Entropic measures of connectivity with an application to intracerebral epileptic signals

Jie Zhu

ANNÉE 2016

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Traitement du Signal et Télécommunications

Ecole doctorale MATISSE
Ecole Doctorale Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Electronique

Jie ZHU

Préparée à l'unité de recherche LTSI - INSERM UMR 1099
Laboratoire Traitement du Signal et de l'Image
ISTIC UFR Informatique et Électronique

Entropic Measures of Connectivity with an Application to Intracerebral Epileptic Signals

Thèse soutenue à Rennes
le 22 juin 2016

devant le jury composé de :

MICHEL Olivier
Professeur des Universités
Grenoble INP / *rapporteur*

WOLF Didier
Professeur des Universités
Université de Lorraine / *rapporteur*

BOUDAOUD Sofiane
Maître de Conférences
Université de Technologie de Compiègne / *examinateur*

PASTOR Dominique
Professeur
Telecom Bretagne / *président du jury*

SHU Huazhong
Professeur des Universités
Université du Sud-Est à Nankin / *examinateur*

LE BOUQUIN JEANNES Régine
Professeur des Universités
Université de Rennes 1 / *directeur de thèse*

Résumé

Les travaux présentés dans cette thèse s'inscrivent dans la problématique de la connectivité cérébrale, connectivité tripartite puisqu'elle sous-tend les notions de connectivité structurelle, fonctionnelle et effective. Ces trois types de connectivité que l'on peut considérer à différentes échelles d'espace et de temps sont bien évidemment liés et leur analyse conjointe permet de mieux comprendre comment structures et fonctions cérébrales se contraignent mutuellement. Notre recherche relève plus particulièrement de la connectivité effective qui permet de définir des graphes de connectivité qui renseignent sur les liens causaux, directs ou indirects, unilatéraux ou bilatéraux via des chemins de propagation, représentés par des arcs, entre les nœuds, ces derniers correspondant aux régions cérébrales à l'échelle macroscopique. Identifier les interactions entre les aires cérébrales impliquées dans la génération et la propagation des crises épileptiques à partir d'enregistrements intracérébraux est un enjeu majeur dans la phase pré-chirurgicale et l'objectif principal de notre travail. L'exploration de la connectivité effective suit généralement deux approches, soit une approche basée sur les modèles, soit une approche conduite par les données comme nous l'envisageons dans le cadre de cette thèse où les outils développés relèvent de la théorie de l'information et plus spécifiquement de l'entropie de transfert, la question phare que nous adressons étant celle de la précision des estimateurs de cette grandeur dans le cas des méthodes développées basées sur les plus proches voisins. Les approches que nous proposons qui réduisent le biais au regard d'estimateurs issus de la littérature sont évaluées et comparées sur des signaux simulés de type bruits blancs, processus vectoriels autorégressifs linéaires et non linéaires, ainsi que sur des modèles physiologiques réalistes avant d'être appliquées sur des signaux électroencéphalographiques de profondeur enregistrés sur un patient épileptique et comparées à une approche assez classique basée sur la fonction de transfert dirigée. En simulation, dans les situations présentant des non-linéarités, les résultats obtenus permettent d'apprécier la réduction du biais d'estimation pour des variances comparables vis-à-vis des techniques connues. Si les informations recueillies sur les données réelles sont plus difficiles à analyser, elles montrent certaines cohérences entre les méthodes même si les résultats préliminaires obtenus s'avèrent davantage en accord avec les conclusions des experts cliniciens en appliquant la fonction de transfert dirigée.

Mesures Entropiques de Connectivité avec Application à l'Épilepsie

Quelle est la structure du cerveau ? Quel en est le fonctionnement physiologique ? Quels sont les dysfonctionnements qui peuvent le dénaturer et le faire passer d'un état sain à un état pathologique ? C'est au travers de la connectivité effective que nous allons chercher des éléments de réponse à ces questions et essayer de comprendre ce que sous-tend ce concept que nous aborderons sous l'angle de la théorie de l'information et plus exactement sous celui de l'entropie de transfert et de son estimation.

Chapitre 1. Contexte Clinique

Notre recherche s'inscrit dans le contexte large d'une compréhension, générique et patient par patient, de certains types d'épilepsie dont les manifestations peuvent être notablement réduites, voire disparaître, à la suite d'un acte chirurgical consistant en une résection d'une région cérébrale qui doit, pour cela, être extrêmement bien délimitée par des examens pré-chirurgicaux incluant l'analyse de signaux électroencéphalographiques, enregistrés en dehors et pendant les périodes critiques. Rappelons que l'épilepsie est une maladie neurologique relativement répandue qui concerne environ 40 millions de personnes dans le monde, dont, en France, 7 individus sur 1000. C'est une affection chronique qui s'étale dans le temps et se caractérise par la survenue de convulsions (ou crises convulsives) qui sont le résultat de décharges électriques paroxystiques. On distingue deux types d'épilepsie, l'épilepsie généralisée, lorsque les décharges ont lieu dans l'ensemble du cortex cérébral et l'épilepsie partielle, lorsque les décharges se produisent dans une partie bien délimitée du cortex cérébral, et qui sont celles concernées ici. L'épilepsie est le résultat d'une activation survenant subitement, de manière simultanée et anor-

malement soutenue, d'un nombre très important de groupes neuronaux, dans certaines régions du cerveau, dites épileptogènes. Elle peut se traduire par des manifestations neurologiques telles que des troubles visuels, olfactifs, auditifs, gustatifs, des pertes de conscience, des convulsions...Si la cause exacte de cette maladie est parfois inconnue, elle peut être par ailleurs la conséquence d'une tumeur au cerveau, d'un accident cérébral vasculaire, d'une intoxication, d'une malformation des vaisseaux cérébraux, de séquelles d'un traumatisme. Différents traitements de la maladie sont envisageables suivant la nature du syndrome, son origine et son intensité : prise de médicaments, stimulation électrique, traitement chirurgical. C'est dans ce dernier cas que se positionne ce travail, lorsque l'épilepsie est pharmaco-résistante (*i.e.* rebelle à tout traitement médicamenteux) et qu'il faut s'en remettre à une chirurgie curative et pratiquer une exérèse d'une partie du cortex cérébral pour supprimer la cause de l'épilepsie. Les patients souffrant d'épilepsies pharmaco-résistantes bénéficient d'une évaluation pré-chirurgicale visant à délimiter la Zone Epileptogène (ZE), responsable de la genèse et de la propagation des crises épileptiques. Une question clé dans la prise en charge de ces patients est donc l'identification de cette ZE. Le service de Neurologie du CHU de Rennes, partenaire du laboratoire, fait partie des rares unités cliniques pratiquant l'enregistrement de signaux intracérébraux (recueillis sur des électrodes de profondeur). Bien qu'invasive, cette procédure permet une exploration plus fine du cerveau, directement au contact des régions épileptiques, en vue de définir au mieux quelle(s) région(s) exciser. Cette définition implique d'identifier des interactions pathologiques entre les aires cérébrales impliquées dans la génération et la propagation des crises. Atteindre ce but à partir d'enregistrements intracérébraux est un enjeu clinique majeur dans la phase pré-chirurgicale et correspond à la motivation principale de notre travail. Etablir ainsi un graphe de connectivité cérébrale à partir de signaux enregistrés en profondeur ou en surface requiert l'application de techniques avancées de traitement du signal. Cette connectivité se décline en trois types. La connectivité anatomique (ou structurelle) réfère à l'ensemble des connexions physiques (*i.e.* axonales) liant des groupes neuronaux plus ou moins grands (de la connexion synaptique entre neurones individuels à l'analyse de paquets de connexions ou d'ensembles synaptiques liant des populations neuronales, voire à grande échelle, à l'analyse des connexions vues comme des chemins liant les grandes régions du cerveau). La connectivité fonctionnelle repose, quant à elle, sur l'étude des liens statistiques entre des signaux reflétant des activités cérébrales dans des régions distinctes, sans ambitionner de mettre en évidence des influences causales. Elle renvoie ainsi au concept de corrélation spatio-temporelle entre activités résultant d'interactions neuronales dynamiques permises par les connexions anatomiques permanentes. Bien évidemment, cette connectivité fonctionnelle entre deux aires cérébrales n'implique pas nécessairement l'existence d'une connexion anatomique directe entre elles, une corrélation mesurée pouvant être le résultat d'une médiation par une structure tierce. Plus avant, la connectivité effective est une notion plus forte en ce

sens qu'elle s'intéresse à l'organisation des flux d'information entre régions cérébrales et est définie comme l'influence directe ou indirecte exercée par un système neuronal sur un second système.

Ces trois types de connectivité que l'on peut considérer à différentes échelles d'espace et de temps sont bien évidemment liés et leur analyse conjointe permet de mieux comprendre comment structures et fonctions cérébrales se contraignent mutuellement. Notre problématique relève plus particulièrement de la connectivité effective qui permet entre autres de définir des graphes de connectivité renseignant sur les liens causaux, directs ou indirects, unilatéraux ou bilatéraux via des chemins de propagation (représentés par des arcs) entre les nœuds, ces derniers correspondant aux régions cérébrales à l'échelle macroscopique. Au-delà de la connectivité fonctionnelle qui renvoie à la notion de couplage statistique entre signaux, il s'agit donc d'établir des graphes orientés, dans lesquels les éléments constitutifs sont pondérés par des coefficients traduisant la densité ou l'efficacité des connexions, ces graphes décrivant des flux d'informations entre populations impliquées. L'enjeu est donc de comprendre les fonctions du cerveau non seulement en identifiant correctement les régions activées mais aussi en décelant les interactions fonctionnelles au cours du temps parmi les ensembles neuronaux stimulés éventuellement éloignés. Notre questionnement concerne donc le sens de circulation de l'information entre différents sites, sens pouvant évoluer durant le décours temporel d'une crise. D'un point de vue applicatif (investigation clinique) ce à quoi nous voulons répondre à long terme se résume à : quelles sont les structures impliquées dans les activités neuronales au cours des crises ? Est-il possible de détecter des structures dominantes dans les réseaux épileptogènes ? Une réponse à ces questions doit dégager la route d'une compréhension des mécanismes mis en jeu au cours d'activités épileptiques afin d'éradiquer la survenue des crises de manière adaptée, patient par patient.

En amont de ces visées cliniques, la détection de causalités significatives à partir de signaux disponibles oblige à recourir à des techniques spécifiques de traitement du signal qui correspondent, après définition de certains indices théoriques candidats à exprimer cette causalité, à la production d'algorithmes d'estimation de ces indices, associés à une mesure de confiance statistique, généralement en termes de biais et de variance. Différents indices et algorithmes sont disponibles dans la littérature, concernant aussi bien l'analyse de mécanismes cérébraux, que celle d'autres phénomènes physiologiques, sans compter avec de nombreuses autres applications, comme par exemple l'analyse de systèmes dynamiques non linéaires couplés en physique. Une grande partie des résultats publiés dans ce champ sont validés au moyen de simulations de systèmes dynamiques aléatoires, linéaires ou non. Deux questionnements nous ont semblé *a priori* légitimes. Le premier concerne l'amélioration des estimateurs d'un type particulier d'indice, l'entropie de transfert, qui tout à la fois présente des vertus de grande généralité mais également

l'inconvénient de souffrir de biais d'estimation notable. Le deuxième concerne l'apport concret de certaines améliorations d'estimateurs d'entropie de transfert, parfois obtenues au prix d'une complexité calculatoire accrue, quand on les applique aux signaux réels rencontrés dans le domaine de l'épilepsie, bien connus pour leur nature complexe.

Chapitre 2. Etat de l'Art

L'exploration de la connectivité effective suit généralement deux approches, soit une approche conduite par les données (incluant les méthodes de type "causalité de Granger" ou basées sur la théorie de l'information) soit une approche basée sur les modèles comme il en va pour la modélisation causale dynamique (Dynamic Causal Modeling (DCM)).

L'idée de base pour la première approche revient à Wiener, bien que connue sous le nom de causalité de Granger (il est de fait plus correct de parler de causalité de Wiener-Granger puisqu'elle fut d'abord introduite par Wiener). Celle-ci consiste à considérer qu'un signal Y cause un signal X si Y contient des informations permettant d'affiner la prédiction de X (où la prédiction est à structure imposée, linéaire, et optimale au sens d'une variance d'erreur de prédiction minimale) relativement à une prédiction basée uniquement sur le passé de X et éventuellement sur d'autres variables contextuelles. Autrement dit, selon Granger, X est causé par Y si la prise en compte supplémentaire de valeurs passées de Y permet de mieux prédire les valeurs de X qu'en se basant uniquement sur les valeurs passées de X.

Depuis son introduction en économétrie par Granger en 1969, cette notion de causalité linéaire en moyenne quadratique a bien évidemment fait l'objet d'un certain nombre de débats, entre autres sur son insuffisance à capter les liens de causalité indirects pouvant exister entre deux variables, dès lors qu'il existe au moins une troisième série dans le système. S'en sont ensuivies les questions de mesures de causalité conditionnelle dans le domaine temporel, étendues par la suite au domaine fréquentiel. Si la causalité de Granger est devenue une mesure très largement utilisée, elle n'en reste pas moins perfectible dans des situations où les signaux traités s'avèrent complexes, et des extensions de cette mesure dans des situations présentant des non-linéarités ont déjà été envisagées. Plus récemment mais souvent dans la même lignée, de nouvelles approches fréquentielles ont été proposées et largement utilisées.

Parallèlement à ces méthodes conduites par les observations, se sont développées des méthodes basées sur des modèles, qui tiennent compte de certaines hypothèses *a priori* portant sur les mécanismes physiologiques sous-jacents, la plus connue étant sans doute l'approche de modèle causal dynamique ainsi nommée et popularisée dans les années 2000 par Friston, initialement pour l'analyse de réseaux cognitifs. Pour cela les popula-

tions neuronales *a priori* concernées dans le réseau étudié sont modélisées par des systèmes d'équations différentielles stochastiques non linéaires incluant des paramètres physiologiquement significatifs et dont les interactions sont représentées par des paramètres dits de connectivité. Les signaux d'observation, chacun attaché à l'une des populations, sont également modélisés, ce qui constitue la sortie du modèle à confronter aux observations. Pour décider de la topologie d'un graphe de connectivité effective entre les populations du réseau, on introduit autant de modèles du réseau que de graphes orientés candidats. Une procédure (coûteuse en temps de calcul) d'estimation paramétrique est ensuite appliquée à chacun des modèles envisagés, procédure qui délivre des valeurs de paramètres de populations et de connectivité. Finalement une procédure statistique de sélection de modèle élit le modèle le plus vraisemblable et donc le graphe de connectivité correspondant. Notons cependant que la connaissance d'un graphe statique (orienté) de connectivité est par essence qualitative et ne quantifie pas en elle-même les transferts dynamiques d'information entre populations. Une évaluation de ces derniers est théoriquement possible, mais nécessiterait une reconstruction des trajectoires d'état pour chaque population, ce qui finit par constituer une méthodologie lourde, d'où l'intérêt *a priori* des méthodes dirigées par les données qui sont les seules considérées dans cette thèse, particulièrement celles qui ont été proposées pour permettre l'analyse de dynamiques non linéaires. Ces dernières comprennent les méthodes s'appuyant sur des régressions non linéaires paramétriques ou pseudo non paramétriques (méthodes à noyau) et les méthodes non paramétriques dites entropiques popularisées depuis l'introduction par Schreiber de l'entropie de transfert en 2002, pour laquelle on peut établir une relation avec l'indice de causalité de Granger quand les observations sont assimilées à des processus aléatoires conjointement gaussiens. L'entropie de transfert est définie initialement seulement pour une paire (ordonnée) de signaux. Il s'avère cependant que cette méthode peut s'étendre sans difficulté, du moins du point de vue de la définition d'un indice théorique, à une forme conditionnelle si tant est que l'on cherche à exclure (ou plus justement à tenir compte de) l'influence d'un signal (ou d'un groupe de signaux) tiers. L'entropie de transfert sera très largement développée dans le chapitre suivant puisqu'elle est au cœur de ce travail de recherche.

Chapitre 3. Méthodes et Matériels

Comme mentionné précédemment, les méthodes basées sur la théorie de l'information, et plus particulièrement l'entropie de transfert, jouent un rôle essentiel dans la détection d'influences causales à partir de signaux d'observation. Lorsque ces derniers sont non stationnaires et que chaque mesure doit s'effectuer sur un intervalle d'observation relativement court (typiquement de l'ordre d'une à quatre secondes), la question qui se pose d'ores et déjà est la précision des estimations produites. Plus exactement, la pierre

d'achoppement dans cette estimation réside dans le biais qui lui est attaché. Nos travaux s'inscrivent donc dans cette problématique, ce chapitre étant consacré à (i) définir les quantités entropiques concernées d'un point de vue formel, (ii) présenter des estimateurs proposés dans la littérature pour ces quantités, et enfin (iii) proposer de nouveaux estimateurs conçus pour abaisser le biais, la stabilité de la variance n'étant contrôlée qu'expérimentalement.

Dans ce chapitre, nous donnons la définition des différentes quantités entropiques (des fonctionnelles à valeurs réelles admettant en argument une ou plusieurs distributions de probabilité) impliquées dans notre travail, nommément l'entropie, les entropies conjointe et conditionnelle, l'information mutuelle, les entropies de transfert standard et conditionnelle. Des points communs entre le calcul de l'information mutuelle et celui de l'entropie de transfert sont soulignés.

Puis, avant de nous focaliser sur l'entropie de transfert elle-même, nous commençons par rappeler les techniques les plus utilisées dans l'estimation de l'entropie, à savoir les approches à noyaux et celles basées sur les plus proches voisins. Ces dernières font l'objet de nos développements ultérieurs. Elles impliquent l'utilisation d'une distance dans l'espace des observations déduite d'une norme qui doit être spécifiée, en général la norme du maximum ou la norme euclidienne. Ainsi, les deux estimateurs d'entropie qui sont considérés par la suite peuvent être implémentés tant pour la norme euclidienne que la norme du maximum. Toutefois, par la suite, sera essentiellement considéré le cas de cette dernière norme, pour laquelle une boule formelle correspond concrètement à un cube. Ces deux estimateurs issus de la littérature sont celui de Kozachenko-Leonenko et celui de Singh, menant à des estimations très proches pour de longues séquences d'observations.

Ce chapitre se poursuit alors par l'estimation de l'information mutuelle qui peut s'écrire comme une somme algébrique d'entropies conjointe et marginales. Cette écriture peut faire espérer une compensation partielle des biais à condition qu'ils soient du même ordre de grandeur. Pour que cette dernière condition soit remplie, la première stratégie adoptée par Kraskov consiste à fixer le nombre k de voisins dans l'espace joint (comme d'ordinaire) et d'exporter dans les espaces marginaux la distance entre le point courant et son $k^{\text{ème}}$ plus proche voisin, *i.e.* le rayon du voisinage-boule dont le volume intervient dans l'estimation, pour y construire des boules de même rayon que dans l'espace d'origine. Dans une deuxième étape, Kraskov suggère de reconsidérer le voisinage dans l'espace joint en remplaçant l'hyper-cube par un hyper-rectangle et en exportant cette fois deux valeurs de distance distinctes dans les deux espaces marginaux, respectivement, conduisant à l'écriture d'un second estimateur de l'information mutuelle.

La suite de ce chapitre pose logiquement la question de l'estimation de l'entropie

de transfert en présentant tout d'abord deux estimateurs issus de la littérature basés sur les travaux précités de Kraskov et en ouvrant sur une discussion qui pose les bases des améliorations qui seront proposées pour les estimateurs déjà existants, soit plus précisément :

- (i) considérant l'estimation de l'information mutuelle sous la forme d'une somme algébrique d'entropies, il est possible d'exprimer le biais résultant comme une même combinaison des biais attachés à leurs estimations respectives, et un choix judicieux des rayons des voisinages dans les espaces marginaux peut permettre de le réduire (c'est la démarche suivie empiriquement par Kraskov qui l'a appliquée au cas de la norme du maximum). Toutefois, cette procédure requiert de disposer d'une expression théorique du biais (pour une norme quelconque) et de pouvoir en déduire les rayons des boules dans les espaces marginaux pour l'annuler ou du moins l'atténuer. Pour cela, les valeurs optimales des rayons marginaux doivent pouvoir s'exprimer en fonction des seules informations disponibles, ce qui exclut de les faire dépendre des fonctions de densité de probabilité et de leurs dérivées ;

- (ii) une seconde idée pour réduire individuellement chaque biais est de considérer comme voisinages les hyper-rectangles de volume minimal incluant les voisins sélectionnés, aussi bien dans l'espace joint que dans les espaces marginaux.

Ces éléments de réflexion sont à l'origine des améliorations proposées dans la suite.

Ainsi, dans un premier temps, nous introduisons une forme analytique du biais pour l'estimation de l'entropie individuelle basée sur un développement de Taylor à l'ordre deux, utilisable aussi bien pour l'approche d'estimation de densité de probabilité par noyau que pour l'approche des k plus proches voisins, à la seule condition que le rayon de la boule-voisinage ne soit pas trop grand. Cette forme analytique du biais est donnée à la fois dans le cas de la norme euclidienne et de la norme du maximum. Dans le cas de signaux indépendants, une relation est alors établie entre les rayons des voisinages dans les espaces marginaux et celui déterminé par l'emplacement du $k^{\text{ème}}$ plus proche voisin dans l'espace joint, pour chacune de ces deux normes, afin d'annuler le biais à la fois dans l'estimation de l'information mutuelle et de l'entropie de transfert. Cependant, la condition d'indépendance s'avère, dans le cas de l'entropie de transfert, peu réaliste, car elle implique pratiquement que le signal subissant potentiellement une influence causale soit une suite de variables indépendantes et identiquement distribuées (un bruit blanc). Même si ce ne sera pas le cas dans ce travail, la stratégie proposée pourrait être généralisée en vue d'être appliquée à l'estimation de l'entropie de transfert conditionnelle. Sous l'hypothèse de non indépendance, il n'est plus possible d'annuler le biais avec la même stratégie et notre choix s'est porté vers des combinaisons linéaires pondérées d'estimateurs que ce soit pour le calcul de l'entropie, de l'information mutuelle ou de l'entropie de

transfert. Il s'agit en fait de faire intervenir dans les différentes grandeurs des nombres de voisins variables différemment pondérés et de trouver le meilleur compromis. Cette procédure conduit aux estimateurs que nous dénommons "estimateurs composés".

Dans une dernière partie, nous repartons des estimateurs d'entropie proposés d'une part par Kozachenko-Leonenko, et d'autre part par Singh, dont nous résumons tout d'abord les développements mathématiques. Nous proposons ensuite de modifier ces deux estimateurs de la même manière, *i.e.* en substituant au voisinage hyper-cubique de volume minimal incluant les k plus proches voisins un voisinage hyper-rectangulaire de volume minimal incluant ces mêmes voisins. Avec ce type de voisinage la probabilité d'avoir plus d'un point sur la frontière (en l'occurrence 2 points dans le cas à deux dimensions) est strictement positive, ce qui nécessite de reprendre les développements mathématiques précédents pour ce nouveau type de voisinage. Finalement, nous obtenons deux nouveaux estimateurs d'entropie, le premier étendant celui de Kozachenko-Leonenko et le second l'estimateur d'entropie de Singh. Ce dernier nécessite de déterminer le nombre de points sur les frontières des hyper-rectangles sans que cela présente un inconvénient pratique. Ces deux estimateurs d'entropie sont ensuite utilisés pour proposer respectivement deux nouveaux estimateurs d'entropie de transfert.

Chapitre 4. Résultats Expérimentaux

Dans le chapitre précédent, nous avons proposé différentes stratégies pour l'estimation de l'information mutuelle et de l'entropie de transfert qu'il s'agit d'évaluer en considérant différentes situations, incluant signaux indépendants ou dépendants, relations linéaires et non linéaires. Les premières simulations portent sur des processus blancs gaussiens mais aussi des modèles vectoriels autorégressifs parfois non linéaires. Ces choix ont été conduits d'une part pour disposer autant que faire se peut de valeurs de référence et d'autre part pour une certaine représentativité de caractéristiques pouvant être celles de signaux réels. Dans un second temps, nous nous sommes intéressés à un modèle physiologique de populations neuronales potentiellement couplées. Pour ce modèle pour lequel nous ne disposons pas de valeur théorique de référence, des tests statistiques sont conduits pour valider nos approches.

Concernant les performances des différents estimateurs d'information mutuelle, comme attendu, le premier modèle (modèle 1) testé sur deux signaux indépendants met en évidence la supériorité des estimateurs récemment proposés dans la litterature comparés à ceux utilisant le même nombre de voisins. Dans le cas de deux signaux dépendants, deux autres modèles sont testés, le premier (modèle 2) pour considérer l'effet de matrices de covariance des observations non diagonales, le second (modèle 3) pour mettre en exergue "le fléau de la dimension". Dans les deux cas, les nouveaux estimateurs présentent un

meilleur comportement au regard des estimateurs issus de la littérature, pour les deux normes, ce résultat étant d'autant plus vrai que la corrélation entre signaux est élevée, leur longueur faible et la dimension importante.

Pour ce qui est de l'entropie de transfert, le premier modèle testé (modèle 4) simule une suite d'observations (chacune rassemblant une valeur à prédire de X, un vecteur correspondant à son passé et un vecteur correspondant au passé de Y) indépendantes, pour lesquelles les matrices de covariance ne sont pas nécessairement diagonales. Les deux estimateurs proposés sont testés et comparés à l'indice de causalité de Granger ainsi qu'aux algorithmes issus de la littérature (l'algorithme standard et l'algorithme étendu). Pour des dimensions faibles et un nombre de voisins suffisant, les algorithmes proposés surpassent les autres estimateurs d'entropie de transfert mais s'avèrent moins performants que l'indice de Granger. Les deux modèles suivants (modèles 5 et 6) décrivent des signaux autorégressifs linéaires respectivement au nombre de 2 et 3 présentant des connectivités bidirectionnelles pour chaque paire de signaux. Là encore, l'indice de Granger se révèle le plus pertinent au prix d'une variance parfois légèrement accrue. Toutefois, parmi les estimateurs d'entropie de transfert, les nouveaux estimateurs ont des comportements extrêmement corrects, l'un d'eux se montrant, de façon quasi-systématique, plus efficace que ceux issus de la littérature. L'intérêt des estimateurs d'entropie de transfert trouve son sens dans les résultats rapportés sur le modèle suivant (modèle 7) qui présente de fortes non-linéarités. Dans ce cas, l'indice de Granger est mis en échec alors que les différents estimateurs d'entropie de transfert continuent à bien se comporter et à tendre vers la valeur théorique pour des longueurs suffisantes de signaux.

Pour finir, quatre estimateurs de transfert d'information (indice de Granger, algorithmes standard et étendu, et premier estimateur d'entropie de transfert proposé) sont appliqués sur le modèle physiologique évoqué plus haut après avoir vérifié la pertinence des estimateurs d'information mutuelle sur la dépendance ou non des populations neuronales générées. Il s'avère que l'indice de Granger et le nouvel estimateur sont les seuls à distinguer avec pertinence les trois situations considérées (populations indépendantes, connectivité unidirectionnelle et bidirectionnelle).

En conclusion de ce chapitre, pour l'information mutuelle, les résultats sur signaux simulés prouvent l'efficacité de la stratégie proposée dans le chapitre précédent. Quant à l'entropie de transfert, si l'un des estimateurs se trouve en difficulté lorsque l'on considère un nombre de voisins faible, les deux estimateurs développés apparaissent appropriés dans une vision de réduction du biais d'estimation, d'autant plus en présence de non-linéarités, et ce pour des temps de calculs comparables à ceux demandés par les estimateurs déjà existants.

Chapitre 5. Analyse de Signaux Réels

Les estimateurs proposés dans le chapitre 3 s'étant avérés pertinents quand on les a évalués sur des simulations (chapitre 4), l'objectif du présent chapitre est de les confronter aux signaux réels enregistrés sur un patient épileptique, afin de vérifier si certaines hiérarchies de performances se maintiennent. Ces signaux réels proviennent d'un sujet souffrant d'une épilepsie temporale trouvant son origine dans l'hémisphère gauche et pour laquelle nous disposons de l'expertise de cliniciens. Les signaux à traiter s'avèrent nettement plus complexes que ceux produits par les modèles décrits dans le chapitre précédent, ne serait-ce que par leur caractère non stationnaire. Cette première analyse a conduit à des catégorisations et facilité l'analyse ultérieure de mesures de causalité. Dans ce cadre, nous avons choisi de tester différents indices, à commencer par l'indice de causalité de Granger. Comme mesures de transfert d'entropie, nous avons retenu les algorithmes issus de la littérature (algorithmes standard et étendu) ainsi que l'un des deux estimateurs proposés dans le chapitre 3. Après une revue de la littérature sur les outils les plus communément utilisés pour détecter les voies d'initiation et de propagation de l'épilepsie, nous avons également choisi de tester la fonction de transfert dirigée qui a souvent fait l'objet d'applications dans ce domaine. De cette mise en compétition, il ressort que, parmi les estimateurs de transfert d'entropie, celui proposé dans cette thèse est le plus efficient. Néanmoins, sur l'ensemble des estimateurs testés, malgré une certaine variabilité, la fonction de transfert dirigée se révèle la plus pertinente.

Cette thèse se conclut par une discussion sur les différentes contributions apportées en indiquant dès à présent des améliorations possibles des estimateurs proposés et ouvre des perspectives de travail sur ce vaste domaine de la connectivité effective.

Acknowledgements

This thesis was performed in the frame of CRIBs (Centre de Recherche en Information Biomédicale sino-français), which is an international associate French-Chinese laboratory (Université de Rennes 1 - France, INSERM - France, SouthEast University - China).

Foremost, I would like to express my special appreciation and thanks to my advisor Professor Régine LE BOUQUIN JEANNES for her continuous support of my Ph.D study and research. Her patience, motivation and enthusiasm really impressed me and will have a huge positive impact on my future life and work. I am also grateful to my co-supervisor Associate Professor Jean-Jacques BELLANGER. He showed me the beauty of mathematics and I benefited a lot from his important knowledge.

Besides, my sincere thanks go to Professor Huazhong SHU for his help during the thesis and his useful suggestions. I also appreciated the fruitful discussions with Isabelle MERLET on experimental signals. I also thank Professor Hongqing ZHU, who insisted to make me go on with further study when I was still an undergraduate student.

Additionally, I would like to thank Professor Olivier MICHEL and Professor Didier WOLF who accepted to review my thesis, as well as Professor Dominique PASTOR and Associate Professor Sofiane BOUDAOUD who accepted to preside at the jury. I really appreciated their insightful comments on the thesis and their hard questions which helped me to widen my research from various perspectives.

I am grateful to all members of LTSI, who gave me help, and not only on the thesis.

Last but not least, I would like to thank my family (my father, mother and little sister) for their spiritual support throughout my life.

Abstract

The work presented in this thesis deals with brain connectivity, including structural connectivity, functional connectivity and effective connectivity. These three types of connectivities are obviously linked, and their joint analysis can give us a better understanding on how brain structures and functions constrain each other. Our research particularly focuses on effective connectivity that defines connectivity graphs with information on causal links that may be direct or indirect, unidirectional or bidirectional. The main purpose of our work is to identify interactions between different brain areas from intracerebral recordings during the generation and propagation of seizure onsets, a major issue in the pre-surgical phase of epilepsy surgery treatment. Exploring effective connectivity generally follows two kinds of approaches, model-based techniques and data-driven ones. In this work, we address the question of improving the estimation of information-theoretic quantities, mainly mutual information and transfer entropy, based on k-Nearest Neighbors techniques. The proposed approaches we developed are first evaluated and compared with existing estimators on simulated signals including white noise processes, linear and nonlinear vectorial autoregressive processes, as well as realistic physiology-based models. Some of them are then applied on intracerebral electroencephalographic signals recorded on an epileptic patient, and compared with the well-known directed transfer function. The experimental results show that the proposed techniques improve the estimation of information-theoretic quantities for simulated signals, while the analysis is more difficult in real situations. Globally, the different estimators appear coherent and in accordance with the ground truth given by the clinical experts, the directed transfer function leading to interesting performance.

Contents

List of Abbreviations

- **ADTF**: Adaptive Directed Transfer Function
- **AEDs**: AntiEpileptic Drugs
- **AIC**: Akaike's Information Criterion
- **AR**: AutoRegressive
- **BIC**: Bayesian Information Criterion
- **CMI**: Conditional Mutual Information
- **CSE**: Causation Entropy
- **CTE**: Conditional Transfer Entropy
- **CT**: Computed Tomography
- **DCM**: Dynamic Causality Modeling
- **DI**: Directed Information
- **DTE**: Decomposed Transfer Entropy
- **DTF**: Directed Transfer Function
- **ECoG**: ElectroCorticoGram
- **EEG**: ElectroEncephaloGraphy
- **EGCI**: Extended Granger Causality Index
- **EPDC**: Extended Partial Directed Coherence
- **EZ**: Epileptogenic Zone

- **fMRI**: functional Magnetic Resonance Imaging
- **FDA**: Food and Drug Administration
- **gOPDC**: generalized Orthogonalized Partial Directed Coherence
- **GC**: Granger Causality
- **GPDC**: Generalized Partial Directed Coherence
- **iEEG**: intracranial (or intracerebral) ElectroEncephaloGraphy
- **IID**: Independent and Identically Distributed
- **ILAE**: International League Against Epilepsy
- **IP**: Inhomogeneous Polynomial
- k**NN**: k-Nearest Neighbors
- **KDE**: Kernel Density Estimation
- **KGC**: Kernel Granger Causality
- **KSG**: Kraskov-Stögbauer-Grassberger
- **LCI**: Local Causality Index
- **LLNAR**: Local Linear Nonlinear AutoRegressive
- **MEG**: MagnetoEncephaloGraphy
- **MI**: Mutual Information
- **MRI**: Magnetic Resonance Imaging
- **MSI**: Magnetic Source Imaging
- **MVAAR**: MultiVariate Adaptive AutoRegressive
- **MVAR**: MultiVariate AutoRegressive
- **NARX**: Nonlinear AutoRegressive with eXogenous inputs
- **NLGC**: NonLinear Granger Causality
- **PC**: Partial Correlation
- **PDC**: Partial Directed Coherence
- **PET**: Positron Emission Tomography

- **PMI**: Partial Mutual Information
- **PSD**: Power Spectrum Density
- **PTE**: Partial Transfer Entropy
- **RKHS**: Reproducing Kernel Hilbert Spaces
- **SDE**: Stochastic Differential Equation
- **SPECT**: Single-Photon Emission Computed Tomography
- **STE**: Symbolic Transfer Entropy
- **TDMI**: Time-Delayed Mutual Information
- **TE**: Transfer Entropy
- **TLMI**: Time-Lagged Mutual Information
- **VAR**: Vectorial AutoRegressive
- **VEPs**: Visual Evoked Potentials
- **VNS**: Vagus Nerve Stimulation
- **WGC**: Wiener-Granger Causality

Introduction

The detection of causal relations plays a fundamental role in many domains, where it is important to identify causal relations between a set of subsystems.

This work takes place in the context of human epileptic seizures. Epilepsy is a neurological disease, which is the fourth most common neurological disorder and affects people of all ages. This disease is characterized by the repetition of seizures, called ictal periods, whose frequency and duration are variable. Many epileptic patients may have other symptoms of neurological problems as well. In about 30% of cases, the patients do not successfully respond to anti-seizure drug therapy, or in other words, the epilepsies remain drug-resistant. In this case, surgery is an alternative for those patients, whose seizures cannot be controlled by medications.

The use of epilepsy surgery increased in the 1980s and 90s, which reflects its effectiveness as an alternative to seizure medicines. However, for this kind of treatment, there is no guarantee of success in controlling seizures, and the benefits of surgery should be weighed carefully against its risks. For a given patient, the foremost difficulty is to determine the epileptogenic zone which is responsible for the seizure, and the surgery should be carried out under the constraint that post-surgical deficits (sensitive, driving or cognitive) are limited. The initiation and the propagation of epileptic activities often take place in a network of neuronal ensembles which are distributed in distant structures. Now, the localization and the characterization of these distributed epileptogenic networks are known as difficult tasks but they are essential to expect eradicating seizures.

Technically, these tasks imply to adopt some quantitative description of causality and to develop algorithms aimed at estimating the strength of influence between different subsystems. In this work, we mainly focus on information-theoretic measures, such as mutual information and transfer entropy, which consider no assumption on the underlying model. These two quantities are generally estimated using a sum of individual entropies. A drawback of these measures is that it is difficult to obtain an accurate estimation with

a limited number of samples, which is common in real applications. Moreover, difficulty arises when those information-theoretic quantities are calculated in high-dimensional spaces, that is commonly called the "curse of dimensionality". So, it is meaningful to make effort to improve the estimation of these quantities. To this end, we consider two different ways in this work, (i) quantify the estimation bias of each individual entropy, and try to cancel out their combination in the summation, (ii) reduce the bias of each individual entropy estimation so as to decrease the total bias in the summation.

In this thesis, we first give some technical discussions on the estimation of information-theoretic quantities, and then apply different causality detection measures to real epileptic signals.

The remainder of this thesis is organized as follows.

Chapter 1 is a general introduction to the research background. In this chapter, firstly, we make a short summary of epilepsy and its treatment, including both medication and surgery. The structure of the human brain is recalled and different types of brain connectivity are presented.

In this work, we are interested in effective connectivity, which plays an important role in the treatment of epilepsy. Chapter 2 presents a detailed study of different effective connectivity estimation methods. We classify the measurements of effective connectivity into two classes, (i) data-driven methods, including Granger causality related and information-theoretic approaches which are the only ones to be addressed in this thesis, and (ii) model-based methods such as dynamic causal modeling (DCM).

Chapter 3 concerns the estimation of information-theoretic quantities. After a detailed summary of previous works, two novel improvements are deeply investigated. First of all, a novel analytical form of the bias for the estimation of entropy is presented. Then, we apply it into the estimation of mutual information and transfer entropy. Secondly, we discuss the estimation of information-theoretic quantities using the maximum norm. For the standard maximum norm, the determined region around the center point is a (hyper-)cube, where the sizes in all directions are identical. In this chapter, we propose to release this restriction, and use a (hyper-)rectangle instead of a (hyper-)cube, so that the sizes in the different directions can be of different lengths. In this chapter, several new mutual information and transfer entropy estimators are proposed.

In chapter 4, the performance of the different algorithms we propose is evaluated through numerical simulations under different situations, including independent and dependent signals, linear and nonlinear relations. Different types of models are tested, including white processes, linear vectorial autoregressive models, nonlinear models as well as a physiology-based model.

In chapter 5, we analyze intracranial (or intracerebral) electroencephalographic (iEEG) signals recorded in the cerebral cortex of an epileptic patient. Both Granger causality and different transfer entropy estimators are tested on these epileptic signals. Since the directed transfer function is a measure widely used in recent literature in the analysis of electroencephalographic (EEG) datasets, we decide to compare a local connectivity index based on our two-channel algorithms.

To conclude this thesis, we summarize our contributions to the field of the effective connectivity analysis and possible directions for further work are also addressed for essential improvement in clinical context.

Chapter 1

Research Background

Our research is performed in the context of epilepsy surgery and, in this chapter, we give a brief introduction to this field. In section 1.1, we present epilepsy and briefly describe the human brain in section 1.2 (including both external and internal morphologies), and, finally, in section 1.3, launch a short discussion on human brain connectivities.

1.1. Epilepsy

1.1.1. Introduction

Epilepsy [Magiorkinis 2010] is a group of neurological disorders, and it includes many different manifestations depending on various factors, like the age of the individual, the part of the brain that is affected, the underlying causes, among others [Smithson 2012]. This kind of disease has a very long history, and the first known detailed record of the disorder can be traced back to more than 3,000 years ago, which is written in a Babylonian cuneiform medical text [WHO 2005].

Nowadays, epilepsy is one of the most serious common neurological diseases, and there is a growing recognition of its harm on modern society. Worldwide, about 50 million people (1% of the world's total population) are affected by epilepsy, and nearly 80% of cases occur in the developing countries. In 1990, about 111,000 people died from epilepsy, and this number increased to 116,000 in 2013 [Naghavi 2015]. In Europe, due to epilepsy, the direct economic lost is around 15.5 billion Euros in 2004 [Nunes 2012]. It also impairs the quality of people's life; as a matter of fact, in many areas of the world, due to the high risk of being involved in a traffic accident, people with epilepsy have restrictions placed on their ability to drive or are not permitted to drive [Devlin 2012].

This kind of disease can be caused by brain injury, stroke, brain tumor, and drug/al-

cohol misuse, and in some rare cases, it is linked to genetic mutations [Malani 2012]. However, currently, the exact cause of most cases of epilepsy remains unknown.

Epilepsy is usually characterized by repetitive seizures, which are the results of excessive and abnormal cortical nerve cell activity in the brain [Fisher 2005]. Practically, the basis for the diagnosis of epilepsy is two or more unprovoked seizures occurring more than 24 hours apart [Engel 2008]. In the document of the international league against epilepsy (ILAE) [Fisher 2005], an epileptic seizure is defined as "a transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain". The duration of an epileptic seizure can vary from nearly undetectable to long periods, where the patient often becomes unconscious and loses control of his body [Browne 2008].

Recurrent and unprovoked epileptic seizures are the main characteristics of epilepsy, but epilepsy is more than seizures. A fundamental definition of epilepsy can be found in the document of ILAE [Fisher 2005]:

> *"A chronic condition of the brain characterized by an enduring propensity to generate epileptic seizures, and by the neurobiological, cognitive, psychological, and social consequences of this condition."*

According to this definition, there is a persistent intrinsic epileptogenic abnormality existing in the brain of epileptic patient, even outside the duration of seizures. So, as illustrated in Fig. 1.1, for the affected patient, epilepsy is not only limited to seizures, but also includes psychological and social consequences.

1.1.2. Epilepsy Treatment

Once a firm diagnosis of epilepsy has been made, the treatment of this disease must be considered, for which the desired goal is "no seizures, no side effects ".

1.1.2.1. Medications

For the majority of patients (about 70% [Eadie 2012]), seizures can be controlled by taking antiepileptic drugs (AEDs), which can contribute to decrease the number and/or duration of seizures. Usually, anticonvulsant medication treatment lasts for one person's entire life. Epilepsy cannot be cured by AEDs, and anticonvulsant prevents seizures from occurring by changing the levels of the chemicals in the brain that conduct electrical impulses. So, seizures may still occur while taking antiepileptic medication.

The choice of anticonvulsant depends on various factors, like seizure type, epilepsy syndrome, or other medications used [Nunes 2012]. Currently, there are about twenty

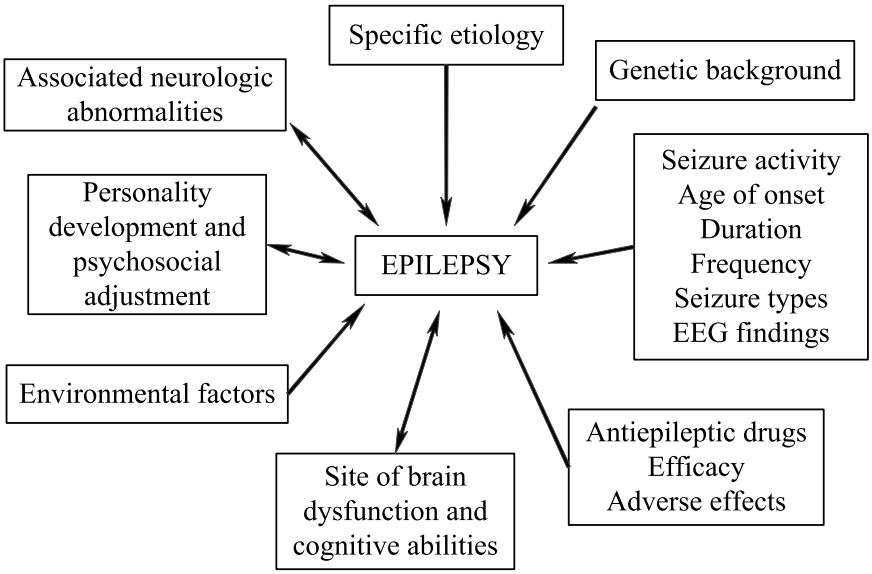


Figure 1.1: Different interacting factors that contribute to the totality of epilepsy [Engel 2008].

medications in the United States, which are approved for the use of epilepsy treatment by the food and drug administration (FDA), and several new drugs are in the clinical testing stages.

It is possible for the antiepileptic medication to cause some side effects, which can be divided into three categories [RCH 2015]: (a) common mild side effects at the beginning of medication, especially with a rapidly increasing dose, like abdominal pain, dizziness, sleepiness, irritability, (b) side effects caused by a too large dose, including unsteadiness, poor concentration, double vision, (c) peculiar side effects in individual medications, involving liver troubles, severe behavior disturbance and worsening of seizure control.

However, for about 30% of patients, medication is not effective or is intolerable (due to the side effects of the medication), and in this case, epilepsy surgery could be an option [Duncan 2006].

1.1.2.2. Surgery

Epilepsy surgery is a brain operation, whose goal is to resect the area of the brain involved in seizures, or restrict the spread of seizure activities (or reduce the seizure frequency or severity), where the corresponding results of such a surgery can be curative or palliative. Depending on the type of seizures and the location of the seizure focus, there are two different types of surgical strategies [Ellen 2015]: (a) curative procedures, such as temporal lobectomy, cortical excision, or hemispherectomy, (b) palliative procedures, such as corpus callosotomy or vagus nerve stimulation (VNS).

This kind of surgery is risky since it can worsen existing problems or create new ones in the brain functions, for instance loss of functions such as vision, speech, memory or movement. Therefore, epilepsy surgery is considered only within certain situations: (a) seizures are uncontrollable with medications, (b) the seizure focus can be clearly identified and is not responsible for any critical functions, such as language, sensation and movement, and (c) the life quality of the patient is significantly affected by the seizures.

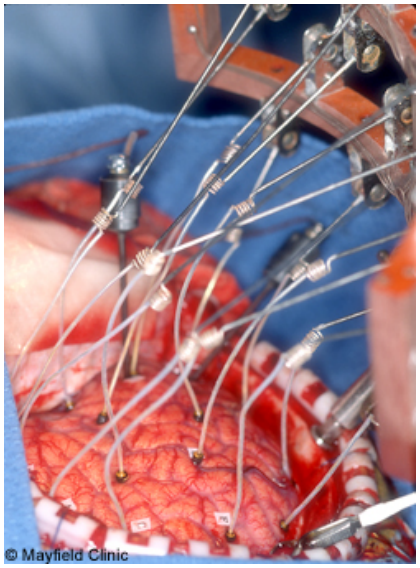


Figure 1.2: Intracerebral electrodes are surgically implanted into the brain tissue to map the seizure activities before removing brain structures [Ellen 2015].

Several modalities are used for the evaluation of seizures, including neuropsychological testing, invasive intracranial monitoring, and neuroimaging such as skull radiography, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), single-photon emission CT (SPECT), and magnetoencephalography/-magnetic source imaging (MEG/MSI).

During intracranial monitoring, the surgical implantation of EEG electrodes is often performed in order to map brain area, or localize the seizure focus. Generally, electrodes placed on the brain surface and/or directly inside the brain can be used to map seizure activity and/or identify important functional areas. As shown in Fig. 1.2, electrodes are surgically implanted into the brain tissue, and this kind of practice, where electroencephalographic signals are recorded via depth electrodes, is called iEEG [Palmini 2006]. Currently, iEEG is the only technique that provides direct access to electrophysiological recordings in the seizure onset zone, and it allows the determination of the depth of epileptogenic areas. It should be mentioned here that, currently, this procedure is risky, and it could lead to brain hemorrhage and infection in rare cases (less than 1%) [Cossu 2005].

During the mapping, the patient has to be awakened to identify functional areas such as language, sensation or vision [Ellen 2015]. These data collected with small electric probes are examined by the experts on a comprehensive epilepsy board, and surgery is processed based on their conclusions. In some rare cases, further surgery is not recommended if a single seizure focus cannot be revealed during the intracranial monitoring.

In some cases, seizures can be completely controlled after surgery, while, for others, the frequency of seizures is significantly reduced. After surgery, most patients need to continue taking anti-seizure medication for at least one year. In 2011, a study, which was performed among 615 adults who underwent epilepsy surgeries, identified long-term outcomes of epilepsy surgery in adults [de Tisi 2011]. According to this study, about 52% of patients remained seizure-free five years after the surgery.

The problem is then to exploit collected iEEG signals to discover strategic informations about the spatio-temporal epileptic propagation mechanisms between involved structures in the brain. This leads to an approach which has its roots in the connectivity paradigm introduced in the next section.

1.2. Human Brain

1.2.1. Introduction

The brain acts as the center of human nervous system, and a good understanding on the internal organization of the brain, which is always the central issue of modern neurobiology, would help in the treatment of brain diseases, especially epilepsy.

1.2.1.1. Morphology of the Brain

The brain weighs about $1300 \sim 1400$ grams (about 2% of total body weight) [Eric 2015] and is protected by the skull. Compared with the brain of other mammals, the human brain has a larger relative size. It is composed of two hemispheres (left and right), which are nearly symmetrical. These two hemispheres are separated by a deep median furrow (longitudinal fissure of the brain, or inter-hemispheric fissure) and interconnected by a very large nerve bundle (the corpus callosum, which crosses the midline above the level of the thalamus) and two other smaller connections, the anterior commissure and hippocampal commissure [Mooshagian 2008]. On the surface of the hemispheres, there is a pallium of very pleated gray matter. This superficial gray matter is the cerebral cortex, which contains many folds. On the cerebral cortex, the deepest folds are called furrows (or fissures). Each hemisphere can be conventionally divided into four lobes: the frontal lobe, parietal lobe, occipital lobe, and temporal lobe. The temporal lobe is on the side of

the brain, the parietal lobe is positioned above the occipital lobe and behind the frontal lobe. Fig. 1.3 and 1.4 display the lateral and medial surfaces of the cerebral hemisphere respectively.

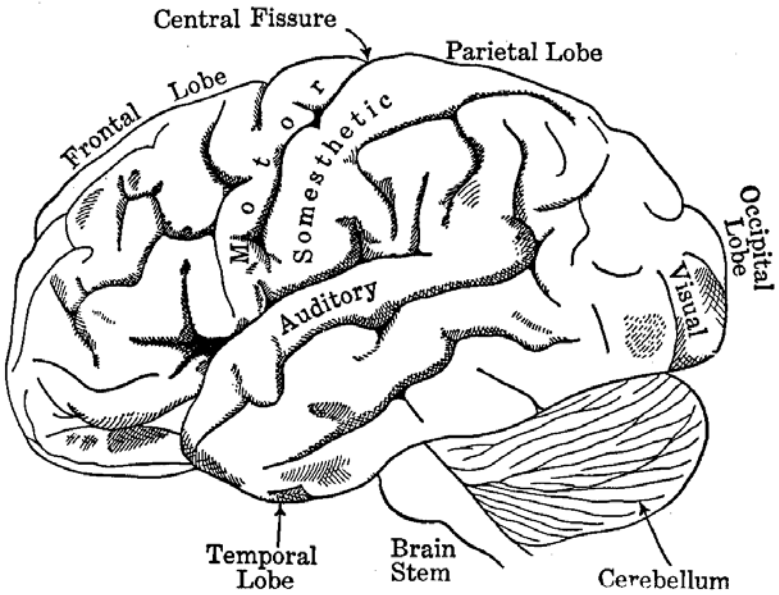


Figure 1.3: Lateral surface of the cerebral hemisphere [Dewey 2007]. On each hemisphere, there are four lobes bounded by fissures: the frontal lobe, parietal lobe, occipital lobe, and temporal lobe.

By cutting the interhemispheric commissures and opening the third ventricle, we have a vision on the medial aspect of the hemisphere. As shown in Fig. 1.4, on the medial side, there is a special cortical convolution, called cingulate gyrus (or limbic gyrus), which is delimited by the cingulate sulcus (calloso-marginal fissure). This convolution is wrapped around the deep part of the hemisphere. Above the limbic convolution, one can distinguish the frontal lobe. The medial surface of the occipital lobe is the cuneus, delimited by the groove parietaloccipital sulcus (internal perpendicular fissure) and the calcarine sulcus (fissure calcarine), which is the cortical projection of vision.

Different regions on the lower face of the hemispheres are shown in Fig. 1.5. In the center of the underside of the brain between two hemispheres, there is the isthmus of the brain, which corresponds to the junction of the brain stem and the brain.

1.2.1.2. Brain Gray and White Matter

The central nervous system contains two different major components: the grey matter and the white matter (as displayed in Fig. 1.6). The first one contains numerous cell bodies and relatively few myelinated axons (myelinization speeds up the neural electric propagation), while the second one is composed of long-range myelinated axon tracts and contains relatively very few cell bodies [Miller 1980].

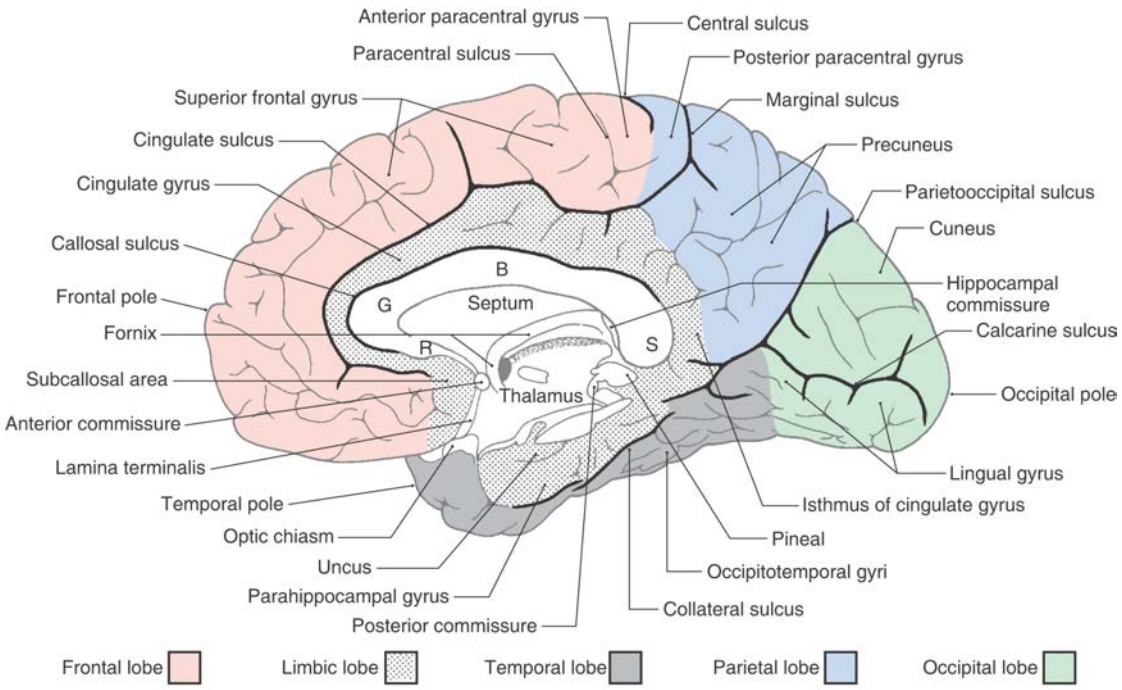


Figure 1.4: Medial surface of the cerebral hemisphere [Haines 2015]. Different lobes and their associated gyri and sulci are marked in this figure.

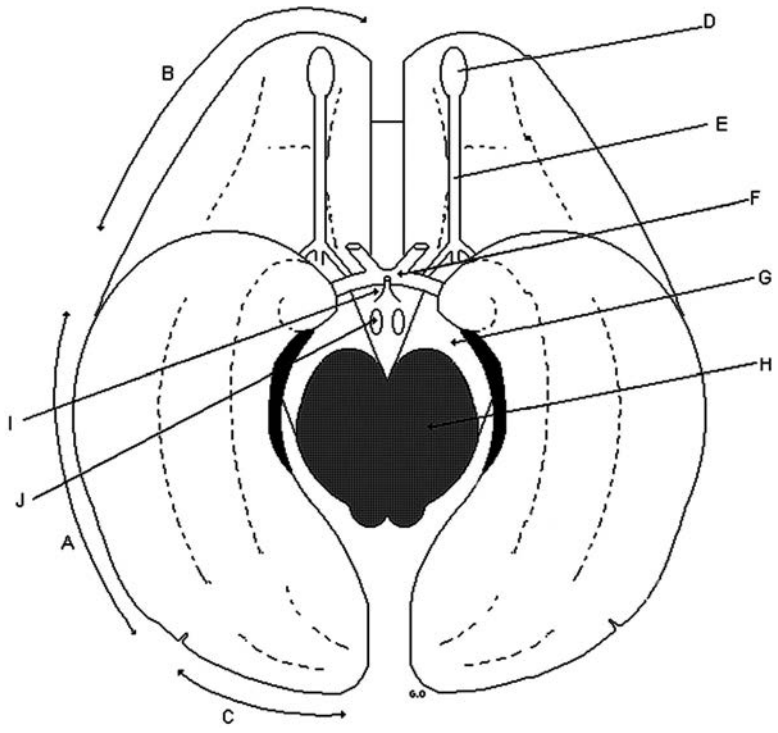


Figure 1.5: Lower face of the hemispheres [Bertrand 2015]. A: temporal lobe, B: frontal lobe, C: occipital lobe, D: olfactory bulb, E: olfactory tract, F: optic chiasm, G: cerebral peduncle, H: brainstem, I: pituitary gland, J: mammillary tubercle.

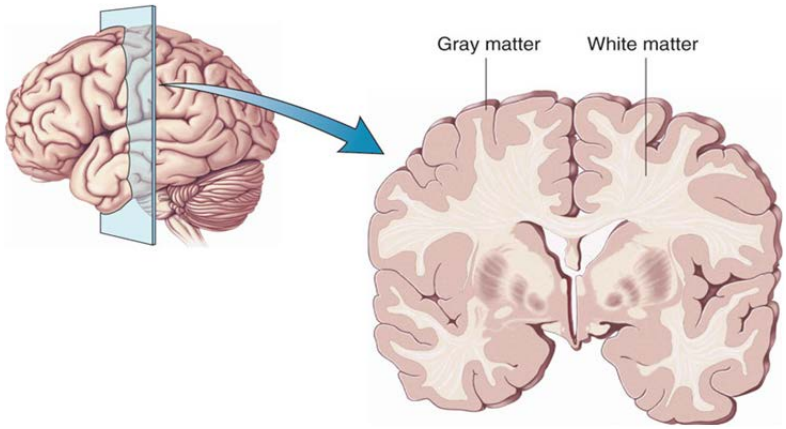


Figure 1.6: Gray and white matter in human brain [Giovannoni 2015].

The neocortex is the largest part of the cerebral cortex, and it covers the entire surface of the two hemispheres. There are two primary types of neurons in the neocortex, excitatory pyramidal neurons (∼80% of neocortical neurons) and inhibitory interneurons (∼20%) [Noback 2005]. The thickness of the neocortex is about 4 mm, and, as shown in Fig. 1.7, it is made up of six horizontal layers segregated principally by cells types and neuronal connections [Kurzweil 2013]. Also, there are numerous vertical structures in the neocortex (with a diameter of about 0.5 mm and a depth of about 2 mm), called cortical columns, where the neurons are arranged vertically. The number of cortical columns in neocortex is approximately 500,000, and each of them contains about 70,000 neurons [EPFL 2015].

White matter, which is located beneath the gray matter, is composed of millions of axons and forms a large part of the fabric of the cerebral hemispheres. Its white color comes from the electrical insulator, myelin, which surrounds axons. Myelin is found in all long nerve fibers and increases the speed of transmission of all nerve signals [Breedlove 2007]. If it is damaged, the nerve conduction will be affected, and thus alter the sensory, movement and cognitive functions. Three different kinds of tracts (or bundles of axons) are observed in the white matter: (a) projection tracts that carry information between the cerebrum and the rest of the body, (b) commissural tracts that cross from one cerebral hemisphere to the other through commissures and (c) association tracts that connect different regions within the same hemisphere of the brain [Kenneth 1998].

1.2.2. Brain Connectivity

As mentioned previously, the neurons in the brain act as information processing units and assembled circuits that perform brain functions, or in other words, they form a distributed network. Therefore, the study of the connectivity between different brain components is always a central issue in neuroscience, which aims to find links between

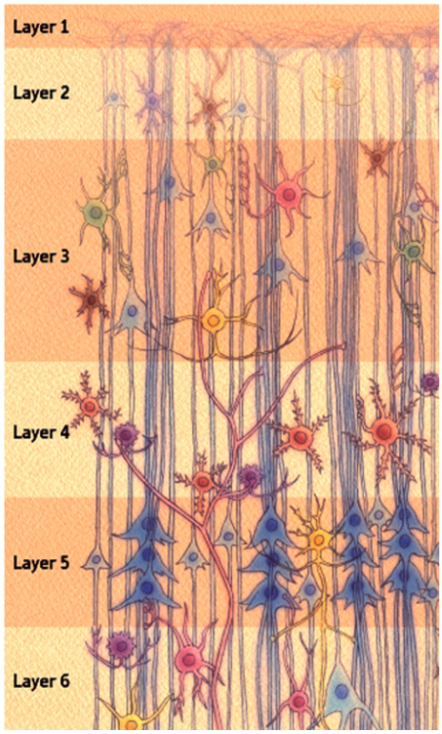


Figure 1.7: Histological structure of the cerebral cortex [Michael 2012]. The neocortex is composed of nerve cells which arranged in six horizontal layers segregated principally by cells types and neuronal connections. Layer 1: cell surface association, Layer 2: cells of intra-hemispheric association, Layer 3: small pyramidal cells, Layer 4: projection of sensitive and sensory cells, Layer 5: large pyramidal cells of Betz (origin of pyramidal tract), Layer 6: inter-hemispheric cells association (callosal fibers).

structures and functions at different levels of description [Alexandre 2013].

