les densité plutôt que sur la distribution empirique, et indique que comparé à certains tests sur les distributions de Blum et al. (1961) le test basé sur les densités est éventuellement moins puissant mais plus aisé à mettre en œuvre en raison de la convergence vers une loi normale au lieu d'une loi aux propriétés inconnues dans le second cas. D'autre part un estimateur de la fonction de répartition de type "espacé" qui serait basé sur une statistique d'ordre (voir le cas de l'estimateur RADICAL p. 75) à une complexité en n plus élevée, de l'ordre de $n \log n$.

Adaptation

En plus de l'étude du seuillage global sur un estimateur U-statistique du contraste L_2, on peut envisager l'application de procédures adaptatives alternatives comme la méthode de Lepski (lissage aléatoire), ainsi que des méthodes d'agrégation ou de sélection de modèles.

Gradient exact du contraste L_2

Dans ce document on a étudié le gradient de l'estimateur empirique du contraste en projection. On peut aussi chercher à déterminer le gradient et le hessien exact du contraste L_2, ce qui revient à s'intéresser à des quantités proches de celles qu'on trouve dans la méthode de la fonctionnelle matricielle. On peut ensuite chercher à estimer ces quantités dérivées selon une méthode non paramétrique à base d'ondelettes.

Mixage non linéaire

Le contraste en ondelette reste valide même si le mixage n'est pas linéaire.

Pour une variable aléatoire $X = F(S)$ où $F: U \subset \mathbb{R}^d \mapsto V \subset \mathbb{R}^d$ n'est plus linéaire mais inversible, la mesure de factorisation s'écrit de la même façon, $\int \|f_F - f_F^*\|^2$ et on en a une estimation par le contraste C_j. Dans le cas non linéaire, la minimisation ne s'opérerait plus sur la sous-variété de Stiefel, mais sur un autre ensemble dépendant de F.

ACI sans ACP

On peut envisager une minimisation sur la sous-variété de Stiefel sans tranformation ACP préalable.

Soit A une matrice inversible ; par factorisation QR, il existe une matrice Q orthogonale telle que $QA = T$ soit triangulaire supérieure. Si la diagonale de T est strictement positive, Q et T sont uniques.

Pour une telle matrice T l'inversion est immédiate.

Soit $x = As$; il existe donc une unique matrice Q orthogonale telle que $y = Qx = QAs = Ts$. La normalisation $\operatorname{var} s = I_d$ implique que $\operatorname{cov} y = TT'$. La matrice symétrique TT' a donc pour terme général $\operatorname{cov}(y^k, y^l) = a_{kl} = \sum_{i \geq l} t_{ki} t_{li}$, quantités estimables à partir des observations et de la matrice Q.

Soit W une matrice orthogonale, si $W = Q$, T^{-1} est estimable à partir de la matrice de covariance observée.

On pourrait donc considérer la fonction de contraste

$$C: Q \mapsto C(\hat{T}^{-1} Qx).$$

En dimension 2, $T = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$, $TT' = \begin{pmatrix} a^2 + b^2 & bc \\ bc & c^2 \end{pmatrix}$, $T^{-1} = \begin{pmatrix} a^{-1} & -b(ac)^{-1} \\ 0 & c^{-1} \end{pmatrix}$.

Soit,

$$c^2 = \operatorname{var} y_2,\ a^2 = \operatorname{var} y_1 - \operatorname{cov}(y_1, y_2)^2 (\operatorname{var} y_2)^{-1},\ b^2 = \operatorname{cov}(y_1, y_2)^2 (\operatorname{var} y_2)^{-1}.$$

et

$$\begin{aligned} s_2 &= c^{-1} y_2 \\ s_1 &= a^{-1}[y_1 - bc^{-1} y_2] \end{aligned}$$

1.7 Organisation du document

Les parties 3 à 5 correspondent à trois articles ayant été soumis à des revues à comité de lecture. La partie 6 contient tous les éléments utiles à l'implémentation de la méthode.

- Dans la partie 3, on montre expérimentalement que la méthode fonctionne, que l'ondelette de Haar convient en dehors de l'usage d'un gradient, que les courbes en dimension 2 sont en général elliptiques si on est à une résolution convenable, qu'il existe une bande de résolutions convenables autour du j optimal, que le contraste en ondelette a un pouvoir discriminant très fin (des rotations de l'ordre de un demi degré sont détectées), et que les temps de calculs sont compétitifs en petite dimension, même pour nombre d'observations très grand.

  Du point de vue théorique, on montre que le mixage par A inversible conserve l'appartenance Besov des densités originales. On relie le critère C_j à une mesure de l'indépendance des composantes de X, et on montre, en adoptant un point de vue volontairement simplifié, que l'erreur en moyenne quadratique (EMQ) du contraste en ondelette est au plus de l'ordre $n^{\frac{-2s}{2s+d}}$, soit la vitesse minimax de l'estimation de densité.

- Dans la partie 4, on développe la relation entre le problème $\int f^2$ et le problème ACI. On montre que la vitesse de convergence de l'EMQ du contraste en ondelettes est en fait au moins de l'ordre $Cn^{\frac{-4s}{4s+d}}$ et on montre que des estimateurs U-statistiques atteignent une vitesse paramétrique pour des régularités $s \geq d/4$, comme dans le problème $\int f^2$.

  La démonstration passe par des lemmes combinatoires permettant de simplifier des calculs tels quels irréalisables. On montre que un estimateur U-statistique pris sur l'échantillon tout entier est légèrement sous-optimal ; et en conséquence on étudie des stratégies de découpage de l'échantillon qui permettent de retrouver l'optimalité.

- Dans la partie 5, on étudie le risque d'un estimateur seuillé du contraste en ondelettes. Des simulations complètent la présentation théorique.

- Dans la partie 6, on a regroupé tout ce qui concerne l'implémentation de la méthode, ce qui comprend à la fois des informations sur la sous-variété de Stiefel et sur la méthode de minimisation contrainte à cette sous-variété, les formulations du gradient et du Hessien et leurs propriétés de filtrage, des éléments pratiques et le détail commenté des parties clefs de la programmation.

Notations et conventions

Dans la suite on peut être amené à employer indifféremment le terme "contraste en ondelettes" pour désigner $C_j \in \mathbb{R}^+$, un estimateur $\hat{C}_j(X_1, \ldots, X_n)$, où X_i est une variable aléatoire de densité f_A, ou bien la fonction $y_1, \ldots y_n \in \mathbb{R}^{d \times n} \mapsto \hat{C}_j(y_1, \ldots, y_n)$ ou encore la

fonction $W \in \mathbb{R}^{d\times d} \mapsto \hat{C}_j(Wx_1, \ldots, Wx_n)$ où $x_i \in \mathbb{R}^d$ est l'observation i.

Dans les parties 4 et 5, C_j est noté C_j^2, et idem pour l'estimateur $\hat{C}_j$, en référence à la norme au carré de $P_j f$.

D'une manière générale, les caractères en exposant désignent les coordonnées d'entités multidimensionnelles, tandis que les caractères en indice désignent des éléments d'un même ensemble.

Autres notations

C_n^p, A_n^p	$n!/p!(n-p)!$, $n!/(n-p)!$
C_j	contraste en ondelette sauf partie 4
C_j^2	contraste en ondelette dans la partie 4
$\tilde{X}$, X_i, X_i^ℓ	échantillon $X_1 \ldots, X_n$, variable numéro i, composante ℓ de la variable i
f_A	$\lvert\det A^{-1}\rvert\, f(A^{-1}x)$
$E_{f_A}^n$	espérance par rapport à la loi jointe du couple $X_1, \ldots, X_n$ de densité $f_A^{\otimes n}$
$f^\star$, $f^{\star\ell}$	produit des marges de f, marge numéro ℓ de f
$a \vee b$, $a \wedge b$	$\max(a,b)$, $\min(a,b)$
$\mathbb{I}\{R\}$, $\mathbb{I}_B$	fonctions indicatrices
Ω_n^m, I_n^m	$\{(i^1, \ldots, i^m) \colon i^\ell \in \mathbb{N},\, 1 \le i^\ell \le n\}$, $\{i \in \Omega_n^m \colon \ell_1 \neq \ell_2 \Rightarrow i^{\ell_1} \neq i^{\ell_2}\}$
Φ_{jk}, Ψ_{jk}, Λ_{jk}	$\varphi_{jk^1}(x^1) \ldots, \varphi_{jk}(x^d)$, voir le rappel sur les ondelettes dans la partie 3, voir p. 26
α_{jk}, β_{jk}	coordonnées du développement en ondelette sur les fonctions Φ_{jk}, Ψ_{jk}
α_{jk^ℓ}, β_{jk^ℓ}	coordonnées du développement en ondelette sur les fonctions φ_{jk^ℓ},
λ_{jk}, δ_{jk}	$\alpha_{jk^1} \ldots \alpha_{jk^d}$, $\alpha_{jk} - \lambda_{jk}$
$\lambda_{jk}^{\langle r \rangle}$	$\alpha_{jk^1}^{p_1} \ldots \alpha_{jk^d}^{p_d}$ pour des p_i tels que $0 \le p_i \le r$, $\sum_{i=1}^d p_i = r$
$\lvert A\rvert$	cardinal de l'ensemble A
N	paramètre de l'ondelette Daubechies $D2N$, N=1 pour Haar
n, d	taille de l'échantillon, dimension du domaine de f
W	candidat à l'inversion de la matrice de mixage A

2. Introduction (english translation)

Given a scatter of n points in dimension d, principal component analysis (PCA) consists of finding the direction of the space carrying the largest part of the total dispersion ; then the orthogonal direction carrying the second largest part, and so on. From a geometrical point of view it is a simple change of basis, resulting in a new representation in directions corresponding to decorrelated variables (in the empirical sense), and thereby obtaining the factor status, likely to explain the information contained in the scatter without the redundancy that could characterize the initial representation.

The emblematic cocktail party example illustrates the point of view of signal processing : to separate the conversations of d persons talking at the same time, set d microphones in the room, each recording a superposition of all conversations, but more clearly those in direct proximity. The problem is to isolate each source to understand what was said.

In psychometrics, one of the disciplines precursory in factorial analysis, one makes use of latent variables, that is to say not directly observable but whose effects are measurable through series of tests, each test revealing, partly and among others, the effect of some variable or other of the latent model. A known example is the five dimensions of personality (Big Five) identified from factorial analysis by researchers in personality evaluation (Roch, 1995).

Other application fields such as numerical imaging, data mining, economy, finance or statistical analysis of texts, make use of those inverse type models aiming to reveal independent factors to be extracted from an accumulation of possibly quite diverse indicators that are globally related to a phenomenon.

Independent component analysis (ICA) is one such tool, with or without dimension reduction, and whose goal is to reveal factors that are independent in the full sense and not only in the sense of decorrelation.

The model is usually stated as follows : let X be a random variable on $\mathbb{R}^d$, $d \geq 2$, such that $X = AS$ where A is a square invertible matrix and S a latent random variable whose components are mutually independent. One tries to estimate A, to reach $\{S_1, \ldots, S_n\}$, given an independent identically distributed random sample $\{X_1, \ldots, X_n\}$ distributed according to the law of X, that is to say such that X_i is independent of X_j for $j \neq i$, but such that components $X_i^1, \ldots, X_i^d$ of a single observation X_i are a priori not mutually independent.

ICA thus only considers linear superpositions of independent signals, resulting from the mixing by A. This is often a legitimate restriction ; for instance transmission systems are linear media where signals act as if they were present independently of each other, they do not interact but add up (Pierce, 1980).

Other formulations take into account the so called post non linear mixtures (Taleb, 1999,

Achard, 2003) consisting in monotone tranformations of linear mixtures and expressed by $X = f(AS)$, f having monotone components and A invertible. In other cases, one does not build upon independence but on the contrary tries to exploit a temporal correlation of the sources. Also exist convolutive models, where the mixing by A is not instantaneous, but of the form $X(t) = \sum_u A(u)S(t-u)$, $A(u)$ designating a sequence of invertible matrices. Finally, one can consider models with noise.

The model under study in what follows is the standard problem defined above. For this problem, in practice, the (empirical) PCA transformation of $d \times n$ matrix $M = (\, x_1 \quad \dots \quad x_n \,)$ containing the observed signal gives the first part of the answer, by linear decorrelation ; and the whole answer if the signals are purely Gaussian. In the alternative, to fully solve the problem, the ordinary procedure consists of minimizing some contrast function $C = C(W)$ that cancels out if and only if components of WX are independent, where W is a $d \times d$ matrix candidate to represent an ICA inverse of A.

The ICA problem is always parametric in A and is parametric or non parametric in S according to the functional hypothesis applied to the probability density of S, which is always supposed to admit a density relative to Lebesgue measure (discrete deconvolution models are also studied, see Gassiat and Gautherat, 1999).

The ICA density model is written in the following way : let f be the density of S relative to Lebesgue measure ; the observed variable $X = AS$ admits density f_A, defined 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 line number ℓ of the matrix $B = A^{-1}$; this writing comes from a change of variable, given that f, the density of S, is the product of its margins $f^1 \dots f^d$.

In the ICA model expressed this way, f and A are the two unknown, and the data consists in an independent, identically distributed sample $\{X_1, \dots, X_n\}$ of f_A. The semi-parametric case corresponds to non parametric hypotheses for f, which is left unspecified except for general regularity assumptions required for estimation.

2.1 Main approaches in the ICA problem

One can find in the extensive ICA literature advantages and drawbacks of the different methods that have been proposed since the eighties and particularly the nineties, when the first effective algorithms were introduced.

It can be noticed that statistical properties (especially convergence rates) of classical models are generally not studied very explicitly, for it is not necessarily in the top preoccupations of researchers of domains under concern, and also because Amari criteria can serve as a reference to validate a method.

The effectiveness of an algorithm is in fact not specifically measured by how much dependence was suppressed, but by its ability to find the inverse of A. The criteria in use is Amari distance (Amari, 1996) normalized between 0 and 100, between matrix A (known in simulations) and the estimation of its ICA inverse W,

$$\frac{100}{2d(d-1)}\left[\sum_{i=1}^{d}(\sum_{j=1}^{d}\frac{|p_{ij}|}{\max_k|p_{ik}|}-1)+\sum_{j=1}^{d}(\sum_{i=1}^{d}\frac{|p_{ij}|}{\max_k|p_{kj}|}-1)\right] \qquad \text{with } WA=(p_{ij}).$$

Both objectives are besides designated by different names (one talks about mixing estimation versus source restitution) and do not coincide in certain types of models where A is not squared invertible.

Most classical methods do not strictly speaking make parametric hypotheses for f, but that does not entail the implementation of a typical non parametric procedure.

The explanation is that these methods make use of substitution contrasts, whose cancellation does not exactly imply mutual independence, and which are from a theoretical viewpoint easier to estimate than an exact criteria. It is a fact that in many cases the contrasts used yield very satisfying ICA inverses of A.

Methods pushing forward their non parametric nature are based on exact contrasts, whose cancellation implies mutual independence (sometimes only pairwise independence), and whose evaluation in full generality is only possible through the specific technical constraints of non parametric estimation.

Taken apart the fact that classical methods are often characterized by an easier implementation and a low numerical complexity, the most obvious distinction between the two groups is certainly the resort or not to a regularization parameter, a key element in function approximation techniques, and an element that is linked to the general performance of the ICA algorithm in cases where the statistical properties are the most clearly established.

Classical methods

Maximum-likelihood methods and contrast functions based on mutual information or other divergence measures between densities are commonly employed. Comon (1994) defined the concept of ICA as maximizing the degree of statistical independence between outputs, using contrast functions based on an Edgeworth expansion of Kullback-Leibler divergence.

In other respects, Hyvärinen and Oja (1997) proposed the FastICA algorithm, which is a minimization method based on a fixed point method (instead of gradient) applicable to several types of ICA contrasts.

One usually identifies four categories of methods (see Hyvärinen et al. 2001) :

- Maximization of non Gaussianity (Hyvärinen, 1999) ;

  The principle is to maximize the non Gaussianity of a linear combination ${}^t bx$ of observation x. The non Gaussian nature is estimated through kurtosis or neg entropy.

  It is a deflationary method, that is to say operating component by component, as opposed to a simultaneous approach where one tries to find directly W allowing the extraction of all components at the same time. This can entail the accumulation of the error in the components that are estimated first, and increase the error in later components (Hyvärinen et al, 2001, p. 271). It is also a method not robust against outliers because of a cubic increase of the kurtosis in $|y|$ (idem, p. 277).

  Maximization is carried out by means of a gradient method or by a fixed point method (FastICA).

- Maximum likelihood and particularly Infomax algorithm (Bell, Snejowski, 1995) ;

  The likelihood of a sample $\{x_1, ..., x_n\}$ is written

$$\frac{1}{n} \log L(x_1, \ldots, x_n, B) = \log |\det B| + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \log f^j(\ {}^t b_j x_i).$$

  In ICA by maximization of the likelihood one seeks after densities f^j in a family $\{p_\theta\ ;\ \theta \in \Theta\}$ where Θ contains only 2 values corresponding one to a supergaussian density, and the other to a subgaussian density.

  This is explained by the fact that for any function G even and C^2, and let $z = NAs$ be the ACP transform of $x = As$, local maxima (resp. minima) in w of $EG(\ {}^t wz)$ under the constraint $\|w\| = 1$ are the lines ℓ of matrix NA, such that $E[s^\ell G'(s^\ell) - G''(s^\ell)] > 0$ (resp. > 0), where s^ℓ is component ℓ of s. The condition comes from a Taylor expansion.

  Function G thus divides the space of densities in two parts according to the sign of the criteria associated with G and s^ℓ, and one representative of each half-space is enough to obtain convergence (see Hyvärinen, 2001, p.201).

  In the case of maximum likelihood, function $G = (G^1, \ldots, G^d)$ is given by

$$G^\ell = \log f^\ell(\ {}^t wz)\ ;$$

  in this expression, any density in the same half-space than f^ℓ is also suitable ; in this way, in the algorithm of Bell and Snejowski (1990s) one takes $(G^\ell_+)'(y) = -2\tanh(y)$ and $(G^\ell_-)'(y) = \tanh(y) - y$.

The Infomax algorithm (Bell and Snejowski, 1995), is based on a neural network and is comparable to a maximum likelihood method (Cardoso, 1997).

It must be noted that methods based on maximization of non Gaussianity and maximum likelihood (using the splitting above) are in the wrong if there is a half-space mismatch at initialisation.

- Tensorial methods, and particularly FOBI and JADE (Cardoso, 1990, 1994).

  To suppress correlations of $x = (x^1, \ldots, x^d)$ up to order 4, one considers cumulant tensors $\mathrm{cum}(x^i, x^j, x^k, x^l)$ that generalize covariance matrices $\mathrm{cov}(x^i, x^j)$.

  Let F be the order 4 tensor, operating on matrices $d \times d$, defined by

$$M, d \times d \mapsto F_{ij}(M) = \sum_{k,l} M_{kl} \,\mathrm{cum}(x^i, x^j, x^k, x^l)\,;$$

  an eigenmatrix of F is a matrix M such that $F(M) = \lambda M$, $\lambda \in \mathbb{R}$.

  It can be shown that orthogonal matrices $W = w_m \, {}^t w_m$, $m = 1 \ldots, n$, are eigenmatrices of F. The associated eigenvalues are the kurtosis of independent components s^ℓ.

  JADE (joint approximate diagonalization of eigenmatrices) builds upon this principle (Cardoso, 1993).

- Non linear decorrelation, algorithm of Jutten and Hérault (1987), Cichocki and Ubenhauen (1992).

  The purpose is to estimate $Ef(y^1)g(y^2)$, ideally for any continuous compactly supported function f, g, thus passing from a simple decorrelation criteria to an orthogonality criteria, equivalent to independence if at least one of the two variables is centered.

  Using a Taylor expansion, the criteria becomes

$$\sum_{i=1}^{p} \sum_{j=1}^{p} \frac{1}{j!i!} f^{(i)}(0) g^{(j)}(0) E(y^1)^i (y^2)^j + R = 0$$

  and the problem is transformed in a decorrelation of the components y^1, y^2 at every power i, j of the development at order p. This amounts to exploit the different moments of the observations, with all involved robustness problems. Jutten and Hérault are historically among the first to present an ICA algorithm. It is known that their method presents convergence problems in certain situations (J.C. Fort, 1991) ; Amari and Cardoso (1997) have proposed an extension of it (estimable functions).

Mutual information

Hyvärinen et al. (2001, p. 274) noticed that the theoretical criteria of mutual information between components of $x = As$ has a universal nature, in that it can be seen as the limit of maximum likelihood principle to which many methods are comparable.

Let $\hat{f}_A$ be an estimator of the density f_A of $X = AS$; the likelihood $L(x_1, \ldots, x_n, \hat{f}_A) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_A(x_i)$ tends to $E_{f_A} \log \hat{f}_A(x)$, which quantity is also written, using substitution $\hat{f}_A = \hat{f}_A f_A^{-1} f_A$,

$$\int f_A \log \hat{f}_A = \int f_A \log f_A + \int f_A \log \frac{\hat{f}_A}{f_A} = -K(f_A || \hat{f}_A) - H(f_A),$$

Differential entropy $H(f_A)$ is a constant independent of A, since it is invariant by invertible transformation (or more exactly by orthogonal transformation, contrary to Shannon entropy – Comon 1994, p.293 –, which supposes that the problem is reduced to the case A orthogonal by PCA pre-whitening), just as Kullback-Leibler divergence K, and since $s = A^{-1}x$,

$$K(f_A(x) || \hat{f}_A(x)) = K(f(s) || \hat{f}(s)) \; ;$$

components of s being independent, divergence can also be decomposed into

$$K(f(s) || \hat{f}(s)) = K(f(s) || f^\star(s)) + K(f^\star(s) || \hat{f}(s))$$

where $f^\star(s)$ is the density product of the components of s. This quantity is minimum when $\hat{f}_A = f_A^\star$ and is equal to $K(f(s) || f^\star(s))$ which is also the mutual information of s components.

Mutual information of the pair (x, y) is also defined by $I(x, y) = H(x) + H(y) - H(x \otimes y)$, where H is differential entropy and $x \otimes y$ is the compound ; This quantity is always positive and equal to zero if and only if x and y are independent.

The criteria is generally used under the form of a marginal entropy minimum : if $Y = BS$ is a d dimensional random variable, $I(Y) = \sum_i H(Y^i) - H(S) - \log|\det B|$, the log term being zero for orthogonal matrices B. So one seeks in this case after an element $W^* = \operatorname{argmin}_W \left[H(Y^1) + \ldots + H(Y^d)\right]$, where Y^ℓ is component ℓ of Y.

It can be noticed that proceeding this way does not allow to know if the minimum obtained is close to mutual independence, since $H(S)$ is not known (and $H(Y)$, integral in dimension d, is not estimated). Marginal entropy minimum thus eliminates the possibility to build a decision rule to accept/reject the minimum reached.

Essentially non parametric methods

These methods can be categorized in three groups : kernel methods, exact contrast methods, and direct methods.

- Group 1 — kernel ICA (Bach and Jordan, 2002) and methods based on a Hilbert-Schmidt norm (Gretton et al., 2003, 2004) which provide pairwise rather than mutual independence

criteria, by means of kernel methods in a reproducing Hilbert space (RKHS). The exact criteria on which kernel methods are based is the classical independence condition of two σ-algebras : X and Y are independent if $Efg = EfEg$ for all f squared integrable on the sets of the σ-algebra generated by X and all g squared integrable on the sets of the σ-algebra generated by Y. It is then implicitely assumed that the RKHS is a subset of $L_2(\sigma(X))$ and $L_2(\sigma(Y))$.

For two real random variables x and y, the kernel ICA contrast is defined as the $\mathcal{F}_\sigma$-correlation $\rho_{\mathcal{F}}$,

$$\rho_{\mathcal{F}_\sigma} = \max_{f,g \in \mathcal{F}_\sigma} \operatorname{corr}(f(x), g(y)).$$

$\mathcal{F}_\sigma$-correlation is zero for x and y independent, and, for a space of functions $\mathcal{F}_\sigma$ big enough, containing for instance functions $x \mapsto e^{iwx}$, a null $\mathcal{F}_\sigma$-correlation implies independence of x and y. In kernel ICA, $\mathcal{F}_\sigma$ stems from the isotropic Gaussian kernel $k(x,y) = e^{-\frac{1}{2\sigma^2}\|x-y\|^2}$, with $\emptyset = \mathcal{F}_0$ and $\mathcal{F}_\sigma$ increase to $L_2(\mathbb{R}^d)$, when $\sigma^2 \to +\infty$.

In Hilbert-Schmidt norm methods, one considers two measured spaces $(\mathcal{X}, \Gamma, p_x)$ and $(\mathcal{Y}, \Lambda, p_y)$, and the product space $(\mathcal{X} \times \mathcal{Y}, \Gamma \otimes \Lambda, p_{xy})$.

One also considers $\mathcal{F}$, a RKHS of functions from $\mathcal{X}$ to $\mathbb{R}$ associated to the positive definite kernel $k(.,.)\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and to characteristic function ϕ, $\phi(x) = k(x,.)$; and one considers $(\mathcal{G}, l(.,.), \psi)$, a second RKHS of functions from $\mathcal{Y}$ to $\mathbb{R}$.

The contrast is defined as the Hilbert-Schmidt norm of the cross-covariance operator

$$C_{xy} = E_{xy}\big[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)\big]$$

where E_{xy} is the expectation relative to the joint law p_{xy}, $\otimes$ is the tensorial product defined by $f \otimes g\colon h \in \mathcal{G} \mapsto f < g, h >_{\mathcal{G}} \in \mathcal{F}$ and μ_x, μ_y are given respectively by $< \mu_x, f >_{\mathcal{F}} = E_x f(x)$ and $< \mu_y, g >_{\mathcal{G}} = E_y g(y)$.

In dimension greater than 2, one operates on pairwise associations.

- Group 2 — mutual information, and methods based on a density estimation, implicit or not ; for instance method Radical (Miller and Fisher III, 2003) based on an estimator of differential entropy introduced initially by Vasicek (1976) ;

$$H_{nm} = n^{-1} \sum_{i=1}^{n} \log \frac{n}{2m} (X_{(i+m)} - X_{(i-m)})$$

where $X_{(i)}$ is the order statistic associated to an independent, identically distributed sample $X_1, \ldots, X_n$.

Another estimator of mutual information used in an ICA context by Boscolo et al. (2001)

is written

$$L(W) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{d} \log f_A^{\star\ell}(w_\ell x_i) + \log |\det W|,$$

where w_ℓ is line ℓ of candidate matrix W, and the authors estimate the marginal densities $f_A^{\star\ell}$ from a classical Gaussian kernel estimator that is substituted afterwards in the above formula.

Pham (2004), also proposed fast algorithms based on the estimation of the criterion $C(W) = \sum_{\ell=1}^{d} H(WX^\ell) - \log\det W$, using a spline kernel estimator of marginal entropies, and score functions related to the gradient of the netropy for minimization. The method also builds upon a discretized estimation of the density times its logarithm, $f \log f$.

- Group 3 — method of the matrix functional (Tsybakov et Samarov, 2001) ; this is one of the most accomplished methods from a theoretical viewpoint, based on a kernel estimator of a matrix functional that provides a direct algebraic solution instead of a contrast to minimize numerically.

  This method gives a consistent estimate of A^{-1} at rate $\sqrt{n}$ without the help of a contrast function to minimize, and using a functional based on the gradient of f. Taken apart some types of densities that are excluded by the method, it is optimal and gives a convergence rate in the estimation of A directly and not the one of an intermediary contrast to be minimized. In practice, the numerical complexity in n is $O(n^2)$.

2.2 Motivation

The distinction between classical methods and non parametric methods (that is to say here with a parameter but for regularization purpose) seems finally rather thin, since in one case one has a biased contrast estimated with some degree of error, and in the other case an exact contrast whose estimation is necessarily biased.

Classical criteria generally tend to mutual information but with no mention of any rate since one does not build upon a functional class defined by a regularity notion such as modulus of continuity (it is sometimes assumed that the density is "sufficiently differentiable"). Often the contrasts are based on moments of the distribution, which quantity does not depend on the form the density.

The motivation of essentially non parametric approaches is then to start with exact independence criteria, from which it is hoped to find the way to a more precise A inversion, and to make use of the extended statistical framework of the non parametric approach for their estimation.

Mutual information criteria shows that the problem of the estimation of an exact ICA contrast is inherently a problem of estimation of a functional of a density.

- Observation 1 ; in many cases the additional precision that is reached on paper is somewhat lowered by difficulties in the implementation or a high numerical complexity that forces to cede back something in the concrete computation.

  In method Radical, seeing that parameter m alone is not enough to take regularity into account, the authors add a second noise-like regularization, replacing each observation X_i by $\sum_{j=1}^{R} \epsilon_j$ where ϵ_j follows a normal law $N(X_i, \sigma^2 I_d)$, and R is to be chosen. If the consistency of Vasicek estimator is shown in the paper of Vasicek (1976) and in a subsequent paper from Song (2000), the second regularization changes the problem.

  Kernel methods build upon a regularization parameter defining the width of the kernel, in general Gaussian and isotropic. And this is again insufficient in practice. In kernel ICA, the generalized eigenvalue problem has no solution without an additional regularization whose statistical effect is not totally clear. The Hilbert-Schmidt method does not formally require a second regularization, but for the contrast to work one needs to fix a significance parameter γ that plays about the same role. From a general standpoint, the incomplete Choleski decomposition of kernel methods introduce additional parameters whose optimal values are not known and are thus set by empirical rules.

- Other general remark, simulations show that non parametric methods give good results in a wide range of situations, but one classical method or another is never very far behind in terms of Amari error (see examples also in the article of Bach and Jordan), and finally no method breaks away from the others in all situations.

In the matrix functional method, it has been noticed less precise results than classical methods in the case of multi-modal densities (Gómez Herrero, 2004, p. 15).

In the Hilbert-Schmidt method, authors claim excellent results compared to classical methods and kernel ICA, except in the case of few observations ($n = 250$).

On the test set in dimension 2 of Bach and Jordan, simulations of Radical show results on average better than FastICA, Jade, Infomax and kernel ICA.

In general, it seems that classical methods are effective in particular for unimodal densities where the notion of supergaussian and subgaussian have a visual signification.

- Observation 3 ; to present an interest in practice, an ICA method must have a moderate numerical complexity, and on this point classical methods have the advantage. It seems impossible to obtain at the same time a theoretical optimality and a low numerical complexity.

  Kernel methods considerably lower contrast complexity thanks to an incomplete Choleski decomposition, something from $O(d^3n^3)$ to $O(nd^2)$. But this entails introduction of new parameters whose optimal values are not known.

  Mutual information methods never estimate the multivariate density to cut down an otherwise lengthy procedure. On the other hand the possibility to know or test if the minimum is attained is lost.

  The matrix functional method is optimal in theory, but the form of the estimator implies a high numerical complexity in n, of the order of $O(n^2)$. Also, one has to use kernels with $d + 3$ vanishing moments, which is not an ideal situation in high dimension.

- Finally, some methods work under conditions, or do not work without the pre-whitening by PCA, which is a data dependent operation. No method takes anisotropy into account.

  For the matrix functional, in simulations run by Gómez Herrero (2004), estimation with window h uniform in all d dimensions is unstable without a pre-whitening, yet not formally planned by the method.

  It is important to note that pre-whitening entirely depends on the current sample, and that its use introduces from the very start a non linear break in the ICA problem. Besides, Cardoso (1994, 1999) showed that the separating power of ICA algorithms with pre-whitening admits a lower bound ; he also underlines that the minimum of an independence criteria does not exactly correspond to a total decorrelation, since independence is never exact in practice ; put in other words, if decorrelation is needed it must be enforced specifically.

In many methods, at most one Gaussian component is allowed for the contrasts to work, and also in the case of the matrix functional.

Identifiability of the ICA model is discussed in a paper of Comon (1994). For a source s with independent component and if $x = As$ also has independent components, $\operatorname{cov} x$ et $\operatorname{cov} s$ are diagonal, $\operatorname{cov} x = A \operatorname{cov} s \, {}^tA$, and A can be identified with any matrix of the form $A = (\operatorname{cov} x)^{\frac{1}{2}} Q (\operatorname{cov} s)^{-\frac{1}{2}}$, where Q is orthogonal. Matrix A of $x = As$ is identifiable up to a permutation or a normalization if and only if at most one component is Gaussian.

Gaussian laws are thus excluded from the field of ICA, but in practice it is not impossible to come across them, and a contrast that keeps a signification even with more than one Gaussian component can be useful, even if a post PCA minimization would not fully work.

▪

The motivation of this project was to study to what extent could a wavelet based contrast bring something in the resolution of the ICA problem, especially seeing the many existing methods.

Wavelet theory was designed to approximate possibly very irregular functions or surfaces and is an efficient tool used in data compression, imagery and signal processing. Its use in statistics also proved very powerful. Computational algorithms are fast ; for instance the projection on a wavelet basis at a given resolution and change of resolution are numerically stable operations and linear in n.

As opposed to a sinusoidal wave, a wavelet has finite duration ; the term thus refers to a frequency and also temporal (or spatial) localization. Spatial irregularities (multi-modes, discontinuities, mix) can be efficiently represented on a wavelet basis and the description of complicated functions generally hold in a small number of coefficients, often lower than in classical Fourier analysis.

Linear projection of a function f on a wavelet basis at resolution j operates a smoothing that boils down to set detail coefficients to zero. Another option consists of keeping only coefficients above some threshold ; the result is a non linear projection called thresholding, that provides adaptive procedures in the case where the regularity of f is not known .

The other interest of wavelets is their articulation with Besov spaces, approximation spaces par excellence, that generalize classical functional spaces, like Sobolev spaces, Hölder spaces, etc., having their own history.

The functional framework of Besov spaces allows to obtain convergence rates precisely linked to the regularity of the latent density and possibly adaptive conditions of use. One hopes this way to dispose of an exact framework that is not diluted later in a succession of tuning parameters to set according to generally empirical rules.

An a posteriori motivation of this project is to propose an alternative to mutual information, which remains considered as difficult to estimate (Bach and Jordan, 2002, p.3).

The mutual information estimator proposed by Vasicek (1976) is obtained after a change of variable $F(x) = p$, $0 \leq p \leq 1$, and deriving $F(F^{-1}(p)) = p$, differential entropy H is written

$$H = -\int_{-\infty}^{+\infty} [\log f(x)] f(x) dx = -\int_0^1 \log f(F^{-1}(p)) dp = \int_0^1 \log \frac{d}{dp} F^{-1}(p) dp.$$

Estimator H_{nm} is obtained with $\hat{F}^{-1}(p) = \inf\{x : \hat{F}_n(x) \geq p\}$ where $\hat{F}_n$ is the empirical estimator, and replacing the derivative by a first difference. ; that is to say $\frac{d}{dp}\hat{F}^{-1}(p) = \frac{n}{2m}(X_{(i+m)} - X_{(i-m)})$ for $(i-1)/n < p \leq i/n$.

Song (2000) noticed that direct estimation of H in function of F is not possible because of the differential operator. H being fundamentally a functional of a density, the technique of estimation is essentially the one of a density, hence the presence of parameter m in the role of the smoothing parameter.

An alternative to mutual information is given by the (squared) distance in L_2 norm between the density and the product of its marginals.

$$C(f_A) = \int (f_A - f_A^\star)^2,$$

where $f_A^\star$ is the product of the margins of f_A.

It can be noted that almost sure equality of f_A and $f_A^\star$, a consequence of $C(f_A) = 0$, does entail mutual independence of the components since the distribution functions F_A and $F_A^\star$ will be equal pointwise. The contrast $C(f_A)$ is then an exact mutual independence criteria. The similitude of form between the two quantities $\int f \log f$ and $\int f^2$ can also be noted, together with the simpler nature of the second quantity.

This factorization measure has been considered initially by Rosenblatt (1975) within the scope of an independence test of the components of a function in dimension 2, and with a kernel method. Also found in Rosenblatt's paper is a test procedure to decide about independence according to a criteria that tends to a normal law. The whole is moreover directly usable in the ICA framework, under a form which could possibly be modernized, and if one sticks to a pairwise component association.

Measure $C(f_A)$ is similar to the standard problem of the estimation of $\int f^2$, for which minimax rate is known under several regularity assumptions. For Hölder type regularity and for the L_2 loss, Bickel and Ritov (1988) have shown that minimax rate for the problem $\int f^2$ in dimension 1 is of parametric type, that is to say n^{-1}, for a regularity $s \geq 1/4$; the rate slows down to $n^{-8s/1+4s}$ for $s \leq 1/4$,.

In dimension d, if $s > d/4$, and for a Besov ball $\Theta = B_{s2q}(M)$ the constant is identified and the minimax rate is expressed by

$$\inf_{\hat{q}} \sup_{f \in \Theta} E_f \left(\hat{q} - \int f^2 \right)^2 = 4A(\Theta) n^{-1} (1 + o(1))$$

where $A(\Theta) = \sup_{f \in \Theta} \int f^2$ is a constant and $4A(\Theta)n^{-1}$ is the inverse of non parametric Fisher information (see Cai et Low, 2005).

One can find in the introduction of a paper of Kerkyacharian & Picard (1996) an example of projection estimator on a wavelet basis that attains minimax rate under Besov assumptions.

Measure $C(f_A)$ is also related to the case $\int (f-g)^2$, for f and g two unrelated functions. This case is developed in a paper from Butucea and Tribouley (2006) about a non parametric homogeneity test based on the L_2 loss.

The factorization measure $\int (f - f^\star)^2$ we take interest in has a slightly different form ; the functional is written $q(f) = \int \left(f - \int_{\star 1} f ... \int_{\star d} f\right)^2$, where $\int_{\star \ell} f$ is margin number ℓ.

The other a posteriori interest of this project is to be able to propose different estimators of the L_2 contrast, that are numerically stable, potentially optimal, or else sub-optimal but with a linear complexity in n, with a minimum of tuning parameters, and within a unique and clearly established statistical scope.

The procedure can also distinguish mutual independence from pairwise independence. This distinction is in theory of no interest in ICA where by hypothesis at most one Gaussian composes the source s ; indeed in this case, according to a result of Comon (1994), pairwise and mutual independence are equivalent. But it recovers a meaning if the model reintegrates noise or if one admits several Gaussian components, taken apart that in this last case A identifiability problems possibly show up ; the distinction could nevertheless provide some information in the case of several Gaussian components.

2.3 Theoretical results

The study focuses on the risk of different estimators of the L_2 factorization measure under a Besov regularity, and on questions related to the use of the wavelet contrast.

Linear mixing and Besov membership

With contrast minimization in scope, we checked that for an invertible matrix A, any transformed f_A belongs to the same Besov space as the original density f, in other words risk computation has a general impact, valid for the whole minimization procedure.

(see propositions 3.2 and 3.3).

Contrast in projection

Let the contrast in projection

$$C_j(f_A) = \int \left(P_j f_A - P_j f_A^\star\right)^2$$

where P_j is the projection operator on the space V_j of a multiresolution analysis of $L^2(\mathbb{R}^d)$.

- The gap between the wavelet contrast and the square of the L_2 norm of $f_A - f_A^\star$ representing the ideal contrast is given by

$$0 \leq \|f_A - f_A^\star\|_2^2 - C_j(f_A - f_A^\star) \leq C2^{-2js}\ ;$$

- The contrast in projection thus provides an approximate factorization criteria, but which is precisely linked to the regularity s

$$\begin{aligned} f \text{ is factorisable} &\implies C_j(f - f^\star) = 0 \\ C_j(f - f^\star) = 0 &\implies P_j f = P_j f^{\star 1} \dots P_j f^{\star d} \quad p.s. \end{aligned}$$

(see also propositions 3.1 and 4.6).

- The plug-in contrast can also be expressed on the detail coefficients (see lemma 5.15),

$$C_{j_1}(f) = C_{j_0}(f) + \sum_{j=j_0}^{j_1} \sum_k (\beta_{jk} - \beta_{jk^1} \dots \beta_{jk^d})^2.$$

where $k = (k^1, \dots, k^d)$ and $\beta_{jk^\ell} = \int f^{\star\ell}(x^\ell)\psi_{jk^\ell}(x^\ell)\,dx^\ell$ and $\beta_{jk} = \int f(x)\Psi_{jk}(x)\,dx$.

Estimators and rates

The table below summarizes the convergence rates that were obtained at optimal resolution, for different estimators detailed later. The first line is about estimators using distinct blocks coming from a split of the original sample ; the estimator appearing at line 4 is adaptive.

Convergence rates			
statistic	$2^{jd} < n$	$2^{jd} \geq n$	choice of resolution j
$\hat{\Delta}_j^2,\ \hat{G}_j^2,\ \hat{F}_j^2$	parametric	$n^{\frac{-8s}{4s+d}}$	$2^j = n^{\frac{2}{d+4s}}\ ,\ s \leq d/4\ ;\quad 2^{jd} \approx n, s \geq d/4$
$\hat{D}_j^2(\tilde{X})$	$n^{-1+\frac{1}{1+4s}}$	$n^{\frac{-8s}{4s+d}}$	idem if $s \leq d/4\ ;\quad 2^j = n^{\frac{1}{1+4s}}\ ,\ s \geq d/4$
$\hat{C}_j(\tilde{X})$	$n^{\frac{-4s}{4s+d}}$	inoperable	$2^j = n^{\frac{1}{4s+d}}$
$\tilde{\hat{C}}_{j_0,j_1}$	$n^{\frac{-2s}{2s+d}} \log n$	inoperable	$2^{j_1 d} = Cn(\log n)^{-1}\ ,\ j_0 = 0$

- An estimator of the factorization measure $C(f_A) = \|f_A - f_A^\star\|_2^2$ and more directly of the wavelet contrast $C_j(f_A - f_A^\star)$ is given by the statistic

$$\hat{C}_j(X_1, \dots, X_n) = \sum_k (\hat{\alpha}_{jk} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d})^2$$

where $\hat{\alpha}_{jk}$ (resp. $\hat{\alpha}_{jk^\ell}$) is the natural estimator of coefficient α_{jk} of f_A (resp. margin number ℓ of f_A) projection on the subspace V_j of a multiresolution analysis of $L_2(\mathbb{R}^d)$; that is to say

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

where X^ℓ is coordinate ℓ of $X \in \mathbb{R}^d$.

We show that the mean squared error of the statistic $\hat{C}_j$ for a Besov class B_{spq} is at most of the order of $n^{\frac{-4s}{4s+d}}$ (see proposition 4.10), that is to say better than the rate in density estimation $n^{\frac{-2s}{2s+d}}$ but less precise than the optimal rate in the estimation of a quadratic functional, n^{-1} for $s \geq \frac{d}{4}$ and $n^{\frac{-8s}{4s+d}}$ for $s \leq \frac{d}{4}$, to which $\int (f_A - f_A^\star)^2$ is related.

- A U-statistic estimator of the wavelet contrast is written

$$\hat{D}_j^2 = \hat{D}_j^2(X_1, \dots, X_n) = \frac{1}{A_n^{2d+2}} \sum_{i \in I_n^{2d+2}} h(X_{i^1}, \dots, X_{i^{2d+2}})$$

with

$$h = \sum_{k \in \mathbb{Z}^d} \left[\Phi_{jk}(X_{i^1}) - \varphi_{jk^1}(X_{i^2}^1) \dots \varphi_{jk^d}(X_{i^{d+1}}^d)\right] \left[\Phi_{jk}(X_{i^{d+2}}) - \varphi_{jk^1}(X_{i^{d+3}}^1) \dots \varphi_{jk^d}(X_{i^{2d+2}}^d)\right]$$

where φ is a scaling function, $\Phi_{jk}(x) = \varphi_{jk^1}(x^1) \dots \varphi_{jk^d}(x^d)$, $I_n^m = \left\{(i^1, \dots, i^m) \colon i^\ell \in \mathbb{N},\, 1 \leq i^\ell \leq n,\, i^{\ell_1} \neq i^{\ell_2} \text{ if } \ell_1 \neq \ell_2\right\}$ and $A_n^m = \frac{m!}{(n-m)!} = |I_n^m|$.

One can simplify the expression of the kernel h defining a function Λ_{jk} in the following way

$$\begin{aligned} \Lambda_{jk}(X_{i_1}, \dots, X_{i_d}) &= \varphi_{jk^1}(X_{i_1}^1) \dots \varphi_{jk^d}(X_{i_d}^d) \quad \forall i \in I_n^d \\ \Lambda_{jk}(X_i) = \Phi_{jk}(X_i) &= \varphi_{jk^1}(X_i^1) \dots \varphi_{jk^d}(X_i^d) \quad \forall i \in \{1 \dots, n\} \end{aligned}$$

where the second line is taken as a convention.

We then split up blocs of $2d+2$ variables X_i with distinct indices in 4 sections : For $i \in I_n^{2d+2}$, define the 4 dummy variables $Y_i = X_{i_1}$, $V_i = (X_{i_2}, \dots X_{i_{d+1}})$, $Z_i = X_{i_{d+2}}$, $T_i = (X_{i_{d+3}}, \dots X_{i_{2d+2}})$; that is to say Y_i and Z_i admit distribution P_{f_A}, V_i and T_i admit distribution $P_{f_A}^d$, and Y_i, V_i, Z_i, T_i are mutually independent under $P_{f_A}^n$.

It is then possible to express $\hat{D}_j^2$ under a form that recalls the order 2 U-statistic estimator of $\int (f-g)^2$ for f and g two unrelated functions defined on $\mathbb{R}^d$ (see Butucea and Tribouley 2006) :

$$\hat{D}_j^2 = \frac{1}{A_n^{2d+2}} \sum_{I_n^{2d+2}} \sum_k \left[\Lambda_{jk}(Y_i) - \Lambda_{jk}(V_i)\right] \left[\Lambda_{jk}(Z_i) - \Lambda_{jk}(T_i)\right].$$

In this way the additional complexity of $\hat{D}_j^2$ is entirely encapsulated in wide sections V_i and T_i, that may each possess up to d common coordinates with the second copy of the kernel in the case of the MSE calculus $E_{f_A}^n\left[\hat{D}_j^2\right]^2$.

Moreover $\hat{D}_j^2$ admits decomposition $\hat{D}_j^2 = \hat{U}_{j\alpha\alpha} - 2\hat{U}_{j\alpha\mu} + \hat{U}_{j\mu\mu}$, with

$$\begin{aligned}
\hat{U}_{j\alpha\alpha} &= \frac{1}{A_n^2}\sum_{I_n^2}\sum_k \Phi_{jk}(X_{i_1})\Phi_{jk}(X_{i_2}) \\
\hat{U}_{j\alpha\mu} &= \frac{1}{A_n^{d+1}}\sum_{I_n^{d+1}}\sum_k \Phi_{jk}(X_{i_0})\Lambda_{jk}(X_{i_1},\ldots,X_{i_d}) \\
\hat{U}_{j\mu\mu} &= \frac{1}{A_n^{2d}}\sum_{I_n^{2d}}\sum_k \Lambda_{jk}(X_{i_1},\ldots,X_{i_d})\Lambda_{jk}(X_{i_{d+1}},\ldots,X_{i_{2d}}),
\end{aligned}$$

each respectively being also an unbiased estimator of $C_{j\alpha\alpha}$, $C_{j\alpha\mu}$ and $C_{j\mu\mu}$, with

$$C_j = \sum_k \alpha_{jk}^2 - 2\sum_k \alpha_{jk}\lambda_{jk} + \sum_k \lambda_{jk}^2 \equiv C_{j\alpha\alpha} - 2C_{j\alpha\mu} + C_{j\mu\mu},$$

which shows that $\hat{C}_j^2$ is the V-statistic associated to $\hat{D}_j^2$.

Note that only $\hat{U}_{j\alpha\alpha}$ has a symmetric kernel ; for this reason most of the computations in this project do not rely on a symmetry of the kernel.

Risk computation can be done on $\hat{D}_j^2$ or else on each of the components $\hat{U}_{j..}$, but in the second case one looses a relatively incidental property of the risk saying that the bound is composed of a constant equal to zero at independence ; the bound is written indeed

$$C^\star 2^j n^{-1} + C 2^{jd} n^{-2},$$

where $C^\star = 0$ at independence (when $A = I$) (proposition 4.11). Such a constant $C^\star$ cannot be obtained with a bound on each component separately.

The presence of the factor 2^j in the expression above is the genesis of sub-optimality of $\hat{D}_j^2$ relative to a U-statistic estimator of $\int f^2$.

- We also used the symmetric kernel U-statistics

$$\begin{aligned}
\hat{B}_j^2(\{X_1,\ldots,X_n\}) &= \sum_k \frac{1}{A_n^2}\sum_{i\in I_n^2} \Phi_{jk}(X_{i^1})\Phi_{jk}(X_{i^2}) \\
\hat{B}_j^2(\{X_1^\ell,\ldots,X_n^\ell\}) &= \sum_{k^\ell} \frac{1}{A_n^2}\sum_{i\in I_n^2} \varphi_{jk^\ell}(X_{i^1}^\ell)\varphi_{jk^\ell}(X_{i^2}^\ell)
\end{aligned}$$

that places us in the scope of the estimation of $\int f^2$ treated in introduction of the paper of Kerkyacharian and Picard (1996), except that in our case estimation takes place in dimension d.

Estimating each plug-in $\hat{\alpha}_{jk}(\tilde{R}^0)$ and $\hat{\alpha}_{jk^\ell}(\tilde{R}^\ell)$, and U-statistic $\hat{B}_j^2(\tilde{R}^0)$, $\hat{B}_j^2(\tilde{R}^\ell)$, $\ell = 1, \dots, d$ on independent segments $\tilde{R}^0, \tilde{R}^1, \dots, \tilde{R}^d$ of the original sample, one can define the mixed plug-in estimator

$$\hat{F}_j^2(\tilde{R}^0, \tilde{R}^1, \dots, \tilde{R}^d) = \hat{B}_j^2(\tilde{R}^0) + \prod_{\ell=1}^{d} \hat{B}_j^2(\tilde{R}^\ell) - 2 \sum_{k \in \mathbb{Z}^d} \hat{\alpha}_{jk}(\tilde{R}^0) \hat{\alpha}_{jk^1}(\tilde{R}^1) \dots \hat{\alpha}_{jk^d}(\tilde{R}^d),$$

with the idea of estimating the quantity

$$\sum_k \alpha_{jk}^2 + \prod_{\ell=1}^{d} \Big(\sum_{k^\ell \in \mathbb{Z}} \alpha_{jk^\ell}^2 \Big) - 2 \sum_k \alpha_{jk} \alpha_{jk^1} \dots \alpha_{jk^d} = C_j^2$$

(see proposition 4.8).

- Second possibility : draw from the original sample $\{X_1, \dots, X_n\}$ an i.i.d. sample of $f_A^\star$, for instance $\{X_1^1 \dots X_d^d,\ X_{d+1}^1 \dots X_{2d}^d,\ \dots,\ X_{([n/d]-1)d+1}^1 \dots X_{[n/d]d}^d\}$, in the way of Hoeffding decomposition. Clearly, this method has the drawback of leaving out a large part of the information available.

  Nevertheless we can assume that we dispose of two independent samples, one for f_A labelled $\tilde{R}$ and one for $f_A^\star$ labelled $\tilde{S}$, with $\tilde{R}$ independent of $\tilde{S}$. Define then the estimators

$$\hat{G}_j^2(\tilde{R}, \tilde{S}) = \hat{B}_j^2(\tilde{R}) + \hat{B}_j^2(\tilde{S}) - 2 \sum_{k \in \mathbb{Z}^d} \hat{\alpha}_{jk}(\tilde{R}) \hat{\alpha}_{jk}(\tilde{S})$$

  and the two-sample U-statistic used in Butucea and Tribouley (2006)

$$\hat{\Delta}_j^2(\tilde{R}, \tilde{S}) = \frac{1}{A_n^2} \sum_{i \in I_n^2} \sum_{k \in \mathbb{Z}^d} \big[\Phi_{jk}(R_{i^1}) - \Phi_{jk}(S_{i^1})\big] \big[\Phi_{jk}(R_{i^2}) - \Phi_{jk}(S_{i^2})\big]$$

  assuming for simplification that both samples have same size n (that is different from the original n).

