Mesures de dépendance pour la séparation aveugle de sources. Application aux mélanges post non linéaires

Sophie Achard

THÈSE

présentée par

SOPHIE ACHARD

pour obtenir le titre de docteur de

L'UNIVERSITÉ JOSEPH FOURIER (GRENOBLE I)

dans la spécialité :

MATHÉMATIQUES APPLIQUÉES

MESURES DE DÉPENDANCE POUR LA SÉPARATION AVEUGLE DE SOURCES

APPLICATION AUX MÉLANGES POST NON LINÉAIRES

Thèse préparée au LABORATOIRE DE MODÉLISATION ET CALCUL et
au LABORATOIRE DES IMAGES ET DES SIGNAUX

soutenue le 2 décembre 2003 devant le jury composé de:

Serge DÉGERINE	Président
Jean-François CARDOSO	Rapporteurs
Elisabeth GASSIAT	
Fabrice GAMBOA	Examinateur
Christian JUTTEN	Directeurs de thèse
Dinh Tuan PHAM	

A mon pépé Némos
A mon pépé Maurice

Remerciements

Je tiens à remercier mes deux directeurs de thèse : Christian Jutten, avec lequel j'ai eu des discussions très variées mélant une approche statistique et une approche traitement du signal et Dinh-Tuan Pham pour l'aide qu'il m'a apportée tout au long de cette thèse.

Je remercie Jean-François Cardoso et Elisabeth Gassiat qui ont bien voulu être rapporteurs de cette thèse.

J'adresse aussi beaucoup de remerciements à Fabrice Gamboa et Serge Dégerine pour leur participation en tant qu'examinateurs dans le jury de cette thèse.

J'ai beaucoup apprécié le soutien de tous les membres du Laboratoire de Modélisation et Calcul. En particulier, je remercie tous les doctorants pour l'ambiance amicale dans laquelle j'ai travaillé durant ces années de thèse. Je tiens à saluer Guillaume Allègre, François Gannaz, Valérie Pham-Trong pour toute l'aide autant morale que matérielle qu'ils m'ont apportée. De longues et enthousiastes discussions avec Gersende Fort et Pierre Lafaye de Micheaux m'ont apporté des motivations supplémentaires.

Mes plus vifs remerciements vont à Sylvie Champier et Jacques Berruyer, qui ont su me faire partager leur passion pour la recherche et l'enseignement.

Ma reconnaissance va en particulier à mes parents et toute ma famille qui m'ont beaucoup soutenue.

Enfin, je tiens à exprimer ma plus grande gratitude pour Jocelyn Étienne, qui a su susciter de nombreuses discussions scientifiques très enrichissantes et qui a eu toute la patience de lire le manuscrit.

Table des matières

Chapitre 1

Introduction

La séparation aveugle de sources au fil des années

Perspectives historiques

Le problème de séparation aveugle de sources a tout d'abord été développé par les travaux de Ans, Hérault et Jutten dans les années 80. Dans [40], Jutten et Taleb décrivent le problème biologique qui a initié les travaux sur la séparation aveugle de sources. Celui-ci consistait à étudier les réponses musculaires émises à l'issue de différentes sortes d'excitation.

Puis les travaux de Comon, en 1994, *"Independent component analysis, A new concept ?"* [20], ont permis de formaliser le lien entre la méthode d'analyse en composantes indépendantes (ICA) et le problème de séparation aveugle de source (BSS) dans le cadre d'un mélange linéaire. En effet, grâce au théorème de Darmois, [27], Comon a montré que l'ICA est équivalente à la BSS dans le cadre de mélanges linéaires inversibles et non bruités à condition qu'il y ait au plus une source gaussienne.

Depuis les années 90, à partir de ces travaux, ont été développées de nombreuses méthodes de résolution du problème de séparation aveugle de sources dans le cadre d'un mélange linéaire qui ont mené à de nombreux algorithmes. Pour un exposé plus précis, on peut consulter [13] ou [37].

Ensuite, en 1999, Taleb et Jutten [69] ont introduit un nouveau type de mélange, les mélanges post non linéaires (PNL).

Enfin, plus récemment, Cardoso [15], propose d'envisager *"the three easy routes to independent component analysis"* pour des mélanges linéaires. En effet, partant de la constatation que le problème d'analyse en composantes indépendantes n'est pas équivalent au problème de séparation aveugle de sources dans le cas de signaux gaussiens indépendants et identiquement distribués (i.i.d.), il propose d'envisager trois cas différents de telle sorte que les deux problèmes d'ICA et de BSS soient équivalents. Ces cas diffèrent par les hypothèses sur le modèle des sources, on peut envisager des

sources non gaussiennes i.i.d, ou bien des sources gaussiennes non stationnaires, ou enfin, des sources gaussiennes corrélées temporellement. Dans [15], Cardoso décrit les différentes méthodes envisagées, en fonction des hypothèses sur le modèle des sources, afin de résoudre le problème de séparation aveugle de sources grâce à l'analyse en composantes indépendantes.

La séparation aveugle de sources sans l'analyse en composantes indépendantes

Bien sûr, l'analyse en composantes indépendantes n'est pas la seule approche possible pour résoudre le problème de séparation aveugle de sources. Dans [20, p. 293], Comon propose de définir une fonction de contraste dans le cadre de mélanges linéaires. Ces fonctions permettent alors la résolution du problème de séparation aveugle de sources en recherchant leur maximum, sans pour autant être des mesures de dépendance. Dans [20] Comon propose une fonction de contraste construite à partir de l'information mutuelle avec une estimation d'Edgeworth de la densité et exprimée à l'aide des cumulants. Il montre que la fonction de contraste définie ainsi est maximale en un point solution du problème de séparation aveugle de sources sous l'hypothèse que les sources sont indépendantes. On peut aussi envisager l'utilisation d'autres propriétés caractérisant l'indépendance dans le cadre de mélange linéaire, voir annexe A.

Puis, dans [45], Krob propose une fonction de contraste pour résoudre le problème de séparation aveugle de sources dans le cadre de mélanges polynomiaux sous l'hypothèse que les sources sont indépendantes. On remarque que ces fonctions de contraste sont adaptées à des conditions très particulières sur le mélange.

Plus récemment, de nombreux auteurs ont proposé d'envisager d'autres hypothèses. On peut par exemple se restreindre à des sources discrètes [22]. Dans ce cadre, Comon *et. al.* proposent une nouvelle fonction de contraste adaptée à ce problème sans l'hypothèse d'indépendance des sources. Ou bien, certains auteurs ont choisi d'exploiter des propriétés de parcimonie du signal [18], [39], c'est en particulier le cas des signaux de paroles. Ou encore, on peut exploiter des propriétés de non stationnarité des sources [15]. On peut sûrement envisager beaucoup d'autres hypothèses ...

Applications

Depuis le début des recherches sur la résolution du problème de séparation aveugle de sources, celui-ci a été lié à des problèmes concrets. De la biologie [78, 74] à l'astronomie [53, 16] en passant par la chimie [52], les algorithmes de séparation aveugle de sources sont utilisés pour la résolution de nombreux problèmes réels.

Le problème de la "cocktail party" est une illustration de situation dans laquelle la séparation aveugle de sources permet d'obtenir des résultats souvent très convaincants.

Dans une pièce où sont présentes 3 personnes qui parlent indépendamment les unes des autres, nous plaçons aussi 3 magnétophones. Il est naturel de penser que sur les bandes des 3 cassettes, on va pouvoir entendre les voix des 3 personnes superposées les unes sur les autres. L'objectif de la séparation aveugle est alors de retrouver les paroles des 3 personnes à partir seulement des 3 enregistrements des magnétophones (voir figure 1.1). Ce problème est très étudié en traitement du signal de parole.

FIG. 1.1 – *Le problème de la séparation de sources tel que défini par Gotlib dans les Dingodossiers [0]*

Un autre problème très étudié actuellement, est l'extraction non-invasive de l'électrocardiogramme du foetus à partir d'électrodes placées sur l'abdomen de la mère [48, 47]. En effet, sur le signal enregistré, on voit apparaître le signal provenant du coeur de la mère et celui provenant du coeur du bébé avec des bruits et d'autres signaux. On peut alors supposer que ces deux signaux sont indépendants et mélangés, ce qui permet d'envisager de le résoudre à partir de techniques de séparation aveugle de sources.

Bien d'autres applications sont étudiées actuellement.

Le problème de séparation aveugle de sources

Dans un contexte tout à fait général, le problème de séparation aveugle de sources se formule de la manière suivante,

Nous disposons de K processus aléatoires appelés *observations*, et notés $X_1(.), \ldots,$ $X_K(.)$ qui proviennent d'un mélange de L processus aléatoires $S_1(.), \ldots, S_L(.)$ appelés *sources*. On note alors $\mathcal{F}$ la transformation liant les sources et les observations :

$$\mathbf{X}(.) = \mathcal{F}\mathbf{S}(.), \tag{1.1}$$

où $\mathbf{X}(.) = (X_1(.), \ldots, X_K(.))$ et $\mathbf{S}(.) = (S_1(.), \ldots, S_L(.))$.

L'originalité du problème de séparation aveugle de sources provient du peu d'hypothèses faites sur le modèle (1.1). En effet, si on suppose la fonction de mélange $\mathcal{F}$ connue et les observations connues, le problème de reconstruction des sources se résout en cherchant l'inverse de $\mathcal{F}$. Ici, nous nous plaçons dans un contexte aveugle, i.e. la fonction de mélange $\mathcal{F}$ est inconnue. Notre problème consiste donc à restituer les sources, en connaissant seulement les observations. Les méthodes d'analyse en composantes principales ou de projection pursuit sont un exemple de résolution de ce genre de problèmes. Elles se différencient de l'approche d'analyse en composantes indépendantes par les hypothèses qui sont faites sur les sources et sur le mélange. Cependant, il est clair que sans hypothèse supplémentaire, la résolution du problème de séparation aveugle de sources est tout à fait impossible. Il est alors apparu nécessaire de faire des hypothèses sur le mélange et sur les sources. Nous allons décrire maintenant quelques situations dans lesquelles, le problème de séparation aveugle de sources peut être résolu.

L'objectif de la séparation aveugle de sources est alors de rechercher des processus aléatoires $Y_1(.), \ldots, Y_L(.)$, appelés *sources reconstituées*, tels que ceux-ci représentent une estimation des sources.

On note alors la structure de séparation,

$$\mathbf{Y}(.) = \mathcal{G}\mathbf{X}(.), \tag{1.2}$$

où $\mathbf{Y}(.) = (Y_1(.), \ldots, Y_L(.))^T$.

Il apparaît alors ici deux problèmes qui peuvent être différents en fonction des hypothèses que l'on fait sur le modèle. En effet, on peut considérer le problème consistant en l'estimation de la structure de séparation, que l'on appelle le problème d'identification du mélange. Ou bien, on peut considérer le problème appelé restitution des sources [32] consistant à proposer une estimation des sources sans pour autant proposer une estimation de la structure de séparation. Bien sûr, dans le cas d'un mélange

inversible, ces deux problèmes coincïdent.

Nous n'envisageons dans cette thèse que la résolution de la séparation aveugle de sources à l'aide de l'analyse en composantes indépendantes. C'est pourquoi, nous détaillons ici dans quelle mesure l'analyse en composantes indépendantes peut permettre la résolution du problème de séparation aveugle de sources. Enfin, nous expliciterons plus précisément le plan de cette thèse.

La Séparation aveugle de sources à l'aide de l'analyse en composantes indépendantes

Nous avons signalé que la résolution du problème de séparation aveugle de sources nécessite des hypothèses supplémentaires. La première approche proposée par Ans, Hérault et Jutten a été d'exploiter l'hypothèse faite sur l'indépendance des sources, $S_1(.), \ldots, S_L(.)$.

C'est pour cela que ces derniers ont rapproché le problème de séparation aveugle de la méthode qu'ils ont appelée l'analyse en composantes indépendantes. Par conséquent, dans cette configuration, l'objectif de la séparation aveugle est de déterminer les processus aléatoires $Y_1(.), \ldots, Y_L(.)$ de telle sorte que ceux-ci soient indépendants. Mais pour que le problème de séparation aveugle de sources soit effectivement résolu, les processus aléatoires $Y_1(.), \ldots, Y_L(.)$ doivent aussi être une estimation des sources. Or ces deux objectifs peuvent, dans certaines conditions, ne pas être résolus en même temps.

En effet, citons l'exemple classique en statistique permettant de construire des variables aléatoires gaussiennes,

Soient R une variable aléatoire uniforme dans [0,1], et Θ une variable uniforme dans $[0, 2\pi]$ de telle sorte qu'elles soient indépendantes.
Alors, les variables $X = R\cos\Theta$ et $Y = R\sin\Theta$ sont indépendantes.

D'autre part, Darmois [26] a montré qu'à partir de variables aléatoires dépendantes, il est toujours possible de construire une transformation non triviale de telle sorte que les variables aléatoires obtenues à la sortie du mélange soient indépendantes.

La seule hypothèse d'indépendance des sources n'est donc pas suffisante à l'unicité de la solution du problème de séparation aveugle de sources (aux indéterminations près de permutations et de facteurs d'échelle). Ceci a conduit alors de nombreux auteurs à se poser la question de savoir quand l'analyse en composantes indépendantes permet effectivement de résoudre le problème de séparation aveugle. Une première réponse a

été apportée par Comon en 1991 [19] en proposant de faire une hypothèse sur la fonction de mélange. Dans [75], Yang *et. al.* exposent un cadre plus générale d'étude de la possibilité de résoudre le problème de séparation aveugle dans le cas de mélanges non linéaires. Nous allons à présent détailler quelques configurations possibles où l'analyse en composantes indépendantes permet effectivement de résoudre le problème de séparation aveugle de sources. Dans cette thèse, nous nous placerons toujours dans le cas de mélanges non bruités.

$\mathcal{F}$ linéaire instantané :

Les premiers travaux sur la séparation aveugle de sources se placent dans le contexte d'un mélange linéaire instantané et sans bruit :
Soient $X_1, \ldots, X_K$ des variables aléatoires, appelées observations et $S_1, \ldots, S_L$ des variables aléatoires indépendantes appelées sources telles que

$$\mathbf{X} = \mathbf{AS}$$

où $\mathbf{X} = (X_1, \ldots, X_K)^T$ et $\mathbf{S} = (S_1, \ldots, S_L)^T$
On appelle généralement la matrice $\mathbf{A}$ matrice de mélange.

Dans le cas où $K \leq L$, on dit que le mélange est sous déterminé, et dans ce cas, il est impossible de résoudre le problème sans hypothèses supplémentaires. Pour un exposé plus détaillé de l'identifiabilité des modèles, on peut consulter [72]. Dans [21], Comon propose une solution en se restreignant à des sources discrètes.

Dans le cas où $L \leq K$, Comon [20, p.294] a démontré à l'aide d'un théorème de Darmois [27] que l'analyse en composantes indépendantes est alors équivalente à la résolution du problème de séparation aveugle de sources à condition de supposer qu'il y ait au plus une source gaussienne. Les sources pourront alors être restituées à un facteur d'échelle près et à une permutation près. Dans ce contexte, de nombreuses approches ont été envisagées, à partir des statistiques d'ordre supérieur [33], [14], à partir de la néguentropie [37] et bien d'autres. De plus, de nombreux algorithmes ont été développés, JADE [14], FastICA [37], SOBI [9], InfoMax [8], ...

Mais ces mélanges ne reflètent pas exactement certains cas réels, il a alors été nécessaire d'envisager d'autres types de mélanges.

$\mathcal{F}$ linéaire convolutive :

Dans ce cas particulier, en se plaçant dans le domaine des fréquences, on peut se ramener à la résolution d'un problème de séparation aveugle de sources dans

le cadre d'un mélange linéaire. On utilise alors les méthodes précédentes. Mais on peut aussi considérer des modèles de mélanges FIR. [76, 28, 71]

$\mathcal{F}$ post non linéaire :

Plus récemment, Taleb et Jutten [67, 69] se sont intéressés à la possibilité de considérer un mélange non linéaire. Cependant, comme nous l'avons évoqué précédemment, il est tout à fait impossible de résoudre le problème de séparation aveugle de sources dans le cadre d'un mélange non linéaire quelconque en utilisant l'analyse en composantes indépendantes. Observant alors que les mélanges réels peuvent être souvent vus comme un mélange linéaire suivi de saturations provoquées par le fonctionnement des capteurs, ils ont proposé d'étudier des mélanges appelés post non linéaires. Nous reviendrons sur ce mélange dans le chapitre 4.

$\mathcal{F}$ post non linéaire convolutive :

Enfin, Babaie-Zadeh a envisagé une nouvelle catégorie de mélanges, les mélanges post non linéaires convolutifs [6]

Grâce à une hypothèse supplémentaire sur la structure du mélange, le problème de séparation aveugle de sources sous la seule hypothèse d'indépendance des sources et la présence d'au plus une source gaussienne peut donc être résolu, entre autres approches, en utilisant l'analyse en composantes indépendantes.

Lignes directrices

Dans cette thèse, nous aborderons le problème de séparation aveugle de sources en utilisant la méthode d'analyse en composantes indépendantes.

Cette méthode impose de définir des mesures permettant de déterminer "à quel point" des variables sont dépendantes. Nous présentons donc deux mesures de dépendance, une bien connue, l'information mutuelle et l'autre, que nous proposons, la mesure de dépendance quadratique (**chapitre 2**). Ces mesures de dépendance ont la propriété d'être toujours positives et de s'annuler lorsque les variables sont indépendantes. Mais elles diffèrent aussi par d'autres propriétés, en particulier concernant les estimations.

Afin de réaliser l'analyse en composantes indépendantes, il est naturel de vouloir chercher leur minimum. Nous envisageons alors d'utiliser une méthode de descente de gradient (**chapitre 3**).
Néanmoins, le calcul de ces mesures de dépendance ne peut-être réalisé sans procéder

à une estimation. Une première possibilité est de calculer le gradient de la mesure de dépendance formellement à partir de son expression théorique, puis de procèder à l'estimation du gradient. La deuxième possibilité consiste à proposer une estimation de la mesure de dépendance puis à en calculer le gradient.

En particularisant notre propos au problème de séparation aveugle de sources dans les mélanges post non linéaires, nous proposons différents algorithmes (**chapitre 4**) (nous nous limitons à l'étude de mélanges instantanés non bruités). Trois approches sont détaillées, une approche non paramétrique, une deuxième approche non paramétrique mais utilisant une propriété des mélanges post non linéaires et des mesures de dépendances, et enfin une approche paramétrique.

Dans ces algorithmes, les mesures de dépendance n'interviennent qu'à travers leurs estimations. En ce qui concerne la mesure de dépendance quadratique, nous détaillons une étude asymptotique de son estimateur (**chapitre 5**). Comme la mesure de dépendance quadratique dépend du choix d'un noyau et d'une taille de fenêtre, cette étude permet de proposer une procédure automatique du choix de la taille de fenêtre en fonction du noyau. Nous étudions les biais des estimateurs obtenus à partir de l'information mutuelle. Ceci nous amène à comparer cette approche par rapport à celle envisagée par Taleb et Jutten [69].

Enfin, le **chapitre 6** illustre le comportement de ces mesures de dépendances dans différentes situations, mélange linéaire, mélange post non linéaire ...

Chapitre 2

Mesures de dépendance

2.1 Introduction

Dans le cadre de l'analyse en composantes indépendantes, il est nécessaire de savoir si des variables aléatoires sont indépendantes.

C'est pourquoi nous envisageons dans ce chapitre d'étudier des mesures de dépendance. Dans la suite, nous ferons la distinction entre la notion de mesure de dépendance ou de critère de séparation. Dans la section 2.2, nous donnons quelques exemples de critère de séparation déjà utilisés en séparation aveugle de sources.

Dans cette thèse, nous définissons les mesures de dépendance comme donnant une indication sur la dépendance des variables aléatoires en fournissant une valeur réelle. Celles-ci sont donc définies de manière tout à fait générale, indépendamment de la résolution du problème de séparation de sources. Ce que nous appellerons un critère de séparation sera la fonction objectif à optimiser du problème de séparation aveugle de sources.

Nous verrons que pour certains mélanges de sources, une mesure de dépendance fournira directement un critère de séparation (section 4.2). C'est le cas en particulier de l'information mutuelle dont nous rappelons les propriétés dans la section 2.3, et de la mesure de dépendance quadratique que nous définissons dans la section 2.4. Néanmoins, signalons qu'il est aussi possible de définir des critères de séparation qui ne sont pas des mesures de dépendance. Nous illustrerons ces propos dans le paragraphe 2.3.3.

2.2 Exemples de critères de séparation

Depuis les années 90, beaucoup d'auteurs se sont attachés à définir des critères de séparation qui se basent sur des mesures de dépendance. En voici quelques uns.

critères de séparation basés sur la corrélation non linéaire :

Ces critères sont basés sur l'utilisation du maximum de la corrélation non linéaire, appelée $\mathcal{F} - correlation$ dans [7], définie par,

$$\rho_{\mathcal{F}} = \max_{f_1, f_2 \in \mathcal{F}} corr(f_1(X_1), f_2(X_2)) = \max_{f_1, f_2 \in \mathcal{F}} \frac{cov(f_1(X_1), f_2(X_2))}{var(f_1(X_1))^{1/2} var(f_2(X_2))^{1/2}}$$

où X_1 et X_2 sont des variables aléatoires réelles.

On remarque alors que pour des choix particuliers de $\mathcal{F}$, $\rho_{\mathcal{F}} = 0$ est équivalent à l'indépendance de X_1 et X_2 [62, p. 265].
Dans [7], Bach et Jordan utilisent le maximum de corrélation comme critère de séparation pour deux variables aléatoires dans le cadre d'un mélange linéaire de sources. Nous verrons aussi dans la section 4.3, que la notion de maximum de corrélation a été utilisée dans le cadre de mélanges post non linéaires, [79], mais en considérant une autre propriété du maximum de corrélation pour deux variables.

critères de séparation basés sur la divergence de Kullback-Leibler :

Ces critères sont issus d'une mesure, l'information mutuelle, permettant de mesurer par une quantité scalaire l'écart entre deux distributions, la densité conjointe et le produit des densités marginales. Dans le cas de mélanges linéaires, Pham [57] étudie un tel critère en montrant les différences avec la méthode de "projection pursuit". Dans [58], on propose un schéma plus général de critères de séparation utilisant la concavité des fonctions utilisées.
Dans les sections 2.3, 3.2.1 et 3.2.3, à partir des techniques utilisées dans [57], nous expliciterons comment à partir de la mesure d'information mutuelle on peut en déduire une méthode de séparation aveugle de sources.
D'autres critères peuvent être vus comme dérivant de l'information mutuelle. Nous détaillerons dans la suite, les liens existant entre ces critères et l'information mutuelle. Citons en particulier,

- critères basés sur les statistiques d'ordre supérieur (HOS):

  Ces critères utilisent les propriétés des cumulants vis à vis de l'indépendance des variables aléatoires, [20], ou bien leurs propriétés dans l'approximation des densités. On peut citer, les méthodes utilisant la maximisation d'une fonction de vraisemblance [14], le principe InfoMax [8]. Ces critères sont utilisés pour la séparation de sources dans le cas de mélanges linéaires car facilement implémentables. Par contre, dans le cas de mélanges non lineaires, ils sont plus difficilement exploitables.

- critères basés sur le caractère gaussien des variables:

  On se place ici aussi dans le cadre de mélanges linéaires. D'après le théorème de la limite centrale, on sait qu'une somme de variables aléatoires indépendantes converge en loi vers une loi normale. Ces critères sont donc basés sur le principe que moins les variables sont gaussiennes, plus elles sont indépendantes. Nous reviendrons sur cette approche intuitive dans le paragraphe 2.3.3. On peut citer les critères basés sur le kurtosis ou la néguentropie [37]. L'usage de ceux-ci présente des avantages dans les méthodes de calcul, mais ils présentent aussi de nombreux inconvénients. Le calcul du kurtosis est fortement dégradé par rapport aux données présentant des artefacts (voir [37]). D'autre part, ces critères ne sont des mesures de dépendance que sous la contrainte de blanchiment préalable.

critères basés sur des mesures quadratiques :

Ces critères prennent aussi en compte la comparaison de la densité conjointe avec le produit des densités marginales. Dans [66], Tjøstheim fait l'inventaire de critères utilisés dans le cadre de tests d'indépendance. En particulier, on peut d'une certaine façon les relier aux critères de séparation basés sur des méthodes de noyaux. Nous introduisons alors dans la section 2.4, une nouvelle mesure de dépendance, appelée dépendance quadratique. Nous verrons comment cette mesure peut être rapprochée des critères basés sur la comparaison de la densité conjointe et du produit des marginales introduits par Rosenblatt, [63], ou bien des critères basés sur la comparaison de la fonction caractéristique conjointe et du produit des fonctions caractéristiques marginales, Feuerverger [31], Kankainen [44].

2.3 Information mutuelle

Nous rappelons ici, la définition de la mesure de dépendance appelée information mutuelle à partir de la divergence de Kullback-Leibler, ainsi que les propriétés qui nous seront utiles pour la suite.

2.3.1 Définitions

Afin de montrer la similitude de l'approche avec la mesure de dépendance quadratique, nous rappelons ici les propriétés caractéristiques de l'information mutuelle. Nous verrons comment d'autres méthodes comme les méthodes utilisant le principe du maximum de vraisemblance ou le principe d'InfoMax peuvent être déduites de ce critère. Puis, dans le chapitre 4, nous introduirons de nouvelles méthodes de séparation aveugle de sources dans le cadre de mélanges post non linéaires utilisant l'information

mutuelle (c.f. 4.1.1, définition des mélanges post non linéaires). Rappelons tout d'abord la propriété suivante,

Soient $T_1, T_2, \ldots, T_K$, K variables aléatoires réelles telles que :
le vecteur aléatoire $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ est de loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^K$. On note $p_{T_1,T_2,\ldots,T_K}$ la densité de celui-ci.
(Par conséquent, pour tout i, $1 \leq i \leq K$, la variable aléatoire T_i est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$. On note p_{T_i} sa densité.)
On a alors la propriété suivante,

$$T_1, T_2, \ldots, T_K \text{ mutuellement indépendantes si et seulement si } p_{T_1,T_2,\ldots,T_K} = \prod_{i=1}^{K} p_{T_i}$$

Remarque 2.3.1 *Nous précisons "mutuellement indépendantes" car il est bien connu que l'indépendance par paire et l'indépendance mutuelle ou dans l'ensemble sont deux notions différentes. Cependant, dans le cadre de mélanges linéaires Comon a démontré [20] que l'indépendance par paire est équivalente à l'indépendance mutuelle. Dans le cadre d'un mélange linéaire, il est alors très utile de ne vérifier que l'indépendance par paire des variables aléatoires.*

Une mesure naturelle de dépendance peut alors se définir par une comparaison de "l'écart" entre la densité conjointe et le produit de ses marginales. Dans ce contexte, la divergence de Kullback-Leibler peut être une bonne candidate (c.f. [46]). Cette dernière se définit de la manière suivante :

Définition 2.3.1 *Soient deux fonctions f et g intégrables par rapport à une même mesure μ. On note la divergence de Kullback-Leibler entre f et g par :*

$$KL(f, g) = \int f \log \frac{f}{g} d\mu \tag{2.1}$$

Notons que cette intégrale est définie dans $\overline{\mathbb{R}}^+$.

On définit alors l'information mutuelle comme la divergence de Kullback-Leibler entre $p_{T_1,T_2,\ldots,T_K}$ et $\prod_{i=1}^{K} p_{T_i}$:

Définition 2.3.2 (Information mutuelle) *Soient $T_1, T_2, \ldots, T_K$, K variables aléatoires réelles telles que :*
le vecteur aléatoire $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ est de loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^K$. On note $p_{T_1,T_2,\ldots,T_K}$ la densité de celui-ci.
(Par conséquent, pour tout i, $1 \leq i \leq K$, la variable aléatoire T_i est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$. On note p_{T_i} sa densité.)

On appelle information mutuelle entre $T_1, T_2, \ldots, T_K$ la fonction à valeurs dans $\mathbb{R}$ suivante :

$$I(T_1, T_2, \ldots, T_K) = KL(p_{T_1,T_2,\ldots,T_K}, \prod_{i=1}^{K} p_{T_i}) =$$
$$= -\int \log \frac{\prod_{i=1}^{K} p_{T_i}(t_i)}{p_{T_1,T_2,\ldots,T_K}(t_1, t_2, \ldots, t_K)} \, p_{T_1,T_2,\ldots,T_K}(t_1, t_2, \ldots, t_K) dt_1 dt_2 \cdots dt_K \tag{2.2}$$

Remarque 2.3.2 *Dans la définition précédente, nous avons pris comme mesure de référence la mesure de Lebesgue pour définir la densité des variables considérées. Mais, il est tout à fait possible d'envisager de prendre une autre mesure de référence. En particulier, dans le cas où les lois sont discrètes, la mesure dénombrement peut-être considérée comme mesure de référence. Cette définition de l'information mutuelle peut donc être utilisée pour des variables aléatoires discrètes ou de loi absolument continue par rapport à n'importe quelle mesure de référence.*

Notons que l'expression de la divergence de Kullback-Leibler (2.1) n'est pas symétrique, on la prend sous cette forme car elle va nous permettre de poursuivre nos calculs.

2.3.2 Propriétés

La divergence de Kullback-Leibler vérifie certaines propriétés, classiques, intéressantes pour notre développement. (Nous ne démontrerons pas ces propriétés, on peut se référer à [46])

Propriété 1 : L'information mutuelle permet de caractériser l'indépendance de variables aléatoires.

Soient deux fonctions f et g intégrables par rapport à une même mesure μ. Alors,

- *$KL(f, g) \geq 0$.*
- *et $KL(f, g) = 0$ si et seulement si $f = g$.*

Ce résultat est une conséquence directe de l'inégalité de Jensen.
En appliquant simplement ce lemme à l'information mutuelle définie ci-dessus, nous obtenons,

Avec les hypothèses de la définition 2.3.2,

- *$I(T_1, T_2, \ldots, T_K) \geq 0$.*

– *et* $I(T_1, T_2, \ldots, T_K) = 0$ *si et seulement si* $T_1, T_2, \ldots, T_K$ *sont mutuellement indépendantes.*

Propriété 2 : L'information mutuelle est invariante par composition par des fonctions inversibles, c'est-à-dire,

Avec les hypothèses de la définition 2.3.2,
pour toutes fonctions $l_1, l_2, \ldots, l_K$ *inversibles, dérivables, d'inverses dérivables,*

$$I(T_1, T_2, \ldots, T_K) = I(l_1(T_1), l_2(T_2), \ldots, l_K(T_K))$$

Il suffit d'écrire les transformations affectant les densités conjointes et marginales.

Propriété 3 : Nous pouvons écrire l'information mutuelle grâce à l'entropie différentielle définie par Shannon.

Soient $T_1, T_2, \ldots, T_K$, K *variables aléatoires réelles telles que, le vecteur aléatoire* $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ *est de loi absolument continue par rapport à la mesure de Lebesgue sur* $\mathbb{R}^K$. *On note* $p_{T_1,T_2,\ldots,T_K}$ *la densité de celui-ci.*
On suppose de plus que l'entropie de Shannon de $\mathbf{T}$,$T_1, T_2, \ldots,$ *et* T_K *existe.*
Alors,

$$I(T_1, T_2, \ldots, T_K) = \sum_{i=1}^{K} H(T_i) - H(\mathbf{T}) \tag{2.3}$$

où

$$H(\mathbf{T}) = -\int p_{T_1,T_2,\ldots,T_K} \log p_{T_1,T_2,\ldots,T_K} \qquad et \qquad H(T_i) = -\int p_{T_i} \log p_{T_i}$$

preuve :

On exprime le logarithme du quotient, et on utilise l'égalité,
pour $i = 1, \ldots, K$, pour tout t_i réel,

$$\int p_{T_1,\ldots,T_K}(t_1, \ldots, t_K) \prod_{1 \leq k \leq K,\ k \neq i} dt_k = p_{T_i}(t_i).$$

■

Remarque 2.3.3 *Il est indispensable de supposer que toutes les entropies de Shannon intervenant sont bien définies car on exprime l'information mutuelle sous la forme d'une différence de deux entropies.*

Propriété 4 : Lien entre entropie et divergence de Kullback-Leibler avec la densité uniforme.

En effet, on a aussi la possibilité d'exprimer tout simplement l'entropie à l'aide de la divergence de Kullback-Leibler.

Soit Y une variable aléatoire de densité p_Y et de support $[0,1]$.
Alors,

$$H(Y) = KL(p_U, p_Y)$$

où U est une variable aléatoire de loi uniforme sur $[0,1]$, $p_U(u) = \mathbb{1}_{[0,1]}(u)$.
preuve :

Il suffit de remarquer que,

$$H(Y) = -\int p_Y(y) \log p_Y(y) dy = -\int p_Y(y) \log \frac{p_Y(y)}{p_U(y)} dy$$

et $p_Y(y) \log p_U(y) = 0, \quad y \in [0,1]$

■

2.3.3 Information mutuelle utilisée en séparation aveugle de sources (BSS) dans le cadre de mélanges linéaires

D'après les résultats précédents, il est naturel de proposer la mesure de l'information mutuelle (2.2) pour caractériser l'indépendance mutuelle de variables aléatoires. Cependant, il n'est pas possible d'utiliser l'information mutuelle sous la forme (2.2) à cause des problèmes dus à l'estimation. Afin de résoudre ce problème, de nombreux auteurs ont envisagés diverses méthodes d'estimation qui ont conduit à l'implémentation de différents critères de séparation.

Rappelons ici le cadre de la séparation aveugle de sources pour les mélanges linéaires.

Etant donné des variables aléatoires réelles $\mathbf{X} = (X_1, \ldots, X_K)^T$ vérifiant le système suivant,

$$\mathbf{X} = \mathbf{AS}$$

où $\mathbf{S} = (S_1, \ldots, S_K)^T$ est un vecteur aléatoire réel tel que ses composantes soient **indépendantes**, et $\mathbf{A}$ une matrice inversible.

Notre objectif est alors de retrouver les sources $S_1, \ldots, S_K$ connaissant les observations $X_1, \ldots, X_K$. On notera dans la suite, $\mathbf{Y} = (Y_1, \ldots, Y_K)^T$ les sources reconstituées vérifiant le système,

$$\mathbf{Y} = \mathbf{BX}.$$

Regardons à présent comment certaines méthodes classiques de séparation aveugle de sources dans le cadre de mélanges *linéaires* utilisent l'information mutuelle soit en faisant des hypothèses supplémentaires (sur la densité des sources par exemple), soit en utilisant différentes méthodes d'estimation. Nous pouvons consulter à ce propos [14]. Comme nous l'avons énoncé précédemment, dans le cadre du mélange linéaire, il est possible de s'attacher à ne retrouver que des variables indépendantes par paires, [20].

- Lien avec la méthode du **maximum de vraisemblance**:

  Différents auteurs se sont intéressés à un critère basé sur l'utilisation d'une fonction de vraisemblance: Gaeta et Lacoume [33], Garat et Pham [32] [60], Cardoso [13], Zaidi [77] ...
  La fonction de vraisemblance permettant de comparer la loi des sources reconstituées et celle des vraies sources peut en fait être reliée à l'information mutuelle, pour plus de détails, on peut consulter [14] et [12]. Donnons ici les grandes lignes de cette méthode.

  On note p_{S_i} la densité de chaque source S_i alors, la densité conjointe vérifie, $p_{\mathbf{S}} = \prod_{i=1}^{K} p_{S_i}$.

  Lemme 2.3.1 *Soient $X_k(1), \ldots, X_k(N)$ un échantillon de la variable X_k pour tout k, $1 \leq k \leq K$.*
  Alors, la log-vraisemblance associée aux variables $X_1, \ldots, X_K$ définie par,

$$\begin{aligned} L_N(\mathbf{A}, p_{\mathbf{S}}) &= \frac{1}{N} \sum_{n=1}^{N} \log p_{\mathbf{X},\mathbf{A},p_{\mathbf{S}}}(\mathbf{X}(n)) \\ &= \frac{1}{N} \sum_{n=1}^{N} \log p_{\mathbf{S}}(\mathbf{A}^{-1}\mathbf{X}(n)) - \log |det\mathbf{A}| \end{aligned}$$

  admet comme limite,

$$\begin{aligned} \lim_{N \to +\infty} L_N(\mathbf{A}, p_{\mathbf{S}}) &= E[\log p_{\mathbf{S}}(\mathbf{A}^{-1}\mathbf{X})] - \log |det\mathbf{A}| \\ &= -H(\mathbf{X}) - KL(p_{\mathbf{A}^{-1}\mathbf{X}}, p_{\mathbf{S}}) \end{aligned}$$

La méthode du maximum de vraisemblance revient donc asymptotiquement à minimiser la divergence de Kullback-Leibler entre la vraie densité des sources et celle obtenue après inversion de la matrice de mélange ($\mathbf{B} = \mathbf{A}^{-1}$) appliquée sur les observations. Cette méthode présente alors l'inconvénient d'avoir besoin de faire des hypothèses sur les densités des sources afin de calculer la log-vraisemblance. On remarque aussi que par ces propriétés asymptotiques, cette méthode est optimale dans le sens qu'elle maximise une vraisemblance.

- Lien avec la méthode **InfoMax** introduite par Bell et Sejnowski [8]:

  Cette méthode peut aussi être dérivée de l'information mutuelle, citons à ce propos [54]. Rappelons brièvement le principe de cette méthode afin de la relier à l'information mutuelle.

  Définition 2.3.3 *Soient ϕ_i, $1 \leq i \leq K$ des fonctions non linéaires fixées telles que pour tout i, $1 \leq i \leq K$, ϕ_i est différentiable et*

  $$\int \phi_i'(x)dx = 1$$

  On définit pour tout i, $1 \leq i \leq K$, $W_i = \phi_i(Y_i)$.
  Alors, la méthode InfoMax consiste à chercher le maximum de

  $$H(\mathbf{W}) = -\int p_{\mathbf{W}}(\omega) \log p_{\mathbf{W}}(\omega) d\omega$$

  par rapport à $\mathbf{B}$.

  Dans [54], les auteurs précisent que souvent les fonctions non linéaires ϕ_i sont prises égales à des sommes de sigmoïdes. Regardons alors où intervient l'information mutuelle, dans [54], on remarque que,

  $$H(\mathbf{W}) = KL(p_{\mathbf{Y}}, \prod_{i=1}^{K} \phi_i').$$

  Il est clair alors que le critère utilisé dans la méthode InfoMax correspond à l'information mutuelle lorsque les dérivées des transformations non linéaires ϕ_i' coïncident avec les densités marginales des variables reconstituées Y_i. La méthode InfoMax peut donc être vue comme une estimation particulière de l'information mutuelle. En effet, pour que la maximisation de l'entropie de $\mathbf{W}$ corresponde à la minimisation de l'information mutuelle, chaque fonction ϕ_i doit coïncider avec la fonction de répartition de Y_i de telle sorte que W_i sera alors une variable uniforme.

- **Néguentropie**:

  Intéressons nous à présent à une autre méthode dérivée aussi de l'information mutuelle. Tout d'abord, rappelons la définition de la néguentropie ([20], [37]) qui est à la base des méthodes explicitées ci-après.

  Définition 2.3.4 *Soit* $\mathbf{T}$ *un vecteur aléatoire réel. On suppose que la densité de* $\mathbf{T}$*,* $p_{\mathbf{T}}$*, admet des moments d'ordre 2 et que sa matrice de covariance* V *est inversible.*
  Alors, la néguentropie de $\mathbf{T}$ *notée* $J(\mathbf{T})$ *est définie par,*

$$J(T_1, \ldots, T_K) = H(\mathbf{T}_g) - H(\mathbf{T})$$

  où $\mathbf{T}_g$ *est le vecteur aléatoire gaussien admettant même moyenne et même matrice de covariance que* $\mathbf{T}$*.*

  On déduit que,

$$I(T_1, \ldots, T_K) = J(T_1, \ldots, T_K) - \sum_{i=1}^{K} J(T_i) + \frac{1}{2} \log \frac{\prod_{i=1}^{K} V_{ii}}{\det V}$$

  En remarquant que la néguentropie est invariante par multiplication par une matrice, le critère considéré par les auteurs Comon [20] et Hyvärinen, Karhunen et Oja [37] prend en compte simplement la somme des néguentropies marginales. Cependant, la présence du terme avec la matrice de covariance V, nécessite de procéder à un pré-blanchiment des données. Nous verrons que les méthodes considérées par la suite ont l'avantage de ne pas avoir cette contrainte.
  Ici aussi, la maximisation de la somme des néguentropies marginales revient à minimiser l'information mutuelle après blanchiment des observations.

  Ces méthodes se rapprochent aussi de celles utilisant la notion de kurtosis. On peut alors les caractériser par le fait qu'elles cherchent des sources reconstituées de telles sorte que ces dernières soient le moins gaussiennes possible.

  En effet, si on note $\mathbf{W}$ la matrice de blanchiment des observations, et $\mathbf{Z} = \mathbf{WX}$. Alors le problème de séparation aveugle de sources consiste en la recherche d'une matrice $\mathbf{U}$ de rotation telle que

$$\sum_{i=1}^{K} H(Y_i) \text{ soit minimale}$$

  où $\mathbf{Y} = \mathbf{UZ}$.

En effet, $I(\mathbf{Y}) = \sum_{i=1}^{K} H(Y_i) - H(\mathbf{Z}) - \log|\det \mathbf{U}| = \sum_{i=1}^{K} H(Y_i) - H(\mathbf{Z})$.

On cherche alors les variables Y_i avec une variance fixée telle que l'entropie respective de chaque Y_i soit minimale. Comme la maximisation de l'entropie avec une variance fixée est réalisée lorsque les variables sont gaussiennes, ceci explique le fait que l'on parle de méthodes qui "éloignent" de la loi gaussienne.

- **Statistiques d'ordre supérieur**:

  En partant du critère de néguentropie, Comon [20], propose une méthode d'estimation de l'information mutuelle utilisant le développement d'Edgeworth de la densité.
  Il est tout à fait connu que l'utilisation des moments d'ordre deux ne permet pas de décider si des variables non gaussiennes sont indépendantes. Par contre, l'utilisation des cumulants croisés de tous ordres permet de décider si des variables sont indépendantes ou non. Pour cela, on utilise la propriété suivante, si tous les cumulants croisés d'un ensemble de variables aléatoires de tous ordres sont nuls, alors les variables aléatoires sont indépendantes. Dans [20], Comon définit une fonction de contraste en utilisant les cumulants d'ordre 4. Il démontre alors géométriquement que cette fonction permet de retrouver les sources provenant d'un mélange linéaire.

2.3.4 Un critère pour l'analyse en composantes indépendantes (ICA)

L'expression de l'information mutuelle (2.2) ne permet pas de définir un critère de séparation. Nous devons proposer une estimation afin de définir une fonction que l'on pourra effectivement optimiser.

D'après la définition de l'information mutuelle, l'estimation de cette dernière conduit à envisager différentes estimations de l'entropie. Mais, quand on considère l'expression de l'information mutuelle précédente (2.2), il apparait la nécessité d'estimer l'entropie conjointe. Or, ceci peut poser des problèmes dus à l'estimation de densité conjointe d'un trop grand nombre de variables. Cependant, il est connu que dans le cas de mélange linéaire ou post non linéaire, grâce aux propriétés de l'entropie, on peut définir un critère de séparation ne dépendant que des entropies marginales de chaque variable. Nous considérerons dans la section 4.4.8 l'étude de différentes estimations de l'information mutuelle ou du critère de séparation envisagé par Taleb et Jutten [69]. Puis dans la section 5.5, nous analyserons les différences de ses estimations en comparant le biais des estimateurs.

Ce petit récapitulatif sur l'information mutuelle, nous a permis de voir comment se situe la mesure d'information mutuelle par rapport aux nombreuses méthodes déjà développées. Nous avons pu voir que celle-ci est à la base de multiples méthodes.
Nous allons à présent introduire une nouvelle mesure de dépendance basée sur une mesure de dépendance quadratique et non sur la distance de Kullback-Leibler.

2.4 Mesure de dépendance quadratique

La mesure de dépendance quadratique que nous définissons ici, est une généralisation de la statisque de test étudiée par Kankainen, [44] et du critère de séparation utilisé par Eriksson *et. al.* dans le cadre de mélange linéaire [29, 30].

Tout d'abord, dans le paragraphe 2.4.1, après quelques lemmes préliminaires, nous définirons la mesure de dépendance quadratique (definition 2.4.1).

Puis, dans le paragraphe 2.4.2, nous étudierons quelques unes de ses propriétés. Les lemmes 2.4.3 et 2.4.4 insistent sur les hypothèses à faire pour que la mesure de dépendance quadratique permette de déterminer si des variables sont indépendantes. Ensuite, le lemme 2.4.5 établit les propriétés d'invariance de la mesure de dépendance quadratique par translation et par mutiplication par un scalaire. A la différence de l'information mutuelle, on note que la mesure de dépendance quadratique ne nécessite pas de faire une estimation précise de la densité des variables aléatoires (c.f. lemme 2.4.3). Cependant, dans certaines conditions, la mesure de dépandance quadratique peut être interprétée comme une mesure comparant la densité conjointe et le produit des densités marginales (c.f. lemme 2.4.6).

Enfin, le lemme 2.4.7, exprime la mesure de dépendance quadratique à l'aide des fonctions caractéristiques jointes et marginales.

2.4.1 Définition

Comme pour la construction de la mesure de dépendance appelée information mutuelle, nous analysons ici aussi une condition permettant de caractériser complètement l'indépendance de variables aléatoires. Considérons alors le lemme suivant :

Lemme 2.4.1 *Soient $T_1, \ldots, T_K$, K variables aléatoires réelles.*

On note $\mathbf{T} = (T_1, \ldots, T_K)^T$ le vecteur aléatoire associé à $T_1, \ldots, T_K$.
Alors, les variables $T_1, \ldots, T_K$ sont mutuellement indépendantes si et seulement si

$$E\left[\prod_{k=1}^{K} f_k(T_k)\right] = \prod_{i=1}^{K} E[f_k(T_k)], \tag{2.4}$$

pour toutes fonctions $f_1, \ldots, f_K$ mesurables et intégrables par rapport aux lois des T_k et telles que $\prod_{k=1}^{K} f_k$ soit intégrable par rapport à la loi de $\mathbf{T}$.

Quelques remarques : (Pour une preuve, on peut consulter [62] ou [51] pour plus de détails.)

En effet, selon un choix particulier des fonctions $f_1, \ldots, f_K$, l'égalité du lemme n'exprime en fait que des conditions bien connues traduisant l'indépendance des variables aléatoires.

– si pour tout t_i réel, $f_{t_i}(x) = \mathbb{1}_{\{x \leq t_i\}}$ pour i, $1 \leq i \leq K$, alors (2.4) s'écrit,

$$F_{\mathbf{T}}(t_1, \ldots, t_K) = \prod_{i=1}^{K} F_{T_i}(t_i)$$

pour tous $t_1, \ldots, t_K$ réels.
où $F_{\mathbf{T}}$ représente la fonction de répartition du vecteur $\mathbf{T}$,

$$F_{\mathbf{T}}(t_1, \ldots, t_K) = P(T_1 \leq t_1, \ldots, T_K \leq t_K) = E\left[\prod_{i=1}^{K} \mathbb{1}_{\{T_i \leq t_i\}}\right]$$

et F_{T_i} représente la fonction de répartition de la variable T_i,

$$F_{T_i}(t_i) = P(T_i \leq t_i) = E\left[\mathbb{1}_{\{T_i \leq t_i\}}\right].$$

On reconnait ici clairement la propriété caractérisant l'indépendance de variables aléatoires définie à partir des fonctions de répartition.

– si pour tout t_k réel, $f_{t_k}(x) = \exp(ixt_k)$, alors (2.4) s'écrit,

$$\psi_{\mathbf{T}}(t_1, \ldots, t_K) = \prod_{k=1}^{K} \psi_{T_k}(t_k)$$

pour tous $t_1, \ldots, t_K$ réels.
où $\psi_{\mathbf{T}}$ désigne la fonction caractéristique du vecteur aléatoire T,

$$\psi_{\mathbf{T}}(t_1, \ldots, t_K) = E\left[\exp\left(i\sum_{k=1}^{K} t_k T_k\right)\right] = E\left[\prod_{k=1}^{K} \exp(it_k T_k)\right]$$

et ψ_{T_k} représente la fonction caractéristique marginale de chaque variable T_k,

$$\psi_{T_k}(t_k) = E[\exp(it_k T_k)]$$

– On peut aussi définir bien d'autres ensembles de fonctions plus restreints qui permettront de satisfaire la condition d'indépendance.

Nous choisissons ici de nous focaliser sur une autre possibilité de définition des fonctions f_i, à partir de noyaux.

On pose pour tout x réel, $f_i(t) = \mathcal{K}_i(x-t)$ où $\mathcal{K}_i$ est un noyau choisi de telle sorte que les quantités utilisées précédemment soient bien définies, i.e. que les différentes espérances existent. Remarquons que l'on se place ici dans le cas général où l'on a choisi K noyaux qui peuvent être différents les uns des autres.

Lemme 2.4.2 (Indépendance avec les noyaux) *Soient $\mathcal{K}_1, \ldots, \mathcal{K}_K$ des noyaux intégrables tels que leurs transformées de Fourier soient différentes de zéro presque partout. On note de plus $dF_{\mathbf{T}}$, la loi du vecteur aléatoire $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ et dF_{T_k} la loi de chaque variable T_k.*

Alors, les variables aléatoires $T_1, \ldots, T_K$ sont indépendantes si et seulement si,

$$E\left[\prod_{i=1}^{K} \mathcal{K}_i(x_i - T_i)\right] = \prod_{i=1}^{K} E[\mathcal{K}_i(x_i - T_i)], \quad \textit{pour tous } x_1, \ldots, x_K \textit{ réels} \tag{2.5}$$

preuve :

Tout d'abord, réécrivons l'égalité (2.5) de la manière suivante :

pour tous $t_1, \ldots, t_K$ réels,

$$\begin{aligned} E\left[\prod_{i=1}^{K} \mathcal{K}_i(t_i - T_i)\right] &= \left[\prod_{i=1}^{K} \mathcal{K}_i\right] * dF_{\mathbf{T}}(t_1, \ldots, t_K) \\ \prod_{i=1}^{K} E[\mathcal{K}_i(t_i - T_i)] &= \prod_{i=1}^{K} [\mathcal{K}_i * dF_{T_i}](t_i) \end{aligned}$$

où * désigne le produit de convolution d'une fonction f et d'une loi $d\mu$, $f * \mu(x) = \int f(x-t)d\mu(t))$

On en déduit donc que (2.5) est équivalente à

$$\left[\prod_{i=1}^{K} \mathcal{K}_i\right] * dF_{\mathbf{T}}(t_1, \ldots, t_K) = \prod_{i=1}^{K} [\mathcal{K}_i * dF_{T_i}](t_i) \quad \text{pour tous } t_1, \ldots, t_K \text{ réels}$$

En prenant la transformée de Fourier de chaque coté de (2.5), on montre encore que cette dernière est équivalente à

$$\prod_{i=1}^{K} \psi_{\mathcal{K}_i}(t_i)\psi_{\mathbf{T}}(t_1,\ldots,t_K) = \prod_{i=1}^{K} \psi_{\mathcal{K}_i}(t_i)\psi_{T_i}(t_i) \quad \text{pour tous } t_1,\ldots,t_K \text{réels}$$

où $\psi_{\mathcal{K}_i}$ désigne la transformée de Fourier de $\mathcal{K}_i$ ($\psi_{\mathcal{K}_i}(x) = \int \mathcal{K}_i(t)\exp(itx)dt$) et $\psi_{\mathbf{T}}$ désigne la transformée de Fourier de $dF_{\mathbf{T}}$, c'est la fonction caractéristique de $\mathbf{T}$.

D'où, si les transformées de Fourier des noyaux $\mathcal{K}_i$ sont non nulles presque partout, (2.5) est équivalente à

$$\psi_{\mathbf{T}}(t_1,\ldots,t_K) = \prod_{i=1}^{K} \psi_{T_i}(t_i)$$

pour presque tous $t_1,\ldots,t_K$ réels.

Comme les fonctions caractéristiques sont continues, [50], cette dernière égalité est équivalente à dire que les variables aléatoires $T_1,\ldots,T_K$ sont indépendantes.

■

Remarquons pour finir que la quantité, définie à partir de (2.5),

$$E\left[\prod_{i=1}^{K} \mathcal{K}_i(x_i - T_i)\right] - \prod_{i=1}^{K} E[\mathcal{K}_i(x_i - T_i)], \quad \text{pour tous } x_1,\ldots,x_K \text{ réels}$$

n'est pas invariante par multiplication des variables $T_1,\ldots,T_K$ par un scalaire. Nous introduisons alors un coefficient de facteur d'échelle, qui peut par exemple être l'écart-type de chaque variable.

En conséquence, la mesure de dépendance quadratique sera définie de la manière suivante :

Définition 2.4.1 (Mesure de dépendance quadratique) *Soient $T_1,\ldots,T_K$ des variables aléatoires et $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$. Soient $\mathcal{K}_k$, pour $1 \leq k \leq K$, des noyaux de carré intégrable, tels que leurs transformées de Fourier soient différentes de zéro presque partout. On définit alors la mesure de dépendance quadratique des K variables aléatoires $T_1,\ldots,T_K$ par,*

$$Q(T_1,\ldots,T_K) = \int D_{\mathbf{T}}(t_1,\ldots,t_K)^2 dt_1 \ldots dt_K.$$

où pour tous $t_1,\ldots,t_K$ réels,

$$D_{\mathbf{T}}(t_1,\ldots,t_K) = E\left[\prod_{k=1}^{K} \mathcal{K}_k\left(t_k - \frac{T_k}{\sigma_{T_k}}\right)\right] - \prod_{k=1}^{K} E\left[\mathcal{K}_k\left(t_k - \frac{T_k}{\sigma_{T_k}}\right)\right]$$

et σ_{T_k} est un coefficient de facteur d'échelle, c'est-à-dire une fonction positive dépendant seulement de la loi des T_k tel que $\sigma_{\lambda T_i} = |\lambda|\sigma_{T_i}$, pour toute constante réelle λ.

Remarque 2.4.1 *Nous remarquons que dans cette définition, il est possible d'utiliser des noyaux différents pour chaque variable aléatoire. Mais, il est naturel de se poser ici la question du choix des noyaux. En effet, théoriquement, il est possible de choisir n'importe quel noyau vérifiant les propriétés de la définition précédente et de choisir K noyaux différents. Il s'avère alors difficile de réellement étudier la mesure de dépendance quadratique dans ces conditions, le nombre de paramètres à faire varier étant trop grand. Par conséquent, dans la suite nous choisirons un seul et même noyau $\mathcal{K}$ pour toutes les variables aléatoires, tout en gardant à l'esprit que ces noyaux peuvent être choisis différents.*

Existence de la définition :

Avant de poursuivre, justifions que les expressions écrites ci-dessus sont bien définies.
Tout d'abord, par un changement de variable, nous pouvons écrire que,

$$Q(T_1,\dots,T_K) = \frac{1}{\prod_{k=1}^K \sigma_{T_k}} \int \left\{ E\left[\prod_{k=1}^K \mathcal{K}\left(\frac{t_k - T_k}{\sigma_{T_k}}\right)\right] - \prod_{k=1}^K E\left[\mathcal{K}\left(\frac{t_k - T_k}{\sigma_{T_k}}\right)\right] \right\}^2 dt_1 \dots dt_K.$$

Or la fonction Q peut aussi s'écrire en introduisant le noyau $\mathcal{K}_{\sigma_{T_k}}$ défini par, pour tout x réel,

$$\mathcal{K}_{\sigma_{T_k}}(x) = \mathcal{K}\left(\frac{x}{\sigma_{T_k}}\right)$$

$$Q(T_1,\dots,T_K) = \frac{1}{\prod_{k=1}^K \sigma_{T_k}} \int \left\{ \left[\prod_{k=1}^K \mathcal{K}_{\sigma_{T_k}}\right] * dF_{\mathbf{T}}(t_1,\dots,t_K) - \prod_{k=1}^K \mathcal{K}_{\sigma_{T_k}} * dF_{T_k}(t_k) \right\}^2 d\mathbf{t}.$$

où $d\mathbf{t} := dt_1 \dots dt_K$.
En reprenant les calculs effectués dans le lemme précédent, les transformées de Fourier de $(t_1,\dots,t_K) \mapsto \left[\prod_{k=1}^K \mathcal{K}_{\sigma_{T_k}}\right] * dF_{\mathbf{T}}(t_1,\dots,t_K)$ et $t_k \mapsto \mathcal{K}_{\sigma_{T_k}} * dF_{T_k}(t_k)$ sont respectivement égales à

$$(t_1,\dots,t_K) \mapsto \prod_{k=1}^K \sigma_{T_k}\psi_{\mathcal{K}}(\sigma_{T_k}t_k)\psi_T(t_1,\dots,t_K) \text{ et } t_k \mapsto \sigma_{T_k}\psi_{\mathcal{K}}(\sigma_{T_k}t_k)\psi_{T_k}(t_k)$$

Or, d'après la formule de Plancherel-Parseval qui montre que la transformée de Fourier conserve la norme dans L^2, on obtient,

$$Q(Y_1,\ldots,Y_K) = \frac{1}{(2\pi)^K \prod_{k=1}^K \sigma_{T_k}} \int \prod_{k=1}^K |\sigma_{T_k}\psi_{\mathcal{K}}(\sigma_{T_k}t_k)|^2 \left|\psi_T(t_1,\ldots,t_K) - \prod_{k=1}^K \psi_{T_k}(t_k)\right|^2 d\mathbf{t}.$$

Cette dernière égalité montre bien que Q sera définie dès que $\psi_{\mathcal{K}}$ est de carré intégrable, puisque les fonctions caractéristiques sont des fonctions bornées. La condition imposant de prendre un noyau de carré intégrable est donc suffisante à l'existence de la mesure de dépendance quadratique Q.

