Contributions à la modélisation statistique et à ses applications en biologie et dans le monde industriel

Frédéric Bertrand

Habilitation à diriger des recherches

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE

UMR 7501

Strasbourg

Université de Strasbourg
Spécialité MATHÉMATIQUES APPLIQUÉES

Frédéric Bertrand

Contributions à la modélisation statistique et à ses applications en biologie et dans le monde industriel

Soutenue le 10 décembre 2018
devant la commission d'examen

Christophe Prud'Homme, garant d'habilitation
Hervé Abdi, rapporteur
Anne-Laure Boulesteix, examinatrice
Marianne Clausel, rapporteuse
Anne Gégout-Petit, rapporteuse
Gilbert Saporta, examinateur

https://irma.math.unistra.fr

À Anaëlle, David et Myriam.

Remerciements

Je tiens en premier lieu à remercier Hervé Abdi, Marianne Clausel et Anne Gégout-Petit d'avoir accepté de prendre le temps de lire ce mémoire et d'en être les rapporteurs. Je les remercie également, ainsi que tous les autres membres du jury Anne-Laure Boulesteix, Christophe Prud'homme et Gilbert Saporta, d'avoir accepté d'assister à la présentation de ce travail. Une merci particulier à Christophe Prud'homme d'avoir accepté d'être mon garant d'habilitation.

Je voudrais remercier Dominique Collombier et Photis Nobelis pour m'avoir transmis, lors de ma thèse, leur passion pour la statistique sous toutes ses formes.

Je tiens à remercier le Labex de mathématiques IRMIA et ses directeurs successifs, Thomas Delzant et Nalini Anantharaman, pour avoir soutenu quelques-unes de mes demandes de moyens pour la recherche et en particulier deux financements de thèse, que j'ai pu co-encadrer. Je souhaite également remercier l'IDEX de l'université de Strasbourg, et plus particulièrement Catherine Florentz, qui a soutenu certaines de mes initiatives de projets transdisciplinaires.

Les aspects statistiques de certains des travaux présentés dans ce mémoire ont été réalisés en collaboration avec une autre chercheuse de l'IRMA, M. Maumy-Bertrand, par deux fois, avec un autre chercheur appartenant au monde industriel, P. Bastien, et quelques fois encore avec l'un ou l'une des quatre étudiants H. Alawieh, N. Jung, J. Magnanensi et T.-A. Nengsih pendant qu'ils préparaient leur thèse de doctorat.

En ce qui concerne les aspects transdisciplinaires de mes recherches, je tiens à remercier les collaborateurs avec lesquels j'ai pu personnellement échanger lors de l'exécution de ces travaux, même si la liste est plus longue : I. Aouadi, S. Bahram, J. Barbe, S. Betoulle, S. Blanc, M. Boos, C. Carapito, R. Carapito, K. El Bayed, A. Forgione, L.-M. Fornecker, L. Fussler, Ph. Georgel, J. Grenèche, S. Giroud, J. Grosman, S. Jacquet, N. Kobes, Ph. Kuntzmann, N. Meyer, I-J. Namer, O. Petit, C. Pionneau, M. Pleynet, A. Rolland,

F. Rubino, S. Savary, C. Schleiss, P. Tassi, B. Thierry, L. Vallat et C. Zimmer.

Je tiens également à remercier J-J. Droesbeke, G. Saporta et C. Thomas-Agnan pour leur confiance qui m'a permis de m'initier à l'activité d'édition scientifique en intégrant plusieurs fois le groupe des éditeurs de l'ouvrage des Journées d'Étude en Statistique.

Je remercie M. Chion qui vient de s'engager dans la préparation d'un doctorat sur un sujet transdisciplinaire à l'interface entre statistique et chimie en co-direction avec C. Carapito que je remercie ici plus particulièrement pour ses discussions et son enthousiasme qui ont permis de concevoir ce projet particulièrement intéressant.

Je remercie le personnel administratif de l'UFR de Mathématique-Informatique, plus particulièrement Sandrine Cerdan, ainsi que le personnel administratif de l'IRMA, pour leur disponibilité, leur efficacité et leur gentillesse. Je n'oublie bien sûr pas de remercier le personnel administratif de la Faculté des Sciences de la Vie, celui de l'École doctorale Vie et Santé ainsi que celui du Collège Doctoral.

J'ai certainement oublié de remercier d'autres personnes. Si quelqu'un se trouve dans cette situation, je le prie de bien vouloir m'excuser et je le remercie chaleureusement lui aussi, ou elle aussi.

Un dernier mot pour les êtres qui me sont chers : ma femme Myriam et mes deux enfants Anaëlle et David, mes parents Aline et Guy, mon frère Guillaume et ma belle-sœur Marie. Je les remercie pour leur soutien permanent et les joies nombreuses qu'ils m'ont apportées ou qu'ils m'apportent encore tous les jours.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Expériences en matière de recherche

Au cours des dix dernières années, mes travaux et activités de recherche se sont concentrés autour de plusieurs thématiques fédératrices fortes. Au cours du temps, certaines ont évoluées et d'autres sont apparues mais toutes se sont révélées passionnantes. À ce jour, la plus importante d'entre elles s'avère être l'inférence statistique dans un contexte de grande ou d'ultra grande dimension en présence éventuelle de censure ou de données manquantes. Je me suis intéressé aussi bien au cas de la régression linéaire, de la régression linéaire généralisée qu'à d'autres contextes de régression, comme la régression bêta ou le modèle de Cox, pour lesquels j'ai non seulement produit de nouveaux critères de choix de modèle mais aussi de sélection de variables. Les contextes d'application envisagés m'ont naturellement incité à m'appuyer sur des approches de régression pénalisée, typiquement *lasso*, *ridge* ou *elasticnet*, de régression par les moindres carrés partiels parcimonieuse ou non. Je développe depuis un an une thématique liée à l'apprentissage statistique et aux réseaux de neurones.

La première thématique que j'ai abordée est la planification expérimentale et en particulier les plans sphériques de force t. J'y ai été confronté lors de ma thèse de doctorat (Bertrand [2007]). Ma préoccupation était alors d'obtenir des résultats précis d'existence de ces objets combinatoires. Afin d'y parvenir j'ai principalement utilisé des résultats d'algèbre et le contexte défini par la statistique algébrique. Bien que la planification expérimentale soit un sujet par essence appliqué, il s'agissait exclusivement dans ce cas d'un travail de nature théorique.

En tant que futur chercheur en statistique, il m'a rapidement semblé primordial d'avoir une perception précise de la manière dont la statistique mathématique

était appliquée ensuite par ses utilisateurs. C'est une des manières de parvenir à proposer des résultats de mathématiques appliquées qui serviront à résoudre des problèmes réels et actuels. J'ai donc également cherché, dès 2005 et alors que j'étais encore en thèse de doctorat, à me confronter à des problèmes que les expérimentateurs considéraient comme ouverts, c'est-à-dire des problèmes qu'aucune des techniques statistiques existantes ne leur avait permis de résoudre (Bertrand et Maumy [2007a]). J'ai également eu très tôt l'opportunité de participer ou de mettre en l'œuvre l'analyse de jeux de données complexes ayant souvent une composante temporelle voire spatio-temporelle comme ceux récoltés par l'observatoire national des maladies du bois de la vigne (Kobes *et al.* [2007], Bertrand *et al.* [2007a], Bertrand *et al.* [2008], Kuntzmann *et al.* [2013]), ceux consacrés à l'étude des assemblages d'espèces de phytoplanctons (Rolland *et al.* [2009], Bertrand et Maumy [2010]) ou à des problématiques d'écophysiologie (Giroud *et al.* [2008]), d'écologie (Zimmer *et al.* [2011]), d'éthologie (Petit *et al.* [2008]) ou de médecine (Grenèche *et al.* [2011b], Grenèche *et al.* [2011a], Grenèche *et al.* [2013], Vallat *et al.* [2013], Carapito *et al.* [2016]). Un projet particulièrement ambitieux m'a occupé ces six dernières années : la modélisation de données multi-omiques (gènes et protéines) avec comme finalité la réalisation d'interventions dirigées dans des cellules cancéreuses. Plusieurs articles sont en cours de rédaction ou de révision (Bertrand *et al.* [2018d], Schleiss *et al.* [2018a], Schleiss *et al.* [2018b], Fornecker *et al.* [2018]) et j'ai produit plusieurs rapports de recherche et d'analyse intermédiaires (Bertrand [2015c], Bertrand [2015b], Bertrand [2015a], Bertrand [2016c], Bertrand [2016b], Bertrand [2017b], Bertrand [2017a]).

Ainsi, depuis le début de mes travaux de recherche j'ai abordé plusieurs problématiques statistiques aussi bien d'un point de vue théorique que computationnel : les plans d'expériences (Bertrand [2008b], Bertrand [2008a], Bertrand [2009], Bertrand [2010]), la régression pénalisée (Vallat *et al.* [2013], Jung *et al.* [2014c], la prépublication Aouadi *et al.* [2018]), la régression des moindres carrés partiels (PLS) (Magnanensi *et al.* [2016a], Magnanensi *et al.* [2017] et la prépublication Nengsih *et al.* [2018]) et certaines de leurs extensions (Meyer *et al.* [2010], Bertrand *et al.* [2013a], Bastien *et al.* [2015] et la prépublication Magnanensi *et al.* [2016b]) ainsi qu'une extension aléatoire de l'ajustement multidimensionnel (*multidimensional fitting*), la prépublication Alawieh *et al.* [2018]. Il ne faut pas non plus oublier l'inférence de réseaux biologiques qui m'a permis aussi bien d'appliquer certains des résultats précédents que d'en développer de nouveaux. Ces quatre thématiques restent actives à ce jour et, depuis un an, je suis en train d'y ajouter une composante formée des réseaux de neurones et plus généralement de l'apprentissage statistique. Mes principaux domaines d'application de cette nouvelle thématique sont la statistique industrielle et les systèmes complexes.

Comme je l'ai indiqué plus haut, les impulsions à l'origine de ces recherches

peuvent se classer en trois catégories.

1. Soit purement théorique : comme le développement d'outils de construction de plans d'expériences exacts et la preuve d'existence de solutions isovariantes exactes dans $\mathbb{R}^3$ ou la recherche d'une preuve de non-existence de certaines structures combinatoires,

2. soit dans un but d'améliorer ou de compléter une méthodologie existante mais qui a montré ses limites (traitement des valeurs manquantes et des problèmes de colinéarité en régression logistique ou pour les modèles de Cox, influence des valeurs manquantes en régression PLS, critère de choix de variables ou du nombre de composantes en régression PLS et ses extensions modèles linéaires généralisés, cas des modèles parcimonieux)

3. soit même dues à la nécessité de concevoir des outils spécifiques pour des expériences innovantes pour lesquelles il fallait créer une solution faite sur mesure (inférence temporelle de réseaux de gènes, inférence conjointe de réseaux aussi bien au niveau population qu'au niveau individuel, détermination de cibles optimales pour prédire une intervention dirigée fiable dans un réseau biologique).

J'essaye, dans la mesure du possible, d'assurer un aller-retour constant entre des développements généraux et des applications de ceux-ci dans l'un des projets transdisciplinaires auquel je participe. Dès le début de mes travaux de recherche, j'ai manifesté un intérêt pour les applications de la statistique et je ne vois pas cette discipline comme étant une discipline « hors-sol ».

Par conséquent, je me suis rendu compte très rapidement que les hypothèses communément faites pour permettre une exploitation scientifiquement rigoureuse des modèles statistiques ne sont qu'exceptionnellement compatibles avec des jeux de données réels, même s'ils sont collectés avec les meilleurs protocoles expérimentaux puisque les problèmes rencontrés tiennent généralement à la nature même des observations et non à la méthodologie de mesure. Or ce sont ces hypothèses qui permettent une gestion rigoureuse des risques d'erreur qui apparaissent dans la théorie des tests de significativité ou des niveaux de confiance présents lors de la construction de régions de confiance. En outre, les séries statistiques ou les échantillons réels présentent souvent des problématiques additionnelles comme l'absence de certaines valeurs, présence de valeurs manquantes avec des mécanismes d'apparition plus ou moins complexes en fonction de l'appareil ou de la méthodologie de collecte des mesures, ou la présence de valeurs atypiques qui peuvent grandement influencer les outils statistiques utilisés.

Ainsi, après une première phase de développement d'un outil dans un cadre classique, typiquement celui de la régression pénalisée ou de la régression PLS, je

m'intéresse à étudier ses propriétés en présence de valeurs manquantes ou à en proposer une extension robuste.

Mon projet de recherche, voir le chapitre 7, propose des axes qui me permettront de continuer de s'intéresser à ces thématiques (régression pénalisée, régression PLS, inférence de réseaux biologiques, modèles pour données protéiques, intelligence artificielle). Plusieurs demandes de financement sur ces axes ont été reçues positivement et ont remporté plusieurs appels à projets (PEPS, allocations doctorales).

1.2 Expériences en matière d'encadrement

1.2.1 Encadrement de doctorants et de chercheurs

- septembre 2018 – : Encadrant, à 100% pour la partie mathématique, de Marie Chion, étudiante en thèse. Christine Carapito, chargée de recherche en chimie, nous apporte son expertise sur la protéomique, thématique d'application de la thèse, et participe donc à la direction de la thèse.

- décembre 2016 – : Co-Encadrant à 80% (100% pour la partie mathématique) de Titin Agustin Nengsih, étudiante en thèse. Le PU-PH N. Meyer assure la part restante de la direction de la thèse.

- mai 2015 – octobre 2015 : Co-supervision à 50% d'une ingénieure d'étude, Khadija Musayeva. Khadija a poursuivi son parcours par une thèse en apprentissage statistique au LORIA.

- octobre 2014 – mai 2016 : Co-supervision à 50% d'un ingénieur de recherche, Marius Kwemou.

- décembre 2013 – novembre 2015 : Supervision à 100% d'un chercheur post doctorant, Théo Rietsch.

- novembre 2012 – décembre 2015 : Co-Encadrant à 80% (100% pour la partie mathématique) de Jérémy Magnanensi, étudiant en thèse. Le PU-PH N. Meyer a assuré la part restante de la direction de la thèse. Jérémy est chercheur en CDI dans une société de biostatistique.

1.2.2 Encadrement de stages et de mémoires d'étudiants

- 2017 – 2018 : Encadrant d'un stage d'IUT deuxième année

- 2016 – 2017 : Encadrant d'un mémoire de Master Recherche deuxième année

- 2015 – 2016 : Encadrant d'un mémoire d'agrégation (M2) et de deux Master deuxième année d'actuariat
- 2014 – 2015 : Encadrant d'un mémoire de Master deuxième année
- 2013 – 2014 : Encadrant d'un stage de deuxième année en école d'ingénieur, d'un mémoire de Master deuxième année
- 2012 – 2013 : Encadrant d'un stage de première année en école d'ingénieur
- 2011 – 2012 : Encadrant d'un stage de fin d'étude en école d'ingénieur
- 2010 – 2011 : Encadrant de deux stages de Master deuxième année, d'un stage de Licence troisième année et d'un de Master première année de Magistère de mathématiques
- 2009 – 2010 : Encadrant de deux stages de Master deuxième année, d'un stage de Licence troisième année de Magistère de mathématiques
- 2008 – 2009 : Encadrant de deux stages et de deux projets Informatiques de Master deuxième Année
- 2007 – 2008 : Encadrant de deux stages et de deux projets Informatiques de Master deuxième Année
- 2006 – 2007 : Encadrant d'un stage et d'un projet Informatique de Master deuxième Année

1.3 Expériences en matière d'enseignement

C'est naturellement que je suis intervenu et que j'interviens encore dans divers composantes (chimie, pharmacie, biologie, service de formation continue) ou école doctorale (Vie et Santé) de l'université de Strasbourg en plus de l'UFR de mathématique-informatique.

Par conséquent j'ai été confronté à des publics variés tous de niveaux master ou plus lorsqu'il s'agit d'intervention dans un module d'école doctorale. Cette diversité a été à la fois très intéressante et stimulante mais aussi très chronophage. Ainsi, par exemple, j'ai dû me former à différents logiciels ou langages informatique en fonction des besoins du public concerné. J'utilise R, SAS, python, Excel et XLStat dans mes enseignements mais j'ai aussi dû utiliser dans le passé SPSS et Statistica.

Depuis 2009, j'ai rédigé onze livres, dont deux ont été réédités, et donc retravaillés, trois fois, pour un public d'étudiants de licence, de master ou en École d'ingénieur (Fredon *et al.* [2009a], Fredon *et al.* [2009b], Fredon *et al.* [2009c], Bertrand

et al. [2011], Bertrand et Maumy-Bertrand [2011], Bertrand et Maumy-Bertrand [2012], Bertrand *et al.* [2013a], Bertrand *et al.* [2016], Bertrand et Maumy-Bertrand [2018c] 3^{e} édition, Bertrand *et al.* [2018b], Meyer *et al.* [2018] 3^{e} édition).

1.4 Expériences en matière d'expertise scientifique

Depuis 2005, je participe, sans interruption, à des projets de recherche en partenariat avec des acteurs du monde industriel. J'ai exercé mon expertise scientifique dans les cas suivants :

1. pour Veolia sur plusieurs sujets (2005-2007 et 2007-2008),
2. pour Lilly (2005-2007),
3. pour Interstat (2006-2010),
4. pour GlaxoSmithKline (2007-2008),
5. pour SkyePharma sur plusieurs sujets (2008 et 2009),
6. pour Avenseo (2009),
7. pour Sorin Group sur plusieurs sujets (2009, 2009-2010 et 2010),
8. pour NaturaConst (2011-2012),
9. pour Merck sur plusieurs sujets (depuis 2014),
10. pour Yourdata à trois reprises (depuis 2016) et
11. pour la SATT sur projet de prématuration (2016-2017) qui a été converti en projet de maturation (depuis 2017).

1.5 Logiciels

Ayant toujours souhaité rendre accessibles au plus grand nombre les outils que j'ai développés, j'ai eu constamment à cœur de les rendre facilement utilisables. Chaque fois que cela a été possible, c'est-à-dire compatible avec les contraintes de confidentialités des projets, j'ai réalisé et distribué des programmes informatiques, le plus souvent des *packages* pour le langage R. En voici la liste :

- hébergés sur le site du CRAN, `https://cran.r-project.org`,

  - `plsRglm`, Bertrand et Maumy-Bertrand [2018g], Bertrand et Maumy-Bertrand [2018f] et Bertrand *et al.* [2014d], site internet `https://cran.r-project.org/web/packages/plsRglm/index.html`,
  - `plsRbeta`, Bertrand *et al.* [2013b] et Bertrand et Maumy-Bertrand [2018d], site internet `https://cran.r-project.org/web/packages/plsRbeta/index.html`,
  - `plsRcox`, Bertrand et Maumy-Bertrand [2018e], Bertrand *et al.* [2014b] et Bastien *et al.* [2015], site internet `https://cran.r-project.org/web/packages/plsRcox/index.html`,

- hébergé ici `http://www.math.unistra.fr/genpred/spip.php?rubrique4`, car la taille des fichiers d'exemple est trop importante pour le CRAN,
  - `Cascade`, Jung *et al.* [2018], Jung *et al.* [2014a], Jung *et al.* [2014c]
- ou en cours de finalisation et bientôt sur le CRAN,
  - `selectboost`, Bertrand *et al.* [2018c] et
  - `Patterns`, Bertrand et Maumy-Bertrand [2018a],

voire des feuilles de calculs Maple lorsque cela s'est avéré nécessaire

- développements d'Edgeworth, Bertrand et Maumy-Bertrand [2018b],
- problèmes de construction polynomiale de plans d'expériences.

J'ai repris la maintenance du package `plsdof` suite au désistement de ses créateurs.

1.6 Divers

J'ai obtenu la PEDR en 2012. Elle a été renouvelée en 2016. J'ai été membre élu du CNU 26ème section de 2011 à 2015, puis réélu en 2015. Je suis membre du conseil des directeurs de l'European Regional Section (ERS) de l'International Association for Statistical Computing (IASC). J'ai également une expérience de l'édition d'ouvrages scientifique (Bertrand *et al.* [2017], Bertrand *et al.* [2019]) et participation à cette édition (Droesbeke *et al.* [2014], Maumy-Bertrand *et al.* [2018]).

Chapitre 2

Contrôle de qualité et plans d'expériences

2.1 Contrôle de qualité

Un des problèmes des laboratoires de contrôle ou de l'industrie chimique, pharmaceutique, agroalimentaire, etc., est l'estimation de la proportion d'une population de mesures se situant entre deux limites. La littérature est relativement abondante sur ce sujet : citons par exemple Boullion *et al.* [1985] et Mee [1988] pour de plus amples détails. Supposons qu'une série de mesures suive une loi normale de moyenne μ et de variance $\sigma^2 > 0$ inconnues et que la proportion des mesures appartenant à l'intervalle $[L,U]$, L pour Lower et U pour Upper, notée π, est le paramètre d'intérêt. Par exemple, dans les laboratoires de contrôle, $[L,U]$ représente l'étendue des valeurs produites acceptables et π est alors la proportion d'unités conformes.

Cette thématique d'estimation de la proportion π m'a été soumise lors d'un séjour de deux mois comme chercheur invité dans l'équipe European Early Phase Statistics d'Eli Lilly à Louvain-la-Neuve, Belgique. Afin d'améliorer les outils existants de contrôle de qualité et de validation de méthodes sur de petits échantillons, plusieurs développements ont été proposés. En premier lieu, dans Bertrand et Maumy [2007a], nous proposons deux estimateurs possibles pour la proportion π et établissons des développements d'Edgeworth, Withers [1983], qui permettent de construire des intervalles de confiance généralement plus précis, voir Brown *et al.* [2001], Brown *et al.* [2002], Brown *et al.* [2003], Cai [2005] pour le cas de la famille exponentielle. Le recours à ces développements se justifie par la faible taille des séries de mesures utilisées, souvent comprise entre 5 et 20, et par conséquent, par le fait que les approximations asymptotiques ne sont plus acceptables dans ce cas.

Le calcul explicite a été réalisé à l'aide de Maple, Maple Team [2014], et un programme général de détermination des développements d'Edgeworth d'ordre 1 ou 2 a été créé. Il a permis de détecter une erreur dans le développement proposé dans le Théorème 13.5 page 194 du livre Das Gupta [2008]. Le calcul explicite des intervalles a été communiqué dans Bertrand et Maumy [2007b].

Ces développements d'Edgeworth doivent être combinées à l'utilisation de techniques de bootstrap, Hall [1992], pour la détermination des intervalles de confiance. Les résultats obtenus avec Maple sont alors combinés avec les fonctionnalités de R : bootstrap, Canty et Ripley [2014]; Davison et Hinkley [1997], et réarrangement croissant pour les développements d'Edgeworth et de Cornish-Fisher Chernozhukov *et al.* [2010]; Graybill *et al.* [2011]. Un tutoriel de vingt pages incluant le calcul des développements de la moyenne et de la variance ainsi qu'un exemple plus complexe d'estimation par maximum de vraisemblance est détaillé dans la prépublication disponible en ligne Bertrand et Maumy-Bertrand [2018b].

En second lieu, nous avons proposé des modifications des règles de décision existantes dans Bertrand et Maumy [2008] en utilisant ces nouveaux intervalles ou en améliorant des intervalles existant à l'aide des résultats de Cai [2005].

2.2 Statistique algébrique, plans d'expérience et recherche de réfutations

En 2016-2017, j'ai proposé un mémoire de Master 2^{e} année sur les plans sphériques de force t avec comme application le problème ouvert de l'existence d'un plan de force 5 ou de force 4 de taille 11 dans $\mathbb{R}^3$. En effet, si $n = 3$, alors il existe :

- des plans sphériques de force 4 de taille égale 12, 14 ou supérieure ou égale à 16, Hardin et Sloane [1992],
- des plans sphériques de force 5 de taille égale 12, 14, 16, 18, 20 ou supérieure ou égale à 22, Reznick [1995], Hardin et Sloane [1996].

En 2009, dans leur article, Bannai et Bannai [2009] indiquait que ce problème est encore ouvert.

Je m'étais précédemment attaché à construire algébriquement des plans sphériques de force t pour les effectifs mentionnés ci-dessus, et bien d'autres encore, dans Bertrand [2008b]. J'avais également montré, dans Bertrand [2008a, 2010], comment lier la non-existence d'un plan sphérique avec l'existence d'une réfutation qui est une équation algébrique particulière. Une telle réfutation peut être recherchée à l'aide d'algorithmes d'algèbre commutative et d'outils de programmation semi-définie positive comme Papachristodoulou *et al.* [2013]. La difficulté

étant la dimension du problème à résoudre. Un programme de calcul a été mis en place, mais, pour l'instant, le temps a manqué pour pouvoir le mettre en œuvre efficacement et en particulier essayer de paralléliser les calculs ou d'en déporter une partie sur des GPU.

2.3 Utilisation des plans d'expérience

2.3.1 Optimalité et approche population en pharmacologie

Dans de nombreuses situations pratiques, l'expérimentateur peut être amené à estimer les constantes inconnues d'un modèle de PK-PD. Pour cela, une des méthodes préconisées est de déterminer un ensemble de points d'échantillonnage optimaux. Il faut noter qu'il s'agit, dans l'étude de cas que nous avons menée, d'une approche population. La connaissance des points support du plan permet alors non seulement de déterminer le nombre de données expérimentales qu'il faut récolter mais aussi les temps d'échantillonnage auxquels il faut faire les mesures.

J'ai été amené à m'intéresser à ce type de problématique lors d'un séjour de deux mois comme chercheur invité dans l'équipe European Early Phase Statistics d'Eli Lilly à Louvain-la-Neuve, Belgique. J'ai encadré le mémoire d'une étudiante de DESS sur ce sujet.

Nous avons en particulier utilisé le logiciel POPT pour déterminer le plan d'expérience cherché. Ce logiciel permettait d'obtenir cet ensemble de points à l'aide de différents procédés numériques comme l'algorithme du simplexe, le recuit simulé ou l'algorithme de Fedorov-Wynn, dans le cas du simple ou double échange. Ces algorithmes sont combinés à la résolution numérique, par la méthode de Runge-Kutta, du système d'équations différentielles définissant le modèle PK-PD lorsque qu'une forme explicite des solutions ne peut être établie. C'est majoritairement le cas lorsque le modèle comporte une partie dynamique, et plus particulièrement, la situation dans laquelle a été réalisée cette étude. Les résultats de ce travail ont été publiés dans les actes d'une conférence Bertrand *et al.* [2007b].

2.3.2 Planification d'expériences pour la détermination d'un biomarqueur

Dans le cadre d'une étude en prématuration et en maturation de la SATT Conectus, j'ai mise en place, en concertation avec les chimistes et les médecins, les deux plans de mesures protéomiques à réaliser. Pour des raisons de confidentialité, il m'est impossible de donner des détails sur ces travaux qui toujours en cours mais qui présentent plusieurs aspects statistiques intéressants comme la présence

de *pools* d'échantillons ou d'observations longitudinales sur certains individus, voir la section 6.5.3.

2.3.3 Interactions et santé

Je suis responsable d'équipe dans le projet REPERTOX - REPERage et hiérarchisation du risque éco-TOXique associé aux pratiques phytosanitaires en secteur agri-viticole, déposé au Programme National de Recherche en Environnement-Santé-Travail. Le projet est actuellement en recherche de financement.

La France demeure le premier utilisateur de pesticides en Europe avec près de 60000 tonnes de substances actives vendues annuellement. L'utilisation de pesticides s'y accroît régulièrement au fil des années (Augmentation moyenne de 5,8% au cours de la période 2011-2014 ; Ministère de l'Agriculture, 2016). L'estimation du niveau de risque écotoxique lié à l'exposition des populations humaines ou animales aux pesticides est délicate dans un contexte correspondant à des expositions chroniques à de faibles doses de mélanges de contaminants.

L'objectif est d'analyser le lien entre l'exposition aux produits phytopharmaceutiques et les effets sur la santé humaine, en prenant en compte au mieux les systèmes de production, les pratiques agricoles et les effets cocktails.

Ce projet rejoint la thématique d'analyse des maladies du bois de la vigne et combine l'utilisation de plans d'expérience pour des expériences contrôlées en laboratoire à des approches factorielles ou multitableaux.

Chapitre 3

Modélisation en biologie

Mes premières collaborations en biologie m'ont permis de me rendre compte des besoins des chercheurs utilisateurs de la statistique, d'être confronté à des jeux de données réels et de proposer plus tard des développements théoriques ou méthodologiques issus de problématiques réelles. Ces collaborations, le plus souvent sur plusieurs années, m'ont également formé à la gestion de projet. En effet, j'ai participé à l'analyse des données d'un observatoire national pendant près de dix ans et à l'encadrement des étudiants qui ont mis en œuvre ce traitement statistique. J'ai procédé moi-même au dépouillement des résultats de quatre thèses pour lesquelles des techniques statistiques sophistiquées étaient requises afin de tirer le meilleur parti des données collectées. J'ai également encadré deux ingénieurs statisticiens en leur indiquant des choix méthodologiques d'analyse puis en les aidant à mettre en forme leurs résultats afin de pouvoir les diffuser auprès des chercheurs.

J'ai commencé par effectuer l'analyse des résultats d'études cliniques réalisées par deux chirurgiens du service de chirurgie digestive de l'IRCAD-EITS. Étant jeune à l'époque, je n'ai pas particulièrement demandé à être associé aux publications qui concernaient trois études dont une sur la possibilité d'un traitement chirurgical du diabète. La difficulté statistique était principalement de travailler avec de (tout) petits échantillons. Ainsi, très rapidement, je me suis rendu compte que les hypothèses usuelles du modèle linéaire gaussien sont pas valides dans un grand nombre de situations expérimentales. En effet, comment tester l'hypothèse de normalité avec des tailles d'échantillon allant d'une dizaine à deux voire dans de rares cas trois dizaines de valeurs. Il est bien connu que les meilleurs de normalité *omnibus* ne commencent à déceler « efficacement » les défauts de normalité que pour des échantillons de taille au moins 30, Thode [2002]. Ces expériences m'ont fait connaître d'autres approches pour sélectionner ou valider des modèles statistiques comme la validation croisée Hastie *et al.* [2008] ou les approches par per-

mutations Good [2005]. Ces influences se sont manifestées lorsque j'ai moi-même développé de nouveaux outils statistiques, voir les chapitres 4, 5 et 6.

3.1 Approches factorielles, modèles linéaires généralisées et techniques multitableaux

3.1.1 Analyse des résultats d'un observatoire national

L'observatoire a duré de 2003 à 2008. Sa création était liée à un problème ayant un fort impact pour la communauté viticole, comme le résume bien l'abstract de Bertrand *et al.* [2008] :

> L'objectif de l'Observatoire National des Maladies du Bois de la Vigne est de dresser un état des lieux de la répartition, de la fréquence et de l'intensité de l'expression des symptômes foliaires des maladies du bois, pour répondre à la question de leur progression dans le vignoble français. En effet, suite à une interdiction de l'utilisation de l'arsénite de soude, les viticulteurs ne disposent plus d'aucune méthode de lutte chimique curative homologuée contre les maladies du bois de la vigne. Cet observatoire collecte, chaque année, depuis 2003, un ensemble de données cohérentes. Le jeu de données est complexe : il comporte des variables quantitatives et qualitatives qui évoluent au cours du temps. La problématique de l'étude est de dégager les grandes tendances en matière d'épidémiologie végétale afin de déterminer quelles sont les mesures prophylactiques à mettre en œuvre collectivement et à grande échelle.

Si la DRAF-SRPV Alsace, chargée de l'étude des résultats de l'observatoire par son Ministère de tutelle, nous a sollicité, c'est avant tout pour avoir un fort appui méthodologique, concernant les problématiques statistiques, dans le traitement de ces données. En effet, devant la complexité de l'analyse à mener, un avis d'expert était nécessaire pour obtenir des résultats fiables. Pour les aspects relevant de la phytopathologie, une collaboration avec l'UMR Santé Végétale du centre INRA de Bordeaux ainsi que la DRAF-SRPV Rhône a été mise en place afin d'essayer de tirer le meilleur parti des données en cours de collecte.

Ma contribution à ce travail a été de proposer une méthodologie pertinente et d'encadrer successivement plusieurs stagiaires sur ce sujet. En effet, l'observatoire a collecté des données pendant plusieurs années successives et une mise à jour annuelle des résultats était donc nécessaire.

Le travail statistique s'est articulé en trois points. Nous avons mis en évidence des relations entre les différentes variables de l'étude, puis nous avons utilisé l'analyse des correspondances multiples, l'analyse en composantes principales et l'analyse factorielle de données mixtes. Ces premiers résultats ont été suffisamment pertinents pour être publiés, non seulement dans une revue de vulgarisation scientifique à l'attention de la communauté viticole, Kobes *et al.* [2007], mais aussi dans la revue de société américaine de phytopathologie spécialisée dans le domaine Fussler *et al.* [2008].

Ensuite, pour tenir compte des différences de nature entre les variables nous avons utilisé l'analyse factorielle des données mixtes, puis pour intégrer le facteur temps, nous avons employé des méthodes d'analyse factorielle de *K-tables CA*. Ces approches par tableaux multiples ont été réalisées dans un second temps puisqu'elles n'ont été possibles qu'à partir du moment où la durée d'observation a été suffisante. Nous avons soumis puis avons été sélectionné pour présenter cette analyse complétée à une session spéciale d'études de cas organisée lors des journées de la statistique en 2007 à Angers, ce qui a amené à la publication Bertrand *et al.* [2007a].

Enfin, afin de préciser les relations décelées, notre choix s'est porté sur des modèles de régressions logistiques binaires et ordinales. Nous avons utilisé des techniques bootstrap pour construire des régions de confiance autour de leurs paramètres. Bien que ces études aient utilisé des outils statistiques existants, elles ont fait l'objet de choix méthodologiques qu'il nous a semblé pertinent de publier dans Bertrand *et al.* [2008].

Suite à ce travail, j'ai été invité à faire partie du comité de pilotage du projet CASDAR, coordonné par l'INRA de Bordeaux : « Épidémiologie de l'esca/BDA de la vigne : Dynamique spatio-temporelle, modélisation et facteurs de risque. »

Enfin, c'est suite à une sollicitation locale qu'une dernière collaboration sur cette thématique des maladies du bois de la vigne a eu lieu. Un nouveau jeu de données a été analysé avec des techniques d'analyse des correspondances multiples mais aussi de la régression moindres carrés partiels (PLS) car il présentait des valeurs manquantes et des variables posant des problèmes de colinéarité. L'analyse a été mise en œuvre par une stagiaire et a donné lieu à une publication Kuntzmann *et al.* [2013].

3.1.2 Dynamiques spatio-temporelles d'espèces

Ce travail s'inscrit dans une collaboration à une thèse de l'UMR INRA 42 (CARRTEL), Station d'Hydrobiologie Lacustre. Les micro-organismes, en particulier les

espèces de phytoplancton, peuvent être considérés comme des indicateurs des changements locaux et plus globaux dans les écosystèmes aquatiques et peuvent donc constituer un excellent biomarqueur de la qualité de l'eau. évaluer l'influence des variables d'environnement biologiques, chimiques et physiques sur la régulation du phytoplancton est une étape clef pour parvenir à comprendre la structure et la dynamique des populations, leur diversité et leur succession afin de proposer, si nécessaire et si possible, une intervention humaine avant que toute prolifération excessive d'algues puisse se produire. Ces questions sont d'un intérêt premier pour les scientifiques et les gestionnaires d'eau afin de permettre aux réservoirs d'eau de grande taille, lacs et étangs, d'atteindre le « bon état écologique » recommandé par la directive-cadre sur l'eau (DCE) d'ici 2015.

Ce projet se concentrait sur l'étude du réservoir Marne (bassin de la Seine), l'un des plus grands réservoirs en Europe occidentale. En 2006, c'est-à-dire la première année du projet, le réservoir a été échantillonné une fois par mois en mars et avril, puis une fois toutes les deux semaines entre mai et septembre. Pour évaluer l'hétérogénéité spatiale, six stations et différentes profondeurs pour chaque station ont été étudiée.

Malheureusement, comme indiqué dans Rolland *et al.* [2009], ces jeux de données posent des problèmes spécifiques dus à leur échantillonnage spatio-temporel. La dynamique et la diversité du phytoplancton sont particulièrement difficiles à analyser, en particulier lorsque

1. la granularité de l'analyse se situe au niveau de l'espèce,
2. la diversité des espèces est élevée,
3. l'étude couvre plusieurs saisons et
4. l'échantillonnage a été réalisé dans de nombreuses stations de l'écosystème.

L'analyse triadique partielle est une méthode d'analyse multitableaux qui est un outil statistique adapté pour obtenir une représentation claire d'une série chronologique, une pour chaque date de prélèvement, de matrices observées (abondances des espèces à chacune des stations). Elle permet l'analyse en composantes principales simultanée de plusieurs matrices et permet de trouver une structure spatiale commune à chacune de ces matrices et d'étudier la stabilité temporelle de celle-ci. L'analyse triadique partielle commence par la recherche d'un tableau moyen appelé compromis. Le tableau compromis est ensuite analysé et sa reproductibilité pour chacune des tables initiales est finalement étudiée.

Cette technique a été exploitée avec succès dans Rolland *et al.* [2009] pour décrypter l'organisation spatio-temporelle des assemblages d'espèce de phytoplanc-

ton. La méthodologie employée pour l'analyse statistique a été publié comme une étude de cas dans Bertrand et Maumy [2010].

3.2 Modèles mixtes

3.2.1 Mesures doublement répétées et splines cubiques

L'originalité statistique des données étudiées lors de cette collaboration à une thèse avec le Laboratoire de Psychologie des Cognitions (EA 4440-UdS), Strasbourg, France provient de la présence de données doublement répétées sur les sujets, mesures répétées lors de sessions elles-mêmes répétées. Le nombre de répétitions au sein d'une session de 32h pouvant être élevée, au maximum toutes les heures après la nuit de repos soit 25 tests pour l'étude de la puissance des ondes cérébrales par électro-encéphalogramme, des modèles mixtes non paramétriques basés sur des splines cubiques ont été utilisés, voir Grenèche *et al.* [2011b].

En effet, cette étude de suivi a été menée avec des tâches à mémoire et répétée chez 12 patients atteints du SAHOS et 10 témoins sains ayant subi trois séances de 32 heures, la première avant la PPC (T0), la deuxième (T3) et la troisième (T6) respectivement après trois et six mois de traitement pour les patients souffrant d'apnée du sommeil et de syndrome d'hypopnée obstructive (SAHOS). Chaque session comprenait une nuit de sommeil suivie de 24 heures d'éveil prolongé pendant lesquelles les deux groupes effectuaient des tâches sollicitant leur mémoire à court terme (STM), y compris des tâches numériques (DS) et des tâches Sternberg.

Peu d'études ont examiné l'impact de la thérapie par pression positive continue (CPAP) sur la mémoire à court terme (STM) par rapport à l'état de veille prolongé chez les patients atteints d'apnée du sommeil et de syndrome d'hypopnée obstructive (SAHOS). Nous avons cherché à savoir si le traitement CPAP pouvait inverser la dégradation de la STM dans un paradigme de veille continue de 24 heures. La quantité de données recueillie et exploitée pendant ce travail de thèse était d'une ampleur très conséquente et a donné lieu à trois publications : Grenèche *et al.* [2011b], Grenèche *et al.* [2011a] et Grenèche *et al.* [2013].

3.2.2 Modèles linéaires généralisés

Écophysiologie

Collaboration à une thèse de l'Institut Pluridisciplinaire Hubert Curien, Département d'Écologie, Physiologie, Éthologie UMR 7178 CNRS. L'analyse de ces données de physiologie animale, recueillies lors de la thèse, a mis en jeu des techniques

de lissages, de détection de rupture ainsi que des modèles mixtes généralisés non linéaires avec une distribution gamma et un lien log.

Les résultats de ces recherches ont été publiés dans l'article Giroud *et al.* [2008].

Éthologie

J'ai participé à l'analyse de données d'éthologie animale lors d'une collaboration avec l'équipe d'Éthologie des Primates, Département Écologie, Physiologie et Éthologie, IPHC, CNRS, Université Louis Pasteur. La nature des données a requis l'utilisation de modèles mixtes linéaires et généralisés combinés à des tests de permutation.

Les résultats de ces recherches ont été publiés dans l'article Petit *et al.* [2008].

Écologie

Collaboration à une thèse de l'Institut Pluridisciplinaire Hubert Curien, Département d'Écologie, Physiologie, Éthologie UMR 7178 CNRS. L'étude portait sur l'analyse, répétée sur les mêmes individus de plusieurs espèces, de durées de comportements ainsi que la variation du nombre de visites. Il a de ce fait été nécessaire d'utiliser des modèles mixtes généralisés avec une modélisation de la réponse suivant des lois gammas et des lois de Poisson.

Les résultats de ces recherches ont été publiés dans l'article Zimmer *et al.* [2011].

3.3 Risques compétitifs en médecine

Collaboration à des recherches coordonnées sur la maladie du greffon contre l'hôte (GvHD) par le Laboratoire d'ImmunoRhumatologie Moléculaire, INSERM Unité Mixte de Recherche S1109[1]. La maladie du greffon contre l'hôte (GvHD) est la complication majeure de la greffe de moelle osseuse provenant d'un donneur étranger (greffe allogénique). L'objectif était d'évaluer s'il était bénéfique pour le sujet greffé d'apparier certains paramètres génétiques (MICA) entre donneur et receveur.

La quantité de données à traiter et la complexité du problème posé a nécessité le travail d'un chercher post doctorant puis d'un ingénieur statisticien que j'ai tous les deux encadrés sur ce sujet. Les modèles utilisés relèvent de l'analyse de survie et des risques compétitifs. Les résultats de ces recherches ont été publiés dans l'article Carapito *et al.* [2016].

[1]Plateforme GENOMAX, Faculté de Médecine, Fédération Hospitalo-Universitaire OMICARE, Fédération de Médecine Translationnelle de Strasbourg et LabEx TRANSPLANTEX, Faculté de Médecine, Université de Strasbourg

Des recherches similaires, concernant un second paramètre génétique, ont été réalisées et un article a été finalisé.

Chapitre 4

Contributions à la régression pénalisée et à la régression PLS

4.1 Introduction

Ce chapitre concerne l'étude des propriétés de modèles de régression existants ainsi que le développement d'outils permettant d'améliorer leur utilisation. Il concerne principalement la régression des moindres carrés partiels classique Wold *et al.* [1983], Wold *et al.* [1984], Wold *et al.* [2001] ou parcimonieuse Lê Cao *et al.* [2009], Chun et Keleş [2010] mais aussi la problématique plus générale de l'amélioration de la spécificité lors des phases de sélection de variables en régression avec en tête l'application aux techniques de régression pénalisée comme le *lasso* Tibshirani [1996], Tibshirani [2011], la régression *ridge*, Hoerl et Kennard [1970], ou l'*elasticnet* Zou et Hastie [2005] ou à nouveau aux moindres carrés partiels parcimonieux Lê Cao *et al.* [2008], Lê Cao *et al.* [2011].