  $\hat{\Delta}_j^2(R, S)$ no more conceals large sections, is only of order 2 and is the exact replication of the statistic used by Butucea and Tribouley (2006) (in dimension d instead of 1) for estimating $\int (f-g)^2$ in the case where f and g are two unrelated functions.

  This procedure is optimal and its risk is $C^\star n^{-1} + 2^{jd} n^{-2}$, with $C^\star = 0$ at independence (when $A = I$) (see proposition 4.9).

- The thresholded estimator that has been studied is written $\tilde{\hat{C}}_{j_0, j_1} = \hat{C}_{j_0} + \tilde{\hat{T}}_{j_0 j_1}$ with

$$\tilde{\hat{T}}_{j_0 j_1} = \sum_{j=j_0}^{j_1} \sum_k (\tilde{\hat{\beta}}_{jk} - \tilde{\hat{\beta}}_{jk^1} \dots \tilde{\hat{\beta}}_{jk^d})^2$$

where $\tilde{\hat{\beta}}_{jk}$ is the thresholded substitute for $\hat{\beta}_{jk}$, and $j_0 = 0$ or some units.

We studied the hard thresholding case

$$\tilde{\hat{\beta}}_{jk} = \hat{\beta}_{jk} I\{\,|\hat{\beta}_{jk}| > t/2\} \text{ and } \tilde{\hat{\beta}}_{jk^\ell} = \hat{\beta}_{jk^\ell} I\{\,|\hat{\beta}_{jk^\ell}| > (t/2)^{\frac{1}{d}}\}, \text{ with } t \approx \sqrt{\frac{\log n}{n}}.$$

The procedure starts at the resolution of density estimation that is to say $2^{j_1 d} = Cn(\log n)^{-1}$.

The term by term thresholded contrast $\hat{C}_j$ admits a convergence rate of the order of $(\log n) n^{\frac{-2s}{2s+d}}$, that is to say slightly less than the minimax rate of density estimation problem in the adaptive case which is $[(\log n) n^{-1}]^{\frac{2s}{2s+d}}$.

In other words term by term thresholding costs more than a log compared to linear contrast with rate in $n^{\frac{-4s}{4s+d}}$ (see proposition 5.18).

This loss of efficiency is due to an inappropriate form of thresholding : it is a term by term thresholding of $\hat{C}_j$, whereas for the problem $\int f^2$, it is known that block thresholding is more effective (see for instance Cai and Low, 2005 p.4). But for the plug-in contrast it was inoperable because block thresholding typically starts at $2^{j_1 d} \approx n^2$ and stops at $2^{j_0 d} \approx n$, two resolutions beyond the technical maximum of $\hat{C}_j$: for $2^{jd} > n$, the algorithm is unstable, since one is never ensured of the convergence of the bound in $2^{jd} n^{-1}$ when n increases.

We did not implement the block thresholded U-statistic estimator of $\int (f_A - f_A^\star)^2$, because of a high complexity in the associated computation, and neither did we try to show the theoretical result since it should be closely related to the standard case $\int f^2$, but it is theoretically the best configuration to take advantage of thresholding.

For a Besov ball $\Theta = B_{s2q}(M)$, it is known in effect that minimax risk of the adaptive estimator $\int f^2$ is unchanged for $s > d/4$, that is to say it remains parametric, and is of the order of $n^{\frac{-8s}{d+4s}} (\log n)^{\frac{4s}{d+4s}}$ when $s \leq d/4$, that is to say is slowed by a log term (Cai et Low 2005). See also Gayraud and Tribouley (1999) for the white noise model.

As an estimator of the thresholded projection contrast itself $\tilde{C}_{j_0,\,j_1}$, defined by $\tilde{C}_{j_0,\,j_1} = C_{j_0} + \tilde{T}_{j_0 j_1}$ with

$$\tilde{T}_{j_0 j_1} = \sum_{j=j_0}^{j_1} \sum_k (\tilde{\beta}_{jk} - \tilde{\beta}_{jk^1} \ldots \tilde{\beta}_{jk^d})^2$$

where $\tilde{\beta}_{jk} = \beta_{jk} I\{\,|\beta_{jk}| > t/2\}$, the statistic $\tilde{\hat{C}}_{j_0,\,j_1}$ admits a rate with no log term. It is interesting en-soi since contrast estimation is only an intermediate objective, and a thresholded contrast omitting small contributions can be enough to give a satisfying inverse of A.

Filtering properties of the gradient and the hessian

In ICA, once estimated at a point f_{WA}, where $d \times d$ matrix W represents current estimation of A^{-1}, the contrast still has to be minimized in W. At this point, we return to the general procedure followed by any other contrast function used in other ICA methods, and particularly Stiefel minimization used by Bach et Jordan.

In the case of a twice differentiable wavelet, so for example a Daubechies at least $D4$, it is possible to give explicit formulas of the gradient and the hessian of the estimator $\hat{C}_j$ as a function of W or of $Y = WX$; moreover the formulation is filter aware which allows an easy change of resolution by discrete wavelet transformation (DWT). See part 6.

2.4 Practical results

Practical results presented in this part concern essentially the plug-in estimator.

As for kernel ICA, concrete estimation of the matrix B rests on a minimization procedure that can turn out well or poorly but is generally resolved in some iterations.

Plug-in contrast usage

Plug-in estimator $\hat{C}_j$ only works at resolution j such that $2^{jd} < n$; indeed, since the risk is in $2^{jd}n^{-1}$, there is no stability by increasing n on the set $\{j, d\colon 2^{jd} >= n\}$ (see propositions 3.4 and 4.10).

Theoretically the optimal choice of j depends upon regularity s and is set up at j_* such that $2^{j_*} = n^{\frac{1}{d+4s}}$ (see proposition 4.7 and see proposition 3.5 based on a suboptimal risk bound). But in practice a whole band of resolutions j works just as well as the optimal j. If regularity s is unknown, one can start at the smallest technical resolution (taking into account the Daubechies filter length in use), and increase j progressively if it seems that minimization does not occur (see simulations in part 3).

The alternative is to use the thresholded contrast, that automatically adapts to the right resolution ; in simulations that were carried out there is no clear-cut advantage of thresholding compared to an average choice of resolution between 0 and $\log n(d \log 2)^{-1}$ (see simulations in part 5 p. 164). It must be noted that an optimal thresholding procedure in a problem of the type $\int f^2$ is a global one. In global thresholding, one starts at a resolution 2^{jd} beyond n, which is not effective for the plug-in contrast.

The contrast is not in the wrong with more than one Gaussian component or if the mixing is not linear. In all cases one has an estimation of the L_2 norm of $f_A - f_A^\star$. Obviously if the mixing is not linear, the risk of the procedure is not the one announced since for instance Besov membership of f_A is maybe not valid anymore, and Stiefel minimization probably

becomes ineffective.

In this project we used pre-whitening by PCA and a minimization on the Stiefel manifold to be close to the model used by Bach et Jordan, but this is not a necessity. Other minimization methods are conceivable and the absence of pre-whitening has no consequence on the numerical stability of the contrast.

In practice, one single parameter must be set, resolution j.

Numerical stability

Based on convolutions and sums, the wavelet contrast is numerically stable and trivial from an algorithmic viewpoint, even if an efficient implementation independent of d is relatively difficult to obtain. Projection on the space V_j is equivalent to a histogram with window 2^{-j} for a $D2$ (Haar) or else approximately $2^{-(j \vee L)}$ with weighting for a $D2N$ (L is the length of dyadic approximation), with an overlapping band because of the overlapping of the wavelet supports ; the actual contrast computation is a sum of squared differences.

This operational simplicity is to be put in perspective with approaches based on linear algebra calculus (generalized eigenvalues, incomplete Choleski decomposition) whose implementation requires additional parameters, sometimes fixed by empirical rules, and whose numerical sensibility can be a problem in limiting cases (d or n high) or if conditions of use are not exactly met (mixing not exactly linear, more than one Gaussian, regularization poorly calibrated...).

In the case of the wavelet contrast, the realization of the computation is only a question of system resources. For jd high, it is possible to split up the allocation space necessary to the projection on V_j, together with the execution of the big contrast loop. Computation lends itself well to parallel or distributed programming.

Numerical complexity

The statistic $\hat{C}_j$ is a plug-in estimator ; its evaluation uses in the first place the complete estimation of the density f_A and of its margins ; which takes a computing time of the order of $O(n(2N-1)^d)$ where N is the order of the Daubechies wavelet, and n the number of observations ; we clarify here what was meant by $O(n)$.

A parameter n of the order of 10000 or 100000 does not saturate the procedure (see p. 101 and following), one thus has, if need be, the possibility to demix vast data sets, for instance in data mining applications (for diverse such advanced applications see for instance Bingham, 2003). In comparison, kernel methods generally reach a ceiling at $n = 5000$, at least in the simulations shown.

In the second place, the actual contrast is a simple function of the $2^{jd} + d2^j$ coefficients that

estimate density f_A and its margins ; the additional computing time is then in $O(2^{jd})$.

One can see here the main numerical drawback of the wavelet contrast in its total formulation — to be of exponential complexity in dimension d of the problem ; but this is by definition the cost of a condition that guarantees mutual independence of the components in full generality : d sets $B_1, \ldots, B_d$ are mutually independent if $P(B_1 \cap \ldots \cap B_d) = PB_1 \ldots PB_d$ for each of the 2^d choices of indices in $\{1, \ldots, d\}$.

Complexity in jd drops down to $O(d^2 2^{2j})$ if one concentrates on a pairwise independence, like in kernel ICA and in the Hilbert-Schmidt norm method (see pp. 71, 73) or in method Radical (see p. 75). Pairwise independence is in fact equivalent to mutual independence in the no noise model and with at most one Gaussian component (Comon, 1994).

U-statistic estimators usage

U-statistic estimators of C_j have complexity at minimum in $O(n^2(2N-1)^{2d})$, that is to say quadratic in n ; on the other hand the complexity in jd is probably lowered since the contrast can be computed by accumulation, without it being necessary to keep all projection in memory, but only a window whose width depends upon the length of the Daubechies filter.

We did not implement the U-statistic estimator, but in a configuration with pairwise minimization, since densities in dimension at most two are estimated, its use could prove to be competitive, not to mention that the U-statistic lends itself to block thresholding.

Haar and direct evaluation at dyadic rationals

The Haar wavelet $(D2N, N = 1)$ is not suitable for a gradient computation, empirical or theoretical ; the histogram variation (projection on the space V_j spanned by the Haar wavelet), in response to a small perturbation of the demixing matrix W is in general nonexistent or imperceptible (see in particular simulations p. 101 and following).

The problem also crops up at initialization by histogram at high resolution, often used in density estimation, with a string of filtering passes down to lower resolutions to follow.

To obtain an empirical gradient, we used a wavelet at least $D4$ and direct calculus $\frac{1}{n}\sum_i \phi_{jk}(X_i)$ with approximation of the values X_i at dyadic rationals from the algorithm explained by Nguyen et Strang (1996). Fortunately, and contrary maybe to a commonly held idea, that calculus is not a problem, if the values taken by the wavelet are preloaded in memory at a given precision. The corresponding computing time is by no means a bottleneck of the procedure in ICA, even less because we get rid of strings of filtering passes in dimension d (see part 6).

Separating power of the contrast

We give below an excerpt of simulations appearing in part 3 ; it is an average result of 100 runs in dimension 2 with 10000 observations, a Daubechies D4, $j = 3$ et $L = 10$ (dyadic precision) for different densities ; colomn `start` indicates Amari distance (scaled from 0 to 100) and the wavelet contrast on entry ; colomn `it` is the average number of iterations in Stiefel minimization. For some densities, after pre-whitening, the position is already close to the minimum, but the contrast still detects a dependency ; the procedure is not executed if the contrast or the empirical gradient are too close to zero, and this practically always corresponds to an Amari error less than 1.

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

It is worth noting that in the case of Gaussian components, minimization stalls.

The two examples below allow to visualize contrast and Amari error variation, at not necessarily the best resolution. Note that Haar (D2) based contrast gives a curve visually correct for the search of a minimum.

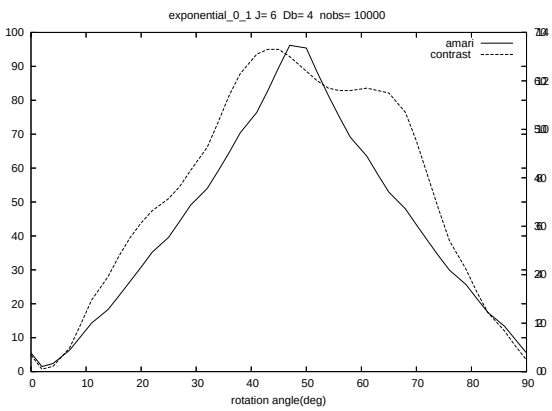

Fig.1. Exponential, D4, j=6, n=10000

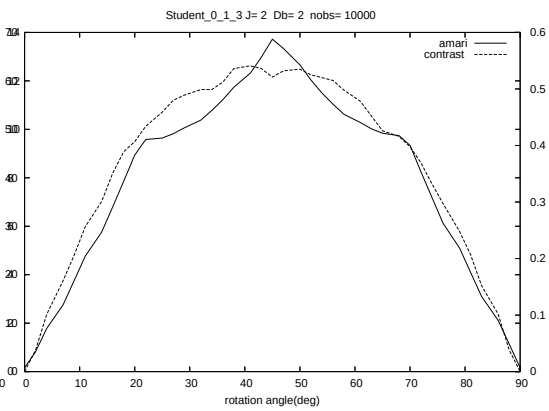

Fig.2. Student, D2, j=2, n=10000

See other examples in part 3.

2.5 Comparison with other methods

The method we propose provides a unique framework whose statistical properties are clearly established, and which leaves freedom of manoeuvre in practice. One has the choice of plug-in or U-statistic estimators ; it can be decided to restrict to pairwise, or triple wise independence, or mutual independence of a subset of components to cut down the entire procedure with higher complexity when its use is not justified (that is to say under normal conditions : at most one Gaussian component and no noise).

Contrary to mutual information criteria used on the margins only, the global minimum is known and there is at least in theory possibility of knowing if the zero was attained.

In this part we review some more precisely non parametric methods already presented above.

Kernel ICA

In Bach et Jordan (2002) kernel ICA , one makes use of a non linear decorrelation carried by a reproducing kernel Hilbert space (RKHS) $\mathcal{F}_\sigma$.

The space $\mathcal{F}_\sigma$ is generated by the isotropic Gaussian kernel $k(x,y) = e^{-\frac{1}{2\sigma^2}\|x-y\|^2}$, with $\emptyset = \mathcal{F}_0$ and $\mathcal{F}_\sigma$ increases to $L_2(\mathbb{R}^d)$, when $\sigma^2 \to +\infty$.

Parameter σ is then equivalent to resolution parameter j in wavelet ICA.

Let W be a potential inverse of matrix A solution of $x = As$, let $y_i = Wz_i$ where z_i is the PCA transform of x_i, observation i.

One makes use of estimator $\widehat{\mathrm{cov}}_{\mathcal{F}_\sigma}(Y^1, Y^2) = \frac{1}{n}\, {}^t\alpha^1 K_1 K_2 \alpha^2$, and

$$\hat{\rho_{\mathcal{F}}}(Y^1, Y^2) = \max_{\alpha^1,\alpha^2} \frac{{}^t\alpha^1 K_1 K_2 \alpha^2}{({}^t\alpha^1 K_1^2 \alpha^1)^{\frac{1}{2}}({}^t\alpha^2 K_1^2 \alpha^2)^{\frac{1}{2}}},$$

where K_1 and K_2 are $n \times n$ Gram matrices given by $K_\ell = \begin{pmatrix} k(y_1^\ell, y_1^\ell) & \dots & k(y_1^\ell, y_n^\ell) \\ \vdots & \ddots & \vdots \\ k(y_n^\ell, y_1^\ell) & \dots & k(y_n^\ell, y_n^\ell) \end{pmatrix}$.

The solution of this canonical correlation problem is the same as the one of the generalized eigenvalue problem

$$\begin{pmatrix} 0 & K_1K_2 \\ K_2K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha^1 \\ \alpha^2 \end{pmatrix} = \rho \begin{pmatrix} K_1^2 & 0 \\ 0 & K_2^2 \end{pmatrix} \begin{pmatrix} \alpha^1 \\ \alpha^2 \end{pmatrix}$$

In the general case one forms a super matrix of Gram matrices K_ℓ, centered beforehand, with dimension $(nd \times nd)$ labelled $\mathcal{K} : \mathcal{K} = (\, K_1 \quad \dots \quad K_d\,)\,(\, K_1 \quad \dots \quad K_d\,)'$.

The numerical complexity of the generalized solution is theoretically of the order of $O(d^3n^3)$, but the authors use an incomplete Choleski decomposition, which boils down to finding a solution of rank $m = h(\eta/n) \ll nd$, where η is a precision parameter and h a function depending on the (unknown) type of decrease at infinity of the underlying densities, that is to say $h(t) = O(\log t)$ for an exponential decrease and $h(t) = t^{1/d+\varepsilon}$ for a polynomial decrease. Numerical complexity is then brought down to $O(d^2n)$, but there are only empirical rules for the optimal choice of m according to the desired precision η.

In the case of the wavelet contrast, the pairwise criteria complexity is in $O(d^2n)+O(d^2 2^{2j}) = O(d^2n)$ for the plug-in statistic and $O(d^2n^2)$ for the U-statistic. There is no supplementary parameter to introduce, and the knowledge of the type of decrease at infinity is useless.

Another point worth noting lies in the necessity to introduce a second regularization parameter $\kappa < 1$ in the diagonal of $\mathcal{K}$, hence K_ℓ^2 is replaced by $(K_\ell + \kappa I)^2$.

Let $\mathcal{D}$ be the matrix whose diagonal is occupied by the square of the $(K_\ell + \kappa I)$; $\mathcal{D} = \mathrm{diag}((K_1+\kappa I)^2, \ldots, (K_d+\kappa I)^2)$. One actually searches for $\hat{\lambda}$, the smallest eigenvalue of equation $\mathcal{K}\alpha = \lambda\mathcal{D}\alpha$ where $\alpha \in R^{dn}$; $\hat{\lambda}$ is labelled first (kernel) canonical correlation.

The regularization parameter has the effect that the problem becomes numerically stable and must be chosen of the form $\kappa = \kappa_0 n$ to obtain a criteria $\hat{\rho}_{\mathcal{F}}$ independent of n, if neglecting the term of order κ^2. The criteria actually minimized is then written

$$\hat{\rho_{\mathcal{F}}}(Y^1, Y^2) = \max_{\alpha^1,\alpha^2} \frac{{}^t\alpha^1 K_1 K_2 \alpha^2}{(\mathrm{var}\, f^1(x^1) + 2\kappa n^{-1}\|f^1\|^2_{\mathcal{F}})^{\frac{1}{2}}(\mathrm{var}\, f^2(x^2) + 2\kappa n^{-1}\|f^2\|^2_{\mathcal{F}})^{\frac{1}{2}}},$$

since ${}^t\alpha(K+\kappa I)^2\alpha = {}^t\alpha K^2\alpha + 2\kappa\, {}^t\alpha K\alpha + \kappa^2\, {}^t\alpha\alpha$, et $n^{-1}\, {}^t\alpha K^2\alpha = \mathrm{var}\, f^1(x^1)$, $n^{-1}\, {}^t\alpha K\alpha = \|f^1\|^2_{\mathcal{F}}$.

In the general case, $d > 2$, $\hat{\lambda}(K_1, \ldots, K_d)$ is an estimate of $\lambda(x_1, \ldots, x_d)$ the $\mathcal{F}$-correlation between d variables. One has $0 \leq \lambda \leq 1$, and $\lambda = 1$ if and only if the variables $f_1(x_1), \ldots, f_d(x_d)$ are not correlated, where $f_i \in \mathcal{F}_\sigma$. It is about pairwise independence, since a correlation matrix gives information about pairs (see also Hilbert-Schmidt method p. 73 that generalizes kernel ICA) ; in usual ICA, it is equivalent to mutual independence (Comon, 1994).

As regards the wavelet contrast, pairwise and mutual independence can be distinguished, when there is a difference, with an increase in complexity.

In sum, kernel ICA needs a choice of four parameters, σ (equivalent to j for wavelets), m and η for incomplete Choleski decomposition, κ for numerical stability, to which also add up contrast minimization parameters.

The wavelet contrast needs only one parameter (apart from minimization parameters) ; resolution j.

In the paper of Bach and Jordan (2002), choices are $\eta = 10^{-3}\kappa$, $\kappa = 10^{-3}n$, $\sigma = 1$ for $n < 1000$ et $\sigma = 1/2$ pour $n > 1000$. There is no theoretical criteria that gives the optimal choices.

For resolution j of the wavelet contrast, there is an adaptive procedure by thresholding ; without thresholding, an average choice of j with $2^{jd} < n$ turns out to be quasi adaptive also in the set of simulations presented with the plug-in estimator. The optimal j is linked to the regularity of the underlying density s.

A very complete set of simulations in dimension 2 with $n = 256$ and 1024 observations show results on average better for kernel ICA as compared to Jade, Infomax, and FastICA.

Yet it looks like the method is in trouble with a high number of observations n, because of the size of the matrix $\mathcal{K}$ depending on n. In simulations presented in high dimension, $d = 8, 16$, the maximum number of observations stops at 4000 which looks relatively small. In these cases could the method achieve better results by increasing n ?

As a standard projection estimator, the computation of the wavelet contrast is not compromised with a high n (100000).

From a theoretical viewpoint, statistical properties of kernel ICA remain to be studied ; for instance there is no order of the convergence rate of the algorithm.

Hilbert-Schmidt norm method

Other approaches based on a reproducing kernel Hilbert space (RKHS) have been studied by Gretton et al. (Gretton et al. 2003, 2004) who present a general criteria of statistical dependence related to Hilbert-Schmidt norm.

Let $\mathcal{F}$ be a RKHS of functions of $\mathcal{X}$ in $\mathbb{R}$ associated with the positive definite kernel $k(.,.)\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and to characteristic function ϕ, $\phi(x) = k(x,.)$, that is to say in short

$$\forall x \in \mathcal{X},\ \exists \phi(x) \in \mathcal{F},\ <\phi(x), \phi(x')>_{\mathcal{F}} = k(x, x') \text{ et } <\phi(x), f>_{\mathcal{F}} = \delta_x f = f(x).$$

Let $(\mathcal{G},\, l(.,.),\, \psi)$ be a second RKHS of functions from $\mathcal{Y}$ to $\mathbb{R}$.

It is also assumed that $\mathcal{X}$ and $\mathcal{Y}$ are measured spaces $(\mathcal{X}, \Gamma, p_x)$ and $(\mathcal{Y}, \Lambda, p_y)$, where p_x p_y are probability measures. One considers also the joint entity $(\mathcal{X} \times \mathcal{Y}, \Gamma \otimes \Lambda, p_{xy})$.

The authors introduce the cross covariance operator

$$C_{xy} = E_{xy}\big[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)\big]$$

where E_{xy} is the expectation relative to the joint law p_{xy}, $\otimes$ is the tensorial product defined by $f \otimes g\colon h \in \mathcal{G} \mapsto f <g, h>_{\mathcal{G}} \in \mathcal{F}$ and μ_x, μ_y are given respectively by $<\mu_x, f>_{\mathcal{F}} = E_x f(x)$ and $<\mu_y, g>_{\mathcal{G}} = E_y g(y)$.

The dependency measure *HSIC* considered is the the square of the Hilbert-Schmidt norm of the linear operator C_{xy}

$$HSIC(p_{xy}, \mathcal{F}, \mathcal{G}) = \|C_{xy}\|^2_{HS}$$

with $\|C_{xy}\|_{HS}^2 = \sum_{i,j} <Cv_i, u_j>_{\mathcal{F}}^2$ and $\{u_j,\, j \in J\}$ orthonormal basis of $\mathcal{F}$, $\{v_i,\, i \in I\}$ orthonormal basis of $\mathcal{G}$.

The authors then show that $\|C_{xy}\|_{HS}^2 = 0$ if and only if x is independent of y, under the assumptions that $\mathcal{X}$ et $\mathcal{Y}$ are compact domains and that $\mathcal{F}$ are $\mathcal{G}$ sets of bounded functions. For the case $d > 2$, the criteria is generalized considering all the C_d^2 pairing of two variables, which boils down to a restriction to pairwise independence, less restrictive than mutual independence, but enough in the usual ICA context.

What was said in kernel ICA about the isotropic Gaussian kernel and incomplete Choleski decomposition remains valid for this method. One finds again the three parameters σ (equivalent to j for wavelets), η and m.

The criteria *HSIC* is expressed with kernels by

$$HSIC(p_{xy}, \mathcal{F}, \mathcal{G}) = E_{xx'yy'}\left[k(x,x')l(y,y')\right] + E_{xx'}k(x,x')E_{yy'}l(y,y') - 2E_{xy}\left[E_{x'}k(x,x')E_{y'}l(y,y')\right]$$

where $E_{xx'yy'}$ is the expectation relative to the law of two independent pairs (x,y), (x',y').

An estimator of HSIC is given by

$$HSIC(Z, \mathcal{F}, \mathcal{G}) = (n-1)^{-2} \text{ trace } KHLH,$$

with $H, K, L \in \mathbb{R}^{n^2}$, $H = I_n - n^{-1}(1_n\ {}^t1_n)$, $K_{ij} = k(x_i, x_j)$, $L_{ij} = l(y_i, y_j)$ and $Z = \{(x_1, y_1), \ldots (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$ is an independent, identically distributed sample of p_{xy}.

Authors show that $HSIC(Z)$ is biased at order n^{-1} and show also, using a large deviation bound for U-statistic that for all $\delta > 0$, for every law p_{xy} and for $n > 1$

$$P_Z\left[|HSIC(p_{xy}) - HSIC(Z)| \geq \sqrt{\frac{\log 6/\delta}{\alpha^2 n}} + \frac{C}{n}\right] < \delta$$

where $\alpha > 0,24$.

There is convergence in probability P_Z of the empirical criteria to the exact criteria with rate at best $n^{-1/2}$, and with great probability one can expect an increase of precision by increasing nn.

For comparison, the consistency of the wavelet contrast is shown in quadratic mean. Also the consistency does not take into account the bias between the RKHS and the overall Hilbert space on which the exact criteria is defined.

To test independence at significance level γ one defines

$$\Delta(Z) = \mathbb{I}\left\{HSIC(Z) > C\sqrt{n^{-1}\log 1/\gamma}\right\}$$

and

$$E_Z[\Delta(Z) = 1] = P_Z\left[|HSIC(Z)| > C\sqrt{n^{-1}\log 1/\gamma}\right] < \gamma.$$

The method does not formally need an additional regularization parameter κ as for kernel ICA. On the other hand the significance level γ must be set, which can be seen as a regularization parameter, together with the constant C.

Resolution uses minimization of the criteria HSIC as in kernel ICA. As for kernel ICA, HSIC is approximated by an incomplete Choleski decomposition.

Wavelet ICA can also make use of a statistical test of independence and works with or without such a test, whereas in the Hilbert-Schmidt method, the computation of the criteria is inseparable from the statistical test.

The authors announce excellent results compared to classical methods and to kernel ICA, except in the case of few observations ($n = 250$).

RADICAL

Miller and Fisher III (2003) proposed a ICA contrast based on mutual information using an estimator initially introduced by Vasicek (1976).

Let $Y_{(1)}, \dots Y_{(n)}$ be the order statistic of an i.i.d. sample of $Y = WX = WAS$. Let F_A be the distribution function of Y.

As indicated above, if $Y = BS$ is a random variable in dimension d, $I(Y) = \sum_i H(Y^i) - H(S) - \log|\det B|$, the log term is zero in the case of B orthogonal. One tries to find $W^* = \text{ argmin }_W \left[H(Y^1) + \ldots + H(Y^d)\right]$, where Y^ℓ is component ℓ of Y.

The proposed estimator is written

$$\hat{H}(X_1, \dots X_n) = \frac{m}{n-1} \sum_{i=0}^{\frac{n-1}{m}-1} \log\left[\frac{n+1}{m}\left(X_{(mi+m+1)} - X_{(mi+1)}\right)\right]$$

where $X_{(i)}$ is the order statistic associated to the sample.

The consistency of this estimator is showed in the paper of Vasicek (1976) and in a later paper from Song (2000).

Parameter m plays the role of a regularization parameter, and it must be ensured that $m \to +\infty$ and $m/n \to 0$. Authors took $m = \sqrt{n}$.

In the method proposed by Miller and Fisher, minimization of $\hat{H}(W)$ is carried out by reviewing all C_d^2 free plans of $\mathbb{R}^d$, since a minimization in dimension 2 is equivalent to finding the minimum of a θ between 0 and $\pi/2$. One then substitutes to a minimization on the Stiefel manifold $S(n, d)$, $d(d-1)/2$ minimizations in $S(n, 2)$. It is equivalent to minimizing pairwise dependencies between components of X, and concretely one applies Jacobi rotations to matrix W to select the plan in which the contrast is to be minimized. The authors propose a systematic calculus along a grid with $K = 150$ points to avoid local minimums problems.

The method thus provides only a pairwise independence criteria, easier to obtain than mutual independence but enough in the usual ICA context.

The same minimization procedure can be used with the wavelet contrast in place of mutual information criteria ; in doing so, one obtains on top a criteria that lends itself to the construction of an independence test. Also a Haar based contrast (numerical complexity Cn with $C = 1$) is usable if minimizing without a gradient.

Seeing that regularization parameter m alone is not enough, the authors add a second regularization by noising, replacing each observation X_i by $\sum_{j=1}^{R} \epsilon_j$ where ϵ_j follows a normal law $N(X_i, \sigma^2 I_d)$.

Choices in simulations are $\sigma = 0,35$ for $n < 1000$ and $\sigma = 0,175$ for $n > 1000$, $K = 150$, $R = 30$, $m = \sqrt{n}$. The algorithm complexity is in $O(KRN \log RN)$, the log term resulting from the necessity to sort the observations (typically complexity $n \log n$).

The complexity is then higher than for the wavelet contrast in its plug-in version.

On the test set of Bach and Jordan in dimension 2, simulations show results on average better than FastICA, Jade, Infomax and kernel ICA.

Matrix functional

Tsybakov and Samarov (2002) proposed a method of estimation of directions b_j, where $B = (b_1 \quad \dots \quad b_d)$ based on non parametric estimation of a matrix functional using the gradient of f_A.

The method allows to estimate matrix $B = A^{-1}$ at parametric rate and also the density f, whose components are independent at a dimension 1 rate. The method is based on kernels having at least $d + 3$ vanishing moments, and the choice of an appropriate window. A characteristic of this method lies in the fact that estimation is carried out directly, without the need to minimize a contrast. It means that one obtains values algebraically, that have to be found numerically in other methods.

In wavelet ICA, estimation of the underlying densities is also given, from the plug-in estimator, although we do not specifically give the risk of this estimation, for which the optimal j is different from the j used in estimating the wavelet contrast. In the same way, also given is an estimation of the signal at all steps of the demixing process. The U-statistic estimator also attains parametric rate, but only in the case $s \geq d/4$. If using pairwise independence criteria, it is equivalent to say that estimation attains parametric rate as soon as $s > 1/2$.

The authors consider the functional

$$T(f) = E_f\big[\nabla f(x)\ {}^t\nabla f(x)\big] = \sum_{j=1}^{d}\sum_{k=1}^{d} c_{jk} b_j\ {}^t b_k,$$

where ∇f is the gradient of f,

$$c_{jk} = (\det B)^2 E\big[\prod_{i\neq j} p_i(\ {}^t x b_i) \prod_{m\neq k} p_m(\ {}^t x b_m) p_j'(\ {}^t x b_j) p_k'(\ {}^t x b_k)\big],$$

and $B = (\,b_1 \quad \dots \quad b_d\,)$.

With condition $\int (f^\ell)'(x)(f^\ell)^2 = 0$, $\ell = 1,\dots,d$, the functional can be simplified to $T(f) = \sum_{j=1}^{d} c_{jj} b_j\ {}^t b_j = BC\ {}^t B$, $C = \mathrm{diag}(c_{jj})$.

T is positive definite, $B\ {}^t T^{-1} B = C^{-1}$ et $B\ {}^t \mathrm{var}(X) B = D$; which implies that $P\ {}^t \Sigma P = \Lambda$ and $P\ {}^t T^{-1} P = I$ with $\Sigma = \mathrm{var}(X)$, $P = BC^{\frac{1}{2}}$ and $\Lambda = C^{\frac{1}{2}} D C^{\frac{1}{2}}$.

A matrix algebra result allows to write Λ as the diagonal matrix of the eigenvalues of $T\Sigma$ and columns of P as eigenvectors of $T\Sigma$.

One searches for the vectors p_j, $j = 1,\dots,d$, solution of $T\Sigma p_j = \lambda_j p_j$, which is equivalent to search for vectors q_j solution of $T^{\frac{1}{2}}\Sigma T^{\frac{1}{2}} q_j = \lambda_j q_j$, where $q_j = T^{-\frac{1}{2}} p_j$ and q_j pairwise orthogonal.

One then estimate $W = T^{\frac{1}{2}}\Sigma T^{\frac{1}{2}}$, then take PCA transformation to estimate the q_j, then the $p_j = T^{\frac{1}{2}} q_j$, then the $b_j = c_{jj}^{-\frac{1}{2}} p_j = p_j \|p_j\|^{-1}$.

The Σ estimator is the classical variance empirical estimator, and an estimator of T is given by

$$\hat{T} = \frac{1}{n}\sum_{i=1}^{n} \nabla \hat{p}_{-i}(X_i)\ {}^t\nabla \hat{p}_{-i}(X_i)$$

where component l of $\nabla \hat{p}_{-i}(X_i)$ is given by

$$\frac{d\hat{p}_{-i}(X_i)}{dx^l} = \frac{1}{(n-1)h^{d+1}} \sum_{j=1,\, j\neq i}^{n} K_1\Big(\frac{X_j^l - X_i^l}{h}\Big) \prod_{k=1,\, k\neq l}^{d} K\Big(\frac{X_j^l - X_i^l}{h}\Big)$$

With Hölder type regularity for the components of f, conditions on the windows h, $nh^{2d+4} \to \infty$ and $nh^{2b-2} \to 0$, $b > d+3$, and b vanishing moments for the kernels K, and K_1, and $E\|X\|^4 < \infty$, the authors show that estimation of directions b_j of matrix B is consistent at rate $\sqrt{n}$.

As for the wavelet contrast, the number of vanishing moments of the wavelet is independent of d, a D4 proved enough in simulations shown. Also consistency is shown in quadratic mean. There is no restriction on the form of the densities.

The matrix functional method also allow to estimate density f from the estimator

$$\hat{f}(x) = \det \hat{B} \prod_{j=1}^{d} \frac{1}{n\tilde{h}_j} \sum_{i=1}^{n} \tilde{K}\Big(\frac{{}^tX_i\hat{b}_j - {}^tx\hat{b}_j}{\tilde{h}_j} \Big)$$

where $\tilde{h}_j \approx n^{\frac{-1}{2s_j+1}}$ and $\tilde{K}$ admits $s = \min s_j$ vanishing moments. This estimation is consistent at rate $n^{\frac{-s}{2s+1}}$, that is the optimal density estimation rate in dimension 1. But the number of vanishing moments of the kernel depends on unknown parameter s.

The method of Tsybakov and Samarov is optimal from a theoretical point of view, has the advantage of providing the consistency of a direct estimator of A, and gave good results in simulations (Gómez Herrero, 2004) yet also has some drawbacks.

Condition $\int (f^\ell)'(x)(f^\ell)^2 = 0$ verified for any density with support in $\mathbb{R}$ or symmetrical density with support in an interval, and condition $E\|X\|^4 < \infty$, de facto exclude densities not meeting the criteria (for instance beta, Cauchy, χ^2, Pareto, triangular,...)

It has been noted less precise results than classical methods in the case of multimodal densities (Gómez Herrero, 2004, p. 15).

The form of the estimator $\hat{T}$ implies a high numerical complexity, at least of the order of $O(\kappa d n^2)$, where κ is a multiplicative factor depending on the kernel with $d+3$ vanishing moments.

The optimal choice of the window h depends on $b \leq \min s_j$, where s_j are the regularities of the unknown densities (just like the j in wavelet ICA, but there are adaptive estimators available).

In simulations run by Gómez Herrero (2004), the estimation with h uniform in the d dimensions is unstable without a PCA pre-whitening, yet not formally specified by the method, but in fact widely used in practice of ICA.

The method gives relatively undetermined results with several Gaussian components that introduce multiple eigenvalues in the spectral decomposition of T. Even if Gaussian laws are out of the field of ICA, in practice it is possible to see some.

In comparison, being a very general independent criteria, the wavelet contrast is not compromised in the presence of Gaussian components, even if the minimization is, which boils down to saying that the consistency of the wavelet method only bears upon the accuracy of the contrast estimation, not A estimation.

2.6 Perspectives

At the end of this project, we present a certain number of possible extensions of computing, statistical or even purely practical nature.

Practical extensions

as regards to computing, the numerical complexity of the minimization could be improved.

Simulations in dimension 2 show that the Haar wavelet produces a visually elliptic contrast curve, often with no noisy local minima (see p. 106). The problem for Haar comes from the gradient computation. Minimization methods involving no gradient, could restore the utility of this contrast. In that case the complexity of the projection on V_j changes from $O(n(2N-1)^d)$ à $O(n)$ ($N=1$ for Haar).

For instance the method tested by Miller et Fisher (2003) and Gretton et al. (2004), consisting in searching for the minimum in each of the C_d^2 free plans of $\mathbb{R}^d$, applying Jacobi rotations to select a particular plan. A search in each plan is equivalent to the case $d=2$, where the problem is to find the minimum in θ of a function on $\mathbb{R}$, for $\theta \in [0, \pi/2]$. To do so, the simplest could be to try out all points from 0 to $\pi/2$ along a grid, or to use bisection type methods if the curve does not possess local minima (as it is the case for the wavelet contrast in the examples p. 106).

It is conceivable to implement the U-statistic estimator that would have a numerical complexity in n^2, and on which block thresholding can also be performed. With few observations as it is often the case in ICA simulations, that would allow to propose an adaptive criteria, potentially more precise than the plug-in, in its linear or adaptive version.

Taking into account anisotropy

If the components of f have very different regularities, the anisotropic case could be studied.

In this project we restricted to a split of space in cubes with sides 2^{-j}, but in the anisotropic case, the method can be modified by considering rectangles with j different according to the dimensions. As far as implementation is concerned, that does not constitute a very difficult extension. From a theoretical point of view, one has to consider anisotropic Besov spaces.

Plug-in type contrast reaching a parametric rate

It could be interesting to study more general contrasts of the form $\sum_k |\hat{\alpha}_{jk} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d}|^p$ or else $\sup_k |\hat{\alpha}_{jk} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d}|$; this last contrast could avoid storing an array of size 2^{jd} since only the current maximum of $|\hat{\alpha}_{jk} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d}|$ is needed in memory.

The L^2 case can indeed be extended in the following way : From Meyer's lemma (lemma 4.8) one obtains,

$$\int |f|^p \leq 2^{p-1}\big(\|P_j f\|_p^p + 2^{-pjs}\big) \leq C 2^{jd(\frac{p}{2}-1)} \sum_k |\alpha_{jk}|^p + C 2^{-pjs}$$

The generalized wavelet ICA contrast is then defined as

$$C_j^p(f_A - f_A^\star) = \sum_{k\in\mathbb{Z}^d} \left| \int (f_A - f_A^\star)\Phi_{jk} \right|^p$$

and the normalized contrast as $C_j^{p\prime}(f_A - f_A^\star) = 2^{\frac{jd}{2}(p-2)} C_j^p(f_A - f_A^\star)$.

Note that for $f \in B_{spq}$, if the generalized contrast is zero, $\int |f_A - f_A^\star|^p \leq C2^{-pjs}$ (and in all cases when the functional C_j^p is equal to zero, so is the generalized contrast).

For the problem of estimating a non quadratic functional, Kerkyacharian and Picard (1996) used an exact development of $(P_j f + f - P_j f)^3$, where P_j is the projection operator associated to the Haar wavelet.

In our case, except for $p = 2$, the generalized contrast C_j^p defined above only uses a convexity inequality. This probably suboptimal approach for estimating a non quadratic functional of a density nevertheless provides a family of contrasts applicable to any Daubechies wavelet.

In the case $p = \infty$, one can define $C_j^\infty(f_A - f_A^\star) = \sup_{k\in\mathbb{Z}^d} \left|\int (f_A - f_A^\star)\Phi_{jk}\right|$ and one has in the same way

$$\begin{aligned}\sup_x |f| &\leq \sup_x |P_j f| + 2^{-js} \\ &\leq C2^{\frac{-jd}{2}} \sup_k |\alpha_{jk}| + C2^{-js}\end{aligned}$$

that is to say if the contrast C_j^∞ is zero, $\sup_x |f| \leq C2^{-js}$.

The first investigations seem to indicate that p above 2 improves the convergence rate of the plug-in contrast. One would then have a result looking like the following :

Let $p \in \mathbb{N}^*$. Let $X_1, \ldots, X_n$ be an i.i.d. sample of f, a compactly supported density defined on $\mathbb{R}^d$. Assume that φ is a Daubechies wavelet $D2N$. Let $\Omega_n^m = \left\{(i^1, \ldots, i^m) : i^\ell \in \mathbb{N},\, 1 \leq i^\ell \leq n\right\}$. Let $I_n^m = \left\{i \in \Omega_n^m : \ell_1 \neq \ell_2 \Rightarrow i^{\ell_1} \neq i^{\ell_2}\right\}$

For $i \in \Omega_n^p$, let $h_i = \sum_k \Phi_{jk}(X_{i^1}) \ldots \Phi_{jk}(X_{i^p})$. Let $\theta = \sum_k \alpha_{jk}^p$. Let $\hat{B}_j^p = (A_n^p)^{-1} \sum_{i\in I_n^p} h_i$ be the U-statistic estimator of θ. Let $\hat{H}_j^p = n^{-p} \sum_{i\in\Omega_n^p} h_i$ be the plug-in estimator of θ.

On the set $\{2^{jd} < n^2\}$,

$$\begin{aligned}E_{f_A}^n |\hat{B}_{j\alpha^p} - \theta|^2 &\leq C2^{jd(2-p)} n^{-1} \mathbb{I}\left\{2^{jd} < n\right\} + C2^{jd} n^{-p} \mathbb{I}\left\{2^{jd} > n\right\} \\ E_{f_A}^n |\hat{H}_{j\alpha^p} - \theta|^2 &\leq C2^{jd} 2^{jd(2-p)} n^{-1} \mathbb{I}\left\{2^{jd} < n\right\} + C2^{jdp} n^{1-2p} \mathbb{I}\left\{2^{jd} > n\right\}\end{aligned}$$

In other words the case $p = 2$ was in fact not enough for the plug-in contrast $(\hat{H}_{j\alpha^p})$, and the choice $p \geq 3$ would ensure parametric rate on the set $2^{jd} < n$.

A more complete study of the generalized case would then allow to rule on this hypothesis and to envisage a strategy of choice of p.

Test of Independence

It is conceivable to build a statistical test to dispose of a decision rule about independence or not in a real situation where Amari criteria is not available.

Butucea et Tribouley (2006) have already proposed a homogeneity test based on the criteria $\int (f-g)^2$, for f and g two functions on $\mathbb{R}$. To test hypothesis

$$H_0 \colon f = g$$

against the alternative

$$H_1 \colon f, g \in V \cap \{f, g \colon \|f-g\|_2^2 \geq C r_{n,m}\}$$

where V is a functional space necessary for estimation, and $r_{n,m}$ a sequence tending to zero when $n \wedge m$ tends to infinity. The authors consider the two sample statistic

$$T_j = [(n \wedge m)(n \wedge m - 1)]^{-1} \sum_{i_1, i_2 = 1}^{n \wedge m} \sum_k \big(\varphi_{jk}(X_{i_1}) - \varphi_{jk}(Y_{i_1})\big)\big(\varphi_{jk}(X_{i_2}) - \varphi_{jk}(Y_{i_2})\big)$$

and a test statistic D,

$$D = D_j = \begin{cases} 0 & \text{si } |T_j| \leq t_{j,n,m} \\ 1 & \text{si } |T_j| > t_{j,n,m} \end{cases}$$

where j and $t_{j,n,m}$ are to be chosen so that to obtain the best test D_j in a family $\{D_j, j \in J\}$, or else in the adaptive case, $D = \max_{j \in J} D_j$.

The authors obtain that the minimax rate of separation of the hypothesis H_0/H_1 at level γ in dimension 1 is of the order of $n^{\frac{-4s}{4s+1}}$ for an appropriate choice of resolution j and of the level of the test t_j.

The procedure should apply mutatis mutandis to the case $g = f^\star$, in view of what was already done in this project (see in particular part p. 26 and following for the similarities between the statistic used by Butucea and Tribouley and our).

Moreover the independence test found in the article of Rosenblatt (1975) should also apply.

Considering in dimension 2 the empirical factorization measure

$$S_n = \int \big[f_n(x) - g_n(x^1) h_n(x^2)\big]^2 dx,$$

where $f_n(x) = \frac{1}{nh^2} \sum_{j=1}^n w\left(\frac{x - X_i}{h}\right)$, $g_n(x^1) = \frac{1}{nh} \sum_{j=1}^n w\left(\frac{x^1 - X_i^1}{h}\right)$ (idem for h) and $w(x) = w^1(x^1) w^2(x^2)$ is a positive definite kernel with 2 vanishing moments, and with other supplementary conditions, Rosenblatt finds that the quantity

$$h^{-1}\Big[- A(h) + nh^2 \int \big[f_n - g_n h_n\big]^2\Big]$$

converges to a normal law with zero mean and variance $2w^{(4)} \int f^2$ when $n \to \infty, nh^2 \to \infty$ and $h = o(n^{-\frac{1}{5}})$, where $A(h)$ is a quantity in $O(1)$.

In his paper, Rosenblatt wonders about the advantages of a test based on densities rather than empirical distributions, and indicates that compared to some tests on distributions by Blum et al. (1961) the density based test is possibly less powerful but easier to implement because of the convergence to a normal law instead of a law with unknown properties in the other case.

Also a spacing estimate of the distribution function that would be based on the order statistic (see the case of estimator Radical p. 75) has higher complexity in n, of the order of $n \log n$.

Adaptation

Beside the study of a global thresholding on a U-statistic estimator of the L_2 contrast, one could envisage the application of alternative adaptive procedures like Lepski's method (random smoothing), together with aggregation methods or model selection.

Exact gradient of the L_2 contrast

In this document we studied the gradient of the projected contrast estimator. We could also try to determine the gradient and the hessien of the exact L_2 contrast. This means we probably come across quantities that are used in the method of the matrix functional of Tsybakov and Samarov. We could then try to estimate these derived quantities with a non parametric wavelet based method.

Non linear mixing

The wavelet contrast is valid even if the mixing is not linear.

For a random variable $X = F(S)$ where $F: U \subset \mathbb{R}^d \mapsto V \subset \mathbb{R}^d$ is no more linear but invertible, the factorization measure is written in the same way , $\int \|f_F - f_F^*\|^2$ and there is an estimation of it through contrast C_j. Minimization would be carried out no longer on the Stiefel manifold but on another set depending on F.

ACI without pre-whitening

One could consider performing a minimization on the Stiefel manifold without any preliminary PCA transformation, that has some annoying side effects.

For an invertible matrix A, there exists by QR factorisation, an orthogonal matrix Q such that $QA = T$ is upper triangular. Q and T are moreover unique is the diagonal of T is strictly positive.

For such a matrix T the inversion is immediate.

Let $x = As$; so there exists a unique orthogonal Q such that $y = Qx = QAs = Ts$. The normalisation $\operatorname{var} s = I_d$ implies that $\operatorname{cov} y = TT'$. The symmetric matrix TT' thus has for general term $\operatorname{cov}(y^k, y^l) = a_{kl} = \sum_{i \geq l} t_{ki} t_{li}$, all quanties estimables from the observations and Q.

Let W be an orthogonal matrix, if $W = Q$, then T^{-1} is estimable from the observed covariance matrix.

One could then consider the contrast function

$$C \colon Q \mapsto C(\hat{T}^{-1} Qx).$$

In dimension 2, $T = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$, $TT' = \begin{pmatrix} a^2 + b^2 & bc \\ bc & c^2 \end{pmatrix}$, $T^{-1} = \begin{pmatrix} a^{-1} & -b(ac)^{-1} \\ 0 & c^{-1} \end{pmatrix}$.

that is to say,

$$c^2 = \operatorname{var} y_2,\ a^2 = \operatorname{var} y_1 - \operatorname{cov}(y_1, y_2)^2 (\operatorname{var} y_2)^{-1},\ b^2 = \operatorname{cov}(y_1, y_2)^2 (\operatorname{var} y_2)^{-1}.$$

and

$$\begin{aligned} s_2 &= c^{-1} y_2 \\ s_1 &= a^{-1}[y_1 - bc^{-1} y_2] \end{aligned}$$

2.7 Organization of the document

Parts 3 to 5 correspond to three articles submitted to journals with reading committee. Part 6 contains all elements used in the implementation.

- In part 3, we show experimentally that the method works, that the Haar wavelet is suitable if using no gradient, that curves in dimension 2 are generally elliptic if at a suitable resolution, that there exists a whole band of suitable resolution about optimal j, that the wavelet contrast has a very thin discriminating power (half-degree rotations are detected), and that computing times are competitive in small dimensions.

  From a theoretical point of view, we show that mixing by invertible A conserves Besov membership of the original densities. We link criteria C_j to a measure of independence of the components of X, and we show, deliberately adopting a simplified point of view, that mean squared error (MSE) of the wavelet contrast is at most of the order of $n^{\frac{-2s}{2s+d}}$, that is to say minimax rate of density estimation.

- In part 4, we develop the relation between the problem $\int f^2$ and the ICA problem. We show that the convergence rate of the contrast MSE is actually at least of the order of $Cn^{\frac{-4s}{4s+d}}$ and we show that U-statistic estimators do attain a parametric rate for regularities $s \geq d/4$, as in the problem $\int f^2$.

The demonstration uses combinatorial lemmas that simplify computations otherwise unachievable. We show that a U-statistic estimator on the full sample is slightly suboptimal ; and consequently we study splitting strategies in an attempt to recover optimality.

- In part 5, we study the risk of a thresholded estimator of the wavelet contrast. Simulations complete the theoretical study.

- In part 6, we regrouped all concerning implementation issues, which includes information on the Stiefel manifold and on the minimization method on this manifold, filter aware formulations of the gradient and the hessian, other practical concerns and the commented code of key parts of the program.

Notations et conventions

In the following we employ indifferently the term "wavelet contrast" to designate $C_j \in \mathbb{R}^+$, an estimator $\hat{C}_j(X_1, \ldots, X_n)$, where X_i is a random variable with density f_A, or else function $y_1, \ldots y_n \in \mathbb{R}^{d\times n} \mapsto \hat{C}_j(y_1, \ldots, y_n)$ or else function $W \in \mathbb{R}^{d\times d} \mapsto \hat{C}_j(Wx_1, \ldots, Wx_n)$ where $x_i \in \mathbb{R}^d$ is observation i.

In parts 4 and 5, C_j is noted C_j^2, and idem for estimator $\hat{C}_j$, in reference to the squared norm of $P_j f$.

In general, superscript designates coordinates of multidimensional entities, whereas subscript designates elements of the same set.

