Apprentissage machine pour l'intégration et l'interaction de données structurées

Marie Szafranski

Université Paris-Saclay

École Doctorale de Mathématiques Hadamard

Spécialité : mathématiques aux interfaces

Habilitation à Diriger les Recherches

présentée par

Marie Szafranski

Apprentissage machine pour l'intégration et l'interaction de données structurées

The adventure

`Struct ured {int* a_tions};`

in Dataland

Présentée et soutenue publiquement le 20/12/2023 devant le jury composé de :

Mathilde Mougeot	PU, ENSIIE	*Présidente*
Stéphane Canu	PU, INSA Rouen	*Rapporteur*
Marie-Laure Martin	DR, INRAE	*Rapporteure*
Nathalie Vialaneix	DR, INRAE	*Rapporteure*
Christophe Ambroise	PU, Université d'Évry Val d'Essonne	*Examinateur*
Cécile Capponi	PU, Aix-Marseille Université	*Examinatrice*
Liva Ralaivola	VP Recherche, Criteo	*Examinateur*

À Ayô et Lisa,
À Matthieu, Maxandre et Maorèn.

Remerciements

Acknowledgement

J'adresse tout d'abord mes très sincères remerciements à Stéphane Canu, Marie-Laure Martin et Nathalie Vialaneix qui ont accepté de rapporter cette habilitation à diriger des recherches, ainsi qu'à Christophe Ambroise, Cécile Capponi, Mathilde Mougeot et Liva Ralaivola qui ont accepté de faire partie du jury.

Côté recherche, je souhaite en particulier remercier Florent Guinot et Kylliann De Santiago avec qui le travail a été ou est particulièrement agréable ainsi que les jeunes chercheur·e·s ou étudiant·e·s avec qui j'ai eu le plaisir de collaborer. Je tiens à remercier à nouveau Christophe Ambroise pour m'avoir chaleureusement accueillie dans l'équipe Statistique et Génome et donné l'opportunité de participer aux encadrements de thèse de Florent et Kylliann.

Côté ensiie, je remercie celles et ceux dont les qualités sont à la hauteur de leur efficacité, en particulier Gaël Thomas qui rend beaucoup de choses possibles, Dimitri Watel et sa patience inouïe, Christophe Mouilleron et son sens de l'à-propos permettant de redresser nombre de situations bancales, Julien Forest et ses (im-)partialités toujours justes et constructives ainsi que son aide et sa présence toujours bien dosée. Merci à Christophe Mouilleron et Valentin Honoré pour les réglages ! Merci aussi à Xavier Urbain et son héritage oursement profitable.

Je remercie également les nombreuses personnes qui m'ont aidée dans les démarches et la préparation liée à cette habilitation et celles venues m'encourager le jour J, parfois de loin, ainsi que celles ayant envoyé un message.

Enfin, je remercie infiniment ma famille pour son soutien sans faille et sa patience immense pendant ce (trop) long processus [1]. Merci à Élisabeth et Georges d'être depuis toujours des parents si attentionnés et soutenants au quotidien. Merci à Nicole et Patrick pour votre accueil très régulier et généreux à Soulac et les moments apaisants. Merci à Maxandre et Maorèn d'être chaque jour une source de joie et d'énergie. Merci à Matthieu, pour tout.

1. Et je ne remercie vraiment pas la COVID-19, ni les politiques successives d'attractivité de la petite enfance ou de l'éducation nationale, entre autres, qui ont permis de normaliser durablement le #télétravailavecteskids…

Préambule

Underground

Parcours professionnel

Ma formation se situe à l'interface de l'informatique et des mathématiques appliquées. J'ai d'abord obtenu un diplôme d'ingénieur en *génie informatique* à l'Université de Technologie de Compiègne (UTC) en parallèle d'un diplôme de master en *sciences et technologies*. Mon stage de fin d'études sur la sélection de variables en classification non supervisée, encadré par Christophe Ambroise et Gérard Govaert, s'est déroulé au laboratoire Heudiasyc. J'ai ensuite entamé en 2005 la préparation d'une thèse, traitant de l'intégration de connaissances sur la structure des données dans les modèles d'apprentissage supervisés, sous la direction d'Yves Grandvalet et Pierre Morizet-Mahoudeaux. J'ai soutenu cette thèse en 2008 quelques mois après avoir quitté Compiègne, professionnellement du moins.

Pour m'arracher à mes racines compiègnoises et me faire venir à Marseille cette même année, et sans nier la condition favorable annexe, il aura fallu le talent de persuasion de Liva Ralaivola. J'ai passé à Aix-Marseille Université (AMU) une année d'ATER exceptionnelle au sein de (la genèse de) l'équipe Qarma du Laboratoire d'Informatique Fondamentale (LIF, devenu LIS). J'y ai poursuivi des travaux sur les méthodes à noyaux et aussi exploré des sujets plus théoriques dans le cadre PAC-Bayesien. En prime, j'ai bénéficié d'une plongée dans la culture marseillaise, bien accompagnée par plusieurs autres membres du LIF ou du Laboratoire d'Analyse Topologie et Probabilités (LATP, devenu I2M).

En 2009, j'ai eu l'opportunité de rejoindre comme maîtresse de conférences l'École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ensiie) à Évry-Courcouronnes, mon second port d'attache. L'objet de ce document laissant peu de place aux aspects pédagogiques, je mentionnerai ici mes collègues de l'école, présents ou passés, et en particulier Julien Forest, Christophe Mouilleron et Xavier Urbain, avec qui j'ai gagné de nombreux points d'expérience sur une palette de compétences [HG23] [2] allant de l'enseignement de l'informatique via des paradigmes appropriés aux structures et à l'utilisation des données jusqu'aux questions variées autour la gestion pédagogique et administrative d'une formation, en passant par le sujet délicat du positionnement de la frontière d'acceptabilité dans le compromis.

2. Spéciale dédicace à la CTI et au MESR.

Après un peu plus de cinq années passées au laboratoire d'informatique IBISC de l'Université d'Évry Val d'Essonne (UÉVE), j'ai été chaleureusement accueillie dans l'équipe Statistique et Génome du Laboratoire de Mathématiques et Modélisation d'Évry (LaMME) au début de l'année 2015. Avec Christophe Ambroise et Julien Chiquet, nous avons débuté des premiers travaux sur l'interaction de données omiques. Ces travaux se sont poursuivis avec Florent Guinot dans la seconde contribution de sa thèse, co-encadrée avec Christophe. Pendant le post-doctorat de Marie Courbariaux, nous nous sommes intéressés à l'intégration de données de suivi clinique et de données génétiques afin d'identifier des sous-types de la maladie de Parkinson. Pour finaliser ces travaux, nous avons été épaulés par Kylliann De Santiago, qui a ensuite débuté une thèse, co-encadrée avec Christophe et Guillaume Andéol de l'Institut de Recherche Biomédicale des Armées (IRBA), sur la stratification de patients atteints de traumatismes sonores aigus à partir de données hétérogènes.

À propos de *machine learning*

Ma thématique de recherche est à la frontière de l'informatique, des statistiques et du traitement mathématique du signal. Bien que perméable, cette frontière se caractérise notamment sur le plan lexical : les informaticiens parleront plutôt d'*apprentissage automatique* ou *artificiel* tandis que les statisticiens et les mathématiciens du signal parleront plutôt d'*apprentissage statistique*. Les distinctions derrière cette sémantique sont aussi débattues dans la communauté internationale depuis de nombreuses années déjà [HG8, et les nombreux points de vue en commentaires].