Mathematically, a network is considered as a graph structure which represents some type of links/interactions between units. In the brain network, the word "unit" may correspond to an individual neuron, a neuronal population, or an anatomically segregated brain region [Sporns 2007]. This graph includes a list of devices (or nodes) and a set of links, grouped in a connectivity matrix. The links can be oriented or not. If they are binary values, we get a pure structural graph. Now, the graph can be valued: each unit and each link may have a value (*i.e.* corresponding respectively to a state of activation and neuronal synaptic weight).

There are three types of brain connectivity [Friston 1994]: structural connectivity (anatomical), functional connectivity (measuring statistical dependence between neuronal activations) and effective connectivity (measuring the causal interactions or information flows on a network). These three types of connectivity are obviously linked, and their joint analysis allows a better understanding of brain structures and functions [Alexandre 2013].

In Fig. 1.8, these three types of brain connectivity [Sporns 2007] are displayed in different forms. In the lower half, there are the connectivity matrices, where (a) structural connectivity forms a sparse and directed matrix, (b) functional connectivity leads to a full symmetric matrix and (c) effective connectivity yields a full non-symmetric matrix. All these matrices can be weighted, with weights representing connection densities or efficacies for structural connectivity, strength of statistical dependence or proximity between two elements (neurons, recording sites, voxels) of the system for functional connectivity, and quantity of information flow for effective connectivity. Applying a threshold to such matrices will yield binary directed graphs with the setting of the threshold controlling the degree of sparsity, where the binary elements indicating the presence or absence of a connection.

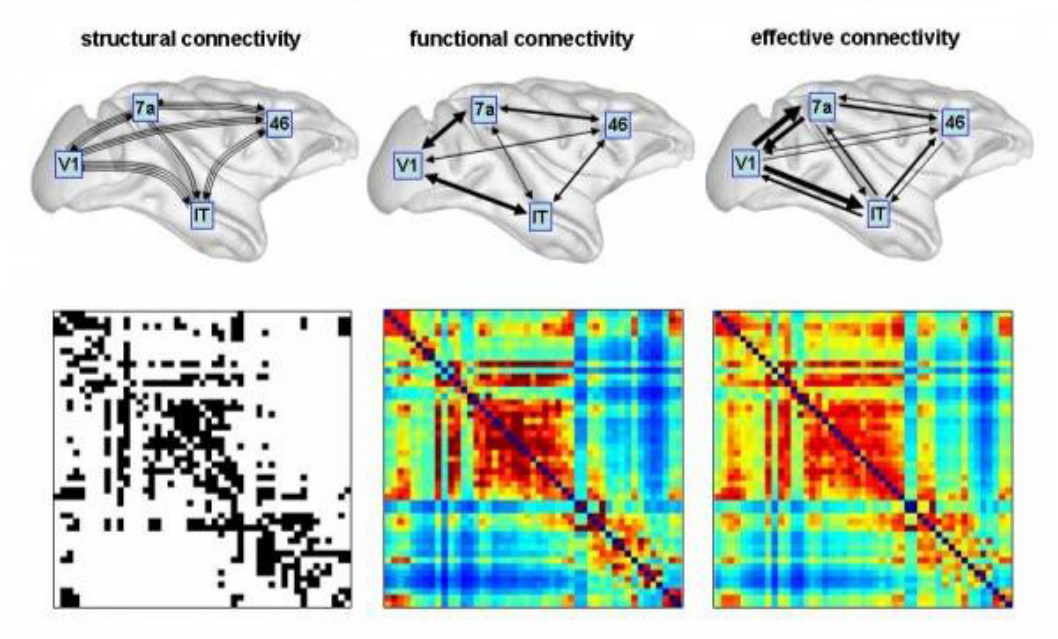


Figure 1.8: Patterns of brain connectivity [Honey 2007, Sporns 2007]. The upper half of this figure illustrates structural connectivity (fiber pathways), functional connectivity (correlations), and effective connectivity (information flow) among four brain regions in macaque cortex. In the lower half of the figure, three different matrices show binary structural connections (left), symmetric mutual information (middle) and non-symmetric transfer entropy (right), respectively.

In this thesis, we essentially focus on functional and effective connectivities.

1.2.2.1. Structural Connectivity

Several studies [Bassett 2006, He 2007, Alexandre 2013] on the structural network have already shown the existence of a small-world network, at several levels, with the presence of clustered areas and long connections that permit rapid communication between these areas. This type of organization allows a compromise between structural cost and

processing efficiency. A good understanding of the structural contributions of individual areas allows us to identify and classify the brain network hubs, which are defined as highly connected brain regions, including areas of parietal and prefrontal cortex. However, due to various reasons (measurement noise or variability), currently, a considerable number of details on the structural networks of the human cerebral cortex are still not clear [Sporns 2005].

1.2.2.2. Functional Connectivity

Functional brain network is another important issue, which is concerned by a large number of studies. The functional connectivity refers to the connectivity between brain regions that share functional properties, and can be defined as a temporal correlation between spatially remote neurophysiological events [Biswal 1997]. The functional connectivity is particularly sensitive to learning, which is based primarily on Hebb's rule [Hebb 2005]. Several studies show that small-world networks facilitate the synchronization of the entire network [Wang 2002]. Additionally, the functional brain network suggests that small-world attributes possibly reflect the underlying structural organizations of anatomical connections [Achard 2006].

1.2.2.3. Effective Connectivity

Effective connectivity describes the influence from one neuronal system (such as individual neuron, neuronal population, or anatomically segregated brain region) to another one, and possibly reflects the causal relation between different systems [Lang 2012]. The study of effective connectivity helps us in the analysis of the dynamic information processing. This can be done by estimating linear/nonlinear, temporal/frequential theoretical correlation indexes. For instance, in cognitive tasks, we are mainly interested in studying the organization of the information flows between different brain regions. Generally, these studies rely on a combination of several techniques, such as fusion of anatomical atlases, electroencephalography or functional magnetic resonance imaging (fMRI) data (for locating activation regions) and transcranial magnetic stimulation to disrupt the network during a task and observe the reactions of the volunteers [Alexandre 2013].

1.3. Problem Statement

As mentioned above, epilepsy is a neurological disease that affects approximately 1% of the population and is characterized by the repetition of seizures (called ictal periods) whose frequency and duration are variable. Epilepsies remain drug-resistant in about 30% cases, which can require a surgical operation.

For a given patient, the key-point is to determine or confirm the accurate boundaries

of the "epileptogenic zone" (EZ), which define the brain areas to be eventually surgically resected to achieve freedom from epileptic seizures, and make sure that, if these areas are partially or completely removed after surgery, the post-surgical deficits (such as vision or language) remain limited. However, such a surgery is challenging, because, in most cases, the epileptogenic zone, which is responsible for the initiation and the propagation of seizure activities, corresponds to a network of neuronal ensembles distributed in distant structures. Due to this distributed characteristic, the localization and characterization of the epileptogenic networks are difficult tasks but they are crucial to help in delimiting the cerebral volume to be resected to avoid epileptic activity.

The long-term objective is to help to understand the mechanisms of the seizure in order to cancel it or at least to stop it. This implies to address the following issues: What are the structures that are involved in such activities? What is the effective connectivity between these structures? Is it possible to detect dominant structures in this epileptogenic network?

Compared to functional connectivity, which underlies the notion of coupling between different structures, the issue of effective connectivity involves the information of directionality and is more challenging. The simplest investigation is the one which detects and quantifies effective connectivity marginally for all oriented pairs of iEEG signals. A more accurate investigation is the one which tries to characterize the directed connectivity from one channel towards another one conditionally to the contextual information provided by the other iEEG channels.

This second approach can be implemented through multichannel causal modeling, for instance VAR modeling, if we limit the causality characterization to causal linear effects. Now, if it is necessary to take nonlinear effects into account, the conditional statistical causal analysis may have to face identifiability condition and computational load.

Chapter 2

State of the Art

2.1. Introduction to Effective Connectivity

During the past several decades, there has been an increasing interest in the research of human brain structure and functions. However, our knowledge on the effective connectivity, which could provide us a deep understanding of the brain, is still poor [Petkov 2015].

The concept of effective connectivity firstly appeared in the work of Aertsen and Preissl [Aertsen 1991]. Shortly later, this concept was studied by several authors on rat/human projects [McIntosh 1991, McIntosh 1992, Grafton 1994, McIntosh 1994, Friston 1994]. So far, the term "effective connectivity" has been defined by various authors [Horwitz 2003], and usually, it is understood as a measure of the impact of one neural system on another, directly or indirectly [Friston 1994]. Therefore, one important feature of effective connectivity is that it contains the information on the directionality of causal influence. More specifically, if an observation of the neuronal activity in one brain region allows us to get a better prediction on the neuronal activity in another brain region, then it is said that the former region has an influence on the latter [Lang 2012]. Fig. 2.1 illustrated the different possibilities of the connectivities among three neural populations [Wendling 2000].

The new causality indexes proposed in this thesis correspond to (semi)nonparametric entropic methods. To position this work among other possible approaches, we give in section 2.2 a list of methods developed in more or less recent literature to address this topic.

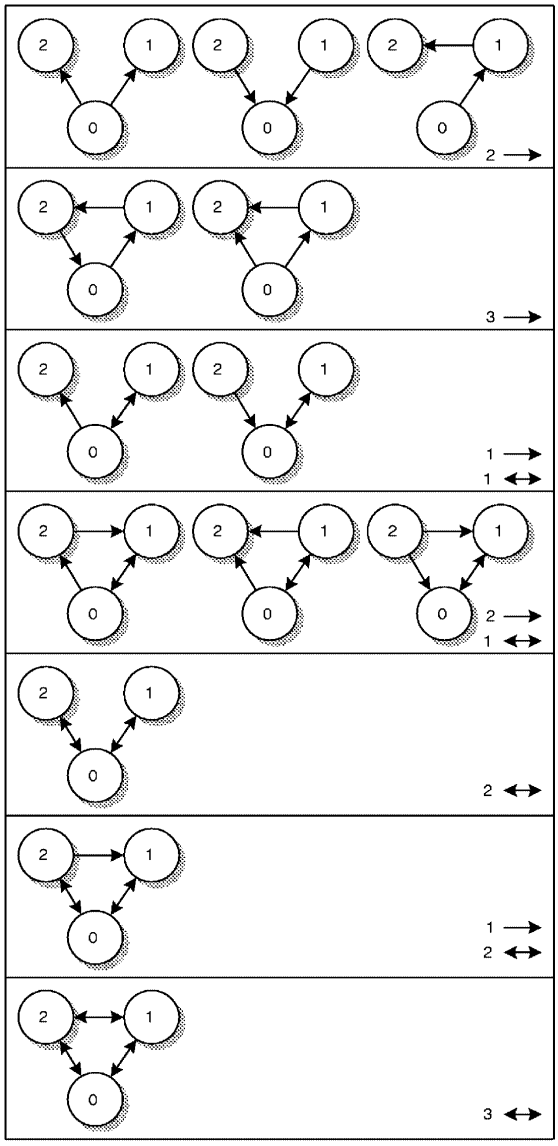


Fig. 2. Building systems from defined underlying organizations. This example gives the possible ways to interconnect three neural populations from unidirectional couplings (*single-sided arrow*) and bidirectional couplings (*double-sided arrows*)

Figure 2.1: Possible connectivities among three neural populations, both unidirectional couplings (single-sided arrow) and bidirectional coupling (double-sided arrows) [Wendling 2009].

2.2. State of the Art

The measurements of effective connectivity can be categorized into two classes: data-driven methods, including Granger causality related [Bressler 2011] and information-theoretic approaches [Hlaváčková-Schindler 2007] (see sections 2.2.1 and 2.2.3), and model-based methods, such as dynamic causal modeling (DCM) [Friston 2003] (see section 2.2.2).

temporal gyrus) structures. The position of each electrode is strategically defined from the analysis of surface EEG signals, clinical data and imaging in order to record structures that play a potential role during the seizures.

3.1 Influence of the excitation/inhibition loop

A Gaussian white noise was used as the model input, $p^n(t)$. The mean and variance (corresponding to a rate of 30–150 pulses/s) were adjusted so that the model produced a signal similar to the spontaneous EEG recorded from neocortical structure electrodes during interictal periods when other parameters of the model are set to standard values (Table 1). This signal (Fig. 3a) reflects a normal activity which resembles that reflected by real SEEG signals (Fig. 3f). Starting from this point,

all parameters were kept st
constant in the excitatory lo
increased. This increase (F
that appear sporadically (fro
(A = 3.6) and finally rhy
average frequency of this rh
increases as A is increased (
Visual inspection of real si
spikes (usually appearing bef
rhythmic discharges of spik
seizures start) can be encou
(Fig. 3g–i). For this type o
that real waves are quite ac
model. From A = 8.5, the r
activity to sinusoidal activit
lope also depends on the val
activity may be reflected
seizures (see the first 5 s of
described below). The exci
also be adjusted to position
between two types of activity
transitions occur (Fig. 4).
presents a real SEEG signal
the beginning of a tempora
SEEG signal shows high an
faster activity (quasi-sinusoi
slows down to resume spi
average frequency. Figure 4b
this period of time. A spon
quasi-sinusoidal activity and
One notices that the simula
tively as the real one; the s
amplitude and higher frequ
frequency, higher-amplitude

3.2 Influence of the coupling

A key advantage of the mult
ability to produce signals r
between underlying populati
easily controlled via the str
the direction of couplings (uni
Figure 5 illustrates the effe
simple example in which th
ered. The first one was mad
(i) its excitation/inhibition r
other populations remained
couplings between these thre
and (ii) signals correspond to norma
sporadic spikes appearing
population and only norma
the two others. When para
senting the unidirectional co
population 2 and the uni
population 2 to popul
(K^{12} = 200, K^{13} = 200), th
from population 1 to popul
simulated (Fig. 5a). As expe
delays, spikes in populations
population 1. Now, introdu

2.2.1. Time Domain Wiener-Granger Method

2.2.1.1. Linear Granger Causality

The notion of "causality" has been discussed by many authors [Friston 1994, Pearl 2009, Valdes-Sosa 2011, Roebroeck 2011a, Roebroeck 2011b], the most conventional meaning of the word "causal" could be "a cause occurs prior to its effect" [Amblard 2012].

The basic idea of Granger causality (GC) can be traced back to the work of Wiener [Wiener 1956] and Granger [Granger 1969]. It measures the statistical dependence between the past of a process and the present of another one. Precisely, given two time series X and Y, including the lagged value of X, if the lagged value of Y provides statistically significant information about the future value of X (through a series of t-test and F-test [Greene 2003]), we can say that Y Granger-causes X. This concept was later generalized by Geweke [Geweke 1982, Geweke 1984], and the index of causality from Y to X was defined as the logarithm of the ratio of the asymptotic mean square error when predicting the future value of X from its own past, to the asymptotic mean square error when making the same prediction from the lagged value of both X and Y.

Granger causality was developed originally in econometrics, and nowadays, it is one of the most popular measures to infer causal interactions between time series. Due to its simplicity and effectiveness, it has been widely used in neuroscience, including both fMRI [Roebroeck 2005, Sato 2006, Deshpande 2009, Deshpande 2010b, Seth 2013] and EEG/MEG [Hesse 2003, Gow 2008, Ploner 2009, Gow 2009, Adhikari 2013]. Besides neuroscience, it has also numerous applications in a variety of research areas including climate [Kaufmann 1997, Triacca 2001, Mosedale 2006, Smirnov 2009], engineering [Kim 2012, Yuan 2014], energy [Pao 2011, Bozoklu 2013, Tiwari 2014], economics [Comincioli 1996, Granger 2000, McCracken 2007, Chiou 2008], and others [Seth 2007].

Since its introduction in 1969 [Granger 1969], discussions on this causal measure never end, and several authors dedicated their research to its implementation [Cui 2008, Seth 2010, Barnett 2014]. In 1984, Geweke [Geweke 1984] introduced the concept of conditional GC. Recently, Ding *et al.* [Ding 2006] demonstrated that pairwise GC cannot distinguish some specific situations, where there are joint dependencies between X and Y and a third set of time series Z (as illustrated in Fig. 2.2), and they defined conditional Granger causality (also called partial Granger causality in the literature [Guo 2008a, Krishna 2008, Guo 2008b, Wu 2012]) to determine whether the causal influence between two channels is direct or mediated by another group of channels. In most applications, Granger causality is used to analyze the relation among multiple channels, which explains the expanding popularity of conditional Granger causality [Zhou 2009, Liao 2010, Gao 2011, Roelstraete 2012, Detto 2013].

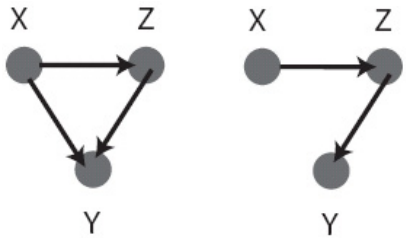


Figure 2.2: Granger causality fails to distinguish these two situations: (a) there is direct influence from X to Y, (b) the influence from X to Y is mediated by Z, which is indirect.

In 2010, Angelini *et al.* [Angelini 2010] indicated that the redundant variables in multivariate Granger causality would lead to under-estimated causalities. In [Barnett 2011], Barnett and Seth discussed the behavior of Granger causality under filtering, and drew the conclusion that Granger causality of a stationary vectorial autoregressive (VAR) process is fully invariant when the data are preprocessed with an arbitrary invertible filter. As pointed out in [Hu 2011], Granger causality is not a perfect measure, and it fails to determine the real strength of causality among channels in certain simulations. Barrett *et al.* [Barrett 2013] gave a deeper discussion to this issue: Granger causality cannot work perfectly with all kinds of signals, and if the observed data are not obtained with straightforward autoregression process, then this causality is only an approximate measure of the real causal influence, or in other words, Granger causality is designed to measure the effect but not the mechanism. In 2013, Davey *et al.* [Davey 2013] delineated the equivalence of linear Granger causality connectivity methods and correlation-based connectivity methods, such as partial correlation (PC).

To sum up, Granger causality, also termed Wiener-Granger causality (WGC) [Bressler 2011], is usually operationalized in the context of VAR model theory [Hamilton 1994, Lütkepohl 2005], thus it reflects only the linear features of the signals. Additionally, it also assumes that the analyzed signals are covariance stationary [Seth 2007]. During the past several decades, Granger causality has been greatly developed and extended in numerous studies [Pereda 2005], some of its extensions being recalled in the following sections.

2.2.1.2. Nonlinear Extensions of Granger Causality

In the standard application of WGC, only the linear features are captured whereas many target systems are known to be nonlinear. Therefore, the exploration of nonlinear information transmission could provide additional information about the system [Marinazzo 2011]. To this end, several studies have been conducted to extend WGC to nonlinear situations.

One possibility of extending linear GC to nonlinear situations is to replace linear autoregression with other nonlinear prediction schemes. The local linear nonlinear autoregressive model (LLNAR) is a generalization of linear autoregression. The nonlinearity of LLNAR is controlled by a parameter, the bandwidth, and LLNAR reduces to the ordinary autoregressive model when the bandwidth tends to infinity. The flexibility of LLNAR in describing nonlinear characteristics of EEG signals has been demonstrated by Hernández *et al.* [Hernández 1996]. In 1999, based on LLNAR, Freiwald *et al.* [Freiwald 1999] introduced a generalized definition of GC, which is valid for both linear and nonlinear systems. With this method, Freiwald *et al.* successfully revealed the existence of both unidirectional and bidirectional influences between neural groups in the macaque inferotemporal cortex.

Another important nonlinear extension of Granger's idea was proposed by Chen *et al.* [Chen 2004] in 2004, and named extended Granger causality index (EGCI). The basic idea of EGCI is that, even in nonlinear systems, one can locally approximate the dynamics linearly, and so apply linear GC to each local neighborhood and get average statistical quantities that properly reflect the nonlinear dynamics. In the experiments presented in [Chen 2004], the effectiveness of EGCI has been proved on artificial data. The same year, Ancona *et al.* [Ancona 2004] proposed a nonlinear extension of Granger causality for bivariate time series, further called nonlinear Granger causality (NLGC) in [Sun 2008] and [Ishiguro 2008]. In NLGC, a nonlinear kernel autoregression scheme is employed, instead of a linear autoregression one. Specifically, Ancona *et al.* [Ancona 2004] argued that not all nonlinear prediction schemes are suitable for the extension of linear GC. These schemes should follow the following property: given two time series X and Y, if Y is statistically independent of X, then the variance of prediction error obtained when predict the future of X with the past information of both X and Y, should be equal to the one predicted with only the past information of X itself, and analogously for the opposite direction. It should be mentioned that, according to this property, the EGCI method proposed in [Chen 2004] satisfied the above property only if the number of points in the local neighborhood, where linear regression is performed, is sufficiently high to obtain reliable statistics, but this assertion is in contradiction with the basic idea of local linearization [Ancona 2004].

In 2008, based on the theory of reproducing kernel Hilbert spaces (RKHS) [Shawe-Taylor 2004], a kernel version of Granger causality (KGC) was proposed by Marinazzo *et al.* [Marinazzo 2008b] to detect nonlinear cause-effect relationship. In this method, the linear GC was performed in the feature space of suitable kernel functions (Gaussian kernels and inhomogeneous polynomial (IP) kernels as those presented in [Marinazzo 2008b]). KGC assumes arbitrary degree of nonlinearity, *i.e.* it is able to handle all orders of nonlinearity [Marinazzo 2011]. This measure was also generalized to the multivariate

case to analyze dynamical networks [Marinazzo 2008a]. As pointed out by the authors, the multivariate version of this approach is still able to reveal the real causalities while complexity of the model increases. Shortly after its introduction, KGC was adopted in the analysis of fMRI data [Liao 2009]. Using both simulated and real fMRI data, this newly proposed approach was shown to capture the effective coupling which was missed by linear GC. The effectiveness of the multivariate KGC was validated by different studies. In [Angelini 2009], Marinazzo's method was applied on several physical systems, and the results showed that the information flow was properly identified. Stramaglia *et al.* [Stramaglia 2011] tested the same approach on EEG data to reveal nonlinear interaction. In [Liao 2011], Liao *et al.* used multivariate KGC and graph theory on resting-state fMRI recordings to reveal the network architecture of the brain network.

In [Bezruchko 2008], Bezruchko *et al.* proposed to use a polynomial of degree p instead of the linear autoregressive model to capture nonlinear causality. This method can be viewed as a special case of KGC with IP kernel functions. The well-known NARX (nonlinear autoregressive with exogenous inputs) model [Billings 2013] is another extension of the linear model, which could be used as the basis of the nonlinear Granger analysis. This possibility has been firstly explored by Li *et al.* [Li 2012]. Recently, in 2013, Zhao *et al.* [Zhao 2013] went further with this idea: different indexes were designed to investigate the linear and nonlinear causalities between time series separately. Moreover, it should be noted that, compared to all the nonlinear extensions of GC mentioned above, these methods based on NARX are time-variant, which is supposed to provide extra important insights into the signals. In both [Li 2012] and [Zhao 2013], the usefulness of their methods was demonstrated with real EEG signals.

We conclude this subsection with a brief presentation of the h^2 index introduced in neuroscience by [Pijn 1990] in the scope of Wiener-Granger causality. This index can be considered as a nonlinear correlation index. It extends the linear r^2 index defined as the square of the Pearson correlation coefficient between $X(t)$ and $Y(t-\tau)$, maximized with respect to τ. This index is also equal to one minus the ratio of the estimated linear mean squared prediction error (when predicting $X(t)$ from $Y(t-\tau^*)$ where τ^* is the optimal lag) to the estimated variance of $X(t)$. If τ is constrained to be strictly positive, $\log\left(1-\left(1/r^2\right)\right)$ can be considered as a degenerate Granger causality index from Y to X where the past of X is reduced to an empty set and the past of Y to $Y(t-\tau^*)$, *i.e.* to only one past scalar value. Now, this scalar value is optimally selected among the values $Y(t-\tau)$, $0<\tau\leq\tau_{\max}$, and so r^2 implicitly depends from all past values in an authorized set specified by $\tau_{\max}$. Clearly, if τ is constrained to be strictly negative, then $\log\left(1-\left(1/r^2\right)\right)$ quantifies the reverse causality, from X to Y. To obtain the h^2 index, the mean square linear prediction error in the expression of r^2 is replaced by the mean square nonlinear regression error when estimating $X(t)$ from $Y(t-\tau)$, maximized by

tuning τ. The regression function is estimated from the observed values of X and Y, and is constrained to be a piecewise affine function on a given partition (union of intervals) of the real line. Finally, when τ^* is selected among the authorized strictly positive values, h^2 can be interpreted as a degenerate nonlinear Wiener-Granger causality index. Compared to the Wiener-Granger approach, even if the past values of X are not considered in the h^2 index, the low dimensionality of the corresponding statistical inference algorithm can lead to a not too spread estimation of the regression error variances, in comparison to more complex nonlinear regression algorithms. This can partly explain its practical efficiency, for instance when it is used to analyze real epileptic signals as in [Caparos 2006, Dorr 2007].

This subsection was devoted to give an overview of nonlinear non-entropic methods that are not considered in the rest of the document.

2.2.2. Spectral Methods

In the past few years, several frequency-domain causality detection methods have been introduced to analyze the directional connectivity in neural systems [Chicharro 2011, Hu 2012]. With these spectral measures, it becomes possible to detect the frequency band(s) where causality occurs [Brovelli 2004, Bressler 2008, Ladroue 2009], which is *a priori* interesting for iEEG signals analysis.

Most spectral methods of causality are related to the spectral form of Granger causality based on transfer functions derived from multivariate autoregressive modeling. As a matter of fact, a great advantage of Granger causality is that it could be decomposed in the frequency domain [Geweke 1982, Geweke 1984]. Compared to its spectral form, this quantity in the time domain may be considered as an average over all frequencies (up to the Nyquist frequency). In other words, given two time series X and Y, the spectral Granger causality from Y to X at frequency ω quantifies the directed contribution of Y to the power of X at frequency ω [Chen 2006, Barrett 2014].

To our best knowledge, the first spectral decomposition of linear GC was proposed by Geweke [Geweke 1982, Geweke 1984], and then easily extended to other different forms: (i) conditional spectral Granger causality, which calculates the causality contribution of Y to the power of X, in addition to the causality contribution of a third group of variables to the power of X [Chen 2006], (ii) multivariate form, which computes the causality contribution from one group of variables to another group [Barrett 2010]. Since its conception, this spectral measure has generated interest among scientific researchers. For example, Bernasconi and König have used this method to find neuronal interactions in the visual cortical areas of the cat [Bernasconi 1999, Bernasconi 2000]. In 2004, Brovelli *et al.* used it to find the Beta oscillatory networks in monkey sensorimotor cortex [Brovelli

2004]. Chen *et al.* used its conditional form to analyze neural field potential time series [Chen 2006]. In 2013, Croux and Reusens [Croux 2013] used this method to investigate the predictive power for the future domestic economic activity that is contained in the domestic stock prices. In [Epstein 2014], to investigate the features of preictal seizure networks, this spectral method was employed for the analysis of iEEG signals recorded at high frequencies (500 or 1000 Hz).

In 2014, one important generalization of spectral GC was introduced by He *et al.* [He 2014]. This method uses the nonlinear NARX [Billings 2013] signals modeling, and can be considered as the spectral decomposition of the nonlinear GC described in [Li 2012, Zhao 2013]. In their experiments, He *et al.* validated the effectiveness of the proposed method with both artificial data and a real human intracranial EEG data set.

In 1975, the concept of causality was also introduced in the study of feedback relations between input and output variables by Caines and Chan [Caines 1975], and generalized later to the multivariate situation by Kamiński and Blinowska [Kamiński 1991] who proposed a new spectral causality measure, the directed transfer function (DTF). Using a multivariate autoregressive model to fit multi-channel signals, the DTF method, which is based on spectral domain functions, is designed to detect the directional influences between any given pair of signals in a multivariate data set. It has been demonstrated that the DTF function could be interpreted in terms of GC [Kamiński 2001]. After its introduction, DTF has been widely investigated, especially in the analysis of EEG signals, for the purpose of localizing epileptogenic foci in patients with partial epilepsy [Ding 2007, Dorr 2007, Wilke 2010, Lu 2012].

In some special cases, for instance cognitive experiments, the signals length may be too short [Ding 2000] to provide good estimates, and DTF becomes ineffective with such data. To overcome this issue, the repetition of the task can improve the final estimate [Ding 2000, Philiastides 2006], and/or the introduction of an important overlapping of the processed windows can allow a better estimation, as is the case in short time DTF [Ding 2000]. Using this technique, data are processed by highly overlapped time windows, and the multivariate autoregressive (MVAR) modeling coefficients are adapted to each window. The window's length must be short enough to consider local stationarity. Ding *et al.* applied this method to a visuomotor integration task, and revealed rapidly changing cortical dynamics. In [Liang 2000], short-time DTF was used to detect causal influence between cortical sites on a fraction-of-a-second time scale. In 2006, Philiastides and Sajda [Philiastides 2006] used this method for the analysis of multi-channel EEG data recorded during a face discrimination task.

Standard DTF algorithm assumes that the signals are stationary. The short-window DTF mentioned above introduces a windowing technique so as to consider local sta-

tionarity, while Wilke *et al.* [Wilke 2007, Wilke 2008] provided another solution, termed adaptive DTF (ADTF), where the coefficients of the multivariate adaptive autoregressive (MVAAR) model are time-dependent. Tested on both simulated and real electrocorticogram (ECoG) data, ADTF appeared more suited for the detection of time-variant connectivities than the standard DTF.

For the time-varying causality detection algorithm mentioned above, the non-zero covariance of the model's residual was used to describe the causality phenomenon. However, in some situations, for example, the blurring of the neuronal activity with the sluggish latent response, non-zero covariance of model residuals can be observed. This zero-lag correlation would lead to spurious time-lagged causality detection [Deshpande 2010a]. To solve this problem, Xu *et al.* [Xu 2014] introduced another adaptive extension of DTF, termed time-lagged adaptive DTF. Compared with the adaptive algorithm proposed in [Wilke 2008], Xu's method employed a more general adaptive model, and took the influence of instantaneous connectivity into consideration. The effectiveness of this method has been demonstrated through the detection of dynamic spectral causality in real visual evoked potentials (VEPs) data.

Another multivariate measure was proposed by Sameshima and Baccalá [Sameshima 1999, Baccalá 2001], called partial directed coherence (PDC). Contrary to DTF, PDC measures the direct causality between two channels discounting the influence of all other recorded channels, while DTF captures the influence of the whole network by taking all the channel transmission pathways into consideration [Sameshima 2014]. Both DTF and PDC assume that the signals are stationary, and for non-stationary data, it could be divided into windows to obtain approximate stationarity [Bressler 2011].

In 2007, Baccalá introduced a scale-invariant form of PDC, the generalized partial directed coherence (GPDC) [Baccald 2007]. The GPDC measure overcomes the drawbacks of PDC (as discussed in [Schelter 2009]): (i) GPDC is not affected by the relation between the given source and a third group of channels, (ii) GPDC is scale invariant, (iii) GPDC allows the analysis of the absolute strength of coupling.

Since its introduction in 2001, PDC has interested a lot of researchers for multiple applications. With both simulated and real data, Astolfi *et al.* [Astolfi 2005, Astolfi 2006] tested PDC for the estimation of human cortical connectivity. In [Schelter 2006], Schelter *et al.* discussed the statistical properties on the estimation of PDC, and tested PDC on EEG and EMG data recorded from an essential tremor patient. Sato *et al.* [Sato 2009] applied PDC for connectivity analysis of multisubject fMRI data using multivariate bootstrap. In [Sun 2009], PDC was used to study the cortical connective network under audiovisual cognitive processes. In [Wang 2015], PDC was used for automatic epileptic seizure detection.

Both standard PDC and GPDC are based on a MVAR model, which only describes the lagged effects among different signals. To overcome this limitation, Faes and Nollo [Faes 2010] proposed an extended version of PDC (EPDC) based on the utilization of an extended MVAR model including both instantaneous and lagged effects. The authors proved the effectiveness of EPDC with two different EEG data sets.

Recently, Omidvarnia *et al.* [Omidvarnia 2012, Omidvarnia 2014] have extended the classical PDC connectivity analysis, and introduced a novel measure, called generalized orthogonalized PDC (gOPDC). In the development of gOPDC, the coefficients of MVAR were orthogonalized. This new measure was supposed to be less affected by the volume conduction and amplitude scaling, and evaluated on both simulated models and newborns' EEG signals.

2.2.3. Model-based Methods

As mentioned above, Granger causality is data-driven and can be estimated without any *a priori* physiological hypothesis. On the contrary, model-based causal measures involve neural population models and take physiological evidences into account.

An important model-based approach was first proposed by Friston [Friston 2003], and called dynamic causal modeling (DCM). The basic idea of DCM is to consider the brain as a dynamic input-state-output system with multiple inputs and outputs, and construct an explicit forward or generative model. In some situations, different dynamic causal models for a system of interest can be constructed, standing for different competing hypotheses about the mechanisms that generate the observed data. In this case, considering both model fitting and relative complexity, Bayesian model selection can be used to identify the most optimal DCM [Penny 2004].

Shortly after its introduction by Friston in 2003 [Friston 2003], this nonlinear and dynamic method has obtained numerous attentions, and several improvements and extensions have been brought to DCM for specific applications, such as fMRI [Kiebel 2007, Stephan 2007, Marreiros 2008, Stephan 2008] and EEG/MEG [Kiebel 2006, Chen 2008, Marreiros 2009, Moran 2009, Daunizeau 2009, Penny 2009].

Both aiming at detecting causality, DCM and GC have been compared in the literature [David 2008]. However, there is fundamental difference between these approaches: DCM employs a biophysical realistic model and focuses on how the observed data are generated, while GC examines for statistical dependencies between different physiological responses [Friston 2009]. Compared with GC, the important number of parameters to estimate in DCM may limit its practical use [Seth 2010]. Now, Friston *et al.* [Friston 2014] demonstrated that the spectral Granger causality can be unreliable on noisy data,

and also showed that this problem can be finessed by deriving spectral causality measures based on the parameters estimated using dynamic causal modeling.

2.2.4. Information Theory Measurement

Largely spread since 1948 [Shannon 1948], information theory concerns quantities that measure the uncertainty in random variables based on a probabilistic concept, and it has been applied in numerous fields, including neuroscience [Dimitrov 2011]. Using information theory based approaches, we can measure the amount of uncertainties reduced in one random process after observing another one. Compared with the methods mentioned above, such as GC or DCM, information-theoretic quantities assume almost no *a priori* information on data modeling and may capture both linear and nonlinear relations between time series [Jin 2010], as illustrated in papers such as [Vicente 2011]: these techniques are fully "model-free".

One well-known information-theoretic quantity is mutual information (MI) [Cover 2012], which is widely used to quantify the amount of uncertainty shared between two random variables. Now, the drawback of this kind of measure is its symmetrical nature, so that it only reflects some unidirectional dependence. Thus, to detect the directed information flow between two variables, the time-lagged form of mutual information (TLMI) (also termed time-delayed mutual information (TDMI) [Na 2002]) was introduced [Hlaváčková-Schindler 2007], which estimates the shared information between one process and the lagged version of another one. Several authors have applied TLMI to the analysis of EEG signals [Jeong 2001, Na 2002, Min 2003]. However, as demonstrated by some authors [Schreiber 2000, Kaiser 2002], TLMI sometimes fails in revealing the actual information flow.

In 2000, Schreiber [Schreiber 2000] introduced an information-theoretic statistic measurement, named transfer entropy (TE), to quantify the amount of time-delayed information between two dynamical systems. Given the past time evolution of a dynamical system $\mathcal{A}$, transfer entropy from another dynamical system $\mathcal{B}$ to the first system $\mathcal{A}$ is the amount of Shannon uncertainty reduction in the future time evolution of $\mathcal{A}$ when including the knowledge of the past evolution of $\mathcal{B}$. As demonstrated in [Schreiber 2000, Kaiser 2002], compared with TLMI, TE is capable to distinguish the information exchanged from shared information due to common history and input signals. Like Granger causality, this method can be extended so as to get its conditional form [Wibral 2014b] to exclude the influence of a third group of variables. Transfer entropy, as classical Granger causality, does not reveal instantaneous coupling. This coupling is taken into account in the more general notion of directed information (DI) [Amblard 2012].

As an information-theoretic implementation of Wiener's principle of observational

causality, TE is related to some other causality measures. As known, GC emphasizes the concept of "prediction", while TE is framed in terms of "resolution of uncertainty". So, TE can be considered as a measurement of the degree to how Y disambiguates the future of X beyond the degree to how X disambiguates its own future [Paluš 2001]. The work by Barnett *et al.* [Barnett 2009] specified the relation of these two causality measures for the first time, and bridges the information-theoretic and autoregressive methods. Under Gaussian assumptions, TE and GC are entirely equivalent, up to a factor of 2. Therefore, Granger causality can be understood as a linear approximation of transfer entropy [Hlinka 2013]. The Gaussian assumption seems to be strict for many biological and physical mechanisms. Based on the results of Barnett *et al.* [Barnett 2009], Hlaváckova-Schindler investigated how the equivalence of the two causality measures can be extended under some conditions on probability density distributions of the data, and generalized this relation to other common types of distributions [Hlaváčková-Schindler 2011]. Later, with a very general class of continuous or discrete Markov models, Barnett and Bossomaier [Barnett 2012] extended TE to a log-likelihood ratio in a maximum likelihood framework.

It has been indicated that [Paluš 2001, Hlaváčková-Schindler 2007] TE is actually an equivalent expression for conditional mutual information with the history of the influenced variable in the condition. In [Chicharro 2011], Chicharro discussed the spectral decomposition of several causality measures, however, he argued that, as most general information-theoretic measures related to GC, TE lacks a spectral representation in terms of the recorded processes. Standard transfer entropy may be sometimes defective. As a matter of fact, when estimating the causality between two systems, according to the definition of TE, a non-zero value of TE is a clue of causality relation. However, for unidirectional relation, it is rather possible to obtain non-zero TE value for both directions, which would cause spurious conclusions. In [Smirnov 2013], Smirnov discussed the following three typical factors leading to such phenomenon (i) unobserved state variables of the driving system, (ii) low temporal resolution, and (iii) observation errors.

Due to its information-theoretic background, TE is model-free (inherently nonlinear) and has already shown its superiority in the detection of effective connectivity for nonlinear interactions [Nichols 2005, Lungarella 2007a, Vicente 2011]. Also, several open source tools are available for its practical use [Wibral 2011b, Lizier 2014, Montalto 2014]. Thus TE obtains wide applications in various fields, including neuroscience [Chávez 2003, Gourévitch 2007, Garofalo 2009, Vakorin 2009, Sabesan 2009, Vakorin 2010, Besserve 2010, Buehlmann 2010, Lüdtke 2010, Wibral 2011a, Lizier 2011, Neymotin 2011, Vakorin 2011, Roux 2013, Pampu 2013, Wibral 2014a, Marinazzo 2014, Faes 2014, Lobier 2014], climate [Pompe 2011], complex system theory [Lizier 2008, Lizier 2010], physiology [Faes 2006, Faes 2011, Faes 2012], economics [Kwon 2008, Kim 2013], and other fields [Materassi

2007, Ver Steeg 2012].

After its introduction by Schreiber in 2000, TE gained its popularity immediately, and many researchers have greatly contributed to its development. Hereafter, are synthesized some important extensions of this approach.

In 2007, Lungarella *et al.* [Lungarella 2007b] extended transfer entropy into a wavelet-based form to measure directional transfer of information between coupled systems at multiple time scales. With this method, the time series were projected into the wavelet space to obtain a new set of variables, and then the causal dependencies between the new variables are extracted. Due to their experimental results, this wavelet-based extension of transfer entropy is capable of analyzing signals at multiple scales, even in the presence of nonstationarities, and it also succeeds in detecting the scale-dependent causal dependencies between coupling systems.

Another important extension of TE is based on symbolizing technique. This technique first comes from the concept of permutation entropy [Bandt 2002a, Bandt 2002b], where the symbolic presentation of the time series determined from ordering the amplitude values instead of these time series themselves is used to reduce noise contributions in observed data. The symbolizing technique could be summarized as follows [Melzer 2014]: start from a time series $X = \{x_i\}$ with value x_i for the ith sampling point. Given an embedding dimension k and a delay τ, the amplitude values are combined to $\{x(i), x(i+\tau), \ldots, x(i+(k-1)\tau)\}$, giving a sequence of length k for the ith point. These sequences are then sorted in ascending order $\{x(i+(t_{i1}-1)\tau) \leq x(i+(t_{i2}-1)\tau) \leq \ldots \leq x(i+(t_{ik}-1)\tau)\}$. A symbol then denotes the order indices $\widehat{x}_i \triangleq (\widehat{t}_{i1}, \widehat{t}_{i2}, \ldots, \widehat{t}_{ik})$ thus mapping this sequence onto one of the possible $k!$ permutations of the number 1 to k, reflecting the successive order of the amplitude values. Staniek and Lehnertz [Staniek 2008] introduced this technique into the estimation of TE and proposed a new method called symbolic transfer entropy (STE). Compared with transfer entropy, STE is computationally faster and more robust to noise [Staniek 2008]. In [Staniek 2009], Staniek and Lehnertz gave another discussion on this measure of directed interactions, and demonstrated its performance in the analysis of human epileptic brain. Melzer and Schella [Melzer 2014] applied STE to analyze the behavior of changed-particle systems, and revealed the information transportation successfully.

In 2008, Bettencourt *et al.* [Bettencourt 2008] demonstrated that, analogous to a Taylor series, mutual information between a stochastic variable and a set of other variables, could be expanded into a sum of information quantities, which allowed to obtain a new view of high-order correlations. Based on this work [Bettencourt 2008], Stramaglia *et al.* [Stramaglia 2012] proposed a formal expansion of transfer entropy, involving irreducible sets of variables which provided information for the future state of the target.

Also, for practical applications, the authors argued that a conspicuous amount of phenomenology in the brain can be explained by linear models, and introduced the assumption of Gaussianity [Barnett 2009] to obtain computational convenience. The application of this proposed expansion was illustrated to both simulated data and two real EEG data sets.

As pointed out in [Sun 2014], transfer entropy often results in erroneous identification of network connections, especially for time-dependent networks. To break this limit, in 2014, Sun and Bollt [Sun 2014] developed a measure called causation entropy (CSE) to obtain reliable identification of true couplings. CSE could be considered as a generalization of transfer entropy. It measures the extra information flow between two processes in addition to the information already provided by a third group of processes. Through numerical simulations, the authors highlighted the superiority of CSE over transfer entropy, where CSE successfully inferred the real causal relationships while TE gave misinterpretations.

The common definition of transfer entropy involves infinite vectors [Schreiber 2000], which leads to the calculation of infinite-dimensional densities. Also, despite its numerous advantages, transfer entropy has been mostly applied in a bivariate scene, since it is difficult to obtain reliable TE estimation in high dimensions due to the "curse of dimensionality". To overcome this limitation, by embedding TE into the framework of graphical models [Dahlhaus 2000, Eichler 2012], Runge *et al.* [Runge 2012] presented a formula that decomposes TE into a sum of finite-dimensional contributions, called decomposed transfer entropy (DTE). Compared with TE, DTE drastically reduces the estimation dimension which leads to a more reliable estimation, and the graphical model also enables a richer picture of causal interactions. In [Runge 2012], the advantages of this approach were demonstrated on observational climate data (sea level pressure).

Despite all these advantages, it remains a tough task to obtain accurate estimations of the information-theoretic measures while carried out on finite sample length signals, particularly in the field of neuroscience [Pereda 2005, Hlaváčková-Schindler 2007], where getting large amounts of stationary data is still problematic. This issue is largely discussed in the next chapter.

Chapter 3

Methods and Materials

As mentioned previously, the information-theoretic approaches, especially transfer entropy, play an important role in the detection of causality. A common problem is how to obtain an accurate estimation of these information-theoretic quantities, which has been proven to be difficult. More precisely, a well-known problem arises when estimating entropy is the difficulty to lower the estimation bias. This chapter presents previous published work and some proposed improvements concerning this problem, and it is arranged in three parts.

In section 3.1, we introduce the mathematical definition of different information-theoretic quantities, and discuss some similarities in the calculation of mutual information and transfer entropy. Some important non-parametric approaches for the estimation of these quantities are reviewed in section 3.2. With these previous works, several questions are raised, and to answer them, we propose two different improvements of the existing approaches. In section 3.3, we introduce an analytical form of bias for the estimation of individual entropy, and then apply it into the estimation of both mutual information and transfer entropy, where the bias reduction strategies of relation-specific distance are proposed. These strategies vary with different norms, and it should be mentioned that, when the strategy with maximum norm is retained, our results well explain the conclusions drawn by Kraskov *et al.* in [Kraskov 2004], which were only derived from numerical experiments. Additionally, to further decrease the bias estimation of mutual information between two dependent variables, a weighted linear combination of distinct mutual information estimators is introduced. In section 3.4, based on the idea proposed in [Kraskov 2004], we deeply discuss the improvement in the estimation of information-theoretic quantities using the maximum norm, when a (hyper-)rectangle is used instead of a (hyper-)cube. Following two different methodologies [Kozachenko 1987, Singh 2003], we extend the existing kNN (k-Nearest Neighbors) entropy estimators to the rectangular

situation, which results in two other estimators. Applying these new entropy estimators into the calculation of transfer entropy, we present two novel TE estimators. Finally, in section 3.5, we discuss some other issues while applying transfer entropy in the analysis of neural signals, such like order selection. For ease of readability, the links between the concepts and methodologies described in this chapter are summarized and illustrated in Fig. 3.6. In this diagram, a box identified by a number n in a circle is designed by box (n) hereafter.

Note that only the bias analysis is addressed in the theoretical developments proposed in this thesis. The variances estimations are investigated only experimentally in other sections.

3.1. Problem Statement

3.1.1. Introduction to Information-theoretic Quantities

In this section, we give mathematical descriptions of different information-theoretic quantities.

3.1.1.1. Entropy, Joint Entropy and Conditional Entropy

Entropy

Initially defined in statistical physics as being proportional to the logarithm of the number of possible configurations of a physical system (typically a system of molecules), the entropy formalized by Shannon [Shannon 1948] measures the amount of "uncertainty" on what can be observed at the output of a random information source. In other words, it quantifies the amount of information needed in order to specify completely this output. Furthermore, other theoretical measures were proposed by Shannon, as conditional entropy and mutual information, called entropic measures, which extended the practical and conceptual usefulness of source entropy. Besides their applications in theoretical and applied information systems, entropic measures have been applied in various fields, like independent component analysis [Pham 2004], image analysis [Chang 2006], genetic analysis [Martins Jr 2008], speech recognition [Jung 2008], manifold learning [Costa 2004], time delay estimation [Benesty 2007], among others [Brillouin 2013]. As explained hereafter, transfer entropy is a more recently introduced entropic measure.

Given a d_X-dimensional random vector X (*i.e.* a measurable function $X : \Omega \to \mathbb{R}^{d_X}$ defined on some underlaying probability space (Ω, τ, P)). To define the entropy of X, we first assume that the image set $\mathbb{X} = X(\Omega)$ (the set of possible realizations of X in $\mathbb{R}^{d_X}$) is finite or enumerable, *i.e.* $\mathbb{X} = \{x_i, i \in I\}$ where I is countable. Then, if we consider that

$\log\left(\frac{1}{P(X=x_i)}\right)$ is the decrease of uncertainty (*i.e.* the information gain) resulting from the observation of the value x_i, the entropy $\mathcal{H}_{\text{dis}}(X)$ associated to the random vector X is defined as the expectation (computed with respect to the probability distribution of X on $\mathbb{X}$) of this information gain. If we denote the function $x \to P(X = x)$ on $\mathbb{X}$ by $p_X(x)$, this expectation can be written

$$\begin{aligned} \mathcal{H}_{\text{dis}}(X) &= \mathrm{E}\left[\log \frac{1}{p_X(X)}\right] \\ &= -\mathrm{E}\left[\log p_X(X)\right] \\ &= -\sum_{i \in I} p_X(x_i) \log p_X(x_i), \end{aligned} \tag{3.1}$$

where $0 \times \log(0) \triangleq 0$. Here, the units of $\mathcal{H}_{\text{dis}}(X)$ are "nats" when the natural logarithm is used, and "bits" for base 2 logarithm. The change of variable $p_X(x_i) = p_i$ leads to $\mathcal{H}_{\text{dis}}(X) = -\sum_{i \in I} p_i \log p_i$ and this underlines that the entropy does not depend on the particular values set $\mathbb{X} = \{x_i, i \in I\}$. It depends only on the probability distribution $\{p_i, i \in I\}$ on this set, so that the random variables X and $Y = h(X)$ admit the same entropy whenever h is an injective transformation. The particular case $\{p_i, i \in I\} = \left\{\frac{1}{2^K}, i = 1, \ldots, 2^K\right\}$, *i.e.* an uniform distribution on a finite set including 2^K elements, leads to $\forall i : \log\left(\frac{1}{P(X=x_i)}\right) = \log\left(2^K\right)$ and to an expected information gain equal to $\mathcal{H}_{\text{dis}}(X) = K$ bits. Now, let us consider a random variable X continuously distributed on $\mathbb{R}^{d_X}$, with a probability density function denoted by $p_X(x)$, $x \in \mathbb{R}^{d_X}$. If we try to introduce a definition of its entropy which originates from the one adopted for a discrete distribution, a natural way could be (i) introduce a discretely distributed random variable $X^{\varepsilon}_{\text{dis}}$ such that $\sup_{\omega \in \Omega} \left\| X(\omega) - X^{\varepsilon}_{\text{dis}}(\omega) \right\|_{\sup} \leq \varepsilon$ which approximate X better if we decrease ε, (ii) impose a small ε value to have a small maximal difference between $X^{\varepsilon}_{\text{dis}}$ and X and (iii) define the entropy of X, approximately, by that of $X^{\varepsilon}_{\text{dis}}$. To follow this idea, let us introduce a partition $\{c_i, i \in \mathbb{N}\}$ of $\mathbb{R}^{d_X}$ (*i.e.* $\bigcup_{i \in \mathbb{N}} c_i = \mathbb{R}^{d_X}$, $i \neq j \Rightarrow c_i \cap c_j = \varnothing$) such that $\sup_{i \in \mathbb{N}} \{\text{diameter}(c_i), i \in \mathbb{N}\} \leq \varepsilon$ and choose for each i an arbitrary point x_i in c_i. For example we can retain for $\{c_i, i \in \mathbb{N}\}$ a coverage of $\mathbb{R}^{d_X}$ obtained from a lattice with the same period ε along each Cartesian axis in $\mathbb{R}^{d_X}$ (*i.e.* such that the sets c_i are disjoint (hyper-)cubes, half open along each direction with an edge length equal to ε). Then $X^{\varepsilon}_{\text{dis}}$ can be defined univocally as a function of X by imposing $X^{\varepsilon}_{\text{dis}}(\omega) = x_i$ if and only if $X(\omega) \in c_i$, where each x_i is arbitrary chosen in c_i. This leads to

$$\begin{aligned} \mathcal{H}_{\text{dis}}(X^{\varepsilon}_{\text{dis}}) &= -\sum_{i \in \mathbb{N}} P\left(X^{\varepsilon}_{\text{dis}} = x_i\right) \log\left(P\left(X^{\varepsilon}_{\text{dis}} = x_i\right)\right) \\ &= -\sum_{i \in \mathbb{N}} P\left(X \in c_i\right) \log\left(P\left(X \in c_i\right)\right). \end{aligned} \tag{3.2}$$

The discrete probability distribution of the discrete random variable $X^{\varepsilon}_{\text{dis}}$ taking its values in the countable set $\{x_i, i \in \mathbb{N}\}$ can be approximated by

$$\left\{P\left(X^{\varepsilon}_{\text{dis}} = x_i\right) \simeq p_X(x_i)v\left(c_i\right),\ i \in \mathbb{N}\right\} \tag{3.3}$$

when ε is small. So, it seems natural to approximate $\mathcal{H}_{\text{dis}}(X^{\varepsilon}_{\text{dis}})$ as follows:

$$\begin{aligned}\mathcal{H}_{\text{dis}}(X^{\varepsilon}_{\text{dis}}) &\simeq -\sum_{i\in\mathbb{N}} p_X(x_i)v\left(c_i\right)\log\left(p_X(x_i)v\left(c_i\right)\right)\\ &= -\sum_{i\in\mathbb{N}} p_X(x_i)\log\left(p_X(x_i)\right)v\left(c_i\right) - \sum_{i\in\mathbb{N}} p_X(x_i)\log\left(v\left(c_i\right)\right)v\left(c_i\right),\end{aligned} \tag{3.4}$$

where $v\left(c_i\right) = (\varepsilon)^{d_X}$ is the Lebesgue measure of c_i. Now, considering the two sums in the second line of Equ. (3.4), when the supremum value of the $v\left(c_i\right)$ is sufficiently small, we have:

$$-\sum_{i\in\mathbb{N}} p_X(x_i)\log\left(p_X(x_i)\right)v\left(c_i\right) \simeq -\int_{x\in\mathbb{R}^{d_X}} p_X(x)\log\left(p_X(x)\right)v\left(\mathrm{d}x\right) \tag{3.5}$$

and

$$\begin{aligned}-\sum_{i\in\mathbb{N}} p_X(x_i)\log\left(v\left(c_i\right)\right)v\left(c_i\right) &\simeq -\log\left(v\left(c_i\right)\right)\int_{x\in\mathbb{R}^{d_X}} p_X(x)v\left(\mathrm{d}x\right)\\ &= -\log\left(v\left(c_i\right)\right).\end{aligned} \tag{3.6}$$

So we can approximate the sum in Equ. (3.4) by

$$\begin{aligned}\mathcal{H}_{\text{dis}}(X^{\varepsilon}_{\text{dis}}) &\simeq -\int_{x\in\mathbb{R}^{d_X}} p_X(x)\log\left(p_X(x)\right)v\left(\mathrm{d}x\right) - \log\left(v\left(c_i\right)\right)\\ &= -\mathrm{E}\left[\log\left(p_X(X)\right)\right] - d_X\log\varepsilon.\end{aligned} \tag{3.7}$$

Now, clearly this approximation is not bounded below when ε decreases and we can note that the divergent term $-d_X\log\varepsilon$ does not depend on the density function p_X. Consequently a continuous-specific entropy definition has been proposed. It is named differential entropy [Papoulis 1985], denoted here by $\mathcal{H}_{\text{cnt}}(X)$ and obtained by keeping only the first term in Equ. (3.7):

$$\mathcal{H}_{\text{cnt}}(X) = -\mathrm{E}\left[\log\left(p_X(X)\right)\right]. \tag{3.8}$$

This definition is used currently when the observed data are modelized as continuously distributed in probability, and will be generally retained in the sequel. It seems strange to neglect $-d_X\log\varepsilon$. But, In fact, the unbounded increase of $-d_X\log\varepsilon$ (when ε decreases)

can be interpreted as the necessary increase of information to localize more precisely a point in $\mathbb{R}^{d_X}$ and can be considered as to be non-relevant to quantify the uncertainty inherent to the shape of the probability distribution of X on $\mathbb{R}^{d_X}$. Moreover, if one wants to compare the uncertainty associated with two random variables X_1 and X_2, each of them continuously distributed on $\mathbb{R}^{d_X}$, we have no clear argument leading to choose two distinct values of ε, ε_1 and ε_2. So, we can consider that it is sufficient to compute the differential entropies to make this comparison. Finally, considering Equ. (3.1) and (3.8), a same formula can be retained, $\mathcal{H}(X) = -\mathrm{E}\left[\log p_X(X)\right]$, where $\mathcal{H}(X)$ corresponds to $\mathcal{H}_{\mathrm{dis}}(X)$ or $\mathcal{H}_{\mathrm{cnt}}(X)$ and $p_X(\cdot)$ corresponds respectively to a discrete distribution or to a density distribution. When the context is clear, this common symbolization will not introduce any confusion. But in the continuous distribution case, it must be clear that $\mathcal{H}(X)$ may be negative contrary to the discrete case. Another important difference between the discrete and continuous cases is the influence of an invertible transformation $h : \mathbb{R}^{d_X} \to \mathbb{R}^{d_X}$ on the value of the entropy $\mathcal{H}(Y)$ of the random vector $Y = h(X)$. In the discrete case, the entropy is invariant under any invertible transformation, *i.e.* $\forall h$ invertible : $\mathcal{H}_{\mathrm{dis}}(Y) = \mathcal{H}_{\mathrm{dis}}(h(X)) = \mathcal{H}_{\mathrm{dis}}(X)$ and in the continuous case this property is no more fulfilled.

It is interesting to introduce the following integral representation of entropy (which will be useful later)

$$\begin{aligned}\mathcal{H}(X) &= -\mathrm{E}\left[\log p_X(X)\right] \\ &= \int_{\mathbb{R}^{d_X}} \log\left(\frac{\mathrm{d}P_X}{\mathrm{d}\mu_r}(x)\right) \mathrm{d}P_X(x),\end{aligned} \tag{3.9}$$

where the sum is defined with respect to the probability measure P_X (on the Borel sets of $\mathbb{R}^{d_X}$) induced by the random vector $X : \Omega \to \mathbb{R}^{d_X}$, and which can be discrete or continuous. Here the derivative of the measure P_X with respect to the reference measure μ_r, $\frac{\mathrm{d}P_X}{\mathrm{d}\mu_r}(x)$, represents the discrete probability distribution or the probability density function (in the discrete case, this reference measure is the uniform counting measure supported by the countable set $X(\Omega) = \{x_i, i \in I\}$, and in the continuous case it corresponds to the Lebesgue measure on $\mathbb{R}^{d_X}$).

Joint Entropy

The entropy of a pair $(X : \Omega \to \mathbb{R}^{d_X}, Y : \Omega \to \mathbb{R}^{d_Y})$ of random vectors must correspond, clearly, to the entropy of the (column) vector $\left[X^T, Y^T\right]^T$. It is named joint entropy and is denoted by $\mathcal{H}(X, Y)$. Consequently, we have:

$$\mathcal{H}(X, Y) = -\mathrm{E}\left[\log p_{X,Y}(X, Y)\right]. \tag{3.10}$$

This definition can be easily extended to a finite set comprising more than two random

vectors. Equ. (3.10) implies that, if X and Y are independent, *i.e.* if $p_{X,Y}(\cdot,\cdot) = p_X(\cdot)p_Y(\cdot)$, we have the following relation

$$\mathcal{H}(X,Y) = \mathcal{H}(X) + \mathcal{H}(Y). \tag{3.11}$$

Conditional Entropy

The conditional entropy $\mathcal{H}(X|Y)$ measures the uncertainty about X when Y is observed beforehand

$$\mathcal{H}(X|Y) = -\mathrm{E}\left[\log p_{X|Y}(X|Y)\right], \tag{3.12}$$

where $p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$ is the conditional discrete distribution or the conditional probability density function. Note that the expectation in Equ. (3.12) is computed with respect to the joint probability distribution of (X,Y) as indicated below

$$\begin{aligned}\mathcal{H}(X|Y) &= -\mathrm{E}\left[\log p_{X|Y}(X|Y)\right] \\ &= \int_{\mathbb{R}^{d_X}} \log\left(\frac{\mathrm{d}P_{X|Y}(\cdot|y)}{\mathrm{d}\mu_r}(x)\right)\mathrm{d}P_{X,Y}(x,y),\end{aligned} \tag{3.13}$$

where μ_r is defined on the Borel sets of $\mathbb{R}^{d_X}$. The following basic property is easy to verify:

$$\mathcal{H}(X|Y) = \mathcal{H}(X,Y) - \mathcal{H}(Y). \tag{3.14}$$

3.1.1.2. Mutual Information

Mutual Information (MI), besides its historical central role in telecommunication theory and engineering, is a widely used information-theoretic independence measurement which has received particular attention during the past few years [Urbanczik 2003, Stögbauer 2004, Wissman 2011, Foster 2011, Dunleavy 2012].

For two discrete random variables X and Y, with outcomes x and y from $\mathbb{X}$ and $\mathbb{Y}$ separately, the mutual information $\mathcal{I}(X,Y)$ is defined as

$$\mathcal{I}_{\mathrm{dis}}(X,Y) = \sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} p_{X,Y}(x,y)\log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right), \tag{3.15}$$

where $p_{X,Y}(x,y) = P(X=x, Y=y)$, $p_X(x) = P(X=x)$ and $p_Y(y) = P(Y=y)$.

