

UNIVERSITÉ CLAUDE BERNARD - LYON 1  
FACULTÉ DE PHARMACIE  
INSTITUT DES SCIENCES PHARMACEUTIQUES ET BIOLOGIQUES

2015

THÈSE n° 55

T H E S E

pour le DIPLÔME D'ÉTAT DE DOCTEUR EN PHARMACIE  
présentée et soutenue publiquement le 19 juin 2015

par

M. MARTIN Olivier

Né le 12 avril 1991  
à Clermont-Ferrand

\*\*\*\*\*

DÉVELOPPEMENT D'UN LOGICIEL IMPLÉMENTANT  
UNE MÉTHODE D'ANALYSE DE DONNÉES POUR LES  
SCIENCES DE LA VIE

\*\*\*\*\*

JURY

Mme MOYRET-LALLE Caroline, Pharm.D., Ph.D., MCU, HDR

M. GOUTELLE Sylvain, Pharm.D., Ph.D., MCU

M. BOURGUIGNON Laurent, Pharm.D., Ph. D.

M. DUCHER Michel, Pharm.D., Ph.D, HDR

## UNIVERSITE CLAUDE BERNARD LYON 1

- Président de l'Université M. François-Noël GILLY
- Vice-Président du Conseil d'Administration M. Hamda BEN HADID
- Vice-Président du Conseil Scientifique M. Germain GILLET
- Vice-Président du Conseil des Etudes et de la Vie Universitaire M. Philippe LALLE

### **Composantes de l'Université Claude Bernard Lyon 1**

#### **SANTE**

- UFR de Médecine Lyon Est Directeur : M. Jérôme ETIENNE
- UFR de Médecine Lyon Sud Charles Mérieux Directeur : Mme Carole BURILLON
- Institut des Sciences Pharmaceutiques et Biologiques Directrice : Mme Christine VINCIGUERRA
- UFR d'Odontologie Directeur : M. Denis BOURGEOIS
- Institut des Techniques de Réadaptation Directeur : M. Yves MATILLON
- Département de formation et centre de recherche en Biologie Humaine Directeur : Anne-Marie SCHOTT

#### **SCIENCES ET TECHNOLOGIES**

- Faculté des Sciences et Technologies Directeur : M. Fabien DE MARCHI
- UFR de Sciences et Techniques des Activités Physiques et Sportives (STAPS) Directeur : M. Yannick VANPOULLE
- Ecole Polytechnique Universitaire de Lyon (ex ISTIL) Directeur : M. Pascal FOURNIER
- I.U.T. LYON 1 Directeur : M. Christophe VITON
- Institut des Sciences Financières et d'Assurance (ISFA) Directeur : M. Nicolas LEBOISNE
- ESPE Directeur : M. Alain MOUGNIOTTE

**UNIVERSITE CLAUDE BERNARD LYON 1  
ISPB -Faculté de Pharmacie Lyon**

**LISTE DES DEPARTEMENTS PEDAGOGIQUES**

**DEPARTEMENT PEDAGOGIQUE DE SCIENCES PHYSICO-CHIMIQUE ET PHARMACIE GALENIQUE**

**• CHIMIE ANALYTIQUE, GENERALE, PHYSIQUE ET MINERALE**

Monsieur Raphaël TERREUX (Pr)  
Monsieur Pierre TOULHOAT (Pr - PAST)  
Madame Julie-Anne CHEMELLE (MCU)  
Monsieur Lars-Petter JORDHEIM (MCU)  
Madame Christelle MACHON (AHU)

**• PHARMACIE GALENIQUE -COSMETOLOGIE**

Madame Marie-Alexandrine BOLZINGER (Pr)  
Madame Stéphanie BRIANCON (Pr)  
Madame Françoise FALSON (Pr)  
Monsieur Hatem FESSI (Pr)  
Monsieur Fabrice PIROT (PU - PH)  
Monsieur Eyad AL MOUAZEN (MCU)  
Madame Sandrine BOURGEOIS (MCU)  
Madame Ghania HAMDI-DEGOBERT (MCU-HDR)  
Monsieur Plamen KIRILOV (MCU)  
Monsieur Damien SALMON (AHU)

**• BIOPHYSIQUE**

Monsieur Richard COHEN (PU – PH)  
Madame Laurence HEINRICH (MCU)  
Monsieur David KRYZA (MCU – PH)  
Madame Sophie LANCELOT (MCU - PH)  
Monsieur Cyril PAILLER-MATTEI (MCU-HDR)  
Madame Elise LEVIGOUREUX (AHU)

**DEPARTEMENT PEDAGOGIQUE PHARMACEUTIQUE DE SANTE PUBLIQUE**

**• DROIT DE LA SANTE**

Monsieur François LOCHER (PU – PH)  
Madame Valérie SIRANYAN (MCU - HDR)

**• ECONOMIE DE LA SANTE**

Madame Nora FERDJAOUUI MOUMJID (MCU - HDR)  
Madame Carole SIANI (MCU – HDR)  
Monsieur Hans-Martin SPÄTH (MCU)

**• INFORMATION ET DOCUMENTATION**

Monsieur Pascal BADOR (MCU - HDR)

**• HYGIENE, NUTRITION, HYDROLOGIE ET ENVIRONNEMENT**

Madame Joëlle GOUDABLE (PU – PH)

**• INGENIERIE APPLIQUEE A LA SANTE ET DISPOSITIFS MEDICAUX**

Monsieur Gilles AULAGNER (PU – PH)  
Monsieur Daniel HARTMANN (Pr)

**• QUALITOLOGIE – MANAGEMENT DE LA QUALITE**

Madame Alexandra CLAYER-MONTEMBAULT (MCU)  
Monsieur Vincent GROS (MCU-PAST)  
Madame Audrey JANOLY-DUMENIL (MCU-PH)  
Madame Pascale PREYNAT (MCU PAST)

- **MATHEMATIQUES – STATISTIQUES**  
Madame Claire BARDEL-DANJEAN (MCU)  
Madame Marie-Aimée DRONNE (MCU)  
Madame Marie-Paule PAULTRE (MCU - HDR)

## **DEPARTEMENT PEDAGOGIQUE SCIENCES DU MEDICAMENT**

- **CHIMIE ORGANIQUE**  
Monsieur Pascal NEBOIS (Pr)  
Madame Nadia WALCHSHOFER (Pr)  
Monsieur Zouhair BOUAZIZ (MCU - HDR)  
Madame Christelle MARMINON (MCU)  
Madame Sylvie RADIX (MCU - HDR)  
Monsieur Luc ROCHEBLAVE (MCU - HDR)
- **CHIMIE THERAPEUTIQUE**  
Monsieur Roland BARRET (Pr)  
Monsieur Marc LEBORGNE (Pr)  
Monsieur Laurent ETTOUATI (MCU - HDR)  
Monsieur Thierry LOMBERGET (MCU - HDR)  
Madame Marie-Emmanuelle MILLION (MCU)
- **BOTANIQUE ET PHARMACOGNOSIE**  
Madame Marie-Geneviève DIJOUX-FRANCA (Pr)  
Madame Marie-Emmanuelle HAY DE BETTIGNIES (MCU)  
Madame Isabelle KERZAON (MCU)  
Monsieur Serge MICHALET (MCU)
- **PHARMACIE CLINIQUE, PHARMACOCINETIQUE ET EVALUATION DU MEDICAMENT**  
Madame Roselyne BOULIEU (PU – PH)  
Madame Magali BOLON-LARGER (MCU - PH)  
Madame Christelle CHAUDRAY-MOUCHOUX (MCU-PH)  
Madame Céline PRUNET-SPANO (MCU)  
Madame Catherine RIOUFOL (MCU- PH-HDR)

## **DEPARTEMENT PEDAGOGIQUE DE PHARMACOLOGIE, PHYSIOLOGIE ET TOXICOLOGIE**

- **TOXICOLOGIE**  
Monsieur Jérôme GUITTON (PU – PH)  
Madame Léa PAYEN (PU-PH)  
Monsieur Bruno FOUILLET (MCU)  
Monsieur Sylvain GOUTELLE (MCU-PH)
- **PHYSIOLOGIE**  
Monsieur Christian BARRES (Pr)  
Monsieur Daniel BENZONI (Pr)  
Madame Kiao Ling LIU (MCU)  
Monsieur Ming LO (MCU - HDR)
- **PHARMACOLOGIE**  
Monsieur Michel TOD (PU – PH)  
Monsieur Luc ZIMMER (PU – PH)  
Monsieur Roger BESANCON (MCU)  
Madame Evelyne CHANUT (MCU)  
Monsieur Nicola KUCZEWSKI (MCU)  
Monsieur Olivier CATALA (Pr-PAST)  
Madame Corinne FEUTRIER (MCU-PAST)  
Madame Mélanie THUDEROZ (MCU-PAST)

## **DEPARTEMENT PEDAGOGIQUE DES SCIENCES BIOMEDICALES A**

- IMMUNOLOGIE**

Monsieur Jacques BIENVENU (PU – PH)  
Monsieur Guillaume MONNERET (PU-PH)  
Madame Cécile BALTER-VEYSSEYRE (MCU - HDR)  
Monsieur Sébastien VIEL (AHU)

- HEMATOLOGIE ET CYTOLOGIE**

Madame Christine TROUILLOT-VINCIGUERRA (PU - PH)  
Madame Brigitte DURAND (MCU - PH)  
Monsieur Olivier ROUALDES (AHU)

- MICROBIOLOGIE ET MYCOLOGIE FONDAMENTALE ET APPLIQUEE AUX BIOTECHNOLOGIE INDUSTRIELLES**

Monsieur Patrick BOIRON (Pr)  
Monsieur Jean FRENEY (PU – PH)  
Madame Florence MORFIN (PU – PH)  
Monsieur Didier BLAHA (MCU)  
Madame Ghislaine DESCOURS (MCU-PH)  
Madame Anne DOLEANS JORDHEIM (MCU-PH)  
Madame Emilie FROBERT (MCU - PH)  
Madame Véronica RODRIGUEZ-NAVA (MCU-HDR)

- PARASITOLOGIE, MYCOLOGIE MEDICALE**

Monsieur Philippe LAWTON (Pr)  
Madame Nathalie ALLIOLI (MCU)  
Madame Samira AZZOUZ-MAACHE (MCU - HDR)

## **DEPARTEMENT PEDAGOGIQUE DES SCIENCES BIOMEDICALES B**

- BIOCHIMIE – BIOLOGIE MOLECULAIRE - BIOTECHNOLOGIE**

Madame Pascale COHEN (Pr)  
Monsieur Alain PUISIEUX (PU - PH)  
Monsieur Karim CHIKH (MCU - PH)  
Madame Carole FERRARO-PEYRET (MCU - PH-HDR)  
Monsieur Boyan GRIGOROV (MCU)  
Monsieur Hubert LINCET (MCU-HDR)  
Monsieur Olivier MEURETTE (MCU)  
Madame Caroline MOYRET-LALLE (MCU – HDR)  
Madame Angélique MULARONI (MCU)  
Madame Stéphanie SENTIS (MCU)  
Monsieur Anthony FOURIER (AHU)

- BIOLOGIE CELLULAIRE**

Madame Bénédicte COUPAT-GOUTALAND (MCU)  
Monsieur Michel PELANDAKIS (MCU - HDR)

- INSTITUT DE PHARMACIE INDUSTRIELLE DE LYON**

Madame Marie-Alexandrine BOLZINGER (Pr)  
Monsieur Daniel HARTMANN (Pr)  
Monsieur Philippe LAWTON (Pr)  
Madame Sandrine BOURGEOIS (MCU)  
Madame Marie-Emmanuelle MILLION (MCU)  
Madame Alexandra MONTEMBIAULT (MCU)  
Madame Angélique MULARONI (MCU)  
Madame Valérie VOIRON (MCU - PAST)

- **Assistants hospitalo-universitaires sur plusieurs départements pédagogiques**

Madame Emilie BLOND  
Madame Florence RANCHON

- **Attachés Temporaires d'Enseignement et de Recherche (ATER)**

Madame Sophie ASSANT 85<sup>ème</sup> section  
Monsieur Benoit BESTGEN 85<sup>ème</sup> section  
Madame Marine CROZE 86<sup>ème</sup> section  
Madame Mylène HONORAT MEYER 85<sup>ème</sup> section

**Pr** : Professeur

**PU-PH** : Professeur des Universités, Praticien Hospitalier

**MCU** : Maître de Conférences des Universités

**MCU-PH** : Maître de Conférences des Universités, Praticien Hospitalier

**HDR** : Habilitation à Diriger des Recherches

**AHU** : Assistant Hospitalier Universitaire

**PAST** : Personnel Associé Temps Partiel

## Remerciements

En premier lieu, je souhaite remercier Michel Ducher pour toutes les heures passées à m'encadrer ainsi que toutes les discussions que nous avons eues. Je suis particulièrement reconnaissant envers Pascal Maire pour son accueil et pour avoir fondé une pharmacie hospitalière axée sur la recherche et l'esprit critique. J'adresse également ma gratitude envers Laurent Bourguinon et Sylvain Goutelle pour leur pédagogie, accessibilité et compétence. Je dois également remercier Caroline Moyret-Lalle de m'avoir orienté au cours de ces deux dernières années et d'avoir accepté de juger ce travail. Ce travail n'aurait pas été possible sans la bonne humeur de toute l'équipe des pharmacies hospitalières Antoine Charial et Pierre Garraud. Je souhaiterais également remercier ma famille pour tout l'amour et la culture qu'ils m'ont transmis. Je souhaite finalement remercier l'ensemble de mes amis.

# Table des matières

<b>I. Introduction .....</b>	<b>7</b>
<b>II. Étude bibliographique .....</b>	<b>10</b>
<b>1. Modélisation mathématique du vivant .....</b>	<b>10</b>
1.1. Définition d'un modèle mathématique .....	11
1.2. Objectifs de la modélisation : prédire et expliquer .....	12
1.2.1. Modèles explicatifs.....	12
1.2.2. Modèle prédictif.....	12
1.2.3. Indépendance entre explication et prédiction .....	13
1.3. Classification mathématique des modèles .....	13
1.3.1. Linéaire vs non-linéaire.....	13
1.3.2. Continu vs discret .....	14
1.3.3. Statique vs dynamique .....	14
1.3.4. Déterministe vs stochastique.....	14
1.3.5. Paramétrique vs non-paramétrique .....	15
1.4. Intérêts scientifiques.....	16
1.4.1. Tester ou générer des hypothèses .....	16
1.4.2. Intervention .....	16
1.4.3. Simulation .....	17
1.4.4. Prévisions .....	17
1.4.5. Classification.....	17
1.4.6. Analyse d'un grand nombre de données.....	18
1.5. Les difficultés de la mathématisation du vivant.....	18
1.6. Exemple de modèle : association statistique.....	20
1.6.1. Causalité et association .....	20
1.6.2. Généralités sur l'association statistique .....	21
1.6.3. Mesures basées sur la variance : coefficient de corrélation de Pearson .....	22
1.6.4. Mesures basées sur les probabilités : chi-deux .....	26
1.6.5. Différence entre mesure globale et locale : Z de Ducher .....	29

<b>2. Outils informatiques : exemple du Zébu .....</b>	<b>33</b>
2.1. Essor des approches computationnelles – bioinformatique.....	34
2.2. Algorithmes et langages de programmation .....	35
2.2.1. Algorithmes : exemple de comparaison de séquences .....	35
2.2.2. Langages de bas-niveau et haut-niveau.....	36
2.2.3. Intérêt de la programmation fonctionnelle .....	38
2.2.4. Choix du langage de programmation .....	39
2.2.5. Interfaces utilisateurs : interfaces graphiques et lignes de commandes.....	39
2.3. Licences logiciels .....	41
2.3.1. Définition et classification.....	41
2.3.2. Intérêt de l'open source en sciences de la vie .....	41
<b>3. Applications des mesures d'association locales.....</b>	<b>43</b>
3.1. Physiologie cardiovasculaire.....	44
3.2. Pharmacocinétique .....	45
3.3. Linguistique computationnelle .....	45
3.4. Analyse d'image .....	45
3.5. Géographie et analyse spatiale .....	46
<b>III. Travail personnel.....</b>	<b>47</b>
1. Conception d'un outil informatique implémentant le Z de Ducher et l'information mutuelle spécifique : Zébu.....	47
2. Exemples d'utilisation du Zébu.....	48
2.1. Relation continue non-monotone : exemple de l'hormèse .....	48
2.2. Variables catégorielles : exemple du tabac .....	50
2.3. Relation trivariée : exemple des études d'association pangénomique.....	52
<b>IV. Discussion .....</b>	<b>55</b>
<b>V. Conclusion.....</b>	<b>57</b>
<b>VI. Glossaire .....</b>	<b>59</b>
<b>VII. Références .....</b>	<b>62</b>
<b>VIII. Annexes : scripts R .....</b>	<b>68</b>
1. Variance et covariance.....	68
2. Exemples d'utilisations du logiciel Zébu .....	70

# Liste d'équations

Équation 1. Coefficient de corrélation de Pearson .....	22
Équation 2. Variance d'une variable aléatoire .....	22
Équation 3. Covariance de deux variables aléatoires .....	23
Équation 4. Indépendance statistique entre deux variables aléatoires.....	27
Équation 5. Résiduels du chi-deux.....	28
Équation 6. Le chi-deux est la somme des résiduels .....	28
Équation 7. Déviation de l'indépendance .....	30
Équation 8. Z de Ducher.....	30
Équation 9. Z global de Ducher .....	31
Équation 10. Équation polynomiale utilisée pour simuler une relation hormétique entre dose et réponse .....	49

# Liste des figures

Figure 1. Interactions des sciences de la vie avec les mathématiques et l'informatique	8
Figure 2. Variance d'une variable aléatoire .....	24
Figure 3. Interprétation géométrique de la covariance .....	25
Figure 4. Deux séquences nucléiques homologues .....	35
Figure 5. Algorithme pseudocodé pour calculer le nombre de différences entre deux séquences de taille identique.....	36
Figure 6. Algorithme en Python pour calculer le nombre de différences entre deux séquences de taille identique.....	37
Figure 7. Interface en ligne de commande utilisant le langage bash pour contrôler une machine tournant sur Mac OS X.....	40
Figure 8. Nuage de points : relation hormétique entre dose et réponse.....	49
Figure 9. Z de Ducher : relation hormétique entre dose et réponse.....	49
Figure 10. Tableau de contingence : relation entre « fumer » et « faire un infarctus du myocarde ».....	50
Figure 11. Z de Ducher : relation entre « fumer » et « faire un infarctus du myocarde » .....	51
Figure 12. Analyse en sous-groupe basé sur la dépendance : distribution de l'âge en fonction de l'association fumeur - infarctus.....	52
Figure 13 : Z de Ducher bivariée : circuit logique XOR.....	54

## Liste des tableaux

Tableau 1. Tableau de contingence pour deux variables binaires .....	26
Tableau 2. Quatre libertés fondamentales du logiciel libre.....	41
Tableau 3. Tables de vérité des circuits logiques communs .....	53
Tableau 4. Z de Ducher trivariée : circuit logique XOR .....	54

# I. Introduction

Nous connaissons aujourd’hui, notamment dans les sciences de la vie, une explosion du volume de données disponibles. En effet, les progrès techniques nous permettent de générer rapidement une quantité importante de données et de les stocker à moindre prix. Afin d’illustrer ces propos, prenons quelques exemples historiques. Alors que le *Human Genome Project* coûta presque 3 milliards de dollars et dura 13 ans, le séquençage haut débit d’un génome humain coûte aujourd’hui moins de 1000 dollars et prend quelques jours (1). De plus, alors qu’en 1957, l’IBM 350 stockait 3,75 mégaoctets, pesait plus d’une tonne et coûtait plus d’un million de dollars actuels (2), aujourd’hui, nous pouvons acheter un petit disque dur d’un téraoctet pour moins de 100 dollars.

Nous nous retrouvons ainsi confrontés à des bases de données de plus en plus grandes. Afin d’en extraire la connaissance, les méthodes et les outils issus des mathématiques et de l’informatique sont d’un grand intérêt. Leur utilisation dans les sciences de la vie a une longue histoire. On pourrait notamment citer la modélisation épidémiologique de la variole faite par Bernoulli au XII<sup>e</sup> siècle ou le perceptron, modèle neuronal issu du *machine learning* (3). La complexité du vivant a souvent été avancée comme argument contre ces approches. Pourtant, celles-ci permettent de simplifier considérablement des problèmes complexes en se concentrant sur l’essentiel et en évitant de se noyer dans un déluge d’information. Ces approches deviennent donc de plus en plus populaires et nous entendons aujourd’hui fréquemment parler de biomathématiques, de biostatistiques et de bioinformatiques. Ces disciplines, situées à la frontière, sont intimement liées à la pratique clinique et aux sciences expérimentales (Figure 1). Les modèles qu’elles produisent peuvent être considérés comme des représentations abstraites de nos connaissances actuelles. La confrontation de ces modèles face au réel permet alors de mieux cibler les lacunes de notre connaissance et ainsi d’y remédier. Les nouvelles connaissances vont alors modifier nos pratiques cliniques et orienter notre recherche expérimentale. Celles-ci vont alors générer de nouvelles données qui pourront être intégrées dans de nouveaux modèles fermant ainsi la boucle.

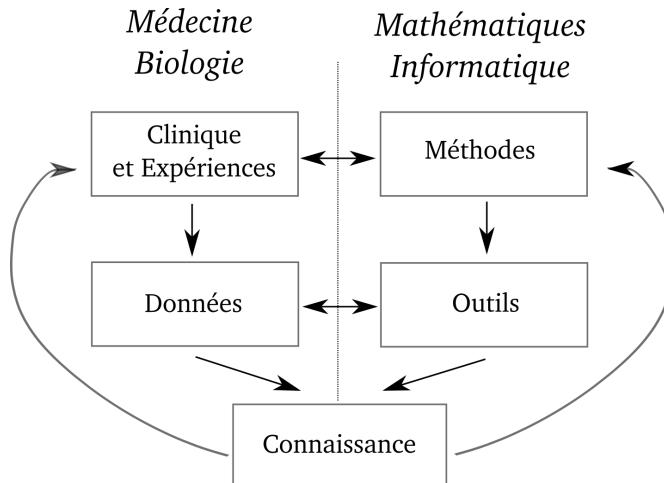


Figure 1. Interactions des sciences de la vie avec les mathématiques et l'informatique

Une difficulté de modélisation caractéristique des sciences de la vie est l'absence d'un déterminisme fort comme l'entend Laplace (4). En effet, les processus biologiques sont souvent intrinsèquement stochastiques (5,6). La prédictibilité de ces systèmes est alors uniquement statistique. Nous parlons donc d'association (ou corrélation) entre les variables. Les mesures d'association, comme le coefficient de corrélation de Pearson, sont là pour apporter une métrique à l'association. Ces mesures peuvent être globales ou locales. Les mesures globales, comme le coefficient de corrélation de Pearson, supposent que la force de l'association est identique pour toutes les modalités des variables. Cette supposition n'est pas systématiquement justifiée en sciences de la vie où l'on observe des relations discontinues comme des effets de seuil ou de saturation. Une alternative est alors d'utiliser des mesures locales comme le Z de Ducher (7). Celles-ci quantifient l'association localement pour les modalités des variables. Elles permettent une meilleure description de l'association entre les phénomènes en évitant d'homogénéiser, et donc de perdre, de l'information.

Aucun logiciel ne permet actuellement le calcul des mesures d'association locale. Or, l'utilisation de nouvelles méthodes a tendance à être retardée par le manque d'outils disponibles (8). L'objectif principal de cette thèse a donc été de développer un logiciel implémentant leur calcul. Celui-ci a été conçu pour être intuitif et directement disponible sur internet.

Ce travail a été divisé en deux parties. Premièrement, nous décrirons plus en détail la place et l'intérêt des méthodes et outils mathématiques et informatiques dans les sciences de la vie. Cette partie n'est pas destinée aux spécialistes, mais au contraire, cherche à introduire le non-initié aux concepts et à donner des exemples d'applications. Nous expliquerons également au cours de cette partie les différentes mesures d'associations. Deuxièmement, nous décrirons le logiciel développé : le Zébu. Celui-ci sera présenté par l'intermédiaire d'un article scientifique. Cette partie sera donc beaucoup plus technique, mais devrait être rendue compréhensible par la lecture de la première partie. Nous invitons également le lecteur à essayer le logiciel en se connectant sur <http://olivmrtn.shinyapps.io/Zebu>. Le lecteur pourra également consulter le code source sur <https://github.com/olivmrtn/Zebu>. Finalement, nous clôturons cette deuxième partie par des exemples d'utilisation montrant l'intérêt du logiciel. Le lecteur trouvera un glossaire en fin de thèse.