De façon très réductrice, l'*apprentissage automatique* est associé à des travaux autour de la modélisation, la définition et l'étude d'algorithmes et la mise en place de protocoles d'évaluation [HG50] tandis que l'*apprentissage statistique* est lié à l'étude théorique des garanties associées aux modèles ou aux protocoles [HG49]. Cependant, la communauté du *machine learning* s'est construite autour de l'ensemble de ses aspects complémentaires et indissociables. Un exemple emblématique, ayant amplement participé à l'essor du machine learning, est celui des *Support Vector Machines*. Cette approche, plébiscitée dans les années 1990 grâce aux succès applicatifs rendus possibles par les algorithmes de résolution proposés par Boser, Guyon et Vapnik [HG7] puis par Cortes et Vapnik [HG15], trouve ses fondements théoriques dans les travaux de Vapnik et Chervonenkis commencés à la fin des années 1960.

Bien que ce document présente des travaux sous l'angle de la modélisation et de méthodes, mon parcours m'a permis de m'intéresser à l'ensemble des aspects associés au machine learning [3]. J'utiliserai donc la traduction littérale ***apprentissage machine*** qui a l'avantage d'englober l'ensemble des points de vues.

3. Un résumé de l'ensemble de mes travaux de recherche est disponible en section 1.2.

Organisation du document

Ce document commence par un récapitualtif de ma production scientifique. Les noms des étudiants ayant fait l'objet d'un encadrement seront en couleur.

Dans la première partie, le chapitre 1 contient une synthèse de l'ensemble de mes activités de recherche et une description générale du cadre scientifique. Le chapitre 2 est une rétrospective contextualisée de mes travaux antérieurs sur l'***intégration d'information structurée*** dans les termes de pénalisation des modèles d'apprentissage.

La seconde partie contient le chapitre 3 dédié à l'***intégration de sources multiples d'information***, sous l'angle des méthodes à noyaux et des modèles de mélange d'experts. Le chapitre 4 est consacré à l'***interaction des composantes*** de source(s) d'information, à travers des approches impliquant des tests statistiques sur les termes de modèles de régression. Cette partie sera conclue par un chapitre présentant quelques perspectives de recherche.

Production scientifique et réalisations

Manuscrits et chapitres

Thèse de doctorat

[MT1] M. Szafranski. *Pénalités hiérarchiques pour l'intégration de connaissances dans les modèles statistiques*. Université de Technologie de Compiègne, 2008.
(Cf. page 19)

Chapitre d'ouvrage

[CM1] F. Guinot, M. Szafranski et C. Ambroise. Compression structurée de l'information génétique et étude d'association pangénomique par modèles additifs. C. Froidevaux, M.-L. Martin-Magniette et G. Rigaill, éditeurs, *Intégration de données biologiques : approches informatiques et statistiques*, chapitre 5, 129-163. ISTE, 2022.
(Cf. pages 7, 48, 52)

Articles

Revues internationales

[RI1] M. Courbariaux, K. De Santiago, C. Dalmasso, F. Danjou, S. Bekadar, J.-C. Corvol, M. Martinez, M. Szafranski et C. Ambroise[4]. A sparse logistic mixture of experts model for disease subtyping with clinical and genetic data. *Frontiers in Genetics*, 13 :1-11, 2022.
(Cf. pages 7, 32, 36)

4. La dernière position est partagée avec Christophe Ambroise, selon les usages en recherche clinique.

[RI2] F. Guinot, M. Szafranski, J. Chiquet, A. Zankarini, C. Le Signor, C. Mougel et C. Ambroise [5]. Fast computation of genome-metagenome interaction effects. *Algorithms for Molecular Biology*, 15 :1-21, 2020.
(Cf. pages 3, 8, 43, 49, 52)

[RI3] F. Guinot, M. Szafranski, C. Ambroise et F. Samson. Learning the optimal scale for GWAS through hierarchical SNP aggregation. *BMC Bioinformatics*, 19 :459-472, 2018.
(Cf. pages 7, 52)

[RI4] C. Brouard, M. Szafranski et F. d'Alché-Buc. Input Output Kernel Regression : supervised and semi-supervised structured output prediction with operator-valued kernels. *Journal of Machine Learning Research*, 17 :1-48, 2016.
(Cf. page 6)

[RI5] L. Ralaivola, M. Szafranski et G. Stempfel. Chromatic PAC-Bayes bounds for non-iid data : Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 11 :1927-1956, 2010.
(Cf. page 5)

[RI6] M. Szafranski, Y. Grandvalet et A. Rakotomamonjy. Composite Kernel Learning. *Machine learning*, 79(1-2) :73-103, 2010.
(Cf. pages 6, 20, 21, 23, 32, 33)

Conférences internationales avec comité de lecture et actes

[CI1] K. De Santiago, M. Szafranski et C. Ambroise. Mixture of Stochastic Block Models for multiview clustering. *To appear in : Proceedings of the 31th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2023)*, pages 1-6, 2023.

[CI2] M. Szafranski et Y. Grandvalet. KEOPS : KErnels Organized into PyramidS. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*, pages 8262-8266, 2014.
(Cf. pages 3, 6, 13, 23, 32, 33, 35)

[CI3] C. Brouard, F. d'Alché-Buc et M. Szafranski. Semi-supervised penalized output kernel regression for link prediction. *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*, pages 593-600, 2011.
(Cf. page 6)

[CI4] M. Kowalski, M. Szafranski et L. Ralaivola. Multiple Indefinite Kernel Learning with mixed norm regularization. *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 545-552, 2009.
(Cf. pages 6, 20, 21, 32, 34, 35)

5. Les contributions des auteurs font l'objet d'une section dans l'article.

[CI5] L. Ralaivola, M. Szafranski et G. Stempfel. Chromatic PAC-Bayes bounds for non-iid data. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS'09)*, pages 416-423, 2009.
(Cf. page 5)

[CI6] M. Szafranski, Y. Grandvalet et A. Rakotomamonjy. Composite Kernel Learning. *Proceedings of the 25th Annual International Conference on Machine Learning (ICML'08)*, pages 1040-1047, 2008.
(Cf. pages 6, 21, 23, 32, 33)

[CI7] M. Szafranski, Y. Grandvalet et P. Morizet-Mahoudeaux. Hierarchical Penalization. *Advances in Neural Information Processing Systems 20 (NIPS'07)*, pages 1057-1464, 2007.
(Cf. pages 5, 6, 21, 23)

Conférences nationales avec comité de lecture et actes

[CN1] M. Courbariaux, M. Szafranski, C. Dalmasso, C. Ambroise et consortium MeMoDeep. Sous-typage de maladie avec étude d'association génétique intégrée. *Actes des 50e Journées de Statistique*, 2018.

[CN2] F. Guinot, M. Szafranski, J. Chiquet et C. Ambroise. Une approche hiérarchique de la recherche d'interactions entre données omiques. *Actes des 50e Journées de Statistique*, 2018.

[CN3] M. Szafranski, M. Kowalski et L. Ralaivola. Apprentissage à partir de noyaux multiples et indéfinis. *Actes du colloque du Groupement de Recherche et d'Étude en Traitement du signal et des Images*, 2013.

[CN4] C. Bourard, M. Szafranski et F. d'Alché-Buc. Régression semi-supervisée à noyaux à valeur opérateur pour la prédiction de liens. *Actes des 13e Journées Ouvertes en Biologie, Informatique et Mathématiques*, 2012.

[CN5] C. Bourard, M. Szafranski et F. d'Alché-Buc. Régression semi-supervisée à sortie noyau pour la prédiction de liens. *Actes de la 13e Conférence d'Apprentissage*, 2011.

[CN6] M. Szafranski et Y. Grandvalet. Pyramides de noyaux. *Actes du colloque du Groupement de Recherche et d'Étude en Traitement du signal et des Images*, 2009.

[CN7] M. Szafranski, Y. Grandvalet et A. Rakotomamonjy. Learning with Groups of Kernels. *Actes de la 10e Conférence d'Apprentissage*, 2008.