Références bibliographiques :

De manière similaire à la notion de densité de probabilité et de fonction de répartition, la notion de fonction caractéristique est très utilisée en statistique. Comme nous l'avons remarqué, en utilisant les fonctions caractéristiques, il est possible de décider si des variables sont indépendantes. Des techniques de test d'indépendance ont été développées. Dans un premier temps, Csörgő [24] a utilisé un test d'indépendance basé sur les fonctions caractéristiques :

$$S_n(\mathbf{t}^{(0)}) = n^{1/2}\{c_n(\mathbf{t}^{(0)}) - \prod_{k=1}^d c_{nk}(t_k^{(0)})\},$$

où $\mathbf{t}^{(0)}$ est la valeur qui maximise la fonction de variance complexe du processus Gaussien centré $S_F(\mathbf{t})$ dans un compact connexe K de $\mathbb{R}^d$. Et, c_n et c_{nk} sont respectivement les estimations des fonctions caractéristiques conjointes et marginales. Mais, dans [44], Kankaiken fait remarquer que ce test n'est pas consistant. Dans un deuxième temps, Feuerverger [31] introduit le test d'indépendance basé sur les fonctions caractéristiques défini par,

$$T'_n = \iint \frac{|\Gamma'_n(s,t)|}{(1-\exp(-s^2))(1-\exp(-t^2))} W(s,t)dsdt,$$

où W est une fonction de poids positive choisie de manière adéquate et $\Gamma'_n(s,t) = c'_n(s,t) - c_n^{X'}(t)c_n^{Y'}(t)$ où X' (resp. Y') sont les normales scores estimées de X (resp. Y). De plus, $c'_n(s,t)$, $c_n^{X'}$ et $c_n^{Y'}$ sont respectivement les estimations des fonctions caractéristiques jointes et marginales des variables X' et Y' considérées. Remarquons que cette procédure ne permet que de faire le test d'indépendance de deux variables aléatoires. Or il est connu que l'indépendance par paire n'implique pas l'indépendance mutuelle (ou dans l'ensemble). Enfin, Kankainen dans [44], introduit alors un test d'indépendance basé sur les fonctions caractéristiques,

consistant et permettant de tester l'indépendance mutuelle. En effet, dans [44], l'auteur introduit le test suivant,

$$T_n = n \int_{\mathbb{R}^d} |c_n(\mathbf{t}) - \prod_{k=1}^{d} c_{nk}(t_k)|^2 g(\mathbf{t}) d\mathbf{t},$$

où c_n et c_{nk} sont respectivement les estimations des fonctions caractéristiques conjointes et marginales, et g est une fonction de poids à choisir de manière adéquate. Ce test d'indépendance a ensuite été utilisé par Eriksson, Kankainen et Koivunen, [29, 30] dans le cadre du problème de séparation de sources dans les mélanges linéaires. Pour des questions de facilité de calculs, dans ces travaux, la fonction de poids g a été choisie comme le produit de densités normales ou de densités de Laplace à une variable. En effet, la nécessité de calculer des intégrales multiples motive l'emploi de fonctions simples à utiliser. La mesure de dépendance quadratique peut donc être vue comme un cas particulier de la définition de ce test d'indépendance en imposant à la fonction de poids d'être le carré du produit de fonctions d'une variable. Dans ce qui suit, nous nous bornerons à choisir des fonctions de poids pouvant s'écrire comme produit de fonctions d'une variable. De plus, avec la définition que nous proposons, il est possible d'envisager n'importe quel produit de fonctions avec certaines précautions en ce qui concerne les hypothèses sur les noyaux utilisés, voir la section 2.5.

2.4.2 Propriétés

Après avoir défini précédemment la mesure de dépendance quadratique et en vue de l'utiliser comme une fonction mesurant la dépendance de variables aléatoires, nous allons à présent détailler certaines de ses propriétés.

Propriété 1 : La mesure de dépendance quadratique définie précédemment est bien une mesure de dépendance dans le sens qu'elle caractérise parfaitement l'indépendance mutuelle de variables aléatoires réelles.

Lemme 2.4.3 (Caractérisation de l'indépendance) *Dans les hypothèses de la définition 2.4.1,*
- $Q(T_1, \ldots, T_K) \geq 0$
- $T_1, \ldots, T_K$ *sont mutuellement indépendantes si et seulement si* $Q(T_1, \ldots, T_K) = 0$.

preuve :

Dans le paragraphe précédent 2.4.1, nous avons fait la constatation suivante,

$$Q(Y_1,\ldots,Y_K) \;=\; \frac{1}{(2\pi)^K \prod_{k=1}^K \sigma_{T_k}} \int \prod_{k=1}^K |\sigma_{T_k}\psi_{\mathcal{K}}(\sigma_{T_k}t_k)|^2 \big|\psi_T(\mathbf{t}) - \prod_{k=1}^K \psi_{T_k}(t_k)\big|^2 d\mathbf{t}.$$

où $\mathbf{t} = (t_1,\ldots,t_K)$.
Ceci montre alors clairement que

- $Q(T_1,\ldots,T_K) \geq 0$
- si $T_1,\ldots,T_K$ sont mutuellement indépendantes alors ,
  $\psi_T(t_1,\ldots,t_K) - \prod_{k=1}^K \psi_{T_k}(t_k) = 0$ pour tous $t_1,\ldots,t_K$ réels.
  Donc $Q(T_1,\ldots,T_K) = 0$
- si $Q(T_1,\ldots,T_K) = 0$ alors $\psi_T(t_1,\ldots,t_K) - \prod_{k=1}^K \psi_{T_k}(t_k) = 0$ pour tous $t_1,\ldots,t_K$ tels que $\prod_{k=1}^K \psi_{\mathcal{K}}(t_k)$ est différent de zéro. Par hypothèse sur les noyaux $\mathcal{K}$, cette condition est vérifiée presque partout, donc $\psi_T(t_1,\ldots,t_K) - \prod_{k=1}^K \psi_{T_k}(t_k) = 0$ pour presque tous $t_1,\ldots,t_K$ réels. Alors comme les fonctions caractéristiques sont continues, (voir par exemple Lukacs [50]), les variables $T_1,\ldots,T_K$ sont mutuellement indépendantes. ■

Remarque 2.4.2 *L'hypothèse sur les points d'annulation des transformations de Fourier des noyaux peut-être affaiblie en supposant la fonction caractéristique conjointe analytique. En effet, on a le résultat suivant :*

Lemme 2.4.4 *Soit $\mathcal{K}$ un noyau de carré intégrable tel que sa transformée de Fourier soit différente de zéro sur un ensemble de la forme $]-a;a[$ contenu dans $\mathbb{R}$ et contenant 0.*
Soient $T_1,\ldots,T_K$ des variables aléatoires telles que leur fonction caractéristique conjointe est analytique.
Alors, $T_1,\ldots,T_K$ sont mutuellement indépendantes si et seulement si $Q(T_1,\ldots,T_K) = 0$.

preuve :
Ceci vient de la propriété suivante, si deux fonctions caractéristiques analytiques coïncident sur $\prod_{i=1}^K]-a_i;a_i[$, elles coïncident partout, c.f. Cuppens, [25].
■

Insistons sur le fait que la condition d'analyticité des fonctions caractéristiques est absolument indispensable. Reprenons pour cela l'exemple de Kankainen, [44], où on définit la fonction caractéristique suivante,

$$c(t_1,t_2) = \begin{cases} \frac{1}{2}(1-|t_1+2|)(1-|t_2+2|), & |t_1+2|\leq 1, |t_2+2|\leq 1, \\ (1-|t_1|)(1-|t_2|), & |t_1|\leq 1, |t_2|\leq 1, \\ \frac{1}{2}(1-|t_1-2|)(1-|t_2-2|), & |t_1-2|\leq 1, |t_2-2|\leq 1, \\ 0, & \text{ailleurs} \end{cases} \tag{2.6}$$

Il est clair que

$$c(t_1, t_2) = c(t_1, 0)c(0, t_2), |t_1| \leq 1, |t_2| \leq 1.$$

Pourtant,

$$c(2, 2) = \frac{1}{2} \neq 0 = c(2, 0)c(0, 2).$$

Et, comme la fonction c est non nulle seulement sur un compact, elle ne peut pas être analytique. Sinon, elle serait nulle partout.
Cette fonction caractéristique c non analytique coïncide donc sur $[-1, 1] \times [-1; 1]$ avec la fonction $(t_1, t_2) \mapsto c(t_1, 0)c(0, t_2)$. Néanmoins, ces deux fonctions ne sont pas égales partout.
Dans cet exemple, la densité conjointe correspondante s'écrit,

$$f(x_1, x_2) = \frac{2}{\pi^2} \frac{(1 - \cos x_1)(1 - \cos x_2)}{x_1^2 x_2^2} \cos^2(x_1 + x_2).$$

Propriété 2 : Grâce au facteur d'échelle introduit dans la définition, la mesure de dépendance quadratique est invariante par changement d'échelle et par translation.

Lemme 2.4.5 (Invariances) *Soient $\lambda_1, \ldots, \lambda_K$ et $\mu_1, \ldots, \mu_K$ des constantes réelles. Dans les hypothèses de la définition 2.4.1,*
$Q(T_1, \ldots, T_K) = Q(\lambda_1 T_1 + \mu_1, \ldots, \lambda_K T_K + \mu_K)$

Remarquons que contrairement à l'information mutuelle, la mesure de dépendance quadratique n'est pas invariante par composition par une transformation inversible.

Propriété 3 : La mesure de dépendance quadratique peut être interprétée comme une mesure comparant la densité conjointe et le produit des densités.
Choisissons un noyau particulier de telle sorte que,

$$\mathcal{K}(u) = \frac{1}{h} \widetilde{\mathcal{K}} \left(\frac{u}{h} \right).$$

où $\widetilde{\mathcal{K}}$ est un noyau à densité (i.e. $\mathcal{K}$ est une application de $\mathbb{R}$ dans $\mathbb{R}$ bornée, positive, intégrable par rapport à la mesure de Lebesgue et d'intégrale égale à 1) et h un paramètre de lissage que l'on nomme aussi taille de fenêtre. Alors nous pouvons énoncer le lemme suivant,

Lemme 2.4.6 (Avec les densités) *Soient $T_1, \ldots, T_K$ des variables aléatoires de telle sorte que le vecteur aléatoire $\mathbf{T} = (T_1, \ldots, T_K)^T$ admette une densité conjointe notée $p_{\mathbf{T}}$. On notera p_{T_k} la densité marginale de chaque T_k. Soit $\widetilde{\mathcal{K}}$ un noyau de Parzen-Rosenblatt.(i.e. $\lim_{|x|\to\infty} |x|\widetilde{\mathcal{K}}(x) = 0$.)*

Alors, pour tout k, $1 \leq k \leq K$, et en tout point $(t_1, \ldots, t_K)$ où la densité conjointe est continue,

$$\lim_{h\to 0} E\left[\prod_{k=1}^{K} \mathcal{K}\left(t_k - \frac{T_k}{\sigma_{T_k}}\right)\right] = p_{\mathbf{T}'}(t_1, \ldots, t_K)$$

De plus, en tout point t_k où la densité marginale p_{T_k} est continue,

$$\lim_{h\to 0} E\left[\mathcal{K}\left(t_k - \frac{T_k}{\sigma_{T_k}}\right)\right] = p_{T'_k}(t_k)$$

où $\mathbf{T}'$ est le vecteur aléatoire normé correspondant à $\mathbf{T}$, c'est-à-dire $\mathbf{T}' = (T_1/\sigma_{T_1}, \ldots, T_K/\sigma_{T_K})^T$. Et, pour tous k, $1 \leq k \leq K$, T'_k est la variable aléatoire normée correspondant à T_k, c'est-à-dire, $T_k = T_k/\sigma_{T_k}$.

preuve :

On applique simplement le lemme de Bochner [10] (c.f. énoncé en annexe D), en remarquant que,

$$E\left[\prod_{k=1}^{K} \mathcal{K}\left(t_k - \frac{T_k}{\sigma_{T_k}}\right)\right] = \widetilde{\mathcal{K}}_h * p_{\mathbf{T}'}(t_1, \ldots, t_K)$$

et

$$E\left[\mathcal{K}\left(t_k - \frac{T_k}{\sigma_{T_k}}\right)\right] = \widetilde{\mathcal{K}}_h * p_{T'_k}(t_k)$$

où on a noté, $\widetilde{\mathcal{K}}_h(u) = \widetilde{\mathcal{K}}(u/h)/h$

■

Ceci indique donc que si le noyau est choisi de manière adéquate, la mesure de dépendance quadratique peut-être vue comme une comparaison entre la densité conjointe et le produit des densités marginales des variables normées. Insistons sur le fait que la mesure de dépendance quadratique garde la propriété de caractérisation de variables indépendantes sans avoir besoin de contraindre le noyau à être un noyau de Parzen-Rosenblatt et sans avoir besoin d'introduire un paramètre de lissage h qui tend vers 0.

Propriété 4 : La mesure de dépendance quadratique s'exprime à l'aide des fonctions caractéristiques.

Lemme 2.4.7 (Avec les fonctions caractéristiques) *Dans les hypothèses de la définition 2.4.1,*

$$Q(T_1,\ldots,T_K) = \int \frac{1}{\prod_{k=1}^{K}\sigma_{T_k}} \prod_{k=1}^{K} \left| \frac{\sigma_{T_k}\psi_{\mathcal{K}_i}(\sigma_{T_k}t_k)}{\sqrt{2\pi}} \right|^2 |D^c_{\mathbf{T}}(\mathbf{t})|^2 dt_1 \ldots dt_K.$$

où $\mathbf{T} = (T_1 \quad \cdots T_K)^T$ *et*

$$D^c_{\mathbf{T}}(u_1,\ldots,u_K) = \psi_{\mathbf{T}}(u_1,\ldots,u_K) - \prod_{k=1}^{K}\psi_{T_k}(u_k) \tag{2.7}$$

2.5 Mesure de dépendance quadratique comme critère pour l'analyse en composantes indépendantes

Dans la définition initiale de la mesure de dépendance quadratique, on remarque que malgré l'estimation de l'espérance par la moyenne empirique, il subsiste un calcul d'intégrale multiple. Or, il est connu que le calcul d'intégrale multiple est assez complexe et couteux. Dans le paragraphe 2.5.1, nous allons voir comment le calcul de ces intégrales multiples peut être évité par l'introduction d'un nouveau noyau. Cette procédure s'apparente au "kernel trick" défini dans la théorie des supports vecteurs machines (SVM). Dans ce même paragraphe, (lemme 2.5.2), nous verrons comment choisir le nouveau noyau pour que la mesure de dépendance quadratique conserve les propriétés énoncées dans 2.4.2. Puis, dans le paragraphe 2.5.2, nous décrirons la conséquence de certaines propriétés des nouveaux noyaux sur la mesure de dépendance quadratique. De plus, le paragraphe 2.5.3 sera consacré à la description de quelques noyaux que nous avons utilisé dans la suite de nos travaux. Enfin, nous présenterons l'estimateur utilisé pour permettre le calcul de la mesure de dépendance quadratique (paragraphe 2.5.4), en développant quelques propriétés asymptotiques (lemme 2.5.6)

2.5.1 Nouvelle écriture de la mesure de dépendance quadratique

Lemme 2.5.1 (Dépendance quadratique en fonction de $\mathcal{K}_2$) *Sous les mêmes hypothèses que dans la définition 2.4.1, la mesure de dépendance quadratique de K variables aléatoires $T_1, \ldots, T_K$ telles que $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ peut-être exprimée par,*

$$Q(T_1, \ldots, T_K) = E[\pi_{\mathbf{T}}(\mathbf{T})] + \prod_{k=1}^{K} E[\pi_{T_k}(T_k)] - 2E\left[\prod_{k=1}^{K} \pi_{T_k}(T_k)\right]$$

où

$$\begin{aligned} \pi_{\mathbf{T}}(u_1, \ldots, u_K) &= E\left[\prod_{k=1}^{K} \mathcal{K}_2\left(\frac{u_k - T_k}{\sigma_{T_k}}\right)\right] \\ \pi_{T_k}(u_k) &= E\left[\mathcal{K}_2\left(\frac{u_k - T_k}{\sigma_{T_k}}\right)\right] \end{aligned}$$

et, pour tout u réel,

$$\mathcal{K}_2(u) = \int \mathcal{K}(v)\mathcal{K}(u+v)dv. \tag{2.8}$$

preuve :
Nous donnons les détails de cette démonstration en annexe, B.3, mais nous précisons ici les grandes lignes.

A partir de la définition 2.4.1, nous prenons $\mathcal{K}$ un noyau de carré intégrable. Puis, l'application du théorème de Fubini permet d'inverser les variables d'intégration, ce qui nous permet d'écrire la mesure de dépendance quadratique seulement en fonction de $\mathcal{K}_2$.

■

Remarque 2.5.1 *On remarque, en appliquant la formule de Plancherel-Parseval,*

$$E[\pi_{\mathbf{T}}(\mathbf{T})] = \int \frac{1}{\prod_{k=1}^{K} \sigma_{T_k}} \prod_{k=1}^{K} \left|\frac{\sigma_{T_k}\psi_{\mathcal{K}_i}(\sigma_{T_k}t_k)}{\sqrt{2\pi}}\right|^2 |\psi_{\mathbf{T}}(t_1, \ldots, t_K)|^2 dt_1 \ldots dt_K$$

et

$$E[\pi_{T_k}(T_k)] = \int \frac{1}{\prod_{k=1}^{K} \sigma_{T_k}} \prod_{k=1}^{K} \left|\frac{\sigma_{T_k}\psi_{\mathcal{K}_i}(\sigma_{T_k}t_k)}{\sqrt{2\pi}}\right|^2 |\psi_{T_k}(t_k)|^2 dt_k.$$

Ce lemme nous conduit alors à utiliser la mesure de dépendance quadratique en définissant un noyau $\mathcal{K}_2$. En effet, il est alors clair qu'après le choix du noyau $\mathcal{K}_2$, il ne faudra procéder à aucune intégration multiple. Mais, pour cela nous devons choisir le noyau $\mathcal{K}_2$ de telle sorte que toutes les propriétés de la mesure de dépendance quadratique soient conservées (paragraphe 2.4.2). Pour cela, établissons le lemme suivant,

Lemme 2.5.2 (Lien entre $\mathcal{K}$ et $\mathcal{K}_2$) *Soit $\mathcal{K}$, une fonction de carré sommable et telle que sa transformée de Fourier soit différente de zéro presque partout. Alors, la fonction $\mathcal{K}_2$ définie par,*

$$\mathcal{K}_2(u) = \int \mathcal{K}(v)\mathcal{K}(u+v)dv.$$

est un noyau réel tel que sa transformée de Fourier soit positive, sommable et différente de zéro presque partout. Et réciproquement, si $\mathcal{K}_2$, est un noyau réel tel que sa transformée de Fourier soit positive, sommable et différente de zéro presque partout. Alors, le noyau $\mathcal{K}$ qui vérifie

$$\mathcal{K}_2(u) = \int \mathcal{K}(v)\mathcal{K}(u+v)dv.$$

est de carré sommable et tel que sa transformée de Fourier est différente de zéro presque partout.

preuve :

sens direct :

On suppose $\mathcal{K}$, une fonction de carré sommable et telle que sa transformée de Fourier soit différente de zéro presque partout.
Tout d'abord, on remarque que la transformée de Fourier de $\mathcal{K}_2$ s'écrit,

$$\psi_{\mathcal{K}_2} = |\psi_{\mathcal{K}}|^2$$

On en déduit alors que la transformée de Fourier de $\mathcal{K}_2$ est sommable, positive et différente de zéro presque partout.
Par ailleurs, par définition de l'expression de $\mathcal{K}_2$ (2.8), celui-ci est pair. De plus, comme sa transformée de Fourier est réelle, le noyau $\mathcal{K}_2$ est réel.
En effet,

$$\overline{\mathcal{K}_2(t)} = \overline{\int |\psi_{\mathcal{K}}(x)|^2 e^{itx}dx} = \int |\psi_{\mathcal{K}}(x)|^2 e^{-itx}dx = \mathcal{K}_2(t)$$

réciproque :

On suppose $\mathcal{K}_2$ un noyau réel tel que sa transformée de Fourier soit positive, sommable et différente de zéro presque partout.

Montrons alors qu'il existe un noyau $\mathcal{K}$ de carré intégrable tel que sa transformée de Fourier soit différente de zéro presque partout.
Comme la transformée de Fourier de $\mathcal{K}_2$ est sommable et positive, la fonction $\sqrt{\psi_{\mathcal{K}_2}}e^{i\theta}$, où θ est une fonction réelle et impaire, est de carré sommable. On définit alors $\mathcal{K}$ comme étant la transformée de Fourier inverse de $\sqrt{\psi_{\mathcal{K}_2}}e^{i\theta}$. On vérifie bien que $\mathcal{K}$ est une fonction de carré sommable dont la transformée de Fourier est différente de zéro presque partout.

■

Nous sommes maintenant amenés à envisager différents choix pour le noyau $\mathcal{K}_2$.

2.5.2 Choix des noyaux $\mathcal{K}_2$: Propriétés

Grâce au lemme 2.5.2, il est clair que la mesure de dépendance quadratique associée à des variables aléatoires ne dépend que du choix du noyau $\mathcal{K}_2$. Mais, afin de conserver les propriétés de la mesure de dépendance quadratique, le noyau $\mathcal{K}_2$ doit satisfaire certaines propriétés provenant de celles du noyau $\mathcal{K}$. Intéressons nous à présent au problème du choix du noyau $\mathcal{K}_2$.

Propriété 1 : La mesure Q définie à l'aide du noyau $\mathcal{K}_2$ garde la propriété de caractérisation d'indépendance des variables aléatoires dès que la transformée de Fourier de $\mathcal{K}_2$ est presque partout différente de zéro.
En effet, on remarque que par le même procédé utilisé dans la justification de l'existence de la mesure de dépendance quadratique (c.f. définition 2.4.1),

$$Q(T_1,\ldots,T_K) = \frac{\prod_{k=1}^K \sigma_{T_k}}{(2\pi)^K}\int\prod_{k=1}^K \psi_{\mathcal{K}_2}(\sigma_{T_k}t_k)\left|\psi_T(t_1,\ldots,t_K)-\prod_{k=1}^K\psi_{T_k}(t_k)\right|^2 d\mathbf{t}.$$

Par la même analogie que précédemment, si les fonctions caractéristiques sont analytiques, nous pouvons supposer simplement que la transformée de Fourier de $\mathcal{K}_2$ est non nulle sur un ouvert borné de $\mathbb{R}$ contenant 0.

Propriété 2 : En supposant que le noyau $\mathcal{K}$ est un noyau à densité concentré autour de 0, le noyau $\mathcal{K}_2$ vérifiera les mêmes propriétés. Et nous pouvons donner une interprétation en terme d'entropie de Rényi.

Définition 2.5.1 (Entropie de Rényi) *Soit* $\mathbf{T}$ *un vecteur aléatoire absolument continue, alors on définit l'entropie de Rényi par,*

$$H_{R_\alpha}(\mathbf{T}) = \frac{1}{\alpha-1}\log\int f_{\mathbf{T}}(\mathbf{t})^\alpha$$

où α *est un réel strictement positif différent de 1.*

On remarque que lorsque α tend vers 1, on retrouve la définition de l'entropie de Shannon.

Lemme 2.5.3 *Soit $\widetilde{\mathcal{K}}_2$ un noyau de Parzen-Rosenblatt positif, tel que*

$$\widetilde{\mathcal{K}}_2(u_1,\ldots,u_p)=\prod_j \mathcal{K}_0(u_j),$$

où $\mathcal{K}_0$ est un noyau à densité symétrique univarié satisfaisant $\int v^2\mathcal{K}_0(v)dv=1$. On définit alors, pour h un réel positif,

$$\mathcal{K}_2(u)=\frac{1}{h}\widetilde{\mathcal{K}}_2\left(\frac{u}{h}\right)$$

On suppose de plus que $\mathbf{T}=(T_1,\ldots,T_K)^T$ est de loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^K$. On note $p_{\mathbf{T}}$ et p_{T_k} respectivement, la densité conjointe et les densités marginales associées aux variables $T_1,\ldots,T_K$ de carré sommable.

$$\begin{aligned}\lim_{h\to 0} R(\mathbf{T}) &= \prod_{k=1}^{K}(\sigma_{T_k})\int p^2_{\mathbf{T}}(\mathbf{t})d\mathbf{t}\\ \lim_{h\to 0} R(T_k) &= \sigma_{T_k}\int p^2_{T_k}(t_k)dt_k\end{aligned}$$

où

$$\begin{aligned}R(\mathbf{T}) &= E[\pi_{\mathbf{T}}(\mathbf{T})]\\ R(T_k) &= E[\pi_{T_k}(T_k)]\end{aligned}$$

preuve :
On utilise ici un résultat énoncé, par exemple, dans [38], et on utilise le fait que, pour tout t_k, $\pi_{T_k}(t_k)=\mathcal{K}_2 * p_{T_k/\sigma_{T_k}}(t_k/\sigma_{T_k})$, alors,

$$R(T_k)=\sigma_{T_k}\int p^2_{T_k}(t_k)dt_k+0.5h^2\int p''_{T_k/\sigma_{T_k}}(t_k/\sigma_{T_k})p_{T_k/\sigma_{T_k}}(t_k/\sigma_{T_k})dt_k+o(h^2)$$

On montre le résultat relatif au vecteur aléatoire $\mathbf{T}$ de manière analogue.

■

2.5.3 Exemples : Différents noyaux possibles

En conclusion, les hypothèses sur les noyaux $\mathcal{K}_2$ seront, pour tous k, $1 \leq k \leq K$.

- $\mathcal{K}_2$ est réel.
- la transformée de Fourier de $\mathcal{K}_2$ est positive et sommable.
- la transformée de Fourier de $\mathcal{K}_2$ est presque partout différente de zéro.

Insistons ici aussi sur le fait que dans la définition de la mesure de dépendance quadratique, il est inutile de restreindre le choix du noyau à une classe trop petite. Les seules propriétés que doivent vérifier les noyaux sont citées ci-dessus.

Dans la définition de la mesure de dépendance quadratique 2.4, à la différence des mesures de dépendance basées sur une estimation de la densité, une fois le noyau $\mathcal{K}_2$ choisi, quelque soit la taille de fenêtre h, le noyau $x \mapsto \mathcal{K}_2(x/h)/h$ pourra aussi être utilisé pour définir la mesure de dépendance quadratique. Cependant, pour des raisons d'efficactité du test d'indépendance, et de la nécessité de minimiser la variance de l'estimateur, nous verrons dans la section 5.3 comment, à partir d'un noyau $\mathcal{K}_2$ fixé, choisir la taille de fenêtre adéquate de telle sorte que l'estimateur de la mesure de dépendance quadratique ainsi défini permette la résolution du problème de séparation aveugle de sources.

Présentons ici quelques exemples de noyaux utilisés dans la suite.

On note $\mathcal{K}_2$ le noyau, et $\psi_{\mathcal{K}_2}$ les transformées de Fourier respectives.

1. Le noyau Gaussien : (noyau 1)
   $\mathcal{K}_2(x) = e^{-x^2}$, $\psi_{\mathcal{K}_2}(t) = \sqrt{\pi}e^{-t^2/4}$
2. Le noyau de Cauchy carré : (noyau 2)
   $\mathcal{K}_2(x) = 1/(1+x^2)^2$, $\psi_{\mathcal{K}_2}(t) = \pi(|t|+1)e^{-|t|}$
3. L'opposé de la dérivée seconde du noyau de Cauchy carré : (noyau 3)
   $\mathcal{K}_2(x) = -(20x^2-4)/(1+x^2)^4$,
   $\psi_{\mathcal{K}_2}(t) = 4t^2\pi^3(|t|+1)e^{-|t|}$.

Les représentations graphiques des noyaux précédents avec leurs transformées de Fourier sont exposées dans la figure 2.1.

Dans toute la suite de ce travail, nous choisirons des noyaux sous la forme

$$\mathcal{K}_2(x) = \widetilde{\mathcal{K}}_2(x/h)/h$$

où $\widetilde{\mathcal{K}}_2$ sera choisi parmi les exemples ci-dessus.

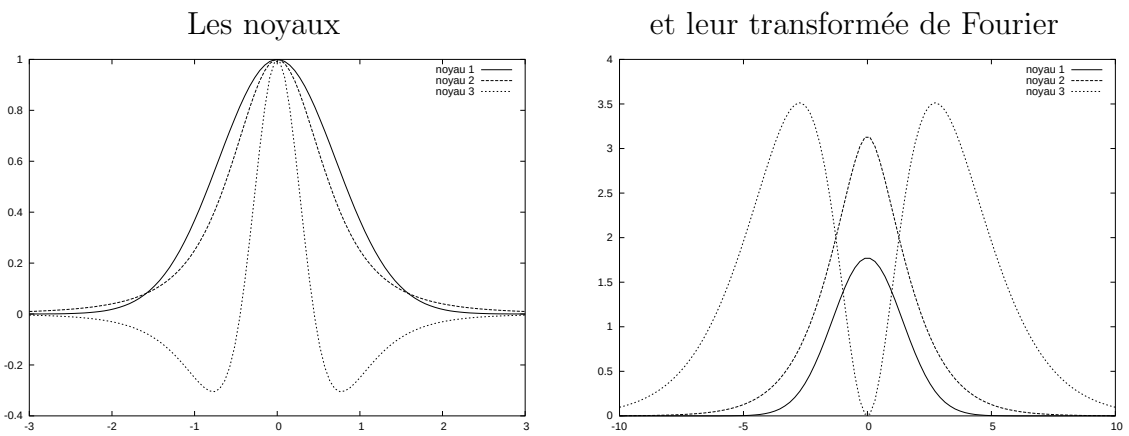


FIG. 2.1 – *Les noyaux 1, 2 et 3 et leurs transformées de Fourier respectives*

2.5.4 Estimation de la mesure de dépendance quadratique

Définition d'un estimateur pour la mesure de dépendance quadratique

Comme nous avons pu le voir dans le lemme 2.5.1, l'estimation de la mesure de dépendance quadratique ne nécessite que l'usage de la moyenne empirique, un estimateur de l'espérance. Nous définirons alors l'estimateur de la mesure de dépendance quadratique de la façon suivante,

Définition 2.5.2 (Estimateur de la dépendance quadratique) *Soient $\mathcal{K}_2$, un noyau réel tel que sa transformée de Fourier soit positive et sommable. Soient K variables aléatoires $T_1, \ldots, T_K$, on note $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$. Pour tout k, $1 \leq k \leq K$, on note $T_k(1), \ldots, T_k(N)$ un échantillon de taille N de la variable aléatoire T_k. Alors l'estimateur de la mesure de dépendance quadratique est défini par,*

$$\widehat{Q}(T_1, \ldots, T_K) = \widehat{E}[\widehat{\pi}_{\mathbf{T}}(\mathbf{T})] + \prod_{k=1}^{K} \widehat{E}[\widehat{\pi}_{T_k}(T_k)] - 2\widehat{E}\left[\prod_{k=1}^{K} \widehat{\pi}_{T_k}(T_k)\right]$$

où $\widehat{E}[\Phi(X)] = \frac{1}{N}\sum_{n=1}^{N} \Phi(X(n))$, pour toute fonction Φ de la variable aléatoire X d'échantillon $X(1), \ldots, X(N)$ et

$$\begin{aligned}
\widehat{\pi}_{\mathbf{T}}(t_1, \ldots, t_K) &= \widehat{E}\left[\prod_{k=1}^{K} \mathcal{K}_2\left(\frac{t_k - T_k}{\widehat{\sigma_{T_k}}}\right)\right] = \frac{1}{N}\sum_{n=1}^{N}\prod_{k=1}^{K} \mathcal{K}_2\left(\frac{t_k - T_k(n)}{\widehat{\sigma_{T_k}}}\right) \\
\widehat{\pi}_{T_k}(T_k) &= \widehat{E}\left[\mathcal{K}_2\left(\frac{t_k - T_k}{\widehat{\sigma_{T_k}}}\right)\right] = \frac{1}{N}\sum_{n=1}^{N} \mathcal{K}_2\left(\frac{t_k - T_k(n)}{\widehat{\sigma_{T_k}}}\right)
\end{aligned}$$

et $\widehat{\sigma_{T_k}}$ est un estimateur de σ_{T_k} ne dépendant que de $T_k(1), \ldots, T_k(N)$ l'echantillon de T_k.

Propriétés de l'estimateur

Propriété 1 : Remarquons que l'estimateur obtenu précédemment peut être déduit de la formule de la mesure de dépendance quadratique en prenant simplement un estimateur des fonctions caractéristiques.

Lemme 2.5.4 (Avec les fonctions caractéristiques estimées) *Soit $\mathcal{K}_2$, un noyau réel tel que sa transformée de Fourier soit positive et sommable. Soient K variables aléatoires $T_1, \ldots, T_K$, on note $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$. Pour tous k, $1 \leq k \leq K$, on note $T_k(1), \ldots, T_k(N)$ un échantillon de taille N de la variable aléatoire T_k. Alors l'estimateur de la mesure de dépendance quadratique s'ecrit aussi,*

$$\widehat{Q}(T_1, \ldots, T_K) = \frac{\prod_{k=1}^{K} \widehat{\sigma_{T_k}}}{(2\pi)^K} \int \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(\widehat{\sigma_{T_k}} t_k) \left| \widehat{\psi}_T(t_1, \ldots, t_K) - \prod_{k=1}^{K} \widehat{\psi}_{T_k}(t_k) \right|^2 dt_1 \ldots dt_K.$$

où

$$\widehat{\psi}_{\mathbf{T}}(t_1, \ldots, t_K) = \widehat{E}\left[\prod_{k=1}^{K} e^{it_k T_k}\right] = \frac{1}{N} \sum_{n=1}^{N} \prod_{k=1}^{K} e^{it_k T_k(n)}$$

$$\widehat{\psi}_{T_k}(T_k) = \widehat{E}\left[e^{it_k T_k}\right] = \frac{1}{N} \sum_{n=1}^{N} e^{it_k T_k(n)}$$

et $\widehat{\sigma_{T_k}}$ est un estimateur de σ_{T_k} ne dépendant que de $T_k(1), \ldots, T_k(N)$ l'echantillon de T_k.

preuve:
Ceci vient tout simplement du développement suivant, pour toute fonction $g(t_1, \ldots, t_k) = \prod_{k=1}^{K} g_k(t_k)$ intégrable,

$$\begin{aligned}
\widehat{Q}(T_1, \ldots, T_K) = & \frac{1}{N^2} \sum_{m=1}^{N} \sum_{n=1}^{N} \prod_{k=1}^{K} \int e^{it_k \left(\frac{T_k(n) - T_k(m)}{\widehat{\sigma_{T_k}}} \right)} g_k(t_k) dt_k \\
& + \prod_{k=1}^{K} \frac{1}{N^2} \sum_{m=1}^{N} \sum_{n=1}^{N} \int e^{it_k \left(\frac{T_k(n) - T_k(m)}{\widehat{\sigma_{T_k}}} \right)} g_k(t_k) dt_k \\
& - 2 \frac{1}{N} \sum_{m=1}^{N} \prod_{k=1}^{K} \frac{1}{N} \sum_{n=1}^{N} \int e^{it_k \left(\frac{T_k(n) - T_k(m)}{\widehat{\sigma_{T_k}}} \right)} g_k(t_k) dt_k
\end{aligned}$$

Pour finir, on remarque que dans notre cas, $g_k(t_k) = \widehat{\sigma_{T_k}}\psi_{\mathcal{K}_2}(\widehat{\sigma_{T_k}}t_k)$, qui est par hypothèse intégrable. On utilise alors le résultat qui précise que la composition de la transformée de Fourier conjugué par la transformée de Fourier est égale à l'opérateur identité.

■

Propriété 2 : L'estimateur défini précédemment vérifie certaines propriétés asymptotiques intéressantes. Présentons maintenant deux lemmes inspirés de la thèse de Kankainen, [44].

Lemme 2.5.5 *Si la fonction de répartition conjointe, $F_{\mathbf{T}}$ du vecteur aléatoire* $\mathbf{T}$ *satisfait la condition suivante,*

$$\int e^{\delta|t|} dF_{\mathbf{T}}(\mathbf{t}) < \infty,$$

pour $\delta > 0$, alors sous l'hypothèse de non indépendance des variables aléatoires $T_1, \ldots, T_K$,

$$\lim_{N\to+\infty} \widehat{Q}(T_1, \ldots, T_K) > 0 \quad p.s.$$

pour tout noyau $\mathcal{K}_2$ de telle sorte que

$$0 < \int \psi_{\mathcal{K}_2}(t)dt < \infty$$

preuve : c.f. [44].

Nous utiliserons plus particulièrement le lemme suivant, qui est aussi directement inspiré des travaux de Kankainen, mais pour lequel nous avons affaibli les hypothèses sur le noyau.

Lemme 2.5.6 (Convergence asymptotique) *Si la transformée de Fourier de $\mathcal{K}_2$ est presque partout différente de zéro, (rappelons qu'elle est déjà positive), Alors sous l'hypothèse de non indépendance des variables aléatoires $T_1, \ldots, T_K$,*

$$\lim_{N\to+\infty} \widehat{Q}(T_1, \ldots, T_K) > 0 \quad p.s.$$

quelle que soit la fonction de répartition de $\mathbf{T}$.

preuve :

Comme les variables $T_1, \ldots, T_K$ ne sont pas indépendantes, on sait que

$$\psi_{\mathbf{T}}(t_1, \ldots, t_K) \neq \prod_{k=1}^{K} \psi_{T_k}(t_k)$$

Alors, il existe un ouvert U borné de mesure strictement positive tel que

$$\inf_{\mathbf{t}\in U} |\psi_{\mathbf{T}}(t_1,\ldots,t_K) - \prod_{k=1}^{K} \psi_{T_k}(t_k)| > 0$$

Notons N l'ensemble de mesure nulle sur lequel s'annule $\prod_{k=1}^{K} \psi_{\mathcal{K}_2}$, alors,

$$\int_{U\setminus N} |\psi_{\mathbf{T}}(t_1,\ldots,t_K) - \prod_{k=1}^{K} \psi_{T_k}(t_k)|^2 \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t} > 0$$

Or, d'après Csörgő, [23],

$$\sup_{\mathbf{t}\in B} \left|\widehat{\psi}_{\mathbf{T}}(\mathbf{t}) - \psi_{\mathbf{T}}(\mathbf{t})\right| \xrightarrow{p.s.} 0, N \to +\infty$$

et

$$\sup_{\mathbf{t}\in B} \Big| \prod_{k=1}^{K} \widehat{\psi}_{T_k}(t_k) - \prod_{k=1}^{K} \psi_{T_k}(t_k)\Big| \xrightarrow{p.s.} 0, N \to +\infty$$

pour tout ensemble B borné, et en particulier pour $B = U$.
Alors, comme les transformées de Fourier des noyaux sont intégrables,

$$\lim_{N\to+\infty} \int_{U\setminus N} \left|\widehat{\psi}_{\mathbf{T}}(\mathbf{t}) - \psi_{\mathbf{T}}(\mathbf{t})\right| \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t} = 0 \quad p.s.$$

et

$$\lim_{N\to+\infty} \int_{U\setminus N} \Big| \prod_{k=1}^{K} \widehat{\psi}_{T_k}(t_k) - \prod_{k=1}^{K} \psi_{T_k}(t_k)\Big| \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t} = 0 \quad p.s.$$

En outre,

$$\begin{aligned}
\int_{U\setminus N} |\widehat{\psi}_{\mathbf{T}}(t_1,\ldots,t_K) - \prod_{k=1}^{K} \widehat{\psi}_{T_k}(t_k)| \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t} \geq \\
\int_{U\setminus N} |\psi_{\mathbf{T}}(t_1,\ldots,t_K) - \prod_{k=1}^{K} \psi_{T_k}(t_k)| \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t} \\
- \int_{U\setminus N} \left|\widehat{\psi}_{\mathbf{T}}(\mathbf{t}) - \psi_{\mathbf{T}}(\mathbf{t})\right| \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t} \\
- \int_{U\setminus N} \Big| \prod_{k=1}^{K} \widehat{\psi}_{T_k}(t_k) - \prod_{k=1}^{K} \psi_{T_k}(t_k)\Big| \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t}
\end{aligned}$$

Finalement, on obtient donc,

$$\liminf_{N\to+\infty} \int_{U\setminus N} |\widehat{\psi}_{\mathbf{T}}(t_1,\ldots,t_K) - \prod_{k=1}^{K} \widehat{\psi}_{T_k}(t_k)| \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t} > 0$$

Enfin, d'après l'inégalité de Cauchy-Schwarz, on en déduit,

$$\liminf_{N\to+\infty} \int_{U\setminus N} |\widehat{\psi}_{\mathbf{T}}(t_1,\ldots,t_K) - \prod_{k=1}^{K} \widehat{\psi}_{T_k}(t_k)|^2 \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t} > 0$$

En effet,

$$\begin{aligned}
&\int_{U\setminus N} |\widehat{\psi}_{\mathbf{T}}(t_1,\ldots,t_K) - \prod_{k=1}^{K} \widehat{\psi}_{T_k}(t_k)|^2 \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t} \\
&\geq \frac{\int_{U\setminus N} |\widehat{\psi}_{\mathbf{T}}(t_1,\ldots,t_K) - \prod_{k=1}^{K} \widehat{\psi}_{T_k}(t_k)| \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t}}{\int_{U\setminus N} \prod_{k=1}^{K} \psi_{\mathcal{K}_2}(t_k) d\mathbf{t}}
\end{aligned}$$

2.6 Conclusion

De par la nécessité de construire un critère de séparation, nous nous sommes intéressés à deux mesures de dépendance, l'une basée sur la comparaison des densités, l'information mutuelle, et l'autre basée sur l'utilisation des fonctions caractéristiques, la mesure de dépendance quadratique.

Ces deux critères possédant des propriétés différentes, des estimations différentes sont envisagées.

Ayant maintenant à notre disposition deux mesures permettant de décider si des variables sont indépendantes ou non, dans le cadre de l'analyse en composantes indépendantes et de la séparation aveugle de sources, nous nous intéressons à diverses méthodes d'optimisation.

Chapitre 3

Optimisation

3.1 Introduction

Dans le chapitre précédent, nous avons défini deux mesures de dépendance qui ont la propriété d'être nulles si et seulement si les variables considérées sont indépendantes. Dans ce contexte, afin de simplement rechercher des composantes indépendantes ou de résoudre le problème de séparation aveugle de sources, il faut chercher, respectivement, des variables aléatoires indépendantes ou une structure de séparation adéquate qui réalise le minimum du critère de séparation considéré. Dans notre cas, nous considérons que le mélange est inversible, il est donc équivalent de chercher la structure de séparation ou les variables aléatoires indépendantes. Nous nous plaçons dans le contexte où la méthode d'analyse en composantes indépendantes permet de résoudre le problème de séparation aveugle de sources.

Pour cela, nous constatons que l'on peut envisager deux stratégies différentes qui vont conduire à deux méthodes distinctes.

(i) «**Estimer ensuite**» La première approche consiste simplement à faire le calcul du gradient relatif à partir du critère de séparation théorique, puis à procéder à la minimisation de celui-ci en faisant une estimation des différentes composantes du gradient relatif. Cette approche correspond à celle étudiée par Cardoso [17] dans le cas d'un mélange linéaire, ou bien par Taleb [69] ou encore par Babaie-Zadeh [6] dans le cadre d'un mélange post non linéaire.

(ii) «**Estimer d'abord**» La nouvelle approche que nous avons envisagée correspond à faire tout d'abord une estimation du critère de séparation, puis nous calculons le gradient relatif pour en déduire une méthode de minimisation. Nous notons que dans [57] Pham avait déjà fait une remarque allant dans ce sens, en constatant que dans le cas d'un mélange linéaire, dans une certaine mesure, l'estimation du gradient théorique coïncide avec le gradient de l'estimateur du critère de séparation.

Rappelons tout d'abord le contexte dans lequel nous nous plaçons.

Nous disposons d'observations $\mathbf{X} = (X_1, \ldots, X_K)^T$, où $\mathbf{X}$ est un vecteur aléatoire. Dans ces conditions, nous cherchons à retrouver une transformation h inversible, telle que les composantes $Y_1, \ldots, Y_K$ du vecteur aléatoire $\mathbf{Y} = h(\mathbf{X})$ soient indépendantes.

Dans notre approche, h sera définie comme réalisant le minimum d'un des deux critères de séparation défini dans le chapitre précédent : l'information mutuelle ou la mesure de dépendance quadratique.

Afin de rechercher ce minimum, nous allons utiliser comme méthode d'optimisation une méthode de descente du gradient. Ce qui nous conduit à calculer pour chacun des deux critères, leur développement limité. Par analogie au cas linéaire, nous reprenons ici la notion de gradient relatif introduite par Cardoso et Laheld [17]. Celle-ci consiste simplement à considérer une variation relative aux variables considérées. Dans le cadre d'un mélange linéaire, ceci permet de construire des méthodes de minimisation ayant des performances invariantes par rapport à la matrice de mélange. Nous verrons que ce n'est pas forcément le cas dans un mélange non linéaire.

Dans un premier temps, nous allons décrire la stratégie (i), «Estimer ensuite» par rapport aux deux mesures de dépendance, l'information mutuelle (paragraphe 3.2.1) et la dépendance quadratique (paragraphe 3.2.2). Nous verrons que cette approche consiste en fait à la résolution d'équations appelées équations d'estimation. Dans un deuxième temps, nous verrons comment estimer les deux mesures de dépendance envisagées, et comment procéder à leur minimisation, c'est la stratégie (ii) «Estimer d'abord» (paragraphes 3.2.3 et 3.2.4). Dans le cas de l'information mutuelle, nous observons que les expressions des gradients obtenues avec les deux stratégies diffèrent par le choix des estimateurs des fonctions scores. Mais, avec la mesure de dépendance quadratique, les expressions des gradients obtenues pour les deux stratégies coïncident. Pour finir, dans la section 3.3, nous ferons quelques commentaires sur les méthodes d'optimisation à utiliser.

3.2 Méthode de descente du gradient

Dans cette partie, nous allons expliquer plus précisément la méthode de minimisation appelée méthode de descente du gradient dans notre cas particulier des deux mesures de dépendance, l'information mutuelle et la dépendance quadratique.

Nous avons remarqué que la résolution de notre problème consiste à chercher le minimum d'une des deux mesures de dépendance en déterminant une transformation inversible qui rend les composantes du vecteur aléatoire de sortie indépendantes.

3.2.1 «Estimer ensuite» avec l'information mutuelle

Nous avons pu voir qu'avec l'information mutuelle, on dispose d'une mesure de dépendance qui en la minimisant rend les variables le plus indépendantes possibles. Afin de procéder à la minimisation de l'information mutuelle, calculons à présent le développement limité de l'information mutuelle. Dans la thèse de Babaie-Zadeh [6], il est explicité le développement au premier ordre. Afin d'établir le développement au second ordre, énonçons alors les lemmes préliminaires suivants,

Lemmes préliminaires

Seule la démonstration du lemme 3.2.1 est explicité en annexe B.1. Les autres ne sont pas détaillées ici, ce sont simplement des extensions au cas multidimensionnels des résultats obtenus dans [1].

Tout d'abord, explicitons le développement concernant la variation de la densité,

Lemme 3.2.1 *Soient $T_1, T_2, \ldots, T_K$, K variables aléatoires réelles telles que : le vecteur aléatoire $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ est de loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^K$. On note $p_{T_1,T_2,\ldots,T_K}$ la densité de celui-ci. (Par conséquent, pour tout i, $1 \leq i \leq K$, la variable aléatoire T_i est de loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$. On note p_{T_i} sa densité.)*

Soit $\boldsymbol{\Delta}_\eta$ une famille de fonctions inversibles, on note $\boldsymbol{\Delta}_\eta = (\Delta_1, \ldots, \Delta_K)^T$, telle que pour tout η, $h + \boldsymbol{\Delta}_\eta \circ h$ est inversible et $\boldsymbol{\Delta}_\eta$ converge simplement[1] vers 0. (La variation relative par rapport à h correspond à l'utilisation du gradient relatif)

Alors, pour tout $\mathbf{t} \in \mathbb{R}^K$,

$$\begin{aligned} p_{\mathbf{T}+\boldsymbol{\Delta}_\eta(\mathbf{T})}(\mathbf{t}) - p_{\mathbf{T}}(\mathbf{t}) &= -\sum_{i=1}^{K} \partial_i(\Delta_i(\mathbf{t}) p_{\mathbf{T}}(\mathbf{t})) \\ &+ \frac{1}{2}\sum_{i,j=1}^{K} \partial_i\partial_j\left(\Delta_i(\mathbf{t})\Delta_j(\mathbf{t}) p_{\mathbf{T}}(\mathbf{t})\right) + o(\eta^2) \end{aligned}$$

et

$$\begin{aligned} p_{T_i+\Delta_i(\mathbf{T})}(t) - p_{T_i}(t) &= -(E[\Delta_i(\mathbf{T})|T_i = t]p_{T_i}(t))' \\ &+ \frac{1}{2}(E[\Delta_i^2(\mathbf{T})|T_i = t]p_{T_i}(t)) + o(\eta^2) \end{aligned}$$

où ∂_i indique la dérivée par rapport à la i-ième variable.

1. $\forall x \in \mathbb{R}^K, \lim_{\eta \to 0} \boldsymbol{\Delta}_\eta(x) = 0$

preuve : c.f. annexe B.1
On en déduit alors le résultat concernant le développement de l'entropie,

Lemme 3.2.2 *Soient $T_1, T_2, \ldots, T_K$, K variables aléatoires réelles telles que :*
le vecteur aléatoire $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ est de loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^K$. Soit $\mathbf{\Delta}_\eta$ une famille de fonctions inversibles, on note $\mathbf{\Delta}_\eta = (\Delta_1, \ldots, \Delta_K)^T$, telle que pour tout η, $h + \mathbf{\Delta}_\eta \circ h$ est inversible et $\mathbf{\Delta}_\eta$ converge simplement vers 0.

Alors,

$$\begin{aligned} H(\mathbf{T} + \mathbf{\Delta}_\eta(\mathbf{T})) - H(\mathbf{T}) &= -\sum_{i=1}^{K} E\left[\partial_i(\log p_{\mathbf{T}})(\mathbf{T})\Delta_i(\mathbf{T})\right] \\ &- \frac{1}{2}\sum_{i,j=1}^{K} E[\partial_i\Delta_i(\mathbf{T})\partial_j\Delta_j(\mathbf{T})] \\ &- \sum_{i,j=1}^{K} E\left[\Delta_i(\mathbf{T})\partial_j\Delta_j(\mathbf{T})\frac{\partial_i p_{\mathbf{T}}(\mathbf{T})}{p_{\mathbf{T}}(\mathbf{T})}\right] \\ &- \frac{1}{2}\sum_{i,j=1}^{K} E\left[\Delta_i(\mathbf{T})\Delta_j(\mathbf{T})\frac{\partial^2_{ij} p_{\mathbf{T}}(\mathbf{T})}{p_{\mathbf{T}}(\mathbf{T})}\right] + o(\eta^2) \end{aligned}$$

et

$$\begin{aligned} H(T_i + \Delta_i(\mathbf{T})) - H(T_i) &= -\sum_{i=1}^{K} E\left[(\log p_{T_i})'(t)\Delta_i(\mathbf{T})\right] \\ &+ \frac{1}{2}E[var(\Delta_i(\mathbf{T})|T_i)(\log p_{T_i})''(T_i) - (E[\Delta_i(\mathbf{T})|T_i])'^2] \\ &+ o(\eta^{\mathbf{2}}) \end{aligned}$$

où ∂^2_{ij} indique la dérivée seconde par rapport aux variables i et j et
$var(\Delta_i(\mathbf{T})|T_i) = E[\Delta_i^2(\mathbf{T})|T_i] - E[\Delta_i(\mathbf{T})|T_i]^2$.

Gradient de l'information mutuelle

Lemme 3.2.3 (Développement limité de l'information mutuelle) *Soient $T_1, T_2, \ldots, T_K$, K variables aléatoires réelles telles que :*
le vecteur aléatoire $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ est de loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^K$. On note $p_{T_1,T_2,\ldots,T_K}$ la densité de celui-ci.
(Par conséquent, pour tout i, $1 \leq i \leq K$, la variable aléatoire T_i est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$. On note p_{T_i} sa densité.) Soit

$\boldsymbol{\Delta}_\eta$ une famille de fonctions inversibles, on note $\boldsymbol{\Delta}_\eta = (\Delta_1, \ldots, \Delta_K)^T$, telle que pour tout η, $h + \boldsymbol{\Delta}_\eta \circ h$ est inversible et $\boldsymbol{\Delta}_\eta$ converge simplement vers 0.
Alors,

$$\begin{aligned} I(\mathbf{T} + \boldsymbol{\Delta}_\eta(\mathbf{T})) &= I(\mathbf{T}) + E\{\boldsymbol{\Delta}_\eta^T(\mathbf{T})\beta_{\mathbf{T}}(\mathbf{T})\} \\ &+ \frac{1}{2}E[var(\Delta_i(\mathbf{T})|T_i)(\log p_{T_i})''(T_i) - (E[\Delta_i(\mathbf{T})|T_i])'^2] \\ &- \frac{1}{2}\sum_{i,j=1}^{K} E[\partial_i\Delta_i(\mathbf{T})\partial_j\Delta_j(\mathbf{T})] \\ &+ \sum_{i,j=1}^{K} E\left[\Delta_i(\mathbf{T})\partial_j\Delta_j(\mathbf{T})\phi_i(\mathbf{T})\right] \\ &- \frac{1}{2}\sum_{i,j=1}^{K} E\left[\Delta_i(\mathbf{T})\Delta_j(\mathbf{T})\frac{\partial_i\partial_j p_{\mathbf{T}}(\mathbf{T})}{p_{\mathbf{T}}(\mathbf{T})}\right] \\ &+ o(\boldsymbol{\eta}^{\mathbf{2}}) \end{aligned}$$

où

$$var[\Delta(\mathbf{T})|T] = E[\Delta^2(\mathbf{T})|T] - (E[\Delta(\mathbf{T})|T])^2$$

et

$$\beta_{\mathbf{T}}(\mathbf{t}) = \psi_{\mathbf{T}}(\mathbf{t}) - \phi_{\mathbf{T}}(\mathbf{t}),$$

et

$$\psi_{\mathbf{T}}(\mathbf{t}) = (\psi_1(t_1), \ldots, \psi_K(t_K))$$

avec

$$\psi_i(t_i) = -\frac{p'_{T_i}(t_i)}{p_{T_i}(t_i)},$$

(appelée la fonction score marginale par Babaie-Zadeh.)

$$\phi_{\mathbf{T}}(\mathbf{t}) = (\phi_1(\mathbf{t}), \ldots, \phi_K(\mathbf{t}))$$

avec

$$\phi_i(\mathbf{t}) = -\partial_i \log p_{\mathbf{T}}(\mathbf{t})$$

(appelée la fonction score jointe par Babaie-Zadeh.)