If X and Y are continuous random variables with dimensions d_X and d_Y respectively,

the definition changes to

$$\mathcal{I}_{\text{cnt}}(X,Y) = \int\limits_{y\in\mathbb{R}^{d_Y}} \mathrm{d}y \int\limits_{x\in\mathbb{R}^{d_X}} p_{X,Y}(x,y)\log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right)\mathrm{d}x. \tag{3.16}$$

Equ. (3.15) and (3.16) can be summarized by the following integral with respect to the measure $P_{X,Y}$

$$\mathcal{I}(X,Y) = \int\limits_{(x,y)\in\mathbb{R}^{d_X}\times\mathbb{R}^{d_Y}} \log\left(\frac{\mathrm{d}P_{X,Y}}{\mathrm{d}\left(P_X\otimes P_Y\right)}(x,y)\right)\mathrm{d}P_{X,Y}(x,y), \tag{3.17}$$

where the derivative in brackets is defined with respect to the tensor product of P_X and P_Y (by definition, $P_X\otimes P_Y$ is equal to the joint probability measure for a pair of random vectors U and V, respectively valuated in $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$, such that $P_U = P_X$ and $P_V = P_Y$).

Theoretically, $\mathcal{I}(X,Y)$ is always non-negative, and $\mathcal{I}(X,Y) = 0$ if and only if X and Y are independent (*i.e.* $p_{X,Y}(\cdot,\cdot) = p_X(\cdot)p_Y(\cdot)$).

Mutual information is symmetric,

$$\mathcal{I}(X,Y) = \mathcal{I}(Y,X), \tag{3.18}$$

and it can also be expressed in terms of entropies

$$\mathcal{I}(X,Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y), \tag{3.19}$$

or, equivalently

$$\mathcal{I}(X,Y) = \mathcal{H}(X) - \mathcal{H}(X|Y). \tag{3.20}$$

According to Equ. (3.20), $\mathcal{I}(X,Y)$ can also be considered as the decrease of uncertainty on X supplied by the observation of Y. The relation between the information-theoretic quantities mentioned above given non-independent random variables X and Y is summarized in Fig. 3.1.

Mutual information is tightly related to the Kullback-Leibler divergence [Latham 2009, Cover 2012], which is a non-symmetrical measure of the dissimilarity between two distributions. Given two probability measures P and Q on the borel sets of $\mathbb{R}^d$, the Kullback-Leibler divergence of P with respect to Q is defined as

$$\mathcal{D}_{\text{kl}}(P||Q) = \int\limits_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right)\mathrm{d}P(x). \tag{3.21}$$

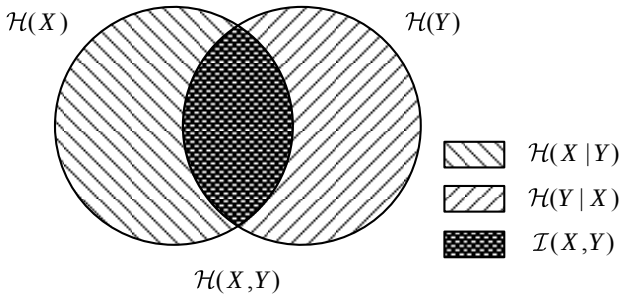


Figure 3.1: The relation between different information-theoretic quantities with non-independent random variables X and Y [Abramson 1963]. The area contained by both circles is the joint entropy $\mathcal{H}(X,Y)$. The left and right circles stand for the individual entropies $\mathcal{H}(X)$ and $\mathcal{H}(Y)$ respectively.

The Kullback-Leibler divergence is also frequently named Kullback distance, improperly, as this measure is not symmetric, *i.e.* $\mathcal{D}_{\mathrm{kl}}(P||Q) \neq \mathcal{D}_{\mathrm{kl}}(Q||P)$. The main property is that $\mathcal{D}_{\mathrm{kl}}(P||Q) \geq 0$ with equality to zero if and only if $P = Q$. According to Equ. (3.21), mutual information can be written as the Kullback-Leibler divergence of the joint measure $P_{X,Y}$ with respect to the measure equal to the tensorial product $P_X \otimes P_Y$:

$$\mathcal{I}(X,Y) = \mathcal{D}_{\mathrm{kl}}(P_{X,Y}||P_X \otimes P_Y). \tag{3.22}$$

Thus, mutual information can be considered as a measure of how close the joint distribution of (X,Y) is to the product of the marginal distributions of X and Y.

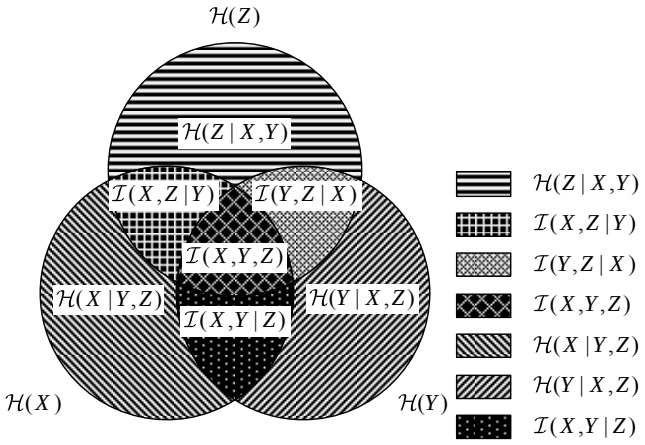


Figure 3.2: An illustration for the relations of different conditional information-theoretic measures for three non-independent random variables X, Y, and Z [Abramson 1963]. The lower-left, lower-right and upper circles, stand for $\mathcal{H}(X)$, $\mathcal{H}(Y)$ and $\mathcal{H}(Z)$ respectively.

For three random variables X, Y and Z, the conditional mutual information $\mathcal{I}(X,Y\,|Z)$ measures the information shared between X and Y conditioned on Z. By applying Equ.

(3.14) and (3.20), it can be presented in terms of joint or conditional entropies,

$$\begin{aligned}\mathcal{I}(X,Y|Z) &= \mathcal{H}(X,Z) + \mathcal{H}(Y,Z) - \mathcal{H}(X,Y,Z) - \mathcal{H}(Z) \\ &= \mathcal{H}(X|Z) - \mathcal{H}(X|Y,Z).\end{aligned} \tag{3.23}$$

Additionally, the concept of mutual information can also be extended to multivariate situation. For a set of random variables $\{X_1, X_2, \ldots, X_M\}$, $\mathcal{I}(X_1, X_2, \ldots, X_M)$ quantifies the information shared among these variables. For $M > 1$, we have the following relation

$$\mathcal{I}(X_1, X_2, \ldots, X_M) = \mathcal{I}(X_1, X_2, \ldots, X_{M-1}) - \mathcal{I}(X_1, X_2, \ldots, X_{M-1}|X_M). \tag{3.24}$$

The relations between the different conditional information-theoretic quantities are illustrated in Fig. 3.2.

3.1.1.3. Transfer Entropy

Transfer Entropy (TE) is an information-theoretic statistical measurement, which aims at measuring the amount of time-directed information between two dynamical systems. As described in the previous chapter, given the past time evolution of a dynamical system $\mathcal{A}$, the transfer entropy from another dynamical system $\mathcal{B}$ to the first system $\mathcal{A}$ is the amount of Shannon uncertainty reduction in the future time evolution of $\mathcal{A}$ when including the knowledge of the past evolution of $\mathcal{B}$.

More precisely, let us suppose that we observe the time sampled output $X_i \in \mathbb{R}$, $i \in \mathbb{Z}$, of some sensors connected to $\mathcal{A}$. If the sequence X is supposed to be a mth order Markov process, *i.e.* if considering subsequences $X_i^{(k)} = (X_{i-k+1}, X_{i-k+2}, \cdots, X_i)$, $k > 0$, the probability measure $\mathcal{P}_X$ (defined on measurable subsets of real sequences) attached to X fulfills the mth order Markov hypothesis

$$\forall i : \forall m' > m : \mathrm{d}\mathcal{P}_{X_{i+1}|X_i^{(m)}}\left(x_{i+1}|x_i^{(m)}\right) = \mathrm{d}\mathcal{P}_{X_{i+1}|X_i^{(m')}}\left(x_{i+1}|x_i^{(m')}\right), \quad x_{i+1} \in \mathbb{R}, x_i^{(k)} \in \mathbb{R}^k, \tag{3.25}$$

then the past information $X_i^{(m)}$ (before time instant $i+1$) is sufficient for a prediction of X_{i+k}, $k \geq 1$, and can be considered as a m-dimensional state vector at time i (note that, to know from X the hidden dynamical evolution of $\mathcal{A}$, we need a one-to-one relation between $X_i^{(m)}$ and the physical state of $\mathcal{A}$ at time i). For sake of clarity, we introduce the following notation: $\left(X_i^p, X_i^-, Y_i^-\right)$, $i = 1, 2, \ldots, N$, is an independent and identically distributed (IID) random sequence, each term following the same distribution as a random vector $(X^p, X^-, Y^-) \in \mathbb{R}^{1+m+n}$ whatever i (in X^p, X^-, Y^-, the superscripts "p" and "$-$" correspond to "predicted" and "past" respectively). This notation substitutes for

the notation $\left(X_{i+1}, X_i^{(m)}, Y_i^{(n)}\right)$, $i = 1, 2, \ldots, N$ and we denote by $\mathcal{S}_{X^p,X^-,Y^-}$, $\mathcal{S}_{X^p,X^-}$, $\mathcal{S}_{X^-,Y^-}$ and $\mathcal{S}_{X^-}$ the spaces in which (X^p, X^-, Y^-), (X^p, X^-), (X^-, Y^-) and X^- are respectively observed.

Now, let us suppose that a causal influence exists from $\mathcal{B}$ to $\mathcal{A}$ and that an auxiliary random process $Y_i \in \mathbb{R}$, $i \in \mathbb{Z}$, recorded from a sensor connected to $\mathcal{B}$ is such that, at each time i and for some $n > 0$, $Y_i^- \triangleq Y_i^{(n)}$ is an image (not necessarily one-to-one) of the physical state of $\mathcal{B}$. The negation of this causal influence implies

$$\forall n > 0 : \forall i : \mathrm{d}\mathcal{P}_{X_{i+1}|X_i^{(m)}}\left(x_{i+1}|x_i^{(m)}\right) = \mathrm{d}\mathcal{P}_{X_{i+1}|X_i^{(m)},Y_i^{(n)}}\left(x_{i+1}|x_i^{(m)}, y_i^{(n)}\right). \tag{3.26}$$

If Equ. (3.26) holds, it is said that there is an absence of information transfer from $\mathcal{B}$ to $\mathcal{A}$. Otherwise the process X can be no more considered strictly as a Markov process. Let us suppose the joint process (X, Y) is Markovian, *i.e.* there exist a given pair (m', n'), a transition function f, and an independent random sequence e_i, $i \in \mathbb{Z}$, such that

$$[X_{i+1}, Y_{i+1}]^T = f\left(X_i^{(m\prime)}, Y_i^{(n')}, e_{i+1}\right), \tag{3.27}$$

where the random variable e_{i+1} is independent of the past random sequence $\left(X_j, Y_j, e_j\right)$, $j \leq i$, whatever i. As $X_i = g\left(X_i^{(m)}, Y_i^{(n)}\right)$ where g is clearly a non-injective function, the pair $\left\{\left(X_i^{(m)}, Y_i^{(n)}\right), X_i\right\}$, $i \in \mathbb{Z}$, corresponds to a hidden Markov Process, and it is well known that this observation process is not generally Markovian.

The deviation from this assumption can be quantified using the Kullback pseudo-metric. To define TE at time i let us consider the Kullback pseudo metric of the measure $\mathcal{P}_{X_i^p|X_i^-,Y_i^-}\left(\cdot|x_i^-, y_i^-\right)$ with respect to $\mathcal{P}_{X_i^p|X_i^-}\left(\cdot|x_i^-\right)$ as a function of $\left(x_i^-, y_i^-\right)$, say $\mathcal{D}_{\mathrm{kl}}\left(\mathcal{P}_{X_i^p|X_i^-,Y_i^-}\left(\cdot|x_i^-, y_i^-\right)||\mathcal{P}_{X_i^p|X_i^-}\left(\cdot|x_i^-\right)\right) = g\left(x_i^-, y_i^-\right)$. It quantifies, conditionally to $\left\{X_i^- = x_i^-, Y_i^- = y_i^-\right\}$, a deviation from the hypothesis in Equ. (3.25). Then TE at time i can be defined at the average of this conditional Kullback distance computed with respect to the joint probability distribution of $\left(X_i^-, Y_i^-\right)$

$$\begin{aligned}\mathrm{TE}_{Y\to X,i} &= \mathrm{E}\left[\mathcal{D}_{\mathrm{kl}}\left(\mathcal{P}_{X_i^p|X_i^-,Y_i^-}\left(\cdot|X_i^-, Y_i^-\right)||\mathcal{P}_{X_i^p|X_i^-}\left(\cdot|X_i^-\right)\right)\right] \\ &= \int\limits_{\mathcal{S}_{X^-,Y^-}} g\left(x_i^-, y_i^-\right)\mathrm{d}\mathcal{P}_{X_i^-,Y_i^-}\left(x_i^-, y_i^-\right),\end{aligned} \tag{3.28}$$

which can be finally written

$$\mathrm{TE}_{Y\to X,i} = \int\limits_{\mathcal{S}_{X^p,X^-,Y^-}} \log\left(\frac{\mathrm{d}\mathcal{P}_{X_i^p|X_i^-,Y_i^-}\left(\cdot|x_i^-, y_i^-\right)}{\mathrm{d}\mathcal{P}_{X_i^p|X_i^-}\left(\cdot|x_i^-\right)}\left(x_i^p\right)\right)\mathrm{d}\mathcal{P}_{X_i^p,X_i^-,Y_i^-}\left(x_i^p, x_i^-, y_i^-\right), \tag{3.29}$$

where the ratio in Equ. (3.29) corresponds to the Radon-Nikodym derivative [Roman 1974] (*i.e.* the density) of the conditional measure $\mathrm{d}\mathcal{P}_{X_i^p|X_i^-,Y_i^-}\left(\cdot|x_i^-,y_i^-\right)$ with respect to the conditional measure $\mathrm{d}\mathcal{P}_{X_i^p|X_i^-}\left(\cdot|x_i^-\right)$. Now, given two observable scalar random time series X and Y with no *a priori* given model (as it is generally the case), if we are interested in defining some causal influence from Y to X through TE analysis, we must specify the dimensions of the past information vectors X^- and Y^-, *i.e.* m and n. Even if we impose them, it is not evident that all the coordinates in $X_i^{(m)}$ and $Y_i^{(n)}$ will be useful. To deal with this issue, variable selection procedures have been proposed in the literature such as uniform and non-uniform embedding algorithms [Kugiumtzis 2013, Montalto 2014].

Here, we consider that the joint probability measure $\mathcal{P}_{X_i^p,X_i^-,Y_i^-}$ is absolutely continuous (with respect to the Lebesgue measure in $\mathbb{R}^{m+n+1}$ denoted by μ^{m+n+1}) with the corresponding probability density function

$$p_{X_i^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)=\frac{\mathrm{d}\mathcal{P}_{X_i^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)}{\mathrm{d}\mu^{n+m+1}\left(x_i^p,x_i^-,y_i^-\right)}. \tag{3.30}$$

Then, we are sure that the following conditional densities probability functions exist:

$$\begin{cases} p_{X_i^p|X_i^-}\left(x_i^p|x_i^-\right)=\dfrac{\mathrm{d}\mathcal{P}_{X_i^p|X_i^-}\left(x_i^p|x_i^-\right)}{\mathrm{d}\mu^1\left(x_i^p\right)} \\ p_{X_i^p|X_i^-,Y_i^-}\left(x_i^p|x_i^-,y_i^-\right)=\dfrac{\mathrm{d}\mathcal{P}_{X_i^p|X_i^-,Y_i^-}\left(x_i^p|x_i^-,y_i^-\right)}{\mathrm{d}\mu^1\left(x_i^p\right)}, \end{cases} \tag{3.31}$$

and Equ. (3.29) yields to

$$\begin{aligned} \mathrm{TE}_{Y\to X,i} &= \int_{\mathbb{R}^{m+n+1}} p_{X^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)\log\left(\frac{p_{X^p|X_i^-,Y_i^-}\left(x_i^p|x_i^-,y_i^-\right)}{p_{X^p|X_i^-}\left(x_i^p|x_i^-\right)}\right)\mathrm{d}x_i^p\mathrm{d}x_i^-\mathrm{d}y_i^- \\ &= \int_{\mathbb{R}^{m+n+1}} p_{X^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)\times \\ &\qquad \log\left(\frac{p_{X^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)p_{X_i^-}\left(x_i^-\right)}{p_{X_i^-,Y_i^-}\left(x_i^-,y_i^-\right)p_{X^p,X_i^-}\left(x_i^p,x_i^-\right)}\right)\mathrm{d}x_i^p\mathrm{d}x_i^-\mathrm{d}y_i^-. \end{aligned} \tag{3.32}$$

Equ. (3.32) can be rewritten

$$\begin{aligned} \mathrm{TE}_{Y\to X,i} = &-\mathrm{E}\left[\log\left(p_{X_i^p,X_i^-}\left(X_i^p,X_i^-\right)\right)\right]-\mathrm{E}\left[\log\left(p_{X_i^-,Y_i^-}\left(X_i^-,Y_i^-\right)\right)\right] \\ &+\mathrm{E}\left[\log\left(p_{X_i^p,X_i^-,Y_i^-}\left(X_i^p,X_i^-,Y_i^-\right)\right)\right]+\mathrm{E}\left[\log\left(p_{X_i^-}\left(X_i^-\right)\right)\right], \end{aligned} \tag{3.33}$$

or

$$\mathrm{TE}_{Y\to X,i} = \mathcal{H}\left(X_i^p, X_i^-\right) + \mathcal{H}\left(X_i^-, Y_i^-\right) - \mathcal{H}\left(X_i^p, X_i^-, Y_i^-\right) - \mathcal{H}\left(X_i^-\right), \tag{3.34}$$

where $\mathcal{H}(U)$ denotes the Shannon differential entropy of a random vector U. Considering "log" as the natural logarithm, $\mathrm{TE}_{Y\to X,i}$ is measured in natural units (nats). Note that, if the processes Y and X are assumed to be jointly stationary, for any real function $g : \mathbb{R}^{m+n+1} \to \mathbb{R}$, the expectation $\mathrm{E}\left[g\left(X_{i+1}, X_i^{(m)}, Y_i^{(n)}\right)\right]$ does not depend on i. Consequently, $\mathrm{TE}_{Y\to X,i}$ does not depend on i (and so can be simply denoted by $\mathrm{TE}_{Y\to X}$), nor all the quantities defined in Equ. (3.28) to (3.34). In theory, TE is never negative and is equal to zero if and only if Equ. (3.26) holds.

According to the definition (in Equ. (3.28)), TE is not symmetric and it can be regarded as a conditional mutual information (CMI) [Hlaváčková-Schindler 2007, Paluš 2001] (sometimes also named partial mutual information (PMI) in the literature [Frenzel 2007]). Recalling that definition of conditional mutual information in Equ. (3.23), TE can be also written as

$$\mathrm{TE}_{Y\to X} = \mathcal{I}\left(X^p, Y^- | X^-\right). \tag{3.35}$$

TE can be considered as a measurement of the degree to how past Y^- of the process Y disambiguates the future X^p of X beyond the degree to how its only past X^- disambiguates its future [Paluš 2001]. It is an information-theoretic implementation of Wiener's principle of observational causality. Hence TE reveals a natural relation to Granger causality. As it is well known, Granger causality emphasizes the concept of reduction of the mean square error of the linear prediction of X_i^p when adding Y_i^- to X_i^- by introducing the Granger causality index

$$\mathrm{GC}_{Y\to X} = \log\left(\frac{\mathrm{var}\left(lpe_{X_i^p|X_i^-}\right)}{\mathrm{var}\left(lpe_{X_i^p|X_i^-,Y_i^-}\right)}\right), \tag{3.36}$$

which is independent of i under the stationary hypothesis and where $lpe_{X_i^p|U}$ is the error when predicting linearly X_i^p from U. TE is framed in terms of reduction of the Shannon uncertainty (entropy) of the predictive probability distribution. When the probability distribution of $\left(X_i^p, X_i^-, Y_i^-\right)$ is assumed to be Gaussian, TE and Granger causality are entirely equivalent, up to a factor of 2 [Barnett 2009]:

$$\mathrm{TE}_{Y\to X} = \frac{1}{2}\mathrm{GC}_{Y\to X}. \tag{3.37}$$

Equ. (3.37) can be used as a reference for the comparison between different TE estimators. Consequently, in the Gaussian case, TE can be easily computed from a

statistical second-order characterization of $\left(X_i^p, X_i^-, Y_i^-\right)$. This Gaussian assumption obviously holds when the processes Y and X are jointly normally distributed and, more particularly, when they correspond to a Gaussian autoregressive (AR) bivariate process. In [Barnett 2009] Barnett *et al.* discussed the relation between these two causality measures and this work bridged information-theoretic methods and autoregressive ones.

TE is used as a pairwise causality detection measure, however, as mentioned in chapter 2, this kind of pairwise approach is not able to distinguish the direct and indirect relations in multivariate systems. Similar as Granger causality, TE could also be extended to the conditional situation, termed conditional transfer entropy (CTE) [Yang 2012] (or partial transfer entropy (PTE) in some literature [Gómez-Herrero 2015]). Considering three random variables X, Y and Z, in order to quantify the information flow from Y to X conditioned on Z, similarly to Equ. (3.28), CTE at time i can be defined as

$$\begin{aligned} \mathrm{TE}_{Y\to X|Z,i} = \int\limits_{\mathbb{R}^{m+n+q+1}} \log\left(\frac{\mathrm{d}\mathcal{P}_{X_i^p|X_i^-,Y_i^-,Z_i^-}\left(x_i^p|x_i^-,y_i^-,z_i^-\right)}{\mathrm{d}\mathcal{P}_{X_i^p|X_i^-,Z_i^-}\left(x_i^p|x_i^-,z_i^-\right)}\right) \\ \mathrm{d}\mathcal{P}_{X_i^p,X_i^-,Y_i^-,Z_i^-}\left(x_i^p,x_i^-,y_i^-,z_i^-\right), \end{aligned} \tag{3.38}$$

where Z^- stands for the past of Z with order q.

Similarly as for transfer entropy, CTE could also be presented by means of conditional mutual information

$$\mathrm{TE}_{Y\to X|Z} = \mathcal{I}\left(X^p, Y^-|X^-, Z^-\right), \tag{3.39}$$

and Equ. (3.39) could be rewritten in terms of joint and marginal entropies

$$\mathrm{TE}_{Y\to X|Z,i} = \mathcal{H}\left(X_i^-, Y_i^-, Z_i^-\right) + \mathcal{H}\left(X_i^p, X_i^-, Z_i^-\right) - \mathcal{H}\left(X_i^p, X_i^-, Y_i^-, Z_i^-\right) - \mathcal{H}\left(X_i^-, Z_i^-\right). \tag{3.40}$$

3.1.2. The Estimator Structures for MI and TE

Let us consider the estimation $\widehat{\mathrm{TE}_{Y\to X}}$ of $\mathrm{TE}_{Y\to X}$ as a function defined on the set of observable occurrences (x_i, y_i), $i = 1, \ldots, N$, of a stationary sequence (X_i, Y_i), $i = 1, \ldots, N$.

From Equ. (3.33), assuming that X and Y are jointly strongly ergodic leads to

$$\begin{aligned} \mathrm{TE}_{Y\to X} = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\Bigg(-\log\left(p_{X_i^-,Y_i^-}\left(X_i^-,Y_i^-\right)\right) - \log\left(p_{X_i^p,X_i^-}\left(X_i^p,X_i^-\right)\right) \\ + \log\left(p_{X_i^p,X_i^-,Y_i^-}\left(X_i^p,X_i^-,Y_i^-\right)\right) + \log\left(p_{X_i^-}\left(X_i^-\right)\right)\Bigg), \end{aligned} \tag{3.41}$$

where the convergence holds with probability one. Hence, a standard estimation $\widehat{\mathrm{TE}_{Y\to X}}$ of $\mathrm{TE}_{Y\to X}$ is given by

$$\begin{aligned}\widehat{\mathrm{TE}_{Y\to X}} &= \widehat{\mathcal{H}\left(X^-,Y^-\right)} + \widehat{\mathcal{H}\left(X^p,X^-\right)} - \widehat{\mathcal{H}\left(X^p,X^-,Y^-\right)} - \widehat{\mathcal{H}\left(X^-\right)} \\ &= -\frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_1}\left(u_{1n}\right)\right)} - \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_2}\left(u_{2n}\right)\right)} + \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_3}\left(u_{3n}\right)\right)} \\ &\quad + \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_4}\left(u_{4n}\right)\right)},\end{aligned} \tag{3.42}$$

where U_1, U_2, U_3 and U_4 stand respectively for (X^-,Y^-), (X^p,X^-), (X^p,X^-,Y^-) and X^-. For each n, $\widehat{\log\left(p_U\left(u_n\right)\right)}$ is an estimated value of $\log\left(p_U\left(u_n\right)\right)$ computed as a function $f_n\left(u_1,\ldots,u_N\right)$ of the observed sequence u_n, $n=1,\ldots,N$. With the kNN approach, $f_n\left(u_1,\ldots,u_N\right)$ explicitly depends only on u_n and on its k NNs (nearest neighbors). So, the calculation of $\widehat{\mathcal{H}(U)}$ is completely specified by the chosen estimation functions f_n. Note that if, for N fixed, these functions correspond respectively to unbiased estimators of $\log\left(p\left(u_n\right)\right)$, then $\widehat{\mathrm{TE}_{Y\to X}}$ is also unbiased, otherwise we can only expect that $\widehat{\mathrm{TE}_{Y\to X}}$ is asymptotically unbiased (for N large). It is like that if the estimators of $\log\left(p_U\left(u_n\right)\right)$ are asymptotically unbiased.

Now, the theoretical derivation and analysis of the most currently used estimators

$$\widehat{\mathcal{H}(U)}\left(u_1,\ldots,u_N\right) = -\frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p(u_n)\right)} \tag{3.43}$$

for the estimation of $\mathcal{H}(U)$ generally suppose that $u_1,\ldots,u_N$ are N independent occurrences of the random vector U, *i.e.* $u_1,\ldots,u_N$ is an occurrence of an IID sequence $U_1,\ldots,U_N$ of random vectors $(\forall i=1,\ldots,N:\mathcal{P}_{U_i}=\mathcal{P}_U)$. Although the IID hypothesis does not apply to our initial problem concerning the measure of TE on stationary random sequences (that are generally not IID), the new methods presented in this thesis are extended from existing ones assuming this IID hypothesis (without relaxing it). However, the chapter 4 will present results not only on IID observations but also on non-IID stationary AR processes as our goal was to verify if some improvements can be nonetheless obtained for non-IID data such as AR data.

If we come back to MI defined by Equ. (3.19) and compare it with Equ. (3.34), it is obvious that estimating MI and TE shares similarities. Hence, similarly to Equ. (3.42) for TE, a basic estimation $\widehat{\mathcal{I}(X,Y)}$ of $\mathcal{I}(X,Y)$ from a sequence (x_i,y_i), $i=1,\ldots,N$, of

N independent trials is

$$\widehat{\mathcal{I}(X,Y)} = -\frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_X\left(x_n\right)\right)} - \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_Y\left(y_n\right)\right)} + \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{X,Y}\left(x_n,y_n\right)\right)}. \tag{3.44}$$

To conclude this section, we reviewed the mathematical definitions of different information-theoretic quantities, and discussed the similarities of the estimator structures of MI and TE. For practical use, it is important to find accurate estimators for these quantities. Deriving a good estimator includes two aspects: unbiasness and accuracy, which are reflected in estimation bias and variance. These issues are largely discussed in the literature. However, the estimation remains a tough task while carried out on finite sample length signals, for example in the field of neuroscience, where getting large amounts of stationary data is problematical [Hlaváčková-Schindler 2007]. Moreover, this estimation suffers from the "curse of dimensionality" [Verleysen 2005], especially for transfer entropy. In the following section, we focus on the estimation of these quantities, and give an overview of some important existing approaches.

3.2. Previous Works

For the estimation of information-theoretic quantities, there are two kinds of approaches, parametric and non-parametric [Venelli 2010]. The parametric approaches make extra assumptions on the data to be processed. For example, they can be assumed to be issued from a known family of distributions, such as normal distribution, and the derived estimator is optimized based on this assumption. In contrast, for the non-parametric estimations, no *a priori* assumption is considered, and the estimators depend on the data themselves. In the remainder of this section, only efficient non-parametric estimation methods are described.

3.2.1. Estimation of Entropy

Coming back to the definition of entropy in Equ. (3.1), given a random variable X with samples x_i, $i = 1,\ldots,N$, the estimation of $\mathcal{H}(X)$, $\widehat{\mathcal{H}(X)}$, can be calculated as

$$\widehat{\mathcal{H}(X)} = -\frac{1}{N}\sum_{i=1}^{N}\log\widehat{p(x_i)}, \tag{3.45}$$

which requires $\widehat{p(x_i)}$ is the estimation of the probability density function $p(\cdot)$ at data point x_i.

To solve this problem, for scalar observations, the most straightforward approach

is the histogram-based method. Let us consider for example the one-dimensional case. The real axis is first partitioned into M bins corresponding to M equal length intervals $a_j = [s + (j-1)\cdot\tau, s + j\cdot\tau]$ with $j = 1, \ldots, M$, where s is the starting value of the first interval a_1, and τ is the length of the intervals (bins). If the number of samples falling into interval a_j is denoted by n_j, and the two following conditions are satisfied

$$\begin{cases} s \leq \min\{x_1, \ldots, x_N\} \\ s + M\cdot\tau \geq \max\{x_1, \ldots, x_N\}, \end{cases} \tag{3.46}$$

then we have

$$\sum_{j=1}^{M} n_j = N \tag{3.47}$$

and the probability $p_X(x)$ is calculated as

$$\widehat{p_X(x)} = \sum_j \frac{n_j}{N\tau} 1_{a_j}(x), x \in \mathbb{R}, \tag{3.48}$$

where $1_E(\cdot)$ is the indicator function of the set E.

In this case, the summation in Equ. (3.45) should be rewritten, such as

$$\widehat{\mathcal{H}(X)}_{\text{his}} = -\sum_{j=1}^{M} \left(\frac{n_j}{\tau N} \log\left(\frac{n_j}{\tau N}\right)\right)\tau, \tag{3.49}$$

which corresponds, when N is large enough and τ small enough to get $\frac{n_j}{\tau N} \simeq p_X(x)$, $x \in a_j$, to a Riemann approximation of the integral $-\int_{x\in\mathbb{R}} p_X(x)\log(p_X(x))\,\mathrm{d}x = \mathcal{H}(X)$.

For the histogram-based method, the choice of the interval length τ (related to the number of bins M) is critical. Even if this method is efficient in a computational point of view, it only gives approximate estimation.

Beyond the histogram-based method, we can go further with density estimation. Given a random variable X, drawn from the unknown density function $p_X(\cdot)$, the probability that a new sample of X falls into a region $\mathcal{L}(x)$ around point x is given by

$$P(X \in \mathcal{L}(x)) = \int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y. \tag{3.50}$$

Suppose that the total number of data points drawn independently from $p_X(\cdot)$ is N and that k points fall into the region $\mathcal{L}(x)$, if N is large enough, this probability can be written

$$P(X \in \mathcal{L}(x)) \approx \frac{k}{N}. \tag{3.51}$$

Moreover, if the diameter of $\mathcal{L}(x)$ is small enough, $p_X(x)$ can be considered as constant in this region. So Equ. (3.50) can be rewritten as

$$\begin{aligned} P(X \in \mathcal{L}(x)) &= \int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y \\ &\approx p_X(u) \int_{\mathcal{L}(x)} \mathrm{d}y \\ &= p_X(u)V, \ u \in \mathcal{L}(x), \end{aligned} \tag{3.52}$$

where V is the volume of the region $\mathcal{L}(x)$. Combining Equ. (3.51) and (3.52), we obtain

$$p_X(u) \approx \frac{k}{NV}, \ u \in \mathcal{L}(x). \tag{3.53}$$

As shown in Fig. 3.3, in order to estimate $p_X(x)$, there are two possible starting points: (i) choose a fixed region size with a fixed volume V and count how many points fall into this region, (ii) choose a fixed value of neighbors k and compute the corresponding volume V of the smallest ball including the k neighbors. In the literature [Bishop 1995], the first way refers to "kernel density estimation" (KDE, here with a Kernel shape corresponding to the indicator function of the chosen region), and the second one to the "k-Nearest Neighbors" approach.

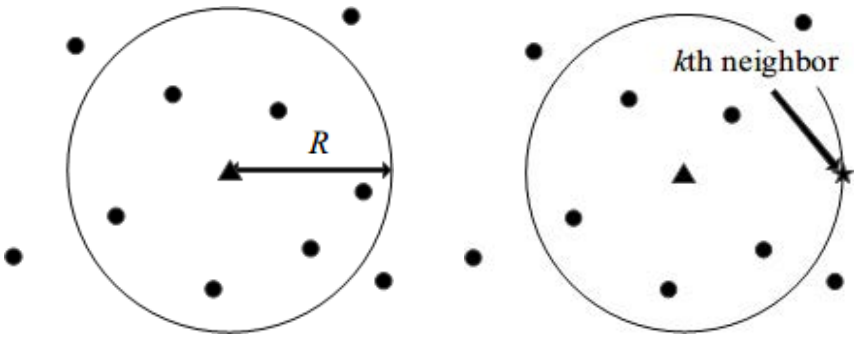


Figure 3.3: An illustration of two different ways to estimate density with a ball shape region $\mathcal{L}(x)$ around the center point, (a) the ball radius $\mathcal{R}$ is imposed, (b) the radius of $\mathcal{L}(x)$ is determined by the distance between kth NN and the center of the ball (in this example, $k = 6$). In this example the considered norm is the standard Euclidean norm.

For KDE approach, the maximum norm, which results in a ball with a cubic shape, is widely used. Formally we can introduce a kernel function $H(u)$, also known as the Parzen window [Parzen 1962], to determine the number k of points falling into the ball. For an unit d-dimensional (hyper-)cube centered at the origin (with an unity edge length) $H(u)$

can be defined as

$$H(u) = \begin{cases} 1, & \text{if } \max\limits_{1 \le i \le d} |u_i| \le \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.54}$$

So, the number k of points x_i included in a (hyper-)cube centered on x and with edge length equal to $\mathcal{R}$ is given by

$$k = \sum_{i=1}^{N} H\left(\frac{x - x_i}{\mathcal{R}}\right). \tag{3.55}$$

Finally, we obtain the following kernel density estimator

$$\widehat{p_X(x)}_{\text{kde}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{V} H\left(\frac{x - x_i}{\mathcal{R}}\right), \tag{3.56}$$

where V is the volume of the region $\mathcal{L}(x)$. For this kind of density estimator, kernel functions $H(\cdot)$ other than the one used in Equ. (3.54) can be chosen. A common choice is the Gaussian kernel function, and the corresponding density estimator is

$$\widehat{p_X(x)}_{\text{kde}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{(2\pi\mathcal{R}^2)^{\frac{d}{2}}} \exp\left(-\frac{(x - x_i)^2}{2\mathcal{R}^2}\right). \tag{3.57}$$

So, with $\widehat{p_X(x)}$ obtained with Equ. (3.56), $\widehat{\mathcal{H}(X)}$ can be calculated with Equ. (3.45). However, for the KDE method, the choice of the fixed width parameter $\mathcal{R}$ is important. If $\mathcal{R}$ is too large, the estimated density is over-smoothed, and, if it is too small, the estimation suffers from too much statistical variability. In Fig. 3.4, three different values of $\mathcal{R}$ are given to highlight their influence on the results.

Now, we consider the kNN approach. Using the kth NN to determine the region $\mathcal{L}(x)$, an unbiased estimator of $p_X(x)$ is given by [Fukunaga 2013]

$$\widehat{p(x)}_{\text{knn}} = \frac{k - 1}{NV}, \tag{3.58}$$

where V is the volume of the neighborhood $\mathcal{L}(x)$. Using this kNN density estimator, we write the entropy estimator as

$$\widehat{\mathcal{H}(X)}_{\text{knn}} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{k - 1}{N v_i}, \tag{3.59}$$

where v_i is the volume of the region $\mathcal{L}(x_i)$ (around center point x_i). Since the estimator in Equ. (3.59) has been proven to be biased [Singh 2003], two other unbiased entropy

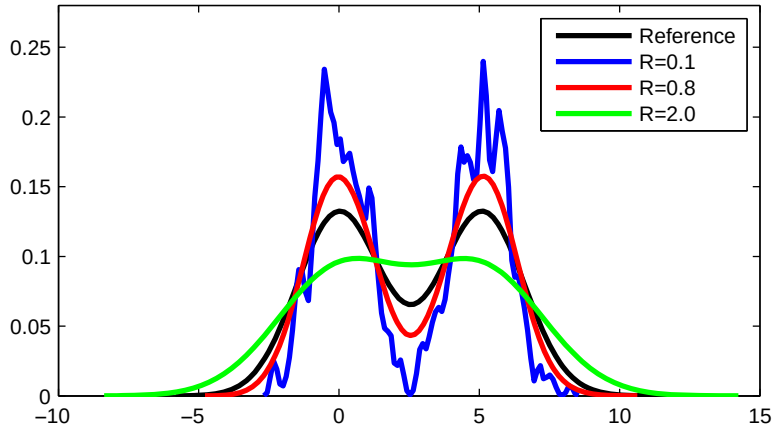


Figure 3.4: An example of using kernel method for density estimation with different widths $\mathcal{R}$ ($\mathcal{R} = 0.1$, $\mathcal{R} = 0.8$ and $\mathcal{R} = 2.0$). Data are generated by the mixture of two equally weighted Gaussian distributions with mean values $\mu_1 = 0$, $\mu_2 = 5$, and $\sigma_1 = \sigma_2 = 1$.

estimators based on the kNN technique have been introduced, as shown below.

Note that these two estimators can be implemented whatever the norm. Now, in this section, when we refer to Fig. 3.6, we only consider the maximum norm, for which the region of interest is a cube (Box ① in Fig. 3.6).

The first kNN entropy estimator is the Kozachenko-Leonenko entropy estimator [Leonenko 2008] (Box ② in Fig 3.6),

$$\widehat{\mathcal{H}(X)}_{\text{kl}} = \psi(N) + \frac{1}{N}\sum_{i=1}^{N}\log(v_i) - \psi(k), \tag{3.60}$$

where v_i is the volume of the smallest ball centered on x_i which includes the k NNs of x_i and $\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)}$ denotes the digamma function.

The second one has been derived by Singh *et al.* in [Singh 2003] and is denoted by $\widehat{\mathcal{H}(X)}_{\text{sg}}$ hereafter (Box ③ in Fig. 3.6),

$$\widehat{\mathcal{H}(X)}_{\text{sg}} = \log(N) + \frac{1}{N}\sum_{i=1}^{N}\log(v_i) - \psi(k), \tag{3.61}$$

where v_i and $\psi(k)$ share the same definition as in Equ. (3.60). Note that, using Equ. (3.60) or (3.61), and for a given chosen norm, the only parameters to be fixed are the number of observations N and the number of neighbors, k. Then, for a given N and a given probability distribution, the choice of k determines all the estimation statistical properties, as bias and variance. Actually, Equ. (3.60) and (3.61) give quite similar results in practical use as, for large N, we have $\log(N) \approx \psi(N)$. These two kNN entropy

estimators (especially the one expressed using Equ. (3.60)) have become popular and have been largely adopted in the estimation of both mutual information and transfer entropy, these two quantities being deeper discussed in the rest of this chapter.

3.2.2. Estimation of Mutual Information

As mentioned previously, mutual information can be computed as a combination of joint and marginal entropies. Let (X, Y) be a pair of multidimensional random variables with a continuous distribution specified by a joint probability density $p_{X,Y}$ with marginal densities p_X and p_Y. Considering the entropy estimator in Equ. (3.45), the mutual information $\mathcal{I}(X, Y)$ can be estimated from a sequence of independent realizations (x_i, y_i) of (X, Y) as

$$\begin{aligned}\widehat{\mathcal{I}(X,Y)} &= \widehat{\mathcal{H}(X)} + \widehat{\mathcal{H}(Y)} - \widehat{\mathcal{H}(X,Y)} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log\widehat{p_X(x_i)} - \frac{1}{N}\sum_{i=1}^{N}\log\widehat{p_Y(y_i)} + \frac{1}{N}\sum_{i=1}^{N}\log\widehat{p_{X,Y}(x_i,y_i)} \\ &= \frac{1}{N}\sum_{i=1}^{N}\log\frac{\widehat{p_{X,Y}(x_i,y_i)}}{\widehat{p_X(x_i)}\,\widehat{p_Y(y_i)}}.\end{aligned} \tag{3.62}$$

In order to calculate the estimator expressed by Equ. (3.62), we can first calculate three individual densities $\widehat{p_X(x_i)}$, $\widehat{p_Y(y_i)}$, and $\widehat{p_{X,Y}(x_i,y_i)}$ separately by using the KDE approach introduced previously. However, KDE remains a hard problem (for instance, the tuning of the kernel) [Suzuki 2008] and could lead to unreliable estimators of the entropic quantities. It is not considered in the scope of this work.

Another possible solution is to calculate $\widehat{\mathcal{H}(X)}$, $\widehat{\mathcal{H}(Y)}$ and $\widehat{\mathcal{H}(X,Y)}$ separately, using the kNN entropy estimators described in Equ. (3.60) or (3.61). In this way, it is possible to adapt the neighborhood determination strategy for each of the three estimators in order to cancel out (more or less) the bias errors in individual estimations and to avoid an adverse accumulation of errors. To this end, Kraskov *et al.* [Kraskov 2004] proposed to use a common neighboring size for both joint and marginal spaces when selecting NNs. This strategy consisted in fixing the number of neighbors in the joint space $\mathcal{S}_{X,Y}$, then projecting the resulting distance (which corresponds, in this space, to the maximum norm distance between the center of the cube and the kth NN) into the marginal spaces $\mathcal{S}_X$ and $\mathcal{S}_Y$, *i.e.* the cubic balls in $\mathcal{S}_X$ and $\mathcal{S}_Y$ have the same radius as the cubic ball in $\mathcal{S}_{X,Y}$ (Box ④ in Fig. 3.6). Following this idea, the following MI estimator was proposed by Kraskov in 2004 [Kraskov 2004] (Box ⑤ in Fig 3.6)

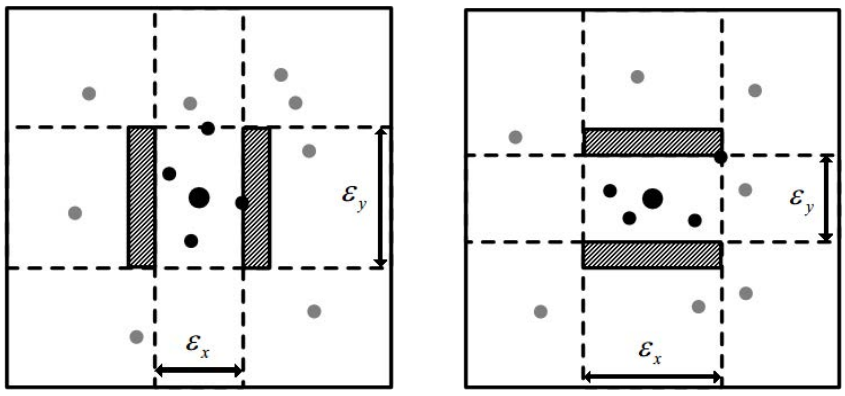
$$\widehat{\mathcal{I}(X,Y)}_{\text{k1}} = \psi(k) - \langle\psi(n_X+1) + \psi(n_Y+1)\rangle + \psi(N), \tag{3.63}$$

where N is the signal length, k is the fixed number of neighbors in $\mathcal{S}_{X,Y}$, $\psi(\cdot)$ denotes the digamma function, the symbol $\langle\cdot\rangle$ stands for an averaging on the sample data set, n_X and n_Y are the numbers of points which fall into the $\mathcal{S}_X$ and $\mathcal{S}_Y$ balls which share a same radius equal to the radius of the $\mathcal{S}_{X,Y}$ ball.

In [Kraskov 2004], Equ. (3.63) was developed with the maximum norm, and one drawback of this estimator was highlighted. As shown in Fig. 3.5, one of the (hyper-)cube determined in $\mathcal{S}_X$ or $\mathcal{S}_Y$ by the distance projection procedure cannot have any point on its border, what is not consistant with the Kozachenko-Leonenko estimation procedure. Therefore, Kraskov *et al.* suggested, for MI estimation, to replace minimal (hyper-)cubes with smaller minimal (hyper-)rectangles equal to the product of two minimal (hyper-)cubes built separately in subspaces $\mathcal{S}_X$ and $\mathcal{S}_Y$ (Box ⑥ in Fig 3.6), to exploit more efficiently the Kozachenko-Leonenko approach. So, considering a (hyper-)rectangle (the development of such a (hyper-)rectangle and the corresponding estimators will be heavily stated in section 3.4.) instead of a (hyper-)cube, Kraskov *et al.* proposed a second MI estimator [Kraskov 2004] (Box ⑦ in Fig. 3.6)

$$\widehat{\mathcal{I}(X,Y)}_{\mathrm{k2}} = \psi(k) - \frac{1}{k} - \langle \psi(n_X) + \psi(n_Y) \rangle + \psi(N). \tag{3.64}$$

According to [Kraskov 2004], Equ. (3.63) and (3.64) give quite comparable results. Based on these kNN MI estimators, an open source toolbox, named MILCA [Sergey 2015], has been made available.



(a) (b)

Figure 3.5: An example of a cube in $\mathcal{S}_{X,Y}$ space (with $d_X = d_Y = 1$) determined by the kth NN with $k = 4$. Denoting (c_x, c_y) the center of the cube in $\mathcal{S}_{X,Y}$, with probability one, only one of the two following cases can arise: (a) $\varepsilon_x < \varepsilon_y$ there is no point on the border of the interval $\left[-c_x + \frac{\varepsilon_y}{2}, c_x + \frac{\varepsilon_y}{2}\right]$, (b) $\varepsilon_x > \varepsilon_y$ there is no point on the border of the interval $\left[-c_y + \frac{\varepsilon_x}{2}, c_y + \frac{\varepsilon_x}{2}\right]$.

In [Kraskov 2004], the effectiveness of this strategy to reduce bias is attested through numerical experiments. This strategy has also been extended to the calculation of other

information theory functionals, such as divergence [Wang 2009] or conditional mutual information [Frenzel 2007].

In [Kraskov 2004], the following interesting conjecture has been raised from simulation results:

$$\mathrm{E}\left[\widehat{\mathcal{I}(X,Y)}_{\mathrm{k1}}\right] = \mathrm{E}\left[\widehat{\mathcal{I}(X,Y)}_{\mathrm{k2}}\right] = 0, \text{ iif } \mathcal{I}(X,Y) = 0. \tag{3.65}$$

In section 3.3, we propose to give some theoretical explanations to justify this result.

3.2.3. Estimation of Transfer Entropy

For the estimation of transfer entropy, using the same notations as in Equ. (3.32), and similar with Equ. (3.62), the estimator of transfer entropy can be written as

$$\begin{aligned}
\widehat{\mathrm{TE}}_{Y\to X} &= \widehat{\mathcal{H}(X^-,Y^-)} + \widehat{\mathcal{H}\left(X^p,X^-\right)} - \widehat{\mathcal{H}\left(X^p,X^-,Y^-\right)} - \widehat{\mathcal{H}(X^-)} \\
&= -\frac{1}{N}\sum_{i=1}^{N}\widehat{\log\left(p_{X^-,Y^-}(x_i^-,y_i^-)\right)} - \frac{1}{N}\sum_{i=1}^{N}\widehat{\log\left(p_{X^p,X^-}(x_i^p,x_i^-)\right)} \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\widehat{\log\left(p_{X^p,X^-,Y^-}(x_i^p,x_i^-,y_i^-)\right)} + \frac{1}{N}\sum_{i=1}^{N}\widehat{\log\left(p_{X^-}(x_i^-)\right)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\log\frac{\widehat{p_{X^p,X^-,Y^-}(x_i^p,x_i^-,y_i^-)}\,\widehat{p_{X^-}(x_i^-)}}{\widehat{p_{X^-,Y^-}(x_i^-,y_i^-)}\,\widehat{p_{X^p,X^-}(x_i^p,x_i^-)}},
\end{aligned} \tag{3.66}$$

where (x_i^p, x_i^-, y_i^-) is an observation (realization) of (X^p, X^-, Y^-). As previously, it is possible to calculate the marginal and joint density probabilities in Equ. (3.66) using KDE approach, and then estimate TE. This method has been adopted by some authors [Sabesan 2007, Yang 2013]. In [Zuo 2013], Zuo *et al.* tried to improve this method by introducing an adaptive bandwidth [Hwang 1994] to estimate the joint density probability with the highest dimension, $\widehat{p_{X^p,X^-,Y^-}(x_i^p,x_i^-,y_i^-)}$.

Inspired from the MI estimators proposed by Kraskov *et al.* (Equ. (3.63) and (3.64)), two different kNN TE estimators have been proposed afterwards. Applying the same strategy to estimate TE, the number of neighbors in the joint space $\mathcal{S}_{X^p,X^-,Y^-}$ is first fixed. Then, for each i, the resulting distance $\varepsilon_i \triangleq d_{(x_i^p,x_i^-,y_i^-),k}$ between (x_i^p, x_i^-, y_i^-) and its kth NN is projected into the three other lower-dimensional spaces, leading to the following TE estimator [Vicente 2011, Lindner 2011, Wibral 2013, Wibral 2014a, Wollstadt 2014, Gómez-Herrero 2015] (Box ⑧ in Fig 3.6)

$$\widehat{\mathrm{TE}}_{Y\to X\,\mathrm{k1}} = \psi(k) + \frac{1}{N}\sum_{i=1}^{N}\left(\psi(n_{X^-,i}+1) - \psi(n_{(X^-,Y^-),i}+1) - \psi(n_{(X^p,X^-),i}+1)\right), \tag{3.67}$$

where $n_{X^-,i}$, $n_{(X^-,Y^-),i}$ and $n_{(X^p,X^-),i}$ denote the number of points which fall into the distance ε_i from x_i^-, $\left(x_i^-, y_i^-\right)$ and $\left(x_i^p, x_i^-\right)$ in the lower-dimensional spaces $\mathcal{S}_{X^-}$, $\mathcal{S}_{X^-,Y^-}$ and $\mathcal{S}_{X^p,X^-}$, respectively. An implementation of this TE estimator is available in the TRENTOOL toolbox [Wollstadt 2015], version 3.0. Another kNN TE estimator is derived from Equ. (3.64) and written as (Box ⑨ in Fig. 3.6)

$$\widehat{\mathrm{TE}}_{Y\to X\,\mathrm{k2}} = \frac{1}{N}\sum_{i=1}^{N}\left(\psi(k) - \frac{2}{k} + \psi(n_{X^-,i}) - \psi(n_{(X^p,X^-),i}) + \frac{1}{n_{(X^p,X^-),i}} - \psi(n_{(X^-,Y^-),i}) + \frac{1}{n_{(X^-,Y^-),i}}\right). \tag{3.68}$$

Similarly as for Equ. (3.64), Equ. (3.68) is based on the idea of rectangle described in Fig. 3.5, and it has been implemented in the JIDT toolbox [Lizier 2014], version 1.2. The same approach inspired by the works of Kraskov has led to propose a kNN approach to estimate DI [Amblard 2014].

Note that, in the following development of this work, transfer entropy estimated by the free TRENTOOL toolbox (corresponding to Equ. (3.67)) is marked as Standard algorithm, and that estimated by JIDT (corresponding to Equ. (3.68)) is marked as Extended algorithm.

3.2.4. Discussion

In this section, different methods for the estimation of information-theoretic quantities have been reported. For the calculation of MI and TE, the most popular approaches are the kNN related methods, and these estimators use similar strategies to reduce bias. Hereafter, we give a short summary on these strategies before raising some fundamental questions. As discussed in section 3.1.2, the estimators of MI and TE present comparable structures, so that we consider here only mutual information but the same reasoning applies to transfer entropy.

Given

$$\widehat{\mathcal{I}(X,Y)} = \widehat{\mathcal{H}(X)} + \widehat{\mathcal{H}(Y)} - \widehat{\mathcal{H}(X,Y)}, \tag{3.69}$$

if we calculate the marginal and joint entropies on the right-hand side of Equ. (3.69) separately, this introduces estimation bias for each term, denoted by $\mathcal{B}_H(X)$, $\mathcal{B}_H(Y)$ and $\mathcal{B}_H(X,Y)$ respectively. Then the bias for $\widehat{\mathcal{I}(X,Y)}$ is expressed as

$$\mathcal{B}_I(X,Y) = \mathcal{B}_H(X) + \mathcal{B}_H(Y) - \mathcal{B}_H(X,Y). \tag{3.70}$$

Our goal is to reduce $\mathcal{B}_I(X,Y)$ as much as possible. To this end, there are two basic

ideas.

(1) First, we can choose proper parameters for the estimation of individual entropies, and try to tend towards the following approximation

$$\mathcal{B}_H(X) + \mathcal{B}_H(Y) \approx \mathcal{B}_H(X, Y). \tag{3.71}$$

In this case, the individual estimation bias would be cancelled out, and $\mathcal{B}_I(X, Y)$ would vanish to zero. The MI estimator in Equ. (3.63) and the TE estimator (Equ. (3.67) follow this idea. To apply this strategy, Kraskov proposed to obtain the distance in the joint space $\mathcal{S}_{X,Y}$, and then project the distance obtained into the marginal spaces $\mathcal{S}_X$ and $\mathcal{S}_Y$. It should be mentioned that, until now, this strategy has been only implemented with the maximum norm.

(2) A second basic idea is to reduce $\mathcal{B}_H(X)$, $\mathcal{B}_H(Y)$ and $\mathcal{B}_H(X, Y)$ as much as possible but in a separate manner. For this purpose, we can calculate the individual entropies using the (hyper-)rectangle region instead of a (hyper-)cube. For the MI estimator in Equ. (3.64) (*resp.* Equ. (3.68) for TE), this idea is applied together with the first one. Of course, in this case, it is impossible to support completely the first idea, because, for a (hyper-)rectangle, the side lengths for different dimensions may be different.

However, there are still several questions to be solved with these strategies.

Firstly, for the point (1) mentioned above (marked as a red arrow in Fig. 3.6), there are three questions to answer. Firstly, the effectiveness of the strategy proposed in [Kraskov 2004] is verified only with numerical experiments. There is currently a lack of theoretical explanation and a deeper analysis of the bias in the entropy estimation is required. For the moment, the most popular bias analysis approach is dedicated to the Edgeworth expansion [Van Hulle 2005], and it could also be used in the estimation of mutual information. However, this method is not suitable for the investigation mentioned here: (a) this method uses the Gaussian distribution and some additional correction terms to approximate the entropy of a distribution, and the size-related parameters (bandwidth for KDE, or number of neighbors for kNN) are not involved in this analysis, (b) according to previous studies [Suzuki 2008], the Edgeworth MI estimator is accurate when the underlying distributions are close to Gaussian distribution, and becomes unreliable when the distributions are far from Gaussian.

Secondly, as pointed out in [Kraskov 2004], the two MI estimators could provide accurate results for $\widehat{\mathcal{I}(X, Y)}$ only when X and Y are independent. The question is now: is it possible to improve the results when X and Y are dependent?

Thirdly, the two kNN estimators (Equ. (3.63) for MI, and Equ. (3.67) for TE) are

only implemented with the maximum norm (MI estimator in MILCA [Sergey 2015], and TE estimator in TRENTOOL [Wollstadt 2014]), where the distances are obtained in the joint space with the highest dimension, and these distances are then projected into the other spaces. Is it possible to apply this strategy to other norms, for instance, the Euclidean one?

These three questions are covered in section 3.3 (see the blue dotted box in Fig. 3.6). Firstly, a new analytical form of bias for the plug-in entropy estimator is introduced (Box ⑩), and using this result, a relation leading to an optimal distance is developed for the estimation of mutual information and transfer entropy for two norms (Euclidean norm and maximum norm, Boxes ⑫ and ⑬ respectively). In the case of the maximum norm, this relation (Box ⑬) actually provides the theoretical explanation for the kNN MI and TE mentioned previously (Boxes ⑤ and ⑧). In Box ⑫, this relation is extended to Euclidean norm. With the relations in Boxes ⑫ and ⑬, we developed the "basic" estimators for both MI (Box ⑭) and TE (Box ⑮), using the kNN density estimator (Box ⑯). Note that the optimized distance relations are developed only under the independence assumption. To futher eliminate the bias for dependent X and Y, novel MI and TE estimators, named "mixed" estimators, are also introduced (Boxes ⑱ and ⑲).

For point (2) (marked as a blue arrow in Fig. 3.6), there is one extra question. In [Kraskov 2004], the development for the idea of product of cubes is based on the Kozachenko-Leonenko entropy estimator (Equ. (3.60)) using the maximum norm. After some extra mathematical development, this idea can be extended to a more general case (idea of rectangle, Box ⑳ in Fig. 3.6), where the determined region in the joint space is a rectangle (the side lengths in two different dimensions can be different). Is it possible to derive the new kNN entropy estimator with the idea of rectangle based on Singh's entropy estimator (Equ. (3.61))? This point is discussed in section 3.4 (marked as a red dotted box in Fig. 3.6). The standard kNN methods using maximum norm for probability density estimation and entropy non-parametric estimation introduce, around each data point, a minimal (hyper-)cube (Box ①), which includes the first k NNs, as it is the case for two already developed entropy estimators, namely the well-known Kozachenko-Leonenko estimator (Box ②) and the less commonly used Singh's estimator (Box ③). The idea of rectangle extends the idea of the product of cubes (Box ⑥). It consists in proposing a different construction of the neighborhoods, which are no longer minimal (hyper-)cubes, nor products of (hyper-)cubes, but minimal (hyper-)rectangles (Box ⑳), with possibly a different length for each dimension, to get two novel entropy estimators (Boxes ㉑ and ㉒), respectively derived from the Kozachenko-Leonenko entropy estimator and Singh's entropy estimator . These two new entropy estimators lead respectively to two new TE estimators (Boxes ㉓ and ㉔) to be compared with the Standard and

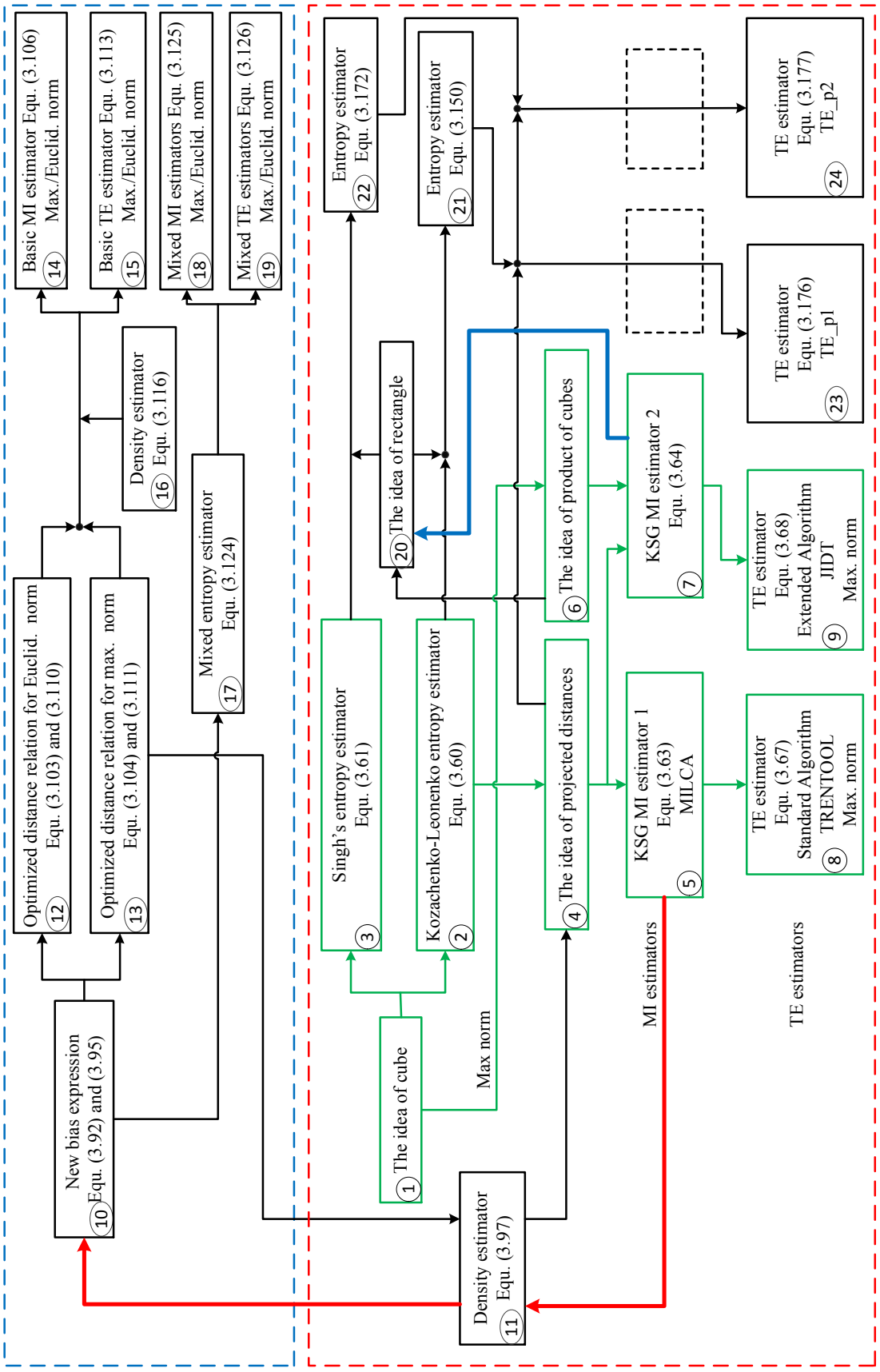


Figure 3.6: Synthetic overview of the different concepts and estimators. References to section 3.2 are marked in green.