[CN8] M. Szafranski, Y. Grandvalet et P. Morizet-Mahoudeaux. Pénalisation hiérarchique. *Actes de la 9e Conférence d'Apprentissage*, 2007.

Codes et serveurs web

[CS1] M. Courbariaux, K. De Santiago, C. Ambroise, M. Szafranski et C. Dalmasso. **DiSuGen** : Disease Subtyping with integrated Genome association. *Package* `R`, 2021.
(Cf. page 7)

[CS2] F. Guinot, J. Chiquet, C. Ambroise et M. Szafranski. **SICOMORE** : Selection of Interaction effects in COmpressed Multiple Omics REpresentations. *Package* `R`, 2020.
(Cf. pages 8, 46)

[CS3] M. Courbariaux, F. Samson, C. Ambroise, M. Szafranski et C. Dalmasso. **MeMoDeepWeb** : Disease subtyping with integrated genome association. *Serveur web*, 2017.
(Cf. page 7)

[CS4] F. Guinot, F. Samson, M. Szafranski et C. Ambroise. **LEOS** : LEarning the Optimal Scale in Genome-Wide Association Studies. *Serveur web*, 2017.
(Cf. page 7)

[CS5] M. Szafranski, Y. Grandvalet et A. Rakotomamonjy. **KEOPS** : KErnels Organized into PyramidS. *Toolbox* `Matlab`, 2014.
(Cf. page 6)

[CS6] J. Kossaifi, M. Kowalski, M. Szafranski et L. Ralaivola. **MIKL** : Multiple Indefinite Kernel Learning. *Toolbox* `Python`, 2012.
(Cf. page 6)

Au pays des merveilles

L'ordre alphabétique reflète une contribution équivalente sur la post-production.

[WB16] M. Kowalski et M. Szafranski. Maxandre, système (fortement) dynamique non linéaire. Janvier 2016.

[WB19] M. Kowalski et M. Szafranski. Maorèn, processus stochastique non stationnaire. Février 2019.

Table des matières

Première partie

Vue d’ensemble et rétrospective

Hindsight ground

1. Synthèse et contexte

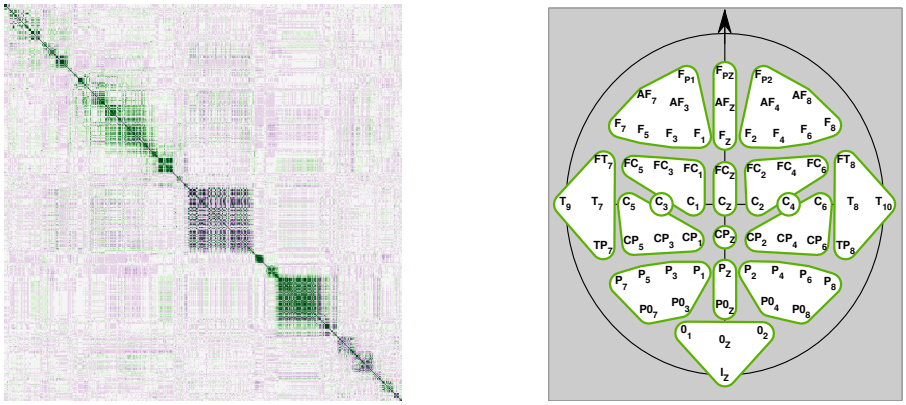
Sommaire

1.1. Problématique

Mes travaux de recherche émanent de problèmes où les données sont à la fois structurées et issues de sources d'informations multiples, voire hétérogènes. Ces caractéristiques vont être illustrées ici à travers deux exemples issus de [CI2] et [RI2].

1.1.1. Données structurées

Certaines données comportent une structure qui peut être explicite ou sous-jacente. Intégrer ce type d'information dans la modélisation peut permettre d'une part de mieux guider le processus d'apprentissage mais aussi, dans une démarche plus exploratoire, de mieux comprendre les mécanismes des problèmes étudiés. La figure 1.1 montre, pour deux applications différentes, deux exemples de structures que l'on peut être amené à considérer.



(a) (b)

FIGURE 1.1. – Exemples de structures. (a) Extrait d'une matrice de corrélations entre marqueurs génétiques. (b) Cartographie de la disposition d'électrodes dans un problème d'interface cerveau-machine.

(a) Dans les études d'associations génétiques, certains marqueurs présentent des structures de corrélations complexes comme illustré sur la figure 1.1a. Ce phénomène, désigné par le terme de *déséquilibre de liaison*, illustre l'organisation spatiale du génome et traduit, de façon simplifiée, une association préférentielle de certains gènes sur des parties spécifiques du chromosome.

(b) Dans d'autres domaines comme en neurosciences, les réponses d'un sujet à un stimulus peuvent être localisées en fonction de zones spécifiques du cerveau. On peut alors utiliser un atlas défini soit à partir d'une connaissance préalable permettant de faire la correspondance entre certaines zones et certaines fonctions

du cerveau, soit par des proximités induites par le procédé d'acquisition des données comme illustré sur la figure 1.1b.

Dans mes travaux, je me suis en particulier intéressée aux structures hiérarchiques permettant de définir des groupes de granularités différentes à chaque niveau d'une arborescence. Toutefois, il existe une palette assez large de structures pouvant être considérées, puis intégrées dans les problèmes d'optimisation. Le chapitre 2 présentera une synthèse sur chacun de ces aspects.

1.1.2. Sources multiples

Les informations disponibles peuvent également être issues de sources diverses. Dans les problèmes d'interfaces cerveau-machine par exemple, les réponses des patients à un stimulus sont mesurées par un ensemble d'électrodes réparties sur le cuir chevelu comme illustré sur la figure 1.1b. Nous sommes ici en présence d'informations de même nature collectées par plusieurs sources : chaque électrode mesure le même type d'information, en l'occurence des signaux EEG pour lesquels on appliquera des prétraitements identiques.

Les données collectées peuvent aussi être de natures différentes, comme dans les suivis cliniques de patients où on peut disposer de questionnaires, d'analyses sanguines, de signaux EEG, d'IRM, de séquences de génomes, etc. Dans un ordre d'idée similaire, les études d'associations génomiques (GWAS pour *Genome Wide Association Studies*) ont longtemps été privilégiées pour rechercher des marqueurs génétiques liés à une pathologie [HG46]. Cependant, de récentes études suggèrent que la part environnementale caractérisée par le métagénome (MWAS pour *Metagenome Wide Association Studies*) permettrait également d'expliquer des pathologies complexes telles que l'obésité ou le diabète [HG45]. L'étude combinée de ces deux sources d'information semble donc pertinente pour évaluer comment influent le facteur individuel (génomique) et le facteur environnemental (métagénomique) dans le développement d'une pathologie ou la construction d'un phénotype.

Ici également, il existe de nombreuses approches permettant d'intégrer les sources de données multiples. Certaines seront abordées aux chapitres 3 et 4.

1.1.3. Axes de recherche

Afin de tirer parti de l'ensemble des informations disponibles sur les données, je me suis intéressée aux moyens d'intégrer dans les problèmes d'apprentissage la structure sous-jacente des données ainsi que les différentes sources d'information, souvent hétérogènes, utilisées pour collecter ces données. Ceci s'est manifesté autour de trois axes principaux, souvent liés :

1. La prise en compte d'une ***structure*** sur les données,
2. L'***intégration*** de sources d'information multiples,
3. L'***interaction*** entre les composantes issues de ces sources.