Propriétés

Babaie-Zadeh en déduit alors la propriété suivante,

Lemme 3.2.4 *Soient $T_1, T_2, \ldots, T_K$, K variables aléatoires réelles telles que : le vecteur aléatoire $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ est de loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^K$.*

Les variables $T_1, T_2, \ldots, T_K$ sont indépendantes si et seulement si $\beta_{\mathbf{T}}(\mathbf{t}) = 0$ pour tout $\mathbf{t} \in \mathbb{R}^K$

preuve : c.f. [6, p.38]

Nous en déduisons alors l'écriture du gradient relatif et du Hessien de l'information mutuelle.

Lemme 3.2.5 *Soient $T_1, T_2, \ldots, T_K$, K variables aléatoires réelles telles que : le vecteur aléatoire $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ est de loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^K$.*

Le gradient relatif de l'information mutuelle s'écrit alors,

$$\Delta_1, \ldots, \Delta_K \mapsto E\{\boldsymbol{\Delta}^T(\mathbf{T})\beta_{\mathbf{T}}(\mathbf{T})\}$$

Et le Hessien de l'information mutuelle s'écrit,

$$\begin{aligned}
\Delta_1, \ldots, \Delta_K \;\mapsto\; & \frac{1}{2}E[var(\Delta_i(\mathbf{T})|T_i)(\log p_{T_i})''(T_i) - (E[\Delta_i(\mathbf{T})|T_i])'^2] \\
- \; & \frac{1}{2}\sum_{i,j=1}^{K} E[\partial_i\Delta_i(\mathbf{T})\partial_j\Delta_j(\mathbf{T})] \\
+ \; & \sum_{i,j=1}^{K} E\left[\Delta_i(\mathbf{T})\partial_j\Delta_j(\mathbf{T})\phi_i(\mathbf{T})\right] \\
- \; & \frac{1}{2}\sum_{i,j=1}^{K} E\left[\Delta_i(\mathbf{T})\Delta_j(\mathbf{T})\frac{\partial_i\partial_j p_{\mathbf{T}}(\mathbf{T})}{p_{\mathbf{T}}(\mathbf{T})}\right]
\end{aligned}$$

Afin de pouvoir implémenter notre méthode, il est indispensable de procéder à une estimation de ce gradient. Ici, il est nécessaire d'estimer une densité conjointe. Nous avons déjà évoqué les problèmes que cela entraîne en trop grande dimension. Cependant ici, l'estimation de la densité porte à la fois sur deux quantités dont on effectue la différence. Babaie-Zadeh fait alors la remarque suivante, si on estime séparément la fonction score jointe et les fonctions scores marginales, les erreurs commises sur les deux estimateurs ne vont pas forcément se compenser. Il montre alors le résultat suivant,

Lemme 3.2.6 *Soient $T_1, T_2, \ldots, T_K$, K variables aléatoires réelles telles que : le vecteur aléatoire $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ est de loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^K$.*

Pour tout i, $1 \leq i \leq K$, et pour tout t réel,

$$\psi_i(t) = E[\phi_{\mathbf{T}}(\mathbf{T})|T_i = t] \tag{3.1}$$

Babaie-Zadeh suggère alors de ne pas estimer les fonctions scores marginales indépendamment des fonctions scores jointes, mais d'utiliser la relation ci-dessus afin de déduire une estimation des fonctions scores marginales à partir des fonctions scores jointes. On peut expliquer l'observation de Babaie-Zadeh de la manière suivante,

Considérons un estimateur de ψ_i, $\widehat{\psi_i}$, défini à l'aide de $\widehat{\phi_i}$ par,

$$\widehat{\psi_i}(x) = \frac{\int \widehat{\phi_i}(\mathbf{x})\widehat{p}_{\mathbf{T}}(\mathbf{x})dx_1 \dots dx_{i-1}dx_{i+1} \dots dx_K}{\widehat{p}_{T_i}(x_i)}$$

où $\widehat{p}_{\mathbf{T}}$ et $\widehat{p}_{T_i}$ sont les estimateurs à noyaux de la densité (c.f. annexe D).

On remarque que $\widehat{p}_{\mathbf{T}}$ vérifie,

$$\widehat{p}_{T_i}(x) = \int \widehat{p}_{\mathbf{T}}(\mathbf{x})dx_1 \dots dx_{i-1}dx_{i+1} \dots dx_K$$

Il est alors clair que si $\widehat{\phi_i}$ présente un biais, celui ci sera aussi présent dans l'estimation de $\widehat{\psi_i}$.

De plus, on remarque qu'en utilisant les estimateurs à noyaux pour la fonction score conjointe et les fonctions scores marginales, ceux-ci vérifient une estimation de l'équation 3.1.

En effet, pour le montrer considérons l'estimateur à noyaux de ϕ_i donné par,

Définition 3.2.1 *Pour tout $i = 1, \dots, K$, on note $T_i(1), \dots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.*

Alors, pour tout $i = 1, \dots, K$, un estimateur à noyaux de ϕ_i sera défini par,

$$\widehat{\phi_i}(\mathbf{x}) = -\frac{\sum_{j=1}^{N} \mathcal{K}'\left(\frac{x_i - T_i(j)}{\widehat{\sigma}_{T_i}}\right) \prod_{k \neq i} \mathcal{K}\left(\frac{x_k - T_k(j)}{\widehat{\sigma}_{T_k}}\right)}{\widehat{\sigma}_{T_i} \sum_{j=1}^{N} \prod_{k=1}^{K} \mathcal{K}\left(\frac{x_k - T_k(j)}{\widehat{\sigma}_{T_k}}\right)}$$

où pour tout $k = 1, \dots, K$, $\widehat{\sigma}_{T_k}$ est un estimateur de la variance de T_k.

Nous en déduisons alors le résultat suivant pour l'estimateur de ψ_i

Lemme 3.2.7 *Pour tout $i = 1, \dots, K$, on note $T_i(1), \dots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.*

Soit $\mathcal{K}$ un noyau intégrable, d'intégrale égale à 1.

Alors, pour tout $i = 1, \dots, K$, un estimateur à noyaux de ψ_i défini par,

$$\widehat{\psi_i}(x) = -\frac{\sum_{j=1}^{N} \mathcal{K}'\left(\frac{x - T_i(j)}{\widehat{\sigma}_{T_i}}\right)}{\widehat{\sigma}_{T_i} \sum_{j=1}^{N} \mathcal{K}\left(\frac{x - T_i(j)}{\widehat{\sigma}_{T_i}}\right)}$$

vérifie,

$$\widehat{\psi}_i(x) = \frac{\int \widehat{\phi}_i(\mathbf{x})\widehat{p}_{\mathbf{T}}(\mathbf{x})dx_1 \dots dx_{i-1}dx_{i+1} \dots dx_K}{\widehat{p}_{T_i}(x_i)}$$

où $\widehat{p}_{\mathbf{T}}$ et $\widehat{p}_{T_i}$ sont les estimateurs à noyaux de la densité.

La méthode de minimisation obtenue ici se réduit en fait à la résolution d'une équation (2.4) qui est équivalente à l'indépendance des variables. Dans le paragraphe 4.4.8, nous explicitons un autre avantage de tenir compte de la relation 3.1 entre les estimations de ϕ_i et ψ_i.

3.2.2 «Estimer ensuite» avec la mesure de dépendance quadratique

La mesure de dépendance quadratique, au même titre que l'information mutuelle fournit un critère de dépendance. Grâce à la minimisation de cette mesure, il est possible de reconstituer des variables indépendantes. Par le même procédé que pour l'information mutuelle, nous allons calculer le développement limité de cette mesure de dépendance quadratique afin d'en déduire l'expression du gradient et du Hessien.

Lemmes préliminaires

Comme la mesure de dépendance quadratique ne dépend que des variables normées, T_k/σ_k, regardons tout d'abord la variation de ces quantités.

Lemme 3.2.8 *Soient $T_1, T_2, \dots, T_K$, K variables aléatoires réelles. On note le vecteur aléatoire $\mathbf{T} = (T_1, T_2, \dots, T_K)^T$. Soit $\mathbf{\Delta}_\eta$ une famille de fonctions inversibles, on note $\mathbf{\Delta}_\eta = (\Delta_1, \dots, \Delta_K)^T$, telle que $h + \mathbf{\Delta}_\eta \circ h$ est inversible et $\mathbf{\Delta}_\eta$ converge simplement vers 0.*

Alors, pour tout $k = 1, \dots, K$,

$$\frac{T_k + \Delta_k(\mathbf{T})}{\sigma_{T_k + \Delta_k(\mathbf{T})}} - \frac{T_k}{\sigma_{T_k}} = \frac{\Delta_k(\mathbf{T})}{\sigma_{T_k}} - \frac{T_k}{\sigma_{T_k}} E[\phi_{T_k}(T_k)\Delta_k(\mathbf{T})] + o(\eta)$$

où ϕ_{T_k} est le gradient de la fonction $\log \sigma_{T_k}$, défini de la manière suivante,
$\log \sigma_{T_k + \Delta_k(\mathbf{T})} = \log \sigma_{T_k} + E[\phi_{T_k}(T_k)\Delta_k(\mathbf{T})] + o(\eta)$

Remarque 3.2.1 *Si on prend par exemple σ_{T_k} comme l'écart-type de T_k, alors $\phi_{T_k}(y) = (y - E[T_k])/\sigma^2_{T_k}$.*

preuve :

En effet, par définition de l'écart-type, $\sigma_{T_k} = \sqrt{E[(T_k - E[T_k])^2]}$, nous choisissons alors par commodité de calculer la variation du $\log \sigma_{T_k}$.

Par un développement classique,

$$\sigma^2_{T_k+\Delta_k(\mathbf{T})} = \sigma^2_{T_k} + 2E[\Delta_k(\mathbf{T})(T_k - E[T_k])] + E[(\Delta_k(\mathbf{T}) - E[\Delta_k(\mathbf{T})])^2]$$

D'où, en prenant le log de chaque côté de l'expression, on obtient,

$$2\log\sigma_{T_k+\Delta_k(\mathbf{T})} = 2\log\sigma_{T_k} + \frac{2}{\sigma^2_{T_k}}E[\Delta_k(\mathbf{T})(T_k - E[T_k])] + o(\eta)$$

Alors, en divisant par -2 on obtient,

$$-\log\sigma_{T_k+\Delta_k(\mathbf{T})} = -\log\sigma_{T_k} - \frac{1}{\sigma^2_{T_k}}E[\Delta_k(\mathbf{T})(T_k - E[T_k])] + o(\eta)$$

Par conséquent le développement à l'ordre 1 dans le cas de l'écart-type s'écrit,

$$\frac{1}{\sigma_{T_k+\Delta_k(\mathbf{T})}} = \frac{1}{\sigma_{T_k}} - \frac{1}{\sigma_{T_k}}E[\phi_k(\mathbf{T})\Delta_k(\mathbf{T})] + o(\eta)$$

Pour la preuve du lemme, il suffit de remarquer que

$$\frac{1}{\sigma_{T_k+\Delta_k(\mathbf{T})}} = \frac{1}{\sigma_{T_k}} - \frac{1}{\sigma_{T_k}}E[\phi_k(\mathbf{T})\Delta_k(\mathbf{T})] + o(\eta)$$

■

Gradient de la mesure de dépendance quadratique

Nous en déduisons alors le développement de la dépendance quadratique suivant,

Lemme 3.2.9 (Développement limité de la dépendance quadratique)
Soient $T_1, \ldots, T_K$ des variables aléatoires telles que $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$. Soient $\mathcal{K}$ un noyau de carré intégrable, tel que sa transformée de Fourier soit différente de zéro presque partout. On suppose de plus que $\mathcal{K}$ est un noyau deux fois dérivable et à support compact.
Soit $\boldsymbol{\Delta}_\eta$ une famille de fonctions inversibles, on note $\boldsymbol{\Delta}_\eta = (\Delta_1, \ldots, \Delta_K)^T$, telle que pour tout η, $h + \boldsymbol{\Delta}_\eta \circ h$ est inversible et $\boldsymbol{\Delta}_\eta$ converge simplement vers 0.
Alors,

$$Q(\mathbf{T} + \boldsymbol{\Delta}_\eta(\mathbf{T})) - Q(\mathbf{T}) = E\left[\sum_{k=1}^{K} G^*_k(\mathbf{T})\Delta_k(\mathbf{T})\right] + o(\eta)$$

où pour $\mathbf{t} \in \mathbb{R}^K$,

$$G_k^*(\mathbf{t}) = G_k(\mathbf{t}) - E[T_k G_k(\mathbf{T})]\phi_k(t_k),$$

$$\begin{aligned} G_k(\mathbf{t}) &= \pi_{k,\mathbf{T}}(t_1,\ldots,t_k) - \pi'_{T_k}(t_k)\prod_{l\neq k}\pi_{T_l}(t_l) + \pi'_{T_k}(t_k)\prod_{l\neq k}E[\pi_{T_l}(T_l)] \\ &\quad - E\left[\frac{1}{\sigma_{T_k}}\mathcal{K}_2'\left(\frac{t_k - T_k}{\sigma_{T_k}}\right)\prod_{l\neq k}\pi_{T_l}(T_l)\right]. \end{aligned}$$

et $\pi_{\mathbf{T}}(\mathbf{z}) = E[\prod_{l=1}^K \mathcal{K}_2((z_l - T_l)/\sigma_{T_l})]$, $\pi_{k,\mathbf{T}}$ *désigne sa dérivée partielle par rapport à la k-ième composante et* $\pi_{T_k}(z_k) = E[\mathcal{K}_2((z_k - T_k)/\sigma_{T_k})]$ *et* π'_{T_k} *désigne sa dérivée.*

On remarque que l'hypothèse restrictive sur le support des noyaux peut être remplacée par l'hypothèse suivante sur la famille de fonction Δ_η :

$$\int\int \mathcal{K}''(x)\Delta_\eta^2(x)dF_{\mathbf{T}}(x) = o(\eta).$$

preuve:

On remarque que l'on peut écrire le gradient $G_k(\mathbf{t})$ de la manière suivante,

$$\begin{aligned} G_k(\mathbf{t}) = -\frac{1}{\sigma_{T_k}}\int \quad & \left\{\mathcal{K}'\left(z_k - \frac{t_k}{\sigma_{T_k}}\right)\left[\prod_{l\neq k}\mathcal{K}\left(z_l - \frac{t_l}{\sigma_{T_l}}\right) - \prod_{l\neq k}E\mathcal{K}\left(z_l - \frac{T_l}{\sigma_{T_l}}\right)\right]\right\} \\ & \left\{E\left[\prod_{k=1}^K \mathcal{K}\left(z_l - \frac{T_l}{\sigma_{T_l}}\right)\right] - \prod_{k=1}^K E\mathcal{K}\left(z_l - \frac{T_l}{\sigma_{T_l}}\right)\right\} dz_1\ldots dz_K \end{aligned}$$

Comme cette formule provient clairement de la différentiation du critère Q écrit comme dans la définition 2.4.1, nous en déduisons le résultat voulu.

■

On en déduit alors les formules du gradient relatif.

Lemme 3.2.10 *Soient* $T_1,\ldots,T_K$ *des variables aléatoires et* $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$. *Soit* $\mathcal{K}$ *un noyau de carré intégrable, tel que sa transformée de Fourier soit différente de zéro presque partout.*

Expression du gradient,

$$\boldsymbol{\Delta} = (\Delta_1,\ldots,\Delta_K)^T \quad \mapsto \quad E\left[\sum_{k=1}^K G_k^*(\mathbf{T})\Delta_k(\mathbf{T})\right]$$

Comme pour l'information mutuelle, nous pouvons faire la remarque suivante,

Lemme 3.2.11 *Soient $T_1, \ldots, T_K$ des variables aléatoires et $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$. Soit $\mathcal{K}_2$ un noyau intégrable et de dérivée intégrable, tel que sa transformée de Fourier soit positive, sommable et différente de zéro presque partout.*

Alors, $T_1, \ldots, T_K$ sont indépendantes si et seulement si, pour tout $k = 1, \ldots, K$, et pour tout $\mathbf{t} \in \mathbb{R}^K$ $G_k^*(\mathbf{t}) = 0$.

preuve:

En effet, si les variables sont indépendantes, il est clair que pour tout $k = 1, \ldots, K$, et pour tout $\mathbf{t} \in \mathbb{R}^K$ $G_k(\mathbf{t}) = 0$.

Intéressons nous alors à l'autre implication. Supposons que pour tout $k = 1, \ldots, K$, et pour tout $\mathbf{t} \in \mathbb{R}^K$ $G_k^*(\mathbf{t}) = 0$. Alors, $G_k(\mathbf{t}) = E[T_k G_k(\mathbf{T})]\phi_k(t_k)$. En dérivant les deux membres de l'équation par rapport à la variable t_l, où $l \neq k$, on obtient,

pour tout $k = 1, \ldots, K$, et pour tout $\mathbf{t} \in \mathbb{R}^K$ $\partial_l G_k(\mathbf{t}) = 0$.

Or ceci est équivalent à dire, d'après l'expression de G_k,

$$
\begin{aligned}
& E\left[\mathcal{K}_2'\left(\frac{t_l - T_l}{\sigma_{T_l}}\right)\mathcal{K}_2'\left(\frac{t_k - T_k}{\sigma_{T_k}}\right)\prod_{h=1,h\neq k,l}^{K}\mathcal{K}_2\left(\frac{t_h - T_h}{\sigma_{T_h}}\right)\right] \\
&= E\left[\mathcal{K}_2'\left(\frac{t_l - T_l}{\sigma_{T_l}}\right)\right] E\left[\mathcal{K}_2'\left(\frac{t_k - T_k}{\sigma_{T_k}}\right)\right]\prod_{h=1,h\neq k,l}^{K} E\left[\mathcal{K}_2\left(\frac{t_h - T_h}{\sigma_{T_h}}\right)\right]
\end{aligned}
$$

En supposant certaines propriétés d'intégrabilité sur les dérivées des noyaux, ceci indique donc bien que les variables sont indépendantes.

■

Enonçons à présent une autre propriété du gradient,

Remarque 3.2.2 *On remarque que, $E[G_k(\mathbf{T})] = 0$. Ceci se déduit du développement limité précédent en remarquant que la mesure de dépendance quadratique est invariante par translation. On peut le montrer aussi en remarquant que les espérances des dérivées de $\pi_{\mathbf{T}}$ et de π_{T_k} sont clairement nulles. Et les espérances des deux autres termes sont égales.*

Estimation des expressions du gradient

L'expression du gradient n'étant pas exploitable algorithmiquement sous cette forme, nous devons tout d'abord proposer une estimation de celui-ci. De la même

manière que pour l'estimation de la dépendance quadratique, nous avons simplement besoin d'utiliser l'estimateur de l'espérance empirique.

Définition 3.2.2 *Pour tout $i = 1, \ldots, K$, on note $T_i(1), \ldots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.*

Un estimateur du gradient s'écrit donc sous la forme,

$$\begin{aligned}\widehat{G}_k(\mathbf{y}) &= \widehat{\pi}'_{k,\mathbf{Y}}(\mathbf{y}) - \widehat{\pi}'_{Y_k}(y_k)\prod_{l\neq k}\widehat{\pi}_{Y_l}(y_l) + \widehat{\pi}'_{Y_k}(y_k)\prod_{l\neq k}\widehat{E}(\widehat{\pi}_{Y_l}(Y_l)) \\ &\quad -\frac{1}{\widehat{\sigma}_{Y_k}}\widehat{E}\left[\mathcal{K}'_2\left(\frac{y_k - Y_k}{\widehat{\sigma}_{Y_k}}\right)\prod_{l\neq k}\widehat{E}(\widehat{\pi}_{Y_l}(Y_l))\right]\end{aligned}$$

avec

$$\widehat{\pi}'_{k,\mathbf{Y}}(\mathbf{y}) = \frac{1}{N\widehat{\sigma}_{Y_k}}\sum_{n=1}^{N}\mathcal{K}'_2\left(\frac{y_k - Y_k(n)}{\widehat{\sigma}_{Y_k}}\right)\prod_{l\neq k}\mathcal{K}_2\left(\frac{y_l - Y_l(n)}{\widehat{\sigma}_{Y_l}}\right),$$

la k-ième composante du gradient de $\widehat{\pi}_{\mathbf{Y}}$ et

$$\widehat{\pi}'_{Y_k}(y_k) = \frac{1}{N\widehat{\sigma}_{Y_k}}\sum_{n=1}^{N}\mathcal{K}'_2\left(\frac{y_k - Y_k(n)}{\widehat{\sigma}_{Y_k}}\right)$$

la dérivée de $\widehat{\pi}_{Y_k}$.

Par les mêmes arguments que pour l'estimation de la mesure de dépendance quadratique, nous pouvons en déduire que le gradient relatif estimé va tendre asymptotiquement vers le gradient relatif théorique. Nous examinerons plus en détail ces convergences dans la partie qui concerne la minimisation du critère empirique.

Termes d'ordre 2

Afin d'utiliser d'autres méthodes de minimisation, nous pouvons nous interroger sur le calcul du Hessien de ce critère. L'expression de ce Hessien étant complexe dans le cas général, nous allons nous restreindre au calcul de ce dernier au point où les variables sont indépendantes. D'autre part, comme nous l'avons signalé ci-dessus, la mesure de dépendance quadratique admet en un point où les variables sont indépendantes un minimum global, ce qui nous assure que le Hessien obtenu en ce point est défini positif. Il serait donc possible de l'introduire dans les méthodes de descente du gradient implémentées afin d'améliorer la convergence de nos algorithmes.

Lemme 3.2.12 *Soient $T_1, \ldots, T_K$ des variables aléatoires et $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$. Soit $\mathcal{K}$ un noyau de carré intégrable, tel que sa transformée de Fourier soit différente de zéro presque partout. Soit $\mathbf{\Delta}_\eta$ une famille de fonctions inversibles, on note $\mathbf{\Delta}_\eta =$*

$(\Delta_1, \ldots, \Delta_K)^T$, telle que pour tout η, $h + \mathbf{\Delta}_\eta \circ h$ est inversible et $\mathbf{\Delta}_\eta$ converge simplement vers 0. On suppose les variables $T_1, \ldots, T_K$ ***indépendantes*** *et on note $\mathbf{T}'$ un vecteur aléatoire de même distribution que $\mathbf{T}$ et indépendant de $\mathbf{T}$.*

Alors,

$$Q(\mathbf{T} + \mathbf{\Delta}(\mathbf{T})) - Q(\mathbf{T}) = \sum_{k=1}^{K} \sum_{j=1}^{K} E[H^*_{jk}(\mathbf{T}, \mathbf{T}')\Delta_j(\mathbf{T})\Delta_k(\mathbf{T}')] + o(\eta^2)$$

où pour $\mathbf{t}, \mathbf{t}' \in \mathbb{R}^K$,

$$\begin{aligned} H^*_{jk}(\mathbf{t}, \mathbf{t}') = \quad & H_{jk}(\mathbf{t}, \mathbf{t}') - E[H_{jk}(\mathbf{t}, \mathbf{T})T_k]\phi_k(t'_k) - E[H_{jk}(\mathbf{T}, \mathbf{t}')T_j]\phi_j(t_j) \\ + \quad & E[H_{jk}(\mathbf{T}, \mathbf{T}')T_jT'_k]\phi_j(t_j)\phi_k(t'_k) \end{aligned}$$

et

$$\begin{aligned} H_{jk}(\mathbf{t}, \mathbf{t}') = \int \quad & \mathcal{K}'\left(z_j - \frac{t_j}{\sigma_{T_j}}\right)\left[\prod_{l\neq j}\mathcal{K}\left(z_l - \frac{t_l}{\sigma_{T_l}}\right) - \prod_{l\neq j} E\mathcal{K}\left(z_l - \frac{T_l}{\sigma_{T_l}}\right)\right] \\ & \mathcal{K}'\left(z_k - \frac{t'_k}{\sigma_{T_k}}\right)\left[\prod_{l\neq j}\mathcal{K}\left(z_l - \frac{t'_l}{\sigma_{T_l}}\right) - \prod_{l\neq j} E\mathcal{K}\left(z_l - \frac{T_l}{\sigma_{T_l}}\right)\right]\frac{dz_1 \ldots dz_K}{\sigma_{T_j}\sigma_{T_k}} \end{aligned}$$

preuve :

On remarque simplement que comme les variables sont indépendantes, les termes liés au gradient sont nuls, et que les termes du développement de la mesure quadratique au second ordre s'écrivent alors,

$$\int \left\{ E\sum_{k=1}^{K} \mathcal{K}'\left(z_k - \frac{T_k}{\sigma_{T_k}}\right)\left[\prod_{l\neq k}\mathcal{K}\left(z_l - \frac{T_l}{\sigma_{T_l}}\right) - \prod_{l\neq k} E\mathcal{K}\left(z_l - \frac{T_l}{\sigma_{T_l}}\right)\right]\Delta^*_k(\mathbf{t}) \right\}^2 \frac{dz_1 \ldots dz_K}{\sigma^2_{T_k}}$$

où $\Delta^*_k(\mathbf{t}) = \Delta_k(\mathbf{t}) - t_k E[\delta_k(T_k)\Delta_k(\mathbf{T})]$

En effet, on remarque que

$$\prod_{k=1}^{K}\mathcal{K}\left(t_l-\frac{T_l+\Delta_l(\mathbf{T})}{\sigma_{T_l+\Delta_l(\mathbf{T})}}\right)-\prod_{k=1}^{K}\mathcal{K}\left(t_l-\frac{T_l}{\sigma_{T_l}}\right)=$$

$$-\sum_{k=1}^{K}\mathcal{K}'\left(t_k-\frac{T_k}{\sigma_{T_k}}\right)\frac{1}{\sigma_{T_k}}(\Delta_k^*(\mathbf{T})+\Delta_k^{**}(\mathbf{T}))\prod_{l\neq k}\mathcal{K}\left(t_l-\frac{T_l}{\sigma_{T_l}}\right)$$

$$-\sum_{k=1}^{K}\sum_{l=1}^{K}\mathcal{K}'\left(t_k-\frac{T_k}{\sigma_{T_k}}\right)\mathcal{K}'\left(t_l-\frac{T_l}{\sigma_{T_l}}\right)\frac{1}{\sigma_{T_k}}(\Delta_k^*(\mathbf{T}))\frac{1}{\sigma_{T_l}}(\Delta_l^*(\mathbf{T}))\prod_{m\neq k,m\neq l}\mathcal{K}\left(t_m-\frac{T_m}{\sigma_{T_m}}\right)$$

$$+\sum_{k=1}^{K}\mathcal{K}''\left(t_k-\frac{T_k}{\sigma_{T_k}}\right)\left(\frac{1}{\sigma_{T_k}}(\Delta_k^*(\mathbf{T}))\right)^2\prod_{l\neq k}\mathcal{K}\left(t_l-\frac{T_l}{\sigma_{T_l}}\right)$$

$$+\ o(\eta^2)$$

où $\Delta_k^*(\mathbf{T})=(T_k+\Delta_k(\mathbf{T})/\sigma_{T_k}-T_kE[\phi_k(\mathbf{T})\Delta_k(\mathbf{T})]/\sigma_{T_k}$ et Δ_k^{**} correspond au terme d'ordre 2 dans le développement de σ_{T_k}. Il n'est pas nécessaire de l'expliciter ici car on ne fait le calcul du Hessien seulement en un point où les variables sont indépendantes. Ces termes n'apparaissent donc pas dans l'expression finale du Hessien.

Nous avons fait l'hypothèse que les variables aléatoires sont indépendantes, alors on a l'égalité suivante,

$$E\left[\prod_{k=1}^{K}\mathcal{K}\left(t_l-\frac{T_l}{\sigma_{T_l}}\right)\right]=\prod_{k=1}^{K}E\left[\mathcal{K}\left(t_l-\frac{T_l}{\sigma_{T_l}}\right)\right]$$

■

Pour pouvoir exploiter cette expression du Hessien, il est à première vue nécessaire de procéder à une intégration multiple. Mais comme précédemment, nous allons réécrire les termes du Hessien à l'aide du noyau $\mathcal{K}_2$ défini par la convolution du noyau $\mathcal{K}$ avec sa fonction miroir (c.f. paragraphe 2.5.1)

Lemme 3.2.13 *L'expression du Hessien devient, pour tous* $\mathbf{t},\mathbf{t}'\in\mathbb{R}^K$,

Pour $j\neq k$ *:*

$$H_{jk}(\mathbf{t},\mathbf{t}')=\frac{1}{\sigma_{T_j}\sigma_{T_k}}\mathcal{K}_2'\left(\frac{t_j'-t_j}{\sigma_{T_j}}\right)\mathcal{K}_2'\left(\frac{t_k'-t_k}{\sigma_{T_k}}\right)\prod_{l\neq j,k}\mathcal{K}_2\left(\frac{t_l'-t_l}{\sigma_{T_l}}\right)$$

$$+\ \pi_{T_j}'(t_j)\pi_{T_k}'(t_k')\prod_{l\neq j,k}R(T_l)$$

$$-\left[\frac{1}{\sigma_{T_k}}\mathcal{K}_2'\left(\frac{t_k'-t_k}{\sigma_{T_k}}\right)\pi_{T_j}'(t_j)\prod_{l\neq j,k}\pi_{T_l}(t_l)+\frac{1}{\sigma_{T_j}}\mathcal{K}_2'\left(\frac{t_j-t_j'}{\sigma_{T_j}}\right)\pi_{T_k}'(t_k)\prod_{l\neq j,k}\pi_{T_l}(t_l')\right]$$

Pour $j = k$ *:*

$$H_{kk}(\mathbf{t}, \mathbf{t}') = -\mathcal{K}_2''\left(\frac{t_j - t_j'}{\sigma_{T_j}}\right)\left[\prod_{l\neq j}\mathcal{K}_2\left(\frac{t_l' - t_l}{\sigma_{T_l}}\right) + \prod_{l\neq j}R(T_l) - \prod_{l\neq j}\pi_{T_l}(t_l) - \prod_{l\neq j}\pi_{T_l}(t_l')\right]$$

où $\mathcal{K}_2$ *est un noyau réel de transformée de Fourier positive, sommable et différente de zéro presque partout,* $R(T_l) = E[\pi_{T_l}(T_l)]$ *et*

$$\pi_{T_l}'(t_l) = \frac{1}{\sigma_{T_l}}E\mathcal{K}_2'\left(\frac{t_l - T_l}{\sigma_{T_l}}\right) = \frac{d\pi_{Y_k}(y_k)}{dy_k}, \qquad \pi_{Y_k}(y_k) = E\mathcal{K}_2\left(\frac{t_l - T_l}{\sigma_{T_l}}\right)$$

La procédure précédente consiste donc en la résolution d'équation d'estimation correspondant à l'annulation du gradient. Comme nous avons pu le voir précédemment, annuler le gradient correspond exactement à trouver des variables indépendantes. Nous envisageons ici une nouvelle procédure. Celle-ci consiste à tout d'abord donner une estimation de la mesure de dépendance utilisée, et ensuite à calculer son gradient afin d'en déduire une méthode de minimisation.

3.2.3 «Estimer d'abord» avec l'information mutuelle

Par définition de l'information mutuelle, on doit envisager l'estimation de l'entropie. Or celle-ci peut-être estimée de deux manières différentes qui vont conduire à deux méthodes assez distinctes. Nous pouvons nous reporter aux travaux de Joe [38] ou de Hall *et. al.* [34]. Dans la section 5.5, nous étudierons plus particulièrement les convergences de certains estimateurs.

Première méthode: Avec l'estimateur de la moyenne empirique

On envisage ici un estimateur de l'information mutuelle à l'aide de la moyenne empirique. Ensuite, on calcule le gradient relatif de celui-ci afin de procéder à sa minimisation.

De par la définition de l'entropie, pour un vecteur aléatoire $\mathbf{T}$ de loi absolument continue de densité $p_{\mathbf{T}}$, l'entropie de $\mathbf{T}$ s'écrit,

$$H(\mathbf{T}) = -E[\log p_{\mathbf{T}}(\mathbf{T})]$$

Nous en déduisons alors l'estimateur suivant,

Définition 3.2.3 *Soient* $T_1, \ldots, T_K$ *des variables aléatoires et* $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ *soit de loi absolument continue.*

Pour tout $i = 1, \ldots, K$*, on note* $T_i(1), \ldots, T_i(N)$ *un échantillon de taille* N *de la variable aléatoire réelle* T_i*.*

Alors, un estimateur de l'entropie sera donné par,

$$\widehat{H}^m(\mathbf{T}) = \frac{1}{N}\sum_{j=1}^{N} \log \widehat{p}^m_{\mathbf{T}}(\mathbf{T}(j))$$

où $\widehat{p}^m_{\mathbf{T}}$ est une estimation à noyaux de la densité de $\mathbf{T}$

L'estimateur correspondant pour l'information mutuelle est alors donné clairement par,

Définition 3.2.4 (Estimateur de l'IM avec la moyenne empirique) *Soient $T_1, \ldots, T_K$ des variables aléatoires telles que $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ soit de loi absolument continue.*

Pour tout $i = 1, \ldots, K$, on note $T_i(1), \ldots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.

Alors, un estimateur de l'information mutuelle des variables $T_1, T_2, \ldots, T_K$ sera donné par,

$$\widehat{I}^m(T_1, T_2, \ldots, T_K) = \sum_{k=1}^{K} \widehat{H}^m(T_k) - \widehat{H}^m(\mathbf{T})$$

Nous ne montrons pas ici les propriétés de convergence de cet estimateur, elles seront développées dans la section 5.5.

Détaillons à présent, le développement limité de ce critère quand on fait varier les échantillons des variables aléatoires.

Fixons tout d'abord quelques notations.

D'après la définition de l'entropie donnée précédemment, on constate que cet estimateur $\widehat{H}^m(\mathbf{T})$ ne dépend que des échantillons de la variable $\mathbf{T}$, $\mathbf{T}(j) = (T_1(j), \ldots, T_K(j))^T$.

Définissons alors,

Définition 3.2.5 *Pour tout $i = 1, \ldots, K$ et pour tout $j = 1, \ldots, N$,*

- $\widehat{\phi}^m_i(\mathbf{T}(j)) = N\partial^2_{ij}\widehat{H}^m(\mathbf{T})$
- $\widehat{\psi}^m_i(T_i(j)) = N\partial^2_{ij}\widehat{H}^m(T_i)$

où ∂^2_{ij} désigne la dérivée par rapport à la variable $T_i(j)$.

Avec ces notations, nous pouvons alors en déduire le développement limité de l'entropie discrétisée puis de l'information mutuelle discrétisée:

Lemme 3.2.14 *Soient $T_1, \ldots, T_K$ des variables aléatoires telles que $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ soit de loi absolument continue.*

Pour tout $i = 1, \ldots, K$, on note $T_i(1), \ldots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.

Soit $\mathbf{\Delta}_\eta$ une famille de fonctions définies comme dans le lemme 3.2.3.

$$\widehat{H}^m(\mathbf{T}+\mathbf{\Delta}_\eta(\mathbf{T})) - \widehat{H}^m(\mathbf{T}) \;=\; \frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\widehat{\phi}_k^m(\mathbf{T}(j))\Delta_k(\mathbf{T}(j)) + o(\eta)$$

Et pour chaque variable aléatoire T_k, on a,

$$\widehat{H}^m(T_k+\Delta_k(\mathbf{T})) - \widehat{H}^m(T_k) \;=\; \frac{1}{N}\sum_{j=1}^{N}\widehat{\psi}_k^m(T_k(j))\Delta_k(\mathbf{T}(j)) + o(\eta)$$

Et par conséquent, en ce qui concerne l'information mutuelle, on obtient,

Lemme 3.2.15 (DL de l'estimateur de l'IM avec la moyenne empirique) *Soient $T_1,\ldots,T_K$ des variables aléatoires telles que $\mathbf{T}=(T_1,T_2,\ldots,T_K)^T$ soit de loi absolument continue.*

Pour tout $i=1,\ldots,K$, on note $T_i(1),\ldots,T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.

Soit $\mathbf{\Delta}_\eta$ une famille de fonctions définies comme dans le lemme 3.2.3.

$$\widehat{I}^m(\mathbf{T}+\mathbf{\Delta}_\eta(\mathbf{T})) - \widehat{I}^m(\mathbf{T}) \;=\; \frac{1}{N}\sum_{j=1}^{N}\mathbf{\Delta}_\eta^T(\mathbf{T}(j))\widehat{\beta}_{\mathbf{T}}^m(\mathbf{T}(j)) + o(\eta)$$

où $\widehat{\beta}_k^m(\mathbf{T}(j)) = \widehat{\psi}_k^m(T_k(j)) - \widehat{\phi}_k^m(\mathbf{T}(j))$, pour tout $k=1,\ldots,K$.

On retrouve ici exactement les mêmes résultats que lors du calcul du gradient de l'information mutuelle théorique avec des estimateurs particuliers pour les fonctions scores jointes et marginales.

Nous renvoyons au paragraphe 5.4.1 où nous démontrons plus précisément comment nous obtenons les expressions de $\widehat{\phi}_i^m$ et $\widehat{\psi}_i^m$. Nous donnons ici seulement leur définition afin d'observer les différences par rapport aux estimateurs généralement utilisés.

Lemme 3.2.16 *Soient $T_1,\ldots,T_K$ des variables aléatoires telles que $\mathbf{T}=(T_1,T_2,\ldots,T_K)^T$ soit de loi absolument continue.*

Pour tout $i=1,\ldots,K$, on note $T_i(1),\ldots,T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.

Les fonctions $\widehat{\phi}_i^m$ et $\widehat{\psi}_i^m$ pour tout $i=1,\ldots,K$ sont données par,

$$\widehat{\phi}_i^m(\mathbf{T}(j)) \;=\; \widetilde{\phi}_i(T(j)) + \frac{T_i(j)-\widehat{E}(T_i)}{\widehat{\sigma}_{T_i}^2}\left[1-\sum_{m=1}^{N}\widetilde{\phi}_i(T(m))T_i(m)\right]$$

$$\widehat{\psi}_i^m(T_i(j)) \;=\; \widetilde{\psi}_i(T_i(j)) + \frac{T_i(j) - \widehat{E}(T_i)}{\widehat{\sigma}_{T_i}^2}\left[1 - \sum_{m=1}^{N} \widetilde{\psi}_i(T_i(m))T_i(m)\right]$$

où

$$\widetilde{\phi}_i(\mathbf{T}(j)) \;=\; \widehat{\phi}_i(T_i(j)) + \sum_{m=1}^{N} \frac{\mathcal{K}'[(T_i(m) - T_i(j))/\widehat{\sigma}_{T_i}]}{N\prod_{k=1}^{K}\widehat{\sigma}_{T_k}\widehat{\sigma}_{T_i}\widehat{p}_{\mathbf{T}}(\mathbf{T}(m))} \prod_{k=1,k\neq i}^{K} \mathcal{K}[(T_k(m) - T_k(j))/\widehat{\sigma}_{T_k}]$$

avec $\widehat{\phi}_i = -\partial_i \log \widehat{p}_{\mathbf{T}}^m$
et

$$\widetilde{\psi}_i(T_i(j)) \;=\; \widehat{\psi}_i(T_i(j)) + \sum_{m=1}^{N} \frac{\mathcal{K}'[(T_i(m) - T_i(j))/\widehat{\sigma}_{T_i}]}{N\widehat{\sigma}_{T_i}^2\widehat{p}_{T_i}(T_i(m))}]$$

avec $\widehat{\psi}_i = -(\log \widehat{p}_{T_i}^m)'$

<u>Deuxième méthode :</u> En utilisant une discrétisation de l'intégration.

Cette méthode est décrite par Pham dans [61]. Nous détaillons ici la méthode pour permettre une comparaison avec la précédente. Nous envisageons un estimateur de l'information mutuelle à l'aide d'une discrétisation d'intégrale. Puis nous calculons le gradient relatif de cet estimateur.

A présent, nous considérons la définition de l'entropie sous la forme,

$$H(\mathbf{T}) = -\int p_{\mathbf{T}}(\mathbf{t}) \log p_{\mathbf{T}}(\mathbf{t}) d\mathbf{t}$$

où $\mathbf{T}$ est un vecteur aléatoire de loi absolument continue et $p_{\mathbf{T}}$ sa densité.

Un estimateur de l'entropie sera alors défini de la manière suivante:

Définition 3.2.6 *Soient $T_1, \ldots, T_K$ des variables aléatoires telles que $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ soit de loi absolument continue.*

Pour tout $i = 1, \ldots, K$, on note $T_i(1), \ldots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.

Alors, un estimateur de l'entropie sera donné par,

$$\widehat{H}^i(\mathbf{T}) = -\sum_{\mathbf{l}} \log[\widehat{p}_{\mathbf{T}}^i(\widehat{E}(\mathbf{T}) + \mathbf{l} : \mathbf{b} : \widehat{\sigma}_{\mathbf{T}})\prod_{k=1}^{K}\widehat{\sigma}_{T_k}]\widehat{p}_{\mathbf{T}}^i(\widehat{E}(\mathbf{T}) + \mathbf{l} : \mathbf{b} : \widehat{\sigma}_{\mathbf{T}})\prod_{k=1}^{K} b_k\widehat{\sigma}_{T_k} + \sum_{k=1}^{K}\log\widehat{\sigma}_{T_k}$$

où

$$\widehat{p}^i_{\mathbf{T}}(\widehat{E}(\mathbf{T})+\mathbf{l}:\mathbf{b}:\widehat{\sigma}_{\mathbf{T}})\prod_{k=1}^{K}\widehat{\sigma}_{T_k}=\frac{1}{N}\sum_{n=1}^{N}\prod_{k=1}^{K}\mathcal{K}\left(l_kb_k-\frac{T_k(n)-\widehat{E}(T_k)}{\widehat{\sigma}_{T_k}}\right)$$

où $\mathbf{b}=(b_1,\ldots,b_K)^T$ *désigne le vecteur des pas de discrétisation. (il est formé de petits nombres positifs), et* $\mathbf{l}=(l_1,\ldots,l_K)^T$*. De plus on définit, pour tous vecteurs* $\mathbf{u}$ *et* $\mathbf{v}$*,* $\mathbf{u}:\mathbf{v}=\mathbf{w}$ *où* $w_k=u_kv_k$*.*

Et pour tout $k=1,\ldots,K$*, on en déduit un estimateur de l'entropie marginale de chaque variable* T_k*.*

$$\widehat{H}^i(T_k)=-\sum_{l_k}\log[\widehat{p}^i_{T_k}(\widehat{E}(T_k)+l_kb_k\widehat{\sigma}_{T_k})\widehat{\sigma}_{T_k}]\widehat{p}^i_{T_k}(\widehat{E}(T_k)+l_kb_k\widehat{\sigma}_{T_k})b_k\widehat{\sigma}_{T_k}+\log\widehat{\sigma}_{T_k}$$

où

$$\widehat{p}^i_{T_k}(\widehat{E}(T_k)+l_kb_k\widehat{\sigma}_{T_k})\widehat{\sigma}_{T_k}=\frac{1}{N}\sum_{n=1}^{N}\mathcal{K}\left(l_kb_k-\frac{T_k(n)-\widehat{E}(T_k)}{\widehat{\sigma}_{T_k}}\right)$$

Il réside ici aussi le problème du choix du noyau dans l'estimation. Pham [61] propose de choisir un noyau spline à support compact. En effet, ceci permet de réduire les coûts de calcul en utilisant les propriétés de récurrence des splines.

On en déduit alors aisément une estimation de l'information mutuelle.

Définition 3.2.7 (Estimateur de l'IM avec discrétisation de l'intégrale)
Soient $T_1,\ldots,T_K$ *des variables aléatoires telles que*
$\mathbf{T}=(T_1,T_2,\ldots,T_K)^T$ *soit de loi absolument continue.*

Pour tout $i=1,\ldots,K$*, on note* $T_i(1),\ldots,T_i(N)$ *un échantillon de taille* N *de la variable aléatoire réelle* T_i*.*

Alors, un estimateur de l'information mutuelle des variables $T_1,T_2,\ldots,T_K$ *sera donné par,*

$$\widehat{I}^i(T_1,T_2,\ldots,T_K)=-\sum_{\mathbf{l}}\log\frac{\widehat{p}^i_{\mathbf{T}}(\widehat{E}(\mathbf{T})+\mathbf{l}:\mathbf{b}:\widehat{\sigma}_{\mathbf{T}})\prod_{k=1}^{K}\widehat{\sigma}_{T_k}}{\prod_{k=1}^{K}\widehat{p}^i_{T_k}(\widehat{E}(T_k)+l_kb_k\widehat{\sigma}_{T_k})\widehat{\sigma}_{T_k}}\widehat{p}^i_{\mathbf{T}}(\widehat{E}(\mathbf{T})+\mathbf{l}:\mathbf{b}:\widehat{\sigma}_{\mathbf{T}})\prod_{k=1}^{K}\widehat{\sigma}_{T_k}$$

Remarquons, que pour certains choix du noyau, en particulier, si le noyau réalise une partition de l'unité, nous en déduisons la propriété suivante:

Lemme 3.2.17 *Avec les notations précédentes, pour tout* $k=1,\ldots,K$*,*

$$\widehat{p}^i_{T_k}(\widehat{E}(T_k)+l_kb_k\widehat{\sigma}_{T_k})\widehat{\sigma}_{T_k}=\sum_{\mathbf{i}:i_k=j}\widehat{p}^i_{\mathbf{T}}(\widehat{E}(\mathbf{T})+\mathbf{l}:\mathbf{b}:\widehat{\sigma}_{\mathbf{T}})\prod_{k=1}^{K}\widehat{\sigma}_{T_k}$$

On déduit alors la propriété suivante de l'information mutuelle, (qui est similaire à la propriété vérifiée théoriquement),

Lemme 3.2.18 *Soient $T_1, \ldots, T_K$ des variables aléatoires telles que $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ soit de loi absolument continue.*

Pour tout $i = 1, \ldots, K$, on note $T_i(1), \ldots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.

Alors, l'estimateur de l'information mutuelle des variables $T_1, T_2, \ldots, T_K$ vérifie,

$$\widehat{I}^i(T_1, T_2, \ldots, T_K) = \sum_{k=1}^{K} \widehat{H}^i(T_k) - \widehat{H}^i(\mathbf{T})$$

La minimisation de l'estimateur de l'information mutuelle conduit donc à déterminer des variables indépendantes. Calculons alors à présent le développement limité de l'information mutuelle.

Pour cela regardons tout d'abord ce qu'il en est du développement limité de l'entropie.

Lemme 3.2.19 *Soient $T_1, \ldots, T_K$ des variables aléatoires telles que $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ soit de loi absolument continue.*

Pour tout $i = 1, \ldots, K$, on note $T_i(1), \ldots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.

Soit $\mathbf{\Delta}_\eta$ une famille de fonctions définies comme dans le lemme 3.2.3.

Alors,

$$\widehat{H}^i(\mathbf{T} + \mathbf{\Delta}_\eta(\mathbf{T})) - \widehat{H}^i(\mathbf{T}) = \frac{1}{N} \sum_{k=1}^{K} \sum_{j=1}^{N} \widehat{\phi}_k^i(\mathbf{T}(j)) \Delta_k(\mathbf{T}(j)) + o(\eta)$$

Et pour chaque variable aléatoire T_k, on a,

$$\widehat{H}^i(T_k + \Delta_k(\mathbf{T})) - \widehat{H}^i(T_k) = \frac{1}{N} \sum_{j=1}^{N} \widehat{\psi}_k^i(T_k(j)) \Delta_k(\mathbf{T}(j)) + o(\eta)$$

où

$$\begin{aligned} \widehat{\phi}_k^i(T_k(j)) &= \widehat{\phi}_k^{\mathbf{b}}(\mathbf{T}(j)) - \frac{1}{N} \sum_{m=1}^{N} \widehat{\phi}_k^{\mathbf{b}}(\mathbf{T}(m)) \\ &+ \sum_{r=1}^{K} \frac{T_r(j) - \widehat{E}(T_r)}{\widehat{\sigma}_{T_r}^2} \left[1 - \sum_{m=1}^{N} \widehat{\phi}_k^{\mathbf{b}}(\mathbf{T}(m)) (T_r(m) - \widehat{E}(T_r)) \right] \end{aligned}$$

où

$$\widehat{\phi}_k^{\mathbf{b}}(\mathbf{T}(j)) = \sum_{\mathbf{l}} [\log \widehat{p}^i_{\mathbf{T}}(\widehat{E}(\mathbf{T}) + \mathbf{l} : \mathbf{b} : \widehat{\sigma}_{\mathbf{T}})\widehat{\sigma}_{\mathbf{T}} + 1]$$
$$\frac{1}{\widehat{\sigma}_{T_k}} \mathcal{K}' \left(\frac{l_k b_k \widehat{\sigma}_{T_k} + T_k(j) - \widehat{E}(T_k)}{\widehat{\sigma}_{T_k}} \right) b_k \prod_{i=1, i \neq K}^{K} \mathcal{K}' \left(\frac{l_i b_i \widehat{\sigma}_{T_i} + T_i(j) - \widehat{E}(T_i)}{\widehat{\sigma}_{T_i}} \right) b_i$$

et

$$\begin{aligned} \widehat{\psi}_k^i(T_k(j)) &= \widehat{\psi}_k^{b_k}(T_k(j)) - \frac{1}{N} \sum_{m=1}^{N} \widehat{\psi}_k^{b_k}(T_k(m)) \\ &+ \frac{T_k(j) - \widehat{E}(T_k)}{\widehat{\sigma}^2_{T_k}} \left[1 - \sum_{m=1}^{N} \widehat{\psi}_k^{b_k}(T_k(m))(T_k(j) - \widehat{E}(T_k)) \right] \end{aligned}$$

où

$$\widehat{\psi}_k^{b_k}(T_k(j)) = \sum_{l_k} [\log \widehat{p}^i_{T_k}(\widehat{E}(T_k) + l_k b_k \widehat{\sigma}_{T_k})\widehat{\sigma}_{T_k} + 1] \frac{1}{\widehat{\sigma}_{T_k}} \mathcal{K}' \left(l_k b_k \widehat{\sigma}_{T_k} + \frac{T_k(j) - \widehat{E}(T_k)}{\widehat{\sigma}_{T_k}} \right)$$

Le détail des calculs afin d'obtenir les expressions explicites de $\widehat{\psi}_k^i$ et $\widehat{\phi}_k^i$ est présenté dans le paragraphe 5.4.2.

On en déduit alors simplement le développement limité de l'estimateur de l'information mutuelle en utilisant une discrétisation de l'intégrale,

Lemme 3.2.20 (DL de l'estimateur de l'IM avec une discrétisation de l'intégrale) *Soient $T_1, \ldots, T_K$ des variables aléatoires telles que $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ soit de loi absolument continue.*

Pour tout $i = 1, \ldots, K$, on note $T_i(1), \ldots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.

Soit $\mathbf{\Delta}_\eta$ une famille de fonctions définies comme dans le lemme 3.2.3.

$$\widehat{I}^i(\mathbf{T} + \mathbf{\Delta}_\eta T) - \widehat{I}^i(\mathbf{T}) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{\Delta}^T(\mathbf{T}(j)) \widehat{\beta}^i_{\mathbf{T}}(\mathbf{T}(j)) + o(\eta)$$

où $\widehat{\beta}_k^i(\mathbf{T}(j)) = \widehat{\psi}_k^i(T_k(j)) - \widehat{\phi}_k^i(\mathbf{T}(j))$, pour tout $k = 1, \ldots, K$.

3.2.4 «Estimer d'abord» avec la mesure de dépendance quadratique

Nous allons ici aussi proposer une estimation de la mesure de dépendance quadratique. Puis nous calculerons son gradient relatif afin de procéder à sa minimisation.

Rappelons ici l'estimateur de la mesure de dépendance quadratique utilisé :

Définition 3.2.8 *Soit $\mathcal{K}_2$ un noyau réel tel que sa transformée de Fourier soit positive et sommable. Soient K variables aléatoires $T_1, \ldots, T_K$ telles que $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$. Pour tout k, $1 \leq k \leq K$, on note $T_k(1), \ldots, T_k(N)$ un échantillon de taille N de la variable aléatoire T_k. Alors l'estimateur de la mesure de dépendance quadratique est défini par,*

$$\widehat{Q}(T_1, \ldots, T_K) = \widehat{E}[\widehat{\pi}_{\mathbf{T}}(\mathbf{T})] + \prod_{k=1}^{K} \widehat{E}[\widehat{\pi}_{T_k}(T_k)] - 2\widehat{E}\left[\prod_{k=1}^{K} \widehat{\pi}_{T_k}(T_k)\right]$$

où $\widehat{E}[\Phi(X)] = \sum_{n=1}^{N} \Phi(X(n))/N$, pour toute fonction Φ de la variable aléatoire X d'échantillon $X(1), \ldots, X(N)$ et

$$\begin{aligned}
\widehat{\pi}_{\mathbf{T}}(t_1, \ldots, t_K) &= \widehat{E}\left[\prod_{k=1}^{K} \mathcal{K}_2\left(\frac{t_k - T_k}{\widehat{\sigma_{T_k}}}\right)\right] = \frac{1}{N}\sum_{n=1}^{N}\prod_{k=1}^{K} \mathcal{K}_2\left(\frac{t_k - T_k(n)}{\widehat{\sigma_{T_k}}}\right) \\
\widehat{\pi}_{T_k}(T_k) &= \widehat{E}\left[\mathcal{K}_2\left(\frac{t_k - T_k}{\widehat{\sigma_{T_k}}}\right)\right] = \frac{1}{N}\sum_{n=1}^{N} \mathcal{K}_2\left(\frac{t_k - T_k(n)}{\widehat{\sigma_{T_k}}}\right)
\end{aligned}$$

et $\widehat{\sigma_{T_k}}$ est un estimateur de σ_{T_k} ne dépendant que de $T_k(1), \ldots, T_k(N)$ l'echantillon de T_k.

Remarquons que celui-ci ne dépend que des variables normalisées, notées $T'_k(n) = T_k(n)/\widehat{\sigma}_{T_k}$.

Détaillons tout d'abord le développement limité des variables normalisées, pour une petite variation de $\mathbf{T}(n)$.

Lemme 3.2.21 *Soient $T_1, \ldots, T_K$ des variables aléatoires.*

Pour tout $i = 1, \ldots, K$, on note $T_i(1), \ldots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.

Soit $\mathbf{\Delta}_\eta$ une famille de fonctions définies comme dans le lemme 3.2.9.

Alors,

$$(\mathbf{T} + \mathbf{\Delta}(\mathbf{T}))'_k(n) - T'_k(n) = \frac{1}{\widehat{\sigma}_{T_k}}\left\{\Delta_k - T_k(n)\widehat{E}[\widehat{\phi}_{T_k}(T_k)\Delta_k]\right\} + o(\eta)$$

où par analogie au cas théorique, $\widehat{\phi}_{T_k}(T_k(n))$ désigne la dérivée partielle de $N \log \widehat{\sigma_{T_k}}$ par rapport à $T_k(n)$ et $(\mathbf{T} + \mathbf{\Delta}(\mathbf{T}))'_k(n) = (\mathbf{T} + \mathbf{\Delta}(\mathbf{T}))_k(n)/\widehat{\sigma}_{(\mathbf{T}+\mathbf{\Delta}(\mathbf{T}))_k}$.

On en déduit alors le développement limité de l'estimateur de la mesure de dépendance quadratique,

Lemme 3.2.22 (DL de l'estimateur de la DQ) *Soient $T_1, \ldots, T_K$ des variables aléatoires.*

Pour tout $i = 1, \ldots, K$, on note $T_i(1), \ldots, T_i(N)$ un échantillon de taille N de la variable aléatoire réelle T_i.

Soit $\boldsymbol{\Delta}_\eta$ une famille de fonctions définies comme dans le lemme 3.2.9.