4.2 Régression sur données qualitatives

Mon premier contact avec la régression des moindres carrés partiel s'est fait pour répondre à un besoin de modélisation dans le domaine des données médicales : l'allélotypage. Un microsatellite est une séquence non-codante de l'ADN. L'allélotypage consiste à rechercher le statut normal ou altéré d'un ensemble prédéfini de microsatellites, en général dans une cellule cancéreuse. Les données d'allélotypage présentent les spécificités suivantes :

- une série de variables explicatives binaires décrivant l'état global des chromosomes de la cellule ;

- une caractéristique à expliquer, elle aussi qualitative binaire, du sujet ou de la tumeur ;
- un nombre de variables pouvant dépasser le nombre de sujets ;
- présence éventuelle de colinéarité entre les variables.

La compréhension des mécanismes de cancérogenèse implique également une description multivariée des données. À l'époque, toutes les publications biomédicales traitant de données d'allélotypage ignoraient la structure réelle des données.

En effet, les analyses étaient faites uniquement sous un angle univarié (voir par exemple Zhu *et al.* [1998]). Les quelques approches multivariées qui avaient été tentées étaient des analyses en cluster comme dans Weber *et al.* [2007], analyses qui ne permettaient pas de répondre à toutes les questions posées par les biologistes. De plus, il est à noter que les méthodes descriptives multivariées n'ont pas pour but de modéliser les données, ce qui limite leur utilisation pour étudier sur le plan biologique les relations possibles entre les voies d'altérations et une caractéristique clinique du patient ou de la tumeur cancéreuse.

Si la caractéristique à expliquer avait été quantitative continue, l'utilisation de la régression PLS Tenenhaus [1998] aurait été naturelle et directe. Dans notre cas d'une réponse binaire, il fallait considérer des variantes PLS des régressions linéaire et logistique comme proposée dans Tenenhaus [1999]. Une spécificité supplémentaire des données d'allélotypage est la présence de variables explicatives toutes qualitatives. Nous avons donc comparé, dans Meyer *et al.* [2010], les performances des variantes PLS des régressions linéaire et logistique sur des données toutes qualitatives.

4.3 Package `plsRglm`

Afin de pouvoir mettre en œuvre les comparaisons réalisées dans l'article Meyer *et al.* [2010], j'avais créé un ensemble de fonctions pour le langage R. Il m'a semblé pertinent de rendre accessible ces fonctions à l'ensemble de la communauté scientifique sous la forme d'un package, nommé `plsRglm`, pour le langage R, Bertrand et Maumy-Bertrand [2018f], `https://cran.r-project.org/web/packages/plsRglm/index.html`.

En effet, l'analyse des jeux de données comportant un grand nombre de variables était, et est toujours, en forte progression dans tous les domaines et particulièrement en médecine et en biologie. Toutefois, ces jeux de données présentent souvent un niveau de complexité supplémentaire due à la présence d'une corrélation linéaire forte entre les variables, ou pire de plus de variables que d'observations.

Il faut alors recourir à des méthodes plus raffinées que le modèle linéaire usuel. L'une de ces méthodes est justement la régression moindres carrés partiels Wold *et al.* [1983], Wold *et al.* [1984], Wold *et al.* [2001].

L'objectif lors du développement du package `plsRglm` était de pouvoir traiter des jeux de données complets, ou présentant des valeurs manquantes Little et Rubin [2002], à l'aide de techniques qui n'existaient pas alors dans R. Il faut souligner que les autres packages de régression PLS n'autorisait pas la présence de valeurs manquantes : soit en les supprimant purement et simplement, soit en se terminant avec une erreur. Il implémente la régression moindres carrées partiels univariée usuelle (PLSR) mais aussi son extension aux modèles de régression linéaires généralisés (PLSGLR) due à Bastien *et al.* [2005] et en particulier aux modèles de régression logistique PLS binaires ou ordinaux.

Pour ces différents modèles, le package propose :

- d'ajuster des modèles de régression PLSR ou PLSGLR Bastien *et al.* [2005] à des jeux de données complets ou incomplets,
- d'utiliser des versions pondérées des modèles PLSR Haaland et Howland [1998] et PLSGLR, une nouveauté,
- de mettre en œuvre, en utilisant différents critères d'évaluation des modèles, des validations croisées *k-fold* pouvant être doublement répétées mais aussi *leave-one-out* sur des jeux de données complets ou incomplets,
- d'appliquer des techniques de bootstrap Lazraq *et al.* [2003] et Bastien *et al.* [2005] pour déterminer des intervalles de confiance pour les prédicteurs d'origine, non seulement dans les cas PLSR mais dans le cas PLSGLR, et ainsi de d'évaluer leur significativité.

J'ai été le premier à intégrer à un package de PLS, la correction pour le calcul des degrés de liberté proposée par Kraemer et Sugiyama [2011] et implémentée dans le package `plsdof`. Suite à l'abandon du package par ses auteurs, je suis m'occupe actuellement de sa maintenance.

Suite à la présentation du package à la conférence UseR ! 2014, Bertrand *et al.* [2014d], le package a été intégré à l'offre de modélisation du package `caret`, Kuhn. [2018]. Pour plus de détails sur les fonctionnalités du package, une prépublication Bertrand et Maumy-Bertrand [2018g] et une vignette détaillée a été rédigée Bertrand *et al.* [2014c].

Le package est toujours activement maintenu et développé.

4.4 Détermination du nombre de composantes

Une des étapes clefs dans l'utilisation de la régression PLS est la détermination correcte du nombre de composantes du modèle. Il s'agissait d'un problème ouvert et important Wiklund *et al.* [2007], Kraemer et Sugiyama [2011]. En effet, compte tenu du manque relatif d'hypothèses faites sur le modèle de régression PLS, qui fait de la PLS une approche de type *soft modelling*, voir Manne [1987], il n'est pas possible de développer des tests statistiques reposant sur des lois de probabilités connues pour tester les paramètres du modèle, Wakeling et Morris [1993]. L'approche généralement retenue est alors d'introduire et de comparer des critères numériques à l'aide de campagnes de simulation. Les critères qui ont été les plus mis en avant, pour cette sélection de modèle, sont basés sur la PRESS, introduite par Allen [1971]. Or pour être évaluée convenablement, cette statistique nécessite l'utilisation d'un jeu de donné test indépendant du jeu de données d'apprentissage. Néanmoins, pour des raisons logistiques, ce jeu de données additionnel n'est que rarement disponible [Efron et Tibshirani, 1993, p. 240].

De ce fait, il est généralement d'usage de recourir à des techniques de validation croisée pour obtenir une estimation de statistiques fonction du PRESS. Or, des difficultés concernant la capacité de la validation croisée à déterminer le pouvoir prédictif des modèles, souvent liées à la grande variabilité des résultats obtenus, ont été mises en avant par [Efron et Tibshirani, 1993, p. 240], [Hastie *et al.*, 2009, p. 249], Wiklund *et al.* [2007], Boulesteix [2014]. Lors de la rédaction de Meyer *et al.* [2010] et de la vignette Bertrand *et al.* [2014c] qui impliquaient l'étude, sur des exemples variés, des propriétés de la PLSR et de la PLSGLR, je me suis rendu compte que le critère du Q^2, pourtant reconnu comme étant le plus performant, posait ces problèmes et d'autres encore : propriétés peu étudiées en présence d'un grand nombre de variables bruitant le signal, comme c'est le cas pour un nombre de plus en plus important de jeux de données biologiques, par exemple génomiques ou protéomiques, ni dans le cas de la PLSGLR, ni en présence de valeurs manquantes. Pour ce dernier point, voir la section 4.5.

J'ai donc proposé un sujet de thèse au LabEx IRMIA sur ce problème de la détermination du nombre de composantes en régression PLSGLR. Le professeur de médecine Nicolas Meyer (PUPH) avait accepté d'être le directeur de cette thèse car je ne pouvais pas l'encadrer seul. Ma proposition de sujet a été retenue et j'ai assuré tout l'encadrement côté mathématique, et l'essentiel de l'encadrement en général, de Jérémy Magnanensi.

Nous avons proposé dans Magnanensi *et al.* [2016a] puis étendu dans Magnanensi *et al.* [2017] un nouveau critère d'arrêt pour déterminer le nombre de composantes en PLSR et en PLSGLR. Il est caractérisé par un grand niveau de stabilité (par rapport au rééchantillonnage) et de robustesse (par rapport au bruit

qui pourrait être présent dans les données). Ce nouveau critère est universel car il est approprié à la fois pour la PLSR et la PLSGLR. Il repose sur l'utilisation de techniques de bootstrap non paramétrique, Efron [1979], et permet de tester l'intérêt de l'ajout de chaque composante supplémentaire au niveau α. La performance et la robustesse de ce critère a été évalué par simulation sur des jeux de données à n individus et p variables dans les cas $n < p$ et $p < n$ avec différents niveaux de bruits, résiduel ou dans les variables explicatives X. La stabilité du critère a, quant à elle, été évaluée en rééchantillonnant un jeu de données réel. Un autre point important est que ce critère donne globalement de meilleures performances que ceux existants dans les cas PLSR et PLSGLR.

Suite à cette étude, nous nous sommes intéressés, dans Magnanensi *et al.* [2016b], aux approches parcimonieuses en régression PLS qui ont récemment attiré beaucoup d'attention dans l'analyse des jeux de données génomiques de grande dimension. En effet, depuis le début des années 2000, des méthodes basées sur la régression par les moindres carrés partiels (PLS) ont été développées pour effectuer une sélection de variables. La plupart de ces techniques reposent aussi sur le choix d'hyperparamètres, souvent déterminés par des méthodes basées sur la validation croisée (CV), ce qui pose à nouveau d'importants problèmes de stabilité.

Pour surmonter cela, nous avons développé une nouvelle méthode dynamique, basée à nouveau sur le bootstrap, pour permettre la sélection des prédicteurs significatifs. Elle est adaptée à la fois à la régression PLS et à son extension aux modèles linéaires généralisés (GPLS). Elle repose sur l'établissement d'intervalles de confiance bootstrap, ce qui permet de tester la signification des prédicteurs à un niveau de risque α fixé à l'avance, et évite l'utilisation de la validation croisée. Nous avons également développé des versions adaptatives de la régression de PLS (SPLS) et de GPLS parcimonieuse (SGPLS) en intégrant le critère d'arrêt que nous avions introduit précédemment dans Magnanensi *et al.* [2017]. Enfin, nous avons comparé la fiabilité et la stabilité de la sélection de variables, celles de la détermination des hyperparamètres du modèle, ainsi que leurs capacités prédictives, en utilisant des données simulées pour la PLS et des données réelles issues des expressions géniques de puces à ADN (*microarrays*) pour la classification logistique PLS.

Par rapport aux autres méthodes évaluées, notre nouvelle méthode dynamique présente la propriété de mieux séparer le bruit aléatoire, présent dans la réponse y, des informations pertinentes, conduisant à une meilleure précision et à une amélioration des capacités prédictives, en particulier en présence de niveaux de bruit non négligeables.

D'un point de vue théorique, une de mes objectifs serait d'obtenir une évaluation des degrés de liberté, dans les modèles issus de ces extensions de la régression

PLS à la régression généralisée, à la manière de Kraemer et Sugiyama [2011], le cas de la régression PLS logistique ou de la régression PLS de Poisson semblant être les plus simples par lequels commencer.

4.5 Influence des valeurs manquantes sur la sélection de composantes en PLS

J'ai proposé à Titin Agustin Nengsih un sujet de thèse, dirigée par le professeur Nicolas Meyer car je ne pouvais pas l'encadrer seul, sur l'influence des valeurs manquantes en régression PLS et la parallélisation à l'aide de GPU. Nous avions commencé à nous intéresser à la manière dont la PLS gérait les valeurs manquantes et s'il était nécessaire, et si oui dans quels cas, d'utiliser des approches plus sophistiquées. L'arrivée de Titin Agustin Nengsih, nous a permis de reprendre ses recherches et de mettre en place une étude par simulation suffisamment conséquente pour aboutir à la rédaction de Nengsih *et al.* [2018].

Les données manquantes, Little et Rubin [2002], sont connues pour être un sujet de préoccupation pour la recherche appliquée, en particulier dans le domaine de la santé ou des études médicales. Plusieurs méthodes ont été développées pour traiter des données incomplètes. La méthode d'imputation est le processus de substitution des données manquantes avant l'estimation des paramètres d'intérêt du modèle.

La régression PLS est un modèle multivarié pour lequel deux algorithmes (SIMPLS ou NIPALS) peuvent être utilisés pour en fournir des estimations des paramètres. La régression PLS a été largement utilisée dans le domaine de la recherche en santé en raison de son efficacité pour analyser les relations entre la réponse et plusieurs composantes.

Toutefois, la gestion des valeurs manquantes lors de l'utilisation de la régression PLS fait toujours l'objet d'un débat. L'algorithme NIPALS a la propriété intéressante de pouvoir fournir des estimations à partir de jeux de données incomplets. La sélection du nombre de composantes pour créer un modèle approprié est une étape clef lors de la régression PLS. Plusieurs approches ont été proposées dans la littérature pour déterminer le nombre de composantes à inclure dans un modèle, tels que le critère Q^2, le critère d'information d'Akaike (AIC) ou le critère d'information bayésien (BIC). L'objectif de notre étude de simulation était d'analyser l'impact de la proportion de données manquantes sous l'hypothèse de données manquantes MCAR et MAR sur l'estimation du nombre de composantes d'une régression PLS.

Nous avons comparé les critères de sélection du nombre de composantes d'une régression PLS sur des données incomplètes avec l'algorithme NIPALS (NIPALS-PLSR) et la régression PLS sur un jeu de données imputé en utilisant trois méthodes d'imputation : l'imputation multiple *Multivariate Imputations by Chained*

Equations (MICE, van Buuren et Groothuis-Oudshoorn [2011]), l'imputation par les k plus proches voisins (KNNimpute, Kowarik et Templ [2016]) et l'imputation basée sur la décomposition en valeurs singulières (SVDimpute, Perry [2015]). Les critères qui ont été comparés sont Q^2-LOO, Q^2-10-*fold*, AIC, AIC-DoF, BIC et BIC-DoF sur différentes proportions de données manquantes et selon l'hypothèse MCAR ou l'hypothèse MAR. La comparaison a été effectuée sur différentes proportions de données manquantes (allant de 5% à 50%).

1. Les données ont été simulées d'après Li *et al.* [2002b]. Le vrai nombre de composantes a été choisi égal à 2, 4 ou 6. Le nombre d'observations n et le nombre de variables p respectent les cinq configurations suivantes :

   - $n = 100$ and $p = 20$,
   - $n = 80$ and $p = 25$,
   - $n = 60$ and $p = 33$,
   - $n = 40$ and $p = 50$,
   - $n = 20$ and $p = 100$.

2. Les données manquantes sont créées sous l'hypothèse d'un mécanisme MCAR ou d'un mécanisme MAR avec un pourcentage de valeurs manquantes allant de 5% à 50% par pas de 5%.

3. Les valeurs manquantes sont imputées en utilisant les méthodes MICE, KNNimpute et SVDimpute.

4. Le nombre de composantes est choisi à l'aide d'une validation croisée *LOO* (Leave One Out) ou 10-*fold* calculée sur les données incomplètes à l'aide des deux méthodes standard et adaptative (qui sélectionne la méthode de prédiction en fonction de la présence de valeurs manquantes dans une ligne du tableau de données, Bertrand et Maumy-Bertrand [2018f]). Pour MICE, le nombre de composantes est le mode des nombres de composantes obtenus par validation croisée pour chacun des m jeux de données imputées où m est égal à 100 x la proportion de valeurs manquantes, White *et al.* [2011].

5. Nous avons aussi fixé à 8 le nombre maximal de composantes pouvant être extraites. Le vrai nombre de composantes est 2, 4 ou 6.

6. Pour chaque combinaison du nombre de vraies composantes, de la proportion de valeurs manquantes, de la configuration ligne-colonne et du mécanisme générateur des valeurs manquantes, 1000 réplicats ont été tirés.

L'étude par simulation a montré que :

- Le Q2-LOO affiche la meilleure performance quelles que soient les méthodes d'imputation. Les performances augmentent lorsque la taille de l'échantillon augmente et diminuent avec une proportion croissante de données manquantes.

- Le nombre de composantes sélectionnées par AIC, AIC-DoF et BIC est presque deux fois plus important que le nombre réel de composants.

- Le nombre réel de composantes d'une régression PLS est difficile à déterminer, en particulier pour un échantillon de petite taille et lorsque la proportion de données manquantes est supérieure à 30%.

- L'exécution de MICE a pris beaucoup de temps. Par exemple, lorsque $n = 100$ et que la proportion de données manquantes = 10%, la durée d'exécution de MICE était environ 11 fois plus lente que celle de NIPALS-PLSR.

Pour plus de détails, voir Nengsih *et al.* [2018].

4.6 Ajustement stochastique multidimensionnel

Les tableaux de données multidimensionnels apparaissent naturellement dans de nombreuses disciplines scientifiques, par exemple en biologie avec l'étude des expressions des gènes, Cheung [2012], Golub *et al.* [1999], en géographie avec l'analyse des données spatiales des séismes, van der Hilst *et al.* [2007], dans l'étude des marchés financiers et en particulier la constitution et l'évaluation de portefeuilles, Jagannathan et Ma [2003], ainsi que dans de nombreux autres domaines. La complexité des jeux de données réels est la plupart du temps telle qu'il n'est pas possible de réduire l'étude des données à celle d'une seule variable. Ainsi une analyse multivariée globale des données, Mardia *et al.* [1979], est généralement nécessaire, voir les chapitres 3, 4, 5, 6 pour des exemples d'application de certaines de ces techniques.

La méthode d'ajustement multidimensionnel (*Multidimensional fitting*, *MDF*) est une méthode d'analyse de données multivariée récemment mise au point et basée sur l'ajustement des distances. Imaginons que nous disposions de deux matrices observées X et D : la première, X, contient les coordonnées des individus et la seconde, D, des distances entre ces individus. À partir de la matrice X, il est également possible de calculer des distances, D_X, entre les individus. L'originalité de l'ajustement multidimensionnel est de proposer des vecteurs de modification des coordonnées des individus, donc la matrice X, afin de faire se rapprocher les distances, D_X, calculées sur ces coordonnées modifiées des distances initialement présentes dans la matrice D.

Cette méthode a été introduite dans Berge *et al.* [2010] et Alawieh *et al.* [2017]. Or dans ces articles, les auteurs ne considèrent qu'un modèle déterministe pour les vecteurs de déplacement alors que des effets aléatoires pourraient se produire durant le processus de modification et avoir un effet sur le résultat de celle-ci.

Nous avons introduit un modèle stochastique pour l'ajustement multidimensionnel afin de trouver des vecteurs de déplacement optimaux dans un contexte bruité. La fonction objectif n'étant plus déterministe, il existe de multiples critères permettant d'évaluer la pertinence des vecteurs de modification du MDF et nous avons proposé deux approches permettant de trouver une solution dans le cas de l'écart quadratique moyen.

La première approche conduit à un problème d'optimisation déterministe et repose sur un cadre gaussien. Le cas indépendant et identiquement distribué est présenté dans Alawieh *et al.* [2018]. Même si elle n'est pas présentée dans Alawieh *et al.* [2018] pour des raisons de concision, nous avons aussi envisagé une approche gaussienne tenant compte de corrélations éventuelles entre les variables qui décrivent les individus dans la matrice X et qui repose sur des développements mathématiques plus raffinés impliquant l'utilisation de lois de Wishart non centrales, Anderson [1946], et le calcul de certains de leurs moments, Letac et Massam [2004]. La seconde approche est une optimisation stochastique qui repose sur l'utilisation d'un algorithme de Metropolis-Hastings, Metropolis *et al.* [1953]. Elle convient à des contextes non gaussiens ou de corrélation entre les variables explicatives mais est plus longue à mettre en œuvre.

L'article Alawieh *et al.* [2018] contient cette extension aléatoire du *Multidimensional fitting*, ainsi qu'une application dans le domaine de la sensométrie.

4.7 Sélection de variables avec confiance

Avec l'émergence de technologies à haut débit, il est possible de mesurer de grandes quantités de données à un coût relativement faible. De telles situations se posent dans de nombreux domaines, des sciences aux sciences humaines, et la sélection de variables peut être très utile pour répondre aux défis spécifiques à chacune d'elles. La sélection des variables peut permettre de connaître, parmi toutes les variables mesurées, celles qui présentent un intérêt ou non.

Une des raisons initiales qui a motivé ce travail est le problème exposé au chapitre 6 dans la section sur le décodage (*reverse engineering*) des réseaux biologiques et plus particulièrement la détermination de cibles pour une intervention dirigée dans les programmes géniques des cellules cancéreuses. Une intervention dirigée dans un réseau de gènes en cascade consiste à agir sur certains gènes en amont de la cascade dans le but d'obtenir l'effet souhaité sur d'autres gènes situés plus en aval. Notre problème était donc non seulement de trouver ces gènes sur lesquels

agir mais aussi de choisir parmi eux, ceux pour en lesquels nous avions la plus grande confiance pour obtenir la modification souhaitée.

De nombreuses méthodes ont été proposées pour traiter ce problème, le *lasso* et d'autres régressions pénalisées en étant des cas particuliers. Ces méthodes échouent dans certains cas et la corrélation linéaire entre les variables explicatives est la plus courante des situations pour laquelle cela se produit. Or, celle-ci apparaît naturellement dans les grands ensembles de données. Dans Jung *et al.* [2015], nous avons présenté un algorithme capable d'améliorer la précision de toute méthode de sélection de variables.

J'ai alors proposé une extension de cet algorithme au cas des modèles linéaires généralisés. J'ai également repris et repensé intégralement le code qui avait été développé initialement afin d'améliorer ses performances et de proposer un package R le contenant, Bertrand *et al.* [2018c]. Avec l'aide d'Ismaïl Aouadi, ingénieur d'étude, nous avons appliqué ce nouvel algorithme, appelé selectboost, à différents jeux de données dont des problèmes de régression logistique binaire pénalisée (typiquement du *lasso*). Puis, nous avons récemment écrit une révision majeure de la première prébublication, Aouadi *et al.* [2018].

Chapitre 5

Extensions de la régression PLS

Les recherches de ce chapitre sont des extensions originales de la régression PLS à deux nouveaux contextes : les réponses bornées, et dont le support est connu avant que l'expérience ne soit mise en œuvre, ainsi que les données de survie. Comme pour les contributions proposées au chapitre 4, je me suis intéressé au problème du choix du nombre de composantes ainsi qu'à celui de la sélection des variables.

5.1 Régression Bêta

5.1.1 Motivation

De nombreuses variables d'intérêt, comme par exemple des résultats expérimentaux, des rendements ou des indicateurs économiques, s'expriment naturellement sous la forme de taux, de proportions ou d'indices dont les valeurs sont nécessairement comprises entre zéro et un ou plus généralement deux valeurs fixes connues à l'avance. La régression Bêta permet de modéliser ces données avec beaucoup de souplesse puisque les fonctions de densité des lois Bêta peuvent prendre des formes très variées. En effet, l'intérêt pratique de la loi Bêta a été plusieurs fois affirmé par exemple par Johnson *et al.* [1995] : "Beta distributions are very versatile and a variety of uncertainties can be usefully modelled by them. This flexibility encourages its empirical use in a wide range of applications." Plusieurs articles récents se sont intéressés à l'étude de la régression Bêta et de ses propriétés. L'article de Ferrari et Cribari-Neto [2004] mérite d'être mentionné comme introduction à ces modèles et ceux de Kosmidis et Firth [2010], Simas *et al.* [2010] et Grün *et al.* [2012] pour des extensions ou des améliorations des techniques d'estimation de ces modèles.

Toutefois, comme tous les modèles de régression usuels, la régression Bêta ne peut s'appliquer directement lorsque les prédicteurs présentent des problèmes de multicolinéarité ou pire lorsqu'ils sont plus nombreux que les observations. Ces situations se rencontrent fréquemment de la chimie à la médecine en passant par l'économie ou le marketing. Pour circonvenir cette difficulté, nous avons formulé une extension de la régression PLS pour les modèles de régression Bêta. Celle-ci, ainsi que plusieurs outils, comme la validation croisée et des techniques bootstrap, est disponible pour le langage R dans le *package* `plsRbeta`. Ce *package* utilise la régression Bêta implémentée dans le *package* `betareg` (Cribari-Neto et Zeileis [2010]) pour le langage R.

La régression PLS, fruit de l'algorithme NIPALS initialement développée par Wold [1966] et exposée en détails par Tenenhaus [1998], avait été étendue avec succès aux modèles linéaires généralisés par Bastien *et al.* [2005] et aux modèles de Cox par Bastien [2008].

5.1.2 Bootstrap

Nous supposons avoir retenu le nombre m adéquat de composantes d'un modèle de régression Bêta PLS de Y sur $x_1, \ldots, x_j, \ldots, x_p$. Nous proposons l'algorithme suivant pour construire des intervalles de confiance et des tests de significativité pour les prédicteurs x_j, $1 \leqslant j \leqslant p$, à l'aide de techniques bootstrap.
Soit $\widehat{F}_{(T|Y)}$ la fonction de répartition empirique étant données la matrice T formée des m composantes PLS et la réponse Y.

Étape 1. Tirer B échantillons de $\widehat{F}_{(T|Y)}$.

Étape 2. Pour tout $b = 1, \ldots, B$, calculer :

$$c^{(b)} = (T^{(b)\prime}T^{(b)})^{-1}T^{(b)\prime}Y^{(b)} \quad \text{et} \quad b^{(b)} = W^* c^{\prime(b)},$$

où $[T^{(b)}, Y^{(b)}]$ est le b^{e} échantillon bootstrap, $c^{\prime(b)}$ est le vecteur des coefficients des composantes et $b^{(b)}$ est le vecteur des coefficients des p prédicteurs d'origine pour cet échantillon et enfin W^* est la matrice fixe des poids des prédicteurs dans le modèle d'origine comportant m composantes.

Étape 3. Pour chaque j, notons Φ_{b_j} l'approximation de Monte-Carlo de la fonction de répartition de la statistique bootstrap de b_j.
Pour chaque b_j, des boîtes à moustaches et des intervalles de confiance peuvent être construits à l'aide des percentiles de Φ_{b_j}. Un intervalle de confiance peut être défini par $I_j(\alpha) =]\Phi_{b_j}^{-1}(\alpha), \Phi_{b_j}^{-1}(1-\alpha)[$ où $\Phi_{b_j}^{-1}(\alpha)$ et $\Phi_{b_j}^{-1}(1-\alpha)$ sont les valeurs obtenues à partir de la fonction de répartition de

la statistique bootstrap de telle sorte qu'un niveau nominal de confiance de niveau $100(1-2\alpha)\%$ soit atteint. Afin d'améliorer la qualité de l'intervalle de confiance en termes de taux de couverture, c'est-à-dire la capacité de $I_j(\alpha)$ à fournir les taux de couverture attendus, il est possible d'utiliser plusieurs techniques de construction : normale, percentile ou BC_a (Efron et Tibshirani [1993] ou Davison et Hinkley [1997]). Les intervalles ainsi obtenus ne sont pas conçus pour servir à réaliser des comparaisons multiples ou deux à deux et doivent être interprétés séparément.

5.1.3 Choix du nombre de composantes

Un problème crucial pour une utilisation correcte de la régression PLS est la détermination du nombre de composantes.

Le nombre de composantes PLS $\boldsymbol{t}_h$ peut être déterminé en régression PLS classique par validation croisée. Une composante $\boldsymbol{t}_h$ est ajoutée si le *PRESS* (*PRedicted Error Sum of Squares*) de l'étape h est nettement plus petit que le *RESS* (*Residual Sum of Squares*) de l'étape $h-1$. Wold propose dans le logiciel SIMCA (Eriksson *et al.* [2006]) d'introduire la h-ième composante si l'indice de Stone-Geisser

$$Q^2 = 1 - \frac{PRESS_h}{RESS_{h-1}}$$

est au moins égal à 0,0975. Le même type d'approche a été introduit, dans Tenenhaus [1999] et Tenenhaus [2005] pour les extensions de la régression PLS à la régression logistique binaire et à la régression logistique ordinale. Dans le cas de la régression logistique binaire, elle repose sur l'utilisation du χ^2 de Pearson défini par

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \pi_i)^2}{\pi_i(1-\pi_i)}$$

où Y_i est la valeur de la variable Y pour l'individu i et π_i la probabilité de l'événement $\{Y=1\}$ pour un individu ayant les caractéristiques de l'individu i. Le χ^2 de l'étape h peut être calculé par substitution en remplaçant π_i par son estimation à l'aide de la régression logistique sur les composantes $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_h$. Il peut aussi être calculé par validation croisée en estimant π_i sans utiliser l'observation i, ou plus généralement en estimant $\pi_{i_1}, \ldots, \pi_{i_k}$ sans utiliser les observations $i_1, \ldots, i_k$. On considère que la composante $\boldsymbol{t}_h$ est significative si le χ^2 calculé à l'étape h par validation croisée est nettement inférieur au χ^2 calculé à l'étape $h-1$ par substitution. En reprenant l'approche de Wold, on décide que la composante $\boldsymbol{t}_h$ est significative si l'indice

$$Q^2 = 1 - \frac{\chi^2_{\text{validation croisée, étape } h}}{\chi^2_{\text{substitution, étape } h-1}}$$

est au moins égal à 0,0975.

Plus généralement, une extension au cas des modèles linéaires généralisés à réponse univariée est possible en considérant que les densités f_i des réponses Y_i font parties d'une famille exponentielle uni-dimensionnelle :

$$f(y_i,\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi/A_i} + c(y_i,\phi/A_i)\right)$$

où θ_i, A_i et ϕ sont des paramètres et les fonctions b et c sont connues. La valeur du χ^2 s'obtient alors à l'aide de la formule

$$\chi^2 = \phi \sum_{i=1}^{n} \frac{(Y_i - \mathbb{E}(Y_i))}{V(Y_i)}.$$

où Y_i est la variable aléatoire réponse Y pour l'individu i, $\mathbb{E}(Y_i)$ l'espérance de Y_i, $V(Y_i)$ sa variance et ϕ un paramètre de dispersion. Dans le contexte de la famille exponentielle uni-dimensionnelle, une expression plus simple du χ^2 ne met en jeu que les des dérivées d'ordre 1 et 2 de la fonction b et les paramètres θ_i et A_i.

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - b'(\theta_i))^2}{b''(\theta_i)/A_i}.$$

Le χ^2 de l'étape h peut être calculé par substitution en remplaçant θ_i par son estimation à l'aide de la régression généralisée sur les composantes $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_h$. Il peut aussi être calculé par validation croisée en estimant π_i sans utiliser l'observation i, ou plus généralement en estimant $\pi_{i_1}, \ldots, \pi_{i_k}$ sans utiliser les observations $i_1, \ldots, i_k$. On considère que la composante $\boldsymbol{t}_h$ est significative si le χ^2 calculé à l'étape h par validation croisée est nettement inférieur au χ^2 calculé à l'étape $h-1$ par substitution. En reprenant l'approche de Wold, on décide que la composante $\boldsymbol{t}_h$ est significative si l'indice

$$Q^2 = 1 - \frac{\chi^2_{\text{validation croisée, étape } h}}{\chi^2_{\text{substitution, étape } h-1}}$$

est au moins égal à 0,0975.

Si, dans le cas de la régression PLS originale, le critère du Q^2 est extrêmement efficace Tenenhaus [1998], ses bonnes propriétés disparaissent malheureusement pour les modèles de régression linéaire généralisée PLS. Une étude par simulation s'impose donc afin de déterminer un critère fonctionnel pour choisir le nombre de composantes. Nous avons comparé les critères suivants AIC et BIC, Cribari-Neto et Zeileis [2010], χ^2 de Pearson, R^2 de Pearson et pseudo-R^2, Ferrari et Cribari-Neto [2004], ou critères $Q^2\chi^2$ et $Q^2\chi^2$ cumulé estimés par validation croisée en 5 groupes (5-CV) ou en 10 groupes (10-CV), Bastien *et al.* [2005].

L'algorithme utilisé pour créer les données simulées est une adaptation directe de l'algorithme de Li *et al.*, Li *et al.* [2002a] qui est lui-même une généralisation multivariée de celui de Naes et Martens, Naes et Martens [1985]. Ce type de généralisation a déjà été utilisé avec succès dans le cas des modèles de régression logistique PLS, Meyer *et al.* [2010].

De manière générale, les résultats de l'étude par simulations montrent que le $Q^2\chi^2$ (5-CV et 10-CV), déjà connu pour son comportement surprenant en régression logistique PLS (Bastien *et al.* [2005], Meyer *et al.* [2010]), ne se comporte guère mieux pour les modèles de régression Bêta PLS. La maximisation des critères du R^2 ou du pseudo-R^2, s'avère également inefficace. Les critères AIC et BIC retiennent systématiquement quelques composantes de trop. Cette tendance est également connue dans le cas de la Régression PLS traditionnelle (Kraemer et Sugiyama [2011]) comme dans celui de la Régression Logistique PLS (Meyer *et al.* [2010]).

5.1.4 Exemples d'application

Médecine

Les tumeurs cancéreuses représentent l'une des trois principales causes de mortalité dans le monde occidental. La compréhension des mécanismes des pathologies cancéreuses repose actuellement sur l'étude des relations mutuelles des anomalies génétiques acquises, apparaissant dans les tissus au cours du processus de la cancérisation. Ces anomalies sont fréquemment analysées par allélotypages, permettant de déterminer pour un nombre plus ou moins important de sites géniques, la présence ou non d'une modification du nombre de copies de chaque gène. La description multivariée de ces anomalies est informative sur le processus de cancérogénèse. Par ailleurs, l'ensemble de ces sites géniques porteur ou non d'anomalie peut être utilisé pour tenter de prédire certaines caractéristiques cliniques ou biologiques de la tumeur telles que le taux de cellules tumorales sur la biopsie d'une lésion. La modélisation dans un modèle statistique de taux, variable dont l'espace de variation est contenu dans l'intervalle fermé $[0; 1]$ comme variable prédite suggère l'utilisation d'une régression Bêta. Par ailleurs, les données d'allélotypage sont caractérisées par une fréquente colinéarité et par une proportion importante de données manquantes. De plus la matrice des données a souvent des dimensions $(i; j)$ telles que $j > i$, ce qui rend la matrice non-inversible, posant des difficultés dans l'ajustement d'un modèle de régression. La régression Bêta de type PLS que nous avons développée est donc particulièrement adaptée pour traiter les données d'allélotypage dans le contexte particulier de la prédiction d'une variable de type taux.

L'exemple, présenté dans Bertrand *et al.* [2013b], est celui de données d'allélotypage obtenues sur une série de 93 patients atteints de différents types de cancer du poumon et comportant 23,2% de valeurs manquantes. La variable prédite est le taux de cellularité tumorale du prélèvement peropératoire de la tumeur. Les variables explicatives sont composées de 56 variables binaires indicatrices de la présence d'une anomalie sur chacun des 56 microsatellites et de trois variables cliniques.

La sélection de variables est, dans cet exemple, très importante car elle permet de définir un sous-ensemble de prédicteurs, c'est-à-dire de sites géniques, capable de prédire le taux de cellules tumorales. En effet, les pathologies cancéreuses sont des pathologies génétiques acquises et certaines de ces anomalies sont la cause et d'autre la conséquence de la pathologie tumorale. Par ailleurs, l'information contenue dans les différents sites géniques microsatellites est potentiellement redondante. La sélection de variable, séparant les variables jouant probablement un rôle moteur dans le développement tumoral des variables ne faisant que traduire un bruit de fond aléatoire induit par des anomalies causées par ce développement tumoral, est alors une aide indispensable à la compréhension des mécanismes sous-jacents de la tumorogenèse.

Chimiométrie

L'objectif est de trouver des composés permettant de prédire le taux d'infiltration de patients en cellules cancéreuses à partir de données de spectrométrie. Un patient sain a 0% de cellules cancéreuses dans une biopsie tandis qu'un patient malade aura dans une biopsie un pourcentage d'autant plus élevé que l'échantillon contient des cellules cancéreuses. L'intérêt de cette expérience est d'essayer de réduire considérablement le temps d'analyse des biopsies en évitant d'avoir recours à un comptage par un médecin spécialiste en anatomie et cytologie pathologique. Une difficulté statistique supplémentaire apparaît dans l'exemple présenté dans Bertrand *et al.* [2013b] : il y a plus de variables (180) que d'individus (80). Plus de détails sur le protocole expérimental suivi sont disponibles dans l'article dans lequel ce jeu de donnés a déjà été initialement publié Piotto *et al.* [2012].

5.1.5 Bilan

Mon objectif a été de proposer une extension de la régression PLS aux modèles de régression Bêta, puis de la mettre à la disposition des utilisateurs du langage libre R. Nous offrons ainsi la possibilité de travailler, pour modéliser des taux ou des proportions, avec des prédicteurs colinéaires, difficulté inévitable dans le cas de la modélisation des mélanges ou lors de l'analyse de spectres, de l'étude de données génétiques, protéomiques ou métabolomiques.

De plus, la régression Bêta PLS peut être aussi appliquée à des jeux de données incomplets. Il est également possible dans ce cas, comme dans celui des données complètes, de sélectionner le nombre de composantes par validation croisée *repeated k-fold cross-validation*. Enfin, nous proposons des techniques bootstrap afin de, par exemple, tester la significativité de chacun des prédicteurs présents dans le jeu de données et ainsi valider les modèles construits. L'étude de deux jeux de données réels a permis aux outils proposés de démontrer leur efficacité.

Plusieurs extensions ont été envisagées : approches robuste, parcimonieuse, inflation de zéro, réponses multivariées (Bry *et al.* [2013]), sélection du nombre de composantes à l'aide du critère d'arrêt introduit dans Magnanensi *et al.* [2017]. Il faudrait également réaliser une étude de l'influence de la proportion et du type de valeurs manquantes similaire à celle de Nengsih *et al.* [2018].

5.2 Données de survie

5.2.1 Motivation et premiers résultats

Le point de départ de ces travaux est la volonté d'analyser un jeu de données original sur la prévision de la survie de patients atteints de cancer du côlon à l'aide de données d'allélotypage. L'allélotypage consiste à rechercher le statut normal ou altéré d'un ensemble prédéfini de microsatellites, séquence non-codante de l'ADN, dans une cellule cancéreuse. La survie devait être étudiée en fonction de 33 microsatellites simultanément, voir section 4.2, et du stade de cancer. Un des intérêts spécifique de la régression PLS, souvent apprécié dans l'étude des jeux de données génomiques ou protéomiques, est la détermination des composantes. Elle peuvent représenter des associations, ici entre microsatellites, qui seront interprétables par le biologiste ou le médecin typiquement comme l'altération d'une fonction.

J'ai commencé par implémenter différents modèles existants à ce jour dont les plus récents venaient d'être présentés dans l'article de Bastien [2008].

- COX-PLS, un modèle de Cox sur des composantes créées à partir de la régression PLS de la durée de survie par rapport aux variables explicatives ;
- LASSO-LARS DR, un modèle de Cox par rapport aux variables explicatives sélectionnées lors de l'application du LASSO aux résidus de la déviance calculés pour un modèle de Cox sans variable explicative.
- PLSDR, un modèle de Cox par rapport aux composantes PLS sélectionnées lors de l'application d'une régression PLS aux résidus de la déviance calculés pour un modèle de Cox sans variable explicative.

- DKPLSDR, un modèle de Cox par rapport aux composantes PLS sélectionnées lors de l'application d'une régression Kernel PLS aux résidus de la déviance calculés pour un modèle de Cox sans variable explicative.

Je les ai comparés entre eux et à un modèle introduit entre temps et qui partageait la même finalité à savoir proposer des modèles de Cox en présence d'un très grand nombre variables explicatives. En effet, l'article Kim *et al.* [2008] a permis à Sohn *et al.* [2009] de développer un modèle qui s'est montré particulièrement efficace : le « L^1 *penalized Cox PH model using the generalized lasso algorithm* ». La qualité prédictive des modèles avait été comparée à l'aide de courbes ROC adaptées aux données censurées introduites par Heagerty *et al.* [2000b] et Heagerty et Zheng [2005a].

Les deux jeux de données réels, qui ont servi à la comparaison, sont : le jeu de données d'allélotypage mentionné ci-dessus et un jeu de données comportant des observations recueillies à l'aide de puces à ADN. Ce dernier avait été introduit dans Alizadeh *et al.* [2000] et étudié par de nombreux autres auteurs, par exemple dans Rosenwald *et al.* [2002], Bastien [2008] et Sohn *et al.* [2009]. Il a été obtenu en utilisant des « *lymphochips* » : des puces conçues pour les pathologies lymphoïdes et qui comprennent environ 18 000 clones d'ADN complémentaires. Il sert généralement à comparer les capacités prédictives de modèles de survie dans le cas d'un patient atteint par un lymphome diffus à grandes cellules B. Ce jeu de données comporte un total de 240 patients atteints de LMNH-B dont 138 patients sont décédés pendant le suivi, durée médiane avant décès de 2,8 ans, et 30% de temps de survie censurés à droite. Les « *lymphochips* » contiennent 7399 zones qui représentent 4128 gènes.

5.2.2 Approches parcimonieuses

Le second jeu de données a fait très nettement ressortir l'intérêt d'intégrer une étape de sélection de variable pour faciliter l'interprétation des résultats par la communauté médicale. C'est de ce point de départ qu'est partie l'idée de ma collaboration avec Philippe Bastien pour le développement de méthodes rapides de sélections de variable pour le modèle de Cox. En effet, estimer des modèles de régression PLS ou sPLS, éventuellement à noyaux, sur les résidus d'un modèle de Cox permet de combiner un temps d'exécution rapide, pour pouvoir traiter des jeux de données de grande dimension comportant des centaines ou des milliers de variables, avec toutefois une prise en compte de la nature censurée des données.

L'importance de ce type de jeux de données transparaît au travers d'une vaste littérature depuis les années 2000 qui est consacrée aux relations entre les profils géniques et le temps, pour un sujet, de survie ou de rechute de son cancer. La

découverte de biomarqueurs à partir de données de grande dimension, telles que les profils transcriptomiques ou SNP, constitue un défi majeur dans la recherche de diagnostics plus précis. Le modèle de régression à risque proportionnel suggéré par Cox, 1972, pour étudier la relation entre le temps avant événement et un ensemble de covariables en présence de censure est le modèle le plus couramment utilisé pour l'analyse des données de survie.

Cependant, comme pour la régression multivariée, cela suppose qu'il existe plus d'observations que de variables, des données complètes et des variables non fortement corrélées. En pratique, lorsqu'il s'agit de données de grande dimension, ces contraintes ne sont pas vérifiées. La colinéarité engendre des problèmes de sur-ajustement et de mauvaise identification des modèles. La sélection de variables peut améliorer la précision de l'estimation en identifiant efficacement le sous-ensemble de prédicteurs pertinents et en améliorant l'interprétabilité du modèle avec une représentation parcimonieuse.