La prise en compte de structure, résumée dans la section 1.2.1, fera l'objet d'une rétrospective ciblée dans le chapitre 2. Elle concernera l'incorporation d'une structure sur les composantes dans le terme de régularisation d'un problème d'apprentissage, cet aspect étant par ailleurs présent dans la quasi-totalité des travaux concernant les deux axes suivants. Les contributions sur l'intégration de sources multiples d'information, résumées dans la section 1.2.2, seront décrites dans le chapitre 3. Enfin, les contributions sur l'interaction entre descripteurs liés aux variables explicatives, résumées dans la section 1.2.3, seront elles développées dans le chapitre 4.

1.2. Activités de recherche

Les travaux de recherche résumés ci-dessous sont le fruit de collaborations avec différentes personnes dont les noms seront mentionnés dans cette synthèse avec les affiliations à l'époque des travaux. Les noms des étudiants ayant fait l'objet d'un encadrement seront en couleur.

1.2.1. Prise en compte d'une structure sur les données

Structure hiérarchique sur les composantes. *2006 – 2009 (doctorat)*

Mes travaux de thèse ont débuté avec la définition d'un modèle linéaire parcimonieux dans lequel les variables d'un problème sont organisées dans une structure arborescente de deux niveaux permettant de constituer des groupes [CI7]. Nous avons proposé un problème d'optimisation et établi les liens entre la pénalité issue de cette formulation et une norme mixte. Ce point de vue a permis de définir les propriétés de convexité et de parcimonie du modèle. Un algorithme fondé sur la notion de contraintes actives a été développé pour résoudre le problème d'optimisation.

Co-auteurs. *Yves Grandvalet (CNRS & UTC) et Pierre Morizet-Mahoudeaux (UTC)*

Structure d'interdépendance sur les observations. *2009 – 2010*

Dans ces travaux, nous nous sommes intéressés à des configurations d'apprentissage dans lesquelles les données sont identiquement distribuées mais plus indépendantes. C'est par exemple le cas avec des données séquentielles, lorsque chaque donnée est observée sur une fenêtre relativement aux données des temps précédents et suivants. Nous avons proposé des bornes d'erreur de type Pac-Bayes permettant de garantir la capacité de généralisation de classifieurs entraînés sur ce type de données. Ces bornes sont établies à partir de la notion de coloration fractionnaire d'un graphe [CI5]. Elles s'appliquent au cadre de l'apprentissage de données séquentielles ou encore celui de l'apprentissage de fonctions pour l'ordonnancement biparti [RI5].

Co-auteurs. *Liva Ralaivola et Guillaume Stempfel (Aix-Marseille Université)*

Structure de connexion sur les étiquettes. *2011 – 2016*

On parle de données structurées en sortie lorsque la cible à prédire se présente sous une plus forme riche qu'une valeur scalaire. Nous nous sommes intéressées au problème de prédiction de liens dans un réseau d'interactions représenté en sortie par une matrice. Pour résoudre ce problème, nous avons proposé une approche fondée sur la théorie des espaces de Hilbert à noyaux à valeur opérateur reproduisants qui offre un cadre adapté pour prédire ce genre de sortie. Nous avons établi un théorème de représentation dans le cadre semi-supervisé [CI3] ainsi qu'une extension pour la sélection de modèle fondée sur la validation croisée généralisée [RI4].

Co-autrices. *Céline Brouard (doctorante, UÉVE) et Florence d'Alché-Buc (UÉVE, IMT ParisTech)*

1.2.2. Intégration de sources d'information multiples

Méthodes à noyaux reproduisants. *2008 – 2010 (doctorat), 2014*

Dans mes seconds travaux de thèse, nous avons étendu notre approche développée pour les modèles linéaires [CI7] à des modèles additifs par le biais de méthodes à noyaux. Cela nous a permis de prendre en compte différentes sources d'information, d'abord avec des arborescences de deux niveaux où le premier niveau représente les sources et le second les variables associées à chacune de ces sources [CI6, RI6]. Nous avons plus tard généralisé cela à des arborescences contenant un nombre arbitraire de niveaux [CI2]. Un algorithme d'optimisation alternée a été établi à partir des conditions d'optimalité du problème à résoudre [CS5]. Ces travaux ont été appliqués au domaine des interfaces cerveau-machine où les électrodes permettant de collecter les signaux EEG représentent les sources.

Co-auteurs. *Yves Grandvalet (CNRS & UTC) et Alain Rakotomamonjy (Université de Rouen)*

Méthodes à noyaux non positifs. *2009 – 2012*

Dans d'autres travaux, nous nous sommes intéressés à une structure croisée portant sur les sources d'information mais également sur les exemples. Une des spécificités de cette approche a consisté à représenter les sources au travers de noyaux non positifs qui permettent par exemple d'utiliser des similarités non symétriques entre observations [CI4]. Nous avons établi des bornes mesurant la capacité de généralisation [1] de la famille de prédicteurs associée à ces noyaux peu classiques en apprentissage. Nous avons également adapté des algorithmes proximaux pour prendre en compte différentes configurations sur la façon d'organiser nos structures [CS6].

Co-auteurs et collaborateur. *Matthieu Kowalski (Université Paris-Saclay), Liva Ralaivola (Aix-Marseille Université) et Jean Kossaifi (stagiaire de 2e année, ensiie)*

1. C'est-à-dire la capacité d'un classifieur à prédire correctement une étiquette sur des données non utilisées pour apprendre le classifieur.

Modèles de mélange d'experts. *2017 – 2022*

Ces travaux consistent à identifier des sous-types de la maladie de Parkinson. Nous disposons pour cela de deux informations : le suivi clinique de plusieurs patients, sur une durée de une à cinq années, associé aux marqueurs génotypiques de ces patients. Nous avons proposé une méthode non supervisée de classification fondée sur un modèle de mélanges à variables concomitantes. La classification des patients s'effectue sur la base des estimations obtenues par un algorithme de type EM (Expectation-Maximization). Les paramètres des lois associées aux variables cliniques et génétiques sont estimés dans l'étape M de l'algorithme. Les probabilités d'appartenance des patients aux classes sont quant-à elles estimées dans l'étape E sur la base des données génétiques [RI1, CS1, CS3].

Co-autrice et co-auteurs.
Méthodologie : Marie Courbariaux (post-doctorante, UÉVE), Kylliann De Santiago (doctorant, UÉVE & Sensorion), Christophe Ambroise (UÉVE), Cyril Dalmasso (UÉVE) et Franck Samson (INRAE)
Clinique : Jean-Christophe Corvol (APHP, ICM[2]), Samir Bekadar (ICM), Fabrice Danjou (APHP, ICM)

1.2.3. Interaction entre composantes de source(s) d'information

Interaction gène $\times$ gène. *2016 – 2020*

À mon arrivée au LaMME, je me suis à nouveau intéressée aux problématiques associées aux structures hiérarchiques dans le contexte d'études pangénomiques à large échelle, où l'on recherche à établir avec des tests statistiques d'association si des relations existent entre marqueurs génétiques (de l'ordre du million) et phénotype. Afin d'améliorer la puissance statistique dans les études dassociations pangénomiques, nous faisons intervenir une structure hiérarchique du génome humain permettant de considérer différents ensembles de gènes agrégés selon la façon dont ils interagissent. Le niveau d'agrégation optimal est obtenu en testant différents niveaux de coupes par le biais d'un modèle linéaire régularisé. De plus, des stratégies heuristiques fondées sur le nombre de gènes impliqués dans une agrégation peuvent permettre de cibler la plage de recherche du niveau de coupe de la hiérarchie [RI3, CS4]. Dans [CM1], nous avons étendu ces travaux à des modèles additifs, les avons mis en perspective dans le contexte des études d'associations pangénomiques et avons présenté des résultats expérimentaux supplémentaires.