Alors,

$$\widehat{Q}(\mathbf{T} + \boldsymbol{\Delta}_\eta(\mathbf{T}) - \widehat{Q}(\mathbf{T}) = \widehat{E}\left[\sum_{k=1}^{K} G_k^*(\mathbf{T})\Delta_k(\mathbf{T})\right] + o(\eta)$$

où

$$G_k^*(\mathbf{t}) = \widehat{G}_k(\mathbf{t}) - \widehat{E}[T_k\widehat{G}_k]\widehat{\phi}_{T_k}(T_k)$$

avec

$$\begin{aligned}\widehat{G}_k(\mathbf{y}) &= \widehat{\pi}'_{k,\mathbf{Y}}(\mathbf{y}) - \widehat{\pi}'_{Y_k}(y_k)\prod_{l\neq k}\widehat{\pi}_{Y_l}(y_l) + \widehat{\pi}'_{Y_k}(y_k)\prod_{l\neq k}\widehat{E}(\widehat{\pi}_{Y_l}(Y_l)) \\ &\quad -\frac{1}{\widehat{\sigma}_{Y_k}}\widehat{E}\left[\mathcal{K}'_2\left(\frac{y_k - Y_k}{\widehat{\sigma}_{Y_k}}\right)\prod_{l\neq k}\widehat{E}(\widehat{\pi}_{Y_l}(Y_l))\right]\end{aligned}$$

avec

$$\widehat{\pi}'_{k,\mathbf{Y}}(\mathbf{y}) = \frac{1}{N\widehat{\sigma}_{Y_k}}\sum_{n=1}^{N}\mathcal{K}'_2\left(\frac{y_k - Y_k(n)}{\widehat{\sigma}_{Y_k}}\right)\prod_{l\neq k}\mathcal{K}_2\left(\frac{y_l - Y_l(n)}{\widehat{\sigma}_{Y_l}}\right),$$

la k-ième composante du gradient de $\widehat{\pi}_{\mathbf{Y}}$ et

$$\widehat{\pi}'_{Y_k}(y_k) = \frac{1}{N\widehat{\sigma}_{Y_k}}\sum_{n=1}^{N}\mathcal{K}'_2\left(\frac{y_k - Y_k(n)}{\widehat{\sigma}_{Y_k}}\right)$$

la dérivée de $\widehat{\pi}_{Y_k}$.

Remarque 3.2.3 *Comme on a remarqué dans le cas théorique, on a une propriété identique vérifiée par $\widehat{G}_k(\mathbf{t})$,*

$$\sum_{n=1}^{N}\widehat{G}_k(\mathbf{T}(n)) = 0$$

On démontre ceci simplement en faisant le calcul sur les termes du gradient $\widehat{G}_k$

Avec ce dernier lemme, on remarque que dans le cas de la mesure de dépendance quadratique, les deux gradients obtenus en utilisant les deux stratégies «Estimer ensuite» et «Estimer d'abord» sont égaux.

3.3 Commentaires

3.3.1 Méthodes de résolution : les stratégies «Estimer ensuite» et «Estimer d'abord»

Nous avons introduit dans ce chapitre une nouvelle méthode de résolution, la stratégie «Estimer d'abord». Cette stratégie permet de mettre en place des méthodes d'optimisation qui recherchent effectivement le minimum de l'estimateur du critère de séparation, alors que la stratégie «Estimer ensuite» fait une descente selon l'estimation d'un gradient, laquelle n'est ni exactement le gradient du critère, ni le gradient de son estimation : la fonction que l'on optimise est inconnue. Avec cette nouvelle méthode, en étudiant la convergence des estimateurs des différents critères de séparation, il est possible de connaître le comportement de la fonction à optimiser. Nous reviendrons sur ce point dans les chapitres 5 et 6. D'autre part, le fait de connaître la fonction à optimiser permet de contrôler les pas dans la méthode de descente du gradient. En effet, dans le chapitre 4, nous construirons un algorithme qui adapte les pas de la méthode de descente du gradient de telle sorte que le critère soit effectivement minimisé. On sait en effet que, dans les algorithmes de minimisation, si les paramètres de descente sont choisis trop grand, on observe des oscillations à la convergence, et dans certains cas, la méthode peut diverger au bout d'un trop grand nombre d'itérations.

3.3.2 Minima locaux

Un des problèmes majeurs pour la méthode de descente du gradient est la présence de minima locaux. Nous avons observé que dans un cadre général, les deux mesures de dépendance n'admettent aucun minimum local. Cependant, lorsqu'on restreint le mélange à une certaine classe, on ne peut rien dire[2]. Nous verrons dans le paragraphe 4.4.8 qu'il existe au moins un point où le gradient s'annule et qui ne corresponde pas à la solution du problème. Bien sûr ces points peuvent correspondre à des points selles ou des maxima. Dans ce cas, ils n'engendreront aucun problème dans la recherche du minimum. De plus, ces méthodes de minimisation sont très dépendantes du point d'initialisation. Dans le cadre de mélanges post non linéaires, Sole et Jutten [65] proposent d'initialiser les non linéarités de telle sorte qu'elles soient égales à la composition de la fonction de répartition de chaque observation correspondante et de l'inverse de la fonction de répartition de la loi gaussienne. Une autre méthode proposée par Babaie-Zadeh [6] consiste à chercher la direction de descente du gradient sans respecter la structure de séparation, puis de projeter la solution dans l'espace des mélanges post non linéaires. Il a appelé cette méthode, minimisation par projection.

2. C'est le problème classique de l'optimisation sous contrainte

Chapitre 4

Mélanges post non linéaires

Depuis quelques années, de nombreux travaux ont été effectués dans le cadre de la séparation de source pour des mélanges linéaires. Nous avons déjà décrit dans les chapitres précédents quelques unes des méthodes développées. Cependant, il est tout à fait naturel de penser que la contrainte de linéarité du mélange peut dans certaines circonstances être trop restrictive – au point de ne pas pouvoir trouver une solution acceptable. Citons par exemple le cas de signaux provenant de satellites qui sont le résultat de la composition "d'un filtre linéaire et d'un amplificateur non linéaire" [69], où Taleb et Jutten ont envisagé un nouveau type de mélange, appelé mélange post non linéaire. Ce type de mélange est non linéaire et permet d'approcher un grand nombre de non linéarités, on peut aussi remarquer qu'il présente des similitudes avec les réseaux de neurones.

Dans un premier temps, nous définirons ce qu'est un mélange post non linéaire, et nous commenterons les problèmes d'identifiabilité liés à ce mélange (sections 4.1 et 4.2). Dans la section 4.3, nous ferons un inventaire non exhaustif des méthodes déjà développées. Puis, nous détaillerons plus particulièrement certaines approches utilisant les propriétés des mélanges post non linéaires (section 4.4). Enfin nous conclurons par l'application de ces méthodes avec les mesures de dépendance définies dans le chapitre 2 : l'information mutuelle et la mesure de dépendance quadratique (paragraphes 4.4.8 et 4.4.9).

4.1 Définition

Ayant remarqué que dans certaines situations, les mélanges linéaires sont tout à fait inadaptés, Taleb et Jutten ont proposé en 1999, [68], la définition de mélanges post non linéaires, qui ont la propriété d'introduire une non linéarité tout en conservant la propriété d'identifiabilité.

Définition 4.1.1 (Mélange post non linéaire) *On dispose des enregistrements de capteurs $X_1, X_2, \ldots, X_K$ provenant des transformations non linéaires sur des mélanges linéaires de K sources* ***indépendantes****, $S_1, S_2, \ldots, S_K$, c'est-à-dire :*

$$X_i = f_i(\sum_{k=1}^{K} \mathbf{A}_{ik} S_k), \qquad i = 1, \ldots, K$$

où $\mathbf{A}_{ik}$ désigne le terme général de la matrice de mélange $\mathbf{A}$ ***inversible****, et $f_1, f_2, \ldots, f_K$ sont K applications* ***monotones****.*

Dans [68], Taleb fait alors remarquer que ce type de mélange peut être observé dans de nombreuses situations en constatant que la distortion non linéaire à la sortie du mélange peut modéliser des caractéristiques inconnues des capteurs. D'autre part, nous pouvons reconnaître dans cette définition, une ressemblance avec les réseaux de neurones, qui fournissent une technique d'approximation de non linéarités. On peut montrer une propriété d'identifiabilité pour ces mélanges.

4.2 Identifiabilité

Nous avons pu remarquer que la méthode d'analyse en composantes indépendantes permet dans certaines situations de résoudre le problème de séparation aveuge de sources, c'est le cas pour des mélanges linéaires. Par contre, ce n'est pas le cas pour toutes sortes de mélanges non linéaires (c.f. Darmois [26], Taleb [69]). Cependant, dans le cadre de mélanges post non linéaires, la propriété d'identifiabilité s'exprime sous la forme du lemme suivant :

Lemme 4.2.1 *Soient $S_1, \ldots, S_K$, K variables aléatoires indépendantes.*

On note $Y_1, \ldots, Y_K$, K variables aléatoires, les transformées de $S_1, \ldots, S_K$ de la manière suivante,

$$\mathbf{Y} = \mathbf{B}h\mathbf{A}\mathbf{S}$$

où $\mathbf{Y} = (Y_1, \ldots, Y_K)^T$, $\mathbf{S} = (S_1, \ldots, S_K)^T$,$\mathbf{A}$ et $\mathbf{B}$ sont deux matrices inversibles, enfin $h(x) = (h_1(x), \ldots, h_K(x))^T$ avec $h_1, \ldots, h_K$ des transformations inversibles.

Avec les hypothèses suivantes,

- $\mathbf{A}$ *est une matrice inversible mélangeante, c'est-à-dire qu'il existe au moins deux valeurs non nulles par colonne ou par ligne.*
- *les transformations $h_1, h_2, \ldots, h_K$ sont inversibles, par convention on les supposera strictement croissantes.*
- *Il existe au plus une source gaussienne.*

alors, si $Y_1, \ldots, Y_K$ sont indépendantes, $h_1, \ldots, h_K$ sont linéaires.

D'après les notations utilisées précédemment (définition 4.1.1), les fonctions h_k correspondent à $g_k \circ f_k$.

Ce lemme a été montré par Taleb [67] avec des hypothèses techniques sur le support des sources, puis par Babaie-Zadeh [6], en faisant l'hypothèse que les supports des distributions des sources sont bornés. Cependant, il serait intéressant de montrer l'identifiabilité des mélanges post non linéaires sans imposer de restrictions sur les densités des sources. Ce problème s'avère difficile et reste ouvert.

Remarque 4.2.1 *Remarquons que l'hypothèses d'inversibilité des transformations non linéaires est indispensable. Ceci signifie qu'il est possible de trouver un mélange post non linéaire dans lequel une des transformations non linéaires n'est pas inversible et qui conserve l'indépendance. En effet, citons l'exemple suivant, (c.f. [41]),*

Lemme 4.2.2 *Soit X une variable aléatoire gaussienne, on note f_X sa densité, $f_X(x) = (1/\sigma)\phi(x/\sigma)$ (où ϕ est la densité de la loi normale centrée réduite). Soit Y une variable indépendante de X et ayant pour densité, $f_Y(x) = (1/2\sigma)[\phi((x - \lambda\sigma^2)/\sigma) + \phi((x + \lambda\sigma^2)/\sigma)]$. Alors, on montre que $X - Y$ et $|X - Y|$ sont indépendantes.*

Naturellement, ce contexte de mélanges non linéaires a suscité la curiosité et différentes méthodes ont été déjà développées.

4.3 Méthodes déjà existantes

Nous décrivons ici succintement les méthodes déjà développées dans le cadre de mélanges non linéaires. Nous verrons alors par la suite, l'apport de la mesure de dépendance quadratique par rapport à ces différentes méthodes.

- Approche bayésienne [73]
  A l'aide de méthodes bayésiennes, cette approche présente une extension au cas non linéaire de l'analyse factorielle "factor analysis". Les non linéarités sont alors approchées par des perceptrons multi couches. Ces méthodes requièrent des hypothèses sur les lois des sources. Généralement, on les suppose gaussiennnes. Bien que coûteuses du point de vue algorithmique, ces méthodes permettent de reconstituer des signaux provenant de mélanges différents du cas post non linéaire.
- MISEP [4]
  Cette méthode développée par Almeida est une généralisation du principe d'InfoMax dans le cadre de mélanges non linéaires. La fonction non linéaire est alors

paramétrée par un perceptron multi couches ce qui rend possible la maximisation du critère défini par le même procédé que dans la méthode InfoMax. L'auteur se base sur l'hypothèse que les non linéarités vérifient certaines propriétés de régularité.

- Géométriques [6, 70]
  Dans sa thèse, [6], Babaie-Zadeh développe une méthode géométrique basée sur l'hypothèse que les sources ont des distributions à support compact. Ainsi, la méthode permet de détecter les contours du nuage de points formées par les observations, et d'en extraire la non linéarité.
- Cartes de Kohonen (SOM, GMT) Cette méthode est sûrement l'une des premières développée pour les mélanges non linéaires.
  La méthode SOM (Self-Organizing Map) [55] a pour but de trouver une représentation des données dans laquelle ces dernières auront une distribution uniforme sur une grille rectangulaire et préservant le mieux possible certaines structures des données. Alors, en constatant que deux variables uniformes indépendantes admettent une distribution conjointe de support rectangulaire, on peut envisager cette méthode dans le but de trouver des variables indépendantes et de densité uniforme [37]. Cette méthode présente certaines difficultés en ce qui concerne l'hypothèse de densité uniforme et les coûts de calculs.
  La méthode GTM (Generative Topographic Mapping) a été introduite par Bishop *et. al.*, elle peut être vue comme un prolongement de la méthode SOM. En effet, dans la méthode GTM, les distributions uniformes conjointes des sources sont modélisées par des sommes de fonctions de Dirac et le mélange non linéaire par une base de fonctions construite à partir des gaussiennes. Puis, dans [56], on présente une version modifiée de la méthode GTM afin de prendre en considération d'autres types de distributions de sources. On peut consulter aussi [37], pour plus de détails sur ces deux méthodes et des comparaisons expérimentales.
- Alternating Conditional Expectation : ACE [79]
  Introduite par Breiman *et. al.* [11] dans le cadre de tests d'indépendance, Ziehe *et. al.* ont utilisé ce principe pour obtenir une méthode de séparation aveugle de sources dans le cadre de mélanges post non linéaires. Ceux-ci proposent de prendre pour inverser les non linéarités du mélange, les fonctions réalisant le maximum de corrélation.
  Dans le cas d'un mélange de deux sources, on pose,

$$\begin{aligned} x_1 &= f_1(a_{11}s_1 + a_{12}s_2) \\ x_2 &= f_2(a_{21}s_1 + a_{22}s_2) \end{aligned}$$

Les approximations des inverses des non linéarités, notées g_1 et g_2, sont alors

définies en cherchant le maximum de $corr(g_1(x_1), g_2(x_2))$.
Comme le font remarquer Breiman *et. al.* [11], Kolmogorov a démontré que si $a_{11}s_1 + a_{12}s_2$ et $a_{21}s_1 + a_{22}s_2$ ont une densité conjointe gaussienne, alors $g_1 \circ f_1$ et $g_2 \circ f_2$ seront linéaires. Dans [79], les auteurs font alors l'hypothèse que $a_{11}s_1 + a_{12}s_2$ et $a_{21}s_1 + a_{22}s_2$ ont une densité conjointe proche de la gaussienne.
Cette méthode possède donc l'avantage de séparer le problème de séparation aveugle de sources dans le cadre de mélanges post non linéaires en deux parties distinctes. Dans l'étape d'inversion des non linéarités, Ziehe *et. al.* n'utilisent pas l'indépendance des sources, mais seulement une approximation de loi gaussienne à la sortie d'un mélange linéaire.

- Information Mutuelle : [69, 6]
  Taleb et Babaie-Zadeh ont exploité les propriétés de l'information mutuelle afin de dégager une méthode de séparation de source par minimisation d'un critère bien choisi. Dans un premier temps, Taleb a exploité les propriétés du mélange post non linéaire pour définir le critère à optimiser. Dans un deuxième temps, Babaie-Zadeh a constaté certaines erreurs qui peuvent être très importantes dans les estimations et a alors envisagé une approche plus générale.

4.4 Méthodes dérivées de l'information mutuelle et de la mesure de dépendance quadratique

Dans le chapitre 2 nous avons introduit deux mesures de dépendance qui peuvent avoir des comportements très différents, comme nous allons le voir. Grâce à ces deux mesures qui permettent de caractériser des variables indépendantes, nous envisageons l'étude d'algorithmes de séparation aveugle de sources dans le cadre particulier de mélanges post non linéaires.

De par la structure des mélanges post non linéaires, définition 4.1.1, nous sommes contraints d'envisager une méthode d'optimisation traitant séparément la minimisation de la partie linéaire et celle de la partie non linéaire. En effet, comme l'a fait remarqué Taleb, [67], la structure des mélanges post non linéaires ne forme pas un groupe et donc ne permet pas de procéder à une optimisation globale du système.

Après avoir particularisé la variation $\mathbf{\Delta}$ du chapitre 3 au cas des mélanges post non linéaires (paragraphe 4.4.1), nous présenterons différents prolongements adaptés à ce type de mélanges particuliers.

Dans le contexte des mélanges post non linéaires, la minimisation de la partie linéaire sera tout à fait semblable au cadre des mélanges linéaires. Cependant, les non linéarités ne sont généralement pas paramétriques. Ceci introduit une difficulté supplémentaire quant à la recherche de ces dernières par des méthodes de minimisation.

Nous expliciterons alors dans un premier temps la forme du gradient relatif de la

partie linéaire (paragraphe 4.4.2).

Dans un deuxième temps, nous nous intéresserons à la partie non linéaire (paragraphe 4.4.3). Dans ce cadre là, nous nous intéresserons à trois différentes méthodes. Nous expliciterons une approche non paramétrique déjà développée par Taleb [67] et Babaie-Zadeh [6]. Puis toujours dans un contexte non paramétrique, nous introduirons une nouvelle méthode basée sur une propriété d'invariance des critères utilisés quand on ajoute une constante aux non linéarités. Enfin, nous développerons une approche paramétrique utilisant les particularités des mélanges post non linéaires. Nous détaillerons une méthode semi-paramétrique basée sur une approximation des non linéarités par des fonctions affines par morceaux. Pour finir, dans les paragraphes 4.4.4, 4.4.5, 4.4.6 et 4.4.7, nous décrirons les algorithmes de minimisation pour ces différentes approches.

Notons que pour ces différentes parties, on détaillera les deux stratégies «Estimer ensuite» et «Estimer d'abord» (c.f. section 3.1) et on explicitera les différences.

4.4.1 $\boldsymbol{\Delta}$ dans le cas de mélanges post non linéaires

Dans le cadre particulier des mélanges post non linéaires, nous sommes en mesure d'expliciter les fonctions $\boldsymbol{\Delta}$ intervenant dans le calcul du gradient relatif (c.f. chapitre 3).
La résolution du problème de séparation aveugle de sources dans le cadre de mélanges post non linéaires consiste alors en la minimisation d'un critère de dépendance, ici l'information mutuelle ou la mesure de dépendance quadratique, sous la contrainte que les observations $X_1, \ldots, X_K$ proviennent d'un mélange post non linéaire de sources $S_1, \ldots, S_K$ indépendantes. Notre problème se résume alors de la manière suivante:

on cherche K applications $g_1, g_2, \ldots, g_K$ et une matrice $\mathbf{B}$ telles que les variables $Y_1, Y_2, \ldots, Y_K$ définies par,

$$Y_i = \sum_{k=1}^{K} \mathbf{B}_{ik} Z_k, \text{ où } Z_k = g_k(X_k), \text{ pour tout } i = 1, \ldots, K.$$

représentent le minimum du critère considéré et donc aussi une estimation des sources $S_1, S_2, \ldots, S_K$. Ce mélange sera appelé dans la suite structure de séparation.

Généralement, on appellera les fonctions $g_1, \ldots, g_K$ non linéarités.

Stratégie «Estimer ensuite»

En vue de l'utilisation d'une méthode de descente du gradient, nous allons définir les fonctions $\boldsymbol{\Delta}$ de la manière suivante,

$$\mathbf{\Delta}(\mathbf{Y}) = \varepsilon\mathbf{Y} + \mathbf{B}\delta(\mathbf{Z}) + \varepsilon\mathbf{B}\delta(\mathbf{Z})$$

où ε désigne la variation relative de la matrice $\mathbf{B}$ et $\delta = (\delta_1, \ldots, \delta_K)^T$ désigne les variations relatives de chaque fonction inversible $g_1, \ldots, g_K$.

Par hypothèse sur la fonction $\mathbf{\Delta}$, ε est tel que $\mathbf{B} + \varepsilon\mathbf{B}$ est inversible et $\delta_1, \ldots, \delta_K$ sont telles que $g_1 + \delta_1 \circ g_1, \ldots, g_K + \delta_K \circ g_K$ sont aussi inversibles.

Rappelons que la variation $\mathbf{\Delta}$ ne peut être définie que séparément par rapport à la partie linéaire et la partie non linéaire du système à cause de la structure des mélanges post non linéaires qui ne forme pas un groupe, [69].

Nous définissons alors les notations suivantes,

- pour tout $i,\ 1 \leq i \leq K,\ \Gamma_i(\mathbf{Y}) = \sum_{k=1}^{K} \varepsilon_{ik} Y_k + \sum_{k=1}^{K} \mathbf{B}_{ik}\delta_k(g_k(X_k))$
- pour tout $i,\ 1 \leq i \leq K,\ \Lambda_i(\mathbf{Y}) = \sum_{k=1}^{K}\sum_{j=1}^{K} \varepsilon_{ij}\mathbf{B}_{jk}\delta_k(g_k(X_k))$

Alors, nous pouvons écrire, pour tout $i = 1, \ldots, K$,

$$\mathbf{\Delta}_i(\mathbf{Y}) = \Gamma_i(\mathbf{Y}) + \Lambda_i(\mathbf{Y}).$$

Stratégie «Estimer d'abord»

En ce qui concerne l'étude du gradient dans le cadre du critère estimé, on a noté $\mathbf{\Delta}$ la variation que l'on fait subir au critère afin d'en déduire son développement. De la même manière, nous particularisons, dans le cadre de mélanges post non linéaires, le vecteur $\mathbf{\Delta}$ par,

$$\mathbf{\Delta}_i(\mathbf{Y}(j)) = \Gamma_i(\mathbf{Y}(j)) + \Lambda_i(\mathbf{Y}(j)),$$

pour tout $i = 1, \ldots, K$ et pour tout $j = 1, \ldots, N$.
où

- pour tout $i = 1, \ldots, K$, $X_i(1), \ldots, X_i(N)$ est un échantillon de taille N de la variable X_i. Et, pour tout $i = 1, \ldots, K$ et pour tout $j = 1, \ldots, N$,

$$Y_i(j) = \sum_{k=1}^{K} \mathbf{B}_{ik} Z_k(j)$$

  et $Z_k(j) = g_k(X_k(j))$ pour tout $k = 1, \ldots, K$ et pour tout $j = 1, \ldots, N$.
- pour tout $i,\ 1 \leq i \leq K,$ et pour tout $j,\ 1 \leq j \leq N$,

$$\Gamma_i(\mathbf{Y}(j)) = \sum_{k=1}^{K} \varepsilon_{ik} Y_k(j) + \sum_{k=1}^{K} \mathbf{B}_{ik}\delta_k(g_k(X_k(j)))$$

– pour tout $i,\ 1 \leq i \leq K,$ et pour tout $j,\ 1 \leq j \leq N,$

$$\Lambda_i(\mathbf{Y}(j)) = \sum_{k=1}^{K} \sum_{l=1}^{K} \varepsilon_{il} \mathbf{B}_{lk} \delta_k(g_k(X_k(j)))$$

4.4.2 Gradient de la partie linéaire

Dans la mesure où la partie linéaire du mélange correspond à une multiplication par une matrice, cette partie est déjà paramétrée.

Le gradient relatif de la partie linéaire s'écrira alors de la manière suivante, quelque soit le critère utilisé, théorique ou empirique.

$$\varepsilon \mapsto \sum_{i \neq j=1}^{K} \sum \varepsilon_{ij} \mathcal{S}_{ij}^{L}$$

où l'expression de $\mathcal{S}_{ij}^{L}$ diffère selon qu'on utilise l'information mutuelle ou la dépendance quadratique. Nous préciserons son expression exacte dans les sections suivantes 4.4.8 et 4.4.9.

Remarque 4.4.1 *On vérifie facilement que dans les cas considérés ci-après, pour tout* $i = 1, \ldots, K$, $\mathcal{S}_{ii}^{L} = 0$.

4.4.3 Gradient de la partie non linéaire

Les différentes approches que nous avons mises en place s'appliquent naturellement à différents critères. Par conséquent, nous présentons ces méthodes dans un cadre tout à fait général puis nous particulariserons ces approches avec les critères utilisés.

Approche non paramétrique

Stratégie «Estimer ensuite»:

Cette méthode a été introduite initialement par Taleb [69], puis reprise par Babaie-Zadeh [6] en utilisant des critères dérivés de l'information mutuelle. Nous donnons ici les détails de cette méthode sans préciser le critère utilisé.

Le gradient relatif de la partie non linéaire du critère sous la contrainte de mélange post non linéaire s'écrit,

$$\delta_1, \ldots, \delta_K \ \mapsto \ \sum_{k=1}^{K} E\left\{\delta_k(Z_k) \mathcal{S}_k^{NL}(Z_k)\right\} \tag{4.1}$$

où $\mathcal{S}_k^{NL}$ est issue des expressions des gradients relatifs obtenus dans la section 3.2. On en déduit les équations d'estimation correspondantes,

$$\left\{ \begin{array}{rcll} \mathcal{S}_{ij}^{L} & = & 0, & 1 \leq i \neq j \leq K. \\ \mathcal{S}_{k}^{NL} & = & 0, & 1 \leq k \leq K. \end{array} \right.$$

Afin de pouvoir implémenter cette méthode, il est nécessaire de procéder à l'estimation des différentes quantités intervenant dans l'expression du gradient. Une fois le gradient estimé, on obtient une représentation implicite des non linéarités dans un espace de dimension finie. Cependant l'expression du gradient estimé est différente en fonction du critère utilisé (c.f. paragraphes 3.2.1 et 3.2.2). La fonction effectivement minimisée par l'algorithme n'est donc pas le critère théorique, puisque la descente se fait sur une *estimation* du gradient et non sur le gradient exact. Cette méthode revient, en réalité, à chercher les solutions des équations d'estimation.

Nous verrons aussi, sections 4.4.4, 4.4.5, 4.4.6 et 4.4.7, que dans l'application des algorithmes il est préférable d'avoir une bonne connaissance du critère minimisé.

Nous détaillerons les estimations utilisées dans les paragraphes, 4.4.8 et 4.4.9, consacrés à l'utilisation d'un critère particulier. De plus nous détaillerons certains aspects théoriques des estimations au cours du chapitre 5.

Stratégie «Estimer d'abord» :

On rappelle ici que l'on a noté,
pour tout $i = 1, \ldots, K$, $X_i(1), \ldots, X_i(N)$ est un échantillon de taille N de la variable X_i. Et alors pour tout $i = 1, \ldots, K$ et pour tout $j = 1, \ldots, N$,

$$Y_i(j) = \sum_{k=1}^{K} \mathbf{B}_{ik} Z_k(j)$$

et $Z_k(j) = g_k(X_k(j))$ pour tout $k = 1, \ldots, K$ et pour tout $j = 1, \ldots, N$.

Le gradient relatif de la partie non linéaire sous la contrainte de mélange post non linéaire s'écrit,

$$\delta_1, \ldots, \delta_K \;\; \mapsto \;\; \sum_{k=1}^{K} \widehat{E} \left\{ \delta_k(Z_k) \widetilde{\mathcal{S}}_k^{NL} \right\}$$

où $\widetilde{\mathcal{S}}_k^{NL}$, découle des expressions du gradient obtenu lors de l'utilisation de l'estimation du critère.(c.f. section 3.2). $\widehat{E}$ désigne l'opérateur d'espérance empirique. Nous verrons dans les sections suivantes, 4.4.8 et 4.4.9, que $\widetilde{\mathcal{S}}_k^{NL}$ peut-être vu comme une estimation

particulière de $\mathcal{S}_k^{NL}$.

On remarque que dans l'expression du gradient relatif de la partie non linéaire, on ne connait les non linéarités qu'en les points de l'échantillon. Le problème de minimisation ne s'exprime donc plus qu'en ces points, et il est impossible sans hypothèse supplémentaire de déterminer les non linéarités en dehors de ces points.

Approche non paramétrique utilisant les dérivées des non linéarités

Avec l'utilisation de l'information mutuelle ou de la dépendance quadratique, on remarque que le fait d'ajouter une constante à chaque non linéarité ne modifie pas le résultat. Ceci montre que les critères de dépendance envisagés dépendent seulement des dérivées des non linéarités.

Cette constatation nous conduit à étudier seulement le développement limité des critères en fonction des dérivées des non linéarités $g_1, \ldots, g_K$. Nous sommes donc naturellement amenés à écrire les fonctionnelles du gradient en termes des dérivées $\delta'_1, \ldots, \delta'_K$ de $\delta_1, \ldots, \delta_K$.

Stratégie «Estimer ensuite»:

De même que précédemment $\delta_1, \ldots, \delta_K$ représentent les variations relatives de chaque fonction inversible $g_1, \ldots, g_K$.

On suppose de plus que $\delta_1, \ldots, \delta_K$ sont telles que $g_1 + \delta_1 \circ g_1, \ldots, g_K + \delta_K \circ g_K$ sont aussi inversibles.

Remarquons la relation suivante, pour tout $k = 1, \ldots, K$,

$$\delta_k(Z_k) = \int \mathbb{1}_+(Z_k - z)\delta'_k(z)dz$$

où $\mathbb{1}_+$ désigne la fonction indicatrice sur la demi-droite réelle $[0, +\infty)$.

Nous en déduisons alors les expressions du gradient en fonction des dérivées $\delta'_1, \ldots, \delta'_K$.

Le gradient relatif de la partie non linéaire du critère sous la contrainte de mélange post non linéaire s'écrit alors,

$$\delta'_1, \ldots, \delta'_K \;\mapsto\; \sum_{k=1}^{K} \int \mathcal{R}_k^{NL}(z)\delta'_k(z)dz \tag{4.2}$$

où $\mathcal{R}_k^{NL}(z) = E\left\{\mathbb{1}_+(Z_k - z)\mathcal{S}_k^{NL}(Z_k)\right\}$ est issu des expressions des gradients relatifs obtenus dans la section 3.2.

On en déduit alors les équations d'estimation correspondantes,

$$\left\{ \begin{array}{rcll} \mathcal{S}_{ij}^{L} & = & 0, & 1 \leq i \neq j \leq K. \\ \mathcal{R}_{k}^{NL} & = & 0, & 1 \leq k \leq K. \end{array} \right.$$

Stratégie «Estimer d'abord»:

Nous utilisons ici aussi le fait que les fonctions $g_1, \ldots, g_K$ ne sont déterminées qu'à une constante près. D'autre part, nous avons remarqué que dans le cas de la minimisation du critère estimé, celui-ci ne dépend des valeurs des fonctions g_k qu'aux points des observations, nous en déduisons donc que la fonctionnelle du gradient ne dépend que de la moyenne des δ_k' dans chacun des intervalles $[Z_k(n:N), Z_k(n+1:N)[$ pour $n = 1, \ldots, N-1$, en utilisant les statistiques d'ordre [1].

En effet, nous avons les égalités suivantes,

$$\begin{aligned} \frac{1}{N}\sum_{n=1}^{N} \delta_k(Z_k(n))\widetilde{\mathcal{S}}_k^{NL}(n) &= \frac{1}{N}\sum_{n=1}^{N} \delta_k(Z_k(o_{k,n}))\widetilde{\mathcal{S}}_k^{NL}(o_{k,n}) \\ &= \frac{1}{N}\sum_{n=1}^{N-1} [\delta_k(Z_k(n+1:N)) - \delta_k(Z_k(n:N))] \sum_{m=n+1}^{N} \widetilde{\mathcal{S}}_k^{NL}(o_{k,m}) \\ &= \frac{1}{N}\sum_{n=1}^{N-1} \int \sum_{m=n+1}^{N} \widetilde{\mathcal{S}}_k^{NL}(o_{k,m}) \mathbb{1}_{[Z_k(n:N), Z_k(n+1:N)[}(z)\delta_k'(z)dz \end{aligned}$$

On suppose que les estimateurs sont choisis de telle sorte qu'ils soient invariants par translation, on a alors,

$$\widehat{E}\{\widetilde{\mathcal{S}}_k^{NL}\} = 0$$

Il apparait alors clairement que le gradient ne dépend que de la moyenne des fonctions δ_k' dans chacun des intervalles $[Z_k(n:N), Z_k(n+1:N)[$ pour $n = 1, \ldots, N-1$.

Nous en déduisons alors les fonctionnelles du gradient associées au critère estimé pour la partie non linéaire.

$$\delta_1, \ldots, \delta_K \;\mapsto\; \sum_{k=1}^{K} \int \widetilde{\mathcal{R}}_k^{NL}(z)\delta_k'(z)dz \tag{4.3}$$

1. Pour tout $k = 1, \ldots, K$, $Z_k(1:N), \ldots, Z_k(N:N)$ sont les statistiques d'ordre de Z_k et $o_{k,1}, \ldots, o_{k,N}$ est la permutation telle que $Z_k(n:N) = Z_k(o_{k,n})$

où

$$\widetilde{\mathcal{R}}_k^{NL}(z) = \frac{1}{N}\sum_{n=1}^{N-1}\sum_{m=n+1}^{N} \widetilde{\mathcal{S}}_k^{NL}(o_{k,m})\mathbb{1}_{[Z_k(n:N),Z_k(n+1:N)[}(z)$$

(Cette fonction est donc constante par morceaux sur chaque intervalle défini par l'échantillon des statistiques d'ordre de $\mathbf{Z}$).

preuve :

En effet, cette écriture se justifie par le fait que la fonction $\widetilde{\mathcal{R}}_k^{NL}(z)$ est constante sur chaque intervalle $[Z_k(n:N), Z_k(n+1:N)[$ pour $n = 1, \ldots, N-1$.

■

Cette approche nous conduit naturellement à envisager une approche paramétrique.

Approche paramétrique

Cette méthode originellement envisagée par Taleb, [69], consistait en l'approximation des non linéarités par des réseaux de neurones. Ici, nous rappellerons la méthode développée par Taleb en donnant une expression générale valable pour toute paramétrisation. Puis nous énoncerons deux paramétrisations, adaptées au cadre des mélanges post non linéaires, une nouvelle basée sur une approximation des non linéarités par des fonctions linéaires par morceaux, l'autre, introduite par Pham [59] et basée sur l'utilisation de la fonction quantile.

Stratégie «Estimer ensuite»:

Cette méthode consiste en la paramétisation de chaque g_k par un vecteur θ_k, on note alors les non linéarités, g_{k,θ_k}.
Afin de pouvoir faire le lien avec les notations précédentes, nous nous intéressons au développement limité de la fonction g_{k,θ_k} par rapport à θ_k:

$$g_{k,\theta_k+\mathrm{d}\theta_k} - g_{k,\theta_k} = g'_{k,\theta_k}\mathrm{d}\theta_k$$

où g'_{k,θ_k} désigne la dérivée de g_{k,θ_k} par rapport à θ_k. Donc, par analogie avec les notations précédentes, nous allons remplacer δ_k par $\dot{g}_{k,\theta_k}\mathrm{d}\theta_k$ où $\dot{g}_{k,\theta_k} = g'_{k,\theta_k} \circ g^{-1}_{k,\theta_k}$.

Le gradient relatif du critère sous la contrainte de mélange post non linéaire s'écrit alors,

$$\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_K \;\mapsto\; \sum_{k=1}^{K} E\left\{\dot{g}_{k,\theta_k}(Z_k)\mathcal{S}_k^{NL}(Z_k)\right\}\mathrm{d}\theta_k$$

où $\mathcal{S}_k^{NL}$ est issue des expressions des gradients relatifs obtenus dans la section 3.2, et $g_{k,\theta_k} = g'_{k,\theta_k} \circ g^{-1}_{k,\theta_k}$ avec g'_{k,θ_k} la dérivée de g_{k,θ_k} par rapport à θ_k.

On en déduit alors les équations d'estimation correspondantes,

$$\left\{ \begin{array}{rcll} \mathcal{S}_{ij}^{L} & = & 0, & 1 \leq i \neq j \leq K. \\ \mathcal{S}_{k}^{NL} & = & 0, & 1 \leq k \leq K. \end{array} \right.$$

Pour les mêmes raisons que précédemment, nous détaillons aussi une méthode basée sur la minimisation du critère estimé.

Stratégie «Estimer d'abord»:

L'utilisation du critère estimé ne change pas les expressions générales du gradient relatif, mais dans certaines circonstances, les algorithmes seront différents par rapport aux estimations envisagées.

Le gradient relatif de la partie non linéaire s'écrit,

$$\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_K \;\; \mapsto \;\; \sum_{k=1}^{K} \widehat{E} \left\{ \dot{g}_{k,\theta_k}(Z_k) \widetilde{\mathcal{S}}_k^{NL} \right\} \mathrm{d}\theta_k$$

où $\widetilde{\mathcal{S}}_k^{NL}$ est issue des expressions des gradients relatifs obtenus dans la section 3.2, et $g_{k,\theta_k} = g'_{k,\theta_k} \circ g^{-1}_{k,\theta_k}$ avec g'_{k,θ_k} la dérivée de g_{k,θ_k} par rapport à θ_k.

En outre, nous pouvons faire à nouveau la remarque suivante : si l'on ajoute une même constante aux $Z_k(1), \ldots, Z_k(N)$, le critère reste inchangé. Celui-ci ne dépend que des écarts successifs entre les statistiques d'ordre de Z_k. Nous pouvons alors exprimer les fonctionnelles du gradient seulement en fonction de ces écartements,

$$\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_K \;\; \mapsto \;\; \sum_{k=1}^{K} \frac{1}{N} \sum_{n=2}^{N} \left\{ \sum_{m=n}^{N} \widetilde{\mathcal{S}}_k^{NL}(o_{j,m}) \right\} \widetilde{g}_{j,\theta_j}(n) \mathrm{d}\theta_k$$

où $\widetilde{g}_{j,\theta_j}(n) = \dot{g}_j(Z_j(n:N)) - \dot{g}_j(Z_j(n-1:N))$

L'approche paramétrique présentée ci-dessus, s'applique naturellement à toutes sortes de paramétrisation. Nous présentons maintenant deux paramétrisations particulières adaptées aux mélanges post non linéaires. La première méthode, appelée approche semi-paramétrique consiste à estimer les non linéarités par des fonctions linéaires par morceaux. La seconde méthode est basée sur l'utilisation de la fonction quantile.

Approche semi-paramétrique:

Dans la méthode non paramétrique utilisant les invariances des critères par rapport aux non linéarités, seule la dérivée des transformations non linéaires intervient. On décide donc d'approcher $g_1, g_2, \ldots, g_K$ par des fonctions continues linéaires par morceaux dont les intervalles de linéarité sont notés $[\xi_{k,m}, \xi_{k,m+1}]$, $m = 1 \ldots M-1$. On peut supposer sans perte de généralité que les fonctions $g_1, g_2, \ldots, g_K$ sont croissantes. Nous verrons comment cette condition peut-être aisément vérifiée lors de l'implémentation de l'algorithme. Remarquons à présent que la composée de deux fonctions linéaires par morceaux est encore linéaire par morceaux. Nous prendrons donc les fonctions δ_k continues linéaires par morceaux dont les intervalles de linéarité sont $[\zeta_{k,m}, \zeta_{k,m+1}]$, $1 \leq m \leq M-1$, où $\zeta_{k,i} = g_k(\xi_{k,i})$, $i = 1 \ldots M$, $k = 1 \ldots K$. Leurs dérivées s'écrivent

$$\delta'_k(z) = \sum_{m=1}^{M-1} d_{k,m} \mathbb{1}_{[\zeta_{k,m};\zeta_{k,m+1}[}(z).$$

Le vecteur θ_k représente alors simplement les pentes des droites constituant l'approximation de g_k. En remplaçant les dérivées δ'_k définies par la formule ci-dessus, dans l'expression du gradient relatif obtenu dans 4.2, on obtient,

$$d_{k,1}, \ldots, d_{k,M-1} \mapsto \sum_{m=1}^{M-1} \int_{\zeta_{k,m}}^{\zeta_{k,m+1}} d_{k,m} \mathcal{R}_k^{NL}(z) dz, \quad 1 \leq k \leq K.$$

On remarque que dans le cadre du critère estimé, on remplace simplement $\mathcal{R}_k^{NL}$ par $\widetilde{\mathcal{R}}_k^{NL}$.

Remarque 4.4.2 *Dans cette méthode, il se pose le problème du choix des intervalles dans l'approximation des non linéarités. Dans les méthodes implémentées, nous avons choisi des intervalles basés sur les valeurs des échantillons. De plus, nous avons choisi entre 10 et 20 intervalles. En effet, le fait de considérer trop peu d'intervalles conduit à une estimation trop grossière des non linéarités. Mais à l'inverse, le fait de considérer un trop grand nombre d'intervalles fait apparaître un trop grand nombre de degré de liberté, et ne convient pas non plus.*

Une méthode de paramétrisation utilisant la fonction quantile:

Décrivons ici schématiquement la méthode développée par Pham [59].

– Paramétrisation de la fonction quantile de Z_k, pour tout $k = 1, \ldots, K$. On la note Q_{Z_k}.(en utilisant les bases de splines)

– Calcul des échantillons correspondant à la variable Z_k par,

$$Z_k(i:N) = Q_{Z_k}((i-0.5)/N), \qquad \text{pour tout } i = 1, \ldots, N.$$

– On en déduit alors les valeurs des échantillons des variables Y_k, pour tout $n = 1, \ldots, N$ et pour tout $k = 1, \ldots, K$, par,

$$Y_k(n) = \sum_{j=1}^{K} \mathbf{B}_{kj} Z_j(n)$$

où $Z_j(n) = Z_j(o_{j,n}^{-1})$ et $o_{j,n},\ n = 1, \ldots, N$ est la permutation qui permet de trier les échantillons de Z_j.

4.4.4 Approche non paramétrique : Algorithmes

Ayant décrit les différentes expressions des gradients utilisés, nous allons décrire à présent les différents algorithmes de minimisation implémentés.

La minimisation est réalisée grâce à la méthode de descente du gradient. On ajuste les variables $\mathbf{B}$ et $g_1, \ldots, g_K$ par le schéma itératif suivant, λ et μ représentent les constantes de réglage :

$$\begin{cases} \mathbf{B} & \mapsto & \mathbf{B} - \lambda \mathbf{D}\mathbf{B} \\ g_k & \mapsto & g_k - \mu h_k \circ g_k \quad , \ 1 \leq k \leq K. \end{cases}$$

où $\mathbf{D}$ et h_k sont déduites des expressions du gradient (4.1, p. 82). On a,

$$\mathbf{D}_{ij} = \begin{cases} \mathcal{S}_{ij}^{L} & 1 \leq i \neq j \leq K, \\ 0 & \text{sinon.} \end{cases}$$

Quant aux h_k, pour $1 \leq k \leq K$, on les obtient de deux manières différentes en fonction de la métrique adoptée :

- avec la métrique normale (associée au produit scalaire de l'espace de Hilbert des fonctions de carré intégrable par rapport à la mesure de Lebesgue), on a :

$$h_k^n(z) = \mathcal{S}_k^{NL}(z)$$

- avec la métrique probabiliste (associée au produit scalaire de l'espace de Hilbert des fonctions de carré intégrable par rapport à la mesure de probabilité), on a :

$$h_k^p(z) = \frac{1}{p_{Z_k}(z)} \mathcal{S}_k^{NL}(z)$$

Nous en déduisons alors l'algorithme suivant,

Initialisations :
$\mathbf{B}$
$g_1, \ldots, g_K$
Boucle : pour t=1, 2, 3, ... (t indique le numéro de l'itération)
$\widehat{\mathcal{S}_{ij}^L}$ estimation de $\mathcal{S}_{ij}^l$,$i \neq j = 1, \ldots, K$.
$\widehat{h}_k$ estimation de h_k, $k = 1, \ldots, K$.
$Z_k^{(t+1)} = [Z_k^{(t)} - \mu \widehat{h}_k(Z_k^{(t)})]$
$k = 1, \ldots, K$

Normalisation de $Z^{(t+1)}$.

$\mathbf{B}^{(t+1)} = \mathbf{B}^{(t)} - \lambda \hat{\mathbf{D}} \mathbf{B}^{(t)}$.
où $\hat{\mathbf{D}}_{ij} = \begin{cases} \widehat{\mathcal{S}}_{ij}^l & 1 \leq i \neq j \leq K, \\ 0 & \text{sinon.} \end{cases}$

$Y^{(t+1)} = \mathbf{B}^{(t+1)} Z^{(t+1)}$.

Normalisation de $Y^{(t+1)}$.
Jusqu'à convergence.

Dans cet algorithme, il est nécessaire de procéder à des normalisations à chaque étape intermédiaire de progression de l'algorithme. En effet, il est bien connu que le problème de séparation aveugle de source ne peut se résoudre qu'à un facteur d'échelle près et à une permutation près. Pour fixer l'indétermination de facteur d'échelle, nous normalisons les variables Y_i à chaque pas de l'algorithme. D'autre part, dans le cadre des mélanges post non linéaires, nous remarquons que les fonctions g_k ne peuvent être reconstruites qu'à un facteur d'échelle près qui sera alors compensé par un facteur d'échelle sur la matrice et à une constante près car les critères envisagés sont invariants par translation, voir paragraphe 4.3. Pour compenser ces indéterminations nous choisissons ici de normaliser les variables Z_k à chaque pas de l'algorithme. Ceci permet à la fois de fixer le facteur d'échelle à 1 et la constante à 0.

Faisons à présent quelques remarques sur ce qui nous pousse à utiliser la minimisation du critère estimé, stratégie «Estimer d'abord». Tout d'abord, la notion de convergence reste à définir, nous ne pouvons a priori pas savoir à quel moment arrêter l'algorithme. Pour cela, nous pouvons envisager à chaque tour de boucle de l'algorithme de calculer un critère de dépendance qui nous dira si on peut accepter l'hypothèse que les variables sont indépendantes, c'est à dire que l'on a atteint le minimum.

Ensuite, un autre problème apparaît aussi dans le choix des paramètres λ et μ de la descente de gradient. Nous proposons ici de les ajuster en fonction du critère

à minimiser. Autrement dit, nous calculons la variation des variables $\mathbf{Z}$ et $\mathbf{Y}$. Puis, nous ajustons les constantes λ et μ de telle sorte que le critère calculé pour les nouvelles variables $\mathbf{Z}$ et $\mathbf{Y}$ soit effectivement inférieur à celui calculé pour les variables à l'itération précédente. Il vient alors naturellement l'idée de minimiser une estimation du critère au lieu d'utiliser une estimation du gradient relatif, afin de minimiser le critère théorique que nous ne pouvons pas évaluer.

Nous avons fait le constat que l'approche non paramétrique appliquée avec le critère estimé n'est pas adaptée. Nous ne développerons donc les algorithmes implémentés avec le critère estimé que dans le cadre de l'approche non paramétrique utilisant les dérivées des non linéarités et de l'approche paramétrique.

4.4.5 Approche non paramétrique utilisant les dérivées des non linéarités : Algorithmes

La minimisation est réalisée grâce à la méthode de descente du gradient. On ajuste les variables $\mathbf{B}$ et $g'_1, \ldots, g'_K$ par le schéma itératif suivant, λ et μ représentent les constantes de réglage:

$$\begin{cases} \mathbf{B} & \mapsto & \mathbf{B} - \lambda \mathbf{D}\mathbf{B} \\ g'_k & \mapsto & g'_k(1 - \mu h'_k \circ g'_k) \quad , \ 1 \leq k \leq K. \end{cases}$$

(Ceci provient du fait que la variation relative des non linéarités g_k s'écrit, $g_k + \delta_k \circ g_k$.)

où $\mathbf{D}$ et h'_k sont déduites des expressions du gradient. On a

$$\mathbf{D}_{ij} = \begin{cases} \mathcal{S}^L_{ij} & 1 \leq i \neq j \leq K, \\ 0 & \text{sinon.} \end{cases}$$

Quant au h'_k, pour $1 \leq k \leq K$, on les obtient de deux manières différentes en fonction de la métrique adoptée :

- avec la métrique normale (associée au produit scalaire de l'espace de Hilbert des fonctions de carré intégrable par rapport à la mesure de Lebesgue), on a :

$$h'^n_k(z) = \mathcal{R}^{NL}_k(z)$$

- avec la métrique probabiliste (associée au produit scalaire de l'espace de Hilbert des fonctions de carré intégrable par rapport à la mesure de probabilité), on a :

$$h'^p_k(z) = \frac{1}{p_{Z_k}(z)} \mathcal{R}^{NL}_k(z)$$

Comme nous l'avons signalé dans le cadre de l'approche non paramétrique, nous préférons utilisé le critère estimé pour procéder à la minimisation. Dans ce qui suit,

nous écrirons les algorithmes avec l'utilisation du critère estimé. Naturellement, quand on utilise le critère théorique, la structure générale des algorithmes ne sera pas modifiée, seule les définitions des expressions du gradient seront différentes.

Initialisations :
$\mathbf{B}$
$g_1, g_2, \ldots, g_K$
Boucle : pour t=1, 2, 3, ... (t indique le numéro de l'itération)
Calcul de $\mathcal{S}_{ij}^L$, $i \neq j = 1, \ldots, K$.
Calcul de $\widetilde{h}'_k$ à partir de $\widetilde{\mathcal{R}}_k^{NL}$, $k = 1, \ldots, K$.
$Z_k^{(t+1)}(m+1:N) - Z_k^{(t+1)}(m:N) = [Z_k^{(t)}(m+1:N) - Z_k^{(t)}(m:N)]$
$\times \left[1 - \mu \widetilde{h}'_k(Z_k^{(t)}(m:N)\right]$, $k = 1, \ldots, K$, $n = 1, \ldots, N$

Normalisation de $Z^{(t+1)}$.

$\mathbf{B}^{(t+1)} \mapsto \mathbf{B}^{(t)} - \lambda \hat{\mathbf{D}} \mathbf{B}^{(t)}$.
où $\hat{\mathbf{D}}_{ij} = \begin{cases} \mathcal{S}_{ij}^l & 1 \leq i \neq j \leq K, \\ 0 & \text{sinon.} \end{cases}$

$Y^{(t+1)} = \mathbf{B}^{(t+1)} Z^{(t+1)}$.

Normalisation de $Y^{(t+1)}$.
Jusqu'à convergence.

Nous procédons ici aux mêmes normalisations que précédemment. Cependant, comme dans ce cas nous travaillons seulement avec les dérivées des non linéarités, il est aussi possible d'envisager une nouvelle normalisation en imposant aux g'_k de vérifier la condition suivante, $E[\log |g'_k(X_k)|] = 0$ dans le cas de la métrique probabiliste et la condition $\int \delta'_k(z)dz$ dans le cas de la métrique de Lebesgue. Cette quantité est facilement évaluée car on travaille dans cet algorithme avec les dérivées des non linéarités qui sont constantes par morceaux. Enfin, la croissance des non linéarités est simplement contrôlée en imposant que le terme $1 - \mu \widehat{h}'_k(Z_{k,t}^{(m)})$ soit positif, ce qui dépend essentiellement du choix du paramètre μ.

4.4.6 Approche paramétrique: Algorithmes

On se place ici aussi dans le cadre de la minimisation du critère estimé. La minimisation est réalisée grâce à la méthode de descente du gradient. On ajuste les variables

$\mathbf{B}$ et $g_1, \ldots, g_K$ par le schéma itératif suivant, λ et μ représentent les constantes de réglage:

$$\begin{cases} \mathbf{B} & \mapsto & \mathbf{B} - \lambda \mathbf{D}\mathbf{B} \\ \theta_k & \mapsto & \theta_k - \mu\gamma_k \quad , \ 1 \leq k \leq K. \end{cases}$$

où $\mathbf{D}$ et γ_k sont déduites des expressions du gradient. On a

$$\mathbf{D}_{ij} = \begin{cases} \mathcal{S}_{ij}^L & 1 \leq i \neq j \leq K, \\ 0 & \text{sinon.} \end{cases}$$

Quant aux γ_k, pour $1 \leq k \leq K$, on les obtient de la manière suivante:

$$\gamma_k = \widehat{E}\left\{ \dot{g}_{k,\theta_k}(Z_k) \widetilde{\mathcal{S}}_k^{NL} \right\}.$$

Nous en déduisons l'algorithme suivant,

Initialisations :
$\mathbf{B}$
$\theta_1, \ldots, \theta_K$
Boucle :pour t=1, 2, 3, ... (t indique le numéro de l'itération)
Calcul de $\mathcal{S}_{ij}^L$, $i \neq j = 1, \ldots, K$.
Calcul de γ_k à partir de $\widetilde{\mathcal{T}}_k^{NL}$, $k = 1, \ldots, K$.
$\theta_k^{(t+1)} = [\theta_k^{(t)} - \mu \mathrm{d}\gamma_k], \qquad k = 1, \ldots, K$

$Z_k^{(t+1)} = g_{k,\theta_k^{(t+1)}}(X_k), \qquad k = 1, \ldots, K$
Normalisation de $Z^{(t+1)}$.

$\mathbf{B}^{(t+1)} = \mathbf{B}^{(t)} - \lambda \hat{\mathbf{D}} \mathbf{B}^{(t)}$.
où $\hat{\mathbf{D}}_{ij} = \begin{cases} \mathcal{S}_{ij}^l(n) & 1 \leq i \neq j \leq K, \\ 0 & \text{sinon.} \end{cases}$

$Y^{(t+1)} = \mathbf{B}^{(t+1)} Z^{(t+1)}$.

Normalisation de $Y^{(t+1)}$.
Jusqu'à convergence.

4.4.7 Approche semi-paramétrique: Algorithmes

Détaillons ici le cadre particulier de l'approche semi-paramétrique:
La minimisation du critère est réalisée en ajustant cette fois les paramètres $d_{k,m}$, $1 \leq k \leq K,\ 1 \leq m \leq M-1$, selon le schéma itératif suivant,

$$\left\{ \begin{array}{lcl} \mathbf{B} & \mapsto & \mathbf{B} - \lambda \mathbf{D}\mathbf{B} \\ d_{k,m} & \mapsto & d_{k,m} - \mu\gamma_{k,m} \end{array} \right. , \ 1 \leq k \leq K, \ 1 \leq k \leq M-1.$$

où $\mathbf{D}$ et $\gamma_{k,m}$ sont déduites des expressions du gradient. On a

$$\mathbf{D}_{ij} = \begin{cases} \mathcal{S}^L ij & 1 \leq i \neq j \leq K, \\ 0 & \text{sinon.} \end{cases}$$

Quant aux $\gamma_{k,m}$, pour $1 \leq k \leq K$, $1 \leq m \leq M-1$, on les obtient de la manière suivante :

– avec le critère théorique et la métrique de Lebesgue associée au produit scalaire, $< d_{k,.}, d'_{k,.} >= \sum d_{k,m} d'_{k,m} (\zeta_{k,m+1} - \zeta_{k,m}))$, on a :

$$\gamma_{k,m} = \frac{1}{\zeta_k^{(m+1)} - \zeta_k^{(m)}} \int_{\zeta_{k,m}}^{\zeta_{k,m+1}} \mathcal{R}_k^{NL}(z) dz.$$

– avec le critère théorique et la métrique probabiliste, associée au produit scalaire, $< d_{k,.}, d'_{k,.} >= \sum d_{k,m} d'_{k,m} P(\zeta_{k,m} \leq Z_k \leq \zeta_{k,m+1}))$, on a :

$$\gamma_{k,m} = \frac{1}{P(\zeta_{k,m} \leq Z_k \leq \zeta_{k,m+1})} \int_{\zeta_{k,m}}^{\zeta_{k,m+1}} \mathcal{R}_k^{NL}(z) dz.$$

– avec le critère empirique :

$$\gamma_{k,m} = \frac{1}{\zeta_k^{(m+1)} - \zeta_k^{(m)}} \int_{\zeta_{k,m}}^{\zeta_{k,m+1}} \widetilde{\mathcal{R}}_k^{NL}(z) dz.$$

Dans le cadre de l'utilisation du critère théorique, nous avons besoin d'estimer les différentes quantités. Pour plus de détails sur l'expression de ces estimations nous renvoyons aux paragraphes suivants utilisant un critère particulier 4.4.8 et 4.4.9. Et en ce qui concerne l'étude de ces estimations, nous renvoyons au chapitre 5.

Afin de permettre une implémentation plus aisée, nous choisissons de travailler avec les statistique d'ordre, $Z_k(1:N), \ldots, Z_k(N:N)$ de Z_k, pour tout $k = 1, \ldots, K$. De plus, on note $\{o_{k,n}\}$ la permutation définie par $Z_k(m:N) = Z_k(o_{k,m})$, pour tout $m = 1, \ldots, N$.