## II. Étude bibliographique

### 1. Modélisation mathématique du vivant

*Un homme cherche ses clefs au pied d'un réverbère. Vous les avez perdues ici? Non, je les ai perdues là-bas dans le noir, mais ici, au moins il y a de la lumière.*

Au premier abord, nous pourrions penser que l'homme au pied du réverbère est un idiot. Cependant, qui défendrait celui qui cherche dans l'obscurité ? Dans cette courte histoire, la lumière est utilisée comme une métaphore pour les moyens que nous avons afin d'accéder à la connaissance. De fait, nous ne pouvons pas chercher la clef de nos problèmes sans lumière. Nous sommes limités par les moyens disponibles au moment de notre recherche. Cependant, en nous rendant compte d'une limite de nos moyens, nous pouvions y remédier en les développant. C'est l'exemple de Newton qui dû inventer le calcul infinitésimal afin de formaliser la théorie de la gravité universelle (9).

La science peut être définie comme un moyen pour accéder à une connaissance sur le monde qui nous entoure. Pour cela, elle doit se fonder sur des observations sur celui-ci. Une discipline scientifique est ainsi toujours uniquement descriptive à sa naissance. C'est l'exemple de Linné et sa classification du vivant aux débuts de la biologie. Cependant, le réel est plus qu'un catalogue d'observations sans liens entre elles. Ces observations et leurs liens peuvent être expliqués. On passe alors d'une science descriptive à une science explicative. Celle-ci impose de synthétiser l'ensemble de nos observations par l'intermédiaire de modèles abstraits qui, à terme, seront incorporés dans une théorie plus générale. Par exemple, la diversité des espèces observée par Linné peut être expliquée par la théorie de l'évolution telle que formulée par Darwin (10). Cette théorie est composée de modèles verbaux et est donc qualitative (11). Elle ne fait pas appel à des modèles quantitatifs basés sur des mathématiques. Effectivement, une science explicative est d'abord qualitative avant de devenir quantitative. La physique a fait ce bond au XV<sup>e</sup> siècle avec Galilée. Nous défendrons l'idée au cours de cette partie que la biologie suivra une évolution

semblable. Il existe néanmoins encore des réticences de par des cliniciens et biologistes à employer des approches mathématiques. Ces réticences sont le plus souvent justifiées par la complexité et stochasticité du vivant que les mathématiques, trop rigides, ne pourraient saisir. Nous nous opposerons à cette conception en soutenant que le vivant n'est pas intrinsèquement non mathématisable, mais que les méthodes mathématiques utilisées ne sont pas toujours adaptées à saisir le vivant (12,13).

Nous commencerons par définir et classifier les modèles mathématiques avant de décrire leurs objectifs et défendre leurs intérêts. Nous pourrons alors discuter des réticences à la mathématisation du vivant et argumenterons qu'elles ne sont pas bien fondées. Nous prendrons un exemple concernant les mesures d'association. Pour cela, nous rappellerons d'abord au lecteur les méthodes classiques pour mesurer une association statistique. Nous montrerons ensuite que ces méthodes ne saisissent pas la complexité du concept d'association et ne sont pas toujours adaptées à comprendre le vivant. Nous décrirons finalement une mesure d'association conçue pour l'étude du vivant : le Z de Ducher.

### 1.1. Définition d'un modèle mathématique

Un modèle mathématique est une « *représentation* des aspects *essentiels* d'un système présentés sous une forme *exploitable* » (14).

Le moyen de *représentation* utilisé est issu du langage mathématique. Les différents paramètres quantifiables (ex. : température) sont reliés entre eux par des opérateurs (ex. : moyenne) à l'intérieur d'équations mathématiques. Cette représentation suppose donc une définition explicite du système.

La modélisation suit le principe de parcimonie : on omettra volontiers des variables impertinentes (ex. : cours de la bourse dans un modèle de croissance bactérienne). Effectivement, un modèle cherche à se concentrer sur l'*essentiel* et non à reproduire fidèlement chaque facette du monde réel. Ce qui compte est ce que le modèle explique ou prédit et non pas ce qu'il laisse de côté.

La formalisation mathématique de nos connaissances à l'avantage d'être facilement *exploitable* par le scientifique. En effet, les conclusions que le modèle nous permet de tirer sont toutes des conséquences logiques du modèle. Grâce à celles-ci, nous pourrons procéder à deux objectifs complémentaires.

## 1.2. Objectifs de la modélisation : prédire et expliquer

Les modèles mathématiques ont deux objectifs principaux et complémentaires : *expliquer* et *prédir* (15). Il existe également des modèles descriptifs dont le but est de représenter concisément un système. Ces modèles sortent du cadre de cette thèse.

### 1.2.1. Modèles explicatifs

Un modèle explicatif cherche à décrire le fonctionnement du système en identifiant les différentes variables causales ainsi que les interactions entre elles. Cette approche cherche donc avant tout à être représentative de la réalité et la comprendre. Il s'agit, par exemple, de la méthodologie d'un essai clinique randomisé où l'on cherche à déterminer si un médicament cause une amélioration de l'état de santé d'un malade.

### 1.2.2. Modèle prédictif.

Un modèle prédictif a pour but de prédire de nouvelles observations. Il ne cherche donc pas nécessairement à être représentatif de la réalité. Pour cela, entre deux modèles ayant un pouvoir prédictif égal, nous sélectionnerons celui qui est le plus simple (principe de parcimonie) en omettant des variables bien qu'elles soient causales du phénomène en question. De même, une variable non causale sera incluse dans le modèle si elle améliore sa puissance de prédiction. Par exemple, la capacité prédictive d'un modèle pronostic d'un patient pour une maladie, peut être augmenté avec la prise en compte de variables comme la classe socioéconomique.

### 1.2.3. Indépendance entre explication et prédition

Les physiciens, comme Bunge, considèrent qu'explication et prédition sont des notions dépendantes : « *A theory can predict to the extent it can describe and explain.* » (16). Il est nécessaire de garder à l'esprit que la biologie a montré l'indépendance de ces aspects (5). Par exemple, la biologie évolutionnaire constitue une excellente description et explication de l'origine des espèces, mais est incapable de prédire l'émergence d'une nouvelle espèce. Au contraire, un modèle, comme un réseau de neurones, peut parfaitement prédire une observation nouvelle sur un système, sans pour autant représenter la réalité correctement.

## 1.3. Classification mathématique des modèles

Les modèles peuvent être opposés avec les qualificatifs suivants : linéaire vs non-linéaire, statique vs dynamique, continu vs discret, déterministe vs stochastique, paramétrique vs non-paramétrique.

### 1.3.1. Linéaire vs non-linéaire

Un modèle linéaire est un modèle qui peut être représenté par une combinaison linéaire de ses paramètres, c'est-à-dire, qui respecte le principe de superposition. Cela signifie plus simplement que l'on peut représenter graphiquement le modèle par une droite. Un exemple est la régression linéaire. Ce sont les modèles les plus simples, les plus intuitifs et les plus utilisés en sciences de la vie. Un modèle non-linéaire est un modèle qui n'est pas linéaire. Autrement dit, la réponse du système n'est pas proportionnelle à l'entrée. Ces modèles non-linéaires ont des dynamiques généralement beaucoup plus complexes et peuvent être associés avec des phénomènes comme le chaos (17). Le chaos est une propriété d'un système déterministe dans lequel il existe une extrême sensibilité aux conditions initiales empêchant ainsi toute prédition sur le long terme. Certains systèmes biologiques le rythme cardiaque ou l'activité neuronale auraient des dynamiques non-linéaires et l'utilisation de méthodes linéaires ne permettraient pas de les modéliser convenablement (18).

### 1.3.2. Continu vs discret

Un modèle continu est un modèle dont ses composantes sont continues, c'est-à-dire, pouvant prendre une infinité de valeurs. (ex. : l'âge d'un patient). Un modèle discret est un modèle dont ses composantes sont discrètes, c'est-à-dire, pouvant prendre un nombre fini de valeurs (ex. : l'âge d'un patient sous forme d'intervalles comme jeune, adulte et âgé). Le plus souvent, l'implémentation informatique d'un modèle requiert que celui-ci soit discret (19). C'est également le cas dans le logiciel que nous avons développé. Le passage d'un modèle continu à un modèle discret porte le nom de discréétisation. Il existe plusieurs moyens de discréteriser une variable continue (19). Les algorithmes les plus connus sont ceux des espaces-égaux ou des fréquences-égales.

### 1.3.3. Statique vs dynamique

Un modèle statique est un modèle qui ne dépend pas du temps. L'étude du système doit donc se faire à l'état d'équilibre que l'on doit supposer. Un exemple est une étude épidémiologique entre un facteur d'exposition et une maladie. Au contraire, un système dynamique dépend du temps. Un exemple est un modèle de croissance bactérienne où le nombre de bactéries est une fonction du temps.

### 1.3.4. Déterministe vs stochastique

Un modèle déterministe est un modèle où il existe une unique conséquence pour chaque état. Très souvent, ces modèles utilisent des équations différentielles. Des exemples sont les modèles de mécanique classique ou les modèles pharmacocinétiques compartimentaux. Un modèle stochastique est un modèle où il existe une distribution de la probabilité de chaque conséquence possible pour chaque état. Ce sont notamment les modèles statistiques dont dérivent les tests statistiques comme le test t de Student. D'autres modèles sont mixtes. C'est le cas des modèles de régression de la forme  $Y = \beta X + \varepsilon$ . Cette équation est composée d'une composante déterministe  $\beta X$  ainsi que d'une composante stochastique  $\varepsilon$  représentant notre manque de connaissance sur le processus ainsi que l'erreur expérimentale. Il est intéressant de remarquer que, du fait de la présence de l'erreur expérimentale, un phénomène

déterministe est équivalent à un phénomène stochastique du point de vue de l'observateur. On parle d'équivalence observationnelle (20). Le modélisateur doit faire des choix concernant l'utilisation des modèles utilisés. Il doit se baser sur des concepts comme la généralisabilité, l'exactitude et l'originalité de ses prédictions, la simplicité et l'intuitivité de son modèle. Un modèle est par définition incomplet, il s'agit uniquement d'une *représentation* de la réalité. On retiendra la citation du statisticien Georges Box : « *Essentially, all models are wrong, but some are useful.* » (21).

### 1.3.5. Paramétrique vs non-paramétrique

Cette distinction concerne uniquement les modèles statistiques. Les modèles paramétriques supposent l'existence d'une loi de probabilité connue qui peut être décrite par un nombre fini de paramètres. Par exemple, on peut décrire une loi normale uniquement à partir d'une moyenne et d'un écart-type. On écrira simplement  $\mathcal{N}(\mu, \sigma)$  pour décrire entièrement une loi normale de moyenne  $\mu$  et d'écart-type  $\sigma$ . Au contraire, les modèles non-paramétriques ne se basent pas sur des paramètres pour décrire la distribution de probabilité et n'assument pas une loi de distribution particulière. La complexité du modèle croît donc avec le nombre d'observations. Le choix du type de modèle repose sur un compromis entre le biais et la variance du modèle (22). Le biais représente l'inadéquation entre la distribution du modèle et celle qui est réellement observée. La variance représente la différence d'estimation de la distribution si l'on utilisait un autre échantillon d'observations. Les modèles paramétriques, en supposant une loi de probabilité, introduisent une rigidité réduisant ainsi la variance du modèle (23). Cependant, cette rigidité pourrait empêcher le modèle de refléter la réalité et donc le biaiser. Au contraire, les modèles non-paramétriques ne supposent pas de distribution sous-jacente et sont donc plus flexibles, réduisant la possibilité de biais. Cependant, cette flexibilité se paie par une plus grande variance et impose généralement donc un plus grand nombre d'observations pour une estimation correcte.

## 1.4. Intérêts scientifiques

Les intérêts de la modélisation mathématique semblent difficiles à résumer succinctement et nous ne chercherons pas à être exhaustifs, mais à donner quelques exemples.

### 1.4.1. Tester ou générer des hypothèses

Les modèles mathématiques sont analogues aux modèles verbaux que les biologistes utilisent le plus souvent (ex : modèles de la théorie de l'évolution telle que formulé par Darwin). Dans les deux cas, nous partons d'un ensemble de suppositions sur le fonctionnement du système et nous suivons un cheminement logique jusqu'aux conclusions. L'utilité des mathématiques ici provient de sa rigueur permettant de réduire la possibilité d'un raisonnement erroné (11). Les modèles mathématiques peuvent donc être utilisés pour tester des hypothèses. Ces modèles permettent également de générer des hypothèses en révélant les relations entre les différentes variables. De plus, s'il existe un désaccord entre le modèle et les observations, nous serons amenés à réaliser de nouvelles investigations. Les modèles mathématiques permettent donc d'orienter notre recherche. La résolution de ce désaccord peut être source de nouvelles connaissances (24).

### 1.4.2. Intervention

Une fois qu'un modèle explicatif est construit, il est possible d'intervenir sur le système rationnellement. Nous pouvons modifier la dynamique du système à notre avantage. Par exemple, en modélisant l'ensemble des interactions protéine-protéine d'un processus physiopathologique sous la forme d'un réseau mathématique, nous pouvons identifier des cibles thérapeutiques potentielles en minimisant le risque d'effets secondaires (25).

#### 1.4.3. Simulation

Les modèles mathématiques permettent également de simuler *in silico* des expériences qui sont non éthiques, trop chères, trop contraignantes ou simplement impossibles à réaliser. Par exemple, afin d'attester la toxicité du tabac, il est non éthique de conduire des essais cliniques randomisés. Grâce aux modèles mathématiques, il est cependant possible d'évaluer cette toxicité à travers des données purement observationnelles (26).

On remarquera ici l'importance de l'informatique dans l'implémentation des simulations. Certaines méthodes de simulation, bien que relativement simples conceptuellement, requièrent de nombreuses itérations. Ces méthodes ont connu une grande popularité avec l'essor de l'informatique. Ce sont, par exemple, les méthodes de simulation Monte-Carlo comme celles de rééchantillonnage (27). Celles-ci permettent de trouver une solution à un problème n'admettant pas de solution analytique, c'est-à-dire, un problème qui ne peut pas être évalué en un nombre fini d'opérations.

#### 1.4.4. Prévisions

Les modèles mathématiques permettent d'anticiper un événement. Ce sont, par exemple, les modèles de pronostic ou de réponse thérapeutique qui sont utilisés en cancérologie (28). Ces modèles permettent ainsi de mieux orienter nos décisions actuelles afin d'atteindre un état futur souhaité.

#### 1.4.5. Classification

Les modèles mathématiques permettent d'attribuer une classe à une combinaison de variables. Ce sont, par exemple, les modèles d'aide au diagnostic qui propose un diagnostic à partir d'un ensemble de caractéristiques du patient et de symptômes. Leur implémentation porte souvent le nom de système expert. Un exemple est Watson, le superordinateur d'IBM (29).

#### 1.4.6. Analyse d'un grand nombre de données

L'explosion récente de la quantité de données impose l'utilisation de modèles mathématiques ainsi que d'algorithmes afin de rendre l'analyse possible. C'est par exemple la recherche d'homologues pour une séquence protéique que nous aurions nouvellement identifiée. La base de données UniProt (30) contient à l'heure actuelle plus de 500 000 séquences qui sont potentiellement homologues à la nôtre (30). Sans des méthodes heuristiques de recherche et des outils informatiques accessibles, cette recherche serait impossible.

### 1.5. Les difficultés de la mathématisation du vivant

Les modèles mathématiques, comme nous avons précédemment vu, possèdent de nombreux intérêts à être utilisés dans les sciences de la vie. Ceux-ci nous permettent effectivement d'accéder à des connaissances qui seraient autrement inaccessibles. Il est donc défendable de penser que les mathématiques avec l'informatique auront un rôle majeur au sein des sciences de la vie dans les années à venir (13). Cependant, il existe encore de nombreuses réticences de la part des cliniciens et biologistes à les utiliser. Ces approches sont marginales et sont généralement limitées à l'application en routine de tests statistiques (31). De plus, on pourrait même croire qu'elles sont mal perçues par les cliniciens et biologistes. Le nombre d'équations mathématiques dans un article serait inversement proportionnel au nombre de citations que celui-ci obtient (32). Ils argumenteront que la complexité et la stochasticité du vivant sont incompatibles avec la rigidité imposée par les modèles mathématiques. Cependant, remarquons qu'il existe une synergie entre les mathématiques et les sciences naturelles. Par exemple, la physique est dépendante des mathématiques depuis Galilée au XV<sup>e</sup> siècle. De nombreuses découvertes auraient sans doute été impossibles sans les mathématiques. De même, la physique a permis de nombreuses avancées en mathématiques en stimulant la recherche avec des problèmes concrets. C'est l'exemple, déjà cité, de Newton qui dû inventer le calcul infinitésimal afin de formaliser la théorie de la gravité universelle. La mathématisation de la biologie a le plus souvent puisé dans les méthodes et outils des sciences physiques et de l'ingénieur. Ces approches, le plus souvent déterministes, sont

effectivement très adaptées lorsque nous adoptons une étude réductionniste. Cependant, nous nous rendons compte aujourd’hui que des approches holistes, comme celle de la biologie des systèmes, sont nécessaires afin de mieux saisir ce qu’est la vie. Ces approches sont nouvelles et il est probable que les méthodes mathématiques adaptées n’existent pas encore, mais devront être développées (12). Le problème n’est pas que la vie n’est pas mathématisable, mais que sa dynamique demande des mathématiques adaptées.

Les systèmes vivants ont souvent une dynamique purement stochastique : ils sont indéterministes. C'est l'exemple de la survenue d'une mutation ponctuelle. De plus, la survenue de cet événement est sans rapport avec son importance. Cette mutation, bien qu'issue du hasard, pourrait conférer à l'espèce porteuse un avantage considérable. Celle-ci impacte donc considérablement le cours du futur et affecte ainsi nos capacités de prédiction. Cependant, comme le remarqua Fisher dans son livre *The Design of Experiments* (33), cela ne signifie pas que nous ne pouvons pas étudier l'incertitude de la biologie avec la rigueur des mathématiques. Une branche des mathématiques, les statistiques, est née en partie grâce à la biologie. C'est notamment Galton qui définit le concept de corrélation pour étudier l'hérédité et Fisher qui développa les méthodes d'analyse de la variance pour l'agronomie (34).

Les systèmes vivants peuvent également avoir une dynamique chaotique. Effectivement, ils résultent de l'association de plusieurs éléments simples (ex. : protéines) associés au sein d'un réseau complexe (ex. : protéome) (25). Pris individuellement, les éléments peuvent généralement être appréhendés par des approches déterministes et réductionnistes. Cependant, l'association de ces éléments dans un réseau a pour conséquence de faire apparaître des propriétés émergentes : « le tout est différent de la somme de ses parties ». La dynamique du système devient alors complexe. Bien qu'elle puisse être fondamentalement déterministe, il est impossible de faire des prédictions précises. La constatation de cette complexité a souvent été utilisée comme argument contre la modélisation. Cependant, cette complexité a résulté en des avancées mathématiques significatives, notamment dans la théorie des réseaux et la théorie du chaos. On pense à l'article d'Alan Turing

décrivant un modèle chaotique permettant d'expliquer la formation des différents motifs sur les pelages des animaux (35). C'est aussi le développement de la théorie des graphes aléatoires par Erdős et Renyi stimulés par la volonté de modéliser les réseaux synaptiques (36).

Nous allons dans la prochaine partie décrire une méthode mathématique conçue pour l'étude du vivant : le Z de Ducher. Cette mesure sera contrastée avec des mesures d'association historiques comme le coefficient de corrélation de Pearson qui ne sont pas toujours adaptées aux systèmes vivants.

## 1.6. Exemple de modèle : association statistique

### 1.6.1. Causalité et association

Le déterminisme causal est une notion épistémologique qui est défini comme suit : « ordre de faits dans lequel chaque phénomène est dépendant de certaines conditions et se produit nécessairement lorsque ces conditions sont satisfaites. » (37). Les mêmes causes produisent donc toujours les mêmes effets. Ce principe de causalité devient alors d'intérêt scientifique, car les causes permettraient *d'expliquer* ou de *prédir* les effets. Le but de la science serait donc d'identifier et de caractériser le processus causal. Le démon de Laplace (4) définit un idéal scientifique dans lequel il serait possible de connaître tous les effets si l'on connaissait toutes les causes. Cependant, cet idéal semble aujourd'hui utopique (38,39). Effectivement, même si nous supposons que le déterminisme fort existe, l'identification de la causalité reste particulièrement difficile, car celle-ci n'est pas directement observable. Seule sa conséquence, sur l'association entre les événements, est observable. Bien que nous sachions que l'association n'implique pas la causalité, l'association peut être utilisée pour nous guider, car la causalité implique une forme d'association (6). Pour cela, la nature et la force des relations d'associations doivent être testées et mesurées correctement. Nous faisons remarquer au lecteur que des méthodes adaptées à l'étude de la causalité sont relativement récentes. On citera les réseaux bayésiens (40) qui peuvent faire appel à des mesures d'association comme l'information mutuelle (41).

### 1.6.2. Généralités sur l'association statistique

L'association statistique peut s'interpréter comme la réduction d'incertitude d'une variable  $X$  par la connaissance d'une deuxième variable  $Y$ . L'association complète est le cas où la variable  $X$  est complètement déterminée par la variable  $Y$ . L'absence d'association ou indépendance est le cas où la variable  $X$  n'apporte aucune information sur la variable  $Y$ . Les mesures d'association sont là pour apporter une définition graduée de l'association. Elles sont généralement normalisées entre 0 et 1 ou -1 et 1. Dans le cas des mesures globales, une valeur absolue de 1 implique l'association parfaite, tandis qu'une valeur de 0 implique l'indépendance. Les valeurs intermédiaires reflètent la force de l'association. Le signe de la mesure reflète le sens de la relation : décroissante (négative) ou croissante (positive).

Il est important de distinguer la force de l'association de sa significativité statistique (mesurée sous la forme d'une valeur-p ou d'un intervalle de confiance). Effectivement, une association forte peut ne pas être significative (ex. : gamme d'étalonnage sur faible effectif), tandis qu'une association faible peut l'être (ex. : étude épidémiologique avec fort effectif). Cette significativité statistique peut être calculée soit analytiquement en connaissant la distribution de la mesure ou bien par des méthodes de rééchantillonnage comme le *bootstrap* (27). Nous utilisons cette dernière méthode dans le Zébu pour le calcul de valeur-p et d'intervalles de confiance.

Les mesures d'association sont nombreuses dans la littérature. On pourrait en dénombrer au moins une cinquantaine (42). Elles partagent cependant des similarités entre elles. Nous ne procéderons donc pas à une description de chacune de ces mesures, mais nous restreindrons à quelques mesures d'intérêts. Nous remarquons cependant qu'il est possible de les classifier en méthodes basées sur la variance, sur la différence de moyenne et sur les probabilités. Il est aussi possible de les distinguer selon qu'elles sont continues ou discrètes, bivariées (deux variables) ou multivariées (plusieurs variables), paramétriques ou non-paramétriques ainsi que globales ou locales. Nous décrirons en détail trois mesures : le coefficient de corrélation de Pearson, le chi-deux et le Z de Ducher.

### 1.6.3. Mesures basées sur la variance : coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson, noté  $r$ , est sans doute la mesure d'association la plus connue. Elle mesure la corrélation, ou association, entre deux variables  $X$  et  $Y$ . Pour cela, elle se base sur la covariance de  $X$  et  $Y$ , notée  $cov(X, Y)$ . La covariance est normalisée entre -1 et 1 par la variance de  $X$  et  $Y$ , notées respectivement  $\sigma_X^2$  et  $\sigma_Y^2$ . La formule du coefficient de corrélation est donnée dans l'Équation 1 où  $\sigma$  représente l'écart-type.

$$r = \frac{cov(X, Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Équation 1. Coefficient de corrélation de Pearson

La variance d'une variable  $X$  est calculée pour un échantillon comme indiqué dans l'Équation 2. Dans cette formule  $\bar{x}$  représente la moyenne et  $x_i$  représente une réalisation de  $X$ . On dénombre  $n$  réalisations.

$$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Équation 2. Variance d'une variable aléatoire

La variance traduit la dispersion des réalisations  $x_i$  autour de la moyenne  $\bar{x}$ . Afin d'avoir des écarts à la moyenne  $(x_i - \bar{x})$  toujours positifs, on prend le carré des écarts  $(x_i - \bar{x})^2$ . Ainsi, la somme des écarts au carré, la variance est toujours positive et comprise entre 0 et l'infini. Cette mesure est alors divisée par le nombre de réalisations  $n$ . Afin de mieux appréhender ce qu'est la variance, des exemples de lois normales centrées ayant des variances différentes sont présentés dans la Figure 2A. Une interprétation géométrique de la variance est donnée dans la Figure 2B.