Depuis l'article de référence de Tibshirani [1997], de nombreuses méthodes basées sur les modèles à risques proportionnels de Cox pénalisés par *lasso* ont été proposées. La régularisation pourrait également être effectuée à l'aide d'une réduction de dimension, comme c'est le cas pour la régression moindres carrés partiels (PLS). Nous avons proposé deux algorithmes originaux nommés sPLSDR et son pendant non linéaire à noyau, DKsPLSDR, voir Rosipal et Trejo [2002], Tenenhaus *et al.* [2007] pour les approches à noyau dans d'autres contextes, en utilisant une régression PLS parcimonieuse (sPLS) basée sur les résidus de déviance. Nous avons comparé leurs performances en termes de prévision avec les meilleurs algorithmes disponibles à l'époque à l'aide d'une vaste campagne de simulation intégrant des jeux de données simulés et des jeux de données réel de référence.

Le résultat de cette campagne de simulation est que les deux nouvelles méthodes proposées sPLSDR et DKsPLSDR ont obtenu des résultats comparables voire meilleurs que les autres méthodes en termes de temps de calcul, de pouvoir prédictif et de sensibilité. De plus, de par leur nature même de régression PLS, elles offrent des possibilités additionnelles intéressantes comme les représentations en *biplot* ou la capacité de s'accommoder naturellement de valeurs manquantes.

Ces nouvelles méthodes, que nous avons considérées comme un ensemble d'outil pertinent pour les utilisateurs du très répandu modèle de Cox, ont été publiées dans Bastien *et al.* [2015].

Outre le fait de rendre accessible non seulement tous ces nouveaux développements méthodologiques aux utilisateurs de R mais aussi d'autres approches plus anciennes comme larsDR, coxPLS, PLScox, les points forts de l'implémentation du package `plsRcox`, Bertrand et Maumy-Bertrand [2018e], tiennent au fait que l'utilisateur a la possibilité de faire de la validation croisée, d'utiliser des techniques

bootstrap et de travailler avec des jeux de données présentant des valeurs manquantes. Ce package a été présenté à la conférence internationale des utilisateurs du langage R, User ! 2014, Bertrand *et al.* [2014b].

5.3 Critères de choix de modèles pour les extensions de la régression moindres carrés partiels au modèle de Cox

L'étude par simulation réalisée dans Bastien *et al.* [2015], m'a amené à m'interroger sur la validité des critères usuels de choix de modèle dans un contexte de régression PLS étendue au modèle de Cox.

5.3.1 L'échec des deux critères classiques

Pour une validation croisée en k échantillons traditionnelle, c'est-à-dire en l'absence d'évènements censurés, chaque échantillon sera à tour de rôle un ensemble de test et permettra de calculer la valeur d'une mesure de qualité d'ajustement du modèle (par exemple log vraisemblance partielle, aire intégrée sur la courbe ROC, erreur de prédiction).

En présence d'évènements censurés et pour le critère de la log vraisemblance partielle cross-validée (CVLL, Verweij et Van Houwelingen [1993]), il est possible d'utiliser plus efficacement les ensembles d'individus à risque de mourir : van Houwelingen *et al.* [2006] conseille de calculer la log vraisemblance partielle cross-validée du j^{e} échantillon par soustraction (en soustrayant la log vraisemblance partielle évaluée sur le jeu de données en entier à celle évaluée sur le jeu de données en entier privé du j^{e}). Il s'agit de la log vraisemblance partielle cross-validée de van Houwelingen (vHCVLL). C'est pourquoi, lors de l'application de la validation croisée à des modèles de Cox ou à ses extensions, le critère le plus souvent utilisé est celui de la log vraisemblance partielle cross-validée suivant un schéma classique ou de van Houwelingen.

De manière surprenante, j'ai observé, sur la base d'une vaste campagne de simulations impliquant trois schémas de simulations différents, que les deux critères CVLL et vHCVLL sous-estiment systématiquement, et échouent donc à sélectionner, le bon nombre de composantes avec les extensions plus ou moins complexes de la régression moindres carrés partiels au modèle de Cox (7 extensions étaient considérées : PLS-Cox, autoPLS-Cox, Cox-PLS, PLSDR, sPLSDR, DKPLSDR and DKsPLSDR). Ce comportement pathologique a été validé par une autre simulation suivant un plan plus simple conçu dans Simon *et al.* [2011].

C'est un résultat assez intéressant pour au moins deux raisons.

Premièrement, plusieurs fonctionnalités intéressantes des modèles basés sur la PLS, telles que la régularisation, l'interprétabilité des composantes, le support des données manquantes, la visualisation des données grâce aux *biplots* des individus et des variables, et même la parcimonie pour les modèles basés sur la sPLS, expliquent l'utilisation fréquente de ces extensions par les statisticiens qui sélectionnent généralement les hyperparamètres de ces modèles en utilisant la validation croisée.

Deuxièmement, ils font presque toujours partie des études comparatives pour évaluer les performances d'une nouvelle technique d'estimation utilisée dans un contexte de grande dimension et présentent souvent de piètres propriétés statistiques.

5.3.2 Mise en lumière de critères pertinents

Quels autres critères ?

Suite à ce constat d'échec, j'ai recherché d'autres critères pour déterminer le nombre de composantes dans ces modèles.

Li [2006] a utilisé *the integrated area under the curves of time-dependent ROC curves* (iAUCsurvROC, Heagerty *et al.* [2000a]) pour mettre en œuvre ses validations croisées, approche implémentée dans le package `survcomp`, Schröder *et al.* [2011].

Mis à part ce critère, (Figure 5.8) j'ai ajouté cinq autres mesures d'AUC intégrée : estimateur de Chambless et Diao [2006] intégré (iAUCCD, Figure 5.3), estimateur de Hung et Chiang [2010] intégré (iAUCHC, Figure 5.4), estimateur de Song et Zhou [2008] intégré (iAUCSH, Figure 5.5), estimateur de Uno *et al.* [2007] intégré (iAUCUno, Figure 5.6) et estimateur de Heagerty et Zheng [2005b]) intégré (iAUCHZ, Figure 5.7) de l'AUC cumulé/dynamique pour données censurées à droite, implémentées dans le package `survAUC` package, Potapov *et al.* [2012], et le package `risksetROC`, Heagerty et packaging by Paramita Saha-Chaudhuri [2012].

J'ai aussi étudié deux versions de deux critères d'erreur de prédiction, le score de Brier intégré pondéré, ou non, (Graf *et al.* [1999], Gerds et Schumacher [2006], iBS(un)w, l'écart quadratique intégré pondéré, ou non, entre les survies prédites et observées, voir équation 5.3 pour une définition mathématique, résultats voir Figures 5.9 et 5.11) et le score de Schmid intégré pondéré, ou non, (Schmid *et al.* [2011], iSS(un)w, l'écart absolu intégré pondéré, ou non, entre les survies prédites et observées, voir équation 5.6 pour une définition mathématique, résultats voir Figures 5.10 et 5.12), également implémentés dans le package `survAUC`, Potapov *et al.* [2012].

En résumé, voici la liste des douze critères étudiés, voir Table 5.1, pour la validation croisée classés suivant leur nature.

- Vraisemblance (2) : Verweij et Van Houwelingen [1993] (classic CVLL), van Houwelingen *et al.* [2006] (vHCVLL).
- Mesures basées sur AUC intégré (6) : Chambless et Diao [2006] (iAUCCD), Hung et Chiang [2010] (iAUCHC), Song et Zhou [2008] (iAUCSH), Uno *et al.* [2007] (iAUCUno), Heagerty et Zheng [2005b] (iAUCHZ), Heagerty *et al.* [2000b] (iAUCsurvROC).
- Évaluation de l'erreur de prédiction (4) : score de Brier intégré pondéré ou non (iBS(un)w, Gerds et Schumacher [2006]) ou score de Schmid intégré pondéré ou non (iSS(un)w, Schmid *et al.* [2011]).

J'ai alors également évalué ces dix autres critères potentiels de validation croisée, soit basés sur l'AUC, soit sur l'erreur de prédiction.

Score de Brier

Le score de Brier est une fonction du temps t. En l'absence d'observations censurées dans le jeu de données, pour un temps donné t et pour une observation i, nous souhaiterions que ce score possède les propriétés suivantes :

1. Si un évènement est survenu pour l'observation i avant le temps t, alors nous prédisons une probabilité de survie proche de 0 pour i.
2. Si un évènement surviendra pour l'observation i après le temps t, alors nous prédisons une probabilité de survie proche de 1 pour i.

Dans ce contexte, pour un temps t donné, le score de Brier est égal au carré de l'écart entre l'estimation de la fonction de survie et la vraie valeur de celle-ci. Plus précisément, en l'absence d'observations censurées, le score de Brier au temps t, $BS_{\text{pas de censure}}(t)$, est défini par :

$$BS_{\text{pas de censure}}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[(0-\hat{S}(t \mid \mathbf{X}_i))^2 I(t_i \leqslant t) + (1-\hat{S}(t \mid \mathbf{X}_i))^2 I(t_i > t)\right] \tag{5.1}$$

où I désigne une fonction indicatrice.

Considérons désormais le cas où de la censure est observée dans le jeu de données. Appelons $G(t)$ la probabilité d'absence de censure $\mathcal{C}$ jusqu'à l'instant t : $G(t) = \mathbb{P}(\mathcal{C} > t)$ et $\hat{G}(t)$ son estimateur de Kaplan-Meier à partir des observations

$(t_i,1-\delta_i)$, où $\delta_i = 1$ si l'évènement a eu lieu et $\delta_i = 0$ si t_i est un des temps censurés.

Le score de Brier pondéré $BSw(t)$ (Brier [1950]; Radespiel-Tröger *et al.* [2003]; Hothorn *et al.* [2004]) est encore une mesure d'erreur de prédiction basée sur le carré de l'écart entre les fonctions de survie comme défini à l'équation 5.1, mais pondéré par $1/\hat{G}(t_i)$ si un évènement est survenu pour l'observation i avant le temps t et par $1/\hat{G}(t)$ si rien ne s'est encore produit pour l'observation i après le temps t et qu'il est possible qu'il se produise quelque chose après le temps t. Notons que si une observation est censurée avant le temps t, elle ne contribue pas au score de Brier. Ce type de pondération est connu sous le nom d'*Inverse Probability of Censoring Weighting* (*IPCW*).

Ce score de Brier pondéré $BSw(t)$ est défini comme une fonction du temps t, $t > 0$ par :

$$BSw(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{S}(t \mid \mathbf{X}_i)^2 I(t_i \leqslant t \wedge \delta_i = 1)}{\hat{G}(t_i)} + \frac{(1-\hat{S}(t \mid \mathbf{X}_i))^2 I(t_i > t)}{\hat{G}(t)}\right] \quad (5.2)$$

La valeur attendue du score de Brier pour un modèle de prédiction ne comportant pas de covariable correspond à l'estimatation de Kaplan-Meier. Pour obtenir la version non pondérée du score de Brier, $BSunw(t)$, il suffit d'éliminer les termes $\hat{G}(t_i)$ et $\hat{G}(t)$ des dénominateurs.

Les valeurs du score de Brier pondéré sont comprises entre 0 et 1. Un bon pouvoir prédictif au temps t est synonyme de faible score de Brier pondéré. Comme le score de Brier pondéré dépend du temps t, il est naturel de définir le score de Brier pondéré intégré ($IBSw$) de la manière suivante :

$$IBSw = \frac{1}{\max(t_i)}\int_0^{\max(t_i)} BSw(t)dt. \quad (5.3)$$

Il s'agit d'un score qui permet d'évaluer la qualité de l'ajustement des fonctions de survie de toutes les observations pour tous les temps t entre 0 et $\max(t_i)$, $i = 1,\ldots,N$. Il est intéressant de remarquer que le score de Brier intégré $IBSw$ reste approprié pour des méthodes de prédiction qui ne sont pas basées sur des modèles de régression de Cox. Il est ainsi devenu une mesure classique de performance pour les modèles de survie (Hothorn *et al.* [2006]; Schumacher *et al.* [2007]). La version non pondérée $IBSunw$ se définit de manière analogue.

Score de Schmid

Pour un temps t, le score de Schmid pondéré $SSw(t)$ (Schmid *et al.* [2011]) est une mesure d'erreur de prédiction basée sur l'écart absolu, au lieu du carré des écarts pour le score de Brier pondéré, entre les fonctions de survie. C'est une

amélioration robuste de la mesure empirique au temps t de l'écart absolu entre fonctions de survie introduite par Schemper et Henderson [2000], notée $SHw(t)$, et égale à :

$$SHw(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{S}(t \mid \mathbf{X}_i)I(t_i \leqslant t \wedge \delta_i = 1)}{\hat{G}(t_i)} + \frac{(1-\hat{S}(t \mid \mathbf{X}_i))I(t_i > t)}{\hat{G}(t)}\right] \quad (5.4)$$

où $\hat{G}$ et I ont été définies ci-dessus.

Le score de Schmid pondéré au temps t, noté $SSw(t)$, est défini par :

$$SSw(t) = \frac{1}{n}\sum_{i=1}^{n}|I(t_i > t) - \hat{S}(t \mid \mathbf{X}_i)|\left[\frac{I(t_i \leqslant t \wedge \delta_i = 1)}{\hat{G}(t_i^-)} + \frac{I(t_i > t)}{\hat{G}(t_i)}\right] \quad (5.5)$$

où t_i^- est une durée de survie légèrement inférieur à t_i. Pour obtenir la version non pondérée du score de Schmid, $SSunw(t)$, il suffit d'éliminer les termes $\hat{G}(t_i-)$ et $\hat{G}(t)$ des dénominateurs.

Les valeurs du score de Schmid pondéré sont comprises entre 0 et 1. Un bon pouvoir prédictif au temps t est synonyme de faible score de Schmid pondéré. Comme le score de Schmid pondéré dépend du temps t, il est naturel de définir le score de Schmid pondéré intégré ($ISSw$) de la manière suivante :

$$ISSw = \frac{1}{\max(t_i)}\int_0^{\max(t_i)} SSw(t)dt. \quad (5.6)$$

Il s'agit d'un score qui permet d'évaluer la qualité de l'ajustement des fonctions de survie de toutes les observations pour tous les temps t entre 0 et $\max(t_i)$, $i = 1, \ldots, N$. Il est intéressant de remarquer que le score de Schmid pondéré intégré $ISSw$ reste approprié pour des méthodes de prédiction qui ne sont pas basées sur des modèles de régression de Cox. La version non pondérée $ISSunw$ se définit de manière analogue.

Résultats

Les résultats des simulations, voir figures 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12 pour le cas d'un modèle à quatre composantes, ont mis en avant l'estimateur de Song et Zhou (iAUCSH), implémenté dans le package `survAUC`, Potapov *et al.* [2012], comme étant le meilleur critère de validation croisée pour les deux méthodes PLS-Cox et autoPLS-Cox même s'il a de piètres performances dans les autres cas.

En ce qui concerne les autres méthodes, ce sont les critères iAUCsurvROC, iAUCUno qui ont eu les meilleurs résultats. Les deux critères non pondérés iBSunw et iSSunw échouent uniformément pour tous les modèles.

Le critère iBSw est trop conservatif et sélectionne à tort des modèles vides dans plus de la moitié des cas pour le lien linéaire et dans la quasi-totalité des cas pour le lien quadratique.

Le critère iSSw n'a pas de bons résultats pour les méthodes Cox-PLS, sPLSDR et DKsPLSDR et seulement des résultats moyens dans le cas des méthodes PLSDR et DKPLSDR.

Les deux modèles SPLSDR et DKSPLSDR nécessite un paramètre additionnel : le seuil η. Celui-ci a également été étudié et des figures similaires ont été produites pour tous les critères, voir Bertrand *et al.* [2018a]. Dans le cas particulier des deux critères iAUCUno et iAUCsurvROC, la dispersion observée montre une dispersion raisonnable pour le paramètre η.

Cette campagne de simulations a permis de trouver de meilleurs critères pour la sélection du nombre correct de composantes lors d'une validation croisée des extensions PLS ou sPLS des modèles de Cox. Voici les recommandations qui en ressortent :

- iAUCsh pour PLS-Cox et autoPLS-Cox.
- iAUCSurvROC et iAUCUno pour Cox-PLS, (DK)PLSDR et (DK)sPLSDR.

Implémentation

Ces recommandations ont été implémentées comme choix par défaut pour la validation croisée de ces modèles dans le package `plsRcox`. Ce sont elles que nous utiliserons par la suite pour réévaluer les performances de ces modèles.

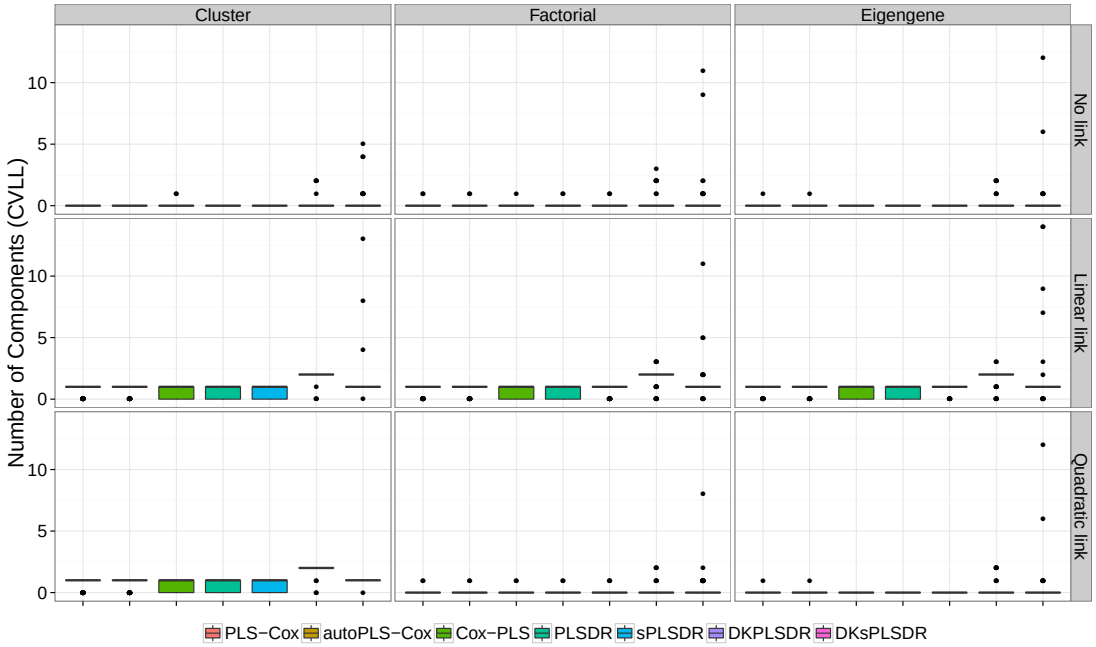


Figure 5.1 : Nombre de composantes retenues, critère LL.

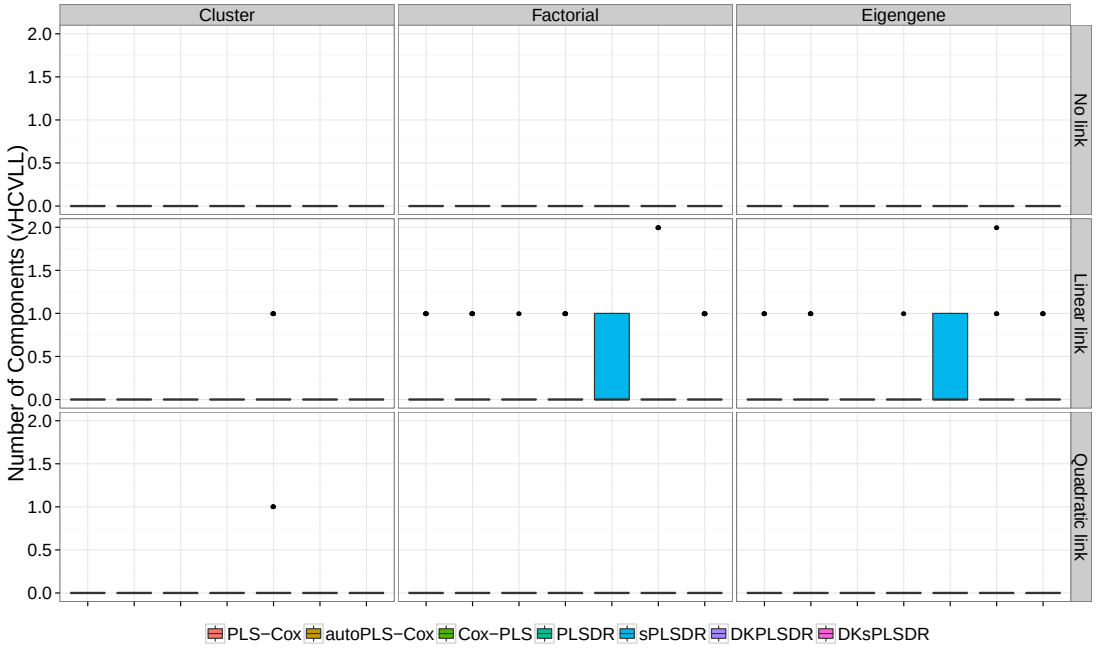


Figure 5.2 : Nombre de composantes retenues, critère vHLL.

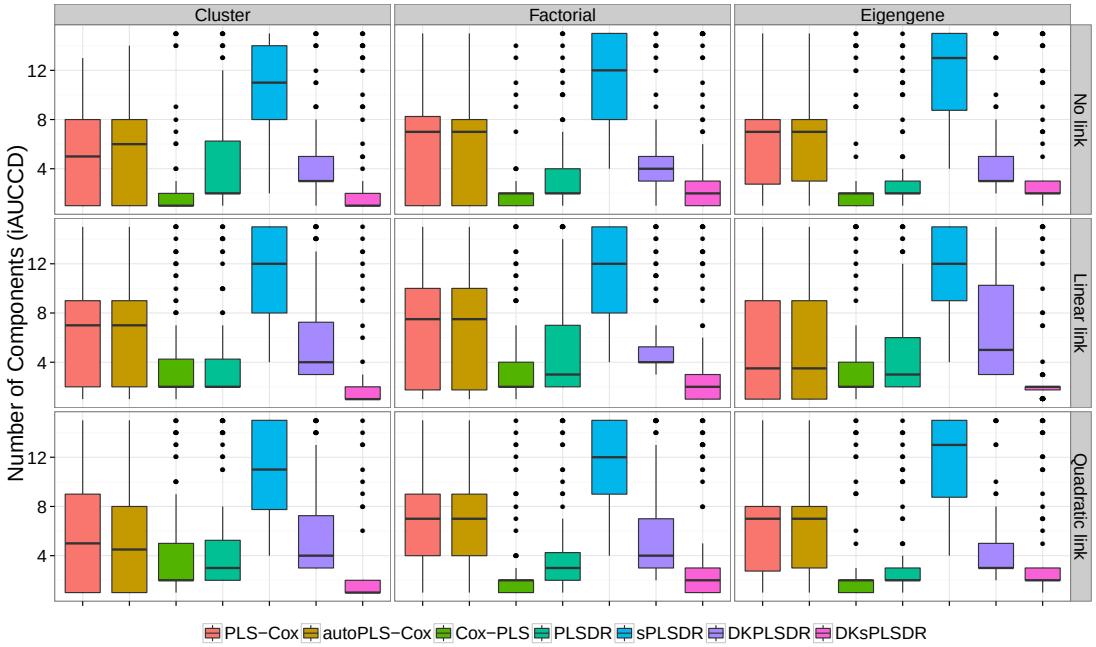


Figure 5.3 : Nombre de composantes retenues, critère iAUCCD.

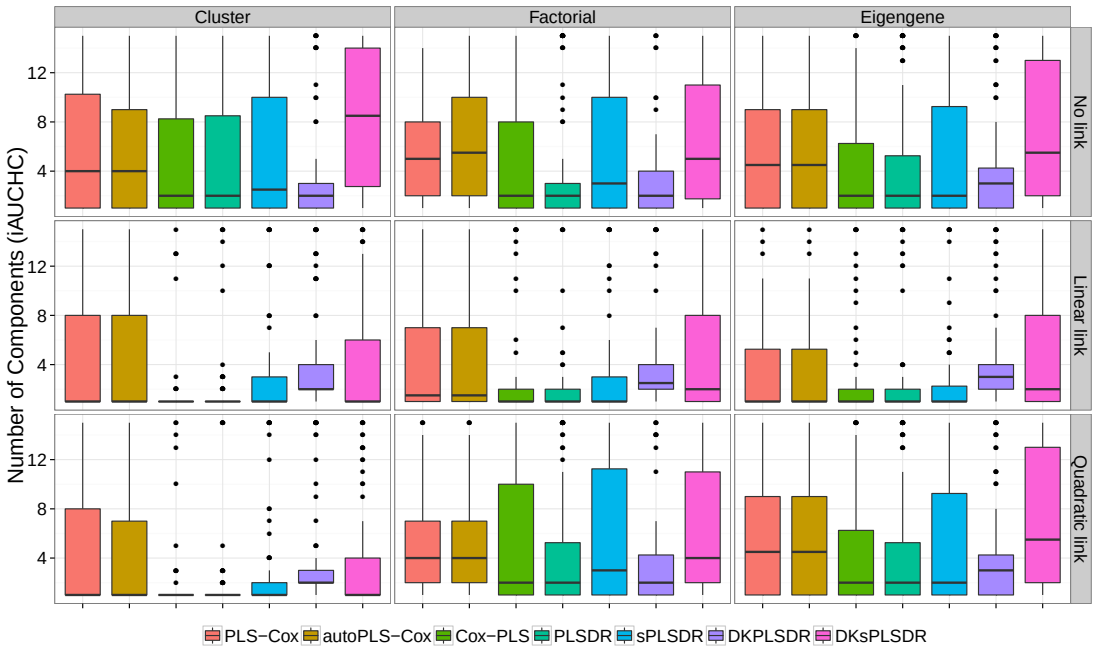


Figure 5.4 : Nombre de composantes retenues, critère iAUCHC.

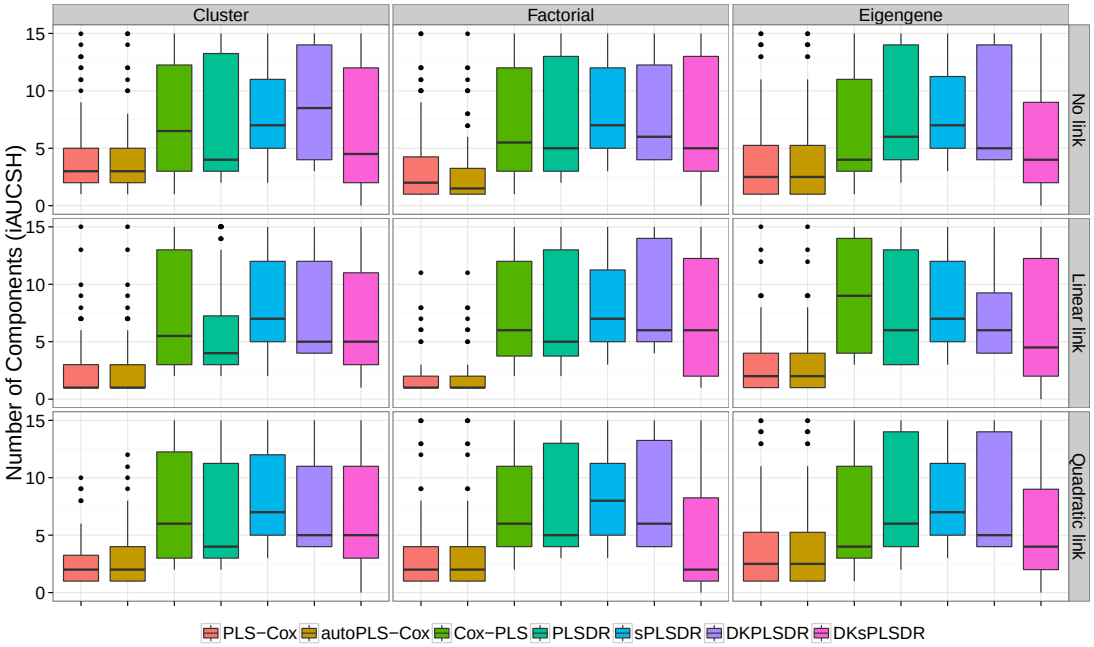


Figure 5.5 : Nombre de composantes retenues, critère iAUCSH.

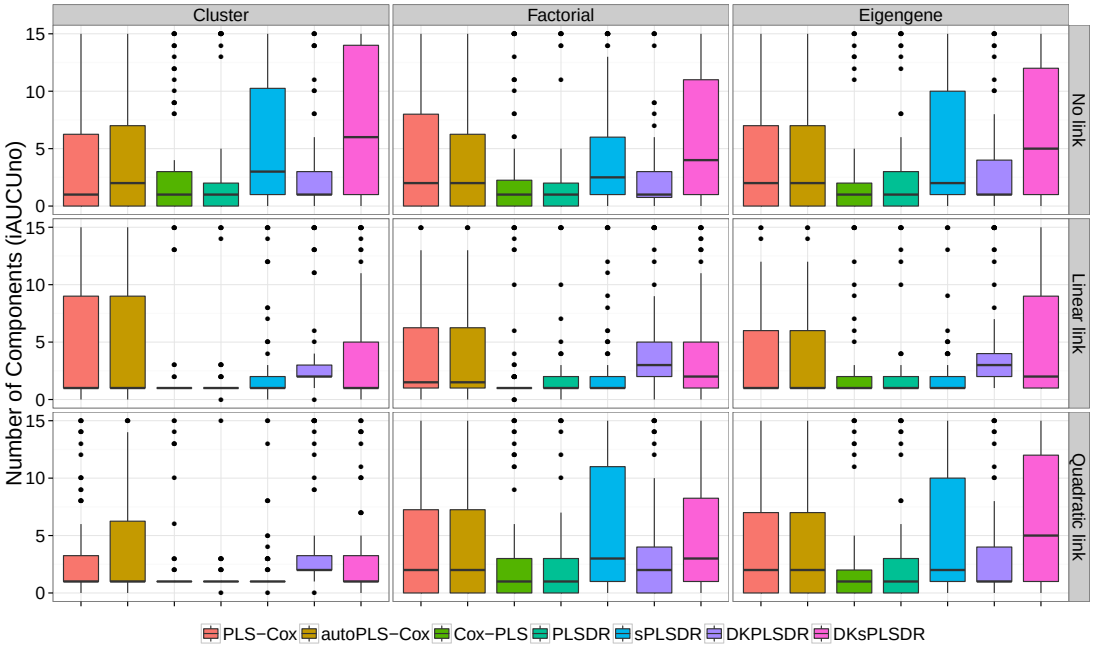


Figure 5.6 : Nombre de composantes retenues, critère iAUCUno.

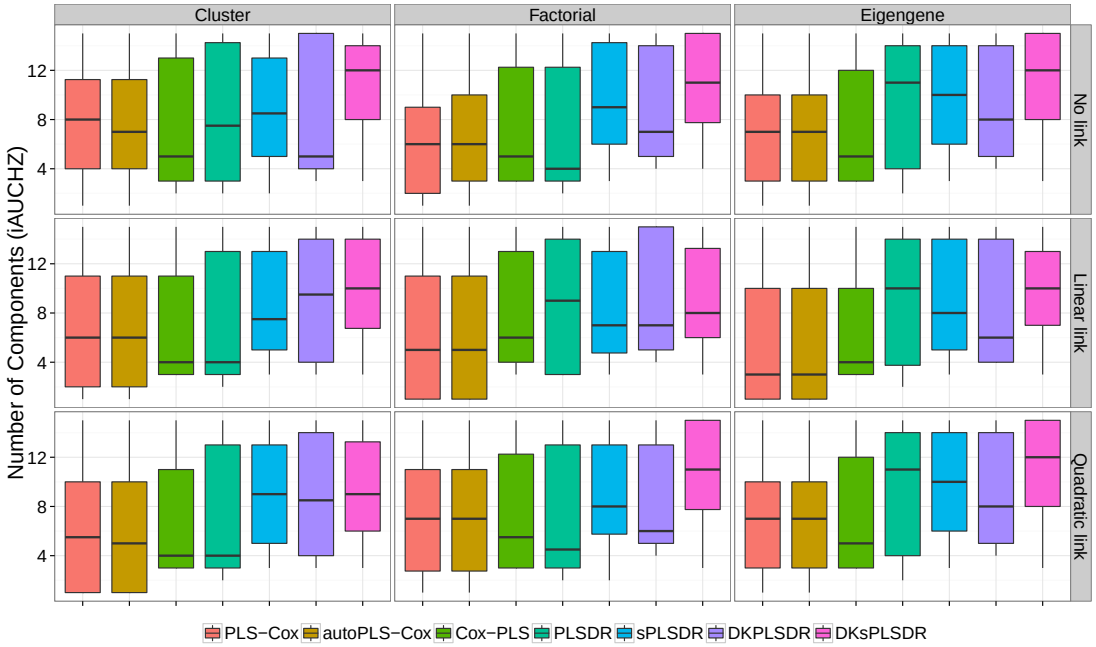


Figure 5.7 : Nombre de composantes retenues, critère iAUCHZ.

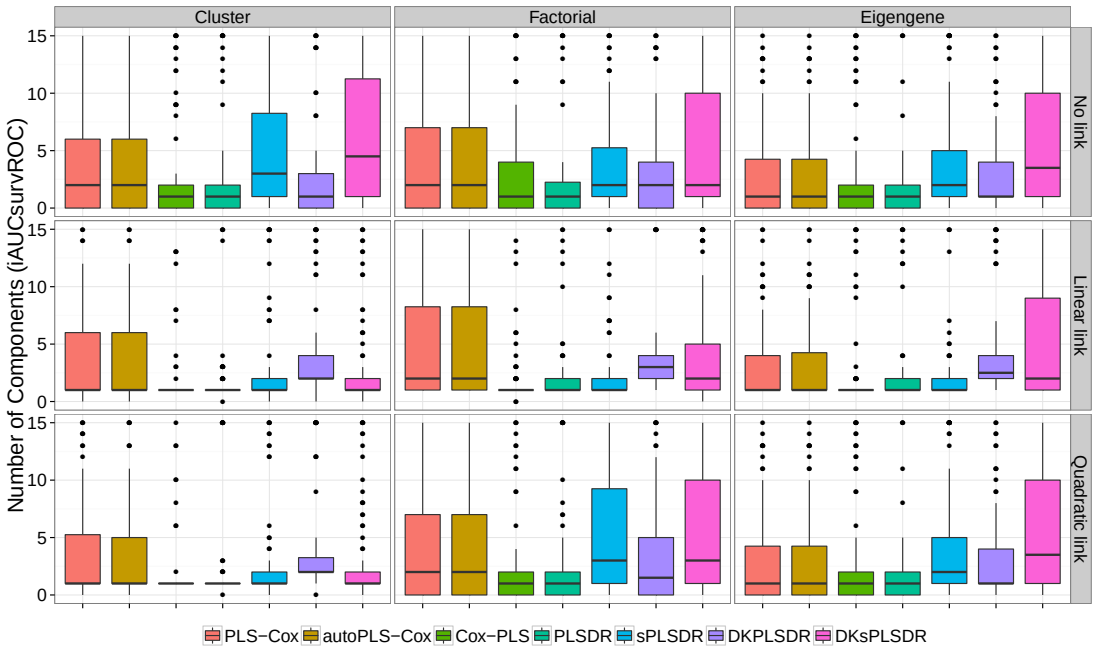


Figure 5.8 : Nombre de composantes retenues, critère iAUCSurvROC.

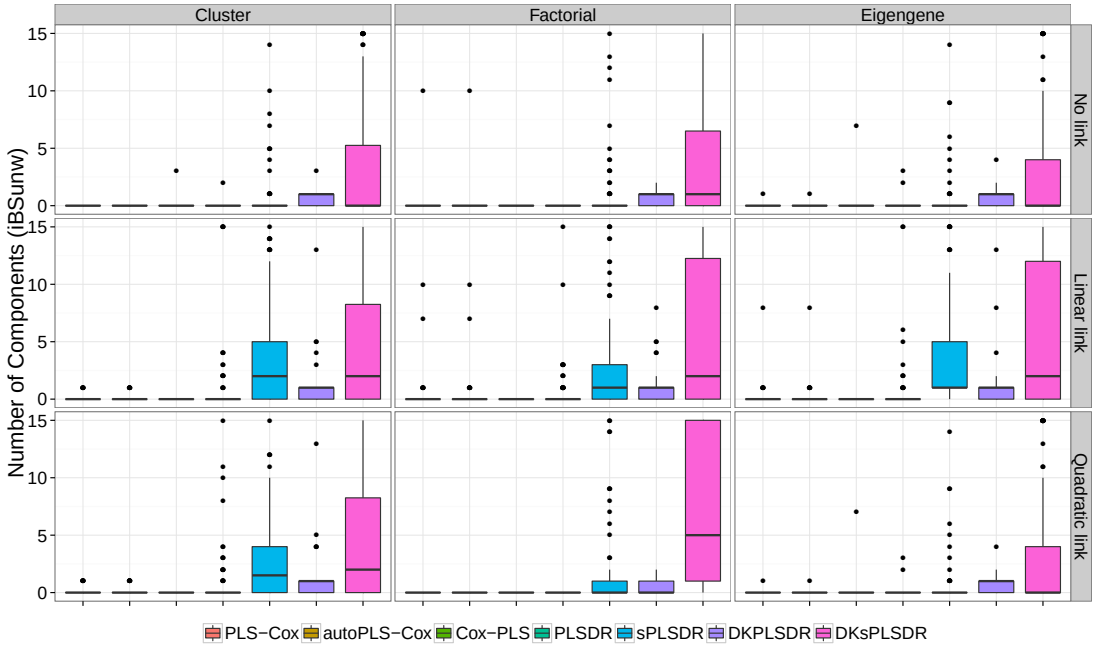


Figure 5.9 : Nombre de composantes retenues, critère iBSunw.

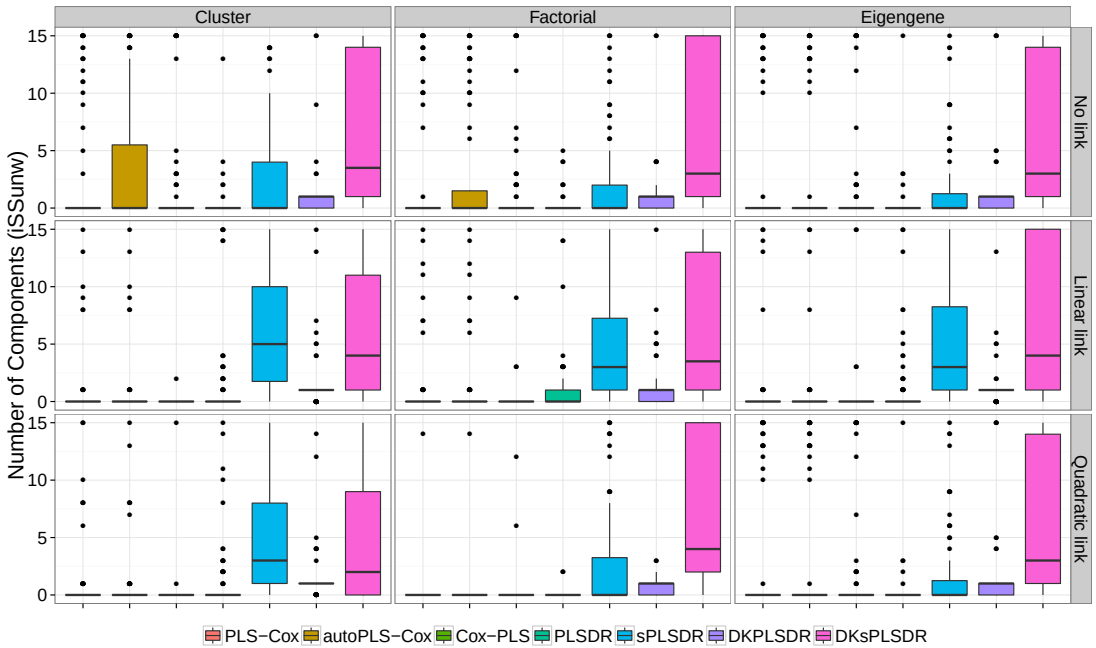


Figure 5.10 : Nombre de composantes retenues, critère iSSunw.

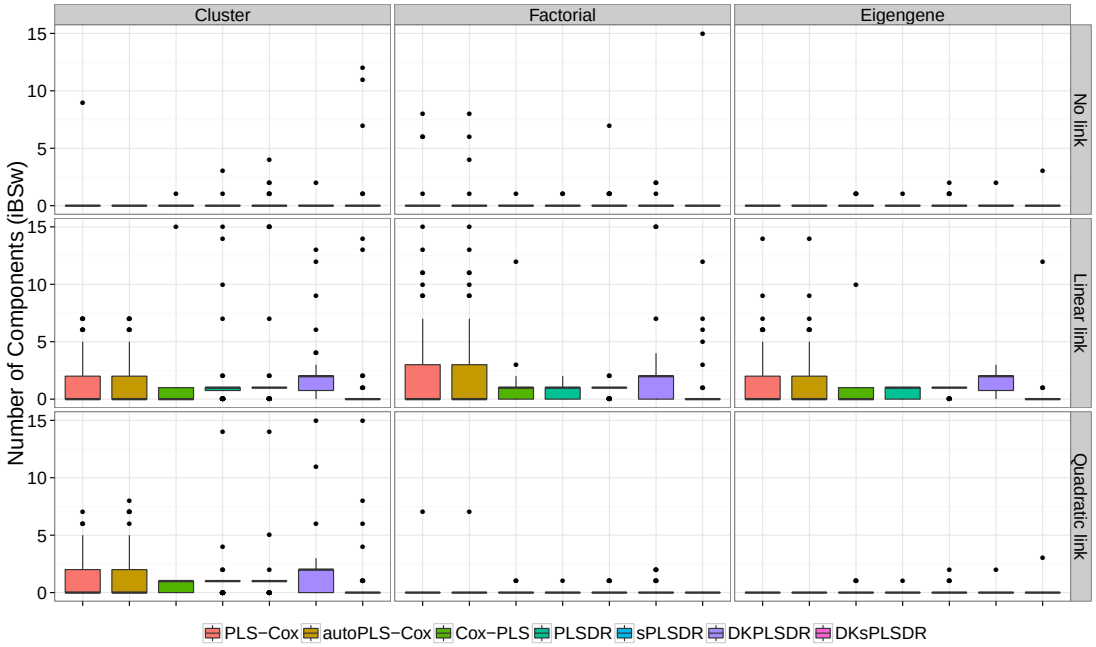


Figure 5.11 : Nombre de composantes retenues, critère iBSw.