Extended TE estimators.

3.3. First Improvement

This section [1] deals with the control of estimation bias when estimating mutual information or transfer entropy from non-parametric approach. We focus on continuously distributed random data and the estimators we developed are based on non-parametric kNN approach for arbitrary metrics. Using a multidimensional Taylor series expansion, a general relationship between the estimation error bias and neighboring size for plug-in entropy estimator is established without any assumption on the data for two different norms. The theoretical analysis based on the maximum norm developed coincides with the experimental results drawn from numerical tests made by Kraskov *et al.* [Kraskov 2004]. To further validate the novel relation, a weighted linear combination of distinct mutual information estimators is proposed.

3.3.1. New Bias Expression for the Plug-in Entropy Estimator

In the following development, we consider a d_X-dimensional random variable X whose outcomes are in $\mathbb{R}^{d_X}$ and with a probability distribution specified by the probability density function $p_X(x)$. $\mathcal{L}(x)$ standing for a small region around x in $\mathbb{R}^{d_X}$, we introduce the volume (Lebesgue measure) of $\mathcal{L}(x)$

$$v(x) = \int_{\mathcal{L}(x)} \mathrm{d}u. \tag{3.72}$$

As mentioned in section 3.2.1, in most existing density estimation algorithms, including either KDE with the Parzen window or kNN, $p_X(x)$ is estimated as

$$\begin{aligned} \widehat{p_X}(x) &= \frac{\widehat{P\left(X \in \mathcal{L}(x)\right)}}{v(x)} \\ &= \frac{\widehat{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}}{v(x)}, \end{aligned} \tag{3.73}$$

where $\widehat{P\left(X \in \mathcal{L}(x)\right)}$ corresponds to an estimation of the probability that X belongs to the set $\mathcal{L}(x)$. If we assume that $P\left(X \in \mathcal{L}(x)\right)$ is perfectly known (but not $p_X(x)$), we

[1] This first improvement was the subject of our contribution in [Zhu 2014, Zhu 2015b].

can use the following approximation

$$\begin{aligned} \log p_X(x) &\approx \log\left(\frac{P\left(X \in \mathcal{L}\left(x\right)\right)}{v\left(x\right)}\right) \\ &= \log\left(\frac{\int_{\mathcal{L}(x)} p_X\left(y\right) \mathrm{d}y}{v\left(x\right)}\right). \end{aligned} \tag{3.74}$$

Given Equ. (3.73), an estimation $\widehat{\log p_X(x)}$ of $\log p_X(x)$ is introduced

$$\begin{aligned} \widehat{\log p_X(x)} &= \log \widehat{p_X}\left(x\right) \\ &= \log \frac{\widehat{P\left(X \in \mathcal{L}(x)\right)}}{v(x)} \\ &= \log\left(\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} + \varepsilon\right), \end{aligned} \tag{3.75}$$

where the random estimation error ε given by

$$\varepsilon = \frac{\widehat{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}}{v(x)} - \frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} \tag{3.76}$$

is zero mean when $\widehat{P\left(X \in \mathcal{L}(x)\right)}$ is unbiased.

From observations X_i (random variables) issued from P_X, the corresponding differential entropy $\mathcal{H}(X)$ can be estimated as

$$\widehat{\mathcal{H}(X)} = -\frac{1}{N}\sum_{i=1}^{N} \widehat{\log p_X}(X_i), \tag{3.77}$$

where N is the number of data used in the averaging. Then, we approximate the probability density $p_X(y)$ using a second-order Taylor approximation around x,

$$p_X(y) \approx p_X(x) + \left(\frac{\partial p_X}{\partial x}(x)\right)^T (y-x) + \frac{1}{2}(y-x)^T \left(\frac{\partial^2 p_X}{\partial x^2}(x)\right)(y-x), \tag{3.78}$$

with the superscript T standing for matrix transposition, and analyze the bias of $\widehat{\mathcal{H}(X)}$ with

$$\begin{aligned} \widehat{\mathcal{H}(X)} &= -\frac{1}{N}\sum_{i=1}^{N} \log \widehat{p_X}(X_i) \\ &= -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\int_{\mathcal{L}(X_i)} p_X(y)\mathrm{d}y}{v(X_i)} + \varepsilon_i\right), \end{aligned} \tag{3.79}$$

where the index i refers to the sample number. Integrating Equ. (3.78) on both sides

and dividing by $v(x)$, we get

$$\begin{aligned}\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} \approx p_X(x) + \left(\frac{\partial p_X}{\partial x}(x)\right)^T \frac{1}{v(x)} \int\limits_{\mathcal{L}(x)} (y-x)\mathrm{d}y \\ + \frac{1}{2v(x)} \int\limits_{\mathcal{L}(x)} (y-x)^T \left(\frac{\partial^2 p_X}{\partial x^2}(x)\right)(y-x)\mathrm{d}y.\end{aligned} \tag{3.80}$$

If we assume that $\mathcal{L}(x)$ admits x as a center of symmetry, then

$$\int\limits_{\mathcal{L}(x)} (y-x)\mathrm{d}y = 0 \tag{3.81}$$

and the first-order term on the right-hand side of Equ. (3.80) is zero. According to the property $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ of the trace operator, denoted $\mathrm{tr}(\cdot)$, applied to the product of two matrices A and B, Equ. (3.80) can be transformed into

$$\begin{aligned}\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} &\approx p_X(x) + \frac{1}{2v(x)} \int\limits_{\mathcal{L}(x)} (y-x)^T \left(\frac{\partial^2 p_X}{\partial x^2}(x)\right)(y-x)\mathrm{d}y \\ &= p_X(x) + \frac{1}{2v(x)} \mathrm{tr}\left(\int\limits_{\mathcal{L}(x)} (y-x)(y-x)^T \mathrm{d}y \frac{\partial^2 p_X}{\partial x^2}(x)\right).\end{aligned} \tag{3.82}$$

Finally, the estimator $\widehat{\log p_X}(x)$ of $\log p_X(x)$ is approximated by

$$\begin{aligned}&\log\left(\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} + \varepsilon\right) \\ \approx &\log\left(p_X(x) + \frac{1}{2v(x)}\mathrm{tr}\left(\left(\int\limits_{\mathcal{L}(x)} (y-x)(y-x)^T\mathrm{d}y\right)\left(\frac{\partial^2 p_X}{\partial x^2}(x)\right)\right) + \varepsilon\right) \\ \approx &\log p_X(x) + \underbrace{\frac{1}{p_X(x)}\frac{1}{2v(x)}\mathrm{tr}\left(\left(\int\limits_{\mathcal{L}(x)} (y-x)(y-x)^T\mathrm{d}y\right)\left(\frac{\partial^2 p_X}{\partial x^2}(x)\right)\right)}_{\approx \mathcal{B}_X} + \frac{1}{p_X(x)}\varepsilon,\end{aligned} \tag{3.83}$$

where the term $\left(\frac{1}{p_X(x)} \cdot \varepsilon\right)$ is zero mean.

The bias $\mathcal{B}_X$ in $\widehat{\mathcal{H}(X)}$ is approximated by the second term in the right-hand side of Equ. (3.83) and used as a correcting term. To build $\mathcal{L}(x)$ which admits x as a center of symmetry, we retain two norms, the Euclidean norm ($\|\cdot\| = \|\cdot\|_{\mathrm{E}}$) and the maximum

norm ($\|\cdot\| = \|\cdot\|_{\mathrm{M}}$):

$$\mathcal{L}(x) = \{y : \|y - x\| \leq \mathcal{R}(x)\}. \tag{3.84}$$

The resulting domain corresponds respectively to a standard ball and to a d_X dimensional cube (a cubic ball). Consequently, the value $\mathcal{R}(x)$ fixes respectively the radius of the standard ball or the half of the edge length of the cube.

Note that $\int_{\mathcal{L}(x)} (y-x)\,(y-x)^T \mathrm{d}y$ is a diagonal matrix, which can be expressed as $\mathcal{T} \cdot I$, where I is the identity matrix and $\mathcal{T}$ is a scalar independent of x. Therefore, we can move $\mathcal{T}$ out of the $\mathrm{tr}(\cdot)$ function

$$\begin{aligned}
\mathcal{B}_X &\approx \frac{1}{p_X(x)} \frac{1}{2v(x)} \mathrm{tr}\left(\left(\int_{\mathcal{L}(x)} (y-x)(y-x)^T \mathrm{d}y\right)\left(\frac{\partial^2 p_X}{\partial x^2}(x)\right)\right) \\
&= \frac{1}{p_X(x)} \frac{1}{2v(x)} \mathrm{tr}\left((\mathcal{T} \cdot I) \cdot \left(\frac{\partial^2 p_X}{\partial x^2}(x)\right)\right) \\
&= \left(\frac{\mathcal{T}}{2v(x)}\right) \cdot \frac{1}{p_X(x)} \cdot \mathrm{tr}\left(\frac{\partial^2 p_X}{\partial x^2}(x)\right).
\end{aligned} \tag{3.85}$$

Now, we derive an expression for $\frac{\mathcal{T}}{2v(x)}$ in Equ. (3.85). Assume a conventional quadratic distance function such as

$$\mathcal{R}^2(x, y) = (y-x)^T A (y-x). \tag{3.86}$$

Using d_X-dimensional spherical coordinates, we have (Equ. (14) in [Fukunaga 1973])

$$\int_{\mathcal{L}(x)} (y-x)(y-x)^T \mathrm{d}y = \frac{1}{(d_X+2)\pi} \Gamma^{\frac{2}{d_X}}\left(\frac{d_X+2}{2}\right) v^{1+\frac{2}{d_X}}(x) |A|^{\frac{1}{d_X}} A^{-1}, \tag{3.87}$$

where $\Gamma(\cdot)$ is the Gamma function, and

$$\begin{aligned}
v(x) &= \int_{\mathcal{L}(x)} \mathrm{d}x \\
&= \frac{\pi^{\frac{d_X}{2}} \mathcal{R}^{d_X}(x)}{|A|^{\frac{1}{2}} \Gamma\left(\frac{d_X+2}{2}\right)}.
\end{aligned} \tag{3.88}$$

From Equ. (3.87) and (3.88), we have

$$\frac{1}{2v(x)} \cdot \int_{\mathcal{L}(x)} (y-x)(y-x)^T \mathrm{d}y = \frac{\mathcal{R}^2(x)}{2(d_X+2)} \cdot A^{-1}. \tag{3.89}$$

The derivation of Equ. (3.89) can be found in Appendix A. For normal Euclidean norm, $A = I$. So

$$\frac{1}{2v(x)} \cdot \int_{\mathcal{L}(x)} (y-x)(y-x)^T \mathrm{d}y = \frac{\mathcal{R}^2(x)}{2(d_X+2)} \cdot I, \tag{3.90}$$

and finally

$$\frac{\mathcal{T}}{2v(x)} = \frac{\mathcal{R}^2(x)}{2(d_X+2)}. \tag{3.91}$$

Therefore, for Euclidean norm, $\mathcal{B}_X$ could be approximated as (Box ⑩ in Fig. 3.6)

$$\mathcal{B}_X(x) \approx \frac{\mathcal{R}^2(x)}{2(d_X+2)} \cdot \frac{1}{p_X(x)} \cdot \mathrm{tr}\left(\frac{\partial^2 p_X}{\partial x^2}(x)\right). \tag{3.92}$$

Now, let us consider the maximum norm, for which the region $\mathcal{L}(x)$ is a d-dimensional (hyper-)cube with side length $2\mathcal{R}(x)$. Also, due to the symmetry of the region, we have

$$\begin{aligned} &\frac{1}{2v(x)} \cdot \int_{\mathcal{L}(x)} (y-x)(y-x)^T \mathrm{d}y \\ &= \frac{1}{2(2\mathcal{R}(x))^{d_X}} \cdot \left(\begin{array}{c|c} \frac{u^3}{3} & \mathcal{R}(x) \\ & -\mathcal{R}(x) \end{array}\right) \cdot (2\mathcal{R}(x))^{d_X-1} \cdot I \\ &= \frac{\mathcal{R}(x)^2}{6} \cdot I, \end{aligned} \tag{3.93}$$

which results in

$$\frac{\mathcal{T}}{2v(x)} = \frac{\mathcal{R}^2(x)}{6}. \tag{3.94}$$

So, using the maximum norm distance, the bias $\mathcal{B}_X$ can be approximated as (Box ⑩ in Fig. 3.6)

$$\mathcal{B}_X(x) \approx \frac{\mathcal{R}^2(x)}{6} \cdot \frac{1}{p_X(x)} \cdot \mathrm{tr}\left(\frac{\partial^2 p_X}{\partial x^2}(x)\right). \tag{3.95}$$

Note that, with the second-order approximation, the bias $\mathcal{B}_X$ increases with larger $\mathcal{R}(x)$ whatever the norm.

Until now, no particular form of density estimator was specified in our bias analysis. In [Kraskov 2004], the Kozachenko–Leonenko estimator [Kozachenko 1987] (Equ. (3.60)) does not use explicitly an estimation of the densities for each sample point. However, since $\psi(N) \approx \log(N)$ for large N, Equ. (3.60) has the following structure

$$\widehat{\mathcal{H}(U)} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{e^{\psi(k)}}{N v_i}\right), \tag{3.96}$$

where the term inside the brackets can be interpreted as a density probability estimation. Using the kNN entropy estimator proposed by Singh (Equ. (3.61)), Equ. (3.96) could be directly derived. Therefore, using Equ. (3.75), (3.79) and (3.96), we consider the following generic density estimator (Box ⑪ in Fig. 3.6)

$$\widehat{p_U(u_i)} = \frac{e^{\psi(k)}}{N v_i}, \tag{3.97}$$

and the kNN entropy estimators mentioned previously (Equ. (3.60) and (3.61)) can be considered as estimators with the same structure as in Equ. (3.79), and so can be discussed under our bias analysis framework.

3.3.2. Bias Reduction of MI/TE Estimators Based on the New Bias Expression

Coming back to the estimation of MI, we can try to decrease its bias by subtracting the bias term highlighted in Equ. (3.83):

$$\begin{aligned}\widehat{\mathcal{I}(X,Y)} = -\frac{1}{N}\sum_{i=1}^{N}\Big(&\log \widehat{p_X}(x_i) + \log \widehat{p_Y}(y_i) - \log \widehat{p_{X,Y}}(x_i,y_i) \\ &- \big(\mathcal{B}_X(x_i) + \mathcal{B}_Y(y_i) - \mathcal{B}_{X,Y}(x_i,y_i)\big)\Big).\end{aligned} \tag{3.98}$$

If X and Y are independent, we obtain for each i:

$$\frac{\mathrm{tr}\left(\frac{\partial^2 p_{X,Y}}{\partial (x,y)^2}(x_i,y_i)\right)}{p_{X,Y}(x_i,y_i)} = \frac{\mathrm{tr}\left(\frac{\partial^2 p_X}{\partial x^2}(x_i)\right)}{p_X(x_i)} + \frac{\mathrm{tr}\left(\frac{\partial^2 p_Y}{\partial y^2}(y_i)\right)}{p_Y(y_i)}. \tag{3.99}$$

In this case, we impose relationship-specific distances for different entropy estimations in Equ. (3.69) to cancel out the bias, *i.e.*

$$\mathcal{B}_X(x_i) + \mathcal{B}_Y(y_i) - \mathcal{B}_{X,Y}(x_i,y_i) = 0. \tag{3.100}$$

With the help of Equ. (3.99), for the Euclidean norm, the left side of Equ. (3.100) can be transformed into

$$\begin{aligned}\mathcal{B}_X(x_i) + \mathcal{B}_Y(y_i) - \mathcal{B}_{X,Y}(x_i,y_i) = &\left(\frac{\mathcal{R}^2(x_i)}{2(d_X+2)} - \frac{\mathcal{R}^2(x_i,y_i)}{2(d_{X,Y}+2)}\right)\cdot\frac{\mathrm{tr}\left(\frac{\partial^2 p_X}{\partial x^2}(x_i)\right)}{p_X(x_i)} \\ &+ \left(\frac{\mathcal{R}^2(y_i)}{2(d_Y+2)} - \frac{\mathcal{R}^2(x_i,y_i)}{2(d_{X,Y}+2)}\right)\cdot\frac{\mathrm{tr}\left(\frac{\partial^2 p_Y}{\partial y^2}(y_i)\right)}{p_Y(y_i)},\end{aligned} \tag{3.101}$$

where d_X, d_Y and $d_{X,Y}$ are the dimensions of the signals X, Y and (X,Y), and $\mathcal{R}(x_i)$, $\mathcal{R}(y_i)$ and $\mathcal{R}(x_i,y_i)$ are the distances used for the estimation of $\widehat{p_X}(x_i)$, $\widehat{p_Y}(y_i)$ and $\widehat{p_{X,Y}}(x_i,y_i)$, respectively. Here, $p_X(x_i)$, $p_Y(y_i)$ and their corresponding second-order derivatives are unknown, so to get Equ. (3.101) equal to zero, the following sufficient pair of conditions is derived

$$\begin{cases} \dfrac{\mathcal{R}^2(x_i)}{2(d_X+2)} = \dfrac{\mathcal{R}^2(x_i,y_i)}{2(d_{X,Y}+2)} \\ \dfrac{\mathcal{R}^2(y_i)}{2(d_Y+2)} = \dfrac{\mathcal{R}^2(x_i,y_i)}{2(d_{X,Y}+2)}. \end{cases} \tag{3.102}$$

Finally, Equ. (3.102) yields to (Box ⑫ in Fig. 3.6)

$$\begin{cases} \mathcal{R}(x_i) = \sqrt{\dfrac{d_X+2}{d_{X,Y}+2}} \cdot \mathcal{R}(x_i,y_i) \\ \mathcal{R}(y_i) = \sqrt{\dfrac{d_Y+2}{d_{X,Y}+2}} \cdot \mathcal{R}(x_i,y_i). \end{cases} \tag{3.103}$$

Similarly, using the maximum norm, we obtain (Box ⑬ in Fig. 3.6)

$$\begin{cases} \mathcal{R}(x_i) = \mathcal{R}(x_i,y_i) \\ \mathcal{R}(y_i) = \mathcal{R}(x_i,y_i). \end{cases} \tag{3.104}$$

Equ. (3.104) formally confirms (as suggested but not proved in [Kraskov 2004]) that, if X and Y are independent, using the maximum norm and constraining the values $\mathcal{R}(x_i)$ and $\mathcal{R}(y_i)$ to be equal to $\mathcal{R}(x_i,y_i)$ allows to decrease the bias $\widehat{\mathcal{I}(X,Y)} - \mathcal{I}(X,Y)$. Equ. (3.103) extends this result when the Euclidean norm is used for the 3 individual spaces. We should mention that Equ. (3.99) no longer holds if signals X and Y are not independent. In this case, only a part of the bias can be expected to be cancelled out.

So, finally, in the case of independence between X and Y, we introduced the following MI estimator

$$\widehat{\mathcal{I}(X,Y)} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log \widehat{p_X}(x_i) + \log \widehat{p_Y}(y_i) - \log \widehat{p_{X,Y}}(x_i,y_i)\right) \tag{3.105}$$

with an (approximately) zero bias by choosing $\mathcal{R}(x_i,y_i)$ and properly defining $\mathcal{R}(x_i)$ and $\mathcal{R}(y_i)$ using Equ. (3.103) or (3.104). When $\mathcal{R}(x_i,y_i)$ results from the kNN approach (*i.e.* when $\mathcal{R}(x_i,y_i) = \|k\text{NN}(x_i,y_i) - (x_i,y_i)\|$ is the distance from (x_i,y_i) to its kth

NN, also denoted $\mathcal{R}_k(x_i, y_i)$), this estimator is denoted by $\widehat{\mathcal{I}(X,Y)}^k_{\text{basic}}$ with

$$\begin{aligned}\widehat{\mathcal{I}(X,Y)}^k_{\text{basic}} &= \widehat{\mathcal{H}(X)}_{\text{basic}} + \widehat{\mathcal{H}(Y)}_{\text{basic}} - \widehat{\mathcal{H}(X,Y)}_{\text{basic}} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{k(x_i)-1}{N\cdot v(x_i)} + \log\frac{k(y_i)-1}{N\cdot v(y_i)} - \log\frac{k(x_i,y_i)-1}{N\cdot v(x_i,y_i)}\right),\end{aligned} \tag{3.106}$$

where $k(x_i, y_i) = k$ (Box ⑭ in Fig. 3.6). Hereafter, this estimator is written as $\widehat{\mathcal{I}(X,Y)}^k_{\text{basic,E}}$ for the Euclidean norm and by $\widehat{\mathcal{I}(X,Y)}^k_{\text{basic,M}}$ for the maximum norm, and called "basic estimator".

Now, we consider the estimation of transfer entropy. Similarly with Equ. (3.98), Equ. (3.34) can be rewritten as

$$\begin{aligned}\widehat{\text{TE}_{Y\to X}} = -\frac{1}{N}\sum_{i=1}^{N}\Bigg(&\log\widehat{p_{X_i^-,Y_i^-}}(x_i^-,y_i^-) + \log\widehat{p_{X_i^p,X_i^-}}(x_i^p,x_i^-) - \log\widehat{p_{X_i^-}}(x_i^-) \\ &- \log\widehat{p_{X_i^p,X_i^-,Y_i^-}}(x_i^p,x_i^-,y_i^-) \\ &- \left(\mathcal{B}_{X_i^-,Y_i^-}(x_i^-,y_i^-) + \mathcal{B}_{X_i^p,X_i^-}(x_i^p,x_i^-) - \mathcal{B}_{X_i^p,X_i^-,Y_i^-}(x_i^p,x_i^-,y_i^-) - \mathcal{B}_{X_i^-}(x_i^-)\right)\Bigg).\end{aligned} \tag{3.107}$$

In the same way as for mutual information, we impose relationship-specific distances for different entropy estimations in Equ. (3.107) to cancel out the bias, and obtain the following relation

$$\mathcal{B}_{X_i^-,Y_i^-}(x_i^-,y_i^-) + \mathcal{B}_{X_i^p,X_i^-}(x_i^p,x_i^-) - \mathcal{B}_{X_i^p,X_i^-,Y_i^-}(x_i^p,x_i^-,y_i^-) - \mathcal{B}_{X_i^-}(x_i^-) = 0. \tag{3.108}$$

If the three random variables X_i^p, X_i^-, Y_i^- are mutually independent, after calculation, for the Euclidean norm, we have

$$\begin{cases}\dfrac{\mathcal{R}^2(x_i^-,y_i^-)}{2(d_{X_i^-,Y_i^-}+2)} = \dfrac{\mathcal{R}^2(x_i^p,x_i^-,y_i^-)}{2(d_{X_i^p,X_i^-,Y_i^-}+2)} \\ \dfrac{\mathcal{R}^2(x_i^p,x_i^-)}{2(d_{X_i^p,X_i^-}+2)} = \dfrac{\mathcal{R}^2(x_i^p,x_i^-,y_i^-)}{2(d_{X_i^p,X_i^-,Y_i^-}+2)} \\ \dfrac{\mathcal{R}^2(x_i^-)}{2(d_{X_i^-}+2)} = \dfrac{\mathcal{R}^2(x_i^p,x_i^-,y_i^-)}{2(d_{X_i^p,X_i^-,Y_i^-}+2)},\end{cases} \tag{3.109}$$

where $\mathcal{R}(x_i^-, y_i^-)$, $\mathcal{R}(x_i^p, x_i^-)$, $\mathcal{R}(x_i^p, x_i^-, y_i^-)$ and $\mathcal{R}(x_i^-)$ are the distances used for the estimation of $\widehat{p_{X_i^-,Y_i^-}}(x_i^-, y_i^-)$, $\widehat{p_{X_i^p,X_i^-}}(x_i^p, x_i^-)$, $\widehat{p_{X_i^p,X_i^-,Y_i^-}}(x_i^p, x_i^-, y_i^-)$ and $\widehat{p_{X_i^-}}(x_i^-)$ at the ith point, $d_{(\cdot)}$ is the dimension of the corresponding space. After simplification, Equ. (3.109) leads to (Box ⑫ in Fig. 3.6)

$$\begin{cases} \mathcal{R}(x_i^-, y_i^-) = \sqrt{\dfrac{d_{X_i^-,Y_i^-}+2}{d_{X_i^p,X_i^-,Y_i^-}+2}} \cdot \mathcal{R}(x_i^p, x_i^-, y_i^-) \\ \mathcal{R}(x_i^p, x_i^-) = \sqrt{\dfrac{d_{X_i^p,X_i^-}+2}{d_{X_i^p,X_i^-,Y_i^-}+2}} \cdot \mathcal{R}(x_i^p, x_i^-, y_i^-) \\ \mathcal{R}(x_i^-) = \sqrt{\dfrac{d_{X_i^-}+2}{d_{X_i^p,X_i^-,Y_i^-}+2}} \cdot \mathcal{R}(x_i^p, x_i^-, y_i^-). \end{cases} \tag{3.110}$$

For the maximum norm, the relation becomes (Box ⑬ in Fig. 3.6)

$$\begin{cases} \mathcal{R}(x_i^-, y_i^-) = \mathcal{R}(x_i^p, x_i^-, y_i^-) \\ \mathcal{R}(x_i^p, x_i^-) = \mathcal{R}(x_i^p, x_i^-, y_i^-) \\ \mathcal{R}(x_i^-) = \mathcal{R}(x_i^p, x_i^-, y_i^-). \end{cases} \tag{3.111}$$

Equ. (3.111) corresponds to the TE estimator (Equ. (3.67)) adapted in the TRENTOOL toolbox [Wollstadt 2015], where the distances are obtained in the joint space with the highest dimension, $\mathcal{S}_{X^p,X^-,Y^-}$, and then projected into the other spaces, $\mathcal{S}_{X^-,Y^-}$, $\mathcal{S}_{X^p,X^-}$ and $\mathcal{S}_{X^-}$. Now we must highlight here that the independence condition between the variables X_i^p, X_i^-, Y_i^-, which is equivalent to the following relations set (the two first conditions in Equ. (3.112) being redundant as being implied by the third one)

$$\begin{cases} p_{X_i^-,Y_i^-}(x_i^-, y_i^-) = p_{X_i^-}(x_i^-) p_{Y_i^-}(y_i^-) \\ p_{X_i^p,X_i^-}(x_i^p, x_i^-) = p_{X_i^p}(x_i^p) p_{X_i^-}(x_i^-) \\ p_{X_i^p,X_i^-,Y_i^-}(x_i^p, x_i^-, y_i^-) = p_{X_i^p}(x_i^p) p_{X_i^-}(x_i^-) p_{Y_i^-}(y_i^-) \end{cases} \tag{3.112}$$

is difficult to justify in a context of causality analysis. Indeed this independence condition implies than the TE value is equal to zero but is not necessary to obtain this zero value. More precisely the second equality in Equ. (3.112) amounts to say that X is a white sequence, what is not generally the case. So actually the TE estimator in Equ. (3.67) seems to us more difficult to justify. To conclude about this question, only a part of the bias can be expected to be cancelled out if Equ. (3.112) is not satisfied.

In the same manner as previously (Equ. (3.106)), we can also define the basic esti-

mator $\widehat{\mathrm{TE}_{Y\to X}}^{k}_{\text{basic}}$ for transfer entropy (Box ⑮ in Fig. 3.6)

$$\begin{aligned}\widehat{\mathrm{TE}_{Y\to X}}^{k}_{\text{basic}} &= \widehat{\mathcal{H}(X^-,Y^-)}_{\text{basic}} + \widehat{\mathcal{H}(X^p,X^-)}_{\text{basic}} - \widehat{\mathcal{H}(X^p,X^-,Y^-)}_{\text{basic}} - \widehat{\mathcal{H}(X^-)}_{\text{basic}} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{k(x_i^-,y_i^-)-1}{N\cdot v(x_i^-,y_i^-)} + \log\frac{k(x_i^p,x_i^-)-1}{N\cdot v(x_i^p,x_i^-)}\right. \\ &\qquad\left. - \log\frac{k(x_i^p,x_i^-,y_i^-)-1}{N\cdot v(x_i^p,x_i^-,y_i^-)} - \log\frac{k(x_i^-)-1}{N\cdot v(x_i^-)}\right),\end{aligned} \tag{3.113}$$

where $k(x_i^p,x_i^-,y_i^-)=k$. Hereafter, this estimator is written as $\widehat{\mathrm{TE}_{Y\to X}}^{k}_{\text{basic,E}}$ for the Euclidean norm and by $\widehat{\mathrm{TE}_{Y\to X}}^{k}_{\text{basic,M}}$ for the maximum norm.

Note that, this development for the strategy of relation-specific distances can be generalized to the estimation of entropy combinations other than MI and TE, such as CTE [Yang 2012].

3.3.3. Bias Reduction for Dependence Situations

Previously, we discussed the bias reduction strategies of relation-specific distances for the estimation of MI and TE. However, these strategies work well in independence situations, and the bias is only partly cancelled out when the independence conditions are not satisfied. To further eliminate the bias in the general case, we still consider the estimation of individual entropies. Removing the bias $\mathcal{B}_X$ in Equ. (3.83) is not an easy task since its mathematical expression depends on the unknown probability density. However, we can expect to cancel it out considering a weighted linear combination [Sricharan 2013]. Consequently, we introduce the following form of an ensemble estimator of entropy:

$$\widehat{\mathcal{H}(X)} = \left(-\frac{1}{N}\sum_{i=1}^{N}\left((1-\alpha_i)\log\widehat{p_X}^{(1)}(x_i)\right)\right) + \left(-\frac{1}{N}\sum_{i=1}^{N}\left(\alpha_i\log\widehat{p_X}^{(2)}(x_i)\right)\right), \tag{3.114}$$

where α_i, $i=1,\ldots,N$ is a sequence of weighting coefficients to be determined, $\widehat{p_X}^{(1)}(\cdot)$ and $\widehat{p_X}^{(2)}(\cdot)$ are two density estimations with the same structure $\left(\widehat{\frac{\int_{\mathcal{L}(x)}p_X(y)\mathrm{d}y}{v(x)}}\right)$ obtained from two distinct definitions of $\mathcal{L}(\cdot)$. Until now, $\mathcal{L}(x)$ was built either from a KDE approach or a kNN approach. In the first case, $\mathcal{R}(x)$ depends on the imposed bandwidth, and, in the second case, $\mathcal{R}(x)$ is deduced from the kth NN. Hereafter, to carry on with the conjecture proposed in [Kraskov 2004], we only consider the kNN approach integrating two steps, (i) the choice of two different numbers of neighbors k_1 and k_2, (ii)

the definition of the probability density estimators,

$$\begin{cases} \widehat{p_X}^{(1)}(x_i) = \widehat{p_{k_1}}(x_i) \\ \widehat{p_X}^{(2)}(x_i) = \widehat{p_{k_2}}(x_i), \end{cases} \tag{3.115}$$

where

$$\widehat{p_{k_j}}(x) = \frac{k_j - 1}{N \cdot v_{k_j}(x)}, \quad j = 1, 2 \tag{3.116}$$

is the standard kNN density estimator as defined in [Fukunaga 2013]. The volume $v_k(x)$ is equal to the Lebesgue measure of

$$\mathcal{L}_k(x) = \{y : \|y - x\| \leq \mathcal{R}_k(x)\}, \tag{3.117}$$

and $\mathcal{R}_k(x_i)$ is the distance between x_i and its kth NN.

Considering each bias term, we write

$$\mathcal{B}_X(x_i) \triangleq (1 - \alpha_i)\mathcal{B}_{k_1}(x_i) + \alpha_i \mathcal{B}_{k_2}(x_i). \tag{3.118}$$

The question arises of how to choose α_i in Equ. (3.118) so that $\mathcal{B}_X(x_i) = 0$. Given the Euclidean norm (Equ. (3.92)), we have

$$\begin{aligned} \mathcal{B}_X(x_i) &= (1 - \alpha_i)\mathcal{B}_{k_1}(x_i) + \alpha_i \mathcal{B}_{k_2}(x_i) \\ &= \frac{(1 - \alpha_i)\mathcal{R}_{k_1}^2(x_i) + \alpha_i \mathcal{R}_{k_2}^2(x_i)}{2\,(d_X + 2)\,p_X(x_i)} \cdot \operatorname{tr}\left(\frac{\partial^2 p_X}{\partial x^2}(x_i)\right). \end{aligned} \tag{3.119}$$

Now, zeroing out Equ. (3.119) for any $i = 1, \ldots, N$ with respect to α_i leads to

$$\alpha_i = \frac{\mathcal{R}_{k_1}^2(x_i)}{\mathcal{R}_{k_1}^2(x_i) - \mathcal{R}_{k_2}^2(x_i)}. \tag{3.120}$$

When starting from Equ. (3.95) instead of Equ. (3.92) to address the maximum norm, Equ. (3.120) still holds.

Considering the estimation of MI, in the dependence case we can apply the same strategy to X, Y and (X, Y) separately with distinct coefficients α_i^x, α_i^y, $\alpha_i^{(x,y)}$ and then compute the ensemble MI estimator using

$$\widehat{\mathcal{I}(X,Y)}_{\text{ens}} = \widehat{\mathcal{H}(X)}_{\text{ens}}^{k_1^x, k_2^x} + \widehat{\mathcal{H}(Y)}_{\text{ens}}^{k_1^y, k_2^y} - \widehat{\mathcal{H}(X,Y)}_{\text{ens}}^{k_1^{(x,y)}, k_2^{(x,y)}}, \tag{3.121}$$

where

$$\widehat{\mathcal{H}(U)}_{\text{ens}}^{k_1^u,k_2^u} = -\frac{1}{N}\sum_{i=1}^{N}\left((1-\alpha_i^u)\log\frac{k_1^u}{N\cdot v_{k_1}(u_i)} + \alpha_i^u \log\frac{k_2^u}{N\cdot v_{k_2}(u_i)}\right), \tag{3.122}$$

with the pairs $\left(k_1^x, k_2^x\right)$, $\left(k_1^y, k_2^y\right)$ and $\left(k_1^{(x,y)}, k_2^{(x,y)}\right)$ chosen independently for X, Y and (X,Y).

In the independence case, the basic strategy can be used. But we note that the values

$$\alpha_i^u = \frac{\mathcal{R}_{k_1}^2(u_i)}{\mathcal{R}_{k_1}^2(u_i) - \mathcal{R}_{k_2}^2(u_i)} \tag{3.123}$$

with u replaced by x, y or (x,y), are identical if we choose $\mathcal{R}_{k_1}^2$, $\mathcal{R}_{k_2}^2$ with the constraint imposed by Equ. (3.103) (or Equ. (3.104)).

Developing Equ. (3.121) with the substitution $\alpha_i^x = \alpha_i^y = \alpha_i^{(x,y)} = \alpha_i$, we have (Box ⑰ in Fig. 3.6)

$$\widehat{\mathcal{H}(U)}_{\text{mixed}}^{k_1,k_2} = -\frac{1}{N}\sum_{i=1}^{N}\left((1-\alpha_i)\log\frac{k_{k_1}(u_i)}{n\cdot v_{k_1}(u_i)} + \alpha_i \log\frac{k_{k_2}(u_i)}{n\cdot v_{k_2}(u_i)}\right). \tag{3.124}$$

Then, we get a mixed mutual information estimator (Box ⑱ in Fig. 3.6)

$$\widehat{\mathcal{I}(X,Y)}_{\text{mixed}}^{k_1,k_2} = \widehat{\mathcal{H}(X)}_{\text{mixed}}^{k_1,k_2} + \widehat{\mathcal{H}(Y)}_{\text{mixed}}^{k_1,k_2} - \widehat{\mathcal{H}(X,Y)}_{\text{mixed}}^{k_1,k_2}. \tag{3.125}$$

In summary, this mixed MI estimator is built following the three steps:

(i) Fix the number of NNs (k_1 and k_2 separately) in the joint space $\mathcal{S}_{X,Y}$ to get the distances between the center point (x_i, y_i) and the particular NNs (k_1th NN and k_2th NN), marked as $\mathcal{R}_{k_1}(x_i, y_i)$ and $\mathcal{R}_{k_2}(x_i, y_i)$

(ii) Use $\mathcal{R}_{k_1}(x_i, y_i)$ and $\mathcal{R}_{k_2}(x_i, y_i)$ to get respectively $\mathcal{R}_{k_1}(x_i)$, $\mathcal{R}_{k_1}(y_i)$, and $\mathcal{R}_{k_2}(x_i)$, $\mathcal{R}_{k_2}(y_i)$, using Equ. (3.103) or (3.104) (depending on the norm) and determine the numbers of points $k_{k_1}(x_i)$, $k_{k_1}(y_i)$, $k_{k_2}(x_i)$ and $k_{k_2}(y_i)$ falling into the corresponding regions

(iii) Estimate $\mathcal{H}(X)$ and $\mathcal{H}(Y)$ with Equ. (3.124), where α_i is given by Equ. (3.123), $\mathcal{H}(X,Y)$ being calculated similarly (with $k_{k_1}(x_i, y_i) = k_1$ and $k_{k_2}(x_i, y_i) = k_2$) and then calculate $\widehat{\mathcal{I}(X,Y)}_{\text{mixed}}^{k_1,k_2}$ by Equ. (3.125). The resulting estimator is named "mixed estimator" and denoted by $\widehat{\mathcal{I}(X,Y)}_{\text{mixed,E}}$ for the Euclidean norm and $\widehat{\mathcal{I}(X,Y)}_{\text{mixed,M}}$ for the maximum norm.

Note that $\widehat{\mathcal{I}(X,Y)}^{k}_{\text{basic}}$ is obtained by replacing Equ. (3.125) by Equ. (3.106) in step (iii).

Similarly, if we consider the estimation of TE, we obtain the following mixed estimator

$$\begin{aligned}\widehat{\text{TE}_{Y\to X}}^{k_1,k_2}_{\text{mixed}} = \widehat{\mathcal{H}(X^-,Y^-)}^{k_1,k_2}_{\text{mixed}} + \widehat{\mathcal{H}\left(X^p,X^-\right)}^{k_1,k_2}_{\text{mixed}} \\ - \widehat{\mathcal{H}\left(X^p,X^-,Y^-\right)}^{k_1,k_2}_{\text{mixed}} - \widehat{\mathcal{H}(X^-)}^{k_1,k_2}_{\text{mixed}},\end{aligned} \tag{3.126}$$

where $\widehat{\mathcal{H}(U)}^{k_1,k_2}_{\text{mixed}}$ is defined in Equ. (3.124).

For the mixed MI/TE estimators, an optimal choice of the parameters k_1 and k_2 is not obvious. In practice, it is possible to tune these two parameters to improve the estimation, but the empirical choice for the parameters remains an issue which explains why we only considered this kind of ensemble estimator in the estimation of mutual information and not in the estimation of transfer entropy.

3.4. Second Improvement

In this section [2], we first give an overview of the original kNN strategies on the estimation of entropy, including the Kozachenko–Leonenko entropy estimator [Kozachenko 1987] (Box ② in Fig. 3.6) and the one by Singh [Singh 2003] (Box ③), then introduce the idea of rectangle (Box ⑳). After that, we discuss the extensions of the existing kNN entropy estimators based on the idea of rectangle, which results in two novel entropy estimators (Boxes ㉑ and ㉒). Based on them, two new TE estimators are proposed (Boxes ㉓ and ㉔).

3.4.1. Original k-Nearest Neighbors Strategies

In this section, we consider a sequence x_i, $i = 1,\ldots,N$ in $\mathbb{R}^{d_X}$ (in our context this sequence corresponds to an outcome of an IID sequence $X_1,\ldots,X_N$ such that the common probability distribution is equal to that of a given random vector X). The set of the k NNs of x_i in this sequence (except for x_i) and the distance between x_i and its kth NN are respectively denoted by χ_i^k and $d_{x_i,k}$. We denote $\mathcal{D}_{x_i}\left(\chi_i^k\right) \subset \mathbb{R}^{d_X}$ a neighborhood of x_i in $\mathbb{R}^{d_X}$ which is the image of $\left(x_i,\chi_i^k\right)$ by a set valued map. For a given norm $\|\cdot\|$ on $\mathbb{R}^{d_X}$ a standard construction

$$\left(x_i,\chi_i^k\right) \in \left(\mathbb{R}^{d_X}\right)^{k+1} \to \mathcal{D}_{x_i}\left(\chi_i^k\right) \subset \mathbb{R}^{d_X} \tag{3.127}$$

[2] This second improvement was the subject of our contribution in [Zhu 2015a].

is the (hyper-)ball of radius equal to $d_{x_i,k}$, *i.e.*

$$\mathcal{D}_{x_i}\left(\chi_i^k\right) = \left\{x : \|x - x_i\| \leq d_{x_i,k}\right\}. \tag{3.128}$$

The (hyper-)volume (*i.e.* the Lebesgue measure) of $\mathcal{D}_{x_i}\left(\chi_i^k\right)$ is then

$$v_i = \int_{\mathcal{D}_{x_i}(\chi_i^k)} \mathrm{d}x, \tag{3.129}$$

where $\mathrm{d}x \triangleq \mathrm{d}\mu^{d_X}(x)$.

3.4.1.1. Kozachenko-Leonenko Entropy Estimator

Recalling the kNN entropy estimator defined in Equ. (3.60), to come up with a concise presentation of this estimator, we give hereafter a summary of the different steps to get it starting from [Kraskov 2004]. First, let us consider the distance $d_{x_i,k}$ between x_i and its kth NN (introduced above) as a realization of the random variable $D_{x_i,k}$ and let us denote by $q_{x_i,k}(x)$, $x \in \mathbb{R}$, the corresponding probability density function (conditioned by $X_i = x_i$). Secondly, let us consider the quantity

$$h^{x_i}(\varepsilon) = \int_{\|u - x_i\| \leq \frac{\varepsilon}{2}} \mathrm{d}P_X(u). \tag{3.130}$$

This is the probability mass of the (hyper-)ball with radius equal to $\frac{\varepsilon}{2}$ and centered on x_i. This probability mass is approximately equal to

$$\begin{aligned} h^{x_i}(\varepsilon) &\simeq p_X(x_i) \int_{\|\xi\| \leq \frac{\varepsilon}{2}} \mathrm{d}\mu^d(\xi) \\ &= p_X(x_i)\, c_d\, \varepsilon^d, \end{aligned} \tag{3.131}$$

if the density function is approximately constant on the (hyper-)ball. The variable c_d is the volume of the unity radius d-dimensional (hyper-)ball in $\mathbb{R}^d$ ($c_d = 1$ with the maximum norm). Furthermore, it can be established (see [Kraskov 2004] for details) that the expectation $\mathrm{E}\left[\log\left(h^{X_i}\left(D_{X_i,k}\right)\right)\right]$, where h^{X_i} is the random variable associated to h^{x_i}, $D_{X_i,k}$ (which must not be confused with the notation $\mathcal{D}_{x_i}\left(\chi_i^k\right)$ introduced previously) denotes the random distance between the kth neighbor selected in the set of random vectors $\{X_k, 1 \leq k \leq N, k \neq i\}$ and the random point X_i, is equal to $\psi(k) - \psi(N)$ and

does not depend on $p_X(\cdot)$. Equating it with $\mathrm{E}\left[\log\left(p_X(X_i)\, c_d\, D_{X_i,k}\right)\right]$ allows to write

$$\begin{aligned}\psi(k) - \psi(N) &\simeq \mathrm{E}\left[\log\left(p_X(X_i)\right)\right] + \mathrm{E}\left[\log\left(c_d\, D^d_{X_i,k}\right)\right] \\ &= -\mathcal{H}(X_i) + \mathrm{E}\left[\log(V_i)\right],\end{aligned} \tag{3.132}$$

or, equivalently

$$\mathcal{H}(X_i) \simeq \psi(N) - \psi(k) + \mathrm{E}\left[\log\left(c_d\, D^d_{X_i,k}\right)\right]. \tag{3.133}$$

Finally, by using the law of large numbers, when N is large we have

$$\begin{aligned}\mathcal{H}(X_i) &\simeq \psi(N) - \psi(k) + \frac{1}{N}\sum_{i=1}^{N}\log(v_i) \\ &= \widehat{\mathcal{H}(X)}_{\mathrm{kl}},\end{aligned} \tag{3.134}$$

where v_i is the realization of the random (hyper-)volume $V_i = c_d\, D^d_{x_i,k}$.

Moreover, as observed in [Kraskov 2004], it is possible to make the number of neighbors k depend on i by substituting the mean $\frac{1}{N}\sum_{i=1}^{N}\psi(k_i)$ for the constant $\psi(k)$ in Equ. (3.134), so that $\widehat{\mathcal{H}(X)}_{\mathrm{kl}}$ becomes:

$$\widehat{\mathcal{H}(X)}_{\mathrm{kl}} = \psi(N) + \frac{1}{N}\sum_{i=1}^{N}\left(\log(v_i) - \psi(k_i)\right). \tag{3.135}$$

3.4.1.2. Singh's Entropy Estimator

Now, let us consider the kNN entropy estimator defined in Equ. (3.61), which was proposed by Singh *et al.* in [Singh 2003]. As mentioned previously, using the approximation $\psi(N) \approx \log(N)$ for large values of N, this estimator is close to that defined by Equ. (3.60). This estimator was derived in [Singh 2003] through the four following steps:

(1) Introduce the classical entropy estimator structure

$$\begin{aligned}\widehat{\mathcal{H}(X)} &\triangleq -\frac{1}{N}\sum_{i=1}^{N}\log\widehat{p_X(X_i)} \\ &= \frac{1}{N}\sum_{i=1}^{N}T_i,\end{aligned} \tag{3.136}$$

where

$$\widehat{p_X(x_i)} \triangleq \frac{k}{N v_i}. \tag{3.137}$$

(2) Assuming that the random variables T_i, $i = 1, \ldots, N$ are identically distributed so

that $\mathrm{E}\left[\widehat{\mathcal{H}(X)}\right] = \mathrm{E}[T_1]$ (note that $\mathrm{E}[T_1]$ depends on N, even if the notation does not make that explicit), compute the asymptotic value of $\mathrm{E}[T_1]$ (when N is large) by firstly computing its asymptotic cumulative probability distribution function and the corresponding probability density p_{T_1}, and finally compute the following expectation

$$\mathrm{E}[T_1] = \int_{\mathbb{R}} t \; p_{T_1}(t)\mathrm{d}t. \tag{3.138}$$

(3) It appears that

$$\begin{aligned} \mathrm{E}[T_1] &= \mathrm{E}\left[\widehat{\mathcal{H}(X)}\right] \\ &= \mathcal{H}(X) + B, \end{aligned} \tag{3.139}$$

where B is a constant which is identified with the bias.

(4) Subtract this bias from $\widehat{\mathcal{H}(X)}$ to get

$$\widehat{\mathcal{H}(X)}_{\mathrm{sg}} = \widehat{\mathcal{H}(X)} - B \tag{3.140}$$

and the formula given in Equ. (3.61).

Note that the cancellation of the asymptotic bias does not imply that the bias obtained with a finite value of N is also exactly cancelled. Now, we explain the origin of the bias for the entropy estimator given in Equ. (3.136). Let us consider the equalities

$$\begin{aligned} \mathrm{E}[T_1] &= -\mathrm{E}\left[\log\left(\widehat{p_X(X_1)}\right)\right] \\ &= -\mathrm{E}\left[\log\left(\frac{k}{NV_1}\right)\right], \end{aligned} \tag{3.141}$$

where V_1 is the random volume for which v_1 is an outcome. Conditionally to $X_1 = x_1$, if we have

$$\frac{k}{NV_1} \xrightarrow[N\to\infty]{pr} p_X(x_1), \tag{3.142}$$

where pr denotes the convergence in probability, we could expect that

$$\mathrm{E}[T_1 | X_1 = x_1] \xrightarrow[N\to\infty]{} -\log(p_X(x_1)) \tag{3.143}$$

and, by deconditioning, that

$$\mathrm{E}[T_1] \xrightarrow[N\to\infty]{} -\mathrm{E}[\log(p_X(X_1))] = \mathcal{H}(X). \tag{3.144}$$

So, the convergence

$$\frac{k}{NV_1} \xrightarrow[N\to\infty]{pr} p_X(x_1) \tag{3.145}$$

could lead to an asymptotically unbiased estimation of $\mathcal{H}(X)$. Now, this convergence in probability does not hold, even if we assume the following convergence of the mean

$$\mathrm{E}\left[\frac{k}{NV_1}|X_1 = x_1\right] \xrightarrow[N\to\infty]{} p_X(x_1), \tag{3.146}$$

because we do not have

$$\mathrm{var}\left(\frac{k}{NV_1}|X_1 = x_1\right) \xrightarrow[N\to\infty]{} 0. \tag{3.147}$$

The ratio $\left(\frac{k}{NV_1}\right)$ remains fluctuating when $N \to \infty$, because the ratio $\left(\frac{\sqrt{\mathrm{var}(V_1)}}{\mathrm{E}[V_1]}\right)$ does not tend to zero even if V_1 tends to be smaller: when N increases, the neighborhoods become smaller and smaller but continue to "fluctuate". This explains informally (see [Zhu 2014] for a more detailed analysis) why the naive estimator given by Equ. (3.136) is not asymptotically unbiased. It is interesting to note that the Kozachenko–Leonenko entropy estimator avoids this problem and so it does not need any bias subtraction asymptotically.

Observe also that, as for the Kozachenko–Leonenko estimator, it is possible to adapt Equ. (3.135) if we want to consider a number of neighbors k_i depending on i. Equ. (3.61) can then be replaced by

$$\widehat{\mathcal{H}(X)}_{\mathrm{sg}} = \log(N) + \frac{1}{N}\sum_{i=1}^{N}(\log(v_i) - \psi(k_i)). \tag{3.148}$$

3.4.2. From Square to Rectangular Neighboring Region for Entropy Estimation

In [Kraskov 2004], to estimate MI, as illustrated in Fig. 3.7, Kraskov *et al.* discussed two different techniques to build the neighboring region to compute $\widehat{\mathcal{I}(X,Y)}$: in the standard technique (square $ABCD$ in Fig. 3.7(a) and 3.7(b)), the region determined by the first k NNs is a (hyper-)cube and leads to Equ. (3.63), and, in the second technique (rectangle $A'B'C'D'$ in Fig. 3.7(a) and 3.7(b)), the region determined by the first k NNs is a (hyper-)rectangle. Note that the TE estimator mentioned in the previous section (Equ. (3.67)) is based on the first situation (square $ABCD$ in Fig. 3.7(a) or 3.7(b)). The introduction of the second technique by Kraskov *et al.* was to circumvent the fact that Equ. (3.135) was not applied rigorously to obtain the terms $\psi(n_{X,i}+1)$ or $\psi(n_{Y,i}+1)$ in Equ. (3.63). As a matter of fact, for one of these terms, no point x_i (or y_i) falls exactly on the border of the (hyper-)cube $\mathcal{D}_{x_i}$ (or $\mathcal{D}_{y_i}$) obtained by the distance projection from the $\mathcal{S}_{X,Y}$ space. As clearly illustrated in Fig. 3.7 (rectangle $A'B'C'D'$ in Fig. 3.7(a) and 3.7(b)), the second strategy prevents from that issue since the border of the (hyper-)cube (in this case an interval of $\mathbb{R}$) after projection from $\mathcal{S}_{X,Y}$ space to $\mathcal{S}_X$

space (or $\mathcal{S}_Y$ space) contains one point. When the dimensions of $\mathcal{S}_X$ and $\mathcal{S}_Y$ are larger than one, this strategy leads to build a (hyper-)rectangle equal to the product of two (hyper-)cubes, one of them in $\mathcal{S}_X$ and the other one in $\mathcal{S}_Y$. If the maximum distance of the kth NN in $\mathcal{S}_{X,Y}$ is obtained in one of the directions in $\mathcal{S}_X$, this maximum distance, after multiplying by two, fixes the size of the (hyper-)cube in $\mathcal{S}_X$. To obtain the size of the second (hyper-)cube (in $\mathcal{S}_Y$), the k neighbors in $\mathcal{S}_{X,Y}$ are first projected on $\mathcal{S}_Y$ and then the largest of the distances calculated from these projections fixes the size of this second (hyper-)cube.

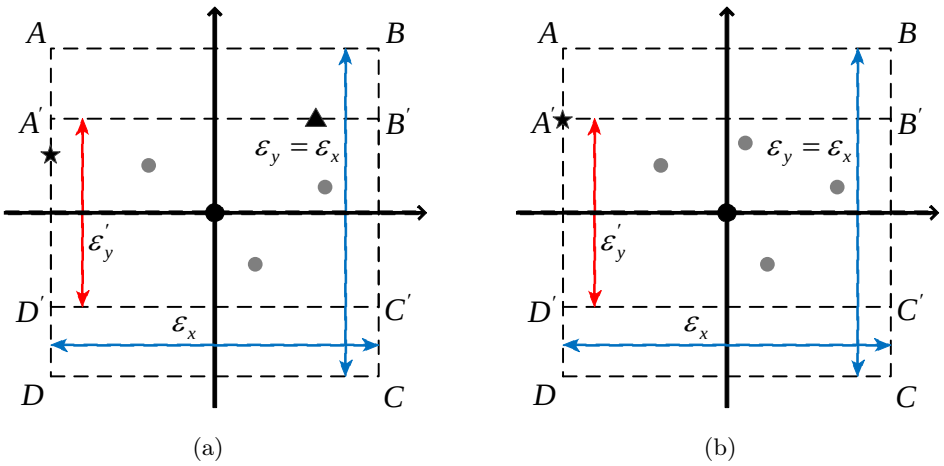


Figure 3.7: In this 2-dimensional example, $k = 5$. The origin of the Cartesian axis corresponds to the current point x_i. Only the 5 NNs of this point, *i.e.* the points in the set χ_i^k, are represented. The 5th NN is symbolized by a star. The neighboring regions $ABCD$, obtained from the maximum norm around the center point, are squares, with equal edge lengths $\varepsilon_x = \varepsilon_y$. Reducing one of the edge lengths, ε_x or ε_y, until one point falls onto the border (in the present case, in the vertical direction), leads to the minimum size rectangle $A'B'C'D'$, where $\varepsilon_x \neq \varepsilon'_y$. Two cases must be considered, illustrated respectively in Fig. 3.7(a) and 3.7(b). For case (a) the 5th NN is not localized on an intersection of two edges, contrary to the case (b). This leads to obtain either two points (respectively the star and the triangle in Fig. 3.7(a)) or only one point (the star in Fig. 3.7(b)) on the border of $A'B'C'D'$. Clearly it is theoretically possible to have more than 2 points on the border of $A'B'C'D'$ but the probability of such an occurrence is equal to zero when the probability distribution of the random points X_j is continuous.

In the remainder of this section, for an arbitrary dimension d, we propose to apply this strategy to estimate the entropy of a single multidimensional variable X observed in $\mathbb{R}^d$. This leads to introduce a d-dimensional (hyper-)rectangle centered on x_i having a minimal volume and including the set χ_i^k of neighbors. Hence the rectangular neighboring is built by adjusting its size separately in each direction in the space $\mathcal{S}_X$. Using this strategy, we are sure that, in any of the d directions, there is at least one point on one of the two borders (and only one with probability one). Therefore, in this approach the (hyper-)rectangle, denoted by $\mathcal{D}_{x_i}^{\varepsilon_1,\ldots,\varepsilon_d}$, where the sizes $\varepsilon_1, \ldots, \varepsilon_d$ in the respective d

directions are completely specified from the neighbors set χ_i^k, is substituted for the basic (hyper-)square

$$\mathcal{D}_{x_i}\left(\chi_i^k\right) = \left\{x : \|x - x_i\| \leq d_{x_i,k}\right\}. \tag{3.149}$$

It should be mentioned that the central symmetry of the (hyper-)rectangle around the center point allows for reducing the bias in the density estimation [Fukunaga 1973] (*cf.* Equ. (3.131) or (3.137)). Note that, when $k < d$, there must exist neighbors positioned on some vertices or edges of the (hyper-)rectangle. With $k < d$ it is impossible that, for any direction, one point falls exactly inside a face (*i.e.* not on its border). For example with $k = 1$ and $d > 1$ the first neighbor is on a vertex and the sizes of the edges of the reduced (hyper-)rectangle are equal to twice the absolute value of its coordinates, whatever the direction.

Hereafter, we propose to extend the entropy estimators by Kozachenko–Leonenko and Singh using the above strategy before deriving the corresponding TE estimators.

3.4.3. Extension of the Kozachenko–Leonenko Method

As indicated before, in [Kraskov 2004], Kraskov *et al.* extended the Kozachenko–Leonenko estimator (Equ. (3.60)) using the rectangular neighboring strategy to derive MI estimator. Now, focusing on entropy estimation, we can obtain another estimator of $\mathcal{H}(X)$, denoted by $\widehat{\mathcal{H}(X)}_{\text{kl2}}$ (Box ㉑ in Fig. 3.6),

$$\widehat{\mathcal{H}(X)}_{\text{kl2}} = \psi(N) + \frac{1}{N}\sum_{i=1}^{N}\log\left(v_i\right) - \psi(k) + \frac{d-1}{k}, \tag{3.150}$$

where v_i is the volume of the minimum volume (hyper-)rectangle around the point x_i.

Hereafter we give the mathematical development to get Equ. (3.150). As illustrated in Fig. 3.7, for $d = 2$ there are two cases to be distinguished: (i) ε_x and ε_y are determined by the same point, (ii) ε_x and ε_y are determined by distinct points.

Considering the probability density $q_{i,k}\left(\epsilon_x, \epsilon_y\right)$, $\left(\epsilon_x, \epsilon_y\right) \in \mathbb{R}^2$ of the pair of random sizes $\left(\varepsilon_x, \varepsilon_y\right)$ (along x and y respectively), we can extend it to the case $d > 2$. Hence let us denote by $q^d_{x_i,k}\left(\varepsilon_1, \ldots, \varepsilon_d\right)$, $\left(\varepsilon_1, \ldots, \varepsilon_d\right) \in \mathbb{R}^d$ the probability density (conditional to $X_i = x_i$) of the d-dimensional random vector whose d components are respectively the d random sizes of the (hyper-)rectangle built from the random k NNs and denote by

$$h^{x_i}\left(\varepsilon_1, \ldots, \varepsilon_d\right) = \int\limits_{u \in \mathcal{D}_{x_i}^{\varepsilon_1,\ldots,\varepsilon_d}} \mathrm{d}P_X\left(u\right) \tag{3.151}$$

the probability mass (conditional to $X_i = x_i$) of the random (hyper-)rectangle $\mathcal{D}_{x_i}^{\epsilon_1,\ldots,\epsilon_d}$.

In [Kraskov 2004] the equality

$$\mathrm{E}\left[\log\left(h^{x_i}\left(D_{x_i,k}\right)\right)\right] = \psi(k) - \psi(N) \tag{3.152}$$

obtained for a (hyper-)cube is extended for the case $d > 2$ to

$$\mathrm{E}\left[\log\left(h^{x_i}\left(\epsilon_1, \ldots, \epsilon_d\right)\right)\right] = \psi(k) - \frac{d-1}{k} - \psi(N). \tag{3.153}$$

So, if p_X is approximately constant on $\mathcal{D}_{x_i}^{\varepsilon_1,\ldots,\varepsilon_d}$ we get

$$h^{x_i}\left(\varepsilon_1, \ldots, \varepsilon_d\right) \simeq v_i\, p_X\left(x_i\right), \tag{3.154}$$

where $v_i = \int_{\mathcal{D}_{x_i}^{\varepsilon_1,\ldots,\varepsilon_d}} \mathrm{d}\mu^d(\xi)$ is the volume of the (hyper-)rectangle, and we obtain

$$\log p_X(x_i) \approx \psi(k) - \psi(N) - \frac{d-1}{k} - \log\left(v_i\right). \tag{3.155}$$

Finally, by taking the experimental mean of the right term in Equ. (3.155) we obtain an estimation of the expectation $\mathrm{E}\left[\log p_X(X)\right]$, *i.e.* Equ. (3.150).

3.4.4. Extension of Singh's Method

In this section, we propose to extend Singh's entropy estimator by using a (hyper-)rectangular domain as we did for the Kozachenko–Leonenko estimator extension introduced in the preceding section. Considering a d-dimensional random vector $X \in \mathbb{R}^d$ continuously distributed according to a probability density function p_X, we aim at estimating the entropy $\mathcal{H}(X)$ from the observation of a p_X-distributed IID random sequence X_i, $i = 1, \ldots, N$. For any specific data point x_i and a fixed number k $(1 \leq k \leq N)$, the minimum (hyper-)rectangle (rectangle $A'B'C'D'$ in Fig. 3.7 is fixed, we denote this region by $\mathcal{D}_{x_i}^{\varepsilon_1,\ldots,\varepsilon_d}$, and its volume by v_i. Let us denote ξ_i $(1 \leq \xi_i \leq \min(k, d))$ the number of points on the border of the (hyper-)rectangle that we consider as a realization of a random variable Ξ_i. In the situation described in Fig. 3.7(a) and 3.7(b), $\xi_i = 2$ and $\xi_i = 1$ respectively. According to [Fukunaga 2013] (chapter 6, page 269), if $\mathcal{D}_{x_i}\left(\chi_i^k\right)$ corresponds to a ball (for a given norm) of volume v_i, an unbiased estimator of $p_X(x_i)$ is given by

$$\widehat{p_X(x_i)} = \frac{k-1}{N v_i}, i = 1, 2, \ldots, N. \tag{3.156}$$

This implies that the classical estimator $\widehat{p_X(x_i)} = \frac{k}{N v_i}$ is biased and that presumably $\log\left(\frac{k}{N v_i}\right)$ is also a biased estimation of $\log\left(p_X(x_i)\right)$ for N large as shown in [Fukunaga 2013].