Initialisations :
$\mathbf{B}$
$g_1, g_2, \ldots, g_K$
Boucle :pour t=1, 2, 3, ... (t indique le numéro de l'itération)
Calcul ou estimation de $\mathcal{S}_{ij}^L$, $i \neq j = 1, \ldots, K$.
Calcul ou estimation de $\gamma_{k,m}$, $k = 1, \ldots, K$.
$Z_k^{(t+1)}(m+1:N) - Z_k^{(t+1)}(m:N) = [Z_k^{(t)}(m+1:N) - Z_k^{(t)}(m:N)] \times [1 - \mu\gamma_{k,m}]$
$k = 1, \ldots, K$, $n = 1, \ldots, N$

Normalisation de $Z^{(t+1)}$.

$\mathbf{B}^{(t+1)} \mapsto \mathbf{B}^{(t)} - \lambda\hat{\mathbf{D}}\mathbf{B}^{(t)}$.
où $\hat{\mathbf{D}}_{ij} = \begin{cases} \mathcal{S}_{ij}^l & 1 \leq i \neq j \leq K, \\ 0 & \text{sinon.} \end{cases}$

$Y^{(t+1)} = \mathbf{B}^{(t+1)} Z^{(t+1)}$.

Normalisation de $Y^{(t+1)}$.
Jusqu'à convergence.

Nous procédons ici aux mêmes normalisations que précédemment.

Nous avons ici présenté de manière tout à fait générale les méthodes que nous avons implémentées par la suite en utilisant un critère particulier. Ces méthodes présentent chacune des avantages et des inconvénients que nous avons observés lors de l'implémentation de celles-ci. Naturellement, le critère utilisé joue aussi un rôle important dans le déroulement de l'algorithme. Par ailleurs, jusqu'ici, nous avons considéré seulement des méthodes de descente de gradient, dans la suite nous verrons que parfois il est possible d'envisager d'autres méthodes, comme des méthodes de gradient conjugué ou de quasi Newton.

4.4.8 Méthodes basées sur l'information mutuelle

Tout d'abord, nous détaillerons principalement certaines de ces méthodes en utilisant un critère basé sur l'information mutuelle. Ce critère a déjà été utilisé par Babaie-Zadeh [6] dans le cadre des mélanges post non linéaires avec une approche non paramétrique. Nous nous intéresserons parallèlement au critère utilisé par Taleb [69]. Ce dernier est obtenu par transformation de l'information mutuelle. Nous montrerons alors les liens que l'on peut établir vis à vis de leurs gradients. Nous verrons alors clairement les différences induites par le critère utilisé par Babaie-Zadeh et celui

utilisé par Taleb.

Dans le chapitre 3, nous avons étudié le gradient de l'information mutuelle. Cependant, dans [69], Taleb remarque que dans le cadre de mélanges post non linéaires, il est possible de simplifier l'expression de l'information mutuelle afin de ne faire apparaître que des termes d'entropie d'une seule variable. Il en déduit alors le critère suivant,

Définition 4.4.1 (Critère de l'IM décomposée dans le cas de mélanges PNL) *Soient K applications $g_1, g_2, \ldots, g_K$ et une matrice $\mathbf{B}$. On définit les variables $Y_1, Y_2, \ldots, Y_K$ par,*

$$Y_i = \sum_{k=1}^{K} \mathbf{B}_{ik} Z_k, \;\; \textit{où } Z_k = g_k(X_k), \;\; \textit{pour tout } i = 1, \ldots, K.$$

On notera alors C le critère utilisé par Taleb pour résoudre le problème de séparation de sources, défini par,

$$C(\mathbf{B}, g_1, \ldots, g_K) = \sum_{i=1}^{K} (H(Y_i) - H(Z_i)) - \log |\det \mathbf{B}|$$

Dans [69], Taleb procède alors à la résolution du problème de séparation de sources en minimisant ce critère.

Nous allons décrire maintenant les différentes approches présentées dans la section 4.4.3, en utilisant l'information mutuelle. Les deux critères, information mutuelle et C, sont égaux à des constantes près. Les fonctionnelles du gradient calculé à partir des critères théoriques sont donc égales.

Approche non paramétrique

Calcul du gradient relatif:

Nous explicitons cette méthode, initialement développée par Babaie-Zadeh [6], avec les notations introduites précédemment,

Soient $T_1, T_2, \ldots, T_K$, K variables aléatoires réelles telles que :
le vecteur aléatoire $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ est absolument continu par rapport à la mesure de Lebesgue sur $\mathbb{R}^K$.

Le gradient relatif de l'information mutuelle sous la contrainte de mélange post non linéaire s'écrit alors,

- Gradient relatif par rapport à $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{i \neq k=1}^{K} \sum \varepsilon_{ij} \underbrace{E\{Y_i \beta_j(\mathbf{Y})\}}_{\mathcal{S}_{ij}^{L}}$$

- Gradient relatif par rapport à $\delta_1, \ldots, \delta_K$:

$$\delta_1, \ldots, \delta_K \;\; \mapsto \;\; \sum_{k=1}^{K} E \left\{ \delta_k(Z_k) \underbrace{\sum_{i=1}^{K} E\left[\beta_i(\mathbf{Y}) B_{ik} | Z_k\right]}_{\mathcal{S}_k^{NL}(z)} \right\}$$

Comme nous l'avons évoqué dans le chapitre 3, Babaie-Zadeh [6] propose différentes estimations de la fonction β. Nous ne détaillerons pas ici les différentes estimations utilisées. Nous reviendrons sur ces différentes notions lors du développement des méthodes basées sur le critère estimé (chapitre 5).

Afin de faire le lien avec le critère C introduit par Taleb, rappelons la définition de la fonction $\beta_{\mathbf{Y}}$.

$$\beta_{\mathbf{Y}}(\mathbf{y}) = \psi_{\mathbf{Y}}(\mathbf{y}) - \phi_{\mathbf{Y}}(\mathbf{y}),$$

et

$$\psi_{\mathbf{Y}}(\mathbf{y}) = (\psi_{Y_1}(y_1), \ldots, \psi_{Y_K}(y_K))$$

avec

$$\psi_{Y_i}(y_i) = -\frac{p'_{Y_i}(y_i)}{p_{Y_i}(y_i)},$$

$$\phi_{\mathbf{Y}}(\mathbf{y}) = (\phi_1(\mathbf{y}), \ldots, \phi_K(\mathbf{y}))$$

avec

$$\phi_i(\mathbf{y}) = -\partial_i \log p_{\mathbf{Y}}(\mathbf{y})$$

D'autre part, on remarque que,

Lemme 4.4.1 *Pour tous* $i, j = 1, \ldots, K$,

$$\left\{ \begin{array}{rcll} E[Y_i \phi_j(\mathbf{Y})] & = & 0, & i \neq j, \\ & = & 1, & i = j \end{array} \right.$$

et
pour tout $k = 1, \ldots, K$

$$\sum_{i=1}^{K} E\left[\phi_i(\mathbf{Y})B_{ik}|Z_k = z\right] = \psi_{Z_k}(Z_k)$$

preuve: c.f. Annexe B.2
De plus, rappelons l'expression du gradient relatif du critère C utilisé par Taleb,

- Le gradient relatif du critère C par rapport à $\mathbf{B}$ est représenté par la forme linéaire:

$$\varepsilon \mapsto \sum_{i\neq k=1}^{K}\sum \varepsilon_{ik} E[\psi_{Y_i}(Y_i)Y_k] \tag{4.4}$$

- Le gradient relatif du critère C par rapport à $\delta_1, \delta_2, \ldots, \delta_K$ est représenté par les fonctionnelles linéaires:

$$\text{Pour tout } k,\ 1 \leq k \leq K,\ \delta_k \mapsto E\left\{\delta_k(Z_k)E\left[\sum_{i=1}^{K}\mathbf{B}_{ik}\psi_{Y_i}(Y_i) - \psi_{Z_k}(Z_k)\middle| Z_k\right]\right\} \tag{4.5}$$

Avec ces deux remarques nous déduisons que les gradients relatifs du critère C utilisé par Taleb et de l'information mutuelle utilisée par Babaie-Zadeh sont égaux. Nous avons montré cette relation dans le cadre d'une approche non paramétrique, mais celle-ci est aussi vérifiée dans le cadre des différentes approches présentées dans la section précédente. Dans la suite nous nous contenterons alors d'écrire les fonctionnelles du gradient en utilisant l'information mutuelle. Nous renvoyons à [2] et [3] pour le détail des calculs en utilisant le critère C.

Points où le gradient s'annule:

Faisons à présent quelques remarques en ce qui concerne ce gradient. Nous avons montré (c.f. [1]) que l'annulation du gradient relatif conduit aux équations d'estimation suivantes,

$$E[Y_j\psi_{Y_i}(Y_i)] = 0, \qquad 1 \leq i \neq j \leq K. \tag{4.6}$$

$$E\left[\sum_{i=1}^{K}\mathbf{B}_{ik}\psi_{Y_i}(Y_i)\middle| Z_k\right] = \psi_{Z_k}(Z_k), \qquad 1 \leq k \leq K. \tag{4.7}$$

Nous avons aussi remarqué (c.f. [1]) que lorsque les variables $Y_1, \ldots, Y_K$ sont indépendantes alors effectivement, ces équations d'estimation sont vérifiées. Par contre, il est intéressant de remarquer qu'il existe des configurations permettant l'annulation du gradient sans que les variables $Y_1, \ldots, Y_K$ soient indépendantes.

En effet, que ce soit dans le cas d'un mélange linéaire ou non linéaire, nous allons montrer qu'il est possible de trouver des situations particulières où le gradient est nul et pourtant les variables à la sortie de la structure de séparation ne sont pas indépendantes. Ces points correspondent donc à des minima locaux, maxima locaux ou points selles.

Nous nous placerons ici dans le cas d'un mélange linéaire.

On cherche alors dans ce contexte une matrice $\mathbf{B}$ de telle sorte que $C(\mathbf{B})$ soit minimale.

Considérons le cas suivant, on prend un mélange linéaire avec comme matrice de mélange,

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

et deux sources S_1 et S_2 de même densité symétrique. De plus, on note $\mathbf{X} = \mathbf{AS}$.

On montre alors que

Lemme 4.4.2

$$E[X_j \psi_{X_i}(X_i)] \;=\; 0, \qquad 1 \leq i \neq j \leq K. \tag{4.8}$$

preuve:

On montre tout d'abord que $E[X_2|X_1] = cte$ (c.f. [43]).

Pour tout x réel,

$$E[X_2|X_1 = x] = \int y \frac{p_{X_1,X_2}(x,y)}{p_{X_1}(x)} dy$$

Alors en prenant la dérivée par rapport à x,

$$(E[X_2|X_1 = x])' = \int y \frac{\partial}{\partial_x} \frac{p_{X_1,X_2}(x,y)}{p_{X_1}(x)} dy$$

Donc, comme les densités de S_1 et S_2 sont symétriques et identiques, et

$$p_{X_1,X_2}(x,y) = \frac{1}{2} f_{S_1}(x+y) f_{S_2}(x-y),$$

cette dernière est aussi symétrique par rapport à chacune de ses variables.

Alors,

$$\frac{\partial}{\partial_x}\frac{p_{X_1,X_2}(x,y)}{p_{X_1}(x)}$$

est symétrique par rapport à y et donc $(E[X_2|X_1 = x])' = 0$

Il s'en suit que, $E[\psi_{X_1}(X_1)X_2] = 0$. En effet, on remarque que

$$E[\psi_Y(Y)Z] = \int (E[Z|Y = x])' p_Y(x)dx.$$

$$\begin{aligned} E[\psi_Y(Y)Z] &= -\int \frac{p'_Y(y)}{p_Y(y)} E[Z|Y = y]p_Y(y)dy \\ &= -\int p'_Y(y)E[Z|Y = y]dy \end{aligned}$$

Puis, par une intégration par partie, comme la densité de Y s'annule à l'infini,

$$E[\psi_Y(Y)Z] = -\int (E[Z|Y = y])' p_Y(y)dy$$

Nous avons alors montré qu'il existe bien des solutions parasites qui peuvent nous empêcher de déterminer effectivement le minimum global du critère C.

Dans le but d'utiliser éventuellement les termes d'ordre 2 des développements limités, nous avons obtenu l'expression des termes du Hessien de chaque critère. Cependant, nous n'avons pas jusqu'à présent développée ce genre de méthodes. Les expressions du Hessien sont détaillées en annexe C.1.

Approche non paramétrique utilisant les dérivées des non linéarités

Comme nous l'avons évoqué précédemment, il n'est pas du tout adéquat d'envisager une méthode non paramétrique pour minimiser un critère estimé. Nous allons à présent nous concentrer sur la méthode non paramétrique utilisant les dérivées des non linéarités. Celle-ci va nous permettre de faire le lien entre l'utilisation du critère théorique et du critère estimé.

Grâce aux développements limités obtenus dans la section 3.2 et au développement de l'approche non paramétrique utilisant les dérivées des non linéarités (paragraphe 4.4.3), nous déduisons que le gradient relatif de l'information mutuelle s'écrit de la même manière que l'on utilise le critère théorique ou bien une estimation. Par contre, ils diffèrent par l'expression des différentes quantités mises en jeu. En effet, rappelons ici les expressions des fonctionnelles du gradient relatif de l'information mutuelle

théorique:

- Gradient relatif par rapport à $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{i \neq k=1}^{K} \sum \varepsilon_{ij} \underbrace{E\{Y_i \beta_j(\mathbf{Y})\}}_{\mathcal{S}_{ij}^L}$$

- Gradient relatif par rapport à $\delta'_1, \ldots, \delta'_K$:

$$\delta'_1, \ldots, \delta'_K \mapsto \sum_{k=1}^{K} \sum_{n=1}^{N} \int_{Z_k(n:N)}^{Z_k(n+1:N)} \underbrace{E\left[\mathbb{1}_+(Z_k - z) \sum_{i=1}^{K} \beta_i(\mathbf{Y}) B_{ik}\right]}_{\mathcal{R}_k^{NL}(z)} \delta'_k(z) dz$$

Par comparaison, détaillons ici les expressions des fonctionnelles du gradient dans le cas de l'utilisation du critère estimé, stratégie «Estimer d'abord» :

- Gradient relatif par rapport à $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{i \neq k=1}^{K} \sum \varepsilon_{ij} \underbrace{\widehat{E}\left[\widehat{\beta}_k^{mk}(\mathbf{Y}) Y_k\right]}_{\widetilde{\mathcal{R}}_{ij}^L}$$

- Gradient relatif par rapport à $\delta'_1, \ldots, \delta'_K$:

$$\delta'_1, \ldots, \delta'_K \mapsto \sum_{k=1}^{K} \sum_{n=1}^{N} \int_{Z_k(n:N)}^{Z_k(n+1:N)} \underbrace{\widehat{E}\left[\mathbb{1}_+(Z_k - z) \sum_{i=1}^{K} \widehat{\beta}_i^m(\mathbf{Y}) B_{ik}\right]}_{\widetilde{\mathcal{R}}_k^{NL}(z)} \delta'_k(z) dz$$

où $\widehat{\beta}^m$ est définie dans le lemme 3.2.15.

Nous voyons apparaître ici clairement les similitudes des deux calculs. Mais d'après le chapitre 3, la fonction β n'est pas du tout estimée de la même manière. Cette différence provient naturellement du fait que l'on ne minimise pas la même quantité.

A l'aide de ces deux représentations, nous pouvons faire l'analogie aussi avec les deux métriques que nous avons considérées dans le cadre de la minimisation du critère théorique.

En effet, en utilisant naturellement les estimations adéquates, la métrique probabiliste est associée à l'estimation par l'opérateur d'espérance empirique et la métrique de Lebesgue est associée à l'estimation en utilisant une intégration numérique.

La méthode non paramétrique a été implémentée par Babaie-Zadeh [6]. D'autre part, Pham [61] a implémenté la méthode utilisant l'approximation de l'information mutuelle par une discrétisation de l'intégrale. Les autres méthodes ont été implémentées par nos soins.

Remarque 4.4.3 *Avec le lemme 3.2.6, Babaie-Zadeh propose une estimation particulière des fonctions scores marginales à partir des fonctions scores jointes :* $\widehat{\psi}_i = E[\widehat{\phi}_i(\mathbf{T})|T_i]$. *En regardant ici les expressions des estimations des gradients, on montre que pour ce choix particulier des estimateurs des fonctions scores, le gradient estimé en utilisant la forme complète de l'information mutuelle (comme Babaie-Zadeh) et celui estimé en utilisant le critère C de Taleb coïncident.*

Il est ici aussi intéressant de vouloir considérer les termes d'ordre 2 des développements limités. Nous renvoyons à l'annexe C.1 pour une expression détaillée des Hessien.

Approche semi-paramétrique

Quand on approche les non linéarités par des fonctions continues linéaires par morceaux, on obtient pour expression du gradient,

- Gradient relatif par rapport à $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{i \neq k=1}^{K} \sum E[Y_j \beta_i(Y_i)] \varepsilon_{ik}$$

- Gradient relatif par rapport à $d_{k,1}, \ldots, d_{k,M-1}$:

$$d_{k,1}, \ldots, d_{k,M-1} \mapsto \sum_{m=1}^{M-1} \int_{\zeta_{k,m}}^{\zeta_{k,m+1}} E\left\{ \mathbb{1}_+(Z_k - z) \sum_{i=1}^{K} \beta_i(Y_i) \mathbf{B}_{ik} \right\} dz \qquad 1 \leq k \leq K.$$

Rappelons ici la formule générale des fonctions δ'_k,

$$\delta'_k(z) = \sum_{m=1}^{M-1} d_{k,m} \mathbb{1}_{[\zeta_{k,m};\zeta_{k,m+1}[}(z).$$

Par analogie, nous donnons les expressions du gradient dans le cas de l'utilisation de l'estimation du critère. Dans ce contexte, les fonctions non linéaires sont approchées par des fonctions linéaires par morceaux. Ceci permet alors d'avoir une approximation plus précise surtout en ce qui concerne la régularité des non linéarités.

Le gradient relatif s'écrit alors,

- Gradient relatif par rapport à $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{i \neq k=1}^{K} \sum \widehat{E}[Y_j \widehat{\beta}_i(Y_i)]\varepsilon_{ik}$$

- Gradient relatif par rapport à $d_{k,1}, \ldots, d_{k,M-1}$:

$$d_{k,1}, \ldots, d_{k,M-1} \mapsto \sum_{m=1}^{M-1} d_{k,m} \int_{\zeta_{k,m}}^{\zeta_{k,m+1}} \widehat{E}\left\{ \mathbb{1}_+(Z_k - z) \left(\sum_{i=1}^{K} \widehat{\beta}_i(Y_i)\mathbf{B}_{ik} \right) \right\} dz, \qquad 1 \leq k \leq K.$$

où $\widehat{\beta}$ désigne l'approximation de la fonction score calculée à partir de la dérivée de l'estimation de l'entropie en utilisant soit l'intégration numérique soit l'opérateur d'espérance empirique (paragraphe 3.2.3).

On reconnait ici aussi exactement les expressions du gradient du critère théorique estimé. Cependant, les estimations des fonctions scores sont obtenues à partir de l'estimation du critère.

Précisons dans ce dernier cas les expressions du gradient relatif de l'estimation du critère C de Taleb,

- Gradient relatif par rapport à $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{i \neq k=1}^{K} \sum \widehat{E}[Y_j \widehat{\psi}_{Y_i}(Y_i)]\varepsilon_{ik}$$

- Gradient relatif par rapport à $d_{k,1}, \ldots, d_{k,M-1}$:

$$d_{k,1}, \ldots, d_{k,M-1} \mapsto \sum_{m=1}^{M-1} d_{k,m} \int_{\zeta_{k,m}}^{\zeta_{k,m+1}} \widehat{E}\left\{ \mathbb{1}_+(Z_k - z) \left(\sum_{i=1}^{K} \widehat{\psi}_{Y_i}(Y_i)\mathbf{B}_{ik} - \widehat{\psi}_{Z_k}(Z_k) \right) \right\} dz, \qquad 1 \leq k \leq K.$$

Evoquons, à présent quelques problèmes rencontrés lors de la mise en place de ces méthodes. Dans le chapitre 6, nous illustrerons plus précisément ces méthodes par des simulations.

Problèmes

- Nous sommes confrontés à des problèmes de minima locaux. En effet, nous avons montré que dans le cas de mélanges linéaires il existe des points où le gradient s'annule sans que la solution soit atteinte. Ceci nous pousse à penser que dans certaines situations peut-être assez fréquentes, la méthode de descente du gradient sera tout à fait incapable de retrouver la solution. Dans le chapitre 6, nous illustrerons cette situation sur quelques exemples précis.
- D'autre part, nous avons essayé de réduire le nombre de degrés de liberté du problème en approchant les non linéarités par des fonctions linéaires par morceaux. Ceci nous permet de garder l'inversibilité des non linéarités. Par contre, les fonctions linéaires par morceaux ne sont pas dérivables. Nous pourrions envisager d'utiliser des fonctions splines.
- Dans le chapitre 5, nous nous intéressons à l'étude statistique des estimateurs définis tout au long de cette partie. Nous remarquons qu'effectivement, il existe quelques différences qui peuvent expliquer certains comportement différents dans les algorithmes.

4.4.9 Méthodes basées sur la mesure de dépendance quadratique

Nous envisageons à présent les méthodes précédentes en utilisant la mesure de dépendance quadratique comme critère de séparation. Nous avons déjà souligné quelques aspects intéressants par rapport à la simplicilité de son estimation. Nous allons voir maintenant le détail des méthodes précédentes en utilisant la mesure de dépendance quadratique.

Approche non paramétrique

Nous avons déjà développé l'expression du gradient de la mesure de dépendance quadratique dans le chapitre 3, intéressons nous alors au cas des mélanges post non linéaires.

Dans ce contexte, les expressions des fonctionnelles du gradient s'écrivent,

- Gradient relatif par rapport à $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{i \neq j=1}^{K} \sum \varepsilon_{jk} \underbrace{[\Gamma_{kj} - \Gamma_{kk}\Sigma_{kj}/\Sigma_{kk}]}_{\mathcal{S}^L_{jk}}$$

- Gradient relatif par rapport à $\delta_1, \ldots, \delta_K$:

$$\delta_1, \ldots, \delta_K \;\mapsto\; \sum_{k=1}^{K} E\left\{ \delta_k(Z_k) \underbrace{E\left[\sum_{i=1}^{K} G_i^*(\mathbf{Y}) \mathbf{B}_{ik} \middle| Z_k \right]}_{\mathcal{S}_k^{NL}(z)} \right\}$$

où $\Gamma_{kj} = E[G_k(\mathbf{Y}) Y_j]$, $\Sigma_{kj} = E[(Y_k - E(Y_k))(Y_j - E(Y_j))]$
et nous rappelons ici l'expression de G^* donnée dans le lemme 3.2.9,
pour $\mathbf{t} \in \mathbb{R}^K$,

$$G_k^*(\mathbf{t}) = G_k(\mathbf{t}) - E[T_k G_k(\mathbf{T})] \phi_k(t_k)$$

et

$$\begin{aligned} G_k(\mathbf{t}) &= \pi_{k,\mathbf{T}}(t_1, \ldots, t_k) - \pi'_{T_k}(t_k) \prod_{l \neq k} \pi_{T_l}(t_l) + \pi'_{T_k}(t_k) \prod_{l \neq k} E[\pi_{T_l}(T_l)] \\ &\quad - E\left[\frac{1}{\sigma_{T_k}} \mathcal{K}_2'\left(\frac{t_k - T_k}{\sigma_{T_k}} \right) \prod_{l \neq k} \pi_{T_l}(T_l) \right]. \end{aligned}$$

et
$\pi_{\mathbf{T}}(\mathbf{z}) = E[\prod_{l=1}^{K} \mathcal{K}_2((z_l - T_l)/\sigma_{T_l})]$, $\pi_{k,\mathbf{T}}$ désigne sa dérivée partielle par rapport à la k-ième composante et $\pi_{T_k}(z_k) = E[\mathcal{K}_2((z_k - T_k)/\sigma_{T_k})]$ et π'_{T_k} désigne sa dérivée.
Les équations d'estimation s'écrivent ici de la manière suivante,

$$\begin{aligned} \Gamma_{kj} - \Gamma_{kk}\Sigma_{kj}/\Sigma_{kk} &= 0, \qquad 1 \leq i \neq j \leq K. \\ E\left[\sum_{i=1}^{K} G_i^*(\mathbf{Y}) \mathbf{B}_{ik} \middle| Z_k = z \right] &= 0, \qquad 1 \leq k \leq K. \end{aligned}$$

Grâce aux résultats du chapitre 3, nous pouvons en déduire l'expression du Hessien en un point où les variables sont indépendantes, les résultats sont exposés en annexe C.2.

Remarque 4.4.4 *Nous remarquons que la mesure de dépendance quadratique présente la propriété suivante, l'estimation de son gradient théorique est égale au gradient de son estimateur. En effet, ceci vient des calculs du gradient du critère théorique (paragraphe 3.2.1) et celui du critère estimé (paragraphe 3.2.4). Donc les stratégies «Estimer ensuite» et «Estimer d'abord» sont identiques dans ce cas.*

Approche non paramétrique utilisant les dérivées des non linéarités

Nous nous limiterons naturellement au cas du critère estimé. Nous écrivons dans ce contexte, les expressions du gradient en fonction de ε et des dérivées $\delta'_1, \ldots, \delta'_K$ que l'on approche par des fonctions linéaires par morceaux,

- Gradient relatif par rapport à $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{j \neq k=1}^{K} \varepsilon_{jk} [\widehat{\Gamma_{kj}} - \widehat{\Gamma_{kk}} \widehat{\Sigma}_{kj} / \widehat{\Sigma}_{kk}]$$

- Gradient relatif par rapport à $\delta'_1, \ldots, \delta'_K$:

$$\delta'_1, \ldots, \delta'_K \mapsto \sum_{k=1}^{K} \int \widehat{E} \left[\mathbb{1}_+(Z_k - z) \sum_{i=1}^{K} \widehat{G}_i^*(\mathbf{Y}) \mathbf{B}_{ik} \right] \delta'_k(z) dz, \quad 1 \leq k \leq K.$$

où $\widehat{\Gamma}_{kj} = \widehat{E}[\widehat{G}_k(\mathbf{Y}) Y_j]$, $\widehat{\Sigma}_{kj} = \widehat{E}[(Y_k - \widehat{E}(Y_k))(Y_j - \widehat{E}(Y_j))]$

En introduisant alors les statistiques d'ordre de Z_k, nous pouvons réécrire les expressions du gradient,

- Gradient relatif relatif par rapport à $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{j \neq k=1}^{K} \widehat{\varepsilon}_{jk} [\widehat{\Gamma_{kj}} - \widehat{\Gamma_{kk}} \widehat{\Sigma}_{kj} / \widehat{\Sigma}_{kk}]$$

- Gradient relatif par rapport à $\delta'_1, \ldots, \delta'_K$:

$$\delta'_1, \ldots, \delta'_K \mapsto \frac{1}{N} \sum_{j=1}^{K} \sum_{n=2}^{N} \left\{ \sum_{m=n}^{N} \sum_{k=1}^{K} \widehat{G}_k^*(\mathbf{Y}(o_{j,m})) \mathbf{B}_{kj} \right] (\delta'_j(Z_j(n-1:N)) - \delta'_j(Z_j(n-1:N)))$$

Nous voyons alors aussi que l'approche paramétrique va s'écrire de la manière suivante,

Approche paramétrique

- Gradient relatif par rapport à $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{j \neq k=1}^{K} \varepsilon_{jk} (\Gamma_{kj} - \Gamma_{kk} \Sigma_{kj} / \Sigma_{kk})$$

- Gradient relatif de $\widehat{C}$ par rapport à $d\theta_1, \ldots, d\theta_K$:

$$d\theta_1, \ldots, d\theta_K \mapsto$$
$$\frac{1}{N}\sum_{j=1}^{K}\sum_{n=2}^{N}\left\{\sum_{m=n}^{N}\sum_{k=1}^{K}\widehat{G}_k^*(\mathbf{Y}(o_{j,m}))\mathbf{B}_{kj}\right\}\{\dot{g}_{j,\theta_j}[Z_j(n:N)] - \dot{g}_{j,\theta_j}[Z_j(n-1:N)]\}\mathrm{d}\theta_j$$

où $g_{k,\theta_k} = g'_{k,\theta_k} \circ g^{-1}_{k,\theta_k}$ avec g'_{k,θ_k} la dérivée de g_{k,θ_k} par rapport à θ_k.

En annexe C.2, nous donnons les expressions du Hessien dans le cadre de l'approche paramétrique.

4.5 Conclusion

Nous avons présenté dans ce chapitre trois approches distinctes permettant la résolution du problème de séparation aveugle de sources dans le cadre de mélanges post non linéaires.

Pour ces trois approches, nous avons détaillé les deux stratégies «Estimer ensuite» et «Estimer d'abord» et exhibé leurs différences.

Ces mélanges sont représentatifs de quelques situations réelles, par exemple le cas de signaux provenant de satellites. Mais on peut leur reprocher de ne pas être assez généraux, et la définition d'un cadre plus large de mélanges non linéaires et identifiables reste un problème ouvert. En particulier, il serait intéressant d'étudier des mélanges post non linéaires bruités.

Chapitre 5

Estimation et convergence

Nous avons constaté que dans les chapitres précédents, le problème de séparation aveugle de source sous la seule hypothèse d'indépendance est fortement lié à la définition d'une mesure de dépendance puis à son estimation. Il s'avère donc intéressant de pouvoir étudier ces mesures de dépendance d'un point de vue statistique, c'est-à-dire de savoir par exemple si l'estimateur défini permet effectivement de décider si des variables sont indépendantes et avec quel taux d'erreur.

Tout d'abord, nous allons étudier plus précisément la mesure de dépendance quadratique. Dans la section 5.1, nous représentons l'estimateur de la mesure de dépendance quadratique à l'aide des U-statistiques introduites par Hoeffding en 1948 [35]. Puis, en reprenant une partie des travaux de Kankainen [44], nous déduisons les lois asymptotiques de la mesure de dépendance quadratique sous l'hypothèse que les variables sont dépendantes et sous l'hypothèse inverse (section 5.2).

Ensuite nous utiliserons les propriétés asymptotiques de la mesure de dépendance quadratique dans la section 5.3 afin de proposer une méthode du choix de la taille de fenêtre en fonction du noyau. Nous illustrerons aussi le comportement de la mesure de dépendance quadratique dans le problème simple de séparation aveugle de sources d'un mélange linéaire (paragraphes 5.3.3 et 5.3.4).

Dans la section 5.4, nous détaillons plus précisément les expressions des estimations des fonctions scores obtenues dans la section 3.2.3. Enfin, nous terminons par un calcul de biais des deux critères de séparation dérivés de l'information mutuelle (c.f. section 4.4.8), le critère C utilisé par Taleb [67], et celui utilisé par Babaie-Zadeh [6] (section 5.5).

5.1 La mesure de dépendance quadratique en termes de U-statistiques

Dans la thèse de Kankainen [44], il est démontré la convergence de l'estimateur noté T_n (on rapelle la définition dans la section 2.4), à l'aide de U-statistiques. Comme nous l'avons fait remarquer auparavant, Kankainen étudie la mesure de dépendance quadratique seulement dans le cas d'un noyau gaussien ou d'un noyau uniforme. A présent, nous reprenons le raisonnement effectué par Kankainen en l'adaptant à la définition de la mesure de dépendance quadratique 2.5.1. Ceci nous permet d'obtenir des expressions plus explicites, et donc plus facilement exploitables.

Rappelons ici l'écriture de la mesure de dépendance quadratique donnée dans le lemme 2.5.1,

$$Q(T_1,\ldots,T_K) = E[\pi_{\mathbf{T}}(\mathbf{T})] + \prod_{k=1}^{K} E[\pi_{T_k}(T_k)] - 2E\left[\prod_{k=1}^{K}\pi_{T_k}(T_k)\right]$$

où $T_1,\ldots,T_K$, sont K variables aléatoires telles que $\mathbf{T} = (T_1, T_2, \ldots, T_K)^T$ et $\mathcal{K}_2$ est un noyau réel et tel que sa transformée de Fourier est positive sommable et non nulle presque partout.

De plus, on a noté,

$$\begin{aligned}
\pi_{\mathbf{T}}(t_1,\ldots,t_K) &= E\left[\prod_{k=1}^{K}\mathcal{K}_2\left(\frac{t_k - T_k}{\sigma_{T_k}}\right)\right] \\
\pi_{t_k}(T_k) &= E\left[\mathcal{K}_2\left(\frac{t_k - T_k}{\sigma_{T_k}}\right)\right]
\end{aligned}$$

Notons par soucis de clarté,

- $\theta_1 = E[\pi_{\mathbf{T}}(\mathbf{T})]$
- $\theta_2 = \prod_{k=1}^{K} E[\pi_{T_k}(T_k)]$
- $\theta_3 = E\left[\prod_{k=1}^{K}\pi_{T_k}(T_k)\right]$

Comme l'a fait remarquer Kankainen [44], sous l'hypothèse que les variables aléatoires $T_1,\ldots,T_K$ sont *dépendan-tes*, la convergence de cet estimateur peut-être étudiée à l'aide des U-statistiques. La notion de U-statistique a été introduite par Hoeffding [35]. Puis celui-ci a utilisé ces U-statistiques afin d'étudier par exemple un test d'indépendance construit à partir des fonctions de répartition, [36]. Pour une vue générale des U-statistiques, on peut aussi consulter [49].

Nous nous plaçons ici dans le cas où les variables aléatoires $T_1,\ldots,T_K$ sont *dépendantes*. Comme l'a fait remarquer Kankainen, cette étude ne peut-être reprise sous l'hypothèse

d'indépendance avec les U-statistiques. Pour cela, Kankainen établit la loi asymptotique de la mesure de dépendance quadratique par d'autres moyens. Afin de pouvoir calculer la puissance de test et d'autres probabilités, nous allons réécrire les différentes quantités intervenant dans l'étude asymptotique de la mesure de dépendance quadratique à l'aide des U-statistiques. Pour chaque terme de la définition de la mesure de dépendance quadratique, nous allons faire apparaître la fonctionnelle de distribution dépendant d'un noyau. Puis nous écrirons alors la U-statistique associée comme l'a définie Hoeffding [35]. Enfin, nous pourrons en déduire la loi asymptotique de la mesure de dépendance quadratique dans la section 5.2.

Afin d'appliquer le théorème 7.3 de Hoeffding [35], nous représentons chaque estimateur U_i' de θ_i par une somme,

$$U_i' = \lambda(N)U_i + \frac{b_N^{(i)}}{\sqrt{N}}$$

où U_i est un estimateur sans biais de θ_i (i.e. la U-statistique associée à θ_i), $\lambda(N) \sim 1$ et $b_N^{(i)}$ tel que $\lim_{N\to+\infty} E[b_N^{(i)2}] = 0$.

5.1.1 Etude de θ_1

Définition de la U-statistique associée

Ecrivons tout d'abord ce terme sous la forme,

$$\begin{aligned} E[\pi_{\mathbf{T}}(\mathbf{T})] &= \iint \prod_{k=1}^{K} \mathcal{K}_2\left(\frac{u_k - t_k}{\sigma_{T_k}}\right) dF_{\mathbf{T}}(\mathbf{u}) dF_{\mathbf{T}}(\mathbf{t}) \\ &= \iint \prod_{k=1}^{K} \mathcal{K}_2\left(u_k - t_k\right) dF_{\mathbf{T}/\sigma_{\mathbf{T}}}(\mathbf{u}) dF_{\mathbf{T}/\sigma_{\mathbf{T}}}(\mathbf{t}) \end{aligned}$$

Le noyau de la fonctionnelle de distribution va s'écrire dans ce cas,

$$\widetilde{K}_2^{(1)}(\mathbf{u},\mathbf{t}) = \prod_{k=1}^{K} \mathcal{K}_2(u_k - t_k)$$

Comme le noyau $\mathcal{K}_2$ est choisi pair, $\widetilde{K}_2^{(1)}$ est symétrique.

On peut donc associer à cette variable une U-statistique,

(où on note un vecteur aléatoire $\mathbf{T}$ qui admet un échantillon composé de N vecteurs $\mathbf{T}(n)$ indépendants suivant la loi $F_{\mathbf{T}}$.)

$$U_1(\mathbf{T}(1),\ldots,\mathbf{T}(N)) = \frac{2}{N(N-1)} \sum_{1\leq i<j\leq N} \prod_{k=1}^{K} \mathcal{K}_2\left(\frac{T_k(i)-T_k(j)}{\widehat{\sigma_{T_k}}}\right)$$

Cependant, dans l'estimateur $\widehat{Q}$ de Q, on ne voit pas exactement apparaître la satistique U_1 définie ci-dessus.

Lien avec l'estimateur $\widehat{Q}$

Afin de faire le lien avec l'estimateur $\widehat{Q}$, on établit le lemme suivant:

Lemme 5.1.1

$$\begin{aligned} U_1'(\mathbf{T}(1),\ldots,\mathbf{T}(N)) := \widehat{E}[\widehat{\pi}_{\mathbf{T}}(\mathbf{T})] &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{N}\sum_{m=1}^{N}\prod_{k=1}^{K}\mathcal{K}_2\left(\frac{T_k(n)-T_k(m)}{\widehat{\sigma_{T_k}}}\right) \\ &= \frac{N-1}{N}U_1 + \frac{b_N^{(1)}}{\sqrt{N}} \end{aligned}$$

où

$$b_N^{(1)} = \frac{1}{\sqrt{N}}\prod_{k=1}^{K}\mathcal{K}_2(0)$$

preuve:

En effet, on a l'égalité suivante,

$$\begin{aligned} \widehat{E}[\widehat{\pi}_{\mathbf{T}}(\mathbf{T})] &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{N}\sum_{m=1}^{N}\prod_{k=1}^{K}\mathcal{K}_2\left(\frac{T_k(n)-T_k(m)}{\widehat{\sigma_{T_k}}}\right) \\ &= \frac{N-1}{N}\left(\frac{2}{N(N-1)}\sum_{1\leq i<j\leq N}\prod_{k=1}^{K}\mathcal{K}_2\left(\frac{T_k(i)-T_k(j)}{\widehat{\sigma_{T_k}}}\right)\right) \\ &+ \frac{1}{N^2}\sum_{n=1}^{N}\prod_{k=1}^{K}\mathcal{K}_2\left(\frac{T_k(n)-T_k(n)}{\widehat{\sigma_{T_k}}}\right) \end{aligned}$$

Ce qui nous amène au résultat demandé.

■

5.1.2 Etude de θ_2

Nous allons voir comment l'estimateur de θ_2 peut-être représenté par une U-statistique.

Définition de la U-statistique associée

Ecrivons tout d'abord le terme ci-dessus d'une manière plus adéquate.
En effet,

$$\begin{aligned}
\prod_{k=1}^{K} E[\pi_{T_k}(T_k)] &= \int \dots \int \prod_{k=1}^{K} \mathcal{K}_2 \left(\frac{u_k - t_k}{\sigma_{T_k}} \right) \prod_{k=1}^{K} dF_{T_k}(u_k) dF_{T_k}(t_k) \\
&= \int \dots \int \prod_{k=1}^{K} \mathcal{K}_2(u_k - t_k) \prod_{k=1}^{K} dF_{T_k/\sigma_{T_k}}(u_k) dF_{T_k/\sigma_{T_k}}(t_k) \\
&= \int \dots \int \prod_{k=1}^{K} \mathcal{K}_2(u_k^{(k)} - t_k^{(k)}) \prod_{k=1}^{K} dF_{\mathbf{T}/\sigma_{\mathbf{T}}}(\mathbf{u}^{(k)}) dF_{\mathbf{T}/\sigma_{\mathbf{T}}}(\mathbf{t}^{(k)})
\end{aligned}$$

où $\mathbf{u}_k^{(k)} = u_k$, les autres termes du vecteur $\mathbf{u}^{(k)}$ n'interviennent pas dans le reste de l'expression sous l'intégrale.

Le noyau associé à cette fonctionnelle de distribution s'écrit,

$$\widehat{K}_2^{(2)}(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(K)}, \mathbf{u}^{(K+1)}, \dots, \mathbf{u}^{(2K)}) = \prod_{k=1}^{K} \mathcal{K}_2(u_k^{(k)} - u_k^{(k+K)})$$

Mais, pour pouvoir lui associer une U-statistique, celui-ci doit être symétrique.

Comme le noyau $\mathcal{K}_2$ est pair, on peut définir le noyau symétrique associé par:

$$\begin{aligned}
\widetilde{K}_2^{(2)}(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(K)}, \mathbf{t}^{(1)}, \dots, \mathbf{t}^{(K)}) &= \frac{1}{K!^2} \sum_{q_2} \widehat{K}_2^{(2)}(\mathbf{u}^{(\alpha_1)}, \dots, \mathbf{u}^{(\alpha_K)}, \mathbf{t}^{(\beta_1)}, \dots, \mathbf{t}^{(\beta_K)}) \\
&= \frac{1}{K!^2} \sum_{q_2} \prod_{k=1}^{K} \mathcal{K}_2(u_k^{(\alpha_k)} - t_k^{(\beta_k)})
\end{aligned}$$

où $\sum_{q_2}$ désigne la somme sur toutes les permutations de $(\alpha_1, \dots, \alpha_K)$ de $(1, \dots, K)$ et $(\beta_1, \dots, \beta_K)$ de $(1, \dots, K)$.

Donc la U-statistique associée s'écrit, (on suppose $2K \leq N$)

$$U_2(\mathbf{T}(1), \dots, \mathbf{T}(N)) = \frac{1}{C_N^{2K}} \sum_{(1 \leq i_1 < \dots < i_K < j_1)} \sum_{(i_K < j_1 < \dots < j_K \leq N)} \sum_{s_2} \frac{1}{K!^2} \prod_{k=1}^{K} \mathcal{K}_2 \left(\frac{T_k(\alpha_k) - T_k(\beta_k)}{\widehat{\sigma_{T_k}}} \right)$$

où $\sum_{s_2}$ désigne l'ensemble des permutations $(\alpha_1, \ldots, \alpha_K)$ de $(i_1, \ldots, i_K)$ et $(\beta_1, \ldots, \beta_K)$ de $(j_1, \ldots, j_K)$.

Lien avec l'estimateur $\widehat{Q}$

Reprenons l'écriture de ce terme dans l'estimateur $\widehat{Q}$ pour obtenir le lien entre les deux.

Lemme 5.1.2

$$\begin{aligned} U_2'(\mathbf{T}(1), \ldots, \mathbf{T}(N)) := \prod_{k=1}^{K} \widehat{E}[\widehat{\pi}_{T_k}(T_k)] &= \prod_{k=1}^{K} \frac{1}{N} \sum_{n=1}^{N} \frac{1}{N} \sum_{m=1}^{N} \mathcal{K}_2 \left(\frac{T_k(n) - T_k(m)}{\widehat{\sigma_{T_k}}} \right) \\ &= 2 \frac{C_N^{2K}}{N^{2K}} U_2 + \frac{b_N^{(2)}}{\sqrt{N}} \end{aligned}$$

où

$$b_N^{(2)} = \frac{1}{\sqrt{N} N^{(K-1)}} \sum_{A} \prod_{k=1}^{K} \mathcal{K}_2 \left(\frac{T_k(i_k) - T_k(j_k)}{\widehat{\sigma_{T_k}}} \right)$$

et $A = \{(i_1, \ldots, i_K, j_1, \ldots, j_K) \in (1, \ldots, N) | \exists k, k' \in (1, \ldots, K), i_k = j_{k'}\}$

5.1.3 Etude de θ_3

Enfin, nous pouvons aussi représenter le dernier terme de la même manière.

Définition de la U-statistique associée

Réécrivons tout d'abord ce terme,

$$\begin{aligned} E[\prod_{k=1}^{K} \pi_{T_k}(T_k)] &= \int \ldots \int \prod_{k=1}^{K} \mathcal{K}_2 \left(\frac{u_k - t_k}{\sigma_{T_k}} \right) \prod_{k=1}^{K} dF_{T_k}(t_k) dF_{\mathbf{T}}(\mathbf{u}) \\ &= \int \ldots \int \prod_{k=1}^{K} \mathcal{K}_2(u_k - t_k) \prod_{k=1}^{K} dF_{T_k/\sigma_{T_k}}(t_k) dF_{\mathbf{T}/\sigma\mathbf{T}}(\mathbf{u}) \\ &= \int \ldots \int \prod_{k=1}^{K} \mathcal{K}_2(u_k - t_k^{(k)}) \prod_{k=1}^{K} dF_{\mathbf{T}/\sigma_{\mathbf{T}}}(\mathbf{t}^{(k)}) dF_{\mathbf{T}/\sigma\mathbf{T}}(\mathbf{u}) \end{aligned}$$

où $\mathbf{t}_k^{(k)} = t_k$.

Ici aussi, nous pouvons écrire le noyau de la fonctionnelle de distribution,

$$\widehat{K}_2^{(3)}(\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(K+1)}) = \prod_{k=1}^{K} \mathcal{K}_2(u_k^{(1)} - u_k^{(k+1)})$$

Toujours pour les mêmes raisons, nous devons considérer un noyau symétrique associé,

$$\begin{aligned}\widetilde{K}_2^{(3)}(\mathbf{u}^{(1)},\mathbf{u}^{(2)},\ldots,\mathbf{u}^{(K+1)}) &= \frac{1}{(K+1)!}\sum_{p_3} \widehat{K}_2^{(3)}(\mathbf{u}^{(\alpha_1)},\ldots,\mathbf{u}^{(\alpha_K)},\mathbf{u}^{(\alpha_{K+1})}) \\ &= \frac{1}{(K+1)!}\sum_{p_3}\prod_{k=1}^{K} \mathcal{K}_2(u_k^{(\alpha_1)} - u_k^{(\alpha_{k+1})})\end{aligned}$$

où $\sum_{p_3}$ désigne la somme sur toutes les permutations $(\alpha_1,\ldots,\alpha_{K+1})$ de $(1,\ldots,K+1)$.

Mais ce noyau peut aussi s'écrire,

$$\begin{aligned}\widetilde{K}_2^{(3)}(\mathbf{u},\mathbf{t}^{(1)},\ldots,\mathbf{t}^{(K)}) &= \frac{1}{(K+1)!}\sum_{q_3} \widehat{K}_2^{(3)}(\mathbf{u},\mathbf{t}^{(\alpha_1)},\ldots,\mathbf{t}^{(\alpha_K)}) \\ &+ \frac{1}{(K+1)!}\sum_{l=1}^{K}\sum_{r_3} \widehat{K}_2^{(3)}(\mathbf{t}^{(\alpha_l)},\mathbf{t}^{(\beta_1)},\ldots,\mathbf{t}^{(\beta_{l-1})},\mathbf{u},\mathbf{t}^{(\beta_{l+1})},\ldots,\mathbf{t}^{(\beta_K)}) \\ &= \frac{1}{(K+1)!}\sum_{q_3}\prod_{k=1}^{K} \mathcal{K}_2(u_k - t_k^{(\alpha_{k+1})}) \\ &+ \frac{1}{(K+1)!}\sum_{l=1}^{K}\sum_{r_3}\prod_{k=1}^{K} \mathcal{K}_2(t_k^{(\alpha_l)} - t_k^{\prime(\beta_k)})\end{aligned}$$

où pour tout $k=1,\ldots,K$, $k\neq l$, $\mathbf{t}'^{(\beta_k)} = \mathbf{t}^{(\alpha_k)}$ et $\mathbf{t}'^{(\beta_l)} = \mathbf{u}$
Et, $\sum_{q_3}$ désigne la somme sur toutes les permutations $(\alpha_1,\ldots,\alpha_K)$ de $(1,\ldots,K)$ et $\sum_{r_3}$ désigne la somme sur toutes les permutations $(\beta_1,\ldots,\beta_{l-1},\alpha_l,\beta_{l+1},\ldots,\beta_K)$ de $(1,\ldots,K)$

On obtient,

$$U_3(\mathbf{T}(1),\ldots,\mathbf{T}(N)) = \frac{1}{C_N^{K+1}}\sum_{1\leq j<i_1<\ldots<i_K}\sum_{s_3}\prod_{k=1}^{K}\mathcal{K}_2\left(\frac{T_k(\alpha)-T_k(\beta_k)}{\widehat{\sigma_{T_k}}}\right)$$

où $\sum_{s_3}$ désigne la somme sur toutes les permutations de $(\alpha,\beta_1,\ldots,\beta_K)$ de $(j,i_1,\ldots,i_K)$.

Lien avec l'estimateur $\widehat{Q}$

Lemme 5.1.3

$$\begin{aligned} U_3'(\mathbf{T}(1),\ldots,\mathbf{T}(N)) := \widehat{E}[\prod_{k=1}^{K}\widehat{\pi}_{T_k}(T_k)] &= \frac{1}{N}\sum_{n=1}^{N}\prod_{k=1}^{K}\frac{1}{N}\sum_{m=1}^{N}\mathcal{K}_2\left(\frac{T_k(n)-T_k(m)}{\widehat{\sigma_{T_k}}}\right) \\ &= 2\frac{C_N^{K+1}}{N^{K+1}}U_3 + \frac{b_N^{(3)}}{\sqrt{N}} \end{aligned}$$

où

$$b_N^{(3)} = \frac{1}{\sqrt{N}N^{(K-1)}}\sum_{A}\prod_{k=1}^{K}\mathcal{K}_2\left(\frac{T_k(i_k)-T_k(j)}{\widehat{\sigma_{T_k}}}\right)$$

et $A = \{(i_1,\ldots,i_K,j_1,\ldots,j_K) \in (1,\ldots,N) | \exists k,k' \in (1,\ldots,K), i_k = j_{k'}\}$

5.2 Etude asymptotique

Après avoir représenté l'estimateur de la mesure de dépendance quadratique sous la forme de U-statistique, nous pouvons en déduire la loi asymptotique sous l'hypothèse où les variables sont dépendantes.

Afin d'appliquer le théorème 7.3 de Hoeffding [35], nous supposons que le noyau $\mathcal{K}_2$ est de carré intégrable. Par conséquent, pour $i = 1,2,3$, $E(\widetilde{\mathcal{K}}^{(i)2}) < \infty$. De plus, nous devons vérifier certaines propriétés sur les variables aléatoires $b_N^{(i)}$, pour $i = 1,2,3$. En effet, on montre que,

pour $i = 1,\ldots,3$,

$$\lim_{N\to+\infty} E[b_N^{(i)2}] = 0. \tag{5.1}$$

Il suffit de revenir à la définition de chaque variable. Il est alors clair que ceux-ci vérifient bien la propriété (5.1).

A présent d'après le théorème 7.3 de Hoeffding [35], on déduit que la distribution conjointe des variables aléatoires :

$$\sqrt{N}(U_1' - \theta_1), \sqrt{N}(U_2' - \theta_2), \sqrt{N}(U_3' - \theta_3)$$

tend vers une distribution normale de moyenne 0 et de matrice de covariance Σ définie par, pour $i,j = 1,\ldots,3$,

$$\Sigma_{ij} = m(i)m(j)\xi_1^{(i,j)}$$

où, $m(1) = 2$, $m(2) = K+1$ et $m(3) = 2K$ et $\xi_1^{(i,j)}$ sont définis dans la suite.

Puis par le corollaire de Serfling [64, p.124], on déduit alors la loi asymptotique de $\widehat{Q}$,

Lemme 5.2.1 (Loi asymptotique de $\widehat{Q}$ sous l'hypothèse de dépendance) *$\sqrt{N}(\widehat{Q} - Q)$ suit asymptotiquement une loi normale de moyenne 0 et de variance σ^2.*

où l'expression de σ^2 est la suivante,

$$\sigma^2 = \Sigma_{11} - 4\Sigma_{12} + 2\Sigma_{13} - 4\Sigma_{23} + 4\Sigma_{22} + \Sigma_{33}$$

et les termes Σ_{ij} sont explicités ci dessous.

D'après l'écriture de σ^2 explicitée ci-dessous, nous remarquons qu'il est tout à fait possible de l'estimer. Il est possible à présent d'envisager une étude de la mesure de dépendance quadratique en tant que test d'indépendance. En effet, en approchant la loi de son estimateur par la loi asymptotique du lemme 5.2.1, nous sommes en mesure de calculer les intervalles de confiance empiriques correspondant à cet estimateur (paragraphe 5.3.1) et d'étudier la puissance du test (paragraphe 5.3.2).

Expression de σ^2 :

Tout d'abord, revenons aux notations de Hoeffding pour en déduire la matrice de covariance de la loi conjointe : Σ.

Pour obtenir l'expression de $\xi_1^{(i,j)}$, nous avons besoin de définir 3 fonctions, une correspondant à U_1, une à U_2 et enfin une à U_3.

$\Psi_1^{(1)}$:

$$\begin{aligned}\Psi_1^{(1)}(\mathbf{x}) &= E[\widetilde{K}_2^{(1)}(\mathbf{x}, \mathbf{T}/\sigma_{\mathbf{T}})] - \theta_1 \\ &= E[\prod_{k=1}^{K} \mathcal{K}_{2,k}\left(x_k - \frac{T_k}{\sigma_{T_k}}\right)] - \theta_1 \\ &= \pi_{\mathbf{T}}(\mathbf{x}\sigma_{\mathbf{T}}) - \theta_1\end{aligned}$$

$\Psi_1^{(2)}$:

$$\begin{aligned}\Psi_1^{(2)}(\mathbf{x}) &= E[\widetilde{K}_2^{(2)}(\mathbf{x}, \mathbf{T}(2)/\sigma_{\mathbf{T}}, \ldots, \mathbf{T}(2K)/\sigma_{\mathbf{T}})] - \theta_2 \\ &= \frac{1}{K}\sum_{l=1}^{K}\prod_{k\neq l} E[\pi_{T_k}(T_k)]\pi_{T_l}(x_l\sigma_{T_l}) - \theta_2\end{aligned}$$

$\Psi_1^{(3)}$:

$$\begin{aligned}
\Psi_1^{(3)}(\mathbf{x}) &= E[\widetilde{K}_2^{(3)}(\mathbf{x}, \mathbf{T}/\sigma_{\mathbf{T}})] - \theta_3 \\
&= \frac{1}{K+1}\prod_{k=1}^{K}\pi_{T_k}(x_k\sigma_{T_k}) \\
&+ \frac{1}{K(K+1)}\sum_{l=1}^{K}\sum_{m=1}^{K}\int\prod_{k\neq l}\pi_{T_k}(t_k^{(m)})\mathcal{K}_{2,l}\left(\frac{t_l^{(m)}}{\sigma_{T_l}} - x_l\right)dF_{\mathbf{T}}(t^{(m)}) - \theta_3 \\
&= \frac{1}{K+1}\prod_{k=1}^{K}\pi_{T_k}(x_k\sigma_{T_k}) - \theta_2 \\
&+ \frac{1}{K+1}\sum_{l=1}^{K}E\left[\prod_{k\neq l}\pi_{T_k}(T_k)\mathcal{K}_{2,l}\left(\frac{T_l}{\sigma_{T_l}} - x_l\right)\right] \\
&= \frac{1}{K+1}\prod_{k=1}^{K}\pi_{T_k}(x_k\sigma_{T_k}) + \frac{1}{K+1}\sum_{l=1}^{K}\widetilde{\pi}_{T_l}(x_l\sigma_{T_l}) - \theta_3
\end{aligned}$$

où $\widetilde{\pi}_{T_l}(t_l) = E\left[\prod_{k\neq l}\pi_{T_k}(T_k)\mathcal{K}_{2,l}\left(\frac{T_l - t_l}{\sigma_{T_l}}\right)\right]$

A partir de ces expressions, nous pouvons désormais obtenir les termes de la matrice de covariance Σ de la loi conjointe.