Co-auteurs. *Florent Guinot (doctorant, UÉVE & BIOptimize), Christophe Ambroise (UÉVE) et Franck Samson (INRAE & UÉVE)*

2. Institut du Cerveau et de la Moëlle épinière.

Interaction gène × environnement. *2015 – 2020*

Nous nous sommes ici intéressés aux interactions entre deux sources d'information biologiques : les marqueurs génotypiques qui caractérisent des individus et les marqueurs métagénomiques qui sont liés à l'environnement dans lequel évoluent ces individus. Ces jeux de données peuvent contenir plusieurs centaines de milliers, voire quelques millions, de variables. Des stratégies de réduction de dimension et / ou d'élimination de variables sont donc nécessaires. Dans notre approche, nous avons considéré des groupes de variables issus d'une hiérarchie au sein d'un modèle linéaire d'interactions multiplicatives. Pour gagner en efficacité, nous avons agi sur trois aspects. Pour réduire la dimension du problème, nous avons défini une compression, propre à chaque source d'information, permettant d'obtenir des variables agrégées représentant les groupes aux différentes échelles d'une hiérarchie. Ces dernières sont ensuite aplanies par le biais d'une pondération représentant les sauts entre niveaux de la hiérarchie, afin de conserver un trace des différentes regroupements possibles tout en s'affranchissant de la recherche du niveau de coupe optimal. Cette étape inclut également un processus de sélection des variables agrégées. Enfin, dans une optique de sélection d'interactions pertinentes, nous avons considéré chaque interaction du modèle au sein d'un test d'hypothèse, dans un schéma respectant les structures de dépendance entre effets simples et interactions [RI2, CS2].

Co-auteurs et co-autrices.
Méthodologie : Florent Guinot (doctorant, UÉVE & BIOptimize), Christophe Ambroise (UÉVE) et Julien Chiquet (UÉVE, INRAE)
Biologie : Anouk Zancarini (INRAE, University of Amsterdam), Christine Le Signor (INRAE) et Christophe Mougel (INRAE)

1.3. Cadre mathématique et terminologie

Mes travaux de recherche se placent dans le cadre de l'***apprentissage machine*** qui regroupe un ensemble de méthodes visant à analyser, interpréter ou encore prédire un phénomène. Le processus s'effectue au travers d'objets observés sur un ensemble de D ***attributs***, appelés ***variables explicatives***, ou encore ***descripteurs*** lorsqu'elles subissent une transformation.

Plus formellement, on représentera ces objets appelés ***observations***, ***exemples*** ou encore ***individus***, sous la forme d'un vecteur ligne de dimension D :

$$\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_D) \in \mathcal{X} \ ,$$

où x_j représente l'observation $\mathbf{x}$ associée à la variable j et $\mathcal{X}$ définit l'espace des variables, généralement $\mathbb{R}^D$.

On peut définir à partir de ces observations différents processus d'apprentissage fonction de l'objectif que l'on souhaite atteindre [HG14]. Dans les travaux présentés dans ce document, on s'intéressera à l'apprentissage ***supervisé*** et ***non supervisé***.

Remarque *Sur le terme de classification* — Pour parler d'apprentissage non supervisé, on utilise les termes de ***classification automatique*** ou encore de ***clustering*** en anglais. Cet anglicisme a pour mérite de lever l'ambiguïté entre ***classification automatique*** et ***classification supervisée***, souvent abrégés l'un comme l'autre en ***classification***. ◇

1.3.1. Apprentissage supervisé

En apprentissage supervisé, on dispose d'une information supplémentaire pour guider le processus d'apprentissage : à chaque observation $\mathbf{x}$ est associée une ***étiquette*** $y \in \mathcal{Y}$, appelée aussi ***réponse*** ou ***variable expliquée***, avec généralement $y \in \mathbb{R}$ ou $y \in \{1, \dots, K\}$ avec $K \in \mathbb{N}$.

Remarque *Échantillon indépendant et identiquement distribué* — L'hypothèse généralement faite est que les couples $(\mathbf{x}, y)$ sont générés de façon indépendante et identiquement distribuée selon une loi de probabilité jointe inconnue $\mathbb{P}(X, Y)$, où X est le vecteur aléatoire associé aux observations et Y la variable aléatoire associée aux réponses. ◇

À partir d'un ensemble d'apprentissage $\boldsymbol{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, on cherche à inférer une relation entre les observations et la réponse à travers :

1. Une ***fonction de prédiction*** $f : \mathcal{X} \to \mathcal{Y}$, appelée également ***prédicteur*** ou ***classifieur*** lorsque $y \in \{1, \dots, K\}$,
2. Une ***fonction de coût*** quantifiant l'erreur commise par f sur $\mathbf{x}$ pour évaluer y, telle que $C : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$,
3. Un ***risque empirique*** $R_{\boldsymbol{S}}(f)$ représentant la moyenne du coût C sur $\boldsymbol{S}$

$$R_{\boldsymbol{S}}(f) = \frac{1}{N} \sum_{i=1}^{N} C(y_i, f(\mathbf{x}_i)) \,,$$

que l'on souhaite minimiser pour trouver l'estimateur optimal de f sur $\boldsymbol{S}$.

La théorie statistique de l'apprentissage de Vapnik étudie les conditions de consistance et de convergence de la minimisation de ce risque empirique [HG43]. En termes succincts, il s'agit de savoir si minimiser $R_{\boldsymbol{S}}(f)$ permet de minimiser le risque théorique $R(f) = \mathbb{E}\left[C(Y, f(X))\right]$ lorsque N, la taille de l'échantillon $\boldsymbol{S}$, tend vers l'infini. Il faut de plus que cela soit applicable, ce qui nécessite de pouvoir borner $R(f)$ par $R_{\boldsymbol{S}}(f)$ pour des échantillons de taille finie.

La dimension de Vapnik-Chervonenkis permet précisément de majorer l'écart entre $R(f)$ et $R_{\boldsymbol{S}}(f)$ en fonction du coût C, de la taille N de l'ensemble d'apprentissage, mais aussi de l'ensemble de fonctions $\mathcal{F}$ auquel la fonction de prédiction f appartient. L'existence de telles bornes conduit à envisager des stratégies d'apprentissage qui ne sont pas uniquement basées sur la minimisation du risque empirique, mais qui tiennent également compte de la *complexité* du modèle. La modulation de cette complexité se manifestera sous la forme d'un terme additionnel de ***pénalisation*** $P(f)$, aussi appelé ***régularisation***, dans le problème d'optimisation.

1.3.2. Apprentissage non supervisé

Le but de l'apprentissage non supervisé est de déterminer automatiquement K ***catégories***, encore appelées ***classes***, ***sous-types*** ou ***clusters***, sur un ensemble d'observations : on souhaite associer à chaque $\mathbf{x}$ une catégorie $z \in \mathcal{Z}$, telle que $z \in \{1, \ldots, K\}$.

Dans ce processus de classification automatique, on cherche à regrouper les exemples en fonction de leur ressemblance dans un espace donné qui peut être une transformation de l'espace initial des variables. L'enjeu consiste à trouver une mesure de ressemblance et un critère de classification satisfaisants.

Lorsque ce processus s'effectue dans le cadre de modèles probabilistes paramétriques, de paramètres Θ, on peut adopter un point de vue proche de l'apprentissage supervisé, où l'ensemble $\boldsymbol{S} = \{\mathbf{x}_i\}_{i=1}^N$ est utilisé dans :

1. Une ***fonction de classification*** $f : \mathcal{X} \to \mathcal{Z}$, appelée également ***classifieur***, permettant d'établir une ***partition*** des observations de $\boldsymbol{S}$ en K catégories,
2. Un ***critère de log-vraisemblance*** $L_{\boldsymbol{S}_c}(f \mid \Theta)$ que l'on cherchera à maximiser afin de trouver les paramètres optimaux Θ de la densité de probablité f, lorsque la vraisemblance est observée à partir de $\boldsymbol{S}$ et complétée par l'estimation du vecteur des catégories $\mathbf{z}$, tel que $\boldsymbol{S}_c = \{\boldsymbol{S}, \mathbf{z}\}$.