La covariance de deux variables  $X$  et  $Y$  est calculée pour un échantillon comme indiqué dans l'Équation 3. Dans cette formule  $\bar{x}$  et  $\bar{y}$  représentent la moyenne et  $x_i$  et  $y_i$  représentent une réalisation de respectivement  $X$  et  $Y$ . On dénombre  $n$  réalisations.

$$cov(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Équation 3. Covariance de deux variables aléatoires

La covariance traduit la variation jointe de deux variables. On remarquera la similarité avec la formule de la variance et le fait que  $cov(X, X) = \sigma_X^2$ . Cependant, la covariance n'est pas minorée par 0 et peut prendre des valeurs négatives. Son signe représente le sens de covariation. Elle est négative si la relation est décroissante et positive si la relation est croissante. Nous proposons également une interprétation géométrique de la covariance qui est donnée dans la Figure 3.

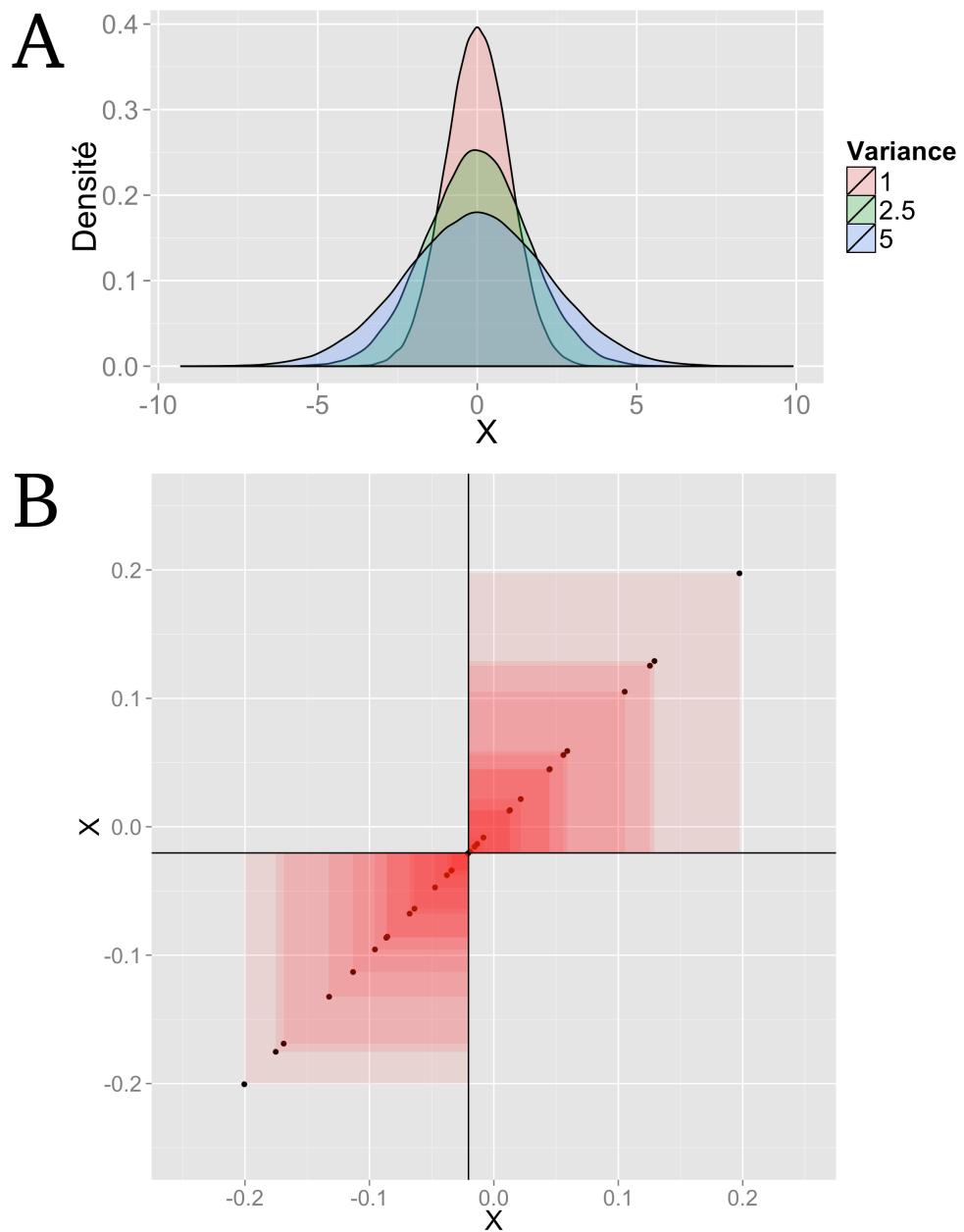


Figure 2. Variance d'une variable aléatoire

A : La variance d'une variable aléatoire  $X$  traduit la dispersion des réalisations  $x_i$  autour de sa moyenne  $\bar{x}$ . Nous avons pris pour exemple des lois normales centrées, c'est-à-dire, de moyenne nulle. Ces lois ont des variances différentes. On remarque que plus la variance est élevée, plus la dispersion des réalisations autour de la moyenne est étendue.

B : Interprétation géométrique de la variance. Nous avons simulé 30 réalisations tirées d'une loi normale centrée de variance 0,1. Dans le plan, les abscisses et les ordonnées représentent les réalisations  $x_i$ . La moyenne  $\bar{x}$  est représentée par les deux lignes noires. Les carrés rouges représentent les carrés des écarts à la moyenne, c'est-à-dire  $(x_i - \bar{x})^2$ . La variance est proportionnelle à la somme de l'aire de tous ces carrés.

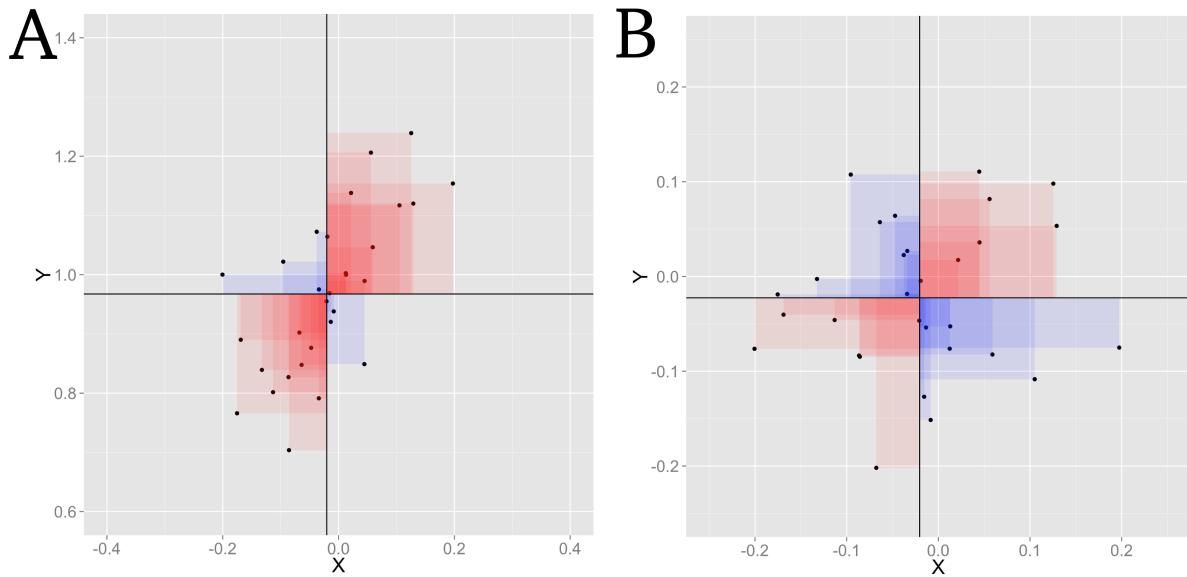


Figure 3. Interprétation géométrique de la covariance

Nous avons simulé 30 réalisations de  $X$  et  $Y$ . Les moyennes  $\bar{x}$  et  $\bar{y}$  sont représentées par les deux lignes noires. Les rectangles représentent les écarts à la moyenne de  $x_i$  et  $y_i$ , c'est-à-dire  $(x_i - \bar{x})(y_i - \bar{y})$ . Les écarts positifs sont en rouge, les écarts négatifs en bleu. La covariance est proportionnelle à la somme de l'aire de tous de ces rectangles. A : Relation linéaire croissante entre  $X$  et  $Y$  avec du bruit gaussien. On remarque qu'il y a plus de rectangles rouges (positifs) que de rectangles bleus (négatifs) ce qui signifie que la covariance est positive. Dans cette simulation :  $Cov(X, Y) = 0,0086$  et  $r = 0,68$ .

B : Absence de relation entre  $X$  et  $Y$ . On remarque qu'il y a presque autant de rectangles rouges (positifs) que de rectangles bleus (négatifs) ce qui signifie que la covariance est proche de zéro. Dans cette simulation :  $Cov(X, Y) = 0,0010$  et  $r = 0,14$ .

Cette interprétation géométrique de la covariance nous permet de comprendre certaines de ses propriétés mathématiques. Par exemple, la covariance dépend de l'échelle utilisée : plus les carrés des écarts sont grands, plus la covariance est grande sans pour autant que l'association soit plus forte. Le coefficient de corrélation permet de normaliser la covariance entre -1 et 1. Dans le cas d'association parfaite, on démontre que  $|Cov(X, Y)| = \sigma_X \sigma_Y$ , ce qui explique la normalisation utilisée. De plus, ces mesures ne sont sensibles qu'à des relations linéaires. Les relations non-linéaires vont créer un mélange de rectangles positifs et négatifs. Ceci peut être problématique en sciences de la vie où l'on retrouve des relations non-linéaires comme des seuils, des saturations ou des relations en U. C'est par exemple les relations dose-réponse hormétique qui ont une forme en U (43). Nous verrons un exemple en III.2.1. Une alternative est le coefficient de corrélation de Spearman qui procède par un recodage

des réalisations sous forme de rangs. Elle est cependant uniquement valide pour des relations monotones, c'est-à-dire, uniquement croissantes ou décroissantes, et donc toujours invalide pour les relations en U. Finalement, on remarquera que ces mesures ne sont utilisables que pour des variables aléatoires continues. Pour des variables catégorielles, on utilisera le plus souvent des mesures dérivées du chi-deux que nous allons maintenant décrire.

#### 1.6.4. Mesures basées sur les probabilités : chi-deux

Le test d'indépendance du chi-deux est un test non-paramétrique cherchant à établir si deux variables  $X$  et  $Y$  sont indépendantes ou non. Pour la simplicité de l'explication, nous considérons que les deux événements  $X$  et  $Y$  n'ont que deux réalisations possibles. Ce sont leurs modalités. Celles-ci seront codées par 0 (non-réalisation) et 1 (réalisation) :  $x_0$ ,  $x_1$ ,  $y_0$  et  $y_1$ . Il devient possible de représenter les combinaisons de ces réalisations dans une matrice ou tableau de contingence (Tableau 1) où  $n_{ij}$  représente le comptage des réalisations avec  $X = i$  et  $Y = j$ . On note  $n_{i\cdot}$  et  $n_{\cdot j}$ , respectivement les comptages marginaux de  $X$  et  $Y$  tel que  $n_{i\cdot} = \sum_j n_{ij}$  et  $n_{\cdot j} = \sum_i n_{ij}$ . On note  $n$  la somme de toutes ces réalisations, soit  $n = \sum_{i,j} n_{ij}$ .

X \ Y	$y_0$	$y_1$	Total
$x_0$	$n_{00}$	$n_{01}$	$n_{0\cdot}$
$x_1$	$n_{10}$	$n_{11}$	$n_{1\cdot}$
Total	$n_{\cdot 0}$	$n_{\cdot 1}$	$n$

Tableau 1. Tableau de contingence pour deux variables binaires

La partie colorée en bleu sont les comptages joints des événements et est proportionnel à la probabilité jointe  $p(X, Y)$ . La partie colorée en orange sont les comptages marginaux et est proportionnel aux probabilités marginales  $p(X)$  et  $p(Y)$ .

Il est possible de définir une fréquence de réalisation telle que :  $f_{ij} = \frac{n_{ij}}{n}$ . La loi des grands nombres nous indique que quand  $n$  tend vers l'infini alors  $f_{ij}$  converge vers la probabilité  $p_{ij}$ . Il est possible d'interpréter ce tableau comme une matrice de probabilité jointe  $p(X, Y)$ . Les probabilités marginales, c'est-à-dire  $p(X)$  et  $p(Y)$ , sont représentées au niveau des marges. À partir de cette matrice de probabilité, il est

possible de mesurer l'association entre les deux variables. La stratégie générale est de calculer l'association localement pour chaque modalité, c'est-à-dire pour chaque cellule, et de faire une somme ou une moyenne afin d'obtenir une mesure globale de l'association. On doit supposer que l'association locale est distribuée de manière homogène entre les modalités et donc que celle-ci peut être ignorée.

L'association entre variables peut aussi être interprétée comme la cooccurrence des événements. Il paraîtrait alors intuitif d'utiliser la probabilité jointe de  $X$  et  $Y$  comme mesure de l'association, c'est-à-dire,  $p(X, Y)$ . Cette mesure est cependant mauvaise, car elle ne fonctionne pas dans le cas d'événements rares. Effectivement, la valeur maximale que peut prendre  $p(X, Y)$  est celle de la plus petite probabilité marginale  $p(X)$  ou  $p(Y)$  :  $p(X, Y) \leq \min(p(X), p(Y))$ . Il semble donc nécessaire de trouver une normalisation à tous les événements, quelle que soit leur probabilité marginale. Un repère possible est celui de l'indépendance statistique. Deux événements sont considérés comme indépendants s'ils respectent l'Équation 4.

$$p(X, Y) = p(X) p(Y)$$

Équation 4. Indépendance statistique entre deux variables aléatoires.

Cette équation peut s'écrire également en utilisant les probabilités conditionnelles dans les formes suivantes :  $p(X|Y) = p(X)$  et  $p(Y|X) = p(Y)$ . La première équation se lit : « la probabilité de  $X$  sachant  $Y$  est égal à la probabilité de  $X$  ». Cela signifie que la connaissance de la réalisation d'une variable ne nous apporte aucune information sur la réalisation de l'autre. C'est exactement le cas qui ne nous s'intéresse pas. Nous pouvons cependant définir deux cas d'intérêt déviant de l'indépendance. Pour l'association positive, les événements cooccurrent plus souvent que l'indépendance :  $p(X, Y) > p(X) p(Y)$ . Pour l'association négative, les événements cooccurrent moins souvent que l'indépendance :  $p(X, Y) < p(X) p(Y)$ . Les mesures de l'association basées sur les probabilités utilisent généralement comme référence l'indépendance statistique.

Dans le cas du test du chi-deux, la mesure locale de l'association prend le nom de résiduel, se note  $R^2$ , et est défini dans l'Équation 5. Le numérateur correspond à une déviation par rapport à l'indépendance. Celui-ci est porté au carré afin que les deux formes d'associations (positive et négative) aient le même signe. Le dénominateur est une manière de standardiser le résultat. On remarquera qu'une standardisation est différente d'une normalisation. Effectivement, une normalisation implique que le résultat soit toujours compris entre deux valeurs (ex : -1 et 1 ou 0 et 1). Ici, les résiduels sont compris entre 0 et l'infini. On obtient une mesure globale de l'association nommée chi-deux en faisant la somme des résiduels (Équation 6).

$$R_{i,j}^2 = n \times \frac{(p(x_i, y_j) - p(x_i) p(y_j))^2}{p(x_i) p(y_j)}$$

Équation 5. Résiduels du chi-deux

$$\chi^2 = \sum_{i,j} R_{i,j}^2.$$

Équation 6. Le chi-deux est la somme des résiduels

Un des désavantages des résiduels et du chi-deux comme mesure est qu'ils ne sont pas normalisés. Ils sont donc incomparables pour des tableaux de contingence de dimensions différentes ainsi que des échantillons de taille différente. Pour cela le chi-deux n'est jamais référé comme étant une mesure. Il est cependant possible d'en dériver des mesures normalisées entre 0 et 1 comme le phi, le V de Cramér et le T de Tschuprow (27). On remarquera au passage qu'il existe de nombreuses autres mesures globales de l'association basées sur les probabilités. C'est notamment le rapport des cotes ou *odds ratio* (44) très utilisé par les épidémiologistes ou les mesures de Goodman et Kruskal (45).

Supposons maintenant que les variables  $X$  et  $Y$  ne soient plus discrètes, mais continus. Théoriquement, il serait toujours possible de calculer les coefficients précédemment vus. En effet, le concept de probabilité existe aussi bien pour des valeurs continues. Cependant, pour des raisons de calculs, cela n'est jamais réalisé. Le

plus souvent, on discrétisera les réalisations continues en des catégories discrètes avant de procéder. Il reste également possible d'utiliser des mesures d'association basée sur la variance comme le coefficient de corrélation de Pearson.

#### 1.6.5. Différence entre mesure globale et locale : Z de Ducher

Jusqu'à maintenant nous avons décrit des mesures globales de l'association, c'est-à-dire, des mesures *uniques* censées représenter la force de l'association entre les variables. Celles-ci correspondent à une moyenne de l'association locale entre les modalités des variables. Elles supposent donc que cette force peut être considérée comme étant uniformément répartie pour toutes les modalités des variables. Bien que cela soit pratique, cela n'est pas systématique justifié. Prenons un exemple concernant la réponse thérapeutique à un médicament. L'association avec une réponse thérapeutique positive sera beaucoup plus forte pour des individus sensibles que pour des individus résistants au médicament. Pour autant, une mesure globale de cette association nous dira juste qu'il existe une tendance générale entre le médicament et la réponse thérapeutique et que celle-ci est de telle force. Elle ne nous révèlera pas la présence de deux types d'individus et sa mesure sera faussement affaiblie par les individus résistants. Cela ne signifie pas que l'on doit arrêter de penser en tendances globales, mais simplement qu'il est parfois nécessaire de décomposer une association globale en ses composantes locales.

L'approche locale peut effectivement permettre de mieux expliquer la relation entre les variables. Il se peut, par exemple, qu'une association globale non significative cache une association locale significative. L'exemple que nous venons de donner pourrait tomber dans cette catégorie. Similairement, une association globale significative peut être composée de régions localement non associées. Nous verrons un exemple dans la partie III.2.2. Bien que l'approche locale ne soit pas nouvelle, elle n'est que rarement considérée de nos jours. Quelques exemples disparates se retrouvent dans la littérature (46–49). Nous offrons ici au lecteur une description d'une des deux mesures implémentées dans le Zébu : le Z de Ducher. Une description de l'autre mesure, l'information mutuelle spécifique, est donnée dans l'article décrivant le Zébu dans la partie III.1. de cette thèse.

Le Z de Ducher est une mesure non-paramétrique de l'association locale. Elle correspond à une déviation normalisée de l'indépendance. Les valeurs prises sont comprises entre -1 et 1. Une valeur positive de 1 signifie que les événements sont toujours associés l'un à l'autre (association positive). Une valeur négative de -1 implique que les événements ne sont jamais associés l'un à l'autre (association négative). Une valeur nulle indique que les événements sont indépendants, c'est-à-dire, non associés. On remarquera que l'interprétation du signe des mesures locales diffère des mesures globales où le signe donne le sens de la relation.

L'Équation 7 ci-dessous définit la déviation de l'indépendance  $\Delta p(x_i, y_j)$  pour les modalités  $x_i$  et  $y_j$ . Dans le cas de l'association positive, c'est-à-dire,  $\Delta p(x_i, y_j) > 0$ ,  $\Delta p(x_i, y_j)$  est maximisée lorsque  $p(x_i, y_j) = \min(p(x_i), p(y_j))$ . Dans le cas de l'association négative, c'est-à-dire,  $\Delta p(x_i, y_j) < 0$ ,  $\Delta p(x_i, y_j)$  est minimisée lorsque  $p(x_i, y_j) = 0$ . Pour le cas de l'indépendance, le Z de Ducher est défini comme étant nul. La normalisation de  $\Delta p(x_i, y_j)$  est donc obtenue avec le système d'Équation 8.

$$\Delta p(x_i, y_j) = p(x_i, y_j) - p(x_i) p(y_j)$$

Équation 7. Déviation de l'indépendance

$$\forall \Delta p(x_i, y_j) > 0, \quad Z(x_i, y_j) = \frac{\Delta p(x_i, y_j)}{\min(p(x_i), p(y_j)) - p(x_i) p(y_j)} \in ]0, 1]$$

$$\forall \Delta p(x_i, y_j) < 0, \quad Z(x_i, y_j) = \frac{\Delta p(x_i, y_j)}{p(x_i) p(y_j)} \in [-1, 0[$$

$$\forall \Delta p(x_i, y_j) = 0, \quad Z(x_i, y_j) = 0$$

Équation 8. Z de Ducher

On obtient avec ses équations une matrice dont les valeurs traduisent l'association pour les modalités des variables  $X$  et  $Y$ . Il est cependant possible de former une mesure globale à partir de ces mesures locales. Nous définissons le Z global de Ducher, noté  $gZ$ , dans l'Équation 9. Il s'agit de la moyenne du Z de Ducher.

$$gZ = \sum_{i,j} p(x_i, y_j) Z(x_i, y_j)$$

Équation 9. Z global de Ducher

On peut vouloir identifier les individus qui dérogent à la tendance générale donnée par une mesure globale. Les mesures de l'association locales peuvent être utilisées comme un filtre de liaison. Effectivement, elles permettent d'extraire les modalités de variables pour lesquels les variables sont fortement dépendantes l'une de l'autre. On définit les couples ayant une association positive. Ceux-ci peuvent être interprétés comme ceux qui respectent la règle ou la tendance générale. On peut également définir les couples ayant une association négative. Ceux-ci peuvent être interprétés comme ceux qui dérogent la règle ou la tendance générale. Ces deux sous-groupes peuvent alors être comparés, par exemple, en s'intéressant à la distribution d'autres variables. On peut ainsi identifier en quoi ces sous-groupes diffèrent. Ce type d'analyse est nommé analyse en sous-groupe et est implémenté dans le Zébu. Nous en donnerons un exemple dans la partie III.2.2.

On remarquera que jusqu'alors nous n'avons considéré que l'association bivariée, c'est-à-dire entre deux variables. Il s'agit cependant d'un concept qui se généralise à plusieurs variables : on parle d'association multivariée (50). Effectivement, il existe des généralisations multidimensionnelles du Z de Ducher et de l'information mutuelle spécifique. Nous les avons implémentées dans le Zébu et elles sont décrites dans l'article correspondant situé dans partie III.1. de cette thèse. Pour l'association multivariée, nous n'obtenons plus une matrice bidimensionnelle comme le tableau de contingence précédent décrit (Tableau 1), mais une matrice multidimensionnelle. Dans certains cas, des variables peuvent être associées dans des dimensions supérieures (ex. : 3 dimensions) alors qu'elles ne le sont pas dans des

dimensions inférieures (ex. : 2 dimensions). C'est l'exemple du circuit logique XOR que nous verrons dans la partie III.2.3. Cette implémentation a donc un intérêt réel et n'est pas une simple curiosité mathématique.

Finalement, on remarquera que l'approche globale est comparable à la médecine contemporaine où un même médicament est supposé guérir tout le monde (51). Au contraire, la médecine personnalisée défend l'idée que chaque patient devrait être traité comme un individu et non pas comme une moyenne statistique. Afin de rendre ce changement de méthodologie possible, il est nécessaire d'avoir des méthodes adéquates. De fait, nous voyons effectivement un changement des méthodes mathématiques utilisées par la communauté scientifique. Par exemple, dans l'interprétation fréquentiste des probabilités, on définit une probabilité comme étant la fréquence d'occurrence d'un événement lorsque celui-ci est répété une infinité de fois. En utilisant cette interprétation, comment définir la probabilité qu'un individu précis contracte une maladie au cours de sa vie si celui-ci n'en possède qu'une et non pas une infinité ? En conséquence, nous avons remarqué un essor des méthodes bayésiennes qui permettent une définition plus claire de probabilités individuelles (52). Similairement l'analyse de l'association locale est un outil qui rentre dans ce cadre d'analyse de l'individu en tant que tel en allant au-delà des moyennes populationnelles.

## 2. Outils informatiques : exemple du Zébu

Un outil peut être défini comme un moyen de mise en œuvre d'une méthode plus générale. Les outils peuvent être matériels comme un microscope ou virtuels comme un logiciel scientifique. Ils sont d'une importance capitale dans les sciences naturelles en nous permettant d'améliorer nos sens ou de faire des calculs autrement impossibles. Il est possible de soutenir que ces moyens techniques sont indissociables de notre connaissance. Effectivement, il semble raisonnable de penser que de nombreuses découvertes auraient été impossibles sans les outils nécessaires. Peut-on imaginer une théorie cellulaire dans un univers qui ne connaît pas le microscope ? Nous pourrions développer le même argumentaire pour un logiciel scientifique. Finalement, l'importance des outils en sciences est également reflétée dans les mesures bibliométriques. Les articles les décrivant font partie des articles scientifiques les plus cités (53).