Other notations

C_n^p, A_n^p	$n!/p!(n-p)!$, $n!/(n-p)!$
C_j	wavelet contrast except part 4
C_j^2	wavelet contrast in part 4
$\tilde{X}$, X_i, X_i^ℓ	sample $X_1 \ldots, X_n$, variable number i, component ℓ of variable i
f_A	$\lvert\det A^{-1}\rvert\, f(A^{-1}x)$
$E_{f_A}^n$	expectation relative to the joint law of the couple $X_1, \ldots, X_n$ with density $f_A^{\otimes n}$
$f^\star$, $f^{\star\ell}$	product of margins of f, margin number ℓ of f
$a \vee b$, $a \wedge b$	$\max(a, b)$, $\min(a, b)$
$\mathbb{I}\{R\}$, $\mathbb{I}_B$	indicator functions
Ω_n^m, I_n^m	$\{(i^1, \ldots, i^m) : i^\ell \in \mathbb{N},\, 1 \le i^\ell \le n\}$, $\{i \in \Omega_n^m : \ell_1 \neq \ell_2 \Rightarrow i^{\ell_1} \neq i^{\ell_2}\}$
Φ_{jk}, Ψ_{jk}, Λ_{jk}	$\varphi_{jk^1}(x^1) \ldots, \varphi_{jk}(x^d)$, see the recall on wavelets in part 3, see p. 26
α_{jk}, β_{jk}	coordinates of the wavelet expansion on functions Φ_{jk}, Ψ_{jk}
α_{jk^ℓ}, β_{jk^ℓ}	coordinates of the wavelet expansion on functions φ_{jk^ℓ},

λ_{jk}, δ_{jk}	$\alpha_{jk^1} \dots \alpha_{jk^d}$, $\alpha_{jk} - \lambda_{jk}$
$\lambda_{jk}^{\langle r \rangle}$	$\alpha_{jk^1}^{p_1} \dots \alpha_{jk^d}^{p_d}$ for some p_i such that $0 \leq p_i \leq r$, $\sum_{i=1}^{d} p_i = r$
$\|A\|$	cardinal of the set A
N	Daubechies wavelet parameter $D2N$, N=1 for Haar
n, d	sample size, dimension of f domain
W	candidate to inversion of mixing matrix A

2.8 Bibliographie de l'introduction

(Achard, 2003) S. Achard. Mesure de dépendance pour la séparation aveugle de sources *Thèse Université Joseph Fourier.*

(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.

(Bell & Sejnowski, 1995) A. J. Bell. T.J. Sejnowski An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7 : 1129-1159, 1995.

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

(Bickel & Ritov, 1988) P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives : sharp best order of convergence estimates. *Sankya Ser* A50, 381-393

(Bingham, 2003) E. Bingham. Advances in Independent component Analysis with applications to datamining. *Helsinki University of Technology, thesis*, 2003

(Blum et al., 1961) J.R. Blum, J. Kieffer, M. Rosenblatt. Distribution free tests of independence based on the sample distribution function. *Ann. Math. statist.* **32** 485–498.

(Boscolo et al. 2001) R. Boscolo, H. Pan, V.P. Roychowdhury. Non parametric ICA. *citeseer.ist.psu.edu/557247.html.* 2001

(Cai & Low, 2005) T. Cai and M. Low. *Optimal adaptive estimation of a quadratic functional.* The Annals of Statistics, to appear.

(Cardoso, 1990) J.F. Cardoso. Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'90)*, pages 2655-2658, 1990.

(Cardoso, 1994) J. F. Cardoso. On the performance of orthogonal source separation algorithm.*Signal Processing VII, Theories and Applications, M.J.J. Holt, C.F.N. Cowan, P.M. Grant, and W.A. Sandham, Eds., Edinburgh, Scotland.* September 1994, pp. 776–779, Elsevier.

(Cardoso, 1997) J. F. Cardoso. Infomax and maximum likelihood for source separation. IEEE Letters on Signal Processing, 4, 112–114.

(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.

(J.C.Fort, 1991) J.C. Fort. Stability of the JH sources separation algo- rithm. *Traitement du Signal*, Vol. 8, No. 1, 1991, pp. 35 42.

(Gassiat et Gautherat, 1999) E. Gassiat, E. Gautherat. Speed of convergence for the blind deconvolution of a linear system with discrete random input *Annals of Stat.*, 27, 1999.

(Gayraud et Tribouley, 1999)Gayraud, G and Tribouley, K. (1999). Wavelet Methods to Estimate an Integrated Quadratic Functional : Adaptivity and Asymptotic Law. *Statistics and Probabilty Letter.* **44.** 109-122.

(Gómez Herrero, 2004) G. Gómez Herrero. *Performance analysis of Nonlinear independent component analysis.* Technical report, Université de Tampere, 2004.

(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.

(Härdle et Marron, 1986)Härdle, W. and Marron J.S. (1986). Random approximations tosome measures of accuracy in nonparametric curve estimation. *Journal ofmultivariate analyse.* **20** 91–113.

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

(Hyvärinen, 1999) Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis.*IEEE Transactions on Neural Networks*, 10(3) : 626?634, May 1999.

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

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

(Kerkyacharian & Picard, 1996) Gérard Kerkyacharian Dominique Picard. Estimating non quadratic functionals of a density using Haar wavelets. *Annals of Statistics*, 24(1996), 485-507.

(Lepski, 1991) Lepskii, O.V. (1991). Asymptotically minimax adaptive estimation I : Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682-697.

(Miller and Fischer III, 2003) E.G. Miller and J.W. Fischer III. ICA using spacings estimates of entropy. *proceedings of the fourth international symposium on ICA and BSS.* 2003.

(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.*

(Numerical recipies 1986) William H. Press, Brian P. Flannery, Saul A. Teukolsky, William T. Vetterling Numerical recipies, cambridge university press, 1986

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

(Pham, 2004) Dinh-Tuan Pham. Fast algorithms for mutual information based independent component analysis. *IEEE transactions on signal processing*, vol. 52, 10, october 2004.

(Pierce, 1980) Pierce, J. An introduction to information theory, Dover, 1980.

(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.

(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.

(Song, 2000) K.S. Song. Limit theorems for non parametric sample entropy estimators. *Statistics & Probability Letters.* 49 (2000) 9–18.

(Stone, 1982) Stone, C. Optimal global rates of convergence for nonparametric estimates. *Ann. Statist.* **10** 1040-1053, 1982.

(Taleb, 1999) A. Taleb. Séparation de Sources dans les Mélanges Non Linéaires. *PhD thesis, I.N.P.G. - Laboratoire L.I.S* 1999.

(Tribouley, 2000) K. Tribouley, Adaptive estimation of integrated functionals. *mathematical methods of statistics* 9(2000) p19-38.

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

(Vasicek, 1976) O. Vasicek. A test of normality based on sampled entropy. *Journal of the Royal Statistical Society.* Series B, vol. 38, 1, pp. 54-79, 1976.

(Wickerhauser, 1994) M. V. Wickerhauser. Adapted wavelet analysis from theory to software. IEEE press, 1994.

3. ICA by Wavelets : the basics

In signal processing, blind source separation consists in the identification of analogical, independent signals mixed by a black-box device. In psychometrics, 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. Hyvärinen 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 ℓth 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 independence 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 which 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.

3.1 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 $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. 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

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

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 (33), 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 $\Psi_{jk}(x) = 2^{\frac{jd}{2}}\Psi(2^j x - k)$ 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 $U_{1j} \ldots \otimes U_{dj}$, 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 function 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 $\alpha_{jk} = \alpha_{jk^1..,k^d} = \int f(x)\Phi_{jk}(x)\,dx$.

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 kth order difference $\Delta_h^k f = \Delta_h \Delta_h^{k-1} f$ $\big(k = 1, 2, \ldots$ with $\Delta_h^0 f = f,\ \Delta_h^1 f = \Delta_h f\big)$.

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)

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

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

$$\Omega_{\mathbb{R}^m}^k(f,\delta)_p = \sup_{|h|=1, h \in \mathbb{R}^m} \omega_h^k(f,\delta)_p.$$

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

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

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

$$\Omega_{\mathbb{R}^m}^k(f^{(\varrho)},\delta)_p = \sup_{|h|=1, h \in \mathbb{R}^m} \omega_h^k(f_h^{(\varrho)},\delta)_p = \sum_{|n|=\varrho} \Omega_{\mathbb{R}^m}^k(f^{(n)},\delta)_p.$$

Let $s = [s] + \alpha$; the Hölder space $H_p^s(\mathbb{R}^d)$ 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 $q = \infty$, $B_{spq}(\mathbb{R}) = H_p^s(\mathbb{R})$; 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).

3.2 Wavelet contrast, Besov membership

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

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

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

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

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

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

The following relation hold :

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

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

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

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.

□

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

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

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,

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

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

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

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

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

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

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

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

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 $B_{spq}(\mathbb{R})$.

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

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

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

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

$$\begin{aligned} \Delta^k_t f^{\star\ell}(x^\ell) &= \sum_{l=0}^{k} (-1)^{l+k} C^l_k f^{\star\ell}(x + t) \\ &= \sum_{l=0}^{k} (-1)^{l+k} C^l_k \int_{\mathbb{R}^{d-1}} f(x + te^\ell) dx^{*\ell} = \int_{\mathbb{R}^{d-1}} \Delta^k_{te^\ell} f(x) dx^{*\ell} \,; \end{aligned}$$

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}.

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 $|A|$,

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

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 $f \in L_p(\mathbb{R}^d) \Longleftrightarrow f^\ell \in L_p(\mathbb{R}), \ell = 1 \dots d$.

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 $u = t|A^{-1}h|$,

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

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

When $1 \leq s$, 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 $s < 1$. □

3.3 Risk upper bound

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

The coordinates α_{jk} in the wavelet contrast are estimated as usual by,

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

The linear wavelet contrast estimator is given by,

$$\hat{C}_j(x_1\ldots,x_n) = \sum_{k^1\ldots,k^d} (\hat{\alpha}_{jk^1\ldots,k^d} - \hat{\alpha}_{jk^1}\ldots\hat{\alpha}_{jk^d})^2 = \sum_{k\in\mathbb{Z}^d} \hat{\delta}^2_{jk}, \tag{8}$$

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

The estimator $\hat{\alpha}_{jk}$is unbiased under $E^n_{f_A}$, but so is not $\hat{\alpha}_{jk^1}\ldots\hat{\alpha}_{jk^d}$ 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 $\sum_k 1 \approx 2^{jd}$.

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.

In the following proposition, the variance rate that is obtained is suboptimal, because the proof is deliberately simplified, relative to the intricate proofs of next chapter, that give the optimal rates. The rate for the bias is the good one though.

Proposition 3.4 (Risk upper bound for $\hat{C}_j$)

The quadratic risk $E^n_{f_A}(\hat{C}_j - C_j)^2$ has a convergence rate in $2^{2jd}O(1/n)$ and the bias $E^n_{f_A}\hat{C}_j - C_j$ has a 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 $2^{2jd}O(1/n)$ and the quadratic risk around zero is $E^n_{f_A}\hat{C}_j^2 = C_j^2 + 2^{2jd}O(1/n)$.

By lemma 3.1, $E^n_{f_A}\hat{C}_j = \sum_k E^n_{f_A}\hat{\delta}^2_{jk} = C_j + 2^{jd}O(n^{-1})$. So upon expansion,

$$\begin{aligned} E^n_{f_A}\left[\hat{C}_j - C_j\right]^2 &= E^n_{f_A}\left[\hat{C}_j\right]^2 - [C_j]^2 + 2^{jd}O(n^{-1}) \\ &= E^n_{f_A}[\hat{C}_j - C_j][\hat{C}_j + C_j] + 2^{jd}O(n^{-1}) \\ &\leq (M + \|f_A\|_2^2)E^n_{f_A}[\hat{C}_j - C_j] + 2^{jd}O(n^{-1}) \\ &= 2^{2jd}O(n^{-1}) \end{aligned}$$

with line 3 using the fact that f_A and Φ are compactly supported hence bounded, so that we can say very roughly that $\hat{C}_j \leq 2^{jd}M$.

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 $E^n_{f_A}\hat{C}_j^2 = (E^n_{f_A}\hat{C}_j)^2 + E^n_{f_A}(\hat{C}_j - E^n_{f_A}\hat{C}_j)^2$.
□

Based on this rate, 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 gives back the minimax $n^{\frac{-2s}{2s+d}}$ of the underlying density estimations (see Kerkyacharian & Picard, 1992).

Proposition 3.5 (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 $2^j \approx n^{\frac{1}{4s+2d}}$ and ensures a quadratic risk converging to zero at rate $n^{-\frac{2s}{2s+d}}$.

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

$$|K_\star(A, f) - C_j(f_A)^{\frac{1}{2}}| \leq K2^{-js} \; ; \tag{9}$$

with K a constant.

Taking power 4 of (9) and using prop. 3.4,

$$R(\hat{C}_j, f_A) + K^\star Q(C_j^{\frac{1}{2}}, K^\star) \leq K2^{-4js} + 2^{2jd}Kn^{-1}\,,$$

with K a placeholder for an unspecified constant, $Q(a,b) = -4a^3 + 6a^2b - 4ab^2 + b^3$, 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 $2^{2jd}Kn^{-1}$.

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 $2^{2jd}n^{-1}$, or $2^{j(2d+4s)}$ with n.

This yields $2^j = O(n^{\frac{1}{4s+2d}})$ and convergence rate $n^{\frac{-4s}{4s+2d}}$ 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. 3.4.

□

Corollary 3.1 (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 $2^j \approx n^{\frac{1}{4s'+2d}}$, with $s' = s + d/2 - d/p$ if $1 \leq p \leq 2$ and $s' = s$ if $p > 2$.

This resolution ensures a quadratic risk converging to zero at rate $n^{-\frac{2s'}{2s'+2d}}$.

If $1 \leq p \leq 2$, using the Sobolev embedding $B_{spq} \subset B_{s'p'q}$ for $p \leq p'$ 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 $\{\epsilon_j\} \in \ell_q$,

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

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

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

Finally with s' as claimed, equation (6) yields again$|K_\star(A,f) - C_j(f_A)^{\frac{1}{2}}| \leq K2^{-js'}$. □

3.4 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 '`linear-choice`', 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 $2 \leq 2^j \leq 2N - 1$, 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 the 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. 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. 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 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 significant 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 of 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 $O(n^3d^3)$. 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 both methods (Bach and Jordan 2002 p.19).

3.5 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 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.

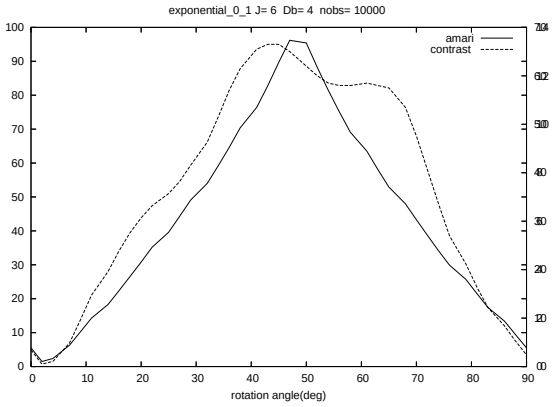

Fig.1. Exponential, D4, j=6, n=10000

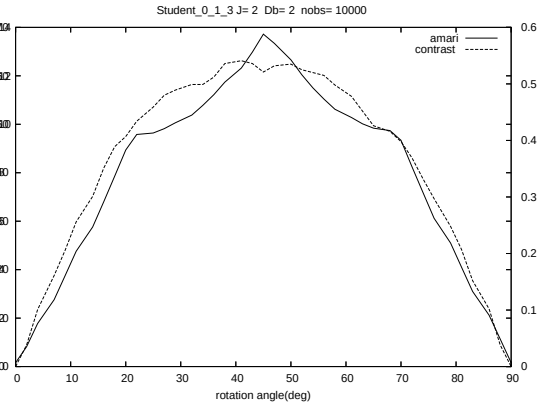

Fig.2. Student, D2, j=2, n=10000

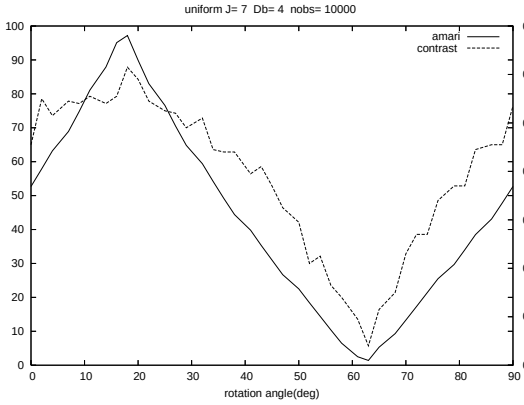

Fig.3. Uniform, D4, j=7, n=10000

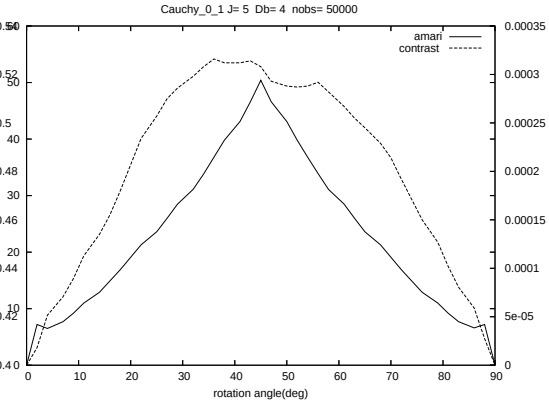

Fig.4. Cauchy, D4, j=5, n=50000

3.6 Appendix

Lemma 3.1 (Bias of $\hat{\alpha}_{jk}^2 - 2\hat{\alpha}_{jk}\hat{\lambda}_{jk} + \hat{\lambda}_{jk}^2$)

Let X be random variables on $\mathbb{R}^d$ with density f_A. Let Φ be the tensorial scaling function of a Daubechies D2N. Let $\hat{\alpha}_{jk} = n^{-1}\sum \Phi_{jk}(X_i)$ and $\hat{\alpha}_{jk^\ell} = n^{-1}\sum \varphi_{jk^\ell} \circ \pi^\ell(X_i)$, $\ell = 1, \ldots, d$. Let $\hat{\lambda}_{jk} = \hat{\alpha}_{jk^1} \ldots \hat{\alpha}_{jk^d}$ and $\lambda_{jk} = \alpha_{jk^1} \ldots \alpha_{jk^d}$.

Then,

$$E_{f_A}^n \hat{\alpha}_{jk}^2 = \alpha_{jk}^2 + O(n^{-1})$$
$$E_{f_A}^n \hat{\alpha}_{jk}\hat{\lambda}_{jk} = \alpha_{jk}\lambda_{jk} + O(n^{-1})$$
$$E_{f_A}^n \hat{\lambda}_{jk}^2 = \lambda_{jk}^2 + O(n^{-1}).$$

For $\hat{\alpha}_{jk}^2$, using lemma 3.3,

$$E_{f_A}^n \hat{\alpha}_{jk}^2 = \frac{1}{n^2}\left[\sum_{i_1=i_2} E_{f_A}^n \Phi_{jk}(X_{i_1})\Phi_{jk}(X_{i_2}) + \sum_{i_1\neq i_2} \alpha_{jk}^2\right] = \frac{1}{n}\Phi_{jk}(X_i)^2 + \frac{n-1}{n}\alpha_{jk}^2 = \alpha_{jk}^2 + O(n^{-1}).$$

For $\hat{\lambda}_{jk}^2$, using lemma 3.2,

$$E_{f_A}^n \hat{\lambda}_{jk}^2 = \lambda_{jk}^2 + O(n^{-1}) + \sum_{k=1}^{2d-1} n(n-1)\dots(n-\kappa+1)E_{f_A}^n Q(2d,k)$$

with $|Q(2d,\kappa)| \leq |\varphi_{jk}(X_{i_1}^{\ell_1})^{p_1}\dots\varphi_{jk}(X_{i_\kappa}^{\ell_\kappa})^{p_\kappa}|$ for some $p_1,\dots,p_\kappa$ with $p_1+\dots+p_\kappa = 2d$, and $\ell_1,\dots,\ell_\kappa \in \{1,\dots,d\}$ corresponding to the maximum in all d dimensions of the term (when there is a match of cardinality c between observation numbers, there is no guarantee that the dimension numbers also match, but we can select each time the dimension that correspond to the maximum of the group of c terms).

So using lemma 3.3, since the overall product factorizes,

$$E_{f_A}^n |Q(2d,\kappa)| \leq C2^{\frac{j}{2}(p_1-2)}\dots 2^{\frac{j}{2}(p_\kappa-2)} = 2^{\frac{j}{2}(2d-2\kappa)}$$

and on $2^{jd} < n$,

$$n^{-2d}\sum_{k=1}^{2d-1} n(n-1)\dots(n-\kappa+1)E_{f_A}^n|Q(2d,k)| \leq n^{-d}\sum_{k=1}^{2d-1}\left(\frac{2^j}{n}\right)^{d-\kappa} \leq Cn^{-d}$$

For $\hat{\alpha}_{jk}\hat{\lambda}_{jk}$, using lemma 3.2,

$$E_{f_A}^n \hat{\alpha}_{jk}\hat{\lambda}_{jk} = \alpha_{jk}\lambda_{jk} + O(n^{-1}) + \sum_{k=1}^{d} n(n-1)\dots(n-\kappa+1)E_{f_A}^n Q(d+1,k)$$

with $|Q(d+1,\kappa)| \leq |\phi_{jk}(X_{i_1})\varphi_{jk}(X_{i_1}^{\ell_1})^{p_1}\dots\varphi_{jk}(X_{i_\kappa}^{\ell_\kappa})^{p_\kappa}|$ for some $p_1,\dots,p_\kappa$ with $p_1+\dots+p_\kappa = d+1$, and $\ell_1,\dots,\ell_\kappa \in \{1,\dots,d\}$ corresponding to the maximum in all d dimensions of the term (we can consider that the big term Φ_{jk} always matches some one dimensional φ_{jk} because the order is higher like this).

So using lemma 3.3, and for the first term using $E_{f_A}^n|\Phi_{jk}(X)\varphi_{jk}(X^\ell)^p| \leq C2^{\frac{j}{2}(p-d)}$, which can be seen by the same means lemma 3.3 is proved, and since again the overall product factorizes,

$$E_{f_A}^n |Q(d+1,\kappa)| \leq C2^{\frac{j}{2}(p_1-d)}\dots 2^{\frac{j}{2}(p_\kappa-2)} = 2^{\frac{j}{2}(d+1-2\kappa-d)}$$

and on $2^{jd} < n$,

$$n^{-1-d}\sum_{k=1}^{d} n(n-1)\dots(n-\kappa+1)E_{f_A}^n|Q(d+1,k)| \leq n^{-d+1/2}\sum_{k=1}^{d}\left(\frac{2^j}{n}\right)^{1/2-\kappa} \leq Cn^{-d+1/2}$$

□

Lemma 3.2 (Decomposition of $\sum F_1(X_{i_1})\dots\sum F_m(X_{i_m})$)

Let $X_1,\dots,X_n$ be a sample of independent identically distributed random variables with distribution P. Let E_P^n denote expectation relative to the joint distribution of $(X_1,\dots,X_n)$. Let $F_1,\dots,F_m$ be bounded functions. Let $S(m,k)$ denote the Stirling number of the second kind.

$$E_P^n \sum_{i_1,\dots,i_m} n^{-m} F_1(X_{i_1})\dots F_m(X_{i_m}) = E_P F_1(X)\dots E_P F_m(X) + O(n^{-1}) + \\ + n^{-m}\sum_{\kappa=1}^{m-1} n(n-1)\dots(n-\kappa+1) E_P^n Q(m,\kappa)$$

where $Q(m,\kappa)$ is a sum of $S(m,\kappa)$ terms each at least factorisable in κ independent products under expectation.

$\sum F_1(X_{i_1})\dots\sum F_m(X_{i_m}) = \sum_{i_1,\dots,i_m} F_1(X_{i_1})\dots F_m(X_{i_m})$ is a sum of n^m terms where the m summation indices can take between 1 and m distinct values to be chosen in the set $\{1\dots,n\}$.

The number of ways for m indices to take κ distinct values is equal to the number of ways to dispatch m objects into κ groups, that is $S(m,\kappa)$, the Stirling number of the second kind ; the so constituted κ groups can be labelled in $n(n-1)\dots(n-\kappa+1)$ different ways.

This amounts to say that $n^m = \sum_{\kappa=1}^m S(m,\kappa) n(n-1)\dots(n-\kappa+1)$, a classical combinatorial formula.

And so for κ fixed, there are $S(d,\kappa)$ ways to factorize the product $F_1(X_{i_1})\dots,F_m(X_{i_m})$ into κ sub-products with unique index i_k, $k=1\dots\kappa$.

Also the expectation of the products factorizes into a product of κ independent sub-products.

Moreover because of the common distribution of the X_is, each of the $S(d,\kappa)$ distinct grouping of indices are invariant under expectation by permutation of the κ indices of the groups, the number of these permutations being $n(n-1)\dots(n-\kappa+1)$;

Hence

$$E^n \sum_{i_1,\dots,i_d} F_1(X_{i_1})\dots F_m(X_{i_d}) = \sum_{\kappa=1}^m n(n-1)\dots(n-\kappa+1) E^n[Q(m,\kappa)]$$

where $Q(m,\kappa)$ is a sum of $S(m,\kappa)$ terms each at least factorisable in κ independent products under expectation.

And then,

$$E^n \frac{1}{n}\sum_i F_1(X_{i_1})\dots\frac{1}{n}\sum_i F_m(X_{i_m}) = \frac{1}{n^m}\Bigg[n(n-1)\dots(n-m+1)EF_1(X)\dots EF_m(X) + \sum_{\kappa=1}^{m-1} n(n-1)\dots(n-\kappa+1)E^n[Q(m,\kappa)]\Bigg] \tag{10}$$

□

Lemma 3.3 **(rth order moment of Φ_{jk})**

Let X be random variables on $\mathbb{R}^d$ with density f. Let Φ be the tensorial scaling function of an MRA of $L_2(\mathbb{R}^d)$. Let $\alpha_{jk} = E_f\Phi_{jk}(X)$. Then for $r \in \mathbb{N}^*$,

$$E_f|\Phi_{jk}(X) - \alpha_{jk}|^r \leq 2^r E_f|\Phi_{jk}(X)|^r \leq 2^r 2^{jd(\frac{r}{2}-1)}\|f\|_\infty\|\Phi\|_r^r.$$

If Φ is the Haar tensorial wavelet then also $E_f\,\Phi_{jk}(X)^r \leq 2^{jd(\frac{r}{2}-\frac{1}{2})}\alpha_{jk}$.

For the left part of the inequality, $\Big(E_f|\Phi_{jk}(X) - \alpha_{jk}|^r\Big)^{\frac{1}{r}} \leq \Big(E_f|\Phi_{jk}(X)|^r\Big)^{\frac{1}{r}} + E_f|\Phi_{jk}(X)|$, and also $E_f|\Phi_{jk}(X)| \leq \Big(E_f|\Phi_{jk}(X)|^r\Big)^{\frac{1}{r}}\Big(E_f 1\Big)^{\frac{r-1}{r}}$.

For the right part, $E_f|\Phi_{jk}(X)|^r = 2^{jdr/2}\int|\Phi(2^jx-k)|^r f(x)dx \leq 2^{jd(\frac{r}{2}-1)}\|f\|_\infty\|\Phi\|_r^r$.

Or also if Φ is positive,

$$\begin{aligned} E_f\Phi_{jk}(X)^r &= 2^{\frac{jd}{2}(r-1)}\int\Phi(2^jx-k)^{r-1}\Phi_{jk}(x)f(x)dx \\ &\leq 2^{\frac{jd}{2}(r-1)}\|\Phi\|_\infty^{r-1}\alpha_{jk}. \end{aligned}$$

□

3.7 References for ICA by wavelets : the basics

(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.* 8 273–303, 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`

4. ICA and estimation of a quadratic functional

In signal processing, blind source separation consists in the identification of analogical, independent signals mixed by a black-box device. In psychometric, one has the notion of structural latent variable whose mixed effects are only measurable through series of tests ; an example are the Big Five identified from factorial analysis by researchers in the domain of personality evaluation (Roch, 1995). Other application fields such as digital imaging, bio medicine, 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 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 $\mathbb{R}^d$, $d \geq 2$; find pairs (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 by minimizing 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.

Let f be the density of the latent variable S relative to Lebesgue measure, assuming it exists. The observed variable $X = AS$ 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 ℓth 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 independence 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.

In the ICA model expressed this way, both f and A are unknown, and the data consists in a random sample of f_A. The semi-parametric case corresponds to f left unspecified, except for general regularity assumptions.

In the semi-parametric case, 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. 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 consider the semi-parametric case and an ICA contrast provided by the factorization measure $\int |f_A - f_A^\star|^2$, with $f_A^\star$ the product of the marginals of f_A. Let's mention 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).

Estimation of a quadratic functional

The problem of estimating nonlinear functionals of a density has been widely studied. In estimating $\int f^2$ under Hölder smoothness conditions, Bickel and Ritov (1988) have shown that parametric rate is achievable for a regularity $s \geq 1/4$, whereas when $s \leq 1/4$, minimax rates of convergence under mean squared error are of the order of $n^{-8s/1+4s}$. This result has been extended to general functionals of a density $\int \phi(f)$ by Birgé and Massart (1995). Laurent (1996) has built efficient estimates for $s > 1/4$.

Let P_j be the projection operator on a multiresolution analysis (MRA) at level j, with scaling function φ, and let $\alpha_{jk} = \int f\varphi_{jk}$ be the coordinate k of f.

In the wavelet setting, given a sample $\tilde{X} = \{X_1, \ldots, X_n\}$ of a density f defined on $\mathbb{R}$, independent and identically distributed, the U-statistic

$$\hat{B}_j^2(\tilde{X}) = \frac{2}{n(n-1)} \sum_{i_1 < i_2} \sum_{k \in \mathbb{Z}} \varphi_{jk}(X_{i_1})\varphi_{jk}(X_{i_2})$$

with mean $\int (P_j f)^2$ is the usual optimal estimator of the quantity $\int f^2$; see Kerkyacharian and Picard (1996), and Tribouley (2000) for the white noise model with adaptive rules.

In what follows, this result is implicitly extended to d dimensions using a tensorial wavelet basis Φ_{jk}, with $\Phi_{jk}(x) = \varphi_{jk^1}(x^1) \ldots \varphi_{jk^d}(x^d)$, $k \in \mathbb{Z}^d, x \in \mathbb{R}^d$; that is to say with $\tilde{X}$ an independent, identically distributed sample of a density f on $\mathbb{R}^d$, the U-statistic $\hat{B}_j^2(\tilde{X}) = \frac{2}{n(n-1)} \sum_{i_1 < i_2} \sum_{k \in \mathbb{Z}^d} \Phi_{jk}(X_{i_1})\Phi_{jk}(X_{i_2})$ with mean $\int (P_j f)^2 = \sum_{k \in \mathbb{Z}^d} \alpha_{jk}^2$ is also optimal in estimating the quantity $\int_{\mathbb{R}^d} f^2$.

In the case of a compactly supported density f, $\hat{B}_j^2$ is computable with a Daubechies wavelet D2N and dyadic approximation of X, but the computational cost is basically in $O(n^2(2N-1)^d)$, which is generally too high in practice.

On the other hand, the plug-in, biased, estimator $\hat{H}_j^2(f) = \sum_k \left[\frac{1}{n} \sum \Phi_{jk}(X_i)\right]^2 = \sum_k \hat{\alpha}_{jk}^2$ enjoys both ease of computation and ease of transitions between resolutions through discrete wavelet transform (DWT), since it builds upon a preliminary estimation of all individual wavelet coordinates of f on the projection space at level j, that is to say a full density estimation. In this setting it is just as easy to compute $\sum_k |\hat{\alpha}_{jk}|^p$ for any $p \geq 1$ or even $\sup |\hat{\alpha}_{jk}|$, with a fixed computational cost in $O(n(2N-1)^d)$ plus sum total, or seek out the max, of a 2^{jd} array.

Both estimators $\hat{H}_j^2$ and $\hat{B}_j^2$ build on the same kernel $h_j(x,y) = \sum_{k\in\mathbb{Z}^d} \Phi_{jk}(x)\Phi_{jk}(y)$ since they are written

$$\hat{H}_j^2(\tilde{X}) = (n^2)^{-1} \sum_{i\in\Omega_n^2} h_j(X_{i^1}, X_{i^2}) \quad \text{and} \quad \hat{B}_j^2(\tilde{X}) = (A_n^2)^{-1} \sum_{i\in I_n^2} h_j(X_{i^1}, X_{i^2}),$$

where, here and in the sequel, $\Omega_n^m = \{(i^1,\ldots,i^m) \colon i^\ell \in \mathbb{N},\, 1 \le i^\ell \le n\}$, $I_n^m = \{i \in \Omega_n^m \colon \ell_1 \neq \ell_2 \Rightarrow i^{\ell_1} \neq i^{\ell_2}\}$ and $A_n^p = n!/(n-p)!$.

The plug-in estimator $\hat{H}_j^2$ is then identified as the Von Mises statistic associated to $\hat{B}_j^2$. In estimating $\sum_k \alpha_{jk}^2$, the mean squared error of unbiased $\hat{B}_j^2$ is merely its variance, while the mean squared error of $\hat{H}_j^2$ adds a squared component $E(\hat{H}_j^2 - \hat{B}_j^2)^2$ because of the inequality $(\hat{H}_j^2 - \sum_k \alpha_{jk}^2)^2 \le 2(\hat{H}_j^2 - \hat{B}_j^2)^2 + 2(\hat{B}_j^2 - \sum_k \alpha_{jk}^2)^2$.

From general results, a U-statistic with finite second raw moment has a variance in Cn^{-1} and under similar conditions, the difference $E|U-V|^r$ between the U-statistic and its associated Von Mises statistic is of the order of n^{-r} (See for instance Serfling, 1980).

In the wavelet case, the dependence of the statistics on the resolution j calls for special treatment in computing these two quantities. This special computation, taking j and other properties of wavelets into account, constitutes the main topic of the paper. In particular whether 2^{jd} is lower than n or not is a critical threshold for resolution parameter j. Moreover, on the set $\{j \colon 2^{jd} > n^2\}$, the statistic $\hat{B}_j^2$, and therefore also $\hat{H}_j^2$, have a mean squared error not converging to zero.

If $\hat{B}_j^2$ and $\hat{H}_j^2$ share some features in estimating $\sum_k \alpha_{jk}^2 = \int (P_j f)^2$, they differ in an essential way : the kernel h_j is averaged in one case over Ω_n^2, the set of unconstrained indexes, and in the other case over I_n^2 the set of distinct indexes. As a consequence, it is shown in the sequel that $\hat{H}_j^2$ has mean squared error of the order of $2^{jd}n^{-1}$, which makes it inoperable as soon as $2^{jd} \ge n$, while $\hat{B}_j^2$ has mean squared error of the order of $2^{jd}n^{-2}$, which is then parametric on the set $\{j \colon 2^{jd} < n\}$. In a general way, this same parallel Ω_n^m versus I_n^m is underpinning most of the proofs presented throughout the paper.

Wavelet ICA

Let f be the latent density in the semi-parametric model introduced above. Let f_A be the mixed density and let $f_A^\star$ be the product of the marginals of f_A.

Assume, as regularity condition, that f belongs to a Besov class $B_{s2\infty}$. It has been checked in previous work (Barbedor, 2005) that f_A and $f_A^\star$, hence $f_A - f_A^\star$ belong to the same Besov space than f.

As usual, the very definition of Besov spaces (here $B_{s2\infty}$) and an orthogonality property of the projection spaces V_j and W_j entails the relation

$$0 \le \int (f_A - f_A^\star)^2 - \int [P_j(f_A - f_A^\star)]^2 \le C2^{-2js}.$$

In this relation, the quantity $\int [P_j(f_A - f_A^\star)]^2$ is recognized as the wavelet ICA contrast $C_j^2(f_A - f_A^\star)$, introduced in a preliminary paper (Barbedor, 2005).

The wavelet ICA contrast is then a factorization measure with bias, in the sense that a zero contrast implies independence of the projected densities, and that independence in projection transfers to original densities up to some bias 2^{-2js}.

Assume for a moment that the difference $f_A - f_A^\star$ is a density and that we dispose of an independent, identically distributed sample $\tilde{S}$ of this difference. Computing the estimators $\hat{B}_j^2(\tilde{S})$ or $\hat{H}_j^2(\tilde{S})$ provides an estimation of $\int (f_A - f_A^\star)^2$, the exact ICA factorization measure. In this case, the j^* realizing the best compromise between the mean squared error in C_j^2 estimation and the bias of the ICA wavelet contrast 2^{-2js}, is exactly the same as the one to minimize the overall risk in estimating the quadratic functional $\int (f_A - f_A^\star)^2$. It is found by balancing bias and variance, a standard procedure in nonparametric estimation. From what was said above $\hat{B}_j^2(\tilde{S})$ would be an optimal estimator of the exact factorization measure $\int (f_A - f_A^\star)^2$.

The previous assumption being heuristic only, and since, in ICA, the data at hand is a random sample of f_A and not $f_A - f_A^\star$, we are lead to consider estimators different from $\hat{B}_j^2$ and $\hat{H}_j^2$, but still alike in some way.

Indeed, let $\delta_{jk} = \int (f_A - f_A^\star)\Phi_{jk}$ be the coordinate of the difference function $f_A - f_A^\star$. In the ICA context, δ_{jk} is estimable only through the difference $(\alpha_{jk} - \alpha_{jk^1} \dots \alpha_{jk^d})$ where $\alpha_{jk} = \int f_A \Phi_{jk}$ is the coordinate of f_A and $\alpha_{jk^\ell} = \int f_A^{\star\ell} \varphi_{jk^\ell}$ refers to the coordinate of marginal number ℓ of f_A, written $f_A^{\star\ell}$.

To estimate $\sum_k \delta_{jk}^2$, estimators of the type $\hat{B}_j^2$ and $\hat{H}_j^2$ are not alone enough. Instead we use the already introduced wavelet contrast estimator (plug-in), $\hat{C}_j^2(\tilde{X}) = \sum_k (\hat{\alpha}_{jk^1,\dots k^d} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d})^2$, and the corresponding U-statistic estimator of order $2d+2$,

$$\hat{D}_j^2(\tilde{X}) = \frac{1}{A_n^{2d+2}} \sum_{i \in I_n^{2d+2}} \sum_{k \in \mathbb{Z}^d} \left[\Phi_{jk}(X_{i^1}) - \varphi_{jk^1}(X_{i^2}^1) \dots \varphi_{jk^d}(X_{i^{d+1}}^d)\right] \left[\Phi_{jk}(X_{i^{d+2}}) - \varphi_{jk^1}(X_{i^{d+3}}^1) \dots \varphi_{jk^d}(X_{i^{2d+2}}^d)\right]$$

with as above $I_n^m = \{(i^1, \dots, i^m) \colon i^\ell \in \mathbb{N},\, 1 \le i^\ell \le n,\, i^{\ell_1} \ne i^{\ell_2} \text{ if } \ell_1 \ne \ell_2\}$ and X^ℓ referring to the dimension ℓ of $X \in \mathbb{R}^d$.

As it turns out, the U-statistic estimator $\hat{D}_j^2$ computed on the full sample $\tilde{X}$ is slightly suboptimal, compared to the rate of a $\hat{B}_j^2$ in estimating a bare quadratic functional.

As an alternative to $\hat{D}_j^2(\tilde{X})$, we are then led to consider various U-statistic and plug-in estimators based on splits of the full sample, which seems the only way to find back the well-known optimal convergence rate of the estimation of quadratic functional, for reasons that will be explained in the course of the proofs.

These additional estimators and conditions of use, together with the full sample estimators $\hat{C}_j^2$ and $\hat{D}_j^2$ are presented in section 3.

Section 2 of the paper recalls some essential definitions for the convenience of the reader not familiar with wavelets and Besov spaces, and may be skipped.

Section 4 is all devoted to the computation of a risk bound for the different estimators presented in section 3.

We refer the reader to a preliminary paper on ICA by wavelets (Barbedor, 2005) which contains numerical simulations, details on the implementation of the wavelet contrast estimator and other practical considerations not repeated here. Note that this paper gives an improved convergence rate in $C2^{jd}n^{-1}$ for the wavelet contrast estimator $\hat{C}_j^2$, already introduced in the preliminary paper.

4.1 Notations

We set here general 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 $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. 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

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

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

A 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

$$\Psi(x)_{i^1\ldots,i^d} = \prod_{\ell=1}^{d} \psi^{(i^\ell)}(x^\ell), \tag{11}$$

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 $\Psi_{jk}(x) = 2^{\frac{jd}{2}} \Psi(2^j x - k)$ 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 $U_{1j} \ldots \otimes U_{dj}$, 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

$$\sum_{k \in \mathbb{Z}^d} \alpha_{j_0 k}(f) \Phi_{j_0 k} + \sum_{j=j_0}^{\infty} \sum_{k \in \mathbb{Z}^d} \beta_{jk}(f) \Psi_{jk} \tag{12}$$

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

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 function 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 a concentration condition verified for compactly supported wavelets, the projection operator can also be written

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

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

- *Besov spaces*

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

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

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

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

Let P_j be the projection operator on V_j and let D_j be the projection operator on W_j. J_{spq} is equivalent to

$$J'_{spq}(f) = \|P_j f\|_p + \left[\sum_{j \geq 0} \left[2^{js} \|D_j f\|_p \right]^q \right]^{\frac{1}{q}}$$

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).

4.2 Estimating the factorization measure $\int (f_A - f_A^\star)^2$

We first recall the definition of the wavelet contrast already introduced in a previous article (Barbedor, 2005).

Let f and g be two functions on $\mathbb{R}^d$ and let Φ be the scaling function of a multiresolution analysis of $L_2(\mathbb{R}^d)$ for which projections of f and g exist.

Define the approximate loss function

$$C_j^2(f-g) = \sum_{k \in \mathbb{Z}^d} \left(\int (f-g)\Phi_{jk} \right)^2 = \|P_j(f-g)\|_2^2.$$

It is clear that $f = g$ implies $C_j^2 = 0$ and that $C_j^2 = 0$ implies $P_j f = P_j g$ almost surely.

Let f be a density function on $\mathbb{R}^d$; denote by $f^{\star\ell}$ the marginal distribution in dimension ℓ

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

and denote by $f^\star$ the product of marginals $f^{\star 1} \ldots f^{\star d}$. The functions f, $f^\star$ and the $f^{\star\ell}$ admit a wavelet expansion on a compactly supported basis (φ, ψ). Consider the projections up to order j, that is to say the projections of f, $f^\star$ and $f^{\star\ell}$ on V_j^d and V_j, namely

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

with $\alpha_{jk}(f^{\star\ell}) = \int f^{\star\ell}\varphi_{jk}$ and $\alpha_{jk}(f) = \int f\Phi_{jk}$. At least for compactly supported densities and compactly supported wavelets, it is clear that $P_j f^\star = P_j^1 f^{\star 1} \ldots P_j^d f^{\star d}$.

Proposition 4.6 (ICA wavelet contrast)

Let f be a compactly supported density function on $\mathbb{R}^d$ and let φ be the scaling function of a compactly supported wavelet.

Define the wavelet ICA contrast as $C_j^2(f-f^\star)$. Then,

$$\begin{aligned} f \textit{ factorizes} &\implies C_j^2(f-f^\star)=0 \\ C_j^2(f-f^\star)=0 &\implies P_j f = P_j f^{\star 1}\dots P_j f^{\star d} \quad a.s. \end{aligned}$$

$f=f^1\dots f^d \implies f^{\star\ell}=f^\ell,\ \ell=1,\dots d.$ □

Wavelet contrast and quadratic functional

Let $f=f_I$ be a density defined on $\mathbb{R}^d$ whose components are independent, that is to say f is equal to the product of its marginals. Let f_A be the mixed density given by $f_A(x)=|\det A^{-1}|f(A^{-1}x)$, with A a $d\times d$ invertible matrix. Let $f_A^\star$ be the product of the marginals of f_A. Note that when $A=I$, $f_A^\star=f_I^\star=f_I=f$.

By definition of a Besov space $B_{spq}(\mathbb{R}^d)$ with a r-regular wavelet φ, $r>s$,

$$f\in B_{spq}(\mathbb{R}^d) \iff \|f-P_jf\|_p = 2^{-js}\,\epsilon_j, \quad \{\epsilon_j\}\in \ell_q(\mathbb{N}^d). \tag{13}$$

So, from the decomposition

$$\begin{aligned} \|f_A-f_A^\star\|_2^2 &= \int P_j(f_A-f_A^\star)^2 + \int \left[f_A-f_A^\star-P_j(f_A-f_A^\star)\right]^2, \\ &= C_j^2(f_A-f_A^\star) + \int \left[f_A-f_A^\star-P_j(f_A-f_A^\star)\right]^2, \end{aligned}$$

resulting from the orthogonality of V_j and W_j, and assuming that f_A and $f_A^\star$ belong to $B_{s2\infty}(\mathbb{R}^d)$,

$$0\le \|f_A-f_A^\star\|_2^2 - C_j^2(f_A-f_A^\star) \le C2^{-2js}, \tag{14}$$

which gives an illustration of the shrinking (with j) distance between the wavelet contrast and the always bigger squared L_2 norm of $f_A-f_A^\star$ representing the exact factorization measure. A side effect of (14) is that $C_j^2(f_A-f_A^\star)=0$ is implied by $A=I$.

Estimators under consideration

Let S be the latent random variable with density f.
Define the experiment $\mathcal{E}^n=\big(\mathcal{X}^{\otimes n}, \mathcal{A}^{\otimes n}, (X_1,\dots,X_n), P_{f_A}^n, f_A\in B_{spq}\big)$, where $X_1,\dots,X_n$ is an iid sample of $X=AS$, and $P_{f_A}^n = P_{f_A}\dots\otimes P_{f_A}$ is the joint distribution of $(X_1\dots,X_n)$.

Define the coordinates estimators

$$\hat\alpha_{jk}=\hat\alpha_{jk^1\dots,k^d}=\frac{1}{n}\sum_{i=1}^n \varphi_{jk^1}(X_i^1)\dots\varphi_{jk^d}(X_i^d) \quad \text{and} \quad \hat\alpha_{jk^\ell}=\frac{1}{n}\sum_{i=1}^n \varphi_{jk^\ell}(X_i^\ell) \tag{15}$$

where X^ℓ is coordinate ℓ of $X\in\mathbb{R}^d$. Define also the shortcut $\hat\lambda_{jk}=\hat\alpha_{jk^1}\dots\hat\alpha_{jk^d}$.

Define the full sample plug-in estimator

$$\hat{C}_j^2 = \hat{C}_j^2(X_1, \dots, X_n) = \sum_{(k^1,\dots,k^d)\in\mathbb{Z}^d} (\hat{\alpha}_{j(k^1,\dots,k^d)} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d})^2 = \sum_{k\in\mathbb{Z}^d} (\hat{\alpha}_{jk} - \hat{\lambda}_{jk})^2 \tag{16}$$

and the full sample U-statistic estimator

$$\begin{aligned}\hat{D}_j^2 = \hat{D}_j^2(X_1, \dots, X_n) = \frac{1}{A_n^{2d+2}} \sum_{i\in I_n^{2d+2}} \sum_{k\in\mathbb{Z}^d} &\left[\Phi_{jk}(X_{i^1}) - \varphi_{jk^1}(X^1_{i^2}) \dots \varphi_{jk^d}(X^d_{i^{d+1}})\right] \\ &\left[\Phi_{jk}(X_{i^{d+2}}) - \varphi_{jk^1}(X^1_{i^{d+3}}) \dots \varphi_{jk^d}(X^d_{i^{2d+2}})\right]\end{aligned} \tag{17}$$

where I_n^m is the set of indices $\{(i^1, \dots, i^m) \colon i^\ell \in \mathbb{N}, 1 \le i^\ell \le n, i^{\ell_1} \ne i^{\ell_2} \text{ if } \ell_1 \ne \ell_2\}$ and $A_n^m = \frac{n!}{(n-m)!} = |I_n^m|$.

Define also the U-statistic estimators

$$\begin{aligned}\hat{B}_j^2(\{X_1, \dots, X_n\}) &= \sum_k \frac{1}{A_n^2} \sum_{i\in I_n^2} \Phi_{jk}(X_{i^1})\Phi_{jk}(X_{i^2}) \\ \hat{B}_j^2(\{X_1^\ell, \dots, X_n^\ell\}) &= \sum_{k^\ell} \frac{1}{A_n^2} \sum_{i\in I_n^2} \varphi_{jk^\ell}(X^\ell_{i^1})\varphi_{jk^\ell}(X^\ell_{i^2}).\end{aligned} \tag{18}$$

Notational remark

Unless otherwise stated, superscripts designate coordinates of multi-dimensional entities while subscripts designate unrelated entities of the same set without reference to multi-dimensional unpacking. For instance, an index k belonging to $\mathbb{Z}^d$ is also written $k = (k^1, \dots, k^d)$, with $k^\ell \in \mathbb{Z}$. Likewise a multi-index i is written $i = (i^1, \dots, i^m)$ when belonging to some $\Omega_n^m = \{i = (i^1, \dots, i^m) \colon i^\ell \in \mathbb{N}, 1 \le i^\ell \le n\}$ or $I_n^m = \{i \in \Omega_n^m \colon \ell_1 \ne \ell_2 \Rightarrow i^{\ell_1} \ne i^{\ell_2}\}$, for some $m \ge 1$; but i_1, i_2 would designate two different elements of I_n^m, so for instance $\left[\sum_{i=1}^n \sum_{k\in\mathbb{Z}^d} \Phi_{jk}(X_i)\right]^2$ is written $\sum_{i_1,i_2} \sum_{k_1,k_2} \Phi_{jk_1}(X_{i_1})\Phi_{jk_2}(X_{i_2})$. Finally X^ℓ is coordinate ℓ of observation $X \in \mathbb{R}^d$ and $\tilde{X}$ refers to a sample $\{X_1, \dots X_n\}$.

▪

As was said in the introduction and as is shown in proposition 4.11, the estimator $\hat{D}_j^2$ computed on the full sample is slightly suboptimal. We now review some possibilities to split the sample so that various alternatives to $\hat{D}_j^2$ on the full sample could be computed in an attempt to regain optimality through block independence.

We need not consider $\hat{C}_j^2$ on independent sub samples because, as will be seen, the order of its risk upper bound is given by the order of the component $\sum_k \hat{\alpha}_{jk}^2 - \alpha_{jk}^2$ which is not improved by splitting the sample (contrary to $\sum_k \hat{\lambda}_{jk}^2 - \lambda_{jk}^2$ and $\sum_k \hat{\alpha}_{jk}\hat{\lambda}_{jk} - \alpha_{jk}\lambda_{jk}$). The rate of $\hat{C}_j^2$ is unchanged compared to what appeared in Barbedor (2005).

Sample split

- Split the full sample $\{X_1, \ldots, X_n\}$ in $d+1$ disjoint sub samples $\tilde{R}^0, \tilde{R}^1, \ldots \tilde{R}^d$ where the sample $\tilde{R}^0$ refers to a plain section of the full sample, $\{X_1, \ldots, X_{[n/d+1]}\}$ say, and the samples $\tilde{R}^1, \ldots, \tilde{R}^d$ refer to dimension ℓ of their section of the full sample, for instance $\{X^\ell_{[n/d+1]\ell+1}, \ldots, X^\ell_{[n/d+1](\ell+1)}\}$.

  Estimate each plug-in $\hat{\alpha}_{jk}(\tilde{R}^0)$ and $\hat{\alpha}_{jk^\ell}(\tilde{R}^\ell)$, and the U-statistics $\hat{B}^2_j(\tilde{R}^0)$, $\hat{B}^2_j(\tilde{R}^\ell)$, $\ell = 1, \ldots, d$ on each independent sub-sample. This leads to the definition of the $d+1$ samples mixed plug-in estimator

$$\hat{F}^2_j(\tilde{R}^0, \tilde{R}^1, \ldots, \tilde{R}^d) = \hat{B}^2_j(\tilde{R}^0) + \prod_{\ell=1}^{d} \hat{B}^2_j(\tilde{R}^\ell) - 2 \sum_{k \in \mathbb{Z}^d} \hat{\alpha}_{jk}(\tilde{R}^0)\hat{\alpha}_{jk^1}(\tilde{R}^1) \ldots \hat{\alpha}_{jk^d}(\tilde{R}^d). \tag{19}$$

  to estimate the quantity $\sum_k \alpha^2_{jk} + \prod_{\ell=1}^{d} \left(\sum_{k^\ell \in \mathbb{Z}} \alpha^2_{jk^\ell} \right) - 2 \sum_k \alpha_{jk}\alpha_{jk^1} \ldots \alpha_{jk^d} = C^2_j$.