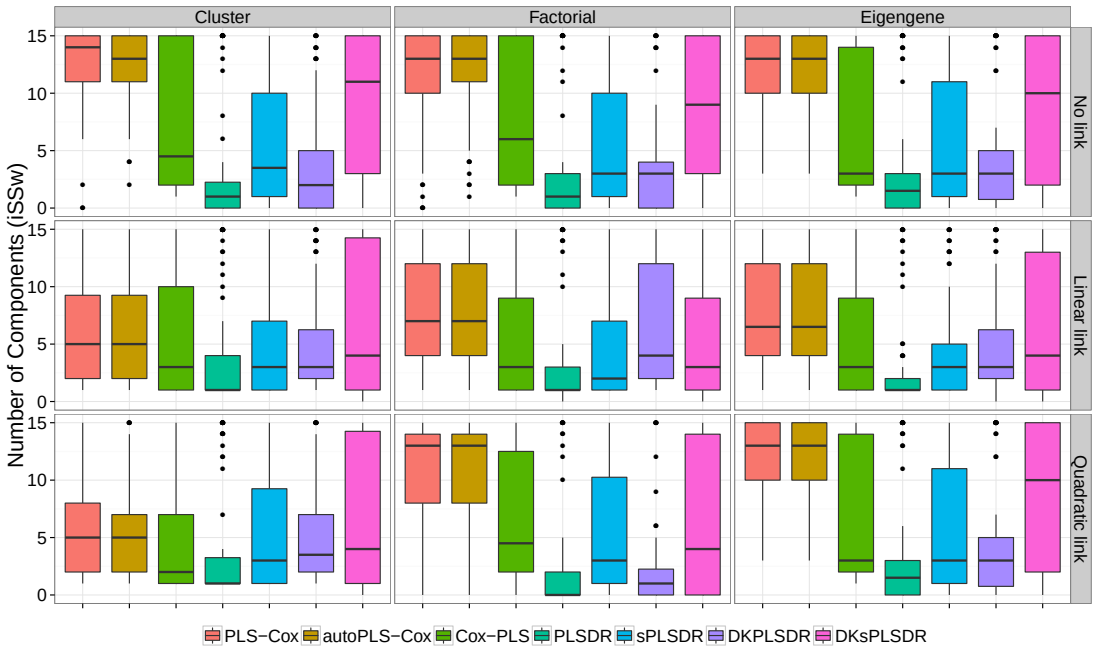
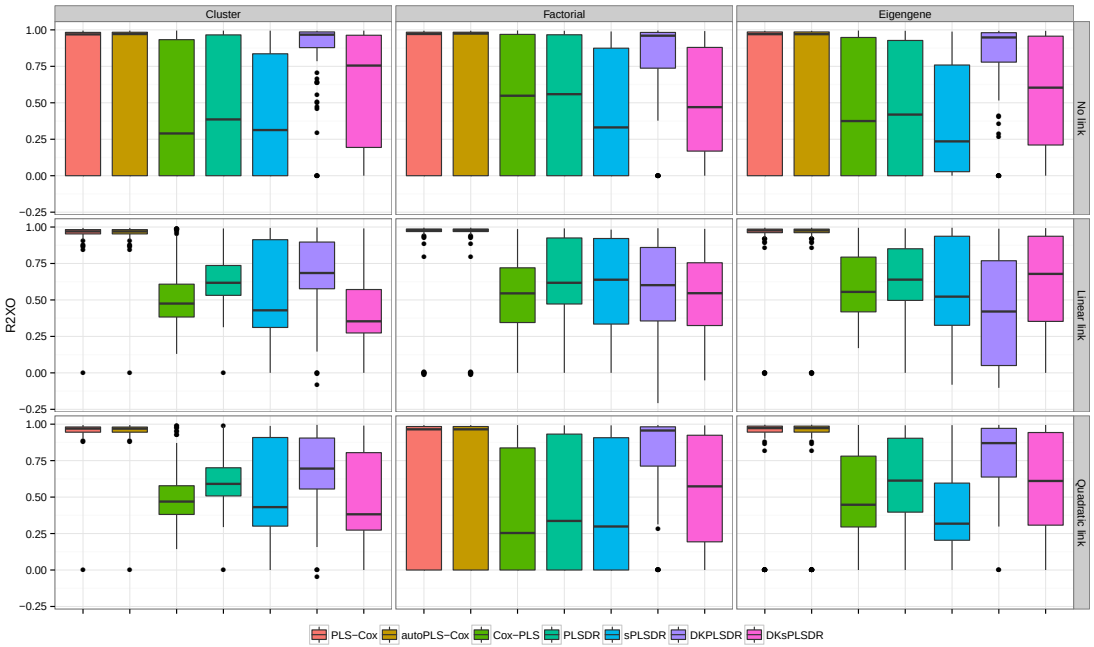
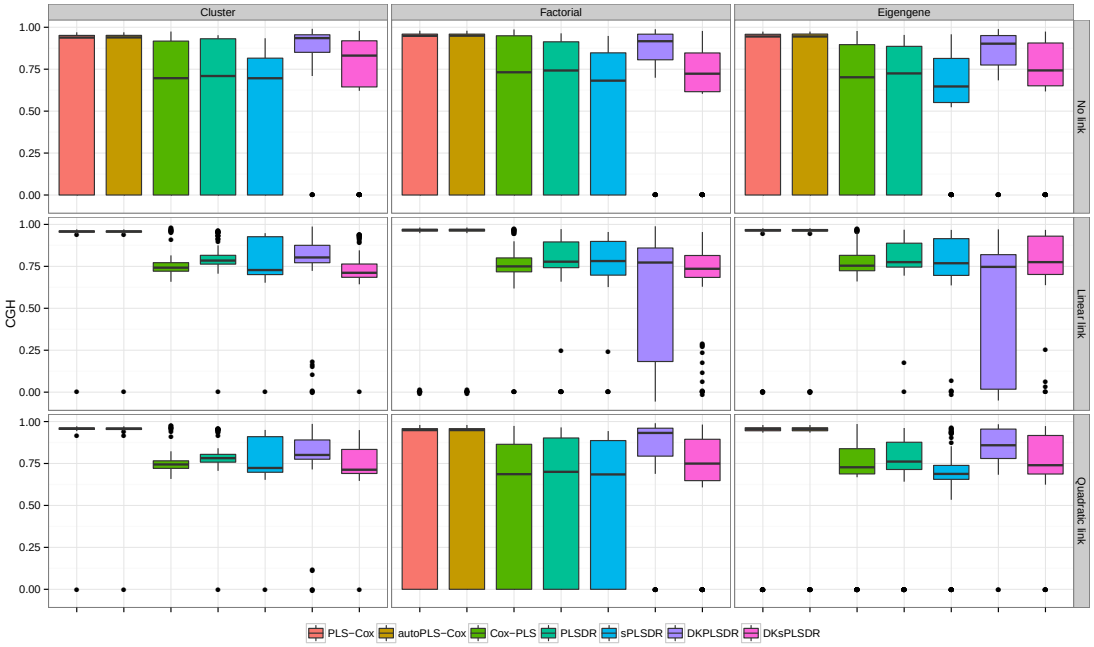


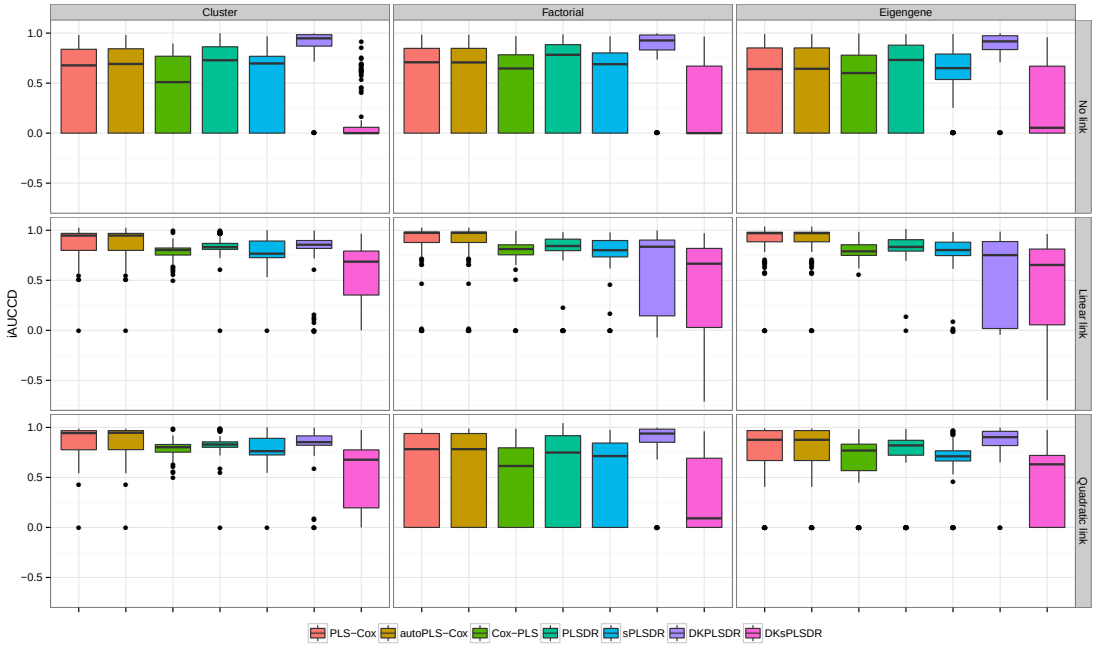
Figure 5.12 : Nombre de composantes retenues, critère iSSw.



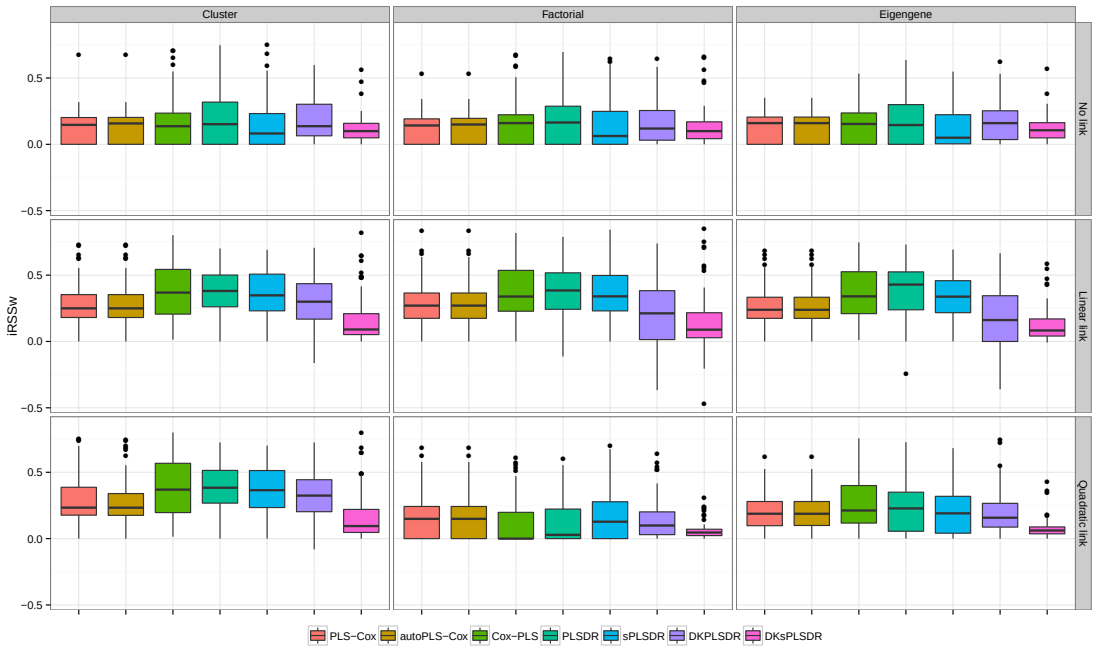
Delta (iAUCSurvROC CV − vHCVLL value). *Figure 5.13 : mesure R2XO.*



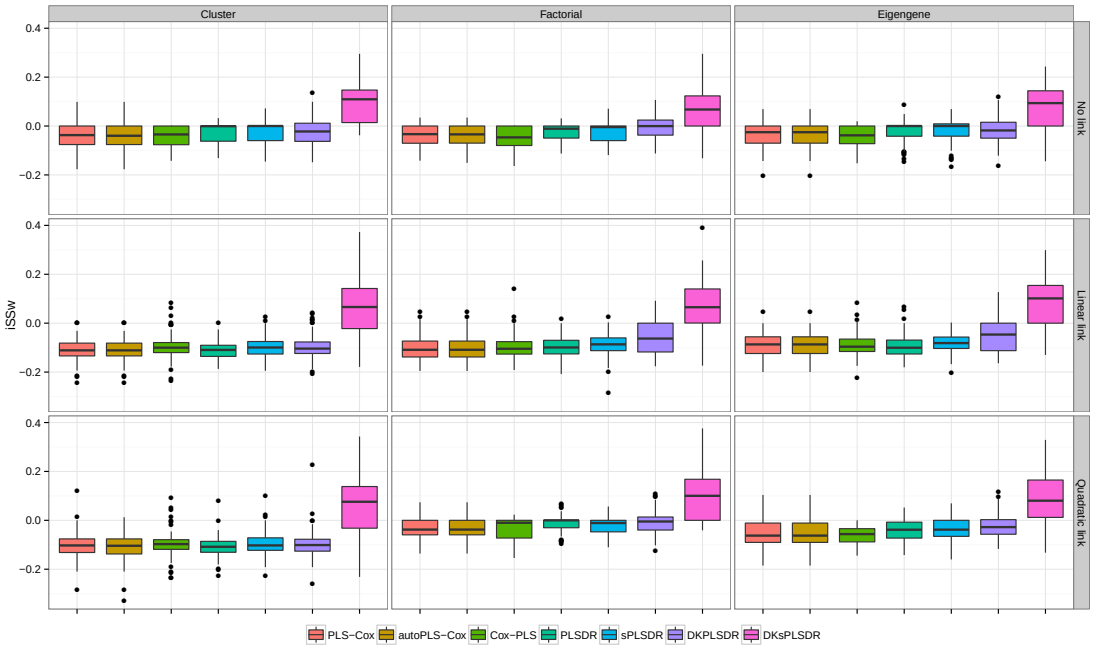
Delta (iAUCSurvROC CV − vHCVLL value). *Figure 5.14 : mesure GHCI.*



Delta (iAUCSurvROC CV − vHCVLL value). *Figure 5.15 : mesure iAUCCD.*



Delta (iAUCSurvROC CV − vHCVLL value). *Figure 5.16 : mesure iRSSw.*



Delta (iAUCSurvROC CV − vHCVLL value). *Figure 5.17 : mesure iSSw.*

5.3.3 Réévaluation des performances des modèles basés sur la PLS

Puis à l'aide de ces nouveaux critères de validation croisée, j'ai ajusté des extensions plus ou moins complexes de la régression moindres carrés partiels au modèle de Cox afin de réévaluer leur performance.

Choix des mesures de performance

Or la détermination de mesures de la précision de la prédiction pour les données de survie n'est pas évidente en présence de données censurées. Pour résoudre ce problème, un nombre varié d'approches différentes ont été suggérées dans la littérature. J'ai sélectionné 23 critères qui se répartissent en trois groupes :

- Approches basées sur la vraisemblance (llrt, varresmart, trois de type R^2).
- Approches basées sur la ROC comme les AUC intégrés (iAUCCD, iAUCHC, iAUCSH, iAUCUno, iAUCHZ, iAUCsurvROC) ou 3 C-index (Harrell, GHCI, UnoC).
- Approches basées sur une distance comme le V de Schemper et Henderson [2000] ou dérivées des scores de Brier ou de Schmid (BS(un)w, SS(un)w et les quatre mesures de type R^2 qui s'en déduisent).

Pour plus de détails voir Bertrand *et al.* [2018a] et pour un résumé voir le tableau 5.1.

Un niveau critère de performance robuste

En notant par BS^0, l'estimateur de Kaplan-Meier construit à partir des $(t_i,\ \delta_i)$, ce qui correspond à une prédiction sans covariable, nous commençons par définir un coefficient R^2 basé sur un score de Brier, pondéré ou non, $R^2_{BS(un)w}$, pour tout $t > 0$:

$$R^2_{BS(un)w}(t) = 1 - \frac{BS(un)w(t)}{BS^0(un)w(t)}. \tag{5.7}$$

Alors le coefficient R^2 basé sur un score de Brier pondéré intégéré, iR2BSw, Graf *et al.* [1999], est défini par :

$$iR2BSw = \frac{1}{\max(t_i)} \int_0^{\max(t_i)} R^2_{BSw}(t)dt. \tag{5.8}$$

Le critère a été utilisé, par exemple, dans Bøvelstad *et al.* [2007] et Lambert-Lacroix et Letué [2011]. Ce critère intégré iR2BSw est peu sensible à la censure, Hielscher *et al.* [2010], et, en tant que mesure basée sur la norme quadratique, n'est pas robuste. La version non pondérée, iR2BSunw, se définit de manière analogue.

J'ai proposé un nouveau critère : le iRSSw. Il s'agit d'une mesure intégrée construite à partir du score de Schmid et qui, au lieu de se baser sur le R^2 traditionnel déduit de la norme quadratique, utilise le coefficient de détermination R de la régression par moindres valeurs absolues, introduit dans McKean et Sievers [1987].

Notons $SS(t)$ la valeur du score de Schmid au temps t pour le modèle étudié et $SS^0(t)$ la valeur du score de Schmid au temps t pour une prédiction réalisée sans covariable. Nous commençons par définir R_{SS} comme une fonction du temps t, avec $t > 0$:

$$R_{SS}(t) = 1 - \frac{SS(t)}{SS^0(t)}. \tag{5.9}$$

Alors le critère intégré iRSSw, est défini par :

$$iRSSw = \frac{1}{\max(t_i)} \int_0^{\max(t_i)} R_{SS}(t)dt. \tag{5.10}$$

La version non pondérée, iR2SSunw, se définit de manière analogue.

Résultats

En utilisant les critères de validation croisée que j'ai sélectionnés à la section 5.3.2 et sur la base de ces 23 mesures de prédiction, j'ai réalisé une réévaluation des performances des extensions PLS-Cox, autoPLS-Cox, Cox-PLS, PLSDR, DKPLSDR, sPLSDR et DKsPLSDR. Elle a montré une nette amélioration de ces méthodes, voir figures 5.13, 5.14, 5.15, 5.16, 5.17, même lorsque comparées avec des approches concurrentes bien connues comme *coxnet*, *coxpath*, *uniCox* and *glcoxph*, voir figures 5.18, 5.19, 5.20, 5.21, 5.22 et encore plus de détails et de figures dans Bertrand *et al.* [2018a].

De ce fait, les critères recommandés améliorent non seulement la précision dans le choix du nombre correct de composantes des modèles mais aussi la performance de ces modèles permettant à certains d'entre eux d'avoir de meilleurs résultats que des approches de référence.

Implémentation

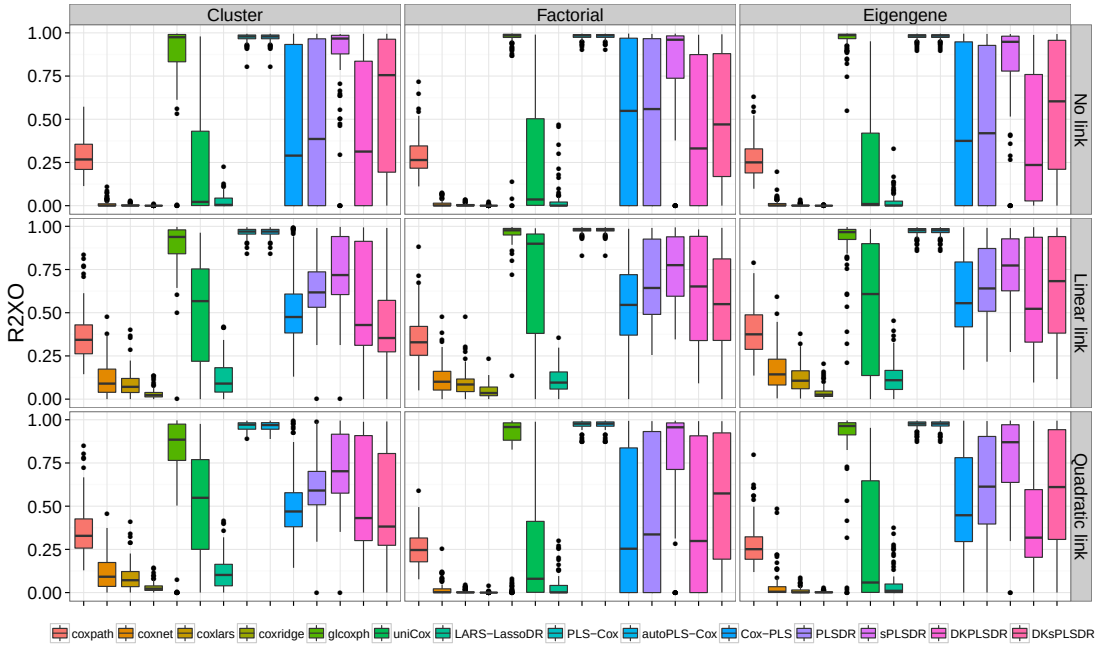
J'ai utilisé ces résultats pour faire évoluer les choix par défaut de validation croisée dans le package `plsRcox`, Bertrand et Maumy-Bertrand [2018e]. J'ai communiqué sur ces résultats, Bertrand *et al.* [2014a], Bertrand *et al.* [2015], et rédigé un article à ce sujet Bertrand *et al.* [2018a].

5.4 Perspectives

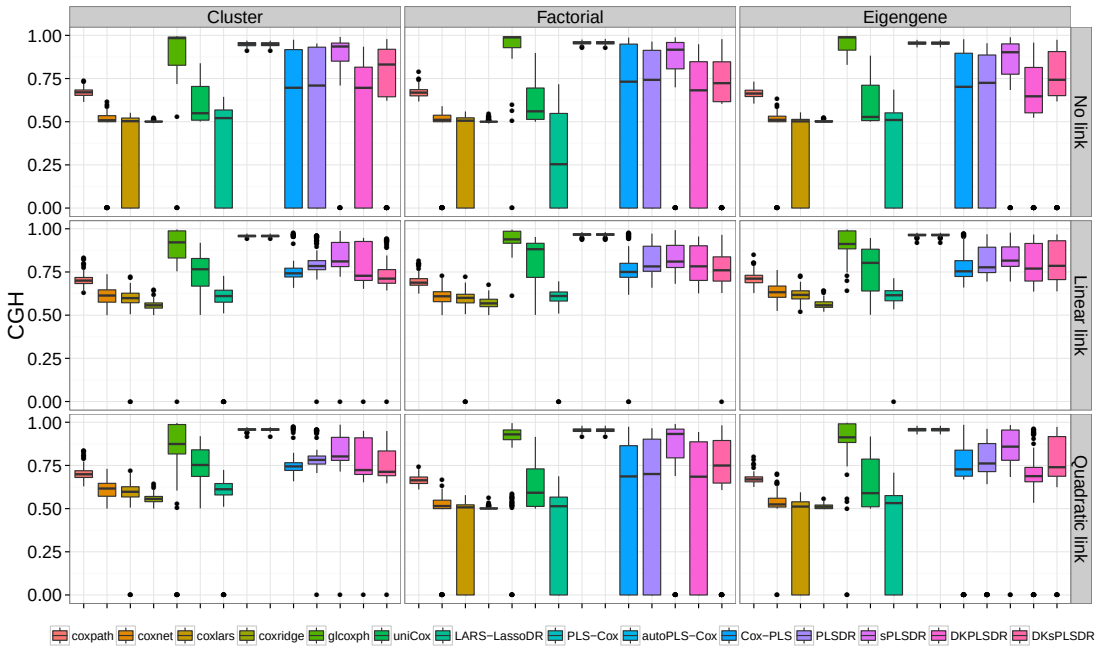
Il serait pertinent de d'étudier la robustesse de ces extensions par rapport à la proportion ou au type de valeurs manquantes, par rapport à la présence de valeurs atypiques, par rapport à la validation croisée, pour déterminer si des approches comme celles proposées dans Magnanensi *et al.* [2017] ou Magnanensi *et al.* [2016b] seraient pertinentes. J'aimerai obtenir, comme je l'ai indiqué à la section 4.4 pour l'extension de la PLS à régression linéaire généralisée, une évaluation des degrés de liberté, à la manière de Kraemer et Sugiyama [2011], dans les modèles issus de ces nouvelles extensions, le cas de la régression Bêta semblant être le plus simple par lequel commencer.

Critère	Type	Pour la validation croisée			Comme mesure de performance		
		Testé	Résultats dans article	Recommandé pour	Est une mesure de perf. ?	Bonnes propriétés statistiques	Résultats dans article
CVLL	LBa	**Oui**	**Oui**		Non	Non	Non
vHCVLL	LBa	**Oui**	**Oui**		Non	Non	Non
LRT p-value	LBa	Non	Non		**Oui**	**Oui**	Non
VarM	LBa	Non	Non		**Oui**	Non	Non
R2Nag	LBa	Non	Non		**Oui**	Non	Non
R2XO	LBa	Non	Non		**Oui**	**Oui**	**Oui**
R2OXS	LBa	Non	Non		**Oui**	Non	Non
iR2BSunw	DBa	Non	Non		**Oui**	Non	Non
iR2BSw	DBa	Non	Non		**Oui**	**Oui**	Non
iRSSunw	DBa	Non	Non		***Nouveau***	Non	Non
iRSSw	DBa	Non	Non		***Nouveau***	**Oui**	**Oui**
iAUCCD	ROCBa	**Oui**	**Oui**		**Oui**	**Oui**	**Oui**
iAUCHC	ROCBa	**Oui**	**Oui**		**Oui**	**Oui**	Non
iAUCSH	ROCBa	**Oui**	**Oui**	PLS−Cox, autoPLS−Cox	**Oui**	**Oui**	Non
iAUCUno	ROCBa	**Oui**	**Oui**	(DK)(s)PLSDR Cox−PLS	**Oui**	**Oui**	Non
iAUCHZ	ROCBa	**Oui**	**Oui**		**Oui**	**Oui**	Non
iAUC SurvROC	ROCBa	**Oui**	**Oui**	(DK)(s)PLSDR Cox−PLS	**Oui**	**Oui**	**Oui**
C	ROCBa	Non	Non		**Oui**	Non	Non
UnoC	ROCBa	Non	Non		**Oui**	Non	Sup. Info.
GHCI	ROCBa	Non	Non		**Oui**	**Oui**	**Oui**
SchemperV	DBa	Non	Non		**Oui**	**Oui**	Non
iBSunw	DBa	**Oui**	**Oui**		**Oui**	Non	Non
iBSw	DBa	**Oui**	**Oui**		**Oui**	**Oui**	Sup. Info.
iSSunw	DBa	**Oui**	**Oui**		**Oui**	Non	Non
iSSw	DBa	**Oui**	**Oui**		**Oui**	**Oui**	**Oui**
Nombre Total	25	12		12	23	14	6 (+2 SI)

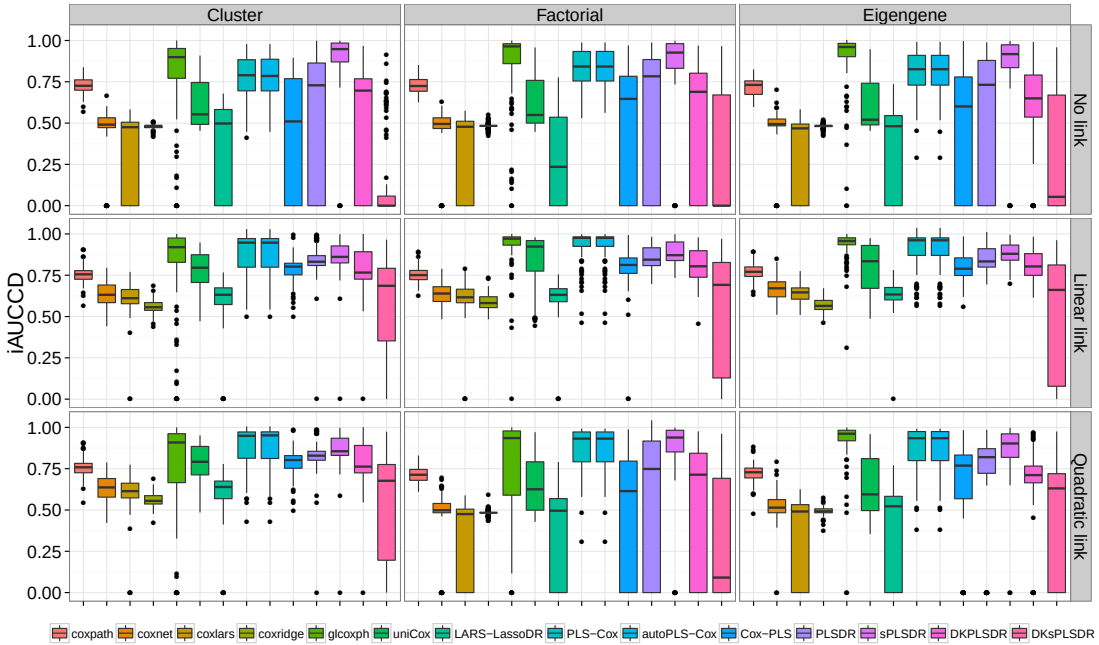
Table 5.1 : *Tableau récapitulatif des différents critères apparaissant dans l'étude et de leur utilisation comme critère de validation croisée ou comme mesure de performance servant à évaluer le modèle.*



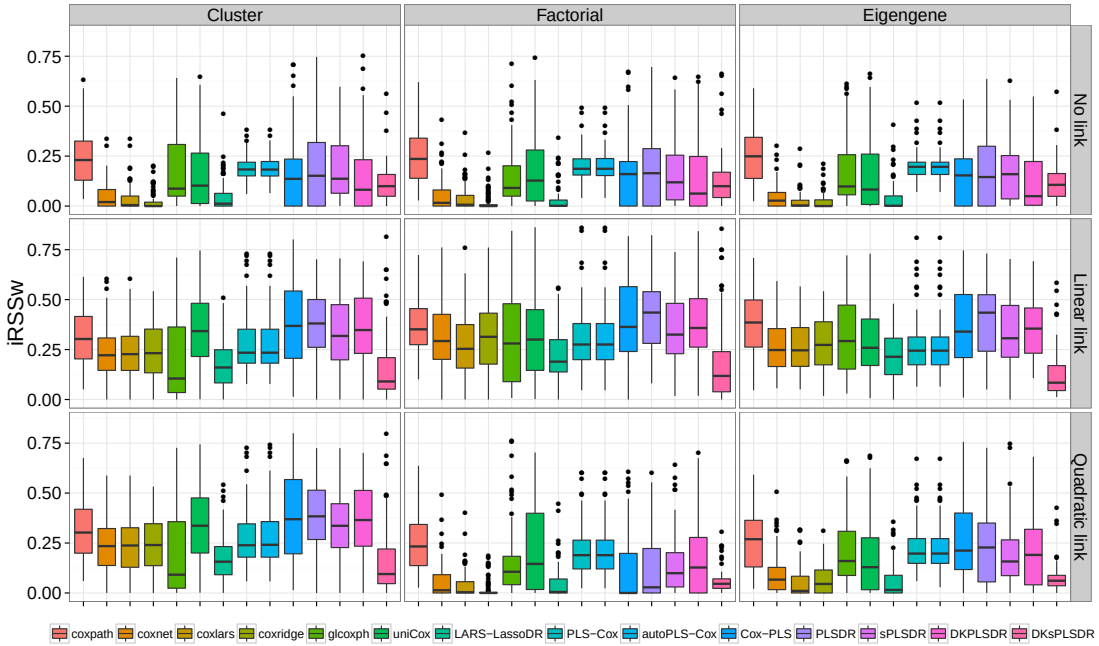
Performance, iAUCsurvROC CV. Figure 5.18 : mesure R2XO



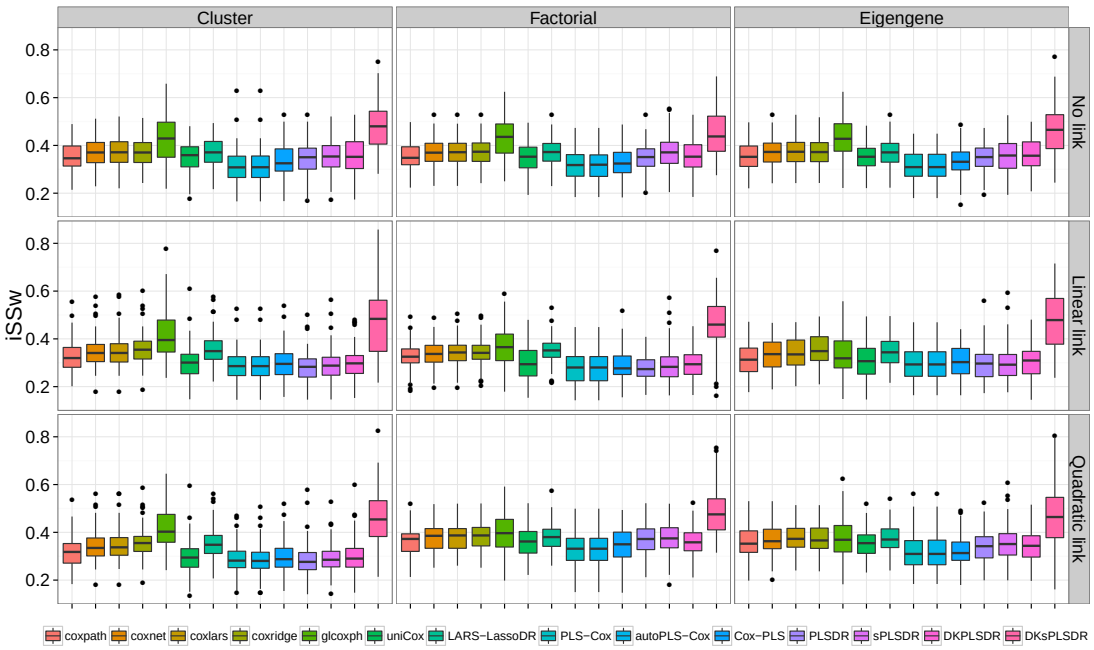
Performance, iAUCsurvROC CV. Figure 5.19 : mesure GHCI.



Performance, iAUCsurvROC CV. *Figure 5.20 : mesure iAUCCD.*



Performance, iAUCsurvROC CV. *Figure 5.21 : mesure iRSSw.*



Performance, iAUCsurvROC CV. *Figure 5.22 : mesure iSSw.*

Chapitre 6

Modélisation des données génomiques et protéomiques

6.1 Intervention dans un réseau de gènes

6.1.1 Contexte

Ce projet s'articule dans une collaboration transdisciplinaire, mathématiques – biologie – médecine, commencée en février 2011 avec l'encadrement d'un stagiaire, Nicolas Jung, de master deuxième année de mathématiques dirigé, d'une part, par Laurent Vallat (MCU-PH) du laboratoire d'ImmunoRhumatologie Moléculaire, UMR_S 1109 INSERM et UDS, LabEx TRANSPLANTEX, et, d'autre part, Myriam Maumy-Bertrand et moi-même. À l'issue de ce stage, nous avons obtenu une bourse de thèse pour Nicolas Jung.

L'objectif est d'obtenir une modélisation mathématique d'un système complexe afin de pouvoir y pratiquer des interventions dirigées, c'est-à-dire de faire évoluer le système dans le sens recherché. La finalité ultime d'application est le traitement de certains cancers.

6.1.2 Présentation du problème biologique

Cette collaboration se concentre autour du sujet suivant : quand une cellule est stimulée, le programme génique qu'elle contient est activé. Plusieurs centaines de gènes mis en action apportent alors une réponse concertée au stimulus sous la forme de molécules d'ARN qui sont, par la suite, traduites en protéines qui détermineront la réponse de la cellule, Crick [1970]. Cette réponse complexe et dynamique peut

être modélisée statistiquement par un réseau dans lequel les nœuds correspondent aux gènes et les liens correspondent à leurs interactions.

Pour inférer le réseau de gènes, il est possible d'étudier le niveau d'expression de ces derniers grâce à des puces à ADN (*microarrays*) qui permettent de mesurer la quantité d'ARN messager produite par chaque gène activé. Afin de pouvoir essayer de s'approcher d'une détermination d'un lien de causalité, il est important de mesurer l'expression des gènes au cours du temps. La causalité est en elle-même un sujet statistique complexe, Pearl [2000], Pearl [2009], Pearl et Mackenzie [2018], et la preuve de relations causales n'est pas simple. Dans notre cas, l'analyse des motifs d'expression temporels des gènes permet de voir ceux dont la variation d'expression précède les autres et permet de trouver des candidats pour une relation de cause à effet. Il faut alors que le biologiste réalise des expériences contrôlées pour valider ces résultats.

Notons que le problème d'inférence est double car il faut non seulement retrouver quels sont les liens possibles entre les gènes, c'est-à-dire la structure du réseau, mais aussi appréhender l'évolution temporelle du signal se propageant dans le réseau.

Dans notre contexte, une stimulation biologique est à réaliser à un instant donné t_0. Pour pouvoir suivre l'évolution de la réponse de la cellule à cette stimulation les cellules ont été réparties en deux lots. D'une part, nous mesurons leur évolution suite à la *vraie* stimulation et d'autre part nous mesurons leur évolution suite à une stimulation *vide* qui servira de contrôle. En effet, le processus de stimulation en lui-même induit déjà une réponse cellulaire qui n'est pas l'objet de cette étude. Les réseaux de gènes impliqués peuvent alors être modélisés sous forme d'interactions en cascade (Luscombe *et al.* [2004], Yosef et Regev [2011]) et Figure 6.1), mais peu d'outils spécifiques ont été développés à ce jour pour appréhender ces phénomènes de régulation en cascade

Néanmoins, depuis l'introduction de technologies à haut débit qui permettent de mesurer simultanément l'expression de milliers de gènes, beaucoup de méthodes statistiques ont été proposées pour l'inférence de ces réseaux de régulation, Bansal *et al.* [2007], Hecker *et al.* [2009] et Bar-Joseph *et al.* [2012]. Ces méthodes peuvent être regroupées en trois catégories principales.

Il y a d'abord les méthodes dites d'interactions, dans lesquelles une mesure de proximité entre les gènes est définie, comme l'entropie dans la méthode ARACNe de Margolin *et al.* [2006]. Ces méthodes sont relativement peu coûteuses en temps de calcul, mais elles ne permettent pas de décrire la dynamique des systèmes biologiques.

Nous trouvons ensuite les méthodes dites d'optimisation dans lesquelles il convient de distinguer les réseaux booléens d'une part (Liang *et al.* [1998]), et les réseaux bayésiens d'autre part (Dondelinger *et al.* [2012]). Dans ces derniers, les

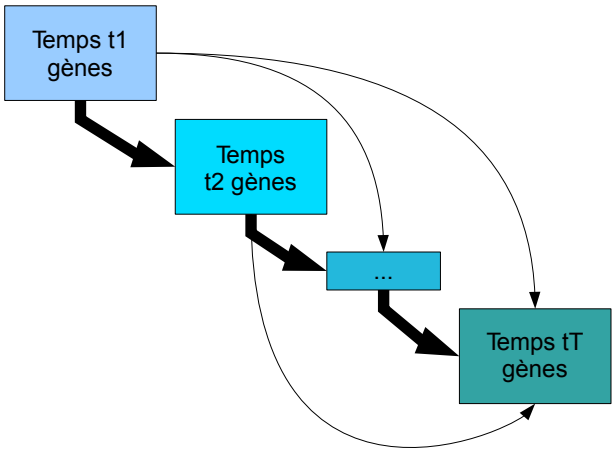


Figure 6.1 : *Principe de la méthode d'inférence des réseaux de gènes provenant de plusieurs états.*

différents gènes régulateurs d'un gène donné sont appelés parents. Des probabilités a priori de chaque gène (sachant ses parents) sont alors définies. Par la formule de Bayes, nous cherchons alors la structure de réseau qui maximise la probabilité a posteriori (sachant les valeurs observées pour les expressions de gènes). Ces méthodes sont particulièrement efficaces dans l'inférence de réseaux de gènes et leur intérêt majeur est de pouvoir distinguer les interactions directes de celles qui sont indirectes (grâce au conditionnement par rapport aux parents). Cependant, ces méthodes dans lesquelles un algorithme de recherche des réseaux possibles est souvent nécessaire, ne sont pas exploitables pour des réseaux contenant plusieurs centaines de gènes, ce qui est le cas dans le modèle biologique que nous étudions, Rau *et al.* [2010] et Rau *et al.* [2012].

Enfin, il y a les méthodes basées sur des équations différentielles ou des régressions, dans lesquelles des techniques spécifiques doivent être utilisées, du fait que le nombre d'observations est souvent largement inférieur au nombre de variables (les gènes). Vu sous cet angle, le problème peut se poser sous la forme d'une sélection de variables. L'approche la plus courante consiste à pénaliser l'estimation des paramètres dans la régression linéaire, comme dans Gustafsson *et al.* [2009], Gustafsson et Hörnquist [2010]. Ces méthodes sont quant à elles particulièrement bien adaptées dans le cadre d'inférence de réseaux de grande taille.

Certaines maladies comme le cancer affectent le programme génique. Il est alors intéressant de chercher à inférer le programme génique des individus sains et des patients malades. Il est légitime de supposer que seule une partie du réseau

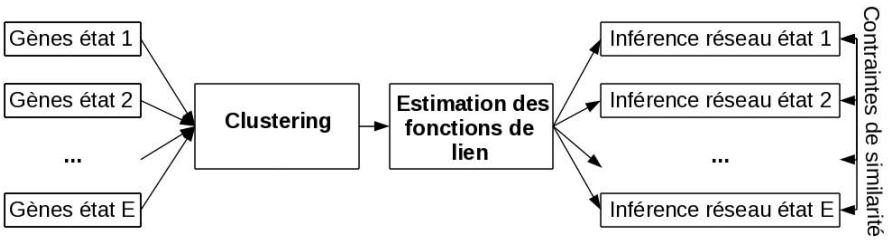


Figure 6.2 : *Principe de la méthode d'inférence des réseaux de gènes provenant de plusieurs états.*

de gènes soit altérée d'un état à l'autre ; par conséquent, l'estimation simultanée du réseau des individus sains et des patients malades permettrait de prendre en compte l'information commune entre les différents états.

Dans les méthodes présentées ci-dessus, seule Dondelinger *et al.* [2012] permet de prendre en compte cette problématique grâce à un réseau dynamique bayésien, dans le cadre de réseaux de taille limitée. Aussi, nous avons proposé une nouvelle méthode dans Vallat *et al.* [2013] permettant l'estimation simultanée ou différentielle de larges réseaux de gènes issus de multiples états à partir de données de puces à ADN collectées dans Vallat *et al.* [2007].