$\xi_1^{(11)}$:

$$\xi_1^{(11)} = E[\pi_{\mathbf{T}}(\mathbf{T})^2] - \theta_1^2$$

$\xi_1^{(12)}$:

$$\xi_1^{(12)} = \xi_1^{(21)} = \frac{1}{K}\sum_{l=1}^{K}E[\prod_{k\neq l}E(\pi_{T_k}(T_k))\pi_{T_l}(T_l)\pi_{\mathbf{T}}(\mathbf{T})] - \theta_1\theta_2$$

$\xi_1^{(13)}$:

$$\begin{aligned}
\xi_1^{(13)} = \xi_1^{(31)} &= \frac{1}{K+1}E[\pi_{\mathbf{T}}(\mathbf{T})\prod_{k=1}^{K}\pi_{T_k}(T_k)] - \theta_1\theta_3 \\
&- \frac{1}{K+1}\sum_{l=1}^{K}E\left[\pi_{\mathbf{T}}(\mathbf{T})\widetilde{\pi}_{T_l}(T_l)\right]
\end{aligned}$$

$\xi_1^{(23)}$:

$$\begin{aligned}\xi_1^{(23)} &= \xi_1^{(32)} = \frac{1}{K(K+1)} \sum_{k=1}^{K} \prod_{l \neq k} E[\pi_{T_k}(T_k)] E[\pi_{T_l}(T_l) \prod_{k=1}^{K} \pi_{T_k}(T_k)] - \theta_2 \theta_3 \\ &+ \frac{1}{K(K+1)} \sum_{l,m=1}^{K} \prod_{l \neq k} E[\pi_{T_k}(T_k)] E[\pi_{T_l}(T_l) \widetilde{\pi}_{T_m}(T_m)]\end{aligned}$$

$\xi_1^{(22)}$:

$$\xi_1^{(22)} = \frac{1}{K^2} \sum_{l,m=1}^{K} \prod_{k \neq l} E[\pi_{T_k}(T_k)] \prod_{k \neq m} E[\pi_{T_k}(T_k)] E[\pi_{T_l}(T_l) \pi_{T_m}(T_m)] - \theta_2^2$$

$\xi_1^{(33)}$:

$$\begin{aligned}\xi_1^{(33)} &= \frac{1}{(K+1)^2} E[\prod_{k=1}^{K} \pi_{T_k}^2(T_k)] + \frac{1}{(K+1)^2} \sum_{l,m=1}^{K} E[\widetilde{\pi}_{T_l}(T_l) \widetilde{\pi}_{T_m}(T_m)] \\ &+ \frac{2}{(K+1)^2} \sum_{l}^{K} E[\prod_{k=1}^{K} \pi_{T_k}(T_k) \widetilde{\pi}_{T_l}(T_l)] - \theta_3^2\end{aligned}$$

5.3 Comment quantifier l'influence de la taille de fenêtre et du choix du noyau

Nous avons pu voir que le critère de dépendance quadratique dépend du choix d'un noyau vérifiant des propriétés spécifiques (c.f. chapitre 2). De plus, pour chaque noyau se pose le problème du choix d'une taille de fenêtre.

D'après l'expression des noyaux, on voit que plus on prend une taille de fenêtre petite, plus les points où le noyau sera significativement différent de zéro seront concentrés autour de l'origine. Et par conséquent la transformée de Fourier de ces noyaux sera étendue.

Les courbes du noyau gaussien pour des tailles de fenêtres de 0.5, 1 et 2 et les transformées de Fourier correspondantes sont représentées dans la figure 5.1.

Malgré le fait que théoriquement, la mesure de dépendance quadratique est toujours discriminante quelque soit la taille de fenêtre, il est naturel de penser que si la taille de fenêtre est trop petite, on va perdre de l'information, et au contraire, si elle est trop grande, on va garder trop d'information. Afin de quantifier ceci, nous pouvons faire une première étude de l'écart-type de l'estimateur de dépendance quadratique sous

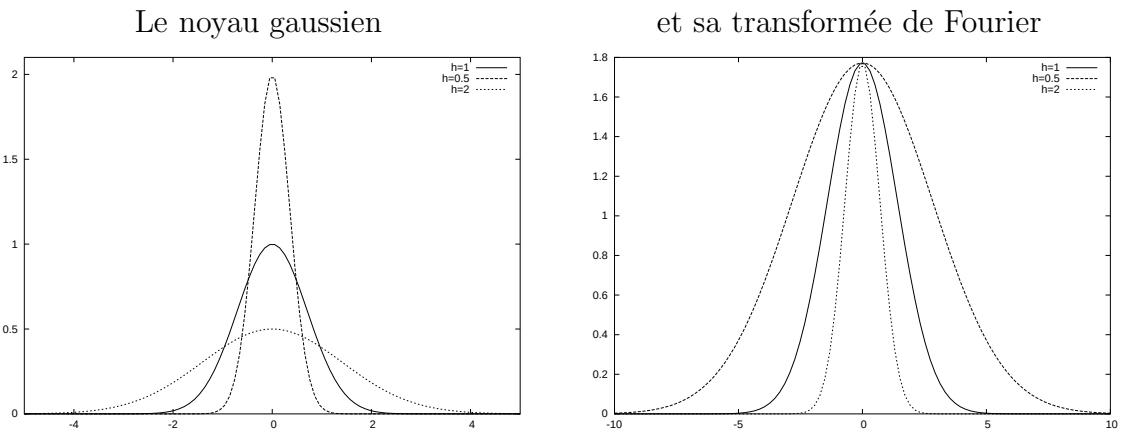


FIG. 5.1 – *Le noyau gaussien et sa transformée de Fourier pour différentes tailles de fenêtre*

l'hypothèse que les variables sont dépendantes (paragraphe 5.3.1). Nous verrons que ceci est insuffisant, nous nous interesserons alors dans un deuxième temps au calcul de la puissance du test en fonction des différentes tailles de fenêtres pour différents noyaux (paragraphe 5.3.2). Enfin, nous conclurons en rapprochant ces résultats du problème de séparation de sources. Dans cette section, nous fixons la taille de l'échantillon à 500, sauf précisions contraire. Les résultats empiriques représentés sont donc relatifs à cette taille d'échantillon.

5.3.1 Etude des intervalles de confiance

En conséquence des résultats précédents, nous pouvons déduire les lemmes suivants, ces lemmes sont dus à Hoeffding [35].

Ici nous nous plaçons dans l'hypothèse où les variables sont dépendantes.

Lemme 5.3.1 *L'estimateur de dépendance quadratique est asymptotiquement sans biais, c'est-à-dire,*

$$E(\widehat{Q}) = Q + O(N^{-1})$$

De plus,

Lemme 5.3.2 *La variance de $\widehat{Q}$ vérifie,*

$$\lim_{N\to\infty} NV(\widehat{Q}) = \sigma^2$$

Enfin, dans la partie précédente, nous avons pu remarquer que sous l'hypothèse que les variables sont dépendantes, l'estimateur de dépendance quadratique converge

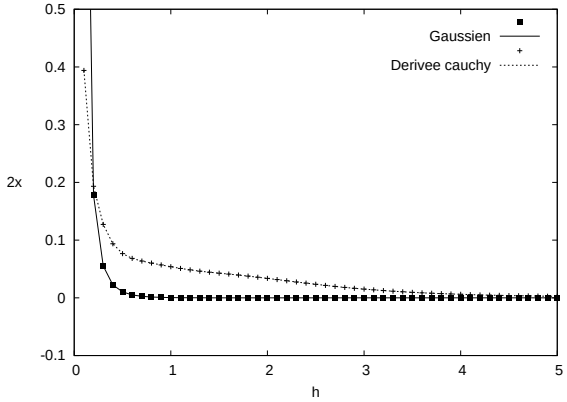


FIG. 5.2 – *Taille de l'intervalle de confiance en fonction de la taille de fenêtre pour différents noyaux.*

en loi vers une loi normale de moyenne la vraie valeur de la mesure de dépendance quadratique et de variance σ^2/N où N est la taille de l'échantillon. Par conséquent, plus la taille de l'échantillon sera grande, plus la variance de la mesure de dépendance quadratique sera faible.

Nous proposons de quantifier ceci en étudiant la taille de l'intervalle de confiance correspondant à un seuil α donné en fonction de la taille de fenêtre.

Nous définissons alors l'intervalle de confiance de la manière suivante, on cherche $x > 0$ de telle sorte que,

$$P(-x \leq \widehat{Q} - Q \leq x) = \alpha$$

sous l'hypothèse que les variables sont dépendantes.

D'après les résultats de la section (5.1), on sait que la loi asymptotique de l'estimateur $\sqrt{N}(\widehat{Q} - Q)$ est une loi normale de moyenne 0 et de variance σ^2 (c.f. lemme 5.2.1). Alors, nous en déduisons la longueur de l'intervalle de confiance de Q,

$$2x = 2\frac{\sigma}{\sqrt{N}}\Phi^{-1}\left(\frac{1-\alpha}{2}\right)$$

où Φ désigne la fonction de répartition de la loi normale centrée réduite.

Les résultats obtenus pour les deux noyaux envisagés (c.f. chapitre 2) sont représentés dans la figure 5.2.

Il serait alors préférable de choisir une taille de fenêtre la plus grande possible. Mais, il apparait vite un deuxième problème, l'ordre de grandeur de la mesure de dépendance diminue lui aussi. Si bien que pour une taille de fenêtre trop grande, l'écart-type et la moyenne de la mesure de dépendance quadratique peuvent être du

même ordre de grandeur. Il est alors naturel de penser que l'on ne pourra pas décider dans une telle mesure de la pertinence du test quant au choix entre les hypothèses de variables indépendantes ou non.

5.3.2 Etude de la puissance du test

Il nous est alors apparu nécessaire d'étudier le test d'indépendance fourni par la mesure de dépendance quadratique dans la situation suivante :

Hypothèses :

H_0:Les variables sont indépendantes, $Q = 0$.
H_1:Les variables sont dépendantes, $Q \neq 0$.

Dans cette situation, Kankainen [44] a montré que ce test d'indépendance est indépendant de la loi des variables.

Statistiques de test : On choisit comme satistique de test, l'estimateur $\widehat{Q}$ de la mesure de dépendance quadratique.

Loi sous H_1 :

D'après ce que nous avons pu voir dans la partie précédente, on va approcher la loi de $\widehat{Q}$ sous H_1 par une loi normale de moyenne Q de variance σ^2/N.

Loi sous H_0 :

Nous nous basons ici sur les travaux de Kankainen [44].
Elle a montré, sous l'hypothèse H_0, que l'estimateur $N\widehat{Q}$ suit une loi $\gamma\chi^2(\beta)$ où γ et β sont définis de la manière suivante,

$$\gamma = \frac{V_1}{2E_1}$$

et

$$\beta = \frac{2E_1^2}{V_1}.$$

E_1 est l'espérance de $\widehat{Q}$ sous H_0,

$$E_1 = \prod_{k=1}^{K} \int \mathcal{K}^2(x)dx - \prod_{k=1}^{K} E[\pi_{T_k}(T_k)] - \sum_{k=1}^{K} (\int \mathcal{K}^2(x)dx - E[\pi_{T_k}(T_k)]) \prod_{l=1, l\neq k}^{K} E[\pi_{T_l}(T_l)]$$

et V_1 la variance de $\widehat{Q}$ sous H_0,

$$\begin{aligned}
V_1 &= 2\prod_{k=1}^{K} E[\pi_{T_k}(T_k)]^2 - 4\prod_{k=1}^{K} E[\pi_{T_k}(T_k)^2] + 4\prod_{k=1}^{K} E[\pi_{2,T_k}(T_k)] \\
&+ 2\sum_{k=1}^{K}(E[\pi_{2,T_k}(T_k)] - E[\pi_{T_k}(T_k)]^2) \prod_{l=1,l\neq k}^{K} E[\pi_{T_l}(T_l)]^2 \\
&- 4\sum_{k=1}^{K}(E[\pi_{2,T_k}(T_k)] - E[\pi_{T_k}(T_k)^2]) \prod_{l=1,l\neq k}^{K} E[\pi_{T_l}(T_l)^2] \\
&+ 2\sum_{k=1}^{K}\sum_{m=1,m\neq k}^{K} (E[\pi_{T_k}(T_k)]^2 E[\pi_{T_m}(T_m)]^2 - 2E[\pi_{T_k}(T_k)^2]E[\pi_{T_m}(T_m)]^2 \\
&+ E[\pi_{T_k}(T_k)^2]E[\pi_{T_m}(T_m)^2]) \prod_{l=1,l\neq k,m}^{K} E[\pi_{T_l}(T_l)]
\end{aligned}$$

Définition de la région critique :

On définit alors l'erreur de première espèce par la probabilité d'accepter H_1 sous l'hypothèse H_0.
On se donne α et on cherche q_α tel que

$$P_{\mathrm{H}_0}(\widehat{Q} > q_\alpha) = 1 - F_{\gamma\chi^2(\beta)}(Nq_\alpha)$$

Puissance du test :

On s'intéresse à l'erreur de deuxième espèce, la probabilité d'accepter H_0 sous l'hypothèse H_1. (Cette probabilité va nous permettre de voir si la valeur obtenue peut-être jugée significativement différente de 0.)
Avec la valeur q_α déterminée auparavant, on détermine alors,

$$P_{\mathrm{H}_1}(\widehat{Q} < q_\alpha) = \Phi\left(\frac{(q_\alpha - Q)\sqrt{N}}{\sigma}\right)$$

La puissance du test sera alors de $1 - P_{\mathrm{H}_1}(\widehat{Q} > q_\alpha)$

Résultats empiriques :

A présent, nous nous intéressons à l'évolution de la puissance du test par rapport à la taille de fenêtre.

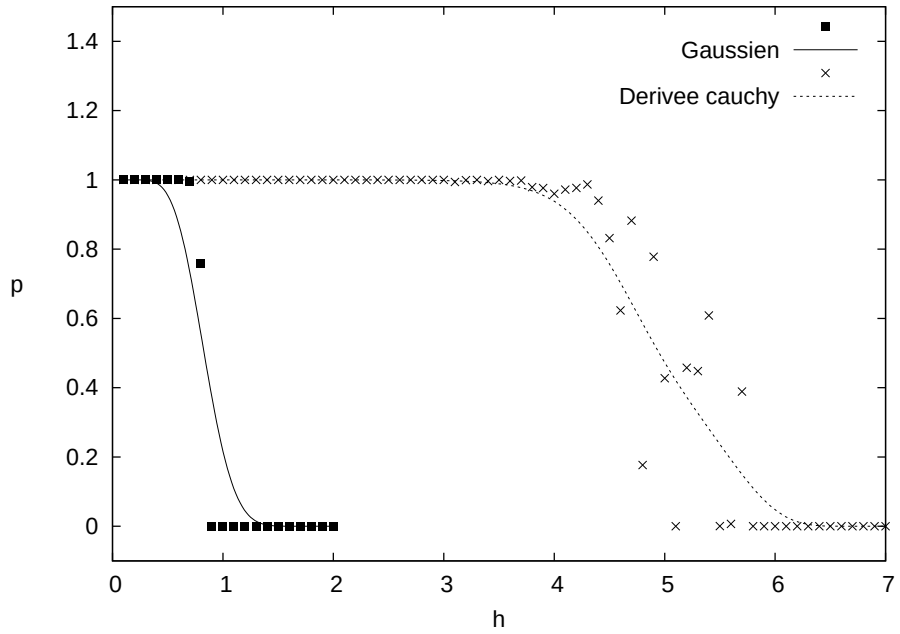


FIG. 5.3 – *Evolution de la puissance du test en fonction de la taille de fenêtre pour différents noyaux.*

Nous nous plaçons dans le cadre d'un mélange linéaire avec 2 sources, On choisit ici deux sources uniformes sur $[-1, 1]$. Les distributions des sources sont donc symétriques.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{bmatrix} \cos(\pi/8) & \sin(\pi/8) \\ -\sin(\pi/8) & \cos(\pi/8) \end{bmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$$

Nous allons alors calculer la puissance de test par rapport aux variables X_1 et X_2. Ce premier graphique 5.3 représente la puissance du test pour des noyaux différents avec en abscisse la taille de fenêtre et en ordonné la puissance correspondante du test.
Sur ce graphique, nous constatons alors que pour chaque noyau, il existe une taille de fenêtre maximale au delà de laquelle, le test d'indépendance fourni par la mesure de dépendance quadratique n'est plus du tout pertinent. Nous verrons dans la dernière section de ce chapitre, une illustration de cette constatation dans le cadre du problème de séparation de sources.

5.3.3 Commentaires par rapport à la taille de l'échantillon

D'après les résultats obtenus dans la section (5.2), nous pouvons remarquer que plus la taille d'échantillon sera grande, plus la mesure de dépendance quadratique

sera précise. Nous allons illustrer ceci en observant le comportement de la mesure de dépendance quadratique dans le contexte d'un mélange linéaire. Plus explicitement, nous choisissons deux variables aléatoires indépendantes uniformes sur $[-1, 1]$. Puis ces variables sont mélangées à l'aide d'une matrice. Nous choisissons de paramétrer cette matrice sous deux formes différentes.

Avec une matrice de rotation

Le mélange est ici défini par,

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$$

où θ est un paramètre que nous allons faire varier.

A partir de ceci, nous voulons étudier pour quelles valeurs de θ, la mesure de dépendance quadratique peut-être capable de dire que les variables X_1 et X_2 sont dépendantes avec un certain taux d'erreur fixé.

Pour cela, nous allons suivre le cheminement suivant,

1. Calcul de la valeur q_β de telle sorte que
$$P_{\mathrm{H}_0}(\widehat{Q}(S_1, S_2) > q_\beta) = \beta$$
où nous prendrons $\beta = 0.05$. (c'est la probabilité de dire que les variables sont dépendantes sous l'hypothèse qu'elles sont indépendantes.)
2. Calcul de la probabilité de dire que les variables sont indépendantes alors que $\theta \neq 0[\pi/2]$.
$$\alpha = P_{\mathrm{H}_1}(\widehat{Q}(X_1, X_2) < q_\beta)$$

Dans les figures 5.4 et 5.5, nous avons tracé l'évolution de la valeur de α en fonction de θ pour différentes tailles d'échantillon.

Avec une matrice symétrique

Le mélange est ici défini par,

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$$

où θ est un paramètre que nous allons faire varier.

Par le même procédé que pour la matrice de rotation, nous représentons la valeur de α en fonction de θ pour des tailles d'échantillon différentes. Nous présentons les résultats dans les figures 5.6 et 5.7.

Ces différentes figures nous permettent de voir que plus on prend une taille d'échantillon grande, plus on pourra discerner des «dépendances» faibles des observations.

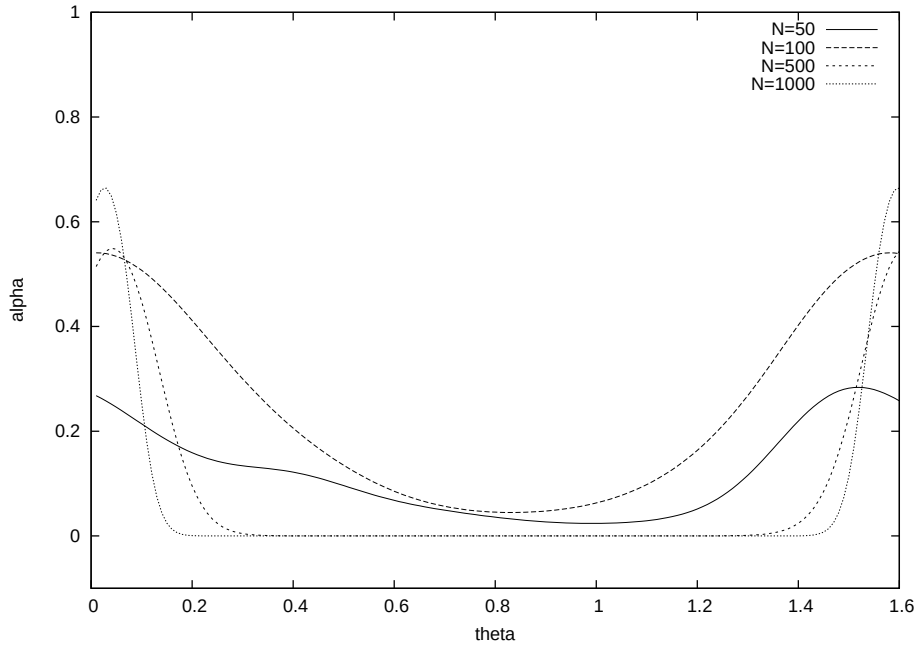


FIG. 5.4 – *Caractère discriminant de la mesure de dépendance quadratique par rapport à une matrice de rotation. Estimation de la mesure de dépendance quadratique par un noyau gaussien avec une taille de fenêtre de 0.5.*

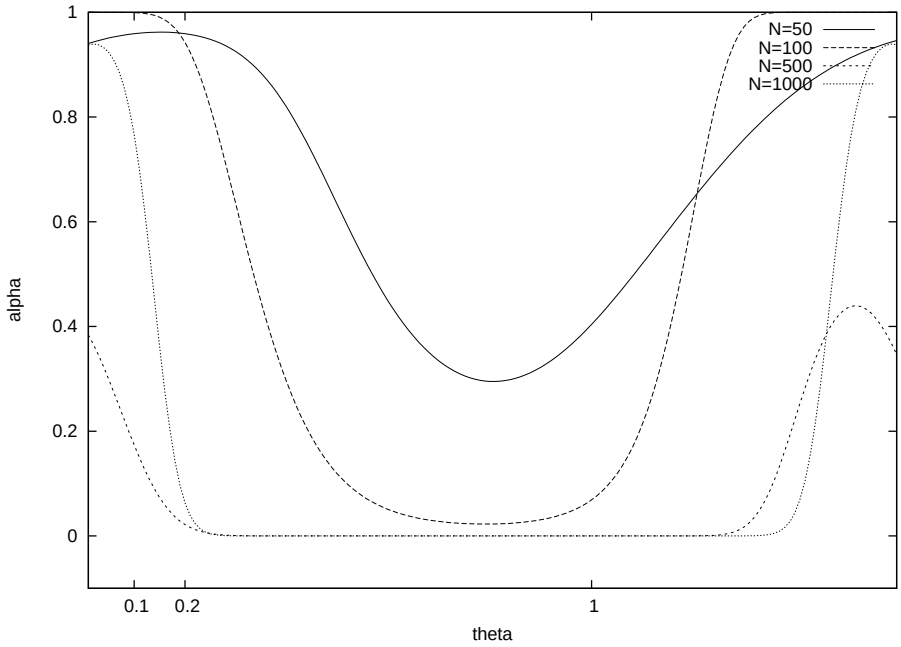


FIG. 5.5 – *Caractère discriminant de la mesure de dépendance quadratique par rapport à une matrice de rotation. Estimation de la mesure de dépendance quadratique par un noyau dérivée seconde de Cauchy carré avec une taille de fenêtre de 3.*

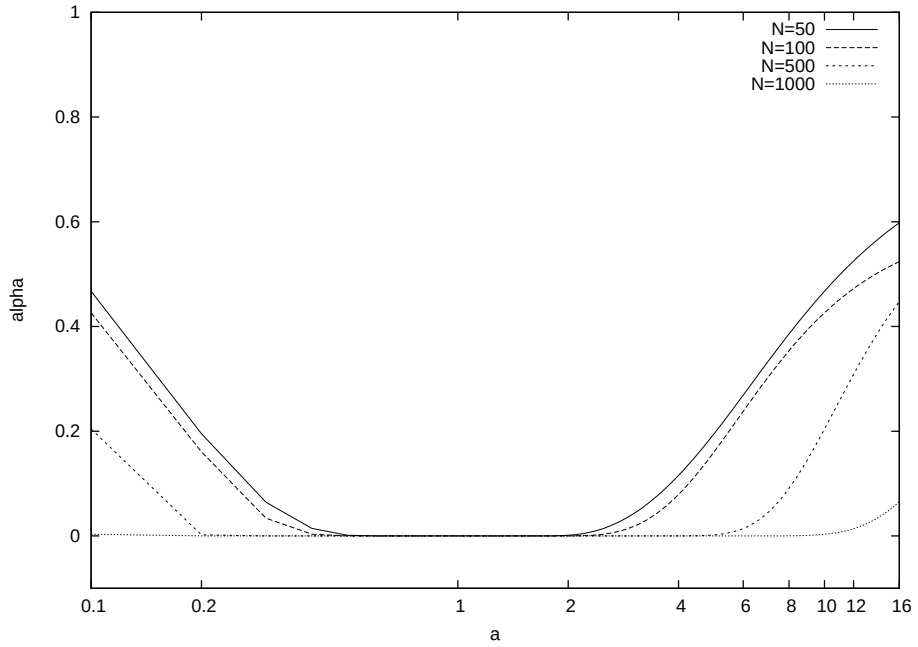


FIG. 5.6 – *Caractère discriminant de la mesure de dépendance quadratique par rapport à une matrice symétrique. Estimation de la mesure de dépendance quadratique par un noyau gaussien avec une taille de fenêtre de 0.5.*

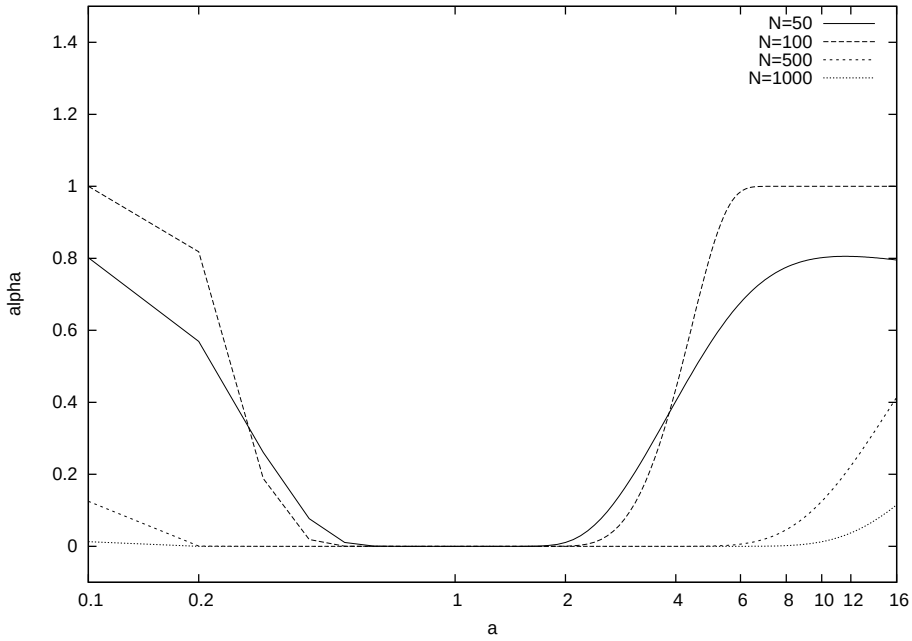


FIG. 5.7 – *Caractère discriminant de la mesure de dépendance quadratique par rapport à une matrice symétrique. Estimation de la mesure de dépendance quadratique par un noyau dérivée seconde de Cauchy carré avec une taille de fenêtre de 3.*

5.3.4 Illustration dans le cadre de la minimisation de la mesure de dépendance quadratique avec un mélange linéaire de sources

Nous nous plaçons dans le cadre d'un mélange linéaire avec 2 sources. On choisit ici deux sources uniformes sur $[-1, 1]$. Les distributions des sources sont donc symétriques.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{bmatrix} \cos(\pi/8) & \sin(\pi/8) \\ -\sin(\pi/8) & \cos(\pi/8) \end{bmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$$

Nous nous intéressons alors à la caractérisation du minimum dans le cas d'un mélange linéaire.

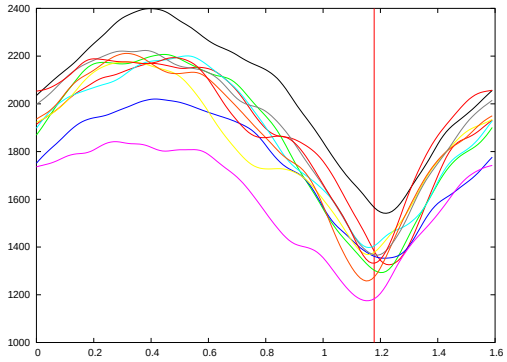
Nous définissons alors à présent, $\mathbf{Y} = (Y_1, Y_2)^T$, par,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

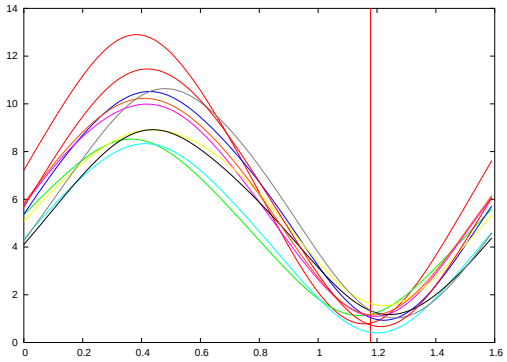
Nous examinons alors comment varie le critère de dépendance quadratique de $\mathbf{Y}$ en fonction de θ. Les graphiques suivants représentent pour différentes tailles de fenêtre et différents noyaux la superposition de 10 courbes correspondant à des échantillons différents. La valeur de $\theta = \pi/8$ est marquée par une barre verticale.

Noyau gaussien :

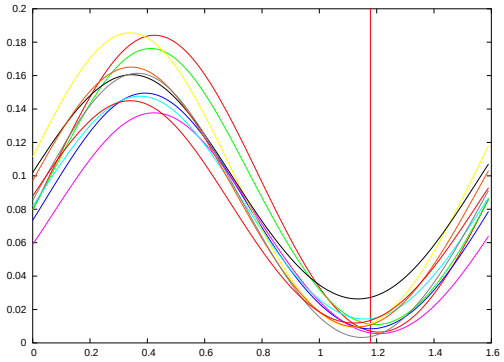
h=0.1



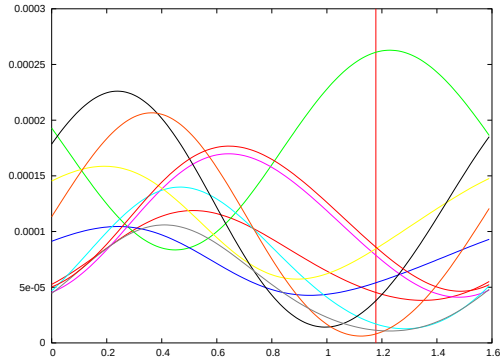
h=0.50625



h=1.139062

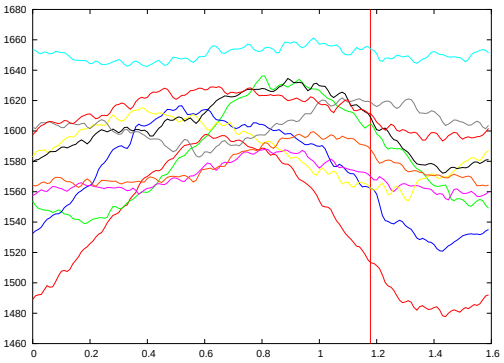


h=2.562889

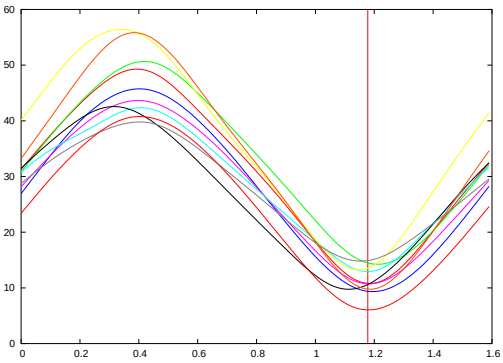


Noyau dérivée Cauchy carré :

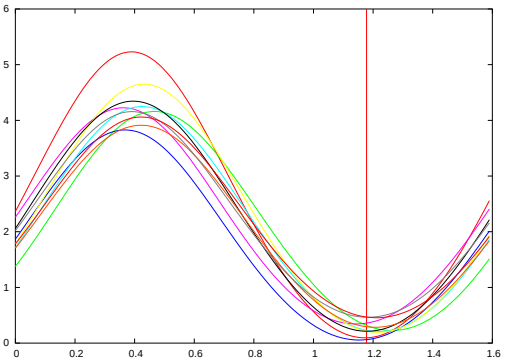
h=0.1



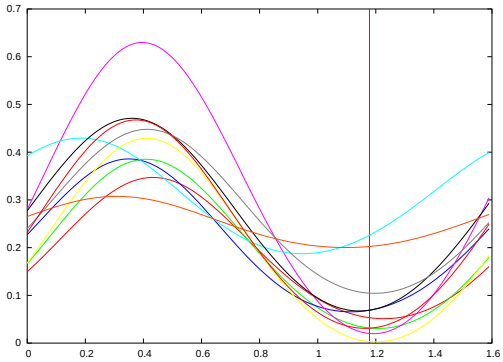
h=1.139062



h=3.844333



h=5.7665



En comparant ces deux catégories de résultats, on observe que pour des tailles de fenêtres trop grandes, le test d'indépendance n'est plus du tout significatif et par conséquent, il est complètement vain de chercher le minimum.

Par contre, pour une taille de fenêtre trop petite, le test d'indépendance reste significatif en pratique, pourtant le minimum est aussi absolument non significatif. Ceci vient du fait qu'avec une taille de fenêtre trop petite, le noyau étant très ressérré au-

tour de l'origine, seulement un petit nombre de variables de l'échantillon seront prises en compte et donc la variance de l'estimateur sera grande. On observe d'ailleurs à ce propos, que pour des petites tailles de fenêtres de l'ordre de 0.1, il est nécessaire de prendre des échantillons de taille 10000 afin de pouvoir utiliser les résultats de convergence en loi.

En conclusion, nous proposons de choisir la taille de fenêtre la plus grande possible pour laquelle la puissance du test est maximale, en écartant les tailles de fenêtres trop petites.

5.4 Estimation des fonctions scores

Nous avons pu voir que dans le cas de l'utilisation du critère d'information mutuelle, les fonctions scores jouent un rôle important. Par conséquent, l'estimation de celles-ci est aussi très importante. Dans [68] on utilise pour estimer les fonctions scores, les estimateurs à noyaux utilisés pour l'estimation de la densité. Puis, dans [1], nous proposons une autre estimation des fonctions scores en utilisant des splines. Pour plus de détails et une comparaison entre les deux méthodes d'estimation, voir [1]. Enfin, dans [6], Babaie-Zadeh propose d'estimer les fonctions scores marginales à partir des fonctions scores jointes en utilisant des estimateurs à noyaux ou des splines. Cependant, dans la section 3.2.3, nous avons proposé de minimiser un critère empirique qui nous a conduit alors à proposer une nouvelle méthode d'estimation des fonctions scores. De plus, nous avons présenté deux méthodes différentes d'estimation, l'une basée sur l'estimation de l'entropie en utilisant l'estimateur d'espérance empirique, et l'autre en utilisant une discrétisation de l'intégrale. Nous détaillons ici les calculs nous permettant d'obtenir les expressions de ces fonctions score.

5.4.1 Estimation de la fonction score en dérivant l'entropie estimée à l'aide de la moyenne empirique

Rappelons ici la définition de cet estimateur,

$$\widehat{\phi}_i^m(\mathbf{T}(j)) = N\partial_{kj}^2 \widehat{H}^m(\mathbf{T})$$

où

$$\widehat{H}^m(\mathbf{T}) = \frac{1}{N}\sum_{j=1}^{N} \log \widehat{p}_{\mathbf{T}}^m(\mathbf{T}(j))$$

et $\widehat{p}^m_{\mathbf{T}}$ est une estimation à noyaux de la densité de $\mathbf{T}$ i.e.

$$\widehat{p}^m_{\mathbf{T}}(x) = \frac{1}{N\prod_{i=1}^K \widehat{\sigma}_{T_i}} \sum_{n=1}^N \prod_{k=1}^K \mathcal{K}\left(\frac{x_k - T_k(n)}{\widehat{\sigma}_{T_k}}\right)$$

Alors en calculant la dérivée de cette expression, on obtient une expression de $\widehat{\phi}^m_i$. Etablissons tout d'abord quelques lemmes qui vont nous conduire au résultat.

Lemme 5.4.1 *Notons $\mathbf{T}'$ la variable normée corespondant à $\mathbf{T}$. On a alors, pour tout $n = 1, \ldots, N$,*

$$\mathbf{T}'(n) = \left(\frac{T_1(n)}{\widehat{\sigma}_{T_1}}, \ldots, \frac{T_K(n)}{\widehat{\sigma}_{T_K}}\right)^T$$

Alors nous pouvons écrire, grâce aux propriétés de l'estimateur d'entropie,

$$\widehat{H}^m(\mathbf{T}) = \widehat{H}^m(\mathbf{T}') + \sum_{k=1}^K \log \widehat{\sigma}_{T_k}$$

Nous en déduisons alors le lemme suivant, en dérivant par rapport à la variable $T_i(j)$

Lemme 5.4.2 *En dérivant par rapport à la variable $T_i(j)$,*

$$\partial_{ij}\widehat{H}^m(\mathbf{T}) = \sum_{k=1}^K \sum_{l=1}^K \frac{\partial}{\partial T'_k(l)} \widehat{H}^m(\mathbf{T}')\partial^2_{ij} T'_k(l) + \partial^2_{ij} \sum_{k=1}^K \log \widehat{\sigma}_{T_k}$$

Regardons à présent en détail chaque terme de ce développement,

Lemme 5.4.3 *Pour tout $i = 1, \ldots, K$ et $j = 1, \ldots, N$,*
Pour tout $k = 1, \ldots, K$, $k \neq i$,

$$\partial^2_{ij} T'_k(l) = 0$$

Et,

$$\partial^2_{ij} T'_i(l) = \frac{1}{\widehat{\sigma}_{T_i}} \left(\delta_{lj} - T_i(l)\frac{\partial \log \widehat{\sigma}_{T_i}}{\partial_{ij}}\right)$$

où

$$\partial_{ij} \log \widehat{\sigma}_{T_i} = \frac{T_i(j) - \widehat{E}(T_i)}{N\widehat{\sigma}^2_{T_i}}$$

Enfin,

Lemme 5.4.4

$$\begin{aligned}\frac{\partial}{\partial T_i'(l)} N\widehat{H}^m(\mathbf{T}') &= -\frac{\sum_{m=1}^N \mathcal{K}'(T_i'(l)-T_i'(m))\prod_{k\neq i}\mathcal{K}(T_k'(l)-T_k'(m))}{\sum_{m=1}^N\prod_{k=1}^K \mathcal{K}(T_k'(l)-T_k'(m))} \\ &+ \sum_{n=1}^N \frac{\mathcal{K}'(T_i'(j)-T_i'(l))\prod_{k\neq i}\mathcal{K}(T_k'(j)-T_k'(l))}{\sum_{m=1}^N\prod_{k=1}^K \mathcal{K}(T_k'(l)-T_k'(m))}\end{aligned}$$

On en déduit alors l'expression de l'estimateur de la fonction score dérivé de l'entropie,

Lemme 5.4.5 *Les fonctions $\widehat{\phi}_i^m$ pour tout $i = 1, \ldots, K$ seront définies par,*

$$\begin{aligned}\widehat{\phi}_i^m(\mathbf{T}(j)) &= N\partial_{ij}^2\widehat{H}^m(\mathbf{T}) \\ &= \widetilde{\phi}_i(T(j)) + \frac{T_i(j)-\widehat{E}(T_i)}{\widehat{\sigma}_{T_i}^2}\left[1-\sum_{m=1}^N \widetilde{\phi}_i(T(m))T_i(m)\right]\end{aligned}$$

où

$$\widetilde{\phi}_i(\mathbf{T}(j)) = \widehat{\phi}_i(T_i(j)) + \sum_{m=1}^N \frac{\mathcal{K}'[(T_i(m)-T_i(j))/\widehat{\sigma}_{T_i}]}{N\prod_{k=1}^K \widehat{\sigma}_{T_k}\widehat{\sigma}_{T_i}\widehat{p}_{\mathbf{T}}(\mathbf{T}(m))}\prod_{k=1,k\neq i}^K \mathcal{K}[(T_k(m)-T_k(j))/\widehat{\sigma}_{T_k}]$$

avec $\widehat{\phi}_i = -\partial_i \log \widehat{p}_{\mathbf{T}}^m$

5.4.2 Estimation de la fonction score en dérivant l'entropie estimée à l'aide d'une discrétisation de l'intégrale

Rappelons ici la définition de cet estimateur,

$$\widehat{\phi}_i^i(\mathbf{T}(j)) = N\partial_{kj}^2\widehat{H}^i(\mathbf{T})$$

où

$$\widehat{H}^i(\mathbf{T}) = -\sum_{\mathbf{l}} \log[\widehat{p}_{\mathbf{T}}^i(\widehat{E}(\mathbf{T})+\mathbf{l}:\mathbf{b}:\widehat{\sigma}_{\mathbf{T}})\prod_{k=1}^K \widehat{\sigma}_{T_k}]\widehat{p}_{\mathbf{T}}^i(\widehat{E}(\mathbf{T})+\mathbf{l}:\mathbf{b}:\widehat{\sigma}_{\mathbf{T}})\prod_{k=1}^K b_k\widehat{\sigma}_{T_k} + \sum_{k=1}^K \log\widehat{\sigma}_{T_k}$$

avec

$$\widehat{p}^i_{\mathbf{T}}(\widehat{E}(\mathbf{T})+\mathbf{l}:\mathbf{b}:\widehat{\sigma}_{\mathbf{T}})\prod_{k=1}^{K}\widehat{\sigma}_{T_k}=\frac{1}{N}\sum_{n=1}^{N}\prod_{k=1}^{K}\mathcal{K}\left(l_kb_k-\frac{\mathbf{T}(n)-\widehat{E}(\mathbf{T})}{\widehat{\sigma}_{\mathbf{T}}}\right)$$

et $\mathbf{b}=(b_1,\ldots,b_K)^T$ désigne le vecteur des pas de discrétisation (il est formé de petits nombres positifs) et pour tous vecteurs $\mathbf{u}$ et $\mathbf{v}$, $\mathbf{u}:\mathbf{v}=\mathbf{w}$ où $w_k=u_kv_k$.

De même que précédemment, grâce à la forme particulière de l'estimateur, nous pouvons en déduire la formule suivante,

Lemme 5.4.6 *Notons $\mathbf{T}'$ la variable normée corespondant à $\mathbf{T}$. On a alors, pour tout $n=1,\ldots,N$,*

$$\mathbf{T}'(n)=\left(\frac{T_1(n)-\widehat{E}(T_1)}{\widehat{\sigma}_{T_1}},\ldots,\frac{T_K(n)-\widehat{E}(T_K)}{\widehat{\sigma}_{T_K}}\right)^T$$

Alors nous pouvons écrire, grâce aux propriétés de l'estimateur d'entropie,

$$\widehat{H}^m(\mathbf{T})=\widehat{H}^m(\mathbf{T}')+\sum_{k=1}^{K}\log\widehat{\sigma_{T_k}}$$

Nous en déduisons alors le lemme suivant, en dérivant par rapport à la variable $T_i(j)$

Lemme 5.4.7 *En dérivant par rapport à la variable $T_i(j)$,*

$$\partial^2_{ij}\widehat{H}^i(\mathbf{T})=\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{\partial}{\partial T'_k(l)}\widehat{H}^i(\mathbf{T}')\partial^2_{ij}T'_k(l)+\partial^2_{ij}\sum_{k=1}^{K}\log\widehat{\sigma}_{T_k}$$

Regardons à présent en détail chaque terme de ce développement, ils vont en effet, être différents de ceux obtenus dans le paragraphe précédent,

Lemme 5.4.8 *Pour tout $i=1,\ldots,K$ et $j=1,\ldots,N$,*
Pour tout $k=1,\ldots,K$, $k\neq i$,

$$\partial^2_{ij}T'_k(l)=0$$

Et,

$$\partial^2_{ij}T'_i(l)=\frac{1}{\widehat{\sigma}_{T_i}}\left(\delta_{lj}-\frac{1}{N}-\left(\frac{T_i(l)-\widehat{E}(T_i)}{\widehat{\sigma}_{T_i}}\right)\partial^2_{ij}\log\widehat{\sigma}_{T_i}\right)$$

où

$$\partial^2_{ij}\log\widehat{\sigma}_{T_i}=\frac{T_i(j)-\widehat{E}(T_i)}{N\widehat{\sigma}^2_{T_i}}$$

Enfin,

Lemme 5.4.9

$$\begin{aligned}\frac{\partial}{\partial T_i'(l)} N\widehat{H}^m(\mathbf{T}') &= \sum_{\mathbf{l}} \left(\log \widehat{p}^i_{\mathbf{T}}(\widehat{E}(\mathbf{T}) + \mathbf{l} : \mathbf{b} : \widehat{\sigma}_{\mathbf{T}}) \prod_{k=1}^{K} \widehat{\sigma}_{T_k} + 1 \right) \\ &\quad \frac{1}{\widehat{\sigma}_{T_k}} \mathcal{K}' \left(\frac{l_k b_k \widehat{\sigma}_{T_k} + T_k(j) - \widehat{E}(T_k)}{\widehat{\sigma}_{T_k}} \right) b_k \\ &\quad \prod_{i=1, i\neq K}^{K} \mathcal{K}' \left(\frac{l_i b_i \widehat{\sigma}_{T_i} + T_i(j) - \widehat{E}(T_i)}{\widehat{\sigma}_{T_i}} \right) b_i\end{aligned}$$

Enfin, l'estimateur de la fonction score est donné par,

$$\begin{aligned}\widehat{\phi}^i_k(T_k(j)) &= \widehat{\phi}^{\mathbf{b}}_k(\mathbf{T}(j)) - \frac{1}{N} \sum_{m=1}^{N} \widehat{\phi}^{\mathbf{b}}_k(\mathbf{T}(m)) \\ &+ \sum_{r=1}^{K} \frac{T_r(j) - \widehat{E}(T_r)}{\widehat{\sigma}^2_{T_r}} \left[1 - \sum_{m=1}^{N} \widehat{\phi}^{\mathbf{b}}_k(\mathbf{T}(m))(T_r(m) - \widehat{E}(T_r)) \right]\end{aligned}$$

où

$$\begin{aligned}\widehat{\phi}^{\mathbf{b}}_k(\mathbf{T}(j)) = \sum_{\mathbf{l}} [\log \widehat{p}^i_{\mathbf{T}}(\widehat{E}(\mathbf{T}) + \mathbf{l} : \mathbf{b} : \widehat{\sigma}_{\mathbf{T}}) \widehat{\sigma}_{\mathbf{T}} + 1] \\ \frac{1}{\widehat{\sigma}_{T_k}} \mathcal{K}' \left(\frac{l_k b_k \widehat{\sigma}_{T_k} + T_k(j) - \widehat{E}(T_k)}{\widehat{\sigma}_{T_k}} \right) b_k \prod_{i=1, i\neq K}^{K} \mathcal{K}' \left(\frac{l_i b_i \widehat{\sigma}_{T_i} + T_i(j) - \widehat{E}(T_i)}{\widehat{\sigma}_{T_i}} \right) b_i\end{aligned}$$

5.5 Comparaison : Estimateurs pour l'Information Mutuelle

D'après les travaux de Joe [38], nous pouvons établir le biais et la variance des estimateurs de l'information mutuelle utilisant un estimateur à noyau pour la densité. Ces calculs de biais sont établis pour des densités vérifiant des propriétés particulières.

5.5.1 Définitions et Hypothèses

Toutes les intégrales seront supposées finies.

Définitions

Définition 5.5.1 *Soit X un vecteur aléatoire de dimension K, nous définissons l'entropie de X par,*

$$H(X) = -\int \log f_X(x) f_X(x) dx,$$

où f_X représente la densité de X.

Nous allons alors considérer l'estimateur de l'entropie de X suivant,

Définition 5.5.2 *Soient $X_1, \ldots, X_N$ un écantillon du vecteur aléatoire X, alors,*

$$\hat{H}^m(X_1, \ldots, X_N) = \frac{1}{n} \sum_{i=1}^{N} \log \hat{f}(X_i)$$

où

$$\hat{f}(x) = (nh^K)^{-1} \sum_{i=1}^{N} k\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{N} k_h(x - X_i)$$

et k est un noyau à K variables avec $k_h(u) = h^{-K} k(u/h)$.

Signalons que l'estimateur considéré par Joe correspond à l'estimateur que nous avons défini dans le paragraphe 3.2.3 dans le cas de variables normées. Cependant, si nous reprenons l'estimateur de l'entropie défini dans le chapitre 3, nous remarquons qu'il vérifie l'égalité suivante:

$$\widehat{H}^m(\mathbf{X}) = \widehat{H}^m(\mathbf{X}') + \sum_{k=1}^{K} \log \widehat{\sigma}_{X_k}$$

où $\mathbf{X}'$ désigne le vecteur normé correspondant à $\mathbf{X}$.

En outre, nous nous intéressons ici au calcul de l'estimateur de l'information mutuelle, nous constatons alors,

$$\begin{aligned} \widehat{I}^m(\mathbf{X}) &= \sum_{i=1}^{K} \widehat{H}^m(X_i) - \widehat{H}^m(\mathbf{X}) \\ &= \sum_{i=1}^{K} \widehat{H}^m(X_i') - \widehat{H}^m(\mathbf{X}') \end{aligned}$$

Par cette constatation, il nous suffit alors d'étudier le biais dans le cas de variables normées, en reprenant les travaux de Joe.

Hypothèses

Nous faisons alors les hypothèses suivantes :

A. S est un ensemble borné de telle sorte que

$$\int_S \log f_X(x) f_X(x) dx \approx \int_{R^K} \log f_X(x) f_X(x) dx$$

B. Les K composantes de X ont la même échelle, c'est-à-dire sans perte de généralité, on les suppose normées.

Cm. f_X admet des dérivées jusqu'à l'ordre m inclus.

D. $\int (\log f_X(x))^2 f_X(x) dx$ existe.

E. Le noyau est paire.

F. Le noyau peut s'écrire, $k(u) = \prod k_0(u_j)$ où k_0 est un noyau symétrique à une dimension satisfaisant, $\int v^2 k_0(v) dv = 1$

G. f_X admet des dérivées première et seconde continues et $|\partial f_X / \partial x_j|, |\partial^2 f_X / \partial x_j^2|$ sont majorées par des fonctions intégrables.

Introduisons à présent quelques notations,

Fonction de répartition empirique: On la notera F_n.
Alors, $\hat{f}_X = \int k_h(x-y) dF_n(y)$.

D'autres notations :

- $f_h(x) = E\hat{f}(x) = \int k_h(x-y) dF(y)$
- $U_n(x) = n^{1/2}(F_n(x) - F(x))$
- $V_n(x) = n^{1/2}(\hat{f}(x) - f_h(x)) = \int k_h(x-y) dU_n(y)$

5.5.2 Etude de l'estimateur d'entropie dans le cas d'une densité à plusieurs variables

Développement asymptotique

D'après les résultats de Joe, on peut écrire que :

$$\begin{aligned} \hat{H}(X_1, \ldots, X_N) &= -\int_S \log(f_h(x)) dF(x) \\ &\quad + n^{-1/2} \int A(x) dU_n(x) + n^{-1} \iint B(x,y) dU_n(x) dU_n(y) \\ &\quad + n^{-3/2} \iiint C(x,y,z) dU_n(x) dU_n(y) dU_n(z) + o(n^{-3/2}) \end{aligned}$$

où

- $A(x) = -\int_S (f_h(y))^{-1} k_h(y-x) dF(y) - \log(f_h(x)) \mathbb{1}_S(x)$
- $B(x,y) = 0.5 \int_S (f_h(y))^{-2} k_h(z-x) k_h(z-y) dF(z) - (f_h(x))^{-1} k_h(x-y) \mathbb{1}_S(x)$
- $C(x,y,z) = -1/6 \int_S (f_h(y))^{-3} k_h(w-x) k_h(w-y) k_h(w-z) dF(w) + 0.5 (f_h(y))^{-2} k_h(x-y) k_h(x-z) \mathbb{1}_S(x)$

Biais de l'estimateur de l'entropie

D'après le lemme 2.1 de Joe, on déduit le biais,

$$E[\hat{H}(X_1, \ldots, X_N)] = -\int_S \log(f_h(x)) dF(x) + n^{-1} a_1(h) + n^{-2} a_2(h) + o(n^{-2})$$

où

- $a_1(h) = 0.5 - h^{-K} k(0) \int_S f(x) (f_h(x))^{-1} dx + 0.5 h^{-K} K_2 \int_S f(x) f_h^*(x) (f_h(x))^{-2} dx$
- $a_2(h) = 1/6 - 1/6 h^{-2K} K_3 \int_S f(x) f_h^{**}(x) (f_h(x))^{-1} dx + 0.5 h^{-2K} k^2(0) \int_S f(x) (f_h^2(x))^{-1} dx + O(h^{-K})$

et

- $K_2 = \int k^2(x) dx$
- $f_h^*(x) = h^{-K} \int l((y-x)/h) dF(x)$, avec $l(u) = k^2(u)/K_2$.
- $K_3 = \int k^3(x) dx$
- $f_h^{**}(x) = h^{-K} \int m((y-x)/h) dF(x)$, avec $m(u) = k^3(u)/K_3$.

De ces expressions, nous pouvons en déduire le biais de l'estimateur de l'entropie dans le cas de variables normées et pour n'importe quel noyau. Nous nous limiterons ici à écrire le biais à l'ordre 1.

$$E[\hat{H}(X_1, \ldots, X_N)] = -\int_S \log(f_h(x)) dF(x) + n^{-1} a_1(h) + o(n^{-1})$$

où $a_1(h) = 0.5 - h^{-K}[0.5 K_2 - k(0)] \int_S f(x) (f_h(x))^{-1} dx + 0.5 h^{-K} K_2 \int_S f(x) (f_h^*(x) - f_h(x)) (f_h(x))^{-2} dx$

D'autre part, d'après l'hypothèse C_2,

$f_h^*(x) - f_h(x) = 0.5 h^2 tr f''(x) \left[1 - \int v^2 l_0(v) dv\right] + o(h^2)$

où $l_0(u) = k_0^2(u)/K_{02}$ et $K_{02} = \int k_0^2(x) dx$

On obtient alors le biais sous la forme,

$$\begin{aligned} E[\hat{H}(X_1,\ldots,X_N)] &= -\int_S \log(f_h(x))dF(x) \\ &+ n^{-1}0.5 - n^{-1}h^{-K}[0.5K_2 - k(0)]\int_S f(x)(f_h(x))^{-1}dx \\ &+ 0.25n^{-1}h^{2-K}K_2\int_S f(x)trf''(x)\left[1-\int v^2 l_0(v)dv\right](f_h(x))^{-2}dx + o(n^{-1}) \end{aligned}$$

5.5.3 Application à l'information mutuelle

Biais de l'estimateur de l'information mutuelle

Cette expression du biais de l'estimateur d'entropie nous permet alors de déduire le biais de l'estimateur de l'information mutuelle correspondante:

$$\begin{aligned} &E[\hat{I}(X_1,\ldots,X_N)] \\ &= \sum_{i=1}^{K} E[\hat{H}(X_i)] - E[\hat{H}(X_1,\ldots,X_N)] \\ &= -\sum_{i=1}^{K}\int_S \log(f_{X_i,h}(x))dF_{X_i}(x) + \int_S \log(f_{X,h}(x))dF_X(x) \\ &\quad +n^{-1}0.5(K-1) - n^{-1}\left\{h^{-1}K[0.5K_{02}-k(0)]\sum_{i=1}^{K}\int_S f_{X_i}(x)(f_{X_i,h}(x))^{-1}dx\right. \\ &\qquad\qquad \left. - h^{-K}[0.5K_2-k(0)]\int_S f_X(x)(f_{X,h}(x))^{-1}dx\right\} \\ &\quad +0.25n^{-1}\left\{hK_{O2}\sum_{i=1}^{K}\int_S f_{X_i}(x)trf''_{X_i}(x)\left[1-\int v^2 l_0(v)dv\right](f_{X_i,h}(x))^{-2}dx\right. \\ &\qquad \left. -h^{2-K}K_2\int_S f_X(x)trf''_X(x)\left[1-\int v^2 l_0(v)dv\right](f_{X,h}(x))^{-2}dx\right\} \\ &+ o(n^{-1}) \end{aligned}$$

Remarque 5.5.1 *Remarquons ici que la quantité définie par*

$$\begin{aligned} F(X_1,\ldots,X_N) &= -\sum_{i=1}^{K}\int_S \log(f_{X_i,h}(x))dF_{X_i}(x) + \int_S \log(f_{X,h}(x))dF_X(x) \\ &= \int \log\left(\frac{f_{X,h}(x)}{\prod_{i=1}^{K} f_{X_i,h}(x)}\right) dF_X(x) \end{aligned}$$

peut-être aussi utilisée comme une mesure de dépendance. Il est alors intéressant de noter que l'estimation de ce critère F est asymptotiquement sans biais.