  Using estimators $\hat{B}^2_j$ places us in the exact replication of the case $\hat{B}^2_j$ found in Kerkyacharian and Picard (1996), except for an estimation taking place in dimension d in the case of $\hat{B}^2_j(R^0)$. The risk of this procedure is given by proposition 4.8.

- Using the full sample $\{X_1, \ldots, X_n\}$ we can generate an identically distributed sample of $f^\star_A$, namely $\tilde{DS} = \cup_{i \in \Omega^d_n} \{X^1_{i^1} \ldots X^d_{i^d}\}$, but is not constituted of independent observations when $A \neq I$.

  But then using a Hoeffding like decomposition, we can pick from $\tilde{DS}$, a sample of independent observations, $\tilde{IS} = \cup_{k=1 \ldots [n/d]} \{X^1_{(k-1)d+1} \ldots X^d_{kd}\}$, although it leads to a somewhat arbitrary omission of a large part of the information available. Nevertheless we can assume that we dispose of two independent, identically distributed samples, one for f_A labelled $\tilde{R}$ and one for $f^\star_A$ labelled $\tilde{S}$, with $\tilde{R}$ independent of $\tilde{S}$. In this setting we define the mixed plug-in estimator

$$\hat{G}^2_j(\tilde{R}, \tilde{S}) = \hat{B}^2_j(\tilde{R}) + \hat{B}^2_j(\tilde{S}) - 2 \sum_{k \in \mathbb{Z}^d} \hat{\alpha}_{jk}(\tilde{R})\hat{\alpha}_{jk}(\tilde{S}) \tag{20}$$

  and the two samples U-statistic estimator

$$\hat{\Delta}^2_j(\tilde{R}, \tilde{S}) = \frac{1}{A^2_n} \sum_{i \in I^2_n} \sum_{k \in \mathbb{Z}^d} \left[\Phi_{jk}(R_{i^1}) - \Phi_{jk}(S_{i^1})\right] \left[\Phi_{jk}(R_{i^2}) - \Phi_{jk}(S_{i^2})\right] \tag{21}$$

  assuming for simplification that both samples have same size n (that would be different from the size of the original sample). $\hat{\Delta}^2_j(R, S)$ is the exact replication (except for dimension d instead of 1) of the optimal estimator of $\int (f-g)^2$ for unrelated f and g found in Butucea and Tribouley (2006). The risk of this optimal procedure is found in proposition 4.9.

-

Bias variance trade-off

Let an estimator $\hat{T}_j$ be used in estimating the quadratic functional $K_\star = \int (f_A - f_A^\star)^2$; using (14), an upper bound for the mean squared error of this procedure when $f_A \in B_{s2\infty}(\mathbb{R}^d)$ is given by

$$E^n_{f_A}(\hat{T}_j - K_\star)^2 \leq 2E^n_{f_A}(\hat{T}_j - C_j^2)^2 + C2^{-4js}, \tag{22}$$

which shows that the key estimation is that of the wavelet contrast $C_j^2(f_A - f_A^\star)$ by the estimator $\hat{T}_j$. Once an upper bound of the risk of $\hat{T}_j$ in estimating C_j^2 is known, balancing the order of the bound with the squared bias 2^{-4js} gives the optimal resolution j. This is a standard procedure in nonparametric estimation.

Before diving into the computation of risk bounds, we give a summary of the different convergence rates in proposition 4.7 below. The estimators based on splits of the full sample are optimal. $\hat{D}_j^2$ is almost parametric on $\{2^{jd} < n\}$ and is otherwise optimal.

Proposition 4.7 (Minimal risk resolution in the class $B_{s2\infty}$ and convergence rates)

Assume that f belongs to $B_{s2\infty}(\mathbb{R}^d)$, and that projection is based on a r-regular wavelet φ, $r > s$. Convergence rates for the estimators defined at the beginning of this section are the following :

Convergence rates		
statistic	$2^{jd} < n$	$2^{jd} \geq n$
$\hat{\Delta}_j^2(\tilde{R}, \tilde{S})$, $\hat{G}_j^2(\tilde{R}, \tilde{S})$, $\hat{F}_j^2(\tilde{R}^0, \tilde{R}^1, \ldots, \tilde{R}^d)$	parametric	$n^{\frac{-8s}{4s+d}}$
$\hat{D}_j^2(\tilde{X})$	$n^{-1+\frac{1}{1+4s}}$	$n^{\frac{-8s}{4s+d}}$
$\hat{C}_j^2(\tilde{X})$	$n^{\frac{-4s}{4s+d}}$	inoperable

Table 7. Convergence rates at optimal $j_\star$

The minimal risk resolution $j_\star$ satisfies, $2^{j_\star d} \approx (<) n$ for parametric cases ; $2^{j_\star d} \approx n^{1+\frac{d-4s}{d+4s}}$ for $\hat{D}_j^2$, $\hat{\Delta}_j^2$, $\hat{G}_j^2$ or $\hat{F}_j^2$ when $s \leq \frac{d}{4}$ and $2^{j_\star d} \approx n^{\frac{d}{d+4s}}$ for $\hat{C}_j^2$.

Besov assumption about f transfers to f_A (see Barbedor, 2005). Using

$$E^n_{f_A}(\hat{H}_j - K_\star)^2 \leq 2E^n_{f_A}(\hat{H}_j - C_j^2)^2 + C2^{-4js},$$

and balancing bias 2^{-4js} and variance of the estimator $\hat{H}_j$, yields the optimal resolution j.

- from proposition 4.10, for estimator $\hat{C}_j^2(\tilde{X})$, the bound is inoperable on $\{2^{jd} > n\}$. Otherwise equating $2^{jd}n^{-1}$ with 2^{-4js} yields $2^j = n^{\frac{1}{d+4s}}$ and a rate in $n^{\frac{-4s}{d+4s}}$.
- from proposition 4.9 and 4.8, for estimators $\hat{F}_j^2(R^0, R^1, \ldots, R^d)$, $\hat{F}_j^2(R, S)$ and $\hat{D}_j^2(R, S)$, on $\{2^{jd} > n\}$ equating $2^{jd}n^{-2}$ with 2^{-4js} yields $2^j = n^{\frac{2}{d+4s}}$ and a rate in $n^{-\frac{8s}{d+4s}}$; on $\{2^{jd} < n\}$ the rate is parametric. Moreover $2^{jd} < n$ implies that $s \geq d/4$ and $2^{jd} > n$ implies that $s \leq d/4$.

- from proposition 4.11, for estimator $\hat{D}_j^2(\tilde{X})$ on $\{2^{jd} > n\}$ equating $2^{jd}n^{-2}$ with 2^{-4js} yields $2^j = n^{\frac{2}{d+4s}}$ and a rate in $n^{-\frac{8s}{d+4s}}$; on $\{2^{jd} < n\}$ the rate is found by equating $2^j n^{-1}$ with 2^{-4js}.

□

4.3 Risk upper bounds in estimating the wavelet contrast

In the forthcoming lines, we 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 $\sum_{k\in\mathbb{Z}^d} 1 \approx 2^{jd}$.

Proposition 4.8 (Risk upper bound, $d+1$ independent samples — f_A, $f_A^{\star 1}, \dots, f_A^{\star d}$)

Let $\{X_1, \dots, X_n\}$ be an independent, identically distributed sample of f_A. Let $\{R_1^\ell, \dots, R_n^\ell\}$ be an independent, identically distributed sample of $f_A^{\star\ell}$, $\ell = 1, \dots, d$. Assume that f is compactly supported and that φ is a Daubechies D2N. Assume that the $d+1$ samples are independent. Let $E_{f_A}^n$ be the expectation relative to the joint distribution of the d+1 samples. Then on $\{2^{jd} < n^2\}$,

$$E_{f_A}^n \left(\hat{F}_j^2(\tilde{X}, \tilde{R}^1, \dots, \tilde{R}^d) - C_j^2\right)^2 \leq Cn^{-1} + C2^{jd}n^{-2}\,\mathbb{I}\left\{2^{jd} > n\right\}.$$

For the U-statistic $\hat{F}_j^2(\tilde{X}, \tilde{R}^1, \dots, \tilde{R}^d)$, with $\hat{\alpha}_{jk} = \hat{\alpha}_{jk}(\tilde{X})$, $\hat{\alpha}_{jk^\ell} = \hat{\alpha}_{jk^\ell}(\tilde{R}^\ell)$ and $\hat{\lambda}_{jk} = \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d}$,

$$(\hat{F}_j^2 - C_j^2)^2 \leq 3\Big[\hat{B}_j^2(\tilde{X}) - \sum_k \alpha_{jk}^2\Big]^2 + 3\Big[\prod_\ell \hat{B}_j^2(\tilde{R}^\ell) - \prod_\ell \sum_{k^\ell} \alpha_{jk^\ell}^2\Big]^2 + 6\Big[\sum_k \hat{\alpha}_{jk}\hat{\lambda}_{jk} - \sum_k \alpha_{jk}\lambda_{jk}\Big]^2.$$

On $\{2^{jd} < n^2\}$, by proposition 4.14 for the term on the left, proposition 4.15 for the middle term, and proposition 4.16 for the term on the right, the quantity is bounded by $Cn^{-1} + C2^{jd}n^{-2}$.

□

Proposition 4.9 (Risk upper bound, 2 independent samples — f_A, $f_A^\star$)

Let $\tilde{X} = \{X_1, \dots, X_n\}$ be an independent, identically distributed sample of X with density f_A. Let $\tilde{R} = \{R_1, \dots, R_n\}$ be an independent, identically distributed sample of R with density $f_A^\star$. Assume that f is compactly supported and that φ is a Daubechies D2N. Assume that the two samples are independent. Let $E_{f_A}^n$ be the the expectation relative to the joint distribution of the two samples.

Then

$$E_{f_A}^n \left(\hat{G}_j^2(\tilde{X}, \tilde{R}) - C_j^2\right)^2 \leq Cn^{-1} + C2^{jd}n^{-2}\,\mathbb{I}\left\{2^{jd} > n\right\}$$
$$E_{f_A}^n \left(\hat{\Delta}_j^2(\tilde{X}, \tilde{R}) - C_j^2\right)^2 \leq C^\star n^{-1} + C2^{jd}n^{-2}.$$

with $C^\star = 0$ at independence.

For the estimator $\hat{G}_j^2(\tilde{X}, \tilde{R})$ the proof is identical to the proof of proposition 4.8, the only difference being that $\hat{\lambda}_{jk}$ and λ_{jk} no more designate a product of d one dimensional coordinates but full fledged d dimensional coordinate equivalent to $\hat{\alpha}_{jk}$ and α_{jk}.

The only new quantity to compute is then $E_{f_A}^n \Big(\sum_k \hat{\alpha}_{jk}(\tilde{X})\hat{\lambda}_{jk}(\tilde{R}) - \sum_k \alpha_{jk}\lambda_{jk}\Big)^2$, coming from the crossed term.

Let $Q = E_{f_A}^n \big(\sum_k \hat{\alpha}_{jk}(\tilde{X})\hat{\lambda}_{jk}(\tilde{R})\big)^2$. Let $\theta = \sum_k \alpha_{jk}\lambda_{jk}$. Recall that $\Omega_n^m = \big\{(i^1, \dots, i^m) \colon i^\ell \in \mathbb{N},\, 1 \le i^\ell \le n\big\}$.

Let $\tilde{\imath}$ be the set of distinct coordinates of $i \in \Omega_n^4$. So that, estimators being plug-in, with a sum on Ω_n^4, with cardinality n^4,

$$\begin{aligned}
Q &= E_{f_A}^n \frac{1}{n^4} \sum_{i \in \Omega_n^4} \sum_{k_1, k_2} \Phi_{jk_1}(X_{i^1})\Phi_{jk_1}(R_{i^2})\Phi_{jk_2}(X_{i^3})\Phi_{jk_2}(R_{i^4}) \\
&\le \frac{1}{n^4}\Bigg[\sum_{|\tilde{\imath}|=4} \theta^2 + \sum_{|\tilde{\imath}|=3} \Big[\theta^2 + (4N-3)^d \sum_k E_{f_A}^n \Phi(X)^2 \lambda_{jk}^2 + (4N-3)^d \sum_k E_{f_A}^n \alpha_{jk}^2 \Phi(R)^2\Big] + \\
&\quad + \sum_{|\tilde{\imath}| \le 2} (4N-3)^d \sum_k E_{f_A}^n \Phi(X)^2 \Phi(R)^2 \Bigg]
\end{aligned}$$

with lines 2 and 3 expressing all possible matches between the coordinates of i, and using lemma 4.10 to reduce double sums in k_1, k_2.

By independence of the samples, using lemma 4.11 and the fact that $|\{i \in \Omega_n^4 \colon |\,\tilde{\imath}\,| = c\}| = O(n^c)$ given by lemma 4.5,

$$Q \le \frac{A_n^4}{n^4}\theta^2 + Cn^{-1}\Big(\theta^2 + C\sum_k \lambda_{jk}^2 + C\sum_k \alpha_{jk}^2\Big) + Cn^{-2}2^{jd}.$$

with $A_n^p = n!/(n-p)!$. So that, with $A_n^4 n^{-4} = 1 - \frac{6}{n} + Cn^{-2}$,

$$Q - \theta^2 \le Cn^{-2} + Cn^{-1} + Cn^{-2}2^{jd}.$$

The rate is thus unchanged for $\hat{F}_j^2$ compared to the $d+1$ sample case in previous proposition.

Case $\hat{\Delta}_j^2(\tilde{X}, \tilde{R})$

Recall that $I_n^m = \big\{(i^1, \dots, i^m) \colon i^\ell \in \mathbb{N},\, 1 \le i^\ell \le n,\, i^{\ell_1} \ne i^{\ell_2} \text{ if } \ell_1 \ne \ell_2\big\}$.

For $i \in I_n^2$, let $h_{jk}(i) = \big[\Phi_{jk}(X_{i^1}) - \Phi_{jk}(R_{i^1})\big]\big[\Phi_{jk}(X_{i^2}) - \Phi_{jk}(R_{i^2})\big]$ and let $\theta = C_j^2$; so that

$$\begin{aligned}
E_{f_A}^n \Big(\hat{\Delta}_j^2(\tilde{X}, \tilde{R}) - \theta\Big)^2 &= -\theta^2 + E_{f_A}^n \frac{1}{(A_n^2)^2} \sum_{i_1, i_2} \sum_{k_1, k_2} h_{jk_1}(i_1) h_{jk_2}(i_2) \\
&= \Big(\frac{\#\{i_1, i_2 \colon |i_1 \cap i_2| = 0\}}{(A_n^2)^2} - 1\Big)\theta^2 + \frac{1}{(A_n^2)^2} \sum_{|i_1 \cap i_2| \ge 1} \sum_{k_1, k_2} E_{f_A}^n h_{jk_1}(i_1) h_{jk_2}(i_2),
\end{aligned}$$

and by lemma 4.6 the quantity in parenthesis on the left is of the order of Cn^{-2}.

Label $Q(i_1, i_2)$ the quantity $E^n_{f_A} \sum_{k_1,k_2} h_{jk_1}(i_1)h_{jk_2}(i_2)$. Let also $\delta_{jk} = \alpha_{jk} - \lambda_{jk}$.

So that with only one matching coordinate between i_1 and i_2,

$$Q(i_1, i_2)\,\mathbb{I}\{|i_1 \cap i_2| = 1\} = E^n_{f_A} \sum_{k_1,k_2} \delta_{jk_1}\delta_{jk_2}\big(\Phi_{jk_1}(X)\Phi_{jk_2}(X) + \Phi_{jk_1}(R)\Phi_{jk_2}(R)\big)$$
$$- 2\sum_k \delta_{jk}\alpha_{jk} \sum_k \delta_{jk}\lambda_{jk}$$

Again by lemma 4.10 and lemma 4.11, for X or R

$$E^n_{f_A} \sum_{k_1,k_2} \delta_{jk_1}\delta_{jk_2}|\Phi_{jk_1}(X)\Phi_{jk_2}(X)| \le (4N-3)^d \sum_k \delta^2_{jk} E^n_{f_A}\Phi_{jk}(X)^2 \le C\sum_k \delta^2_{jk} \le C$$

and since all other terms are bounded by a constant not depending on j, by lemma 4.6 $(A^2_n)^{-2}\sum_{i_1,i_2} Q(i_1, i_2)\,\mathbb{I}\{\,|i_1 \cap i_2| = 1\} \le Cn^{-1}$.

Likewise, the maximum order of $Q(i_1, i_2)\,\mathbb{I}\{|i_1 \cap i_2| = 2\}$ is $\sum_k [E^n_{f_A}\Phi_{jk}(X)^2]^2$, and the corresponding bound is $2^{jd}n^{-2}$.

□

Proposition 4.10 (Full sample $\hat{C}^2_j$ risk upper bound)

Let $\tilde{X} = X_1, \ldots, X_n$ be an independent, identically distributed sample of f_A. Assume that f is compactly supported and that φ is a Daubechies D2N. Let $E^n_{f_A}$ be the the expectation relative to the joint distribution of the sample $\tilde{X}$. Let $\hat{C}^2_j$ be the plug-in estimator defined in (16), Then on $\{2^{jd} < n^2\}$

$$E^n_{f_A}\Big(\hat{C}^2_j(\tilde{X}) - C^2_j\Big)^2 \le C2^{jd}n^{-1}.$$

Use the decompostion

$$E^n_{f_A}\big[\hat{C}^2_j - C^2_j\big]^2 \le E^n_{f_A} 3\big(\sum_k \hat{\alpha}^2_{jk} - \alpha^2_{jk}\big)^2 + 3\big(\sum_k \hat{\lambda}^2_{jk} - \lambda^2_{jk}\big)^2 + 3\big(4\sum_k \hat{\alpha}_{jk}\hat{\lambda}_{jk} - \alpha_{jk}\lambda_{jk}\big)^2.$$

By proposition 4.12 the first term is of the order of $2^{jd}n^{-1}$. By proposition 4.13 the two other terms are of the order of $Cn^{-1} + 2^j n^{-1}\,\mathbb{I}\big\{2^{jd} < n^2\big\}$.

□

As is now shown, the rate of $\hat{D}^2_j(\tilde{X})$ computed on the full sample is slower than the one for $\hat{\Delta}^2_j(\tilde{R}, \tilde{S})$ in the two samples setting.

The reason is that we cannot always apply lemma 4.10 allowing to reduce double sums in k_1, k_2 to a sum on the diagonal $k_1 = k_2$ for translates of the same φ functions. Indeed, when

a match between multi indices i_1 and i_2 involves terms corresponding to margins, it is not guaranteed that a match on observation numbers also corresponds to a match on margin numbers ; that is to say, in the product $\varphi(X^{\ell_1}-k_1)\varphi(X^{\ell_2}-k_2)$, only once in a while $\ell_1=\ell_2$; so most of the time we can say nothing about the support of the product, and the sum spans many more terms, hence the additional factor 2^j in the risk bound for $\hat{D}_j^2$ on the full sample.

Proposition 4.11 (Risk upper bound, full sample — f_A)

Let $X_1,\dots,X_n$ be an independent, identically distributed sample of f_A. Assume that f is compactly supported and that φ is a Daubechies D2N. Let $\hat{D}_j^2$ be the U-statistic estimator defined in (17), Then

$$E_{f_A}^n\left(\hat{D}_j^2(\tilde{X})-\sum_{k\in\mathbb{Z}^d}\delta_{jk}^2\right)^2\leq C2^{jd}n^{-2}+C^\star 2^j n^{-1}$$

with δ_{jk} the coordinate of $f_A-f_A^\star$ and $C^\star=0$ at independence, when $f_A=f_A^\star$.

$E_{f_A}^n\left[\hat{D}_j^2(\tilde{X})-\sum_{k\in\mathbb{Z}^d}\delta_{jk}^2\right]^2=E_{f_A}^n[\hat{D}_j^2(\tilde{X})]^2-\left(\sum_{k\in\mathbb{Z}^d}\delta_{jk}^2\right)^2.$

To make $\hat{D}_j^2(\tilde{X})$ look more like the usual U-estimator of $\int(f-g)^2$ for unrelated f and g, we define for $i\in I_n^{2d+2}$, the dummy slice variables $Y_i=X_{i^1}$, $V_i=(X_{i^2},\dots X_{i^{d+1}})$, $Z_i=X_{i^{d+2}}$, $T_i=(X_{i^{d+3}},\dots X_{i^{2d+2}})$; so that Y_i and Z_i have distribution P_{f_A}, V_i and T_i have distribution $P_{f_A^\star}=P_{f_A^{\star 1}}\dots P_{f_A^{\star d}}$ (once canonically projected), and Y_i, V_i, Z_i, T_i are independent variables under $P_{f_A}^n$. Next, for $k\in\mathbb{Z}^d$, define the function Λ_{jk} as

$$\begin{aligned}\Lambda_{jk}(X_{i^1},\dots,X_{i^d})&=\varphi_{jk^1}(X_{i^1}^1)\dots\varphi_{jk^d}(X_{i^d}^d)\quad\forall i\in\Omega_n^d\\ \Lambda_{jk}(X_i)=\Phi_{jk}(X_i)&=\varphi_{jk^1}(X_i^1)\dots\varphi_{jk^d}(X_i^d)\quad\forall i\in\Omega_n^1=\{1\dots,n\}\end{aligned}\tag{23}$$

with second line taken as a convention.

So that $\hat{D}_j^2(\tilde{X})$ can be written under the more friendly form

$$\hat{D}_j^2(\tilde{X})=\frac{1}{A_n^{2d+2}}\sum_{i\in I_n^{2d+2}}\sum_{k\in\mathbb{Z}^d}\big[\Lambda_{jk}(Y_i)-\Lambda_{jk}(V_i)\big]\big[\Lambda_{jk}(Z_i)-\Lambda_{jk}(T_i)\big],$$

with $I_n^m=\big\{(i^1,\dots,i^m)\colon i^\ell\in\mathbb{N},\,1\leq i^\ell\leq n,\,i^{\ell_1}\neq i^{\ell_2}\ \text{ if } \ell_1\neq\ell_2\big\}$.

Following the friendly notation, let $h_{ik}=\big[\Lambda_{jk}(Y_i)-\Lambda_{jk}(V_i)\big]\big[\Lambda_{jk}(Z_i)-\Lambda_{jk}(T_i)\big]$ be the kernel of $\hat{D}_j^2(\tilde{X})$ at fixed k. Then,

$$[\hat{D}_j^2(\tilde{X})]^2=|I_n^{2d+2}|^{-2}\sum_{i_1,i_2\in I_n^{2d+2}\times I_n^{2d+2}}\ \sum_{k_1,k_2\in\mathbb{Z}^d\times\mathbb{Z}^d}h_{i_1k_1}h_{i_2k_2}.$$

Consider the partitioning sets $M_c=\{i_1,i_2\in I_n^{2d+2}\times I_n^{2d+2}\colon|i_1\cap i_2|=c\}$, $c=0\dots,2d+2$, that is to say the set of pairs with c coordinates in common. Equivalently, M_c can be defined as the set $\{i_1,i_2\in I_n^{2d+2}\times I_n^{2d+2}\colon|i_1\cup i_2|=4d+4-c\}$.

According to the partitioning, with $h_i = \sum_k h_{ik}$,

$$E^n_{f_A}[\hat{D}^2_j(\tilde{X})]^2 = |I^{2d+2}_n|^{-2} \sum_{c=0}^{2d+2} \sum_{(i_1,i_2)\in M_c} E^n_{f_A} h_{i_1} h_{i_2}.$$

Let $\lambda_{jk} = \alpha_{jk^1} \dots \alpha_{jk^d}$ and $\delta_{jk} = \alpha_{jk} - \lambda_{jk}$.

- On M_0, with no match,

$$E^n_{f_A} h_{i_1} h_{i_2} \mathbb{I}\{M_0\} = \sum_{k_1,k_2} (\alpha_{jk_1} - \lambda_{jk_1})^2 (\alpha_{jk_2} - \lambda_{jk_2})^2 = \left(\sum_k \delta^2_{jk}\right)^2$$

By lemma 4.6, the ratio $|M_0|/|I^{2d+2}_n|$ is lower than $1 + Cn^{-2}$. So that

$|I^{2d+2}_n|^{-2} \sum_{M_0} E^n_{f_A} h_{i_1} h_{i_2} - \left(\sum_k \delta^2_{jk}\right)^2 = |I^{2d+2}_n|^{-2} |M_0| E^n_{f_A} h_{i_1} h_{i_2} \mathbb{I}\{M_0\} \le Cn^{-2}$.

- On M_1, assuming the match involves Y_{i_1} and Y_{i_2},

$$\begin{aligned} E^n_{f_A} h_{i_1} h_{i_2} \mathbb{I}\{M_1\} &= \sum_{k_1,k_2} \delta_{jk_1}\delta_{jk_2} E^n_{f_A}\big(\Phi_{jk_1}(Y_{i_1}) - \Lambda_{jk_1}(V_{i_1})\big)\big(\Phi_{jk_2}(Y_{i_2}) - \Lambda_{jk_2}(V_{i_2})\big) \\ &= \sum_{k_1,k_2} \delta_{jk_1}\delta_{jk_2}\big(E^n_{f_A}\Phi_{jk_1}(X)\Phi_{jk_2}(X) - \lambda_{jk_1}\alpha_{jk_2} - \delta_{jk_1}\lambda_{jk_2}\big) \\ &= E^n_{f_A}\left(\sum_k \delta_{jk}\Phi_{jk}(X)\right)^2 - C^2_j \sum_k \lambda_{jk}\delta_{jk} - \left(\sum_k \lambda_{jk}\delta_{jk}\right)\left(\sum_k \alpha_{jk}\delta_{jk}\right) \end{aligned} \tag{24}$$

with $C^2_j = \sum_k \delta^2_{jk}$.

Next by (44) in lemma 4.10 for the first line, the double sum in k under expectation is bounded by a constant times the sum restricted to the diagonal $k_1 = k_2$ because of the limited overlapping of translates φ_{jk} ; using also lemma 4.11,

$$E^n_{f_A}\left(\sum_k \delta_{jk}\Phi_{jk}(X)\right)^2 \le (4N-3)^d \sum_k \delta^2_{jk} E^n_{f_A}\Phi_{jk}(X)^2 \le (4N-3)^d \sum_k C\delta^2_{jk}.$$

Since all other terms in (24) are clearly bounded by a constant not depending on j, we conclude by symmetry that $E^n_{f_A} h_{i_1} h_{i_2} \mathbb{I}\{M_1\} \le C$ for any match of cardinality 1 between narrow slices (Y_{i_1} Y_{i_2} or Z_{i_1} Z_{i_2} or Y_{i_1} Z_{i_2} or Z_{i_1} Y_{i_2}). Moreover $C = 0$ when $f_A = f^*_A$ *i.e.* at independence, because of the omnipresence of δ_{jk}, the coordinate of $f_A - f^\star_A$.

- On M_1, if the match is between Y_{i_1} and V_{i_2}, a calculus as in (24) yields,

$$E^n_{f_A} h_{i_1} h_{i_2} \mathbb{I}\{M_1\} = -\sum_{k_1,k_2} \delta_{jk_1}\delta_{jk_2} E^n_{f_A}\Phi_{jk_1}(Y_{i_1})\Lambda_{jk_2}(V_{i_2}) + C^2_j \sum_k \alpha_{jk}\delta_{jk} + \left(\sum_k \lambda_{jk}\delta_{jk}\right)^2 ;$$

which can also be found from line 2 of (24) using the swap $\Phi_{jk}(Y_{i_2}) \longleftrightarrow -\Lambda_{jk}(V_{i_2})$ and $\alpha_{jk} \longleftrightarrow -\lambda_{jk}$.

Next, for some $\ell \in \{1, \dots, d\}$,

$$\sum_{k_1,k_2} \delta_{jk_1}\delta_{jk_2} E^n_{f_A} \Phi_{jk_1}(Y_{i_1})\Lambda_{jk_2}(V_{i_2}) = \sum_{k_1,k_2} \delta_{jk_1}\delta_{jk_2}\lambda^{\langle d-1\rangle}_{jk_2} E^n_{f_A}\Phi_{jk_1}(X)\varphi_{jk_2^\ell}(X^\ell)$$

with special notation $\lambda^{\langle r\rangle}_{jk} = \alpha^{p_1}_{jk^1}\dots\alpha^{p_d}_{jk^d}$ for some p_i, $0 \le p_i \le r$, $\sum_{i=1}^d p_i = r$.

In the present case $\Phi_{jk_1}(X)\varphi_{jk_2^\ell}(X^\ell) = \Phi_{jk_1}(X)\varphi_{jk_2^\ell}(X^\ell)\,\mathbb{I}\left\{|k_1^\ell - k_2^\ell| < 2N-1\right\}$ does not give any useful restriction of the double sum because the coefficient α_{jk} hidden in δ_{jk} is not guaranteed to factorize under any split of dimension unless $A = I$; and lemma 4.10 is useless. This is a difficulty that did not raise in propositions 4.8 and 4.9 because we could use the fact that these kind of terms were estimated over independent samples.

Instead write $E^n_{f_A}|\Phi_{jk_1}(X)\varphi_{jk_2}(X^\ell)| \le 2^{\frac{j}{2}}\|\varphi\|_\infty E^n_{f_A}|\Phi_{jk_1}(X)| \le C2^{\frac{j}{2}}2^{-\frac{jd}{2}}$ using lemma 4.11. So that when multiplied by $\sum_k \delta_{jk}\sum_k \delta_{jk}\lambda^{\langle d-1\rangle}_{jk}$, using Meyer's lemma, the final order is 2^j.

By symmetry, for any match of cardinality 1 between a narrow and a wide slice (Y or T or equivalent pairing), $E^n_{f_A}|h_{i_1}h_{i_2}|\,\mathbb{I}\{M_1\} \le C2^j$, with $C = 0$ at independence.

- On M_1, if the match is between V_{i_1} and V_{i_2}, by symmetry with (24) or using the swap defined above,

$$E^n_{f_A} h_{i_1}h_{i_2}\,\mathbb{I}_{\{M_1\}} = \sum_{k_1,k_2}\delta_{jk}\delta_{jk'}E^n_{f_A}\Lambda_{jk}(V_i)\Lambda_{jk'}(V_{i_2}) - C_j^2\sum_k \alpha_{jk}\delta_{jk} - \left(\sum_k \lambda_{jk}\delta_{jk}\right)\left(\sum_k \alpha_{jk}\delta_{jk}\right),$$

and for some not necessarily matching $\ell_1, \ell_2 \in \{1, \dots, d\}$ (*i.e.* lemma 4.10 not applicable),

$$\begin{aligned}\sum_{k_1,k_2}\delta_{jk_1}\delta_{jk_2}E^n_{f_A}\Lambda_{jk_1}(V_{i_1})\Lambda_{jk_2}(V_{i_2}) &= \sum_{k_1,k_2}\delta_{jk_1}\delta_{jk_2}\lambda^{\langle d-1\rangle}_{jk_1}\lambda^{\langle d-1\rangle}_{jk_2}E^n_{f_A}\varphi_{jk_1^{\ell_1}}(X^{\ell_1})\varphi_{jk_2^{\ell_2}}(X^{\ell_2})\\ &\le \left(\sum_k \delta_{jk}\lambda^{\langle d-1\rangle}_{jk}\right)^2 = C2^j\end{aligned}$$

with last line using Meyer's lemma, and having reduced the term under expectation to a constant by Cauchy-Schwarz inequality and lemma 4.11.

And we conclude again that, for any match of cardinality 1 between two wide slices (V or T or equivalent), $E^n_{f_A}h_{i_1}h_{i_2}\,\mathbb{I}\{M_1\} \le C2^j$, with $C = 0$ at independence.

By lemma 4.6, the ratio $|M_1|/|I_n^{2d+2} \times I_n^{2d+2}| \approx n^{-1}$, so in summary, the bound for M_1 has the order $C^\star 2^j n^{-1}$, with $C^\star = 0$ at independence.

- On M_c, $c = 2\dots 2d+2$.

Fix the pair of indexes $(i_1, i_2) \in I_n^{2d+2} \times I_n^{2d+2}$, we need to bound a term having the form

$$Q(i_1, i_2) = E^n_{f_A}\sum_{k_1,k_2}\Lambda_{jk}(R_{i_1})\Lambda_{jk}(S_{i_1})\Lambda_{jk_2}(R'_{i_2})\Lambda_{jk_2}(S'_{i_2})$$

where both slices $R_{i_1} \neq S_{i_1}$ unrelated with both slices $R'_{i_2} \neq S'_{i_2}$ are chosen among any of the dummy Y, V, Z, T.

— *Narrow slices only.* For a match spanning four narrow slices exclusively, that is to say $(Y_{i_1} = Y_{i_2}) \cap (Z_{i_1} = Z_{i_2})$ or $(Y_{i_1} = Z_{i_2}) \cap (Z_{i_1} = Y_{i_2})$, a case possible on M_2 only, the general term of higher order is written $\sum_{k_1,k_2} E^n_{f_A} \Phi_{jk_1}(X)\Phi_{jk_2}(X) E^n_{f_A} \Phi_{jk_1}(X)\Phi_{jk_2}(X)$. By lemma 4.10 this is again lower than $(4n-3)^d \sum_k \left[E^n_{f_A} \Phi_{jk}(X)^2\right]^2$, that is $C2^{jd}$. By lemma 4.6, this case thus contributes to the general bound up to $C2^{jd}n^{-2}$.

Three narrow slices only is not possible and two narrow slices correspond to the case M_1 treated above.

— *Wide slices only.* For a match spanning wide slices on M_c, $c = 2, \dots 2d$, a general term with higher order is written $\sum_{k_1,k_2} E^n_{f_A} \Lambda_{jk_1}(V_{i_1})\Lambda_{jk_1}(T_{i_1})\Lambda_{jk_2}(V_{i_2})\Lambda_{jk_2}(T_{i_2})$, with $|i_1 \cap i_2| = c$, (an equivalent is obtained by swapping one V with a T). Since the slices are wide, it is not possible to distribute expectation any further right now : if V_{i_1} is always independent of T_{i_1}, both terms may depend on V_{i_2}, say. Also matching coordinates on i_1, i_2 do not necessarily correspond to matching dimensions X^ℓ of the observation, and then lemma 4.10 is not applicable. Instead write,

$$Q(i_1, i_2) = \sum_{k_1,k_2} \lambda^{\langle 2d-c \rangle}_{jk_1} \lambda^{\langle 2d-c \rangle}_{jk_2} E^n_{f_A} \left[\Lambda^{\langle c \rangle}_{jk_1}(V_{i_1}, T_{i_1}) \Lambda^{\langle c \rangle}_{jk_2}(V_{i_2}, T_{i_2})\right],$$

with $\Lambda^{\langle c \rangle}_{jk}(V_i, T_i)$ a product of c independent terms of the form $\varphi_{jk^\ell}(X^\ell)$ spanning at least one of the slices V_i, T_i.

By definition of i_1 and i_2, the product of $2c$ terms under expectation can be split into c independent products of two terms. So, using $E^n_{f_A}|\varphi_{jk^\ell}(X)^2| \leq C$ on each bi-term, the order at the end is $C\big(\sum_k \lambda^{\langle 2d-c \rangle}_{jk}\big)^2$; and using Meyer's lemma, the bound is then of the order of $C2^{jc}$.

Finally, using lemma 4.6 as above, the contribution of this kind of term to the general bound is $\sum_{c=1}^{2d} 2^{jc} n^{-c}$.

On $\{2^j < n\} \supset \{2^{jd} < n^2\} \supset \{2^{jd} < n\}$, this quantity is bounded by $C2^j n^{-1} < C2^{jd} n^{-2}$ and on $\{2^j > n\}$ it is unbounded.

— *Narrow and wide slices* Reusing the general pattern above, with $c_w \leq 2d$ matching coordinates on wide slices and $c_r \leq 2$ on narrow slices

$$Q(i_1, i_2) = \sum_{k_1,k_2} \lambda^{\langle 2d-c_w \rangle}_{jk_1} \lambda^{\langle 2d-c_w \rangle}_{jk_2} \alpha^{2-c_r}_{jk_1} \alpha^{2-c_r}_{jk_2} E^n_{f_A} \left[\Lambda^{\langle c \rangle}_{jk_1}(Y_{i_1}, V_{i_1}, Z_{i_1}, T_{i_1}) \Lambda^{\langle c \rangle}_{jk_2}(Y_{i_2}, V_{i_2}, Z_{i_2}, T_{i_2})\right],$$

with $\Lambda^{\langle c \rangle}_{jk}(Y_i, V_i, Z_i, T_i)$ a product of c independent terms of the form $\varphi_{jk^\ell}(X)$ or $\Phi_{jk}(X)$ spanning at least one of the slices V_i, T_i and one of the slices Y_i, Z_i. As above, the bracket is a product of independent bi-terms, each under expectation bounded by some constant C, by lemma 4.11, using Cauchy-schwarz inequality if needed. So this is bounded by

$$Q(i_1, i_2) \leq C \sum_{k_1,k_2} \lambda^{\langle 2d-c_w \rangle}_{jk_1} \lambda^{\langle 2d-c_w \rangle}_{jk_2} \alpha^{2-c_r}_{jk_1} \alpha^{2-c_r}_{jk_2} = C\Big(\sum_k \lambda^{\langle 2d-c_w \rangle}_{jk} \alpha^{2-c_r}_{jk}\Big)^2 ;$$

using Cauchy-Schwarz inequality and Meyer's lemma this is bounded by the quantity $2^{\frac{j}{2}(c_w-d)}2^{\frac{jd}{2}(c_r-1)}$ and, with lemma 4.6, the contribution to the general bound on $\{2^j < n^2\} \supset \{2^{jd} < n^2\}$ is

$$2^{-jd}\sum_{a=1}^{2}\sum_{b=1}^{2d} 2^{\frac{jb}{2}} n^{-b} 2^{\frac{jda}{2}} n^{-a}\, \mathbb{I}\left\{2^j < n^2\right\} \leq Cn^{-1}$$

Finally on $\{2^{jd} < n^2\}$, $E^n_{f_A}\hat{B}^2_j - \left(\sum_k \delta^2_{jk}\right)^2 \leq C^\star 2^j n^{-1} + 2^{jd}n^{-2}$. $\square$

4.4 Appendix 1 – Propositions

Proposition 4.12 (2nd moment of $\sum_k \hat{\alpha}^2_{jk}$ about $\sum_k \alpha^2_{jk}$)

Let $X_1, \ldots, X_n$ be an independent, identically distributed sample of f, a compactly supported function defined on $\mathbb{R}^d$. Let $\hat{\alpha}_{jk} = \frac{1}{n}\sum_{i=1}^n \varphi_{jk^1}(X^1_i)\ldots\varphi_{jk^d}(X^d_i)$, $k \in \mathbb{Z}^d$. Assume that φ is a Daubechies D2N.

Then $E^n_{f_A}\left(\sum_k \hat{\alpha}^2_{jk} - \sum_k \alpha^2_{jk}\right)^2 = C2^{jd}n^{-1} + C2^{2jd}n^{-2}\,\mathbb{I}\left\{2^{jd} > n\right\}$

For the mean, using lemma 4.11,

$$\begin{aligned} E^n_{f_A}\sum_k \hat{\alpha}^2_{jk} &= \frac{1}{n^2}\sum_{i_1=i_2}\sum_k E^n_{f_A}\Phi_{jk}(X_{i_1})\Phi_{jk}(X_{i_2}) + \frac{1}{n^2}\sum_{i_1\neq i_2}\sum_k \alpha^2_{jk} \\ &= \frac{1}{n}\sum_k \Phi_{jk}(X_i)^2 + \frac{n-1}{n}\sum_k \alpha^2_{jk} = \sum_k \alpha^2_{jk} + O(\frac{2^{jd}}{n}). \end{aligned}$$

For the second moment, let $M_c = \{i_1, i_2, i_3, i_4 \in \{1, \ldots, n\}\colon |\{i_1\} \cup \ldots \cup \{i_4\}| = c\}$.

$$E^n_{f_A}\big(\sum_k \hat{\alpha}^2_{jk}\big)^2 = \frac{1}{n^4}\sum_{c=1}^{4}\sum_{i_1,\ldots,i_4} E^n_{f_A}\sum_{k_1,k_2}\Phi_{jk_1}(X_{i_1})\Phi_{jk_1}(X_{i_2})\Phi_{jk_2}(X_{i_3})\Phi_{jk_2}(X_{i_4})\,\mathbb{I}\{M_c\}$$

On $c = 1$, the kernel is equal to $\sum_{k_1,k_2}\Phi_{jk_1}(X)^2\Phi_{jk_2}(X)^2 \leq (4N-3)^d\sum_k \Phi_{jk}(X)^4$ by lemma 4.10. And by lemma 4.11, $E^n_{f_A}\sum_k \Phi_{jk}(X)^4 \leq \sum_k C2^{jd} = C2^{2jd}$.

On $c = 2$, the kernel takes three generic forms : (a) $\sum_{k_1,k_2}\Phi_{jk_1}(X)\Phi_{jk_1}(Y)\Phi_{jk_2}(X)\Phi_{jk_2}(Y)$ or (b) $\sum_{k_1,k_2}\Phi_{jk_1}(X)^2\Phi_{jk_2}(Y)^2$ or (c) $\sum_{k_1,k_2}\Phi_{jk_1}(X)\Phi_{jk_1}(Y)\Phi_{jk_2}(Y)^2$. In cases (a) and (c), using lemma 4.10, the double sum can be reduced to the diagonal $k_1 = k_2$. So using also lemma 4.11,

$$(a)\quad E^n_{f_A}|\sum_{k_1,k_2}\Phi_{jk_1}(X)\Phi_{jk_1}(Y)\Phi_{jk_2}(X)\Phi_{jk_2}(Y)| \leq E^n_{f_A}(4N-3)^d\sum_k \Phi_{jk}(X)^2\Phi_{jk}(Y)^2 \leq C2^{jd}$$

$$(b)\quad E^n_{f_A}\sum_{k_1,k_2}\Phi_{jk_1}(X)^2\Phi_{jk_2}(Y)^2 \leq C2^{2jd}$$

$$(c)\quad E^n_{f_A}|\sum_{k_1,k_2}\Phi_{jk_1}(X)\Phi_{jk_1}(Y)\Phi_{jk_2}(Y)^2| \leq E^n_{f_A}(4N-3)^d\sum_k |\Phi_{jk}(X)\Phi_{jk}(Y)^3| \leq C2^{jd}.$$

On $c=3$ the only representative form is

$$E^n_{f_A}\sum_{k_1,k_2}\Phi_{jk_1}(X)\Phi_{jk_1}(Y)\Phi_{jk_2}(Z)^2=\sum_k\alpha^2_{jk}\sum_k E^n_{f_A}\Phi_{jk}(X)^2\le C2^{jd},$$

and on $c=4$ the statistic is unbiased equal to $\left(\sum_k\alpha^2_{jk}\right)^2$ under expectation.

Next, since $|M_4|=A^4_n$ and, using lemma 4.5, $|M_c|=O(n^c)$,

$$\begin{aligned}E^n_{f_A}\Big(\sum_k\hat\alpha^2_{jk}\Big)^2&\le A^4_n n^{-4}\Big(\sum_k\alpha^2_{jk}\Big)^2+C2^{2jd}n^{-3}+n^{-2}2^{2jd}+n^{-1}2^{jd}\\&\le\Big(\sum_k\alpha^2_{jk}\Big)^2+Cn^{-2}+Cn^{-1}2^{jd}\mathbb{I}\left\{2^{jd}<n\right\}+Cn^{-2}2^{2jd}\mathbb{I}\left\{2^{jd}>n\right\}\end{aligned}$$

with $A^4_n n^{-4}=1-\frac{6}{n}+Cn^{-2}$.

Finally

$$\begin{aligned}E^n_{f_A}\Big(\sum_k\hat\alpha^2_{jk}-\sum_k\alpha^2_{jk}\Big)^2&=E^n_{f_A}\Big(\sum_k\hat\alpha^2_{jk}\Big)^2+\Big(\sum_k\alpha^2_{jk}\Big)^2-2E^n_{f_A}\sum_k\hat\alpha^2_{jk}\sum_k\alpha^2_{jk}\\&\le Cn^{-2}+Cn^{-1}2^{jd}+Cn^{-2}2^{2jd}\mathbb{I}\left\{2^{jd}>n\right\}\end{aligned}$$

□

Proposition 4.13 (2nd moments $\sum_k\hat\lambda^2_{jk}$ about $\sum_k\lambda^2_{jk}$ and $\sum_k\hat\lambda_{jk}\hat\alpha_{jk}$ about $\sum_k\lambda_{jk}\alpha_{jk}$)

Let $X_1,\dots,X_n$ be an independent, identically distributed sample of f, a compactly supported function defined on $\mathbb{R}^d$. Let $\hat\lambda_{jk}=\frac{1}{n^d}\sum_{i=1}^n\varphi_{jk^1}(X^1_i)\dots\sum_{i=1}^n\varphi_{jk^d}(X^d_i)$, $k\in\mathbb{Z}^d$. Assume that φ is a Daubechies D2N.

Then on $\{2^{jd}<n^2\}$

$$E^n_{f_A}\Big(\sum_k\hat\lambda_{jk}\hat\alpha_{jk}-\sum_k\lambda_{jk}\alpha_{jk}\Big)^2\le O(n^{-2})+C\frac{2^j}{n}$$

$$E^n_{f_A}\Big(\sum_k\hat\lambda^2_{jk}-\sum_k\lambda^2_{jk}\Big)^2\le O(n^{-2})+C\frac{2^j}{n}$$

$$E^n_{f_A}\Big(\sum_k\hat\lambda^2_{jk}-\lambda^2_{jk}\Big)^2=E^n_{f_A}\left[\Big(\sum_k\hat\lambda^2_{jk}\Big)^2-2\sum_k\lambda^2_{jk}\sum_k\hat\lambda^2_{jk}+\Big(\sum_k\lambda^2_{jk}\Big)^2\right]$$

For $i\in\Omega^{2d}_n$, let V_i be the slice $(X^1_{i^1},X^1_{i^2},\dots,X^d_{i^{2d-1}},X^d_{i^{2d}})$. Let the coordinate-wise kernel function Λ_{jk} be given by $\Lambda_{jk}(V_i)=\varphi_{jk^1}(X^1_{i^1})\varphi_{jk^1}(X^1_{i^2})\dots\varphi_{jk^d}(X^d_{i^{2d-1}})\varphi_{jk^d}(X^d_{i^{2d}})$.

Let $|i|$ be the shortcut notation for $|\{i^1\}\cup\dots\cup\{i^{2d}\}|$. Let $W^{2d}_n=\{i\in\Omega^{2d}_n\colon|i|<2d\}$, that is to say the set of indices with at least one repeated coordinate.

Then the mean term is written

$$\begin{aligned}
E^n_{f_A}\sum_k \hat{\lambda}^2_{jk} &= n^{-2d}\sum_{i\in\Omega^{2d}_n}\sum_k \Lambda_{jk}(V_i)\\
&= n^{-2d}\sum_{W^{2d}_n}\sum_k E^n_{f_A}\Lambda_{jk}(V_i) + A^{2d}_n n^{-2d}\sum_k \lambda^2_{jk}\\
&= Q_1 + A^{2d}_n n^{-2d}\theta
\end{aligned}$$

Let $M_c = \{i \in \Omega^{2d}_n : |i| = c\}$ be the set indices with c common coordinates. So that Q_1 is written

$$Q_1 = n^{-2d}\sum_{c=1}^{2d-1}\mathbb{I}\{M_c\}\sum_{M_c}\sum_k E^n_{f_A}\Lambda_{jk}(V_i) = \sum_k Q_{1jk}$$

By lemma 4.7 with lemma parameters $(d = 1, m = 2d, r = 1)$, $E^n_{f_A}|\Lambda_{jk}(V_i)|\,\mathbb{I}\{M_c\} \le C2^{\frac{j}{2}(2d-2c)}$ and by lemma 4.5, $|M_c| = O(n^c)$. Hence,

$$Q_{1jk} \le \sum_{c=1}^{2d-1} n^{-2d+c}C2^{j(d-c)} = 2^{-jd}\sum_{c=1}^{2d-1} C\left(\frac{2^j}{n}\right)^{(2d-c)}$$

which on $\{2^{jd} < n\}$ has maximum order $2^{j(1-d)}n^{-1}$ when $d - c$ is minimum *i.e.* $c = 2d - 1$. Finally $|Q_1| \le \sum_k C2^{j(1-d)}n^{-1} \le C2^j n^{-1}$.

Next, the second moment about zero is written

$$\begin{aligned}
E^n_{f_A}\Big(\sum_k \hat{\lambda}^2_{jk}\Big)^2 &= n^{-4d}\sum_{i_1,i_2\in(\Omega^{2d}_n)^2}\sum_{k_1,k_2}\Lambda_{jk_1}(V_{i_1})\Lambda_{jk_2}(V_{i_2})\\
&= n^{-4d}\sum_{W^{4d}_n}\sum_{k_1,k_2} E^n_{f_A}\Lambda_{jk_1}(V_{i_1})\Lambda_{jk_2}(V_{i_2}) + A^{4d}_n n^{-4d}\Big(\sum_k \lambda^2_{jk}\Big)^2\\
&= Q_2 + A^{4d}_n n^{-4d}\theta^2
\end{aligned}$$

with $W^{4d}_n = \{i_1, i_2 \in (\Omega^{2d}_n)^2 : |i_1 \cup i_2| < 4d\}$, that is to say the set of indices with at least one repeated coordinate somewhere.

Let this time $M_c = \{i_1, i_2 \in (\Omega^{2d}_n)^2 : |i_1 \cup i_2| = c\}$ be the set indices with overall c common coordinates in i_1 and i_2. So that Q_2 is written

$$Q_2 = n^{-4d}\sum_{c=1}^{4d-1}\mathbb{I}\{M_c\}\sum_{M_c}\sum_{k_1,k_2} E^n_{f_A}\Lambda_{jk_1}(V_{i_1})\Lambda_{jk_2}(V_{i_2}) = \sum_{k_1,k_2} Q_{2j_1k_1j_2k_2}$$

By lemma 4.9, unless $c = 1$, it is always possible to find indices i_1, i_2 with no match between the observations falling under k_1 and those falling under k_2, so that there is no way to reduce the double sum in k_1, k_2 to a sum on the diagonal using lemma 4.10. Note that if $c = 1$, $E^n_{f_A}\Lambda_{jk}(V_i)\Lambda_{jk}(V_i) = E^n_{f_A}\Phi_{jk}(X)^4$ has order $C2^{jd}$.

So coping with the double sum, by lemma 4.7 with lemma parameters $(d = 1, m = 2d, r = 2)$, $E^n_{f_A}|\Lambda_{jk}(V_{i_1})\Lambda_{jk}(V_{i_2})| \le C2^{\frac{j}{2}(4d-2c)}$, and again by lemma 4.5 $|M_c| = O(n^c)$, so $E^n_{f_A}|Q_{2j_1k_1j_2k_2}| \le$

$\sum_{c=1}^{4d-1} n^{c-4d} C 2^{\frac{j}{2}(4d-2c)}$, which on $\{2^{jd} < n\}$ has maximum order $2^{j(1-2d)}n^{-1}$ when $c = 4d-1$. Finally, $E^n_{f_A} Q_2 \leq \sum_{k_1,k_2} C 2^{j(1-2d)} n^{-1} \leq C 2^j n^{-1}$.