6.1.3 Modélisation mathématique

L'article publié dans PNAS, Vallat *et al.* [2013], porte la double mention biologie et mathématiques car il présente conjointement une nouvelle modélisation mathématique et son application à des données de biologie. Le besoin était de développer une modélisation adaptée au contexte biologique qui est celui de l'étude de la propagation en cascade d'un signal dans un réseau de gènes suite à une stimulation pulsée. Cette méthode se base sur un modèle de régression linéaire estimée à l'aide du *lasso*, Tibshirani [1996] et précédé d'une étape de classification, initialement réalisée avec un modèle de mélange de lois de Laplace ajusté par l'algorithme EM puis à l'aide du package `limma`, Ritchie *et al.* [2015], pour la détection de gènes différentiellement exprimés.

Un enrichissement à l'aide de gènes possédant des motifs attendus dans de tels réseaux en cascade est également effectué pour les premiers temps. En effet, les premiers premiers gènes activés, aussi appelés facteurs transcriptionnels, ont souvent des niveaux d'expression plus faibles que des gènes qui s'expriment à des temps plus tardifs.

Ainsi la méthode se décompose, comme montré dans la Figure 6.2, en plusieurs étapes.

Commençons par fixer quelques notations en rappelant quelques propriétés de l'estimateur *lasso* du modèle linéaire.

Estimation par le *lasso*

Soient les couples $(\boldsymbol{x_{i.}},y_i)_{i=1,\cdots,N}$. Les $\boldsymbol{x_{i.}} = (x_{i1},\cdots,x_{ip})^T$ sont les valeurs des prédicteurs alors que les y_i sont celles de la réponse. Le modèle de régression linéaire est :

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \eta_i, \tag{6.1}$$

où les bruits η_i sont indépendantes et identiquement distribuées suivant une loi admettant une variance.

Supposons de plus que les prédicteurs soient centrés et réduits et que la réponse soit centrée. L'estimateur *lasso* est alors donné par :

$$\hat{\boldsymbol{\beta}}^L(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \|\boldsymbol{\beta}\|_1 \right], \tag{6.2}$$

où λ est un scalaire positif qui fixe le niveau de contrainte voulu par l'utilisateurs. Remarquons que :

- Lorsque $\lambda = 0$, $\hat{\boldsymbol{\beta}}^L$ est l'estimateur ordinaire de Gauss-Markov.
- Lorsque $\lambda = +\infty$, nous avons $\hat{\boldsymbol{\beta}}^L = \mathbf{0}_p$.

L'estimateur *lasso* pour la régression linéaire a deux avantages principaux :

1. il permet d'ajuster des modèles à des jeux de données *a priori* pathologiques où le nombre d'observations est strictement inférieur au nombre de variables,
2. il effectue de la sélection de variable : en fonction de la valeur de λ, le vecteur $\hat{\boldsymbol{\beta}}^L(\lambda)$ sera plus ou moins parcimonieux, c'est-à-dire possèdera plus ou moins de coordonnées nulles.

L'estimateur *lasso* pour le modèle de régression linéaire peut aussi être écrit de la manière suivante :

$$\hat{\boldsymbol{\beta}}^L(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p \quad \|\boldsymbol{\beta}\|_1 \leqslant \tilde{\lambda}}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right]. \tag{6.3}$$

Ces deux formulations (équation (6.2) qui est la formulation pénalisée et équation (6.3) qui est la formulation sous contraintes) sont équivalentes au sens où pour chaque valeur positive de λ, il existe une valeur positive de $\tilde{\lambda}$ amenant à la même solution.

Modèle mathématique de décodage de réseaux

Considérons N gènes observés sur P individus à T temps ; notons x_{npt} l'expression du gène n sur l'individu p au temps t. Puisque chaque gène sera considéré exactement une seule fois en tant que variable réponse, notre modèle comporte N modèles de régression linéaire. Pour pouvoir tenir compte du fait que l'action d'un gène sur un autre n'est pas instantanée, nous définissons :

$$\tilde{\boldsymbol{x}}_{np.} = \begin{pmatrix} x_{npt_2} \\ \vdots \\ x_{npt_T} \end{pmatrix} \quad \text{and} \quad \check{\boldsymbol{x}}_{np.} = \begin{pmatrix} x_{npt_1} \\ \vdots \\ x_{npt_{T-1}} \end{pmatrix},$$

Notons que le vecteur $\tilde{\boldsymbol{x}}_{np.}$ comporte les temps allant de t_2 jusqu'au temps t_T, alors que le vecteur $\check{\boldsymbol{x}}_{np.}$ comporte les temps allant de t_1 jusqu'au temps t_{T-1}. Dans la suite, lorsqu'un gène n joue le rôle de la variable réponse dans le modèle nous utiliserons le vecteur $\tilde{\boldsymbol{x}}_{np.}$ pour le représenter, par contre, lorsqu'un gène joue le rôle d'un prédicteur dans le modèle, c'est le vecteur $\check{\boldsymbol{x}}_{np.}$ que nous utiliserons pour le représenter.

Chacun des T groupes est associé de manière bijective à l'un des T temps d'observation. Nous supposons de plus avoir pu associer chacun des gènes à l'un des T groupes temporels de telle sorte que le groupe associé au temps t, $1 \leqslant t \leqslant T$, contienne les gènes qui s'activent au temps t, c'est-à-dire les gènes dont l'expression différentielle par rapport à la stimulation de contrôle « s'éloigne trop » de 0. Cette éloignement est généralement quantifié par un test avec un niveau de significativité donné ou sur la base d'un rang calculé à partir du profil d'expression du gène.

Nous pouvons alors utiliser le modèle de régression linéaire suivant :

$$\tilde{\boldsymbol{x}}_{np.} = \sum_{n'=1}^{N} \omega_{n'n} \boldsymbol{F}_{m(n')m(n)} \check{\boldsymbol{x}}_{n'p.} + \boldsymbol{\varepsilon}_{np},$$

où :

- $n \mapsto m(n)$ est la fonction qui à un gène n associe son groupe temporel,
- $\boldsymbol{F}_{m(n')m(n)}$ est la matrice carrée d'ordre $T-1$ qui décrit l'action des gènes du groupe $m(n')$ sur les gènes du groupe $m(n)$,

- le terme général $\omega_{n'n}$ de la matrice $\boldsymbol{\omega}$ est la force de l'effet, positif ou négatif, du gène n' sur le gène n,
- $\boldsymbol{\varepsilon}_{np}$, $1 \leqslant p \leqslant P$, est un vecteur de bruit de dimension $T-1$.

Nous choisissons un estimateur *lasso* de ces modèles de régression linéaire :

$$(\hat{\boldsymbol{\omega}},(\hat{\boldsymbol{F}}_{a,b})_{1\leqslant a,b\leqslant T}) = \underset{\substack{\omega_{n'n}\in\mathbb{R},\ 1\leqslant n',n\leqslant N \\ \boldsymbol{F}_{ab}\in\mathcal{M}_{T-1}(\mathbb{R}),1\leqslant a,b\leqslant T}}{\operatorname{argmin}} \left[\sum_{n=1}^{N}\sum_{p=1}^{P}\left\|\tilde{\boldsymbol{x}}_{np.} - \sum_{n'=1}^{N}\boldsymbol{F}_{m(n')m(n)}\omega_{n'n}\check{\boldsymbol{x}}_{n'p.}\right\|_2^2\right]$$

avec la contrainte

$$\forall n = 1,...,N, \quad \sum_{n'=1}^{N}\omega_{n'n} \leqslant \lambda_n.$$

Ainsi, n est le gène régulé et $\{n',\ n' = 1,\ \ldots,\ N\}$ est l'ensemble des gènes potentiellement régulateurs de n. Remarquons que c'est la matrice $F_{m(n')m(n)}$ qui capture l'évolution au cours du temps du lien entre les gènes n' et n. Afin de pouvoir respecter la causalité temporelle de la cascade, nous avons ajouté les contraintes temporelles suivantes :

1. $m(n') \geqslant m(n) \Rightarrow \boldsymbol{F}_{m(n')m(n)} = 0$: ceci assure qu'un gène appartenant au groupe temporel du temps t_k a une influence sur un gène du groupe temporel $t_{k'}$ si et seulement si $k < k'$,
2. les matrices $(\boldsymbol{F}_{a,b})_{1\leqslant a,b\leqslant T}$ sont toutes triangulaires inférieures : ceci assure que l'expression d'un gène au temps t_k peut avoir un effet sur un autre gène au temps $t_{k'}$ si et seulement si $k < k'$.

Mêmes lorsqu'elles sont nulles, les sous-diagonales et la diagonale des matrices $(\boldsymbol{F}_{a,b})_{1\leqslant a,b\leqslant T}$ sont choisies constantes, voir figure 6.3. Par conséquent, l'effet des expressions d'un des gènes sur celles d'un autre gène ne dépend que de la différence entre les indices des temps de mesures et non de la différence réelle entre ces temps de mesure.

Ce problème d'estimation est résolu par une approche de type *coordinate ascent* en supposant, à tout de rôle, les matrices $(\boldsymbol{F}_{a,b})_{1\leqslant a,b\leqslant T}$ connues ou la matrice $\boldsymbol{\omega}$ connue. Le résultat de l'optimisation est un réseau de liens donnés par les éléments non-nuls $\hat{\omega}_{n'n}(obs)$ de $\hat{\boldsymbol{\omega}}(obs)$ ainsi que l'ensemble des matrices des interactions

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} a_{ij}^F & 0 & 0 \\ b_{ij}^F & a_{ij}^F & 0 \\ c_{ij}^F & b_{ij}^F & a_{ij}^F \end{pmatrix} \end{array}$$

Figure 6.3 : *Cellule* $\boldsymbol{F}_{ij}$

entre groupes, dépendant du temps, donné par l'ensemble des $(\hat{\boldsymbol{F}}_{a,b}(obs))_{1 \leqslant a,b \leqslant T}$.

Si les groupes sont assez homogènes, l'inférence des matrices $\boldsymbol{F}_{m(n')m(n)}$ ne dépend pas des gènes actifs (c'est-à-dire ceux pour lesquels $\omega_{n'n} \neq 0$). C'est pourquoi un algorithme non itératif est également proposé dans lequel l'estimation des matrices $(\boldsymbol{F}_{a,b})_{1 \leqslant a,b \leqslant T}$ précède celle de la matrice $\boldsymbol{\omega}$.

Pour obtenir un résultat plus robuste, à chaque pas, l'estimation des matrices $(\boldsymbol{F}_{a,b})_{1 \leqslant a,b \leqslant T}$ est faite de manière répétée selon une validation croisée. De plus, afin de réduire les problèmes de convergence, la nouvelle solution est une combinaison linéaire entre la solution précédente et celle qui vient d'être calculée.

6.1.4 Résultats

Comme nous l'avons indiqué dans Vallat *et al.* [2013], le modèle mathématique proposé avait des résultats supérieurs aux approches généralistes connues à l'époque puisqu'il tirait parti de la structure en cascade des réseaux de gènes étudiés.

Ce modèle a été en mesure de prédire une intervention biologique réelle : nous avons prédit avec le modèle mathématique le résultat d'une intervention (inhibition) sur le gène DUSP1 et les chercheurs en biologie ont, quant à eux, réalisé cette intervention en réalité et observé l'effet qu'elle a produit dans le réseau de régulation. Nous avons alors montré qu'il existait une association significative entre ces résultats prédits et ceux observés.

Il s'agit de la première étape pour une prise de contrôle de ce type de systèmes complexes : comprendre le réseau et prédire l'effet d'une intervention sur l'un ou plusieurs des gènes de celui-ci. Nous nous sommes ensuite intéressés à franchir une seconde étape, celle de l'intervention dirigée, voir section 6.2.

6.2 Intervention dirigée dans un réseau de gènes

6.2.1 Objectif

La prolongation naturelle des résultats obtenus avec la modélisation des réseaux de gènes de la section 6.1 est celle d'intervention dirigée. Ici, au lieu de prédire l'effet d'une intervention sur l'un des gènes du réseau, il faut partir d'un ensemble de gènes du réseau, noté M et appelé ensemble de marqueurs, et trouver l'ensemble des meilleurs candidats, appelé ensemble de cibles d'intervention C, en amont dans la cascade pour pouvoir faire évoluer les expressions des gènes de M dans le sens voulu, c'est-à-dire de la manière prédéfinie par l'expérimentateur.

Si, nous nous intéressons plus particulièrement à la fin de la cascade de gènes, c'est parce que ce sont ces gènes qui sont impliqués dans la création des protéines qui elles-mêmes détermineront la réponse de la cellule. Ainsi nous espérons pouvoir changer la réponse de la cellule en agissant sur les gènes au début de la cascade de gènes.

6.2.2 Modélisation

Nous disposions des expressions des gènes chez deux groupes de patients, sains et malades indolents. Dans un premier temps et avec l'aide des médecins et chercheurs en biologie, nous avons recherché les meilleurs marqueurs, aux temps tardifs de la cascade, de la différence de réponse cellulaire entre les patients malades indolents et les patients malades agressifs. Ayant déterminé ces marqueurs entre les deux groupes, nous avons recherché, au sein de la cascade de gènes estimée dans le groupe des malades agressifs, les cibles d'intervention C. L'expression d'une cible d'intervention doit elle aussi être différente entre les groupes de patients indolents et agressifs afin de pouvoir espérer qu'une modification de celle-ci aura l'effet attendu.

La complexité mathématique dans cette situation provient d'une tendance bien connue du *lasso* à choisir les prédicteurs au hasard lorsqu'il est utilisé sur des jeux de données avec des variables corrélées. Or c'était typiquement la situation qui s'offrait à nous compte tenu du faible nombre de mesures disponibles (24) par rapport aux très grand nombre de gènes mesurés (54675 sondes). Jusqu'à présent cette difficulté n'avait pas été cruciale car nous avions pu choisir le gène sur lequel intervenir directement et avions pris un gène qui était bien représenté dans le réseau.

6.2.3 Confiance

Afin de répondre à cette nouvelle problématique, nous avons donc commencé par chercher à intégrer, dès 2013, une notion de confiance lors de la sélection des régulateurs d'un gène, c'est-à-dire des variables sélectionnées par le *lasso*. Pour cela, nous avons considéré plusieurs solutions existantes comme la régression pénalisée *ridge*, Hoerl et Kennard [1970], ou *elasticnet*, Zou et Hastie [2005] mais également la régression PLS parcimonieuse, *sparse pls* Chun et Keleş [2010], l'approche *stability selection*, Meinshausen et Bühlmann [2010] ainsi que développé notre propre méthode, elle spécialement conçue pour tenir compte de l'inévitable corrélation entre les variables grâce à des techniques de réechantillonnage, *selectboost*, Jung *et al.* [2015] puis étendue dans Aouadi *et al.* [2018], plus de détails à la section 4.7. L'objectif était de remplacer l'étape faisant appel à l'estimateur *lasso* des modèles linéaires de régulation des gènes par un estimateur qui permettrait de calculer un niveau de confiance pour chacun des liens entre les gènes pour pouvoir choisir les liens les plus sûrs entre eux.

Une sélection de marqueurs et de cibles a été proposée aux médecins et aux chercheurs en biologie. Ceux-ci en ont extrait des candidats expérimentaux en intégrant de surcroît des contraintes techniques liées à la mise en œuvre pratique des inhibitions des gènes cibles. De même un choix entre les différents marqueurs qui étaient associés à ces gènes cibles a été fait en fonction de leur sens biologique. D'autres gênes qui ne devaient pas être influencés par l'intervention sur ces marqueurs, une inhibition dans notre cas, ont également été sélectionnés comme contrôles.

6.2.4 Validation biologique

Nous avons rencontré des difficultés lors de la validation biologique des résultats du modèle mathématique. Ces difficultés sont indépendantes du modèle en lui-même mais proviennent du fait que les chercheurs en biologie ont surestimé leur capacité à inhiber la cible qu'ils ont retenue. Ainsi au lieu de faire disparaître complètement l'expression du gène, celle n'a été diminuée que de 10 à 20% dans le meilleur des cas, rendant impossible l'observation d'un effet sur les marqueurs en aval dans la cascade. De plus, l'utilisation de cellules « réelles » de patients a aussi considérablement rallongé le processus de collecte des résultats et ainsi plus d'une année a été nécessaire pour pouvoir détecter ce problème.

Dans un second temps, les échantillons collectés ont été analysés une nouvelle fois mais avec une autre technique, très récente, pour mesurer l'expression des gènes : les ampliseq, Li *et al.* [2015]. S'agissant d'une nouvelle technique, j'ai procédé moi-même à une recherche bibliographique pour déterminer quelle était la méthodologie à employer pour étudier convenablement ces données, puis com-

mencé des analyses préliminaires multivariées, Bertrand [2016a], en s'appuyant sur Law *et al.* [2014]. Cette nouvelle méthode de mesure a confirmé l'absence d'effet de l'inhibition sur le gène ciblé. En conséquence le travail de modélisation mathématique, qui est pourtant achevé depuis plusieurs années, n'a pas pu être valorisé, pour l'instant, à sa juste valeur puisque les expériences de validation biologique n'ont pas pu être réalisées en accord avec ce qui avait été prévu lors de la conception du modèle mathématique. Pire encore, l'absence complète d'effet de l'inhibition de la cible nous empêche même de mettre à jour le modèle initial d'intervention dirigée pour tenir compte d'un effet diminué de l'inhibition.

6.3 Réseau multi-omiques 1 : gènes et protéines sur des individus différents

6.3.1 Contexte

Comme nous l'avons vu à la section 6.1.2, le dogme biologique, Crick [1970], indique que ce sont les protéines qui génèrent la réponse cellulaire et le phénotype observé. Notre travail jusqu'à maintenant s'est concentré sur l'étude d'un jeu de données d'expressions de gènes et il est nécessaire d'essayer de relier la modélisation des expressions des gènes à celles des abondances protéiques puisque nous souhaitons à terme pouvoir influer sur la réponse cellulaire.

Le point de départ dont nous disposions est un jeu de données d'expressions temporelles de protéines, Perrot *et al.* [2011], qui n'avait pas encore été exploité dans le cadre d'une inférence statistique. Il a été mesuré selon le même protocole que celui comportant les expressions de gènes, Vallat *et al.* [2007], mais sur trois patients différents. Sur la base de la similarité de critères médicaux en comparaison avec ces trois patients, en premier lieu le stade de la maladie, nous avons sélectionné cinq patients du jeu de données d'expressions de gènes qui avait été utilisé dans notre modèle cellulaire, Vallat *et al.* [2013]. D'après le biologiste, le fait d'utiliser des mesures collectées sur des individus différents donnera plus de force aux similarités trouvées entre les deux types de données.

6.3.2 Problème méthodologique

Le problème méthodologique consiste donc à réussir à associer les variations d'expression des gènes à celles des protéines alors que ces deux jeux de données n'ont pas été observées sur les mêmes individus. L'objectif est d'identifier le cœur du programme génique associé à la prolifération des cellules cancéreuses en combinant les données temporelles des gènes et des protéines. Compte tenu du faible nombre d'observations (24 observations : 3 sujets, 4 temps, 2 conditions), et du

petit nombre de protéines observées (2046) puis sélectionnées et identifiées avec succès (147) par rapport au nombre de gènes (19884), il est souhaitable de vouloir intégrer de l'information biologique externe accessible dans des bases de données. Ce déséquilibre entre le nombre de gènes et le nombre de protéines identifiées vient de la technique de protéomique utilisée, 2D-DIGE, qui non seulement ne permet que l'observation d'une petite fraction des protéines présentes dans l'échantillon mais nécessite également une seconde étape d'identification manuelle de la composition de chacun des spots polypeptidiques obtenus, cette identification pouvant échouer ou être rendue très difficilement exploitable si plusieurs protéines sont détectées au sein du même spot polypeptidique.

6.3.3 Modélisation

J'ai mené cette recherche seul et ai procédé en plusieurs étapes.

- Sélectionner les *spots* polypeptidiques à identifier sur la base des différences d'expression au cours du temps ou de leurs profils et demander au chercheur en protéomique de les identifier.

- Sélectionner des couples gènes/protéines sur la base de leurs profils d'expression, Bertrand [2015a].

- Utiliser l'information biologique pour sélectionner les couples gènes/protéines vraisemblablement impliqués dans la réponse cellulaire, Bertrand [2015b].

- Regrouper les couples gènes/protéines trouvés à l'étape précédente, Bertrand [2015c].

- Inférer le réseau de régulation des couples gènes/protéines précédents en utilisant la répartition en groupe obtenue et une extension de notre méthode de modélisation des cascades qui permet d'utiliser un nombre quelconque de groupes, dans ce cas plus de groupes que de temps observés dans la cascade. Ce réseau représentera alors le cœur du programme génique de la prolifération qui était recherché.

L'utilisation de la correspondance gènes-protéines est censée améliorer l'inférence du réseau de gènes car il intègrera aussi, dans une certaine mesure, les contraintes révélées par l'analyse du réseau protéique. Si plusieurs auteurs (Vanunu *et al.* [2010], Yang *et al.* [2011]) avaient proposé des techniques pour inférer des réseaux protéiques, personne n'avait pour l'instant construit une méthode d'inférence tirant parti simultanément de la connaissance biologique accumulée, des réseaux protéiques et des réseaux géniques.

Le modèle de puce à ADN (HG Hu 133+2.0) utilisé dans cette étude est constituées de 54675 sondes. Certaines sont redondantes : plusieurs mesurent l'expression du même gêne mais avec une efficacité variable. D'autres sont inutiles. J'ai donc fait un premier tri à l'aide de la méthodologie proposée par Li *et al.* [2011], une des rares sources d'information sur la qualité des sondes existante à l'époque, qui calcule pour chacune d'entre elle un score basé sur la *spécificité*, la *splice isoform coverage* et la *robustness against degradation*, afin de ne retenir que les 19884 meilleures sondes associées à des gènes uniques. Puis j'ai normalisé les données d'expression de gènes avec *dchip*, Li et Wong [2001]. La sélection des gènes et des protéines différentiellement exprimées a été faite à l'aide du package `limma`, Ritchie *et al.* [2015]. Pour les protéines, la création de groupes a été faite avec l'algorithme DIRECT, Fu *et al.* [2013], particulièrement bien adapté à nos données d'expressions temporelles et répétées sur plusieurs sujets. Par contre les gènes étaient trop nombreux (19884) pour pouvoir leur appliquer cet algorithme.

Un couple gène/protéine a été retenu pour l'inférence de réseau si :

- le gène est différentiellement exprimé à l'un des temps ou globalement ;
- la protéine est différentiellement exprimée à l'un des temps ou globalement ;
- le profil d'expression est attendu dans un réseau en cascade (pic à certains temps et 0 sinon : profils T1, T2, T3, T4, T1+T2, T1+T2+OA, T3+T4, T3+T4+OA, T1+T3+T4 et toujours exprimé T1+T2+T3+T4)
- connaissance biologique a priori au travers du package `pwomics`, Wachter et Beißbarth [2015].

Au final, 882 couples ont été retenus pour l'inférence du réseau.

L'utilisation du package `pwomics` a été particulièrement complexe à mettre en œuvre. J'ai créé N_C groupes pour cette sélection à l'aide du package `MFuzz`, Kumar et E Futschik [2007]. Puis j'ai analysé l'enrichissement de ces groupes en terme GO, KEGG ou DO à l'aide du package `clusterprofiler`, Yu *et al.* [2012]. En me basant sur la proportion de gènes différentiellement exprimés aux temps 1, 2, 3 ou 4, j'ai affecté chaque cluster à un temps dans la cascade. Ce temps correspond au premier temps à partir duquel les couples gènes/protéines du groupe peuvent agir sur les couples gènes/protéines d'autres groupes. C'est un élément nécessaire pour pouvoir créer les matrices $(\boldsymbol{F}_{a,b})_{1\leqslant a,b\leqslant N_C}$ appropriées, voir section 6.1.3, à utiliser pour l'inférence du réseau. Celle-ci a été réalisée avec le package `Patterns`, voir section 6.6.2.

L'analyse statistique est terminée et un article, dont la nouveauté principale est justement cette méthodologie puisque les données utilisées avaient déjà été publiées séparément, a été écrit avec les chercheurs en biologie, Bertrand *et al.* [2018d]. Il est entré dans sa phase de relecture finale et devrait être soumis prochainement.

6.4 Réseau multi-omiques 2 : gènes et protéines sur les mêmes individus

6.4.1 Introduction

Contexte

À nouveau, j'ai mené les aspects mathématiques de ces recherches seul. Elles constituent le point d'orgue du projet transdisciplinaire mathématiques – médecine – biologie et s'appuient sur deux jeux de données qui ont pu être collectés grâce à des financements multiples ($\geqslant$ 350k€) obtenus suite à plusieurs appels à projet réussis et dans lesquels j'avais rôle de premier plan. Le caractère novateur de la recherche en matière de modélisation statistique a été particulièrement appréciée lors de l'évaluation de ces projets.

Un doctorant en biologie a pu être recruté dans le but exclusif de récolter ces données et il a été aidé au moins par trois autres doctorants ainsi qu'un biologiste.

Les données protéomiques ont cette fois-ci été mesurées par le laboratoire de Spectrométrie de Masse BioOrganique (LSMBO) et ce travail a marqué le début d'une collaboration fructueuse avec ce laboratoire, voir section 6.5.

À la différence près des temps d'observation, le même plan d'expérience a été utilisé pour la collecte des expressions des gènes et celle des abondances des protéines. En effet, si la première mesure en T0, qui servira de contrôle, des gènes et des protéines a été réalisée simultanément, un décalage de 30 minutes a été imposé entre la collecte des mesures des gènes et celle des protéines afin de laisser le temps à la machinerie cellulaire de traduire l'information contenue dans les expressions des gènes en protéines. Malheureusement, rien ne garantit que ce laps de temps soit suffisant ou ne soit déjà trop long, pire encore la cinétique de transformation dépend du couple gène et protéine considéré. C'est l'avis d'expert des chercheurs en biologie qui a fixé ce délai, qui compte tenu de la technique de mesure utilisée, doit être le même pour chaque couple gène et protéine. Le plan de l'expérience est reproduit sur la figure 6.4 et comporte six sujets mesurés à neuf temps différents, ce qui en faisait l'une des plus longues série d'observation existante.

Le contexte est le même qu'à la section 6.3 à la différence que les expressions des gènes et les abondances protéiques ont été observées sur les mêmes individus, ce qui permettra une modélisation conjointe gène/protéine jusqu'au niveau de l'individu. Toutefois, à la différence du plan d'expérience utilisé dans la section 6.3, les contrôles ne sont pas répétés. Comme dans toutes les expériences ayant un coût

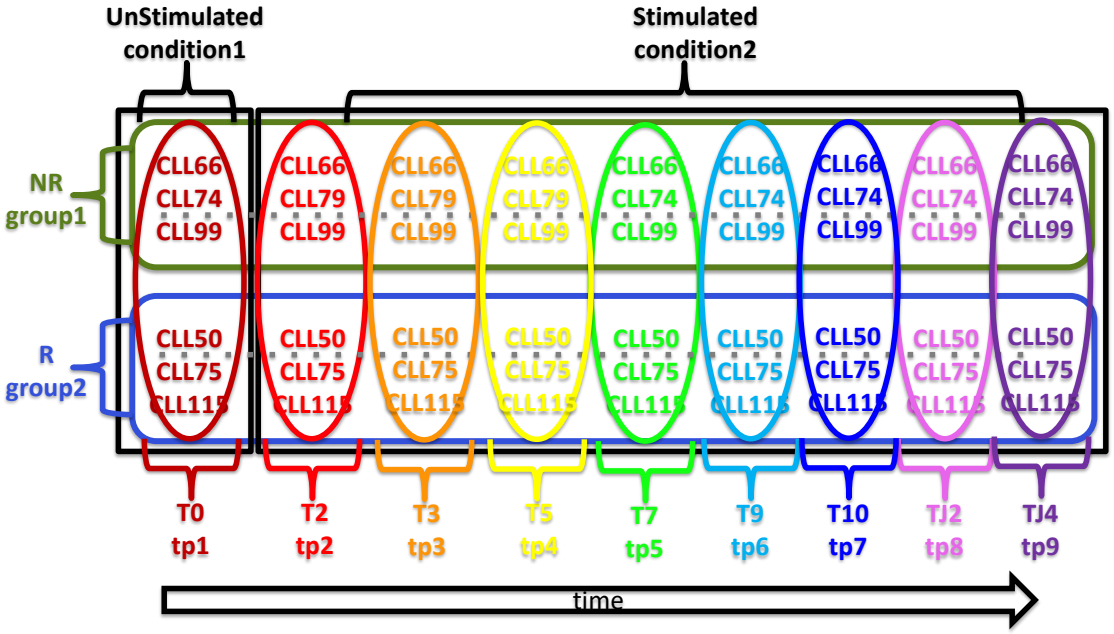


Figure 6.4 : Plan de l'expérience.

élevé, près de 200k€, il faut faire des choix dans la planification de l'expérience. Ici, la non répétition des contrôles, c'est-à-dire utiliser toujours comme contrôle la mesure non stimulée à T0, permet de presque doubler les temps de mesure et de s'intéresser aux temps ultérieurs, par rapport à ceux utilisés aux sections 6.1, 6.2 et 6.3, de la cascade d'expressions des gènes et des protéines. C'est d'un intérêt primordial car le comportement pathologique des cellules, la prolifération cellulaire, n'apparaît que deux jours après la stimulation. Ainsi pour déterminer les protéines puis les gènes directement responsables de cette prolifération, la durée d'observation devait être allongée. Nous avons donc analysé à nouveau les données existantes des sections 6.1 et 6.3 en utilisant comme contrôle toujours le même échantillon : la mesure à T0. La différence entre ces résultats et ceux que nous avions obtenus aux sections 6.1 et 6.3 était suffisamment limité pour que nous décidions de ne plus répéter les contrôles.

Du côté biologique, une étude a été menée, dont j'ai dépouillé les résultats, pour améliorer le processus de stimulation des cellules. Un article a été écrit, Schleiss *et al.* [2018a], mais est en attente de publication.

Problèmes méthodologiques

Le problème méthodologique était de loin le plus complet parmi ceux auxquels j'ai été confronté lors de cette collaboration. À ma connaissance, une modélisation au niveau individuel de l'évolution temporelle des expressions et abondances des couples gènes et protéines n'avait pas été encore réalisée. Plusieurs verrous de modélisation devaient être levés pour y parvenir et une recherche bibliographique montrait l'absence ou la faiblesse d'outils appropriés pour les points, énumérés ci-après, pour lesquels j'ai proposé des solutions de modélisation adéquates innovantes. :

1. Inférer les abondances protéiques.
2. Choix des acteurs à modéliser parmi les 23442 gènes et les 4664 protéines mesurés.
3. Introduire deux types de mesures de nature complètement différente dans le même modèle pour traduire les actions des gènes sur les protéines et des protéines sur les gènes sachant que, de surcroît, pour un grand nombre de gènes, la valeur de la protéine associée n'a pas été mesurée.
4. Recueillir et utiliser de l'information biologique adéquate pour pondérer l'inférence.
5. Retenir dans le réseau les liens entre gènes et protéines qui sont les plus sûrs.

6. Inférer un réseau de gènes qui respecte la structure en cascade due à la stimulation mais qui tient également compte du fait que les temps de mesure sont irrégulièrement espacés et que le décalage entre les mesures des gènes et des protéines n'est pas constant, voir table 6.1.

Ces recherches ont requis la réalisation d'une étude pilote ainsi que la rédaction de plusieurs rapports intermédiaires, Bertrand [2016b], Bertrand [2016c], Bertrand [2017a] et Bertrand [2017b], et sont désormais terminées : les résultats ont été remis aux médecins et chercheurs en biologie pour interprétation. Un article reposant pleinement sur mes recherches et cette nouvelle modélisation statistique est en cours d'écriture Schleiss *et al.* [2018b].

6.4.2 Développements méthodologiques

Inférence des abondances protéiques

Notons que, par rapport à la mesure des expressions des gènes, l'inférence des abondances des protéines comporte un degré de complexité supplémentaire puisqu'il s'agit d'une mesure indirecte passant par celle de l'intensité des peptides obtenus en digérant les protéines à l'aide d'une enzyme. D'un point de vue statistique, ces mesures de peptides présentent des problèmes spécifiques : valeurs manquantes, forte corrélation, valeurs atypiques, Goeminne *et al.* [2015], Goeminne *et al.* [2016].

Une recherche bibliographique et des tests préliminaires m'ont amené à choisir, voir la section 6.5, la solution basée sur une approche *ridge* robuste implémentée dans le package `MSqRob`, Goeminne *et al.* [2017], couplée à un calcul du *FDR* qui tienne compte du caractère spécifique des distributions des p-valeurs des tests, Giai Gianetto *et al.* [2016] implémenté dans le package `cp4p`. La problématique de la gestion des valeurs manquantes ou de l'optimisation de l'utilisation des résultats expérimentaux connus m'a amené à proposer, en collaboration avec Christine Carapito chercheuse en chimie, un sujet de doctorat qui a été financé par le LabEx IRMIA, voir la section 7.4.

	tp1	tp2	tp3	tp4	tp5	tp6	tp7	tp8	tp9
	T0	T2	T3	T5	T7	T9	T10	TJ2	TJ4
gènes	0h	1h	1h30	3h30	6h30	12h	24h	48h	96h
ΔTgènes		1h	0h30	2h	3h	5h30	12h	24h	48h
protéines	0h	1h	2h	4h	7h	12h30	24h30	48h30	96h30
ΔTprotéines		1h	1h	2h	3h	5h30	12h	24h	48h
$T_P - T_G$	0h	0h	0h30	0h30	0h30	0h30	0h30	0h30	0h30

Table 6.1 : *Tableau récapitulatif des temps de mesure.*

Choix des acteurs à modéliser

La série de mesure a été réalisé en deux temps et j'ai commencé par m'intéresser aux résultats d'une étude pilote comprenant

- cinq temps pour les gènes ;
  - sans stimulation (T0, 0h) ;
  - 4 temps après stimulation (identiques à ceux des sections 6.1, 6.2 et 6.3) : 1h, 1h30, 3h30 et 6h30.
- neuf temps pour les protéines ;
  - sans stimulation (T0, 0h) ;
  - 8 temps après stimulation (qui comprennent ceux des sections 6.1, 6.2 et 6.3) : 1h, 2h, 4h, 7h, 12h30, 24h30, J2+30min, J4+30min.

Des techniques de positionnement multidimensionnel (*multidimensional scaling*), Cox et Cox [2000], ont montré que, globalement, les expressions des gènes et les abondances des protéines permettaient de mettre en évidence une séparation aussi bien entre les deux groupes de patients que liée au régime transitoire dû à la propagation du signal au cours du temps. L'analyse des expressions différentielles a également montré l'activation de cascade dans les deux groupes ainsi que des différences entre ces deux groupes. La série complète de mesures a été alors générée et à nouveau des techniques de *multidimensional scaling* ont montré que, globalement, les expressions des gènes et les abondances de protéines permettaient une séparation non seulement des groupes de patients et mais aussi au cours du temps.

L'analyse des expressions différentielles des gènes et des protéines a été réalisée avec le package `edgeR`, Robinson *et al.* [2010] et McCarthy *et al.* [2012], après une étape d'*independent filtering* comme recommandé dans Bourgon *et al.* [2010] afin d'améliorer le FDR et mise en œuvre grâce à la technique `HTSFilter`, spécifiquement développée pour les données de RNAseq, proposée dans Rau *et al.* [2013] . Compte tenu du plan de l'expérience, nombre répété de temps de mesures, étude au sein des patients répondants, des patients non-répondants et en différentiel entre les deux groupes de patients, des corrections additionnelles pour les tests multiples ont été mises en œuvre.

Un ensemble de gènes et de protéines a été sélectionné sur cette base puis enrichi sur la base des profils d'expression des gènes et des protéines. Au final, ce sont 5733 gènes (dont 5722 observés donc pour lesquels l'expression est connue) et

2548 groupes protéiques qui ont été retenus.

Plusieurs approches ont été utilisées pour trouver des motifs d'association spécifiques entre les activations ou inhibitions des gènes et celles des protéines. Par exemple, j'ai utilisé l'algorithme *apriori* du *package* `arules`, Hahsler *et al.* [2005], Hahsler *et al.* [2011] et Hahsler *et al.* [2018] pour rechercher des règles d'associations entre les temps auxquels les gènes et les protéines étaient différentiellement exprimées et ainsi remis en évidence le signal cascade au début de l'expérience et le démarrage de la production stable de protéines aux temps plus tardifs.

J'ai créé des groupes pour cette sélection à l'aide du package `MFuzz`, Kumar et E Futschik [2007] et en partant des *log fold change*, la différence des logarithmes des expressions ou des abondances, estimés entre les groupes. Ces groupes ont été créés pour permettre un partitionnement plus simple de la matrice des actions intergroupes que nous utiliserons dans la suite, comme expliqué en 6.4.2. Ainsi, 20 groupes ont été constitués suite à un clustering des 5722 gènes sur la seule base de leurs expressions puis, pour les 2140 gènes pour lesquels les valeurs des abondances des protéines étaient aussi connues, j'ai raffiné ce clustering. 21 groupes supplémentaires ont été créés sur cette base combinant les expressions des gènes et les abondances des protéines. J'ai dû éliminer 125 abondances protéiques dont les expressions étaient constantes au sein des deux groupes et vérifié que les groupes obtenus présentaient bien les pics et les plateaux attendus compte de la nature de la stimulation biologique employée.

L'étape suivante dans la modélisation est de déterminer, sur la base des profils d'expression dans chaque cluster i, le premier temps d'action possible, C_i, pour le cluster dans le réseau. Ainsi dans la part de la modélisation qui concerne les actions clusters à clusters, un groupe d'acteurs (gènes ou protéines) ne peut avoir une action sur d'autres acteurs (gènes ou protéines) qu'à un temps strictement postérieur à celui de sa propre activation.

Effets des gènes et des protéines

Je suis alors passé à l'élaboration du modèle qui devait non seulement tenir compte de la nature longitudinale des données observées mais aussi du dogme biologique, voir section 6.1.2 et Crick [1970] : un gène X sert à produire la protéine X qui elle peut par contre agir sur d'autres gènes ou protéines. Ainsi au niveau de la modélisation, il convient de privilégier l'abondance de la protéine X pour quantifier l'effet du couple (gène X, protéine X) sur les autres gènes ou protéines. Par contre, si la mesure de la protéine X n'a pas pu être observée, c'est la mesure du gène X qui sera utilisée mais qui est moins fiable car des mécanismes, dits de régulation

post-transcriptionnelle, peuvent intervenir et brouiller la relation entre l'expression du gène X et l'abondance protéique de X. Ainsi, la manière dont les groupes ont été constitués garantit un comportement uniforme au sein d'un groupe et nous pouvons ainsi parler, au niveau d'un groupe entier, d'action de gènes vers protéines (GversP), gènes vers gènes (GversG), protéines vers gènes (PversG) et protéines vers protéines (PversP).

Utilisation de l'information biologique et pondération

Contrairement aux approches que j'avais utilisées précédemment, voir les sections 6.1, 6.2 et 6.3, j'ai intégré des connaissances biologiques pour l'établissement des liens entre les différents acteurs dans le modèle. La difficulté avec l'utilisation de connaissances biologiques est qu'elles ne sont généralement pas obtenues sur les mêmes tissus ni dans les mêmes conditions. Les connaissances tirées de l'observation du fonctionnement de cellules saines, ou souffrant d'autres pathologies, ne peuvent évidemment pas décrire exactement le comportement de nos cellules cancéreuses.

Par conséquent, nous avons cherché à favoriser ou à défavoriser certains liens pour intégrer cette connaissance biologique en réalisant une inférence pondérée. La valeur d'un poids peut varier entre 0, présence systématique du lien, 1, valeur neutre, et $+\infty$, interdiction systématique du lien.

Comme source de connaissance biologique, nous avons utilisé la base de données RegNetwork (341 207 liens, Liu *et al.* [2015]), qui fournit deux types d'information : confiance dans le lien (forte, moyenne, faible) et preuve (observation expérimentale, liaison prédite par un modèle). Les poids que nous en avons déduits ont été modulés par ces deux données.

Ces poids ont également permis d'intégrer d'autres contraintes biologiques :

- Lorsqu'elle est connue, utiliser l'abondance protéique pour l'action de X et non l'expression de X ;
- Pas d'action au sein d'un même groupe : les acteurs d'un groupe (de gènes ou de protéines) ne peuvent être utilisés pour inférer les expressions ou les abondances des autres membres du même groupe. Cette restriction est raisonnable puisque les acteurs d'un même groupe ont les mêmes profils temporels et, par conséquent, sont activés au moment dans la cascade.
- Pas de rétroaction, ou rétrocontrôle, (d'un gène sur lui-même ou d'une protéine sur elle-même).

Structure en cascade

Le cœur du modèle statistique combine une matrice $\boldsymbol{F}$, carrée d'ordre $T * G = 8 \times 62 = 496$, qui décrit l'évolution, en fonction du temps, des actions entre les groupes et une matrice $\boldsymbol{\omega}$, carrée d'ordre $N = 7747$, qui saisit l'intensité et le sens d'un lien entre deux acteurs particuliers. La matrice $\boldsymbol{F}$ permettra de décrire la propagation du signal dans le réseau que la matrice $\boldsymbol{\omega}$ permettra, quant à elle, de reconstituer. En effet j'ai été confronté au double problème d'inférer à la fois la propagation du signal dans le réseau et le réseau lui-même.

Plus précisément, une matrice cellule $\boldsymbol{F}_{ij}$, carrée d'ordre $T = 8$, de la matrice $\boldsymbol{F}$ décrit l'effet de du groupe i sur le groupe j. Il faut ici remarquer que si un acteur n_0 appartient au groupe i et si un acteur n_1 appartient au groupe j, c'est la matrice $\boldsymbol{F}_{ij}$ qui gère la dépendance temporelle de n_1 en fonction de n_0. À nouveau, pour tenir compte de la nature en cascade du signal, j'ai imposé deux contraintes supplémentaires à toutes les matrices $(\boldsymbol{F}_{ij})_{1 \leqslant i,j \leqslant 62}$ qui ne décrivent pas une action d'un cluster de gènes sur un cluster de protéines.

- Un groupe i ne peut pas agir sur un groupe j si $C_i \geqslant C_j$. Ainsi, si $C_i \geqslant C_j$, alors $\boldsymbol{F}_{ij} = 0$. La figure 6.5 montre, en noir, les cellules non-nulles des deux matrices $\boldsymbol{F}_R$ et $\boldsymbol{F}_{NR}$ utilisées pour le décodage du réseau dans le groupe R et le groupe NR. Dans notre application, ces deux matrices sont identiques mais auraient très bien pu différer.

- Si $C_i < C_j$, alors la matrice $\boldsymbol{F}_{ij}$, carrée d'ordre $T = 8$, est triangulaire inférieure stricte sauf dans le cas d'une action GversP. Cette forme signifie que la mesure d'un acteur à l'instant t_k influence la mesure d'un autre acteur au temps t_{k_0} si et seulement si $k < k_0$.