Now, in case $\mathcal{D}_{x_i}\left(\chi_i^k\right)$ is the minimal (*i.e.* with minimal (hyper-)volume) (hyper-)rectangle $\mathcal{D}_{x_i}^{\varepsilon_1,\ldots,\varepsilon_d}$ including χ_i^k, more than one point can belong to the border, and a more general estimator $\widetilde{p_X(x_i)}$ of $p_X(x_i)$ can be *a priori* considered

$$\widetilde{p_X(x_i)} = \frac{\tilde{k}_i}{N v_i}, \tag{3.157}$$

where $\tilde{k}_i$ is some given function of k and ξ_i. The corresponding estimation of $\mathcal{H}(X)$ is then

$$\begin{aligned} \widehat{\mathcal{H}(X)} &= -\frac{1}{N}\sum_{i=1}^{N}\widetilde{\log\left(p_X(x_i)\right)} \\ &= \frac{1}{N}\sum_{i=1}^{N} t_i, \end{aligned} \tag{3.158}$$

with

$$t_i = \log\left(\frac{N v_i}{\tilde{k}_i}\right), i = 1, 2, \ldots, N, \tag{3.159}$$

t_i being realizations of random variables T_i and $\tilde{k}_i$ being realizations of random variables $\widetilde{K}_i$. We have

$$\forall i = 1, \ldots, N : \mathrm{E}\left[\widehat{\mathcal{H}(X)}\right] = \mathrm{E}\left[T_i\right] = \mathrm{E}\left[T_1\right]. \tag{3.160}$$

Our goal is to derive

$$\mathrm{E}\left[\widehat{\mathcal{H}(X)}\right] - \mathcal{H}(X) = \mathrm{E}\left[T_1\right] - \mathcal{H}(X) \tag{3.161}$$

for N large to correct the asymptotic bias of $\widehat{\mathcal{H}(X)}$, according to steps (1) to (3) explained in section 3.4.1.2. To this end, we must consider an asymptotic approximation of the conditional probability distribution $\mathcal{P}\left(T_1 \leq r | X_1 = x_1, \Xi_1 = \xi_1\right)$ before computing the asymptotic difference between the expectation $\mathrm{E}\left[T_1\right] = \mathrm{E}\left[\mathrm{E}\left[T_1 | X_1 = x_1, \Xi_1 = \xi_1\right]\right]$ and the true entropy $\mathcal{H}(X)$.

Let us consider the random Lebesgue measure V_1 of the random minimal (hyper-)rectangle $\mathcal{D}_{x_1}^{\epsilon_1,\ldots,\epsilon_d}$ ($(\epsilon_1, \ldots, \epsilon_d)$ denotes the random vector for which $(\varepsilon_1, \ldots, \varepsilon_d) \in \mathbb{R}^d$ is a realization) and the relation

$$T_1 = \log\left(\frac{N V_1}{\widetilde{K}_1}\right). \tag{3.162}$$

For any $r > 0$, we have

$$\begin{aligned}
&\mathcal{P}\left(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1\right) \\
=&\mathcal{P}\left(\log\left(\frac{NV_1}{\widetilde{K}_1}\right) > r | X_1 = x_1, \Xi_1 = \xi_1\right) \\
=&\mathcal{P}\left(V_1 > v_r | X_1 = x_1, \Xi_1 = \xi_1\right),
\end{aligned} \tag{3.163}$$

where $v_r = e^r \frac{\tilde{k}_1}{N}$ since, conditionally to $\Xi_1 = \xi_1$, we have $\widetilde{K}_1 = \tilde{k}_1$.

Property 1: For N large,

$$\mathcal{P}\left(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1\right) \simeq \sum_{i=0}^{k-\xi_1} \binom{N-\xi_1-1}{i} (p_X(x_1)v_r)^i (1 - p_X(x_1)v_r)^{N-\xi_1-1-i}. \tag{3.164}$$

(see Appendix B for proof of property 1).

The Poisson approximation (when $N \to \infty$ and $v_r \to 0$) of the binomial distribution summed in Equ. (3.164) leads to a parameter λ, such that

$$\lambda = (N - \xi_1 - 1)\, p_X(x_1) v_r. \tag{3.165}$$

As N is large compared to $\xi_1 + 1$, we obtain

$$\lambda \simeq \tilde{k}_1 e^r p_X(x_1) \tag{3.166}$$

and we get the approximation

$$\lim_{N\to\infty} \mathcal{P}\left(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1\right) \simeq \sum_{i=0}^{k-\xi_1} \frac{\left(\tilde{k}_1 e^r p_X(x_1)\right)^i}{i!} e^{-\tilde{k}_1 e^r p_X(x_1)}. \tag{3.167}$$

Since

$$\mathcal{P}(T_1 \leq r | X_1 = x_1, \Xi_1 = \xi_1) = 1 - \mathcal{P}(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1), \tag{3.168}$$

we can get the density function of T_1, noted $g_{T_1}(r)$, by deriving $\mathcal{P}\left(T_1 \leq r | X_1 = x_1, \Xi_1 = \xi_1\right)$. After some mathematical developments, we obtain (see Appendix C for details):

$$\begin{aligned} g_{T_1}(r) &= \mathcal{P}'(T_1 \leq r | X_1 = x_1, \Xi_1 = \xi_1) \\ &= -\mathcal{P}'(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1) \\ &= \frac{\left(\tilde{k}_1 e^r p_X(x_1)\right)^{(k-\xi_1+1)}}{(k-\xi_1)!} e^{-\tilde{k}_1 e^r p_X(x_1)}, \ r \in \mathbb{R}, \end{aligned} \tag{3.169}$$

and consequently (see Appendix D for details),

$$\begin{aligned} &\lim_{N\to\infty} \mathrm{E}\left[T_1 | X_1 = x_1, \Xi_1 = \xi_1\right] \\ &= \int_{-\infty}^{\infty} r \frac{(\tilde{k}_1 p_X(x_1) e^r)^{(k-\xi_1+1)}}{(k-\xi_1)!} e^{-\tilde{k}_1 p_X(x_1) e^r} \mathrm{d}r \\ &= \psi(k-\xi_1+1) - \log\left(\tilde{k}_1\right) - \log p_X(x_1). \end{aligned} \tag{3.170}$$

With the definition of differential entropy $\mathcal{H}(X_1) = \mathrm{E}[-\log(p_X(X_1))]$, we come to

$$\begin{aligned} \lim_{N\to\infty} \mathrm{E}\left[T_1\right] &= \lim_{N\to\infty} \mathrm{E}\left[\mathrm{E}\left[T_1 | X_1, \Xi_1\right]\right] \\ &= \mathrm{E}\left[\psi(k - \Xi_1 + 1) - \log\left(\widetilde{K}_1\right)\right] + \mathcal{H}(X_1). \end{aligned} \tag{3.171}$$

Thus, the estimator expressed by Equ. (3.158) is asymptotically biased. Therefore, we consider a modified version, denoted by $\widehat{\mathcal{H}(X)}_{\mathrm{sg2}}$ obtained by subtracting an estimation of the bias $\mathrm{E}\left[\psi(k - \Xi_1 + 1) - \log\left(\widetilde{K}_1\right)\right]$ given by the empirical mean $\frac{1}{N}\sum_{i=1}^{N} \psi(k-\xi_i+1) + \frac{1}{N}\sum_{i=1}^{N} \log\left(\tilde{k}_i\right)$ (according to the law of large numbers), and we obtain finally (Box ㉒ in Fig. 3.6)

$$\begin{aligned} \widehat{\mathcal{H}(X)}_{\mathrm{sg2}} &= \frac{1}{N}\sum_{i=1}^{N} t_i - \frac{1}{N}\sum_{i=1}^{N} \psi(k-\xi_i+1) + \frac{1}{N}\sum_{i=1}^{N} \log\left(\tilde{k}_i\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{N v_i}{\tilde{k}_i}\right) - \frac{1}{N}\sum_{i=1}^{N} \psi(k-\xi_i+1) + \frac{1}{N}\sum_{i=1}^{N} \log\left(\tilde{k}_i\right) \\ &= \log(N) + \frac{1}{N}\sum_{i=1}^{N} \log(v_i) - \frac{1}{N}\sum_{i=1}^{N} \psi(k-\xi_i+1). \end{aligned} \tag{3.172}$$

In comparison with the development of Equ. (3.150), we followed the same methodology except that we took into account (through a conditioning technique) the influence of the number of points on the border.

We observe that, after cancellation of the asymptotic bias, the choice of the function of k and ξ_i to define $\tilde{k}_i$ in Equ. (3.157) does not have any influence in the final result. By this way, we obtain an expression for $\widehat{\mathcal{H}(X)}_{\mathrm{sg2}}$ which simply takes into account the

values ξ_i that could *a priori* influence the entropy estimation.

Note that, as for the original Kozachenko–Leonenko (Equ. (3.60)) and Singh (Equ. (3.61)) entropy estimators, both new estimation functions (Equ. (3.150) and (3.172)) hold for any value of k such that $k \ll N$, and we do not have to choose a fixed k while estimating entropy in lower-dimensional spaces. So, under the framework proposed in [Kraskov 2004], we built two different TE estimators using Equ. (3.150) and (3.172) respectively.

3.4.5. Computation of the Border Points Number and of the (Hyper-) Rectangle Sizes

We explain more precisely hereafter how to determine the numbers of points ξ_i on the border. Let us denote $x_i^j \in \mathbb{R}^d$, $j = 1, \ldots, k$, the k NNs of $x_i \in \mathbb{R}^d$ and let us consider the $d \times k$ array D_i such that for any $(p, j) \in \{1, \ldots, d\} \times \{1, \ldots, k\}$, $D_i(p, j) = \left| x_i^j(p) - x_i(p) \right|$ is the distance (in $\mathbb{R}$) between the pth component $x_i^j(p)$ of x_i^j and the pth component $x_i(p)$ of x_i. For each p, let us introduce $J_i(p) \in \{1, \ldots, k\}$ defined by

$$D_i(p, J_i(p)) = \max\left(D_i(p, 1), \ldots, D_i(p, k)\right) \tag{3.173}$$

and which is the value of the column index of D_i for which the distance $D_i(p, j)$ is maximum in the row number p. Now, if there exists more than one index $J_i(p)$ which fulfills this equality, we select arbitrary the lowest one, hence avoiding the $\max(\cdot)$ function to be multi-valued. The MATLAB implementation of the $\max(\cdot)$ function selects such a unique index value. Then, let us introduce the $d \times k$ Boolean array B_i defined by

$$\begin{cases} B_i(p, j) = 1, & \text{if} \quad j = J_i(p), \\ B_i(p, j) = 0, & \text{otherwise.} \end{cases} \tag{3.174}$$

Then

(1) The d sizes ε_p, $p = 1, \ldots, d$ of the (hyper-)rectangle $\mathcal{D}_{x_i}^{\varepsilon_1, \ldots, \varepsilon_d}$ are equal respectively to $\varepsilon_p = 2D_i\left(p, J_i(p)\right)$, $p = 1, \ldots, d$.

(2) We can define ξ_i as the number of non-null column vectors in B_i. For example, if the kth NN x_i^k is such that

$$\forall j \neq k, \forall p = 1, \ldots, d : \left| x_i^j(p) - x_i(p) \right| < \left| x_i^k(p) - x_i(p) \right|, \tag{3.175}$$

i.e. when the kth NN is systematically the farthest from the central point x_i for each of the d directions, then all the entries in the last column of B_i are equal to

one while all other entries are equal to zero: we have only one column including values different from zero and, so, only one point on the border ($\xi_i = 1$), what generalizes the case depicted in Fig. 3.7(b) for $d = 2$.

N.B.: this determination of ξ_i may be incorrect when there exists a direction p such that the number of indices j for which $D_i(p, j)$ reaches the maximal value is larger than one: the value of ξ_i obtained with our procedure can then be underestimated. However, we can argue that, theoretically, this case occurs with a probability equal to zero (because the observations are continuously distributed in probability) and so it can be *a priori* discarded. Now, in practice, the measure quantification errors and the round-off errors are unavoidable and this probability will differ from zero (although remaining small when the aforesaid errors are small): theoretically distinct values $D_i(p, j)$ on the row p of D_i may be erroneously confounded after quantification and rounding. But the $\max(\cdot)$ function then selects on the row p only one value for $J_i(p)$ and so acts as an error correcting procedure. The fact that the maximum distance in the concerned p directions can then be allocated not to the right neighbor index has no consequence for the correct determination of ξ_i.

Given the entropy estimators derived from the Kozachenko-Leonenko estimator (Equ. (3.60)) and from Singh's estimator (Equ. (3.61)), we can now derive new TE estimators.

3.4.6. New Estimators of Transfer Entropy

From an observed realization $\left(x_i^p, x_i^-, y_i^-\right) \in \mathcal{S}_{X^p,X^-,Y^-}$, $i = 1, 2, \ldots, N$ of the IID random sequence $\left(X_i^p, X_i^-, Y_i^-\right)$, $i = 1, 2, \ldots, N$ and a number k of neighbors, the procedure could be summarized as follows (distances are from the maximum norm)

(1) similarly as MILCA [Sergey 2015] and TRENTOOL [Wollstadt 2015] toolboxes, normalize, for each i, the vectors x_i^p, x_i^- and y_i^-;

(2) in joint space $\mathcal{S}_{X^p,X^-,Y^-}$, for each point $\left(x_i^p, x_i^-, y_i^-\right)$, calculate the distance $d_{(x_i^p,x_i^-,y_i^-),k}$ between $\left(x_i^p, x_i^-, y_i^-\right)$ and its kth neighbor, then construct the (hyper-)rectangle with sizes $\varepsilon_1, \ldots, \varepsilon_d$ (d is the dimension of the vectors $\left(x_i^p, x_i^-, y_i^-\right)$), for which the (hyper-)volume is $v_{(X^p,X^-,Y^-),i} = \varepsilon_1 \times \ldots \times \varepsilon_d$ and the border contains $\xi_{(X^p,X^-,Y^-),i}$ points;

(3) for each point (x_i^p, x_i^-) in subspace $\mathcal{S}_{X^p,X^-}$, count the number $k_{(X^p,X^-),i}$ of points falling within the distance $d_{(x_i^p,x_i^-,y_i^-),k}$, then find the smallest (hyper-)rectangle which contains all these points and for which $v_{(X^p,X^-),i}$ and $\xi_{(X^p,X^-),i}$ are respectively the volume and the number of points on the border. Repeat the same procedure in subspaces $\mathcal{S}_{X^-,Y^-}$ and $\mathcal{S}_{X^-}$.

From Equ. (3.150), our first proposed TE estimator named $\mathrm{TE}_{\mathrm{p1}}$can be written as (Box ㉓ in Fig. 3.6)

$$\begin{aligned}\widehat{\mathrm{TE}_{Y\to X\,\mathrm{p1}}} &= \frac{1}{N}\sum_{i=1}^{N}\log\frac{v_{(X^p,X^-),i}\cdot v_{(X^-,Y^-),i}}{v_{(X^p,X^-,Y^-),i}\cdot v_{X^-,i}} \\ &+ \frac{1}{N}\sum_{i=1}^{N}\Bigg(\psi(k)+\psi(k_{X^-,i})-\psi(k_{(X^p,X^-),i})-\psi(k_{(X^-,Y^-),i}) \\ &+ \frac{d_{X^p}+d_{X^-}-1}{k_{(X^p,X^-),i}}+\frac{d_{X^-}+d_{Y^-}-1}{k_{(X^-,Y^-),i}}-\frac{d_{X^p}+d_{X^-}+d_{Y^-}-1}{k}-\frac{d_{X^-}-1}{k_{X^-,i}}\Bigg),\end{aligned} \tag{3.176}$$

where $d_{X^p}=\dim(\mathcal{S}_{X^p})$, $d_{X^-}=\dim(\mathcal{S}_{X^-})$, $d_{Y^-}=\dim(\mathcal{S}_{Y^-})$ and, with Equ. (3.172), our second proposed estimator named $\mathrm{TE}_{\mathrm{p2}}$ (Box ㉔ in Fig. 3.6) is written

$$\begin{aligned}\widehat{\mathrm{TE}_{Y\to X\,\mathrm{p2}}} &= \frac{1}{N}\sum_{i=1}^{N}\log\frac{v_{(X^p,X^-),i}\cdot v_{(X^-,Y^-),i}}{v_{(X^p,X^-,Y^-),i}\cdot v_{X^-,i}} \\ &+ \frac{1}{N}\sum_{i=1}^{N}\Bigg(\psi(k-\xi_{(X^p,X^-,Y^-),i}+1)+\psi(k_{X^-,i}-\xi_{X^-,i}+1) \\ &- \psi(k_{(X^p,X^-),i}-\xi_{(X^p,X^-),i}+1)-\psi(k_{(X^-,Y^-),i}-\xi_{(X^-,Y^-),i}+1)\Bigg).\end{aligned} \tag{3.177}$$

In Equ. (3.176) and (3.177) the volumes $v_{(X^p,X^-),i}$, $v_{(X^-,Y^-),i}$, $v_{(X^p,X^-,Y^-),i}$, $v_{X^-,i}$ are obtained by computing, for each of them, the product of the edges lengths of the (hyper-)rectangle, *i.e.* the product of d edges lengths, d being respectively equal to $d_{X^p}+d_{X^-}$, $d_{X^-}+d_{Y^-}$, $d_{X^p}+d_{X^-}+d_{Y^-}$ and d_{X^-}. In a given subspace and for a given direction, the edge length is equal to twice the largest distance between the corresponding coordinate of the reference point (at the center) and each of the corresponding coordinates of the k NNs. Hence a generic formula is $v_U=\prod_{j=1}^{\dim(U)}\varepsilon_{Uj}$ where U is one of the symbols (X^p,X^-), (X^-,Y^-), (X^p,X^-,Y^-), X^- and the ε_{Uj} are the edges lengths of the (hyper-)rectangle.

The new TE estimator $\widehat{\mathrm{TE}_{Y\to X\,\mathrm{p1}}}$ can be compared with the TE estimator proposed in [Wibral 2014a] (Equ. (3.68), implemented in the JIDT toolbox [Lizier 2014], version 1.2, referred as the Extended algorithm). The main difference with our $\widehat{\mathrm{TE}_{Y\to X\,\mathrm{p1}}}$ estimator is that our algorithm uses a different length for each sub-dimension within a variable rather than one length for all sub-dimensions within the variable. We introduced this approach to make the tightest possible (hyper-)rectangle around the k NNs.

Equ. (3.68) differs from Equ. (3.176) in two ways. Firstly, the first summation in the right-hand side of Equ. (3.176) does not exist. Secondly, compared with Equ. (3.176),

the numbers of neighbors $k_{X^-,i}$, $k_{(X^p,X^-),i}$ and $k_{(X^-,Y^-),i}$ included in the rectangular boxes, are replaced respectively with $n_{X^-,i}$, $n_{(X^p,X^-),i}$ and $n_{(X^-,Y^-),i}$ which are obtained differently. More precisely, the preceding step (2) in the Extended TE algorithm (Equ. (3.68)) becomes:

(2′) for each point (x_i^p, x_i^-) in subspace $\mathcal{S}_{X^p,X^-}$, $n_{(X^p,X^-),i}$ is the number of points falling within a (hyper-)rectangle equal to the Cartesian product of two (hyper-)cubes, the first one in $\mathcal{S}_{X^p}$ and the second one in $\mathcal{S}_{X^-}$, whose edge lengths are equal, respectively, to

$$d_{x_i^p}^{\max} = 2 \times \max \left\{ \left\| x_k^p - x_i^p \right\| : \left(x^p, x^-, y^-\right)_k \in \chi^k_{(x^p,x^-,y^-)_i} \right\} \tag{3.178}$$

and

$$d_{x_i^-}^{\max} = 2 \times \max \left\{ \left\| x_k^- - x_i^- \right\| : \left(x^p, x^-, y^-\right)_k \in \chi^k_{(x^p,x^-,y^-)_i} \right\}, \tag{3.179}$$

i.e.

$$n_{(X^p,X^-),i} = \operatorname{card} \left\{ \left(x_j^p, x_i^-\right) : j \in \{\{1,\ldots,N\} - \{i\}\} \ \& \ \left\| x_j^p - x_i^p \right\| \leq d_{x_i^p}^{\max} \ \& \ \left\| x_j^- - x_i^- \right\| \leq d_{x_i^-}^{\max} \right\}. \tag{3.180}$$

Denote by $v_{(X^p,X^-),i}$ the volume of this (hyper-)rectangle. Repeat the same procedure in subspaces $\mathcal{S}_{X^-,Y^-}$ and $\mathcal{S}_{X^-}$.

Note that the important difference between the construction of the neighborhoods used in $\widehat{\mathrm{TE}}_{Y\to X\,\mathrm{k2}}$ and in $\widehat{\mathrm{TE}}_{Y\to X\,\mathrm{p1}}$ is that, for the first case, the minimum neighborhood including the k neighbors is constrained to be a Cartesian product of (hyper-)cubes and, in the second case, this neighborhood is a (hyper-)rectangle whose edges lengths can be completely different.

3.5. Discussion and Conclusion

In this chapter, we deeply discussed the estimation of information-theoretic quantities, especially mutual information and transfer entropy. Beginning with the mathematical definition of different information-theoretic quantities, we showed similarities between the estimations of mutual information and transfer entropy. Based on previous work, especially [Kraskov 2004], several questions have been raised. To answer them, we developed new calculation strategies following two different guidelines.

In section 3.3, an analytical form of the bias for the estimation of individual entropy was proposed. Once the new bias expression derived, we discussed bias cancellation strategies in the estimation of both mutual information and transfer entropy, and introduced bias reduction strategies based on an optimal choice of the neighborhood radius. Dealing with the maximum norm, our strategy explained the conclusions drawn by Kraskov *et al.* in [Kraskov 2004] and derived from numerical experiments. According to these conclusions, we proposed a "basic" estimator both for mutual information and transfer entropy, for different norms. The development of these strategies is based on the independence assumption provided that the bias could be partly cancelled when this hypothesis is not satisfied. So, to further eliminate the bias in dependence case, a weighted linear combination of distinct mutual information estimators, named "mixed" MI estimator, was introduced. In the same manner, we derived a "mixed" TE estimator.

In section 3.4, we focused on an idea already developed in existing literature [Kraskov 2004], where a (hyper-)rectangle was used instead of a (hyper-)cube. Following two different methodologies [Kozachenko 1987, Singh 2003], we extended the existing kNN entropy estimators to the rectangular situation, which resulted in two new entropy estimators we considered in the proposal of two novel TE estimators.

The different concepts and methodologies involved in this chapter were illustrated in Fig. 3.6, and the next chapter is devoted to the experimental comparison of all these new estimators, including both mutual information and transfer entropy, with existing techniques.

Note that one important problem has not been addressed in this chapter. It concerns the selection of the predictors memory sizes (m and n) we have to deal with. Clearly, these sizes must be specified in order to define the transfer entropy before estimating it. This choice was beyond the scope of our investigations which, for a given pair (m, n), focused on how to improve statistical performance (bias) of existing transfer entropy estimators. The mean conditional Kullback distance introduced in Equ. (3.28) clearly indicates that TE calculation is equivalent to compare two predictive (conditional) probability distributions to characterize a directional causal link, and can allow to choose between two classes of such distributions, the one which only depends on m values in the past of X and the one which depends on m values in the past of X but also on n values in the past of Y. The averaged log-likelihood ratio in Equ. (3.29) could be considered as an averaged generalized log-likelihood ratio depending on the unknown parameters m and n introduced to test the hypothesis in Equ. (3.26). That would imply the tuning of these parameters to maximize the ratio. This more complicated and general problem is not really addressed in the literature. The currently retained sub-optimal method introduces two steps (i) order selection (estimation) and (ii) TE computation for the chosen orders.

For Granger causality which imposes linear autoregressive modelization, the first step (i) is implemented with standard AIC or BIC algorithms [Aho 2014] (see Appendix E). When observed data are supposed to be generated by nonlinear mechanisms, the step (i) is generally implemented following the state reconstruction approach [Ponten 2007] which first subsamples the observation signals X and Y according to an "optimal" subsampling ratio, before selecting the n and m values. For some of the experimental results presented in chapter 5 on real signals, we used subsampling without any strict optimality requisite.

Chapter 4

Experimental Results

Previously, we proposed different strategies for the estimation of mutual information and transfer entropy. In this chapter, the performances of different algorithms are evaluated through numerical simulations considering various situations, including independent and dependent signals, linear and nonlinear relations. When we compare different estimators, we consider both the mean value and the corresponding standard deviation. For the physiology-based model for which the theoretical value cannot be derived, we use statistical hypothesis testing to validate if our approach reveals the ground truth.

4.1. Database

First of all, we present the 8 simulation models which are used hereafter to evaluate the performance of the algorithms presented in chapter 3. The first seven models are denoted Model 1,..., Model 7 and are abstract models. The eighth model is named physiology-based model because it is built from electrophysiological hypothesis concerning the origin of the iEEG signals. The first three models are dedicated to test mutual information estimators in the case of IID observations. The remaining five models are dedicated to test transfer entropy estimators. Except the physiology-based model, for all other models tested, the theoretical value of the estimated measures can be derived (Models 1 to 6) or computed by a Monte Carlo numerical approach (Model 7). Only Model 7 and the physiology-based model are nonlinear models. Nevertheless, linear models are relevant to compare statistical performance (bias and variance) of different existing and proposed MI or TE estimators.

4.1.1. Abstract Models

This section is divided in two subsections, the first one dealing with MI estimation (Models 1 to 3), the second one with TE estimation (Models 4 to 7).

4.1.1.1. Models for MI Estimation

For the comparison of different mutual information estimators, we first consider the independence situation *i.e.* we first generated two independent d-dimensional IID random sequences $(X_t)_t$ and $(Y_t)_t$ such that both X_t and Y_t followed a zero mean Gaussian distribution $\mathcal{N}(0, \mathcal{C})$, where $\mathcal{C}$ was a Toeplitz matrix with first line $[1, \alpha, \ldots, \alpha^{d-1}]$. For our simulations, we used sequences of N independent samples $(X_t, Y_t)_{t=1,\ldots,N}$. This model is named as Model 1:

Model 1

$$\begin{aligned} &X \text{ and } Y \text{ independent, } (X_t)_{t=1,\ldots,N} : \text{IID, } (Y_t)_{t=1,\ldots,N} : \text{IID} \\ &X_t \sim \mathcal{N}(0, \mathcal{C}),\ Y_t \sim \mathcal{N}(0, \mathcal{C}),\ \mathcal{C} = \text{toeplitz}\left(1, \alpha, \ldots, \alpha^{d-1}\right) \end{aligned} \tag{4.1}$$

Clearly, whatever the value of $\alpha \in [0, 1[$, the mutual information $\mathcal{I}(X_t, Y_t)$ is theoretically equal to zero.

Additionally, in order to briefly investigate the effect of non-independence on the bias of MI estimation when applying the different strategies, we also considered two dependence situations. First, we replaced the independent pairs (X_t, Y_t) mentioned above (see Model 1) by dependent pairs, (X_t, Y_{1t}), where X_t was the same as previously and Y_t was replaced by Y_{1t}:

Model 2

$$\begin{aligned} &X \text{ and } Y \text{ as in Model 1} \\ &Y_{1t} = \cos\theta \cdot X_t + \sin\theta \cdot Y_t,\ t = 1, \ldots, N \end{aligned} \tag{4.2}$$

The parameter θ, $\theta \in \left[0, \frac{\pi}{2}\right]$, allowed to tune the dependence between X_t and Y_{1t}. Note that, for $\theta = \frac{\pi}{2}$, X_t and Y_{1t} are independent. This model is named as Model 2 and the theoretical value of $\mathcal{I}(X_t, Y_{1t})$ is equal to $-d \log(\sin\theta)$. The derivation of this theoretical value can be found in Appendix F.

In a second dependence situation, data samples $(X_t)_{t=1,\ldots,N}$ and $(Y_t)_{t=1,\ldots,N}$ were generated by the following linear model more specifically to focus on the impact of the increase in dimension of the simulated random vectors X_t and Y_t. To this end, for both of them, their components were mutually independent. This model is denoted Model 3:

Model 3

$$\begin{aligned}
&\text{Random sequences } X \text{ and } e \text{ independent, } (X_t)_{t=1,\dots,N} : \text{IID}, (e_t)_{t=1,\dots,N} : \text{IID} \\
&X_t \sim \mathcal{N}(0, I),\ e_t \sim \mathcal{N}(0, I) \\
&Y_t = X_t + \beta \cdot e_t,\ \beta \in \mathbb{R}
\end{aligned} \tag{4.3}$$

where X_t and e_t were two independent d-dimensional random vectors, and both of them followed a zero mean Gaussian distribution $\mathcal{N}(0, I)$ (I is the identity matrix, and in this case, it is easier to test the effect of dimensionality). β is a scalar coefficient. Clearly, when β decreases, the dependence between X_t and Y_t increases. The theoretical value of mutual information $\mathcal{I}(X_t, Y_t)$ is equal to $\frac{d}{2} \log\left(\frac{1+\beta^2}{\beta^2}\right)$ as detailed in Appendix G.

4.1.1.2. Models for TE Estimation

For transfer entropy, we tested both Gaussian IID and Gaussian AR models, as well as linear and nonlinear situations.

Model 4 was proposed to simulate an IID sequence $\left(X_i^p, X_i^-, Y_i^-\right)_i$ as introduced in chapter 3 except that the temporal statistical dependence inherent in the temporal correlation of the pair of observed processes (X, Y) has been deliberately destroyed in order to fulfill the IID observations hypothesis imposed in the theoretical derivation of the entropy estimators.

Model 4

$$\begin{aligned}
&X_t = aY_t + bZ_t + W_t,\ W_t \in \mathbb{R},\ Y \in \mathbb{R}^{d_Y},\ Z \in \mathbb{R}^{d_Z}, \\
&Y, Z, W :\ \text{mutually independent processes} \\
&Y_t \sim \mathcal{N}(0, \mathcal{C}_Y),\ Z_t \sim \mathcal{N}(0, \mathcal{C}_Z),\ W_t \sim \mathcal{N}\left(0, \sigma_W^2\right) \\
&\mathcal{C}_Y = \text{toeplitz}(1, \alpha, \dots, \alpha^{d_Y - 1}),\ \mathcal{C}_Z = \text{toeplitz}(1, \alpha, \dots, \alpha^{d_Z - 1})
\end{aligned} \tag{4.4}$$

For the matrix $\mathcal{C}_Y$, we chose $\alpha = 0.5$, and, for $\mathcal{C}_Z$, $\alpha = 0.2$. The standard deviation σ_W was set to 0.5. The vectors a and b were such that $a = 0.1 * [1, 2, \dots, d_Y]$ and $b = 0.1 * [d_Z, d_Z - 1, \dots, 1]$.

With this model, we aimed at estimating $\mathcal{H}(X_t|Y_t) - \mathcal{H}(X_t|Y_t, Z_t)$ to test if the knowledge of Y_t and Z_t could improve the prediction of X_t compared to only the knowledge of Y_t. The triplet (X_t, Y_t, Z_t) corresponds to the triplet $\left(X_t^p, X_t^-, Y_t^-\right)$ introduced previously in chapter 3 to define transfer entropy. Here the theoretical value of TE is $\mathcal{H}(X_t|Y_t) - \mathcal{H}(X_t|Y_t, Z_t) = \mathcal{H}(bZ_t + W_t) - \mathcal{H}(W_t)$, *i.e.* $\text{TE} = \mathcal{H}\left(\mathcal{N}\left(0, bC_Z b^T + \sigma_W^2\right)\right) - \mathcal{H}\left(\mathcal{N}\left(0, \sigma_W^2\right)\right)$ and can be easily computed.

The two following models (Model 5 and Model 6) are two VAR models made up of

either two or three one-dimensional signals. For both models, there exists a bidirectional relation between each pair of signals. The PSD of these signals can be found in Appendix H. The models coefficients have been tuned in order to obtain signals displaying narrow bounds PSD shapes that can be retrieved in real epileptic signals.

The first vectorial AR model (marked as Model 5) was as follows:

Model 5

$$\begin{gathered} e_X,\ e_Y,\ X,\ Y:\ \text{random real scalar sequences} \\ e_X,\ e_Y:\ \text{independent } \mathcal{N}(0,1) \text{ white sequences} \\ \begin{cases} X_t = 0.45\sqrt{2}X_{t-1} - 0.9X_{t-2} - 0.6Y_{t-2} + e_{X,t} \\ Y_t = 0.6X_{t-2} - 0.175\sqrt{2}Y_{t-1} + 0.55\sqrt{2}Y_{t-2} + e_{Y,t} \end{cases}, t = 1, \ldots, N \end{gathered} \tag{4.5}$$

The second vectorial AR model (marked as Model 6) was given by:

Model 6

$$\begin{gathered} e_X,\ e_Y,\ e_Z,\ X,\ Y,\ Z:\ \text{random real scalar sequences} \\ e_X,\ e_Y,\ e_Z:\ 3 \text{ independent } \mathcal{N}(0,1) \text{ white sequences} \\ \begin{cases} X_t = -0.25X_{t-2} - 0.35Y_{t-2} + 0.35Z_{t-2} + e_{X,t} \\ Y_t = -0.5X_{t-1} + 0.25Y_{t-1} - 0.5Z_{t-3} + e_{Y,t} \\ Z_t = -0.6X_{t-2} - 0.7Y_{t-2} - 0.2Z_{t-2} + e_{Z,t} \end{cases}, t = 1, \ldots, N \end{gathered} \tag{4.6}$$

In Section 4.2, in order to estimate both TE and Granger causality index, the prediction orders m and n will be equal to the corresponding regression orders of the AR models. For example, when estimating $\mathrm{TE}_{Y \to X}$, we set $m = 2$, $n = 2$ and $\left(X_t^p, X_t^-, Y_t^-\right)$ corresponds to $\left(X_{t+1}, X_t^{(2)}, Y_t^{(2)}\right)$. As the 3 stochastic processes X, Y, Z are jointly Gaussian distributed the 6 theoretical TE values can be obtained from theoretical calculation of the corresponding Granger causality indexes. These latter can be computed from theoretical covariances of (X, Y, Z) or from their estimation on a sufficiently large time interval.

The following Model 7 was introduced as an example of nonlinear causal contribution of Y_i^- onto X_i^p to illustrate the interest of entropic methods versus Granger causality in nonlinear situations. An *a priori* natural approach should have been to consider nonlinear VAR processes. Now, with this type of model the theoretical computation of TE values should have been cumbersome. With the proposed model we were able to obtain a precise numerical approximation of the theoretical value of TE.

Model 7

$$\begin{aligned} &W, Y, Z: \text{ 3 independent } \mathcal{N}(0,1) \text{ scalar real white sequences} \\ &Y_t \sim \mathcal{N}(0,1), \ Z_t \sim \mathcal{N}(0,1), \ W_t \sim \mathcal{N}(0,1) \text{ uniformly distributed on } [0,\theta] \subset \mathbb{R} \\ &X_t = K \cdot \left(rep \cdot Y_t + \sqrt{(1-rep^2)} Z_t^2 \right) + W_t, \ t = 1, \ldots, N, \ K \in \mathbb{R}, \ rep \in \mathbb{R} \end{aligned} \tag{4.7}$$

The parameter K allowed to weight the influences of (Y_t, Z_t). The parameter rep was a weighting coefficient to modify the influences of Y_t and Z_t. In this model, (X_t, Y_t, Z_t) corresponds to $\left(X_t^p, X_t^-, Y_t^-\right)$ in chapter 3, *i.e.* to $\left(X_{t+1}, X_t^{(n)}, Y_t^{(m)}\right)$ with $m = n = 1$. An illustration of the nonlinear statistical link between X_t and Z_t is given in Fig. 4.1. The visualized dimensional distribution clearly illustrates that a strong statistical dependence does not prevent from a null correlation, leading to a Granger causality index equal to zero.

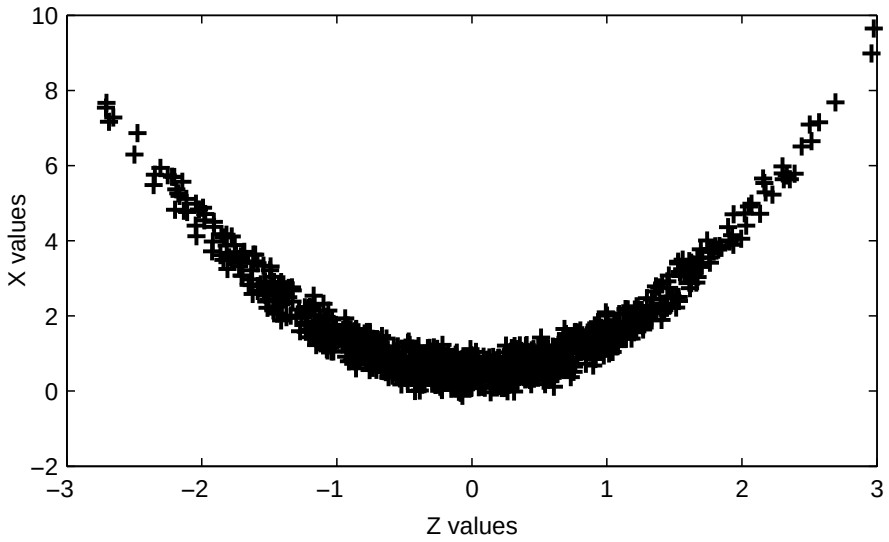


Figure 4.1: Experimental marginal distribution for (X_t, Z_t) with $K = 1$, $rep = 0.1$, $\theta = 1$.

To obtain the theoretical value of TE we started from the following expression

$$\begin{aligned} \mathrm{TE}_{Z \to X} &= \mathcal{H}(X_t | Y_t) - \mathcal{H}(X_t | Y_t, Z_t) \\ &= \mathcal{H}(U_t) - \mathcal{H}(W_t) \end{aligned} \tag{4.8}$$

where $U_t = K\sqrt{(1-rep^2)} Z_t^2 + W_t$.

The entropy $\mathcal{H}(W_t)$ is known as W_t follows a Gaussian distribution. To compute $\mathcal{H}(U_t)$ we first obtain an approximation $\tilde{p}_{U_t}$ of the probability density function of U_t by convolving numerically the respective density functions of W_t and $K\sqrt{(1-rep^2)} Z_t^2$

(the theoretical density function of Z_t^2 is known as $Z_t \sim \mathcal{N}(0,1)$). Then, given a large number N_{MC} of realizations (z_t, w_t) of (Z_t, W_t) we computed the following Monte Carlo approximation:

$$\mathcal{H}(U_t) \simeq -\frac{1}{N_{\mathrm{MC}}} \sum_{t=1}^{N_{\mathrm{MC}}} \tilde{p}_{U_t}\left(K\sqrt{(1-rep^2)}z_t^2 + w_t\right) \times \log\left(\tilde{p}_{U_t}\left(K\sqrt{(1-rep^2)}z_t^2 + w_t\right)\right) \tag{4.9}$$

which led to an accurate estimation of $\mathcal{H}(U_t)$.

4.1.2. Physiology-based Model

This model aims at simulating more realistic iEEG signals as it is based on structural and functional hypotheses on brain neural populations organization and electrophysiological activity. As the firing mean activity of a neural population is a nonlinear response to afferent synaptic excitation/inhibition, nonlinear static sigmoidal operators are included in the model.

4.1.2.1. Presentation

For the physiology-based model, we used a time continuous SDE (stochastic differential equations) model simulated in discrete time to represent the electrical activity of two distant, and possibly coupled, neuronal populations denoted $Pop_{\boldsymbol{X}}$ and $Pop_{\boldsymbol{Y}}$. It was based on the physiology and introduced in [Wendling 2005] to produce outputs similar to intracranial electroencephalographic signals as those recorded with proximal electrodes in hippocampus.

For each population, this type of model generates a mean population membrane potential that is converted to an iEEG signal [Wendling 2005]. Each population is composed of three neuronal subpopulations that mutually interact: a main pyramidal cells subpopulation of excitatory neurons and two inhibitory subpopulations (Fig. 4.2). The main subpopulation has an excitatory feedback loop. The first inhibitory subpopulation projects onto the dendritic region of the main population and onto the second inhibitory one. As for the second inhibitory subpopulation, it projects onto the somatic area of primary neurons. The corresponding mathematical representation of this population model is given in graphical form in Fig. 4.3 [Frogerais 2008]. In this figure, the three subpopulations, denoted P_e, P_{s_i} and P_{f_i} respectively, appear in three boxes underlined by dotted lines. The input $W(t)$ represents the random influence of distant afferent neurons and is classed as time continuous white Gaussian noise (*i.e.* the formal derivation of a Brownian process). In each subpopulation box the synaptic transduction and dendritic time

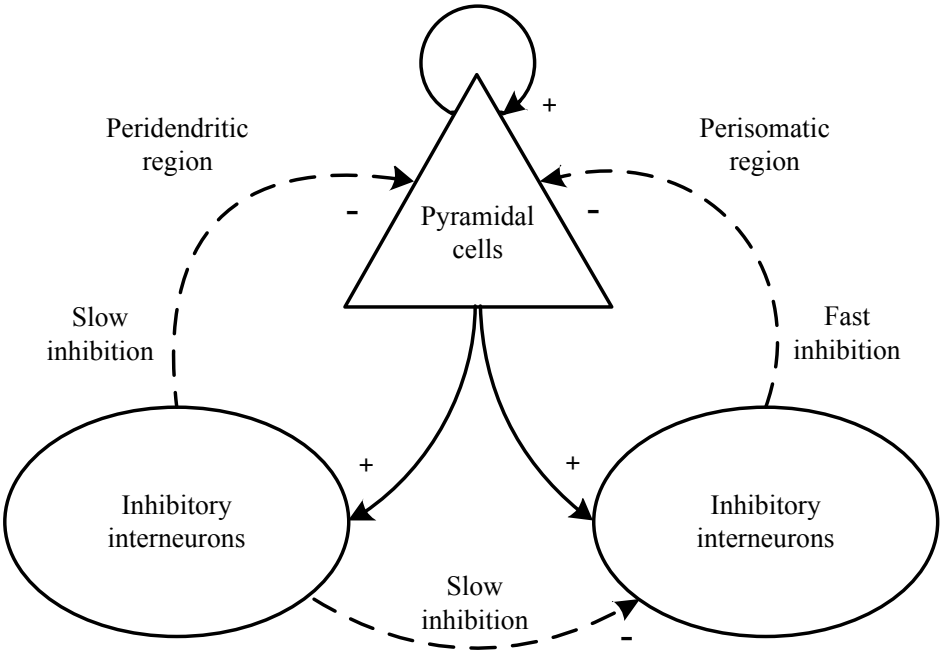


Figure 4.2: Interactions between three neuronal subpopulations of the hippocampus [Wendling 2005]. The symbols "+" and "-" represent excitatory and inhibitory afferences respectively.

constant effects are represented by linear transfer functions corresponding to 3 types of time continuous impulse responses: $h_e(t)$ for the excitatory kinetics, $h_{fi}(t)$ for the fast somatic inhibitory kinetics, $h_{si}(t)$ for the slow inhibitory kinetics. Sigmoidal functions $S(\cdot)$ are also included in the subpopulations models, as a conversion law from mean neuronal membrane potential to mean rate of axonal action potentials. The impulse response $G_{PH}h_{ph}$ (where G_{PH} is the static gain) is that of an instrumentation high-pass filter, whose output is sampled at 256 Hz. Its transfer function is $sG_{PH}/(1+\tau s)$ where s is the Laplace variable. The coefficients C_i, $i = 1, \ldots, 7$, represent the average numbers of synaptic connections from a subpopulation to another. All impulse responses are of the form $h(t) = \alpha t \exp(-\alpha t)$, $t \geq 0$, α being the inverse of a time constant, noted a for the excitation, b for the slow inhibition and g for the fast one. Their input-output relations can be characterized by a pair of first-order differential equations. The coefficients A, B, G represent synaptic efficiency (synaptic gains) for excitation, dendritic inhibition and somatic inhibition respectively. Only these three parameters are supposed to vary during a transition from a normal process to the epileptic seizure and are the ones to be tuned to simulate different types of activity, from healthy activity to paroxystic one.

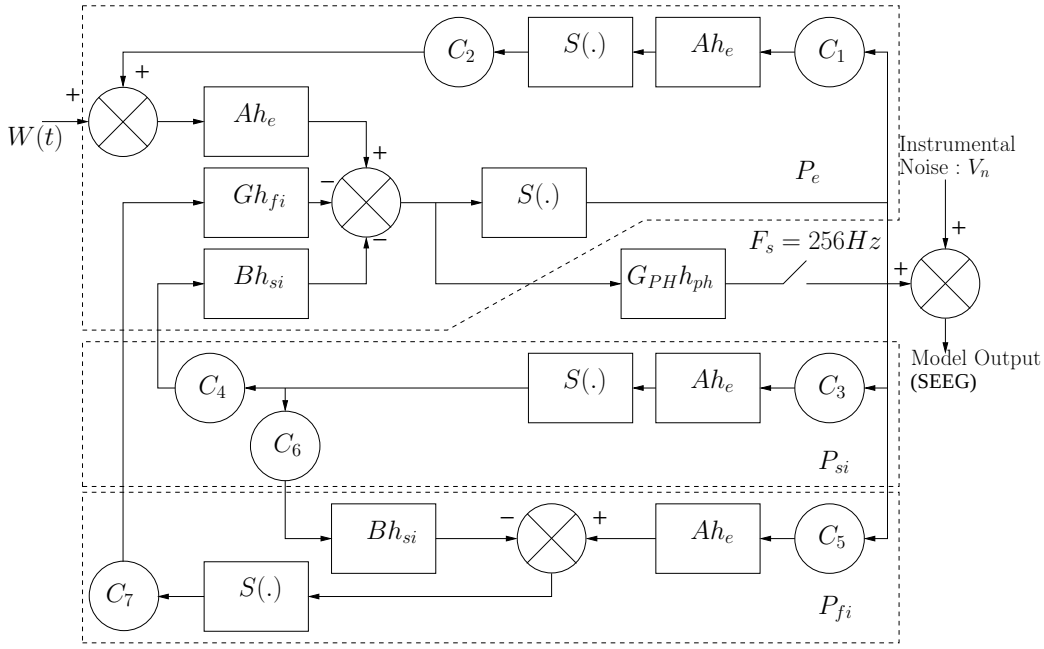


FIG. 2.9 – Voir figure 2.10

Figure 4.3: Interactions between different neuronal subpopulations of the hippocampus [Wendling 2005, Frogerais 2008].

dèle mathématique proprement dit, elle est donnée sous forme graphique dans la figure 2.9, les variables étant explicitées dans la figure 2.10. Ce modèle comporte en fait 4 sous-populations car ce qui est modélisé pour le feed-back positif de la population principale sur elle-même ne correspond pas à un feed-back d'une cellule principale directement sur elle-même. Le système d'équations correspondant est donné ci-dessous.

For one population the complete system of equations can be written

$$\begin{cases} \mathrm{d}x_i = x_{i+5}\mathrm{d}t & i = 0,\ldots,4 \\ \mathrm{d}x_5 = \left(AaS\left(x_1 - x_2 - x_3\right) - 2ax_5 - a^2x_0\right)\mathrm{d}t \\ \mathrm{d}x_6 = \left(Aa\left(m_p + C_2S\left(C_1x_0\right)\right) - 2ax_6 - a^2x_1\right)\mathrm{d}t + Aa\mathrm{d}\beta \\ \mathrm{d}x_7 = \left(BbC_4S\left(C_3x_0\right) - 2bx_7 - b^2x_2\right)\mathrm{d}t \\ \mathrm{d}x_8 = \left(GgC_7S\left(C_5x_0 - C_6x_4\right) - 2gx_8 - g^2x_3\right)\mathrm{d}t \\ \mathrm{d}x_9 = \left(BjS\left(C_3x_0\right) - 2jx_9 - j^2x_4\right)\mathrm{d}t \\ \mathrm{d}x_{10} = \left(G_{PH}\left(x_6 - x_7 - x_8\right) - \frac{1}{\tau}x_{10}\right)\mathrm{d}t. \end{cases} \tag{4.10}$$

(2.2)

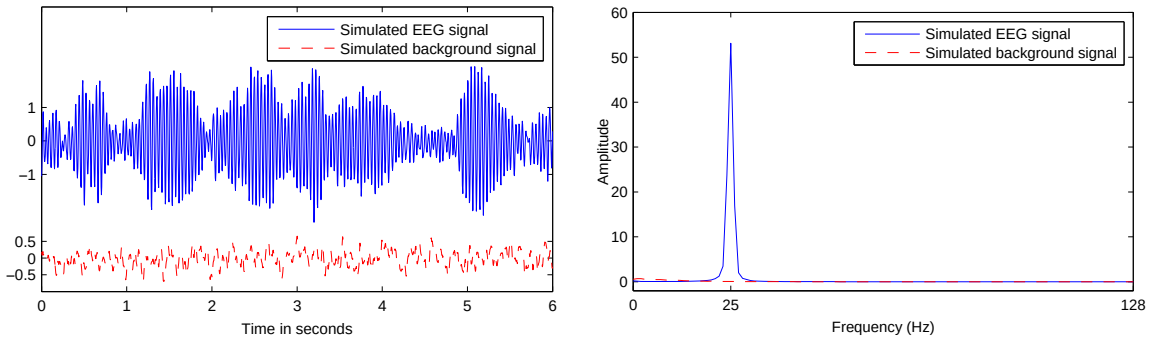
On remarque que ce système est d'ordre 11. Or si on dénombre les fonctions de transfert d'ordre 2 des filtres pré-somatiques dans la figure 2.9 on en trouve 7, ce qui devrait mener à l'introduction de 14 composantes d'état. Cependant on constate que les 3 fonctions de transfert Ah_e apparaissant sur la droite de la figure 2.9 peuvent être fusionnées à la simple condition de les commuter avec les gains C_1, C_2, C_4, ce qui économise 4 composantes d'état. La fonction de transfert du filtre passe-haut de sortie étant d'ordre 1 elle ne nécessite qu'une composante d'état complémentaire et on arrive ainsi à 14-4+1=11 pour la dimension du vecteur d'état et celle du système différentiel. En notant $W(t)\mathrm{d}t = \mathrm{d}\beta(t)$, où $\beta(t)$ désigne un processus brownien de coefficient de diffusion égal à un (voir aussi section 3.2), le système peut alors s'écrire sous la forme d'une équation différentielle stochastique à 11 composantes $\mathrm{d}X(t) = f(X(t),\theta,\gamma)\mathrm{d}t + G\mathrm{d}\beta(t)$. Les vecteurs θ et γ regroupent respectivement les paramètres inconnus et les paramètres connus. Les paramètres liés aux efficacités de l'excitation et des deux types d'inhibition (A, B, G) sont à placer dans θ mais il faut évidemment leur

The connectivity constants C_i, $i = 1,\ldots,7$, the synaptic time constants $1/a$, $1/b$ and $1/g$ and the mean and standard deviation of the input Gaussian process were fixed as in [Wendling 2001, Wendling 2005]. Their values are reported in Tab. 4.1. Generated signals and their corresponding PSD are shown in Fig. 4.4, for two triplets $\{A, B, G\}$ to simulate either background activity or narrow band activity observed at a seizure onset.

Since the pyramidal cells project their axons to other areas of the brain, the model output is the membrane potential of the main cells of one population Pop_X as an excitatory input to the main cells inputs of a second population Pop_Y. In addition, this connection from population Pop_X to Pop_Y

Synaptic time constants		
$1/a$	Excitatory	$1/100$
$1/b$	Slow inhibitory	$1/30$
$1/g$	Fast inhibitory	$1/350$
Connectivity constants		
C_1	$p_e - p_e$	135
C_2	$p_e - p_e$	108
C_3	$p_e - p_{si}$	33.8
C_4	$p_{si} - p_e$	33.8
C_5	$p_e - p_{fi}$	40.5
C_6	$p_{si} - p_{fi}$	13.5
C_7	$p_{fi} - p_e$	121.5
White Gaussian noise (input)		
m_p	mean	90
σ	diffusion	30
Sigmoid		
e_0		2.5 s^{-1}
v_0		6 mV
r		0.56 mV^{-1}

Table 4.1: Example of model constants of hippocampus.



(a) (b)

Figure 4.4: Example of signals generated by the model. (a) simulated EEG signal ($A = 3.67$, $B = 2$ and $G = 22.45$) and simulated background signal ($A = 2.8$, $B = 1$ and $G = 40$). (b) corresponding PSD.

is represented by a parameter K^{XY} which is proportional to the number of corresponding active axonal links for a given type of cerebral activity. An appropriate setting of this parameter allows for building systems where the neuronal populations are coupled either unidirectionally or bidirectionally. The other parameters of this model are internal parameters (inside the population itself). They include excitatory and inhibitory gains in the feedback loops as well as coefficients related to the number of synaptic contacts between subpopulations. They are adjusted to control the intrinsic activity of each population (normal background activity versus epileptic activity).

For the physiology-based model with two populations Pop_X and Pop_Y, we consider 3 situations: 1) Pop_X has an influence on Pop_Y (or the inverse, which corresponds to an unidirectional propagation), 2) both Pop_X and Pop_Y have influence on the other one (bidirectional propagation), 3) Pop_X and Pop_Y are independent. For these 3 cases, the values of the parameters A, B, G and g can be found in the following table.

		Pop_X				Pop_Y			
		A	B	G	g	A	B	G	g
Case 1	$K^{XY} = 1500$, $K^{YX} = 0$	5	3	20	250	3.5	3.5	84	250
Case 2	$K^{XY} = K^{YX} = 1500$	2.8	1	40	250	3.2	1	32.5	250
Case 3	$K^{XY} = K^{YX} = 0$	5	3	20	250	3.67	2.3	22.45	250

Table 4.2: Parameters of the physiology-based model. Case 1: unidirectional situation, $X \rightarrow Y$, case 2: bidirectional situation, $X \leftrightarrows Y$, case 3: independence situation (see Appendix H for the PSD of these simulated EEG signals).

4.1.2.2. Surrogate Strategy

For the physiology-based model, for which the theoretical value of TE is not available for two connected populations, the situations are more complicated. As a matter of fact, for two populations Pop_X and Pop_Y generating signals X and Y for which we measure transfer entropy $\text{TE}_{Y \rightarrow X}$ (given that a model of connectivity is available for these two populations), the problem is to evaluate the deviation from the H_0 hypothesis (X and Y independent), since it is difficult to obtain a theoretical distribution of $\text{TE}_{Y \rightarrow X}$ under H_0. This difficulty can be overcome using surrogate data synthesized from the original data and guaranteeing their independence to get a reference statistics under H_0. To this end, we must develop a strategy to modify the observed signals to make them independent while preserving their marginal frequential characteristics (the variance of any statistics computed from these observations depends on these characteristics). In this work, surrogate pairs were generated with the following strategy. Let us give a series of M independent realizations (x^m, y^m), $m = 1, \ldots, M$, obtained with the same model (same structure, same parameters). To build independent pairs (X, Y') preserving marginal laws, we introduced new pairs $(X^m, Y^{m'})$ where $m \neq m'$. According to this strategy, two different sets of TE values can be obtained, $\text{TE}_{Y \rightarrow X}$ and $\text{TE}_{Y' \rightarrow X}$. If there is an influence from signal Y to signal X, the distributions of $\text{TE}_{Y \rightarrow X}$ and $\text{TE}_{Y' \rightarrow X}$ are different. Otherwise, $\text{TE}_{Y \rightarrow X}$ and $\text{TE}_{Y' \rightarrow X}$ follow the same distribution. We used Student's t-test to decide whether two sets of values are significantly different. This surrogate strategy was also used to test if MI was null or strictly positive.

4.1.2.3. Model Order Selection

To determine approximately parameters m and n in TE estimators, we imposed $m = n$ and computed AIC and BIC indexes versus order for a VAR model. As shown in Fig. 4.5, for both AIC and BIC, the curves decrease rapidly from the starting point, and then remain quite constant. In this case, a very large optimal order can be selected. This situation should be avoided, because (i) a very large order will lead to large estimation error due to the "curse of dimensionality" , (ii) the computation time increases dramatically with an increasing model order. Hence, the model order was set to 6 in this experiment. For MI estimators the size of the vectors, to be tested as to be independent or not, was fixed to the same value m used for TE estimators.

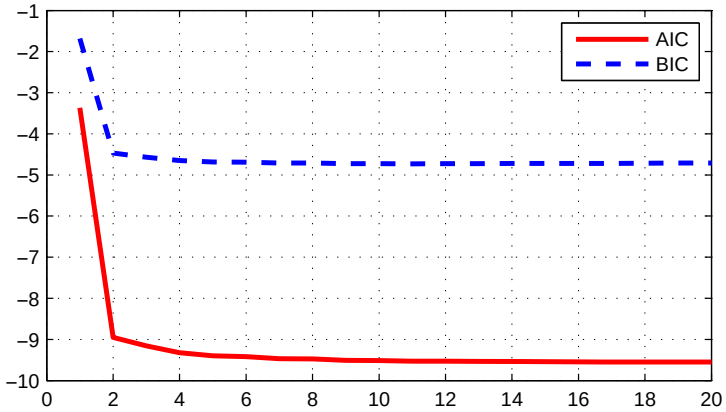


Figure 4.5: AIC and BIC indexes versus order computed under VAR hypothesis on the physiology-based model outputs (case 1: unidirectional connectivity).

4.2. Simulation Results

In this section, we present the results of simulation of the different algorithms we tested. Models 1∼3 are used for the evaluation of different MI estimators (section 4.2.1), and different TE estimators are compared using Models 4∼7 (section 4.2.2). For the physiology-based model, both MI and TE estimators are tested (section 4.2.3). Here, both the mean value and the corresponding standard deviation are plotted in the figures. Hence, for clarity, a small decay is set among all the curves in each figure.

4.2.1. Results on Mutual Information

For mutual information, we tested the 4 different MI estimators $\widehat{\mathcal{I}(X,Y)}_{\text{basic,E}}$ (Equ. (3.106) with the Euclidean norm), $\widehat{\mathcal{I}(X,Y)}_{\text{basic,M}}$ (Equ. (3.106) with the maximum norm), $\widehat{\mathcal{I}(X,Y)}_{\text{mixed,E}}$ (Equ. (3.113) with the Euclidean norm) and $\widehat{\mathcal{I}(X,Y)}_{\text{mixed,M}}$ (Equ. (3.113) with the maximum norm) to estimate $\mathcal{I}(X,Y)$. We also ran the MI estimator

algorithm freely available from the MILCA toolbox [Sergey 2015], simply denoted by MILCA. Here, both $\widehat{\mathcal{I}(X,Y)}_{\text{basic,M}}$ and MILCA use maximum norm and the same distance for all spaces, including both joint space and marginal spaces. Note that they differ in the way that $\widehat{\mathcal{I}(X,Y)}_{\text{basic,M}}$ estimates the probability density for each data sample, while MILCA calculates MI directly (based on the kNN entropy estimator, Equ. (3.60)). For comparison, the strategy of same k was also tested with both maximum norm and Euclidean norm, where the same number of neighbors k was imposed for the 3 individual entropies. Throughout the experiments, we chose $k = 6$ (the default k value in MILCA toolbox) for the basic estimators, and $k_1 = 6$, $k_2 = 20$ for the mixed estimators. The statistical mean and variance of these estimators were estimated by an averaging on 100 trials.

4.2.1.1. Model 1

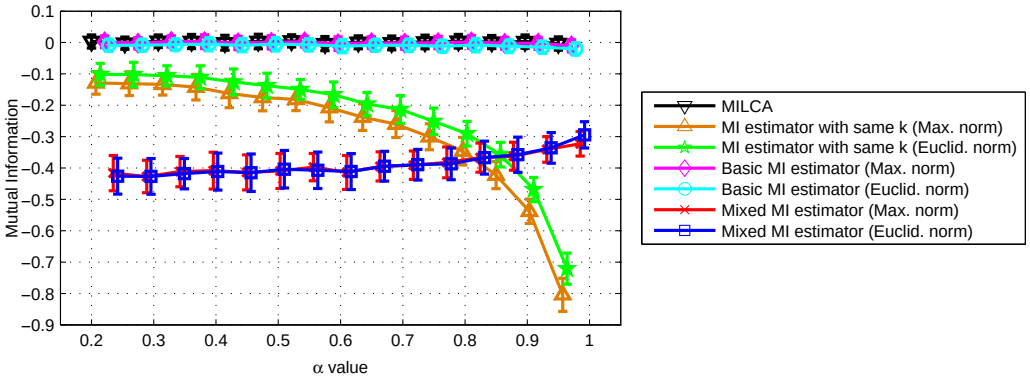
Fig. 4.6 displays the performance of different approaches in the independence case (Model 1) with the two norms (maximum and Euclidean norms). For the two estimators using the same k, the performance drastically falls with increasing α (Fig. 4.6(a)) and high dimensions (greater than 4) (Fig. 4.6(b)). The estimators with a given neighborhood size (basic estimators and MILCA) clearly outperform the former significantly whatever the norm, in terms of estimation bias and standard deviation. Unsurprisingly, in such an independence situation, the mixed estimators, which are designed for the dependence situation, suffer from a large bias whatever the norm. In other words, the new justified strategy (Boxes ⑫ and ⑬ in Fig. 3.6) provides reliable mutual information values for independence test, even with short signal lengths or high dimensional signals.