Cela nous laisse penser que la cadence des avancées scientifiques peut être ralentie en absence d'outils. Par exemple, des méthodes mathématiques très simples conceptuellement, mais demandant de nombreux calculs comme les méthodes Monte-Carlo, n'ont pas été utilisées avant l'avènement de l'informatique (54). Nous remarquons similairement que le Z de Ducher, qui n'a pas d'outil de calcul dédié, a été peu utilisé, à l'exception de ses auteurs (7) et d'un groupe indépendant de chercheurs (55). Des remarques analogues pourraient être faites sur les autres mesures d'association locale. Ceci justifie l'objectif de cette thèse : le développement d'un logiciel répondant à cette lacune afin de stimuler l'utilisation des mesures locales.

Dans cette deuxième partie, nous nous intéresserons aux outils informatiques et, en particulier, aux logiciels scientifiques. Nous commencerons par décrire les relations entre les sciences de la vie et l'informatique. Afin d'expliquer la conception du Zébu, nous donnerons au lecteur quelques notions d'algorithmique, de programmation et de génie logiciel. Nous clôturons cette partie en définissant les licences et en discutant de l'intérêt de l'open source et de l'open-data en sciences de la vie.

## 2.1. Essor des approches computationnelles – bioinformatique

L'utilisation de l'ordinateur en sciences de la vie n'est pas récente. On attribue la première utilisation à un article de Fisher datant de 1950 (33). Néanmoins, l'explosion de données récentes à laquelle nous avons déjà fait référence a rendu ces approches nécessaires. L'informatique apparaît effectivement comme un moyen très efficace de stocker et rendre accessible ces nombreuses données. L'une des préoccupations de ces approches, plus particulièrement de la bioinformatique, est de maintenir des bases de données comme *UniProt* (30) ou la *Protein Data Bank* (56). En plus du problème de la quantité de données, les méthodes d'analyse de données utilisées en sciences de la vie ont rarement des solutions analytiques (54). Le calcul de solutions demande de nombreuses itérations qu'une machine n'aura pas de mal à exécuter. Finalement, les méthodes computationnelles permettent de réaliser des expériences qui, autrement, seraient non éthiques, trop chères, trop contraignantes ou simplement impossibles à réaliser. En somme, l'utilisation des ordinateurs permet d'ouvrir des perspectives autrement inaccessibles, ce qui explique un gain d'intérêt pour les méthodes *in silico*.

Cependant, ces approches vont au-delà de la simple application de l'informatique à la biologie. Elles sont novatrices pour l'analyse et l'interprétation de l'information biologique. Elles développent des méthodes et outils spécifiquement pour la biologie. C'est notamment des algorithmes comme le *BLAST* qui permet de chercher rapidement des homologues de séquences dans de grandes bases de données.

Finalement, ces approches se chargent également du développement de logiciels scientifiques (57). Ceux-ci permettent à la communauté d'utiliser des méthodes qui leur seraient autrement difficiles à réaliser. Nous allons maintenant brièvement décrire quelques aspects du développement de ces logiciels.

## 2.2. Algorithmes et langages de programmation

### 2.2.1. Algorithmes : exemple de comparaison de séquences

Un logiciel est composé d'un ensemble d'algorithmes. Un algorithme est défini comme une « méthode de calcul qui indique la démarche à suivre pour résoudre une série de problèmes équivalents en appliquant dans un ordre précis une suite finie de règles » (37). Autrement dit, un algorithme applique un ensemble de règles du type « additionner ceci avec cela » afin de transformer le problème (entrée) en solution (sortie). Pour rendre cette définition plus claire, prenons un exemple. Nous allons écrire un algorithme pour compter le nombre de différences entre deux séquences d'ADN de taille identique (on parle également de distance de Hamming). Nous avons ci-dessous (Figure 4), deux séquences nucléiques homologues. Celles-ci ont 10 nucléotides chacune et possèdent deux nucléotides de différence. Ceci se remarque facilement à l'œil nu, mais il serait beaucoup plus difficile de faire la même opération pour toutes les séquences d'un génome humain. L'automatisation de ce comptage se fait naturellement par un algorithme. Celui-ci pourrait longer le long des séquences en comparer les nucléotides entre eux. Il enregistrerait alors le nombre de différences et le retournerait à l'utilisateur à la fin de la séquence. Nous avons retrouvé cet algorithme sous forme de pseudocode (Figure 5), c'est-à-dire sans faire une référence spécifique à un langage de programmation.

<b>0 1 2 3 4 5 6 7 8 9</b>
A C T A A C C A A A
A G T A A C C A T A

Figure 4. Deux séquences nucléiques homologues  
Les deux séquences ont 10 nucléotides. Elles possèdent deux différences représentées en rouge.

```

1. séquence1 = ACTAACCAAA
2. séquence2 = AGTAACCATA
3. différences = 0
4.
5. FOR index = 1 STEP index = index + 1 UNTIL index > séquence1.length:
6.     IF séquence1[index] != séquence2[index]:
7.         différences = différences + 1
8.
9. PRINT différences

```

Figure 5. Algorithme pseudocodé pour calculer le nombre de différences entre deux séquences de taille identique

Les lignes de 1 à 3 permettent de définir nos variables d'intérêt. Ici, ce sont deux chaînes de caractères représentant nos séquences et le nombre de différences entre celles-ci qui est initialisée à zéro. Les lignes de 5 à 7 constituent une boucle d'itération comme indiqué par le mot-clé FOR. Cette structure est très commune en informatique. On initialise d'abord un index (FOR index = 1) qui ici correspond à la position dans notre séquence. On va alors comparer les nucléotides de cette position entre les séquences. Si ceux-ci ne sont pas identiques (IF séquence1[index] != séquence2[index]), alors on incrémente le nombre de différences (différences = différences + 1). A chaque itération, on incrémente l'index de 1 (STEP index = index + 1) et on répète jusqu'à ce qu'on arrive à la fin de la séquence (UNTIL index > sequence1.length). Finalement, la ligne 9 indique à l'algorithme afficher la solution au problème sur l'écran.

## 2.2.2. Langages de bas-niveau et haut-niveau

Le pseudocode utilisé a l'avantage d'être lisible par un être humain. Cependant, il est ininterprétable tel quel par une machine. Effectivement, une machine ne possède qu'un alphabet à deux caractères : 0 et 1. De plus, nous faisons complète abstraction de l'architecture de la machine. Un code interprétable pour la machine devrait faire une référence précise aux emplacements où seront stockées nos variables et ne devrait pas se restreindre à les déclarer comme nous avons fait. Finalement, nous serions restreints aux opérations élémentaires qu'une machine sait faire (ex. : faire une addition, appliquer un circuit logique). Les langages informatiques qui restent à ce faible niveau d'abstraction sont dits de bas-niveau. Ce sont le langage machine qui est une chaîne de 0 et 1 appelée bits et le langage assembleur dans lequel les bits sont représentés par des symboles plus proches du langage naturel.

Pour des raisons évidentes, les langages de bas-niveau sont peu faciles à utiliser. Il existe par opposition des langages dits de haut-niveau qui ressemblent plus au pseudocode que nous avons écrit (Figure 5). Nous proposons donc le même algorithme écrit dans un langage de haut-niveau nommé Python (Figure 6). Mis à part le fait qu'on commence à compter à partir de zéro, on remarquera une grande similitude avec le pseudocode. Une petite modification a cependant été apportée et nous reviendrons dessus d'ici peu.

```

1 # Definir fonction calculant le nombre de differences entre deux sequences
2 def calculer_differences(sequence1, sequence2):
3
4     differences = 0
5
6     for index in xrange(0, len(sequence1)):
7         if (sequence1[index] != sequence2[index]):
8             differences = differences + 1
9
10    return differences
11
12 # Appliquer la fonction a nos deux sequences
13 print calculer_differences("ACTAACCAAA", "AGTAACCATA")
```

Figure 6. Algorithme en Python pour calculer le nombre de différences entre deux séquences de taille identique

Nous avons déjà indiqué que les machines ne comprennent pas les langages de haut-niveau. Il y a effectivement un traducteur qui retranscrit le programme en langage machine (chaine de 0 et de 1). Ceux-ci procèdent soit par interprétation, soit par compilation. Les programmes interprétés, aussi appelés scripts, sont exécutés par un autre programme nommé interpréteur qui traduit les instructions une à une avant de les donner à la machine, pour qu'elles les exécutent au fur et à mesure. Le langage Python que nous avons utilisé est un exemple de langage interprété. C'est également le cas du R sur lequel repose le Zébu. Ce sont des langages qui sont généralement très lisibles. De plus, ils ont l'avantage d'être compatibles avec n'importe quelle machine ayant un interpréteur : on dit qu'ils sont portables. Au contraire, un langage compilé est traduit intégralement en langage machine par un programme nommé compilateur avant d'être exécuté. La séparation entre traduction et exécution fait que les programmes compilés sont plus rapides que les programmes interprétés. C'est le cas

du C et du FORTRAN connut pour leur vitesse d'exécution difficilement égalable. Cependant, les langages compilés sont généralement de plus bas-niveau (moins abstrait) que les langages interprétés, ce qui les rend moins lisibles. L'écriture d'un programme avec un langage compilé demande donc plus de temps qu'avec un langage interprété. De plus, un langage compilé n'est pas aussi portable qu'un langage interprété (ex. : un programme compilé sous Windows ne tournera pas sur Mac OS X). Nous commençons donc déjà à voir que tous les langages ne sont pas égaux et que le développeur devra faire un choix. Revenons maintenant sur notre script Python afin d'expliquer l'intérêt de la programmation fonctionnelle.

### 2.2.3. Intérêt de la programmation fonctionnelle

Dans notre script Python (Figure 6), nous avons défini en ligne 2 une fonction nommée « `calculer_differences` ». Une fonction est un sous-programme qui prend des entrées nommées arguments. Notre fonction prend pour arguments nos deux séquences. Après l'exécution d'une fonction, celle-ci nous retourne alors le résultat (ligne 10). Les fonctions ont l'intérêt de simplifier considérablement le code en réduisant sa taille et en augmentant sa lisibilité. Nous pouvons effectivement appeler notre fonction sous la forme d'une unique ligne de code (ligne 13). Cela permet également au programmateur de fonctionner avec un niveau d'abstraction plus élevé et donc de penser au problème qu'il essaie de résoudre et non pas au côté technique du programme uniquement.

La puissance des langages haut-niveau, autre que leur lisibilité, réside dans ces fonctions préimplémentées. Tout langage de haut-niveau contient des fonctions de base comme des fonctions de tri ou de recherche. Cependant, chaque langage est développé dans un but précis et des fonctions plus avancées seront ou non proposées. Par exemple, le langage R (58) a été développé pour faire des calculs statistiques. Il implémente donc un ensemble de techniques statistiques et graphiques. Nous retrouvons, par exemple, une fonction permettant de réaliser des tests statistiques préimplémentés dans le langage (ex : `t.test()` pour réaliser un test t de Student). En plus, les informaticiens ont tendance à partager le code source de leurs fonctions. Dans le cas du langage R, ces codes partagés portent le nom de *package*. Ceux-ci nous

permettant, grâce au travail des autres, de faire des calculs extrêmement complexes avec quelques lignes de codes seulement. Par exemple, nous prendrions sans doute beaucoup de temps à développer un programme implémentant des réseaux bayésiens. Cependant, nous prendrions beaucoup moins de temps à télécharger un *package* R comme *bnlearn* (59) et à apprendre à nous en servir.

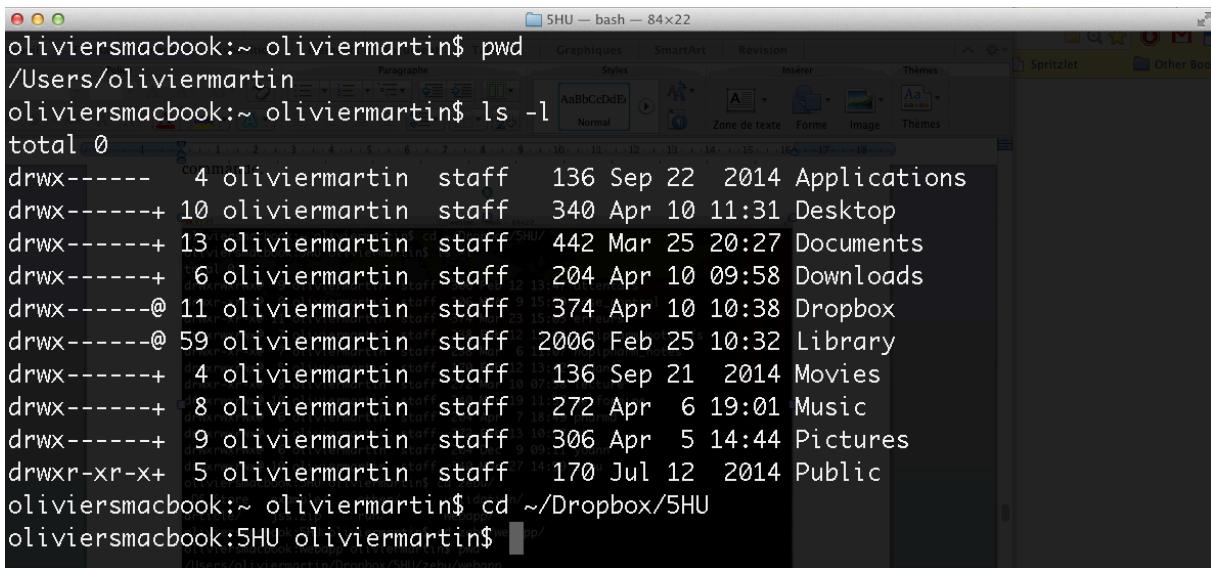
#### 2.2.4. Choix du langage de programmation

De nombreuses autres nuances entre les langages pourraient être décrites et discutées. Nous pourrions par exemple parler longuement sur les paradigmes de programmation (terme technique pour parler des « styles » de programmation) ou distinguer les langages internet avec les notions de « côté serveur » et « côté client ». Nous épargnons le lecteur d'une telle lecture. Cette partie a pour but de montrer au lecteur que tous les langages ont été développés dans un but différent et qu'ils possèdent ainsi des avantages et des inconvénients qu'il faut connaître. Le choix du ou de la combinaison de langage(s) est donc d'une importance capitale dans le développement d'un programme. Nous avons choisi pour le Zébu le langage R. Celui-ci est effectivement conçu pour les calculs statistiques. De plus, un *package* nommé *shiny* (60) permet au programmateur de réaliser facilement des applications web.

#### 2.2.5. Interfaces utilisateurs : interfaces graphiques et lignes de commandes

L'interface utilisateur est le moyen par lequel l'utilisateur interagit avec la machine. Il est possible de distinguer deux grandes formes d'interfaces : les interfaces graphiques et les interfaces en lignes de commandes. Le lecteur sera sans doute familier avec les interfaces graphiques qui sont présentes dans tous les systèmes d'exploitation modernes. Celles-ci sont caractérisées par l'usage de la souris, la présence de fenêtres et la possibilité de faire des actions comme le « glisser-déposer » appelé *drag and drop* par les Anglo-saxons. Elles sont caractérisées par leur ergonomie, leur esthétique et leur intuitivité d'utilisation. Un programme ayant une interface graphique sera donc généralement qualifié de *user-friendly*. Au contraire, une interface à ligne de commande est plus dure d'utilisation et peut effrayer les non-initiés. L'utilisateur communique avec la machine par l'intermédiaire d'un langage informatique. Par exemple, pour communiquer avec les systèmes UNIX comme Linux

et Mac OS X, on utilise le bash (Figure 7). Il est également possible d'utiliser des langages interprétés comme le Python ou le R. Ces interfaces ont l'avantage de rendre facile de réaliser de nombreuses tâches sans intervention de l'utilisateur : on parle de traitement par lots ou de *batch processing*. Imaginons, par exemple, que nous voulons renommer une centaine de fichiers. Cela prendra uniquement une ligne de commande, contrairement à plusieurs minutes d'opération répétitives avec une interface graphique. Le programmeur doit donc choisir entre une interface graphique et un ensemble de fonctions (comme nous avons précédemment défini) qui peuvent être appelées par une interface en ligne de commande. Un des objectifs du logiciel est de stimuler l'usage des mesures de l'association locale. Il nous semblait donc plus logique d'utiliser une interface attrayante pour l'utilisateur. Nous avons donc choisi une interface graphique.



```

oliviersmacbook:~ oliviermartin$ pwd
/Users/oliviermartin
oliviersmacbook:~ oliviermartin$ ls -l
total 0
drwx----- 4 oliviermartin staff 136 Sep 22 2014 Applications
drwx-----+ 10 oliviermartin staff 340 Apr 10 11:31 Desktop
drwx-----+ 13 oliviermartin staff 442 Mar 25 20:27 Documents
drwx-----+ 6 oliviermartin staff 204 Apr 10 09:58 Downloads
drwx-----@ 11 oliviermartin staff 374 Apr 10 10:38 Dropbox
drwx-----@ 59 oliviermartin staff 2006 Feb 25 10:32 Library
drwx-----+ 4 oliviermartin staff 136 Sep 21 2014 Movies
drwx-----+ 8 oliviermartin staff 272 Apr 6 19:01 Music
drwx-----+ 9 oliviermartin staff 306 Apr 5 14:44 Pictures
drwxr-xr-x+ 5 oliviermartin staff 170 Jul 12 2014 Public
oliviersmacbook:~ oliviermartin$ cd ~/Dropbox/5HU
oliviersmacbook:5HU oliviermartin$ 
```

Figure 7. Interface en ligne de commande utilisant le langage bash pour contrôler une machine tournant sur Mac OS X

Les commandes utilisées sont : « *pwd* » signifiant *print working directory* et permettant d'imprimer le répertoire de travail ; « *ls* » signifiant *list* et permettant de lister le contenu d'un répertoire ; « *cd* » signifiant *change directory* et permettant de changer de répertoire de travail.

## 2.3. Licences logiciels

### 2.3.1. Définition et classification

Une licence spécifie les droits d'utilisation d'un logiciel (61). Il peut s'agir d'un contrat entre le donneur et le prenant comme les « conditions d'utilisation » que nous devons accepter avant d'installer le logiciel. Il peut également être question simplement d'une notice qui indique les droits de l'utilisateur. On distingue deux types de licences : propriétaire et libre. Les licences libres donnent à l'utilisateur les quatre libertés fondamentales numérotées de 0 à 3 (Tableau 2) (62). A l'inverse, une licence propriétaire est une licence qui déroge à au moins une de ces libertés.

0. La liberté d'exécuter le programme comme vous voulez, pour n'importe quel usage (liberté 0) ;
1. La liberté d'étudier le fonctionnement du programme, et de le modifier pour qu'il effectue vos tâches informatiques comme vous le souhaitez (liberté 1) ; l'accès au code source est une condition nécessaire ;
2. La liberté de redistribuer des copies, donc d'aider votre voisin (liberté 2) ;
3. La liberté de distribuer aux autres des copies de vos versions modifiées (liberté 3) ; en faisant cela, vous donnez à toute la communauté une possibilité de profiter de vos changements ; l'accès au code source est une condition nécessaire.

Tableau 2. Quatre libertés fondamentales du logiciel libre

### 2.3.2. Intérêt de l'open source en sciences de la vie

Nous avons précédemment argumenté que l'absence d'outils disponibles pouvait retarder considérablement l'avancée scientifique. Les licences libres facilitent l'accès aux outils et sont donc d'un intérêt scientifique majeur. On se rappellera d'une citation de l'informaticien Alfred N. White : « *Civilization advances by extending the number of important operations that we can do without thinking about them.* ». De plus, en rendant le code source disponible, le fonctionnement des logiciels libres devient transparent. Il est possible d'étudier le logiciel et de s'assurer que les résultats sont sans erreurs. On rentre dans une logique de reproductibilité qui est fondamentale à la science. Finalement, nous rappelons que les articles décrivant des logiciels font partie des articles les plus cités. Ceux-ci ont donc une influence considérable sur la réputation de l'auteur. Il existe ainsi de nombreuses bibliothèques de codes sources

comme le CRAN ou Bioconductor pour les *packages R* que nous avons déjà cités. On retrouve également des initiatives pour développer des langages informatiques open source comme R ou comme Biopython (63) qui comporte de nombreuses fonctions utiles aux bioinformaticiens. Le Zébu est un autre exemple de logiciel libre. Effectivement, il possède une licence GNU et son code source est intégralement téléchargeable sur <https://github.com/olivmrtn/Zebu>.

Similairement aux logiciels open source, on parle d'*open dada* pour désigner une base de donnée garantissant les quatre libertés. Ceci est cependant rarement observé dans le domaine compétitif du biomédical. Les données d'un groupe de recherche restent généralement confinées à celui-ci et rares sont les articles accompagnés des données exploitées. Cette conception est opposée à celle de scientifiques comme les biologistes. Effectivement, dans ces milieux, les bases de données fleurissent régulièrement et depuis longtemps. Par exemple, la *Protein Data Bank* existe depuis les années 1970 et regroupe aujourd'hui plus de 100 000 structures de protéines. De plus, tous les journaux de cristallographie obligent leurs auteurs à publier la structure nouvellement découverte dans la *Protein Data Bank* avant de leur soumettre un article (56). Cette approche open-data permet d'accélérer la cadence des découvertes scientifique comme fait le logiciel open source. De plus, elle semble être la base de la notion de la reproductibilité si importante à la science (64). Pour cela certains journaux médicaux comme *Annals of Internal Medicine* ont commencé à promouvoir cette mentalité de partage (65).

Cependant, l'approche de l'*open data* n'est pas sans limites dans le domaine biomédical. Le problème de la confidentialité des patients est réel, notamment pour les maladies rares (64). Il est possible que des solutions existent avec l'utilisation de données synthétiques qui permettraient l'anonymisation des données (66). Nous retrouvons ici l'intérêt des modèles mathématiques. Effectivement, à partir des modèles mathématiques construits, il est possible de simuler (ou synthétiser) de nouveaux patients. Ces patients synthétiques ne correspondent plus à aucun patient réel, mais représentent certaines de leurs caractéristiques qui peuvent être étudiées. Ces méthodes restent cependant non appliquées de nos jours en sciences de la vie.

### 3. Applications des mesures d'association locales

Ce que nous venons de décrire au cours de ces deux dernières parties est incomplet sans une description des champs d'applications des mesures d'association locale. Effectivement, comme le dirait Richard Hamming : « *The purpose of computing is insight, not numbers.* ». Les approches locales sont récentes et sont encore peu connues des cliniciens et scientifiques. Effectivement, l'analyse des résiduels du chi-deux pourrait être datée à 1973 (67), l'information mutuelle spécifique de 1990 (48) et le Z de Ducher de 1992 (7) et les *Local Indicators of Spatial Association* de 1995 (49). De plus, le calcul des mesures locales résulte en une matrice de valeurs riche en informations. L'interprétation est plus difficile que celle d'une valeur unique, ce qui limite leur usage face aux méthodes globales. Par ailleurs, le calcul de ces matrices est lourd à réaliser manuellement et on remarque donc que les textes d'origine font référence à l'essor des méthodes computationnelles (48,49,68). Finalement, l'approche locale semble être fondamentalement opposée à l'approche globale plus communément utilisée. C'est notamment le cas dans la médecine contemporaine, qui recherche à décrire ou agir sur les populations hétérogènes par des valeurs uniques. L'approche que nous prônons semble donc opposée au paradigme contemporain biomédical.

Bien que leur application reste marginale, il est intéressant de remarquer que celle-ci concerne des domaines qui semblent très éloignés les uns des autres. C'est notamment la physiologie cardiovasculaire avec l'étude du baroréflexe (5,41,52–56), la pharmacocinétique avec l'étude de la relation entre les paramètres du modèle (74,75), la linguistique computationnelle avec l'extraction des colocations (48,50,76), l'analyse d'image avec l'extraction des contours d'une image (47) et la géographie avec l'analyse spatiale et l'identification des agrégats (*clusters*) (49,77). Cette diversité traduit le fait que les problèmes rencontrés dans différentes disciplines peuvent être en réalité très proches. Ici, le problème commun est la discontinuité de l'association entre les variables. Par exemple, les relations à effet de seuil lors l'excitabilité neuronale. Le passage du seuil entraîne l'apparition d'une association précédemment inexistante entre les variables. L'association locale n'est pas continue entre les

différentes modalités des variables et les mesures locales permettent de mettre en évidence cette discontinuité. Nous décrirons succinctement l'ensemble de ces applications.