Putting all together, and since $A^p_n n^{-p} = 1 - \frac{(d+1)(d+2)}{2n} + O(n^{-2})$,

$$\begin{aligned} E^n_{f_A}\Big(\sum_k \hat{\lambda}^2_{jk} - \lambda^2_{jk}\Big)^2 &= Q_2 + A^{4d}_n n^{-4d}\theta^2 - 2\theta(Q_1 + A^{2d}_n n^{-2d}\theta) + \theta^2 \\ &= Q_2 - 2\theta Q_1 + \theta^2(1 + A^{4d}_n n^{-4d} - 2A^{2d}_n n^{-2d}) \leq |Q_2| + 2\theta|Q_1| + O(n^{-2}) \\ &\leq C 2^j n^{-1} \end{aligned}$$

For the cross product,

As above, for $i \in \Omega^{d+1}_n$, let V_i be the slice $(X_{i^0}, X^1_{i^1}, \ldots, X^d_{i^d})$. Let the coordinate-wise kernel function Λ_{jk} be given by $\Lambda_{jk}(V_i) = \Psi_{jk}(X_{i^0})\psi_{jk^1}(X^1_{i^1})\ldots\psi_{jk^d}(X^d_{i^d})$. Let $\theta = \sum_k \alpha_{jk}\lambda_{jk}$.

Let $W^{d+1}_n = \{i \in \Omega^{d+1}_n : |i| < d+1\}$, that is to say the set of indices with at least one repeated coordinate.

So that, $E^n_{f_A} \sum_k \hat{\alpha}_{jk}\hat{\lambda}_{jk} = Q_1 + A^{d+1}_n n^{-d-1}\theta$ with $Q_1 = n^{-d-1}\sum_{W^{d+1}_n}\sum_k E^n_{f_A}\Lambda_{jk}(V_i)$ and likewise

$$E^n_{f_A}\Big(\sum_k \hat{\alpha}_{jk}\hat{\lambda}_{jk}\Big)^2 = Q_2 + A^{2d+2}_n n^{-2d-2}\theta^2$$

with $Q_2 = n^{-2d-2}\sum_{W^{2d+2}_n}\sum_{k_1,k_2} E^n_{f_A}\Lambda_{jk_1}(V_{i_1})\Lambda_{jk_2}(V_{i_2})$. And we obtain in the same way,

$$E^n_{f_A}\Big(\sum_k \hat{\alpha}_{jk}\hat{\lambda}_{jk} - \alpha_{jk}\lambda_{jk}\Big)^2 \leq |Q_2| + 2\theta|Q_1| + O(n^{-2})$$

Let $M_c = \{i \in \Omega^{d+1}_n : |i| = c\}$ be the set indices with c common coordinates. So that Q_1 is written

$$Q_1 = n^{-d-1}\sum_{c=1}^{d} \mathbb{I}\{M_c\}\sum_{M_c}\sum_k E^n_{f_A}\Lambda_{jk}(V_i) = \sum_k Q_{1jk}$$

By lemma 4.7 with lemma parameters $(m_d = 1, m_1 = d, r = 1)$,

$$E^n_{f_A}|\Lambda_{jk}(V_i)|\,\mathbb{I}\{M_c\} \leq C 2^{\frac{jd}{2}(1-2c_d)} 2^{\frac{j}{2}(d-2c_1)}$$

with $c_1 + c_d = c$, $0 \leq c_1 \leq d$, $1 \leq c_d \leq 1$ and by lemma 4.5, $|M_c| = O(n^c)$. Hence,

$$Q_{1jk} \leq \sum_{c=1}^{d} n^{-d-1+c} C 2^{j(d-dc_d-c_1)} = 2^{j(-1+(1-d)c_d)}\sum_{c=1}^{d} C\left(\frac{2^j}{n}\right)^{(d+1-c)}$$

which on $\{2^{jd} < n\}$ has maximum order $C 2^{j(1-d)}n^{-1}$ when $d+1-c$ is minimum *i.e.* $c = d$. Finally $|Q_1| \leq \sum_k C 2^{j(1-d)} n^{-1} \leq C 2^j n^{-1}$.

Next, as above $Q_2 = \sum_{k_1,k_2} Q_{2jk_1jk_2}$, and again by lemma 4.9, unless $c = 1$, it is always possible to find indices i_1, i_2 with no matching coordinates corresponding also to matching

dimension number, so that there is no way to reduce the double sum in k_1, k_2 to a sum on the diagonal using lemma 4.10.

So coping once more with the double sum, by lemma 4.7 with lemma parameters ($m_d = 1, m_1 = d, r = 2$), $E^n_{f_A}|\Lambda_{jk}(V_{i_1})\Lambda_{jk}(V_{i_2})| \leq C2^{\frac{jd}{2}(2-2c_d)}2^{\frac{j}{2}(2d-2c_1)}$, with $c_1 + c_d = c$, $1 \leq c_d \leq 2$, $0 \leq c_1 \leq 2d$, and again by lemma 4.5 $|M_c| = O(n^c)$, so

$$E^n_{f_A}|Q_{2j_1k_1j_2k_2}| \leq \sum_{c=1}^{2d+1} n^{c-2d-2}C2^{j(d-dc_d+d-c_1)} = 2^{j(-2+(1-d)c_d)}\sum_{c=1}^{2d+1} C\left(\frac{2^j}{n}\right)^{(2d+2-c)},$$

which on $\{2^{jd} < n\}$ has maximum order $C2^{-jd}n^{-1}$ when $c = 2d+1$. Then either $c_d = 1$, which means that the two terms $\Phi_{jk_1}(X_{i_1})\Phi_{jk_2}(X_{i_2})$ match on the observation number, in which case the sum in k_1, k_2 can be reduced ; either $c_d = 2$. In the first case the order is $E^n_{f_A}Q_2 \leq (4N-3)^d\sum_k C2^{-jd}n^{-1} \leq Cn^{-1}$ and in the second case $E^n_{f_A}Q_2 \leq \sum_{k_1,k_2} C2^{1-2jd}n^{-1} \leq C2^jn^{-1}$.
□

Proposition 4.14 (Variance of $\sum_k \hat{B}^2_j$)

Let $\{X_1, \ldots, X_n\}$ be an i.i.d. sample with density f. Assume that f is compactly supported and that φ is a Daubechies $D2N$.

Let $\hat{B}^2_j = \sum_k \frac{1}{A^2_n}\sum_{i\in I^2_n}\Phi_{jk}(X_{i^1})\Phi_{jk}(X_{i^2})$ be the U-statistic estimator of $\sum_k \alpha^2_{jk}$.

Then on $\{2^{jd} < n^2\}$,

$$E^n_{f_A}\left(\hat{B}^2_j - \sum_k \alpha^2_{jk}\right)^2 \leq Cn^{-1} + 2^{jd}n^{-2}$$

Write that,

$$E^n_{f_A}\left[\hat{B}^2_j(\tilde{X})\right]^2 = n^{-2}(n-1)^{-2}\sum_{i_1,i_2\in I^2_n}\sum_{k_1,k_2}\Phi_{jk_1}(X_{i^1_1})\Phi_{jk_1}(X_{i^2_1})\Phi_{jk_2}(X_{i^1_2})\Phi_{jk_2}(X_{i^2_2})$$

On $M_4 = \{i_1, i_2 \in I^2_n : |i_1 \cup i_2| = 4\}$, *i.e.* with no match between the two indices, the kernel $h_{i_1}h_{i_2} = \sum_{k_1,k_2}\Phi_{jk_1}(X_{i^1_1})\Phi_{jk_1}(X_{i^2_1})\Phi_{jk_2}(X_{i^1_2})\Phi_{jk_2}(X_{i^2_2})$ is unbiased, equal under expectation to $(\sum_k \alpha^2_{jk})^2$.

On $M_c, c = 2, 3$, with at least one match between i_1 and i_2 lemma 4.10 is applicable to reduce the double sum in k_1, k_2 and,

$$\begin{aligned}
E^n_{f_A}h_{i_1}h_{i_2}\mathbb{I}\{M_2 \cup M_3\} &= \sum_{i_1,i_2\in I^2_n}\sum_{k_1,k_2}\Phi_{jk_1}(X_{i^1_1})\Phi_{jk_1}(X_{i^2_1})\Phi_{jk_2}(X_{i^1_2})\Phi_{jk_2}(X_{i^2_2})\mathbb{I}\{M_2 \cup M_3\} \\
&\leq \sum_{M_2,M_3}(4N-3)^d\sum_k|\Phi_{jk}(X_{i^1_1})\Phi_{jk}(X_{i^2_1})\Phi_{jk}(X_{i^1_2})\Phi_{jk}(X_{i^2_2})| \\
&\leq \sum_{M_2,M_3} C\sum_k 2^{jd(2-|i_1\cup i_2|)} = C\sum_{M_2,M_3} 2^{jd(3-|i_1\cup i_2|)},
\end{aligned}$$

using lemma 4.7 with parameter $m = 2$ and $r = 2$ for line 3.

Next, by lemma 4.5, $|M_c| = O(n^c)$ and $|M_4|$ divided by $(A_n^2)^2$ is more precisely equal to $1 - 4n^{-1} + Cn^{-2}$. So that

$$E_{f_A}^n \big[\hat{B}_j^2(\tilde{X})\big]^2 \leq (1 + Cn^{-2})\big(\sum_k \alpha_{jk}^2\big)^2 + C\sum_{c=2}^{3} n^c n^{-4} 2^{jd(3-c)} = \big(\sum_k \alpha_{jk}^2\big)^2 + Cn^{-1} + C2^{jd}n^{-2}.$$

□

Proposition 4.15 (Variance of multisample $\prod \sum_k \hat{B}_j^2(\tilde{R}^\ell)$)

Let $\{R_1^\ell, \dots, R_n^\ell\}$ be an i.i.d. sample of $f^{\star\ell}$, $\ell = 1, \dots, d$. Assume that f is compactly supported and that φ is a Daubechies D2N. Assume that the d samples are independent.

Let $\hat{B}_j^2(R^\ell) = \sum_k \frac{1}{A_n^2} \sum_{i \in I_n^2} \Phi_{jk}(R_{i^1}^\ell)\Phi_{jk}(R_{i^2}^\ell)$ be the U-statistic estimator of $\sum_k \alpha_{jk^\ell}^2$, $\ell = 1 \dots d$.

Then on $\{2^{jd} < n^2\}$,

$$E_{f_A}^n \Big(\prod_{\ell=1}^{d} \hat{B}_j^2(R^\ell) - \sum_{k^1,\dots,k^d} \alpha_{jk^1}^2 \dots \alpha_{jk^d}^2\Big)^2 \leq Cn^{-1} + C\frac{2^j}{n^2}.$$

Successive application of $ab - cd = (a - c)b + (b - d)c$ leads to

$$a_1 \dots a_d - b_1 \dots b_d = \sum_{\ell=1}^{d} (a_\ell - b_\ell) b_1 \dots b_{\ell-1} a_{\ell+1} \dots a_d. \tag{25}$$

So applying (25),

$$\begin{aligned}
\sum_k \hat{\lambda}_{jk}^2 - \lambda_{jk}^2 &= \sum_{k^1 \dots k^d} \hat{\alpha}_{jk^1}^2 \dots \hat{\alpha}_{jk^d}^2 - \alpha_{jk^1}^2 \dots \alpha_{jk^d}^2 \\
&= \sum_{k^1 \dots k^d} \sum_{\ell=1}^{d} (\hat{\alpha}_{jk^\ell}^2 - \alpha_{jk^\ell}^2) \alpha_{jk^1}^2 \dots \alpha_{jk^{\ell-1}}^2 \hat{\alpha}_{jk^{\ell+1}}^2 \dots \hat{\alpha}_{jk^d}^2 \\
&= \sum_{\ell=1}^{d} C \sum_{k^\ell} (\hat{\alpha}_{jk^\ell}^2 - \alpha_{jk^\ell}^2) \sum_{k^{\ell+1}} \hat{\alpha}_{jk^\ell}^2 \dots \sum_{k^d} \hat{\alpha}_{jk^d}^2
\end{aligned}$$

And

$$\Big(\sum_k \hat{\lambda}_{jk}^2 - \lambda_{jk}^2\Big)^2 \leq d \sum_{\ell=1}^{d} C \Big(\sum_{k^\ell} (\hat{\alpha}_{jk^\ell}^2 - \alpha_{jk^\ell}^2) \sum_{k^{\ell+1}} \hat{\alpha}_{jk^\ell}^2 \dots \sum_{k^d} \hat{\alpha}_{jk^d}^2\Big)^2$$

Label $Q = E_{f_A}^n \Big(\sum_k \hat{\lambda}_{jk}^2 - \lambda_{jk}^2\Big)^2$.

If the d samples are independent, if $2^{jd} < n^2$, and by proposition 4.14 with parameter $d = 1$,

$$Q \leq \sum_{\ell=1}^{d-1} \left[C(Cn^{-1} + \frac{2^j}{n^2}) \prod_{l=\ell+1}^{d-1} (C + Cn^{-1} + \frac{2^j}{n^2}) \right] + C(Cn^{-1} + \frac{2^j}{n^2})$$
$$\leq Cn^{-1} + C\frac{2^j}{n^2}$$

□

Proposition 4.16 (Variance of multi sample $\sum_k \hat{\alpha}_{jk}\hat{\lambda}_{jk}$)

Let $\{X_1, \ldots, X_n\}$ be an independent, identically distributed sample of f_A. Let $\{R_1^\ell, \ldots, R_n^\ell\}$ be an independent, identically distributed sample of $f^{\star\ell}$, $\ell = 1, \ldots, d$. Assume that f is compactly supported and that φ is a Daubechies D2N. Assume that the $d+1$ samples are independent. Let $E_{f_A}^n$ be the expectation relative to the joint samples.

Then

$$E_{f_A}^n \Big(\sum_k \hat{\alpha}_{jk}(\tilde{X})\hat{\lambda}_{jk}(\tilde{R}^1, \ldots \tilde{R}^d) - \sum_k \alpha_{jk}\lambda_{jk} \Big)^2 \leq Cn^{-1}\, \mathbb{I}\left\{2^j < n\right\} + C2^{jd}n^{-d-1}\, \mathbb{I}\left\{2^j > n\right\}$$

Let $Q = E_{f_A}^n \left(\sum_{k\in\mathbb{Z}^d} \hat{\alpha}_{jk}\hat{\lambda}_{jk} \right)^2$; expanding the statistic,

$$Q = E_{f_A}^n \sum_{k_1,k_2} \frac{1}{n^{2d+2}} \sum_{i\in\Omega_n^{2d+2}} \Phi_{jk_1}(X_{i^1})\Phi_{jk_2}(X_{i^2})\varphi_{jk_1^1}(R_{i^3}^1)\varphi_{jk_2^1}(R_{i^4}^1)\ldots\varphi_{jk_1^d}(R_{i^{2d+1}}^d)\varphi_{jk_2^d}(R_{i^{2d+2}}^d).$$

By independence of the samples, we only need to consider local constraints on the coordinates of $i \in \Omega_n^{2d+2}$.

Let a be a subset of $\{0, 1, \ldots d\}$. Let $J_a = \{i \in \Omega_n^{2d+2} \colon \ell \in a \Rightarrow i^{2\ell+1} = i^{2\ell+2}\ ;\ \ \ell \notin a \Rightarrow i^{2\ell+1} \neq i^{2\ell+2}\}$. It is clear that $|J_a| = \big(n(n-1)\big)^{d+1-|a|} n^{|a|}$ and that the J_a s define a partition of Ω_n^{2d+2} when a describes the 2^{d+1} subsets of $\{0, 1, \ldots d\}$. One can check that there are C_{d+1}^c distinct sets a such that $|a| = c$, and that $\sum_{c=0}^{d+1} C_{d+1}^c n^c (n(n-1))^{d+1-c} = n^{d+1} \sum_{c=0}^{d+1} C_{d+1}^c (n-1)^{d+1-c} = n^{2d+2}$.

On $J_\emptyset$ the kernel is unbiased. On J_a, $0 \in a$, with the first two coordinates matching, the sum in k_1, k_2 can be reduced to a sum on the diagonal by lemma 4.10. If $0 \notin a$, but some $\ell \in a$ the sum can be reduced only on dimension ℓ, $k_1^\ell = k_2^\ell$, but to no purpose as will be seen below.

So Q is written $Q = n^{-2d-2} \sum_{a\in\mathcal{P}(\{0,\ldots,d\})} Q_{0a} + Q_{1a}$, with

$$Q_{0a} \leq C_1 \sum_{i\in J_a,\, 0\in a} \sum_{k\in\mathbb{Z}^d} E_{f_A}^n \Phi_{jk}(X)^2 E_{f_A}^n \varphi_{jk^{\ell_1}}(R^{\ell_1})^2 \ldots E_{f_A}^n \varphi_{jk^{\ell_{|a|-1}}}(R^{\ell_{|a|-1}})^2 \alpha_{jk^{l_1}}^2 \ldots \alpha_{jk^{l_{d-|a|+1}}}^2$$

and

$$Q_{1a} = \sum_{i\in J_a,\, 0\notin a} \sum_{k_1,k_2} \alpha_{jk_1}\alpha_{jk_2} E_{f_A}^n \varphi_{jk_1^{\ell_1}}(R^{\ell_1})\varphi_{jk_2^{\ell_1}}(R^{\ell_1}) \ldots E_{f_A}^n \varphi_{jk_1^{\ell_{|a|-1}}}(R^{\ell_{|a|-1}})\varphi_{jk_2^{\ell_{|a|-1}}}(R^{\ell_{|a|-1}})$$
$$\alpha_{jk_1^{l_1}}\alpha_{jk_2^{l_1}} \ldots \alpha_{jk_1^{l_{d-|a|+1}}}\alpha_{jk_2^{l_{d-|a|+1}}}$$

for some all distinct $\ell_1, \dots \ell_{|a|-1}$ and $l_1, \dots l_{d-|a|+1}$ whose union is $\{1, \dots d\}$ and with $C_1 = (4N-3)^d$. The bound for Q_{0a} is also written

$$(4N-3)^d \sum_{i \in J_a,\, 0 \in a} \sum_{k \in \mathbb{Z}^d} C \lambda_{jk}^{\langle 2d-2|a|+2 \rangle}$$

with special notation $\lambda_{jk}^{\langle r \rangle} = \alpha_{jk^1}^{p_1} \dots \alpha_{jk^d}^{p_d}$ for some integers $p_1, \dots, p_d$, $0 \le p_i \le r$ with $\sum_{i=1}^d p_i = r$. And so, by Meyer's lemma this is also bounded by $\sum_{i \in J_a,\, 0 \in a} C 2^{j(|a|-1)}$.

For Q_{1a} with $|a| \ge 1$, the sum in k_1, k_2 could be split in $k_1^{l_1} \dots k_1^{l_{d-|a|+1}}, k_2^{l_1} \dots k_2^{l_{d-|a|+1}}$ where no concentration on the diagonal is ensured, and $k^{\ell_1} \dots k^{\ell_{|a|-1}}$ where lemma 4.10 is applicable, but precisely the multidimensional coefficient $\alpha_{jk} = \alpha_{jk^1 \dots k^d}$ is not guaranteed factorisable under any split, unless $A = I$. So we simply fall back to

$$Q_{1a} \le \sum_{i \in J_a,\, 0 \notin a} \sum_{k_1, k_2} \left[\alpha_{jk_1} \alpha_{jk_2}\right] \left[\alpha_{jk_1^{l_1}} \alpha_{jk_2^{l_1}} \dots \alpha_{jk_1^{l_{d-|a|+1}}} \alpha_{jk_2^{l_{d-|a|+1}}}\right] \left[C 2^{\frac{j}{2}} E_{f_A}^n |\varphi_{jk_1^\ell}(R^\ell)|\right]^{|a|-1}.$$

This is also written, using Meyer's lemma at the end,

$$Q_{1a} \le \sum_{i \in J_a,\, 0 \notin a} \Big(\sum_k \alpha_{jk} \lambda_{jk}^{\langle d-|a|+1 \rangle} \Big)^2 \le \sum_{i \in J_a,\, 0 \notin a} C 2^{j(|a|-1)}$$

Finally, with $\sum_{i \in J_a} 1 = |J_a|$ given above, the general bound is written,

$$Q \le n^{-2d-2} \left[\sum_{a \ne \emptyset} C 2^{j(|a|-1)} n^{d+1} (n-1)^{d+1-|a|} + n^{d+1} (n-1)^{d+1} \Big(\sum_k \alpha_{jk} \lambda_{jk} \Big)^2 \right]$$

and so

$$\begin{aligned} Q - \Big(\sum_k \alpha_{jk} \lambda_{jk} \Big)^2 &\le 2^{-j} \sum_{c=1}^{d+1} 2^{jc} (n-1)^{-c} + C n^{-2} \\ &\le C n^{-1} \mathbb{I}\left\{2^j < n\right\} + 2^{jd} n^{-d-1} \mathbb{I}\left\{2^j > n\right\} \end{aligned}$$

□

4.5 Appendix 2 – Lemmas

Lemma 4.4 (Property set)

Let $A_1, \dots, A_r$ be r non empty subsets of a finite set Ω. Let J be a subset of $\{1, \dots, r\}$.

Define the property set $B_J = \{x \in \cup A_j : x \in \cap_{j \in J} A_j \,;\ x \notin \cup_{j \in J^c} A_j\}$, that is to say the set of elements belonging exclusively to the sets listed through J. Let $b_J = |B_J|$ and $b_\kappa = \sum_{|J|=\kappa} b_J$.

Then $\sum_{\kappa=0}^r \sum_{|J|=\kappa} B_J = \Omega$, and

$$|A_1| \vee \dots |A_r| \le \sum_{\kappa=1}^r b_\kappa = |A_1 \cup \dots A_r| \le |A_1| + \dots |A_r| = \sum_{\kappa=1}^r \kappa b_\kappa$$

with equality for the right part only if $b_\kappa = 0$, $\kappa = 2\ldots, r$ i.e. if all sets are disjoint, and equality for the left part if one set A_i contains all the others.

It follows from the definition that no two different property sets intersect and that the union of property sets defines a partition of $\cup A_i$, hence a partition of Ω with the addition of the missing complementary $\Omega - \cup A_i$ denoted by $B_\emptyset$. The B_J are also the atoms of the Boolean algebra generated by $\{A_1, \ldots, A_r, \Omega - \cup A_i\}$ with usual set operations.

With $B_\emptyset$, an overlapping of r sets defines a partition of Ω with cardinality at most 2^r ; there are C_r^κ property sets satisfying $|J| = \kappa$, with $\sum_{\kappa=0}^r C_r^\kappa = 2^r$.
□

Lemma 4.5 (Many sets matching indices)

Let $m \in \mathbb{N}$, $m \geq 1$. Let Ω_n^m be the set of indices $\{(i^1, \ldots, i^m) : i^j \in \mathbb{N}, 1 \leq i^j \leq n\}$. Let $r \in \mathbb{N}$, $r \geq 1$. Let $I_n^m = \left\{i \in \Omega_n^m : \ell_1 \neq \ell_2 \Rightarrow i^{\ell_1} \neq i^{\ell_2}\right\}$.

For $i = (i^1, \ldots, i^m) \in \Omega_n^m$, let $\tilde{\imath} = \cup_{j=1}^m \{i^j\} \subset \{1, \ldots, n\}$ be the set of distinct integers in i.

Then, for some constant C depending on m,

$$\#\left\{(i_1, \ldots, i_r) \in \left(\Omega_n^m\right)^r, : |\,\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_r\,| = a\right\} = O(n^a) I\,\{\,|\tilde{\imath}_1| \vee \ldots \vee |\tilde{\imath}_r| \leq a \leq mr\}$$

and in corollary $\#\left\{(i_1, \ldots, i_r) \in \left(I_n^m\right)^r : |\,i_1 \cup \ldots \cup i_r\,| = a\right\} = O(n^a) I\,\{m \leq a \leq mr\}$.

In the setting introduced by lemma 4.4, building the compound $(\tilde{\imath}_1, \ldots, \tilde{\imath}_r)$ while keeping track of matching indices is achieved by drawing $b_{\{1\}}^1 = |\,\tilde{\imath}_1|$ integers in the 2^0–partition $b_\emptyset^0 = \{1, \ldots, n\}$ thus constituting $\tilde{\imath}_1$, then $b_{\{1,2\}}^2 + b_{\{2\}}^2 = |\tilde{\imath}_2|$ integers in the 2^1–partition $\{b_{\{1\}}^1, b_\emptyset^1\}$ thus constituting two subindexes from which to build $\tilde{\imath}_2$, then $b_{\{1,2,3\}}^3 + b_{\{2,3\}}^3 + b_{\{1,3\}}^3 + b_{\{3\}}^3 = |\,\tilde{\imath}_3|$ integers in the 2^2–partition $\{b_{\{1,2\}}^2, b_{\{1\}}^2, b_{\{2\}}^2, b_\emptyset^2\}$ thus constituting 2^2 subindexes from which to build $\tilde{\imath}_3$, and so on, up to $b_{\{1,\ldots,r\}}^r + \ldots + b_{\{r\}}^r = |\,\tilde{\imath}_r|$ integers in the cardinality 2^{r-1} partition $\{b_{\{1,\ldots,r-1\}}^{r-1} \ldots, b_\emptyset^{r-1}\}$ thus constituting 2^{r-1} subindexes from which to build $\tilde{\imath}_r$.

The number of ways to draw the subindexes composing the r indexes is then

$$A_{b_\emptyset^0}^{b_{\{1\}}^1}\, A_{b_{\{1\}}^1}^{b_{\{1,2\}}^2} A_{b_\emptyset^1}^{b_{\{2\}}^2} \ldots A_{b_{\{1,\ldots,r-1\}}^{r-1}}^{b_{\{1,\ldots,r\}}^r} \ldots A_{b_\emptyset^{r-1}}^{b_{\{r\}}^r} \tag{26}$$

with the nesting property $b_J^j = b_J^{j+1} + b_{J\cup\{j+1\}}^{j+1}$ (provided J exists at step j) and $A_n^m = \frac{n!}{(n-m)!}$.

At step j, the only property set with cardinality equivalent to n, is $B_\emptyset^{j-1}$, while all others have cardinalities lower than m ; so picking integers inside these light property sets involve cardinalities at most in $m!$ that go in the constants, while the pick in $B_\emptyset^{j-1}$ entails a cardinality $A_{b_\emptyset^{j-1}}^{b_{\{r\}}^j} = A_{n-|\,\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_{j-1}|}^{b_{\{r\}}^j} \approx n^{b_{\{r\}}^j}$.

Note that, at step $j-1$, $b_\emptyset^{j-1} = n - |\,\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_{j-1}\,|$, because, at step j, $b_{\{j\}}^j$ designates the number of integers in $\tilde{\imath}_j$ not matching any previous index $\tilde{\imath}_1, \ldots, \tilde{\imath}_{j-1}$; so that also $\sum_{j=1}^r b_{\{j\}}^j = |\,\tilde{\imath}_i \cup \ldots \cup \tilde{\imath}_r|$; and incidentally $\sum_{J \ni j_0} b_J^j = |\,\tilde{\imath}_{j_0}|$.

The number of integers picked from the big property set at each step is

$$A_{b_\emptyset^0}^{b_{\{1\}}^1} A_{b_\emptyset^1}^{b_{\{2\}}^2} \ldots A_{b_\emptyset^{r-1}}^{b_{\{r\}}^r}$$

with $b_\emptyset^j = n - |\,\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_{j-1}|$, $b_\emptyset^0 = n$ and $\sum_{j=1}^r b_{\{j\}}^j = |\,\tilde{\imath}_i \cup \ldots \cup \tilde{\imath}_r|$.

For large n this is equivalent to $n^{|\,\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_r|}$.

Having drawn the subindexes, building the indexes effectively is a matter of iteratively intermixing two sets of a and b elements ; an operation equivalent to highlighting b cells in a line of $a+b$ cells, which can be done in C_{a+b}^b ways, with $C_n^p = A_n^p/p!$.

Intermixing the subindexes thus involve cardinalities at most in $m!$, that go in the constant C.

Likewise, passing from $\tilde{\imath}$ to i involve cardinalities at most in $C_m^{|\tilde{\imath}|}$ and no dependence on n.

For the corollary, if $i \in I_n^m$ then $\tilde{\imath} = i$ and $|\tilde{\imath}| = m$. If moreover $i^1 < \ldots < i^r$, the number of ways to draw the subindexes is given by replacing occurrences of 'A' by 'C' in (26), with $C_n^m = \frac{n!}{m!(n-m)!}$, which does not change the order in n. Also there is only one way to intermix subindexes, because of the ordering constraint.

□

Lemma 4.6 (Two sets matching indices [Corollary and complement])

Let I_n^m be the set of indices $\{(i^1, \ldots, i^m) \colon i^j \in \mathbb{N},\, 1 \le i^j \le n,\, i^j \ne i^\ell$ if $i \ne \ell\}$, and let I'^m_n be the subset of I_n^m such that $\{i^1 < \ldots < i^m\}$.

Then for $0 \le b \le m$,

$$\#\Big\{(i_1, i_2) \in I_n^m \times I_n^m : |\,i_1 \cap i_2\,| = b\Big\} = A_n^m A_m^b A_{n-m}^{m-b} C_m^b = O(n^{2m-b})$$
$$\#\Big\{(i_1, i_2) \in I'^m_n \times I'^m_n : |\,i_1 \cap i_2\,| = b\Big\} = C_n^m C_m^b C_{n-m}^{m-b} = O(n^{2m-b})$$

In corollary, with P (resp. P') the mass probability on $(I_n^m)^2$ (resp. $(I'^m_n)^2$), $P(|i_1 \cap i_2| = b) \approx P'(|i_1 \cap i_2| = b) = O(n^{-b})$ and $P(|i_1 \cap i_2| = 0) = P'(|i_1 \cap i_2| = 0) \le 1 - m^2 n^{-1} + C n^{-2}$.

For $i_1, i_2 \in I_n^m$, the equivalence $|i_1 \cap i_2| = b \Longleftrightarrow |i_1 \cup i_2| = 2m - b$ gives the link with the general case of lemma 4.5.

Reusing the pattern of lemma 4.5 in a particular case : there are A_n^m ways to constitute i_1, there are A_m^b ways to draw b unordered integers from i_1 and A_{n-m}^{m-b} ways to draw $m-b$ unordered integers from $\{1, \ldots, n\} - i_1$.

To constitute i_2, intermixing both subindexes of b and $m-b$ integers is equivalent to highlighting b cells in a line of m cells ; there are C_m^b ways to do so. On I'^m_n, by definition, having drawn the b then $m-b$ ordered distinct integers, intermixing is uniquely determined.

Incidentally, one can check that $\sum_{b=0}^m A_m^b A_{n-m}^{m-b} C_m^b = A_n^m$, and that $\sum_{b=0}^m C_m^b C_{n-m}^{m-b} = C_n^m$.

Dividing by $\left(A_n^m\right)^2$ or $\left(C_n^m\right)^2$, both equivalent to n^{2m}, gives the probabilities. Finally for the special case $b=0$, use the fact that

$$\frac{A_n^m}{A_{n-c}^m} = (1-\frac{c}{n})\dots(1-\frac{c}{n-m+1}) \leq (1-\frac{c}{n})^m$$

□

Lemma 4.7 (Product of r kernels of degree m)

Let $r \in \mathbb{N}^*$. Let $m \geq 1$. Let $(X_1 \dots, X_n)$ be an independent, identically distributed sample of a random variable on $\mathbb{R}^d$. Let Ω_n^m be the set of indices $\{(i^1, \dots, i^m) \colon i^j \in \mathbb{N}, 1 \leq i^j \leq n\}$.

For $i \in \Omega_n^m$, define

$$\begin{aligned} a_{ik} &= \Phi_{jk}(X_{i^1}) \dots \Phi_{jk}(X_{i^m}) \\ b_{ik} &= \varphi_{jk}(X_{i^1}^{\ell_1}) \dots \varphi_{jk}(X_{i^{m_1}}^{\ell_{m_1}}) \Phi_{jk}(X_{i^{m_1+1}}) \dots \Phi_{jk}(X_{i^{m_1+m_d}}). \end{aligned}$$

Let $\tilde{\imath}$ be the set of distinct coordinates in i and let $c = c(\tilde{\imath}_1, \dots \tilde{\imath}_r) = |\tilde{\imath}_1 \cup \dots \cup \tilde{\imath}_r|$ be the overall number of distinct coordinates in r indices $(i_1, \dots i_r) \in (\Omega_n^m)^r$.

Then

$$\begin{aligned} E_f^n |a_{i_1 k_1} \dots a_{i_r k_r}| &\leq C 2^{\frac{jd}{2}(mr-2c)} \\ E_f^n |b_{i_1 k_1} \dots b_{i_r k_r}| &\leq C 2^{\frac{jd}{2}(m_d r - 2c_d)}\, 2^{\frac{j}{2}(m_1 r - 2c + 2c_d)} \end{aligned}$$

with $c_d = c_d(\tilde{\imath}_1, \dots \tilde{\imath}_r) \leq c$ the fraction of c corresponding to products with at least one $\Phi(X)$ term and $1 \leq c_d \leq m_d r$, $0 \leq c - c_d \leq m_1 r$, $1 \leq c \leq (m_1 + m_2) r$.

Using lemma 4.4, one can see that the product $a_{i_1 k_1} \dots a_{i_r k_r}$, made of mr terms, can always be split into $|\tilde{\imath}_1 \cup \dots \cup \tilde{\imath}_r|$ independent products of $c(l)$ dependent terms, $1 \leq l \leq |\tilde{\imath}_1 \cup \dots \cup \tilde{\imath}_r|$, with $c(l)$ in the range from $|\tilde{\imath}_1| \vee \dots \vee |\tilde{\imath}_r|$ to mr and $\sum_l c(l) = mr$.

Using lemma 4.11, a product of $c(l)$ dependent terms, is bounded under expectation by $C 2^{\frac{jd}{2}(c(l)-2)}$. Accumulating all independent products, the overall order is $C 2^{\frac{jd}{2}(mr - 2|\tilde{\imath}_1 \cup \dots \tilde{\imath}_r|)}$.

For $b_{i_1 k_1} \dots b_{i_r k_r}$ make the distinction between groups containing at least one $\Phi(X)$ term and the others containing only $\varphi(X^\ell)$ terms. This splits the number $|\tilde{\imath}_1 \cup \dots \cup \tilde{\imath}_d|$ into $g_{\Phi,\varphi} + g_\varphi$. Let $c_\varphi(l)$ be the number of φ terms in a product of $c(l)$ terms, mixed or not.

On the $g_{\Phi,\varphi}$ groups containing Φ terms, first bound the product of $c_\varphi(l)$ terms by $C 2^{\frac{j}{2} c_\varphi(l)}$, and the remaining terms by $C 2^{\frac{jd}{2}(c(l) - c_\varphi(l) - 2)}$. On the g_φ groups with only φ terms, bound the product by $C 2^{\frac{j}{2}(c_\varphi(l) - 2)}$.

The overall order is then

$$C2^{\frac{jd}{2}\left[\left(\sum_{l=1}^{g_{\Phi,\varphi}} c(l)-c_\varphi(l)\right)-2g_{\Phi,\varphi}\right]}\ 2^{\frac{j}{2}\sum_{l=1}^{g_{\Phi,\varphi}} c_\varphi(l)}\ 2^{\frac{j}{2}\left[\left(\sum_{l=1}^{g_\varphi} c_\varphi(l)\right)-2g_\varphi\right]}.$$

The final bound is found using $\sum_{l=1}^{g_\varphi} c_\varphi(l) + \sum_{l=1}^{g_{\Phi,\varphi}} c_\varphi(l) = m_1 r$ and $\sum_{l=1}^{g_{\Phi,\varphi}} c(l) - c_\varphi(l) = m_d r$.

Rename $c_d = g_{\Phi,\varphi}$ and $c - c_d = g_\varphi$.

As for the constraints, in the product of $(m_1 + m_d)r$ terms, it is clear that Φ terms have to be found somewhere, so $c_d \geq 1$, which also implies that $c - c_d = 0$ when $c = 1$ (in this case there are no independent group with only ϕ terms, but only one big group with all indices equal). Otherwise $c_d \leq m_d r$ and $c - c_d \leq m_1 r$ since there are no more that this numbers of Φ and ϕ terms in the overall product.

□

Lemma 4.8 (Meyer)

Let $V_j, , j \in \mathbb{Z}$ an r-regular multiresolution analysis of $L_2(\mathbb{R}^n)$ and let $\varphi \in V_0$ be the father wavelet.

There exist two constant $c_2 > c_1 > 0$ such that for all $p \in [1, +\infty]$ and for all finite sum $f(x) = \sum_k \alpha(k)\varphi_{jk}(x)$ one has,

$$c_1\|f\|_p \leq 2^{jd(\frac{1}{2}-\frac{1}{p})}\left(\sum_k |\alpha(k)|^p\right)^{\frac{1}{p}} \leq c_2\|f\|_p$$

See Meyer (1997)

We use the bound under a special form.

First note that if $f \in B_{sp\infty}$, $\|f\|_{sp\infty} = \|P_j f\|_p + \sup_j 2^{js}\|f - P_j f\|_p$ so that $\|f - P_j f\|_p \leq C\|f\|_{sp\infty}2^{-js}$. So using (13),

$$\begin{aligned}\sum_k |\alpha_{jk}|^p &\leq C2^{jd(1-p/2)}\|P_j f\|_p^p \leq C2^{jd(1-p/2)}2^{p-1}\big(\|f\|_p^p + \|f - P_j f\|_p^p\big)\\ &\leq C2^{jd(1-p/2)}2^{p-1}\big(\|f\|_p^p + C\|f\|_{sp\infty}^p 2^{-jps}\big)\\ &\leq C2^{jd(1-p/2)}\|f\|_{sp\infty}^p.\end{aligned}$$

When applying the lemma to special coefficient $\lambda_{jk}^{\langle r\rangle} = \alpha_{jk^1}^{p_1}\dots\alpha_{jk^d}^{p_d}$ for some integers $p_1, \dots, p_d$, $0 \leq p_i \leq r$ with $\sum_{i=1}^d p_i = r$, we use

$$\begin{aligned}\sum_{k\in\mathbb{Z}^d} |\lambda_{jk}^{\langle r\rangle}| &= \sum_{k^1\in\mathbb{Z}} |\alpha_{jk^1}^{p_1}|\dots\sum_{k^d\in\mathbb{Z}} |\alpha_{jk^d}^{p_d}|\\ &\leq C2^{\frac{j}{2}(2-p_1)}\|f^{\star 1}\|_{sp_1\infty}^{p_1}\dots 2^{\frac{j}{2}(2-p_d)}\|f^{\star d}\|_{sp_d\infty}^{p_d}\\ &\leq C2^{\frac{j}{2}(2d-r)}\|\max_\ell f^{\star\ell}\|_{sr\infty}^r\end{aligned}$$

so that even if some p_ℓ was zero, the result is a 2^j, which returns the effect of $\sum_{k^\ell} 1$.
□

Lemma 4.9 (Path of non matching dimension numbers)

Let $r \in \mathbb{N}$, $r \geq 2$. Let $\Omega_n^m = \{(i^1, \ldots, i^m) : i^\ell \in \mathbb{N}, 1 \leq i^\ell \leq n\}$. For $i \in \Omega_n^d$, let $\Lambda_{jk}(V_i) = \varphi_{jk}(X_{i^1}^1) \ldots \varphi_{jk}(X_{i^d}^d)$. Let $\tilde{\imath}$ be the set of distinct coordinates of i.

In the product

$$\Big(\sum_j \sum_k \frac{1}{n^d} \sum_{i \in \Omega_n^d} \Lambda_{jk}(V_i)\Big)^r = \frac{1}{n^{dr}} \sum_{i_1, \ldots, i_r \in (\Omega_n^d)^r} \sum_{j_1 \ldots j_r} \sum_{k_1 \ldots, k_r} \Lambda_{j_1 k_1}(V_{i_1}) \ldots \Lambda_{j_r k_r}(V_{i_r})$$

unless $|\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_r| < r$, it is always possible to find indices $(i_1, \ldots, i_r)$ such that no two functions φ_{jk} $\varphi_{jk'}$ match on observation number.

Let $c = |\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_r|$. For $1 \leq \ell \leq n$, let $\ell^{\otimes d} = (\ell, \ldots, \ell) \in \Omega_n^d$.

With r buckets of width d defined by the extent of each index $k_1 \ldots, k_r$, and only $c < r$ distinct observation numbers, once c buckets have been stuffed with terms $V_{\ell^{\otimes d}}$, some already used observation number must be reused in order to fill in the remaining $r - c$ buckets. So that $r - c$ buckets will match on dimension and observation number allowing to reduce the sum to only c distinct buckets.

Once $c > r$, starting with a configuration using $V_{\ell_1^{\otimes d}}, \ldots V_{\ell_r^{\otimes d}}$ we can always use additional observation numbers to fragment further the $\ell^{\otimes d}$ terms, which preserves the empty intersection between buckets.
□

Lemma 4.10 (Daubechies wavelet concentration property)

Let $r \in \mathbb{N}$, $r \geq 1$. Let φ be the scaling function of a Daubechies wavelet $D2N$. Let h_k be the function on $\mathbb{R}^m$ defined as a product of translations of φ

$$h_k(x_1, \ldots, x_m) = \varphi(x_1 - k^1) \ldots \varphi(x_m - k^m),$$

with $k = (k^1, \ldots, k^m) \in \mathbb{Z}^m$.

Then for a Haar wavelet $\left[\sum_k h_k(x_1, \ldots, x_m)\right]^r = \sum_k h_k(x_1, \ldots, x_m)^r$.

For any D2N,

$$\Big(\sum_k |h_k(x_1, \ldots x_m)|\Big)^r \leq (4N-3)^{m(r-1)} \sum_k |h_k(x_1, \ldots x_m)|^r \tag{27}$$

With a Daubechies Wavelet $D2N$, whose support is $[0, 2N-1]$ with $\varphi(0) = \varphi(2N-1) = 0$ (except for Haar where $\varphi(0) = 1$), one has the relation

$$x \mapsto \varphi(x-k)\varphi(x-\ell) = 0, \qquad \text{for } |\ell - k| \geq 2N - 1\,;$$

when k is fixed, the cardinal of the set $|\ell - k| < 2N-1$ is equal to $(4N-3)$.

So that, with $k_1, \dots k_r$ denoting r independent multi-index,

$$\big(\sum_k h_k\big)^r = \sum_{k_1} \sum_{k_2 \dots k_r} h_{k_1} \dots h_{k_r} \, I(\Delta)$$

with $\Delta = \{|k_{i_1}^{\ell_1} - k_{i_2}^{\ell_2}| < (2N-1)\,; \quad i_1, i_2 = 1 \dots r\,; \quad \ell_1, \ell_2 = 1 \dots m\}$. Once k_1 say, is fixed, the cardinal of Δ is not greater than $(4N-3)^{m(r-1)}$ and is exactly equal to 1 for Haar, when all $k_1 = \dots = k_r$.

For any Daubechies wavelet, and $r \geq 1$, using the inequality $(|h_{k_1}|^r \dots |h_{k_r}|^r)^{\frac{1}{r}} \leq \frac{1}{r} \sum_i |h_{k_i}|^r$,

$$\begin{aligned}
\big(\sum_k |h_k|\big)^r &\leq \sum_{k_1, \dots, k_r} \frac{1}{r} \big(|h_{k_1}|^r + \dots + |h_{k_r}|^r \big) \, \mathbb{I}\{\Delta\} \\
&= \frac{1}{r} \Big[\sum_{k_1, \dots, k_r} |h_{k_1}|^r \, \mathbb{I}\{\Delta\} + \dots + \sum_{k_1, \dots, k_r} |h_{k_r}|^r \, \mathbb{I}\{\Delta\} \Big] \\
&\leq (4N-3)^{m(r-1)} \sum_k |h_k|^r,
\end{aligned}$$

□

Lemma 4.11 **(rth order moment of Φ_{jk})**

Let X be random variables on $\mathbb{R}^d$ with density f. Let Φ be the tensorial scaling function of an MRA of $L_2(\mathbb{R}^d)$. Let $\alpha_{jk} = E_f \Phi_{jk}(X)$. Then for $r \in \mathbb{N}^*$,

$$E_f |\Phi_{jk}(X) - \alpha_{jk}|^r \leq 2^r E_f |\Phi_{jk}(X)|^r \leq 2^r 2^{jd(\frac{r}{2}-1)} \|f\|_\infty \|\Phi\|_r^r.$$

If Φ is the Haar tensorial wavelet then also $E_f \, \Phi_{jk}(X)^r \leq 2^{jd(\frac{r}{2}-\frac{1}{2})} \alpha_{jk}$.

For the left part of the inequality, $\Big(E_f |\Phi_{jk}(X) - \alpha_{jk}|^r\Big)^{\frac{1}{r}} \leq \Big(E_f |\Phi_{jk}(X)|^r\Big)^{\frac{1}{r}} + E_f |\Phi_{jk}(X)|$, and also $E_f |\Phi_{jk}(X)| \leq \Big(E_f |\Phi_{jk}(X)|^r\Big)^{\frac{1}{r}} \Big(E_f 1\Big)^{\frac{r-1}{r}}$.

For the right part, $E_f |\Phi_{jk}(X)|^r = 2^{jdr/2} \int |\Phi(2^j x - k)|^r f(x) dx \leq 2^{jd(\frac{r}{2}-1)} \|f\|_\infty \|\Phi\|_r^r$.

Or also if Φ is positive,

$$\begin{aligned}
E_f \Phi_{jk}(X)^r &= 2^{\frac{jd}{2}(r-1)} \int \Phi(2^j x - k)^{r-1} \Phi_{jk}(x) f(x) dx \\
&\leq 2^{\frac{jd}{2}(r-1)} \|\Phi\|_\infty^{r-1} \alpha_{jk}.
\end{aligned}$$

□

4.6 References for ICA and estimation of a quadratic functional

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

(Barbedor, 2005) P. Barbedor. Independent components analysis by wavelets. *Technical report, LPMA Université Paris 7*, PMA-995, 2005.

(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.

(Bickel & Ritov, 1988) P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives : sharp best order of convergence estimates. *Sankya Ser* A50, 381-393

(Butucea & Tribouley, 2006) C. Butucea and K. Tribouley. Nonparametric homogeneity tests. *Journal of Statistical Planning and Inference*, 136(2006), 597–639.

(Birgé & Massart, 1995) L. Birgé and P. Massart. Estimation of integral functionals of a density. *Annals of Statistics* 23(1995), 11-29

(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.

(Kerkyacharian & Picard, 1996) Gérard Kerkyacharian Dominique Picard. Estimating non quadratic functionals of a density using Haar wavelets. *Annals of Statistics*, 24(1996), 485-507.

(Laurent, 1996) B. Laurent. Efficient estimation of a quadratic functional of a density. *Annals of Statistics* 24 (1996), 659 -681.

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

(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.

(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.* 8 273–303, 1972.

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

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

(Tribouley, 2000) K. Tribouley, Adaptive estimation of integrated functionals. *mathematical methods of statistics* 9(2000) p19-38.

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

5. Towards thresholding

Given an independent, identically distributed sample $\tilde{X} = \{X_1, \dots, X_n\}$ of a random variable X on $\mathbb{R}^d$, $d \geq 2$, with presumably linearly mixed components, that is to say with $X = AS$, where A is a $d \times d$ invertible matrix, ICA aims at recovering the original source S, whose components are mutually independent.

In the density model it is assumed that the observed signal 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 ℓth 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 the ICA model expressed this way, both f and A are unknown, and the data consists in a random sample of f_A. The semi-parametric case corresponds to f left unspecified, except for general regularity assumptions.

Demixing is carried out by means of 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 the unknown mixing matrix A.

An exact ICA contrast is provided by the comparison in the L_2 norm of the density f_A with the products of its marginals, an idea that can be traced back to Rosenblatt (1975). The estimation of the comparison $\int |f_A - f_A^\star|^2$, with $f_A^\star$ the product of the marginals of f_A, has been studied in a previous work (Barbedor, 2006) and share many features with the estimation a quadratic functional.

Under Besov smoothness conditions, the previously introduced wavelet contrast defined as $C_j^2(f_A) = \|P_j(f_A - f_A^\star)\|_2^2$, with P_j the projection operator of a multiresolution analysis at level j, yields a factorization measure with bias, in the sense that a zero contrast implies independence of the projected densities, but that independence in projection transfers to original densities up to some bias 2^{-2js}. This is a consequence of the very definition of Besov spaces (here $B_{s2\infty}$) and of an orthogonality property of the projection spaces V_j and W_j which entails the relation

$$0 \leq \int (f_A - f_A^\star)^2 - \int [P_j(f_A - f_A^\star)]^2 \leq C 2^{-2js}.$$

The wavelet contrast estimator $\hat{C}_j^2(\tilde{X}) = \sum_k (\hat{\alpha}_{jk^1,\dots k^d} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d})^2$, with $\hat{\alpha}_{jk^1,\dots k^d}$ the moment estimator of coordinate $(k^1, \dots k^d)$ of f_A and $\hat{\alpha}_{jk^\ell}$ the moment estimator of coordinate k^ℓ of marginal ℓ of f_A, has been used for estimating the exact squared L_2 norm of $f_A - f_A^\star$. It has been shown to have a risk converging to zero at rate $n^{\frac{-4s}{4s+d}}$, for a choice of resolution j^* such that $2^{j^* d} \approx n^{\frac{d}{d+4s}}$. The wavelet contrast estimator is inoperable for a choice of resolution j such that $2^{jd} > n$. This fact rules out block thresholding at levels $j \,:\, 2^{jd} > n$ that is known to be a better approach for the estimation of a quadratic functional (Cai and Low, 2005).

On the other hand, optimal U-statistics estimators of the wavelet contrast are available, with convergence rate the one of the estimation a quadratic functional, but these estimators have computational cost at least in $O(n^2)$ which makes them a priori less practical.

The Plug-in, biased, wavelet contrast estimator enjoys both ease of computation and ease of transitions between resolutions through discrete wavelet transform (DWT), since it builds upon a preliminary estimation of all individual wavelet coordinates of f on the projection space at level j, that is to say a full density estimation. The evaluation of $C_j^2(f_A)$ thus relies on the procedures of wavelet density estimation which are found in a series of articles from Kerkyacharian and Picard (1992) and Donoho et al. (1996).

The filter aware nature of the wavelet contrast estimator opens the way to term-by-term thresholding as is done in density estimation. This is the object of the present paper. We apply a thresholding procedure to the wavelet contrast estimator, making it adaptive in the sense that when f belongs to the Besov class B_{spq}, no knowledge of the regularity parameter s is assumed in the choice of the projection level j.

Follows some implementation issues. Numerical simulations show how the thresholded wavelet contrast estimator behaves as compared to the linear one, and other methods.

5.1 Estimating $\int (f_A - f_A^\star)^2$ with term-by-term thresholding

We first recall the definition of the wavelet contrast already introduced in a previous work (Barbedor, 2005).

Let f and g be two functions on $\mathbb{R}^d$ and let Φ be the scaling function of a multiresolution analysis of $L_2(\mathbb{R}^d)$ for which projections of f and g exist.

Define the approximate loss function

$$C_j^2(f-g) = \sum_{k\in\mathbb{Z}^d} \left(\int (f-g)\Phi_{jk} \right)^2 = \|P_j(f-g)\|_2^2.$$

It is clear that $f = g$ implies $C_j^2 = 0$ and that $C_j^2 = 0$ implies $P_j f = P_j g$ almost surely.

Let f be a density function on $\mathbb{R}^d$; denote by $f^{\star\ell}$ the marginal distribution in dimension ℓ

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

and denote by $f^\star$ the product of marginals $f^{\star 1} \dots f^{\star d}$. The functions f, $f^\star$ and the $f^{\star\ell}$ admit a wavelet expansion on a compactly supported basis (φ, ψ). Consider the projections up to order j, that is to say the projections of f, $f^\star$ and $f^{\star\ell}$ on V_j^d and V_j, namely

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

with $\alpha_{jk}(f^{\star\ell}) = \int f^{\star\ell}\varphi_{jk}$ and $\alpha_{jk}(f) = \int f\Phi_{jk}$. At least for compactly supported densities and compactly supported wavelets, it is clear that $P_j f^\star = P_j^1 f^{\star 1} \dots P_j^d f^{\star d}$.

Proposition 5.17 (ICA wavelet contrast)

Let f be a compactly supported density function on $\mathbb{R}^d$ and let φ be the scaling function of a compactly supported wavelet.