4.2.1.2. Model 2

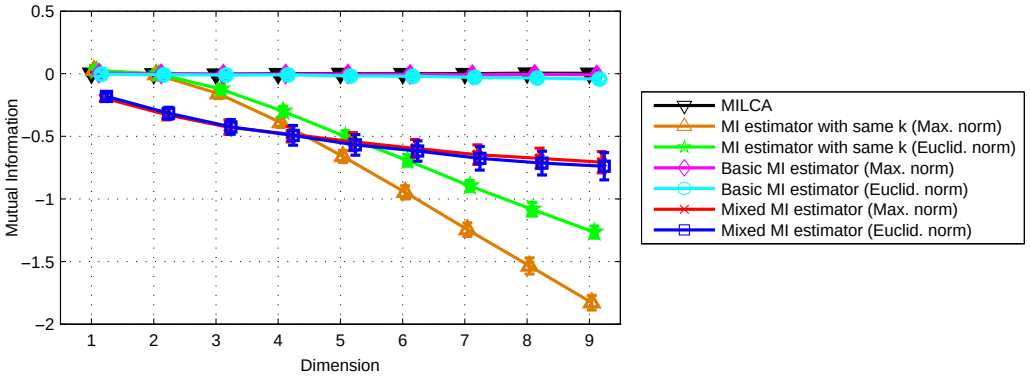
For Model 2, results are displayed in Fig. 4.7. Whatever the norm, when the signals are dependent (θ close to $\frac{\pi}{16}$), the mixed estimators are more accurate. Now, when the dependence between the signals decreases (θ close to $\frac{\pi}{2}$), the experimental mutual information values provided by the basic estimators and MILCA become very close to the theoretical one. It should be mentioned that all these three estimators follow the strategy derived in chapter 3 (based on the independence assumption, Boxes ⑫ and ⑬ in Fig. 3.6).

4.2.1.3. Model 3

Fig. 4.8(a) displays the performance of different algorithms, for a given dimension ($d = 3$), a number of points equal to $N = 512$, and different values of β. The correlation between X and Y is all the more important as β is low. It comes out that all estimators



(a)



(b)

Figure 4.6: (Model 1) Mutual information $\widehat{\mathcal{I}(X,Y)}$ estimation for independent signals using different strategies with 100 trials. (a) Mutual information $\widehat{\mathcal{I}(X,Y)}$ (in nats) estimated with varying α, $d = 3$, $N = 512$. (b) Mutual information $\widehat{\mathcal{I}(X,Y)}$ (in nats) estimated versus dimension, $\alpha = 0.3$, $N = 512$.

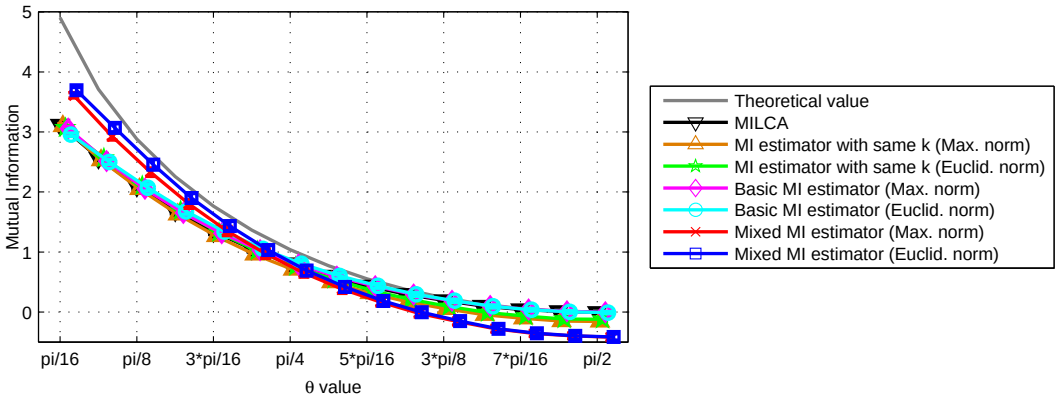
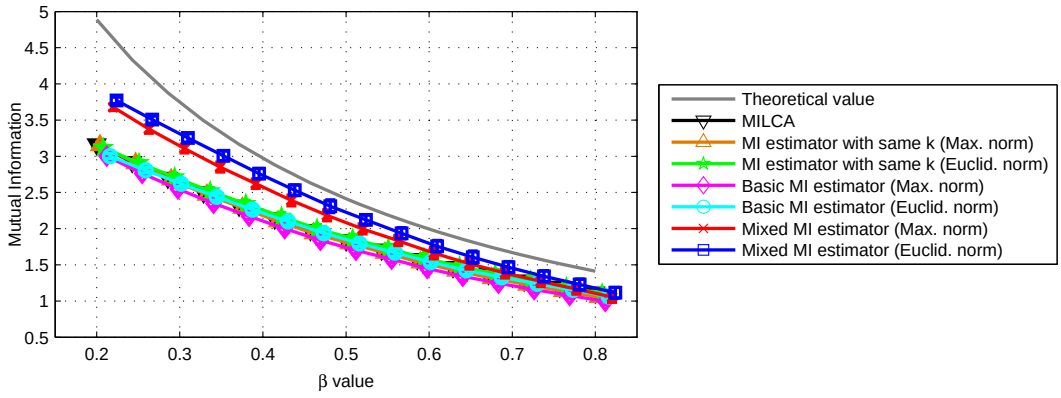
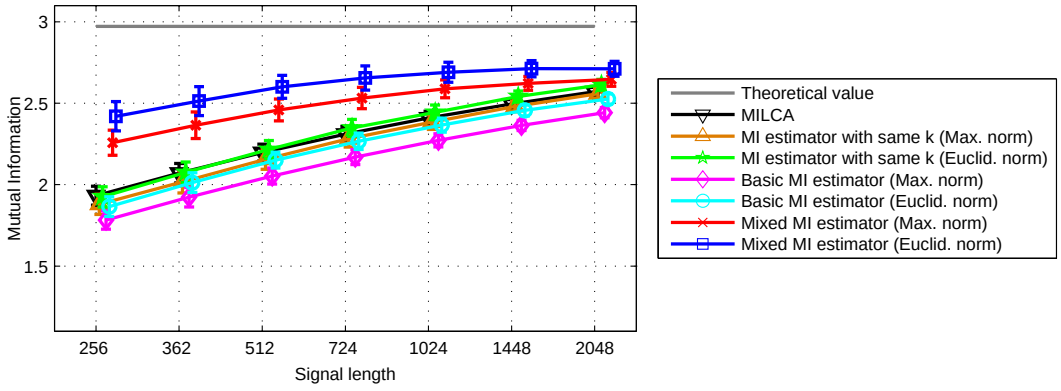


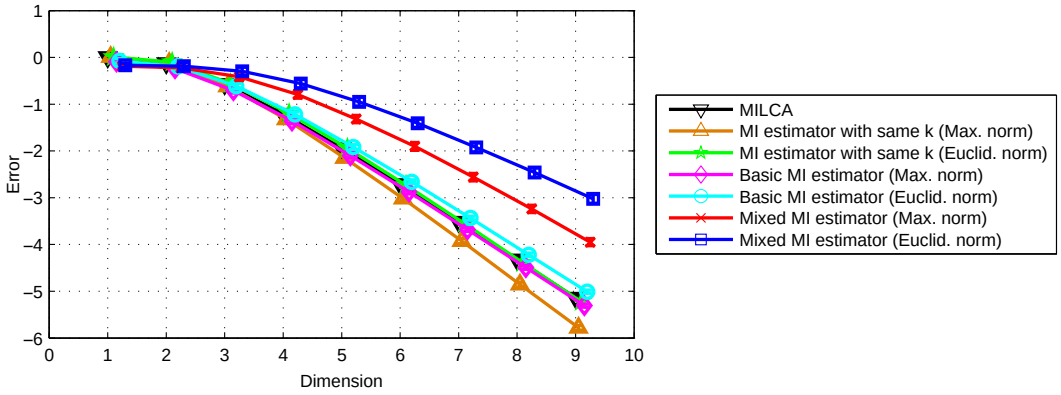
Figure 4.7: (Model 2) Mutual information $\widehat{\mathcal{I}(X,Y_1)}$ (in nats) estimation using different strategies with varying θ, $\alpha = 0.4$, $d = 3$, $N = 512$, 100 trials.



(a)



(b)



(c)

Figure 4.8: (Model 3) Mutual information and mean estimation error (estimation bias) using different estimators with 100 trials. (a) Mutual information (in nats) estimated with varying β, $d = 3$, $N = 512$. (b) Mutual information (in nats) estimated with different signals lengths, $\beta = 0.4$, $d = 3$. (c) Mean estimation error $\widehat{\mathcal{I}(X,Y)} - \mathcal{I}(X,Y)$ (in nats) with varying dimension, $\beta = 0.5$, $N = 512$.

are comparable when β reaches 0.8 (corresponding to a correlation coefficient around 0.78 between X and Y). When the signals are highly correlated (low values of β), the basic estimators still show identical behaviors, but, in this case, the two new mixed estimators clearly outperform the former whatever the norm. Even if all results are not presented here, we find that the two new estimators outperform the basic ones using either $k = 6$ or $k = 20$.

As displayed in Fig. 4.8(b), we also tested the five estimators for different lengths of the time series for given values of β and d. The two new mixed estimators behave better whatever the length of the signals (ranging from 256 until 2048 points), the improvement being all the more important that the signal length is short.

When computing the mean error between the different estimators and the theoretical value, for a given value of β ($\beta = 0.5$) corresponding to a correlation coefficient between the signals equal to 0.89, and an increasing dimension, the same conclusion globally holds, as displayed in Fig. 4.8(c). The new mixed estimators clearly outperform the basic ones (which display comparable behavior) especially for high dimensions. However, for very low dimensions ($d = 1$ or $d = 2$), the original estimators may be preferred. Clearly, for all estimators, the error grows along with the dimension, the best result being systematically obtained with the mixed estimator based on the Euclidean norm. Since the standard deviations are quite low, they are not shown in these figures. Using the basic estimators (or MILCA), the standard deviation varies from 0.03 to 0.06 which is extremely low compared to the estimated values of mutual information (approximately from 1 to 5). As for the mixed estimators, the standard deviation varies from 0.04 to 0.09. The increasing in standard deviation can be considered as negligible in comparison to the accuracy of the estimation.

4.2.2. Results on Transfer Entropy

For a complete comparison, beyond the theoretical value of TE, we also computed the Granger causality index as a reference (as indicated previously, in the case of linear Gaussian signals, transfer entropy and Granger causality index are equivalent up to a factor of 2). In the following figures, GCi/2 corresponds to Granger causality index GC divided by 2, transfer entropy estimated by the free TRENTOOL toolbox (corresponding to Equ. (3.67)) is marked as Standard algorithm, that estimated by JIDT (corresponding to Equ. (3.68)) is marked as Extended algorithm, $\mathrm{TE}_{\mathrm{p1}}$ is the transfer entropy estimator given by Equ. (3.176) and $\mathrm{TE}_{\mathrm{p2}}$ is the transfer entropy estimator given by Equ. (3.177). Additionally, the basic TE estimators (Equ. (3.113) with both Euclidean and maximum norms) are also tested for comparison. Concerning the basic TE estimator with maximum norm and the Standard algorithm, they use the same distance for both

joint space and marginal spaces. The former one estimates the probability density for each data sample, whereas the Standard algorithm calculates TE directly (based on the kNN entropy estimator, Equ. (3.60)).

For all results, the statistical means and the standard deviations of the different estimators have been estimated on 100 trials. In Fig. 4.12, to demonstrate the influence of the number of neighbors, the value of k is set to 3 and 4, while in all other figures, we fix $k = 8$.

4.2.2.1. Model 4

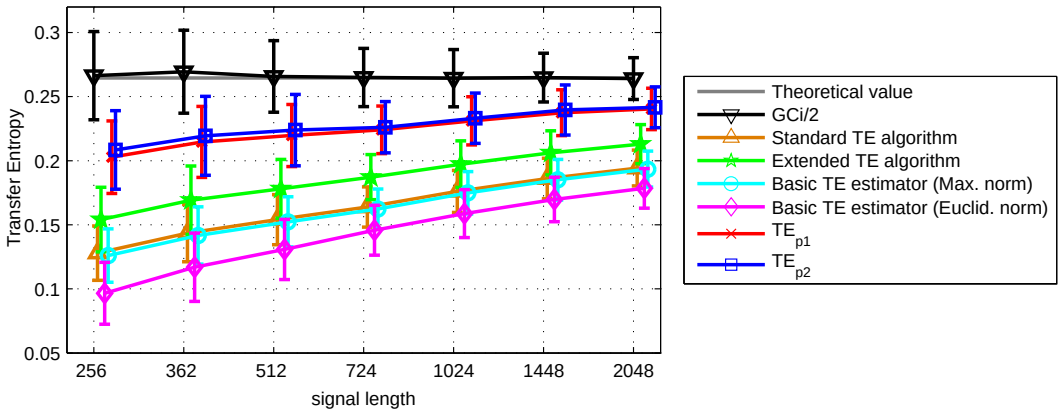
Results for Model 4 are reported in Fig. 4.9 where the dimensions d_Y and d_Z are identical. We observe that, for a low dimension and a sufficient number of neighbors (Fig. 4.9(a)), all TE estimators tend all the more to the theoretical value (around 0.26) that the length of the signals is large, the best estimation being obtained by the two new estimators. Compared to Granger causality, these estimators display a greater bias, but a slightly lower variance. Due to the "curse of dimensionality", with an increasing dimension (see Fig. 4.9(b)), it becomes much more difficult to obtain an accurate estimation of TE. For a high dimension, all estimators reveal a non-negligible bias, even if the two new estimators still behave better than the two reference ones (Standard and Extended algorithms). If we zoom in on Fig. 4.9(b), we observe that standard deviations are within the same order of magnitude and even smaller.

4.2.2.2. Model 5 and Model 6

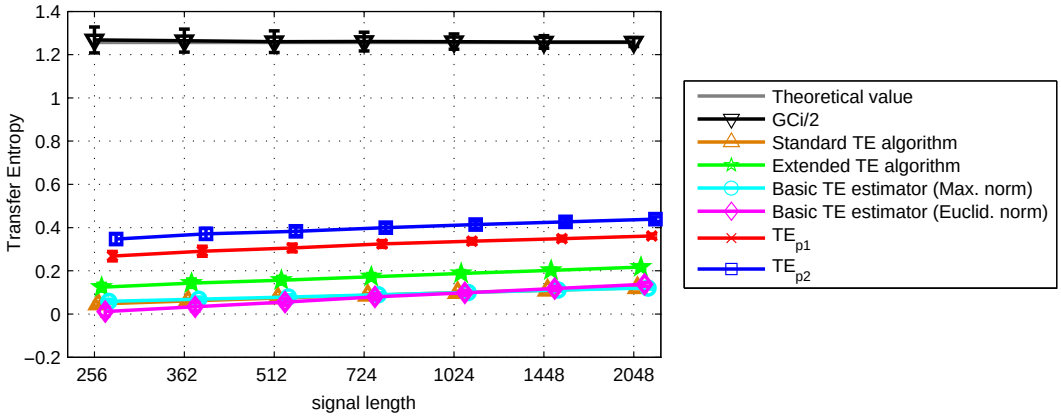
In this section, we firstly displayed the results of these two AR models with $k = 8$, and then, using Model 5 as an example, we discuss how the choice of k influences the performance of the proposed TE estimators.

As previously, for $k = 8$ (in Fig. 4.10 and 4.11), we observe that all the transfer entropy estimators converge towards the theoretical value. This result is all the more true when the signal length increases. As expected in such linear models, Granger causality outperforms the TE estimators at the expense of a slightly larger variance. Contrary to Granger causality, TE estimators are clearly more impacted by the signal length even if their standard deviations remain lower. Here again, when comparing the different TE estimators, it appears that the $\mathrm{TE}_{\mathrm{p1}}$ and $\mathrm{TE}_{\mathrm{p2}}$ estimators achieve improved behavior compared to the Standard and Extended algorithms for $k = 8$.

For Model 5 (Fig. 4.10), for both directions (from X to Y and from Y to X), the proposed TE estimators ($\mathrm{TE}_{\mathrm{p1}}$ and $\mathrm{TE}_{\mathrm{p2}}$) systematically outperform the other estimators. For Model 6, for all directions, this hierarchy is globally preserved. In both Fig. 4.10



(a)

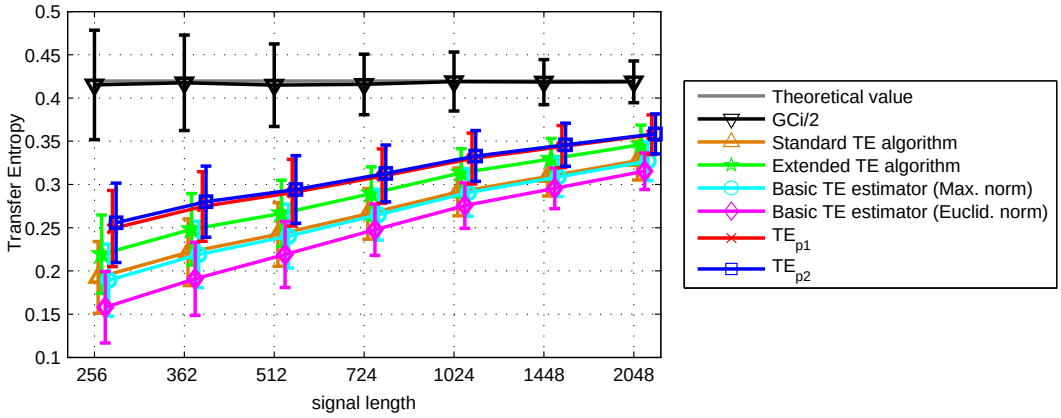


(b)

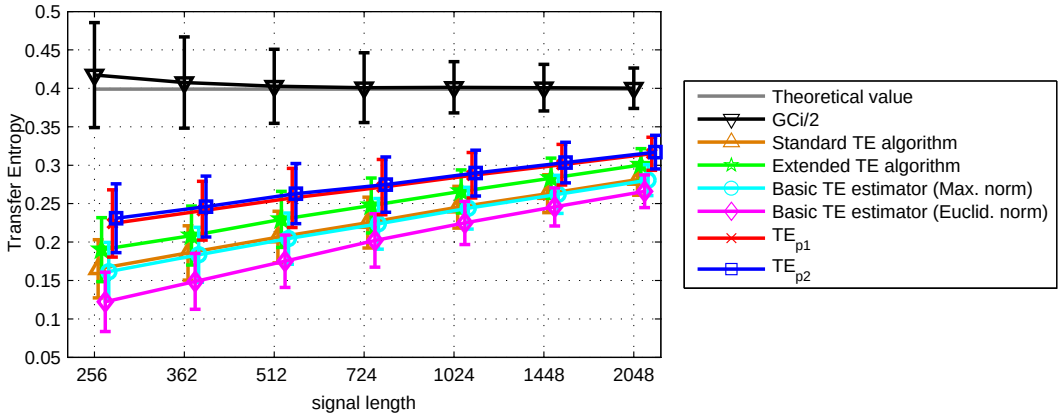
Figure 4.9: (Model 4) Information transfer (in nats) from Z to X estimated for two different dimensions with $k = 8$. The figure displays the mean values and the standard deviations, (a) $d_Y = d_Z = 3$, (b) $d_Y = d_Z = 8$.

and 4.11, the basic estimator with Euclidean norm (Box ⑮ in Fig. 3.6) always gives the poorest performance.

In the scope of kNN algorithms, the choice of k must be a tradeoff between the estimation of bias and variance. Globally, when the value of k decreases, the bias decreases for the Standard and Extended algorithms and for the new estimator $\mathrm{TE}_{\mathrm{p1}}$. Now, for the second proposed estimator $\mathrm{TE}_{\mathrm{p2}}$, it is much more sensitive to the number of neighbors (as can be seen when comparing Fig. 4.12(a) and 4.12(b)). As shown in Fig. 4.10 and 4.11, the results obtained using $\mathrm{TE}_{\mathrm{p2}}$ and $\mathrm{TE}_{\mathrm{p1}}$ are quite comparable when the value of k is large ($k = 8$). Now, when the number of neighbors decreases, the second estimator we proposed, $\mathrm{TE}_{\mathrm{p2}}$, is much less reliable than all the other ones (Fig. 4.12(b)). Concerning the variance, it remains relatively stable when the number of neighbors falls from 8 to 3. More details on this phenomenon are given in Appendix I.



(a)

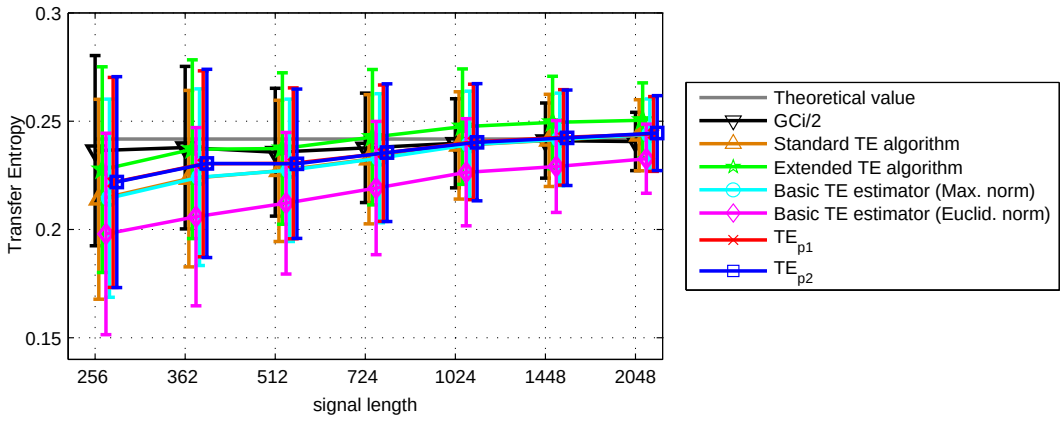


(b)

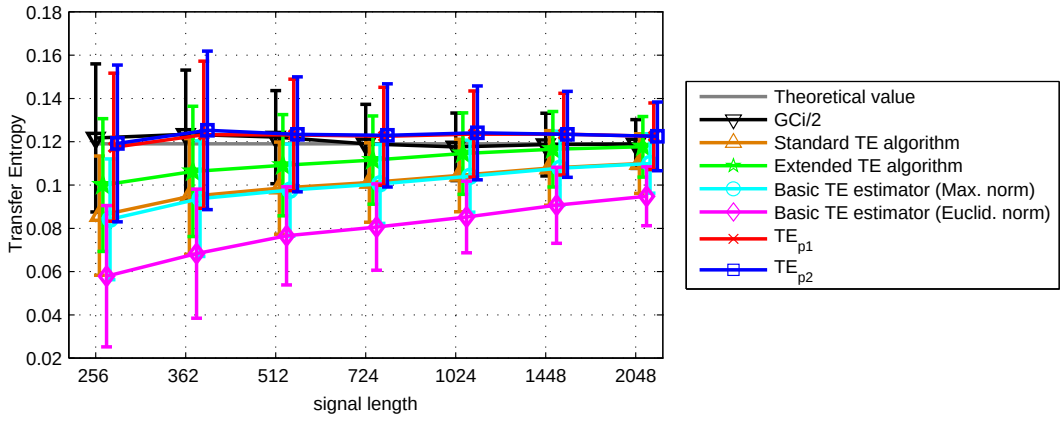
Figure 4.10: (Model 5) Information transfer (in nats), mean values and standard deviations, $k = 8$. (a) From X to Y. (b) from Y to X.

4.2.2.3. Model 7

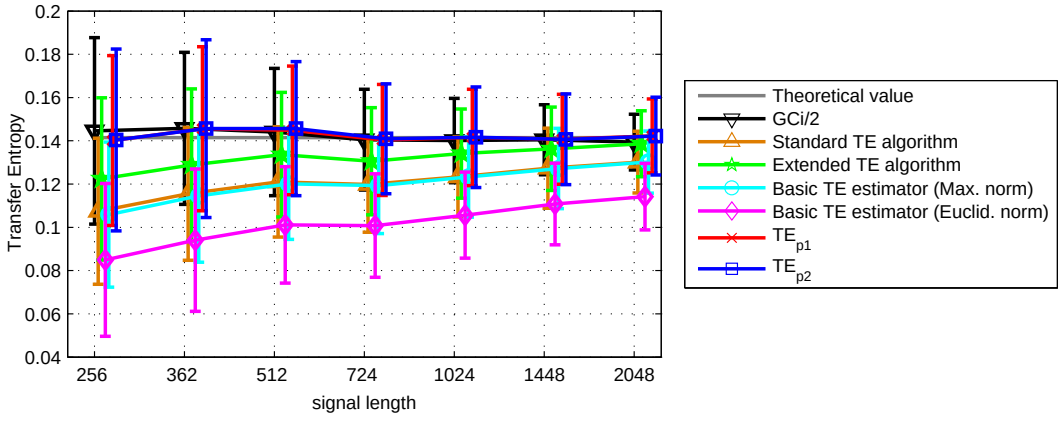
For this nonlinear model, we tuned the parameters to obtain a strong coupling between X and Z. In this situation, the Granger causality index failed in detecting the information flow and remained equal to zero for the different sets of tested parameters (see Fig. 4.13). We observed the same issue as that pointed in [Gao 2015], *i.e.* a very slow convergence of the kNN-based estimator when the number of observations increases, and noticed that all the 6 TE estimators revealed comparable performance (as displayed in Fig. 4.13). In this difficult case, the newly proposed methods do not outperform the existing ones. For this type of strong coupling perhaps further improvement could be obtained at the expense of an increasing computational complexity as that proposed in [Gao 2015].


Transfer Entropy
0.3
0.25
0.2
0.15
256
362
512
724
1024
1448
2048
signal length
Theoretical value
GCi/2
Standard TE algorithm
Extended TE algorithm
Basic TE estimator (Max. norm)
Basic TE estimator (Euclid. norm)
TE_p1
TE_p2

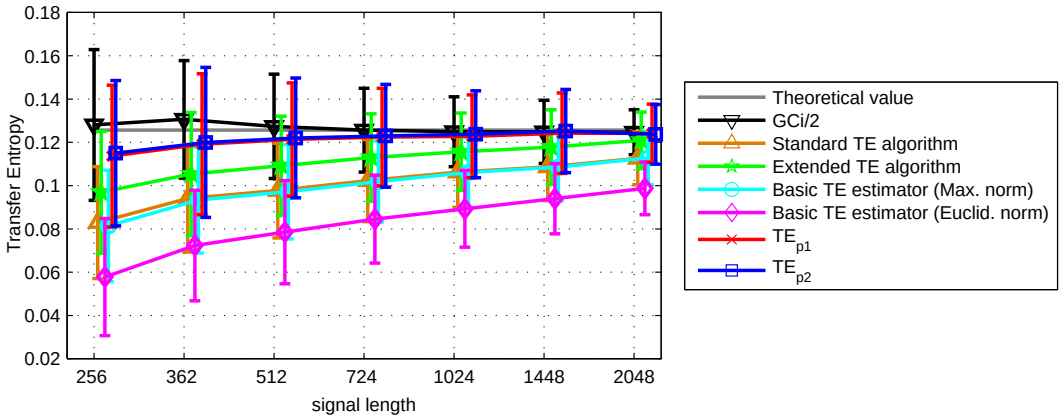
(a)


Transfer Entropy
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
256
362
512
724
1024
1448
2048
signal length
Theoretical value
GCi/2
Standard TE algorithm
Extended TE algorithm
Basic TE estimator (Max. norm)
Basic TE estimator (Euclid. norm)
TE_p1
TE_p2

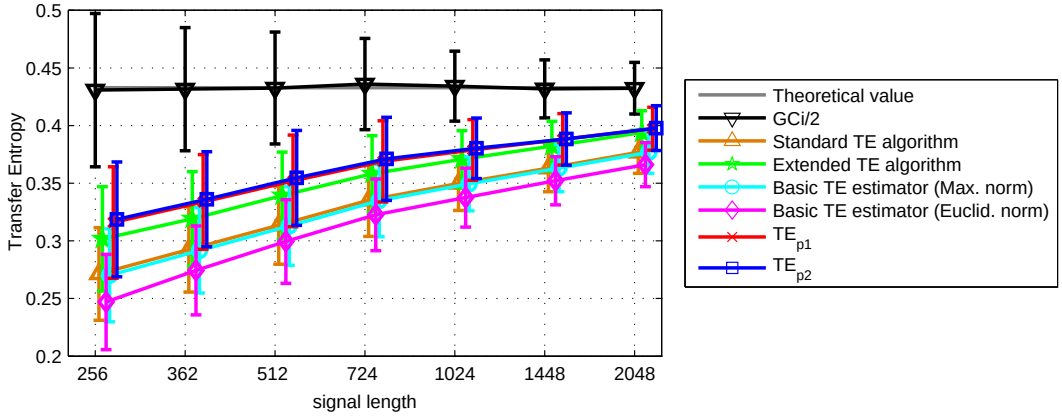
(b)


Transfer Entropy
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
256
362
512
724
1024
1448
2048
signal length
Theoretical value
GCi/2
Standard TE algorithm
Extended TE algorithm
Basic TE estimator (Max. norm)
Basic TE estimator (Euclid. norm)
TE_p1
TE_p2

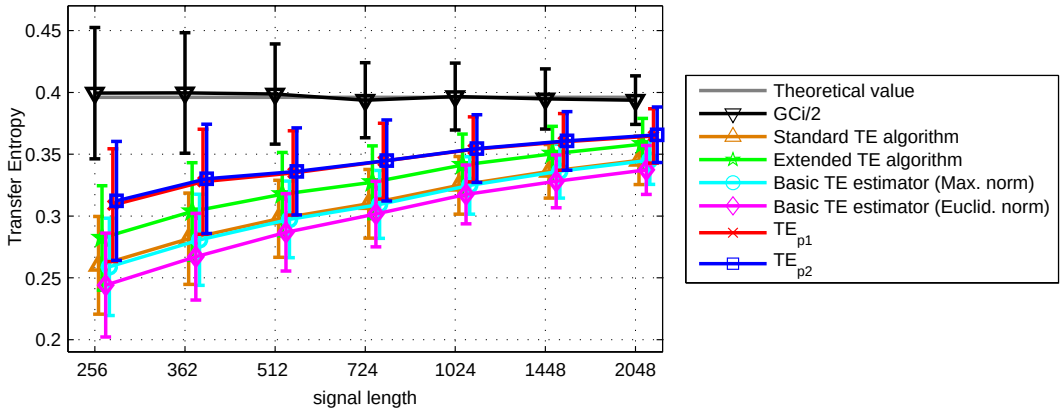
(c)



(d)

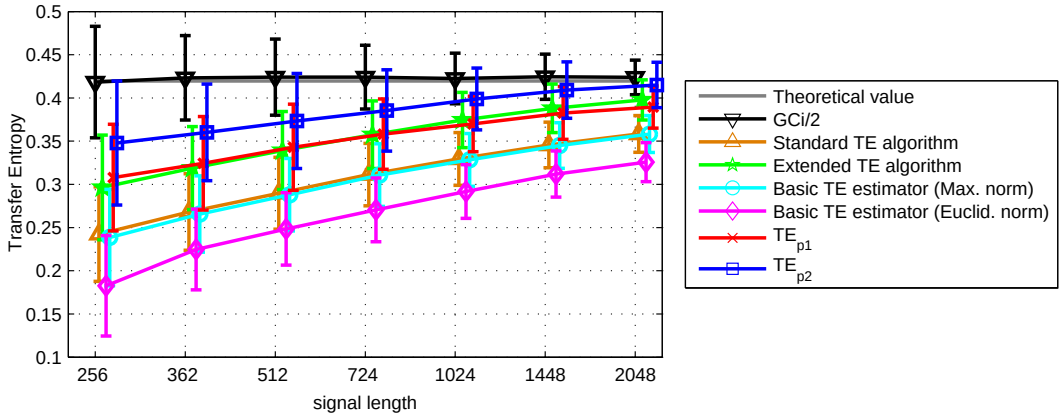


(e)

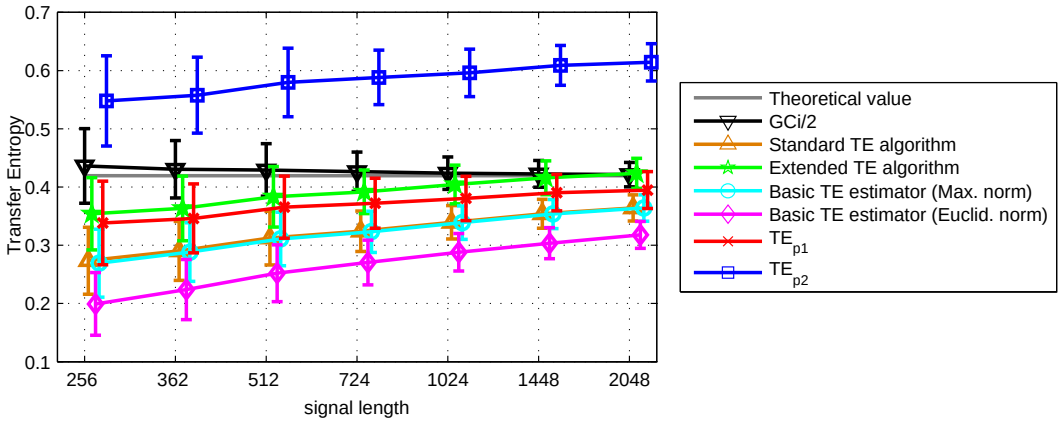


(f)

Figure 4.11: (Model 6) Information transfer (in nats), mean values and standard deviations, $k = 8$. (a) From X to Y. (b) from Y to X. (c) from X to Z. (d) from Z to X. (e) from Y to Z. (f) from Z to Y.



(a)



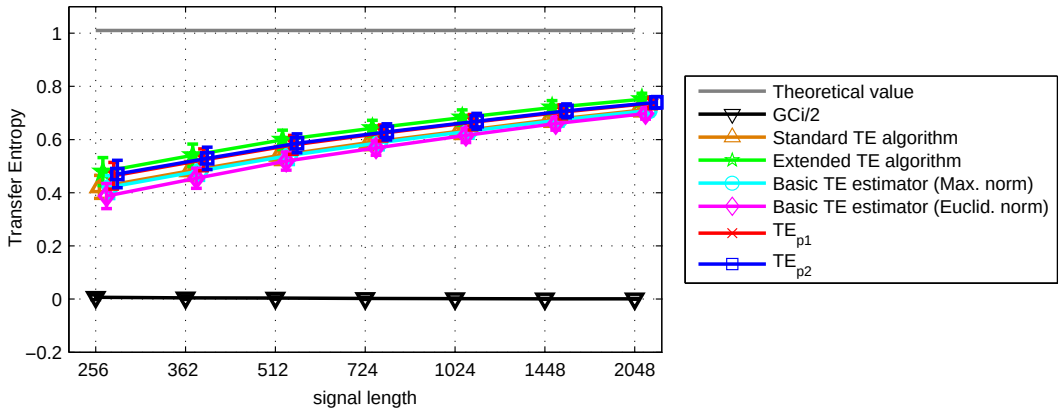
(b)

Figure 4.12: (Model 5 with different k) Information transfer (in nats) from X to Y, mean values and standard deviations. (a) $k = 4$. (b) $k = 3$.

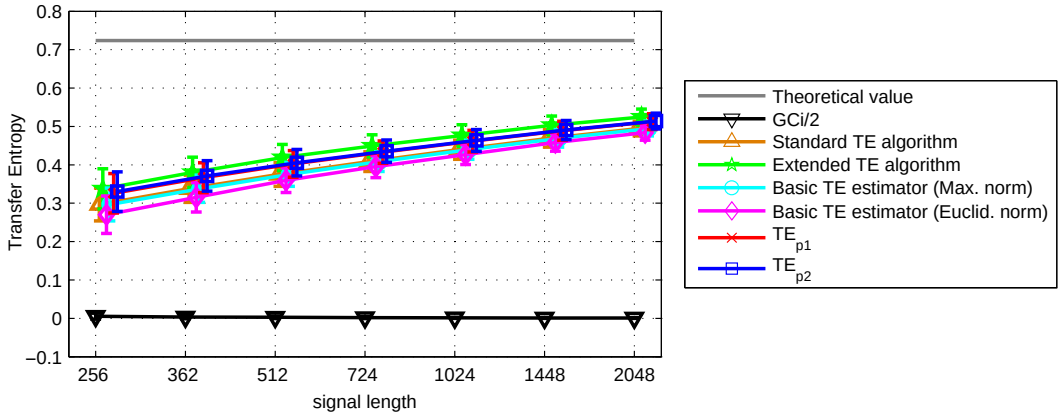
4.2.3. Results on the Physiology-based Model

For the physiology-based model, simulated signals with a Runge-Kutta numerical time step corresponding to a 512 Hz sampling frequency were generated. Surrogate data were obtained following the strategy described in section 4.1.2.2. For each set of surrogate data, both dependency and information flow between two populations were considered.

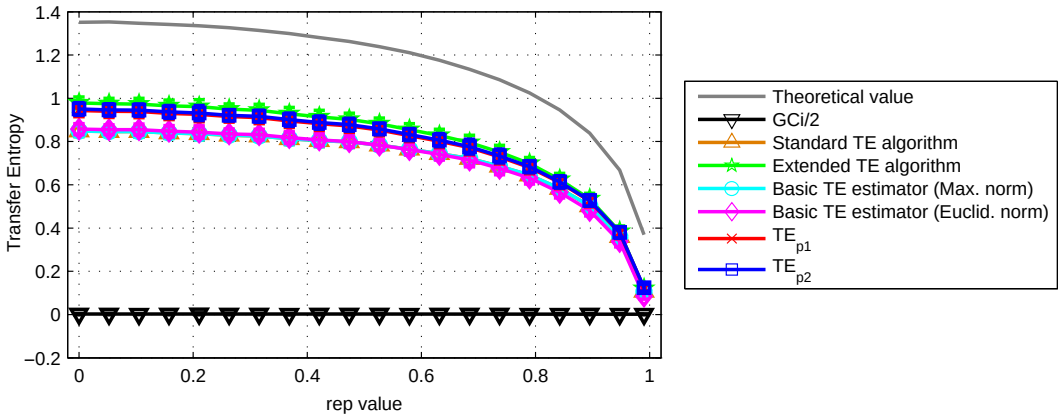
Firstly, basic MI estimators with both maximum and Euclidean norms (Box ⑭ in Fig. 3.6), together with the MILCA toolbox, were tested for independency (see Fig. 4.14 and Tab. 4.3). Secondly, Granger causality, Standard algorithm, Extended algorithm and the proposed algorithm $\mathrm{TE}_{\mathrm{p1}}$ were tested to detect information flow (see Fig. 4.15 and Tab. 4.4). All indexes in this section were calculated on 100 trials with 1024-point length signals ($N = 1024$). For MI and TE, and for each pair {estimator type, connectivity case} boxplots were obtained for both original data and surrogate data as displayed in Fig. 4.14 and Fig. 4.15. In cases 1 and 2 (introduced in section 4.1.2), Pop_X and Pop_Y are



(a)



(b)



(c)

Figure 4.13: (Model 7) Information transfer from Z to X, mean values and standard deviations, $k = 8$, $\theta = 1$, 100 trials. (a) $K = 1$, $rep = 0.8$. (b) $K = 0.6$, $rep = 0.8$. (c) $K = 1$, rep varies, signal length $N = 1024$.

dependent. Hence, with the t-test at the 99% confidence level, all the three MI algorithms give the right conclusion ($p < 0.01$). In case 3, corresponding to independence of Pop_X and Pop_Y, for any MI algorithm the same test quite rightly accepts the independence hypothesis (see Tab. 4.3). For the information flow measures with 99% confidence level, only the Granger causality index and the proposed algorithm are reliable and successfully distinguish the different situations ($p < 0.01$). (see Tab. 4.4).

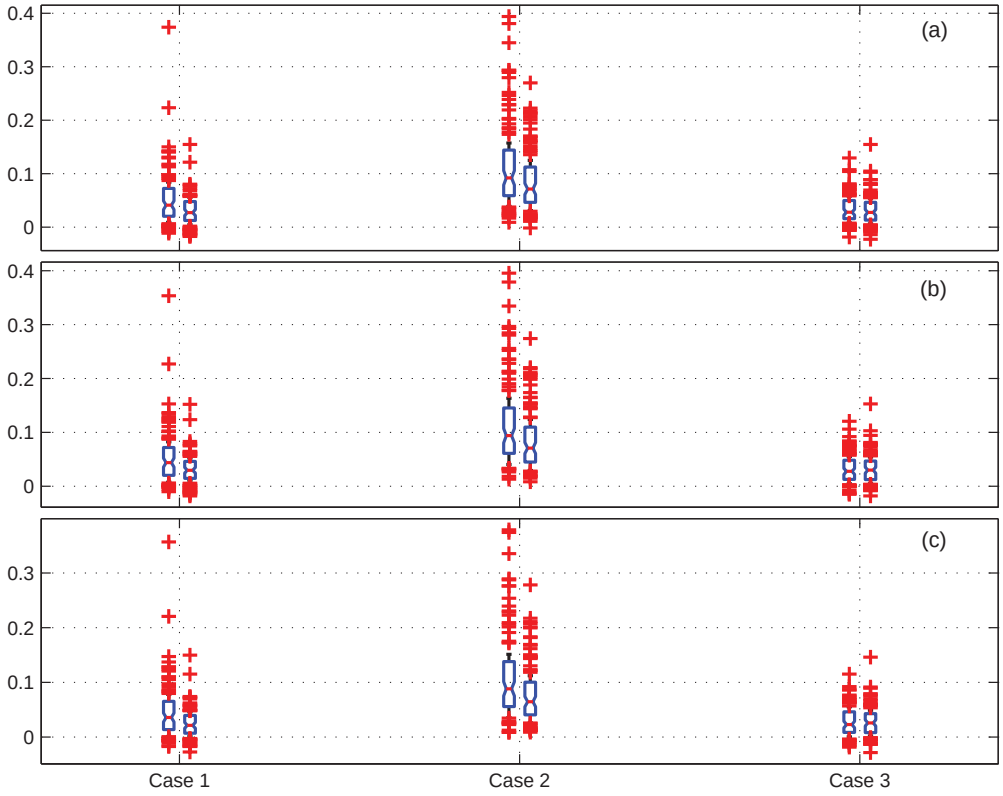


Figure 4.14: (Physiology-based model) Boxplots of mutual information estimated values, $k = 6$. For each case (1, 2 or 3) boxplots are displayed for original data (left-hand side) and surrogate data (*i.e.* under H_0 hypothesis, right-hand side). (a) MILCA toolbox. (b) Basic MI estimator with maximum norm. (c) Basic MI estimator with Euclidean norm.

p-value	MILCA toolbox	Basic MI estimator with Euclidean norm	Basic MI estimator with maximum norm
Case 1	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**
Case 2	**0.007**	**0.006**	**0.004**
Case 3	0.88	0.97	0.99

Table 4.3: (Physiology-based model) p-value of Student's t-test for each couple {MI estimator, case} when comparing MI estimated values obtained from original data and surrogate data.

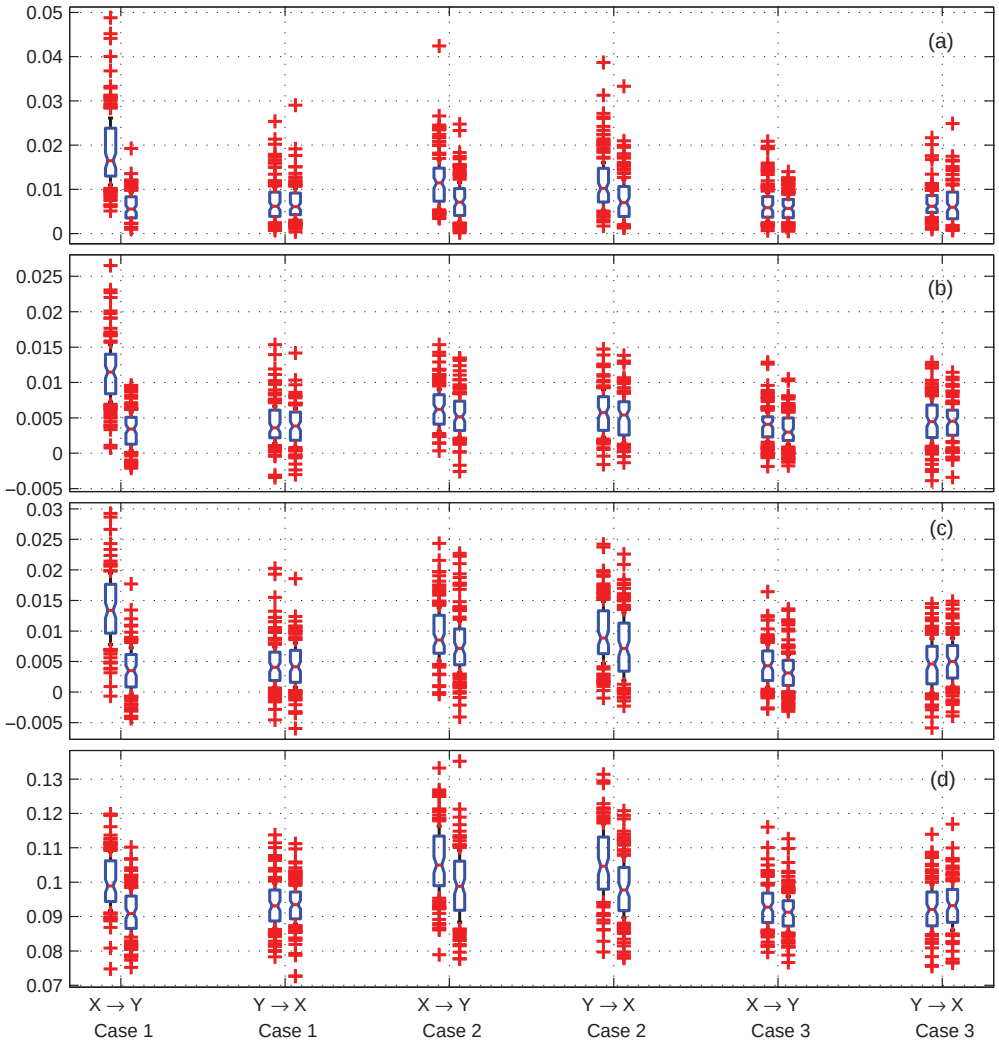


Figure 4.15: (Physiology-based model) Boxplots of information flow estimated values, $k = 30$. For each case (1, 2 or 3) boxplots are displayed for original data (left-hand side) and surrogate data (*i.e.* under H_0 hypothesis, right-hand side). (a) Granger causality. (b) Standard algorithm, $k = 30$. (c) Extended algorithm, $k = 30$. (d) Proposed algorithm ($\mathrm{TE}_{\mathrm{p1}}$), $k = 30$.

4.2.4. Computational Costs

Computation time is also an important issue. In the computation of kNN-based estimators, the most time-consuming part is the procedure of nearest neighbor searching. Compared to Standard and Extended algorithms, both $\mathrm{TE}_{\mathrm{p1}}$ and $\mathrm{TE}_{\mathrm{p2}}$ involve supplementary information, such as the maximum distance of the first kth nearest neighbor in each dimension and the number of points on the border. However, most currently used neighbor searching algorithms, such as kd tree (k-dimensional tree), ATRIA [Merkwirth 2000], provide not only information on the kth neighbor, but also on the first $(k-1)$ nearest neighbors. So, in terms of computation cost, there is no significant difference

p-value		Granger causality	Standard algorithm	Extended algorithm	Proposed algorithm
Case 1	$X \to Y$	$< \mathbf{0.001}$	$< \mathbf{0.001}$	$< \mathbf{0.001}$	$< \mathbf{0.001}$
	$Y \to X$	0.51	0.3	0.57	0.87
Case 2	$X \to Y$	$< \mathbf{0.001}$	$\mathbf{0.005}$	0.011	$< \mathbf{0.001}$
	$Y \to X$	$< \mathbf{0.001}$	0.16	$\mathbf{0.005}$	$< \mathbf{0.001}$
Case 3	$X \to Y$	0.1	0.083	0.051	0.11
	$Y \to X$	0.87	0.78	0.49	0.38

Table 4.4: (Physiology-based model) p-value of Student's t-test for each couple {information flow estimator, case} when comparing information flow estimated values obtained from original data and surrogate data.

	$k = 4$	$k = 16$	$k = 30$
Granger causality	47.31	-	-
Standard algorithm	2.39	3.03	3.88
Extended algorithm	29.14	33.69	35.89
Basic TE estimator (Max. norm)	42.03	42.74	44.38
Basic TE estimator (Euclid. norm)	41.49	42.75	44.98
TE_{p1}	50.87	50.51	53.28
TE_{p2}	62.05	62.97	64.60

Table 4.5: Computational cost of the different algorithms (seconds). This table refers to Model 5 ($X \to Y$) using $N = 1024$ and 50 trials. Our implementation of Granger causality involves matrix inversion and an implementation based on MATLAB system identification toolbox would reduce the computation time. Theoretically, except for the Standard algorithm, TE algorithms require comparable computation times, their implementation being based on "for" loop structures, explaining slow process compared to the Standard algorithm.

among these kNN TE estimators. A practical order of magnitude of the relative computation times is given in Tab. 4.5 which provides the computation times required by the different algorithms according to the number of neighbors (except for Granger causality).

4.3. Discussion and Conclusion

In this chapter, we tested the MI and TE estimators proposed in chapter 3 with different models, including linear and nonlinear structures and simulating dependent and non-independent observations.

For mutual information, the experimental results prove the effectiveness of the strategy proposed in chapter 3, and both MILCA and basic estimators (whatever the norm, Box ⑭ in Fig. 3.6) outperform the MI estimators using the same k in case of independent signals. However, with models simulating non-independent signals, there are no significant differences among the results. The mixed MI estimators provide good performance on dependent signals, but suffer from large bias in independence situations. It should be

noted that the mixed estimators are sensitive to the choice of the number of neighbors and require some empirical tuning for practical use. To merge the respective advantages of basic MI estimators and mixed estimators a two-step procedure could be proposed to measure MI: in a first step the independence hypothesis would be tested with a basic type statistic and in case of independence rejection MI would be measured with a mixed type statistic.

Concerning transfer entropy, for Gaussian distributions, experimental results show the effectiveness of the new estimators on IID data as well as on correlated AR signals in comparison with the standard KSG algorithm estimator. This conclusion still holds when comparing the new algorithms with the extended KSG estimator. Globally, all TE estimators satisfactorily converge to the theoretical TE value, *i.e.* to half the value of the Granger causality index, while the newly proposed TE estimators show lower bias for sufficiently large k (in comparison with the reference TE estimators) and comparable variance estimation errors. Now, one of the new TE estimators, $\mathrm{TE}_{\mathrm{p2}}$, suffers from noticeable error when the number of neighbors is small. Some experiments allowed us to verify that this issue already existed when estimating entropy of a random vector: when the number of neighbors k falls below the value of dimension d, the bias drastically increases. As expected, experiments with Model 4 showed that all the TE estimators under examination suffer from "curse of dimensionality", which makes it difficult to obtain accurate estimation of TE with high dimension data. When tested on a model introducing a strong nonlinearity in the statistical link between the two simulated signals, the new algorithms did not longer display better performance than classical algorithms. This point would require deeper investigations, perhaps in line with existing works as [Gao 2015]. When testing on the physiology-based model, the experiments did not show a clear advantage of TE approaches compared to the Granger causality. Now, as the nonlinearities included in this type of model are smooth nonlinearities, the performance of linear approach is not surprising.

Chapter 5

Analysis of Real Signals

In the previous chapter, the algorithms we developed have been tested on simulated signals. Now, we apply the proposed causality measures to real signals recorded in the cerebral cortex of an epileptic patient. To our knowledge entropic methods, and particularly transfer entropy, have not been applied yet to analyze effective connectivity in epileptic brain in the course of seizures (about 10 seconds before the onset of the seizure, during and after the seizure), as they have been more investigated in cognitive tasks. Beyond a comparison between the causality measures we developed in chapter 3, it seems also interesting to compare these techniques tested on real signals with some reference method. In most works in the literature dedicated to epileptic seizures analysis, the direct transfer function (DTF) [Mierlo 2011, Jung 2011, Mierlo 2013, Zhang 2015] remains a widely used measure. Hereafter, a N-variate statistic, called local causality index (LCI), is introduced. Hence, besides the comparison of Granger causality with different TE estimators discussed in the previous chapter, we also compare them with the directed transfer function through the LCI index to discuss the information they provide with a view to a better understanding of the patient seizure organization. To this end, we use boxplot-based visualization of LCI values computed either with the causality indexes detailed in chapter 3 or with DTF. For each tested causality index, to highlight the effect of each epileptic phase, *i.e.* before, during or after the epileptic seizure onset, a Student's t-test is carried out to compare statistically this phase with a reference phase corresponding to background activity. Based on the ground truth provided by the clinical experts, the performances of the different algorithms are discussed.

5.1. Database

First we present the database to be analyzed hereafter. This database is composed of 72-second length iEEG signals recorded on 20 channels in the cerebral cortex of an epileptic patient. These signals were recorded in brain structures whose activities are *a priori* interesting to investigate according to preliminary clinical and electrophysiological examinations. A schematic diagram of the placement of 12 iEEG electrodes on a lateral view (left hemisphere) is displayed in Fig. 5.1. In this figure the symbol A' means "electrode in anatomical region of type A in the left hemisphere". Around 10 to 15 sensors were placed along each electrode, for example A'1 to A'15. Each iEEG signal is bipolar, obtained by the difference of the potentials recorded on two adjacent sensors. All channels (signals) were sampled at 256 Hz. As shown in Fig. 5.2, a seizure onset up to 32 seconds was recorded in this database. This recording can be divided into three parts: pre-ictal phase (0s∼20s), ictal phase (20s∼52s) and post-ictal phase (52s∼72s) (some examples of PSD of these signals are displayed in Appendix H, Fig. H.3). Additionally, to get a reference in our experiments, another segment of iEEG signals has been recorded on the same 20 channels, far apart from the seizure onset. This additional recording is called hereafter "baseline". According to the clinical experts, the 20 channels can be categorized into three groups, as listed in Tab. 5.1, named respectively Onset group (group O), Propagation group (group P) and Not-involved group (group N). The signals associated with the Onset group are supposed to be linked to activities in brain regions from which the seizure starts. The Propagation group contains electrodes corresponding to brain structures acting as relays which are stimulated with some delay after the beginning of the seizure and which possibly stimulate other structures later. The Not-involved group corresponds to structures which are neutral with respect to epileptic processes.

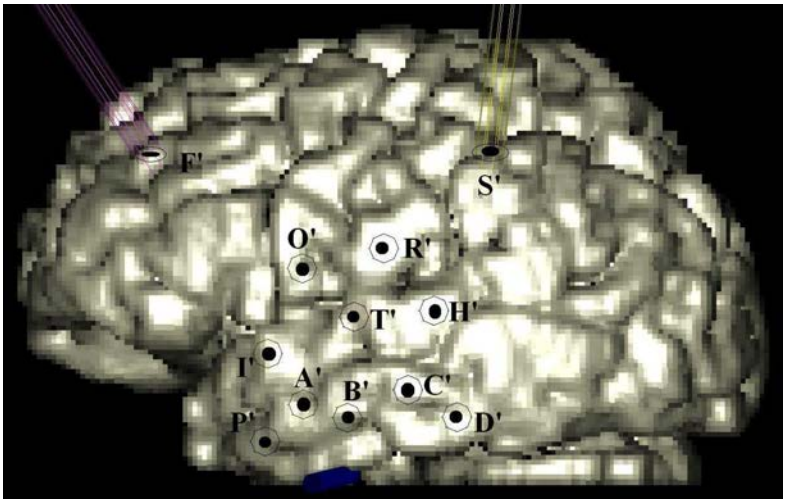


Figure 5.1: Schematic diagram of iEEG electrodes placement (left hemisphere).

Group	Channels
Onset group (group O)	Cp1, Cp4, Pp1, Pp4, Ap2, Ap6, Bp1
Propagation group (group P)	Cp9, Pp8, Dp1, Dp5, Tp1, Fp2
Not-involved group (group N)	Ap11, Bp6, Bp11, Tp8, Hp2, Ip2, Fp8

Table 5.1: Categories of the channels in the database. Each channel corresponds to a bipolar signal. For instance, Api represents the difference between the potentials recorded respectively from the ith sensor of electrode A' and its $(i+1)$th sensor.

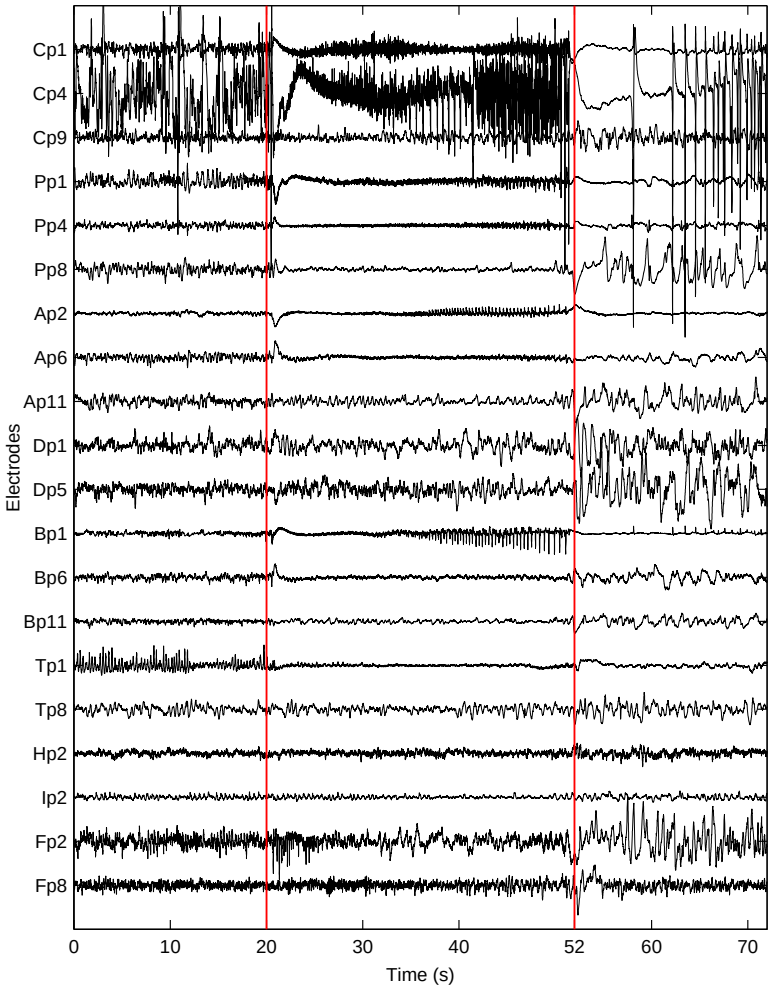


Figure 5.2: A 72-second length iEEG recording with a seizure onset up to 32 seconds. The two red vertical lines separate this recording into three segments: pre-ictal phase (0s∼20s), ictal phase (20s∼52s) and post-ictal phase (52s∼72s).

5.2. Method

The idea is to determine from the 20 recorded signals and a current observation time interval if a given channel can be considered as belonging to either a brain region implied in the seizure initialization mechanisms, or a secondary region relaying the seizure propagation, or a region not involved in the epileptic activities. To this end, we must introduce

some statistics measuring the global transfer information from the tested signal to the others.

5.2.1. Local Connectivity Index

In this work, to highlight the sensitivity of the causality statistics in the different epileptic phases and in the expert electrode labeling, we propose an analysis using a boxplot-based visualization. The target is to visualize the effect of the seizure phase factor (baseline/pre-ictal/ictal/post-ictal) and that of the type of involvement of the brain structures during the seizure onset (group O/group P/group N), which corresponds to a second factor in the proposed statistics.

To proceed with this analysis, we define a generic time dependent local connectivity (scalar) index, LCI, calculated separately for each recorded channel and parameterized by one of the causality indexes (Granger causality, transfer entropy,...) or by the DTF index detailed in Appendix J:

$$LCI_n(t) = \sum_{m \in \text{out}(n)} CI_{m/n}(t) - \sum_{m \in \text{in}(n)} CI_{n/m}(t) \tag{5.1}$$

where LCI_n is the local connectivity parameter corresponding to channel $n \in \{1, \ldots, 20\}$ when the channels listed in Fig. 5.2 numbered from 1 to 20, t is the index of the current analysis time window, $CI_{m/n}$ is either one of the causality indexes parameters defined in chapter 3 for the pair (m, n) in the direction $n \to m$ or the DTF index for the same direction, $\text{out}(n)$ is the channels indices subset (among the channels considered excepted n, *i.e.* 19 channels) obtained by selecting the numbers m such that $CI_{m/n}$ is significantly different from zero, $\text{in}(n)$ is the channels indices subset (among the 20 channels excepted n) obtained by selecting the numbers m such that $CI_{n/m}$ is also significantly different from zero.