En ce qui concerne, les matrices $\boldsymbol{F}_{ij}$ qui décrivent une action d'un cluster de gènes sur un cluster de protéines, ($\boldsymbol{F}_{ij}$, $21 \leqslant i \leqslant 41$, $42 \leqslant j \leqslant 62$), seuls les matrices ($\boldsymbol{F}_{ij}$, (21,41), ..., ($i$,$i + 21$), ..., (42,62)) sont non-nulles car elles décrivent l'action d'un groupe de gènes sur les protéines associées à ces gènes.

Ces matrices non-nulles ont également une forme particulière, triangulaire inférieure, mais qui n'est pas triangulaire inférieure stricte pour pouvoir traduire l'action, au même temps, d'un gène sur sa protéine.

Cette possibilité d'adapter la forme de la matrice $\boldsymbol{F}$ à la nature des liens biologiques entre les acteurs découle de la manière dont j'ai réparti les acteurs en groupes. C'est d'ailleurs une de mes motivations pour avoir construit cette répartition ainsi.

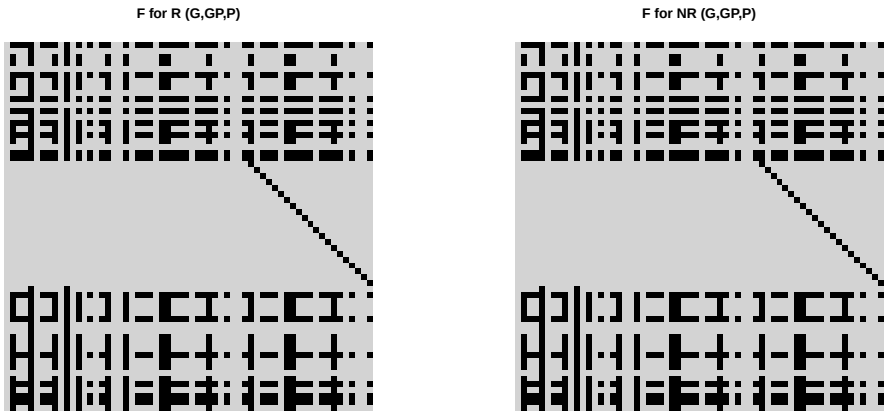


Figure 6.5 : *Matrices* $\boldsymbol{F}$ *pour les groupes R (à gauche) et NR (à droite).*

Mesures irrégulièrement espacées

À nouveau contrairement à ce qui avait été fait avec les modèles utilisés précédemment aux sections 6.1, 6.2 et 6.3, il est nécessaire de prendre en compte la forte irrégularité des temps de mesure (espacés au départ de l'expérience, pour les temps après stimulation, de 30 min et à la fin de celle-ci de 2 jours, voir table 6.1 pour le détail des espacements temporels entre les mesures). Ainsi les sous-diagonales des matrices $\boldsymbol{F}_{ij}$ ne peuvent plus être supposées constantes mais seulement constantes par morceaux pour les mesures pour lesquelles l'écart temporel est d'une durée similaire.

La figure 6.6 représente une matrice $\boldsymbol{F}_{ij}$ dans les cas des liaisons GversG, PversG ou PversP. La figure 6.7 représente une matrice $\boldsymbol{F}_{ij}$ dans le cas de la liaison GversP. Ces matrices permettent de tenir compte des spécificités du plan d'expérience. Par exemple, pour la matrice GversP de la figure 6.7 :

- le délai entre les mesures des gènes et des protéines n'est pas le même pour le temps 1 et les 7 autres temps. Deux coefficients différents ont ainsi été introduits p_{ij}^F et r_{ij}^F.

- l'écart entre la première mesure et la seconde est si court qu'il faut aussi tenir compte de cette valeur pour inférer l'abondance au temps 2, d'où la présence du coefficient q_{ij}^F.

$$
\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array}
\begin{array}{c}
\begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\
\left(\begin{array}{cccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
a_{ij}^F & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
b_{ij}^F & a_{ij}^F & 0 & 0 & 0 & 0 & 0 & 0 \\
c_{ij}^F & b_{ij}^F & a_{ij}^F & 0 & 0 & 0 & 0 & 0 \\
d_{ij}^F & c_{ij}^F & b_{ij}^F & a_{ij}^F & 0 & 0 & 0 & 0 \\
0 & e_{ij}^F & f_{ij}^F & i_{ij}^F & k_{ij}^F & 0 & 0 & 0 \\
0 & 0 & 0 & j_{ij}^F & l_{ij}^F & m_{ij}^F & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & n_{ij}^F & o_{ij}^F & 0
\end{array} \right)
\end{array}
$$

Figure 6.6 : *Cellule* $\boldsymbol{F}_{ij}$ *pour une action GversG, PversG ou PversP*

$$
\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array}
\begin{array}{c}
\begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\
\left(\begin{array}{cccccccc}
p_{ij}^F & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
q_{ij}^F & r_{ij}^F & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & r_{ij}^F & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & r_{ij}^F & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & r_{ij}^F & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & r_{ij}^F & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & r_{ij}^F & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & r_{ij}^F
\end{array} \right)
\end{array}
$$

Figure 6.7 : *Cellule* $\boldsymbol{F}_{ij}$ *pour une action GversP*

Modèle mathématique

Rappelons que $N = 7747$ acteurs ont été mesurés sur $P = 3$ individus et les expressions différentielles ont été calculées pour $T = 8$ temps. Notons x_{npt} la valeur différentielle (calculée à partir des expressions observées du gène ou des abondances protéiques inférées) de l'acteur n mesurée sur l'individu p au temps t et $\tilde{\boldsymbol{x}}_{np\cdot}$ le vecteur de taille $T = 8$ formé des valeurs différentielles observées pour l'acteur n mesurées le sujet p pour les huit temps.

Pour chaque acteur n fixé, choisi parmi les $N = 7747$, le modèle proposé s'écrit :

$$\tilde{\boldsymbol{x}}_{np\cdot} = \sum_{n'=1}^{N} \omega_{n'n} \boldsymbol{F}_{m(n')m(n)} \tilde{\boldsymbol{x}}_{n'p\cdot} + \boldsymbol{\epsilon}_{np}, \qquad 1 \leqslant p \leqslant P.$$

1. N est le nombre total d'acteurs ;
2. $k \mapsto m(k)$ est la fonction qui associe son groupe à un des acteurs ;
3. $\boldsymbol{F}_{ij}$ est une matrice carrée qui décrit l'action des acteurs du groupe i sur ceux du groupe j ;
4. le terme général ω_{kl} de la matrice $\boldsymbol{\omega}$ est la force du lien de l'acteur k vers l'acteur l ;
5. $\boldsymbol{\epsilon}_{np}$, $1 \leqslant p \leqslant P$, est un vecteur aléatoire à T composantes, centré et de variance $\boldsymbol{I_T}$.

Dans ce modèle, n est l'acteur régulé et les n_0, $1 \leqslant n_0 \neq n \leqslant 7747$, sont les régulateurs ($n_0 \neq n$ interdit les autorégulations).

Sélection des liens les plus sûrs, indices de confiance sur les liens

L'ajustement du modèle a été conduit en utilisant une version pondérée de l'algorithme *stability selection*, Meinshausen et Bühlmann [2010], afin de ne retenir que les liens les plus sûrs dans la matrice $\boldsymbol{\omega}$ combinés à des *non negative least squares* pour estimer les paramètres de $\boldsymbol{F}$. Chacune de ces deux matrices est alternativement supposée connue et les estimations itérées jusqu'à convergence. La convergence est d'autant plus rapide que les groupes sont homogènes en termes de profils d'évolution temporelle et ne prend généralement que 1 à 4 itérations.

J'ai également utilisé l'algorithme *selectboost*, voir 4.7 et Aouadi *et al.* [2018], à la place de *stability selection*, qui permet d'obtenir un réseau pour lequel nous disposons d'un indice de confiance pour chacun des liens. Le calcul de ces indices

de confiance en présence de variables corrélées est la raison première de la création de cet algorithme, bien que nous l'ayons présenté dans un cadre plus général dans l'article.

Le résultat obtenu s'interprète de trois manières :

1. un réseau de connectivité avec les éléments non-nuls de $\hat{\boldsymbol{\omega}}$: $\hat{\omega}_{nn'} \neq 0$ signifie qu'une action de n' sur n a été détectée ;

2. chaque matrice $\hat{\boldsymbol{F}}_{ij}$ modélise les actions entre groupes et les temps où elles se produisent ;

3. l'évolution au cours du temps de l'action d'un acteur n sur un acteur n', qui correspond à la propagation du signal au travers du réseau. Elle s'obtient en effectuant le produit $\hat{\boldsymbol{F}}_{m(n')m(n)}\hat{\omega}_{n'n'}$.

Performances du modèle

J'ai procédé à des tests de performance, sensibilité, valeur prédictive positive et F-score, du modèle et en particulier de l'impact d'une bonne, ou d'une mauvaise spécification de la pondération des liens entre les acteurs du réseau.

La comparaison a été effectuée entre l'algorithme basé sur *stability selection*, pour plusieurs paramétrages différents, notre ancien algorithme Cascade, Jung *et al.* [2014c], une version non-pondérée de l'algorithme basé sur *stability selection*, une version non pondérée, pondérée ou mal pondérée de notre algorithme mais basé sur le *lasso*.

Afin de simuler des mesures pour les acteurs du réseau qui soient susceptibles de se rapprocher de celles provenant d'un réseau de régulation, j'ai utilisé un algorithme inspiré de l'approche par attachement préférentiel de Albert et Barabási [2002] et Barabási et Oltvai [2004]. Il a été ensuite adapté aux réseaux temporels emboités (cascades). Les conditions initiales des acteurs dans le réseau ont été définies à l'aide de lois de Laplace comme souvent dans ce contexte.

Les résultats sont représentés sur les Figures 6.8, 6.9, 6.10, 6.11, 6.12 et 6.13. J'ai constaté qu'une pondération correcte améliore grandement les performances de l'algorithme de décodage. De plus, le choix du seuil optimal, en dessous duquel on met à zéro les coefficients $\hat{\boldsymbol{\omega}}$, en est simplifié. Dans notre contexte de simulation les versions pondérées des algorithmes *stability selection* et *selectboost*, non représenté ici dans sa version pondérée, ont obtenu les meilleurs résultats pour les trois critères simultanément.

J'ai également obtenu une estimation du modèle avec le *lasso* non pondéré et comparé les résultats à ceux obtenus sur le jeu de données de Vallat *et al.* [2013].

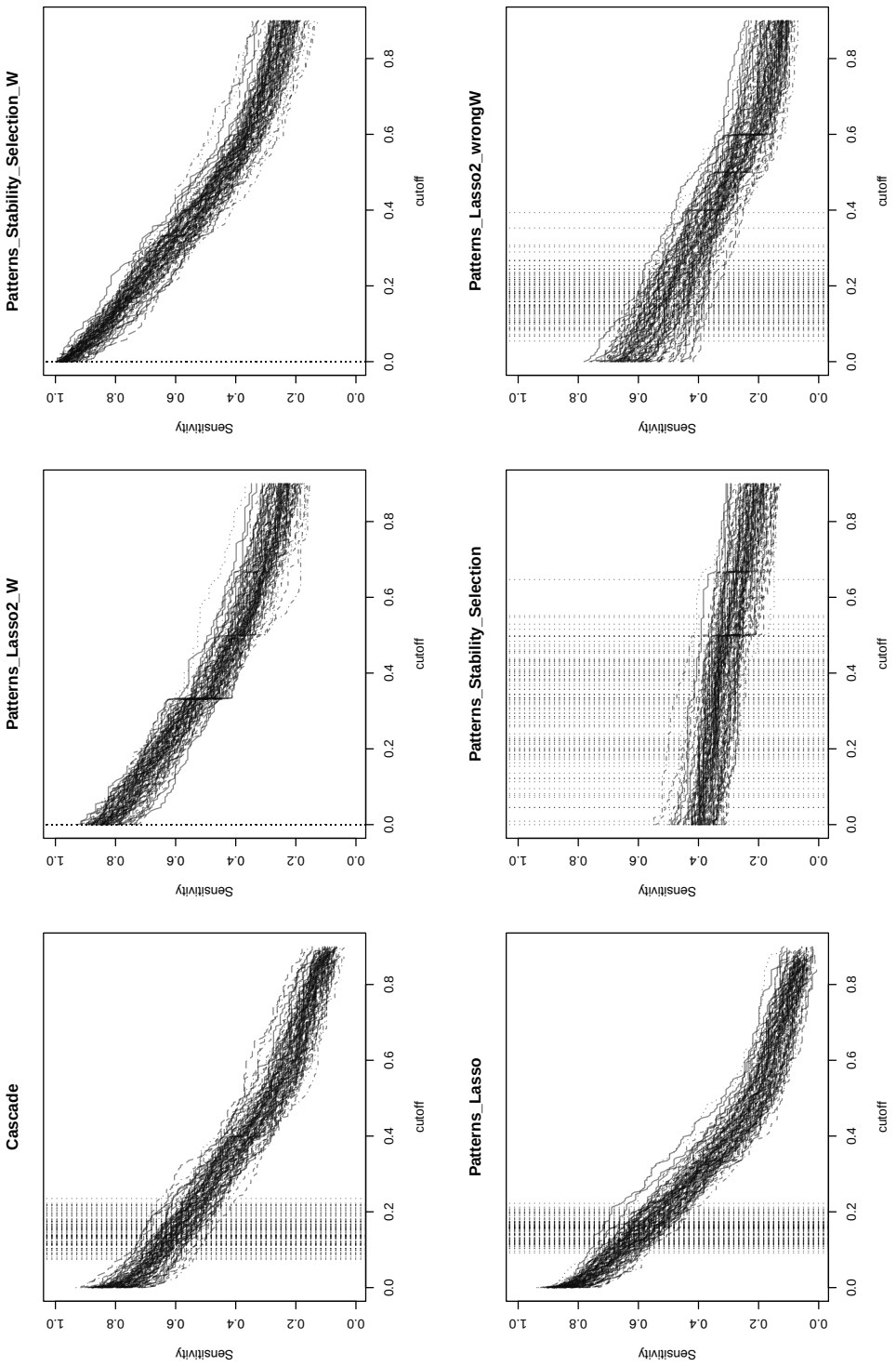


Figure 6.8 : *Sensibilité des méthodes de décodage de réseau.*

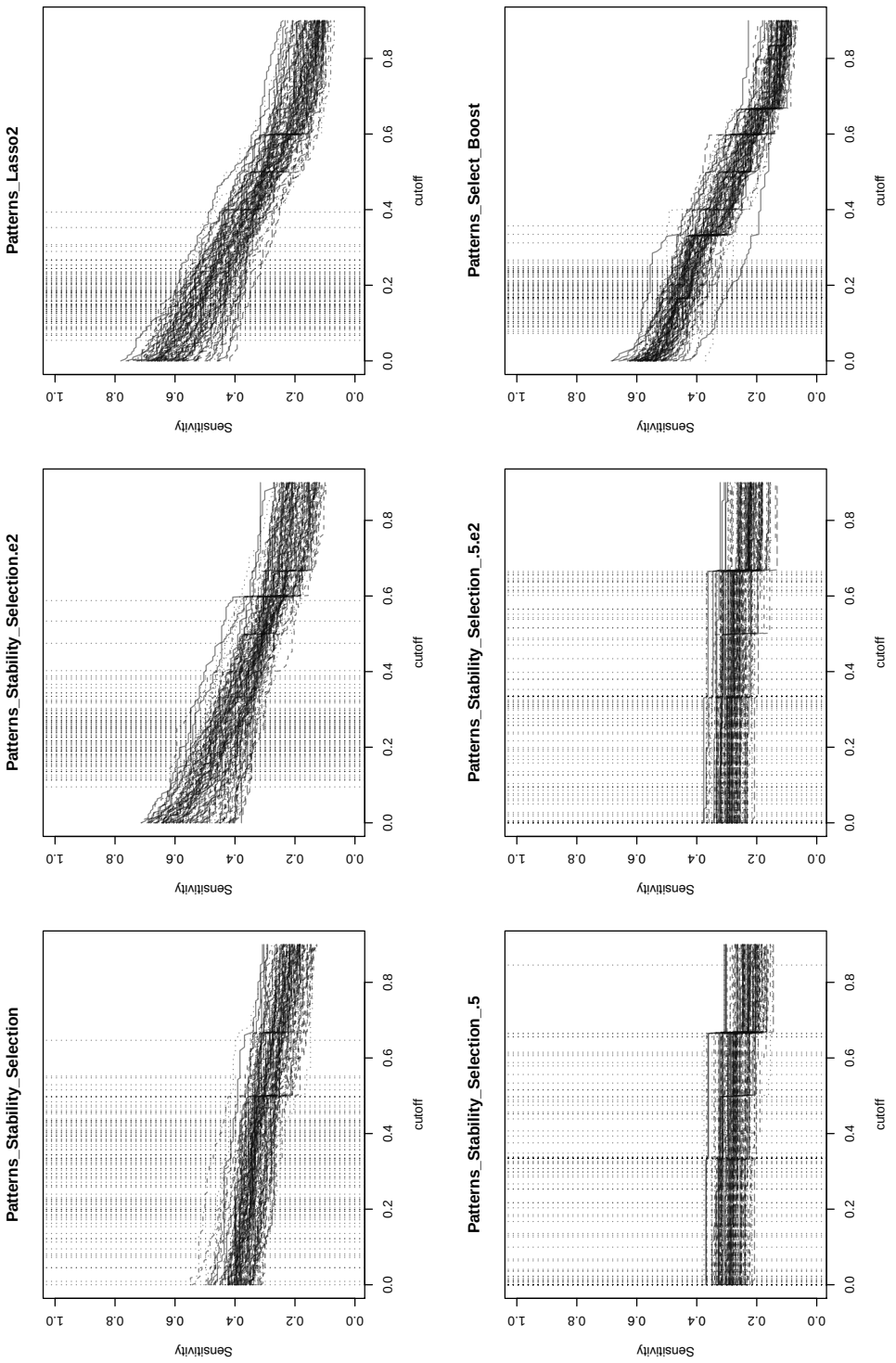


Figure 6.9 : *Sensibilité des méthodes de décodage de réseau.*

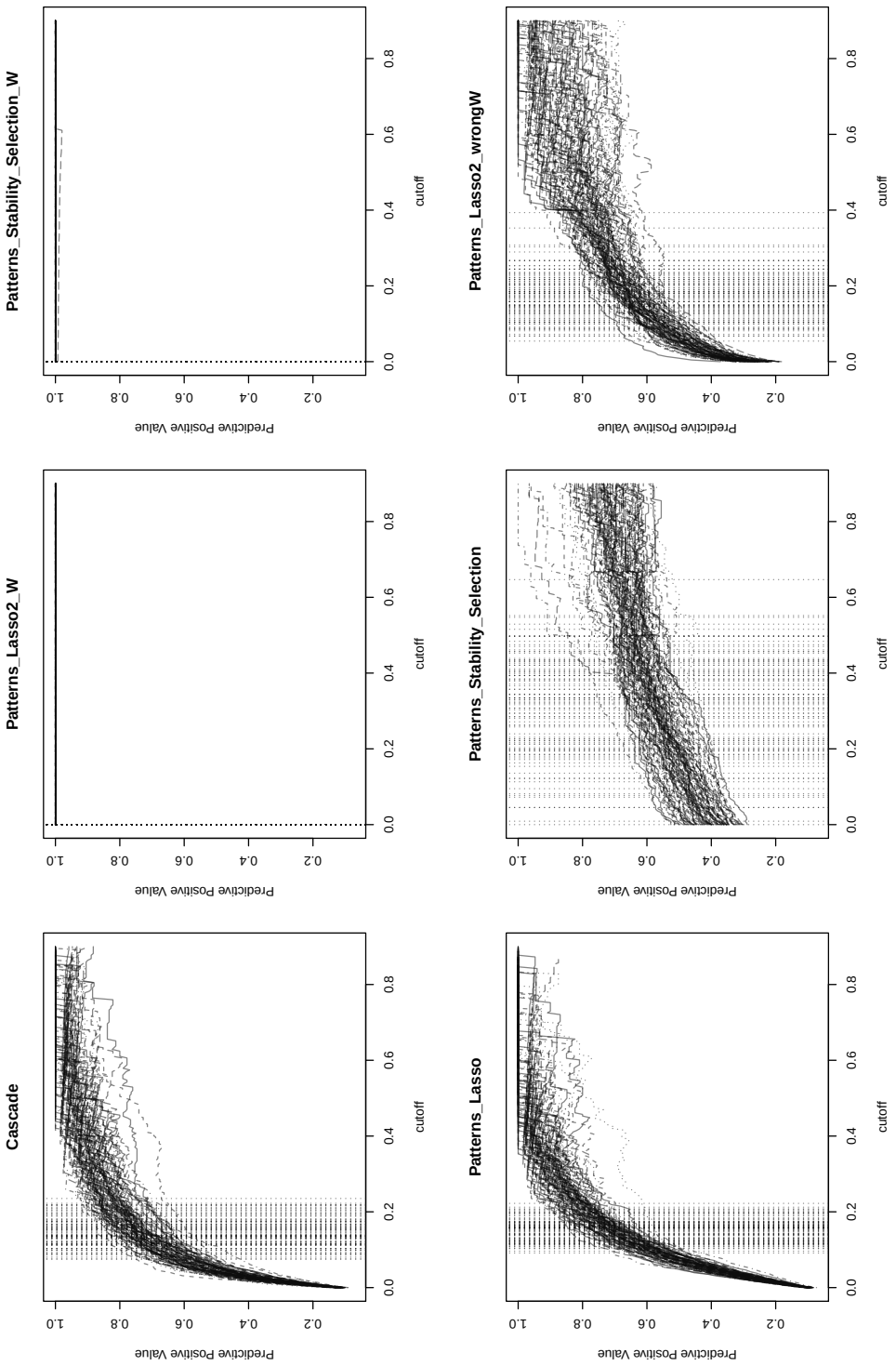


Figure 6.10 : *Valeur prédictive positive des méthodes de décodage de réseau.*

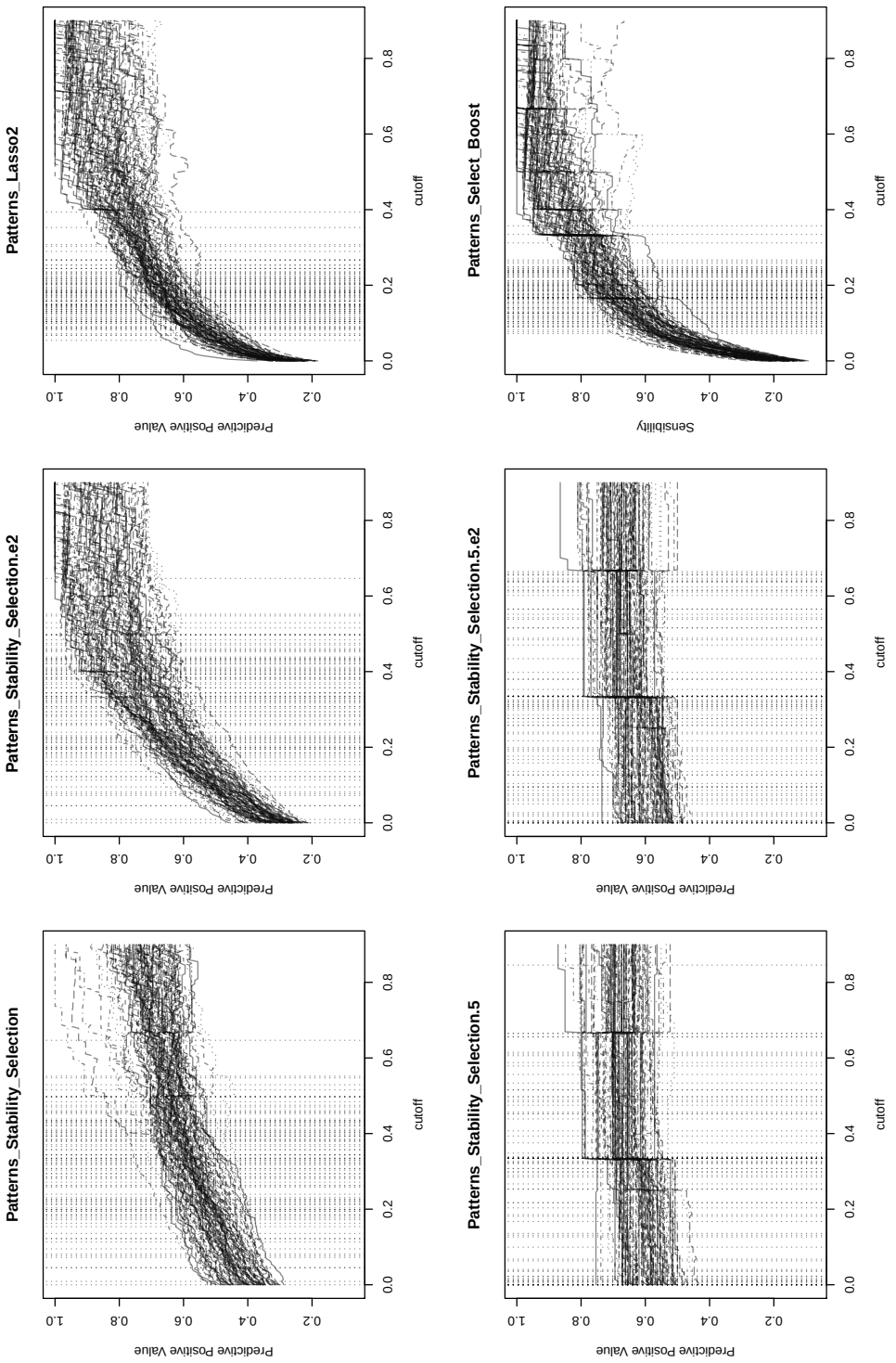


Figure 6.11 : *Valeur prédictive positive des méthodes de décodage de réseau.*

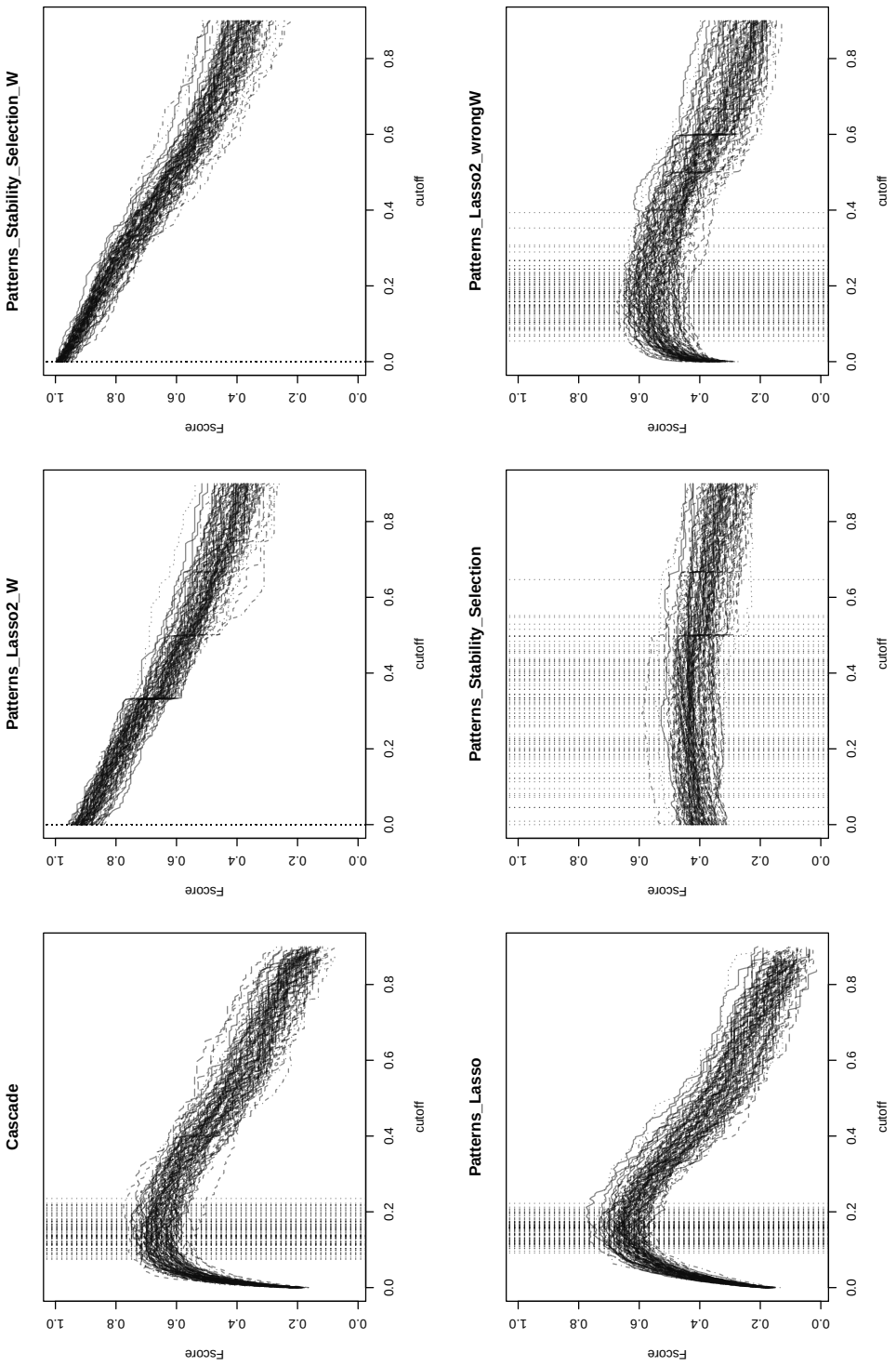


Figure 6.12 : *F-score des méthodes de décodage de réseau.*

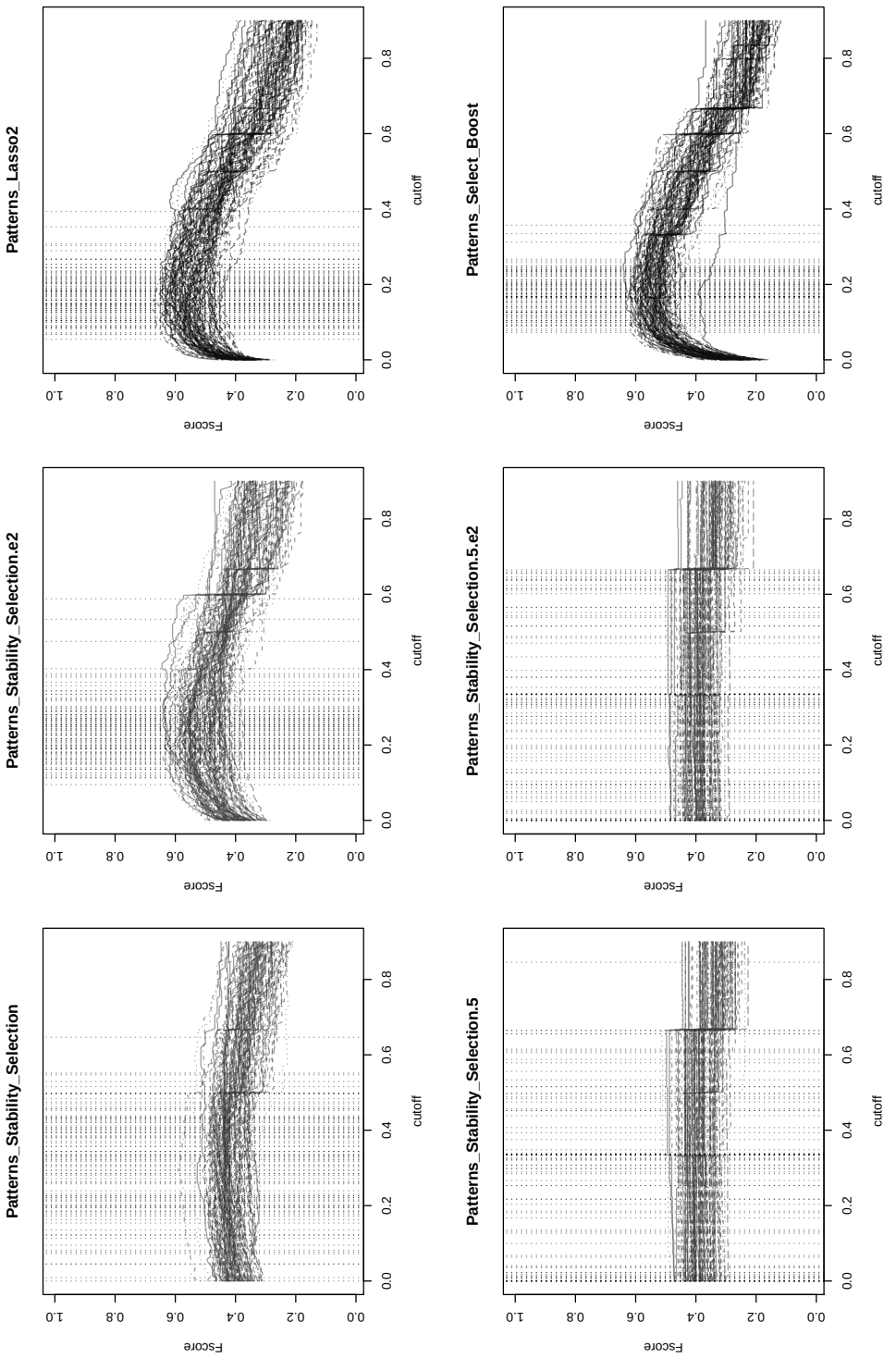


Figure 6.13 : *F-score des méthodes de décodage de réseau.*

Validation biologique

Des cellules du même patient ont été collectées lorsque celui-ci était encore dans une forme indolente de la maladie puis quand il est passé dans une forme agressive de celle-ci. Il s'agit de matériel très intéressant car il limite la variabilité interindividuelle et a servi pour vérifier si 83 acteurs, qui avaient été sélectionnés lors du décodage réalisé à la section 6.4.2, ont un comportement cohérent avec leur rôle supposément associé à la prolifération cellulaire, c'est-à-dire à la forme agressive de la maladie. J'ai utilisé en place des tests de permutation pour tester cette cohérence.

Ces recherches sont achevées depuis plusieurs mois et les résultats de l'inférence des réseaux des deux groupes de patients ont été transmis au médecin et au biologiste. Nous sommes en train d'écrire un article, Schleiss *et al.* [2018b]. J'ai réfléchi à diverses extensions des modèles, voir section 7.3. Elles sont à l'étude.

6.5 Modélisation statistique en protéomique

6.5.1 Évaluation des performances des modèles d'inférence protéiques

Une des approches couramment utilisées pour la détection des différences entre les abondances des protéines est la combinaison de *MaxLFQ*, Cox *et al.* [2014], et de *Perseus*, Tyanova *et al.* [2016]. Elle fonctionne au niveau des protéines alors que les données sont collectées au niveau des peptides.

Or, dans leur article, Suomi *et al.* [2015] ont montré qu'il est préférable de réaliser la détection différentielle des abondances des protéines directement sur les valeurs des intensités peptidiques, en particulier en présence d'un petit nombre d'échantillons. Le modèle proposé dans Goeminne *et al.* [2015], Goeminne *et al.* [2016] et Goeminne *et al.* [2017], est une approche *ridge* et robuste à partir des données peptidiques.

Dans notre cas, nous sommes intéressés par l'inférence des abondances protéiques car nous ne souhaitons pas travailler au niveau des peptiques mais à celui des protéines pour pouvoir faire des associations avec des gènes lors d'études multiomiques. Ainsi, je n'ai pas considéré les approches fonctionnant uniquement au niveau peptidique comme celle proposée dans Ning *et al.* [2016] ou une combinaison de *MaxLFQ*, Cox *et al.* [2014], et de `limma`, Ritchie *et al.* [2015].

Dans tous les cas il est souhaitable d'utiliser Giai Gianetto *et al.* [2016] pour un calcul du *FDR* qui tiendrait compte du caractère spécifique des distributions des p-valeurs des tests, Giai Gianetto *et al.* [2016] implémenté dans le package `cp4p`.

J'ai comparé ces méthodes entre elles à l'aide d'un jeu de données de référence

créé par le LSMBO et pour lequel les abondances protéiques étaient connues.

Les résultats ont confirmé ceux de Goeminne *et al.* [2015] et Goeminne *et al.* [2016] et ont montré que leur approche avait les meilleures performances aussi bien en termes de centrage que de dispersion des estimations des abondances des protéines.

6.5.2 Application à une étude réelle

L'objectif est de trouver des marqueurs qui pourraient permettre de détecter de manière précoce les patients chémoréfractaires, c'est-à-dire présentant une résistance aux traitements chimiques. Ces recherches portent sur des données génomiques et protéomiques qui ont été recueillies sur deux groupes de sujets. J'ai réalisé la partie protéomique de l'analyse statistique, ce qui m'a permis d'éprouver, à l'avance et sur un jeu de données réelles, la méthodologie que j'avais choisie à la section 6.5.1 pour l'étude présentée à la section 6.4.

Plusieurs facteurs explicatifs potentiels de la chimiorésistance ont été trouvés ainsi que de nouvelles cibles thérapeutiques possible. Les résultats de ces recherches sont présentés dans l'article Fornecker *et al.* [2018].

6.5.3 Application à un projet SATT

Pour des raisons de confidentialité, je ne peux que résumer les grandes lignes de ce travail.

À l'aide d'une première étude réalisée avec un plan d'expérience bien choisi et une méthodologie innovante, que j'ai conçus en concertation avec les médecins et chimistes, des peptides marqueurs du bon état de santé de patients ont été sélectionnés. Une seconde étude, en cours de réalisation, permettra de valider ces marqueurs et de construire un modèle statistique qui servira de base à la conception d'un test médical et qui sera, par exemple, basé sur les *quantile regression forest*, Meinshausen [2006] et Meinshausen [2017], ou le prometteur *transformation forests*, voir Hothorn et Zeileis [2017] et Hothorn [2018].

6.5.4 Perspectives futures

Ces travaux, ainsi que ceux associés au projet exposé 6.4, m'ont permis de rentrer dans le cœur des problématiques liées à l'inférence protéiques. Celles-ci ont été mises en évidence au cours de notre collaboration qui a débuté en 2013. Ainsi, suite à de nombreuses discussions et, en collaboration avec Christine Carapito chercheuse au LSMBO, j'ai proposé un sujet de thèse au LabEx IRMIA sur des problématiques propres à la protéomique. Le LabEx IRMIA a retenu cette proposition de sujet, voir la section 7.4 pour plus de détails, et Marie Chion a débuté sa

thèse en septembre 2018 après avoir passé son stage de M2 au LSBMO à s'intéresser à des problématiques de planification d'expériences et d'estimation de composants de la variance.

6.6 Outils de modélisation

6.6.1 Cascade

Afin de permettre une utilisation simplifiée des modélisations de réseaux en cascade introduites dans Vallat *et al.* [2013], j'ai proposé la réalisation d'un package pour le langage R. Celui-ci présentait de nouvelles fonctionnalités, comme le choix automatique du seuillage du réseau pour que le réseau présente une propriété d'invariance d'échelle (*scalefree*) qui est attendue pour ce type de réseau biologique, la visualisation de la propagation du signal dans le réseau ou la sélection des regroupements de gènes à l'aide du package `limma`, Ritchie *et al.* [2015], ou en spécifiant des motifs précis recherchés.

Outre le travail nécessaire sur le code pour obtenir le niveau de généralité souhaité dans un package, nous avons aussi écrit deux vignettes présentant l'application du package à deux jeux de données réels, Jung *et al.* [2014d], Jung *et al.* [2014b].

Les fonctionnalités proposées par le package sont complètes car, en plus de permettre le décodage du réseau, il est possible de procéder à la sélection des gènes, à la simulation d'expressions dans des réseaux en cascade ou bien encore de prédire l'effet d'une intervention dans le réseau de la même manière que ce que nous avions fait dans Vallat *et al.* [2013]. En résumé, ce sont les suivantes :

- Sélection de gènes et répartition des gènes en clusters ;
- Décodage du réseau ;
- Prédiction ;
- Simulation.

Le package `Cascade`, Jung *et al.* [2018], a fait l'objet d'une *application note*, Jung *et al.* [2014c].

Pour aller encore plus loin dans la mise à disponibilité de l'outil et suite à la demande de mes collaborateurs chercheurs en biologie, j'ai encadré le stage d'un étudiant ingénieur dont l'objectif a été la mise en place d'une interface web pour permettre une utilisation du package sans avoir à écrire aucune ligne de commande. Le site a été fonctionnel pendant quatre ans.

6.6.2 Patterns

Afin de pouvoir attaquer la problématique des interventions dirigées, j'ai dû commencer par généraliser le package Cascade dans deux directions.

- Permettre de gérer un nombre de groupes de gènes qui n'est plus exactement le même que le nombre de temps de mesures dans l'expérience. Nous avions en effet remarqué que les groupes de gènes formés dans Vallat *et al.* [2013] n'étaient pas homogènes en termes de profils d'expressions temporels et pensé qu'un découpage de ces groupes en sous-groupes améliorerait vraisemblablement la qualité de l'ajustement du modèle.

- Proposer à l'utilisateur de se servir de sa propre méthode d'estimation de la matrice des liens dans le réseau, $\boldsymbol{\omega}$, voir section 6.1.3, ainsi que des méthodes d'estimation supplémentaires que nous avions considérées : *elasticnet*, Zou et Hastie [2005], *sparse pls*, Chun et Keleş [2010], *stability selection*, Meinshausen et Bühlmann [2010] et notre propre méthode, elle spécialement conçue pour tenir compte de la corrélation entre les variables, *selectboost*, Aouadi *et al.* [2018].

- Outre les deux extensions précédentes qui étaient nécessaires au projet sur les interventions dirigées dans les réseaux, voir section 6.2. J'ai également ajouté la souplesse de permettre une forme de matrice $\boldsymbol{F}$, voir section 6.1.3, complètement déterminée par l'utilisateur là où Cascade ne laissait aucun choix possible. Ceci permet en particulier d'analyser non seulement des réseaux autres que des cascades mais aussi des réseaux en cascade pour lesquels l'espacement temporel entre les mesures ne serait pas régulier ou les acteurs pas tous de la même nature, voir la section 6.4 pour un contexte où ces deux situations sont simultanément réalisées et pour lesquels le package a été utilisé.

- Possibilité de récupérer de l'information biologique à partir de la base RegNetwork (341 207 liens, Liu *et al.* [2015]) pour pondérer les inférences *lasso*, *elasticnet*, *stability selection* ou *selectboost*.

- Possibilité de fixer certaines valeurs des paramètres des matrices $\boldsymbol{F}$ pour pouvoir concevoir des scénarios ou faire des tests statistiques de modèles emboîtés.

Ces trois changements majeurs ont nécessité une réécriture complète des fonctions d'inférence du package. Son code a évolué significativement depuis celui de cascade et il n'est plus uniquement dédié à l'étude des réseaux de gènes. Ce nouveau package a été nommé `Patterns`.

Le package `Patterns`, Bertrand et Maumy-Bertrand [2018a], est actuellement fonctionnel mais sa documentation est en cours de finalisation. Dès qu'elle sera achevée, il sera mis à disposition de la communauté scientifique sur le CRAN, le réseau de miroirs du langage R.