Define the wavelet ICA contrast as $C_j^2(f - f^\star)$. Then,

$$\begin{aligned} f \text{ factorizes} &\implies C_j^2(f - f^\star) = 0 \\ C_j^2(f - f^\star) = 0 &\implies P_j f = P_j f^{\star 1} \dots P_j f^{\star d} \quad a.s. \end{aligned}$$

Proof $f = f^1 \dots f^d \implies f^{\star\ell} = f^\ell$, $\ell = 1, \dots d$. □

Wavelet contrast in place of the quadratic functional $\int (f_A - f_A^\star)^2$

Let $f = f_I$ be a density defined on $\mathbb{R}^d$ whose components are independent, that is to say f is equal to the product of its marginals. Let f_A be the mixed density given by $f_A(x) = |\det A^{-1}| f(A^{-1}x)$, with A a $d \times d$ invertible matrix. Let $f_A^\star$ be the product of the marginals of f_A. Note that when $A = I$, $f_A^\star = f_I^\star = f_I = f$.

Assume as a regularity condition that f belongs to some Besov space B_{spq}. It has been checked in previous work that Besov membership of f transfers to the mixed density f_A and to the products of the marginals $f_A^\star$ (Barbedor, 2005).

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

$$J_{spq}(f) = \|\alpha_{0.}\|_{\ell_p} + \left[\sum_{j \geq 0} \left[2^{js} 2^{dj(\frac{1}{2} - \frac{1}{p})} \|\beta_{j.}\|_{\ell_p} \right]^q \right]^{\frac{1}{q}} < \infty, \tag{28}$$

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

Also, by definition of a Besov space $B_{spq}(\mathbb{R}^d)$ with a r-regular wavelet φ, $r > s$,

$$f \in B_{spq}(\mathbb{R}^d) \iff \|f - P_j f\|_p = 2^{-js}\, \epsilon_j, \quad \{\epsilon_j\} \in \ell_q(\mathbb{N}^d). \tag{29}$$

So, from the decomposition

$$\begin{aligned} \|f_A - f_A^\star\|_2^2 &= \int P_j(f_A - f_A^\star)^2 + \int \left[f_A - f_A^\star - P_j(f_A - f_A^\star) \right]^2, \\ &= C_j^2(f_A - f_A^\star) + \int \left[f_A - f_A^\star - P_j(f_A - f_A^\star) \right]^2, \end{aligned}$$

resulting from the orthogonality of V_j and W_j, and assuming more precisely that f_A and $f_A^\star$ belong to $B_{s2q}(\mathbb{R}^d)$,

$$0 \leq \|f_A - f_A^\star\|_2^2 - C_j^2(f_A - f_A^\star) \leq C 2^{-2js}, \tag{30}$$

which gives an illustration of the shrinking (with j) distance between the wavelet contrast and the always bigger squared L_2 norm of $f_A - f_A^\star$ representing the exact factorization measure. A side effect of (30) is that $C_j^2(f_A - f_A^\star) = 0$ is implied by $A = I$.

Hard-thresholded estimator

Let S be the latent random variable with density f.
Define the experiment $\mathcal{E}^n = \big(\mathcal{X}^{\otimes n}, \mathcal{A}^{\otimes n}, (X_1, \dots, X_n), P_{f_A}^n, f_A \in B_{spq}\big)$, where $X_1, \dots, X_n$ is an iid sample of $X = AS$, and $P_{f_A}^n = P_{f_A} \dots \otimes P_{f_A}$ is the joint distribution of $(X_1 \dots, X_n)$.

Define the coordinates estimators

$$\hat{\alpha}_{jk} = \hat{\alpha}_{jk^1 \dots, k^d} = \frac{1}{n} \sum_{i=1}^n \varphi_{jk^1}(X_i^1) \dots \varphi_{jk^d}(X_i^d) \quad \text{and} \quad \hat{\alpha}_{jk^\ell} = \frac{1}{n} \sum_{i=1}^n \varphi_{jk^\ell}(X_i^\ell) \tag{31}$$

where X^ℓ is coordinate ℓ of $X \in \mathbb{R}^d$. Define also the shortcut $\hat{\lambda}_{jk} = \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d}$.

Define the linear plug-in wavelet contrast estimator

$$\hat{C}_j^2 = \hat{C}_j^2(X_1, \dots, X_n) = \sum_{(k^1, \dots, k^d) \in \mathbb{Z}^d} (\hat{\alpha}_{j(k^1, \dots, k^d)} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d})^2 = \sum_{k \in \mathbb{Z}^d} (\hat{\alpha}_{jk} - \hat{\lambda}_{jk})^2 \tag{32}$$

Recall that a function f admits a wavelet expansion on the basis (Φ, Ψ) if the series

$$\sum_{k \in \mathbb{Z}^d} \alpha_{j_0 k}(f) \Phi_{j_0 k} + \sum_{j=j_0}^{\infty} \sum_{k \in \mathbb{Z}^d} \beta_{jk}(f) \Psi_{jk} \tag{33}$$

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

For thresholding purpose, we need to express the wavelet contrast with β_{jk} coefficients, with β_{jk} defined in (33) ; this is done in lemma 5.15, and so the wavelet contrast is written

$$C_{j_1}(f) = C_{j_0}(f) + \sum_{j=j_0}^{j_1} \sum_k (\beta_{jk} - \beta_{jk^1} \dots \beta_{jk^d})^2.$$

where $k = (k^1, \dots, k^d)$ and $\beta_{jk^\ell} = \int f^{\star \ell}(x^\ell) \psi_{jk^\ell}(x^\ell)\, dx^\ell$ and $\beta_{jk} = \int f(x) \Psi_{jk}(x)\, dx$.

The term-by-term thresholded wavelet contrast will then be estimated by the quantity $\tilde{\hat{C}}_{j_0, j_1} = \hat{C}_{j_0} + \tilde{\hat{T}}_{j_0 j_1}$ with

$$\tilde{\hat{T}}_{j_0 j_1} = \sum_{j=j_0}^{j_1} \sum_{k^1 \dots, k^d} (\tilde{\hat{\beta}}_{jk} - \tilde{\hat{\beta}}_{jk^1} \dots \tilde{\hat{\beta}}_{jk^d})^2 = \sum_{j=j_0}^{j_1} \sum_{k \in \mathbb{Z}^d} \tilde{\hat{\zeta}}_{jk}^2$$

with $\tilde{\hat{\beta}}_{jk}$ the thresholded substitute for $\hat{\beta}_{jk}$, $j_0 = j_s$ depending on s, when it is known, or otherwise $j_0 = 0$ or some units for an adaptive estimate ; and with ζ_{jk} defined as the difference $\beta_{jk} - \beta_{jk^1} \dots \beta_{jk^d}$, and likewise for $\hat{\zeta}_{jk}$ and $\tilde{\hat{\zeta}}_{jk}$.

In the following we study the case of a hard-thresholded procedure of the type

$$\tilde{\hat{\beta}}_{jk} = \hat{\beta}_{jk} I\{\,|\hat{\beta}_{jk}| > t/2\} \text{ and } \tilde{\hat{\beta}}_{jk^\ell} = \hat{\beta}_{jk^\ell} I\{\,|\hat{\beta}_{jk^\ell}| > (t/2)^{\frac{1}{d}}\}, \text{ with } t \approx \sqrt{\frac{\log n}{n}}.$$

We finally 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 $\sum_k 1 \approx 2^{jd}$.

Block threshold

For the record, we set here what would be the rule for block-thresholding the wavelet contrast estimator. Unfortunately, for cases $s < \frac{d}{4}$, the rule implies to choose a resolution j such that $2^{jd} > n$, which is beyond the technical limit $2^{jd} < n$ required for a small risk contrast.

Denote by f_Δ the difference $f_A - f_A^{\star 1}, \dots f_A^{\star d}$. Following our notations, one has the decomposition

$$f_\Delta = \sum_k \delta_{j_0 k} + \sum_{j=j_0}^{\infty} \sum_k \zeta_{jk}.$$

We first determine what should be the level of the threshold for entire blocks at level j, $\sum_k \hat{\zeta}_{jk}^2$.

Assuming that f, hence f_Δ, belongs to $B_{s2\infty}(\mathbb{R}^d)$, one has the property,

$$\sum_k \zeta_{jk}^2 \leq C 2^{-2js} \quad \forall j\,;$$

this is the incurred bias when removing slices at level j.

For $s \geq d/4$, and $\forall j \geq j_s$, such that $2^{j_s d} \approx n$, one has

$$2^{-2js} \leq 2^{-2j_s s} \leq 2^{-(d/2)j_s} \leq C 2^{-(d/2)j_s} 2^{j_s d} n^{-1} \leq 2^{jd/2} n^{-1}\,;$$

so to remove all slices above the optimal resolution j_s, namely $2^{j_s d} \approx n$, we can take the value $t(j,n) = C 2^{jd/2} n^{-1}$.

For $s \leq d/4$ and $\forall j \geq j_s$, such that $2^{j_s d} \approx n^{\frac{2d}{d+4s}}$, one has

$$2^{-2js} \leq 2^{-2j_s s} = C n^{\frac{-4s}{4s+d}} = C 2^{j_s d/2} n^{-1}\,;$$

so that $t(j,n)$ is the smallest threshold depending only on j and n and satisfying,

$$f \in B_{s2\infty}(\mathbb{R}^d) \Longrightarrow \exists C > 0, \forall j \geq j_s, \sum_k \zeta_{jk}^2 \leq C t_1(j,n)$$

Let $L_{jn} = \{j : \sum_k \hat{\zeta}_{jk}^2 > t(j,n)\}$; we set the block thresholded estimator to

$$\bar{\hat{C}}_{j_0 j_1} = \sum_k \hat{\delta}_{j_0 k}^2 + \sum_{j=j_0}^{j_1} I_{L_{jn}} \sum_k \hat{\zeta}_{jk}^2.$$

To cover the regular cases with the linear part of the estimator, we set $2^{j_0 d} \approx n$, and to cover the cases with small regularity, we set $2^{j_1 d} \approx n^2$ (corresponding to worst case $s = 0$).

5.2 Risk upper bound

Let the thresholded estimator $\tilde{\hat{C}}_{j_0,j_1}$ be used in estimating the quadratic functional $K_\star = \int (f_A - f_A^\star)^2$; using (30), an upper bound for the mean squared error of this procedure when $f_A \in B_{s2q}(\mathbb{R}^d)$ is given by

$$\begin{aligned} E^n_{f_A}(\tilde{\hat{C}}_{j_0,j_1} - K_\star)^2 &\leq 2E^n_{f_A}(\tilde{\hat{C}}_{j_0,j_1} - C_j^2)^2 + 2(C_j^2 - \|f_A - f_A^\star\|_2^2)^2 \\ &\leq 2E^n_{f_A}(\tilde{\hat{C}}_{j_0,j_1} - C_j^2)^2 + 2(\|f_A - f_A^\star - P_j(f_A - f_A^\star)\|_2^2)^2 \\ &\leq 2E^n_{f_A}(\tilde{\hat{C}}_{j_0,j_1} - C_j^2)^2 + C2^{-4j_1 s}. \end{aligned}$$

If $f \in B_{spq}$ with $p \geq 2$, since we consider compactly supported densities,

$$\|f_A - f_A^\star - P_j(f_A - f_A^\star)\|_2 \leq \|f_A - f_A^\star - P_j(f_A - f_A^\star)\|_p \leq C2^{-j_1 s}$$

and the rate is unchanged ; otherwise if $1 \leq p \leq 2$, using the Sobolev embedding $B_{spq} \subset B_{s'p'q}$ for $p \leq p'$ 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 $\{\epsilon_j\} \in \ell_q$,

$$\|f_A - f_A^\star - P_j(f_A - f_A^\star)\|_2 \leq \epsilon_j 2^{-j_1(s+d/2-d/p)},$$

and the rate is unchanged except that s' is substituted for s.

Since $2^{j_1 d} \approx n(\log n)^{-1}$ the bias of the wavelet contrast 2^{-4js} is of the order of $[n(\log n)^{-1}]^{\frac{-4s}{d}}$.

As is shown in the next proposition, the risk of $\tilde{\hat{C}}_{j_0,j_1}$ in estimating the wavelet contrast is at best in $Cn^{\frac{-2s}{2s+d}}$. So that since $\frac{4s}{d} > \frac{2s}{2s+d}$, the remainder in 2^{-4js} is negligible.

Proposition 5.18 (Risk of a thresholded procedure)

Assume that f belongs to the Besov class $B_{spq}(\mathbb{R}^d)$. Set $2^{j_1 d} \approx n(\log n)^{-1}$. Set $t \approx c_t\sqrt{n^{-1}\log n}$, with c_t some constant. Set $\tilde{\hat{\beta}}_{jk} = \hat{\beta}_{jk} I\{|\hat{\beta}_{jk}| > t\}$ and for the marginals, $\tilde{\hat{\beta}}_{jk^\ell} = \hat{\beta}_{jk^\ell} I\{|\hat{\beta}_{jk^\ell}| > t^{\frac{1}{d}}\}$.

The risk about the true contrast $C_{j_0 j_1}$ is

$$E^n_{f_A}[\tilde{\hat{C}}_{j_0 j_1} - C_{j_0 j_1}]^2 \leq C(\log n)n^{\frac{-2s}{2s+d}}.$$

A convergence rate for the risk about the thresholded true contrast $\tilde{C}_{j_0 j_1}$ is

$$E^n_{f_A}[\tilde{\hat{C}}_{j_0 j_1} - \tilde{C}_{j_0 j_1}]^2 \leq Cn^{\frac{-2s}{2s+d}}.$$

First note that the value of j_1 ensures that $2^{jd} < n$, for $0 \leq j \leq j_1$.

Define $\mu_{jk} = \beta_{jk^1} \dots \beta_{jk^d}$ and $\tilde{\hat{\zeta}}_{jk} = \tilde{\hat{\beta}}_{jk} - \tilde{\hat{\beta}}_{jk^1} \dots \tilde{\hat{\beta}}_{jk^d} = \tilde{\hat{\beta}}_{jk} - \tilde{\hat{\mu}}_{jk}$; the risk of $\tilde{\hat{C}}_{j_0 j_1+1}$ about $C_{j_0 j_1+1}$ is written,

$$\begin{aligned} E^n_{f_A}\left[\tilde{\hat{C}}_{j_0 j_1+1} - C_{j_0 j_1+1}\right]^2 &= E^n_{f_A}\left[\sum_k (\hat{\delta}^2_{j_0 k} - \delta^2_{j_0 k}) + \sum_{j=j_0}^{j_1}\sum_k (\tilde{\hat{\zeta}}^2_{jk} - \zeta^2_{jk})\right]^2 \\ &\le 2E^n_{f_A}\left[\sum_k (\hat{\delta}^2_{j_0 k} - \delta^2_{j_0 k})\right]^2 + 2E^n_{f_A}\left[\sum_{j=j_0}^{j_1}\sum_k (\tilde{\hat{\zeta}}^2_{jk} - \zeta^2_{jk})\right]^2 \\ &= LR + TR \end{aligned} \tag{34}$$

The first term LR is bounded by $C2^{j_0 d}n^{-1}$ (see Barbedor, 2006).

For the second term, we now expand further $(\tilde{\hat{\zeta}}^2_{jk} - \zeta^2_{jk})$ using the sets

$$\begin{aligned} B = \{j,k \ : \ |\beta_{jk}| > t/2\}, &\quad \hat{B} = \{j,k \ : \ |\hat{\beta}_{jk}| > t\} \\ B^\ell = \{j,k^\ell \ : \ |\beta_{jk^\ell}| > \frac{t^{\frac{1}{d}}}{2}\}, &\quad \hat{B}^\ell = \{j,k^\ell \ : \ |\hat{\beta}_{jk^\ell}| > t^{\frac{1}{d}}\} \\ B_\mu = B^1 \cap \dots \cap B^d \subset \{j,k \ : \ |\mu_{jk}| > t/2\}, &\quad \hat{B}_\mu = \hat{B}^1 \cap \dots \cap \hat{B}^d \subset \{j,k \ : \ |\hat{\mu}_{jk}| > t\} \end{aligned}$$

and their complementary sets (denoted by S) in the enclosing set $\Omega_{j_0 j_1} = \{(j,k)\colon j_0 \le j \le j_1,\ 0 \le k^\ell < 2^j, l = 1 \dots d\}$. Note that $(j,k) \mapsto \beta_{jk}$ or $(j,k) \mapsto \hat{\beta}_{jk}$ define two random variables on $\Omega_{j_0 j_1}$.

Throughout the following lines, C is a placeholder for an unspecified constant ; idem for $C_\star$ but with a constant vanishing under independence.

Using the relation $\tilde{\hat{\zeta}}^2_{jk} = \hat{\zeta}^2_{jk}\,\mathbb{I}(\hat{B}\hat{B}_\mu) + \hat{\beta}^2_{jk}\,\mathbb{I}(\hat{B}\hat{S}_\mu) + \hat{\mu}^2_{jk}\,\mathbb{I}(\hat{S}\hat{B}_\mu)$, and likewise partitioning ζ^2_{jk} with the help of B and B_μ yields

$$\begin{aligned} \tilde{\hat{\zeta}}^2_{jk} - \zeta^2_{jk} = \hat{\zeta}^2_{jk}\,\mathbb{I}(\hat{B}\hat{B}_\mu) - \zeta^2_{jk}\,\mathbb{I}(BB_\mu) + \hat{\beta}^2_{jk}\,\mathbb{I}(\hat{B}\hat{S}_\mu) - \zeta^2_{jk}\,\mathbb{I}(BS_\mu) \\ + \hat{\mu}^2_{jk}\,\mathbb{I}(\hat{S}\hat{B}_\mu) - \zeta^2_{jk}\,\mathbb{I}(SB_\mu) - \zeta^2_{jk}\,\mathbb{I}(SS_\mu). \end{aligned}$$

Upon cross partitioning this is also written,

$$\begin{aligned} \tilde{\hat{\zeta}}^2_{jk} - \zeta^2_{jk} =& (\hat{\zeta}^2_{jk} - \zeta^2_{jk})\,\mathbb{I}(\hat{B}\hat{B}_\mu BB_\mu) + (\hat{\beta}^2_{jk} - \beta^2_{jk})\,\mathbb{I}(\hat{B}\hat{S}_\mu BS_\mu) + (\hat{\mu}^2_{jk} - \mu^2_{jk})\,\mathbb{I}(\hat{S}\hat{B}_\mu SB_\mu) \\ &- S_{jk} + D_{jk}\ ; \end{aligned} \tag{35}$$

with $S_{jk} = \zeta^2_{jk}\,\mathbb{I}(SS_\mu)$ and D_{jk}, composed of terms weighted by an indicator on one of the non matching sets $B\hat{S}$, $B_\mu\hat{S}_\mu$, $S\hat{B}$ or $S_\mu\hat{B}_\mu$, and given by,

$$\begin{aligned} D_{jk} = \hat{\zeta}^2_{jk}&\Big[\mathbb{I}(\hat{B}\hat{S}_\mu)\,\mathbb{I}(S \cup B_\mu) + \mathbb{I}(\hat{B}\hat{B}_\mu)\,\mathbb{I}(S \cup S_\mu) + \mathbb{I}(\hat{S}\hat{B}_\mu)\,\mathbb{I}(B \cup S_\mu)\Big] \\ - \zeta^2_{jk}&\Big[\mathbb{I}(BS_\mu)\,\mathbb{I}(\hat{S} \cup \hat{B}_\mu) + \mathbb{I}(BB_\mu)\,\mathbb{I}(\hat{S} \cup \hat{S}_\mu) + \mathbb{I}(SB_\mu)\,\mathbb{I}(\hat{B} \cup \hat{S}_\mu)\Big]. \end{aligned} \tag{36}$$

Label the first three terms in (35) T_{1jk}, T_{2jk} and T_{3jk}. Note that the term S_{jk} vanishes if in (34) the risk is computed against the thresholded true contrast $\tilde{C}_{j_0,j_1+1}$ instead of C_{j_0,j_1+1},

meaning that small terms contributions, produced by necessarily small departures from independence are left out.

So,

$$TR \leq 5E^n_{f_A}\left[\Big(\sum_{j,k} T_{1jk}\Big)^2 + \Big(\sum_{j,k} T_{2jk}\Big)^2 + \Big(\sum_{j,k} T_{3jk}\Big)^2 + \Big(\sum_{j,k} S_{jk}\Big)^2 + \Big(\sum_{j,k} D_{jk}\Big)^2\right]$$

- Term $\sum_{j,k} T_{2jk}$

With $0 \leq \mathbb{I}(\hat{B}\hat{S}_\mu BS_\mu) \leq \mathbb{I}(B)$,

$$E^n_{f_A}\Big(\sum_{j,k} T_{2jk}\Big)^2 \leq E^n_{f_A}\left[\Big(\sum_{j,k}\hat{\beta}^2_{jk}\,\mathbb{I}(B)\Big)^2 - 2\sum_{j,k}\beta^2_{jk}\,\mathbb{I}(B)\sum_{j,k}\hat{\beta}^2_{jk}\,\mathbb{I}(B) + \Big(\sum_{j,k}\beta^2_{jk}\,\mathbb{I}(B)\Big)^2\right]$$

We now on omit the term $\mathbb{I}(B)$ until it is needed, but it is implicitly present under every $\sum_j\sum_k$ sign. The following lines repeat closely the proof of the linear case (Barbedor, 2006) ; the course of the proof forks only at the very end of the paragraph when we make use of the indicator function $\mathbb{I}(B)$.

So, for the mean,

$$\begin{aligned} E^n_{f_A}\sum_j\sum_k\hat{\beta}^2_{jk} &= \frac{1}{n^2}\sum_{i_1=i_2}\sum_j\sum_k E^n_{f_A}\Psi_{jk}(X_{i_1})\Psi_{jk}(X_{i_2}) + \frac{1}{n^2}\sum_{i_1\neq i_2}\sum_j\sum_k\beta^2_{jk} \\ &= \frac{1}{n}\sum_j\sum_k E^n_{f_A}\Psi_{jk}(X_i)^2 + \frac{n-1}{n}\sum_j\sum_k\beta^2_{jk}. \end{aligned}$$

For the second moment, let $M_c = \{i_1, i_2, i_3, i_4 \in \{1,\ldots,n\}\colon |\{i_1\}\cup\ldots\cup\{i_4\}| = c\}$.

$$E^n_{f_A}\big(\sum_{j,k}\hat{\beta}^2_{jk}\big)^2 = \frac{1}{n^4}\sum_{c=1}^{4}\sum_{i_1,\ldots,i_4} E^n_{f_A}\sum_{j_1k_1,j_2k_2}\Psi_{j_1k_1}(X_{i_1})\Psi_{j_1k_1}(X_{i_2})\Psi_{j_2k_2}(X_{i_3})\Psi_{j_2k_2}(X_{i_4})\,\mathbb{I}\{M_c\}$$

On M_1, the kernel is equal to $\sum_{j_1k_1,j_2k_2}\Psi_{j_1k_1}(X)^2\Psi_{j_2k_2}(X)^2 \leq [2^{J+1}(4N-3)]^d\sum_{j,k}\Psi_{jk}(X)^4$ by lemma 4.10. And by lemma 5.18, $E^n_{f_A}\sum_{j,k}\Psi_{jk}(X)^4 \leq \sum_{j,k} C2^{jd}$.

On M_2, the kernel takes three generic forms :

(a) $\sum_{j_1k_1,j_2k_2}\Psi_{j_1k_1}(X)\Psi_{j_1k_1}(Y)\Psi_{j_2k_2}(X)\Psi_{j_2k_2}(Y)$ or (b) $\sum_{j_1k_1,j_2k_2}\Psi_{j_1k_1}(X)^2\Psi_{j_2k_2}(Y)^2$ or (c) $\sum_{j_1k_1,j_2k_2}\Psi_{j_1k_1}(X)\Psi_{j_1k_1}(Y)\Psi_{j_2k_2}(Y)^2$.

In cases (a) and (c), using lemma 4.10, the double sum can be reduced to the diagonal

$k_1 = k_2, j_1 = j_2$. So using also lemma 5.18,

$$
\begin{aligned}
(a) \quad & E^n_{f_A} \sum_{j_1k_1,j_2k_2} |\Psi_{j_1k_1}(X)\Psi_{j_1k_1}(Y)\Psi_{j_2k_2}(X)\Psi_{j_2k_2}(Y)| \\
& \leq E^n_{f_A}[2^{J+1}(4N-3)]^d \sum_{j,k} \Psi_{jk}(X)^2\Psi_{jk}(Y)^2 \leq \sum_j \sum_k C \\
(b) \quad & E^n_{f_A} \sum_{j_1k_1,j_2k_2} \Psi_{j_1k_1}(X)^2\Psi_{j_2k_2}(Y)^2 \leq \Big(\sum_j \sum_k C\Big)^2 \\
(c) \quad & E^n_{f_A} |\sum_{j_1k_1,j_2k_2} \Psi_{j_1k_1}(X)\Psi_{j_1k_1}(Y)\Psi_{j_2k_2}(Y)^2| \leq E^n_{f_A}[2^{J+1}(4N-3)]^d \sum_j \sum_k |\Psi_{jk}(X)\Psi_{jk}(Y)^3| \\
& \leq \sum_j \sum_k C.
\end{aligned}
$$

On M_3 the only representative form is

$$
E^n_{f_A} \sum_{J_1k_1,j_2k_2} \Psi_{j_1k_1}(X)\Psi_{j_1k_1}(Y)\Psi_{j_2k_2}(Z)^2 = \sum_j \sum_k \beta^2_{jk} \sum_j \sum_k E^n_{f_A} \Psi_{jk}(X)^2 \leq \sum_j \sum_k C,
$$

and on M_4 the statistic is unbiased equal to $\big(\sum_j \sum_k \beta^2_{jk}\big)^2$ under expectation.

Next, since $|M_4| = A^4_n$ and, using lemma 5.13, $|M_c| = O(n^c)$,

$$
\begin{aligned}
E^n_{f_A}\big(\sum_j \sum_k \hat{\beta}^2_{jk}\big)^2 &\leq A^4_n n^{-4}\big(\sum_{j,k} \beta^2_{jk}\big)^2 + C\sum_j \sum_k (2^{jd}n^{-3} + Cn^{-2} + Cn^{-1}) + \big(\sum_j \sum_k C\big)^2 n^{-2} \\
&\leq \big(\sum_{j,k} \beta^2_{jk}\big)^2 + Cn^{-2} + \sum_j \sum_k Cn^{-1} + \big(\sum_j \sum_k Cn^{-1}\big)^2
\end{aligned}
$$

using the fact that $2^{jd} < n$ and with $A^4_n n^{-4} = 1 - \frac{6}{n} + Cn^{-2}$.

Finally, putting all together and with the implicit term $\mathbb{I}(B)$ now made explicit on line 3,

$$
\begin{aligned}
E^n_{f_A}\Big(\sum_{j,k} T_{2jk}\Big)^2 &\leq \big(\sum_j \sum_k \beta^2_{jk}\big)^2 + Cn^{-2} + \sum_j \sum_k Cn^{-1} + \big(\sum_j \sum_k Cn^{-1}\big)^2 \\
&\quad - 2\sum_{j,k} \beta^2_{jk} \left[\frac{1}{n}\sum_j \sum_k E^n_{f_A} \Psi_{jk}(X_i)^2 + \frac{n-1}{n}\sum_j \sum_k \beta^2_{jk}\right] + \Big(\sum_{j,k} \beta^2_{jk}\Big)^2 \\
&\leq Cn^{-2} + \sum_j \sum_k Cn^{-1}\,\mathbb{I}(B) + \big(\sum_j \sum_k Cn^{-1}\big)^2 \mathbb{I}(B) + \frac{2}{n}\Big(\sum_j \sum_k \beta^2_{jk}\Big)^2 \mathbb{I}(B)
\end{aligned}
$$

And since by lemma 5.17 $\sum_j \sum_k \mathbb{I}(B) \leq Cn^{\frac{d}{2s+d}}$ the overall term is bounded by $Cn^{\frac{-2s}{2s+d}}$, which is recognized as the usual rate in density estimation.

- Term $\sum_{j,k} T_{3jk}$

Likewise,

$$
E^n_{f_A}\Big(\sum_{j,k} T_{3jk}\Big)^2 \leq E^n_{f_A}\Big(\sum_{j,k} \hat{\mu}^2_{jk}\,\mathbb{I}(B_\mu)\Big)^2 - 2\sum_{j,k} \mu^2_{jk}\,\mathbb{I}(B_\mu) \sum_{j,k} \hat{\mu}^2_{jk}\,\mathbb{I}(B_\mu) + \Big(\sum_{j,k} \mu^2_{jk}\,\mathbb{I}(B_\mu)\Big)^2
$$

As above, we omit the term $\mathbb{I}(B_\mu)$ until it is needed, but it is implicitly present under every $\sum_j \sum_k$ sign.

For $i \in \Omega_n^{2d}$, let V_i be the slice $(X_{i^1}^1, X_{i^2}^1, \ldots, X_{i^{2d-1}}^d, X_{i^{2d}}^d)$. Let the coordinate-wise kernel function Λ_{jk} be given by $\Lambda_{jk}(V_i) = \psi_{jk^1}(X_{i^1}^1)\psi_{jk^1}(X_{i^2}^1)\ldots\psi_{jk^d}(X_{i^{2d-1}}^d)\psi_{jk^d}(X_{i^{2d}}^d)$.

Let $|i|$ be the shortcut notation for $|\{i^1\} \cup \ldots \cup \{i^{2d}\}|$. Let $W_n^{2d} = \{i \in \Omega_n^{2d} \colon |i| < 2d\}$, that is to say the set of indices with at least one repeated coordinate.

Then the mean term is written

$$\begin{aligned} E_{f_A}^n \sum_j \sum_k \hat{\mu}_{jk}^2 &= n^{-2d} \sum_{i \in \Omega_n^{2d}} \sum_j \sum_k \Lambda_{jk}(V_i) \\ &= n^{-2d} \sum_{W_n^{2d}} \sum_j \sum_k E_{f_A}^n \Lambda_{jk}(V_i) + A_n^{2d} n^{-2d} \sum_j \sum_k \mu_{jk}^2 \\ &= Q_1 + A_n^{2d} n^{-2d} \theta \end{aligned}$$

Let $M_c = \{i \in \Omega_n^{2d} \colon |i| = c\}$ be the set indices with c common coordinates. So that Q_1 is written

$$Q_1 = n^{-2d} \sum_{c=1}^{2d-1} \mathbb{I}(M_c) \sum_{M_c} \sum_j \sum_k E_{f_A}^n \Lambda_{jk}(V_i) = \sum_j \sum_k Q_{1jk}$$

By proposition 5.20 with parameters $(d = 1, m = 2d, r = 1)$, $E_{f_A}^n |\Lambda_{jk}(V_i)| \mathbb{I}(M_c) \leq C 2^{\frac{j}{2}(2d-2c)}$ and by lemma 5.13, $|M_c| = O(n^c)$. Hence,

$$Q_{1jk} \leq \sum_{c=1}^{2d-1} n^{-2d+c} C 2^{j(d-c)} = 2^{-jd} \sum_{c=1}^{2d-1} C \left(\frac{2^j}{n}\right)^{(2d-c)}$$

which on $\{2^{jd} < n\}$ has maximum order $2^{j(1-d)} n^{-1}$ when $d - c$ is minimum *i.e.* $c = 2d - 1$. Finally $|Q_1| \leq \sum_j \sum_k \mathbb{I}(B_\mu) C n^{-1} \leq C n^{\frac{-2s}{2s+d}}$ by lemma 5.17.

Next, the second moment about zero is written

$$\begin{aligned} E_{f_A}^n \Big(\sum_j \sum_k \hat{\mu}_{jk}^2\Big)^2 &= n^{-4d} \sum_{i_1, i_2 \in (\Omega_n^{2d})^2} \sum_{j_1, j_2} \sum_{k_1, k_2} \Lambda_{j_1 k_1}(V_{i_1}) \Lambda_{j_2 k_2}(V_{i_2}) \\ &= n^{-4d} \sum_{W_n^{4d}} \sum_{j_1, j_2} \sum_{k_1, k_2} E_{f_A}^n \Lambda_{j_1 k_1}(V_{i_1}) \Lambda_{j_2 k_2}(V_{i_2}) + A_n^{4d} n^{-4d} \Big(\sum_j \sum_k \mu_{jk}^2\Big)^2 \\ &= Q_2 + A_n^{4d} n^{-4d} \theta^2 \end{aligned}$$

with $W_n^{4d} = \{i_1, i_2 \in (\Omega_n^{2d})^2 \colon |i_1 \cup i_2| < 4d\}$, that is to say the set of indices with at least one repeated coordinate somewhere.

Let this time $M_c = \{i_1, i_2 \in (\Omega_n^{2d})^2 \colon |i_1 \cup i_2| = c\}$ be the set indices with overall c common coordinates in i_1 and i_2. So that Q_2 is written

$$Q_2 = n^{-4d} \sum_{c=1}^{4d-1} \mathbb{I}(M_c) \sum_{M_c} \sum_{j_1, j_2} \sum_{k_1, k_2} E_{f_A}^n \Lambda_{j_1 k_1}(V_{i_1}) \Lambda_{j_2 k_2}(V_{i_2}) = \sum_{j_1, j_2} \sum_{k_1, k_2} Q_{2 j_1 k_1 j_2 k_2}$$

By lemma 4.9, unless $c = 1$, it is always possible to find indices i_1, i_2 with no matching coordinates corresponding also to matching dimension number, so that there is no way to reduce the double sum in j_1, k_1, j_2, k_2 to a sum on the diagonal using lemma 4.10.

So coping with the double sum, by proposition 4.7 with lemma parameters $(d = 1, m = 2d, r = 2)$, $E^n_{f_A}|\Lambda_{jk}(V_{i_1})\Lambda_{jk}(V_{i_2})| \leq C2^{\frac{j}{2}(4d-2c)}$, and again by lemma 5.13 $|M_c| = O(n^c)$, so $E^n_{f_A}|Q_{2j_1k_1j_2k_2}| \leq \sum_{c=1}^{4d-1} n^{c-4d}C2^{\frac{j}{2}(4d-2c)}$, which on $\{2^{jd} < n\}$ has maximum order $2^{j(1-2d)}n^{-1}$ when $c = 4d-1$. Finally, $E^n_{f_A}Q_2 \leq \sum_{j_1,j_2,k_1,k_2} C2^{j(1-2d)}n^{-1}\,\mathbb{I}(B_\mu) \leq \sum_{j,k} C2^{j(1-d)}n^{-1}\,\mathbb{I}(B_\mu) = Cn^{\frac{-2s}{2s+d}}$ using lemma 5.17.

Putting all together, and since $A^p_n n^{-p} = 1 - \frac{(d+1)(d+2)}{2n} + O(n^{-2})$,

$$\begin{aligned}
E^n_{f_A}\Big(\sum_j\sum_k \hat{\mu}^2_{jk} - \mu^2_{jk}\Big)^2 &= Q_2 + A^{4d}_n n^{-4d}\theta^2 - 2\theta(Q_1 + A^{2d}_n n^{-2d}\theta) + \theta^2 \\
&= Q_2 - 2\theta Q_1 + \theta^2(1 + A^{4d}_n n^{-4d} - 2A^{2d}_n n^{-2d}) \leq |Q_2| + 2\theta|Q_1| + O(n^{-2}) \\
&\leq Cn^{\frac{-2s}{2s+d}}
\end{aligned}$$

- Term $\sum_{j,k} T_{1jk}$

From $\hat{\zeta}^2_{jk} - \zeta^2_{jk} = (\hat{\beta}^2_{jk} - \beta^2_{jk}) + (\hat{\mu}^2_{jk} - \mu^2_{jk}) - 2(\hat{\beta}_{jk}\hat{\mu}_{jk} - \beta_{jk}\mu_{jk})$ and since T_1 implies that $\mathbb{I}\{BB_\mu\} = 1$, the only new quantity to compute is $E^n_{f_A}\Big(\sum_j\sum_k \hat{\beta}_{jk}\hat{\mu}_{jk} - \beta_{jk}\mu_{jk}\Big)^2$, which is decomposed into $E^n_{f_A}\Big(\sum_j\sum_k \hat{\beta}_{jk}\hat{\mu}_{jk}\Big)^2 - 2\sum_j\sum_k \beta_{jk}\mu_{jk}E^n_{f_A}\sum_j\sum_k \hat{\beta}_{jk}\hat{\mu}_{jk} + \Big(\sum_j\sum_k \beta_{jk}\mu_{jk}\Big)^2$.

As above, for $i \in \Omega^{d+1}_n$, let V_i be the slice $(X_{i^0}, X^1_{i^1}, \ldots, X^d_{i^d})$. Let the coordinate-wise kernel function Λ_{jk} be given by $\Lambda_{jk}(V_i) = \Psi_{jk}(X_{i^0})\psi_{jk^1}(X^1_{i^1})\ldots\psi_{jk^d}(X^d_{i^d})$. Let $\theta = \sum_j\sum_k \beta_{jk}\mu_{jk}$.

Let $W^{d+1}_n = \{i \in \Omega^{d+1}_n : |i| < d+1\}$, that is to say the set of indices with at least one repeated coordinate.

So that, $E^n_{f_A}\sum_j\sum_k \hat{\beta}_{jk}\hat{\mu}_{jk} = Q_1 + A^{d+1}_n n^{-d-1}\theta$ with $Q_1 = n^{-d-1}\sum_{W^{d+1}_n}\sum_j\sum_k E^n_{f_A}\Lambda_{jk}(V_i)$ and

$$E^n_{f_A}\Big(\sum_j\sum_k \hat{\beta}_{jk}\hat{\mu}_{jk}\Big)^2 = Q_2 + A^{2d+2}_n n^{-2d-2}\theta^2$$

with $Q_2 = n^{-2d-2}\sum_{W^{2d+2}_n}\sum_{j_1k_1,j_2k_2} E^n_{f_A}\Lambda_{j_1k_1}(V_{i_1})\Lambda_{j_2k_2}(V_{i_2})$. And we obtain in the same way,

$$E^n_{f_A}\Big(\sum_j\sum_k \hat{\beta}_{jk}\hat{\mu}_{jk} - \beta_{jk}\mu_{jk}\Big)^2 \leq |Q_2| + 2\theta|Q_1| + O(n^{-2})$$

Let $M_c = \{i \in \Omega^{d+1}_n : |i| = c\}$ be the set indices with c common coordinates. So that Q_1 is written

$$Q_1 = n^{-d-1}\sum_{c=1}^{d}\mathbb{I}(M_c)\sum_{M_c}\sum_j\sum_k E^n_{f_A}\Lambda_{jk}(V_i) = \sum_j\sum_k Q_{1jk}$$

By proposition 5.20 with parameters $(m_d = 1, m_1 = d, r = 1)$,

$$E^n_{f_A}|\Lambda_{jk}(V_i)|\,\mathbb{I}(M_c) \leq C2^{\frac{jd}{2}(1-2c_d)}2^{\frac{j}{2}(d-2c_1)}$$

with $c_1 + c_d = c$, $0 \le c_1 \le d$, $1 \le c_d \le 1$ and by lemma 5.13, $|M_c| = O(n^c)$. Hence,

$$Q_{1jk} \le \sum_{c=1}^{d} n^{-d-1+c} C 2^{j(d-dc_d-c_1)} = 2^{j(-1+(1-d)c_d)} \sum_{c=1}^{d} C \left(\frac{2^j}{n}\right)^{(d+1-c)}$$

which on $\{2^{jd} < n\}$ has maximum order Cn^{-1} when $d+1-c$ is minimum *i.e.* $c = d$. Finally $|Q_1| \le \sum_j \sum_k \mathbb{I}(B_\mu) C n^{-1} \le C n^{\frac{-2s}{2s+d}}$ by lemma 5.17.

Next, as above $Q_2 = \sum_{j_1,j_2} \sum_{k_1,k_2} Q_{2j_1k_1j_2k_2}$, and again by lemma 4.9, unless $c = 1$, it is always possible to find indices i_1, i_2 with no matching coordinates corresponding also to matching dimension number, so that there is no way to reduce the double sum in j_1, k_1, j_2, k_2 to a sum on the diagonal using lemma 4.10.

So coping once more with the double sum, by proposition 4.7 with parameters $(m_d = 1, m_1 = d, r = 2)$, $E^n_{f_A} |\Lambda_{jk}(V_{i_1})\Lambda_{jk}(V_{i_2})| \le C 2^{\frac{jd}{2}(2-2c_d)} 2^{\frac{j}{2}(2d-2c_1)}$, with $c_1 + c_d = c$, $1 \le c_d \le 2$, $0 \le c_1 \le 2d$, and again by lemma 5.13 $|M_c| = O(n^c)$, so

$$E^n_{f_A} |Q_{2j_1k_1j_2k_2}| \le \sum_{c=1}^{2d+1} n^{c-2d-2} C 2^{j(d-dc_d+d-c_1)} = 2^{j(-2+(1-d)c_d)} \sum_{c=1}^{2d+1} C \left(\frac{2^j}{n}\right)^{(2d+2-c)},$$

which on $\{2^{jd} < n\}$ has maximum order $C 2^{-jd} n^{-1}$ when $c = 2d+1$ and $c_d = 1$. Finally, $E^n_{f_A} Q_2 \le \sum_{j_1,j_2,k_1,k_2} C 2^{-jd} n^{-1} \mathbb{I}(B_\mu) \le \sum_{j,k} C n^{-1} \mathbb{I}(B_\mu) = C n^{\frac{-2s}{2s+d}}$ using lemma 5.17.

And so all three terms T_1, T_2, T_3 have same convergence rate.

- For S^2_{jk},

  On $S \cap S_\mu$, since $|\zeta_{jk}| \le |\beta_{jk}| + |\mu_{jk}|$,

$$\sum_{j=j_0}^{j_1} \sum_k S_{jk} = \sum_{j=j_0}^{j_1} \sum_k \zeta^2_{jk} \mathbb{I}(SS_\mu) \le 2 \sum_{j=j_0}^{j_1} \sum_k \beta^2_{jk} \mathbb{I}(S) + 2 \sum_{j=j_0}^{j_1} \sum_k \mu^2_{jk} \mathbb{I}(S_\mu)$$

  Let $2^{j_s d} \approx n^{\frac{d}{4s+d}}$ be the optimal resolution depending on s and lying somewhere between 0 and j_1,

$$\begin{aligned} \sum_{j=j_0}^{j_1} \sum_k \beta^2_{jk} \mathbb{I}(S) &\le 2^{j_s d} t^2 + \sum_{j=j_s}^{j_1} \sum_k \beta^2_{jk} \mathbb{I}(S) \\ &\le (\log n) n^{\frac{-4s}{4s+d}} + \sum_{j=j_s}^{j_1} \sum_k \beta^2_{jk} \end{aligned}$$

  with the term on the right lower than $2^{-2j_s} \approx n^{\frac{-2s}{4s+d}}$ using the B_{s2q} membership of f_A.

  If we take $2^{j_s d} \approx n^{\frac{d}{2s+d}}$, the optimal resolution in density estimation, we get the density estimation adaptive rate $(\log n) n^{\frac{-2s}{2s+d}}$, which is better than the one above.

- For D^2_{jk}, all terms fall under a large deviation bound. On non matching events that are of the type $\hat{B}S$ or $B\hat{S}$, we have $|\hat{\beta}_{jk} - \beta_{jk}| \ge \big| |\hat{\beta}_{jk}| - |\beta_{jk}| \big| > t/2$, and the corresponding probability

is bounded by lemma 5.16 ; likewise on non matching events that are of the type $\hat{B}_\mu \hat{S}_\mu$ or $S_\mu \hat{B}_\mu$, we have $|\hat{\mu}_{jk} - \mu_{jk}| \geq \big| |\hat{\mu}_{jk}| - |\mu_{jk}| \big| > t/2$.

Let $\Delta_{1jk} = \hat{B}S \cup \hat{B}\hat{S}_\mu B_\mu \cup \hat{B}_\mu S_\mu \cup \hat{S}\hat{B}_\mu B$ and $\Delta_{2jk} = B\hat{S} \cup BS_\mu \hat{B}_\mu \cup B_\mu \hat{S}_\mu \cup SB_\mu \hat{B}$. Since by lemma 5.16 the large deviation rates for $\hat{\beta}_{jk^\ell}$ is negligible compared to the one of $\hat{\beta}_{jk}$ and since by this same lemma also $P^n_{f_A}[\,|\hat{\beta}_{jk} - \beta_{jk}| > \frac{c_t}{2} t\,] \leq 2n^{-c_t/C}$,

$$\begin{aligned} P^n_{f_A}(\Delta_{1jk}) + P^n_{f_A}(\Delta_{2jk}) &\leq C P^n_{f_A}(\{|\hat{\beta}_{jk} - \beta_{jk}| > t/2\}) + C P^n_{f_A}(\{|\hat{\mu}_{jk} - \mu_{jk}| > t/2\}) \\ &\leq C P^n_{f_A}(\{|\hat{\beta}_{jk} - \beta_{jk}| > t/2\}) \\ &\leq C n^{-c_t/C}. \end{aligned}$$

Next, from the expansion

$$\Big(\sum_{j,k} D_{jk})\Big)^2 = \Big(\sum_{j,k} \hat{\zeta}^2_{jk}\, \mathbb{I}(\Delta_{1jk})\Big)^2 + \Big(\sum_{j,k} \zeta^2_{jk}\, \mathbb{I}(\Delta_{2jk})\Big)^2 - 2\sum_j \sum_k \hat{\zeta}^2_{jk}\, \mathbb{I}(\Delta_{1jk}) \sum_j \sum_k \zeta^2_{jk}\, \mathbb{I}(\Delta_{2jk})$$

we compute, using Hölder inequality, and lemma 5.19,

$$\begin{aligned} E^n_{f_A}\Big(\sum_{j,k} \hat{\zeta}^2_{jk}\, \mathbb{I}(\Delta_{1jk})\Big)^2 &= \sum_{j_1,j_2,k_,1,k_2} E^n_{f_A} \hat{\zeta}^2_{j_1k_1} \hat{\zeta}^2_{j_2k_2}\, \mathbb{I}(\Delta_{1j_1k_1})\, \mathbb{I}(\Delta_{1j_2k_2}) \\ &\leq \sum_{j_1,j_2,k_,1,k_2} \big[E^n_{f_A} \hat{\zeta}^8_{j_1k_1} E^n_{f_A} \hat{\zeta}^8_{j_2k_2} E^n_{f_A}\, \mathbb{I}(\Delta_{1j_1k_1}) E^n_{f_A}\, \mathbb{I}(\Delta_{1j_2k_2})\big]^{\frac{1}{4}} \\ &\leq \Big(\sum_{j,k} C[P^n_{f_A}(\Delta_{1jk})]^{\frac{1}{4}}\Big)^2 \leq 2^{j_1 d} C n^{-c_t/2C} \leq C n^{1-c_t/2C} \end{aligned}$$

which can be made arbitrary small by raising c_t. Other terms follow in the same way.

Finally, with $j_0 = 0$ or some unities, and $2^{j_1 d} \approx (n \log n)^{-1}$

$$\begin{aligned} E^n_{f_A}[\tilde{\hat{C}}_{j_0j_1} - C_{j_0j_1}]^2 &\leq 2^{j_0 d} n^{-1} + 2^{-4j_1 s} + E^n_{f_A} \sum_j \sum_k (\tilde{\hat{\zeta}}^2_{jk} - \zeta^2_{jk})^2 \\ &\leq C n^{-1} + C(\log n) n^{\frac{-2s}{2s+d}} \end{aligned}$$

□

5.3 Practical issues

In the context of ICA, once estimated, the factorization measure still needs to be minimized. This is generally resolved through constrained minimization on a Stiefel manifold, because a preliminary whitening of the data through principal component analysis (PCA) restricts the problem to orthogonal mixing matrices.

At this stage, the need to compute an empirical gradient practically rules out the Haar wavelet which is much less responsive to a small perturbation of the mixing matrix than

a more regular D2N. Indeed, projection with a Haar wavelet is equivalent to building a histogram, a procedure intrinsically immune to small perturbations of the data. But a minimization method not gradient based may decide otherwise, as it has been noted on dimension 2 examples that contrast curves based on the Haar wavelet have a visually well localized minimum (Barbedor, 2005).

In density estimation, it is customary to load the projection starting with a Haar (D2) wavelet at some high resolution, then apply some filtering passes down to an appropriate resolution, using the filter of a more regular wavelet, D4 for instance. In the thresholding version, the projection is moreover decomposed in its β_{jk} coefficients, to be able to apply the threshold, then reconstituted at the desired resolution.

In ICA, avoiding the use of the Haar wavelet means that computation of the projection with a $D2N$ must rely on approximation at dyadics and direct computation ; a rather unused method which nevertheless present no major inconvenience.

During the minimization process, the parameter W, $d \times d$, representing the current ICA-inverse of A, is moving all along the descent path on the Stiefel manifold as we strive to minimize $\hat{C}_j^2(f_{WA} - f^\star_{WA})$ based on observation WX, where X, $d \times n$, is the random sample of f_A at hand. It has been checked in the preliminary paper (Barbedor, 2005) that mixing by an invertible A conserves Besov membership. Consequently, to produce a risk bound that encompass the entire procedure, the only needed Besov assumption is that f belongs to some B_{spq}, and membership will be automatic at any point f_{WA}.

Concerning implementation, the complexity of the wavelet contrast is linear in the sample size but exponential in the dimension d of the problem that is : $O(n(2N-1)^d)$ for projection, plus $O(2^{jd})$ to compute the contrast itself (N is the order of the Daubechies D2N, N=2 is enough in our simulations, d the dimension of the problem) ; this is on account of the implicit multivariate density estimation. Using a U-statistic estimator would entail a complexity in $O(n^2(2N-1)^d)$ with no exponential complexity in j but square complexity in n.

Simulations show that, if not the fastest available, computing times are not prohibitive, depending on the context of application. In particular, the computation of wavelet values at dyadics, at some given precision, is not a bottleneck of the procedure (Barbedor, 2005).

In compensation to the computational load in high dimension, the wavelet contrast shows a very good sensitivity to small departures from independence, and encapsulates all practical tuning in a single parameter j. In the linear case, for all practical purpose, choosing j such that $2^{jd} < n$ is generally enough to obtain convergence. The thresholded version builds upon the automatic starting resolution $j_1 \colon 2^{j_1 d} \approx n(\log n)^{-1}$, this choice would be convenient also in the linear case.

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

Practical points used in the introductory paper on ICA by wavelets (Barbedor, 2005) are not repeated here, this include computation of $\varphi_{jk}(x)$ at dyadic rational numbers, signal

relocation and whitening step.

For what concerns the discrete wavelet transform, there is some point in detailing the implementation because an algorithm in dimension more than 2 is more rarely found in existing packages.

The algorithm is the usual pyramidal algorithm in dimension one applied to all rows in the cube found by fixing all but one dimension index. An efficient implementation relies on a flat array representation of the cube. Let us take the convention that $A(k^1, \ldots, k^d)$ is stored in a vector sized 2^{jd} at position $k^1 2^{j(d-1)} + \ldots + k^{(d-1)} 2^j + k^d$, so that elements in the last dimension (k^d) are stored contiguously in memory.

- Apply the pyramidal algorithm to the $2^{j(d-1)}$ 2^j-long sections in dimension k^d that are readily found in A at offsets $i2^j$, $i = 0, 2^{j(d-1)} - 1$.
- apply a cyclic transposition so that $A(k^1, \ldots, k^d)$ is now in the order $A(k^d, k^1, \ldots, k^{d-1})$ (see an algorithm for that in Wickerhauser, 1994) ;
- return to step one while the number of transpositions is lower than d (the dth transposition gives back the original order).

Since the pyramidal algorithm moves the α_{jk} in the first half of the vector and the β_{jk} in the second half, we see that starting with the second dimension, we have automatic composition of filters GG, GH, HG, HH thus matching the intended effect of multidimensional filtering.

A second application of the DWT needs to filter only the first half of all sections still found at offsets $i2^j$, $i = 0, 2^{j(d-1)} - 1$, but this time only 2^{j-1}-long.