The addressed question should have been *a priori* to classify the values $LCI_n(t)$ without supervision or in one of the 3 classes *a priori* defined by the experts (group O, group P, group N). Now, building a classifier, with or without supervision, requires a sufficiently large number of examples to set up a learning group and, for the supervised case, a test group. Here, this number was too limited to expect to succeed by this way and so we chose the following procedure.

Let us denote $BXP(CI, s, T_e)$ a boxplot computed on the set of scalar values $\{LCI_n(t), n \in s, t \in T_e\}$ where $s \in \{\text{group O}, \text{group P}, \text{group N}\}$ and e symbolizes one epoch in a set defined by the experts to capture the dynamic behavior of the epileptic propagation. Thus, besides the baseline, pre-ictal and post-ictal phases (named epochs afterwards), the ictal one (20s∼52s) was divided into three overlapped epochs (named ic-

tal 1, ictal 2 and ictal 3 epochs (see Tab. 5.2)), T_e including all the time indices necessary to cover the complete duration of e. The first argument of BXP, CI, indicates the type of causality index (Granger index, one of the TE indexes or DTF index). Concerning the analysis of DTF results, the corresponding boxplots are denoted $BXPDTF\,(s, T_e)$.

5.2.2. Experimental Protocol

Each boxplot is built on an interval of 16 seconds (indexed by t) which is divided into eight 2-second length windows without overlapping. The relations between signals vary over time. Tab. 5.2 displays the starting and ending time for each epoch. This partition was adopted in accordance with the clinician.

Epoch	Starting and ending points (in time)
Pre-ictal	2s∼18s
Ictal 1	22s∼38s
Ictal 2	28s∼44s
Ictal 3	34s∼50s
Post-ictal	54s∼70s

Table 5.2: Starting and ending points of each epoch.

Now, we consider the question of the significance of a causality index computed from one channel to another one. For Granger causality, as explained in Appendix K, it is possible to derive theoretically a threshold value leading approximately to a given probability of false positive (a causal link is decided whereas this link does not exist). For TE and DTF, we choose to use an adaptive threshold to retain the same number of significant links as for Granger causality for each t value. Fig. 5.3 gives an example on this point. This method is clearly not completely satisfying but it avoids making surrogate replications, which is time consuming for TE estimators.

After building the boxplots, a Student's t-test is used to determine if the result obtained on each interval (*i.e.* pre-ictal, ictal 1, ictal 2, ictal 3, post-ictal) is significantly different from that on the baseline or not.

To compute the sizes of the past values vectors in Granger and TE causality index, and also to choose the order of the multivariate AR model in DTF we must estimate the Markovian memory length of the bivariate observed signal (suppose that the model is approximately a Markov process). For the selection of this model order, it is not obvious to find an optimal order using the widely used AIC and BIC criteria, as displayed in Fig. 5.4. Theoretically the estimated order corresponds to the value on the x axis for which the criterion value is minimum but here this x value can be too large to be used in the estimation algorithms. However we observe that the curves slopes are very small when the

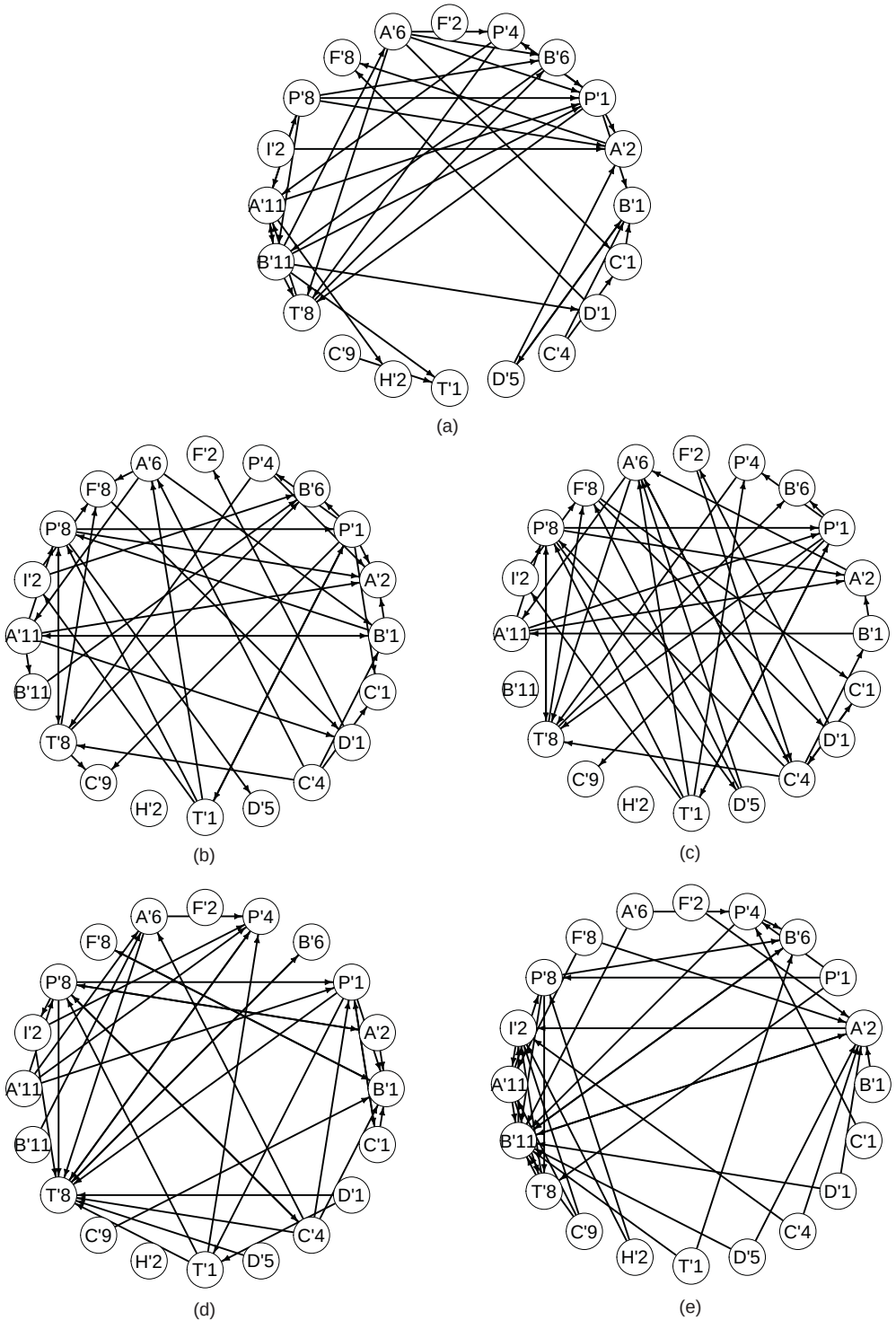


Figure 5.3: An example of the determined significant links obtained by the different indexes for one 2-second length sliding window (2s∼4s). (a) For Granger causality, using the methods introduced in Appendix K, 45 links are considered as significant, so that we kept the same number of links for the other 4 measures, (b) Standard algorithm, (c) Extended algorithm, (d) proposed algorithm, (e) DTF algorithm.

x value exceeds 6. Finally, taking account of orders estimated from many cases (different epochs, different seizures) the model order was uniformly set to 6. Additionally, to avoid the "curse of dimensionality", we chose to keep the temporal memory depth (six times the time sampling period corresponding to 256 Hz) but to subsample with a subsampling ratio equal to 0.5. Thus, given two signals X and Y, to predict x_t, only $[x_{t-2}, x_{t-4}, x_{t-6}]$ and $[y_{t-2}, y_{t-4}, y_{t-6}]$ were used.

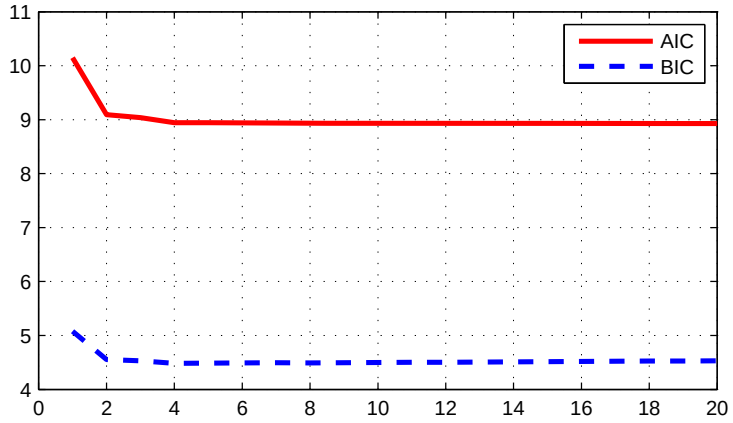


Figure 5.4: AIC and BIC indexes values versus order (computed on channel Cp1 and on the 72-second length signal).

Concerning LCI index, besides Granger causality and DTF, three different TE estimators are tested hereafter: the Standard algorithm (Box ⑧ in Fig. 3.6), the Extended algorithm (Box ⑨ in Fig. 3.6) and the first proposed algorithm (named TE_{p1}, Box ㉓ in Fig. 3.6).

5.3. Experimental Results

This section is devoted to the collection and discussion on the results. The index defined by Equ. (5.1) is expected to be large when the sum of the significant influences of channel n onto the other channels is much greater than the sum of the significant influences from the other channels on it. Consequently, this index reveals whether the channel n drives the others in some manner. For large negative values of this index, the channel n is considered as being more influenced by the other nodes than influencing the others. Note that the magnitude of the values taken by this index is not important if we essentially want to observe its variations with respect to the seizure phase and the tested channel n in Equ. (5.1).

Accordingly, what we expect with the LCI and DTF boxplot values is that: (i) for group O, there should be significant increase during the ictal part, (ii) for group N, there should be no significant mean deviation compared with the baseline. With the definition in Equ. (5.1), for group P, it could be expected to observe significant mean

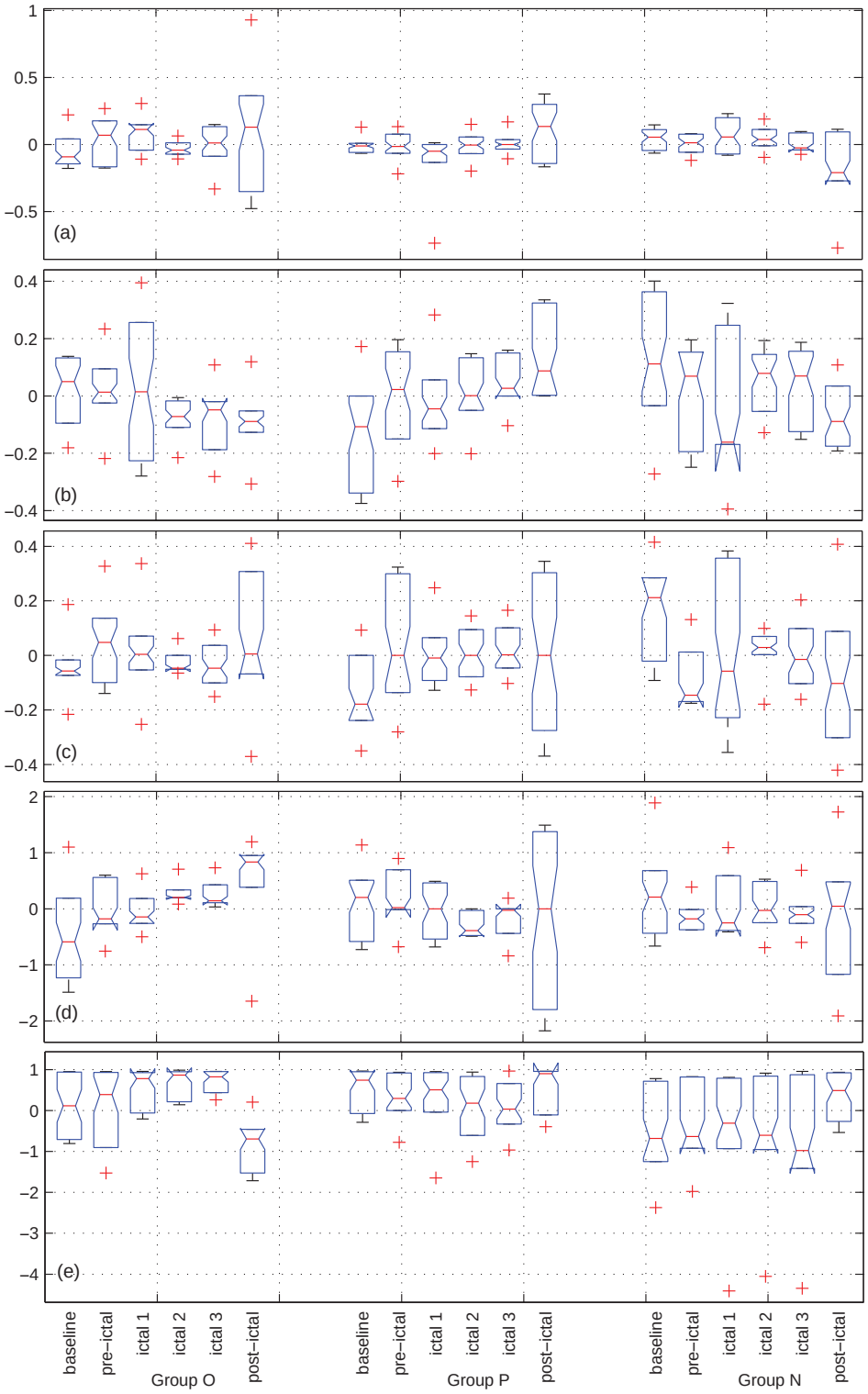


Figure 5.5: Boxplot of different indexes, (a) Granger causality, (b) Standard algorithm, (c) Extended algorithm, (d) proposed algorithm, (e) DTF algorithm.

deviations only in the second part of the seizure, since the corresponding neural groups are stimulated after the beginning of the seizure by group O and possibly stimulate other structures later. Additionally, according to the clinical experts, the number of functional connectivity links in the regions involved in the epileptic seizure drastically grows during the post-ictal period.

		Granger causality	Standard algorithm	Extended algorithm	Proposed algorithm	DTF algorithm
Group O	Pre-ictal	$<$ **0.001**	0.495	0.057	**0.007**	0.144
	Ictal 1	$<$ **0.001**	0.41	0.16	**0.033**	$<$ **0.001**
	Ictal 2	0.888	**0.001**	0.542	$<$ **0.001**	$<$ **0.001**
	Ictal 3	0.491	$<$ **0.001**	0.965	$<$ **0.001**	$<$ **0.001**
	Post-ictal	0.109	$<$ **0.001**	**0.036**	$<$ **0.001**	$<$ **0.001**
Group P	Pre-ictal	0.603	**0.016**	**0.002**	0.373	**0.013**
	Ictal 1	**0.001**	$<$ **0.001**	$<$ **0.001**	0.273	**0.012**
	Ictal 2	0.745	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**
	Ictal 3	0.333	$<$ **0.001**	$<$ **0.001**	**0.001**	$<$ **0.001**
	Post-ictal	**0.012**	$<$ **0.001**	$<$ **0.001**	**0.05**	0.967
Group N	Pre-ictal	**0.005**	**0.001**	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**
	Ictal 1	0.318	**0.02**	**0.004**	0.181	0.574
	Ictal 2	0.74	0.125	$<$ **0.001**	**0.02**	0.633
	Ictal 3	**0.015**	**0.036**	$<$ **0.001**	**0.008**	0.118
	Post-ictal	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**	**0.006**	$<$ **0.001**

Table 5.3: p-value of Student's t-test for different indexes.

Tab. 5.3 reports the p-values obtained with the Student's t-test. For each entry of the table, the t-test is applied to compare the values in $BXP\left(CI, s, T_{e\neq\text{baseline}}\right)$ and the values in $BXP\left(CI, s, T_{e=\text{baseline}}\right)$ or the values in $BXPDTF\left(s, T_{e\neq\text{baseline}}\right)$ and the values in $BXPDTF\left(s, T_{e=\text{baseline}}\right)$. A small p-value indicates that the means of the two samples are significantly different. If, for instance, we choose a probability of false positive equal to 0.05, then we accept the hypothesis that the theoretical means are different if the p-value is smaller than 0.05.

When we analyze Fig. 5.5 and Tab. 5.3, we come to the following conclusions: (i) the Standard algorithm fails in reflecting any variation during the first ictal epoch (ictal 1) for group O, and, on the contrary, there is a significant LCI values decrease for ictal 2 and ictal 3 epochs. Besides, it shows important changes for group N (significant decrease for pre-ictal, ictal 1, ictal 3 and post-ictal epochs); (ii) the Extended algorithm is globally not better than the Standard one for group O (no significant changes during the ictal epochs), and it leads to poor performance on group N (unexpected significant decrease on all epochs); (iii) among the three TE algorithms, the proposed algorithm demonstrates the best performance. Firstly, there is a significant increase for group O on the three ictal epochs. Secondly, for group N, the decrease for all epochs is not globally

more marked than with the two other TE algorithms; (iv) among all indexes, DTF gives interesting results: significant increase during the three ictal epochs for group O, and less significant changes for group N; (v) concerning the Granger causality index, despite a good performance for group O in the beginning of the ictal phase (ictal 1), results are globally not satisfying, as is the case for group N. For the analysis of the group P, it appears difficult to derive general conclusions, partly due to the fuzzy nature of this class of structure, which prevents us from detecting the most relevant methods.

5.4. Discussion and Conclusion

In this chapter, we applied both Granger causality and different transfer entropy estimators discussed previously to analyze human epileptic signals. Since DTF remains popular in the analysis of epileptic signals, it has been also tested. Based on the ground truth provided by the clinical experts, (i) among all the three tested TE estimators, the proposed algorithm outperforms the two others, (ii) DTF appears as the most reliable measure and gives interesting performance. Even if lots of time and effort have been spent on the analysis of real signals, we must acknowledge that the relevance of information-theoretic quantities is not so evident.

To detect significant causality index values, we introduced a decision threshold which was derived theoretically for the Granger causality index (see Appendix K). Now, to test transfer entropy, defining such a threshold is not so trivial. In this experiment, for a more effective comparison among different algorithms, we used an adaptive threshold to make sure that all algorithms retain the same number of links. Another critical issue is the "curse of dimensionality", since large dimensions always bring troubles to TE estimations, including both computation time and estimation accuracy. Here, we tried to avoid this problem by time subsampling the predictor vectors.

In our experiments, all indexes have been calculated on the whole frequency band. In recent literature [Ponten 2007, Mierlo 2011, Varotto 2012, Mierlo 2013], the analysis of epileptic data is performed on different frequency bands, and this could be a perspective to this work.

Conclusion

Our work concerns the detection of effective connectivity to be applied in the context of epilepsy without addressing the question of direct or indirect links.

In chapter 1, we presented a brief introduction to the research background, including the disease itself, the human brain structure and different types of brain connectivity.

As mentioned in our survey of the state of the art in chapter 2, various algorithms have been proposed to quantify the strength of influence between different subsystems. Among these methods, Granger causality and transfer entropy are two well-known approaches that are linked under Gaussian assumption. For Granger causality, a linear model is considered to fit the data and its calculation is relatively simple. However, for information-theoretic approaches, such as transfer entropy, their computation becomes relatively difficult, especially in high-dimensional spaces. This is the main issue preventing us from applying these methods in their conditional form taking into account the environmental network.

In chapter 3, we deeply discussed the estimation of information-theoretic quantities based on k-Nearest Neighbors (kNN) techniques. The estimation of mutual information (MI) and transfer entropy (TE) is always an important issue, especially in neuroscience, where getting large amounts of stationary data is problematic. Both MI and TE can be calculated as a summation of different individual entropy estimations. Until now, the most widely used MI/TE estimator follows the kNN strategy proposed in [Kraskov 2004]: compute the distance in the highest-dimensional space, and use the same distance in other marginal spaces. In [Kraskov 2004], the effectiveness of this strategy was proved by numerical simulations. Using multi-dimensional Taylor expansion, we introduced a novel analytical form of the individual entropy estimation bias depending on the norm. In the case of the maximum norm, we obtained the same conclusion as in [Kraskov 2004], *i.e.* using the same distance for different spaces can be the optimal choice. We got the proof that, with this kind of strategy, the bias could vanish if and only if the independence

assumption was satisfied. For mutual information, to further reduce the bias in the case of dependent signals, we proposed mixed estimators, for which the individual entropy was calculated with a linear combination of two different individual estimations.

Another important improvement is the use of a (hyper-)rectangle instead of a (hyper-)cube while calculating TE with maximum norm. This idea was firstly proposed in [Kraskov 2004], where the (hyper-)rectangle was used only in the joint space and not in the marginal spaces. In chapter 3, we extended this idea and proposed to use (hyper-)rectangles in both joint space and marginal spaces, where the maximum distances in different directions could be different. To this end, we first investigated the estimation of Shannon entropy based on the kNN technique including a rectangular neighboring region and introduced two different kNN entropy estimators. We derived mathematically these new entropy estimators by extending the results and methodology developed in [Kraskov 2004] and [Singh 2003]. Given the new entropy estimators, two novel TE estimators have been proposed, implying no extra computation cost compared to existing similar kNN algorithms.

In chapter 4, the new MI/TE estimators were tested with various kinds of models. To validate the performance of the proposed estimators, we compared them with the MI/TE estimators available in existing toolboxes: (i) Kraskov-Stögbauer-Grassberger (KSG) MI estimator available in MILCA toolbox, (ii) Standard TE estimator and Extended TE estimator available in TRENTOOL and JIDT toolboxes respectively. For mutual information, as expected, in independence situations, the MI estimators following the proposed strategy performed very well (mutual information very close to zero). For time correlated observations sequences, the new mixed estimators (with both Euclidean and maximum norms) gave the best results. However, these mixed MI estimators were very sensitive to the selection of the number of neighbors, and so more difficult to use. For transfer entropy, under Gaussian assumption, experimental results proved the effectiveness of the new estimators for IID data in comparison with the standard TE estimator. This conclusion still held when comparing the new algorithms with the extended TE estimator. Globally, all TE estimators satisfactorily converged to the theoretical TE value, *i.e.* to half value of Granger causality, while the newly proposed TE estimators showed lower bias for a number of neighbors sufficiently large (in comparison with the reference TE estimators) with comparable variance estimation errors. We noticed that the two proposed TE algorithms produced quite comparable results when the number of neighbors was sufficiently large. However, one of the new TE ($\mathrm{TE}_{\mathrm{p2}}$) estimators suffered from noticeable error when the number of neighbors was small. Finally, experimental results on simulated iEEG signals showed that (i) all tested MI algorithms gave the expected results, (ii) for the detection of information flow, only Granger causality and the first proposed TE algorithm ($\mathrm{TE}_{\mathrm{p1}}$) successfully distinguished the three following situations:

independency, unidirectional and bidirectional propagation flows, at 99% confident level.

In chapter 5, we applied different causal approaches, including Granger causality, transfer entropy and directed transfer function, to analyze real signals, for which the ground truth was given by clinical experts. We first proposed a boxplot-based visualization to compare the different algorithms. A test derived from the Granger-Wald test was introduced for Granger causality to determine if the relation between two channels was significant or not. For TE and DTF, the adaptive threshold was used to retain the same number of links as Granger causality. According to the results, the proposed algorithm TE_{p1} outperformed all other tested TE estimators, Granger causality and DTF being slightly better.

Concerning the computation time of the different TE algorithms, we mainly have to consider the procedure of NN searching, which is the most time-consuming part in the computation of kNN-based estimators. The newly proposed TE estimators (TE_{p1} and TE_{p2}) involve supplementary information, (i) the maximum distance of the first k NNs in each dimension (also used in the Extended TE algorithm), and (ii) the number of points on the border. The most widely used neighbor searching algorithms, such as kd tree, provide not only information on the kth neighbor but also on the first k NNs. So, these informations required here can be retained without additional cost. Therefore, it can be considered that there is no significant increase in computation cost for the newly proposed kNN TE estimators.

This work is a first step in a more general context of connectivity investigation for neurophysiological activities obtained either from nonlinear physiology-based modeling or from real human epileptic recordings.

Several possible directions can be considered in a future work. As a matter of fact, in our study, transfer entropy was seen as a pairwise approach. In our experiments, when we estimated the causality between two channels, the influence of the other channels has not been taken into account contrary to either DTF approach or other causality methods based on N-channel VAR models. To remedy this shortcoming, the conditional transfer entropy can be further investigated. However, this conditional measure also brings new challenges, *e.g.* (i) to reduce the complexity, we would have to decide which part of the environment influences causal relations between two specific channels, (ii) to escape the "curse of dimensionality" due to the involvement of more channels, we could consider causality graph following the approach proposed in [Runge 2012].

Moreover, all the causal indexes (including GC, TE and DTF) were calculated on full frequency band in this work. However, in recent literature [Ponten 2007, Mierlo 2011, Varotto 2012, Mierlo 2013], the epileptic signals are analyzed on different frequency

bands. So, in a further work, it would be interesting to discover the corresponding causal relations on different frequency bands.

To summarize, on the one hand we obtained interesting improvements of entropic bivariate measures of causality based on kNN approach that have been validated on simulated signals. On the other hand, in order to retrieve conclusions given by clinical experts, we applied these measures to human intracerebral epileptic signals by integrating them in a proposed local causality index derived from the bivariate causality algorithms, to be compared with the popular DTF causality index. Our preliminary conclusion offered a mixed picture. The strong non-stationarity of epileptic signals is *a priori* an additional challenge using entropic methods which require relatively long observation intervals and which have been developed in the literature more particularly for the analysis of cognitive networks. Now, the performance of the proposed entropic measures is not so poor compared to that of the DTF reference measure. Finally, the expected advantage of entropic methods being some enhancement in nonlinear connectivity, the existence of this nonlinear connectivity should be questioned. Even if the epileptic signals considered individually could be produced by strongly nonlinear mechanisms, it is not obvious that the connectivity mechanisms are strongly nonlinear and necessarily require nonlinear methods to be correctly detected. Supplementary investigation on sub-band signals should be considered to discuss this idea.

Appendix

A. Derivation of Equ. (3.89)

$$
\begin{aligned}
&\frac{1}{2v(x)} \cdot \int_{\mathcal{L}(x)} (y-x)(y-x)^T \mathrm{d}y \\
&= \frac{1}{2(d_X+2)\pi} \Gamma^{\frac{2}{d_X}} \left(\frac{d_X+2}{2}\right) v^{\frac{2}{d_X}}(x) |A|^{\frac{1}{d_X}} A^{-1} \\
&= \frac{1}{2(d_X+2)\pi} \Gamma^{\frac{2}{d_X}} \left(\frac{d_X+2}{2}\right) |A|^{\frac{1}{d_X}} A^{-1} \cdot \frac{\pi \mathcal{R}^2(x)}{|A|^{\frac{1}{d_X}} \Gamma^{\frac{2}{d_X}} \left(\frac{d_X+2}{2}\right)} \\
&= \frac{\mathcal{R}^2(x)}{2(d_X+2)} \cdot A^{-1}
\end{aligned}
\tag{A.1}
$$

B. Proof of Property 1

Let us introduce the (hyper-)rectangle $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ centered on x_1 for which the random sizes along the d directions are defined by $(\epsilon'_1,\ldots,\epsilon'_d) = (\epsilon_1,\ldots,\epsilon_d) \times \left(\frac{v_r}{\epsilon_1\times\ldots\times\epsilon_d}\right)^{\frac{1}{d}}$ so that $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ and $\mathcal{D}_{x_1}^{\epsilon_1,\ldots,\epsilon_d}$ are homothetic and $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ has a (hyper-)volume constrained to the value v_r. We have

$$\int\limits_{x\in\mathcal{D}_{x_1}^{\varepsilon_1,\ldots,\varepsilon_d}} \mathrm{d}\mu^d(x) > v_r \Leftrightarrow \mathcal{D}_{x_1}^{\varepsilon'_1,\ldots,\varepsilon'_d} \subset \mathcal{D}_{x_1}^{\varepsilon_1,\ldots,\varepsilon_d} \Leftrightarrow \mathrm{card}\left\{x_j : x_j \in \mathcal{D}_{x_1}^{\varepsilon'_1,\ldots,\varepsilon'_d}\right\} \leq k-\xi_1, \tag{B.1}$$

where the first equivalence (the inclusion is a strict inclusion) is clearly implied by the construction of $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ and the second equivalence expresses the fact that the (hyper-)volume of $\mathcal{D}_{x_1}^{\epsilon_1,\ldots,\epsilon_d}$ is larger than v_r if and only if the normalized domain $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ does not contain more than $(k-\xi_1)$ points x_j (as ξ_1 of them are on the border of $\mathcal{D}_{x_1}^{\epsilon_1,\ldots,\epsilon_d}$ which is necessarily not included in $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$). These equivalences imply the equalities between conditional probability values

$$\begin{aligned}
&\mathcal{P}\left(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1\right)\\
&= \mathcal{P}\left(\log\left(\frac{NV_1}{\widetilde{K}_1}\right) > r | X_1 = x_1, \Xi_1 = \xi_1\right)\\
&= \mathcal{P}\left(V_1 > v_r | X_1 = x_1, \Xi_1 = \xi_1\right)\\
&= \mathcal{P}\left(\mathrm{card}\left\{X_j : X_j \in \mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}\right\} \leq k-\xi_1\right).
\end{aligned} \tag{B.2}$$

Only $(N-1-\xi_1)$ events $\left\{X_j : X_j \in \mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}\right\}$ are to be considered because the variable X_1 and the ξ_1 variable(s) on the border of $\mathcal{D}_{x_1}^{\epsilon_1,\ldots,\epsilon_d}$ must be discarded. Moreover, these events are independent. Hence the probability value in Equ. (B.2) can be developed as follows

$$\begin{aligned}
\mathcal{P}\left(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1\right) \simeq \sum_{i=0}^{k-\xi_1} \binom{N-\xi_1-1}{i} \left(\mathcal{P}\left(X \in \mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}\right)\right)^i\\
\times \left(1-\mathcal{P}\left(X \in \mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}\right)\right)^{N-\xi_1-1-i}.
\end{aligned} \tag{B.3}$$

If $p_X(x_1)$ is approximately constant on $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$, we have $\mathcal{P}\left(X \in \mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}\right) \simeq p_X(x_1)v_r$ (note that the randomness of $(\epsilon'_1,\ldots,\epsilon'_d)$ does not influence this approximation

as the (hyper-)volume of $\mathcal{D}_{x_1}^{\epsilon'_1,\dots,\epsilon'_d}$ is imposed to be equal to v_r). Finally, we can write

$$\mathcal{P}\left(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1\right) \simeq \sum_{i=0}^{k-\xi_1} \binom{N-\xi_1-1}{i} (p_X(x_1)v_r)^i (1-p_X(x_1)v_r)^{N-\xi_1-1-i}. \tag{B.4}$$

C. Derivation of Equ. (3.169)

With $\mathcal{P}(T_1 \leq r|X_1 = x_1, \Xi_1 = \xi_1) = 1 - \mathcal{P}(T_1 > r|X_1 = x_1, \Xi_1 = \xi_1)$, we take the derivative of $\mathcal{P}(T_1 \leq r|X_1 = x_1, \Xi_1 = \xi_1)$ to get the conditional density function of T_1

$$
\begin{aligned}
&\mathcal{P}'(T_1 \leq r|X_1 = x_1, \Xi_1 = \xi_1) \\
&= -\mathcal{P}'(T_1 > r|X_1 = x_1, \Xi_1 = \xi_1) \\
&= -\left(\sum_{i=0}^{k-\xi_1} \frac{(\tilde{k}_1 p_X(x_1)e^r)^i}{i!} e^{-\tilde{k}_1 p_X(x_1)e^r}\right)' \\
&= -\sum_{i=0}^{k-\xi_1} \left(\left(\frac{(\tilde{k}_1 p_X(x_1)e^r)^i}{i!}\right)' e^{-\tilde{k}_1 p_X(x_1)e^r} + \frac{(\tilde{k}_1 p_X(x_1)e^r)^i}{i!}\left(e^{-\tilde{k}_1 p_X(x_1)e^r}\right)'\right) \\
&= -\sum_{i=0}^{k-\xi_1} \left(\frac{i(\tilde{k}_1 p_X(x_1)e^r)^{i-1}(\tilde{k}_1 p_X(x_1)e^r)}{i!} e^{-\tilde{k}_1 p_X(x_1)e^r}\right. \\
&\qquad\qquad \left. + \frac{(\tilde{k}_1 p_X(x_1)e^r)^i}{i!} e^{-\tilde{k}_1 p_X(x_1)e^r}(-\tilde{k}_1 p_X(x_1)e^r)\right) \\
&= -\sum_{i=0}^{k-\xi_1} e^{-\tilde{k}_1 p_X(x_1)e^r} \left(\frac{(\tilde{k}_1 p_X(x_1)e^r)^i}{(i-1)!} - \frac{(\tilde{k}_1 p_X(x_1)e^r)^{i+1}}{i!}\right).
\end{aligned}
\tag{C.1}
$$

Defining

$$
a(i) = \frac{(\tilde{k}_1 p_X(x_1)e^r)^i}{(i-1)!} \quad \text{and} \quad a(0) = 0, \tag{C.2}
$$

we have

$$
\begin{aligned}
\mathcal{P}'(T_1 \leq r) &= -\sum_{i=0}^{k-\xi_1} e^{-\tilde{k}_1 p_X(x_1)e^r}\,(a(i) - a(i+1)) \\
&= -e^{-\tilde{k}_1 p_X(x_1)e^r}\,(a(0) - a(k-\xi_1+1)) \\
&= e^{-\tilde{k}_1 p_X(x_1)e^r} a(k-\xi_1+1) \\
&= \frac{\left(\tilde{k}_1 p_X(x_1)e^r\right)^{(k-\xi_1+1)}}{(k-\xi_1)!} e^{-\tilde{k}_1 p_X(x_1)e^r}.
\end{aligned}
\tag{C.3}
$$

D. Derivation of Equ. (3.170)

$$
\begin{aligned}
&\lim_{N\to\infty} \mathrm{E}\left[T_1 | X_1 = x_1, \Xi_1 = \xi_1\right] \\
&= \int_{-\infty}^{\infty} r \frac{(\tilde{k}_1 p_X(x_1) e^r)^{(k-\xi_1+1)}}{(k-\xi_1)!} e^{-\tilde{k}_1 p_X(x_1) e^r} \mathrm{d}r \\
&= \int_0^{\infty} \left(\log(z) - \log\left(\tilde{k}_1\right) - \log p_X(x_1)\right) \frac{z^{k-\xi_1}}{(k-\xi_1)!} e^{-z} \mathrm{d}z \\
&= \frac{1}{\Gamma(k-\xi_1+1)} \int_0^{\infty} \left(\log(z) z^{k-\xi_1} e^{-z}\right) \mathrm{d}z - \log\left(\tilde{k}_1\right) - \log p_X(x_1) \\
&= \frac{1}{\Gamma(k-\xi_1+1)} \int_0^{\infty} \left(\log(z) z^{(k-\xi_1+1)-1} e^{-z}\right) \mathrm{d}z - \log\left(\tilde{k}_1\right) - \log p_X(x_1) \\
&= \frac{\Gamma'(k-\xi_1+1)}{\Gamma(k-\xi_1+1)} - \log\left(\tilde{k}_1\right) - \log p_X(x_1) \\
&= \psi(k-\xi_1+1) - \log\left(\tilde{k}_1\right) - \log p_X(x_1).
\end{aligned}
\tag{D.1}
$$

E. AIC and BIC Algorithms

Akaike's information criterion (AIC), proposed by Akaike [Akaike 1973] in 1973, is a classical rule of model selection. For a collection of competing models introduced to interpret observed data, AIC considers the trade-off between the goodness of model fitting and the complexity of the model. It quantifies the quality of each model, relatively to the other models in the collection. Bayesian information criterion (BIC) is another widely used model selection criterion proposed by Schwarz in 1978 [Schwarz 1978]. These two approaches often appear in literature as competing methods [Burnham 2002].

Now, we give brief mathematical descriptions of these two model selection measures. Consider a d-dimensional vectorial autoregressive process $X_t = [x_{1t}, x_{2t}, \ldots, x_{dt}]^T$ with model order p

$$X_t = c + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \ldots + \Phi_p X_{t-p} + \varepsilon_t, \tag{E.1}$$

where Φ_i, $i = 1, \ldots, p$ is $d \times d$ coefficient matrices, ε_t is an $d \times 1$ unobservable zero mean white noise process with time invariant covariance matrix Σ, and c is a constant. Here, for the sake of convenience, we assume $c = 0$.

Given an observed N-length data sequence $(X_1, X_2, \ldots, X_N)$, using the least-squares method [Brockwell 2013], we can fit a q-order VAR model, leading to an estimation of this model,

$$X_t = \widehat{\Phi_1} X_{t-1} + \widehat{\Phi_2} X_{t-2} + \ldots + \widehat{\Phi_q} X_{t-q} + \widehat{\varepsilon_t}. \tag{E.2}$$

Following the model (Equ. (E.1)), we write the predicted value $\widehat{X_t}$ as

$$\widehat{X_t} = \widehat{\Phi_1} X_{t-1} + \widehat{\Phi_2} X_{t-2} + \ldots + \widehat{\Phi_q} X_{t-q}. \tag{E.3}$$

Then, the experimental residual covariance matrix of the input noise of the VAR model can be calculated as

$$\widehat{\Sigma} = \frac{1}{N-q} \sum_{t=q+1}^{N} \left(X_t - \widehat{X_t}\right)\left(X_t - \widehat{X_t}\right)^T, \tag{E.4}$$

where T stands for the transpose operator.

Note that the above procedure can be carried out with any selected order q. For a given q, the AIC value is defined as

$$\mathrm{AIC}(q) = N \log\left(\det\left(\widehat{\Sigma}\right)\right) + 2d^2 q, \tag{E.5}$$

where $\log(\cdot)$ is the natural logarithm, and $\det(\cdot)$ stands for the matrix determinant. Similarly, the BIC value can be calculated as

$$\mathrm{BIC}(q) = N \log\left(\det\left(\widehat{\Sigma}\right)\right) + d^2 q \log(N). \tag{E.6}$$

For practical use, we compute the AIC (or BIC) values for a given limited set of q values, and the selected value of q is the one leads to the minimum AIC (or BIC) value.

F. Development of the Theoretical MI Value for Model 2

Consider the linear system described by Equ. (4.2), the parameter θ has no influence on the marginal covariance matrices $\mathcal{C}_X$ and $\mathcal{C}_{Y1}$, where $\mathcal{C}_X = \mathcal{C}_{Y1} = \mathcal{C}$. The joint covariance matrix $\mathcal{C}_{X,Y_1}$ can be calculated as

$$\mathcal{C}_{X,Y_1} = \begin{bmatrix} \mathcal{C} & \cos\theta \cdot \mathcal{C} \\ \cos\theta \cdot \mathcal{C} & \mathcal{C} \end{bmatrix}. \tag{F.1}$$

According to

$$\det\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right) = \det(A) \cdot \det\left(D - C\,A^{-1}B\right), \tag{F.2}$$

we have

$$\begin{aligned} \det\left(\mathcal{C}_{X,Y_1}\right) &= \begin{vmatrix} \mathcal{C} & \cos\theta \cdot \mathcal{C} \\ \cos\theta \cdot \mathcal{C} & \mathcal{C} \end{vmatrix} \\ &= \det(\mathcal{C}) \cdot \det\left(\mathcal{C} - \cos^2\theta \cdot \mathcal{C}\right) \\ &= \det^2(\mathcal{C}) \cdot \sin^{2d}\theta. \end{aligned} \tag{F.3}$$

Considering the theoretical entropy value of Gaussian data

$$\mathcal{H}(U) = \frac{1}{2}\log\left(\det\left(2\pi e \cdot \mathcal{C}_U\right)\right), \tag{F.4}$$

we have

$$\begin{aligned} \mathcal{I}(X, Y_1) &= \mathcal{H}(X) + \mathcal{H}(Y_1) - \mathcal{H}(X, Y_1) \\ &= \frac{1}{2}\log\left(\frac{\det\left(2\pi e \cdot \mathcal{C}_X\right) \cdot \det\left(2\pi e \cdot \mathcal{C}_{Y_1}\right)}{\det\left(2\pi e \cdot \mathcal{C}_{X,Y_1}\right)}\right) \\ &= \frac{1}{2}\log\left(\frac{(2\pi e)^{d_X + d_{Y_1}} \cdot \det\left(\mathcal{C}_X\right) \cdot \det\left(\mathcal{C}_{Y_1}\right)}{(2\pi e)^{d_X + d_{Y_1}} \cdot \det\left(\mathcal{C}_{X,Y_1}\right)}\right) \\ &= \frac{1}{2}\log\left(\frac{\det^2(\mathcal{C})}{\det^2(\mathcal{C}) \cdot \sin^{2d}\theta}\right) \\ &= -d\log(\sin\theta). \end{aligned} \tag{F.5}$$

G. Development of the Theoretical MI Value for Model 3

For the system described in Equ. (4.3), we have $\mathcal{C}_X = \mathcal{C}_e = I$, $\mathcal{C}_Y = \left(1+\beta^2\right) I$, and

$$\mathcal{C}_{X,Y} = \begin{bmatrix} I & I \\ I & \left(1+\beta^2\right) I \end{bmatrix}, \tag{G.1}$$

where $\mathcal{C}_{(\cdot)}$ stands for the covariance matrix.

According to

$$\det\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right) = \det\left(A\right) \cdot \det\left(D - C\,A^{-1}B\right), \tag{G.2}$$

we obtain

$$\begin{aligned} \det\left(\mathcal{C}_{X,Y}\right) &= \begin{vmatrix} I & I \\ I & \left(1+\beta^2\right) I \end{vmatrix} \\ &= \det(I) \cdot \det\left(\left(1+\beta^2\right) I - I\right) \\ &= \det\left(\beta^2 I\right). \end{aligned} \tag{G.3}$$

Considering the theoretical entropy value of Gaussian data

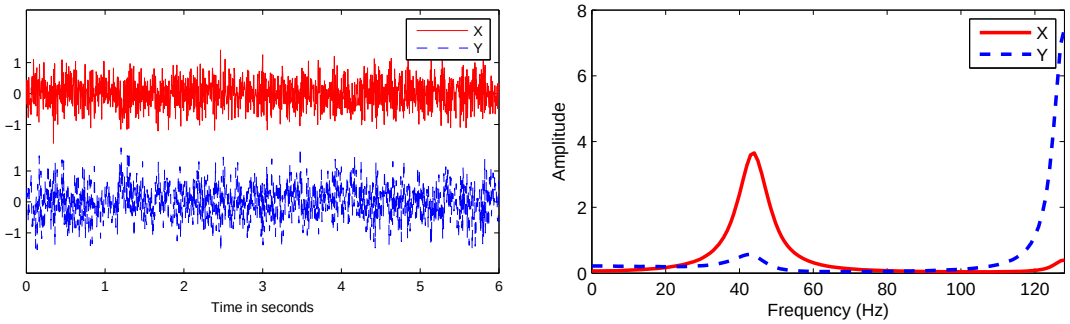
$$\mathcal{H}(U) = \frac{1}{2}\log\left(\det\left(2\pi e \cdot \mathcal{C}_U\right)\right), \tag{G.4}$$

finally, we get

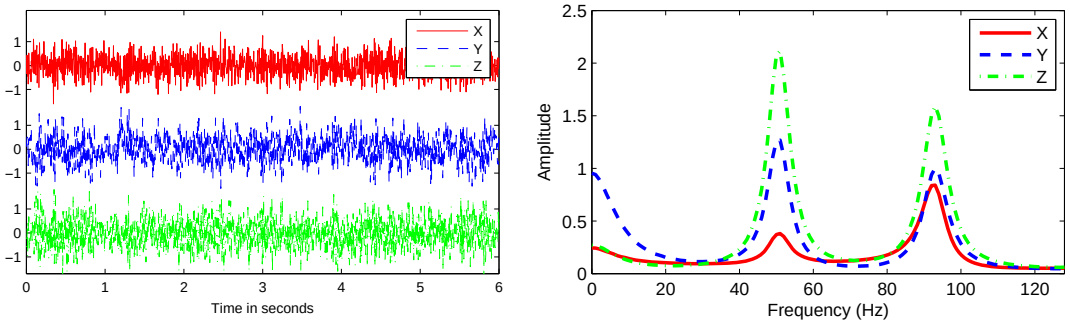
$$\begin{aligned} \mathcal{I}(X,Y) &= \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y) \\ &= \frac{1}{2}\log\left(\frac{\det\left(2\pi e \cdot \mathcal{C}_X\right) \cdot \det\left(2\pi e \cdot \mathcal{C}_Y\right)}{\det\left(2\pi e \cdot \mathcal{C}_{X,Y}\right)}\right) \\ &= \frac{1}{2}\log\left(\frac{(2\pi e)^{2d} \cdot \det\left(\mathcal{C}_X\right) \cdot \det\left(\mathcal{C}_Y\right)}{(2\pi e)^{2d} \cdot \det\left(\mathcal{C}_{X,Y}\right)}\right) \\ &= \frac{1}{2}\log\left(\frac{\det\left(I\right) \cdot \det\left(\left(1+\beta^2\right) I\right)}{\det\left(\beta^2 I\right)}\right) \\ &= \frac{d}{2}\log\left(\frac{1+\beta^2}{\beta^2}\right). \end{aligned} \tag{G.5}$$

H. Power Spectral Densities of the Signals

In this appendix, we display the power spectral densities of signals generated by Model 5, Model 6, the physiology-based model and real signals.

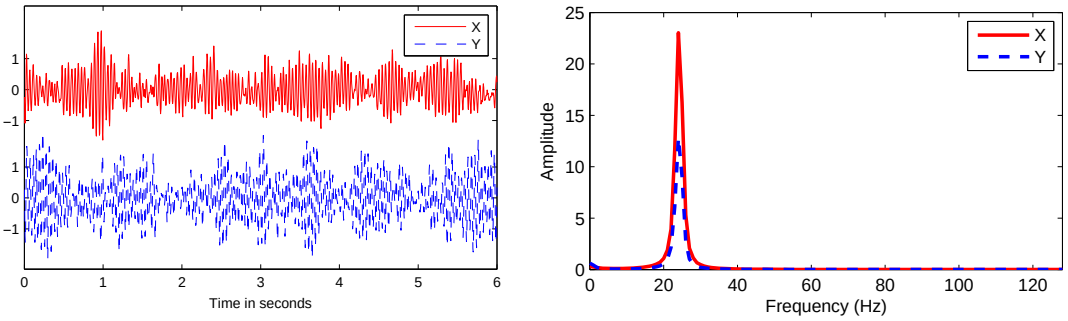


(a)

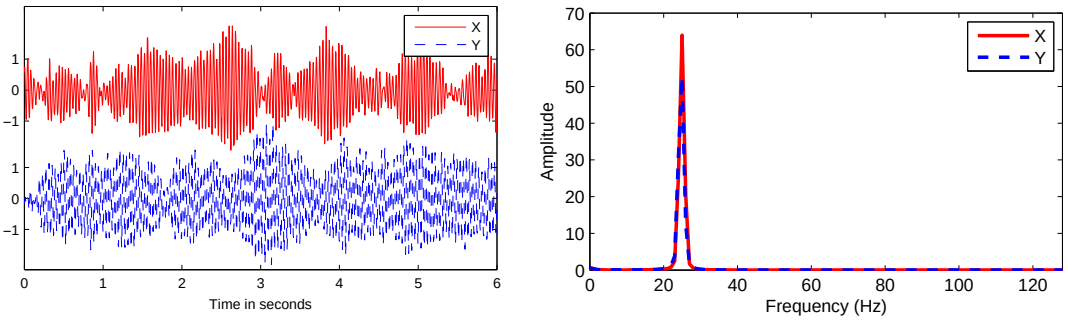


(b)

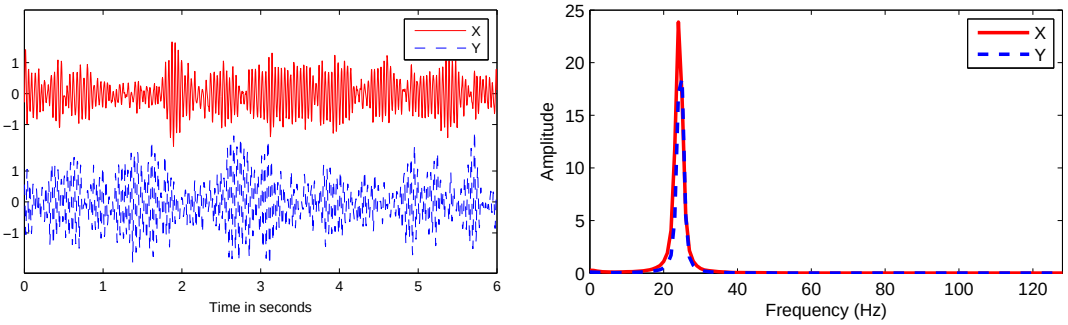
Figure H.1: Gaussian AR signals (left) and the corresponding PSD (right). (a) Model 5. (b) Model 6.



(a)

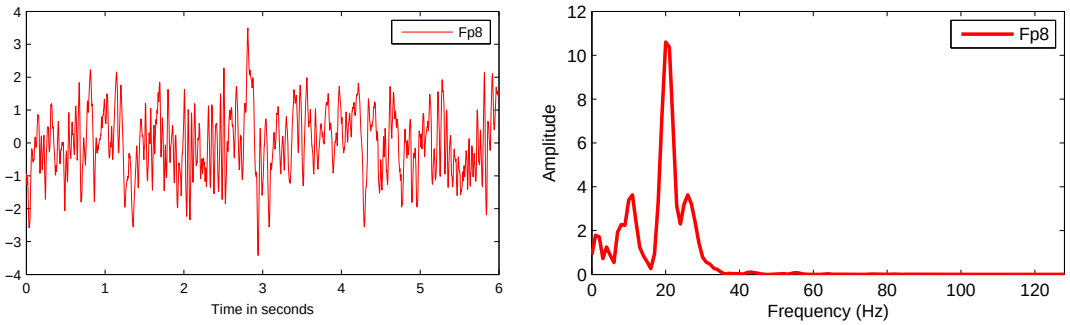


(b)

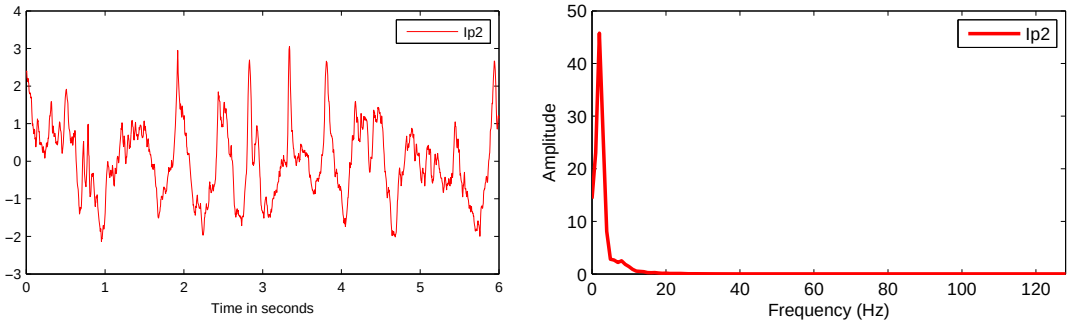


(c)

Figure H.2: Simulated EEG signals generated by the physiology-based model (left) and the corresponding PSD (right). (a) unidirectional situation, $K^{XY} = 1500$, $K^{YX} = 0$, $Pop_{\boldsymbol{X}}$: $A = 5$, $B = 3$ and $G = 20$, $Pop_{\boldsymbol{Y}}$: $A = 3.5$, $B = 3.5$ and $G = 84$. (b) bidirectional situation, $K^{XY} = K^{YX} = 1500$, $Pop_{\boldsymbol{X}}$: $A = 2.8$, $B = 1$ and $G = 40$, $Pop_{\boldsymbol{Y}}$: $A = 3.2$, $B = 1$ and $G = 32.5$. (c) independence situation, $K^{XY} = K^{YX} = 0$, $Pop_{\boldsymbol{X}}$: $A = 5$, $B = 3$ and $G = 20$, $Pop_{\boldsymbol{Y}}$: $A = 3.67$, $B = 2.3$ and $G = 22.45$.



(a)

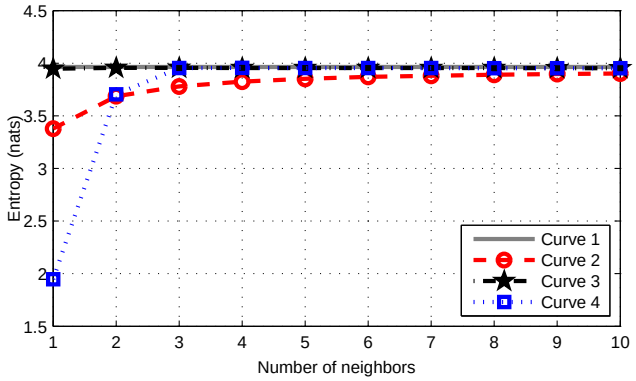


(b)

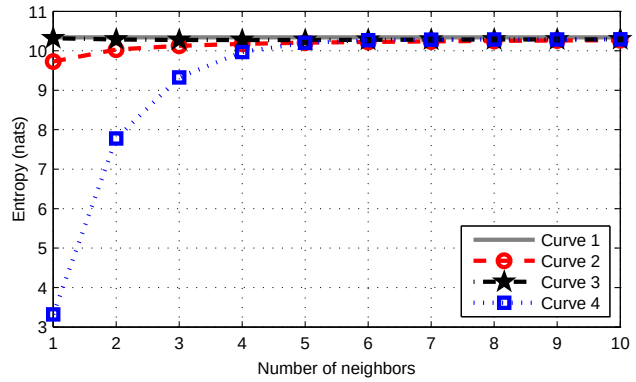
Figure H.3: Real signals recorded during the pre-ictal phase (left) and the corresponding PSD (right). When focusing on different channels or different epochs, very different PSD shapes are encountered. Here are two examples. (a) an example of fast activity (peak around 23 Hz) in channel Fp8 as in [Wendling 2001]. This type of activity that we deliberately generated in the physiology-based model did not occur frequently in the pre-ictal phase of the analyzed seizure. (b) another type of activity localized at lower frequencies in channel Ip2.

I. Comparison between Entropy Estimators

Here, we try to explain the behavior of the estimator $\mathrm{TE}_{\mathrm{p2}}$ in Fig. 4.12(b).



(a)



(b)

Figure I.1: Comparison between four entropy estimators, (a) $d = 3$, (b) $d = 8$. The covariance matrix of the signals is a Toeplitz matrix with first line $\beta^{[0:d-1]}$, where $\beta = 0.5$. "Curve 1" stands for the true value, "Curve 2", "Curve 3" and "Curve 4" correspond to the values of entropy obtained using respectively Equ. (3.60), (3.150) and (3.172).

Fig. I.1 displays the values of entropy for a Gaussian d-dimensional vector as a function of the number of neighbors k, for $d = 3$ in Fig. I.1(a) and $d = 8$ in Fig. I.1(b), obtained with different estimators. The theoretical entropy value is compared with its estimation from the Kozachenko-Leonenko reference estimator (Equ. (3.60), "curve 2", red circles, Box ② in Fig. 3.6), its extension (Equ. (3.150), "curve 3 ", black stars, Box ㉑), and the extension of Singh's estimator (Equ. (3.172), "curve 4", blue squares, Box ㉒). It appears clearly that, for the extended Singh's estimator, the bias (true value minus estimated value) drastically increases when the number of neighbors decreases under a threshold slightly lower than the dimension d of the vector. This allows us to interpret

some apparently surprising results obtained with this estimator in the estimation of TE, as reported in Fig. 4.12(b). TE estimation is a sum of four separate vector entropy estimations, $\widehat{\mathrm{TE}_{Y\to X}} = \widehat{\mathcal{H}(X^-, Y^-)} + \widehat{\mathcal{H}(X^p, X^-)} - \widehat{\mathcal{H}(X^p, X^-, Y^-)} - \widehat{\mathcal{H}(X^-)}$. Here, the dimensions of the four vectors are $d_{X^-,Y^-} = m + n = 4$, $d_{X^p,X^-} = 1 + m = 3$, $d_{X^p,X^-,Y^-} = 1 + m + n = 5$, $d_{X^-} = m = 2$ respectively. Note that, if we denote by X_{M2} and Y_{M2} the two components in Model 5, the general notation (X^p, X^-, Y^-) corresponds to $\left(Y^p_{\mathrm{M2}}, Y^-_{\mathrm{M2}}, X^-_{\mathrm{M2}}\right)$ because in Fig. 4.12(b) the analyzed direction is $X \to Y$ and not the reverse. We see that, when considering the estimation of $\mathcal{H}(X^p, X^-, Y^-) = \mathcal{H}\left(Y^p_{\mathrm{M2}}, Y^-_{\mathrm{M2}}, X^-_{\mathrm{M2}}\right)$, we have $d = 5$ and $k = 3$ which is the imposed neighbors number in the global space. Consequently, from the results shown in Fig. I.1, we can expect that in Model 5 the quantity $\mathcal{H}(X^p, X^-, Y^-)$ will be drastically underestimated. For the other components $\widehat{\mathcal{H}(X^-, Y^-)}$, $\widehat{\mathcal{H}(X^p, X^-)}$, $\widehat{\mathcal{H}(X^-)}$, the numbers of neighbors to consider are generally larger than 3 (as a consequence of Kraskov's technique which introduces projected distances) and $d \leq 5$, so that we do not expect any underestimation of these terms. So, globally, when summing the four entropy estimations, the resulting positive bias observed in Fig. 4.12(b) is understandable.

J. DTF Algorithm Used in Chapter 5

In this appendix, we give brief mathematical descriptions of the DTF index used as a reference index in chapter 5. Let $X(t) = [x_1(t), x_2(t), \ldots, x_d(t)]^T$ denote a d-dimensional (d channels) vector process. An order p VAR model of X is given by

$$X(t) = \sum_{i=1}^{p} A(i) X(t-i) + E(t), \tag{J.1}$$

where E stands for an innovation process vector (which is ideally a random white process). Representing Equ. (J.1) in the frequency domain, we get

$$A(f) X(f) = E(f), \tag{J.2}$$

where $A(f) = -\sum_{n=0}^{p} A(n) e^{-i2\pi n(f/f_s)}$, f_s is the sampling frequency, and $A(0) = -I$ (I being the identity matrix). Equ. (J.2) can be we rewritten as

$$\begin{aligned} X(f) &= A^{-1}(f) E(f) \\ &= H(f) E(f), \end{aligned} \tag{J.3}$$

where H is the transfer matrix from the innovation process to the observed process. The basic DTF from channel i to channel j at frequency f was defined in [Kamiński 1991] as follows

$$\mathrm{DTF}^2_{i\to j}(f) = \frac{|H_{ji}(f)|^2}{\sum_{k=1}^{d} |H_{jk}(f)|^2}. \tag{J.4}$$

To compare this frequency domain index with Granger causality and transfer entropy, we sum it over a chosen frequency band $[f_1,\ f_2]$ and then normalize the sum. Finally, the DTF index (from channel i to channel j) we used in chapter 5 is defined and denoted as

$$\mathrm{DTF}_{i\to j} \triangleq \frac{\sum_{f=f_1}^{f_2} |H_{ji}(f)|^2}{\sum_{k=1}^{d} \sum_{f=f_1}^{f_2} |H_{jk}(f)|^2}. \tag{J.5}$$

For the real signals (sampled at 256 Hz), $[f_1,\ f_2] \subset \{0, \ldots, f_s/2\}$ is set to $[0,\ 128]$ and $\mathrm{DTF}_{i\to j}$ is calculated on a frequency grid with a 1 Hz step. Note that $0 \le \mathrm{DTF}_{i\to j} \le 1$, and $\mathrm{DTF}_{i\to j}$ can be considered as a causality index which reveals a direct effective connectivity which is not sensitive to spurious indirect causality paths (in the connectivity graph built on the d nodes, corresponding to the d channels). Indeed, Equ. (J.1) corresponds to a global model which takes into account the contextual channels set including the d components except for the two components i and j.

K. Independence Test for Granger Causality

In this appendix, we develop the chi-square threshold for Granger causality. First of all, we make an introduction to the notations used in this appendix:

- N is the number of sample points extracted from each of the 2 time series X and Y.
- $\widehat{\sigma}^2_{X|X}$ is the empirical prediction error obtained with an L order AR model for X alone (marked as model (1)).
- $\widehat{\sigma}^2_{X|X,Y}$ is the empirical prediction error obtained when X is predicted from both its own L order past and the L order past of Y (marked as model (2)).
- $\rho = \frac{\widehat{\sigma}^2_{X|X}}{\widehat{\sigma}^2_{X|X,Y}}$.
- $GC \triangleq \log(\rho)$ is the Granger causality index, while $GW \triangleq N(\rho - 1)$ defines another statistic.

Under the H_0 hypothesis, *i.e.* if X and Y are stochastically independent, then the probability distribution of the statistic GW can be approximated by a χ^2_L distribution (centered chi-square distribution with parameter L) when N is sufficiently large (see for example [Hlaváčková-Schindler 2007]). Note that it is easy to compute GW from GC as, clearly, we have $GW = N\left(e^{GC} - 1\right)$.