Chapitre 7

Éléments de projet de recherche

7.1 Introduction

Depuis le début de mes travaux de recherche j'ai abordé plusieurs problématiques statistiques aussi bien d'un point théorique que computationnel : les plans d'expériences, voir chapitre 2, la régression pénalisée, voir chapitre 4, la régression des moindres carrés partiels (PLS), voir chapitre 4, et certaines de leurs extensions, voir chapitre 5, ainsi que l'inférence de réseaux biologiques, voir chapitre 6. À ces quatre thématiques auxquelles je me consacre encore actuellement, je suis en train d'en ajouter une nouvelle impliquant des réseaux de neurones et plus généralement de l'apprentissage statistique.

Comme je l'ai déjà mentionné à la section 1.1 du chapitre 1, les impulsions à l'origine de ces recherches peuvent se classer en trois catégories :

- soit purement théorique ;
- soit dans un but d'améliorer ou de compléter une méthodologie existante mais qui a montré ses limites ;
- soit parfois même dues à la nécessité de concevoir des outils spécifiques pour des expériences innovantes pour lesquelles il était nécessaire de créer une solution sur mesure.

J'ai essayé d'effectuer des allers-retours constants entre des développements plus généraux et des applications de ceux-ci dans l'un des projets transdisciplinaires auxquels j'ai participé, ce qui m'a permis d'identifier des problématiques non résolues. J'ai regroupé dans ce chapitre certaines des directions dans lesquelles je souhaite poursuivre mes recherches. Je souhaite en particulier développer des approches plus robustes aux observations atypiques ou à la présence de valeurs

manquantes. En effet, comme je l'ai indiqué dans la section 1.1, suite à ma confrontation avec des jeux de données réels, même collectés avec les meilleurs protocoles expérimentaux possibles, j'ai vite atteint les limites du cadre traditionnellement défini pour l'inférence statistique.

7.2 Régression par les moindres carrés partiels

7.2.1 Contexte

Au cours des dernières années, j'ai contribué de différentes manières à l'enrichissement des modèles à réponse univariée de régression par les moindres carrés partiels (PLS), c'est donc ce cas qui est présenté ci-après. Considérons les variables centrées $\boldsymbol{y}$, $\boldsymbol{x_1}$, ..., $\boldsymbol{x_j}$, ..., $\boldsymbol{x_p}$. Soit $\boldsymbol{X}$ la matrice des prédicteurs $\boldsymbol{x_1}$, ..., $\boldsymbol{x_j}$, ..., $\boldsymbol{x_p}$. La régression PLS est bien connue et décrite de manière exhaustive notamment par Höskuldsson [1988] et Wold *et al.* [2001]. La présentation classique de la régression PLS est sous forme algorithmique. La régression PLS est un modèle non linéaire qui permet de construire des composantes orthogonales t_h obtenues en maximisant les quantités cov($\boldsymbol{y}$,$\boldsymbol{t_h}$). Soit $\boldsymbol{T}$ la matrice formée de ces composantes, nous avons :

$$\boldsymbol{y} = \boldsymbol{T}^t\boldsymbol{c} + \epsilon, \tag{7.1}$$

où $\boldsymbol{\epsilon}$ est le vecteur des résidus et ${}^t\boldsymbol{c}$ le vecteur des coefficients des composantes, ^{t} désignant la transposée.
En posant $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}^*$, où $\boldsymbol{W}^*$ est la matrice des coefficients des variables $\boldsymbol{x_j}$ dans chaque composante $\boldsymbol{t_h}$, nous avons l'expression directe de la réponse $\boldsymbol{y}$ à l'aide des prédicteurs $\boldsymbol{x_j}$:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{W}^{*t}\boldsymbol{c} + \boldsymbol{\epsilon}. \tag{7.2}$$

En développant le membre de droite de l'équation (7.2), nous obtenons pour chaque composante y_i de $\boldsymbol{y}$:

$$y_i = \sum_{h=1}^{H} \left(c_h w^*_{1h} x_{i1} + \cdots + c_h w^*_{ph} x_{ip}\right) + \epsilon_i, \tag{7.3}$$

H étant le nombre de composantes retenues dans le modèle final avec $H \leqslant \text{rang}(\boldsymbol{X})$, H étant en général très inférieur au rang de la matrice $\boldsymbol{X}$ et p étant égal au nombre de variables contenues dans la matrice $\boldsymbol{X}$. Les coefficients $c_h w^*_{jh}$, où $1 \leqslant j \leqslant p$, suivant la notation avec $*$ de Wold *et al.* [2001], traduisent la relation entre le vecteur $\boldsymbol{y}$ et les variables $\boldsymbol{x_j}$ à travers les composantes $\boldsymbol{t_h}$.

J'ai tout d'abord créé un package pour le langage libre `R`, Bertrand et Maumy-Bertrand [2018f] reprenant les extensions proposées par Bastien *et al.* [2005] et participé à un premier travail pour étudier ces modèles dans le cas des données d'allélotypages Meyer *et al.* [2010]. J'ai alors introduit de nouvelles extensions des modèles PLS, à la régression Bêta dans Bertrand *et al.* [2013a] et, avec Bastien, aux modèles de Cox dans Bastien *et al.* [2015]. L'objectif suivant a été de chercher des critères de choix du nombre de composantes, étape cruciale, pour ces modèles. Elle a été franchie, en PLSR, dans Magnanensi *et al.* [2016a], en PLSGLR, dans Magnanensi *et al.* [2017] et les prépublications Magnanensi *et al.* [2016b], pour la sPLS, et Bertrand *et al.* [2018a], pour les extensions au modèle de Cox. Une étude plus complète de l'influence des valeurs manquantes en PLS linéaire a été réalisée dans la prépublication Nengsih *et al.* [2018].

7.2.2 Développements

Mon prochain objectif est de formaliser des travaux que j'ai entrepris sur

- les méthodes à noyaux en PLS GLR, Bertrand *et al.* [2012b], et en PLS Bêta ,Bertrand *et al.* [2012a] ;
- les approches robustes pour les extensions GLM, Bêta et de Cox de la PLS afin d'obtenir des modèles qui limiteront l'influence des données atypiques en les comparant à l'utilisation de techniques de prétraitement de données comme la winsorisation ;
- l'influence des valeurs manquantes dans les extensions GLR, Bêta ou au modèle de Cox de la régression PLS ;
- le choix de composantes par bootstrap pour les extensions Bêta ou au modèle de Cox de la régression PLS pour éliminer les problèmes d'instabilité observés avec la validation croisée ;

et d'un point de vue de l'implémentation d'achever les extensions suivantes

- l'utilisation de techniques parallèles et/ou GPU pour permettre le traitement de jeux de données de taille plus importante et accélérer les nouvelles techniques de détermination du nombre de composantes que nous avons introduites ;
- tout en développant en parallèle les mêmes outils pour le nouveau langage julia (Bezanson *et al.* [2017]) qui est sans doute un des standards de demain pour tous les chercheurs mettant en œuvre des calculs intensifs.

7.3 Inférences de réseaux biologiques

Depuis 2011 et comme je l'ai exposé à la section 4.7 du chapitre 4 et tout au long du chapitre 6, j'ai recherché de nouvelles méthodes d'inférence de réseaux de gènes, de réseaux de protéines puis d'inférence conjointe de réseaux de gènes et protéines au niveau global ou d'un individu.

7.3.1 Contexte

Le comportement cellulaire en réponse aux signaux de l'environnement est soutenu par différents programmes géniques complexes qui sont progressivement altérés dans des situations pathologiques, notamment dans le cancer.

Des méthodes statistiques ont été proposées pour décrire les réseaux de régulation géniques, structures sous-jacentes de ces programmes. D'autres développements ont été nécessaires pour permettre des modulations orientées de ces systèmes. Notre but était de développer une méthode d'inférence permettant de prédire l'impact d'une intervention dans un programme génique. Une telle capacité prédictive est nécessaire pour obtenir une modulation dirigée à différents niveaux biologiques : génique, protéomique et fonctionnel, voir chapitre 6. Elle permettra à (long) terme d'opérer la modulation dirigée d'un programme génique qui soutient le comportement cellulaire cancéreux dans une perspective thérapeutique.

7.3.2 Approches robustes

Comme cela a été montré par Marbach *et al.* [2012] et Hsiao et Lee [2012], il est intéressant de disposer d'une méthode statistique d'inférence robuste pour les jeux de données géniques. C'est également vraisemblablement le cas dans mon contexte puisque les données auxquelles s'appliquent les modèles que je développe ont été observées sur des patients à l'hôpital et présentent donc potentiellement une variabilité interindividuelle importante.

Je propose donc de créer une extension robuste des méthodologies que j'ai proposées au chapitre 6 qui sont essentiellement basée sur de la régression pénalisée et la détection de gènes différentiellement exprimés à l'aide d'une approche publiée par Ritchie *et al.* [2015] et Phipson *et al.* [2016]. Outre l'écriture d'un article pour une revue internationale spécialisée en statistique, j'envisage de créer un package pour le langage `R` ou pour l'initiative `BioConductor` pour diffuser ces outils.

Plus précisément, les modèles de réseaux qui sont décrits au chapitre 6 sont calculés à partir de l'application élémentaire de régression pénalisées ou de régres-

sion par les moindres carrés partiels à des sous-ensembles de gènes choisis pour être le plus homogène possible.

Pourquoi ne pas remplacer l'étape d'inférence pénalisée par une inférence pénalisée robuste ou appliquer, au préalable, une approche robuste par *winsorization* au jeu de données pour essayer de limiter l'influence des observations influentes. Ces aspects rejoignent certains de ceux évoquées à la section 7.2.2 au sujet de la régression PLS et pourront être menés conjointement.

7.3.3 Des mesures de nature très différente

Une dernière voie d'amélioration concerne spécifiquement l'inférence conjointe des réseaux. Compte tenu de la nature différente, comptage pour les données génomiques et intensité pour les données protéiques, des observations effectuées, il serait intéressant d'étudier l'apport de l'ajout d'une étape de transformation non paramétrique robuste (régression locale robuste, *splines* robustes) afin d'améliorer la compatibilité des profils temporels d'expressions des gènes et des protéines et ainsi la détection d'actions entre gènes et protéines.

7.3.4 Parallélisation du code

L'ajustement de ces modèles serait grandement accéléré par l'utilisation de techniques de calcul haute performance : lors de l'inférence des estimations complexes sont réalisées en parallèle pour chacun des acteurs parmi les milliers d'acteurs impliqués. Ainsi, la nature même des modèles retenus permet d'espérer que l'utilisation de gpu sera profitable. J'envisage de développer également les mêmes outils pour le nouveau langage julia (Bezanson *et al.* [2017]).

7.3.5 Autre adaptation aux nouvelles données biologiques

Avec l'arrivée des techniques « *nextgen* » en biologie, la nature même des variables à modéliser s'est vue modifiée passant d'une variable quantitative continue à un dénombrement. Par conséquent, nous pourrions penser que les modèles développés pour la technologie précédente, les puces à ADN (*microarrays*), et qui reposent sur l'utilisation de lois continues (normale, student, ...) ne sont plus pertinents. Or, la transformation `voom`, Law *et al.* [2014], permet de justifier de continuer à utiliser de ces anciens modèles avec le nouveau type de données une fois transformées. C'est cette voie que j'ai utilisée jusqu'à présent. Néanmoins, il serait pertinent de se tourner également vers des lois discrètes comme les mélanges de lois de Poisson ou de lois binomiales négatives. En conséquence, la méthode d'inférence que

j'ai proposé au chapitre 6 devra être retravaillée et s'appuyer sur l'ajustement de modèles linéaires généralisés par des estimateurs *lasso*, *ridge* ou *elasticnet*, Friedman *et al.* [2010], ainsi que des techniques de maximum de vraisemblance avec contraintes de positivité sur les paramètres.

7.4 Modèles statistiques pour données protéomiques

7.4.1 Contexte

Ce thème de recherches a fait l'objet d'une demande de contrat doctoral auprès du LabEx IRMIA en avril 2018 qui a été acceptée en juin 2018. La thèse a débuté en septembre 2018 et voici le sujet qui a été proposé.

La spectrométrie de masse permet de mesurer très précisément les masses de molécules d'intérêt et d'en caractériser la structure chimique. L'analyse protéomique consiste à étudier l'ensemble des protéines exprimées par une cellule, un tissu, un organe ou un organisme à un moment donné et sous des conditions données, appelé le protéome. Au même titre que la génomique et la transcriptomique, l'analyse protéomique est devenue aujourd'hui un outil incontournable pour l'étude des systèmes biologiques complexes et s'est révélée particulièrement prometteuse, entres autres, pour la découverte et validation de biomarqueurs de pathologies.

7.4.2 Problématique

L'objectif visé dans cette thématique de recherches est double et consiste en la proposition d'une approche novatrice pour traiter les données de quantification et l'utilisation d'outils d'apprentissage statistique pour tirer parti de la très grande masse de spectres récoltés jusqu'à présent sans avoir pu être exploités (en moyenne 75% des centaines de millions de spectres acquis à ce jour et répertoriés dans l'archive PRIDE, Martens *et al.* [2005], Vizcaíno *et al.* [2016], restent ininterprétés, Griss *et al.* [2016], the « *dark mater in proteomics* »).

7.4.3 Valeurs manquantes

Les techniques les plus performantes pour déterminer l'abondance des protéines passent par la mesure des intensités peptidiques. Une gamme variée de modèles statistiques a déjà été utilisée pour mener à bien celle-ci : modèles linéaires mixtes,

modèles non linéaires, modèles pour données censurées. En effet, ces données peptidiques présentent des difficultés majeures : corrélations entre peptides, valeurs manquantes non MCAR et présence d'observations atypiques influentes. Le modèle actuellement le plus performant repose sur l'utilisation, pour chaque protéine, d'un modèle linéaire mixte régularisé (*ridge*) et robuste ajusté aux intensités peptidiques. Celle-ci est combinée avec une estimation mutualisée de la variance des protéines à la manière de ce qui a été retenu pour les expressions des gènes dans les puces à ADN (*microarrays*), Smyth [2004].

Cette approche, qui est la plus performante en termes d'erreur quadratique moyenne, n'est actuellement applicable qu'à la comparaison des protéines entre deux conditions expérimentales pour lesquelles un nombre minimal de peptides a été quantifié dans chacune d'entre elle. Or, l'absence de quantification des peptides dans l'une des deux conditions est elle-même informative d'une différence potentielle entre les abondances des protéines.

L'objectif est de proposer une méthodologie statistique pour permettre une gestion plus satisfaisante des valeurs manquantes et des cas de protéines absentes/présentes au sein des conditions à comparer. D'un point de vue statistique, ce sont des modèles linéaires généralisés mixtes, des modèles avec inflation de zéros, modèles bêta-binomiaux pour données de dénombrement qui pourront être utilisés. Une source d'inspiration intéressante pourra être constituée par les modèles linéaires généralisés mixtes actuellement utilisés pour la modélisation du dénombrement des expressions de gènes en RNA-Seq. Dans un second temps, il faudra essayer de combiner cette approche avec celle déjà existante basée sur l'utilisation des données d'intensité à l'aide d'une approche bivariée/multivariée ou de modèles *hurdle*. L'objet de cette première partie consistera donc avant tout d'un travail méthodologique accompagné de la mise au point d'une bibliothèque de fonctions (*package*) pour le langage `R`, plus particulièrement le projet `Bioconductor`, qui sera mise à la disposition de la communauté scientifique.

7.4.4 Apprentissage statistique

L'utilisation de techniques d'apprentissage statistique permettra de s'intéresser à des problématiques diverses mais toutes d'un intérêt premier pour le chimiste.

La première d'entre elle est l'identification de spectres non encore identifiés. Il s'agit de détecter, parmi la masse de spectres disponibles, ceux qui sont associés à des groupes protéiques (fragments de protéines) mais qui n'ont pas été encore identifiés. Il n'est en effet pas possible d'identifier systématiquement tous les spectres produits, certains étant mêmes associés à du bruit et donc difficilement voire non

identifiables. De plus, l'identification actuelle est issue de l'utilisation d'une banque traduite à partir d'un seul génome, séquence consensus de référence, alors qu'il est connu qu'il existe une grande diversité entre les génomes des individus même lorsqu'ils présentent le même gène, variants de séquences individuels. Ces variants donnent naissance à des formes différentes de protéines qu'il convient d'identifier (banque et médecine personnalisées). La difficulté résidera dans la séparation des spectres de groupes protéiques non encore identifiés du bruit et la méthodologie proposée s'appuiera sur l'utilisation combinée d'algorithmes de clustering et de réseaux de neurones de type GAN (*generative adversory networks*).

La deuxième problématique consiste en la prédiction des spectres à partir des séquences protéiques, c'est-à-dire de données fonctionnelles à partir d'un mot protéique constitué d'une suite d'acides aminés (lettres). Compte tenu de la nature du prédicteur (un mot) et de l'objectif, ici seulement une prédiction, des techniques d'apprentissage statistique comme les séparateurs à vaste marge (SVM) ou les réseaux de neurones semblent être des outils pertinents. Un soin particulier sera apporté à l'étude d'une méthodologie permettant d'obtenir une synthèse « additive » des spectres.

Enfin la dernière problématique qui pourra être développée est la réduction de bruit, qu'elle résulte de l'application directe de filtres, comme ceux de Kalman, du modèle de synthèse additive précédent, ou de l'application directe de techniques d'apprentissage statistique. Cette élimination du bruit peut être réalisée à différents niveaux : le tri en spectres de groupes protéiques ou de bruit de fond mais aussi élimination du bruit chimique et électronique sur les spectres MS/MS.

Il est important de noter que toutes les données nécessaires à la mise en œuvre de ces développements :

- aussi bien les jeux de données mentionnés dans le premier objectif et qui consistent en la mesure d'échantillons dits *spikés*, parfaitement maîtrisés avec ajouts de protéines en quantités connues ;

- que la base de données publique regroupant un très grand nombre de spectres recueillis par la communauté de spectrométrie de masse, Deutsch *et al.* [2017], sont déjà disponibles.

7.5 De la fouille des processus à l'intelligence des processus

7.5.1 Contexte

Cette thématique de recherche a été financée par un PEPS IA (8k€) en mars 2018. Elle s'inscrit dans le cadre d'une collaboration de recherche entre la société Your Data Consulting, jeune entreprise innovante (JEI) basée à Paris et les détails de celle-ci doivent demeurer confidentiels.

La fouille de processus est une discipline de recherche qui se sert, d'une part, des techniques de l'intelligence artificielle et de la fouille de données et, d'autre part, de la modélisation et de l'analyse des processus. Elle facilite ainsi l'analyse des processus d'entreprises sur la base des journaux d'événements, extraits des systèmes informatiques, voir l'article de Van Der Aalst *et al.* [2011]. Elle est comparable au *Data Mining*, mais l'objectif est surtout axé sur l'acquisition de la connaissance des processus, voir l'article de van der Aalst *et al.* [2015]. Les approches existantes permettent de découvrir le modèle de processus, de détecter des modifications du modèle initialement conçu, de trouver des corrélations entre les données du processus et les différentes variantes du modèles (voir Delias *et al.* [2015]), d'analyser et de prédire des aspects inefficaces (voir Grigori *et al.* [2004]).

Même si beaucoup d'approches de fouille de processus ont été proposées dans la littérature, les applications actuelles posent de nouveaux défis (voir Beheshti *et al.* [2018]) :

- le volume des événements stockés dans le journal est très grand ;
- les processus génèrent des données stockées dans différents systèmes et formats posant donc un problème d'intégration de données ;
- des parties de processus peuvent s'exécuter via des échanges de messages électroniques ;
- certains processus sont très flexibles, non–structurés ou *ad-hoc*.

7.5.2 Problématique

La fouille de processus (*process mining*) cherche à comprendre et à prédire la succession des évènements auxquels sont soumis les unités lorsqu'elles parcourent un processus. Cet enchaînement d'étapes s'appelle une trace et se représente naturellement sous la forme d'une suite finie c'est-à-dire d'un mot. C'est en tirant

partie de cette nature séquentielle que, depuis 2016, quelques premiers travaux de recherche cherchant à combiner fouille de processus et apprentissage profond, LeCun *et al.* [2015], ont déjà été menés par, entre autres, Evermann *et al.* [2016], Tax *et al.* [2017], Evermann *et al.* [2017a], Di Francescomarino *et al.* [2017] et Evermann *et al.* [2017b]. Mais ils restent exploratoires et très spécifiques puisque limités à l'utilisation d'un seul type de réseau de neurones : les réseaux de neurones récurrents à mémoire court et long terme, *LSTM*, introduits par Hochreiter et Schmidhuber [1997]. Toutefois, les mêmes auteurs ont montré que ces approches permettent déjà d'obtenir de meilleurs résultats pour prédire le futur d'une trace que les techniques utilisées jusqu'alors.

7.5.3 Développements

Je propose dans cet axe de recherche de concevoir des solutions basées à la fois sur la fouille de processus et l'intelligence artificielle pour analyser le plus finement possible le parcours clients/produits/colis et de prédire un parcours client/produit/colis en poursuivant deux objectifs principaux complémentaires.

- Les jeux de données récoltés pour servir à la fouille de processus contiennent souvent des informations auxiliaires qui sont capitales pour prédire de manière adéquate le devenir des unités dans le processus.

  Il est très surprenant que la quasi-totalité des modèles prédictifs de fouille de processus ne s'appuie malheureusement pas sur ces covariables. À ma connaissance, en février 2018, un seul modèle permettrait d'intégrer des covariables. Il est basé sur des séparateurs à vaste marge (SVM) mais son implémentation n'est pas publique. À ce jour, aucun de ceux combinant fouille de processus et apprentissage profond ne le permet.

  J'ai conçu un modèle qui y parvient et je l'ai implémenté. Il reste à évaluer ses performances.

- Une ouverture vers d'autres outils pour dépasser les performances des modélisations existantes. L'utilisation de nouveaux types de réseaux de neurones récurrents, comme les *Phased LSTM* de Neil *et al.* [2016], ou de nouveaux algorithmes pour accélérer la vitesse d'apprentissage des réseaux, comme l'*importance sampling* de Katharopoulos et Fleuret [2017].

Bibliographie

Alawieh, H., Bertrand, F., Maumy-Bertrand, M., Wicker, N. et Ayoubi, B. A. [2018]. A random model for multidimensional fitting method. URL http://arxiv.org/abs/1810.05042.

Alawieh, H., Wicker, N., Al Ayoubi, B. et Moulinier, L. [2017]. Penalized multidimensional fitting for protein movement detection. *Journal of Applied Statistics*, **44**(15), 2697–2715. ISSN 0266-4763. URL https://www.tandfonline.com/doi/full/10.1080/02664763.2016.1261811.

Albert, R. et Barabási, A.-L. [2002]. Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**(1), 47–97. ISSN 0034-6861. URL https://link.aps.org/doi/10.1103/RevModPhys.74.47.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. et Staudt, L. M. [2000]. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**(6769), 503–511. ISSN 0028-0836. URL http://www.nature.com/articles/35000501.

Allen, D. M. [1971]. *The prediction sum of squares as a criterion for selecting predictor variables*. Technical report 23 - Department of Statistics. University of Kentucky.

Anderson, T. W. [1946]. The Non-Central Wishart Distribution and Certain Problems of Multivariate Statistics. *The Annals of Mathematical Statistics*, **17**(4), 409–431. ISSN 0003-4851. URL http://projecteuclid.org/euclid.aoms/1177730882.

Aouadi, I., Jung, N., Carapito, R., Vallat, L., Bahram, S., Maumy-Bertrand, M. et Bertrand, F. [2018]. selectBoost : a general algorithm to enhance the performance of variable selection methods in correlated datasets. URL http://arxiv.org/abs/1810.01670.

Bannai, E. et Bannai, E. [2009]. A survey on spherical designs and algebraic combinatorics on spheres. *European Journal of Combinatorics*. ISSN 01956698.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A. et di Bernardo, D. [2007]. How to infer gene networks from expression profiles. *Molecular systems biology*, **3**, 78. ISSN 1744-4292. URL http://www.ncbi.nlm.nih.gov/pubmed/17299415.

Bar-Joseph, Z., Gitter, A. et Simon, I. [2012]. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, **13**(8), 552–564.

Barabási, A.-L. et Oltvai, Z. N. [2004]. Network biology : understanding the cell's functional organization. *Nature Reviews Genetics*, **5**(2), 101–113. ISSN 1471-0056. URL http://www.nature.com/articles/nrg1272.

Bastien, P. [2008]. Deviance residuals based PLS regression for censored data in high dimensional setting. *Chemometrics and Intelligent Laboratory Systems*, **91** (1), 78–86.

Bastien, P., Bertrand, F., Meyer, N. et Maumy-Bertrand, M. [2015]. Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. *Bioinformatics*, **31**(3), 397–404. ISSN 14602059.

Bastien, P., Vinzi, V. E. et Tenenhaus, M. [2005]. PLS generalised linear regression. *Computational Statistics & Data Analysis*, **48**(1), 17–46. ISSN 01679473. URL http://www.sciencedirect.com/science/article/pii/S0167947304000271.

Beheshti, S.-M.-R., Benatallah, B., Sakr, S., Grigori, D., Motahari-Nezhad, H. R., Barukh, M. C., Gater, A. et Ryu, S. H. [2018]. *Process Analytics : concepts and techniques for querying and analyzing process data.* Springer, New-York. ISBN 9783319797243.

Berge, C., Froloff, N., Kalathur, R. K. R., Maumy, M., Poch, O., Raffelsberger, W. et Wicker, N. [2010]. Multidimensional fitting for multivariate data analysis. *Journal of computational biology : a journal of computational molecular cell biology*, **17**(5), 723–32. ISSN 1557-8666. URL http://www.ncbi.nlm.nih.gov/pubmed/20175691.

Bertrand, F. [2007]. *Plans sphériques de force t et applications en statistique.* Thèse de doctorat, université Louis Pasteur - Strasbourg I.

Bertrand, F. [2008a]. Problèmes de construction de type polynomial I – Caractérisations polynomiales des propriétés usuelles d'un plan. *Comptes Rendus Mathematique*, **346**(21–22). ISSN 1631073X.

Bertrand, F. [2008b]. Problèmes de construction de type polynomial II - Quelques résultats d'existence de plans sphériques isovariants exacts. *Comptes Rendus Mathematique*, **346**(23–24). ISSN 1631073X.

Bertrand, F. [2009]. Rotatable designs and G-weakly invariant designs | G-invariance faible et isovariance en planification expérimentale. *Comptes Rendus Mathematique*, **347**(1-2), 93–98. ISSN 1631073X.

Bertrand, F. [2010]. Weakly Invariant Designs, Rotatable Designs and Polynomial Designs. Dans *Algebraic Methods in Statistics and Probability II*, volume 516 de *Contemporary mathematics*, 49–60. American Mathematical Society. ISBN 0271-4132.

Bertrand, F. [2015a]. Marginal and conjoint temporal clusterings of proteins and genes. Rapport technique, université de Strasbourg.

Bertrand, F. [2015b]. Pathway-based level-specific data comparison of coupled human proteomic and genomic/transcriptomic. Rapport technique, université de Strasbourg.

Bertrand, F. [2015c]. Protein and gene joint temporal pattern analysis. Rapport technique, université de Strasbourg.

Bertrand, F. [2016a]. AmpliSeq Processing Data and Multivariate Analysis. Rapport technique, université de Strasbourg.

Bertrand, F. [2016b]. Temporal CLL RNASeq analysis : differential expressions and clustering. Rapport technique, université de Strasbourg.

Bertrand, F. [2016c]. Temporal proteomic analysis : differential abundancies and clustering. Rapport technique, université de Strasbourg.

Bertrand, F. [2017a]. Gene and protein joint selection : selecting actors for joint modelling. Rapport technique, université de Strasbourg.

Bertrand, F. [2017b]. Joint modelling of gene and proteins. Rapport technique, université de Strasbourg.

Bertrand, F., Bastien, P. et Maumy-Bertrand, M. [2014a]. Cross-validated partial least squares models and their extensions with censored data. Dans *21st International Conference on Computational Statistics, Genève, Suisse*.

Bertrand, F., Bastien, P. et Maumy-Bertrand, M. [2015]. Cross validating extensions of kernel, sparse or regular partial least squares regression models to censored data. Dans *CMStatistics 2015, London*.

Bertrand, F., Bastien, P., Meyer, N. et Maumy-Bertrand, M. [2014b]. plsRcox, Cox-Models in a high dimensional setting in R. Dans *Proceedings of User2014 !, Los Angeles*, 152.

Bertrand, F., Bastien, P. et Maumy-Bertrand, M. [2018a]. Cross validating extensions of kernel, sparse or regular partial least squares regression models to censored data. URL http://arxiv.org/abs/1810.02962.

Bertrand, F., Droesbeke, J.-J., Saporta, G. et Thomas-Agnan, C., éditeurs [2017]. *Model Choice and Model Aggregation*. Technip, Paris.

Bertrand, F., Fredon, D. et Maumy-Bertrand, M. [2016]. *Mathématiques Licence 1 - Exercices et méthodes*. Dunod, Paris.

Bertrand, F., Fredon, D., Rabba-Idi, Y. et Maumy-Bertrand, M. [2018b]. *Mathématiques Licence 2 - Exercices et méthodes*. Dunod, Paris.

Bertrand, F., Magnanensi, J., Meyer, N. et Maumy-Bertrand, M. [2014c]. *plsRglm : Algorithmic insights and applications*. Vignette of the package.

Bertrand, F., Magnanensi, J., Meyer, N. et Maumy-Bertrand, M. [2014d]. plsRglm, PLS generalized linear models for R. Dans *Proceedings of User2014 !, Los Angeles*, 150.

Bertrand, F., Maumy, M., Fussler, L., Kobes, N., Savary, S. et Grosman, J. [2008]. Étude statistique des données collectées par l'Observatoire National des Maladies du Bois de la Vigne. *Journal de la Société Française de Statistique*, **149**(4), 73–106.

Bertrand, F. et Maumy, M. [2007a]. Développements d'Edgeworth de deux estimateurs d'une proportion de mesures. *Comptes Rendus Mathematique*, **345**(7), 399–404. ISSN 1631073X.

Bertrand, F. et Maumy, M. [2007b]. Intervalles de confiance bilatéraux et unilatéraux d'une proportion de mesures. Dans *Chimiométrie 2007*, 82–85, Lyon, France. URL https://hal.archives-ouvertes.fr/hal-00193173. 4 pages.

Bertrand, F. et Maumy, M. [2008]. Decision rules based on the estimate of a proportion of measurements. Dans *11th Conference on Chemometrics in Analytical Chemistry*, volume 2, 245 – 249, Montpellier, France. URL https://hal.archives-ouvertes.fr/hal-00287748. 5 pages.

Bertrand, F. et Maumy, M. [2010]. Using Partial Triadic Analysis for Depicting the Temporal Evolution of Spatial Structures : Assessing Phytoplankton Structure and Succession in a Water Reservoir. *Case Studies In Business, Industry And Government Statistics*, **4**(1), 23–43. ISSN 2152-372X. URL http://journal-sfds.fr/index.php/csbigs/article/view/286.

Bertrand, F., Maumy, M., Fussler, L., Kobes, N., Savary, S. et Grosman, J. [2007a]. Using Factor Analyses to explore data generated by the National Grapevine Wood Diseases Survey. *Case Studies in Business, Industry and Government Statistics*, **1**(2). URL http://hal.archives-ouvertes.fr/hal-00166970.

Bertrand, F. et Maumy-Bertrand [2018a]. *Patterns : reverse-engineering temporal biological networks*. URL `http://www-irma.u-strasbg.fr/~fbertran/`. R package version 0.5.

Bertrand, F. et Maumy-Bertrand, M. [2011]. *Maxi fiches de Statistique. En 80 fiches*. Dunod, Paris.

Bertrand, F. et Maumy-Bertrand, M. [2012]. *Mathématiques : Concours des catégories A et B*. Dunod, Paris.

Bertrand, F. et Maumy-Bertrand, M. [2018b]. A Sheet of Maple to Compute Second-Order Edgeworth Expansions and Related Quantities of any Function of the Mean of an iid Sample of an Absolutely Continuous Distribution. URL `http://arxiv.org/abs/1810.00289`.

Bertrand, F. et Maumy-Bertrand, M. [2018c]. *Initiation à la statistique avec R : Cours, exemples, exercices et problèmes corrigés*. Dunod, Paris, 3e édition.

Bertrand, F. et Maumy-Bertrand, M. [2018d]. *Partial Least Squares Regression for Beta Regression Models*. URL `http://www-irma.u-strasbg.fr/~fbertran/`. R package version 0.2.3.

Bertrand, F. et Maumy-Bertrand, M. [2018e]. *Partial Least Squares Regression for Cox Models and Related Techniques*. URL `http://www-irma.u-strasbg.fr/~fbertran/`. R package version 1.7.3.1.

Bertrand, F. et Maumy-Bertrand, M. [2018f]. *Partial Least Squares Regression for Generalized Linear Models*. URL `http://www-irma.u-strasbg.fr/~fbertran/`. R package version 1.2.3.

Bertrand, F. et Maumy-Bertrand, M. [2018g]. plsRglm : Partial least squares linear and generalized linear regression for processing incomplete datasets by cross-validation and bootstrap techniques with R. URL `http://arxiv.org/abs/1810.01005`.

Bertrand, F., Maumy-Bertrand, M. et Aouadi, I. [2018c]. *selectboost : A General Algorithm to Enhance the Performance of Variable Selection Methods in Correlated Datasets*. URL `http://www-irma.u-strasbg.fr/~fbertran/`. R package version 0.5.0.

Bertrand, F., Maumy-Bertrand, M., Ferrigno, S., Muller-Gueudin, A. et Marx, D. [2013a]. *Mathématiques pour les sciences de l'ingénieur - Tout le cours en fiches*. Dunod, Paris. ISBN 9782100570614. URL `http://www.dunod.com/sciences-techniques/sciences-fondamentales/mathematiques/licence/mathematiques-pour-les-sciences-de-lingenieur-tout-le-c`.

Bertrand, F., Maumy-Bertrand, M. et Meyer, N. [2012a]. Kernel pls beta regressions. Dans *Proceedings of CAC 2012, Paris, France*.

Bertrand, F., Maumy-Bertrand, M. et Meyer, N. [2012b]. Kernel pls glm regressions. Dans *Proceedings of ENBIS 2012, Ljublana, Slovenia*.

Bertrand, F., Maumy-Bertrand, M. et Périnel, E. [2011]. *Économétrie, Statistiques et Probabilités : Concours des catégories A et B*. Dunod, Paris.

Bertrand, F., Meyer, N., Beau-Faller, M., Bayed, K. E., Izzie-J., N. et Maumy-Bertrand, M. [2013b]. Régression Bêta PLS [PLS Beta regression]. *Journal de la Société Française de Statistique*, **154**(3), 143–159.

Bertrand, F., Ourliac, A. et Boulanger, B. [2007b]. Recherche numérique de plans D-optimaux pour des problèmes de pharmacocinétique et pharmacodynamique : une étude de cas. Dans *39ème Journées de statistique de la SFdS*, 34, Angers, France. Société Française de Statistique. URL `https://hal.archives-ouvertes.fr/hal-00160194`. 6 pages ; 2 figures.

Bertrand, F., Saporta, G. et Thomas-Agnan, C., éditeurs [2019]. *Statistique et Causalité*. Technip, Paris. à paraître.

Bertrand, F., Schleiss, C., Pionneau, C., Mauvieux, L., Herbrecht, R., Maumy-Bertrand, M., S., B. et Vallat, L. [2018d]. Core transcriptional and proteomic program of aggressive CLL B-cells after BCR engagement. en préparation.

Bezanson, J., Edelman, A., Karpinski, S. et Shah, V. B. [2017]. Julia : A Fresh Approach to Numerical Computing. *SIAM Review*, **59**(1), 65–98. ISSN 0036-1445. URL `https://epubs.siam.org/doi/10.1137/141000671`.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, O., Frigessi, A. et Lingjærde, O. C. [2007]. Predicting survival from microarray data - A comparative study. *Bioinformatics*, **23**, 2080–2087.

Boulesteix, A.-L. [2014]. Accuracy estimation for PLS and related methods via resampling-based procedures. Dans *PLS'14 Book of Abstracts*, 13–14.

Boullion, T. L., Cascio, G. C. et Keating, J. P. [1985]. Comparison of estimators of the fraction defective in the normal distribution. *Communications in Statistics - Theory and Methods*, **14**(7), 1511–1529. ISSN 0361-0926. URL `http://www.tandfonline.com/doi/abs/10.1080/03610928508828993`.

Bourgon, R., Gentleman, R. et Huber, W. [2010]. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(21), 9546–51. ISSN 1091-6490. URL `http://www.pnas.org/content/107/21/9546.long`.

Brier, G. W. [1950]. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, **78**(1), 1–3. ISSN 0027-0644. URL `http://journals.ametsoc.org/doi/abs/10.1175/1520-0493%281950%29078%3C0001%3AVOFEIT%3E2.0.CO%3B2`.

Brown, L. D., Cai, T. T. et Das Gupta, A. [2001]. Interval Estimation for a Binomial Proportion. *Statistical Science*, **16**(2), 101–133.

Brown, L. D., Cai, T. T. et Das Gupta, A. [2002]. Confidence Intervals for a Binomial Proportion and Asymptotic Expansions. *The Annals of Statistics*, **30** (1), 160–201.

Brown, L. D., Cai, T. T. et Das Gupta, A. [2003]. Interval Estimation in Exponential Families. *Statistica Sinica*, **13**, 19–49.

Bry, X., Trottier, C., Verron, T. et Mortier, F. [2013]. Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, **119**, 47–60. ISSN 0047259X. URL `http://www.sciencedirect.com/science/article/pii/S0047259X13000407`.

Cai, T. T. [2005]. One-Sided Confidence Intervals in Discrete Distributions. *Journal of Statistical Planning and Inference*, **131**, 63–88.

Canty, A. et Ripley, B. [2014]. *boot : Bootstrap R (S-Plus) Functions*. URL `http://cran.r-project.org/package=boot`.

Carapito, R., Jung, N., Kwemou, M., Untrau, M., Michel, S., Pichot, A., Giacometti, G., Macquin, C., Ilias, W., Morlon, A., Kotova, I., Apostolova, P., Schmitt-Graeff, A., Cesbron, A., Gagne, K., Oudshoorn, M., Van Der Holt, B., Labalette, M., Spierings, E., Picard, C., Loiseau, P., Tamouza, R., Toubert, A., Parissiadis, A., Dubois, V., Lafarge, X., Maumy-Bertrand, M., Bertrand, F., Vago, L., Ciceri, F., Paillard, C., Querol, S., Sierra, J., Fleischhauer, K., Nagler, A., Labopin, M., Inoko, H., Von Dem Borne, P., Kuball, J., Ota, M., Katsuyama, Y., Michallet, M., Lioure, B., De Latour, R., Blaise, D., Cornelissen, J., Yakoub-Agha, I., Claas, F., Moreau, P., Milpied, N., Charron, D., Mohty, M., Zeiser, R., Socié, G. et Bahram, S. [2016]. Matching for the nonconventional MHC-I MICA gene significantly reduces the incidence of acute and chronic GVHD. *Blood*, **128** (15). ISSN 15280020.

Chambless, L. E. et Diao, G. [2006]. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*, **25**, 3474–3486. ISSN 02776715.

Chernozhukov, V., Fernández-Val, I. et Galichon, A. [2010]. Rearranging Edgeworth-Cornish-Fisher Expansions. *Economic Theory*, **42**(2), 419–435.

Cheung, L. W.-K. [2012]. Classification Approaches for Microarray Gene Expression Data Analysis. 73–85. Humana Press. URL `http://link.springer.com/10.1007/978-1-61779-400-1{_}5`.

Chun, H. et Keleş, S. [2010]. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical*

Society. Series B, Statistical methodology, **72**(1), 3–25. ISSN 1369-7412. URL http://www.ncbi.nlm.nih.gov/pubmed/20107611.

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N. et Mann, M. [2014]. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP*, **13**(9), 2513–26. ISSN 1535-9484. URL http://www.mcponline.org/content/13/9/2513.long.

Cox, T. F. et Cox, M. A. [2000]. *Multidimensional scaling*. Chapman and hall/-CRC, 2\ieme édition. ISBN 978-1-4129-1611-0 1-4129-1611-9.

Cribari-Neto, F. et Zeileis, A. [2010]. Beta Regression in R. *Journal of Statistical Software*, **34**(2), 1–24.

Crick, F. [1970]. Central Dogma of Molecular Biology. *Nature*, **227**(5258), 561–563. ISSN 0028-0836. URL http://www.nature.com/articles/227561a0.

Das Gupta, A. [2008]. *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer New York, New York, NY. ISBN 978-0-387-75970-8. URL http://link.springer.com/10.1007/978-0-387-75971-5.

Davison, A. C. et Hinkley, D. V. [1997]. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge.

Delias, P., Grigori, D., Mouhoub, M. L. et Tsoukias, A. [2015]. Discovering Characteristics that Affect Process Control Flow. 51–63. Springer, Cham. URL http://link.springer.com/10.1007/978-3-319-21536-5{_}5.

Deutsch, E. W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., Campbell, D. S., Bernal-Llinares, M., Okuda, S., Kawano, S., Moritz, R. L., Carver, J. J., Wang, M., Ishihama, Y., Bandeira, N., Hermjakob, H. et Vizcaíno, J. A. [2017]. The ProteomeXchange consortium in 2017 : supporting the cultural change in proteomics public data deposition. *Nucleic acids research*, **45**(D1), D1100–D1106. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/27924013.

Di Francescomarino, C., Ghidini, C., Maggi, F. M., Petrucci, G. et Yeshchenko, A. [2017]. An Eye into the Future : Leveraging A-priori Knowledge in Predictive Business Process Monitoring. 252–268. Springer, Cham. URL http://link.springer.com/10.1007/978-3-319-65000-5{_}15.

Dondelinger, F., Husmeier, D. et Lèbre, S. [2012]. Dynamic Bayesian networks in molecular plant science : Inferring gene regulatory networks from multiple gene expression time series. *Euphytica*. ISSN 00142336.

Droesbeke, J.-J., Maumy-Bertrand, M., Saporta, G. et Thomas-Agnan, C. [2014]. *Approches statistiques du risque*. Éditions Technip, Paris.

Efron, B. [1979]. Bootstrap methods : another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

Efron, B. et Tibshirani, R. J. [1993]. *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC, 2000 N.W. Corporate Blvd, Boca Raton, Florida 33431, US.

Ériksson, L., Byrne, T., Johansson, E., Trygg, J. et Wikström, C. [2006]. *Multi- and Megavariate Data Analysis Basic Principles and Applications*. Umetrics, Umeå, troisième édition. ISBN 978-91-973730-5-0. URL `https://webshop.umetrics.com/products/multi-and-megavariate-data-analysis-basic-principles-and-applications-third-revised-edition`.

Evermann, J., Rehse, J.-R. et Fettke, P. [2016]. Predicting Process Behaviour using Deep Learning. URL `http://dx.doi.org/10.1016/j.dss.2017.04.003`.