Notons cependant que contrairement à l'apprentissage supervisé, on ne dispose ici d'aucune information *a priori* sur les catégories, notamment sur leur nombre K qu'il faut donc également estimer. Celeux et al. établissent dans [HG10] un état de l'art des critères permettant cela dans le cadre des modèles de mélanges. L'évaluation de la qualité de la partition générée par le classifieur repose sur des critères de validation dits internes et externes ou encore relatifs [HG9]. Les critères internes vont concerner des mesures de compacité ou de séparation entre les différents clusters tandis que les critères externes reposent sur l'évaluation, à partir de simulations, des écarts entre les partitions réelles et générées [HG48]. En particulier, les indices d'associations, tels que l'indice de Rand ajusté (ARI, pour *Adjusted Random Index*) [HG26], utilisent la comparaison de paires d'observations pour se défaire des permutations aléatoires intervenant sur la numérotation des catégories. Les critères relatifs, tels que le critère de stabilité [HG44], permettront quant-à eux de s'assurer qu'un classifieur, dans une condition initiale fixée, génère des clusters de façon consistante sur différentes données respectant cette condition initiale.

1.3.3. Notations et conventions

Les problèmes d'optimisation en apprentissage seront souvent formulés via les notations matricielles ou vectorielles liées aux données. Dans ce document, les vecteurs de scalaires seront notés en gras et les matrices en majuscule et en gras. En particulier, on utilisera les notations suivantes.

Observations La matrice $\mathbf{X}$ est associée aux observations d'un ensemble d'apprentissage supervisé ou non supervisé, avec $\mathbf{X} \in \mathcal{M}^{N \times D}$, où N représente le nombre d'observations de l'ensemble d'apprentissage et D représente la dimension de $\mathcal{X}$.

Le vecteur $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})$ représente la i^{e} observation en ligne de $\mathbf{X}$ et $\mathbf{x}^j = (x_{1j}, \ldots, x_{Nj})^\top$ la j^{e} variable explicative en colonne de $\mathbf{X}$.

Composantes On utilisera le terme générique de ***composantes*** pour faire référence aux m éléments de la combinaison additive $f(\mathbf{x}) = \sum_{m=1}^{M} f_m(\mathbf{x})$ et on notera $\boldsymbol{f} = (f_1(\mathbf{x}), \ldots, f_M(\mathbf{x}))^\top$ le vecteur associé aux composantes.

Ces composantes peuvent être des coefficients associés aux variables explicatives, aux sources d'information ou encore les éléments d'une base de représentation des observations.

Étiquettes En apprentissage supervisé, $\mathbf{y}$ désigne le vecteur des réponses associées à la matrice d'observation $\mathbf{X}$, avec $\mathbf{y} \in \mathcal{Y}^N$.

En fonction de l'espace $\mathcal{Y}$, on parlera de régression lorsque $y \in \mathbb{R}$ et de classification ou discrimination lorsque $y \in \{1, \ldots, K\}$.

Catégories En apprentissage non supervisé, et selon les situations, on utilisera soit $\mathbf{z} \in \mathcal{Z}^N$ pour désigner le vecteur d'appartenance des observations aux catégories, dont le terme général $z_i \in \{1, \ldots, K\}$, soit $\mathbf{Z}$ la matrice indicatrice d'appartenance des observations aux catégories de taille $N \times K$, que l'on pourra aussi appeler ***partition***, telle que $\mathbf{z}_i \in \{0, 1\}^K$, avec $z_{ik} = 1$ si l'observation $\mathbf{x}_i$ appartient à la k^{e} catégorie et 0 sinon.

Remarque *Indices* — Pour alléger les notations, les variations des indices seront omises.

(a) Pour les observations et les étiquettes, on utilisera l'indice $i : 1 \leq i \leq N$.
(b) Pour les composantes, on utilisera l'indice $m : 1 \leq m \leq M$. Lorsque ces composantes feront référence aux variables explicatives, on les indexera par $j : 1 \leq j \leq D$.
(c) Pour les catégories, on utilisera l'indice $k : 1 \leq k \leq K$.
(d) Pour les structures, les groupes seront indexés par $g : 1 \leq g \leq G$ et les différentes hauteurs d'une arborescence par $h : 0 \leq h \leq H$.

◇

2. Caractérisation et intégration de structures

Sommaire

2.1. Définition d'une structure

Nous avons vu, à travers les deux exemples évoqués dans la section 1.1 du chapitre 1, que certaines données pouvaient se caractériser par des structures relatives à l'organisation, spatiale notamment, des composantes du problème. Cette organisation peut être connue *a priori* ou déduite à l'aide d'algorithmes.

2.1.1. Structure connue

Il est parfois possible de structurer les composantes en fonction de connaissances préalables d'experts ou d'un plan d'acquisition des données, voire de hiérarchiser cette structure lorsque différents niveaux de raffinement peuvent être proposés.

Par exemple, dans le problème d'interface cerveau-machine traité dans [CI2], le plus haut niveau représente une cartographie tenant compte des hémisphères avant, arrière, gauche et droit du cerveau (cf. figure 1.1b), le niveau intermédiaire se définit par rapport aux électrodes et le niveau terminal par rapport aux mesures temporelles effectuées à partir de chaque électrode. Notons que ce type de mesures aussi peut faire l'objet d'une structuration, que ce soit par le biais d'un fenêtrage sur la temporalité ou par celui d'un découpage de type temps-fréquence.

Les travaux cités dans [HG37] permettent de voir que les domaines d'application dans lesquels sont utilisés des structures prédéfinies s'étendent des neurosciences à la génomique, en passant par les problèmes fréquemment rencontrés en traitement du signal tels que la segmentation de la parole ou la vision par ordinateur.

Dans tous les cas, une expertise sur le problème lui-même ou sur le processus d'acquisition des données est ici nécessaire pour construire la structure.

2.1.2. Construction préalable d'une structure

Dans d'autres cas, il peut être nécessaire d'extraire cette structure, en amont du problème d'optimisation, lorsque les connaissances ne permettent pas de la définir ou lorsque cette définition n'est pas suffisamment fine pour tenir compte de caractéristiques liées au problème.

Dans le cas de données génétiques par exemple, on peut organiser les marqueurs en fonction de leurs positions sur les chromosomes. Cependant ce type de représentation ne tient pas forcément compte de mécanismes d'association tels que le déséquilibre de liaison évoqué précédemment. Il est toutefois possible d'utiliser, ou de définir, des techniques de classification automatique intégrant des contraintes adaptées à ces particularités [HG17]. Une illustration sera présentée au chapitre 4.

Là encore, cette définition préalable de la structure, même si elle peut s'envisager de façon non supervisée, nécessite bien souvent un certain niveau d'expertise sur les données pour pouvoir influencer de façon pertinente le problème d'apprentissage.

2.1.3. Construction embarquée d'une structure

Enfin, il existe des approches qui permettent de déduire une structure au sein même du problème d'optimisation. Certaines pénalités rencontrées dans les problèmes tels que l'*elastic net* [HG53] favorisent la sélection de groupes de variables corrélées. Cependant, elles ne permettent pas de définir explicitement une structure.

La recherche explicite d'une structure sur les M composantes est un problème d'optimisation combinatoire dont la complexité croit en fonction de M et du type de la structure. Ces aspects représentent un champ de recherche à part entière. Toutefois, on peut mentionner les travaux de [HG31] qui proposent une approche astucieuse en minimisant une pénalité structurée (cf. section 2.3.2) s'adaptant aux données pour trouver une partition sur les M composantes dont l'algorithme itératif en $\mathcal{O}(M)$ possède certaines garanties de convergence vers la solution optimale.