### 3.1. Physiologie cardiovasculaire

Le Z de Ducher a été conçu par des physiologistes pour être adapté à la complexité du vivant. Effectivement, les auteurs avaient remarqué que les méthodes basées sur la variance ne détectaient que les relations linéaires et n'apportaient qu'une information globale (7). Or, la relation entre pression artérielle systolique et fréquence cardiaque n'est pas linéaire. Effectivement, sa régulation est assurée en partie par deux systèmes antagonistes : le système sympathique qui augmente la pression artérielle ; le baroréflexe qui inhibe cette augmentation. Il en résulte une relation en forme de X (46). La méthode a, d'abord, été utilisée pour quantifier l'association locale entre les différentes modalités de la pression artérielle et de la fréquence cardiaque. Elle a également montré qu'elle pouvait quantifier conjointement le rôle des deux systèmes de régulation sur la variabilité de la pression artérielle (46). La méthode a alors été utilisée pour mesurer de manière non-invasive la sensibilité du baroréflexe chez le rat (69), puis chez l'Homme (78). Ceci constitue une avancée, car les méthodes utilisées jusqu'alors nécessitaient l'injection de drogues vaso-actives. Outre l'aspect non invasif, ces résultats sont d'intérêt, car une diminution de la sensibilité au baroréflexe est associée avec des états d'altération du système nerveux autonome comme l'hypertension ou le diabète. La méthode permettrait ainsi un dépistage précoce en clinique. La méthode a été également utilisée pour étudier la relation entre la consommation de sel et la pression artérielle, et a montré que cette relation ne concernait sans doute qu'une sous-population sensible au sel (73). La méthode a eu d'autres applications concernant l'étude de la physiopathologie de l'hypertension chez l'animal et chez l'homme, que nous ne décrirons pas ici (55,70–72)

### 3.2. Pharmacocinétique

Le Z de Ducher a également été utilisé pour l'étude des relations entre les paramètres pharmacocinétiques décrivant le comportement d'un médicament en fonction des caractéristiques des patients. L'ensemble des études s'est concentré sur l'amikacine, un aminoside utilisé comme antibiotique. Les relations étudiées sont celles entre 1. clairance de l'amikacine avec l'âge (74), ainsi que 2. le volume apparent de distribution de l'amikacine avec le poids du patient (75). Ces études concluent que les relations entre les paramètres pharmacocinétiques et les variables biologiques et anthropométriques ne sont pas applicables à tous les patients et que d'autres variables doivent être utilisées pour mieux modéliser la pharmacocinétique de l'amikacine.

### 3.3. Linguistique computationnelle

En linguistique, une colocation est une association arbitraire de mots qui apparaît plus souvent qu'attendue par le hasard (79). Des exemples sont « argument ferme » ou « voix suave ». Elles mettent en évidence les relations qu'entretiennent les mots entre eux. Les mesures de l'association locale, en particulier, l'information mutuelle spécifique permet de les extraire automatiquement d'un corpus de texte (48,50,76). Ceci s'explique bien par le fait que ces mesures sont normalisées par la fréquence qui serait attendue par le hasard (l'indépendance statistique) et donc indépendante de la fréquence observée du mot dans le corpus. De plus, les relations qu'entretiennent les mots entre eux sont discontinues justifiant une approche locale.

### 3.4. Analyse d'image

L'information mutuelle spécifique a été employée pour extraire automatiquement les contours des différents objets présents sur une même image numérisée. Effectivement, les pixels appartenant à un même objet ont une forte association statistique (47). Les auteurs affirment que leur méthode a une performance bien plus élevée que les méthodes précédentes basées sur des caractéristiques globales de l'image. Ceci ne semble pas étonnant quand on se rend compte de la discontinuité qui existe entre deux objets.

### 3.5. Géographie et analyse spatiale

La géographie inclue dans son étude l'analyse de l'association spatiale, c'est-à-dire, l'étude de la covariance de caractéristiques (ex : sociales, démographiques) avec l'espace étudiée (ex : pays, région). Pour cela, on retrouve l'utilisation de mesures de l'association spatiales. Celles-ci peuvent être globales en mettant en évidence une tendance générale. Elles peuvent aussi être locales et portent alors le nom de *Local Indicators of Spatial Association* ou LISA (49). Elles ont notamment pour but de mettre en évidence des agrégats (*clusters*) locaux géographiques, c'est-à-dire, d'identifier les régions qui contribuent le plus à la tendance globale.

### **III. Travail personnel**

#### **1. Conception d'un outil informatique implémentant le Z de Ducher et l'information mutuelle spécifique : Zébu.**

Le logiciel Zébu est décrit dans l'article scientifique qui suit. Celui-ci adopte un vocabulaire plus technique que le reste de cette thèse. Il utilise des notions générales abordées lors de la première partie.

Les articles scientifiques décrivant des logiciels adoptent des plans atypiques : ils ne suivent généralement pas le modèle IMRED. Il en est de même pour notre article qui est divisé comme suit :

1. Introduction
2. Description des méthodes
3. Guide d'utilisation
4. Exemple d'utilisation
5. Discussion - Conclusion

# Zebu: A Web Application Designed for Computation of Local Association Measures

Olivier MF Martin Michel Ducher

## Abstract

### Background

Global association measures, such as Pearson's r or Cramer's V, suppose that the strength of association is identical for all modalities of variables. This assumption does not hold for discontinuous relationships such as threshold mechanisms frequently found in biology and health sciences. Local association measures quantify association for modalities of variables. They allow a better description of association by taking into account valuable local information. Nonetheless, software actually available only allows computation of global measures.

### Results

Zebu is a user-friendly web application that computes both global and local forms of multi-information (a multivariate generalization of mutual information) and Ducher's Z. It is written in the R language using the shiny web application framework. It was designed with simplicity in mind for scientists and clinicians with no programming knowledge. It is provided free-of-charge under a GLP-3 license and can be accessed directly online at [olivmrtn.shinyapps.io/Zebu](http://olivmrtn.shinyapps.io/Zebu) or as a standalone application with a simple R command ([github.com/olivmrtn](https://github.com/olivmrtn)). Results and graphics can be downloaded and included in publications. Moreover, we show that local association measures can be used as a criterion for subgroup analysis.

### Conclusion

Zebu is the first available software to compute local association measures and thus fills an unmet need. It is of interest to a wide range of scientific disciplines and can be used by anyone with an internet connection.

**Keywords:** measures of association, subgroup analysis, pointwise mutual information, Ducher's Z, R, shiny, web application

## Introduction

Science is concerned with the explanation of phenomena. This involves establishing causal relationships between variables. However, the underlying causes are not directly observable. To reveal them, one must conduct carefully planned experiments to observe the consequences of causal processes: complicated patterns of independence and association between variables. Although association is not a sufficient condition to establish causation, it can be used as a guide for further investigation. Indeed, causation always implies some pattern of association [Shipley, 2000]. For this reason, the nature and strength of correlations have to be thoroughly described and measured. Measures of association are thus of interest. Consequently, numerous tests and measures of association, such as Pearson's  $r$  and the chi-squared test of independence, have been described in the literature.

However most of them only provide a global measure of association. They quantify association between variables for all possible modalities of the variables and suppose that the strength of association is uniform. This assumption does not hold for discontinuous relationships such as threshold mechanisms where the relationship only becomes valid at a certain critical value. Local association measures allow to better describe association between variables. These quantify association for modalities of the variables. Examples are pointwise mutual information and Ducher's  $Z$ . Although software and R packages are readily available to compute global association measures (e.g. `infotheo` [Meyer, 2014]), computation of local association measures rely on *ad hoc* methods.

`Zebu` is an association measure calculator written in R [R Core Team, 2014]. The shiny package [RStudio and Inc., 2014] makes it a user-friendly web application. Users can access software directly online ([olivmrtn.shinyapps.io/Zebu](http://olivmrtn.shinyapps.io/Zebu)). This means no hazardous installation processes or previous knowledge of R is required. More experienced users can also download the source code from the author's Github [github.com/olivmrtn](https://github.com/olivmrtn) account and run it locally.

Three data analysis modules were included in `Zebu`. The first module concerns basic data preprocessing namely the discretization of continuous variables. The next module allows computing both global and local association measures. We have implemented multi-information (a multivariate generalization of mutual information) and Ducher's  $Z$ . Results are displayed using the `ggplot2` [Wickham, 2009] and `latticeExtra` [Sarkar and Andrews, 2013] packages making them elegant and easily interpretable. The last module is concerns subgroup analysis. Two subgroups of are formed on the basis of local association measures: the positively associated and the negatively associated subgroups. The distribution of a variable not included in the association model is then compared between subgroups. This helps better characterize individuals for which variables are associated. `Zebu` is thus aimed at simplifying exploratory analysis and facilitating hypothesis generation.

We will start by providing a brief overview and comparison of association measures. Here we will more deeply explain the choice of implemented measures in `Zebu`. Next we will describe how these measures are computed and describe how these can be used to define subgroups. Then we will describe how to run the software and provide an example using simulated data concerning cancer. Finally we will discuss the use of local association measures and detail limitations of measures and software.

## Background

### Overview of measures of association

Intuitively, in the bivariate case, association can be interpreted as a reduction in uncertainty in one variable by knowing another. Complete association is the case where one variable completely determines the other. Absence of association implies that knowing one doesn't tell us anything about the other. Both of these extreme situations are rarely observed when studying real samples. Association measures are here to bring a more subtle definition. Here we will try to give an overview of different manners of measuring association and explain why multi-information and Ducher's Z were implemented.

Measures of association are quite numerous in the scientific literature. Nonetheless, they share similarities between each other. It is possible to classify them according to the following criterions: theoretical foundations, continuous vs. categorical variables, bivariate vs. multivariate generalization, parametric vs. non-parametric, global vs. local, normalized vs. non-normalized.

Measures of association are based either on variance or probability. Variance-based methods try to quantify the amount of variance of one variable explained by the other. Since variance is a continuous concept, these measures are only applicable to continuous variables. Examples are Pearson's r and Spearman's rho. Probability-based methods are based on the statistical definition of independence. They measure the deviation of observed probability from a theoretical probability for which variables are independent. They are generally only applied to discrete variables mainly for computational efficiency. Examples are Cramer's V, proportional reduction in error, mutual information and Ducher's Z.

Association is often measured only between two variables. This kind of analysis fails in more complex relationships involving interaction between variables [Jakulin and Bratko, 2003]. An example is the XOR gate where looking at bivariate relations provides no evidence of association although variables are clearly linked. Association is not a concept restricted to bivariate relationship. Multivariate association measures exist and should be considered in order to detect interactions between variables.

Parametric measures of association often suppose linear or at least monotone relationships. These are namely methods based on the proportion of explained variance. Although intuitive and convenient, this assumption is not always justified in biology and health sciences. Indeed, living systems often have non-linear dynamics such as thresholds and saturation. An alternative would be to use more robust non-parametric methods such as those based on statistical independence.

Most measures of association were developed for predictive purposes and provide only a global measure of association. They quantify association for all possible modalities of variables and thus neglect valuable local information. Local association measures quantify association for modalities of variables. They allow pinpointing regions of association and thus give a more precise picture of the relationship between variables. Furthermore, global measures suppose that the strength of association is uniform for all possible modalities of the variables. This assumption does not hold for discontinuous relationships. Local association measures are thus of interest. This is best illustrated

from their present field of application. Pointwise measures has been used in computational linguistics for collocation extraction [Van de Cruys, 2011] and in image processing for border detection [Isola et al., 2014]. Ducher's Z as been used in cardiology to assess the baroreflex sensitivity by non-invasive means [Cerutti et al., 1995, Sapoznikov et al., 2013], in pharmacokinetics to study the relation between antibiotic elimination and ageing [Ducher et al., 2001] and in epidemiology to study the relation between salt consumption and blood pressure [Ducher et al., 2003].

Normalized variants of measures often exist. Their goal is to allow comparisons of association for different variables or modalities of variables. By convention, global measures take values from 0 to +1 inclusive or from -1 to 1 inclusive. Independence implies a null association. Total dependence implies an absolute value of 1. In the case of a normalization going from -1 to 1, the sign of the measure defines the direction of the relationship [Goodman and Kruskal, 1979]. Local measures are normalized between -1 to 1 inclusive. While independence still implies a null association the interpretation of non-null values is different. Events that happen less then expected by independence have negative association. Similarly, events that happen more then expected by independence have positive association. Normalized pointwise measures are widely unused. This approach dates back to the beginning of the 90s with PMI<sup>k</sup> [Daille, 1994] and Ducher's Z [Ducher et al., 1992]. It is interesting to note that Ducher's Z is the only normalized multivariate local associate measure we are aware of.

Methods implemented in Zebu are multi-information and Ducher's Z. These are measures based on statistical independence making them applicable for categorical variables as well as continuous variables after discretization. They are well adapted for biology and health science research. Indeed they do not suppose any pattern of association and are valid all sorts of relationships. Non-linear, non-monotonous and discontinuous relationships are not an exemption in this area of research and the adapted tools must be readily available. Furthermore, they have global as well as local measures of association allowing a better understanding of the relationship. Finally, they have multivariate generalizations which allow more complex relationships to be analyzed. Since both of these measures are based on statistical independence, a brief review will be given before these measures are further described.

## Statistical independence

One way to think about association is as events co-occurring. For example, if event A always shows up with event B, then event A is associated with event B. An intuitive measure of association could be the joint probability  $p(A, B)$ . However, this measure fails for rare events. Joint probability of rare events are always as small as events are rare. As a consequence, it is necessary to compare the observed joint probabilities to theoretical probabilities in which the variables are considered independent. This theoretical model is based on the statistical definition of independence defined below.

Let  $\mathbf{X}$  be a N-dimensional random variable vector such as  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ . Modalities are represented by  $x_n$  where  $n$  is the  $n$ 'th random variable. The observed joint probability for one modality is represented by  $p(x_1, x_2, \dots, x_n)$ . The expected probability if events were independent is the factor of marginalized probabilities:  $p(x_1)p(x_2) \dots p(x_n)$ . These will be respectively referred to as observed and expected probabilities. Independence is then defined by the following mathematical relation.

$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2) \dots p(x_n)$$

Independence implies that knowing one or more variables does not give us any information about the others. This is exactly what no one is interested in. It is however possible define two interesting cases where the former equality does not hold: positive association and negative association. Positive association is defined to be events showing up more often than expected. Local association measures are positive. Negative association is defined to be events showing up less often than expected. Local association measures are negative.

Positive association

Negative association

$$p(x_1, x_2, \dots, x_n) > p(x_1)p(x_2) \dots p(x_n) \quad p(x_1, x_2, \dots, x_n) < p(x_1)p(x_2) \dots p(x_n)$$

## Multi-information

Mutual information is founded on information theory. It is most commonly measured in bits. Many multivariate generalization of mutual information exist. The most known are interaction information [McGill, 1954] and multi-information (also known as total correlation) [Watanabe, 1960]. Contrary to interaction information, multi-information is always positive. In Zebu, multi-information measured in bits was implemented. Multi-information  $MI$  is computed in the following manner.

$$MI(X_1, X_2, \dots, X_n) = \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \dots \sum_{x_n \in X_n} p(x_1, x_2, \dots, x_n) \log_2 \left( \frac{p(x_1, x_2, \dots, x_n)}{p(x_1)p(x_2) \dots p(x_n)} \right)$$

Its maximal value  $MI_{max}$  can be used to defined a normalized measure  $NMI$  bounded by  $[0, 1]$ . This maximal value is dependent on the joint entropy  $H$ .

$$H(X_1, X_2, \dots, X_n) = - \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \dots \sum_{x_n \in X_n} p(x_1, x_2, \dots, x_n) \log_2(p(x_1, x_2, \dots, x_n))$$

$$MI_{max}(X_1, X_2, \dots, X_n) = \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \dots \sum_{x_n \in X_n} p(x_1, x_2, \dots, x_n) H(x_1, x_2, \dots, x_n) - \max(H(x_1, x_2, \dots, x_n))$$

$$NMI(X_1, X_2, \dots, X_n) = \frac{MI(X_1, X_2, \dots, X_n)}{MI_{max}(X_1, X_2, \dots, X_n)} \in [0, 1]$$

It is also possible to define a local measure of association known as pointwise multi-information  $pmi$  [Van de Cruys, 2011].

$$pmi(x_1, x_2, \dots, x_n) = \log_2\left(\frac{p(x_1, x_2, \dots, x_n)}{p(x_1)p(x_2)\dots p(x_n)}\right)$$

Bouma [Bouma, 2009] proposed one normalization of pointwise mutual information  $npmi$  bounded by  $[-1, 1]$ . It is only valid in the bivariate case.

$$npmi(x_i, x_j) = \frac{pmi(x_i, x_j)}{-\log_2(p(x_i, x_j))} \in [-1, 1]$$

Computing a confidence interval and p-value for pointwise mutual information is of interest to access statistical significance. Its distribution is however currently unknown making it impossible to compute exact values. It is nonetheless possible to approximate these using the bootstrap method. This is how these values are computed in Zebu. P-value is defined as  $Pr(pmi \leq 0)$  in case of positive association and  $Pr(pmi \geq 0)$  in case of negative association of events. False discovery rate is controlled using the Benjamini and Hochberg method [Benjamini and Hochberg, 1995]. This is equally applicable to its normalized variant  $npmi$ . A confidence interval can also be computed for global measures  $MI$  and  $NMI$ .

## Ducher's Z

Of all association measures, one of the least known is Ducher's Z. Contrary to all other measures who were first developed as a global measure of association with local measures only latter derived, Ducher's Z was developed as a local measure. To the extent of our knowledge, when it was first published in 1992 [Ducher et al., 1992], it was the only normalized local association measure that existed. Its application have been restricted to the fields of cardiovascular, epidemiology and pharmacokinetics research [Ducher et al., 2001, Ducher et al., 2002, Ducher et al., 2003, Sapoznikov et al., 2013, Rughoo et al., 2014].

Ducher's Z is a normalized measure of deviation from independence. Deviation from independence  $\Delta p$  is defined in the following manner.

$$\Delta p(x_1, x_2, \dots, x_n) = p(x_1, x_2, \dots, x_n) - p(x_1)p(x_2)\dots p(x_n)$$

For positive association,  $\Delta p$  is maximized when  $p(x_1, x_2, \dots, x_n) = \min(p(x_1), p(x_2), \dots, p(x_n))$ . For negative association,  $\Delta p$  is minimized when  $p(x_1, x_2, \dots, x_n) = 0$ . In case of independence, Ducher's Z is defined to be null. Normalization of  $\Delta p$  is thus achieved by the following system of equations.

$$\begin{aligned} \forall \Delta p > 0 \quad Z &= \frac{\Delta p(x_1, x_2, \dots, x_n)}{\min(p(x_1), p(x_2), \dots, p(x_n)) - p(x_1)p(x_2)\dots p(x_n)} \in ]0, 1] \\ \forall \Delta p < 0 \quad Z &= \frac{\Delta p(x_1, x_2, \dots, x_n)}{p(x_1)p(x_2)\dots p(x_n)} \in [-1, 0[ \\ \forall \Delta p = 0 \quad Z &= 0 \end{aligned}$$

Similarly to pointwise mutual information, computing a confidence interval and p-value for local Z is of interest to access statistical significance. The Z distribution is also unknown. It is thus impossible to compute exact values. It is however possible to approximate these using the bootstrap method. This is how these values are computed in Zebu. P-value is defined as  $Pr(Z \leq 0)$  in case of positive association and  $Pr(Z \geq 0)$  in case of negative association of events. False discovery rate is controlled using the Benjamini-Hochberg method.

Although a global measure has never before been defined, it is possible to use its expected value  $gZ$ . This is similar to the manner mutual information is computed from pointwise mutual information. A confidence interval of this measure can also be compute by bootstrap.

$$gZ(X_1, X_2, \dots, X_n) = \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \dots \sum_{x_n \in X_n} p(x_1, x_2, \dots, x_n) Z(x_1, x_2, \dots, x_n)$$

## Subgroup analysis of association

### An introductory example: salt consumption, blood pressure and salt-resistance

Before attempting to formalize this methodology, an illustrative example concerning salt consumption and blood pressure will be discussed. This example is widely inspired from Ducher [Ducher et al., 2003]. Blood pressure is thought to be linearly related to salt consumption. However evidence supporting this association of variables is widely contradictory [Freedman and Petitti, 2001]. This suggests that a global relationship may not be applicable to all individuals but rather only to a subgroup of salt-sensitive individuals. These are to be opposed to salt-resistant individuals for whom no relationship can be established [Kaplan, 2010, p. 57]. Global association measures may not be sensitive enough because salt-resistant individuals completely dilute the association that exists for salt-sensitive individuals.

Local association measures allow to quantify association for modalities of salt consumption and blood pressure. These modalities are either positively or negatively associated. Individuals can be classified into two corresponding subgroups accordingly. Positively associated modalities constitute the subset of modalities that are well explained by the global association of variables (*e.g.* low blood pressure and low salt consumption). The corresponding subgroup will thus be composed individuals statistically sensitive to salt. Negatively associated modalities constitute the subset of modalities badly explained by the global relationship (*e.g.* low blood pressure and high salt consumption). The corresponding subgroup will thus be composed of individuals statistically resistant to salt. The defined subgroups can then be compared in order to determine what distinguishes salt-sensitive from salt-resistant individuals (*e.g.* genetics).

## A more formal definition

Let  $\mathbf{X}$  be a N-dimensional random variable vector such as  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  (*e.g.* salt consumption and blood pressure) and  $A(\mathbf{X})$  be a local measure of association. Accordingly, two subsets of  $\mathbf{X}$  can be defined. The first is the positive association subset  $\mathbf{P} \subset \mathbf{X}$  for which  $A(\mathbf{X}) > p$  where  $p > 0$  (*e.g.* salt-sensitive). The second is the negative association subset  $\mathbf{N} \subset \mathbf{X}$  for which  $A(\mathbf{X}) < n$  where  $n < 0$  (*e.g.* salt-resistant). In Zebu, association thresholds  $p$  and  $n$  are parameters defined by the user.

The distributions of a variable not included in the model are then compared between subgroups  $\mathbf{P}$  and  $\mathbf{N}$ . Let  $Y_i$  be a random variable not included in the association model  $A(\mathbf{X})$ . If  $Y_i$  is independent from  $\mathbf{X}$  then  $Y_i$  is independent from both  $\mathbf{P}$  and  $\mathbf{N}$ .

$$\begin{aligned} p(Y_i|\mathbf{P}) &= p(Y_i) \\ p(Y_i|\mathbf{N}) &= p(Y_i) \end{aligned}$$

This implies that no difference in distribution of  $Y_i$  should be observed between subgroups.

$$p(Y_i|\mathbf{P}) = p(Y_i|\mathbf{N})$$

However if  $Y_i$  is dependent on  $\mathbf{X}$  then this last equation does not hold and a difference in distribution may exist. This difference allows us to distinguish individuals for which the global association exists from individuals where it does not. In other words, it allows to characterize the population in which the global association exists. Accessing differences in distribution of  $Y_i$  will depend on whether it is continuous or categorical. In Zebu, continuous variables are plotted as box plots and difference in means is tested by the Mann-Whitney U test. Categorical variables are plotted as tile plots and differences will be tested by the chi-squared independence test.

## Program handling

### Availability

The program is available as a web application at [olivmrtn.shinyapps.io/Zebu](http://olivmrtn.shinyapps.io/Zebu). Source code can be downloaded at [github.com/olivmrtn](https://github.com/olivmrtn) and runned locally. This can be pipelined by following command in the R console after installation of shiny.

```
shiny:::runGitHub('Zebu', 'olivmrtn')
```

### Import file

Supported file formats are CSV, TSV and XLSX files. The user must specify the file path, characteristics (*i.e.* format, column and decimal separator) and press the Import button. If the file is correctly imported, a dynamic jQuery data table will be displayed. If this fails, Zebu will display general information explaining how the data file should be formatted.

## Preprocessing

This panel allows basic preprocessing of the data. User have can edit variable names, decide how to handle missing data and how to discretize continuous variables. If the data is correctly preprocessed, a dynamic jQuery data table will appear. Changes are only applied after pressing the "Apply" button.

### Change names

Variables names are editable. White spaces are allowed. Missing or redundant variable names are not allowed.

### Missing data

Computing probabilities implies having no missing data. The number and proportion of missing cells is displayed. Missing data is handled by data deletion. There are two manners to proceed: listwise and pairwise deletion. In listwise deletion, a record is fully dropped if it contains at least one missing variable. The dataset is consistent, it does not depend on the specified variables (*i.e.* variables used to compute association measures). In pairwise deletion, a record is dropped if at least one of the specified variables is missing. The dataset can change according to the specified variable. This method allows to use more data.

### Variable kind

Three kinds of variables are defined in Zebu. Continuous variables are numeric variables with at least 10 unique values. These must be discretized. Categorical variables are numeric or text variables with less than 10 unique values. Variables with non-numeric entries and with more than 10 unique values are excluded from analysis. Variables that should not be analyzed (*e.g.* identification variable) can be excluded.