For a signal at resolution $j > 0$, we can deconstruct the signal by successive DWT up to j times and obtain finally the projection on $V_j = V_0$ consisting of just one coefficient located at $A(0)$.

For thresholding, after complete deconstruction, we just need to set to zero the vector A where it is lower than the threshold (except at position 0), and compute the contrast on the β_{jk}.

See References below for an implementation.

Choice of the threshold in practice

Figure 1 shows the quantiles of the distribution of β_{jk} coefficients (in absolute value) for the densities taken in example in this section, with $j = 1$ to 7, $k = 0 \ldots 2^{jd}$; corresponding figures appear in Table 1 below. All results in this section can be reproduced using the program `icalette2t`.

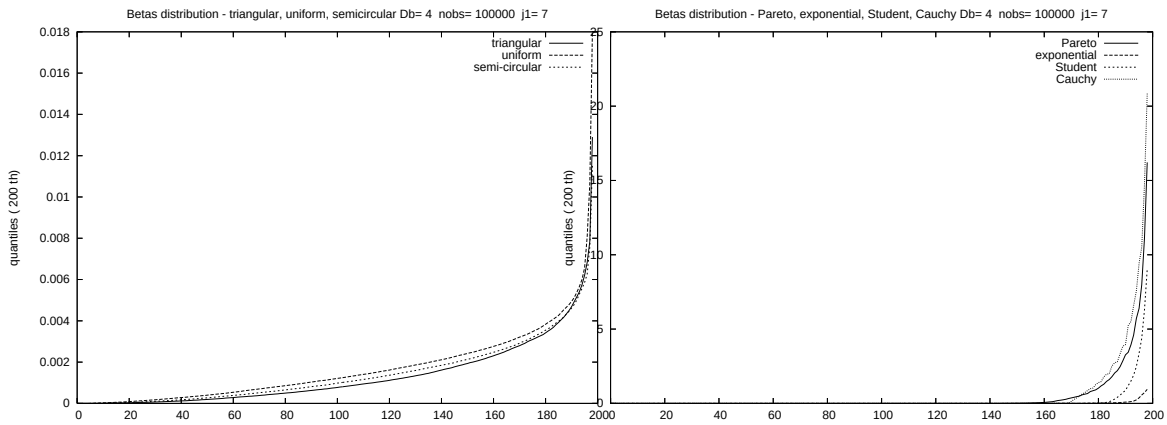

Figure 1. Quantiles of nonzero beta coefficients for some densities in dimension 2 , signals relocated in $]\frac{-1}{2}, \frac{1}{2}[$, N=100000, initialization at j=7 with D4 at L=10, deconstruction with D4 down to j=1, max. beta not shown.

qu.	uniform	semici.	triang.	Pareto	expone.	Student	Cauchy
30	0.52E-03	0.37E-03	0.28E-03	0.14E-05	0.26E-04	0.10E-05	0.38E-05
50	0.12E-02	0.96E-03	0.76E-03	0.20E-04	0.15E-03	0.15E-04	0.19E-04
80	0.27E-02	0.24E-02	0.23E-02	0.55E-01	0.26E-02	0.16E-03	0.85E-03
99	0.11E-01	0.76E-02	0.79E-02	0.11E+02	0.72E+00	0.63E+01	0.15E+02
max	0.50E-00	0.50E+00	0.50E+00	0.35E+02	0.22E+01	0.39E+02	0.58E+02
rsup	0.49E-00	0.47E-00	0.47E+00	0.49E+00	0.45E+00	0.49E+00	0.49E+00
wsup	0.20E-06	0.14E-06	0.27E-06	0.31E-04	0.16E-05	0.17E-04	0.32E-04

Table 1. Figures corresponding to Figure 1

We can see that uniform, semi-circular and triangular densities (on the left of Fig.1) present a similar profile, which can be characterized as noise-like coefficients up to 95%, with a break on the very last coefficients only, and a maximum 100 times bigger than the bulk of the distribution, meaning that very few β_{jk} hold the core of the density shape.

The break happens sooner for distributions with a peak (on the right of Fig.1) : Pareto (80%), Cauchy(85%), Student(90%), and exponential(95%), and the maximum is only 10 times bigger than the bulk of the distribution, meaning that significantly more β_{jk} are needed to fit the density shape. Choosing a threshold below those breaks will likely produce no visual effects on the contrast curves.

This illustrates the action of a fixed threshold ; more coefficients are chopped off for the densities on the left (low frequency) than for those on the right, and more coefficients for

the exponential distribution than for the Pareto distribution. The adaptive property of the fixed threshold comes from the adaptive properties of the β_{jk}.

The exact level of that overall adaptive threshold can also be adjusted for the density at hand.

Recalling equations 41 and 40, we know that the adjusted parameter c_t is related to $\|f_A\|_\infty$; since we deal with relocated and spread into $]\frac{-1}{2}, \frac{1}{2}[$ signals, $\|f_A\|_\infty$ will rather unpredictably be close to $\frac{1}{2}$ as shown in Table 1 (line rsup) ; so we take the supremum norm before relocation but still after whitening, thus removing some size effects (line wsup). We can then automatically lower the overall threshold when applied to densities with a peak (and higher supremum norm after whitening), or raise it for flat densities, which is generally speaking a desirable behavior.

Alternately, one could just decrease the threshold, starting from the raw threshold value, and decreasing one distinct ordered beta at a time, until the shape of the contrast curve shows a satisfactory pattern for minimization to follow. In higher dimensions, this can also be done on each of the C_d^2 free plans of $\mathbb{R}^d$.

Remarks

1. it seems better to choose an n such that in the rule $2^{jd} \approx n/\log n$, j as a floating point number is already close to the rounded integer value that will be used for resolution, not .5 away from it.

2. j must be sufficiently high to cover less regular densities ; if the linear contrast fails at j because of too a low resolution, there is no hope to improve things with thresholding. Since the choice of j is determined by the number of observations n, according to the rule just mentioned above, n should be chosen around 30000 to reach j=6 in dimension 2, and 50000 is a better choice according to remark 1. For $j = 7$, n=190000 gives best results. But it is to be noted that with that much observations, the linear contrast is just as effective, at least in the examples below, that may be rather noise free because artificially generated and not specially designed to contain noise.

▪

Figures below show the thresholded and linear wavelet contrast for $j = 7$ and 180000 observations for the set of densities taken in example, and with $c_t = 6.10^{-5}\text{wsup}^{-1}$ as read in Table 1.

The general effect of an aggressive thresholding seems to be a slope decrease near maximum Amari distance and a slope increase near the minimum. This is most visible on the example of the triangular density and perceptible on the uniform and semicircular (Fig. 1 to 3). In a way, this makes the task harder for a line search descent algorithm. So one can consider that if the curve shows this pattern, the threshold is already too high.

When the threshold is non aggressive (Fig. 4 to 7) the curve looks pretty much like the linear one, even with a rather high percentage of betas chopped off. This is probably

because we took signals with few very low noise level.

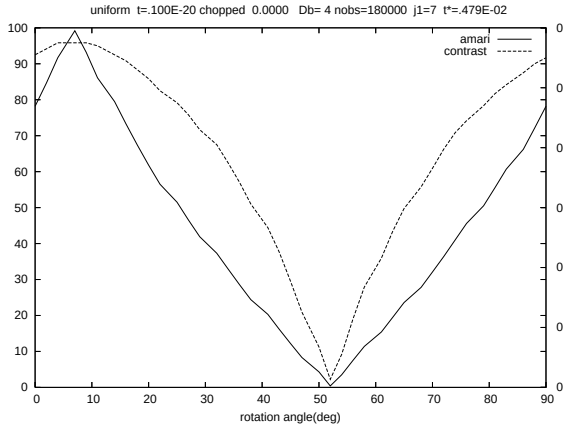

Fig.1. Uniform, linear

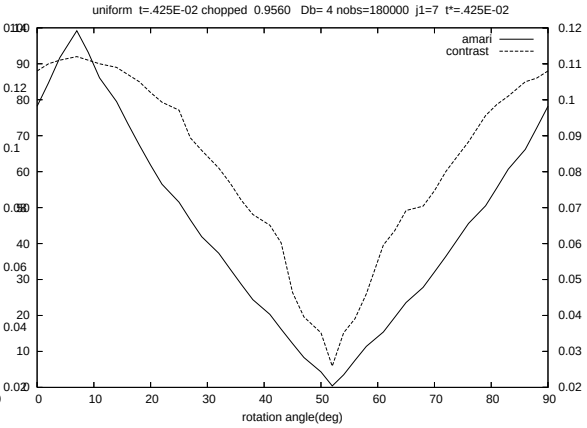

Fig.1t. Uniform, t=.00425, 95,6% off

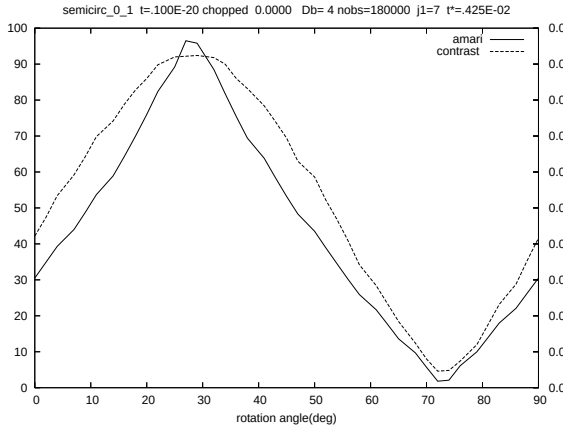

Fig.2. Semi-circular, linear

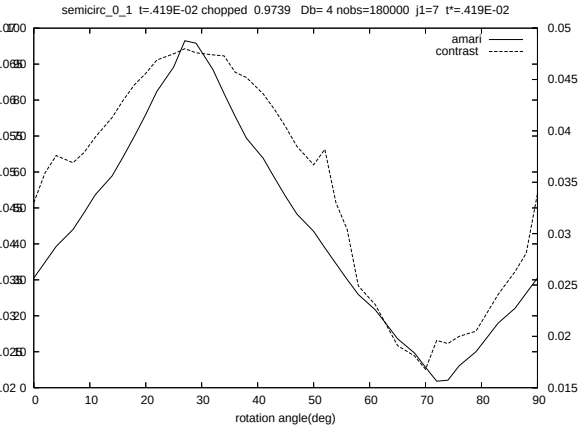

Fig.2t. Semi-circular, t=.00419, 97,4% off

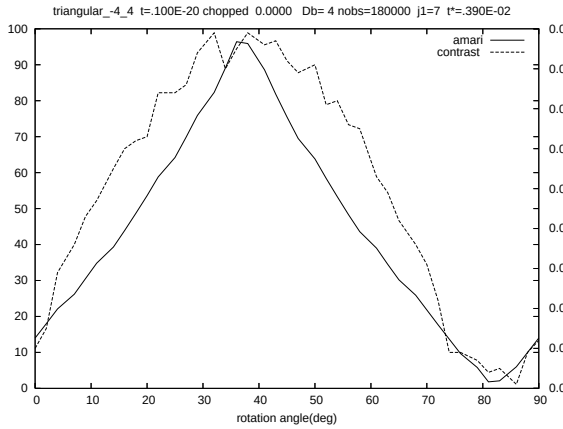

Fig.3. Triangular, linear

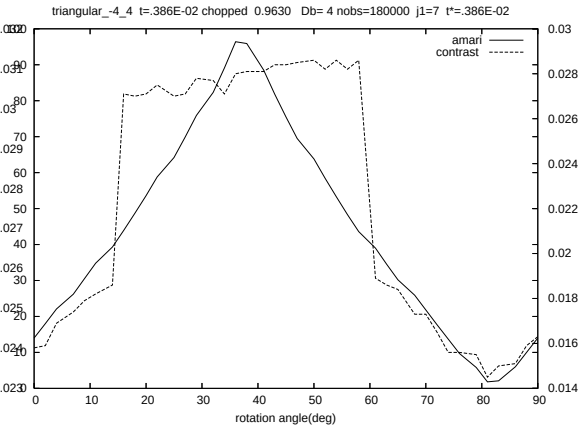

Fig.3t. Triangular, t=.00386, 96,3% off

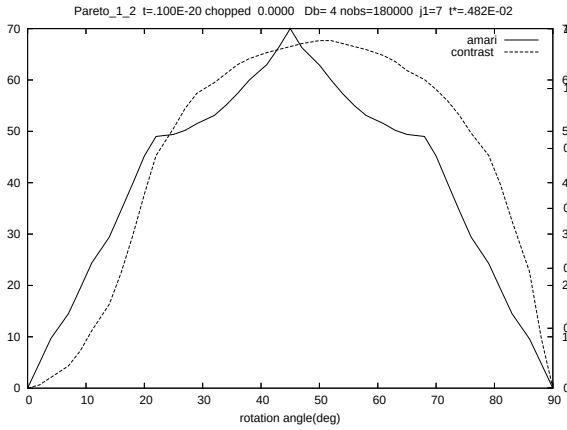

Fig.4. Pareto, linear

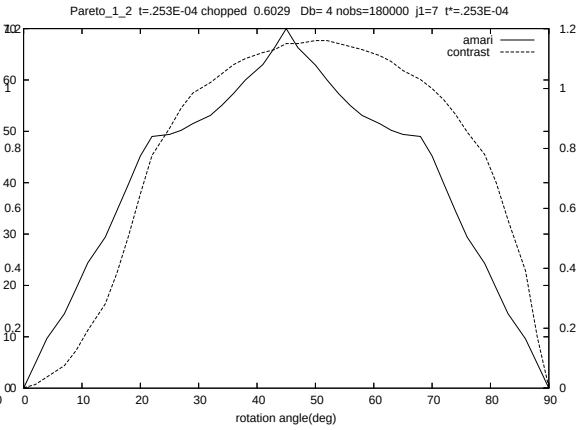

Fig.4t. Pareto, t=.0000253, 60,3% off

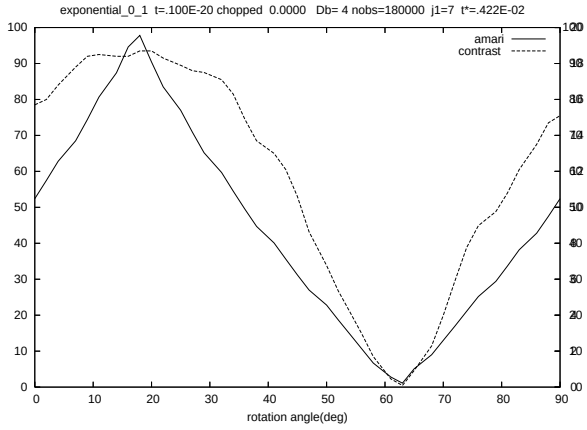

Fig.5. Exponential, linear

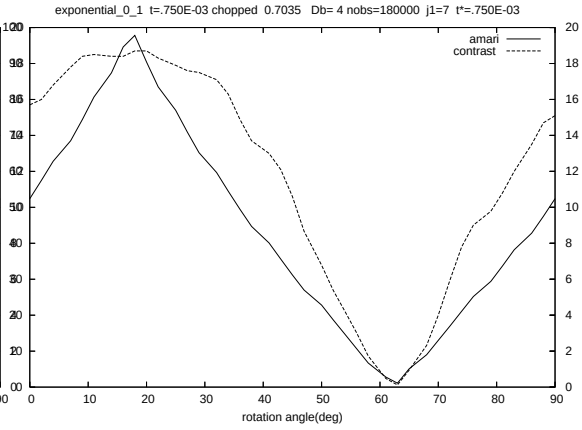

Fig.5t. Exponential, t= .00075, 70,3% off

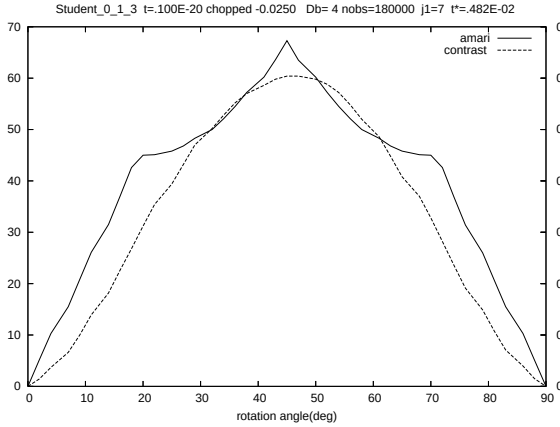

Fig.6. Student, linear

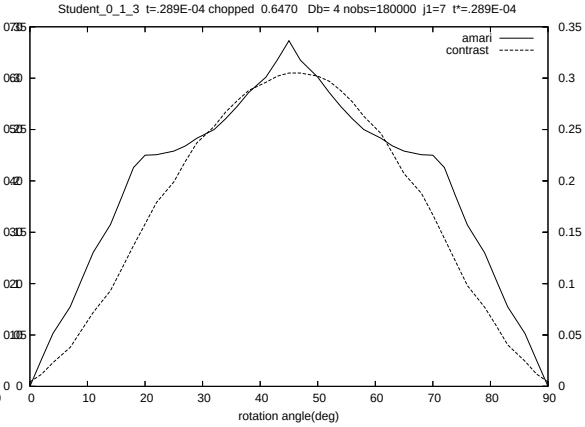

Fig.6t. Student, t= .0000289, 64.7% off

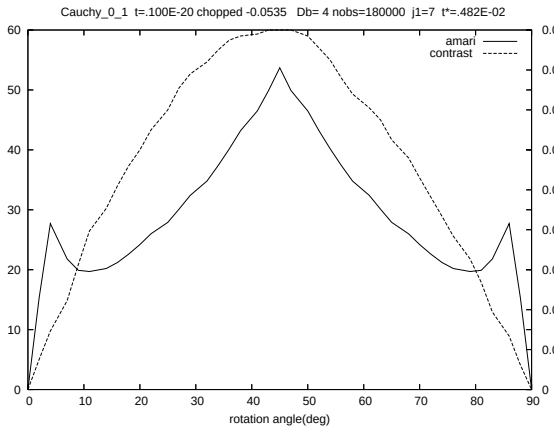

Fig.7. Cauchy, linear

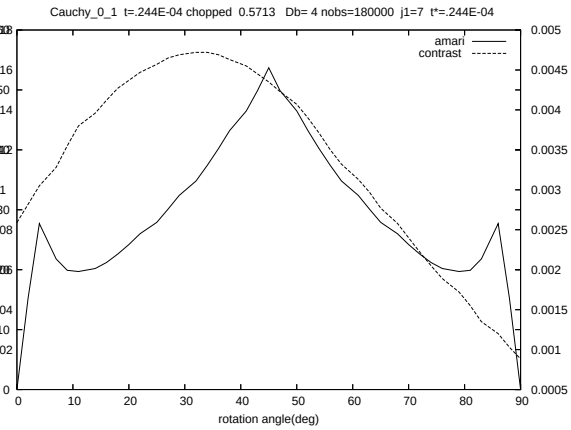

Fig.7t. Cauchy, t= .0000244, 57.1% off

5.4 Appendix 1 – Propositions

Proposition 5.19 **(rth moment of $\hat{\beta}_{jk}$ and $\hat{\mu}_{jk}$)**

Let $(X_1 \ldots, X_n)$ be an i.i.d. sample of a random variable on $\mathbb{R}^d$.

Then ,

$$E^n_{f_A}\hat{\beta}^r_{jk} = \beta^r_{jk} + O(n^{-1}) + O(2^{jd(\frac{r}{2}-1)}n^{1-r})\,\mathbb{I}\left\{2^{jd} > n\right\}$$
$$E^n_{f_A}|\hat{\beta}_{jk} - \beta_{jk}|^r = O(n^{-1}) + O(2^{jd(\frac{r}{2}-1)}n^{1-r})\,\mathbb{I}\left\{2^{jd} > n\right\}$$

and

$$E^n_{f_A}\hat{\mu}^r_{jk} = \mu^r_{jk} + O(n^{-1}) + O(2^{j(\frac{rd}{2}-1)}n^{1-rd})\,\mathbb{I}\left\{2^{j} > n\right\}$$
$$E^n_{f_A}|\hat{\mu}_{jk} - \mu_{jk}|^r = O(n^{-1}) + O(2^{j(\frac{rd}{2}-1)}n^{1-rd})\,\mathbb{I}\left\{2^{j} > n\right\}$$

Let $\Omega^m_n = \left\{(i^1, \ldots, i^m) \colon i^\ell \in \mathbb{N}, 1 \le i^\ell \le n\right\}$. Let $M_c = \{i \in \Omega^m_n \colon |\{i^1\} \cup \ldots \cup \{i^m\}| = c\}$. Let $W^m_n = \bigcup_{c=1}^{m-1} M_c$.

For the raw moment,

$$E^n_{f_A}\hat{\beta}^r_{jk} = \frac{1}{n^r}E^n_{f_A}\sum_{i\in\Omega^r_n}\Psi_{jk}(X_{i^1})\ldots\Psi_{jk}(X_{i^r}) = \frac{1}{n^r}E^n_{f_A}\sum_{i\in W^r_n}\Psi_{jk}(X_{i^1})\ldots\Psi_{jk}(X_{i^r}) + A^r_n n^{-r}\beta^r_{jk},$$

and since by lemma 5.13 $|M_c| = O(n^c)$, and by proposition 4.7 with $(m = 1, r = r)$

$$E^n_{f_A}|\Psi_{jk}(X_{i^1})\ldots\Psi_{jk}(X_{i^r})|\,\mathbb{I}\big(M_c\big) \le 2^{\frac{jd}{2}(r-2c)},$$

$$\frac{1}{n^r}E^n_{f_A}\sum_{i\in W^r_n}|\Psi_{jk}(X_{i^1})\ldots\Psi_{jk}(X_{i^r})| \le C\sum_{c=1}^{r-1}n^{c-r}2^{\frac{jd}{2}(r-2c)} = C2^{-jd\frac{r}{2}}\sum_{c=1}^{r-1}\left(\frac{2^{jd}}{n}\right)^{(r-c)},$$

which on $\{2^{jd} < n\}$ is of the order of $C2^{-jd\frac{r}{2}}2^{jd}n^{-1} \leq Cn^{-1}$, for $r \geq 2$. And which on $\{2^{jd} > n\}$ is of the order of $C2^{jd(\frac{r}{2}-1)}n^{1-r}$.

It remains to note that $A_n^r n^{-r} = 1 + O(n^{-1})$

For the central moment,

$$\begin{aligned}
E_{f_A}^n(\hat{\beta}_{jk} - \beta_{jk})^r &= E_{f_A}^n \sum_{\ell=0}^{r} (-1)^{r-\ell} \hat{\beta}_{jk}^{\ell} \beta_{jk}^{r-\ell} \\
&= \sum_{\ell=2}^{r} (-1)^{r-\ell} E_{f_A}^n \frac{1}{n^\ell} \sum_{i \in W_n^\ell} \Psi_{jk}(X_{i^1}) \dots \Psi_{jk}(X_{i^\ell}) \beta_{jk}^{r-\ell} \\
&+ \sum_{\ell=0}^{r} (-1)^{r-\ell} (1 + O(n^{-1})) \beta_{jk}^{\ell} \beta_{jk}^{r-\ell},
\end{aligned}$$

where line 3 is of the order of $O(n^{-1})$.

Next from what precedes for $\ell = 2, \dots r$,

$$E_{f_A}^n \frac{1}{n^\ell} \sum_{i \in W_n^\ell} |\Psi_{jk}(X_{i^1}) \dots \Psi_{jk}(X_{i^\ell})| = O\big(2^{jd(1-\frac{\ell}{2})} n^{-1}\big) \mathbb{I}\left\{2^{jd} < n\right\} + O\big(2^{jd(-1+\frac{\ell}{2})} n^{1-\ell}\big) \mathbb{I}\left\{2^{jd} > n\right\}$$

and since $|\beta_{jk}| = O(2^{-jd/2})$,

$$\begin{aligned}
E_{f_A}^n |\hat{\beta}_{jk} - \beta_{jk}|^r &= \sum_{\ell=2}^{r} |\beta_{jk}|^{r-\ell} O\big(2^{jd(1-\frac{\ell}{2})} n^{-1}\big) \mathbb{I}\left\{2^{jd} < n\right\} + |\beta_{jk}|^{r-\ell} O\big(2^{jd(-1+\frac{\ell}{2})} n^{1-\ell}\big) \mathbb{I}\left\{2^{jd} > n\right\} \\
&= \sum_{\ell=2}^{r} O\big(2^{jd(1-\frac{r}{2})} n^{-1}\big) \mathbb{I}\left\{2^{jd} < n\right\} + O\big(2^{jd(-1+-\frac{r}{2}+\ell)} n^{1-\ell}\big) \mathbb{I}\left\{2^{jd} > n\right\} \\
&= O\big(2^{jd(1-\frac{r}{2})} n^{-1}\big) \mathbb{I}\left\{2^{jd} < n\right\} + O\big(2^{jd(-1+\frac{r}{2})} n^{1-r}\big) \mathbb{I}\left\{2^{jd} > n\right\}
\end{aligned}$$

For $\hat{\mu}_{jk} = \hat{\beta}_{jk^1} \dots \hat{\beta}_{jk^d}$, Consider for $i \in \Omega_n^d$ the slice $V_i = (X_{i^1}^1, \dots, X_{i^d}^d)$ and the kernel $\Lambda_{jk}(V_i) = \psi_{jk}(X_{i^1}^1) \dots \psi_{jk}(X_{i^d}^d)$.

Then by symmetry with the case above,

$$E_{f_A}^n \hat{\mu}_{jk}^r = \frac{1}{n^{rd}} E_{f_A}^n \sum_{i \in (\Omega_n^d)^r} \Lambda_{jk}(V_{i^1}) \dots \Lambda_{jk}(V_{i^r}) = \frac{1}{n^{rd}} E_{f_A}^n \sum_{i \in W_n^{rd}} \Lambda_{jk}(V_{i^1}) \dots \Lambda_{jk}(V_{i^r}) + A_n^{rd} n^{-rd} \mu_{jk}^r$$

and likewise, using proposition 5.20 with parameters $(m_1 = d, m_d = 0, r = r)$, we see that $E_{f_A}^n |\Lambda_{jk}(V_{i^1}) \dots \Lambda_{jk}(V_{i^r})| \mathbb{I}\big(M_c\big) \leq 2^{\frac{j}{2}(rd-2c)}$, and so

$$\frac{1}{n^{rd}} E_{f_A}^n \sum_{i \in W_n^{rd}} |\Lambda_{jk}(V_{i^1}) \dots \Lambda_{jk}(V_{i^r})| \leq C \sum_{c=1}^{rd-1} n^{c-rd} 2^{\frac{j}{2}(rd-2c)} = C2^{-jd\frac{r}{2}} \sum_{c=1}^{rd-1} \left(\frac{2^j}{n}\right)^{(rd-c)},$$

which on $\{2^j < n\}$ is of the order of $C2^{-jd\frac{r}{2}}2^j n^{-1} \leq Cn^{-1}$, for $r \geq 2$. And which on $\{2^j > n\}$ is of the order of $C2^{j(\frac{rd}{2}-1)}n^{1-rd}$.

The remaining of the proof follows in the same way as above.

□

Proposition 5.20 (Product of r kernels of degree m)

Let $r \in \mathbb{N}^*$. Let $m \geq 1$. Let $(X_1 \ldots, X_n)$ be an i.i.d. sample of a random variable on $\mathbb{R}^d$. Let Ω_n^m be the set of indices $\{(i^1, \ldots, i^m) : i^j \in \mathbb{N}, 1 \leq i^j \leq n\}$. Let Ψ be the tensorial wavelet of a Daubechies D2N.

For $i \in \Omega_n^m$, define the kernels

$$a_{ik} = \Psi_{jk}(X_{i^1}) \ldots \Psi_{jk}(X_{i^m})$$
$$b_{ik} = \psi_{jk}(X_{i^1}^{\ell_1}) \ldots \psi_{jk}(X_{i^{m_1}}^{\ell_{m_1}}) \Psi_{jk}(X_{i^{m_1+1}}) \ldots \Psi_{jk}(X_{i^{m_1+m_d}}).$$

Let $\tilde{\imath}$ be the set of distinct coordinates in i and let $c = c(\tilde{\imath}_1, \ldots \tilde{\imath}_r) = |\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_r|$ be the overall number of distinct coordinates in r indices $(i_1, \ldots i_r) \in (\Omega_n^m)^r$.

Then

$$E_f^n |a_{i_1 k_1} \ldots a_{i_r k_r}| \leq C 2^{\frac{jd}{2}(mr - 2c)}$$
$$E_f^n |b_{i_1 k_1} \ldots b_{i_r k_r}| \leq C 2^{\frac{jd}{2}(m_d r - 2c_d)} \, 2^{\frac{j}{2}(m_1 r - 2c + 2c_d)}$$

with $c_d = c_d(\tilde{\imath}_1, \ldots \tilde{\imath}_r) \leq c$ the fraction of c corresponding to products with at least one $\Psi(X)$ term and $1 \leq c_d \leq m_d r$, $0 \leq c - c_d \leq m_1 r$, $1 \leq c \leq (m_1 + m_2) r$.

Using lemma 4.4, one can see that the product $a_{i_1 k_1} \ldots a_{i_r k_r}$, made of mr terms, can always be split into $|\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_r|$ independent products of $c(l)$ dependent terms, $1 \leq l \leq |\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_r|$, with $c(l)$ in the range from $|\tilde{\imath}_1| \vee \ldots \vee |\tilde{\imath}_r|$ to mr and $\sum_l c(l) = mr$.

Using lemma 5.18, a product of $c(l)$ dependent terms, is bounded under expectation by $C 2^{\frac{jd}{2}(c(l)-2)}$. Accumulating all independent products, the overall order is $C 2^{\frac{jd}{2}(mr - 2|\tilde{\imath}_1 \cup \ldots \tilde{\imath}_r|)}$.

For $b_{i_1 k_1} \ldots b_{i_r k_r}$ make the distinction between groups containing at least one $\Psi(X)$ term and the others containing only $\psi(X^\ell)$ terms. This splits the number $|\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_d|$ into $g_{\Psi,\psi} + g_\psi$. Let $c_\psi(l)$ be the number of ψ terms in a product of $c(l)$ terms, mixed or not.

On the $g_{\Psi,\psi}$ groups containing Ψ terms, first bound the product of $c_\psi(l)$ terms by $C 2^{\frac{j}{2} c_\psi(l)}$, and the remaining terms by $C 2^{\frac{jd}{2}(c(l) - c_\psi(l) - 2)}$. On the g_ψ groups with only ψ terms, bound the product by $C 2^{\frac{j}{2}(c_\psi(l) - 2)}$.

The overall order is then

$$C 2^{\frac{jd}{2}\left[\left(\sum_{l=1}^{g_{\Psi,\psi}} c(l) - c_\psi(l)\right) - 2 g_{\Psi,\psi}\right]} \, 2^{\frac{j}{2} \sum_{l=1}^{g_{\Psi,\psi}} c_\psi(l)} \, 2^{\frac{j}{2}\left[\left(\sum_{l=1}^{g_\psi} c_\psi(l)\right) - 2 g_\psi\right]}.$$

The final bound is found using $\sum_{l=1}^{g_\psi} c_\psi(l) + \sum_{l=1}^{g_{\Psi,\psi}} c_\psi(l) = m_1 r$ and $\sum_{l=1}^{g_{\Psi,\psi}} c(l) - c_\psi(l) = m_d r$.

Rename $c_d = g_{\Psi,\psi}$ and $c - c_d = g_\psi$.

As for the constraints, in the product of $(m_1 + m_d)r$ terms, it is clear that Ψ terms have to be found somewhere, so $c_d \geq 1$, which also implies that $c - c_d = 0$ when $c = 1$ (in this case there are no independent group with only ϕ terms, but only one big group with all indices equal). Otherwise $c_d \leq m_d r$ and $c - c_d \leq m_1 r$ since there are no more that this numbers of Ψ and ϕ terms in the overall product.

□

5.5 Appendix 2 – Combinatorial lemmas

Lemma 5.12 (Property set)

Let $A_1, \ldots, A_r$ be r non empty subsets of a finite set Ω. Let J be a subset of $\{1, \ldots, r\}$.

Define the property set $B_J = \{x \in \cup A_j : x \in \cap_{j \in J} A_j \; ; \; x \notin \cup_{j \in J^c} A_j\}$, that is to say the set of elements belonging exclusively to the sets listed through J. Let $b_J = |B_J|$ and $b_\kappa = \sum_{|J|=\kappa} b_J$.

Then $\sum_{\kappa=0}^{r} \sum_{|J|=\kappa} B_J = \Omega$, and

$$|A_1| \vee \ldots |A_r| \leq \sum_{\kappa=1}^{r} b_\kappa = |A_1 \cup \ldots A_r| \leq |A_1| + \ldots |A_r| = \sum_{\kappa=1}^{r} \kappa b_\kappa$$

with equality for the right part only if $b_\kappa = 0$, $\kappa = 2 \ldots, r$ i.e. if all sets are disjoint, and equality for the left part if one set A_i contains all the others.

It follows from the definition that no two different property sets intersect and that the union of property sets defines a partition of $\cup A_i$, hence a partition of Ω with the addition of the missing complementary $\Omega - \cup A_i$ denoted by $B_\emptyset$.

With $B_\emptyset$, an overlapping of r sets defines a partition of Ω with cardinality at most 2^r ; there are C_r^κ property sets satisfying $|J| = \kappa$, with $\sum_{\kappa=0}^{r} C_r^\kappa = 2^r$.

□

Lemma 5.13 (Many sets matching indices)

Let $m \in \mathbb{N}$, $m \geq 1$. Let Ω_n^m be the set of indices $\{(i^1, \ldots, i^m) : i^j \in \mathbb{N}, 1 \leq i^j \leq n\}$. Let $r \in \mathbb{N}$, $r \geq 1$. Let $I_n^m = \left\{i \in \Omega_n^m : \ell_1 \neq \ell_2 \Rightarrow i^{\ell_1} \neq i^{\ell_2}\right\}$.

For $i = (i^1, \ldots, i^m) \in \Omega_n^m$, let $\tilde{\imath} = \cup_{j=1}^{m} \{i^j\} \subset \{1, \ldots, n\}$ be the set of distinct integers in i.

Then, for some constant C depending on m,

$$\#\left\{(i_1, \ldots, i_r) \in \left(\Omega_n^m\right)^r, : |\, \tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_r \,| = a\right\} = O(n^a) I\left\{\, |\tilde{\imath}_1| \vee \ldots \vee |\tilde{\imath}_r| \leq a \leq mr\right\}$$

and in corollary $\#\left\{(i_1, \ldots, i_r) \in \left(I_n^m\right)^r : |\, i_1 \cup \ldots \cup i_r \,| = a\right\} = O(n^a) I\left\{m \leq a \leq mr\right\}$.

In the setting introduced by lemma 4.4, building the compound $(\tilde{\imath}_1, \ldots, \tilde{\imath}_r)$ while keeping track of matching indices is achieved by drawing $b^1_{\{1\}} = |\,\tilde{\imath}_1|$ integers in the 2^0–partition $b^0_\emptyset = \{1, \ldots, n\}$ thus constituting $\tilde{\imath}_1$, then $b^2_{\{1,2\}} + b^2_{\{2\}} = |\tilde{\imath}_2|$ integers in the 2^1–partition $\{b^1_{\{1\}}, b^1_\emptyset\}$ thus constituting two subindexes from which to build $\tilde{\imath}_2$, then $b^3_{\{1,2,3\}} + b^3_{\{2,3\}} + b^3_{\{1,3\}} + b^3_{\{3\}} = |\,\tilde{\imath}_3|$ integers in the 2^2–partition $\{b^2_{\{1,2\}}, b^2_{\{1\}}, b^2_{\{2\}}, b^2_\emptyset\}$ thus constituting 2^2 subindexes from which to build $\tilde{\imath}_3$, and so on, up to $b^r_{\{1,\ldots,r\}} + \ldots + b^r_{\{r\}} = |\,\tilde{\imath}_r|$ integers in the cardinality 2^{r-1} partition $\{b^{r-1}_{\{1,\ldots,r-1\}} \ldots, b^{r-1}_\emptyset\}$ thus constituting 2^{r-1} subindexes from which to build $\tilde{\imath}_r$.

The number of ways to draw the subindexes composing the r indexes is then

$$A^{b^1_{\{1\}}}_{b^0_\emptyset}\, A^{b^2_{\{1,2\}}}_{b^1_{\{1\}}} A^{b^2_{\{2\}}}_{b^1_\emptyset} \ldots A^{b^r_{\{1,\ldots,r\}}}_{b^{r-1}_{\{1,\ldots,r-1\}}} \ldots A^{b^r_{\{r\}}}_{b^{r-1}_\emptyset} \tag{37}$$

with the nesting property $b^j_J = b^{j+1}_J + b^{j+1}_{J\cup\{j+1\}}$ (provided J exists at step j) and $A^m_n = \frac{n!}{(n-m)!}$.

At step j, the only property set with cardinality equivalent to n, is $B^{j-1}_\emptyset$, while all others have cardinalities lower than m ; so picking integers inside these light property sets involve cardinalities at most in $m!$ that go in the constants, while the pick in $B^{j-1}_\emptyset$ entails a cardinality $A^{b^j_{\{r\}}}_{b^{j-1}_\emptyset} = A^{b^j_{\{r\}}}_{n-|\,\tilde{\imath}_1\cup\ldots\cup\tilde{\imath}_{j-1}|} \approx n^{b^j_{\{r\}}}$.

Note that, at step $j-1$, $b^{j-1}_\emptyset = n - |\,\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_{j-1}\,|$, because, at step j, $b^j_{\{j\}}$ designates the number of integers in $\tilde{\imath}_j$ not matching any previous index $\tilde{\imath}_1, \ldots, \tilde{\imath}_{j-1}$; so that also $\sum_{j=1}^r b^j_{\{j\}} = |\,\tilde{\imath}_i \cup \ldots \cup \tilde{\imath}_r|$; and incidentally $\sum_{J\ni j_0} b^j_J = |\,\tilde{\imath}_{j_0}|$.

The number of integers picked from the big property set at each step is

$$A^{b^1_{\{1\}}}_{b^0_\emptyset} A^{b^2_{\{2\}}}_{b^1_\emptyset} \ldots A^{b^r_{\{r\}}}_{b^{r-1}_\emptyset}$$

with $b^j_\emptyset = n - |\,\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_{j-1}|$, $b^0_\emptyset = n$ and $\sum_{j=1}^r b^j_{\{j\}} = |\,\tilde{\imath}_i \cup \ldots \cup \tilde{\imath}_r|$.

For large n this is equivalent to $n^{|\,\tilde{\imath}_1\cup\ldots\cup\tilde{\imath}_r|}$.

Having drawn the subindexes, building the indexes effectively is a matter of iteratively intermixing two sets of a and b elements ; an operation equivalent to highlighting b cells in a line of $a+b$ cells, which can be done in C^b_{a+b} ways, with $C^p_n = A^p_n/p!$.

Intermixing the subindexes thus involve cardinalities at most in $m!$, that go in the constant C.

Likewise, passing from $\tilde{\imath}$ to i involve cardinalities at most in $C^{|\tilde{\imath}|}_m$ and no dependence on n.

For the corollary, if $i \in I^m_n$ then $\tilde{\imath} = i$ and $|\tilde{\imath}| = m$. If moreover $i^1 < \ldots < i^r$, the number of ways to draw the subindexes is given by replacing occurrences of 'A' by 'C' in (37), with $C^m_n = \frac{n!}{m!(n-m)!}$, which does not change the order in n. Also there is only one way to intermix subindexes, because of the ordering constraint.

□

Lemma 5.14 (Path of non matching dimension numbers)

Let $r \in \mathbb{N}$, $r \geq 2$. Let $\Omega_n^m = \{(i^1, \ldots, i^m) : i^\ell \in \mathbb{N}, 1 \leq i^\ell \leq n\}$. For $i \in \Omega_n^d$, let $\Lambda_{jk}(V_i) = \psi_{jk}(X_{i^1}^1) \ldots \psi_{jk}(X_{i^d}^d)$. Let $\tilde{\imath}$ be the set of distinct coordinates of i.

In the product

$$\Big(\sum_j \sum_k \frac{1}{n^d} \sum_{i \in \Omega_n^d} \Lambda_{jk}(V_i)\Big)^r = \frac{1}{n^{dr}} \sum_{i_1, \ldots, i_r \in (\Omega_n^d)^r} \sum_{j_1 \ldots j_r} \sum_{k_1 \ldots, k_r} \Lambda_{j_1 k_1}(V_{i_1}) \ldots \Lambda_{j_r k_r}(V_{i_r})$$

unless $|\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_r| < r$, it is always possible to find indices $(i_1, \ldots, i_r)$ such that no two functions ψ_{jk} $\psi_{jk'}$ matching observation number also match dimension number.

Let $c = |\tilde{\imath}_1 \cup \ldots \cup \tilde{\imath}_r|$. For $1 \leq \ell \leq n$, let $\ell^{\otimes d} = (\ell, \ldots, \ell) \in \Omega_n^d$.

With r buckets of width d defined by the extent of each index $k_1 \ldots, k_r$, and only $c < r$ distinct observation numbers, once c buckets have been stuffed with terms $V_{\ell^{\otimes d}}$, some already used observation number must be reused in order to fill in the remaining $r - c$ buckets. So that $r - c$ buckets will match on dimension and observation number allowing to reduce the sum to only c distinct buckets.

Once $c > r$, starting with a configuration using $V_{\ell_1^{\otimes d}}, \ldots V_{\ell_r^{\otimes d}}$ we can always use additional observation numbers to fragment further the $\ell^{\otimes d}$ terms, which preserves the empty intersection between buckets.

□

5.6 Appendix 3 – Thresholding related lemmas and others

Lemma 5.15 (wavelet contrast in β_{jk})

With $j_0 \leq j_1$ one has,

$$C_{j_1+1}(f) = C_{j_0}(f) + \sum_{j=j_0}^{j_1} \sum_k (\beta_{jk} - \beta_{jk^1} \ldots \beta_{jk^d})^2.$$

where $k = (k^1, \ldots, k^d)$ and $\beta_{jk^\ell} = \int f^{\star \ell}(x^\ell) \psi_{jk^\ell}(x^\ell)\, dx^\ell$ and $\beta_{jk} = \int f(x) \Psi_{jk}(x)\, dx$.

Recall the reconstruction part of the cascade algorithm :

$$\alpha_{j+1,k} = \sum_\ell c_{k-2\ell} \alpha_{j\ell} + d_{k-2\ell} \beta_{j\ell}$$

with c_k and d_k two orthogonal quadrature filters satisfying for $n \in \mathbb{N}$, $\sum_k c_k c_{k+2n} = \delta(n)$, $\sum_k d_k d_{k+2n} = \delta(n)$ – with $\delta(.)$ the Kronecker delta – and $\sum_k c_k d_{k+2n} = \sum_k d_k c_{k+2n} = 0$ (Daubechies, 1992, p.137).

These relations extends easily to the multidimensional case.

Write $\dot{\alpha}_{jk} = \alpha_{jk^1} \dots \alpha_{jk^d}$ and $\dot{\beta}_{jk} = \beta_{jk^1} \dots \beta_{jk^d}$ for the coordinates on V_j^d and W_j^d of the projections of $f^{\star 1} \dots f^{\star d}$; so one has,

$$\delta_{j+1,k} = \sum_{\ell} c_{k-2\ell}(\alpha_{j\ell} - \dot{\alpha}_{j\ell}) + d_{k-2\ell}(\beta_{j\ell} - \dot{\beta}_{j\ell}) \, ;$$

so that,

$$\begin{aligned}\delta_{j+1,k}^2 = \sum_{\ell,\ell'} & c_{k-2\ell}c_{k-2\ell'}(\alpha_{j\ell} - \dot{\alpha}_{j\ell})(\alpha_{j\ell'} - \dot{\alpha}_{j\ell'}) + d_{k-2\ell}d_{k-2\ell'}(\beta_{j\ell} - \dot{\beta}_{j\ell})(\beta_{j\ell'} - \dot{\beta}_{j\ell'}) \\ & + 2c_{k-2\ell}d_{k-2\ell'}(\alpha_{j\ell} - \dot{\alpha}_{j\ell})(\beta_{j\ell'} - \dot{\beta}_{j\ell'}),\end{aligned}$$

and,

$$\begin{aligned}C_{j+1} &= \sum_{\ell,\ell'} \Big[(\alpha_{j\ell} - \dot{\alpha}_{j\ell})(\alpha_{j\ell'} - \dot{\alpha}_{j\ell'}) \sum_k c_{k-2\ell}c_{k-2\ell'} + (\beta_{j\ell} - \dot{\beta}_{j\ell})(\beta_{j\ell'} - \dot{\beta}_{j\ell'}) \sum_k d_{k-2\ell}d_{k-2\ell'} \\ &\qquad + 2(\alpha_{j\ell} - \dot{\alpha}_{j\ell})(\beta_{j\ell'} - \dot{\beta}_{j\ell'}) \sum_k c_{k-2\ell}d_{k-2\ell'} \Big] \\ &= \sum_{\ell} (\alpha_{j\ell} - \dot{\alpha}_{j\ell})^2 + (\beta_{j\ell} - \dot{\beta}_{j\ell})^2 \\ &= C_j + \sum_k (\beta_{jk} - \beta_{jk^1} \dots \beta_{jk^d})^2,\end{aligned}$$

since $\sum_k c_{k-2\ell}c_{k-2\ell'} = \sum_k c_k c_{k+2(\ell-\ell')} = \delta(\ell - \ell')$, and likewise $\sum_k d_{k-2\ell}d_{k-2\ell'} = \delta(\ell - \ell')$ and $\sum_k c_{k-2\ell}d_{k-2\ell'} = 0$.

Other summing ranges in j are obtained by recurrence.
□

Lemma 5.16 (Large deviation for term by term thresholding)

Let $\hat{\beta}_{jk}$ and $\hat{\beta}_{jk^\ell}$ be the moment estimator of the wavelet coefficients as defined in (7), and suppose that $2^{jd} < n/\log n$; then

$$P_{f_A}^n [\, |\hat{\beta}_{jk} - \beta_{jk}| > \frac{c_t}{2}\sqrt{\frac{\log n}{n}} \,] \leq 2n^{-c_t/C} \tag{38}$$

and,

$$P_{f_A}^n [\, |\hat{\beta}_{jk^\ell} - \beta_{jk^\ell}| > \frac{c_t}{2} \left[\frac{\log n}{n}\right]^{\frac{1}{2d}} \,] \leq 2 \exp\left(-\frac{c_t}{C} n\right)$$

Applying Bernstein inequality to $Y_i = \Psi_{jk}(X_i) - E_{f_A}^n \Psi_{jk}(X_i)$, with,

$$\begin{aligned}|Y_i| \leq \|Y\|_\infty \leq 2^{jd/2}\|\Psi\|_\infty + \beta_{jk} \leq 2\|\Psi\|_\infty 2^{jd/2}, \\ E_{f_A}^n Y_i^2 \leq E_{f_A}^n \Psi_{jk}(X_i)^2 \leq \|f_A\|_\infty, \\ \lambda = \frac{c_t}{2}\sqrt{\frac{\log n}{n}},\end{aligned}$$

we find that,

$$P^n_{f_A}[\,|\hat{\beta}_{jk}-\beta_{jk}|>\frac{c_t}{2}\sqrt{\frac{\log n}{n}}\,]\leq 2\exp\left(\frac{-c_t^2\log n}{8(\|f_A\|_\infty+2\|\Psi\|_\infty\frac{c_t}{6}[\frac{\log n}{n}2^{jd}]^{\frac{1}{2}})}\right)\;; \tag{39}$$

next since $j\leq j_1$, we have $2^{jd}\leq 2^{j_1 d}\leq C\frac{n}{\log n}$, and so,

$$P^n_{f_A}[\,|\hat{\beta}_{jk}-\beta_{jk}|>\frac{c_t}{2}\sqrt{\frac{\log n}{n}}\,]\leq 2n^{\left(\frac{-c_t^2}{8(\|f_A\|_\infty+2\|\Psi\|_\infty C\frac{c_t}{6})}\right)},$$

and we can choose c_t such that $\|f_A\|_\infty\leq Cc_t$; so that it simplifies further to,

$$P^n_{f_A}[\,|\hat{\beta}_{jk}-\beta_{jk}|>\frac{c_t}{2}\sqrt{\frac{\log n}{n}}\,]\leq 2n^{-c_t/C}. \tag{40}$$

Likewise applying Bernstein inequality to $Y_i=\Psi_{jk^\ell}(X_i^\ell)-E^n_{f_A}\Psi_{jk^\ell}(X_i^\ell)$, with,

$$\begin{aligned}|Y_i|\leq\|Y\|_\infty&\leq 2^{j/2}\|\psi\|_\infty+\beta_{jk^\ell}\leq 2\|\psi\|_\infty 2^{j/2},\\ E^n_{f_A}Y_i^2&\leq E^n_{f_A}\psi_{jk}(X_i)^2\leq\|f_A^{\star\ell}\|_\infty,\\ \lambda&=\frac{c_t}{2}\left[\frac{\log n}{n}\right]^{\frac{1}{2d}},\end{aligned}$$

we find that,

$$P^n_{f_A}[\,|\hat{\beta}_{jk^\ell}-\beta_{jk^\ell}|>\frac{c_t}{2}\left[\frac{\log n}{n}\right]^{\frac{1}{2d}}\,]\leq 2\exp\left(\frac{-c_t^2(\log n)^{\frac{1}{d}}n^{1-1/d}}{8(\|f_A^{\star\ell}\|_\infty+2\|\psi\|_\infty\frac{c_t}{6}\left[\frac{\log n}{n}^{1/d}2^j\right]^{\frac{1}{2}})}\right)\;; \tag{41}$$

again since $j\leq j_1$, we have $2^{jd}\leq 2^{j_1 d}\leq C\frac{n}{\log n}$, and so,

$$P^n_{f_A}[\,|\hat{\beta}_{jk^\ell}-\beta_{jk^\ell}|>\frac{c_t}{2}\left[\frac{\log n}{n}\right]^{\frac{1}{2d}}\,]\leq 2\exp\left(\frac{-c_t^2(\log n)^{\frac{1}{d}}n^{1-1/d}}{8(\|f_A^{\star\ell}\|_\infty+2\|\psi\|_\infty\frac{c_t}{6}C}\right)$$

and we can choose c_t such that $\|f_A\|_\infty\leq Cc_t$; so that it simplifies further to,

$$\begin{aligned}P^n_{f_A}[\,|\hat{\beta}_{jk^\ell}-\beta_{jk^\ell}|>\frac{c_t}{2}\left[\frac{\log n}{n}\right]^{\frac{1}{2d}}\,]&\leq 2\exp\left(-c_t/C(\log n)^{\frac{1}{d}}n^{1-1/d}\right)\\ &\leq 2\exp\left(-\frac{c_t}{C}n\right)\end{aligned}$$

□

The third lemma ensures that for $f\in B_{srq}(\mathbb{R}^d)$ and with a choice of resolution j_1 suitably related to the threshold t, the number of big coefficients in the expansion truncated at j_1 is increasing at a rate slower than n.

Lemma 5.17 (Number of big coefficients)

Assume that f_A belongs to the Besov class $B_{spq}(\mathbb{R}^d)$, and choose $2^{j_1 d} = C\frac{n}{\log n}$ and $t = C\sqrt{\frac{\log n}{n}}$; Set

$$B = \{j,k \ : \ |\beta_{jk}| > t/2\}, \quad B^\ell = \{j,k^\ell \ : \ |\beta_{jk^\ell}| > (t/2)^{\frac{1}{d}}\}$$
$$B_\mu = B^1 \cap \ldots \cap B^d \subset \{j,k \ : \ |\mu_{jk}| > t/2\},$$

with $\mu_{jk} = \beta_{jk^1} \ldots \beta_{jk^d}$.