Biais de l'estimateur C dérivé de l'information mutuelle pour les mélanges post non linéaires

Nous pouvons alors ici faire la différence entre l'estimation de l'information mutuelle et celle du critère obtenu par Taleb dans le cadre de mélange post non linéaire (les définitions sont rappelées dans le paragraphe 4.4.8). En effet, ici il est nécessaire de faire le lien avec l'estimateur de la densité, ce qui peut entraîner l'existence d'un biais.

$$\begin{aligned} &E[\hat{C}(Y_1,\ldots,Y_N)] \\ &= \sum_{i=1}^{K}(E[\hat{C}(Y_i)] - E[\hat{H}(Z_i)]) - \log|det\mathbf{B}| \\ &= -\sum_{i=1}^{K}\left(\int_S \log(f_{Y_i,h}(x))dF_{Y_i}(x) - \int_S \log(f_{Z_i,h}(x))dF_{Z_i}(x)\right) \\ &\quad -n^{-1}h^{-1}K[0.5K_{02}-k(0)]\sum_{i=1}^{K}\int_S \left\{f_{Y_i}(x)(f_{Y_i,h}(x))^{-1} - \int_S f_{Z_i}(x)(f_{Z_i,h}(x))^{-1}\right\}dx \\ &\quad +0.25n^{-1}hK_{O2}\left[1-\int v^2 l_0(v)dv\right] \\ &\quad \sum_{i=1}^{K}\int_S \left\{f_{Y_i}(x)trf''_{Y_i}(x)(f_{Y_i,h}(x))^{-2} - f_{Z_i}(x)trf''_{Z_i}(x)(f_{Z_i,h}(x))^{-2}\right\}dx + o(n^{-1}) \end{aligned}$$

D'autre part, introduisons le développement par rapport à la densité,

$$-\int_S (\log f_h - \log f)dF = 0.5h^2\int_S f''d\mu + o(h^2)$$

.

Le biais de l'estimateur C des mélanges post non linéaires s'écrit alors,

$$
\begin{aligned}
&E[\hat{C}(Y_1,\ldots,Y_N)] \\
&= \sum_{i=1}^{K}(E[\hat{C}(Y_i)] - E[\hat{H}(Z_i)]) - \log|det\mathbf{B}| \\
&= -\sum_{i=1}^{K}\left(\int_S \log(f_{Y_i}(x))dF_{Y_i}(x) - \int_S \log(f_{Z_i}(x))dF_{Z_i}(x)\right) \\
&\quad +0.5h^2\sum_{i=1}^{K}\int_S (f''_{Y_i}(x) - f''_{Z_i}(x))dx \\
&\quad -n^{-1}h^{-1}K[0.5K_{02} - k(0)]\sum_{i=1}^{K}\int_S \left\{f_{Y_i}(x)(f_{Y_i,h}(x))^{-1} - \int_S f_{Z_i}(x)(f_{Z_i,h}(x))^{-1}\right\}dx \\
&\quad +0.25n^{-1}hK_{O2}\left[1 - \int v^2 l_0(v)dv\right] \\
&\quad \sum_{i=1}^{K}\int_S \left\{f_{Y_i}(x)trf''_{Y_i}(x)(f_{Y_i,h}(x))^{-2} - f_{Z_i}(x)trf''_{Z_i}(x)(f_{Z_i,h}(x))^{-2}\right\}dx + o(n^{-1})
\end{aligned}
$$

Il apparaît dans cette expression des termes seulement en h, indépendants de n. On remarque qu'en faisant certaines hypothèses sur les densités, certains de ces termes peuvent être nuls.

Commentaires

- Le calcul du biais de l'information mutuelle montre que l'estimateur avec la moyenne empirique utilisé dans 3.2.3 est un estimateur asymptotiquement sans biais d'une mesure de dépendance notée F. De plus on remarque que F est une mesure de dépendance quelque soit le choix de la taille de fenêtre h.

  En ce qui concerne le critère de séparation C, le calcul du biais montre que certains termes seulement en fonction de h apparaissent. Cependant, ces termes doivent être étudiés plus précisément en fonction des densités considérées. Il est possible que pour certaines densités, quelques uns de ces termes soient non nuls, ce qui induirait un biais sur l'estimateur de C. Par ailleurs, on remarque que pour la recherche de la solution du problème de séparation aveugle de sources, nous nous intéressons au minimum de ce critère. Il serait donc plus intéressant d'étudier comment se comporte le biais de l'estimateur de C au voisinage de la solution.
- Enfin, dans [38], Joe remarque que si le noyau est choisi de manière particulière, certains termes du calcul du biais peuvent être annulés.

Chapitre 6

En pratique

Nous allons à présent illustrer certains résultats et certains problèmes soulevés dans les chapitres précédents.

Dans ce travail, la résolution du problème de séparation aveugle de sources consiste comme nous avons pu le voir, en la minimisation d'un critère. Dans le chapitre 3, nous avons fait la différence entre deux stratégies, «Estimer ensuite» et «Estimer d'abord». Dans le cadre de la stratégie «Estimer d'abord», comme la fonction à optimiser est une estimation du critère, il est tout à fait possible de la calculer en tous points. Il est donc envisageable, dans le cadre de mélanges paramétriques, de représenter la fonction à minimiser. Dans ce chapitre, nous allons étudier la difficulté de la mise en oeuvre de cette stratégie.

Dans un premier temps, nous observerons le comportement des différents critères de séparation utilisés pour la minimisation dans le cadre d'un mélange linéaire (section 6.1). Nous verrons que dans ce cas, une méthode de descente de gradient permet d'atteindre la solution quelque soit l'initialisation utilisée. Dans un deuxième temps, nous nous intéresserons aux mélanges non linéaires (section 6.2). Nous constaterons dans ce cas, la présence de minima locaux.

6.1 Avec un mélange linéaire

Nous nous plaçons ici dans le cadre d'un mélange linéaire de deux sources. Le modèle envisagé est donc un modèle paramétrique.

On choisit deux sources uniformes entre -1 et 1. Les distributions des sources sont donc symétriques.

Le mélange est défini par,

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta_0) & \sin(\theta_0) \\ -\sin(\theta_0) & \cos(\theta_0) \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$$

Nous prendrons en pratique, $\theta_0 = \pi/8$

Ici, nous avons choisi de travailler avec une matrice de rotation qui peut être paramétrée à l'aide d'une seule donnée. Cependant, nous allons présenter deux types de résultats. Tout d'abord, nous nous placerons dans l'hypothèse où la matrice de mélange est supposée être une matrice de rotation. Dans ce cas, la matrice de séparation sera paramétrée par θ de telle sorte que,

$$\mathbf{B} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}.$$

Les critères de séparation pourront alors être représentés en fonction d'un seul paramètre, l'angle de la matrice (paragraphe 6.1.1).

Ensuite, aucune hypothèse ne sera faite sur la matrice de mélange. Nous regarderons ce qu'il se passe lorsque la matrice de séparation est définie de manière tout à fait générale par ses termes anti diagonaux. Comme sous la seule hypothèse d'indépendance, le problème de séparation aveugle de sources est résolu à un facteur d'échelle près, nous poserons ainsi, sans perte de généralité, les termes diagonaux de la matrice égaux à 1 (paragraphe 6.1.2). Et la matrice de la structure de séparation s'écrira,

$$\mathbf{B} = \begin{pmatrix} 1 & a \\ b & 1 \end{pmatrix}$$

6.1.1 En 1 dimension

Nous nous plaçons ici dans un cadre particulier, i.e. nous modélisons la matrice de séparation par une matrice de rotation. La structure de séparation sera alors définie par,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Nous allons alors représenter les valeurs du critère en fonction du paramètre θ de la matrice.

Les figures suivantes 6.1, 6.2, 6.3 correspondent au tracé du critère en fonction de l'angle θ de la matrice de séparation pour différentes tailles de fenêtre, et différents noyaux. De plus, on représente θ sur une période de $\pi/2$, entre 0 et $\pi/2$.

On remarque la présence d'un maximum correspondant au point où le gradient s'annule mais où la solution n'est pas atteinte. Ici, c'est le point de coordonnée $\theta = \pi/8$.

En effet, d'après le lemme 4.4.2, θ vérifie en ce point,

$$\begin{bmatrix} \cos(\theta_0) & \sin(\theta_0) \\ -\sin(\theta_0) & \cos(\theta_0) \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

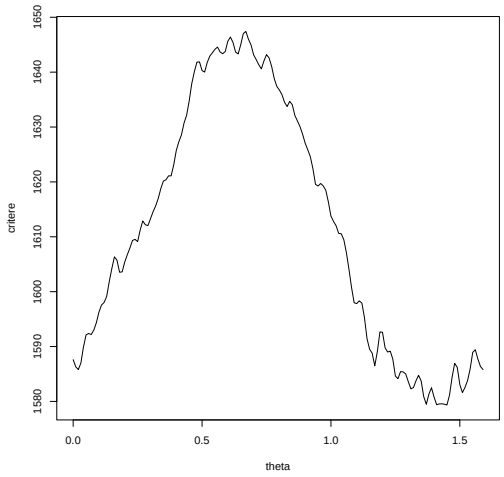


FIG. 6.1 – *Représentation de la mesure de dépendance quadratique avec un noyau dérivée seconde de Cauchy carré et une taille de fenêtre de 0.1 en fonction du mélange*

Par ailleurs, on constate que la minimisation du critère de séparation est rendue difficile par les erreurs dues aux estimations. En effet, l'estimation des critères de séparation étudiés dépend du choix du noyau et de la taille de fenêtre. Ceci peut avoir des conséquences sur la possiblité effective d'atteindre la solution. En effet, dans les figures 6.1, 6.2 et 6.3, on remarque la présence de minima locaux. La figure 6.2 correspond au choix d'une taille de fenêtre trop grande. Les figures 6.3 et 6.1 illustrent la situation où la taille de fenêtre est choisie trop petite. Dans ce cas, à cause de la variance trop grande, on voit apparaître des oscillations dans la courbe représentant le critère de séparation.

6.1.2 En 2 dimensions et plus

Dans ce cadre là, nous ne faisons plus d'hypothèse sur la matrice de mélange. Sans perte de généralité, la structure de séparation est donc définie de la manière suivante,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 & a \\ b & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Le fait de fixer les termes diagonaux de la matrice de séparation égaux à 1 ne particularise pas la recherche de solution du problème. En effet, le problème de séparation aveugle de sources étant résolu à un facteur d'échelle près et à une permutation près, nous fixons tout simplement le facteur d'échelle.

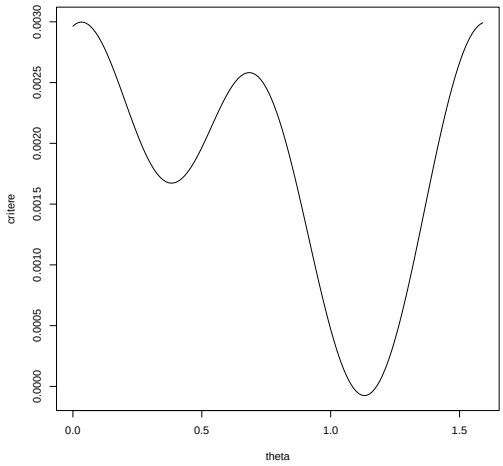


FIG. 6.2 – *Représentation du critère C avec un noyau gaussien et une taille de fenêtre de 1 en fonction du mélange*

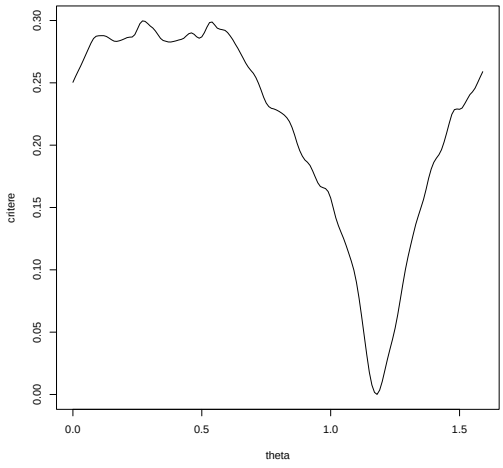


FIG. 6.3 – *Représentation du critère C avec un noyau spline de degré 3 et une taille de fenêtre de 0.1 en fonction du mélange*

Nous allons alors faire varier les paramètres a et b afin d'observer la forme du critère. Dans la suite, nous noterons $\mathbf{B}$ la matrice de séparation du système.

Points où le gradient s'annule

Minima globaux :

Dans les figures 6.4, 6.5, et 6.6, nous observons la présence de deux minima globaux qui correspondent à la solution du problème de séparation de sources. En effet, les points a et b solutions du problème de séparation de sources vérifient l'un des deux systèmes d'équation suivants, où λ représente le facteur d'échelle,

$$\begin{pmatrix} \cos(\pi/8) & \sin(\pi/8) \\ -\sin(\pi/8) & \cos(\pi/8) \end{pmatrix} \begin{pmatrix} 1 & a \\ b & 1 \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

ce qui correspond à $a = -\tan(\pi/8) = -0.41$ et $b = \tan(\pi/8) = 0.41$ (point noté P2)
ou bien

$$\begin{pmatrix} \cos(\pi/8) & \sin(\pi/8) \\ -\sin(\pi/8) & \cos(\pi/8) \end{pmatrix} \begin{pmatrix} 1 & a \\ b & 1 \end{pmatrix} = \begin{pmatrix} 0 & \lambda \\ \lambda & 0 \end{pmatrix}$$

ce qui correspond à $a = 1/\tan(\pi/8) = 2.41$ et $b = -1/\tan(\pi/8) = -2.41$ (point noté P4)
Ces deux points correspondent donc aux minima globaux de la mesure de dépendance et donc aux solutions du problème de séparation aveugle de sources.

Points selles :

Dans le chapitre 4, nous avons remarqué qu'il existe au moins un point où le gradient s'annule sans pour autant avoir atteint la solution. On peut les voir sur les figures 6.4, 6.5 et 6.6. Il correspond à $a = 0.66$ et $b = -0.66$. En effet, d'après le résultat de la section 4.4.8, les coordonnées du point P3 sont solutions du système suivant,

$$\begin{pmatrix} \cos(\pi/8) & \sin(\pi/8) \\ -\sin(\pi/8) & \cos(\pi/8) \end{pmatrix} \begin{pmatrix} 1 & a \\ b & 1 \end{pmatrix} = \begin{pmatrix} \lambda & 1 \\ -1 & \lambda \end{pmatrix}$$

où λ représente l'indétermination du facteur d'échelle.
On en déduit alors $a = (1-\sin(\pi/8))/\cos(\pi/8)$ et $b = -(1-\sin(\pi/8))/\cos(\pi/8)$
On remarque que ce point correspond à un point selle et donc ne perturbe pas la recherche du minimum par une méthode de descente du gradient.

Algorithme de descente du gradient

Dans la figure 6.4 nous retraçons l'évolution de l'algorithme de descente du gradient dans la minimisation de la mesure de dépendance quadratique. Notons, P0 de

coordonnées $a = 0$ et $b = 0$ (correspond à la matrice identité), P1 de coordonnées $a = -0.3$ et $b = 0.3$. Pour les points P0, P1, P2, nous représentons les distributions des sources reconstituées.

Point d'initialisation de la méthode d'optimisation

Quelque soit le critère utilisé, on remarque que dans le cas d'un mélange linéaire, les points où le déterminant de la matrice de séparation s'annule jouent un rôle particulier. En effet, le déterminant est nul lorsqu'une ligne de la matrice est proportionnelle à une autre. Ceci signifie qu'à la sortie de la structure de séparation, les deux variables calculées pour ces deux lignes sont nécessairement dépendantes puisque proportionnelles. Dans ce cas, la valeur du critère n'est pas minimale. C'est ce que nous pouvons observer sur les figures (6.4 et 6.5). Remarquons de plus, que lorsqu'on utilise l'information mutuelle décomposée à l'aide du mélange linéaire, cette dernière n'est pas définie en les points où le déterminant de la matrice est nulle. En effet, l'expression du critère dans ce cadre là est, $\sum H(Y_i) - \log|\det B|$.

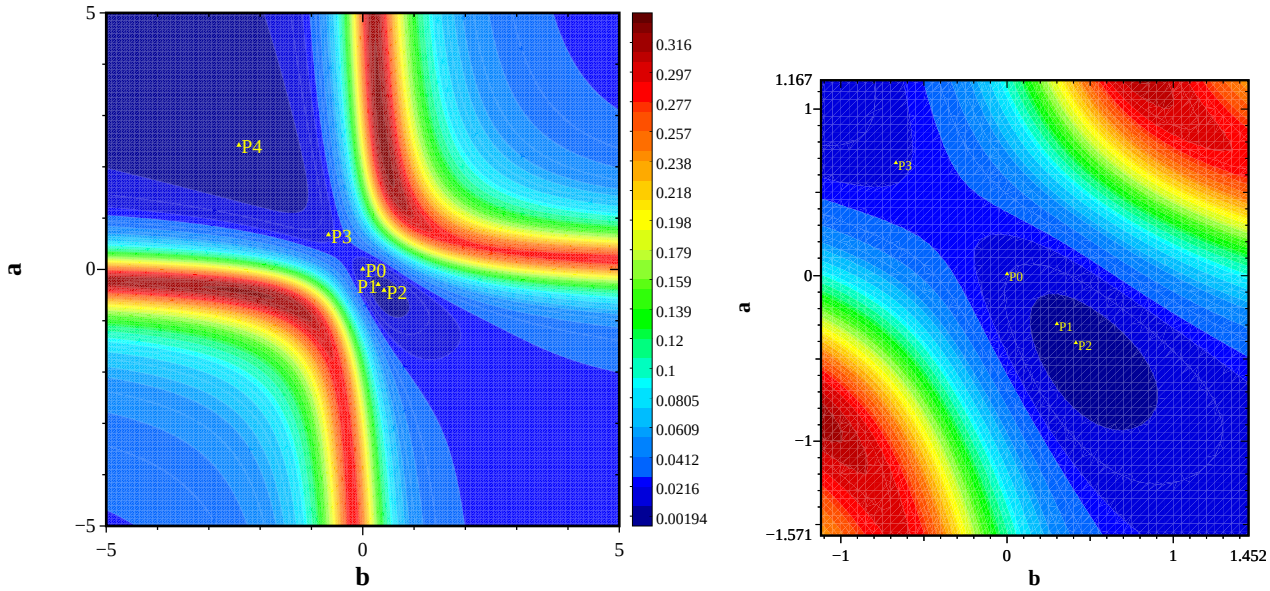
Dans les figures 6.4, 6.5, 6.6, le coefficient a de la matrice de séparation $\mathbf{B}$ est représenté par l'axe des ordonnées. Et le coefficient b par l'axe des abscisses.

Dans notre exemple, en dimension 2, nous pouvons remarquer que les figures 6.4, 6.5 et 6.6 présentent toutes la même allure générale. C'est-à-dire que le plan est découpé en trois zones distinctes, séparées par l'hyperbole d'équation $ab = 1$. Ce qui correspond bien dans notre exemple aux points où le déterminant de la matrice s'annule. D'autre part, cette hyperbole sépare la partie du plan où le déterminant de $\mathbf{B}$ sera positif (partie du plan séparé par les deux branches de l'hyperbole qui contient l'origine) des zones où le déterminant de $\mathbf{B}$ sera négatif (les deux parties restantes). Nous remarquons alors que la partie de plan où le déterminant de la matrice de séparation est positif est connexe.

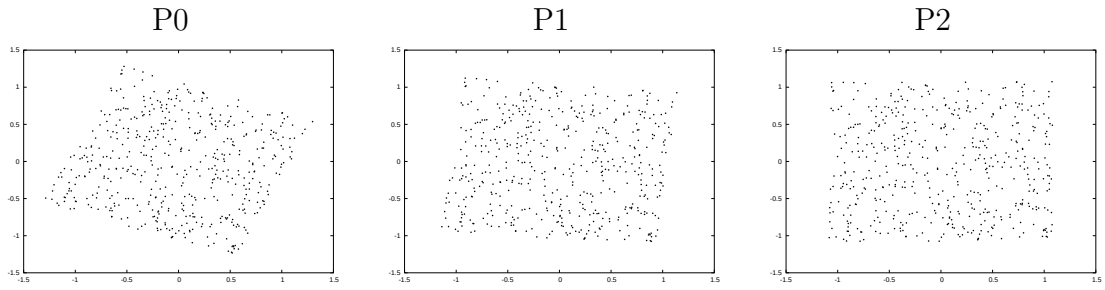
Par ailleurs, le critère de séparation est invariant par permutation de $\mathbf{B}$ et par multiplication de $\mathbf{B}$ par une matrice diagonale, le critère admettra donc toujours un minimum dans le domaine du plan où le déterminant de $\mathbf{B}$ est positif. C'est pourquoi, l'initialisation de l'algorithme de descente du gradient à la matrice identité peut-être choisie.

Plus généralement, montrons le résultat suivant, qui va nous permettre d'affirmer que nous pouvons choisir comme initialisation la matrice identité et ceci indépendamment de la matrice de mélange.

Lemme 6.1.1 *On note* $\mathbf{B}$ *une matrice solution du problème de séparation aveugle de source dans le cadre d'un mélange linéaire. Alors il existe une matrice* $\widetilde{\mathbf{B}}$ *telle que celle-ci est aussi solution du problème de séparation aveugle de source, son déterminant est positif et ses termes diagonaux sont égaux à 1.*



Distributions conjointes des points P0, P1, P2:



Rapports signal/bruit (en dB) des différents points,

P0	P1	P2	P3	P4
13.2	39.2	308.2	0.9	307.1

FIG. 6.4 – *Représentation de la mesure de dépendance quadratique en utilisant un noyau gaussien et une taille de fenêtre de 0.5 en fonction du mélange. (P3 est un point selle théorique)*

preuve :

En effet, si la matrice $\mathbf{B}$ a son déterminant positif, alors on pose $\widetilde{\mathbf{B}} = \Lambda\mathbf{B}$ où Λ est une matrice diagonale de telle sorte que les termes diagonaux de $\widetilde{\mathbf{B}}$ soient égaux à 1. Alors, $\widetilde{\mathbf{B}}$ vérifie les conditions requises par le lemme. Sinon, on pose $\widetilde{\mathbf{B}} = \Lambda\mathbf{P}\mathbf{B}$ où $\mathbf{P}$ est une matrice de permutation de telle sorte que en permutant deux lignes de la matrice, le déterminant change de signe, mais $\widetilde{\mathbf{B}}$ reste toujours solution du problème de séparation de sources.

■

Il nous reste alors à montrer qu'il est toujours possible de relier la matrice identité et la matrice solution tout en restant dans le sous ensemble des matrices de déterminant strictement positif (c'est-à-dire la connexité).

Soit $\mathbf{D}$ une matrice carrée de dimension $K \times K$ telle que ses termes diagonaux sont égaux à 1. On la notera sous la forme suivante:

$$\begin{pmatrix} 1 & d_1 & \dots & d_{K-1} \\ d_K & 1 & \dots & d_{2K-2} \\ \vdots & & \ddots & \\ & & \dots & 1 \end{pmatrix}$$

On note $\mathcal{C}$ le sous ensemble de ces matrices dont le déterminant est nul, $\mathcal{A}$ le sous ensemble de ces matrices de déterminant strictement positif et $(e_1, \dots, e_{K^2-K})$ les directions canoniques de $\mathbb{R}^{K^2-K}$. On note aussi $\mathbf{B}$ la matrice solution du problème de séparation aveugle de source que l'on définit par,

$$\mathbf{B} = \begin{pmatrix} 1 & b_1 & \dots & b_{K-1} \\ b_K & 1 & \dots & b_{2K-2} \\ \vdots & & \ddots & \\ & & \dots & 1 \end{pmatrix}.$$

On note de plus $\beta := \sum_{i=1}^{K^2-K} b_i e_i$, le point solution.

Remarquons tout d'abord la chose suivante,

Lemme 6.1.2 *Pour tout réel d_i, la fonction: $d_i \mapsto \det \mathbf{D}$ ne change de signe qu'une seule fois.*

En effet, cette fonction est un polynôme de degré 1 en d_i

A l'aide de ce résultat, nous démontrons alors,

Lemme 6.1.3 *Il existe un chemin contenu dans $\mathcal{A}$ qui relie n'importe quelle matrice $\mathbf{D}$ avec la matrice identité.*

preuve :
On raisonne par récurrence finie sur les e_i.

initialisation : Pour i=1
On constate que le long de l'axe e_1, le déterminant reste toujours positif. On peut donc se placer sur le point $x_1 = b_1 e_1$. On note alors $r = 2$.

pas de la récurence : Pour $2 \leq i \leq K^2 - K - 2$
On considère alors le plan $\mathcal{P}$ défini par les directions e_r et e_{i+1} passant par x_{i-1}. On note alors $\mathcal{E}$ la droite contenue dans le plan $\mathcal{P}$ passant par x_{i-1} et parallèle à l'axe e_r. On envisage alors deux cas disjoints,

(i) Le segment de droite $\mathcal{E}$ reliant les points x_{i-1} et $x_{i-1} + b_r e_r$ est entièrement contenu dans $\mathcal{A}$. Alors, on note $x_i = x_{i-1} + b_r e_r$. On passe au pas suivant en considérant les directions e_{i+1} et e_{i+2}, i.e. $r = i + 1$.

(ii) Le segment de la droite $\mathcal{E}$ coupe la courbe $\mathcal{C}$ en un point y. On considère alors la droite $\mathcal{D}$, contenue dans le plan $\mathcal{P}$ passant par y et parallèle à l'axe e_{i+1} (c.f. figure 6.7). Cette droite n'est pas tangente à $\mathcal{C}$ du fait du lemme 6.1.2. Et toujours par ce lemme, le segment de droite $\mathcal{D}$ reliant le point y au point $x_i = x_{i-1} + b_{i+1} e_{i+1}$ est entièrement contenu dans $\mathcal{A}$. On passe au pas suivant en considérant les directions e_r et e_{i+2}, i.e. r reste inchangé.

fin : Pour $i = K^2 - K - 1$, il reste alors deux directions à explorer, e_{K^2-K} et e_r, i.e. on a $x_{i-1} = \beta - b_r e_r - b_{i+1} e_{i+1}$. On envisage alors toujours les deux cas précédents,

- Le segment de droite $\mathcal{E}$ reliant les points x_{i-1} et $x_{i-1} + b_r e_r$ est entièrement contenu dans $\mathcal{A}$. Alors, on note $x_i = x_{i-1} + b_r e_r$. Et d'après le lemme 6.1.2, le segment de droite reliant x_i au point β solution, est entièrement contenu dans $\mathcal{A}$. En effet, si ce n'était pas le cas alors le long de cette droite, le signe du déterminant de la matrice changerait plus d'une fois.
- Le segment de droite $\mathcal{E}$ coupe la courbe $\mathcal{C}$ en un point y. On considère alors la droite $\mathcal{D}$, contenue dans le plan $\mathcal{P}$ passant par y et parallèle à l'axe e_{K^2-K}. Cette droite n'est pas tangente à $\mathcal{C}$ du fait du lemme 6.1.2. Et toujours par ce lemme, le segment de droite $\mathcal{D}$ reliant le point y au point $x_i = x_{i-1} + b_{i+1} e_{i+1}$ est entièrement contenu dans $\mathcal{A}$. Pour finir, le segment de droite reliant les points x_i et β est entièrement contenu dans $\mathcal{A}$ toujours par le lemme 6.1.2.

Nous avons alors montré que dans le cadre d'un mélange linéaire, si le critère utilisé ne présente pas de minima locaux, l'initialisation de l'algorithme de descente du gradient à la matrice identité permet d'atteindre la solution en un nombre fini de pas.

6.2 Avec un mélange post non linéaire

Dans le chapitre 4, nous avons étudié le problème de séparation aveugle de source dans le cadre d'un mélange post non linéaire. A présent, nous allons illustrer certains problèmes rencontrés dans la recherche de solution du problème de séparation aveugle de sources pour des mélanges post non linéaires.

Dans un premier temps, nous expliciterons dans quelle mesure nous pouvons affirmer que la recherche de la solution du problème de séparation aveugle de sources dans le cadre de mélange post non linéaire dépend de la matrice de mélange (paragraphe 6.2.1).

Dans un deuxième temps, nous donnerons des exemples par rapport à des distributions de sources différentes (paragraphe 6.2.2).

Enfin, nous terminerons, dans le paragraphe 6.2.3, par l'illustration de différents problèmes liés à la recherche du minimum, en particulier la présence de minima locaux.

Nous avons considéré ici des mélanges post non linéaires paramétrés, que l'on note,

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} f_{1,\lambda} \\ f_{2,\mu} \end{pmatrix} \circ \mathbf{A}_{\theta_0} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$$

où la matrice de mélange sera soit une matrice de rotation d'angle θ_0, soit une matrice symétrique de paramètre a_0 (avec les termes diagonaux égaux à 1, ou encore, une matrice antisymétrique de paramètre a_0 (avec les termes diagonaux égaux à 1). Quant aux non linéarités, on a choisi une seule sorte de non linéarité :

$$f_\lambda(x) = \frac{sign(x)}{2\lambda}(-1 + \sqrt{1 + 4\lambda|x|}).$$

Pour $\lambda > 0$, l'inverse de cette fonction est donc donné par,

$$f_\lambda^{-1}(x) = x + \lambda x|x|.$$

La structure de séparation sera donc paramétrée en fonction du mélange choisi.

On la notera,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \mathbf{B}_\theta \begin{pmatrix} g_{1,\lambda}(X_1) \\ g_{2,\mu}(X_2) \end{pmatrix}$$

où la matrice de séparation sera choisie de la même manière que la matrice de mélange :

– Si $\mathbf{A}_{\theta_0}$ est une matrice de rotation,

$$\mathbf{B}_\theta = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$$

– Si $\mathbf{A}_{a_0}$ est une matrice symétrique,

$$\mathbf{B}_a = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}$$

– Si $\mathbf{A}_{a_0}$ est une matrice antisymétrique,

$$\mathbf{B}_a = \begin{bmatrix} 1 & a \\ -a & 1 \end{bmatrix}$$

Et les non linéarités seront choisies en fonction des non linéarités de la structure de mélange.

Nous allons à présent étudier le comportement des différents critères de séparation lorsque les paramètres θ, λ et μ varient.

6.2.1 Influence de la matrice de mélange

Avec la mesure de dépendance quadratique

Nous avons vu dans les sections 5.3.1 et 5.3.2 que dans certains cas, en utilisant la mesure dépendance quadratique, l'hypothèse d'indépendance sera acceptée alors que les variables sont dépendantes. Nous allons voir ici comment doit-être conditionnée la matrice de mélange afin que la mesure de dépendance quadratique permettent effectivement de ne pas se tromper sur le fait que les variables sont dépendantes.

Matrice de rotation :

Prenons tout d'abord une matrice de rotation paramétrée par l'angle θ. On note

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$$

où S_1 et S_2 sont des variables indépendantes.
Pour $\theta = 0$, on fixe alors le paramètre $q_{0.05}$ de telle sorte $P_{H_0}(\widehat{Q} > q_{0.05}) = 0.05$. Puis on étudie alors $\alpha = P_{H_1}(\widehat{Q} < q_{0.05})$ en fonction de θ. La figure 6.8 représente la courbe obtenue.
On remarque alors clairement que pour des angles compris entre 0 et $\pi/2$, inférieurs à 0.2 ou supérieurs à $\pi/2 - 0.2$, la mesure de dépendance quadratique avec une taille d'échantillon de 500 ne permet pas de dire que les variables sont dépendantes avec un taux d'erreur acceptable. Par ailleurs, on remarque que si $\theta = 0$, la matrice de mélange est la matrice identité. Et, si $\theta = \pi/2$, la matrice de mélange est la matrice identité dont les lignes sont permutées. Donc, si l'angle de la matrice est compris entre 0 et $\pi/2$, inférieur à 0.2 ou supérieur à $\pi/2 - 0.2$, il y aura une forte probabilité pour que la mesure de dépendance

quadratique affirme que les variables mélangées sont indépendantes. Dans ce cas, on a déjà atteint la solution du problème de séparation aveugle de sources.

Matrice symétrique :

Envisageons à présent le cas d'un mélange de la forme suivante,
On note

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$$

Nous procédons alors au même tracé que précédemment, figure 6.9.
Nous remarquons ici aussi, que si a est inférieur à 0.1, ou supérieur à 8, la mesure de dépendance quadratique ne permet pas de savoir si les variables sont indépendantes[1]. Ici aussi, pour un type de mélange de cette sorte, il ne sera pas possible de retrouver les sources avec la mesure de dépendance quadratique.

Etude de l'erreur au voisinage de la solution

Dans le cadre de la résolution du problème de séparation de sources pour des mélanges post non linéaires, nous constatons qu'effectivement, la matrice de mélange a une importance dans l'étude de l'erreur au voisinage de la solution.

En effet, plaçons nous au voisinage de la solution, i.e. $f \circ g = i + \delta$ où δ est une petite fonction, et $\mathbf{B} = (I + \epsilon)\mathbf{A}^{-1}$. Alors, les sources reconstituées $Y_1, \ldots, Y_K$ vont s'écrire,

$$Y_k = S_k + (\mathbf{A}^{-1}\delta(\mathbf{AS}))_k + (\epsilon\mathbf{S})_k + (\epsilon(\mathbf{A}^{-1}\delta(\mathbf{AS})))_k.$$

L'erreur commise sur les sources reconstituées dépend donc directement de la matrice de mélange, mais non des transformations non linéaires. Afin d'illustrer ceci, on considère la mesure de dépendance quadratique estimée à l'aide d'un noyau gaussien, d'une taille de fenêtre de 0.5. Dans les figures 6.10 et 6.11, on a représenté la mesure de dépendance quadratique en fonction des paramètres de séparation.

6.2.2 Influence de la distribution des sources

On peut aussi s'intéresser à l'influence de la distribution des sources sur les comportements des critères de séparation. Dans les figures 6.12, 6.13 et 6.15, nous avons représenté la mesure de dépendance quadratique utilisant un noyau de dérivée de Cauchy carrée avec une taille de fenêtre de 3. Ici, nous avons considéré la matrice de rotation comme matrice de mélange.

1. En effet, quand $a \to +\infty, \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} \to \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ et donc Y_1 et Y_2 vont tendre vers des variables aléatoires indépendantes.

6.2.3 Résolution du problème de minimisation

Minima locaux

Nous avons remarqué que dans le cas d'un mélange linéaire, si il n'y a pas de minima locaux, la solution est toujours atteignable à partir de la matrice identité. Dans le cadre de mélanges post non linéaires, de nouvelles difficultés apparaissent. En effet, on observe qu'en fonction de l'estimateur utilisé, de la densité des sources et de la matrice de mélange, le problème de minimisation peut-être complètement différent. Illustrons tout ceci par quelques figures, 6.10 et 6.16.

Evolution des signaux reconstitués

Nous avons représenté dans les figures 6.13 et 6.14 les signaux issus de la structure de séparation en certains points du graphe.

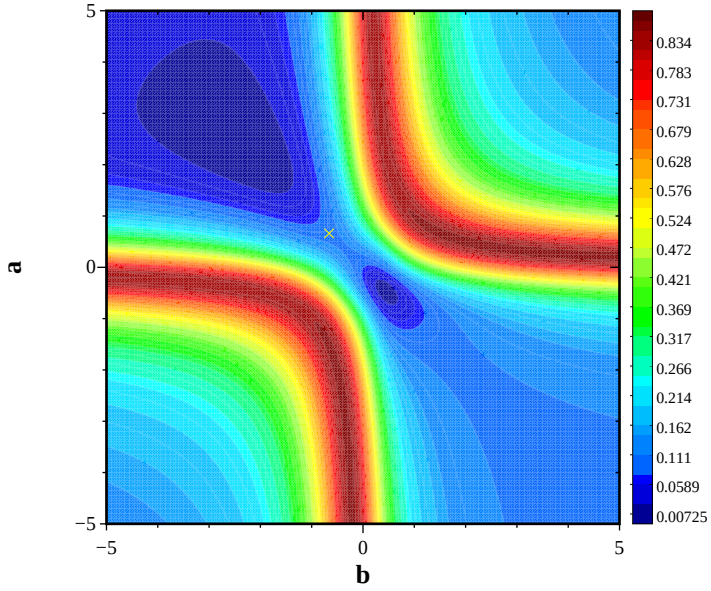


FIG. 6.5 – *Représentation de la mesure d'information mutuelle en utilisant un noyau gaussien et une taille de fenêtre de 0.5 en fonction du mélange (× est un point selle théorique)*

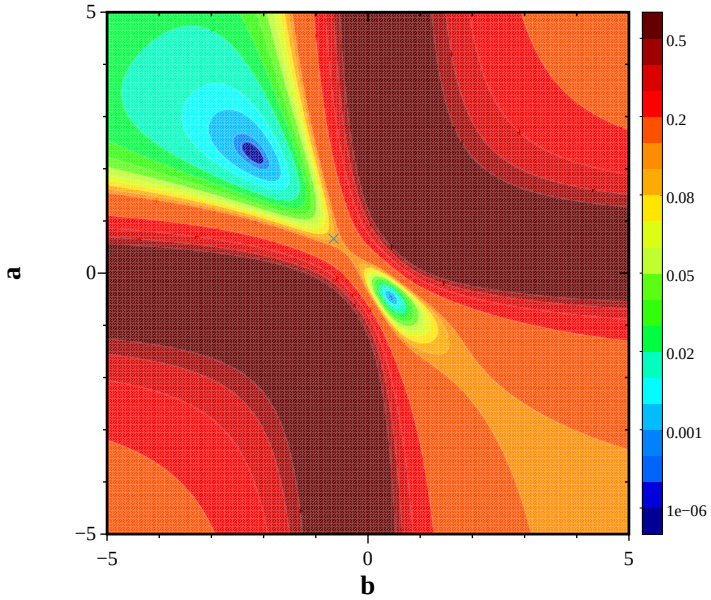


FIG. 6.6 – *Représentation du critère de séparation dérivé de la mesure d'information mutuelle décomposée pour le mélange linéaire en utilisant un noyau gaussien et une taille de fenêtre de 0.5 en fonction du mélange (× est un point selle théorique) (Les valeurs supérieures à 0.5 ont été tronquées, car ce critère tend vers l'infini pour* $\det \mathbf{B} \to 0$*)*

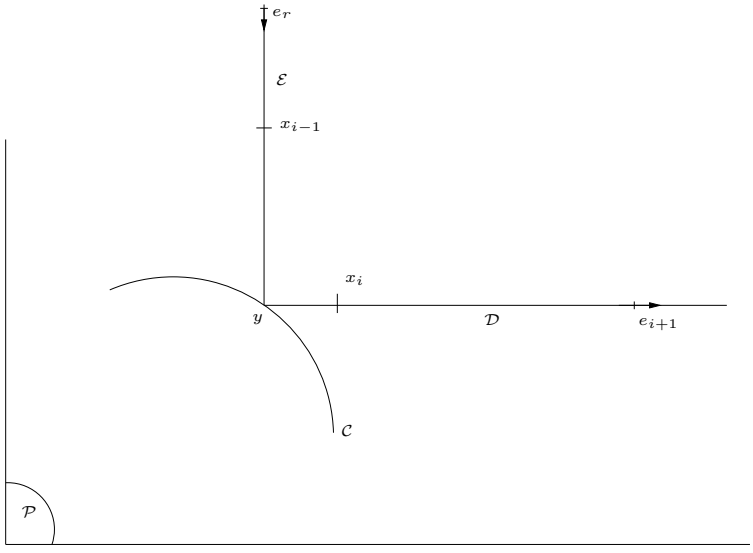


FIG. 6.7 – *Illustration du pas de la récurrence (ii)*

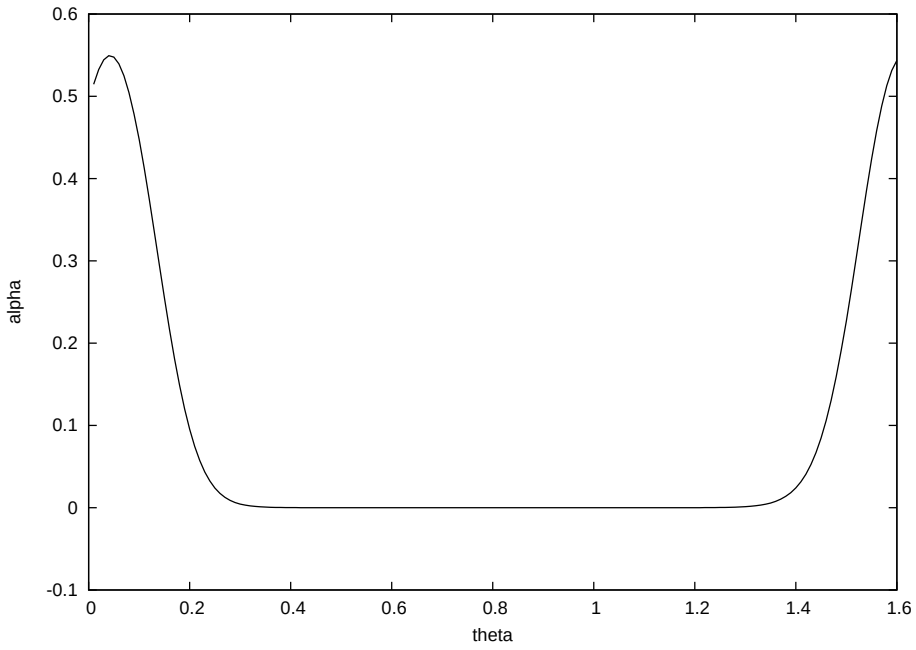


FIG. 6.8 – α *pour N=500,* $\alpha = P_{H_1}(\widehat{Q} < q_{0.05})$*, en fonction des paramètres du mélange,* θ *exprimé en radian*

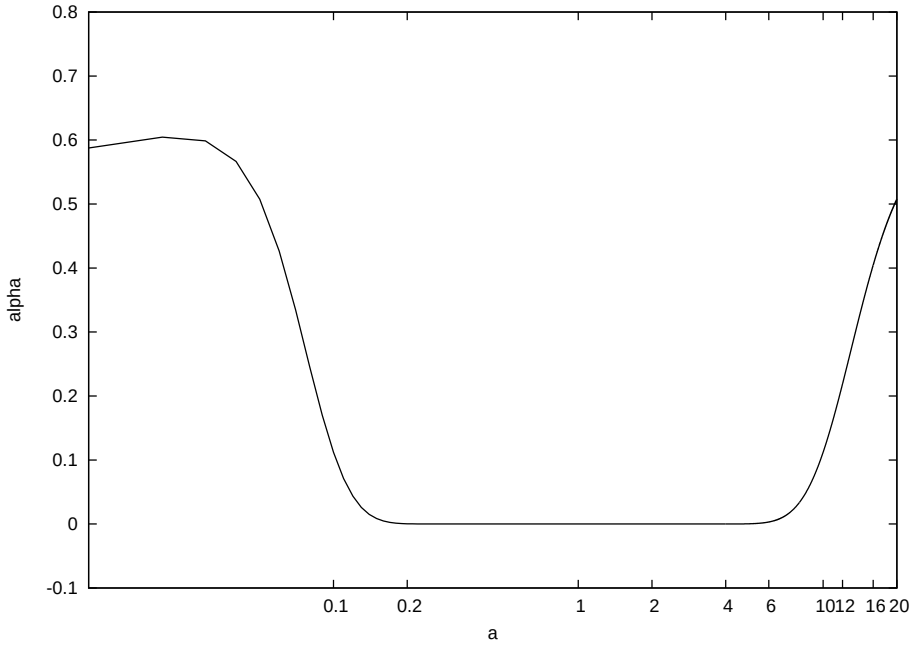


FIG. 6.9 – α *pour N=500,* $\alpha = P_{H_1}(\widehat{Q} < q_{0.05})$*, en fonction des paramètres du mélange*

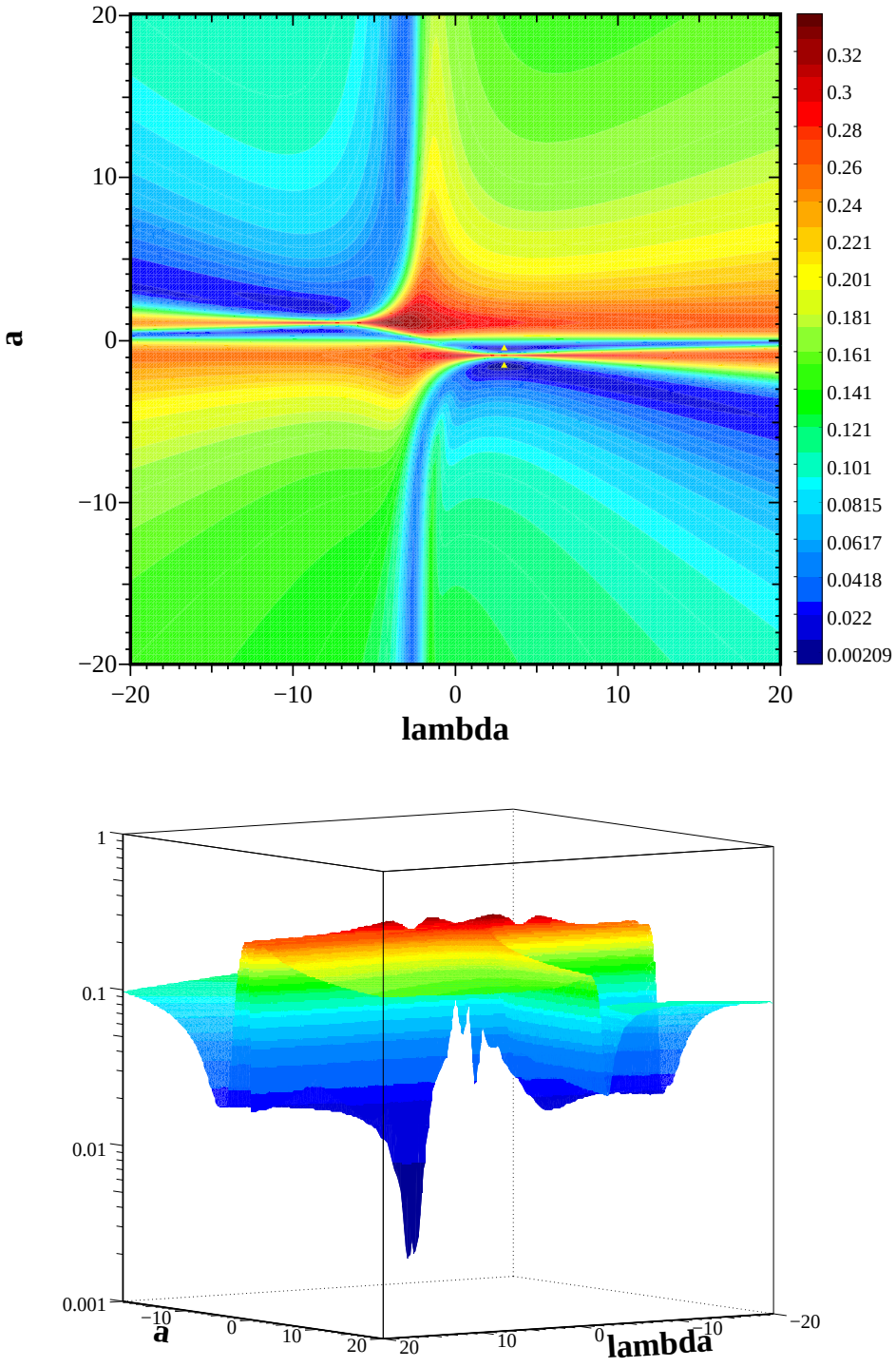


FIG. 6.10 – *Représentation de la mesure de dépendance quadratique lorsque la matrice de mélange est une matrice symétrique de paramètre 0.6 et f_1 est non linéaire de paramètre 3 et f_2 est linéaire avec deux sources uniformes, en fonction des paramètres de séparation.(▲ est un minimum global théorique)*

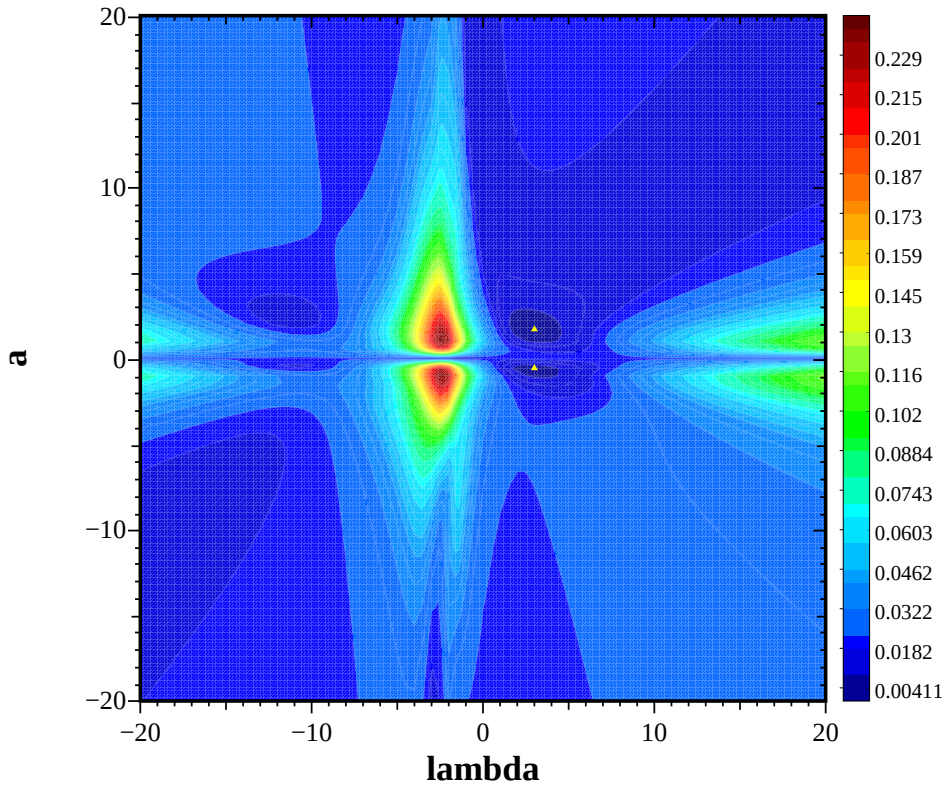


FIG. 6.11 – *Représentation de la mesure de dépendance quadratique lorsque la matrice de mélange est une matrice antisymétrique de paramètre 0.6 et f_1 est non linéaire de paramètre 3 et f_2 est linéaire avec deux sources uniformes, en fonction des paramètres de séparation. (▲ est un minimum global théorique)*

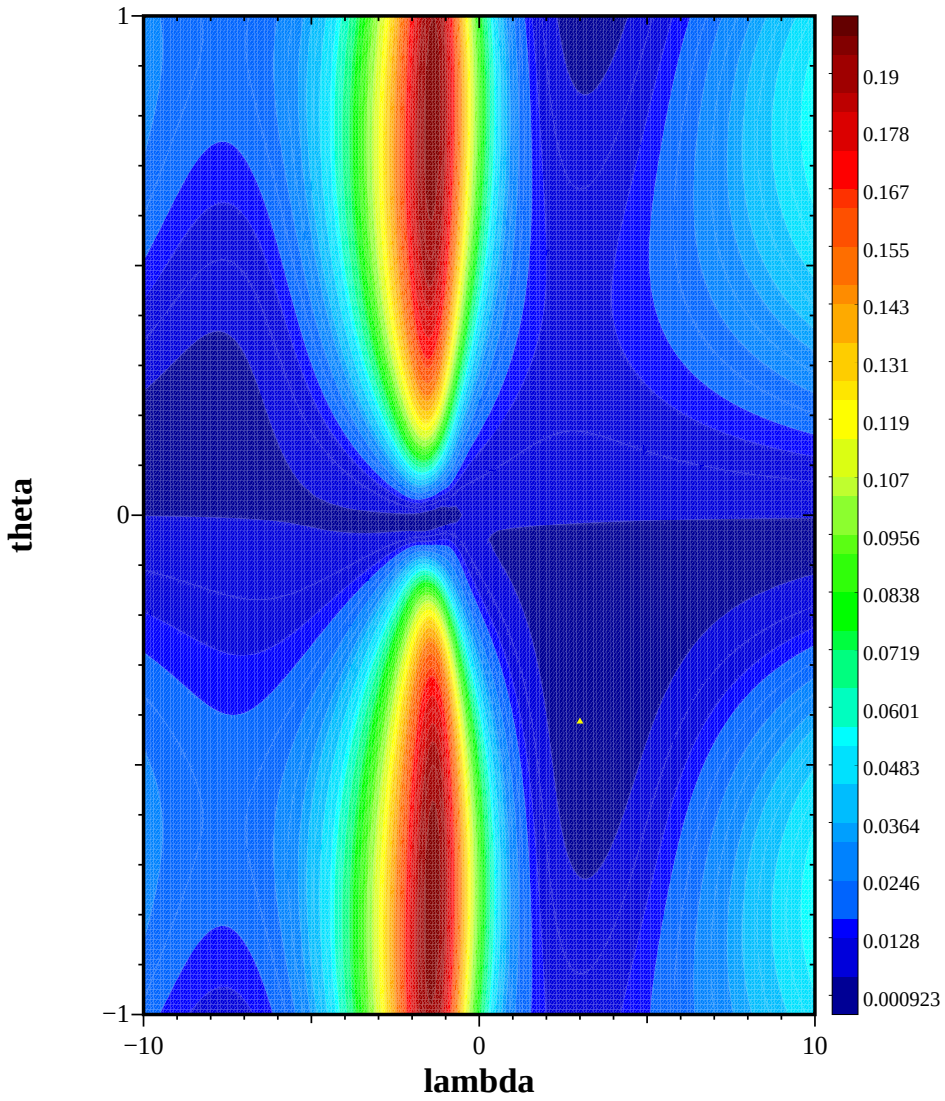


FIG. 6.12 – *Représentation de la mesure de dépendance quadratique lorsque la matrice de mélange est une matrice de rotation d'angle $\pi/8$ et f_1 est non linéaire de paramètre 3 et f_2 est linéaire, avec deux sources super gaussiennes (distribution de Laplace), en fonction des paramètres de séparation (▲ est un minimum global théorique).*

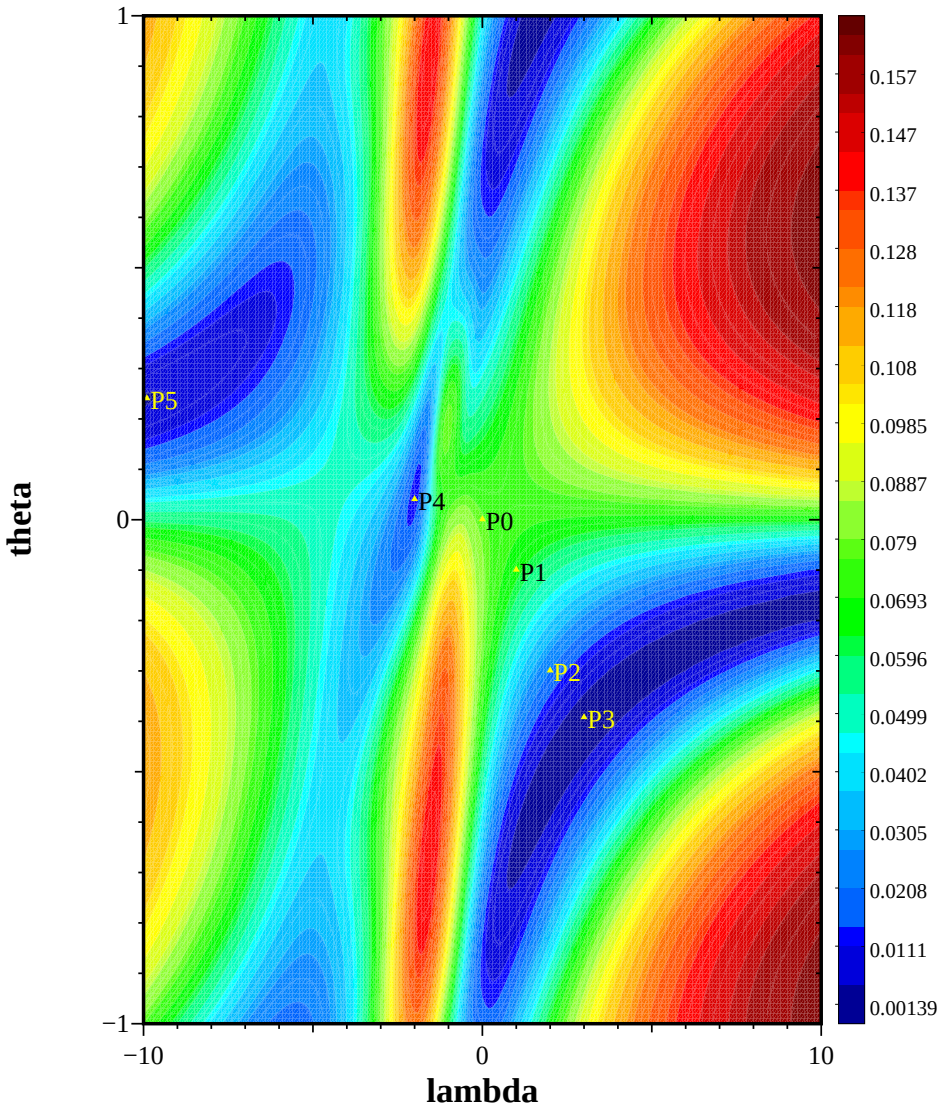
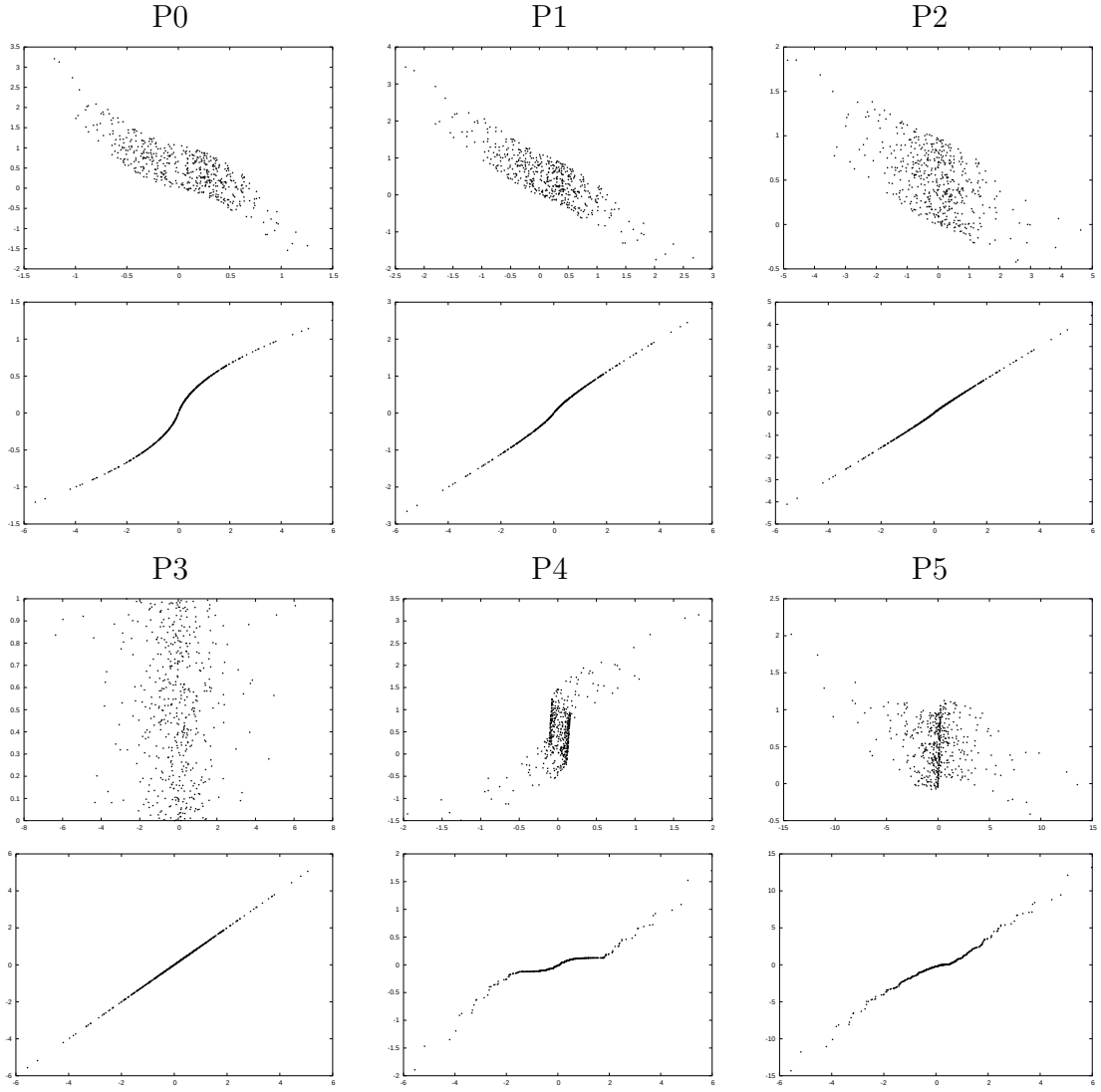


FIG. 6.13 – *Représentation de la mesure de dépendance quadratique lorsque la matrice de mélange est une matrice de rotation d'angle $\pi/8$ et f_1 est non linéaire de paramètre 3 et f_2 est linéaire, avec une source super gaussienne (distribution de Laplace) et une autre sous gaussienne (distribution uniforme), en fonction des paramètres de séparation (P3 est un minimum global théorique).*



Rapports signal/bruit (en dB) des différents points,

P0	P1	P2	P3	P4	P5
5.5	6.2	13.6	288.2	2.1	18.6

FIG. 6.14 – *Distributions conjointes des points P0, P1, P2, P3, P4 et P5 de la figure 6.13 ainsi que la composée des non linéarités,* $f_1 \circ g_1$.