### Discretization

Discretization is the transformation of continuous variables into discrete variables. In other words, a variable that can take an infinite number of values (*e.g.* age) is partitioned into a finite number of bins (*e.g.* child, adult and elderly). This is a necessary step to compute the implemented association measures. Two ways of discretization were implemented: equal-width and user-defined. In equal-width discretization, bins have the same width and only the number of breaks must be defined using the slider. The minimum number of breaks is set to 2 and maximum to 10. To use user-defined discretization, the box next to the variable must be checked. A text input will then appear. Break values must separated by a comma (*e.g.* 0,10,20,50 will generate bins [0,10], ]10,20], ]20,50])). If the user incorrectly formats breaks, an error message will appear in the concerned variable text input.

## Association

### Overview

This panel is used to compute probabilities and association measures. The user must select something to compute and at least two variables. All association measures have multivariate generalizations. It is thus possible to select as many variables as the user wants. Computation do however become more and more intensive.

The results can then be represented in different manners. For bivariate analysis: Results can be visualized in two different ways. The default representation is a tile plot using ggplot2. This is a more graphical representation of a confusion table. Tiles are colored from white to black representing the minimum and maximum of values of the bounds of the normalization or observed values. The user can also see results displayed as a latticeExtra 3D histogram.

No intuitive representation is implemented for multivariate analysis. Results are displayed as a dynamic jQuery datatable. It is possible to define the range of values taken by the association measure for which cases will be displayed.

Computed data and results can be downloaded. Data is exported as a CSV file. This file contains all variables, before and after preprocessing, as well as observed and expected probabilities and selected association measures. Plots are saved as PNG. Resolution can be modified in the "Options" panel.

### Observed and expected probabilities and samples sizes

In the bivariate case, no 3D histogram is available. Sample sizes are specified in parentheses under probabilities.

### Conditional probabilities

In the bivariate case, no 3D histogram is available. It is possible to choose which variables are conditional to all others using the input. Conditional probabilities may not be computable due to divisions by zero.

### Ducher's Z

Confidence intervals and p-values can be computed by bootstrap. Users have to choose to display one of them at the time. P-value is defined as  $Pr(Z \leq 0)$  in case of positive association and  $Pr(Z \geq 0)$  in case of negative association of events. False discovery rate is controlled using the Benjamini-Hochberg method. By default 1000 bootstrap samples and 95% confidence interval are created. This can be modified in the "Options" panel.

### Multi-information

In the bivariate case, the user can choose between pointwise multi-information and its normalized variant. In higher dimensions, only pointwise multi-information is computable. Pointwise multi-information may not be computable due to divisions by zero. Confidence intervals and p-values can be computed by bootstrap. Users have to choose to display one of them at the time. P-value is defined as  $Pr(pmi \leq 0)$  or  $Pr(np mi \leq 0)$  in case of positive association and  $Pr(pmi \geq 0)$  or  $Pr(np mi \geq 0)$  in case of negative association of events. False discovery rate

is controlled using the Benjamini-Hochberg method. By default 1000 bootstrap samples and 95% confidence interval are created. This can be modified in the "Options" panel.

## Subgroups

### Description

After an association measure has been computed in the association panel, the user can use these results for subgroup analysis. Individuals will be classified in two subgroups according to the local association measure: positive and negative association. Positive and negative association thresholds can be defined using corresponding sliders. The distribution of variables not included to compute the association measure can then be compared. Continuous variables are plotted as box plots and difference in mean is accessed with Mann-Whitney U test. Categorical variables are plotted as tiles plots and difference between subgroups is accessed by Pearson's chi-squared test. This allows to better characterize individuals in which the global association exists.

## Options

If the user is not satisfied with default options, he can modify them using this panel. He can define the confidence level and number of iterations the bootstrap algorithm will go through. Minimum value is set to 100 and maximum to 10 000. Increasing the number of iterations makes results more accurate but increases computation time. It is also possible to define the number of displayed decimals and the plot resolution in pixels per inch (ppi).

## Case study

### Simulation of data

To illustrate usage of Zebu, data simulated using a bayesian network will be used. The example presented in figure 1 is borrowed from Pearl [Pearl, 1988]. Five thousand cases with 5% of missing data was simulated using SamIAN. To illustrate how Zebu handles continuous variables, calcium concentrations were simulated. "Not increased" cases were randomly sampled from a normal distribution with mean 2.4 and standard deviation 0.05. "Increased" cases were randomly sampled from a negative exponential distribution with rate 10 to which the value 2.6 was added. An identification variable for each case was also added.

### Analysis of data

The first step is importing the dataset in the "Import file" panel. Once this has been achieved, data can be visualized (Figure 2).

Data must then be preprocessed in the "Preprocess" panel. Variables were renamed by replacing underscores by white spaces. Missing data is handled by pair-wise deletion. The "ID" variable is excluded from analysis. Calcium concentrations are discretized into two user-selected bins ]2.2, 2.6] and ]2.6,

3.4]. These are meant to reflect "non increased" and "increased" concentrations. Once these choices have been applied, the preprocessed dataset can be visualized (Figure 3).

Computation of probabilities and association measures now becomes possible in the "Association" panel. To illustrate this, the relationship between "Brain Tumor" and "Coma" will be studied. Ducher's Z between these two variables was computed and displayed as a tile plot (Figure 4). Results suggest that patients with brain tumors are more often comatose and that patients without brain tumors are less often comatose. Bootstrap computation of confidence interval and p-values for Ducher's Z that these measures are statistically significant. Similar conclusions can be reached using pointwise mutual information.

According to the data generating model, high calcium concentrations can also be responsible for comas in patients without brain tumors. Some patients will be comatose although they do not have a brain tumor. This can be identified using association based subgroup analysis in the "Subgroup" panel. Two subgroups are defined according to the sign of the association measure: the positive association and the negative association subgroup. Here Ducher's Z with association thresholds set to zero were used. Continuous calcium concentrations was plotted for subgroups (Figures 5). The positive association subgroup has significantly lower calcium concentrations than the negatively associated subgroup. This means that association between brain tumors and comas is well explained given that calcium concentration is not increased. Similar conclusions can be reached using pointwise mutual information or categorical calcium concentrations.

## Conclusion

Software presently available only allows computation of global association measures. Zebu provides a user-friendly manner to compute both global and local multivariate association measures. This software can be used by scientists and clinicians with no former knowledge of programming. Furthermore, we have presented a method to define subgroups allowing to characterize the population in which the global association exists.

Local association measures offer a different view on what association is by breaking the myth that it has the same strength for all modalities of variables. This makes theses measures particularly adapted to describe and model discontinuous relationships which are far from being an exception, notably in biology [Schmid et al., 1994]. In terms of quantity, their use is however still limited in the scientific literature. Nonetheless, the diversity of fields interested in these measures show us that they are of interest. Indeed applications are found in computation linguistics, image processing, cardiology, pharmacokinetics and epidemiology.

Local association measures are issued from empirical research. Although these have proven their interest in diverse applications, theoretical studies of their mathematical properties are sparse. For example, only Monte Carlo simulations of Ducher's Z behavior are available [Ducher et al., 1994]. A more theoretical approach to these measures could be of interest. Moreover, improvements in Zebu are also possible. The first concerns discretization, a necessary step for continuous variables. We have restrained ourselves to very simple discretization

methods: equal-width and user-defined. Other discretization algorithms exist [Dash et al., 2011] and may be more adapted for computation of association measures. These will have to be considered in future versions of Zebu. Furthermore, the bootstrap function in Zebu is based on an iterative procedure. These are particularly slow in R. To speed this up, writing the bootstrap function in C or Fortran and calling it from R could be a reliable solution. Finally, Zebu has been conceived for people with no programming knowledge. However, a R package for more experienced users could be of use for more sophisticated analysis.

## Availability and requirements

- Project name: Zebu
- Project home page: [github.com/olivmrtn/Zebu](https://github.com/olivmrtn/Zebu) and [olivmrtn.shinyapps.io/Zebu](https://olivmrtn.shinyapps.io/Zebu)
- Operating system: Platform independent
- Programming languages: R, HTML/CSS/Javascript
- Other requirements: Internet connection, internet browser
- License: GNU General Public 3
- Any restrictions to use by non-academics: None

## Abbreviations

H: Entropy MI: Mutual Information NMI: Normalized Mutual Information npmi: Normalized Pointwise Mutual Information pmi: Pointwise Mutual Information

## Competing interests

The authors declare that they have no competing interests.

## Author's contributions

MD conceived this project. OMF wrote the software code. MD contributed to software development by testing and providing constructive critical comments about the user interface. OMF wrote the manuscript. MD had the primary responsibility for the final content. All authors read and approved the final manuscript.

## Acknowledgements

The authors are grateful to Pascal Maire for making this project possible and to Cécile Serre for reading the draft.

## References

- [Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.
- [Bouma, 2009] Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference*, pages 31–40.
- [Cerutti et al., 1995] Cerutti, C., Ducher, M., Lantelme, P., Gustin, M. P., and Paultre, C. (1995). Assessment of spontaneous baroreflex sensitivity in rats a new method using the concept of statistical dependence. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, 268(2):R382–R388.
- [Daille, 1994] Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Universite Paris 7.
- [Dash et al., 2011] Dash, R., Paramguru, R. L., and Dash, R. (2011). Comparative analysis of supervised and unsupervised discretization techniques. *International Journal of Advances in Science and Technology*, 2(3).
- [Ducher et al., 2002] Ducher, M., Bertram, D., Sagnol, I., Cerutti, C., Thivolet, C., and Fauvel, J. (2002). Limits of clinical tests to screen autonomic function in diabetes type 1. *Journal of the Peripheral Nervous System*, 7(2):138–138.
- [Ducher et al., 1994] Ducher, M., Cerutti, C., Gustin, M., and Paultre, C. (1994). Statistical relationships between systolic blood pressure and heart rate and their functional significance in conscious rats. *Medical and Biological Engineering and Computing*, 32(6):649–655.
- [Ducher et al., 1992] Ducher, M., Cerutti, C., Gustin, M. P., and Paultre, C. Z. (1992). A new method to assess statistical dependence application to the relationships between systolic blood pressure and heart rate. *Genetic Hypertension*, 218:189–191.
- [Ducher et al., 2003] Ducher, M., Fauvel, J., Maurin, M., Laville, M., Maire, P., Paultre, C., and Cerutti, C. (2003). Sodium intake and blood pressure in healthy individuals. *Journal of hypertension*, 21(2):289–294.
- [Ducher et al., 2001] Ducher, M., Maire, P., Cerutti, C., Bourhis, Y., Foltz, F., Sorensen, P., Jelliffe, R., and Fauvel, J. (2001). Renal elimination of amikacin and the aging process. *Clinical Pharmacokinetics*, 40(12):947–953.
- [Freedman and Petitti, 2001] Freedman, D. A. and Petitti, D. B. (2001). Salt and blood pressure conventional wisdom reconsidered. *Evaluation Review*, 25(3):267–287.
- [Goodman and Kruskal, 1979] Goodman, L. A. and Kruskal, W. H. (1979). Measures of association for cross classifications. *Measures of Association for Cross Classifications*, pages 2–34.

- [Isola et al., 2014] Isola, P., Zoran, D., Krishnan, D., and Adelson, E. H. (2014). Crisp boundary detection using pointwise mutual information. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision ECCV 2014*, number 8691 in Lecture Notes in Computer Science, pages 799–814. Springer International Publishing.
- [Jakulin and Bratko, 2003] Jakulin, A. and Bratko, I. (2003). Analyzing attribute dependencies. In Lavrac, N., Gamberger, D., Todorovski, L., and Blockeel, H., editors, *Knowledge Discovery in Databases: PKDD 2003*, number 2838 in Lecture Notes in Computer Science, pages 229–240. Springer Berlin Heidelberg.
- [Kaplan, 2010] Kaplan, N. M. (2010). *Kaplan's clinical hypertension*. Lippincott Williams & Wilkins, 111 edition.
- [McGill, 1954] McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, 19(2):97–116.
- [Meyer, 2014] Meyer, P. E. (2014). *infotheo: Information-Theoretic Measures*. R package version 1.2.0.
- [Pearl, 1988] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.
- [R Core Team, 2014] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [RStudio and Inc., 2014] RStudio and Inc. (2014). *shiny: Web Application Framework for R*. R package version 0.10.2.1.
- [Rughoo et al., 2014] Rughoo, L., Bourguignon, L., Maire, P., and Ducher, M. (2014). Study of relationship between volume of distribution and body weight application to amikacin. *European journal of drug metabolism and pharmacokinetics*, 39(2):87–91.
- [Sapoznikov et al., 2013] Sapoznikov, D., Dranitzki Elhalel, M., and Rubinger, D. (2013). Heart rate response to blood pressure variations: Sympathetic activation versus baroreflex response in patients with end-stage renal disease. *PLoS ONE*, 8(10):e78338.
- [Sarkar and Andrews, 2013] Sarkar, D. and Andrews, F. (2013). *latticeExtra: Extra Graphical Utilities Based on Lattice*. R package version 0.6-26.
- [Schmid et al., 1994] Schmid, B., Polasek, W., Weiner, J., Krause, A., and Stoll, P. (1994). Modeling of discontinuous relationships in biology with censored regression. *The American Naturalist*, 143(3):494–507.
- [Shipley, 2000] Shipley, B. (2000). *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press, Cambridge.

- [Van de Cruys, 2011] Van de Cruys, T. (2011). Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, pages 16–20, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Watanabe, 1960] Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82.
- [Wickham, 2009] Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

## Additional Files

### **Additional file 1 — cancer.net**

This is the bayesian network used to generate simulated data. It is represented in figure 1 and is borrowed from Pearl [Pearl, 1988]. It can be read using a software as SamIan.

### **Additional file 2 — serum\_calcium.R**

This is the R script used to generate continuous data from discrete data in order to illustrate how Zebu handles such variables.

### **Additional file 3 — cancer.csv**

This is the simulated dataset used in the case study.

## Figures

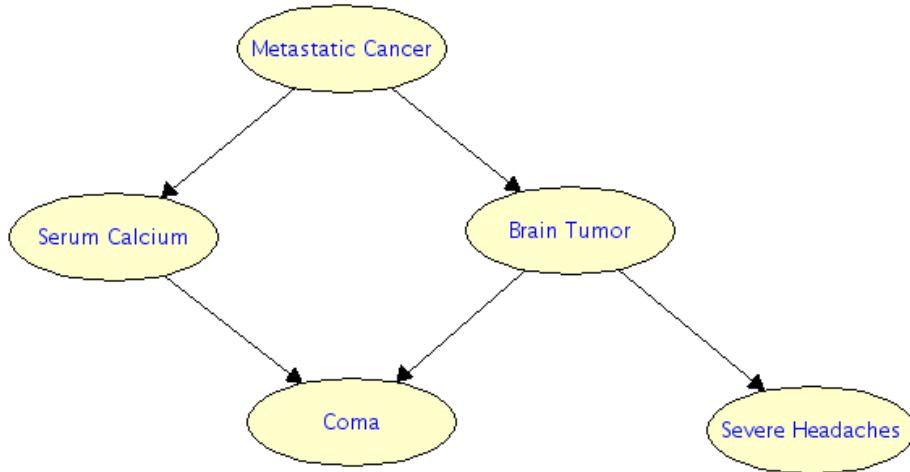


Figure 1: Bayesian network used to generate data

The screenshot shows the Zebu software interface with a correctly imported dataset. The top navigation bar includes 'Zebu', 'Import file', 'Preprocess', 'Association', 'Subgroups', 'Options', and 'About'. The main area displays a table of data with the following columns: Metastatic\_Cancer, Serum\_Calcium, Brain\_Tumor, Coma, Severe\_Headaches, and Serum\_Calcium\_Concentration. The table shows 10 entries of binary data (Absent or Present) and a numerical concentration value. The left sidebar contains import parameters: Header selected, File format CSV selected, Character used as CSV separator Colon (:) selected, and Character used as decimal point Comma (,) selected. An 'Import' button is at the bottom of the sidebar.

Metastatic_Cancer	Serum_Calcium	Brain_Tumor	Coma	Severe_Headaches	Serum_Calcium_Concentration
Absent	Increased	Present	Present	Present	2.652267
Absent	Not_Increased	Absent	Absent	Present	2.466206
Absent	Increased	Absent	Present	Absent	2.622019
Absent	Not_Increased	Absent	Absent	Absent	2.308341
Absent	Not_Increased	Absent	Absent	Present	2.424620
Present	Increased		Present	Present	2.745427
	Not_Increased	Absent	Absent	Present	2.305003
Absent	Increased	Absent	Present	Present	2.670038
Absent	Not_Increased	Absent	Absent	Absent	2.427350
Absent	Increased	Absent	Present	Present	2.702990

Figure 2: Correctly imported dataset

Zebu Import file Preprocess Association Subgroups Options About

#### Change names

Metastatic Cancer  
Serum Calcium  
Brain Tumor  
Coma  
Severe Headaches  
Serum Calcium Concentration  
Excluded  
id

#### Missing data

1.714 % of data is missing.  
How should this be handled?

Pairwise deletion  
 Listwise deletion

#### Discretization breaks

Serum Calcium Concentration  
2.2.2.6.3.4

#### Apply modifications

Apply

**Bins**  
Serum Calcium Concentration: [2.2.2.6], [2.6.3.4]

Variable kind

**Continuous**  
Serum Calcium Concentration

**Categorical**  
Metastatic Cancer Serum Calcium  
Brain Tumor Coma Severe Headaches

Show 10 entries

Metastatic Cancer	Serum Calcium	Brain Tumor	Coma	Severe Headaches	Serum Calcium Concentration	Id
Absent	Increased	Present	Present	Present	(2.6.3.4]	1
Absent	Not_Increased	Absent	Absent	Present	(2.2.2.6]	2
Absent	Increased	Absent	Present	Absent	(2.6.3.4]	3
Absent	Not_Increased	Absent	Absent	Absent	(2.2.2.6]	4
Absent	Not_Increased	Absent	Absent	Present	(2.2.2.6]	5
Present	Increased	Present	Present	Present	(2.6.3.4]	6
	Not_Increased	Absent	Absent	Present	(2.2.2.6]	7
Absent	Increased	Absent	Present	Present	(2.6.3.4]	8
Absent	Not_Increased	Absent	Absent	Absent	(2.2.2.6]	9
Absent	Increased	Absent	Present	Present	(2.6.3.4]	10

Showing 1 to 10 of 5,000 entries

Search:

Metastatic Cancer Serum Calcium Brain Tumor Coma Severe Headaches Serum Calcium Concentration Id

Previous 1 2 3 4 5 ... 500 Next

Figure 3: Preprocessed data

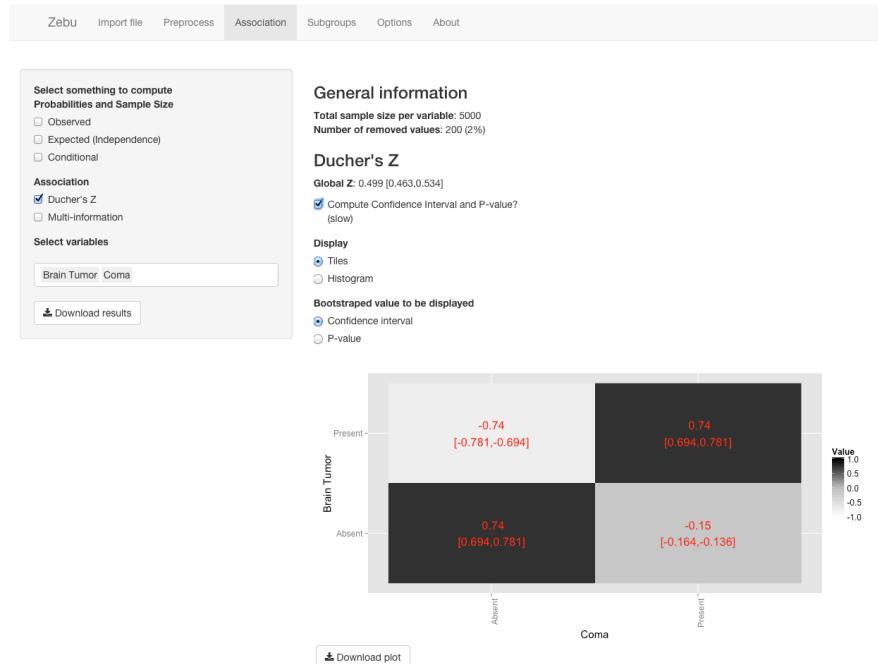


Figure 4: Ducher's Z of Coma and Brain Tumor

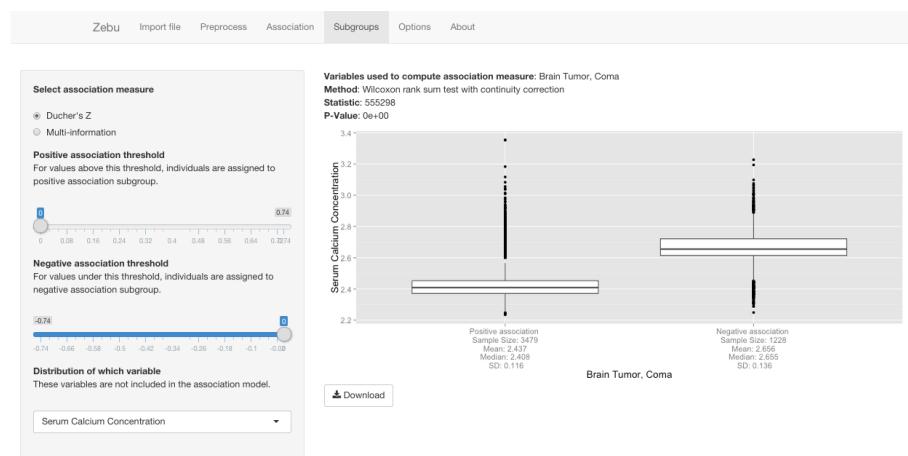


Figure 5: Subgroup analysis of continuous calcium concentrations given Ducher's Z of Coma and Brain Tumor

## 2. Exemples d'utilisation du Zébu

Afin d'illustrer l'utilité des mesures de l'association locale, et donc les limites des mesures globales, nous allons étudier trois relations atypiques. Nous utiliserons, selon nos ressources, des données simulées ou réelles. La mesure de l'association locale utilisée sera le Z de Ducher, mais des résultats analogues peuvent être obtenus avec d'autres mesures locales. Les exemples sont les suivants : une relation non-monotone entre deux variables continues, une relation entre deux variables catégorielles et une relation entre trois variables catégorielles.

### 2.1. Relation continue non-monotone : exemple de l'hormèse

Notre premier exemple concerne une relation bivariée et non-monotone, c'est-à-dire, une relation entre deux variables qui n'est ni strictement croissante ni strictement décroissante. Il s'agit donc d'une relation non-linéaire. Nous avons pris pour illustration des données simulant une relation dose-réponse hormétique, c'est-à-dire, une relation biphasique avec un effet bénéfique à faible dose et un effet délétère à forte dose (43). Un exemple est l'exercice physique : l'absence ou l'excès est délétère pour la santé tandis qu'une quantité modérée est bénéfique (80). Ces relations prennent approximativement l'allure d'une courbe en U inversée que nous avons cherché à reproduire. Nous avons pour cela généré un échantillon de 1000 couples dose-réponse en utilisant une équation polynomiale (Équation 10) où  $\varepsilon$  est une loi normale centrée de variance 0,05. Les résultats de la simulation sont dans la Figure 8.