It is easy to test the hypothesis H_0 corresponding to the acceptation of model (1), with a probability of false positive equal to a given value α_0. If we consider the threshold value $\lambda(\alpha_0)$ defined by $P\left(\chi^2_L > \lambda(\alpha_0)\right) = \alpha_0$, then we have the testing procedure

$$\text{If } GW > \lambda(\alpha_0), \text{ then reject } H_0, \tag{K.1}$$

which is equivalent to

$$\text{If } GC > \log\left(\frac{\lambda(\alpha_0)}{N} + 1\right), \text{ then reject } H_0. \tag{K.2}$$

Now suppose we observe K occurrences $\rho_1, \ldots, \rho_K$ of ρ obtained from K non-overlapping time windows, each of them including N sampling points. If the time correlation of (X, Y) is small comparatively to N, then $\rho_1, \ldots, \rho_K$ can be considered as resulting from K independent trials. Furthermore, if we suppose that the observation (X, Y) is approximately stationary on each time interval supporting data points used to

compute the respective values of ρ, then, under the H_0 hypothesis, the statistic

$$GW_1^K = \sum_{k=1}^{K} GW_k = \sum_{k=1}^{K} N\left(e^{GC_k} - 1\right) \tag{K.3}$$

corresponds approximately to a sum of K independent realizations

$$GW_k = N\left(\rho_k - 1\right),\ k = 1, \ldots, K \tag{K.4}$$

of a random variable following a same χ^2_L distribution. Consequently the statistic GW_1^K follows (approximately) a χ^2_{LK} distribution.

N.B.: If the pair (X, Y) is locally stationary on each of time window, the fact that the AR model could be different on distinct windows has no impact when the order L remains constant.

Finally the following procedure can be proposed to test H_0 from K non-overlapping time windows:

- Determine $\lambda_K(\alpha_0)$ such that $P\left(\chi^2_{LK} > \lambda_K(\alpha_0)\right) = \alpha_0$;
- Compute $GW_1^K = \sum_{k=1}^{K} GW_k = \sum_{k=1}^{K} N\left(e^{GC_k} - 1\right)$;
- If $GW_1^K > \lambda_K(\alpha_0)$ then reject H_0.

Bibliography

[Abramson 1963] Abramson, N. (1963). *Information theory and coding.* McGraw-Hill electronic sciences series. McGraw-Hill.

[Achard 2006] Achard, S., Salvador, R., Whitcher, B., Suckling, J., and Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *The Journal of neuroscience*, 26(1):63–72.

[Adhikari 2013] Adhikari, B. M., Epstein, C. M., and Dhamala, M. (2013). Localizing epileptic seizure onsets with Granger causality. *Physical Review E*, 88(3):030701.

[Aertsen 1991] Aertsen, A. and Preissl, H. (1991). Dynamics of activity and connectivity in physiological neuronal networks. *Nonlinear dynamics and neuronal networks*, 2:281–301.

[Aho 2014] Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of aic and bic. *Ecology*, 95(3):631–636.

[Akaike 1973] Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265.

[Alexandre 2013] Alexandre, F. (2013). Comprendre les réseaux cérébraux. Research Report RR-8219, INRIA.

[Amblard 2012] Amblard, P. O. and Michel, O. J. (2012). The relation between Granger causality and directed information theory: a review. *Entropy*, 15(1):113–143.

[Amblard 2014] Amblard, P. O. and Michel, O. J. (2014). Causal conditioning and instantaneous coupling in causality graphs. *Information Sciences*, 264:279–290.

[Ancona 2004] Ancona, N., Marinazzo, D., and Stramaglia, S. (2004). Radial basis function approach to nonlinear Granger causality of time series. *Physical Review E*, 70(5):056221.

[Angelini 2009] Angelini, L., Pellicoro, M., and Stramaglia, S. (2009). Granger causality for circular variables. *Physics Letters A*, 373(29):2467–2470.

[Angelini 2010] Angelini, L., De Tommaso, M., Marinazzo, D., Nitti, L., Pellicoro, M., and Stramaglia, S. (2010). Redundant variables and Granger causality. *Physical Review E*, 81(3):037201.

[Astolfi 2005] Astolfi, L., Cincotti, F., Mattia, D., Lai, M., de Vico Fallani, F., Salinari, S., Nocchi, F., Baccalà, L., Ursino, M., Zavaglia, M., et al. (2005). Causality estimates among brain cortical areas by partial directed coherence: simulations and application to real data. *Int. J. Bioelectromagn*, 7(1).

[Astolfi 2006] Astolfi, L., Cincotti, F., Mattia, D., Marciani, M. G., Baccala, L., Fallani, F., Salinari, S., Ursino, M., Zavaglia, M., Babiloni, F., et al. (2006). Assessing cortical functional connectivity by partial directed coherence: simulations and application to real data. *IEEE Transactions on Biomedical Engineering*, 53(9):1802–1812.

[Baccalá 2001] Baccalá, L. A. and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biological cybernetics*, 84(6):463–474.

[Baccald 2007] Baccald, L. and de Medicina, F. (2007). Generalized partial directed coherence. In *15th IEEE International Conference on Digital Signal Processing (DSP)*, pages 163–166.

[Bandt 2002a] Bandt, C., Keller, G., and Pompe, B. (2002). Entropy of interval maps via permutations. *Nonlinearity*, 15(5):1595–1602.

[Bandt 2002b] Bandt, C. and Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. *Physical Review Letters*, 88(17):174102.

[Barnett 2009] Barnett, L., Barrett, A. B., and Seth, A. K. (2009). Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701.

[Barnett 2011] Barnett, L. and Seth, A. K. (2011). Behaviour of Granger causality under filtering: theoretical invariance and practical application. *Journal of neuroscience methods*, 201(2):404–419.

[Barnett 2012] Barnett, L. and Bossomaier, T. (2012). Transfer entropy as a log-likelihood ratio. *Physical review letters*, 109(13):138105.

[Barnett 2014] Barnett, L. and Seth, A. K. (2014). The mvgc multivariate Granger causality toolbox: a new approach to Granger-causal inference. *Journal of neuroscience methods*, 223:50–68.

[Barrett 2010] Barrett, A. B., Barnett, L., and Seth, A. K. (2010). Multivariate Granger causality and generalized variance. *Physical Review E*, 81(4):041907.

[Barrett 2013] Barrett, A. B. and Barnett, L. (2013). Granger causality is designed to measure effect, not mechanism. *Frontiers in neuroinformatics*, 7.

[Barrett 2014] Barrett, A. B. and Seth, A. K. (2014). Directed spectral methods. In *Encyclopedia of Computational Neuroscience*, pages 1–5.

[Bassett 2006] Bassett, D. S. and Bullmore, E. (2006). Small-world brain networks. *The neuroscientist*, 12(6):512–523.

[Benesty 2007] Benesty, J., Huang, Y., and Chen, J. (2007). Time delay estimation via minimum entropy. *IEEE Signal Processing Letters*, 14(3):157–160.

[Bernasconi 1999] Bernasconi, C. and KoÈnig, P. (1999). On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings. *Biological cybernetics*, 81(3):199–210.

[Bernasconi 2000] Bernasconi, C., von Stein, A., Chiang, C., and KoÈnig, P. (2000). Bi-directional interactions between visual areas in the awake behaving cat. *Neuroreport*, 11(4):689–692.

[Bertrand 2015] Bertrand, B. and Gérard, O. (2015). http://www.anatomie-humaine.com/Le-Cerveau-1.html/.

[Besserve 2010] Besserve, M., Schölkopf, B., Logothetis, N. K., and Panzeri, S. (2010). Causal relationships between frequency bands of extracellular signals in visual cortex revealed by an information theoretic analysis. *Journal of computational neuroscience*, 29(3):547–566.

[Bettencourt 2008] Bettencourt, L. M., Gintautas, V., and Ham, M. I. (2008). Identification of functional information subgraphs in complex networks. *Physical review letters*, 100(23):238701.

[Bezruchko 2008] Bezruchko, B. P., Ponomarenko, V. I., Prokhorov, M. D., Smirnov, D. A., and Tass, P. A. (2008). Modeling nonlinear oscillatory systems and diagnostics of coupling between them using chaotic time series analysis: applications in neurophysiology. *Physics-Uspekhi*, 51(3):304–310.

[Billings 2013] Billings, S. A. (2013). *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons.

[Bishop 1995] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

[Biswal 1997] Biswal, B. B., Kylen, J. V., and Hyde, J. S. (1997). Simultaneous assessment of flow and bold signals in resting-state functional connectivity maps. *NMR in Biomedicine*, 10(45):165–170.

[Bozoklu 2013] Bozoklu, S. and Yilanci, V. (2013). Energy consumption and economic growth for selected oecd countries: Further evidence from the Granger causality test in the frequency domain. *Energy Policy*, 63:877–881.

[Breedlove 2007] Breedlove, S. M., Watson, N. V., and Rosenzweig, M. R. (2007). *Biological psychology*. Sinauer Associates, Incorporated Publishers.

[Bressler 2008] Bressler, S. L., Tang, W., Sylvester, C. M., Shulman, G. L., and Corbetta, M. (2008). Top-down control of human visual cortex by frontal and parietal cortex in anticipatory visual spatial attention. *The Journal of Neuroscience*, 28(40):10056–10061.

[Bressler 2011] Bressler, S. L. and Seth, A. K. (2011). Wiener–Granger causality: a well established methodology. *Neuroimage*, 58(2):323–329.

[Brillouin 2013] Brillouin, L. (2013). *Science and information theory*. Courier Corporation.

[Brockwell 2013] Brockwell, P. J. and Davis, R. A. (2013). *Time series: theory and methods*. Springer Science & Business Media.

[Brovelli 2004] Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., and Bressler, S. L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9849–9854.

[Browne 2008] Browne, T. R. and Holmes, G. L. (2008). *Handbook of epilepsy*. Jones & Bartlett Learning.

[Buehlmann 2010] Buehlmann, A. and Deco, G. (2010). Optimal information transfer in the cortex through synchronization. *PLoS computational biology*, 6(9):e1000934.

[Burnham 2002] Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

[Caines 1975] Caines, P. E. and Chan, C. (1975). Feedback between stationary stochastic processes. *IEEE Transactions on Automatic Control*, 20(4):498–508.

[Caparos 2006] Caparos, M., Louis Dorr, V., Wendling, F., Maillard, L., and Wolf, D. (2006). Automatic lateralization of temporal lobe epilepsy based on scalp eeg. *Clinical neurophysiology*, 117(11):2414–2423.

[Chang 2006] Chang, C. I., Du, Y., Wang, J., Guo, S. M., and Thouin, P. (2006). Survey and comparative analysis of entropy and relative entropy thresholding techniques. In *IEE Proceedings on Vision, Image and Signal Processing*, volume 153, pages 837–850.

[Chávez 2003] Chávez, M., Martinerie, J., and Le Van Quyen, M. (2003). Statistical assessment of nonlinear causality: application to epileptic eeg signals. *Journal of neuroscience methods*, 124(2):113–128.

[Chen 2008] Chen, C., Kiebel, S. J., and Friston, K. J. (2008). Dynamic causal modelling of induced responses. *Neuroimage*, 41(4):1293–1312.

[Chen 2004] Chen, Y., Rangarajan, G., Feng, J., and Ding, M. (2004). Analyzing multiple nonlinear time series with extended Granger causality. *Physics Letters A*, 324(1):26–35.

[Chen 2006] Chen, Y., Bressler, S. L., and Ding, M. (2006). Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data. *Journal of neuroscience methods*, 150(2):228–237.

[Chicharro 2011] Chicharro, D. (2011). On the spectral formulation of Granger causality. *Biological cybernetics*, 105(5-6):331–347.

[Chiou 2008] Chiou Wei, S. Z., Chen, C. F., and Zhu, Z. (2008). Economic growth and energy consumption revisited–evidence from linear and nonlinear Granger causality. *Energy Economics*, 30(6):3063–3076.

[Comincioli 1996] Comincioli, B. (1996). The stock market as a leading indicator: An application of Granger causality. *University Avenue Undergraduate Journal of Economics*, 1(1):1.

[Cossu 2005] Cossu, M., Cardinale, F., Castana, L., Citterio, A., Francione, S., Tassi, L., Benabid, A. L., and Russo, G. L. (2005). Stereoelectroencephalography in the presurgical evaluation of focal epilepsy: a retrospective analysis of 215 procedures. *Neurosurgery*, 57(4):706–718.

[Costa 2004] Costa, J., Hero, A. O., et al. (2004). Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221.

[Cover 2012] Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[Croux 2013] Croux, C. and Reusens, P. (2013). Do stock prices contain predictive power for the future economic activity? a Granger causality analysis in the frequency domain. *Journal of Macroeconomics*, 35:93–103.

[Cui 2008] Cui, J., Xu, L., Bressler, S. L., Ding, M., and Liang, H. (2008). Bsmart: a matlab/c toolbox for analysis of multichannel neural time series. *Neural Networks*, 21(8):1094–1104.

[Dahlhaus 2000] Dahlhaus, R. (2000). Graphical interaction models for multivariate time series1. *Metrika*, 51(2):157–172.

[Daunizeau 2009] Daunizeau, J., Kiebel, S. J., and Friston, K. J. (2009). Dynamic causal modelling of distributed electromagnetic responses. *NeuroImage*, 47(2):590–601.

[Davey 2013] Davey, C. E., Grayden, D. B., Gavrilescu, M., Egan, G. F., and Johnston, L. A. (2013). The equivalence of linear gaussian connectivity techniques. *Human brain mapping*, 34(9):1999–2014.

[David 2008] David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., and Depaulis, A. (2008). Identifying neural drivers with functional mri: an electrophysiological validation. *PLoS Biol*, 6(12):e315.

[de Tisi 2011] de Tisi, J., Bell, G. S., Peacock, J. L., McEvoy, A. W., Harkness, W. F., Sander, J. W., and Duncan, J. S. (2011). The long-term outcome of adult epilepsy surgery, patterns of seizure remission, and relapse: a cohort study. *The Lancet*, 378(9800):1388–1395.

[Deshpande 2009] Deshpande, G., LaConte, S., James, G. A., Peltier, S., and Hu, X. (2009). Multivariate Granger causality analysis of fmri data. *Human brain mapping*, 30(4):1361–1373.

[Deshpande 2010a] Deshpande, G., Sathian, K., and Hu, X. (2010a). Assessing and compensating for zero-lag correlation effects in time-lagged Granger causality analysis of fmri. *IEEE Transactions on Biomedical Engineering*, 57(6):1446–1456.

[Deshpande 2010b] Deshpande, G., Sathian, K., and Hu, X. (2010b). Effect of hemodynamic variability on Granger causality analysis of fmri. *Neuroimage*, 52(3):884–896.

[Detto 2013] Detto, M., Bohrer, G., Nietz, J. G., Maurer, K. D., Vogel, C. S., Gough, C. M., and Curtis, P. S. (2013). Multivariate conditional Granger causality analysis for lagged response of soil respiration in a temperate forest. *Entropy*, 15(10):4266–4284.

[Devlin 2012] Devlin, A. L., Odell, M., Charlton, J. L., and Koppel, S. (2012). Epilepsy and driving: Current status of research. *Epilepsy research*, 102(3):135–152.

[Dewey 2007] Dewey, R. A. (2007). *Psychology: an introduction*. Russ Dewey.

[Dimitrov 2011] Dimitrov, A. G., Lazar, A. A., and Victor, J. D. (2011). Information theory in neuroscience. *Journal of computational neuroscience*, 30(1):1–5.

[Ding 2007] Ding, L., Worrell, G. A., Lagerlund, T. D., and He, B. (2007). Ictal source analysis: localization and imaging of causal interactions in humans. *Neuroimage*, 34(2):575–586.

[Ding 2000] Ding, M., Bressler, S. L., Yang, W., and Liang, H. (2000). Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment. *Biological cybernetics*, 83(1):35–45.

[Ding 2006] Ding, M., Chen, Y., and Bressler, S. L. (2006). Granger causality: basic theory and application to neuroscience. *Handbook of time series analysis: recent theoretical developments and applications*, page 437.

[Dorr 2007] Dorr, V. L., Caparos, M., Wendling, F., Vignal, J.-P., and Wolf, D. (2007). Extraction of reproducible seizure patterns based on eeg scalp correlations. *Biomedical Signal Processing and Control*, 2(3):154–162.

[Dorr 2007] Dorr, V. L., Caparos, M., Wendling, F., Vignal, J. P., and Wolf, D. (2007). Extraction of reproducible seizure patterns based on eeg scalp correlations. *Biomedical Signal Processing and Control*, 2(3):154–162.

[Duncan 2006] Duncan, J. S., Sander, J. W., Sisodiya, S. M., and Walker, M. C. (2006). Adult epilepsy. *The Lancet*, 367(9516):1087–1100.

[Dunleavy 2012] Dunleavy, A. J., Wiesner, K., and Royall, C. P. (2012). Using mutual information to measure order in model glass formers. *Physical Review E*, 86(4):041505.

[Eadie 2012] Eadie, M. J. (2012). Shortcomings in the current treatment of epilepsy. *Expert review of neurotherapeutics*, 12(12):1419–1427.

[Eichler 2012] Eichler, M. (2012). Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, 153(1-2):233–268.

[Ellen 2015] Ellen, A. and David, F. (2015). http://www.mayfieldclinic.com/PE-EpilepsySurg.htm/.

[Engel 2008] Engel, J., Pedley, T. A., and Aicardi, J. (2008). *Epilepsy: a comprehensive textbook*, volume 3. Lippincott Williams & Wilkins.

[EPFL 2015] EPFL (2015). http://bluebrain.epfl.ch/.

[Epstein 2014] Epstein, C. M., Adhikari, B. M., Gross, R., Willie, J., and Dhamala, M. (2014). Application of high-frequency Granger causality to analysis of epileptic seizures and surgical decision making. *Epilepsia*, 55(12):2038–2047.

[Eric 2015] Eric, C. (2015). Brain facts and figures. https://faculty.washington.edu/chudler/facts.html/.

[Faes 2006] Faes, L. and Nollo, G. (2006). Bivariate nonlinear prediction to quantify the strength of complex dynamical interactions in short-term cardiovascular variability. *Medical and Biological Engineering and Computing*, 44(5):383–392.

[Faes 2010] Faes, L. and Nollo, G. (2010). Extended causal modeling to assess partial directed coherence in multiple time series with significant instantaneous interactions. *Biological cybernetics*, 103(5):387–400.

[Faes 2011] Faes, L., Nollo, G., and Porta, A. (2011). Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. *Physical Review E*, 83(5):051112.

[Faes 2012] Faes, L., Nollo, G., and Porta, A. (2012). Non-uniform multivariate embedding to assess the information transfer in cardiovascular and cardiorespiratory variability series. *Computers in biology and medicine*, 42(3):290–297.

[Faes 2014] Faes, L. and Porta, A. (2014). Conditional entropy-based evaluation of information dynamics in physiological systems. In *Directed information measures in neuroscience*, pages 61–86.

[Fisher 2005] Fisher, R. S., Boas, W. v. E., Blume, W., Elger, C., Genton, P., Lee, P., and Engel, J. (2005). Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe). *Epilepsia*, 46(4):470–472.

[Foster 2011] Foster, D. V. and Grassberger, P. (2011). Lower bounds on mutual information. *Physical Review E*, 83(1):010101.

[Freiwald 1999] Freiwald, W. A., Valdes, P., Bosch, J., Biscay, R., Jimenez, J. C., Rodriguez, L. M., Rodriguez, V., Kreiter, A. K., and Singer, W. (1999). Testing nonlinearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. *Journal of neuroscience methods*, 94(1):105–119.

[Frenzel 2007] Frenzel, S. and Pompe, B. (2007). Partial mutual information for coupling analysis of multivariate time series. *Physical review letters*, 99(20):204101.

[Friston 1994] Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping*, 2(1-2):56–78.

[Friston 2003] Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302.

[Friston 2009] Friston, K. J. (2009). Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS biology*, 7(2):220.

[Friston 2014] Friston, K. J., Bastos, A. M., Oswal, A., van Wijk, B., Richter, C., and Litvak, V. (2014). Granger causality revisited. *NeuroImage*, 101:796–808.

[Frogerais 2008] Frogerais, P. (2008). *Modélisation et identification en épilepsie: De la dynamique des populations neuronales aux signaux EEG*. PhD thesis, Université Rennes 1.

[Fukunaga 2013] Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Academic press.

[Fukunaga 1973] Fukunaga, K. and Hostetler, L. D. (1973). Optimization of k nearest neighbor density estimates. *Information Theory, IEEE Transactions on*, 19(3):320–326.

[Gao 2011] Gao, Q., Duan, X., and Chen, H. (2011). Evaluation of effective connectivity of motor areas during motor imagery and execution using conditional Granger causality. *Neuroimage*, 54(2):1280–1288.

[Gao 2015] Gao, S., Steeg, G. V., and Galstyan, A. (2015). Efficient estimation of mutual information for strongly dependent variables. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*.

[Garofalo 2009] Garofalo, M., Nieus, T., Massobrio, P., and Martinoia, S. (2009). Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks. *PloS one*, 4(8):e6482.

[Geweke 1982] Geweke, J. F. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304–313.

[Geweke 1984] Geweke, J. F. (1984). Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79(388):907–915.

[Giovannoni 2015] Giovannoni, G. (2015). a brief beginner's guide to the brain and mri. http://multiple-sclerosis-research.blogspot.com/2015/01/education-whats-mri.html/.

[Gómez-Herrero 2015] Gómez-Herrero, G., Wu, W., Rutanen, K., Soriano, M. C., Pipa, G., and Vicente, R. (2015). Assessing coupling dynamics from an ensemble of time series. *Entropy*, 17(4):1958.

[Gourévitch 2007] Gourévitch, B. and Eggermont, J. J. (2007). Evaluating information transfer between auditory cortical neurons. *Journal of Neurophysiology*, 97(3):2533–2543.

[Gow 2008] Gow, D. W., Segawa, J. A., Ahlfors, S. P., and Lin, F. H. (2008). Lexical influences on speech perception: a Granger causality analysis of meg and eeg source estimates. *Neuroimage*, 43(3):614–623.

[Gow 2009] Gow, D. W., Keller, C. J., Eskandar, E., Meng, N., and Cash, S. S. (2009). Parallel versus serial processing dependencies in the perisylvian speech network: a Granger analysis of intracranial eeg data. *Brain and language*, 110(1):43–48.

[Grafton 1994] Grafton, S. T., Sutton, J., Couldwell, W., Lew, M., and Waters, C. (1994). Network analysis of motor system connectivity in parkinson's disease: modulation of thalamocortical interactions after pallidotomy. *Human Brain Mapping*, 2(1-2):45–55.

[Granger 1969] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

[Granger 2000] Granger, C. W., Huangb, B. N., and Yang, C. W. (2000). A bivariate causality between stock prices and exchange rates: evidence from recent asianflu. *The Quarterly Review of Economics and Finance*, 40(3):337–354.

[Greene 2003] Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall, 5. edition.

[Guo 2008a] Guo, S., Seth, A. K., Kendrick, K. M., Zhou, C., and Feng, J. (2008a). Partial Granger causality–eliminating exogenous inputs and latent variables. *Journal of neuroscience methods*, 172(1):79–93.

[Guo 2008b] Guo, S., Wu, J., Ding, M., and Feng, J. (2008b). Uncovering interactions in the frequency domain. *PLoS Comput Biol*, 4(5):e1000087.

[Haines 2015] Haines, D. and Mihailoff, G. (2015). The telencephalon. http://clinicalgate.com/the-telencephalon/.

[Hamilton 1994] Hamilton, J. D. (1994). *Time series analysis*, volume 2. Princeton university press Princeton.

[He 2014] He, F., Wei, H.-L., Billings, S. A., and Sarrigiannis, P. G. (2014). A nonlinear generalization of spectral Granger causality. *IEEE transactions on bio-medical engineering*, 61(6):1693–1701.

[He 2007] He, Y., Chen, Z. J., and Evans, A. C. (2007). Small-world anatomical networks in the human brain revealed by cortical thickness from mri. *Cerebral cortex*, 17(10):2407–2419.

[Hebb 2005] Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology Press.

[Hernández 1996] Hernández, J. L., Valdés, P. A., and Vila, P. (1996). eeg spike and wave modelled by a stochastic limit cycle. *NeuroReport*, 7(13):2246.

[Hesse 2003] Hesse, W., Möller, E., Arnold, M., and Schack, B. (2003). The use of time-variant eeg Granger causality for inspecting directed interdependencies of neural assemblies. *Journal of neuroscience methods*, 124(1):27–44.

[Hlaváčková-Schindler 2007] Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., and Bhattacharya, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46.

[Hlaváčková-Schindler 2011] Hlaváčková-Schindler, K. (2011). Equivalence of Granger causality and transfer entropy: a generalization. *Applied Mathematical Sciences*, 5(73):3637–3648.

[Hlinka 2013] Hlinka, J., Hartman, D., Vejmelka, M., Runge, J., Marwan, N., Kurths, J., and Paluš, M. (2013). Reliability of inference of directed climate networks using conditional mutual information. *Entropy*, 15(6):2023–2045.

[Honey 2007] Honey, C. J., Kötter, R., Breakspear, M., and Sporns, O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Sciences*, 104(24):10240–10245.

[Horwitz 2003] Horwitz, B. (2003). The elusive concept of brain connectivity. *Neuroimage*, 19(2):466–470.

[Hu 2011] Hu, S., Dai, G., Worrell, G., Dai, Q., Liang, H., et al. (2011). Causality analysis of neural connectivity: critical examination of existing methods and advances of new methods. *IEEE Transactions on Neural Networks*, 22(6):829–844.

[Hu 2012] Hu, S. and Liang, H. (2012). Causality analysis of neural connectivity: New tool and limitations of spectral Granger causality. *Neurocomputing*, 76(1):44–47.

[Hwang 1994] Hwang, J. N., Lay, S. R., and Lippman, A. (1994). Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing*, 42(10):2795–2810.

[Ishiguro 2008] Ishiguro, K., Otsu, N., Lungarella, M., and Kuniyoshi, Y. (2008). Comparison of nonlinear Granger causality extensions for low-dimensional systems. *Physical Review E*, 77(3):036217.

[Jeong 2001] Jeong, J., Gore, J. C., and Peterson, B. S. (2001). Mutual information analysis of the eeg in patients with alzheimer's disease. *Clinical Neurophysiology*, 112(5):827–835.

[Jin 2010] Jin, S. H., Lin, P., and Hallett, M. (2010). Linear and nonlinear information flow based on time-delayed mutual information method and its application to corticomuscular interaction. *Clinical Neurophysiology*, 121(3):392–401.

[Jung 2008] Jung, G. J. and Oh, Y.-H. (2008). Information distance-based subvector clustering for asr parameter quantization. *IEEE Signal Processing Letters*, 15:209–212.

[Jung 2011] Jung, Y. J., Kang, H. C., Choi, K. O., Lee, J. S., Kim, D. S., Cho, J. H., Kim, S. H., Im, C. H., and Kim, H. D. (2011). Localization of ictal onset zones in lennox-gastaut syndrome using directional connectivity analysis of intracranial electroencephalography. *Seizure*, 20(6):449–457.

[Kaiser 2002] Kaiser, A. and Schreiber, T. (2002). Information transfer in continuous processes. *Physica D: Nonlinear Phenomena*, 166(1):43–62.

[Kamiński 1991] Kamiński, M. and Blinowska, K. J. (1991). A new method of the description of the information flow in the brain structures. *Biological cybernetics*, 65(3):203–210.

[Kamiński 2001] Kamiński, M., Ding, M., Truccolo, W. A., and Bressler, S. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological cybernetics*, 85(2):145–157.

[Kaufmann 1997] Kaufmann, R. K. and Stern, D. I. (1997). Evidence for human influence on climate from hemispheric temperature relations. *Nature*, 388(6637):39–44.

[Kenneth 1998] Kenneth, S. S. and Carol, M. (1998). *Anatomy and physiology: The unity of form and function.* McGraw-Hill. Boston, Massachusetts, USA.

[Kiebel 2006] Kiebel, S. J., David, O., and Friston, K. J. (2006). Dynamic causal modelling of evoked responses in eeg/meg with lead field parameterization. *NeuroImage*, 30(4):1273–1284.

[Kiebel 2007] Kiebel, S. J., Klöppel, S., Weiskopf, N., and Friston, K. J. (2007). Dynamic causal modeling: a generative model of slice timing in fmri. *Neuroimage*, 34(4):1487–1496.

[Kim 2013] Kim, J., Kim, G., An, S., Kwon, Y. K., and Yoon, S. (2013). Entropy-based analysis and bioinformatics-inspired integration of global economic information transfer. *PloS one*, 8(1):e51986.

[Kim 2012] Kim, M. (2012). Time-series dimensionality reduction via Granger causality. *IEEE Signal Processing Letters*, 19(10):611–614.

[Kozachenko 1987] Kozachenko, L. F. and Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16.

[Kraskov 2004] Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical review E*, 69(6):066138.

[Krishna 2008] Krishna, R. and Guo, S. (2008). A partial Granger causality approach to explore causal networks derived from multi-parameter data. volume 5307 of *Lecture Notes in Computer Science*, pages 9–27.

[Kugiumtzis 2013] Kugiumtzis, D. (2013). Direct-coupling information measure from nonuniform embedding. *Physical Review E*, 87(6):062918.

[Kurzweil 2013] Kurzweil, R. (2013). *How to Create a Mind: The Secret of Human Thought Revealed*. Penguin Books, New York, NY, USA.

[Kwon 2008] Kwon, O. and Yang, J.-S. (2008). Information flow between stock indices. *Europhysics Letters*, 82(6):68003.

[Ladroue 2009] Ladroue, C., Guo, S., Kendrick, K., and Feng, J. (2009). Beyond element-wise interactions: identifying complex interactions in biological processes. *PLoS One*, 4(9):e6899.

[Lang 2012] Lang, E. W., Tomé, A., Keck, I. R., Górriz-Sáez, J., and Puntonet, C. (2012). Brain connectivity analysis: a short survey. *Computational intelligence and neuroscience*, 2012:8.

[Latham 2009] Latham, P. E. and Roudi, Y. (2009). Mutual information. *Scholarpedia*, 4(1):1658.

[Leonenko 2008] Leonenko, N., Pronzato, L., Savani, V., et al. (2008). A class of rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5):2153–2182.

[Li 2012] Li, Y., Wei, H. L., Billings, S. A., and Liao, X. F. (2012). Time-varying linear and nonlinear parametric model for Granger causality analysis. *Physical Review E*, 85(4):041906.

[Liang 2000] Liang, H., Ding, M., Nakamura, R., and Bressler, S. L. (2000). Causal influences in primate cerebral cortex during visual pattern discrimination. *Neuroreport*, 11(13):2875–2880.

[Liao 2009] Liao, W., Marinazzo, D., Pan, Z., Gong, Q., and Chen, H. (2009). Kernel Granger causality mapping effective connectivity on fmri data. *IEEE Transactions on Medical Imaging*, 28(11):1825–1835.

[Liao 2010] Liao, W., Mantini, D., Zhang, Z., Pan, Z., Ding, J., Gong, Q., Yang, Y., and Chen, H. (2010). Evaluating the effective connectivity of resting state networks using conditional Granger causality. *Biological cybernetics*, 102(1):57–69.

[Liao 2011] Liao, W., Ding, J., Marinazzo, D., Xu, Q., Wang, Z., Yuan, C., Zhang, Z., Lu, G., and Chen, H. (2011). Small-world directed networks in the human brain: multivariate Granger causality analysis of resting-state fmri. *Neuroimage*, 54(4):2683–2694.

[Lindner 2011] Lindner, M., Vicente, R., Priesemann, V., and Wibral, M. (2011). Trentool: A matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC neuroscience*, 12(1):119.

[Lizier 2008] Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2008). Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E*, 77(2):026110.

[Lizier 2010] Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2010). Information modification and particle collisions in distributed computation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(3):037109.

[Lizier 2011] Lizier, J. T., Heinzle, J., Horstmann, A., Haynes, J.-D., and Prokopenko, M. (2011). Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fmri connectivity. *Journal of Computational Neuroscience*, 30(1):85–107.

[Lizier 2014] Lizier, J. T. (2014). Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1:11.

[Lobier 2014] Lobier, M., Siebenhühner, F., Palva, S., and Palva, J. M. (2014). Phase transfer entropy: a novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions. *Neuroimage*, 85:853–872.

[Lu 2012] Lu, Y., Yang, L., Worrell, G. A., and He, B. (2012). Seizure source imaging by means of fine spatio-temporal dipole localization and directed transfer function in partial epilepsy patients. *Clinical Neurophysiology*, 123(7):1275–1283.

[Lüdtke 2010] Lüdtke, N., Logothetis, N. K., and Panzeri, S. (2010). Testing methodologies for the nonlinear analysis of causal relationships in neurovascular coupling. *Magnetic resonance imaging*, 28(8):1113–1119.

[Lungarella 2007a] Lungarella, M., Ishiguro, K., Kuniyoshi, Y., and Otsu, N. (2007a). Methods for quantifying the causal structure of bivariate time series. *International journal of bifurcation and chaos*, 17(03):903–921.

[Lungarella 2007b] Lungarella, M., Pitti, A., and Kuniyoshi, Y. (2007b). Information transfer at multiple scales. *Physical Review E*, 76(5):056117.

[Lütkepohl 2005] Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Publishing Company, Incorporated.

[Magiorkinis 2010] Magiorkinis, E., Sidiropoulou, K., and Diamantis, A. (2010). Hallmarks in the history of epilepsy: epilepsy in antiquity. *Epilepsy and Behavior*, 17(1):103–108.

[Malani 2012] Malani, P. N. (2012). Harrison's principles of internal medicine. *JAMA*, 308(17):1813–1814.

[Marinazzo 2008a] Marinazzo, D., Pellicoro, M., and Stramaglia, S. (2008a). Kernel-Granger causality and the analysis of dynamical networks. *Physical review E*, 77(5):056215.

[Marinazzo 2008b] Marinazzo, D., Pellicoro, M., and Stramaglia, S. (2008b). Kernel method for nonlinear Granger causality. *Physical Review Letters*, 100(14):144103.

[Marinazzo 2011] Marinazzo, D., Liao, W., Chen, H., and Stramaglia, S. (2011). Nonlinear connectivity by Granger causality. *Neuroimage*, 58(2):330–338.

[Marinazzo 2014] Marinazzo, D., Wu, G., Pellicoro, M., and Stramaglia, S. (2014). Information transfer in the brain: Insights from a unified approach. In *Directed Information Measures in Neuroscience*, pages 87–110.

[Marreiros 2008] Marreiros, A. C., Kiebel, S. J., and Friston, K. J. (2008). Dynamic causal modelling for fmri: a two-state model. *Neuroimage*, 39(1):269–278.

[Marreiros 2009] Marreiros, A. C., Kiebel, S. J., Daunizeau, J., Harrison, L. M., and Friston, K. J. (2009). Population dynamics under the laplace assumption. *Neuroimage*, 44(3):701–714.

[Martins Jr 2008] Martins Jr, D. C., Braga-Neto, U. M., Hashimoto, R. F., Bittner, M. L., and Dougherty, E. R. (2008). Intrinsically multivariate predictive genes. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):424–439.

[Materassi 2007] Materassi, M., Wernik, A., and Yordanova, E. (2007). Determining the verse of magnetic turbulent cascades in the earth's magnetospheric cusp via transfer entropy analysis: preliminary results. *Nonlinear Processes in Geophysics*, 14(2):153–161.

[McCracken 2007] McCracken, M. W. (2007). Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics*, 140(2):719–752.

[McIntosh 1991] McIntosh, A. and Gonzalez-Lima, F. (1991). Structural modeling of functional neural pathways mapped with 2-deoxyglucose: effects of acoustic startle habituation on the auditory system. *Brain research*, 547(2):295–302.

[McIntosh 1992] McIntosh, A. and Gonzalez-Lima, F. (1992). Structural modeling of functional visual pathways mapped with 2-deoxyglucose: effects of patterned light and footshock. *Brain Research*, 578(1):75–86.

[McIntosh 1994] McIntosh, A., Grady, C., Ungerleider, L. G., Haxby, J., Rapoport, S., and Horwitz, B. (1994). Network analysis of cortical visual pathways mapped with pet. *The journal of neuroscience*, 14(2):655–666.

[Melzer 2014] Melzer, A. and Schella, A. (2014). Symbolic transfer entropy analysis of the dust interaction in the presence of wakefields in dusty plasmas. *Physical Review E*, 89(4):041103.

[Merkwirth 2000] Merkwirth, C., Parlitz, U., and Lauterborn, W. (2000). Fast nearest-neighbor searching for nonlinear signal processing. *Physical Review E*, 62(2):2089.

[Michael 2012] Michael, D., Jerry, L., and Hugh, H. (2012). The effects of cryopreservation on the cat. http://chronopause.com/chronopause.com/index.php/2012/02/21/the-effects-of-cryopreservation-on-the-cat-part-3/index.html/.

[Mierlo 2011] Mierlo, P., Carrette, E., Hallez, H., Vonck, K., Van Roost, D., Boon, P., and Staelens, S. (2011). Accurate epileptogenic focus localization through time-variant functional connectivity analysis of intracranial electroencephalographic signals. *Neuroimage*, 56(3):1122–1133.

[Mierlo 2013] Mierlo, P., Carrette, E., Hallez, H., Raedt, R., Meurs, A., Vandenberghe, S., Roost, D., Boon, P., Staelens, S., and Vonck, K. (2013). Ictal-onset localization through connectivity analysis of intracranial eeg signals in patients with refractory epilepsy. *Epilepsia*, 54(8):1409–1418.

[Miller 1980] Miller, A., Alston, R., and Corsellis, J. (1980). Variation with age in the volumes of grey and white matter in the cerebral hemispheres of man: measurements with an image analyser. *Neuropathology and applied neurobiology*, 6(2):119–132.

[Min 2003] Min, B. C., Jin, S. H., Kang, I. H., Lee, D. H., Kang, J. K., Lee, S. T., and Sakamoto, K. (2003). Analysis of mutual information content for eeg responses to odor stimulation for subjects classified by occupation. *Chemical senses*, 28(9):741–749.

[Montalto 2014] Montalto, A., Faes, L., and Marinazzo, D. (2014). Mute: a matlab toolbox to compare established and novel estimators of the multivariate transfer entropy. *PloS one*, 9(10):e109462.

[Mooshagian 2008] Mooshagian, E. (2008). Anatomy of the corpus callosum reveals its function. *The Journal of Neuroscience*, 28(7):1535–1536.

[Moran 2009] Moran, R. J., Stephan, K. E., Seidenbecher, T., Pape, H. C., Dolan, R. J., and Friston, K. J. (2009). Dynamic causal models of steady-state responses. *NeuroImage*, 44(3):796–811.

[Mosedale 2006] Mosedale, T. J., Stephenson, D. B., Collins, M., and Mills, T. C. (2006). Granger causality of coupled climate processes: Ocean feedback on the north atlantic oscillation. *Journal of Climate*, 19(7):1182–1194.

[Na 2002] Na, S. H., Jin, S. H., Kim, S. Y., and Ham, B. J. (2002). eeg in schizophrenic patients: mutual information analysis. *Clinical Neurophysiology*, 113(12):1954–1960.

[Naghavi 2015] Naghavi, M., Wang, H., Lozano, R., Davis, A., Liang, X., Zhou, M., Vollset, S. E., Ozgoren, A. A., Abdalla, S., Abd-Allah, F., et al. (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013. *Lancet*, 385(9963):117–171.

[Neymotin 2011] Neymotin, S. A., Jacobs, K. M., Fenton, A. A., and Lytton, W. W. (2011). Synaptic information transfer in computer models of neocortical columns. *Journal of computational neuroscience*, 30(1):69–84.

[Nichols 2005] Nichols, J., Seaver, M., Trickey, S., Todd, M., Olson, C., and Overbey, L. (2005). Detecting nonlinearity in structural systems using the transfer entropy. *Physical Review E*, 72(4):046217.

[Noback 2005] Noback, C. R., Strominger, N. L., Demarest, R. J., and Ruggiero, D. A. (2005). *The human nervous system: structure and function*. Number 744. Springer Science & Business Media.

[Nunes 2012] Nunes, V. D., Sawyer, L., Neilson, J., Sarri, G., and Cross, J. H. (2012). Diagnosis and management of the epilepsies in adults and children: summary of updated nice guidance. *BMJ*, 344.

[Omidvarnia 2012] Omidvarnia, A. H., Azemi, G., Boashash, B., O'Toole, J. M., Colditz, P., and Vanhatalo, S. (2012). Orthogonalized partial directed coherence for functional connectivity analysis of newborn eeg. In *Neural Information Processing*, pages 683–691.

[Omidvarnia 2014] Omidvarnia, A. H., Azemi, G., Boashash, B., O'Toole, J. M., Colditz, P. B., and Vanhatalo, S. (2014). Measuring time-varying information flow in scalp eeg signals: orthogonalized partial directed coherence. *IEEE Transactions on Biomedical Engineering*, 61(3):680–693.

[Palmini 2006] Palmini, A. (2006). The concept of the epileptogenic zone: a modern look at penfield and jasper's views on the role of interictal spikes. *Epileptic disorders*, 8(2):10–15.

[Paluš 2001] Paluš, M., Komárek, V., Hrnčíř, Z., and Štěrbová, K. (2001). Synchronization as adjustment of information rates: detection from bivariate time series. *Physical Review E*, 63(4):046211.

[Pampu 2013] Pampu, N. C., Vicente, R., Muresan, R. C., Priesemann, V., Siebenhuhner, F., and Wibral, M. (2013). Transfer entropy as a tool for reconstructing interaction delays in neural signals. In *2013 IEEE International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4.

[Pao 2011] Pao, H. T. and Tsai, C. M. (2011). Multivariate Granger causality between co2 emissions, energy consumption, fdi (foreign direct investment) and gdp (gross domestic product): evidence from a panel of bric (brazil, russian federation, india, and china) countries. *Energy*, 36(1):685–693.

[Papoulis 1985] Papoulis, A. and Pillai, S. U. (1985). Probability, random variables, and stochastic processes. *McGraw-Hill*.

[Parzen 1962] Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, pages 1065–1076.

[Pearl 2009] Pearl, J. (2009). *Causality: models, reasoning, and inference*. Cambridge university press.

[Penny 2004] Penny, W. D., Stephan, K. E., Mechelli, A., and Friston, K. J. (2004). Modelling functional integration: a comparison of structural equation and dynamic causal models. *Neuroimage*, 23:S264–S274.

[Penny 2009] Penny, W. D., Litvak, V., Fuentemilla, L., Duzel, E., and Friston, K. J. (2009). Dynamic causal models for phase coupling. *Journal of neuroscience methods*, 183(1):19–30.

[Pereda 2005] Pereda, E., Quiroga, R. Q., and Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Progress in neurobiology*, 77(1):1–37.

[Petkov 2015] Petkov, C. I., Kikuchi, Y., Milne, A. E., Mishkin, M., Rauschecker, J. P., and Logothetis, N. K. (2015). Different forms of effective connectivity in primate frontotemporal pathways. *Nature communications*, 6.

[Pham 2004] Pham, D. T. (2004). Fast algorithms for mutual information based independent component analysis. *IEEE Transactions on Signal Processing*, 52(10):2690–2700.

[Philiastides 2006] Philiastides, M. G. and Sajda, P. (2006). Causal influences in the human brain during face discrimination: a short-window directed transfer function approach. *IEEE Transactions on Biomedical Engineering*, 53(12):2602–2605.

[Pijn 1990] Pijn, J. (1990). Quantitative evaluation of eeg signals in epilepsy. *University of Amsterdam, Amsterdam*.

[Ploner 2009] Ploner, M., Schoffelen, J. M., Schnitzler, A., and Gross, J. (2009). Functional integration within the human pain system as revealed by Granger causality. *Human brain mapping*, 30(12):4025–4032.

[Pompe 2011] Pompe, B. and Runge, J. (2011). Momentary information transfer as a coupling measure of time series. *Physical Review E*, 83(5):051122.

[Ponten 2007] Ponten, S., Bartolomei, F., and Stam, C. (2007). Small-world networks and epilepsy: graph theoretical analysis of intracerebrally recorded mesial temporal lobe seizures. *Clinical neurophysiology*, 118(4):918–927.

[RCH 2015] RCH (2015). Antiepileptic medications. http://www.rch.org.au/neurology/patient_information/antiepileptic_medications/.

[Roebroeck 2005] Roebroeck, A., Formisano, E., and Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fmri. *Neuroimage*, 25(1):230–242.

[Roebroeck 2011a] Roebroeck, A., Formisano, E., and Goebel, R. (2011a). The identification of interacting networks in the brain using fmri: model selection, causality and deconvolution. *Neuroimage*, 58(2):296–302.

[Roebroeck 2011b] Roebroeck, A., Seth, A. K., and Valdes-Sosa, P. (2011b). Causal time series analysis of functional magnetic resonance imaging data. *Causality in Time Series Challenges in Machine Learning, Volume 5*, page 35.

[Roelstraete 2012] Roelstraete, B. and Rosseel, Y. (2012). Does partial Granger causality really eliminate the influence of exogenous inputs and latent variables? *Journal of neuroscience methods*, 206(1):73–77.

[Roman 1974] Roman, P. (1974). *Some modern mathematics for physicists and other outsiders : an introduction to algebra, topology, and functional analysis*. New York : Pergamon Press.

[Roux 2013] Roux, F., Wibral, M., Singer, W., Aru, J., and Uhlhaas, P. J. (2013). The phase of thalamic alpha activity modulates cortical gamma-band activity: evidence from resting-state meg recordings. *The Journal of Neuroscience*, 33(45):17827–17835.

[Runge 2012] Runge, J., Heitzig, J., Petoukhov, V., and Kurths, J. (2012). Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Physical review letters*, 108(25):258701.

[Sabesan 2007] Sabesan, S., Narayanan, K., Prasad, A., Iasemidis, L., Spanias, A., and Tsakalis, K. (2007). Information flow in coupled nonlinear systems: Application to the epileptic human brain. In *Data Mining in Biomedicine*, pages 483–503.

[Sabesan 2009] Sabesan, S., Good, L. B., Tsakalis, K. S., Spanias, A., Treiman, D. M., and Iasemidis, L. D. (2009). Information flow and application to epileptogenic focus localization from intracranial eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 17(3):244–253.

[Sameshima 1999] Sameshima, K. and Baccalá, L. A. (1999). Using partial directed coherence to describe neuronal ensemble interactions. *Journal of neuroscience methods*, 94(1):93–103.

[Sameshima 2014] Sameshima, K. and Baccala, L. A. (2014). *Methods in Brain Connectivity Inference through Multivariate Time Series Analysis*. CRC press.

[Sato 2006] Sato, J. R., Junior, E. A., Takahashi, D. Y., de Maria Felix, M., Brammer, M. J., and Morettin, P. A. (2006). A method to produce evolving functional connectivity maps during the course of an fmri experiment using wavelet-based time-varying Granger causality. *Neuroimage*, 31(1):187–196.

[Sato 2009] Sato, J. R., Takahashi, D. Y., Arcuri, S. M., Sameshima, K., Morettin, P. A., and Baccalá, L. A. (2009). Frequency domain connectivity identification: an application of partial directed coherence in fmri. *Human brain mapping*, 30(2):452–461.

[Schelter 2006] Schelter, B., Winterhalder, M., Eichler, M., Peifer, M., Hellwig, B., Guschlbauer, B., Lücking, C. H., Dahlhaus, R., and Timmer, J. (2006). Testing for directed influences among neural signals using partial directed coherence. *Journal of neuroscience methods*, 152(1):210–219.

[Schelter 2009] Schelter, B., Timmer, J., and Eichler, M. (2009). Assessing the strength of directed influences among neural signals using renormalized partial directed coherence. *Journal of neuroscience methods*, 179(1):121–130.

[Schreiber 2000] Schreiber, T. (2000). Measuring information transfer. *Physical review letters*, 85(2):461.

[Schwarz 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Sergey 2015] Sergey, A., Peter, G., Alexander, K., and Harald, S. (2015). Milca toolbox. http://www.ucl.ac.uk/ion/departments/sobell/Research/RLemon/MILCA/MILCA/.

[Seth 2007] Seth, A. K. (2007). Granger causality. *Scholarpedia*, 2(7):1667.

[Seth 2010] Seth, A. K. (2010). A matlab toolbox for Granger causal connectivity analysis. *Journal of neuroscience methods*, 186(2):262–273.

[Seth 2013] Seth, A. K., Chorley, P., and Barnett, L. C. (2013). Granger causality analysis of fmri bold signals is invariant to hemodynamic convolution but not downsampling. *Neuroimage*, 65:540–555.

[Shannon 1948] Shannon, C. E. (1948). A mathematical theory of communities. *Techn. J*, 27:379–423.

[Shawe-Taylor 2004] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

[Singh 2003] Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A., and Demchuk, E. (2003). Nearest neighbor estimates of entropy. *American journal of mathematical and management sciences*, 23(3-4):301–321.

[Smirnov 2009] Smirnov, D. A. and Mokhov, I. I. (2009). From Granger causality to long-term causality: Application to climatic data. *Physical Review E*, 80(1):016208.

[Smirnov 2013] Smirnov, D. A. (2013). Spurious causalities with transfer entropy. *Physical Review E*, 87(4):042917.

[Smithson 2012] Smithson, W. H. and Walker, M. C. (2012). *ABC of Epilepsy*, volume 201. John Wiley & Sons.

[Sporns 2007] Sporns, O. (2007). Brain connectivity. *Scholarpedia*, 2(10):4695.

[Sporns 2005] Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS Comput Biol*, 1(4):e42.

[Sricharan 2013] Sricharan, K., Wei, D., and Hero III, A. O. (2013). Ensemble estimators for multivariate entropy estimation. *IEEE Transactions on Information Theory*, 59(7).

[Staniek 2008] Staniek, M. and Lehnertz, K. (2008). Symbolic transfer entropy. *Physical Review Letters*, 100(15):158101.

[Staniek 2009] Staniek, M. and Lehnertz, K. (2009). Symbolic transfer entropy: inferring directionality in biosignals. *Biomedizinische Technik/Biomedical Engineering*, 54(6):323–328.

[Stephan 2007] Stephan, K. E., Weiskopf, N., Drysdale, P. M., Robinson, P. A., and Friston, K. J. (2007). Comparing hemodynamic models with dcm. *Neuroimage*, 38(3):387–401.

[Stephan 2008] Stephan, K. E., Kasper, L., Harrison, L. M., Daunizeau, J., den Ouden, H. E., Breakspear, M., and Friston, K. J. (2008). Nonlinear dynamic causal models for fmri. *Neuroimage*, 42(2):649–662.

[Stögbauer 2004] Stögbauer, H., Kraskov, A., Astakhov, S. A., and Grassberger, P. (2004). Least-dependent-component analysis based on mutual information. *Physical Review E*, 70(6):066123.

[Stramaglia 2011] Stramaglia, S., Angelini, L., Pellicoro, M., and Marinazzo, D. (2011). Nonlinear Granger causality for brain connectivity. In *2011 IEEE International Workshop on Medical Measurements and Applications Proceedings (MeMeA)*, pages 197–201.

[Stramaglia 2012] Stramaglia, S., Wu, G. R., Pellicoro, M., and Marinazzo, D. (2012). Expanding the transfer entropy to identify information circuits in complex systems. *Physical Review E*, 86(6):066211.

[Sun 2014] Sun, J. and Bollt, E. M. (2014). Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D: Nonlinear Phenomena*, 267:49–57.

[Sun 2008] Sun, X. (2008). Assessing nonlinear Granger causality from multivariate time series. In *Machine Learning and Knowledge Discovery in Databases*, pages 440–455.

[Sun 2009] Sun, Y., Zhang, H., Feng, T., Qiu, Y., Zhu, Y., and Tong, S. (2009). Early cortical connective network relating to audiovisual stimulation by partial directed coherence analysis. *IEEE Transactions on Biomedical Engineering*, 56(11):2721–2724.

[Suzuki 2008] Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation. *FSDM*, 4:5–20.

[Tiwari 2014] Tiwari, A. K. (2014). The asymmetric Granger-causality analysis between energy consumption and income in the united states. *Renewable and Sustainable Energy Reviews*, 36:362–369.

[Triacca 2001] Triacca, U. (2001). On the use of Granger causality to investigate the human influence on climate. *Theoretical and Applied Climatology*, 69(3-4):137–138.

[Urbanczik 2003] Urbanczik, R. (2003). Learning curves for mutual information maximization. *Physical Review E*, 68(1):016106.

[Vakorin 2010] Vakorin, V. A., Kovacevic, N., and McIntosh, A. R. (2010). Exploring transient transfer entropy based on a group-wise ica decomposition of eeg data. *Neuroimage*, 49(2):1593–1600.

[Vakorin 2009] Vakorin, V. A., Krakovska, O. A., and McIntosh, A. R. (2009). Confounding effects of indirect connections on causality estimation. *Journal of neuroscience methods*, 184(1):152–160.

[Vakorin 2011] Vakorin, V. A., Mišić, B., Krakovska, O., and McIntosh, A. R. (2011). Empirical and theoretical aspects of generation and transfer of information in a neuromagnetic source network. *Frontiers in systems neuroscience*, 5.

[Valdes-Sosa 2011] Valdes-Sosa, P. A., Roebroeck, A., Daunizeau, J., and Friston, K. (2011). Effective connectivity: influence, causality and biophysical modeling. *Neuroimage*, 58(2):339–361.

[Van Hulle 2005] Van Hulle, M. M. (2005). Edgeworth approximation of multivariate differential entropy. *Neural computation*, 17(9):1903–1910.

[Varotto 2012] Varotto, G., Tassi, L., Franceschetti, S., Spreafico, R., and Panzica, F. (2012). Epileptogenic networks of type ii focal cortical dysplasia: a stereo-eeg study. *Neuroimage*, 61(3):591–598.

[Venelli 2010] Venelli, A. (2010). Efficient entropy estimation for mutual information analysis using b-splines. In *Information Security Theory and Practices. Security and Privacy of Pervasive Systems and Smart Devices*, pages 17–30.

[Ver Steeg 2012] Ver Steeg, G. and Galstyan, A. (2012). Information transfer in social media. In *Proceedings of the 21st international conference on World Wide Web*, pages 509–518.

[Verleysen 2005] Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *Computational Intelligence and Bioinspired Systems*, pages 758–770.

[Vicente 2011] Vicente, R., Wibral, M., Lindner, M., and Pipa, G. (2011). Transfer entropy–a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67.

[Wang 2015] Wang, G., Sun, Z., Tao, R., Li, K., Bao, G., and Yan, X. (2015). Epileptic seizure detection based on partial directed coherence analysis. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–1.

[Wang 2009] Wang, Q., Kulkarni, S. R., and Verdú, S. (2009). Divergence estimation for multidimensional densities via-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405.

[Wang 2002] Wang, X. F. and Chen, G. (2002). Synchronization in small-world dynamical networks. *International Journal of Bifurcation and Chaos*, 12(01):187–192.

[Wendling 2005] Wendling, F., Hernandez, A., Bellanger, J. J., Chauvel, P., and Bartolomei, F. (2005). Interictal to ictal transition in human temporal lobe epilepsy: insights from a computational model of intracerebral eeg. *Journal of Clinical Neurophysiology*, 22(5):343.

[Wendling 2000] Wendling, F., Bellanger, J. J., Bartolomei, F., and Chauvel, P. (2000). Relevance of nonlinear lumped-parameter models in the analysis of depth-eeg epileptic signals. *Biological cybernetics*, 83(4):367–378.

[Wendling 2001] Wendling, F., Bartolomei, F., Bellanger, J. J., and Chauvel, P. (2001). Interpretation of interdependencies in epileptic signals using a macroscopic physiological model of the eeg. *Clinical neurophysiology*, 112(7):1201–1218.

[WHO 2005] WHO (2005). Atlas: epilepsy care in the world. http://www.who.int/mental_health/neurology/Epilepsy_atlas_r1.pdf/.

[Wibral 2011a] Wibral, M., Rahm, B., Rieder, M., Lindner, M., Vicente, R., and Kaiser, J. (2011a). Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks. *Progress in biophysics and molecular biology*, 105(1):80–97.

[Wibral 2011b] Wibral, M., Vicente, R., Priesemann, V., and Lindner, M. (2011b). Trentool: an open source toolbox to estimate neural directed interactions with transfer entropy. *BMC Neuroscience*, 12:200.

[Wibral 2013] Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Lindner, M., Lizier, J. T., and Vicente, R. (2013). Measuring information-transfer delays. *PloS one*, 8(2):e55809.

[Wibral 2014a] Wibral, M., Vicente, R., and Lindner, M. (2014a). Transfer entropy in neuroscience. In *Directed Information Measures in Neuroscience*, pages 3–36.

[Wibral 2014b] Wibral, M., Vicente, R., and Lizier, J. T. (2014b). *Directed information measures in neuroscience*. Springer.

[Wiener 1956] Wiener, N. (1956). The theory of prediction. *Modern mathematics for engineers*, 1:125–139.

[Wilke 2007] Wilke, C., Ding, L., and He, B. (2007). An adaptive directed transfer function approach for detecting dynamic causal interactions. In *29th Annual International IEEE Conference of Engineering in Medicine and Biology Society (EMBS)*, pages 4949–4952.

[Wilke 2008] Wilke, C., Ding, L., and He, B. (2008). Estimation of time-varying connectivity patterns through the use of an adaptive directed transfer function. *IEEE Transactions on Biomedical Engineering*, 55(11):2557–2564.

[Wilke 2010] Wilke, C., Van Drongelen, W., Kohrman, M., and He, B. (2010). Neocortical seizure foci localization by means of a directed transfer function method. *Epilepsia*, 51(4):564–572.

[Wissman 2011] Wissman, B., McKay-Jones, L., and Binder, P.-M. (2011). Entropy rate estimates from mutual information. *Physical Review E*, 84(4):046204.

[Wollstadt 2014] Wollstadt, P., Martínez Zarzuela, M., Vicente, R., Díaz Pernas, F. J., and Wibral, M. (2014). Efficient transfer entropy analysis of non-stationary neural time series. *PloS one*, 9(7):e102833.

[Wollstadt 2015] Wollstadt, P., Lindner, M., Vicente, R., Wibral, M., Pampu, N., and Martinez Zarzuela, M. (2015). Trentool toolbox. www.trentool.de/.

[Wu 2012] Wu, X., Wang, W., and Zheng, W. X. (2012). Inferring topologies of complex networks with hidden variables. *Physical Review E*, 86(4):046106.

[Xu 2014] Xu, H., Lu, Y., Zhu, S., and He, B. (2014). Assessing dynamic spectral causality by lagged adaptive directed transfer function and instantaneous effect factor. *IEEE Transactions on Biomedical Engineering*, 61(7):1979–1988.

[Yang 2012] Yang, C. (2012). *Contribution à l'analyse de connectivité effective en épilepsie*. PhD thesis, Université Rennes 1.

[Yang 2013] Yang, C., Le Bouquin Jeannes, R., Bellanger, J. J., and Shu, H. (2013). A new strategy for model order identification and its application to transfer entropy for eeg signals analysis. *IEEE Transactions on Biomedical Engineering*, 60(5):1318–1327.

[Yuan 2014] Yuan, T. and Qin, S. J. (2014). Root cause diagnosis of plant-wide oscillations using Granger causality. *Journal of Process Control*, 24(2):450–459.

[Zhang 2015] Zhang, C. H., Sha, Z., Mundahl, J., Liu, S., Lu, Y., Henry, T. R., and He, B. (2015). Thalamocortical relationship in epileptic patients with generalized spike and wave discharges–a multimodal neuroimaging study. *NeuroImage: Clinical*, 9:117–127.

[Zhao 2013] Zhao, Y., Billings, S. A., Wei, H., He, F., and Sarrigiannis, P. G. (2013). A new narx-based Granger linear and nonlinear casual influence detection method with applications to eeg data. *Journal of neuroscience methods*, 212(1):79–86.

[Zhou 2009] Zhou, Z., Chen, Y., Ding, M., Wright, P., Lu, Z., and Liu, Y. (2009). Analyzing brain networks with pca and conditional Granger causality. *Human brain mapping*, 30(7):2197–2206.

[Zhu 2014] Zhu, J., Bellanger, J. J., Shu, H., Yang, C., and Le Bouquin Jeannès, R. (2014). Bias reduction in the estimation of mutual information. *Physical Review E*, 90(5):052714.

[Zhu 2015a] Zhu, J., Bellanger, J. J., Shu, H., and Le Bouquin Jeannès, R. (2015a). Contribution to transfer entropy estimation via the k-nearest-neighbors approach. *Entropy*, 17(6):4173–4201.

[Zhu 2015b] Zhu, J., Bellanger, J. J., Shu, H., and Le Bouquin Jeannès, R. (2015b). Investigating bias in non-parametric mutual information estimation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3971–3975.

[Zuo 2013] Zuo, K., Zhu, J., Bellanger, J. J., and Le Bouquin Jeannès, R. (2013). Adaptive kernels and transfer entropy for neural connectivity analysis in eeg signals. *IRBM*, 34(4):330–336.

VU :

Le Directeur de Thèse
(Nom et Prénom)

VU :

Le Responsable de l'École Doctorale

VU pour autorisation de soutenance

Rennes, le

Le Président de l'Université de Rennes 1

David ALIS

VU après soutenance pour autorisation de publication :

Le Président de Jury,
(Nom et Prénom)