Then :

$$a) \quad \sum_{j=0}^{j_1} \sum_k I(B) = \#\{(j,k), |\beta_{jk}| \geq \frac{t}{2}\} \leq Cn^{\frac{d}{2s+d}},$$

and likewise,

$$b) \quad \sum_{j=0}^{j_1} \sum_k I(B_\mu) \leq \#\{(j,k), |\mu_{jk}| \geq \frac{t}{2}\} \leq Cn^{\frac{d}{2s+d}}.$$

For a), applying Markov inequality to the random variable $X\colon (j,k) \mapsto |\beta_{jk}|$ defined on the set $\Omega_{j_s j_1} = \{(j,k)\colon j_s \leq j \leq j_1,\ 0 \leq k^l < 2^j, l = 1 \ldots d\}$ we see that,

$$\#\Omega_{j_s j_1}\, Pr(X \geq \frac{t}{2}) = \#\{(j,k), |\beta_{jk}| \geq \frac{t}{2}\} \leq \left(\frac{2}{t}\right)^r \sum_{j=j_s}^{j_1} \sum_k |\beta_{jk}|^r,$$

with # the notation for set cardinal.

So that, using the B_{srq} membership of f_A as expressed by (28), *i.e.* the fact in particular that,

$$\sum_{j=j_s}^{j_1} \sum_k 2^{jr(s+\frac{d}{2}-\frac{d}{r})} |\beta_{jk}|^r < C,$$

we have,

$$\begin{aligned} \sum_{j=j_s}^{j_1} \sum_k I(B) &\leq 2^r t^{-r} 2^{-j_s r(s+\frac{d}{2}-\frac{d}{r})} \sum_{j=j_s}^{j_1} \sum_k 2^{jr(s+\frac{d}{2}-\frac{d}{r})} |\beta_{jk}|^r \\ &\leq \left(\frac{n}{\log n}\right)^{\frac{r}{2}} 2^{-j_s r(s+\frac{d}{2}-\frac{d}{r})} \\ &\leq Cn^{r/2} 2^{-j_s r(s+\frac{d}{2}-\frac{d}{r})}. \end{aligned} \tag{42}$$

And also for the remaining sum down to j_0,

$$\sum_{j=j_0}^{j_s} \sum_k I(B) \leq \frac{2^{(j_s+1)d}-1}{2^d-1} \leq 2^{j_s d} \approx 2^{j_s d}. \tag{43}$$

Take $2^{j_s} = n^{\frac{1}{2s+d}}$, the last line of (42) changes into $Cn^{\frac{d}{2s+d}}$, which has constant rate for all r ; next in (43), the sum is also bounded by $n^{\frac{d}{2s+d}}$.

For b), since we assumed that any marginal is in $B_{spq}(\mathbb{R})$, the product of marginal distributions is in $B_{spq}(\mathbb{R}^d)$ (see Barbedor, 2005), and so using the same argument than in (42), we have

$$\sum_{j=j_s}^{j_1}\sum_k I(B_\mu) \leq \#\{(j,k^\ell), |\beta_{jk^1}\dots\beta_{jk^d}| \geq \frac{t}{2}\} \leq \frac{2^r}{t}\sum_j\sum_{k^\ell}|\beta_{jk^1}\dots\beta_{jk^d}|^r\ ;$$

so that, with the same choice $2^{j_s} = n^{\frac{1}{2s+d}}$, and (43) also true for B_μ, the sum of indicators is of the order of $Cn^{\frac{d}{2s+d}}$. □

Lemma 5.18 (rth order moment of Ψ_{jk})

Let X be random variables on $\mathbb{R}^d$ with density f. Let Ψ be the tensorial wavelet of an MRA of $L_2(\mathbb{R}^d)$. Let $\beta_{jk} = E_f\Psi_{jk}(X)$. Then for $r \in \mathbb{N}^*$,

$$E_f|\Psi_{jk}(X) - \beta_{jk}|^r \leq 2^r E_f|\Psi_{jk}(X)|^r \leq 2^r 2^{jd(\frac{r}{2}-1)}\|f\|_\infty\|\Psi\|_r^r.$$

For the left part of the inequality, $\Big(E_f|\Psi_{jk}(X) - \beta_{jk}|^r\Big)^{\frac{1}{r}} \leq \Big(E_f|\Psi_{jk}(X)|^r\Big)^{\frac{1}{r}} + E_f|\Psi_{jk}(X)|$, and also $E_f|\Psi_{jk}(X)| \leq \Big(E_f|\Psi_{jk}(X)|^r\Big)^{\frac{1}{r}}\Big(E_f 1\Big)^{\frac{r-1}{r}}$.

For the right part, $E_f|\Psi_{jk}(X)|^r = 2^{jdr/2}\int|\Psi(2^jx-k)|^r f(x)dx \leq 2^{jd(\frac{r}{2}-1)}\|f\|_\infty\|\Psi\|_r^r$.
□

Lemma 5.19 (Daubechies wavelet concentration property)

Let $r \in \mathbb{N}$, $r \geq 1$. Let ψ be a Daubechies $D2N$ wavelet. Let h_{jk} be the function on $\mathbb{R}^m$ defined as a product of translations of ψ

$$h_{jk}(x_1,\dots,x_m) = \psi(2^jx_1 - k^1)\dots\psi(2^jx_m - k^m),$$

with $k = (k^1,\dots,k^m) \in \mathbb{Z}^m$.

Then,

$$\Big(\sum_{j=0}^J\sum_k|h_{jk}(x_1,\dots x_m)|\Big)^r \leq \big[2^{J+1}(4N-3)\big]^{m(r-1)}\sum_{j=0}^J\sum_k|h_{jk}(x_1,\dots x_m)|^r \tag{44}$$

With a compactly supported Daubechies wavelet $D2N$, whose support is $[0, 2N-1]$, the product $x \mapsto \psi_{jk}(x)\psi_{j'\ell}(x)$ is zero for $\frac{k+2N-1}{2^j} < \frac{\ell}{2^{j'}}$ that is to say for $|2^{j-j'}\ell - k| > 2N-1$, with $j \geq j'$ say.

When j and k are fixed, and $0 \leq j, j' \leq J$, the cardinal of the set $|2^{j-j'}\ell - k| \leq 2N-1$ is at most equal to $(4N-3) + (4N-3)2^{\epsilon 1} + \dots(4N-3)2^{\epsilon|j-j'|}$, with $\epsilon =$ sign $j'-j$. That is to say

there are twice more points at each higher resolution and twice less points at each lower resolution. The cardinal is then lower than $2^{J+1}(4N-3)$ in any case.

So that,

$$\Big(\sum_{j=0}^{J}\sum_{k} h_{jk}\Big)^r = \sum_{j_1,\dots j_r}\sum_{k_1,\dots k_r} h_{j_1k_1}\dots h_{j_rk_r}\, I(\Delta)$$

with $\Delta = \{2^{j_{i_1}-j_{i_2}}|k_{i_3}^{\ell_1} - k_{i_4}^{\ell_2}| < (2N-1)\,;\ \ i_1, i_2, i_3, i_4 = 1\dots r\,;\ \ \ell_1,\ell_2 = 1\dots m\,;\ \}$. Once k_1 and j_1 say, is fixed, the cardinal of Δ is not greater than $[2^{J+1}(4N-3)]^{m(r-1)}$.

Using $(|h_{j_1k_1}|^r\dots|h_{j_rk_r}|^r)^{\frac{1}{r}} \le \frac{1}{r}\sum_i |h_{j_ik_i}|^r$, for $r \ge 1$,

$$\begin{aligned}\Big(\sum_j\sum_k |h_{jk}|\Big)^r &\le \sum_{j_1,k_1,\dots,j_r,k_r} \frac{1}{r}\big(|h_{j_1k_1}|^r + \dots + |h_{j_rk_r}|^r\big)\,\mathbb{I}\{\Delta\}\\ &= \frac{1}{r}\Bigg[\sum_{j_1,k_1,\dots,j_r,k_r} |h_{j_1k_1}|^r\,\mathbb{I}\{\Delta\} + \dots + \sum_{j_1,k_1,\dots,j_r,k_r} |h_{j_rk_r}|^r\,\mathbb{I}\{\Delta\}\Bigg]\\ &\le \big[2^{J+1}(4N-3)\big]^{m(r-1)}\sum_j\sum_k |h_{jk}|^r,\end{aligned}$$

□

Lemma 5.20 (Bernstein inequality)

if $Y_1,\dots,Y_n$ are i.i.d. bounded random variables such that $E(Y_i) = 0$, $E(Y_i^2) \le \sigma^2$, and $|Y_i| \le \|Y\|_\infty < +\infty$, then,

$$Pr\left[|\frac{1}{n}\sum_i Y_i| > \lambda\right] \le 2\exp\left(\frac{-n\lambda^2}{2(\sigma^2 + \|Y\|_\infty\lambda/3)}\right).$$

□

5.7 References for Towards thresholding

(Barbedor, 2005) P. Barbedor. Independent component analysis by wavelets. *Technical Report PMA-995*, Laboratoire de probabilités et Modèles aléatoires, Université Paris VII, 2005

(Cai & Low, 2005) T. Cai and M. Low. *Optimal adaptive estimation of a quadratic functional.* The Annals of Statistics, to appear.

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

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

(Giné, 1996) E. Giné. *Decoupling and limit theorems for U-statistics and U-processes.* Lectures on probability theory and statistics, Saint-Flour, 1996, 1-35.

(Giné et al. 2000) E. Giné, R. Latała, J. Zinn. *Exponential and moments inequalities for U-statistics*. High dimensional probability II, Progress in Probability, 47, (2000), 13-38.

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

(Kerkyacharian & Picard, 1996) Gérard Kerkyacharian Dominique Picard. Estimating non quadratic functionals of a density using Haar wavelets. *Annals of Statistics*, 24(1996), 485-507.

(Koroljuk & Borovskich, 1994) V. S. Koroljuk and Yu. V. Borovskich. Theory of U-statistics. *Kluwer academic press*, 1994.

(Rosenthal, 1972) Rosenthal, H. P. On the span in lp of sequences of independent random variables. *Israel J. Math.* 8 273–303, 1972.

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

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

6. Stiefel manifold, optimization and implementation issues

The estimator $\hat{C}_j$ 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.

6.1 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 φ_{jk} whose support contains x, namely $\varphi_{je_j} \dots \varphi_{je_j-2N+2}$, 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 α_{jk}, we use circular shifting).

So one needs to compute for each x, $2^{\frac{j}{2}}\varphi(f_j), \dots 2^{\frac{j}{2}}\varphi(f_j + 2N - 2)$, 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`$[i], \dots,$ `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 $k \in \{\lfloor 2^j x \rfloor - k', \quad k' = 0 \dots 2N-2\}$

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

and the index will point to the exact same values in the 2^b times larger table of precomputed φ 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 $b \geq L$.

6.2 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, ℓ) 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^ℓ the (normed) line vector ℓ of W, and $(Y_c)_i$ the column vector i of Y_c.

This way an initial relocation covers all the minimization process.

6.3 Wavelet contrast differential and implementation issues

Recall the notation,

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

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

Proposition 6.21 **($\hat{C}_j$ differential in Y)**

If the wavelet φ 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,

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

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

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

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

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

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

$$\begin{aligned}
D_i^\ell \hat{\delta}_{jk} =& D_i^\ell \left[\frac{1}{n} \sum_i \Phi_{jk}(y_i) - \prod_h \frac{1}{n} \sum_j \varphi_{jk_h}(y_j^h) \right] \\
=& D_i^\ell \left[\frac{1}{n} \sum_i \prod_h \varphi_{jk_h}(y_i^h) - \prod_h \frac{1}{n} \sum_j \varphi_{jk_h}(y_j^h) \right] \\
=& \frac{1}{n} \varphi'_{jk_\ell}(y_i^\ell) \prod_{h \neq \ell} \varphi_{jk_h}(y_i^h) - \frac{1}{n^d} \varphi'_{jk_\ell}(y_i^\ell) \prod_{h \neq \ell} \sum_j \varphi_{jk_h}(y_j^h) \\
=& \varphi'_{jk_\ell}(y_i^\ell) \left(\frac{1}{n} \prod_{h \neq \ell} \varphi_{jk_h}(y_i^h) - \frac{1}{n^d} \prod_{h \neq \ell} \sum_j \varphi_{jk_h}(y_j^h) \right),
\end{aligned} \tag{45}$$

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

$D_i^\ell \hat{\delta}_{jk}$ 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^ℓ,

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

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 6.22 **($\hat{C}_j$ differential in W)**

Under the notation of Prop. 6.21, with $C_j(W) \equiv C_j(Wx_1, \dots, Wx_n)$ and $X = (\, x_1 \quad \dots \quad x_n \,)$, the $d \times n$ matrix arrangement of the observation $\{x_1 \dots, x_n\}$, 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,

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

with D_i^ℓ, the abridged notation for $D_i^j\hat{\delta}_{jk}$ given in Prop. 6.21.

Under column major order,

$$dC_j(W) = 2\sum_i X_i \otimes \sum_k \hat{\delta}_{jk} D_i \tag{46}$$

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

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

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

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

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

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

$(Wx)_{ij} = \sum_{k=1\dots,d} W_{ik} x_j^k$, 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,

$$dT(W) = dT = \begin{pmatrix} {}^tX & & & \\ & {}^tX & & \\ & & \ddots & \\ & & & {}^tX \end{pmatrix} \tag{47}$$

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 ${}^tX$.

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

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

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

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

And,

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

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

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

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

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 6.23 **($\hat{C}_j$ 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 $(i,\ell) = (i',\ell')$,

$$\begin{aligned}
D_i^{\ell} D_i^{\ell}\hat{\delta}_{jk} &= \varphi''_{jk^{\ell}}(y_i^{\ell})\left(\frac{1}{n}\prod_{h\neq\ell}\varphi_{jk_h}(y_i^h) - \frac{1}{n^d}\prod_{h\neq\ell}\sum_j \varphi_{jk_h}(y_j^h)\right)\\
&= \varphi''_{jk^{\ell}}(y_i^{\ell}) Q_1(i,\ell)
\end{aligned} \tag{48}$$

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

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

$$\begin{aligned}
D_{i'}^{\ell'} D_i^{\ell}\hat{\delta}_{jk} &= \varphi'_{jk^{\ell}}(y_i^{\ell})\varphi'_{jk^{\ell'}}(y_i^{\ell'})\left[\frac{1}{n}\prod_{h\neq\ell,h\neq\ell'}\varphi_{jk^h}(y_i^h) - \frac{1}{n^d}\prod_{h\neq\ell,h\neq\ell'}\sum_j \varphi_{jk^h}(y_j^h)\right]\\
&= \varphi'_{jk^{\ell}}(y_i^{\ell})\varphi'_{jk^{\ell'}}(y_i^{\ell'}) Q_2(i,\ell,\ell')
\end{aligned} \tag{49}$$

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

$$\begin{aligned} D^{\ell'}_{i'} D^{\ell}_{i} \hat{\delta}_{jk} &= \varphi'_{jk^\ell}(y^\ell_i)\varphi'_{jk^{\ell'}}(y^{\ell'}_{i'}) \left[-\frac{1}{n^d} \prod_{h\neq\ell, h\neq\ell'} \sum_j \varphi_{jk^h}(y^h_j) \right] \\ &= \varphi'_{jk^\ell}(y^\ell_i)\varphi'_{jk^{\ell'}}(y^{\ell'}_{i'}) Q_3(\ell,\ell') \end{aligned} \tag{50}$$

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

Proposition 6.24 **($\hat{C}_j$ 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 (48) ;

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 (49), and whose terms $(ii') = \varphi'_{jk^\ell}(y^\ell_i)\varphi'_{jk^{\ell'}}(y^{\ell'}_{i'}) Q_3(\ell,\ell')$ are given by (50) ; and with $(D^\ell_i \hat{\delta}_{jk})_{(i\ell)}$ the $nd \times 1$ vector given in (45).

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,

$$\nabla^2 C_j(Y) = 2\sum_k \begin{pmatrix} D^1\, {}^tD^1 & \dots & D^1\, {}^tD^d \\ & \ddots & \\ D^d\, {}^tD^1 & \dots & D^d\, {}^tD^d \end{pmatrix} + \hat{\delta}_{jk} \begin{pmatrix} B^{11} & \dots & B^{1d} \\ & \ddots & \\ B^{d1} & \dots & B^{dd} \end{pmatrix} \tag{51}$$

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 (51) 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 6.25 **($\hat{C}_j$ 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 $d^2 C_j(W)(H_1, H_2) = d^2 C_j(Y)(H_1 X, H_2 X)$.

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

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

where dT is the matrix given in (47) 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,

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

So we have,

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

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

6.4 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 accommodates 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,

$$\begin{aligned} \varphi_{jk}(t) &= 2^{\frac{j}{2}}\varphi(2^j t - k) \\ &= 2^{\frac{j}{2}}\sqrt{2}\sum_\ell c_\ell \varphi(2(2^j t - k) - \ell) \\ &= \sum_\ell c_\ell \varphi_{j+1,2k+\ell}(t) = \sum_\ell c_{\ell-2k}\varphi_{j+1,\ell}(t) \end{aligned} \tag{52}$$

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 α_{jk}, and to $\hat{\delta}_{jk} = \hat{\alpha}_{jk} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d}$:

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

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

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

Proposition 6.26 (filter aware gradient formulation)

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

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

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

The following relation holds,

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

and starting from (45), the partial derivative can be expressed by,

$$\begin{aligned}
D^\ell_i\hat{\delta}_{jk} =& \frac{\varphi'_{jk_\ell}(X^\ell_i)}{\varphi_{jk_\ell}(X^\ell_i)}\left[\hat{\alpha}_{jk} - \frac{n-1}{n}\hat{\alpha}^{(i)}_{jk} - (\hat{\alpha}_{jk^\ell} - \frac{n-1}{n}\hat{\alpha}^{(i)}_{jk^\ell})\prod_{h\neq\ell}\hat{\alpha}_{jk^h}\right] \\
=& \frac{\varphi'_{jk_\ell}(X^\ell_i)}{\varphi_{jk_\ell}(X^\ell_i)}\left[\hat{\delta}_{jk} - \frac{n-1}{n}\left[\hat{\alpha}^{(i)}_{jk} - \hat{\alpha}^{(i)}_{jk^\ell}\prod_{h\neq\ell}\hat{\alpha}_{jk^h}\right]\right].
\end{aligned} \tag{53}$$

□

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 (53), 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 φ'/φ will produce zero, and from expression (46) 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 φ''_{jk}, as we see next.

Proposition 6.27 (filter aware hessian formulation)

The matrices $B^{\ell\ell}$ composing the hessian can be written as a function of Jackknife of the original α_{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,

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

with the bracket also written as,

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

And,

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

Use the jackknife substitution. □

6.5 Wavelet contrast implementation

In the full formulation, that provides a true mutual independence criteria, the programmation of the wavelet contrast requires a good amount of memory and must be based on a flat implementation of entities in dimension d, if one wishes a programmation independent of d.

Indeed with 1 Go (2^{30} bytes) and in double precision (floating point numbers on 8 bytes), one obtains a theoretical maximum of $jd = 27$, very easely reached.

Flat representation routines

The projection space is a cube with side 2^j and volume 2^{jd}, to which is added d segments of length 2^j representing the margins.

The flat representation of the cube is a segment of length 2^{jd} in C order, that is to say with indices of the last dimension contiguous in memory, while indices of other dimensions are separated by offsets ranging from 2^j for the next to last, to 2^{jd-1} for the first one. This is in fact the underlying representation of any multi-dimensional array since a computer memory is a long segment of dimension 1.

To program the DWT in dimension d, you need a routine to transpose a flat d dimensional array. This routine can be found in the book of Wickerhauser (1990), and has been used under a simplified version, taking into account that the array is a cube, with the same numbers of elements along any dimension.

- To transform a developed coordinate $m = (m_1, \ldots, m_d)$ in linear offset, perform a scalar product with the translator `magicv` $= (2^{j(d-1)}, \ldots, 2^j, 1)$. Conversely, to pass from a linear offset i to a developed coordinate m, use the integer division modulo 2^j,

  $$m = (i/2^{j(d-1)}\ mod\ 2^j, \ldots, i/2^j\ mod\ 2^j, i\ mod\ 2^j)$$

  The two operations are written for instance in fortran `mod( i/magicv, dpj)`, with `dpj` equal to 2^j, and `dot_product( magicv, m)`.

- To find all offsets within `dxN` positions of a fixed offset o in any of the d dimensions of the cube, define a minicube with side `dxN` whose translator is `magic_deuxN` $= ((\texttt{dxN}-1)^{(d-1)}, \ldots, 1)$ and loop over the $(\texttt{dxN}-1)^d$ values to constitue the developed coordinates

```
do ii = 0, (dxN - 1) ** d -1
indice = mod( ii / magic_deuxN, dxN - 1)
enddo
```

that next can be added to the developed coordinates of o, with wrap around in case the computed coordinate passes the limits of the general cube.

This process is used in the computation of the projection on the space V_j (see below), where `dxN` is the parameter of the Daubechies $D2N$.

- To transpose a flat cube, we use the routine explained in Wickerhauser (1994, p316).

```
void transpose(double * vec, int vol, int dpj)[
   int i, t, s ;
   char * hasmoved ;
   double temp ;
   hasmoved = (char *) malloc( (size_t) vol * sizeof(char)) ;
   memset(hasmoved, 0, vol) ;  // set region to character '0'
   for (i=1 ;  i < vol -2 ;  i++)[ // first and last index unchanged
        if (hasmoved[i]=='1') continue ;
        temp = vec[i] ;
        t = i ;
        s = i * dpj % (vol-1) ;
        while (s > i) [
        hasmoved[s] ='1' ;
        vec[t] = vec[s] ;
        t = s ;
        s = s * dpj % (vol -1) ;
        ]
        vec[t] = temp ;
   ]
   free(hasmoved) ;

]
```

DWT in dimension 1

The classical formula $\alpha_{jk} = \sum_\ell c_{\ell-2k}\alpha_{j+1,\ell} = (\alpha_{j+1} * \bar{c})_{2k}$ and its inverse, the only one cited in most of the mathematical books on wavelets, does not allow on its own a real programmation of the DWT.

A complete algorithm of the DWT in dimension 1 appears in Numerical recipies (1986). One can find also in the book of Wickerhauser (1994) an explicitation of the relationship between the DWT and a classical filtering. The filter of the Daubechies wavelet is an instance of a quadratic orthogonal filter and the DWT is a standard convolution and decimation operation of periodic type.

We reproduce the routine adapted to our usage below.

```
void dwt(double * vec, int lvec, const double * filt, const double * rfilt, int lfilt,
     int direction)[
```

`lvec` power of two, length of the signal
`vec` workspace of length at least `lvec`
`filt, rfilt` Daubechies filter with length `lfilt`
`direction` 1 for deconstruction , -1 for reconstruction

```
double * wksp ;
int nh = lvec / 2, i, iw, j, wrap = lvec - 1 ;
```

* `lvec` being a power of 2, `wrap` is written in binary digits only with 1 ; & designating the bitwise "and", and let $w = 2^n - 1$, $c \leq w \Rightarrow c \& w = c$, and $c > w \Rightarrow c \& w = c \, mod \, (w+1)$; by this process on obtains an infinite periodic extension of `vec[0,...,lvec-1]` with no memory consumption.

* dwt needs a workspace `wksp` of size `lvec`

```
allovec( wksp, lvec) ;
setvec( wksp, lvec, 0.) ;
if ( direction >=0)
```

* increment of 2 for index `i` of `vec`, to obtain the decimation effect

```
     for ( iw=0, i=0 ;  i < lvec ;  i += 2, iw++)
          for ( j=0 ;  j < lfilt ;  j++)[
```

* term by term multiplication of the filter and `vec` at offset `i` ; `& wrap` is active when the filter passes the right bound of `vec`

```
          wksp[ iw] += filt[ j] * vec[ (i + j) & wrap ] ;
```

* above, the alphas are grouped in the first part of `wksp`

* the betas are placed in the second part (offset `nh`)

```
          wksp[ iw + nh] += rfilt[ j] * vec[ (i + j) & wrap] ;
          ]
```

* idem in reverse direction when `direction` $= -1$

```
else
     for (iw=0, i=0 ;  i < lvec ;  i+=2, iw++)
          for ( j=0 ;  j < lfilt ;  j++)
               wksp[ (i + j) & wrap ] += filt[ j] * vec[iw] + rfilt[ j] * vec[ iw + nh] ;
```

* copy wksp in vec and free memory

```
copyvec(wksp, lvec, vec) ;
freevec(wksp) ;
]
```

DWT in dimension d

The DWT of a sequence (a_k), $k \in \{0, 2^j - 1\}^d$, considered as periodic, consists of applying the DWT in dimension 1 to all sections of sequences (a_{k^ℓ}) extracted from $(a_{k_1,\ldots,k^\ell,\ldots,k^d})$ by fixing $d-1$ indices k.

The program below thus builds entirely on the DWT in dimension 1. The key point of the procedure (line 10) is the transposition, which consists of arranging contiguously indices in dimension ℓ, while all indices in other dimensions are placed at offets $2^j, 2^{2j}, \ldots 2^{j(d-1)}$. Indices in dimension ℓ being arranged contiguously in memory ; one easely applies routine `dwt` in dimension ℓ.

```
void dwtd(double * vec, int j1, int d, const double * filt, const double * rfilt,
          int lfilt, int j0, int direction)[
1    int i, j, k, nesvec, dpj, lvec ;
2    if ( j0 >= j1 ) return ;
3    dpj = pow( 2, j1) ;
```

* total length of the memory segment pointed by `vec`

```
4    lvec = pow(2, j1 * d) ;
```

* for deconstruction `vec`

```
5    if ( direction > 0) [
6         nesvec = pow( 2, j1) ;
```

*as many times as level of resolutions from j1 to j0

```
7         for ( k=1 ;  k <= j1 - j0 ;  k++)[
```

* once for each of the d states of transposition of the cube ; on entry, indices of the last dimension of `vec` are contiguous in memory (C order)

```
8             for (j = 0 ;  j < d ;  j++) [
```

* once for each of the $2^{j(d-1)}$ =`lvec/dpj` segments of length 2^j in dimension l, supposed contiguously arranged in memory thanks to transposition ($2^{j(d-1)}$ is the number of ways to fix the $d-1$ indices of the other dimensions). Pass to routine `dwt` the address of `vec` at offset `i * dpj` which is the starting position of segment number `i` ; the length of the sub-segment containing the alphas is given by `nesvec`, that is to say 2^j at current resolution `j` between `j1` et `j0`.

```
9                 for (i=0 ;  i < lvec / dpj ;  i++)
                  dwt( vec + i * dpj, nesvec, filt, rfilt, lfilt, 1) ;
```

* transpose the cube before leaving loop in `j` ; after d transpositions one finds again the starting order, and on exit of the routine `dwtd` the original order is restored

```
10                transpose (vec, lvec, dpj) ;
11            ]
```

* divide nesvec by two before leaving the loop in k, since there are twice less coefficients

alphas arranged first for the next sweep of DWT

```
12            nesvec /=2 ;
13        ]
```

* same thing in reverse order in direction of signal reconstruction except incrementaion of nesvec must take place at first

```
14   ] else if (direction < 0) [
15        nesvec = pow( 2, j0) ;
16        for ( k=1 ;  k <= j1 - j0 ;  k++)[
17             nesvec *=2 ;
18             for (j = 0 ;  j < d ;  j++) [
19                  for (i=0 ;  i < lvec / dpj ;  i++)
                    dwt( vec + i * dpj, nesvec, filt, rfilt, lfilt, -1) ;
20                  transpose (vec, lvec, dpj) ;
21             ]
22        ]
23   ]

]
```

Thresholding in dimension d

Thresholding consists of setting to zero all beta coefficients below the threshold. In the routine below, one supposes that deconstruction is complete down to level $2^{j_0} = 1$; the thresholding instruction (line 29) applies to all the vector except the first position containing the only alpha coefficient left at $j_0 = 0$.

```
void threshold(double * vec, int j1, double t, int d, const double * filt,
               const double * rfilt, int lfilt, int j0)[
24   int k, nesvec, dpj, lvec ;
25   if (j1 < j0) return ;
26   dpj = nesvec = pow(2,j1) ;
27   lvec = pow(2, j1 * d) ;

* complete deconstruction, assuming j0=0
28   dwtd( vec, j1, d, filt, rfilt, lfilt, j0, 1) ;

* thresholding at j0 = 0
29   for (k=1 ;  k < lvec ;  k++) if ( abs(vec[k]) < t ) vec[k] = 0. ;

* complete reconstruction, assuming j0=0
30   dwtd( vec, j1, d, filt, rfilt, lfilt, j0, -1) ;

]
```

Computation of φ values at dyadic rationals

This computation requires a base 2 conversion routine and a second routine to compute φ

from a base 2 expansion. Routine `dbtabulate`, in third position, is used to constitute the array of precomputed values to be used when performing the projection on V_j.

The base 2 conversion routine is standard, but we added a zero padding to a prescribed width (`pad`), for easier use of the subsequent routine.

```
char * int2bin (const int x, int pad)[
  int myx, p ;
```

* static declaration so that the segment is accessible from outside

```
  static char bin[MAXP + 2] ;
```

* `MAXP` is a constant equal to 31 for 32 bit integers

```
  bin[ MAXP + 1] = '\0' ;
  myx = abs(x) ;
  p = MAXP ;
  while ( p) [
      bin[ p] = myx % 2 == 1 ? '1' : '0' ;
      myx /= 2 ;
```

* when `myx` reaches zero, continue the loop for left padding

```
      if ( myx == 0 && MAXP - (p - 1) >= pad) break ;
      p-- ;
  ]
```

* returns the address of the static segment positioned at first significant value

```
  return bin + p ;
]
```

The routine `phi_bin` stores in `phiv` the values obtained by alternate application of `M0` or `M1` (the squared `lM x lM` matrices cited below) in function of `binexp`, the binary expansion of the dyadic number. It is assumed that `binexp` already contains the binary expansion, that `wksp` contains the vector $\varphi(0), \ldots, \varphi(2N-1)$ which initializes the recursion and that `M0` and `M1` have been initialized.

```
void phi_bin( const char *binexp, const double phiv[], int lphiv, double wksp[]) {
  int i ;
```

* initialization of `phiv`

```
  copyvec( phiv, lphiv, wksp) ;
```

* for each letter of the binary word, starting by the end

```
  for ( i = strlen( binexp) - 1 ;  i >= 0 ;  i--){
```

* `mv` is a matrix vector multiplication routine

```
      if ( binexp[ i] == '0') mv( M0, lM, lM, wksp) ;
      else mv( M1, lM, lM, wksp) ;
  }

}
```

The next procedure tabulates all values of φ between 0 and $2N-1$ at a dyadic precision L, thus constituting an array of size `lphivals`$= 2^L(2N-1)$. By convention the value of $\varphi(i2^{-L}), i = 0, \ldots, 2^L(2N-1)$ is placed at offset `i` in the array.

```
void dbtabulate ( int N2 , int L, double phivals[], int lphivals, int diff) {
  int i, j, dpL=(int) pow(2, L), pad ;
  char * devel ;
  const double * finalphi ;
  static double wksp[ 2 * DB_MAXN -1] ;  // max N=4 ;

[ ...]
```

* initialization of `M0, M1` and of other variables accessed by the current procedure

```
  set_daubechies(N2) ;
```

* for φ or its differential, just change the initialization vector in the recursion

```
  switch (diff) {
      case 1 : finalphi = phinp ;  break ;
      case 2 : finalphi = phinpp ;  break ;
      default : finalphi = phin ;  break ;
  }

[...]
```

* by this first call, `pad` is assigned the width of the largest binary word ; complement with zeros on this width for subsequent calls

```
  pad = strlen(int2bin(lphivals, 1)) ;  //probe call for length of '0' padded left binexp
```

* dyadic numbers with same fractional part, shifted from 0 to $2N-1$, are obtained by the same set of M0, M1 multiplication, but are read at position 0 to $2N-1$ of the resulting vector. Looping on the 2^L first indices `i` one has generated the $2^L(2N-1)$ distinc values.

```
  for ( i = 0 ;  i < dpL ;  i++) {
      devel = int2bin( i, pad) ;
```

* `pad - L` positions the dot in the binary expansion ; we take only the decimal part, that is the L last characters ; we see here the advantage of padding

```
      phi_bin( devel + pad - L, finalphi, N2 - 1, wksp) ;  // only L last characters needed
      for ( j = 0 ;  j < N2 - 1 ;  j++) phivals[ i + j * dpL] = wksp[j] ;
  }

}
```

For convenience we recall from the book of Nguyen et Strang (1996) how to set matrices $M0$ and $M1$ and the initializing vectors.

The scaling function φ of an AMR satisfies the equation $\varphi(x) = \sqrt{2}\sum_k c_k\varphi(2x-k)$, which is translated in the frequency domain by $\hat{\varphi}(\omega) = m_0(\omega/2)\hat{\varphi}(\omega/2)$, where $m_0(\omega) = 2^{-1/2}\sum_k c_k exp(-ik\omega)$ is the discrete Fourier transformed of $2^{-1/2}(c_k)_{k\in\mathbb{Z}}$.

In the case of a compactly supported Daubechies wavelet D2N, the expansion of φ contains only $2N$ non zero coefficients c_k ; $\sum_{k=0}^{2N-1} c_k = \sqrt{2}$ and $\sum_{k=0}^{2N-1}(-1)^k c_k = 0$, from the relations $m_0(0) = 1$ and $|m_0(\omega)|^2 + |m_0(\omega+\pi)|^2 = 1$.

The function φ being zero outside of $[0, 2N-1]$, and zero at the point $2N-1$, one has :

$$\begin{pmatrix} \varphi(0) \\ \vdots \\ \varphi(2N-2) \end{pmatrix} = M_0 \begin{pmatrix} \varphi(0) \\ \vdots \\ \varphi(2N-2) \end{pmatrix}$$

where M_0 is the matrix $(2N-1) \times (2N-1)$ given by

$$M_0 = \sqrt{2} \begin{pmatrix} c_0 & & & & & & & \\ c_2 & c_1 & c_0 & & & & & \\ \vdots & \vdots & \vdots & & c_1 & c_0 & & \\ c_{2N-2} & c_{2N-3} & c_{2N-4} & \dots & c_3 & c_2 & c_1 & c_0 \\ & c_{2N-1} & c_{2N-2} & & \vdots & \vdots & \vdots & \vdots \\ & & & & c_{2N-1} & c_{2N-2} & c_{2N-3} & c_{2N-4} \\ & & & & & & c_{2N-1} & c_{2N-2} \end{pmatrix}$$

with $\sum_{2k} c_k = \sum_{2k+1} c_k = \frac{\sqrt{2}}{2}$ from the properties of $m_0(\omega)$. Each column of M_0 sums to 1, so $(1 \quad \dots \quad 1)$ is a left eigenvector, associated to eigenvalue 1.

The fixed point equation has a solution and $\varphi = (\varphi(0) \quad \dots \quad \varphi(2N-2))^t$ is the eigenvector associated to eigenvalue 1 of M_0 (the eigenvector being determined up to a sign flip, we choose the one whose biggest component is positive).

For half integers of the interval $[0, 2N-1]$, one obtains similarly :

$$\begin{pmatrix} \varphi(1/2) \\ \vdots \\ \varphi(2N-2+1/2) \end{pmatrix} = M_1 \begin{pmatrix} \varphi(0) \\ \vdots \\ \varphi(2N-2) \end{pmatrix}$$

where M_1 is the matrix $(2N-1) \times (2N-1)$ given by

$$M_1 = \sqrt{2} \begin{pmatrix} c_1 & c_0 & & & & & & \\ c_3 & c_2 & c_1 & c_0 & & & & \\ \vdots & \vdots & \vdots & & & c_1 & c_0 & \\ c_{2N-1} & c_{2N-2} & c_{2N-3} & c_{2N-4} & \dots & c_3 & c_2 & c_1 \\ & & c_{2N-1} & c_{2N-2} & & \vdots & \vdots & \vdots \\ & & & & & c_{2N-1} & c_{2N-2} & c_{2N-3} \\ & & & & & & & c_{2N-1} \end{pmatrix}$$

The value of φ for any dyadic integer $d = n/2^j$ inside $[0, 2N-1]$ is deduced by successive application of M_0 or M_1 to the starting vector Φ, solution of the fixed point equation.

The binary expansion of a dyadic number d provides the sequence of M_0 and M_1 to apply to Φ together with the position of $\varphi(d)$ in the resulting vector.

For instance, 23/16 is written 1.1101 in base 2 and $\varphi(23/16) = M_1 M_1 M_0 M_1 \Phi(\text{position } 1)$; 7/2 is written 11.1 and $\varphi(7/2) = M_1 \Phi(\text{position 11 in base 2} = \text{position 3})$. $\varphi(7/2)$ is then equal to 0 for a D4 ($\varphi_{D4}(3) = 0$) and a D2 (no position 3 in vector Φ).

Likewise, for the expression of $\varphi'(x)$ and of $\varphi''(x)$, by differentiation of the scaling equation, one obtains : $\varphi'(x) = 2\sqrt{2}\sum_k c_k \varphi'(2x-k)$.

Up to a factor, one obtains the same fixed point equation than above :

$$\begin{pmatrix} \varphi'(0) \\ \vdots \\ \varphi'(2N-2) \end{pmatrix} = 2M_0 \begin{pmatrix} \varphi'(0) \\ \vdots \\ \varphi'(2N-2) \end{pmatrix}$$

The values of φ' at integer points are thus given by the eigenvector Φ' associated to eigenvalue $1/2$ of M_0, which requires a wavelet at least D4 or more (D6 or more to be sure that φ is C^1). For any dyadic rational inside the support of φ, proceed by recurrence as above ; only the starting vector has changed.

For the second derivative, same principle after having derived the scaling equation once more.

Contrast computation

The code excerpt below illustrates the mode of computation of the contrast when the projection is stored in variables `cube` and `margins`.

The linear offset `i` is translated in developed coordinates $(k^1, \dots k^d)$ through `magicv`. Margins appear in the segment `margins` one after another ; we use the dispatching offset `shs` to read the values of all margins at once in cells in front of k

```
real(wp) function wavelet_contrast( kit)
    type(projection_kit), intent( in), target : : kit
    real(wp), pointer : : cube( : ), margins( : )
    integer : : shs (kit % d), magicv (kit % d), i, dpJ
    real(wp) : : temp

! aliases and flat array handling
    cube => kit % cube
    margins => kit % margins
    dpJ = 2**kit % J
    shs = dpJ * (/(i, i=0, kit % d - 1)/)
    magicv = (/ ( dpJ**i, i = kit % d - 1, 0, -1) /)

!IBM* ASSERT (NODEPS) !IBM* INDEPENDENT, REDUCTION(temp) ! sum generally accurate
    temp = 0._wp
    do i = 0 , 2**( kit % J * kit % d) - 1
        temp = temp + ( cube(i) - product ( margins( mod( i / magicv, dpJ) + shs)))**2
    end do
    wavelet_contrast = temp

end function wavelet_contrast
```

The computation is easely distributable on several machines since it is essentially a loop on a big array. In loop 11-13, there is no side effects, no order in the loop traversal, and any by part summation gives the result modulo machine epsilon.

The matrix equivalent of this operation is expressed this way : substract from the cube the tensorial product of the margins, take the sum of squares of the values stored in the cube. Unfortunately with large arrays, this is an operation that easely overflows memory.

Haar projection (Daubechies $D2N, N = 1$)

```
subroutine project_by_histogram ( indata, kit)
```

* `kit` encapsulates some information

```
1    type( projection_kit), target, intent( inout) : : kit
```

* `indata` array d x n containing data

```
2    real( wp), intent( in) : : indata( : , : )
3    real(wp), pointer : : cube( : ), margins( : )
4    integer : : i, k, magicv(kit % d), shs(kit % d), indice(kit % d), dpJ
5    real(wp) : : fact1, factd
```

* pointer to segments cube and margins, initialization

```
6    cube => kit % cube
7    margins => kit % margins
8    cube = 0._wp
9    margins = 0._wp

! binning
10   dpJ =2**kit % J
11   magicv = (/ ( dpJ** i, i = kit % d - 1, 0, -1) /)
12   shs = dpJ* (/(i, i=0, kit % d - 1)/)
13   factd = sqrt( real( 2**( kit % J * kit % d ), wp))
14   fact1 = sqrt( real( 2**( kit % J ), wp))

!IBM* ASSERT (NODEPS)
```

* loop on observations, from 1 to n

```
15   do i = 1, size(indata,2) ! = number of observation
! for a signal between -1/2 and 1/2 cst is placed inside the int function
! or else there is an unintended shift of 1
```

* on entry `indata` is supposed to be whitened and localised between -1/2 and +1/2. translation to 0 and 1 by CST. The integer part of $2^j x$ gives the developed coordinate of x_i in $[0,1]^d$ at resolution j. k is the corresponding linear offset. For margins, use also the dispatching offset `shs`. all $d+1$ memory cells are incremented by 1.

```
16        indice = int( dpJ * (indata( : , i) + CST))
17        k = dot_product( magicv, indice)
18        cube( k) = cube( k) + 1._wp
19        margins( indice + shs) = margins( indice + shs) + 1._wp
20   end do
```

* final scaling $2^{jd/2}\varphi(2^j x - k)$, $x \in \mathbb{R}^d$.

```
! multiplicative factor 1/n 2**Jd/2 and 1/n 2**J/2
  cube = cube * (factd / size( indata, 2))
  margins = margins * (fact1 / size( indata, 2))

end subroutine project_by_histogram
```

Projection on V_j spanned by a Daubechies $D2N$, general case

In the general case, φ values are read in the array `dbtabulate` where they were stored in advance and we take into account the dyadic approximation parameter `L`.

```
subroutine project( indata, kit, dbkit)
  type(projection_kit), target, intent( inout) : : kit
  type(daubechies_kit), intent(in) : : dbkit
  real(wp), intent( in) : : indata( : , : )
  real(wp), dimension( kit % d) : : dpjx
  real(wp), dimension( 0 : kit % d * ( dbkit % deuxn - 1) - 1) : : phival_values
  integer, dimension( 0 : kit % d -1) : : magicv, magic_deuxn
  integer, dimension( kit % d) : : shs2, indice, ifjx, ejx
  integer, dimension( 0 : kit % d * ( dbkit % deuxn - 1) -1) : : conc_indices, longshs
  integer, dimension( dbkit % deuxn - 1 , kit % d ) : : translates
  integer : : d, dxN, i, k, dpL, dpJ, upper_phi, ii

[...] ! aliases and special variables for flat array handling
 dpJ = 2** kit % J
  dpL = 2** dbkit % L
  d = kit % d
  dxN = dbkit % deuxN
  magic_deuxN = (/ ( (dxN -1)** i, i = d -1, 0, -1) /)
  magicv = (/ ( dpJ**i, i = d - 1, 0, -1) /)
  shs2 = ( dxN - 1) * (/ (i, i=0, d - 1) /)
  longshs = reshape( spread( dpJ * (/ (i, i=0, d - 1) /), ncopies=dxN -1, dim=1), (/ d * (dxN -1) /))
  translates = spread( (/ ( i, i=0, dxN - 2) /), dim=2, ncopies= d)
  upper_phi = ubound( dbkit % phivals, 1)
  kit % cube = 0._wp
  kit % margins = 0._wp

! loop on all observations

!IBM* ASSERT (NODEPS)
```

* loop on observations

```
  do i = 1, size( indata, 2)
```

* signal on entry is between -1/2 + 1/2 ; translation to 0, 1 through CST ;
* multiplication by 2^j performs the dilation

```
      dpjx = dpJ * (indata( : ,i) + CST)
```

* keep the integer part yielding developed coordinate

```
26        ejx = int(dpjx)
```

* find the offset (in array containing precomputed dyadic values) of fractional part of $2^j x$; the integer part gives a shift to right by a multiple of 2^L. Since there are at most $2N-1$ possible shift of 2^L to the right (after which the index goes out of the support) we take them all diretly at line 28

```
27        ifjx = int(dpL * (dpJx - ejx) ) ! index of fractional part of 2**jx
```

* add all possible translations of 2^L to `ifjx` and flatten the array so that the $2N-1$ first values concern dimension 1, next $2N-1$ dimension 2, etc...

```
! concatenated list of all phival indices to retrieve for each dimensions
28        conc_indices = reshape( spread( ifjx, dim=1, ncopies = dxN -1) + dpL * translates, (/ d * ( dxN - 1) /))
```

* since we translated regardless of bounds in array `upper_phi`, we now truncate indices at the maximum index. Since `dbkit % phivals(upper_phi)` is equal to zero, the assignation is a way to do that. At line 30, `phivals_values` contains values read at all positions stored in `conc_indices`, in the right order (shortcut in fortran 90). At this stage `phivals_values` contains all values needed for the contribution of observation `i` to the projection. It remains to determine exactly to which cell coordinates the values must be added

```
! for out of support, last index always contains zero.
29        where ( conc_indices > upper_phi) conc_indices = upper_phi
30        phival_values = dbkit % phivals(conc_indices)
```

* `phivals_values` being initialized, we reuse the space `conc_indices` to compute the different translates of φ concerned by `dpjx` : $2^j x - k \in [0, 2N-1] \Leftrightarrow 2^j x - (2N-1) \leq k \leq 2^j x$. We then build a vector of length $d(2N-1)$ containing $2N-1$ positions k in first dimension, followed by $2N-1$ positions k in second dimension, etc...

```
! concatenated list of all unidimensional k ! conc_indices reused
31        conc_indices = reshape(spread( ejx , dim=1, ncopies = dxN -1) - translates, (/ d * ( dxN - 1) /))
```

* in some rare circumstances despite proper relocation [0,1], rounding effects produce out of bounds

```
! needed in some cases in single precision despite proper relocation
32        where (conc_indices >= dpJ) conc_indices = dpJ -1
```

* if some indices were negative, wrap around

```
! circular shift at borders with high indices ;  ! cannot go out of bound by low indices !
33        forall (ii = 0 : d * ( dxN - 1) -1 , conc_indices(ii) < 0)
34        conc_indices (ii) = mod(abs(dpJ + conc_indices (ii)), dpJ)
35        end forall
```

* value for translated of zero, `phivals[ifjx]`, for translated of -1, `phivals[ifjx + 1 x dpL]`, for translated of -2, `phivals[ifjx + 2 x dpL]` etc... in other words everything is ready in

phival_values so that there is only an addition to execute. Distribution in the dimensions thanks offset longshs taking into account that the $2N-1$ first values concern dimension 1, etc...

```
! one-shot add for the margins ! sum is rarely inaccurate
36        kit % margins(conc_indices + longshs) = kit % margins(conc_indices + longshs) + phival_values
```

* for the cube, compute in linear offset the region, of volume $(2N-1)^d$, affected by contribution of observation i. indice contains successively all combinations from $(0,\ldots,0)$ to $(2N-1,\ldots 2N-1)$ taken from conc_indices, that are next translated in linear offset; the value to add is the product of the margins in front of offset k. It's the same basic information that the one in the margins line 36 above, but multiplied and spread in the cube.

```
! the affected multi dimensional coordinates
! is the set of all tensorial combinations of uni dimensional coordinates
! summing is inaccurate for about 10% of observation

! !IBM* ASSERT (NODEPS)
37        do ii = 0, (dxN - 1) ** d -1
38        indice = mod( ii / magic_deuxN, dxN - 1)
39        k = dot_product( magicv, conc_indices ( indice + shs2))
40        kit % cube( k) = kit % cube( k) + product( phival_values(indice + shs2) )
41        enddo
42   end do

! final scaling
43   kit % cube = kit % cube * ( sqrt( real( 2**( kit % J * kit % d ), wp)) / size( indata, 2))
44   kit % margins = kit % margins * ( sqrt( real( 2**( kit % J ), wp)) / size( indata, 2))
45  end subroutine project
```

6.6 Optimization on the Stiefel manifold

A standard method to resolve ICA consist of first prewhitening the observed signal X, and then minimize the contrast $C(WX)$ as a function of W, under the constraint ${}^tWW = I_d$, corresponding to the fact that W belongs to the Stiefel manifold $S(d,d) = O(d)$, the group of orthogonal matrices. In the ICA context, $SO(d) \subset O(d)$ is a sufficient restriction, which is equivalent to ignoring reflections, since an ICA solution is defined up to a permutation of axes.

The routines below implement the Plumbley algorithm recalled p. 104.

```
real(wp) function son_contrast(t, W, paramdata, pjkit, dbkit, H)
```

* this routine returns the contrast in $W' = \exp(-tH)W$, where t is the amplitude of the move in direction H in $so(d)$.

```
1    real(wp), intent(in) : : t, W( : , : ), H( : , : ), paramdata( : , : )
2    type(projection_kit), intent(inout) : : pjkit
```

```
3    type(daubechies_kit), intent(in) : : dbkit
4    real(wp) : : geo (size(W,1), size(W,2))
```

* padm computes the exponential of a matrix see Sidje (1988)

```
5    geo = padm(real(t,8), real( H ,8))
6    geo = matmul( geo, W)
7    call project( matmul( geo, paramdata), pjkit, dbkit)
8    son_contrast = wavelet_contrast( pjkit)

end function son_contrast

subroutine stiefelmin(W, iter, fret, paramdata, pjkit, dbkit, N, A)
9    use libica, only : amari
10   use differential, only : empgrad
11   integer, intent(out) : : iter
12   real(wp), intent(out) : : fret
13   real(wp), dimension( : , : ), intent(inout) : : W, N, A
14   real(wp), intent(in) : : paramdata( : , : )
15   type(projection_kit), intent(inout) : : pjkit
16   type(daubechies_kit), intent(in) : : dbkit
17   integer, parameter : : ITMAX = 5
18   real(wp), parameter : : STPMX = 100.0_wp, gtol = 1.E-6_wp, ftol = 1.E-5_wp

! optimize minimization
19   integer : : its
20   real(wp) : : fp, stpmax, ax, bx, cx, fa , fb, fc, tk, normg
21   real(wp), dimension( size( W, 1), size( W, 2)) : : g, h, wnew

! initial contrast
22   call project( matmul( W, paramdata), pjkit, dbkit)
23   fret = wavelet_contrast( pjkit)
24   fp = 0._wp
25   normg = 0._wp
26   tk=0._wp
27   do its = 0, ITMAX ! Main loop
28        iter = its

[...]
```

* computes the empirical gradient with forward and backward differences

```
29        g = empgrad(W, paramdata, pjkit, dbkit, forwardonly=.false.)
```

gradient for the canonical metric in Lie algebra so(n), $\nabla_B J = \nabla J\ {}^tW - W\ {}^t\nabla J$

```
30        g = matmul(g, transpose( W)) - matmul( W, transpose( g))
```

* if the gradient norm is too small, exit

```
31        normg = canonical_metric( g, g)
32        if (normg < gtol .or. abs( fp - fret) < ftol) then
```

```
33        [...]
34        return
35        endif
36        fp = fret
```

* direction of descent

```
37        H = -g / normg
```

* standard routine to bracket a minimum see Numerical recipes (1986). `ax` et `bx` are two initialization parameters such that `fa` >`fb`. We seek a `cx`> `bx` such that `fc` >`fb`. Here function $f(t)$ is `son_contrast` that maps back in $SO(d)$, that is to say $f(t) = exp(-tH)W$

```
! brackets a minimum in direction H with initialization from ax to bx
38        ax = 0._wp
39        bx = .001_wp
40        call mnbrak(ax, bx, cx, fa, fb, fc, son_contrast, W, paramdata, pjkit, dbkit, H)
```

* with a braket of a minimum (`ax, bx, cx` initialized above) call another standard routine to find that minimum, by bisection see Numerical recipes (1986).

```
! find the minimum using golden section with ax, bx, cx returned by mnbrak
41        fret = golden(ax, bx, cx, son_contrast, W, paramdata, pjkit, dbkit, H, tol=1.e-5_wp, xmin=tk)
```

* the minimum `tk` of $f(t) = exp(-tH)W$, being fixed compute the new matrix W at this point and iteration

```
! compute the new W through premultiplication by the exponential of
! what was found to be the best move in so(n)
42        Wnew = matmul(padm(real(tk,8), real( H ,8)), W)

! update and loop
43        W = Wnew
44   end do
45   write (*, 100) iter, fret, amari(matmul(W,N),A), tk, normg , abs(fret-fp)

[...]

end subroutine stiefelmin
```