2.2. Typologie des structures

Indépendamment de la façon dont elles sont définies, nous allons nous intéresser ici à différentes structures pour lesquelles nous allons formaliser les notations qui seront illustrées sur les figures 2.1, 2.2 et 2.3.

Nous considérons que l'ensemble $\mathcal{M} = \{1, \dots, M\}$ des composantes du problème peuvent être organisées selon :

- $(\mathcal{P})$ Une structure de groupes formant une partition de $\mathcal{M}$.
- $(\mathcal{R})$ Une structure de groupes recouvrants sur les composantes de $\mathcal{M}$.
- $(\mathcal{D})$ Une structure hiérarchique sous forme de graphe acyclique dirigé (DAG pour *Directed Acyclic Graph*) décrivant la filiation entre les composantes de $\mathcal{M}$.
- $(\mathcal{T})$ Une structure hiérarchique arborescente où chaque niveau définit des groupes de granularité différentes sur les composantes de $\mathcal{M}$.
- $(\mathcal{N})$ Une structure de graphe définissant un voisinage entre les composantes de $\mathcal{M}$.

Remarque *Rappel de notations* — Les compostantes sont indicées par $1 \leq m \leq M$, les groupes par $1 \leq g \leq G$ et les hauteurs des arborescences par $1 \leq h \leq H$. ◇

2.2.1. Structures de groupes

Partitions $(\mathcal{P})$. Une partition $\mathcal{P} = \{\mathcal{G}_1, \dots, \mathcal{G}_G\}$ est structure de groupe disjoints, où chaque groupe $\mathcal{G}_g$ contient les indices de D_g composantes, avec $\sum_g D_g = M$, et où $\mathcal{G}_g \cap \mathcal{G}_{g'} = \emptyset, \forall (g, g')$.

Groupes recouvrants $(\mathcal{R})$**.** Dans une structure de groupes $\mathcal{R} = \{\mathcal{G}_1, \dots, \mathcal{G}_G\}$ recouvrants sur les composantes de $\mathcal{M}$, chaque groupe $\mathcal{G}_g$ contient D_g composantes, avec $1 \leq D_g \leq M$.

Remarque *Passage de $\mathcal{R}$ à $\mathcal{P}$ —* Il est toujours possible de passer d'une structure de groupe recourants à une partition en dupliquant les composantes appartenant à plusieurs groupes, selon le mécanisme représenté sur la figure 2.1. Notons cependant que cela peut considérablement augmenter la taille du problème, notamment lorsque les recouvrements sont importants et dans les problèmes de grande dimension. ◇

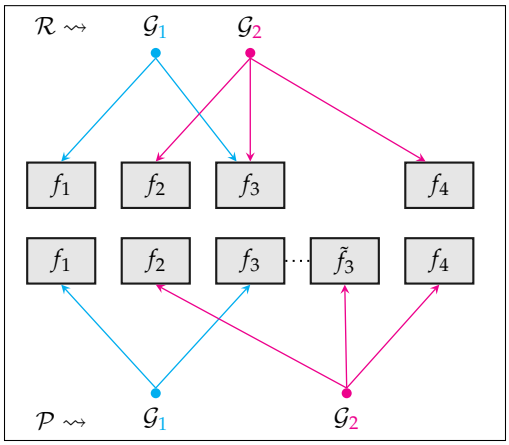
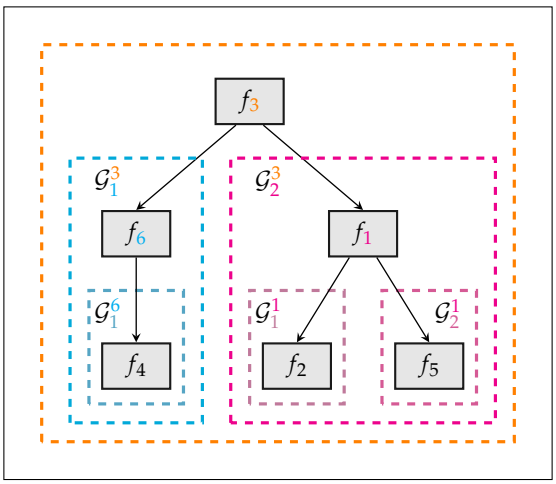


FIGURE 2.1. – Exemple de structures de groupes. En haut, une structure recouvrante $\mathcal{R}$ pour $G = 2$ et $M = 5$ et en bas, la partition $\mathcal{P}$ correspondante la composante f_3 dupliquée.

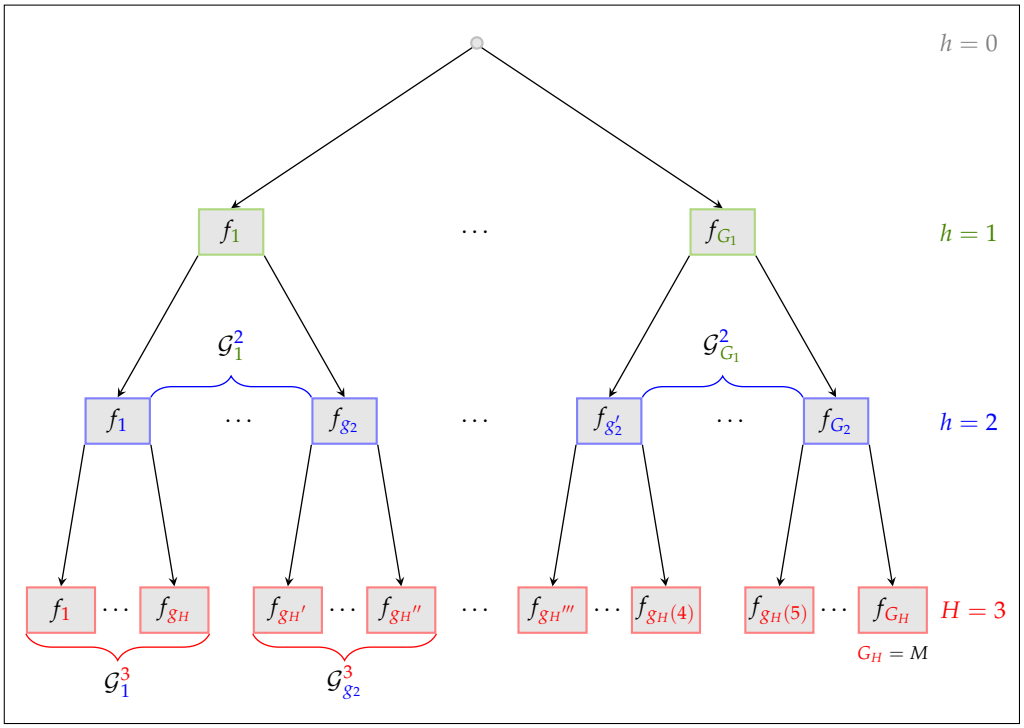
2.2.2. Structures hiérarchiques

Graphes acycliques dirigés $(\mathcal{D})$**.** Dans une structure hiérarchique $\mathcal{D} = \left\{\mathcal{D}^1, \dots, \mathcal{D}^M\right\}$ sous forme de DAG, l'ensemble $\mathcal{D}^m = \left\{\mathcal{G}_1^m, \dots, \mathcal{G}_{G_m}^m\right\}$ correspond aux (groupes de) composantes descendant de la composante m. Chaque branche du graphe définit une structure imbriquée telle que $\mathcal{D}^m \subset \cdots \subset \mathcal{D}^{m'}$, comme décrit sur la figure 2.2a.