La force de l'association entre deux variables reliées non-linéairement ne peut pas être quantifiable par une mesure basée sur la covariance comme le coefficient de corrélation de Pearson. Celui-ci est nul ici ( $r = 0,001$ ). Les méthodes non-paramétriques basées sur les probabilités sont plus adaptées. Effectivement, le Z de Ducher quantifie facilement l'association aussi bien globalement que localement (Figure 9). On remarquera, par ailleurs, que l'association locale n'est pas uniformément répartie. Effectivement, elle est beaucoup plus forte au niveau du maximum de la réponse.

$$Réponse = -8 \times (Dose - 0,5)^2 + 1 + \varepsilon$$

Équation 10. Équation polynomiale utilisée pour simuler une relation hormétique entre dose et réponse

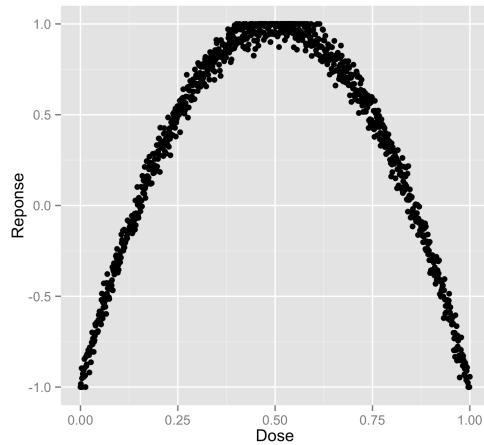


Figure 8. Nuage de points : relation hormétique entre dose et réponse

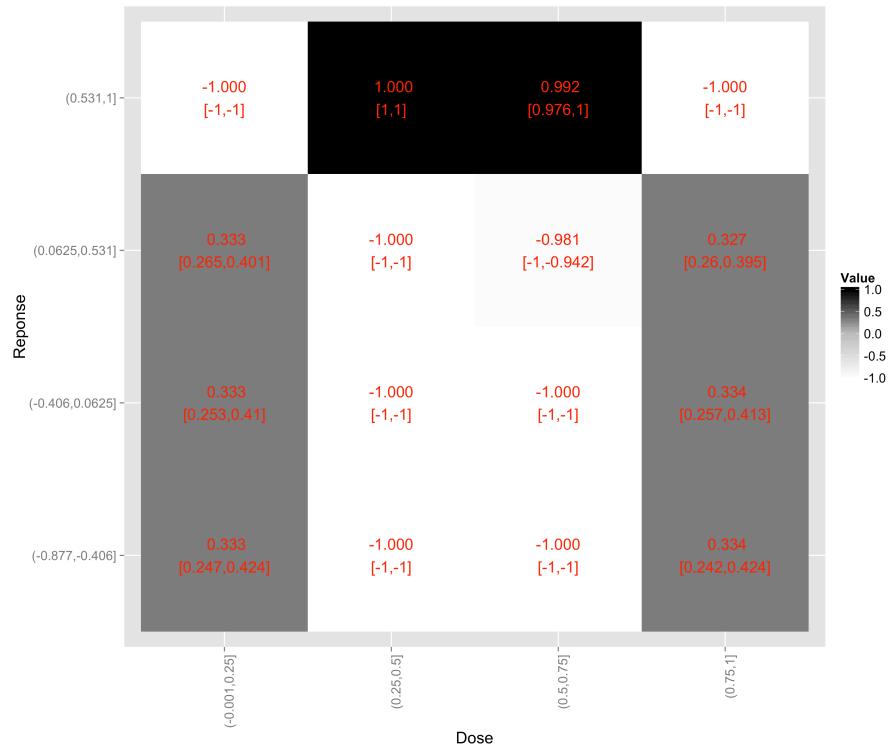


Figure 9. Z de Ducher : relation hormétique entre dose et réponse.  
Les intervalles de confiance à 95% sont représentés entre les crochets. Le Z global est de 0,59 avec un intervalle de confiance à 95% de [0,558 ; 0,626].

## 2.2. Variables catégorielles : exemple du tabac

Notre deuxième exemple concerne deux variables catégorielles dichotomiques et est issu de la base de données INDANA (81). Il s'agit plus précisément de la relation entre « fumer » et « faire un infarctus du myocarde ». Le tableau de contingence est représenté dans la Figure 10 et le traitement par le Z de Ducher dans la Figure 11. La force d'association est généralement résumée en épidémiologie sous la forme d'un rapport des cotes ou *odds ratio*. Ici, celui-ci est de 1,76, ce qui signifie approximativement que les fumeurs ont une probabilité 1,76 fois plus élevée de faire un infarctus du myocarde que les non-fumeurs. Le Z de Ducher nous indique également que les fumeurs sont associés localement et positivement au développement d'un infarctus et que les non-fumeurs sont associés localement et positivement au non-développement d'un infarctus ( $Z = 0,159$  pour les deux). Il nous indique aussi que « ne pas faire un infarctus du myocarde » est quasiment indépendant de « fumer » ( $Z = -0,02$ ). Il est donc nécessaire de considérer d'autres variables explicatives de l'infarctus. Une association globale significative (*odds ratio*) peut cacher une composante locale indépendante (Z de Ducher).

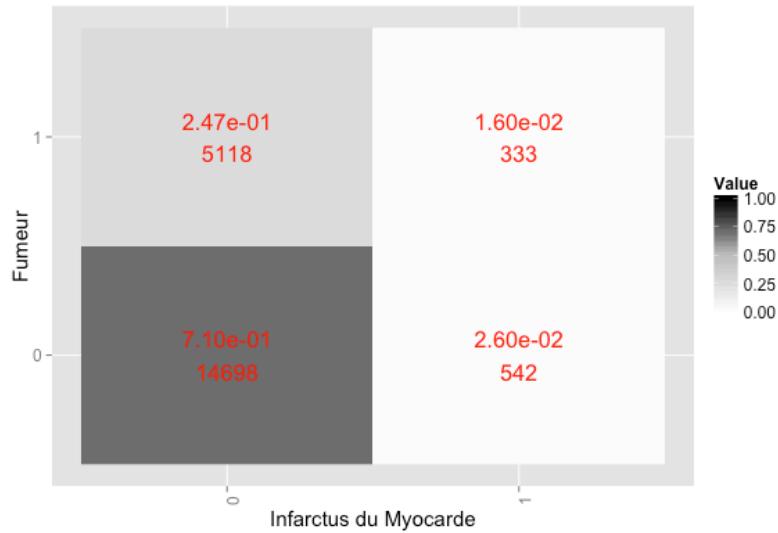


Figure 10. Tableau de contingence : relation entre « fumer » et « faire un infarctus du myocarde ». Les probabilités jointes sont au-dessus des comptages

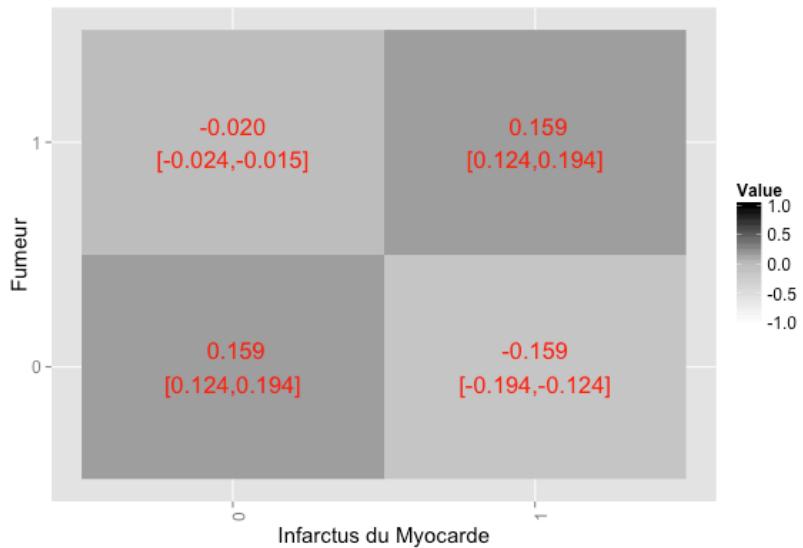


Figure 11. Z de Ducher : relation entre « fumer » et « faire un infarctus du myocarde »

Les mesures de l'association locales nous permettent également d'extraire les couples de modalités de variables pour lesquels les variables sont fortement dépendantes l'une de l'autre. Ici il s'agit des couples (non-fumeur, pas d'infarctus) et (fumeur, infarctus). Ces couples peuvent être interprétés comme une règle ou une tendance générale. On peut ainsi comparer les individus qui sont compris dans cette tendance avec ceux qui en dérogent, c'est-à-dire, les couples (non-fumeur, infarctus) et (fumeur, pas d'infarctus). On peut comparer la distribution d'autres variables entre ces deux sous-groupes pour essayer d'identifier en quoi ils diffèrent et donc identifier des caractéristiques générales aux deux groupes. La distribution de la variable âge est donnée sous la forme de boîtes à moustache dans la Figure 12. On remarque que les patients qui dérogent à la tendance ont un âge significativement plus jeune. Effectivement, l'âge est un facteur de risque de l'infarctus du myocarde. Ceci peut expliquer en partie l'indépendance entre « ne pas faire un infarctus du myocarde » et « fumer ». Ces individus n'ont pas eu une exposition suffisamment longue au tabac pour développer un infarctus du myocarde.

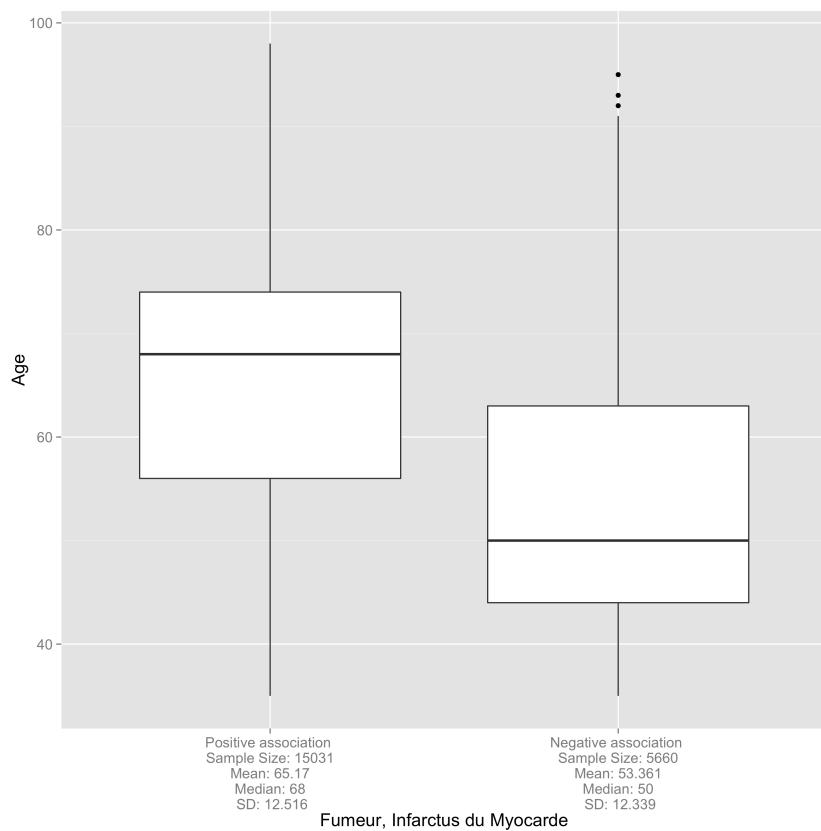


Figure 12. Analyse en sous-groupe basé sur la dépendance : distribution de l'âge en fonction de l'association fumeur - infarctus

### 2.3. Relation trivariée : exemple des études d'association pangénomique

Jusqu'alors, nous n'avons considéré que l'association entre deux variables. C'est effectivement ce qui est le plus souvent étudié. Par exemple, pour les études d'association pangénomique (*genome-wide association study*), on suppose qu'un allèle est associé à une maladie et on mesure l'association entre l'allèle et la maladie (82). On ne considère pas (ou rarement) les interactions entre deux allèles et la maladie. Cela s'explique par le fait que les mesures d'association les plus utilisées, comme le coefficient de corrélation de Pearson, ne sont applicables qu'à deux variables. Or, l'association est un concept qui se généralise aisément à plusieurs variables et il existe des mesures de l'association multivariée dont le Z de Ducher fait partie.

Les relations entre les variables d'un circuit logique sont des exemples de relations multivariées. Ces circuits exécutent une opération algébrique sur deux variables binaires et retournent une variable binaire (0 ou 1). Les sorties en fonction des combinaisons possibles d'entrées sont représentées dans une table de vérité. Les exemples les plus connus sont le « ET » dit *AND*, le « OU inclusif » dit *OR* et le « OU exclusif » dit *XOR*. Les tables de vérité sont données dans le Tableau 3.

AND			OR			XOR		
Entrée 1	Entrée 2	Sortie	Entrée 1	Entrée 2	Sortie	Entrée 1	Entrée 2	Sortie
0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	0	1	1
1	0	0	1	0	1	1	0	1
1	1	1	1	1	1	1	1	0

Tableau 3. Tables de vérité des circuits logiques communs

Il est intéressant de remarquer que dans le cas du circuit logique *XOR*, il est impossible d'identifier une relation entre une entrée prise individuellement et la sortie. Pour illustrer ceci, nous avons simulé 1000 triplets. Effectivement, l'association bivariée entre une entrée et la sortie nous apparaît comme indépendante l'une de l'autre (Figure 13). Il est nécessaire de prendre en compte l'association entre les trois variables pour faire apparaître la relation (Tableau 4).

Il n'y a rien de surprenant à penser que des relations similaires existent dans la nature. Elles sont probablement méconnues, car les mesures multidimensionnelles sont peu utilisées et donc les relations n'ont jamais été identifiées. Imaginons une relation *XOR* entre deux allèles. Il suffirait que les allèles aient des effets antagonistes. Dans le cas où les allèles seraient présents l'un sans l'autre, on observerait effectivement un effet (sortie = 1). Au contraire, dans le cas où ils seraient tous les deux présents ensemble, leurs effets s'annuleraient (sortie = 0). Bien sûr, si les deux sont absents, il n'y a pas d'effet (sortie = 0). Les mesures d'association multidimensionnelles se révèlent alors être un outil indispensable pour l'identification de ces interactions.

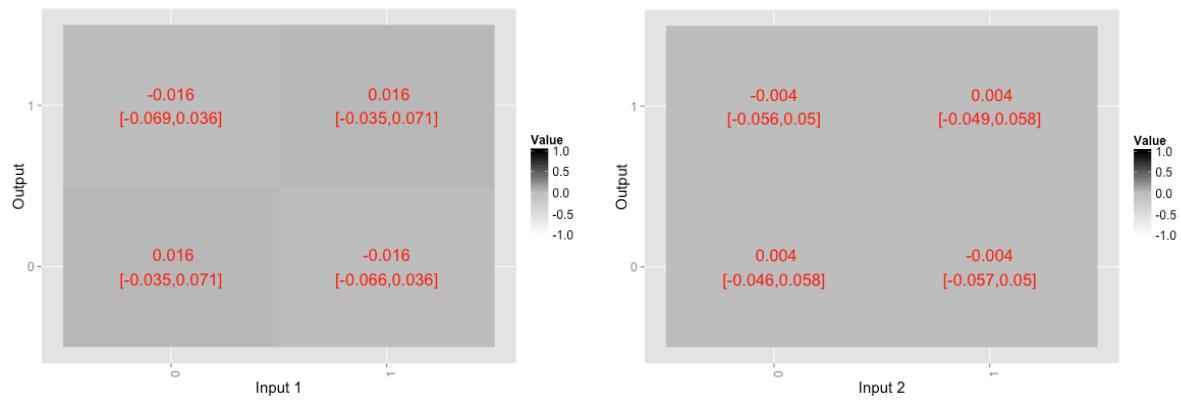


Figure 13 : Z de Ducher bivariée : circuit logique XOR

Input 1	Input 2	Output	Z de Ducher	Intervalle de Confiance 95%
0	0	0	0.348	0.329
0	1	1	0.328	0.307
1	0	1	0.339	0.316
1	1	0	0.339	0.315

Tableau 4. Z de Ducher trivariée : circuit logique XOR

Nous n'affichons que les triplets de valeurs positivement associées ( $Z > 0$ ).

## IV. Discussion

Ce travail a conduit à l'implémentation d'un logiciel libre permettant le calcul de mesures d'association locale ainsi que leur expression globale. Ce logiciel a été conçu pour être intuitif et facile d'accès. Il s'agit donc d'une application web directement utilisable à l'aide d'un navigateur internet. Son code source est facilement accessible par la plateforme GitHub. Cela permet aux utilisateurs plus avancés d'exécuter localement le logiciel sur leur machine ou bien de le distribuer, de l'étudier, de l'améliorer ou de le réutiliser. Il est intéressant de remarquer que bien que ces mesures datent des années 1990 (7,48,49), il s'agit du premier logiciel publié permettant ces calculs. Notre logiciel répond donc à une lacune et devrait stimuler l'utilisation des mesures d'association locales. De plus, son côté intuitif lui permet d'être utilisé comme un outil pédagogique, afin de pouvoir mieux expliquer les subtilités de l'association entre les variables.

Les mesures locales que nous avons implémentées permettent une description plus fine de la relation de dépendance entre variables. En décomposant l'association globale en ses composantes locales, elles sont complémentaires aux mesures globales qui nous informent uniquement d'une tendance générale. On pourra alors identifier des associations locales significatives alors qu'il n'existe pas de tendance générale. De plus, on remarque qu'une association locale non significative peut être masquée par une relation globale significative. Par ailleurs, ces mesures sont non-paramétriques. Elles ne font donc pas d'hypothèse *a priori* sur la forme de la relation. Cela leur permet de quantifier l'association pour des relations non-linéaires. Ceci n'est pas le cas de mesures basées sur la variance comme le r de Pearson. En outre, ces mesures possèdent une généralisation multidimensionnelle. Cela leur permet de capturer des relations multidimensionnelles entre plusieurs variables qui ne sont pas détectables dans des dimensions inférieures. Finalement, le Zébu implémente une méthode originale d'analyse en sous-groupe basée sur la dépendance. Celle-ci permet d'identifier plus facilement les caractéristiques des individus dérogeant à la tendance globale.

Nous remarquerons, maintenant, quelques limitations techniques à l'implémentation que nous proposons. Nous avons justifié le choix de l'implémentation sous la forme d'une application web par le côté accessible et intuitif de ces logiciels. Cependant, le choix d'une interface graphique rend l'automatisation de calculs impossible. Il serait donc souhaitable de proposer une solution contrôlable par ligne de commande. Nous recommandons la construction d'un *package R*. Celle-ci serait facilement réalisable à partir des codes sources du Zébu qui sont disponibles. Par ailleurs, la gestion des valeurs manquantes se fait par déletion. Cette méthode provoque une perte de pouvoir statistique et peut être à l'origine de biais (83). Nous n'avons pas implémenté la méthode de référence, à savoir, l'imputation multiple, car celle-ci demande de réaliser des tests de diagnostics statistiques. Nous aurions pu implémenter un module d'importation des jeux de données générées par une telle méthode. Autre limitation, l'unique méthode automatique de discréétisation des variables continues implantée est celle des écarts-égaux. Cette méthode est surpassée, en certains aspects, par d'autres méthodes de discréétisation (19). Celles-ci pourraient également être implantées. On constatera également que les propriétés mathématiques de ces mesures sont méconnues. Leur connaissance est essentiellement empirique. Par exemple, la distribution du Z de Ducher n'a été abordée qu'à travers des simulations Monte-Carlo (46). Il serait d'intérêt de mieux connaître leurs propriétés à travers des études plus théoriques, par exemple, en établissant des formules de leur variance ou de leur biais d'estimation.

Pour finir, nous remarquerons que nous ne sommes pas limités à implémenter ces mesures sans apporter des nouveautés. Effectivement, le calcul de la significativité de ces mesures à travers des valeur-p et des intervalles de confiance n'avait jamais été réalisé au préalable. Celui-ci a été facilement implanté sous la forme d'une méthode de rééchantillonnage nommée *bootstrap*. De plus, aucune généralisation multivariée n'avait encore été publiée pour le Z de Ducher. Celle-ci est proposée dans l'article décrivant le Zébu. Finalement, à travers nos exemples d'utilisation, nous proposons certains champs d'application potentiels de ces mesures en science de la vie.

## V. Conclusion

THÈSE SOUTENUE PAR : M. MARTIN Olivier

L'accélération du volume de données disponibles ouvre de nombreux opportunités et défis pour les sciences de la vie. Afin d'extraire la connaissance de ces bases de données, les méthodes et les outils issus des mathématiques (ex. : modèles mathématiques) et de l'informatique (ex. : logiciels scientifiques) sont d'intérêt. Ceux-ci permettent, par exemple, de conduire *in silico* des expériences autrement impossibles à réaliser. Cependant, ces méthodes s'inspirent de l'approche déterministe des sciences physiques et de l'ingénieur. Elles ne réussissent donc pas toujours à saisir la complexité et la flexibilité de la vie.

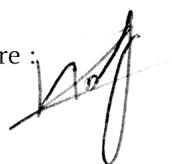
L'étude de la vie requiert des méthodes qui lui sont adaptées. Pour défendre cette idée, nous prenons pour exemple les mesures d'association qui ne sont pas systématiquement adaptées aux sciences de la vie. Par exemple, les mesures de l'association globales, comme le coefficient de corrélation de Pearson, quantifient une tendance générale en supposant que l'association est continue pour l'ensemble des modalités des variables. Cette supposition n'est pas toujours justifiée en sciences de la vie où l'on observe des relations discontinues comme des effets de seuil ou de saturation. Il est parfois nécessaire de décomposer l'association globale en ses composantes locales. Pour chaque combinaison des modalités des variables, on mesure l'association locale grâce à une mesure comme le Z de Ducher. Une autre difficulté provient des relations non-linéaires où, l'association entre variables n'est pas quantifiable par des mesures paramétriques. Il est nécessaire d'utiliser des mesures non-paramétriques qui ne font pas d'hypothèse *a priori* sur la forme de la relation. Finalement, la plupart des mesures de l'association sont bivariées, c'est-à-dire, elles ne mesurent que l'association entre deux variables. Cependant, certaines formes d'associations multidimensionnelles (ex. : 2 causes antagonistes et 1 effet) ne sont pas mesurables dans des dimensions inférieures (ex. : 1 cause et 1 effet). Il est donc nécessaire d'utiliser des mesures ayant une généralisation multidimensionnelle.

Le Z de Ducher et l'information mutuelle spécifique sont des mesures non-paramétriques et multivariées de l'association locale. Elles sont donc particulièrement adaptées à mesurer l'association en sciences de la vie. Cependant, aucun logiciel ne permet actuellement leur calcul. Nous avons donc développé un logiciel afin de répondre à cette lacune. Le logiciel Zébu est d'une application web libre-source directement utilisable à l'aide d'un navigateur internet. Il est intuitif ce qui permet à des cliniciens et scientifiques ne sachant pas programmer de mettre en place une étude adoptant une approche locale. De plus, son code source est facilement accessible par la plateforme GitHub. Cela permet aux utilisateurs plus avancés d'exécuter localement le logiciel sur leur machine ou bien de le distribuer, de l'étudier, de l'améliorer ou de le réutiliser. Par ailleurs, le Zébu implémente une méthode originale d'analyse en sous-groupe basée sur la dépendance. Celle-ci permet d'identifier plus facilement les caractéristiques des individus dérogeant à la tendance globale. Ce logiciel devrait donc, à terme, stimuler l'utilisation de ces mesures d'association locales

**Le Président de la thèse,**

Nom : MOYRET-LALLE Caroline

Signature :



Vu et permis d'imprimer, Lyon, le 21 mai 2015

Vu, la Directrice de l'Institut des Sciences Pharmaceutiques et Biologiques, Faculté de Pharmacie

Pour le Président de l'Université Claude Bernard Lyon 1,



Professeure C. VINCIGUERRA

## VI. Glossaire

### **Association**

Toute relation entre au moins deux variables qui les rend statistiquement dépendantes. Autrement dit, pour deux variables, l'observation de l'une nous informe sur la réalisation de l'autre. Le concept d'association est mesurable : voir mesure de l'association globale et mesure de l'association locale.

Exemple : relation linéaire entre deux variables aléatoires.

Synonyme : corrélation.

### **Circuit logique**

Circuit implémentant une opération de l'algèbre de Boole, c'est-à-dire l'algèbre des variables logiques (vrai ou faux). Les opérations les plus connus sont le « ET », le « OU inclusif » et le « OU exclusif ». Ces opérations s'appliquent sur au moins une variable logique et retournent une unique variable logique. Voir partie III.2.3.

### **Corrélation**

Synonyme d'association.

### **Covariance**

Mesure quantifiant la variabilité jointe de deux variables aléatoires. Sa forme normalisée est le coefficient de corrélation de Pearson. Voir partie II.1.6.3.

### **Discontinuité**

La notion de discontinuité est formellement définie en mathématiques pour les fonctions. Nous l'employons, dans cette thèse, comme étant un changement d'état de la relation entre les variables. Des exemples sont les relations à effet de seuil comme l'excitabilité neuronale. Le passage du seuil entraîne l'apparition d'une relation entre les variables.

## **Hormèse (hormétique)**

Relation dose-réponse biphasique avec un effet bénéfique à faible dose et un effet délétère à forte dose (43). La courbe dose-réponse prend l'allure d'un U inversé. Voir partie III.2.1.

## **Indépendance statistique**

Absence d'association statistique entre au moins deux variables. Autrement dit, pour deux variables, l'observation de l'une ne nous informe pas sur la réalisation de l'autre. Voir partie II.1.6.4.

## **Itération**

En informatique, répétition d'une séquence d'instructions réalisée jusqu'à ce qu'une condition soit remplie.

Exemples : les boucles « FOR » et « WHILE » dans de nombreux langages de programmation.

Voir partie II.2.2.1.