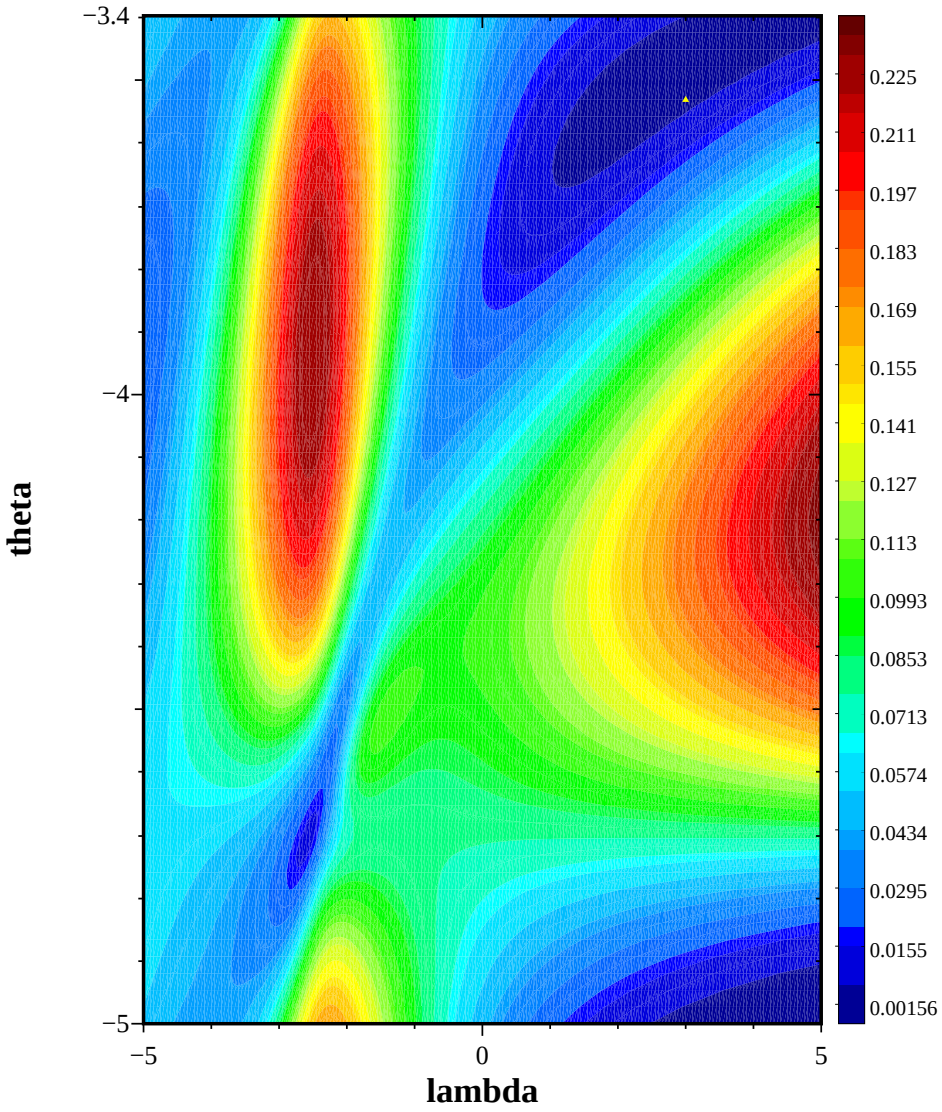


FIG. 6.15 – *Représentation de la mesure de dépendance quadratique lorsque la matrice de mélange est une matrice de rotation d'angle $\pi/8$ et f_1 est non linéaire de paramètre 3 et f_2 est linéaire, avec deux sources sous gaussiennes (distribution uniforme et sinus déterministe), en fonction des paramètres de séparation (▲ est un minimum global théorique).*

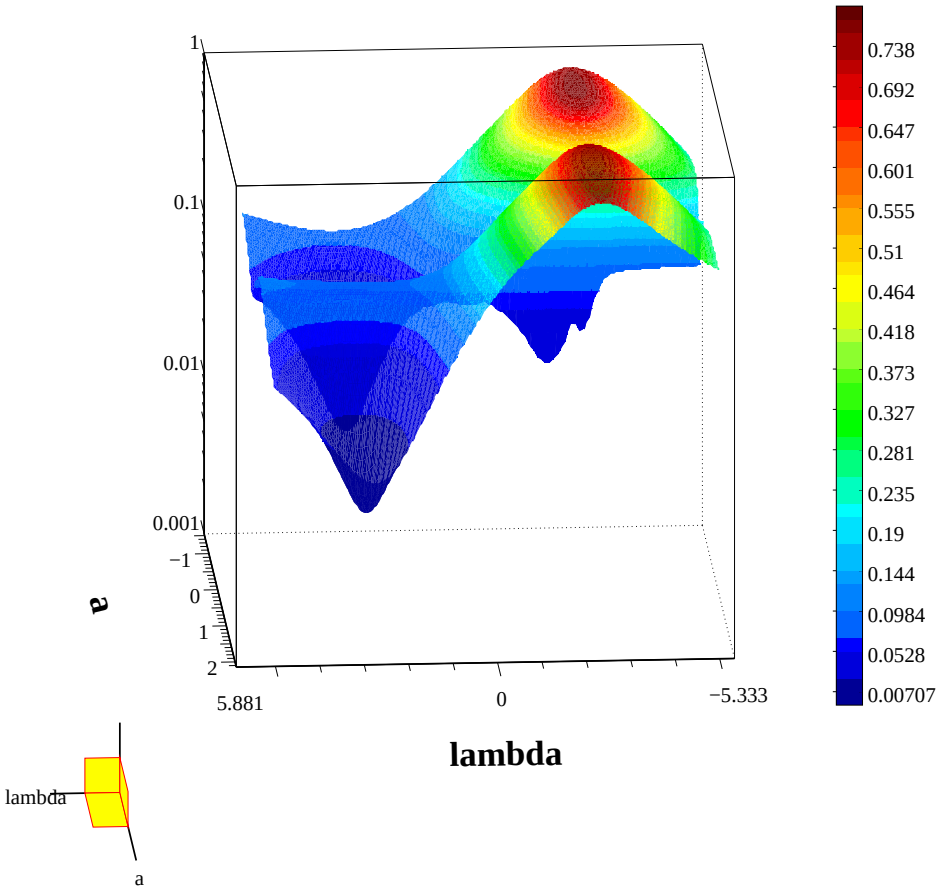


FIG. 6.16 – *Représentation de la mesure d'information mutuelle avec des sources uniformes, la matrice de mélanges est une matrice antisymétrique de paramètre 0.6, f_1 est non linéaire de paramètre 3 et f_2 est linéaire, en fonction des paramètres du mélange.*

Chapitre 7

Conclusion

Une approche de la séparation aveugle de sources

Bilan

Nous avons développé dans cette thèse deux méthodes pour la séparation aveugle de sources à l'aide d'une analyse en composantes indépendantes.

Tout d'abord, nous avons proposé une nouvelle mesure de dépendance, la mesure de dépendance quadratique. Cette mesure de dépendance se distingue de celle utilisée par Eriksson [30] par le fait qu'elle peut être calculée pour un choix plus large de noyaux. Grâce à une expression simple de l'estimateur de cette mesure de dépendance quadratique, on peut résoudre le problème de séparation aveugle de sources dans le cas de mélanges non linéaires.

Ensuite, nous nous sommes intéressés aux méthodes d'optimisation de ces mesures de dépendances. Nous avons proposé de procéder directement à la minimisation de l'estimateur de ces mesures (stratégie «estimer d'abord»). Ceci nous a permis de proposer des méthodes de contrôle de convergence des algorithmes de descente du gradient.

Puis, nous nous sommes focalisés sur la résolution du problème de séparation aveugle de sources dans le cadre de mélanges post non linéaires. Dans cette configuration, il a été développé une nouvelle approche non paramétrique utilisant les propriétés d'invariance par translation des mesures de dépendance. Ainsi, les mesures de dépendance étudiées ne s'expriment qu'en fonction des seules dérivées des non linéarités présentes dans le mélange post non linéaire.

Nous avons pu mener, sur la base des travaux de Kankainen [44], l'étude de la mesure de dépendance quadratique vue en tant que test d'indépendance. Nous avons alors

eu la possiblité d'envisager plusieurs types de noyaux et choisi une taille de fenêtre adéquate de telle sorte que le test d'indépendance construit à partir de la mesure de dépendance quadratique soit le plus puissant possible, mais aussi que cet estimateur ne présente pas une variance trop grande. Puis, à l'aide de cette étude asymptotique de la mesure de dépendance quadratique, nous avons pu donner des conditions sur la fonction de mélange pour que celle-ci permette de résoudre le problème de séparation de sources par une analyse en composantes indépendantes. Enfin, par rapport à l'information mutuelle, nous avons montré qu'il existe effectivement une différence entre le critère de séparation obtenu par Taleb et Jutten [69] et le critère de séparation en estimant la forme complète de l'information mutuelle. L'étude du biais de l'estimateur de l'information mutuelle nous a permis de montrer que celui-ci conserve asymptotiquement la propriété de caractérisation des variables indépendantes.

Enfin, une illustration graphique des résultats théoriques a été donnée. Nous avons pu voir, comment l'expression de la matrice de mélange ou la distribution des sources interviennent dans la représentation des mesures de dépendance en fonction des paramètres du modèle. Ont été signalées les difficultés engendrées par la présence de minima locaux dus aux erreurs d'estimation ou bien à la complexité du modèle.

Perspectives

Nous avons remarqué que dans le cadre d'un mélange linéaire, il existe d'autres méthodes de caractérisation de l'indépendance en comparant certaines espérances conditionnelles. Nous pouvons alors nous interroger sur la possiblité d'utiliser ces caractérisations dans le cadre de la séparation aveugle de sources pour les mélanges linéaires. De plus, beaucoup de travaux ont été effectués en probabilité sur des caractérisations de lois gaussiennes dans différents types de mélanges (polynomiaux ...). Il serait intéressant de les reprendre dans le cadre de la séparation aveugle de sources. Ceci pourrait éventuellement conduire à la proposition de nouveaux mélanges pour lesquels on pourrait envisager des méthodes pour restituer les sources.

Ensuite, d'un point de vue plus algorithmique, nous avons pu voir que la mesure de dépendance quadratique a un coût de calcul assez élevé. Il serait alors avantageux de choisir comme noyau, un noyau à support compact, tel qu'une fonction spline. De plus, grâce à la formule de récurrence vérifiée par les fonctions splines, le coût de calcul pourrait être encore réduit.

Par ailleurs, l'optimisation des mesures de dépendance étudiée présente des difficultés par la présence de minima locaux. Afin de réduire l'influence des minima locaux dus aux erreurs d'estimation, nous pouvons proposer une méthode qui consisterait à choisir une taille de fenêtre volontairement trop grande pour les premières itérations de l'algorithme. Puis quand on se rapproche du minimum, la taille de fenêtre serait

réduite afin d'affiner le résultat.

Enfin, nous avons pu voir que dans la résolution du problème de séparation de sources, seul le minimum du critère de séparation est recherché. Il serait alors intéressant de mener une étude asymptotique du minimum de l'estimateur du critère de séparation. En utilisant l'information mutuelle, nous avons utilisé la notion de fonction score. Le rôle prépondérant de l'estimation de celles-ci dans le cadre de la séparation aveugle de sources pour les mélanges post non linéaires a souvent été évoqué. Il serait alors intéressant de proposer une étude plus approfondie des estimateurs de ces fonctions.

Au delà de la séparation aveugle de sources

Ce travail peut aussi être vu dans un cadre plus large que la séparation de sources.

On peut par exemple penser à employer la méthode d'analyse en composantes indépendantes afin de déterminer la densité conjointe de variables aléatoires. Déjà l'estimation de densité d'une seule variable aléatoire est un problème complexe, et cette complexité augmente rapidement avec le nombre de variables. De la même façon qu'Antoniadis, Amato et Grégoire [5] ont proposé d'utiliser l'analyse en composantes principales pour préconditionner le problème, nous pouvons imaginer qu'une analyse en composantes indépendantes permette de faciliter l'estimation de densité conjointe de variables aléatoires.

Une autre perspective est d'envisager l'utilisation de la mesure de dépendance quadratique comme test d'indépendance. Ces tests permettent de mesurer l'influence d'une variable sur des observations. Par exemple on peut s'intéresser à l'étude de lien de causalité, comme déterminer les symptômes provenant d'une certaine maladie. Notre étude de la puissance de test d'indépendance construite à partir de la mesure de dépendance quadratique est un premier pas dans cette direction. Elle doit encore être affiner et mise en pratique pour plus de deux variables.

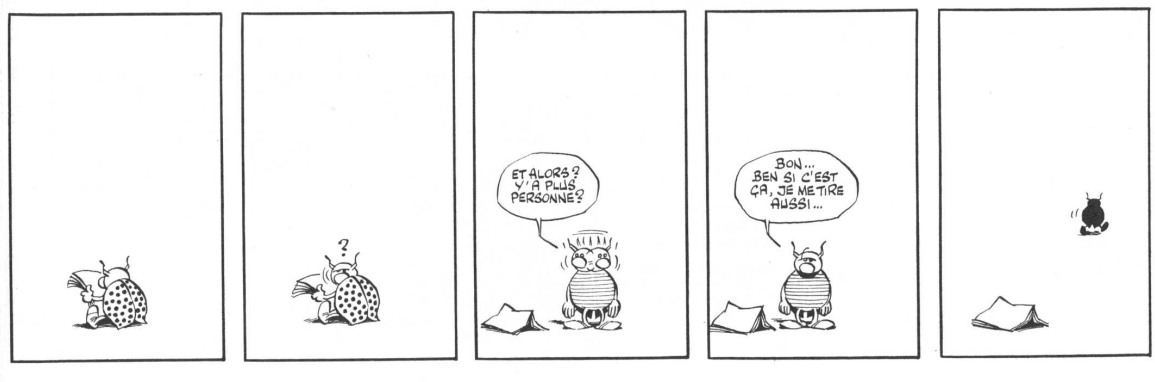


FIG. 7.1 – *Encore un qui ne lira pas les annexes !*

Annexe A

Caractérisation de l'indépendance dans les mélanges linéaires

Ces résultats sont basés sur les travaux de Kagan *et. al.* [43] et [42]

Lemme A.0.1 *Soient des variables aléatoires $X_1, \ldots, X_N$ indépendantes telles que $E(X_j^2) < \infty$.*

On pose $L_1 = a_1X_1 + \ldots + a_NX_N$ et $L_2 = b_1X_1 + \ldots + b_NX_N$.

Alors, si $E[L_2|L_1] = cte$ et $E[L_2^2|L_1] = cte$, on a pour tous $j = 1, \ldots, N$ tels que $a_jb_j \neq 0$, X_j est gaussienne.

Ce lemme permet alors avec le même raisonnement que pour montrer la séparabilité du mélange linéaire [20] de montrer le résultat suivant,

Lemme A.0.2 *Soient des variables aléatoires $X_1, \ldots, X_N$ indépendantes telles que $E(X_j^2) < \infty$ et telles qu'il y ait au plus une variable gaussienne. Soit $\mathbf{A}$ une matrice inversible ayant au moins deux éléments non nuls par colonne.*

Alors, on pose $L = \mathbf{A}X$.

Dans ce contexte, on a l'équivalence suivante,

$L_1, \ldots, L_N$ indépendantes si et seulement si, pour tout $i < j$, $E[L_j|L_i] = cte$ et $E[L_j^2|L_i] = cte$

Annexe B

Preuves

B.1 Lemme 3.2.1

Soit ϕ une fonction test, i.e. C^∞ à support compact. On note *supp* ϕ son support. Alors,

$$\begin{aligned}&\int \phi(\mathbf{t})(p_{\mathbf{T}+\mathbf{\Delta}_\eta(\mathbf{T})}(\mathbf{t}) - p_{\mathbf{T}}(\mathbf{t})d\mathbf{t})\\ &= \int (\phi(\mathbf{t}+\mathbf{\Delta}_\eta(\mathbf{t})) - \phi(\mathbf{t}))p_{\mathbf{T}}(\mathbf{t})d\mathbf{t}\end{aligned}$$

Comme ϕ est C^∞, on effectue un développement limité pour tout $\mathbf{t}$ au voisignage de $\mathbf{\Delta}_\eta(\mathbf{t})$. Il existe $\mathbf{x}_0(\mathbf{t}) \in$ *supp* ϕ tel que,

Alors,

$$\begin{aligned}&\int \phi(x)(p_{\mathbf{T}+\mathbf{\Delta}_\eta(\mathbf{T})}(\mathbf{t}) - p_{\mathbf{T}}(\mathbf{t})d\mathbf{t})\\ &= \int \left[\sum_{i=1}^{K} \partial_i(\phi(\mathbf{t}))\mathbf{\Delta}_\eta(\mathbf{t}) + \frac{1}{2}\sum_{i,j=1}^{K} \partial_i\partial_j(\phi(\mathbf{t}))\mathbf{\Delta}_\eta^2(\mathbf{t}) + \frac{1}{6}\sum_{i,j,k=1}^{K} \partial_i\partial_j\partial_k(\phi(\mathbf{x}_0(\mathbf{t})))\mathbf{\Delta}_\eta^3(\mathbf{t})\right] p_{\mathbf{T}}(\mathbf{t})d\mathbf{t}\end{aligned}$$

Comme ϕ est à support compact, on déduit,

$$\begin{aligned}&\int \phi(x)(p_{\mathbf{T}+\mathbf{\Delta}_\eta(\mathbf{T})}(\mathbf{t}) - p_{\mathbf{T}}(\mathbf{t})d\mathbf{t})\\ &= \sum_{i=1}^{K} \int \partial_i(\phi(\mathbf{t}))\mathbf{\Delta}_\eta(\mathbf{t})p_{\mathbf{T}}(\mathbf{t})d\mathbf{t} + \frac{1}{2}\sum_{i,j=1}^{K} \int \partial_i\partial_j(\phi(\mathbf{t}))\mathbf{\Delta}_\eta^2(\mathbf{t})p_{\mathbf{T}}(\mathbf{t})d\mathbf{t}\\ &\quad + \frac{1}{6}\sum_{i,j,k=1}^{K} \int_{supp\phi} \partial_i\partial_j\partial_k(\phi(\mathbf{x}_0(\mathbf{t})))\mathbf{\Delta}_\eta^3(\mathbf{t})p_{\mathbf{T}}(\mathbf{t})d\mathbf{t}\end{aligned}$$

Comme le support de ϕ est compact, la convergence simple de $\mathbf{\Delta}_\eta$ implique la convergence uniforme sur *supp* ϕ. Par conséquent,

$$\frac{1}{6}\sum_{i,j,k=1}^{K}\int_{supp\phi}\partial_i\partial_j\partial_k(\phi(\mathbf{x}_0)(\mathbf{t}))\mathbf{\Delta}_\eta^3(\mathbf{t})p_{\mathbf{T}}(\mathbf{t})d\mathbf{t} = o(\eta^2).$$

En ce qui concerne les autres termes, on montre par intégration par partie que

$$\int \partial_i(\phi(\mathbf{t}))\mathbf{\Delta}_\eta(\mathbf{t})p_{\mathbf{T}}(\mathbf{t})d\mathbf{t} = -\int \phi(\mathbf{t})\partial_i\{\mathbf{\Delta}_\eta(\mathbf{t})p_{\mathbf{T}}(\mathbf{t})\}d\mathbf{t}$$

et

$$\int \partial_i\partial_j(\phi(\mathbf{t}))\mathbf{\Delta}_\eta^2(\mathbf{t})p_{\mathbf{T}}(\mathbf{t})d\mathbf{t} = \int \phi(\mathbf{t})\partial_i\partial_j\{\mathbf{\Delta}_\eta^2(\mathbf{t})p_{\mathbf{T}}(\mathbf{t})\}d\mathbf{t}$$

■

B.2 Lemme 4.4.1

Supposons tout d'abord $i = j$,
alors, en utilisant le lemme 3.2.6,

$$\begin{aligned} E[Y_i\phi_i(\mathbf{Y})] &= E\{Y_iE[\phi_i(\mathbf{Y})|Y_i]\} \\ &= E\{Y_i\psi_i(Y_i)\} \end{aligned}$$

On en déduit alors une partie du résultat,
Pour tout $i = 1, \ldots, K$, $E[Y_i\phi_i(\mathbf{Y})] = 1$.

D'autre part, pour $i \neq j$,

$$E[Y_i\phi_j(\mathbf{Y})] = E\{Y_iE[\phi_j(\mathbf{Y})|Y_i]\}$$

De plus, on remarque que, $E[\phi_j(\mathbf{Y})|Y_i] = 0$.

En effet,

$$\begin{aligned}
E[\phi_j(\mathbf{Y})|Y_i = y] &= \frac{1}{p_{Y_i}}(y) \int \phi_j(\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) dy_1, \ldots, dy_{i-1}, dy_{i+1}, \ldots, dy_K \\
&= \frac{1}{p_{Y_i}}(y) \int \frac{\partial}{\partial_j} p_{\mathbf{Y}}(\mathbf{y}) dy_1, \ldots, dy_{i-1}, dy_{i+1}, \ldots, dy_K \\
&= \frac{1}{p_{Y_i}(y)} \int \int \frac{\partial}{\partial_j} p_{\mathbf{Y}}(\mathbf{y}) dy_1, \ldots, dy_{i-1}, dy_{i+1}, \ldots, dy_{j-1}, dy_{j+1}, \ldots, dy_K dy_j \\
&= \frac{1}{p_{Y_i}}(y) \int \frac{\partial}{\partial_j} \int p_{\mathbf{Y}}(\mathbf{y}) dy_1, \ldots, dy_{i-1}, dy_{i+1}, \ldots, dy_{j-1}, dy_{j+1}, \ldots, dy_K dy_j \\
&= \frac{1}{p_{Y_i}(y)} \int \frac{\partial}{\partial_j} p_{Y_i, Y_j}(y_i, y_j) dy_j \\
&= 0
\end{aligned}$$

Cette dernière égalité provient du fait que les limites de la densité en $\pm\infty$ vaut 0.

Montrons à présent la deuxième égalité du lemme,
pour tout $k = 1, \ldots, K$

$$\begin{aligned}
\sum_{i=1}^{K} E\left[\phi_i(\mathbf{Y})\mathbf{B}_{ik}|Z_k = z\right] &= E\left[\sum_{i=1}^{K} \phi_i(\mathbf{Y})\mathbf{B}_{ik}|Z_k = z\right] \\
&= \int \frac{\sum_{i=1}^{K} \frac{\partial}{\partial_i} p_{\mathbf{Y}}(\mathbf{B}z)\mathbf{B}_{ik}}{p_{Z_k}(z_k)} dz_1, \ldots, dz_{k-1}, dz_{k+1}, \ldots, dz_K \\
&= \int \frac{\frac{\partial}{\partial_{z_k}} p_{\mathbf{Y}}(\mathbf{B}z)\mathbf{B}_{ik}}{p_{Z_k}(z_k)} dz_1, \ldots, dz_{k-1}, dz_{k+1}, \ldots, dz_K \\
&= \int \frac{\partial}{\partial_{z_k}} \log p_{\mathbf{Y}}(\mathbf{B}z) \frac{p_{\mathbf{Y}}(\mathbf{B}z)}{p_{Z_k}(z_k)} dz_1, \ldots, dz_{k-1}, dz_{k+1}, \ldots, dz_K \\
&= \int \frac{\partial}{\partial_k} \log p_{\mathbf{Z}}(z) \frac{p_{\mathbf{Z}}(z)}{p_{Z_k}(z_k)} dz_1, \ldots, dz_{k-1}, dz_{k+1}, \ldots, dz_K
\end{aligned}$$

La dernière égalité provient de l'hypothèse de mélange post non linéaire, $\mathbf{Y} = \mathbf{BZ}$. Et alors,

$$p_{\mathbf{Y}}(\mathbf{B}z) = \frac{p_{\mathbf{Z}}(z)}{|\det \mathbf{B}|}$$

On obtient alors,

$$\begin{aligned}\sum_{i=1}^{K} E\left[\phi_i(\mathbf{Y})\mathbf{B}_{ik}|Z_k = z_k\right] &= E\left[\frac{\partial}{\partial_k}\log p_{\mathbf{Z}}(\mathbf{Z})\frac{p_{\mathbf{Z}}(\mathbf{Z})}{p_{Z_k}(z_k)}\right] \\ &= \psi_{Z_k}(z_k)\end{aligned}$$

■

B.3 Lemme 2.5.1

A partir de la définition 2.4.1, prenons $\mathcal{K}$ un noyau de carré intégrable.

On sait alors que la mesure de dépendance quadratique de K variables aléatoires $T_1, \ldots, T_K$ s'écrit,

$$Q(T_1, \ldots, T_K) = \int \left\{ E\left[\prod_{k=1}^{K} \mathcal{K}\left(t_k - \frac{T_k}{\sigma_{T_k}}\right)\right] - \prod_{k=1}^{K} E\left[\mathcal{K}\left(t_k - \frac{T_k}{\sigma_{T_k}}\right)\right] \right\}^2 dt_1 \ldots dt_K.$$

Par ailleurs, en développant l'expression au carré sous l'intégrale, on obtient, (dans la suite, on notera, $d\mathbf{u} = du_1 \ldots du_K$),

$$\begin{aligned} D_{\mathbf{T}}(t_1, \ldots, t_K)^2 = & \\ & \left\{ E\left[\prod_{k=1}^{K} \mathcal{K}\left(t_k - \frac{T_k}{\sigma_{T_k}}\right)\right] - \prod_{k=1}^{K} E\left[\mathcal{K}\left(t_k - \frac{T_k}{\sigma_{T_k}}\right)\right] \right\}^2 \\ = & \left\{ \int \prod_{k=1}^{K} \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) dF_{\mathbf{T}}(u_1, \ldots, u_K) - \prod_{k=1}^{K} \int \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) dF_{T_k}(u_k) \right\}^2 \\ = & \iint \prod_{k=1}^{K} \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) \mathcal{K}\left(t_k - \frac{v_k}{\sigma_{T_k}}\right) dF_{\mathbf{T}}(\mathbf{u}) dF_{\mathbf{T}}(\mathbf{v}) \\ + & \prod_{k=1}^{K} \iint \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) \mathcal{K}\left(t_k - \frac{v_k}{\sigma_{T_k}}\right) dF_{T_k}(u_k) dF_{T_k}(v_k) \\ - & 2 \int \prod_{k=1}^{K} \int \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) \mathcal{K}\left(t_k - \frac{v_k}{\sigma_{T_k}}\right) dF_{T_k}(v_k) dF_{\mathbf{T}}(\mathbf{u}) \end{aligned}$$

De plus, on vérifie les hypothèses permettant d'appliquer le théorème de Fubini qui permet d'intervertir les variables d'intégration,

$$\begin{aligned}
&\iiint \prod_{k=1}^{K} \left| \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) \mathcal{K}\left(t_k - \frac{v_k}{\sigma_{T_k}}\right) \right| dF_{\mathbf{T}}(\mathbf{u}) dF_{\mathbf{T}}(\mathbf{v}) d\mathbf{t} \\
&\leq \iint \int \prod_{k=1}^{K} \left| \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) \mathcal{K}\left(t_k - \frac{v_k}{\sigma_{T_k}}\right) \right| d\mathbf{t} dF_{\mathbf{T}}(\mathbf{u}) dF_{\mathbf{T}}(\mathbf{v}) \\
&\leq \iint dF_{\mathbf{T}}(\mathbf{u}) dF_{\mathbf{T}}(\mathbf{v}) \prod_{k=1}^{K} \int |\mathcal{K}|^2 < \infty
\end{aligned}$$

En appliquant la même remarque pour les autres termes du développement, on constate que,

pour tout $k = 1, \ldots, K$,

$$\int \iint \left| \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) \mathcal{K}\left(t_k - \frac{v_k}{\sigma_{T_k}}\right) \right| dF_{T_k}(u_k) dF_{T_k}(v_k) d\mathbf{t} \quad < \quad \infty$$

et

$$\int \int \prod_{k=1}^{K} \int \left| \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) \mathcal{K}\left(t_k - \frac{v_k}{\sigma_{T_k}}\right) \right| dF_{T_k}(v_k) dF_{\mathbf{T}}(\mathbf{u}) d\mathbf{t} \quad < \quad \infty$$

On peut alors intervertir les variables dans les intégrations et on obtient alors le résultat demandé.

$$\begin{aligned}
Q(T_1, \ldots, T_K) &= \int D_{\mathbf{T}}(t_1, \ldots, t_K)^2 dt_1 \ldots dt_K \\
&= \iint \prod_{k=1}^{K} \int \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) \mathcal{K}\left(t_k - \frac{v_k}{\sigma_{T_k}}\right) dt_k \;\; dF_{\mathbf{T}}(\mathbf{u}) dF_{\mathbf{T}}(\mathbf{v}) \\
&+ \prod_{k=1}^{K} \iint \int \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) \mathcal{K}\left(t_k - \frac{v_k}{\sigma_{T_k}}\right) dt_k \;\; dF_{T_k}(u_k) dF_{T_k}(v_k) \\
&- 2 \int \prod_{k=1}^{K} \int \int \mathcal{K}\left(t_k - \frac{u_k}{\sigma_{T_k}}\right) \mathcal{K}\left(t_k - \frac{v_k}{\sigma_{T_k}}\right) dt_k \;\; dF_{T_k}(v_k) dF_{\mathbf{T}}(\mathbf{u})
\end{aligned}$$

où on a posé,

$$\mathcal{K}_2(u) = \int \mathcal{K}(v)\mathcal{K}(u+v)dv. \tag{B.1}$$

Nous sommes alors en mesure de réécrire l'expression de la mesure de dépendance quadratique Q par,

$$\begin{aligned} Q(T_1,\ldots,T_K) &= \iint \prod_{k=1}^{K} \mathcal{K}_2\left(\frac{u_k - v_k}{\sigma_{T_k}}\right) dF_{\mathbf{T}}(\mathbf{u})dF_{\mathbf{T}}(\mathbf{v}) \\ &+ \prod_{k=1}^{K} \iint \mathcal{K}_2\left(\frac{u_k - v_k}{\sigma_{T_k}}\right) dF_{T_k}(u_k)dF_{T_k}(v_k) \\ &- 2\int \prod_{k=1}^{K} \int \mathcal{K}_2\left(\frac{u_k - v_k}{\sigma_{T_k}}\right) dF_{T_k}(v_k)dF_{\mathbf{T}}(\mathbf{u}) \end{aligned}$$

C'est à dire,

$$Q(T_1,\ldots,T_K) = E[\pi_{\mathbf{T}}(\mathbf{T})] + \prod_{k=1}^{K} E[\pi_{T_k}(T_k)] - 2E\left[\prod_{k=1}^{K} \pi_{T_k}(T_k)\right]$$

avec les notations définies dans le lemme 2.5.1.

Annexe C

Expresssions du Hessien

C.1 Expression du Hessien de l'information mutuelle

Approche non paramétrique:

Nous pouvons aussi en déduire du chapitre 3 l'écriture du Hessien. Nous nous limitons à écrire ce Hessien en un point où les variables sont indépendantes car dans le cas général, l'expression est trop compliquée pour être exploitable. Nous remarquons ici aussi que l'expression du Hessien du critère C utilisé par Taleb peut aussi être déduite de l'expression du Hessien de l'information mutuelle (paragraphe 4.4.8).

Le Hessien de l'information mutuelle s'écrit en un point où les variables aléatoires sont indépendantes,

$$\begin{aligned}
\varepsilon, \delta_1, \ldots, \delta_K \;\mapsto\; & \sum_{i=1}^{K} \sum_{j=1, j\neq i}^{K} [\varepsilon_{ij}^2 \mathrm{var}(Y_j) E(\psi'_{Y_i}(Y_i) + \varepsilon_{ij}\varepsilon_{ji}] \\
& + 2 \sum_{i=1}^{K} \sum_{k=1, k\neq i}^{K} \sum_{j=1}^{K} \varepsilon_{ik} E\{\mathbf{B}_{kj} \delta_j(Z_j) \psi_{Y_i}(Y_i) + \mathbf{B}_{ij} cov[\delta_j(Z_j), Y_k | Y_i] \psi'_{Y_i}(Y_i)\} \\
+ \; & \sum_{i=1}^{K} E[\delta'_i(Z_i)] + \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} \mathbf{B}_{ij}\mathbf{B}_{ik} E\{cov[\delta_j(Z_j), \delta_k(Z_k)|Y_i] \psi'_{Y_i}(Y_i) \\
& - E[\delta'_j(Z_j)|Y_i] E[\delta'_k(Z_k)|Y_i] \mathbf{B}_{ji}^{-1} \mathbf{B}_{ki}^{-1}\}
\end{aligned}$$

où $cov(X, Y|Z) = E(XY|Z) - E(X|Z)E(Y|Z)$ pour toutes variables aléatoires X, Y, Z.

Approche non paramétrique utilisant les dérivées des non linéarités:

Nous pouvons ici aussi envisager l'étude de méthodes d'optimisation d'ordre 2. Donnons ici l'expression du Hessien de ce critère en fonction de la matrice $\mathbf{B}$ et des dérivées $g'_1, \ldots, g'_K$. Nous donnons l'expression du Hessien seulement en un point où les variables sont indépendantes, à cause de la difficulté de faire les calculs dans le cas général et nous sommes alors assurés d'obtenir une forme quadratique non négative.

$$\begin{aligned}\varepsilon, \delta'_1, \ldots, \delta'_K \quad \mapsto \quad & \sum_{i=1}^{K} \sum_{j=1, j\neq i}^{K} [\varepsilon_{ij}^2 \mathrm{var}(Y_j) E(\psi'_{Y_i}(Y_i) + \varepsilon_{ij}\varepsilon_{ji}] \\ & + 2 \sum_{i=1}^{K} \sum_{k=1, k\neq i}^{K} \sum_{j=1}^{K} \varepsilon_{ik} \int H_{ik,j}(z) \delta'_j(z) dz \\ & + \sum_{i=1}^{K} \int p_{Z_i}(z) \delta'_i(z)^2 dz + \sum_{j=1}^{K} \sum_{k=1}^{K} \iint H_{jk}(z,u) \delta'_j(z) \delta'_k(u) dz du\end{aligned}$$

où

$$H_{ik,j}(z) = E\{\mathbf{B}_{kj} \mathbb{1}_+(Z_j - z) \psi_{Y_i}(Y_i) + \mathbf{B}_{ij} \mathrm{cov}[\mathbb{1}_+(Z_j - z), Y_k | Y_i] \psi'_{Y_i}(Y_i)\}$$

et

$$\begin{aligned}H_{jk}(z,u) = \sum_{i=1}^{K} \mathbf{B}_{ij}\mathbf{B}_{ik} \quad E\{ \quad & \mathrm{cov}[\mathbb{1}_+(Z_j - z), \mathbb{1}_+(Z_k - u) | Y_i] \psi'_{Y_i}(Y_i) \\ & - p_{Z_j|Y_i}(z|Y_i) p_{Z_k|Y_i}(u|Y_i) \mathbf{B}_{ji}^{-1} \mathbf{B}_{ki}^{-1}\end{aligned}$$

et $p_{Z_j|Y_i}(z|y)$ désigne la densité conditionnelle de Z_j au point z, sachant $Y_i = y$.

C.2 Expression du Hessien de la mesure de dépendance quadratique

Approche non paramétrique:

L'expression du Hessien en un point où les variables sont indépendantes s'écrit,

$$\begin{aligned}\varepsilon, \delta_1, \ldots, \delta_k \quad \mapsto \quad & \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{i=1}^{K}\sum_{i'=1}^{K} \varepsilon_{ji}\varepsilon_{ki'} E[H_{jk}(Y, Y')Y_i Y_{i'}] \\ + \quad & \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{i=1}^{K}\sum_{i'=1}^{K} \mathbf{B}_{ji}\mathbf{B}_{ki'} E[H_{jk}(Y, Y')\delta_i(Z_i)\delta_{i'}(Z_{i'})]\end{aligned}$$

En pratique, nous suggérons de ne garder que les termes diagonaux pour éviter des calculs trop compliqués.

Approche non paramétrique:

Afin d'implémenter une méthode d'optimisation d'ordre 2, nous nous intéressons à présent aux Hessien du critère dont nous conseillons d'utiliser seulement les termes diagonaux.

– Eléments diagonaux du Hessien par rapport à ε:

$$\varepsilon_{ij} \mapsto \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N} \widehat{H}^*_{ii}[\mathbf{Y}(n), \mathbf{Y}(m)]Y_j^2(n)\varepsilon_{ij}, \quad \text{pour tout} i, j = 1, \ldots, K$$

– Eléments diagonaux du Hessien par rapport à $d\theta_1, \ldots, d\theta_K$:

$$\begin{aligned}& d\theta_1, \ldots, d\theta_K \mapsto \\ & \quad \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N} \left\{ \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbf{B}_{ij}\mathbf{B}_{ik}\widehat{H}^*_{jk}[\mathbf{Y}(n), \mathbf{Y}(m)] \right\} \dot{g}_{i,\theta_i}[Z_i(n)]\dot{g}^T_{i,\theta_i}[Z_i(m)]d\theta_i\end{aligned}$$

Annexe D

Commentaires sur les estimateurs à noyaux

Nous reprenons ici des extraits du livre de Bosq et Lecoutre [10].

Le terme de noyau utilisé ici fait référence aux noyaux de convolution utilisé dans les définitions des estimateurs de la densité. Alors que le terme de noyau que nous avons utilisé dans la section 2.4 fait référence à la fonction intervenant dans la définition de la mesure de dépendance quadratique. Dans ce contexte, le noyau défini pour la mesure de dépendance quadratique ne vérifie pas les mêmes propriétés que celui utilisé dans l'estimation de la densité.

D.1 Définitions

Définition D.1.1 (Un noyau) *Un noyau $\mathcal{K}$ est une application de $\mathbb{R}^s$ dans $\mathbb{R}$, bornée, intégrable par rapport à la mesure de Lebesgue et d'intégrale égale à 1.*

Définition D.1.2 (Un noyau de Parzen-Rosenblatt) *On dit qu'un noyau est de Parzen-Rosenblatt si*

$$\lim_{||x||\to\infty} ||x||^s \mathcal{K}(x) = 0$$

Définition D.1.3 (Estimateur de la densité associé au noyau $\mathcal{K}$) *L'estimateur f_n associé au noyau $\mathcal{K}$ (ou la densité de probabilité empirique de noyau $\mathcal{K}$) et à l'échantillon $X_1, \ldots, X_n$ est défini par*

$$f_n(x) = \frac{1}{nh_n^s} \sum_{j=1}^{n} \mathcal{K}\left(\frac{x - X_j}{h_n}\right) = (\mathcal{K}_{h_n} * \mu_n)(x); \qquad x \in \mathbb{R}^s$$

où h_n est un nombre réel positif dépendant de n (window-width) et

$$\mathcal{K}_{h_n}(y) = \frac{1}{h_n^s}\mathcal{K}\left(\frac{y}{h_n}\right); \qquad y \in \mathbb{R}^s$$

Pour que l'estimateur d'une densité soit aussi une densité, on est souvent amené à considérer des noyaux positifs.

Remarque D.1.1 *On peut donner quelques exemples de noyau :*

- $\mathbb{1}_{[-1/2;1/2]}$ *est un noyau de Parzen-Rosenblatt positif.*
- $\exp(-x^2/2)/\sqrt{(}2\pi)$ *est aussi un noyau de Parzen-Rosenblatt positif défini sur* $\mathbb{R}$.

D.2 Lemmes

Le lemme de Bochner traduit une propriété asymptotique qui exprime le fait que, lorsqu h est petit, la convolution avec $\mathcal{K}_h$ perturbe peu une fonction de L^1. Il s'énonce de la manière suivante :

Lemme D.2.1 *1. Soit $\mathcal{K}$ un noyau de Parzen-Rosenblatt et $g \in L^1$. Alors en tout point x où g est continue,*

$$\lim_{h\to 0}(g * K_h)(x) = g(x)$$

2. Soit maintenant $\mathcal{K}$ un noyau quelconque; si $g \in L^1$ est uniformément continue, alors

$$\lim_{h\to 0} d(g * K_h, g) = 0$$

Index

Bibliographie

[1] S. Achard. Initiation à la séparation aveugle de sources dans les mélanges post non linéaires. Master's thesis, UJF, Grenoble, 2000.

[2] S. Achard, D. T. Pham, and C. Jutten. Séparation aveugle de source dans les mélanges post non linéaires. In *Proc. Gretsi 2001*, Sep. 2001.

[3] S. Achard, D.T. Pham, and C. Jutten. Quadratic dependence measure for non linear blind souces separation. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2003*, pages 263–268, Nara, Japan, Apr. 2003.

[4] L. Almeida. Linear and nonlinear ICA based on mutual information. In *Proc. IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pages 117–122, Lake Louise, Canada, Oct. 2000.

[5] A. Antoniadis, U. Amato, and G. Grégoire. Independent component discriminant analysis. *International Mathematical Journal*, to appear.

[6] M. Babaie-Zadeh. *On blind source separation in convolutive and nonlinear mixtures*. PhD thesis, I.N.P.G. - Laboratoire L.I.S., 2002.

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

[8] A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[9] A. Belouchrani, K. Abed Meraim, J. F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

[10] D. Bosq and J.P. Lecoutre. *Théorie de l'estimation fonctionnelle*. Paris : Economica, 1987.

[11] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of American Statistical Association*, 80(391):580–598, sept. 1985.

[12] J.F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE letters on Signal Processing*, 4(4):112–114, Apr. 1997.

[13] J.F. Cardoso. Blind signal separation : Statistical principles. *Proceedings IEEE*, 86(10):2009–2025, Oct. 1998.

[14] J.F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11:157–192, 1999.

[15] J.F. Cardoso. The three easy routes to independent component analysis; contrasts and geometry. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2001*, pages 1–6, San Diego, California, Dec 2001.

[16] J.F. Cardoso. Independent component analysis of the cosmic microwave background. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2003*, Nara, Japan, April 2003.

[17] J.F. Cardoso and B.H. Laheld. Equivariant adaptative source separation. *IEEE Transactions on Signa Processing*, 44(12):3017–3029, Dec. 1996.

[18] J.F. Cardoso and D.T. Pham. Séparation de sources par l'indépendance et la parcimonie. In *19e Colloque GRETSI sur le traitement du signal et des images*, pages 109–112, Paris, France, Sept. 2003.

[19] P. Comon. Independent Component Analysis. In *Proc. Int. Sig. Proc. Workshop on Higher-Order Statistics*, Chamrousse, France, July 1991. Republished in *Higher-Order Statistics*,J.L.Lacoume ed., Elsevier, 1992, pp 29–38.

[20] P. Comon. Independent component analysis, a new concept ? *Signal Processing*, 3(36):287–314, Apr. 1994.

[21] P. Comon and O. Grellier. Non linear inversion of underdetermined mixtures. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA'99*, pages 461–465, Aussois, France, Jan. 1999.

[22] P. Comon and J. Lebrun. Critères de contrastes déterministes pour la séparation de sources. In *Gretsi 2003*, pages 121–124, Paris, France, Sept. 2003.

[23] S. Csörgő. Limit behaviour of the empirical characteristic function. *The Annals of Probability*, 9(1):130–144, 1981.

[24] S. Csörgő. Testing for independence by the empirical characteristic function. *Journal of Multivariate Analysis*, 16:290–299, 1985.

[25] R. Cuppens. *Decomposition of multivariate probability.* New York: Academic press, 1975.

[26] G. Darmois. Analyse des liaisons de probabilités. In *Proceedings Int. Stat. Conferences 1947*, volume III A, page 231, Washington (D.C.), 1951.

[27] G. Darmois. Analyse générale des liaisons stochastiques. *Rev. Inst. Internat. Stat.*, 21:2–8, 1953.

[28] N. Delfosse and Ph. Loubaton. Adaptative blind separation of independent sources: A deflation approach. *Signal Processing*, 45:59–83, 1995.

[29] J. Eriksson, A. Kankainen, and V. Koivunen. Novel characteristic function based criteria for ICA. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2001*, pages 108–113, San Diego, California, Dec. 2001.

[30] J. Eriksson and V. Koivunen. Characteristic function based independent component analysis. *Elsevier Science*.

[31] A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61(3):419–433, 1993.

[32] P. Garat. *Approche statistique pour la séparation aveugle de sources*. PhD thesis, Université Joseph Fourier (Grenoble I), 1994.

[33] M. Gaéta and J.L. Lacoume. Source separation without a priori knowledge: the maximum likelihood solution. In *Proc. EUSIPCO 90*, pages 621–624, Barcelona, Spain, 1990.

[34] P. Hall and S. C. Morton. On the estimation of entropy. *Ann. Inst. Statist. Math.*, 45(1):69–88, 1993.

[35] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.*, 19:293–325, 1948.

[36] W. Hoeffding. A non-parametric test of independence. *Ann. Math. Stat.*, 19:546–557, 1948.

[37] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. New York: John Wiley & Sons, 2001.

[38] H. Joe. Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.*, 41(4):683–697, 1989.

[39] C. Jutten and R. Gribonval. L'analyse en composantes indépendantes : un outil puissant pour le traitement de l'information. In *Gretsi 2003*, pages 11–14, Paris, France, Sept. 2003.

[40] C. Jutten and A. Taleb. Source separation from dusk till dawn. In *Independent Component Analysis*, pages 15–26, Helsinki, June 2000.

[41] A. Kagan, R.C. Laha, and V. Rohatgi. Independence of the sum and absolute difference of independent random variables does not imply their normality. *Mathematical Methods of Statistics*, 6(2):263–265, 1997.

[42] A.M. Kagan. New classes of dependent random variables and a generalization of the darmois-skitovich theorem to several forms. *Theory Probab. Appl.*, 33(2):286–295, 1988.

[43] A.M. Kagan, YU.V. Linnik, and C.R. Rao. *Characterization problems in mathematical statistics*. John Wiley & Sons, 1973.

[44] A. Kankainen. *Consistent testing of total independence based on empirical characteristic functions*. PhD thesis, Unversity of Jyväskylä, 1995.

[45] M. Krob. *Identification aveugle de modèles non linéaires à l'aide des statistiques d'ordre supérieur*. PhD thesis, Université Paris XI Orsay, 1994.

[46] S. Kullback. *Information theory and statistics*. John Wiley & Sons, 1959.

[47] A. Larue. Séparation de sources markoviennes. Master's thesis, INPG Grenoble, 2003.

[48] L. De Lathauwer, D. Callaerts, and B. De Moor et. al. Fetal electrocardiogram extraction by source subspace separation. In *Proc. Int. Workshop on HOS*, pages 134–138, Girona, Spain, June 1995.

[49] A.J. Lee. *U-statistics*. statistics, 1990.

[50] E. Lukacs. *Characteristic functions*. London: Griffin, (2nd edition) edition, 1970.

[51] M. Metivier and J. Neveu. *Cours de probabilités*. Ecole Polytechnique, 1983.

[52] D. Nuzillard. Comment la chimie analytique peut contribuer à l'étude des images astronomiques. In *19e Colloque GRETSI sur le traitement du signal et des images*, pages 23–26, Paris, France, Sept. 2003.

[53] D. Nuzillard and A. Bijaoui. Blind source separation and analysis of multispectral astronomical images. *Astron. Astrophys. Suppl.*, Ser. 147:129–138, August 2000.

[54] D. Obradovic and G. Deco. Information maximization and independent component analysis: Is there a difference? *Neural Computation*, 10:2085–2101, 1998.

[55] P. Pajunen, A. Hyvarinen, and J. Karhunen. Nonlinear blind source separation by self-organizing maps. In *Proc. Int. Conf. on Neural Information Processing*, pages 1207–1210, Hong Kong, 1996.

[56] P. Pajunen and J. Karhunen. A maximum likelihood approach to nonlinear blind source separation. In *Proc. of the 1997 Int. Conf. on Artificial Neural Networks, ICANN'97*, pages 541–546, Lausanne, Switzerland, 1997.

[57] D.-T. Pham. Blind separation of instantaneous mixture of sources via an independent component analysis. *IEEE Transactions on Signal Processing*, 44(11):2768–2779, Nov. 1996.

[58] D.-T. Pham. Contrast functions for blind separation and deconvolution of the sources. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2001*, pages 37–42, San Diego, California, Dec. 2001.

[59] D.-T. Pham. Flexible parametrisation of postnonlinear mixture model in blind source separation. *IEEE Signal Processing Letters*, to appear.

[60] D.-T. Pham and P. Garat. Blind separation of mixture of independent sources through a quasi-maximum lixelihood approach. *IEEE Transactions on Signal Processing*, 45(7):1712–1725, Jul. 1997.

[61] D.T. Pham. Fast algorithm for estimating mutual information, entropies and score functions. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2003*, Nara, Japan, Apr. 2003.

[62] A. Rényi. *Calcul des probabilités avec un appendice sur la théorie de l'information*. Editions Jaques Gabay, 1966.

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

[64] R. J. Serfling. *Approximation theorems of mathematical statistics*. New York: John Wiley & Sons, 1980.

[65] J. Solé, C. Jutten, and D.T. Pham. Improving algorithm speed in PNL mixture separation and wiener system inversion. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2003*, pages 639–644, Nara, Japan, Apr. 2003.

[66] D. Tjøstheim. Measures of dependance and tests of independence. *Statistics*, 28:249–284, 1996.

[67] A. Taleb. *Séparation de Sources dans les Mélanges Non Linéaires*. PhD thesis, I.N.P.G. - Laboratoire L.I.S., 1999.

[68] A. Taleb and C. Jutten. Batch algorithm for source separation in postnonlinear mixtures. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA1999*, pages 279–284, Aussois, France, Jan. 1999.

[69] A. Taleb and C. Jutten. Sources separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 10(47):2807–2820, Oct. 1999.

[70] F. J. Theis, C. Puntonet, and E. W. Lang. Nonlinear geometric ICA. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2003*, pages 275–280, Nara, Japan, April 2003.

[71] H.-L. Nguyen Thi and C. Jutten. Blind source separation for convolutive mixtures. *Signal Processing*, 45:209–229, 1995.

[72] L. Tong, R.-W. Liu, V. C. Soon, and Y.-F. Huang. Indeterminacy and indentifiability of blind identification. *IEEE Transactions on circuits systems*, 38(5):499–509, May 1991.

[73] H. Valpola. *Bayesian Ensemble Learning for Nonlinear Factor Analysis*. PhD thesis, Helsinki University of Technology, 2000.

[74] R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component analysis approach to the analysis of EEG and MEG recordings. *IEEE transactions biomedical engineering*, 47(5):589–593, 2000.

[75] H. H. Yang, S.-I. Amari, and A. Cichoki. Information-theoretic approach to blind sepration of sources in non-linear mixture. *Signal Processing*, 64(3):291–300, 1998.

[76] D. Yellin and E. Weinstein. Criteria for multichannel signal separation. *IEEE Transactions on Signal Processing*, 42:2158–2168, August 1994.

[77] A. Zaidi. *Séparation aveugle d'un mélange instantané de sources autorégressives gaussiennes par la méthode du maximum de vraisemblance exact*. PhD thesis, L.M.C.-I.M.A.G., 2000.

[78] V. Zarsozo and A.-K. Nandi. Noninvasive fetal electrogardiogram extraction : blind source separation versus adaptative noise cancellation. *IEEE Transactions on biomedical engineering*, 48(1):12–18, Jan. 2001.

[79] A. Ziehe, M. Kawanabe, S. Harmeling, and K.R. Müller. Separation of post-nonlinear mixtures using ace and temporal decorrelation. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2001*, pages 433–438, San Diego, California, Dec. 2001.