Evermann, J., Rehse, J.-R. et Fettke, P. [2017a]. A Deep Learning Approach for Predicting Process Behaviour at Runtime. 327–338. Springer, Cham. URL `http://link.springer.com/10.1007/978-3-319-58457-7{_}24`.

Evermann, J., Rehse, J.-R. et Fettke, P. [2017b]. XES Tensorflow - Process Prediction using the Tensorflow Deep-Learning Framework. URL `http://arxiv.org/abs/1705.01507`.

Ferrari, S. L. P. et Cribari-Neto, F. [2004]. Beta Regression for Modeling Rates and Proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

Fornecker, L.-M., Muller, L., Bertrand, F., Paul, N., Pichot, A., Herbrecht, R., Chenard, M.-P., Mauvieux, L., Vallat, L., Bahram, S., Cianférani, S., Carapito, R. et Carapito, C. [2018]. Multi-omics dataset to decipher the complexity of drug resistance in diffuse large B-cell lymphoma. en préparation.

Fredon, D., Maumy, M. et Bertrand, F. [2009a]. *Mathématiques L1/L2 : Algèbre/-Géométrie en 30 fiches*. Express Sup. Dunod, Paris.

Fredon, D., Maumy, M. et Bertrand, F. [2009b]. *Mathématiques L1/L2 : Analyse en 30 fiches*. Express Sup. Dunod, Paris.

Fredon, D., Maumy, M. et Bertrand, F. [2009c]. *Mathématiques L1/L2 : Statistique et Probabilités en 30 fiches*. Express Sup. Dunod, Paris.

Friedman, J., Hastie, T. et Tibshirani, R. [2010]. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, **33** (1), 1–22. ISSN 1548-7660. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2929880{&}tool=pmcentrez{&}rendertype=abstract`.

Fu, A. Q., Russell, S., Bray, S. J. et Tavaré, S. [2013]. Bayesian clustering of replicated time-course gene expression data with weak signals. *The Annals of*

Applied Statistics, **7**(3), 1334–1361. ISSN 1941-7330. URL http://projecteuclid.org/euclid.aoas/1380804798.

Fussler, L., Kobes, N., Bertrand, F., Maumy, M., Grosman, J. et Savary, S. [2008]. A characterization of grapevine trunk diseases in France from data generated by the National Grapevine Wood Diseases Survey. *Phytopathology*, **98**(5). ISSN 0031949X.

Gerds, T. A. et Schumacher, M. [2006]. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, **48**, 1029–1040. ISSN 03233847.

Giai Gianetto, Q., Combes, F., Ramus, C., Bruley, C., Couté, Y. et Burger, T. [2016]. Calibration plot for proteomics : A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. *Proteomics*, **16**(1), 29–32. ISSN 1615-9861. URL http://www.ncbi.nlm.nih.gov/pubmed/26572953.

Giroud, S., Blanc, S., Aujard, F., Bertrand, F., Gilbert, C. et Perret, M. [2008]. Chronic food shortage and seasonal modulations of daily torpor and locomotor activity in the grey mouse lemur (Microcebus murinus). *American journal of physiology. Regulatory, integrative and comparative physiology*, **294**(6), R1958–R1967.

Goeminne, L. J. E., Argentini, A., Martens, L. et Clement, L. [2015]. Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics : Performance, Pitfalls, and Data Analysis Guidelines. *Journal of proteome research*, **14**(6), 2457–65. ISSN 1535-3907. URL http://www.ncbi.nlm.nih.gov/pubmed/25827922.

Goeminne, L. J. E., Gevaert, K. et Clement, L. [2016]. Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics. *Molecular & cellular proteomics : MCP*, **15**(2), 657–68. ISSN 1535-9484. URL http://www.mcponline.org.scdrproxy.u-strasbg.fr/content/15/2/657.

Goeminne, L. J., Gevaert, K. et Clement, L. [2017]. Experimental design and data-analysis in label-free quantitative LC/MS proteomics : A tutorial with MSqRob. *Journal of Proteomics*. ISSN 18743919. URL http://www.ncbi.nlm.nih.gov/pubmed/28391044.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. et Lander, E. S. [1999]. Molecular classification of cancer : Class discovery and class prediction by gene expression monitoring. *Science*. ISSN 00368075.

Good, P. [2005]. *Permutation , Parametric and Bootstrap Tests of Hypotheses*. Springer, New York, third édition.

Graf, E., Schmoor, C., Sauerbrei, W. et Schumacher, M. [1999]. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, **18**(17-18), 2529–2545. ISSN 0277-6715. URL `http://www.ncbi.nlm.nih.gov/pubmed/10474158`.

Graybill, W., Chen, M., Chernozhukov, V., Fernandez-Val, I. et Galichon, A. [2011]. *Rearrangement : Monotonize Point and Interval Functional Estimates by Rearrangement*. URL `http://cran.r-project.org/package=Rearrangement`.

Grenèche, J., Krieger, J., Bertrand, F., Erhardt, C., Maumy, M. et Tassi, P. [2011a]. Short-term memory performances during sustained wakefulness in patients with obstructive sleep apnea-hypopnea syndrome. *Brain and Cognition*, **75**(1). ISSN 02782626 10902147.

Grenèche, J., Krieger, J., Bertrand, F., Erhardt, C., Maumy, M. et Tassi, P. [2013]. Effect of continuous positive airway pressure treatment on short-term memory performance over 24h of sustained wakefulness in patients with obstructive sleep apnea-hypopnea syndrome. *Sleep Medicine*, **14**(10). ISSN 13899457 18785506.

Grenèche, J., Krieger, J., Bertrand, F., Erhardt, C., Muzet, A. et Tassi, P. [2011b]. Effect of continuous positive airway pressure treatment on the subsequent EEG spectral power and sleepiness over sustained wakefulness in patients with obstructive sleep apnea-hypopnea syndrome. *Clinical Neurophysiology*, **122**(5). ISSN 13882457.

Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M. et Shan, M.-C. [2004]. Business Process Intelligence. *Computers in Industry*, **53**(3), 321–343. ISSN 0166-3615. URL `https://www.sciencedirect.com/science/article/pii/S0166361503001994?via{%}3Dihub`.

Griss, J., Perez-Riverol, Y., Lewis, S., Tabb, D. L., Dianes, J. A., Del-Toro, N., Rurik, M., Walzer, M. W., Kohlbacher, O., Hermjakob, H., Wang, R. et Vizcaíno, J. A. [2016]. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature methods*, **13**(8), 651–656. ISSN 1548-7105. URL `http://www.ncbi.nlm.nih.gov/pubmed/27493588`.

Grün, B., Kosmidis, I. et Zeileis, A. [2012]. Extended Beta Regression in R : Shaken, Stirred, Mixed and Partitioned. *Journal of Statistical Software*, **48**(11), 1–25.

Gustafsson, M. et Hörnquist, M. [2010]. Gene expression prediction by soft integration and the elastic net - Best performance of the DREAM3 gene expression challenge. *PLoS ONE*. ISSN 19326203.

Gustafsson, M., Hörnquist, M., Lundström, J., Björkegren, J. et Tegnér, J. [2009]. Reverse engineering of gene networks with LASSO and nonlinear basis functions. *Annals of the New York Academy of Sciences*. ISSN 17496632.

Haaland, D. M. et Howland, J. D. T. [1998]. Weighted partial least squares method to improve calibration precision for spectroscopic noise-limited data. Dans *The eleventh international conference on fourier transform spectroscopy*, volume 430 de *AIP Conference Proceedings*, 253–256. American Institute of Physics, Melville.

Hahsler, M., Buchta, C., Gruen, B. et Hornik, K. [2018]. *arules : Mining Association Rules and Frequent Itemsets*. URL `https://CRAN.R-project.org/package=arules`. R package version 1.6-1.

Hahsler, M., Chelluboina, S., Hornik, K. et Buchta, C. [2011]. The arules r-package ecosystem : Analyzing interesting patterns from large transaction datasets. *Journal of Machine Learning Research*, **12**, 1977–1981. URL `http://jmlr.csail.mit.edu/papers/v12/hahsler11a.html`.

Hahsler, M., Gruen, B. et Hornik, K. [2005]. arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, **14**(15), 1–25. ISSN 1548-7660. URL `http://dx.doi.org/10.18637/jss.v014.i15`.

Hall, P. [1992]. *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer New York, New York, NY. ISBN 978-0-387-94508-8. URL `http://link.springer.com/10.1007/978-1-4612-4384-7`.

Hardin, R. H. et Sloane, N. J. A. [1996]. McLaren's improved snub cube and other new spherical designs in three dimensions. *Discrete & Computational Geometry*, **15**(4), 429–441. ISSN 0179-5376. URL `http://link.springer.com/10.1007/BF02711518`.

Hardin, R. et Sloane, N. [1992]. New spherical 4-designs. *Discrete Mathematics*, **106-107**, 255–264. ISSN 0012-365X. URL `https://www.sciencedirect.com/science/article/pii/0012365X9290552Q`.

Hastie, Tibshirani et Friedman [2008]. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. 1–763. ISSN 09641998.

Hastie, Tibshirani et Friedman [2009]. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. 1–763. ISSN 09641998.

Heagerty, P. J., Lumley, T. et Pepe, M. S. [2000a]. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, **56**(2), 337–344. ISSN 0006-341X.

Heagerty, P. J., Lumley, T. et Pepe, M. S. [2000b]. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. ISSN 0006341X.

Heagerty, P. J. et packaging by Paramita Saha-Chaudhuri [2012]. *risksetROC : Riskset ROC curve estimation from censored survival data*. URL `http://CRAN.R-project.org/package=risksetROC`. R package version 1.0.4.

Heagerty, P. J. et Zheng, Y. [2005a]. Survival model predictive accuracy and ROC curves. *Biometrics*. ISSN 0006341X.

Heagerty, P. J. et Zheng, Y. [2005b]. Survival model predictive accuracy and ROC curves. *Biometrics*, **61**(1), 92–105. ISSN 0006341X.

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E. et Guthke, R. [2009]. Gene regulatory network inference : data integration in dynamic models ?a review. *Biosystems*, **96**(1), 86–103.

Hielscher, T., Zucknick, M., Werft, W. et Benner, A. [2010]. On the Prognostic Value of Gene Expression Signatures for Censored Data. Dans Fink, A., Lausen, B., Seidel, W. et Ultsch, A., éditeurs. *Advances in Data Analysis, Data Handling and Business Intelligence Studies in Classification, Data Analysis, and Knowledge Organization*, 663–673. Springer, Berlin. ISBN 978-3-642-01043-9.

Hochreiter, S. et Schmidhuber, J. [1997]. Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780. ISSN 0899-7667. URL `http://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735`.

Hoerl, A. E. et Kennard, R. W. [1970]. Ridge Regression : Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**(1), 55. ISSN 00401706. URL `https://www.jstor.org/stable/1267351?origin=crossref`.

Höskuldsson, A. [1988]. PLS regression methods. *Journal of Chemometrics*, **2**(3), 211–228.

Hothorn, T. [2018]. *trtf : Transformation Trees and Forests*. URL `https://CRAN.R-project.org/package=trtf`. R package version 0.3-3.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. et van der Laan, M. J. [2006]. Survival ensembles. *Biostatistics (Oxford, England)*, **7**(3), 355–73. ISSN 1465-4644. URL `http://biostatistics.oxfordjournals.org/content/7/3/355.full`.

Hothorn, T., Lausen, B., Benner, A. et Radespiel-Tröger, M. [2004]. Bagging survival trees. *Statistics in medicine*, **23**(1), 77–91. ISSN 0277-6715. URL `http://www.ncbi.nlm.nih.gov/pubmed/14695641`.

Hothorn, T. et Zeileis, A. [2017]. Transformation Forests. URL `http://arxiv.org/abs/1701.02110`.

Hsiao, Y.-T. et Lee, W.-P. [2012]. Inferring robust gene networks from expression data by a sensitivity-based incremental evolution method. *BMC bioinformatics*, **13 Suppl 7**(Suppl 7), S8. ISSN 1471-2105. URL http://www.ncbi.nlm.nih.gov/pubmed/22595005.

Hung, H. et Chiang, C. T. [2010]. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, **38**, 8–26. ISSN 03195724. URL <GotoISI>://000276598600003.

Jagannathan, R. et Ma, T. [2003]. Risk Reduction in Large Portfolios : Why Imposing the Wrong Constraints Helps. *The Journal of Finance*, **58**(4), 1651–1683. ISSN 00221082. URL http://doi.wiley.com/10.1111/1540-6261.00580.

Johnson, N. L., Kotz, S. et Balakrishnan, N. [1995]. *Continuous Univariate Distributions*, volume 2. Wiley, New York, 2nd édition.

Jung, N., Bertrand, F., Bahram, S., Vallat, L. et Bertrand, M. [2014a]. Cascade : a r-package to study, predict and simulate the diffusion of a signal through a temporal gene network. Dans *Proceedings of User2014 !, Los Angeles*, 153.

Jung, N., Bertrand, F., Bahram, S., Vallat, L. et Maumy-Bertrand, M. [2014b]. *Additional application of the Cascade package to E-MTAB-1475 dataset*. Vignette of the package.

Jung, N., Bertrand, F., Bahram, S., Vallat, L. et Maumy-Bertrand, M. [2014c]. Cascade : A R package to study, predict and simulate the diffusion of a signal through a temporal gene network. *Bioinformatics*. ISSN 13674803.

Jung, N., Bertrand, F., Bahram, S., Vallat, L. et Maumy-Bertrand, M. [2014d]. *Introduction to the Cascade package with application to the GSE39411 dataset*. Vignette of the package.

Jung, N., Bertrand, F. et Maumy-Bertrand, M. [2015]. AcSel : selecting variables with accuracy in correlated datasets. URL http://arxiv.org/abs/1512.03307.

Jung, N., Bertrand, F., Maumy-Bertrand, M. et Vallat, L. [2018]. *Selection, Reverse-Engineering and Prediction in Cascade networks*. URL http://www-irma.u-strasbg.fr/~fbertran/. R package version 1.1.

Katharopoulos, A. et Fleuret, F. [2017]. Biased Importance Sampling for Deep Neural Network Training. URL http://arxiv.org/abs/1706.00043.

Kim, J., Kim, Y. et Kim, Y. [2008]. A Gradient-Based Optimization Algorithm for LASSO. *Journal of Computational and Graphical Statistics*, **17**(4), 994–1009. ISSN 1061-8600. URL http://www.tandfonline.com/doi/abs/10.1198/106186008X386210.

Kobes, N., Fussler, L., Pleynet, M., Savary, S., Bertrand, F. et Maumy, M. [2007]. Vignes, maladies du bois, des facteurs clefs. Premiers résultats de l'analyse statistique des données de l'Observatoire national. *PHYTOMA - La Défense des Végétaux*, **604**, 33–37.

Kosmidis, I. et Firth, D. [2010]. A Generic Algorithm for Reducing Bias in Parametric Estimation. *Journal of Chemometrics*, **4**, 1097–1112.

Kowarik, A. et Templ, M. [2016]. Imputation with the R package VIM. *Journal of Statistical Software*, **74**(7), 1–16.

Kraemer, N. et Sugiyama, M. [2011]. The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association*, **106**(494), 697–705.

Kuhn., M. [2018]. *caret : Classification and Regression Training*. URL `https://CRAN.R-project.org/package=caret`. R package version 6.0-80.

Kumar, L. et E Futschik, M. [2007]. Mfuzz : a software package for soft clustering of microarray data. *Bioinformation*, **2**(1), 5–7. ISSN 0973-2063. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2139991{&}tool=pmcentrez{&}rendertype=abstract`.

Kuntzmann, P., Barbe, J., Maumy-Bertrand, M. et Bertrand, F. [2013]. Late harvest as factor affecting esca and Botryosphaeria dieback prevalence of vineyards in the Alsace region of France. *Vitis - Journal of Grapevine Research*, **52**(4). ISSN 00427500.

Lambert-Lacroix, S. et Letué, F. [2011]. Partial Least Squares and Cox model with application to gene expression. URL `http://hal.archives-ouvertes.fr/hal-00568234`. Technical report.

Law, C. W., Chen, Y., Shi, W. et Smyth, G. K. [2014]. Voom : precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, **15** (2), R29. ISSN 1465-6914. URL `http://genomebiology.com/2014/15/2/R29`.

Lazraq, A., Cleroux, R. et Gauchi, J.-P. [2003]. Selecting both latent and explanatory variables in the PLS1 regression model. *Chemometrics and Intelligent Laboratory Systems*, **66**(2), 117–126.

Lê Cao, K., Rossouw, D., Robert-Granié, C. et Besse, P. [2008]. A Sparse PLS for Variable Selection when Integrating Omics data. *Stat Appl Genet Mol Biol*, **7**, Article 35.

Lê Cao, K.-A., Boitard, S. et Besse, P. [2011]. Sparse PLS discriminant analysis : biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*. ISSN 1471-2105.

Lê Cao, K.-A. A., Martin, P. G. G., Robert-Granié, C. et Besse, P. [2009]. Sparse canonical methods for biological data integration : Application to a

cross-platform study. *BMC Bioinformatics*, **10**(1), 34. ISSN 14712105. URL http://www.biomedcentral.com/1471-2105/10/34.

LeCun, Y., Bengio, Y. et Hinton, G. [2015]. Deep learning. *Nature*, **521**(7553), 436–444. ISSN 0028-0836. URL http://www.nature.com/articles/nature14539.

Letac, G. et Massam, H. [2004]. A tutorial on the non central Wishart distribution. URL http://www.math.univ-toulouse.fr/~letac/Wishartnoncentrales.pdf.

Li, B., Morris, J. et Martin, E. [2002a]. Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **64**, 79–89.

Li, B., Morris, J. et Martin, E. B. [2002b]. Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*. ISSN 01697439.

Li, C. et Wong, W. H. [2001]. Model-based analysis of oligonucleotide arrays : expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(1), 31–6. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=14539{&}tool=pmcentrez{&}rendertype=abstract.

Li, L. [2006]. Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*, **22**(4), 466–471. ISSN 13674803.

Li, Q., Birkbak, N. J., Gyorffy, B., Szallasi, Z. et Eklund, A. C. [2011]. Jetset : selecting the optimal microarray probe set to represent a gene. *BMC bioinformatics*, **12**(1), 474. ISSN 1471-2105. URL http://www.biomedcentral.com/1471-2105/12/474.

Li, W., Turner, A., Aggarwal, P., Matter, A., Storvick, E., Arnett, D. K. et Broeckel, U. [2015]. Comprehensive evaluation of AmpliSeq transcriptome, a novel targeted whole transcriptome RNA sequencing methodology for global gene expression analysis. *BMC Genomics*, **16**, 1069. ISSN 1471-2164. URL http://www.biomedcentral.com/1471-2164/16/1069.

Liang, S., Fuhrman, S. et Somogyi, R. [1998]. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, 18–29. ISSN 1793-5091. URL http://www.ncbi.nlm.nih.gov/pubmed/9697168.

Little, R. J. A. et Rubin, D. [2002]. *Statistical analysis with missing data*. ISBN 9780471183860. URL https://www.wiley.com/en-us/Statistical+Analysis+with+Missing+Data{%}2C+2nd+Edition-p-9780471183860.

Liu, Z.-P., Wu, C., Miao, H. et Wu, H. [2015]. RegNetwork : an integrated database of transcriptional and post-transcriptional regulatory networks in hu-

man and mouse. *Database*, **2015**(August 2018), bav095. ISSN 1758-0463. URL `https://academic.oup.com/database/article-lookup/doi/10.1093/database/bav095`.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A. et Gerstein, M. [2004]. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**(7006), 308–312.

Magnanensi, J., Bertrand, F., Maumy-Bertrand, M. et Meyer, N. [2017]. A new universal resample-stable bootstrap-based stopping criterion for PLS component construction. *Statistics and Computing*, **27**(3). ISSN 15731375.

Magnanensi, J., Maumy-Bertrand, M., Meyer, N. et Bertrand, F. [2016a]. A new Bootstrap-Based stopping criterion in PLS components construction. Dans *Springer Proceedings in Mathematics and Statistics*, volume 173.

Magnanensi, J., Maumy-Bertrand, M., Meyer, N. et Bertrand, F. [2016b]. New developments in Sparse PLS regression. URL `http://arxiv.org/abs/1601.03281`.

Manne, R. [1987]. Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, **2**(1), 187–197.

Maple Team [2014]. *Maple 18*. Maplesoft, a division of Waterloo Maple Inc., Waterloo.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Aderhold, A., Allison, K. R., Bonneau, R., Camacho, D. M., Chen, Y., Collins, J. J., Cordero, F., Costello, J. C., Crane, M., Dondelinger, F., Drton, M., Esposito, R., Foygel, R., de la Fuente, A., Gertheiss, J., Geurts, P., Greenfield, A., Grzegorczyk, M., Haury, A.-C., Holmes, B., Hothorn, T., Husmeier, D., Huynh-Thu, V. A., Irrthum, A., Kellis, M., Karlebach, G., Küffner, R., Lèbre, S., De Leo, V., Madar, A., Mani, S., Marbach, D., Mordelet, F., Ostrer, H., Ouyang, Z., Pandya, R., Petri, T., Pinna, A., Poultney, C. S., Prill, R. J., Rezny, S., Ruskin, H. J., Saeys, Y., Shamir, R., Sîrbu, A., Song, M., Soranzo, N., Statnikov, A., Stolovitzky, G., Vega, N. N. M., Vera-Licona, P., Vert, J.-P., Visconti, A., Wang, H., Wehenkel, L., Windhager, L., Zhang, Y., Zimmer, R., Kellis, M., Collins, J. J. et Stolovitzky, G. [2012]. Wisdom of crowds for robust gene network inference. *Nat Methods*. ISSN 1548-7091.

Mardia, K. V., Kent, J. T. et Bibby, J. M. [1979]. *Multivariate Analysis*. Academic Press, London. ISBN 0124712525.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D. et Califano, A. [2006]. ARACNE : An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*. ISSN 14712105.

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J. et Apweiler, R. [2005]. PRIDE : The proteomics identifications database. *PROTEOMICS*, **5**(13), 3537–3545. ISSN 1615-9853. URL http://doi.wiley.com/10.1002/pmic.200401303.

Maumy-Bertrand, M., Saporta, G. et Thomas-Agnan, C. [2018]. *Apprentissage statistique et données massives*. Éditions Technip, Paris.

McCarthy, D. J., Chen, Y. et Smyth, G. K. [2012]. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, **40**(10), 4288–97. ISSN 1362-4962. URL http://nar.oxfordjournals.org/content/40/10/4288.

McKean, J. W. et Sievers, G. L. [1987]. Coefficients of determination for least absolute deviation analysis. *Statistics & Probability Letters*, **5**(1), 49–54. ISSN 01677152. URL http://www.sciencedirect.com/science/article/pii/0167715287900265.

Mee, R. W. [1988]. Estimation of the percentage of a normal distribution lying outside a specified interval. *Communications in Statistics - Theory and Methods*, **17**(5), 1465–1479. ISSN 0361-0926. URL http://www.tandfonline.com/doi/abs/10.1080/03610928808829692.

Meinshausen, N. [2006]. Quantile Regression Forests. *Journal of Machine Learning Research*. ISSN 15410420.

Meinshausen, N. [2017]. *quantregForest : Quantile Regression Forests*. URL https://CRAN.R-project.org/package=quantregForest. R package version 1.3-7.

Meinshausen, N. et Bühlmann, P. [2010]. Stability selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **72**(4), 417–473. ISSN 13697412. URL http://doi.wiley.com/10.1111/j.1467-9868.2010.00740.x.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. et Teller, E. [1953]. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092. ISSN 0021-9606. URL http://aip.scitation.org/doi/10.1063/1.1699114.

Meyer, N., Fredon, D., Maumy-Bertrand, M. et Bertrand, F. [2018]. *Toute l'UE4 en fiches. Evaluation des méthodes d'analyse appliquées aux sciences de la vie et de la santé*. Dunod, Paris, 3e édition.

Meyer, N., Maumy-Bertrand, M. et Bertrand, F. [2010]. Comparaison de variantes de régressions logistiques PLS et de régression PLS sur variables qualitatives : application aux données d'allélotypage. *Journal de la Société Française de Statistique*, **151**(2), 1–18.

Naes, T. et Martens, H. [1985]. Comparison of prediction methods for multicollinear data. *Communications in Statistics – Simulation and Computation*, **14**, 545–576.

Neil, D., Pfeiffer, M. et Liu, S.-C. [2016]. Phased LSTM : Accelerating Recurrent Network Training for Long or Event-based Sequences. Dans *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3889–3897, USA. Curran Associates Inc. ISBN 978-1-5108-3881-9. URL `http://dl.acm.org/citation.cfm?id=3157382.3157532`.

Nengsih, T. A., Bertrand, F., Maumy-Bertrand, M. et Meyer, N. [2018]. Determining the Number of Components in PLS Regression on Incomplete Data. URL `http://arxiv.org/abs/1810.08104`.

Ning, Z., Zhang, X., Mayne, J. et Figeys, D. [2016]. Peptide-Centric Approaches Provide an Alternative Perspective To Re-Examine Quantitative Proteomic Data. *Analytical chemistry*, **88**(4), 1973–1978. ISSN 1520-6882. URL `http://dx.doi.org/10.1021/acs.analchem.5b04148`.

Papachristodoulou, A., adn G. Valmorbida, J. A., Prajna, S., Seiler, P. et Parrilo, P. A. [2013]. *SOSTOOLS : Sum of squares optimization toolbox for MATLAB*. `http://arxiv.org/abs/1310.4716`. Available from `http://www.cds.caltech.edu/sostools`, `http://www.mit.edu/~parrilo/sostools` and `http://www.eng.ox.ac.uk/control/sostools`.

Pearl, J. [2000]. *Causality : Models, Reasoning, and Inference*. Cambridge University Press. 2nd edition : 2009.

Pearl, J. [2009]. Causal inference in statistics : an overview. *Statistics Surveys*, **3**, 96–146.

Pearl, J. et Mackenzie, D. [2018]. *The Book of Why : The New Science of Cause and Effect*. Basic Books.

Perrot, A., Pionneau, C., Nadaud, S., Davi, F., Leblond, V., Jacob, F., Merle-Béral, H., Herbrecht, R., Béné, M.-C., Gribben, J. G., Bahram, S. et Vallat, L. [2011]. A unique proteomic profile on surface IgM ligation in unmutated chronic lymphocytic leukemia. *Blood*, **118**(4), e1–15. ISSN 1528-0020. URL `http://www.ncbi.nlm.nih.gov/pubmed/21602524`.

Perry, P. O. [2015]. *bcv : Cross-Validation for the SVD (Bi-Cross-Validation)*. R package version 1.0.1.

Petit, O., Bertrand, F. et Thierry, B. [2008]. Social play in crested and Japanese Macaques : Testing the covariation hypothesis. *Developmental Psychobiology*, **50**(4), 399–407.

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S. et Smyth, G. K. [2016]. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The Annals of Applied Statistics*, **10**(2), 946–963. ISSN 1932-6157. URL http://projecteuclid.org/euclid.aoas/1469199900.

Piotto, M., Moussallieh, F.-M., Neuville, A., Bellocq, J.-P., Elbayed, K. et Namer, I. J. [2012]. Towards real-time metabolic profiling of a biopsy specimen during a surgical operation by 1H high resolution magic angle spinning nuclear magnetic resonance : a case report. *Journal of Medical Case Reports*, **6**(1).

Potapov, S., Adler, W. et Schmid, M. [2012]. *survAUC : Estimators of prediction accuracy for time-to-event data*. URL http://CRAN.R-project.org/package=glmpath. R package version 1.0-5.

Radespiel-Tröger, M., Rabenstein, T., Schneider, H. et Lausen, B. [2003]. Comparison of tree-based methods for prognostic stratification of survival data. *Artificial Intelligence in Medicine*, **28**(3), 323–341. ISSN 09333657. URL http://www.aiimjournal.com/article/S0933365703000605/fulltext.

Rau, A., Gallopin, M., Celeux, G. et Jaffrézic, F. [2013]. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, **29**(17), 2146–2152.

Rau, A., Jaffrézic, F., Foulley, J. L. et Doerge, R. W. [2012]. Reverse engineering gene regulatory networks using approximate Bayesian computation. *Statistics and Computing*. ISSN 09603174.

Rau, A., Jaffrézic, F., Foulley, J. L. et Doerge, R. W. [2010]. An empirical bayesian method for estimating biological networks from temporal microarray data. *Statistical Applications in Genetics and Molecular Biology*. ISSN 15446115.

Reznick, B. [1995]. Some constructions of spherical 5-designs. *Linear Algebra and its Applications*, **226-228**, 163–196. ISSN 0024-3795. URL https://www.sciencedirect.com/science/article/pii/002437959500101V.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. et Smyth, G. K. [2015]. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7), e47.

Robinson, M. D., McCarthy, D. J. et Smyth, G. K. [2010]. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140. ISSN 1367-4811. URL http://www.ncbi.nlm.nih.gov/pubmed/19910308.

Rolland, A., Bertrand, F., Maumy, M. et Jacquet, S. [2009]. Assessing phytoplankton structure and spatio-temporal dynamics in a freshwater ecosystem using a powerful multiway statistical analysis. *Water Research*, **43**(13). ISSN 00431354.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., López-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. et Staudt, L. M. [2002]. The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. *New England Journal of Medicine*, **346**(25), 1937–1947. ISSN 0028-4793. URL `http://www.nejm.org/doi/abs/10.1056/NEJMoa012914`.

Rosipal, R. et Trejo, L. J. [2002]. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. *Journal of Machine Learning Research*, **2**, 97–123. ISSN 15324435. URL `http://www.crossref.org/jmlr{_}DOI.html`.

Schemper, M. et Henderson, R. [2000]. Predictive accuracy and explained variation in Cox regression. *Biometrics*, **56**(1), 249–255. ISSN 0006-341X. URL `http://www.ncbi.nlm.nih.gov/pubmed/10783803`.

Schleiss, C., Ilias, W., Tahar, O., Guler, Y., Miguet, L., Mayeur-Rousse, C., Mauvieux, L., Fornecker, L.-M., Toussaint, E., Herbrecht, R., Bertrand, F., Maumy-Bertrand, M., Martin, T., Fournel, S., Georgel, P., Bahram, S. et Vallat, L. [2018a]. BCR-associated factors driving chronic lymphocytic leukemia cells proliferation ex vivo. *Scientific Reports*, **in print**.

Schleiss, C., Muller, L., Paul, N., Carapito, C., Mauvieux, L., Herbrecht, R., Carapito, R., Maumy-Bertrand, M., Georgel, P., Bahram, S., Bertrand, F. et Vallat, L. [2018b]. Characterization of proliferative program induced by antigen receptor (BCR) stimulation in aggressive form CLL lymphocytes. en préparation.

Schmid, M., Hielscher, T., Augustin, T. et Gefeller, O. [2011]. A Robust Alternative to the Schemper-Henderson Estimator of Prediction Error. *Biometrics*, **67**, 524–535. ISSN 0006341X.

Schröder, M. S., Culhane, A. C., Quackenbush, J. et Haibe-Kains, B. [2011]. survcomp : An R/BioconductoR package for performance assessment and comparison of survival models. *Bioinformatics*, **27**(22), 3206–3208. ISSN 13674803.

Schumacher, M., Binder, H. et Gerds, T. [2007]. Assessment of survival prediction models based on microarray data. *Bioinformatics (Oxford, England)*, **23**(14),

1768–74. ISSN 1367-4811. URL http://bioinformatics.oxfordjournals.org/content/23/14/1768.full.

Simas, A. B., Barreto-Souza, W. et Rocha, A. V. [2010]. Improved Estimators for a General Class of Beta Regression Models. *Computational Statistics & Data Analysis*, **54**(2), 348–366.

Simon, N., Friedman, J., Hastie, T. et Tibshirani, R. [2011]. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, **39**(5), 1–13. ISSN 15487660. URL http://www.jstatsoft.org/v39/i05/.

Smyth, G. K. [2004]. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**(1), 1–25. ISSN 1544-6115. URL http://www.ncbi.nlm.nih.gov/pubmed/16646809.

Sohn, I., Kim, J., Jung, S.-H. et Park, C. [2009]. Gradient lasso for Cox proportional hazards model. *Bioinformatics*, **25**(14), 1775–1781. ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp322.

Song, X. et Zhou, X.-h. [2008]. A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statistica Sinica*, **18**, 947–965. ISSN 10170405.

Suomi, T., Corthals, G. L., Nevalainen, O. S. et Elo, L. L. [2015]. Using Peptide-Level Proteomics Data for Detecting Differentially Expressed Proteins. *Journal of Proteome Research*, **14**(11), 4564–4570. ISSN 1535-3893. URL http://pubs.acs.org/doi/10.1021/acs.jproteome.5b00363.

Tax, N., Verenich, I., La Rosa, M. et Dumas, M. [2017]. Predictive Business Process Monitoring with LSTM Neural Networks BT. Dans Dubois, E. et Pohl, K., éditeurs. *Advanced Information Systems Engineering*, 477–492, Cham. Springer International Publishing. ISBN 978-3-319-59536-8. URL http://dx.doi.org/10.1007/978-3-319-59536-8{_}30.

Tenenhaus, A., Giron, A., Viennet, E., Béra, M., Saporta, G. et Fertil, B. [2007]. Kernel logistic PLS : A tool for supervised nonlinear dimensionality reduction and binary classification. *Computational Statistics and Data Analysis*, **51**(9), 4083–4100. ISSN 01679473.

Tenenhaus, M. [1998]. *La régression PLS, Théorie et pratique*. Technip, Paris.

Tenenhaus, M. [1999]. La regression logistique PLS. Dans *Proceedings of the 32èmes journées de Statistique de la Société française de Statistique*, 721–723. FES.

Tenenhaus, M. [2005]. La régression logistique PLS. Dans Droesbeke, J.-J., Lejeune, M. et Saporta, G., éditeurs. *Modèles statistiques pour données qualitatives*, 263–275. Technip, Paris.

Thode, H. C. [2002]. *Testing for normality*. Marcel Dekker. ISBN 9780824796136. URL https://www.crcpress.com/Testing-For-Normality/Thode/p/book/9780824796136.

Tibshirani, R. [1996]. Regression Selection and Shrinkage via the Lasso. ISSN 00359246.

Tibshirani, R. [1997]. The lasso method for variable selection in the cox model. *Statistics in Medicine*. ISSN 02776715.

Tibshirani, R. [2011]. Regression shrinkage and selection via the lasso : A retrospective. *Journal of the Royal Statistical Society. Series B : Statistical Methodology*. ISSN 13697412.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M. et Cox, J. [2016]. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, **13**(9), 731–740. ISSN 1548-7091. URL http://www.nature.com/articles/nmeth.3901.

Uno, H., Cai, T., Tian, L. et WEI, L.-J. [2007]. Evaluating Prediction Rules for t-Year Survivors With Censored Regression Models. *Journal of the American Statistical Association*, **102**, 527–537. ISSN 0162-1459. URL papers2://publication/doi/10.1198/016214507000000149.

Vallat, L., Kemper, C. a., Jung, N., Maumy-Bertrand, M., Bertrand, F., Meyer, N., Pocheville, A., Fisher, J. W., Gribben, J. G. et Bahram, S. [2013]. Reverse-engineering the genetic circuitry of a cancer cell with predicted intervention in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(2), 459–64. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3545767{&}tool=pmcentrez{&}rendertype=abstract.

Vallat, L. D., Park, Y., Li, C., Gribben, J. G., Malavasi, F., Cosulich, M. E. et Ferrarini, M. [2007]. Temporal genetic program following B-cell receptor cross-linking : altered balance between proliferation and death in healthy and malignant B cells. *Blood*, **109**(9), 3989–97. ISSN 0006-4971. URL http://www.ncbi.nlm.nih.gov/pubmed/10666191.

van Buuren, S. et Groothuis-Oudshoorn, K. [2011]. mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**(3), 1–67. ISSN 1548-7660. URL http://www.jstatsoft.org/v45/i03/.

Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T.,

Blickle, T., Bose, J. C., Van Den Brand, P., Brandtjen, R., Buijs, J. *et al.* [2011]. Process mining manifesto. Dans *International Conference on Business Process Management*, 169–194. Springer.

van der Aalst, W. M., Low, W. Z., Wynn, M. T. et ter Hofstede, A. H. [2015]. Change your history : Learning from event logs to improve processes. Dans *2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 7–12. IEEE. ISBN 978-1-4799-2002-0. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7230925`.

van der Hilst, R. D., de Hoop, M. V., Wang, P., Shim, S.-H., Ma, P. et Tenorio, L. [2007]. Seismostratigraphy and thermal structure of Earth's core-mantle boundary region. *Science (New York, N.Y.)*, **315**(5820), 1813–7. ISSN 1095-9203. URL `http://www.ncbi.nlm.nih.gov/pubmed/17395822`.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J. et Wessels, L. F. A. [2006]. Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, **25**(18), 3201–3216. ISSN 02776715.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. et Sharan, R. [2010]. Associating Genes and Protein Complexes with Disease via Network Propagation. *PLoS Computational Biology*, **6**(1), e1000641. ISSN 1553-7358. URL `http://dx.plos.org/10.1371/journal.pcbi.1000641`.

Verweij, P. J. et Van Houwelingen, H. C. [1993]. Cross-validation in survival analysis. *Statistics in medicine*, **12**(24), 2305–2314. ISSN 0277-6715.

Vizcaíno, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.-W., Wang, R. et Hermjakob, H. [2016]. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research*, **44**(D1), D447–D456. ISSN 0305-1048. URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1145`.

Wachter, A. et Beißbarth, T. [2015]. pwOmics : an R package for pathway-based integration of time-series omics data using public database knowledge. *Bioinformatics (Oxford, England)*, **31**(18), 3072–4. ISSN 1367-4811. URL `http://bioinformatics.oxfordjournals.org/content/31/18/3072`.

Wakeling, I. N. et Morris, J. J. [1993]. A test of significance for partial least squares regression. *Journal of Chemometrics*, **7**(4), 291–304.

Weber, J. C., Meyer, N., Pencreach, E., Schneider, A., Guérin, E., Neuville, A., Stemmer, C., Brigand, C., Bachellier, P., Rohr, S., Kedinger, M., Meyer, C., Guenot, D., Oudet, P., Jaeck, D. et Gaub, M. P. [2007]. Allelotyping analyses of synchronous primary and metastasis CIN colon cancers identified different subtypes. *International Journal of Cancer*. ISSN 00207136.

White, I. R., Royston, P. et Wood, A. M. [2011]. Multiple imputation using chained equations : Issues and guidance for practice. *Statistics in Medicine*. ISSN 02776715.

Wiklund, S., Nilsson, D., Eriksson, L., Sjöström, M., Wold, S. et Faber, K. [2007]. A randomization test for PLS component selection. *Journal of Chemometrics*, **21**(10-11), 427–439. ISSN 08869383. URL `http://doi.wiley.com/10.1002/cem.1086`.

Withers, C. S. [1983]. Expansions for the Distribution and Quantiles of a Regular Functional of the Empirical Distribution with Applications to Nonparametric Confidence Intervals. *The Annals of Statistics*, **11**(2), 577–587.

Wold, H. [1966]. Estimation of Principal Components and Related Models by Iterative Least Squares. Dans Krishnaiah, P. R., éditeur. *Multivariate Analysis*, 391–420. Academic Press, New York.

Wold, S., Martens, H. et Wold, H. [1983]. The multivariate calibration problem in chemistry solved by the PLS method. *Proc. Conf. Matrix pencils*, 286–293.

Wold, S., Ruhe, A., Wold, H. et Dunn, III, W. J. [1984]. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, **5**(3), 735–743. ISSN 0196-5204. URL `http://epubs.siam.org/doi/abs/10.1137/0905052`.

Wold, S., Sjöström, M. et Eriksson, L. [2001]. PLS-regression : a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, **58**(2), 109–130. ISSN 01697439.

Yang, P., Li, X., Wu, M., Kwoh, C.-K. et Ng, S.-K. [2011]. Inferring Gene-Phenotype Associations via Global Protein Complex Network Propagation. *PLoS ONE*, **6**(7), e21502. ISSN 1932-6203. URL `http://dx.plos.org/10.1371/journal.pone.0021502`.

Yosef, N. et Regev, A. [2011]. Impulse control : temporal dynamics in gene transcription. *Cell*, **144**(6), 886–896.

Yu, G., Wang, L.-G., Han, Y. et He, Q.-Y. [2012]. clusterProfiler : an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS : A Journal of Integrative Biology*, **16**(5), 284–287. ISSN 1536-2310.

Zhu, J. J., Santarius, T., Wu, X. Y., Tsong, J., Guha, A., Wu, J. K., Hudson, T. J. et Black, P. M. [1998]. Screening for loss of herterozygosity and microsatellite instability in oligodendrogliomas. *GENES CHROMOSOMES & CANCER*. ISSN 1045-2257.

Zimmer, C., Boos, M., Bertrand, F., Robin, J.-P. et Petit, O. [2011]. Behavioural

adjustment in response to increased predation risk : A study in three duck species. *PLoS ONE*, **6**(4). ISSN 19326203.

Zou, H. et Hastie, T. [2005]. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **67**(2), 301–320. ISSN 1369-7412. URL `http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x`.

Les travaux et activités de recherche présentés dans ce mémoire se sont concentrés autour de plusieurs thématiques fédératrices fortes.

La plus importante d'entre elles s'avère être l'inférence statistique dans un contexte de grande ou d'ultra grande dimension en présence éventuelle de censure ou de données manquantes. Cette thématique traite aussi bien des cas de la régression linéaire, de la régression linéaire généralisée que d'autres contextes de régression, comme la régression bêta ou le modèle de Cox, pour lesquels ont été produits non seulement de nouveaux critères de choix de modèle mais aussi de sélection de variables. Les contextes d'application envisagés ont naturellement incité à l'utilisation d'approches de régression pénalisée, typiquement *lasso*, *ridge* ou *elasticnet*, de régression par les moindres carrés partiels parcimonieuse ou non.

La suite du mémoire décrit plusieurs approches de modélisation de données génomiques et protéomiques, la plus aboutie d'entre elles permettant la modélisation conjointe des gènes et des protéines au niveau intra et inter-individuel afin de décoder plus efficacement les réseaux biologiques mettant en jeu ces acteurs, par exemple ceux présents dans des cellules cancéreuses.

À ces sujets de recherche, s'ajoute depuis un an une thématique supplémentaire liée à l'apprentissage statistique et aux réseaux de neurones. Pour chacune de ces thématiques, plusieurs perspectives de recherches, aussi bien théoriques que plus appliquées, sont détaillées à la fin du mémoire.

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE
UMR 7501
Université de Strasbourg et CNRS
7 Rue René Descartes
67 084 STRASBOURG CEDEX

Tél. 03 68 85 01 29
Fax 03 68 85 03 28

https://irma.math.unistra.fr
irma@math.unistra.fr

Institut de Recherche Mathématique Avancée

IRMA 2018/010
ISSN 0755-3390
https ://tel.archives-ouvertes.fr/tel-XXX