## **Mesure de l'association globale**

Mesure unique et moyenne de l'association entre au moins deux variables. Le plus souvent normalisée entre -1 et 1 ou 0 et 1. La valeur absolue reflète la force de l'association et le signe le sens de l'association.

Exemples : covariance, coefficient de corrélation de Pearson, chi-deux, V de Cramér, rapport des cotes, information mutuelle, Z global de Ducher.

Voir partie II.1.6.

## **Mesure de l'association locale**

Décomposition de l'association globale en ses composantes locales. Pour chaque combinaison des modalités des variables, on obtient une mesure de l'association. L'ensemble est donc représenté dans une matrice. Le plus souvent

normalisé entre -1 et 1. Les valeurs positives reflètent le cas où les variables cooccurrent plus souvent que si elles étaient indépendantes. Les valeurs négatives le cas où les variables cooccurrent moins souvent que si elles étaient indépendantes.

Exemple : résiduel du chi-deux, information mutuelle spécifique, Z de Ducher.

Voir partie II.1.6.5.

## Modalité

Ensemble des réalisations possibles d'une variable aléatoire.

Exemple : pour une variable binaire, ses modalités peuvent être codées 0 et 1.

## Méthode Monte-Carlo

Méthode dépendant des réalisations de variables aléatoires.

Exemple : méthodes de rééchantillonnage.

## Réalisation

Résultat d'une variable aléatoire au cours d'une expérience probabiliste.

Exemple : au cours d'une expérience probabiliste, les réalisations d'une variable aléatoire binaire pourraient être 0, 0, 0, 1, 1, 0, 1.

## Rééchantillonnage

Méthode Monte-Carlo permettant notamment d'estimer la distribution de variables aléatoires en faisant des tirages aléatoires répétés. Permet également de construire des tests de significativité.

Exemple : *bootstrap*

## Variable aléatoire

Toute variable dont le résultat dépend d'une expérience probabiliste.

Exemple : le résultat d'un lancé de pièce est une variable aléatoire.

## VII. Références

1. Hayden EC. Is the \$1,000 genome for real? *Nat News.* 15 janv 2014;
2. [En ligne]. IBM Archives: 650 RAMAC announcement press release; 23 janv 2003 [cité le 25 mars 2015]. Disponible: [https://www-03.ibm.com/ibm/history/exhibits/650/650\\_pr2.html](https://www-03.ibm.com/ibm/history/exhibits/650/650_pr2.html)
3. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev.* 1958;65(6):386-408.
4. Laplace P-S. *Essai philosophique sur les probabilités.* 1814.
5. Mayr E. Cause and Effect in Biology Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science.* 10 nov 1961;134(3489):1501-6.
6. Shipley B. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference.* Cambridge: Cambridge University Press; 2000.
7. Ducher M, Cerutti C, Gustin MP, Paultre CZ. A new method to assess statistical dependence application to the relationships between systolic blood pressure and heart rate. *Genet Hypertens.* 1992;218:189-91.
8. Prlić A, Procter JB. Ten Simple Rules for the Open Development of Scientific Software. *PLoS Comput Biol.* 6 déc 2012;8(12):e1002802.
9. Newton I. *Philosophiae Naturalis Principia Mathematica.* 1687.
10. Darwin C. *On the origin of species by means of natural selection.* 1859.
11. Servedio MR, Brandvain Y, Dhole S, Fitzpatrick CL, Goldberg EE, Stern CA, et al. Not Just a Theory—The Utility of Mathematical Models in Evolutionary Biology. *PLoS Biol.* 9 déc 2014;12(12):e1002017.
12. Cohen JE. Mathematics Is Biology's Next Microscope, Only Better; Biology Is Mathematics' Next Physics, Only Better. *PLoS Biol.* 14 déc 2004;2(12):e439.
13. Stewart I. *The Mathematics of Life.* Basic Books; 2011.
14. Eykhoff P. *System Identification: Parameter and State Estimation.* Wiley; 1974.
15. Shmueli G. To Explain or To Predict? *Statistical Science.* 24 mai 2010;

16. Bunge M. Metascientific Queries. 1959.
17. Higgins JP. Nonlinear systems in medicine. *Yale J Biol Med.* 2002;75(5-6):247-60.
18. A Lesne A. Chaos in biology. *Riv Biol. déc* 2006;99(3):467-81.
19. Dash R, Paramguru RL, Dash R. Comparative analysis of supervised and unsupervised discretization techniques. *Int J Adv Sci Technol.* 2011;2(3).
20. Werndl C. On choosing between deterministic and indeterministic models: underdetermination and indirect evidence. *Synthese.* 11 juin 2011;190(12):2243-65.
21. Box G. Empirical Model-Building and Response Surfaces. 1987.
22. Domingos P. A few useful things to know about machine learning. *Commun ACM.* 1 oct 2012;55(10):78.
23. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013.
24. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, et al. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell.* 20 juill 2012;150(2):389-401.
25. Archakov AI, Govorun VM, Dubanov AV, Ivanov YD, Veselovsky AV, Lewi P, et al. Protein-protein interactions as a target for drugs in proteomics. *PROTEOMICS.* 1 avr 2003;3(4):380-91.
26. Tobacco Control Evaluation Center. Tips & Tools #8: Observation Methods [En ligne]. Center for Evaluation and Research; 2009. Disponible: [http://tobaccoeval.ucdavis.edu/files/Tips\\_Tools\\_08\\_2009.pdf](http://tobaccoeval.ucdavis.edu/files/Tips_Tools_08_2009.pdf)
27. Sheskin DJ. Handbook of Parametric and Nonparametric Statistical Procedures. 4<sup>e</sup> éd. Chapman & Hall/CRC; 2007.
28. Quaranta V, Weaver AM, Cummings PT, Anderson ARA. Mathematical modeling of cancer: The future of prognosis and treatment. *Clin Chim Acta.* 24 juill 2005;357(2):173-9.
29. [En ligne]. IBM - Watson - France; 31 déc 2011 [cité le 10 mars 2015]. Disponible: <http://www-05.ibm.com/fr/watson/>
30. Consortium TU. UniProt: a hub for protein information. *Nucleic Acids Res.* 28 janv 2015;43(D1):D204-12.

31. Scotch M, Duggal M, Brandt C, Lin Z, Schiffman R. Use of statistical analysis in the biomedical informatics literature. *J Am Med Inform Assoc.* 1 janv 2010;17(1):3-5.
32. Fawcett TW, Higginson AD. Heavy use of equations impedes communication among biologists. *Proc Natl Acad Sci.* 17 juill 2012;109(29):11735-9.
33. Fisher RA. Gene Frequencies in a Cline Determined by Selection and Diffusion. *Biometrics.* 1 déc 1950;6(4):353-61.
34. Ranganath HA. Nothing in biology makes sense without the flavour of mathematics. *Resonance.* 1 mars 2003;8(3):49-56.
35. Turing AM. The Chemical Basis of Morphogenesis. *Philos Trans R Soc Lond B Biol Sci.* 14 août 1952;237(641):37-72.
36. Erdős P, Rényi A. On random graphs. *Publ Math Debr.* 1959;6:290-7.
37. Académie Française. *Dictionnaire de l'académie française.* 9<sup>e</sup> éd. 1992.
38. Olsen J. What characterises a useful concept of causation in epidemiology? *J Epidemiol Community Health.* 1 févr 2003;57(2):86-8.
39. Pépin F. The randomness of life: A philosophical approach inspired by the Enlightenment. *Prog Biophys Mol Biol.* sept 2012;110(1):121-8.
40. Pearl J. An Introduction to Causal Inference. *Int J Biostat.* 26 févr 2010;6(2).
41. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. *Mach Learn.* 1 nov 1997;29(2-3):131-63.
42. Baldwin T, Evert S, Krenn B, Pecina P, Anastasiou D, Carl M, et al. A Machine Learning Approach to Multiword Expression Extraction.
43. Mattson MP. Hormesis Defined. *Ageing Res Rev.* janv 2008;7(1):1-7.
44. Szumilas M. Explaining Odds Ratios. *J Can Acad Child Adolesc Psychiatry.* août 2010;19(3):227-9.
45. Goodman LA, Kruskal WH. Measures of Association for Cross Classifications. *Meas Assoc Cross Classif.* 1 janv 1979;2-34.
46. Ducher M, Cerutti C, Gustin MP, Paultre CZ. Statistical relationships between systolic blood pressure and heart rate and their functional significance in conscious rats. *Med Biol Eng Comput.* 1994;32(6):649-55.
47. Isola P, Zoran D, Krishnan D, Adelson EH. Crisp Boundary Detection

Using Pointwise Mutual Information. Fleet D, Pajdla T, Schiele B, Tuytelaars T, rédacteurs. Comput Vis – ECCV 2014. 1 janv 2014;799–814.

48. Church KW, Hanks P. Word Association Norms, Mutual Information, and Lexicography. *Comput Linguist*. mars 1990;16(1):22–9.

49. Anselin L. Local Indicators of Spatial Association—LISA. *Geogr Anal*. 1 avr 1995;27(2):93–115.

50. Van de Cruys T. Two Multivariate Generalizations of Pointwise Mutual Information. Dans: Proceedings of the Workshop on Distributional Semantics and Compositionality [En ligne]. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011 [cité le 15 déc 2014]. p. 16–20. Disponible: <http://dl.acm.org/citation.cfm?id=2043121.2043124>

51. Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New Biotechnol*. 15 sept 2012;29(6):613–24.

52. Davidoff F. Standing Statistics Right Side Up. *Ann Intern Med*. 15 juin 1999;130(12):1019–21.

53. Richard van Noorden, Brendan Maher, Regina Nuzzo. The top 100 papers. 30 oct 2014;514(7524):550–3.

54. Metropolis N. The beginning of the Monte Carlo method.

55. Sapoznikov D, Dranitzki Elhalel M, Rubinger D. Heart Rate Response to Blood Pressure Variations: Sympathetic Activation versus Baroreflex Response in Patients with End-Stage Renal Disease. *PLoS ONE*. 4 oct 2013;8(10):e78338.

56. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, et al. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res*. janv 2011;39(Database issue):D392–401.

57. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, et al. Best Practices for Scientific Computing. *PLoS Biol*. 7 janv 2014;12(1):e1001745.

58. R Core Team. R: A Language and Environment for Statistical Computing [En ligne]. Vienna, Austria: R Foundation for Statistical Computing; 2015. Disponible: <http://www.R-project.org/>

59. Scutari M. Learning Bayesian Networks with the bnlearn R Package. *J Stat Softw*. 2010;35(3):1–22.

60. RStudio, Inc. shiny: Easy web applications in R. 2015.
61. Techopedia.com [En ligne]. What is Software Licensing? - Definition from Techopedia; [cité le 10 avr 2015]. Disponible: <http://www.techopedia.com/definition/2558/software-licensing>
62. [En ligne]. Free Software Foundation, Inc. Qu'est-ce que le logiciel libre ?; [cité le 16 avr 2015]. Disponible: <https://www.gnu.org/philosophy/free-sw.fr.html>
63. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and. Bioinformatics. 1 juin 2009;25(11):1422-3.
64. Eichler H-G, Abadie E, Breckenridge A, Leufkens H, Rasi G. Open Clinical Trial Data for All? A View from Regulators. PLoS Med. 10 avr 2012;9(4):e1001202.
65. Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible Research: Moving toward Research the Public Can Really Trust. Ann Intern Med. 20 mars 2007;146(6):450-3.
66. Rubin D. "Discussion: Statistical Disclosure Limitation. 1993;9(2):461-8.
67. Haberman SJ. The Analysis of Residuals in Cross-Classified Tables. Biometrics. mars 1973;29(1):205.
68. Ducher M. Le Traitement des Signaux Cardiovasculaires et L'utilisation d'une Nouvelle Méthode de Mesure de la Liaison Statistique. [Lyon]: Université Claude Bernard Lyon 1; 1995.
69. Cerutti C, Ducher M, Lantelme P, Gustin MP, Paultre C. Assessment of spontaneous baroreflex sensitivity in rats a new method using the concept of statistical dependence. Am J Physiol - Regul Integr Comp Physiol. 1995;268(2):R382-8.
70. Ducher M, Siche J, Fauvel J, Gustin M, Pozet N, Paultre C, et al. Comparison of three methods for the estimation of spontaneous cardiac baroreflex sensitivity in normotensive and hypertensive subjects. Arch Mal Coeur Vaiss. 1995;88(8):1233-6.
71. Ducher M, Zhang Z, Cerutti C, Julien C, Gustin M, Paultre C. Spontaneous cardiac baroreceptor reflex and regional circulations in conscious rats. J

Hypertens. 1996;14(7):865-9.

72. Lantelme P, Cerutti C, Lo M, Paultre C, Ducher M. Mechanisms of spontaneous baroreflex impairment in Lyon hypertensive rats. Am J Physiol - Regul Integr Comp Physiol. 1998;275(3):R920-5.
73. Ducher M, Fauvel J, Maurin M, Laville M, Maire P, Paultre C, et al. Sodium intake and blood pressure in healthy individuals. J Hypertens. 2003;21(2):289-94.
74. Ducher M, Maire P, Cerutti C, Bourhis Y, Foltz F, Sorensen P, et al. Renal Elimination of Amikacin and the Aging Process. Clin Pharmacokinet. 2001;40(12):947-53.
75. Rughoo L, Bourguignon L, Maire P, Ducher M. Study of relationship between volume of distribution and body weight application to amikacin. Eur J Drug Metab Pharmacokinet. 2014;39(2):87-91.
76. Bouma G. Normalized (pointwise) mutual information in collocation extraction. Proceedings of the Biennial GSCL Conference. 2009;31-40.
77. Karlström A, Ceccato V. A new information theoretical measure of global and local spatial association. août 2000;
78. Ducher M, Fauvel J, Gustin M, Cerutti C, Najem R, Cuisinaud G, et al. A new non-invasive statistical method to assess the spontaneous cardiac baroreflex in humans. Clin Sci. 1995;88:651455.
79. Williams G. Sur les Caractéristiques de la Colocation. Dans: Tours; 2001.
80. Radak Z, Chung HY, Koltai E, Taylor AW, Goto S. Exercise, oxidative stress and hormesis. Ageing Res Rev. janv 2008;7(1):34-42.
81. Gueyffier F, Boutitie F, Boissel JP, Cope J, Cutler J, Ekbom T, et al. INDANA: a meta-analysis on individual patient data in hypertension. Protocol and preliminary results. Thérapie. août 1995;50(4):353-62.
82. Manolio TA. Genomewide Association Studies and Assessment of the Risk of Disease. N Engl J Med. 8 juill 2010;363(2):166-76.
83. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1 févr 1999;8(1):3-15.

## VIII. Annexes : scripts R

### 1. Variance et covariance

```
library(ggplot2)
set.seed(2)

# Fonctions ----
draw_squares = function(n, x, y, xlim = NULL, ylim = NULL, xlab = NULL, ylab = NULL) {

  ggdata = data.frame(x, y) colnames(ggdata) = c("x", "y")
  meanx = mean(x)
  meany = mean(y)

  # Définition des rectangles représentant  $(xi - E(x))(yi - E(y))$ 
  rect = data.frame(xmin = apply(data.frame(x, meanx), 1, min),
                     xmax = apply(data.frame(x, meanx), 1, max),
                     ymin = apply(data.frame(y, meany), 1, min),
                     ymax = apply(data.frame(y, meany), 1, max))

  # Définition des couleurs des rectangles (position: rouge, négatif: bleu)
  colour = apply(ggdata, 1, function(row) {
    d = (row[1] - meanx) * (row[2] - meany)
    if (sign(d) == 1) return("red")
    else return("blue")
  })

  # Nuages de points avec rectangles
  p = ggplot(data = ggdata) +
    geom_point(mapping = aes(x = x, y = y), ) +
    geom_rect(data = rect, mapping = aes(xmin=xmin, xmax=xmax, ymin=ymin,
                                         ymax=ymax), alpha = 0.1, fill = colour) +
    geom_vline(xintercept = meanx) +
    geom_hline(yintercept = meany) +
    theme(text = element_text(size=20))

  # Définition des limites des abscisses et ordonnées du graphe
  if (!is.null(xlim)) p = p + xlim(xlim)
  if (!is.null(ylim)) p = p + ylim(ylim)
  if (!is.null(xlab)) p = p + xlab(xlab)
  if (!is.null(ylab)) p = p + ylab(ylab)

  return(p)}
```

```

# Main ----

# Dispersion d'une distribution en fonction de la variance --
n = 100 000
x1 = rnorm(n, 0, 1)
x2 = rnorm(n, 0, sqrt(2.5))
x3 = rnorm(n, 0, sqrt(5))
ggdata = data.frame(i = factor(rep(c("1", "2.5", "5"), each = n)), x = c(x1, x2, x3))

ggplot(ggdata, aes(x = x, fill = i)) +
  geom_density(alpha = 0.3) +
  ylab("Densité") +
  xlab("X") +
  guides(fill = guide_legend(title="Variance")) +
  theme(text = element_text(size=20))

ggsave("dispersion_variance.png", width = 8, height = 4)

# Interprétation géométriques -
# Paramètres généraux
n = 30
x = rnorm(n, 0, 0.1)
xlim = c(-0.25, 0.25)

# Variance
draw_squares(n, x = x, y = x, xlim = xlim, ylim = xlim, xLab = "X", yLab = "X")
ggsave("variance.png", width = 8, height = 8)

# Covariance
y1 = x + 1 + rnorm(n, 0, 0.1) # Relation linéaire avec bruit gaussien
print(cor(x, y1, method = "pearson")) # r de Pearson
draw_squares(n = n, x = x, y = y1, xLab = "X", yLab = "Y", xlim = c(-0.40, 0.40), ylim = c(0.6, 1.4))
ggsave("covariance_high.png", width = 8, height = 8)

y2 = rnorm(n, 0, 0.1) # Absence de relation
cor(x,y2) # r de Pearson
draw_squares(n = n, x = x, y = y2, xLab = "X", yLab = "Y", xlim = xlim, ylim = xlim)
ggsave("covariance_low.png", width = 8, height = 8)

```

## 2. Exemples d'utilisations du logiciel Zébu

```
# Init ---
library(ggplot2)
set.seed(1)

# Relation dose réponse ---
dose = seq(0,1,0.001)
response = - 8 * (dose - 0.5)^2 + 1 # Equation polynomiale
response = response + rnorm(length(dose), 0, 0.05) # Bruit gaussien
response[response < -1] = -1 # Normalisation de la réponse entre -1 et 1
response[response > 1] = 1
print(cor(dose, response, method = "pearson")) # r de Pearson
hormesis = data.frame(dose, response)
write.csv(hormesis, quote = FALSE, row.names = FALSE, "hormesis.csv") # Ecrire résultats

ggplot(hormesis, aes(dose, response)) +
  geom_point() +
  xlab("Dose") +
  ylab("Réponse") +
  xlim(0, 1) +
  ylim(-1, 1) # Nuage de points
  ggsave("hormesis.png")

# Circuit logique XOR ---
n = 1000 # Taille de l'échantillon
input1 = sample(c(0,1), size = n, replace = TRUE)
input2 = sample(c(0,1), size = n, replace = TRUE)
output = as.numeric((input1 | input2) & ! (input1 & input2))
xor = data.frame(input1, input2, output)
write.csv(xor, quote = FALSE, row.names = FALSE, "xor.csv")
```

La Faculté de Pharmacie de Lyon et l'Université Claude Bernard Lyon 1 n'entendent donner aucune approbation ni improbation aux opinions émises dans les thèses ; ces opinions sont considérées comme propres à leurs auteurs.

## ISPB-FACULTE DE PHARMACIE

Rapport du Président du jury de la thèse de Mr MARTIN Olivier

Olivier Martin a réalisé son manuscrit de thèse à partir de ses travaux de recherche centrés sur l'intérêt de la modélisation mathématique et l'utilisation et la création d'outils informatiques dans l'analyse de résultats expérimentaux en sciences de la vie et particulièrement dans des contextes physiologiques et pathologiques. Il apparaît aujourd'hui indispensable d'utiliser des outils mathématiques et informatiques face à la quantité de données obtenues à partir des techniques « omiques ». Les modèles mathématiques permettent également de simuler *in silico* des expériences parfois difficiles à réaliser *in vitro* et *in vivo*.

Dans la première partie de son manuscrit, Olivier Martin présente de façon non exhaustive bien sur, des modèles mathématiques prédictifs ou explicatifs. En particulier il présente différents modèles d'association ou de classification de données tel que linéaire vs non-linéaire, statique vs dynamique, continu vs discret, déterministe vs stochastique, paramétrique vs non-paramétrique, etc.

Dans cette première partie du manuscrit, Olivier Martin présente également l'utilisation d'outils informatiques en présentant des notions d'algorithmique, de programmation et de génie logiciel. L'informatique apparaît effectivement comme un moyen très efficace de stocker et rendre accessible ces nombreuses données. L'une des préoccupations de ces approches, plus particulièrement de la bioinformatique, est de maintenir des bases de données comme UniProt, et l'analyse avec l'algorithme BLAST. La bioinformatique s'impose de plus en plus dans tous les laboratoires en sciences de la vie.

Une deuxième partie correspond au travail de recherche réalisé par Olivier Martin et qui a donné lieu à une publication. Ces travaux ont consisté au développement d'un outil informatique dérivé du Z de Ducher, mesure non-paramétrique de l'association locale. Elle correspond à une déviation normalisée de l'indépendance. D'autres mesures correspondent à des mesures globales de l'association, c'est-à-dire, des mesures uniques censées représenter la force de l'association entre les variables. Celles-ci correspondent à une moyenne de l'association locale entre les modalités des variables. Elles supposent donc que cette force peut être considérée. Le Z de Ducher a été conçu par des physiologistes pour être adapté à la complexité de la physiologie cardiovasculaire. Les travaux réalisés ont consisté à concevoir un outil informatique implémentant le Z de Ducher et l'information mutuelle spécifique : Zebu. Les résultats sont présents dans une publication " A Web Application Designed for Computation of Local Association Measures »

L'ensemble des travaux de recherche réalisés par Olivier Martin montre l'importance des outils mathématiques et informatiques dans le domaine de la santé, aussi bien dans l'analyse et la compréhension des mécanismes physiologiques que dans les mécanismes pathologiques. Le mémoire rédigé par Olivier Martin est remarquablement clair et très didactique pour des non mathématiciens, et recommandé pour une diffusion sur internet comme un ouvrage de référence. Les travaux réalisés apportent des perspectives nombreuses d'application.

Lyon, le 20 mai 2015

**Le Président de la thèse,**



**MARTIN Olivier**

Développement d'un logiciel implémentant une méthode d'analyse de données en sciences de la vie.

Th. D. Pharm., Lyon 1, 2015.

**RÉSUMÉ**

Les tentatives de mathématisation du vivant connaissent de nombreux obstacles de par la complexité et la flexibilité qui caractérisent la vie. Ces propriétés sont souvent utilisées comme argument contre toute approche mathématique. Cependant, les méthodes le plus souvent utilisées sont issues des sciences physiques et de l'ingénieur et sont donc inadaptées pour saisir la vie. Il est néanmoins possible de développer des méthodes adaptées pour son étude. Dans cette thèse, nous prenons un exemple concernant les mesures d'association. Nous les différencions en mesures globales et locales. Les mesures de l'association globales, comme le coefficient de corrélation de Pearson, quantifient une tendance générale en supposant que l'association est continue pour l'ensemble des modalités des variables. Cette supposition n'est pas systématiquement justifiée en sciences de la vie où l'on observe des relations discontinues comme des effets de seuil ou de saturation. Au contraire, les mesures de l'association locale, comme le Z de Ducher, décomposent l'association globale en ses composantes locales. Elles permettent ainsi une meilleure description de la relation de dépendance. Cependant, aucun logiciel ne permet actuellement leur calcul ce qui limite leur utilisation.

Nous présentons le Zébu : une application web permettant le calcul des formes globales et locales de l'information mutuelle et du Z de Ducher. Ce logiciel est facile d'utilisation et est directement accessible sur internet. Il permet à des cliniciens et scientifiques ne sachant pas programmer de mettre en place une étude adoptant une approche locale. Nous espérons que ce logiciel, à terme, stimule l'utilisation de ces mesures d'association locales.

**MOTS CLEFS**

Bioinformatique  
Biostatistiques  
Application web  
Mesures d'association

**JURY**

Mme MOYRET-LALLE Caroline, Pharm.D., Ph.D., MCU, HDR  
M. GOUTELLE Sylvain, Pharm.D., Ph.D., MCU  
M. BOURGUIGNON Laurent, Pharm.D., Ph.D.  
M. DUCHER Michel, Pharm.D., Ph.D, HDR

**DATE DE SOUTENANCE**

Vendredi 19 Juin 2015

**ADRESSE DE L'AUTEUR**

250, cours Lafayette – 69003 Lyon