Arborescences $(\mathcal{T})$**.** Une structure arborescente $\mathcal{T} = \left\{\mathcal{T}^0, \dots, \mathcal{T}^H\right\}$, définit à chaque niveau h une partition $\mathcal{T}^h = \left\{\mathcal{G}_1^h, \dots, \mathcal{G}_{G_h}^h\right\}$, de sorte qu'il existe une relation d'inclusion entre des composantes de groupes définis à $\mathcal{T}^{h+1}$ par rapport à ceux définis à $\mathcal{T}^h$, comme illustré sur la figure 2.2b.



(a) Graphe acyclique dirigé $\mathcal{D}$ pour $M = 6$, inspiré de [HG29]. Ici, $\mathcal{D}^3 = \{\mathcal{G}_1^3, \mathcal{G}_2^3\}$. Sur la branche gauche, $\mathcal{D}^6 = \{\mathcal{G}_1^6\}$ et $\mathcal{D}^6 \subset \mathcal{D}^3$. Sur la branche droite, $\mathcal{D}^1 = \{\mathcal{G}_1^1, \mathcal{G}_2^1\}$ et $\mathcal{D}^1 \subset \mathcal{D}^3$.



(b) Arborescence $\mathcal{T}$ de hauteur $H = 3$ et notations associées. Ici, $\mathcal{G}_g^3 \subset \mathcal{G}_1^2$ pour $1 \leq g \leq g_2$.

FIGURE 2.2. – Exemple de structures arborescentes.

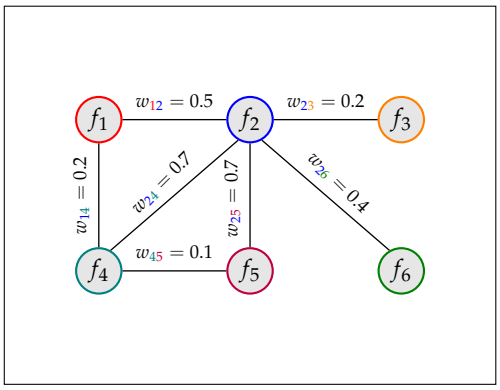
2.2.3. Structures de graphe

Graphes pondérés de voisinage $(\mathcal{N})$**.** Les relations de voisinage entre les composantes M peuvent se décrire par le biais d'un graphe pondéré $\mathcal{N} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, où l'ensemble des sommets $\mathcal{V} = \mathcal{M}$ et les arrêtes $\mathcal{E}$ sont pondérés par $\mathcal{W} = \{w_{mm'} \mid (m, m') \in \mathcal{E}\}$, où $\forall m, (m, m) \notin \mathcal{E}$.

Remarque *Graphes orientés* — Lorsque les graphes sont orientés, il existe $(m, m') \in \mathcal{E}$ tel que $w_{mm'} \neq w_{m'm}$. Toutefois, on considère ici le cadre de graphes non orientés. ◇

Matrices associées. Plusieurs matrices décrivant certaines propriétés de $\mathcal{N}$ peuvent lui être associées. On définit notamment la matrice d'adjacence pondérée $\mathbf{W}$ de terme général $(\mathbf{W})_{mm'} = w_{mm'}$, $\forall (m, m') \in \mathcal{E}$ et 0 sinon, ainsi que la matrice diagonale des degrés $\mathbf{D}$ de terme général $(\mathbf{D})_{mm} = \sum_{m' \neq m} w_{mm'}$. Finalement, la matrice Laplacienne $\mathbf{L} = \mathbf{D} - \mathbf{W}$ correspondante de terme général est définie par

$$(\mathbf{L})_{mm'} = \begin{cases} \sum_{l=1}^{M} w_{lm} & \text{pour } m = m' \\ -w_{mm'} & \text{pour } m \neq m' \text{et } (m, m') \in \mathcal{E} \\ 0 & \text{sinon} \end{cases} .$$



(a) Graphe de voisinage $\mathcal{N}$ pour $M = 6$.

$$\begin{pmatrix} 0.7 & -0.5 & 0.0 & -0.2 & 0.0 & 0.0 \\ -0.5 & 2.5 & -0.2 & -0.7 & -0.7 & -0.4 \\ 0.0 & -0.2 & 0.2 & 0.0 & 0.0 & 0.0 \\ -0.2 & -0.7 & 0.0 & 1.0 & -0.1 & 0.0 \\ 0.0 & -0.7 & 0.0 & -0.1 & 0.8 & 0.0 \\ 0.0 & -0.4 & 0.0 & 0.0 & 0.0 & 0.4 \end{pmatrix}$$

(b) Matrice Laplacienne $\mathbf{L}$ associée à $\mathcal{N}$.

FIGURE 2.3. – Exemple de structures de voisinage.

Nous allons voir comment intégrer ces différentes configurations de structures à la pénalité du problème régularisé (2.1) présenté dans la section suivante.

2.3. Pénalités structurées

Dans la continuité des développements sur les modèles parcimonieux, les travaux sur la prise en compte de structures dans les termes de régularisation des problèmes d'optimisation ont connu un essor important dans la communauté de l'apprentissage à partir du milieu des années 2000 [HG3]. Une synthèse de Qiao et al. [HG37] donne un aperçu des principales méthodes et de leurs algorithmes de résolution, ainsi que des domaines d'application dans lesquels elles sont utilisées.

Remarque *Références sur les pénalités structurées* — La littérature sur ce sujet est pléthorique et concerne des travaux allant des mathématiques appliquées jusqu'au traitement du signal. Il est illusoire de prétendre pouvoir les couvrir tous. Dans cette section, nous nous concentrerons sur les travaux principaux de la communauté de l'apprentissage machine. Le lecteur intéressé pourra trouver de nombreuses références supplémentaires notamment dans [HG2, HG3, HG37]. ◇

2.3.1. Formulation

Dans un cadre supervisé, ces structures peuvent être intégrées de différentes manières à travers une norme sur f, dans le terme de pénalité du problème régularisé caractérisé par

$$\min_{f \in \mathcal{F}} \quad R_S(f) + \lambda P(f)\,, \tag{2.1}$$

où $P : \mathcal{Y} \to \mathbb{R}^+$ représente une pénalité tenant compte de la structure des données et où le paramètre de régularisation $\lambda \in \mathbb{R}^+$ contrôle le compromis entre l'adéquation du modèle aux données $R_S(f)$ et l'opérateur de pénalisation $P(f)$.

Remarque *Cadre non supervisé* — Les problèmes non supervisés peuvent se ramener à un schéma similaire, notamment dans le cadre des modèles de mélange, en maximisant sur S_c un critère pénalisé de log-vraisemblance complète

$$\max_{f \in \mathcal{F}} \quad L_{S_c}(f \mid \Theta) + \lambda P(f \mid \Theta)\,.$$

On restera toutefois dans le cadre supervisé pour conserver l'unité de cet exposé. ◇

Remarque *Composantes versus variables* — Une large partie des travaux sur les pénalités structurées ont été abordés dans le cadre de modèles linéaires agissant directement sur les D variables explicatives du problème : $f(\mathbf{x}) = \sum_j \beta_j x_j$. Toutefois, ils s'étendent assez naturellement à d'autres cadres. Pour conserver le caractère générique de la présentation, nous continuerons avec un modèle additif sur M composantes qui inclut le modèle linéaire avec $f_m(\mathbf{x}) = f_j(\mathbf{x}) = \beta_j$ et $\boldsymbol{f} = \boldsymbol{\beta} \in \mathbb{R}^D$. ◇

Bach et al. [HG3] présentent plusieurs types de structures, permettant de considérer des liens entre les composantes, qui peuvent être intégrées dans le problème (2.1) au sein de familles de pénalités $P(f)$ communes faisant intervenir des normes.

Définition 2.1 **Normes ℓ_γ** — *Pour $\boldsymbol{f} = (f_1(\mathbf{x}), \ldots, f_M(\mathbf{x}))^\top$ et $1 \leq \gamma < \infty$, la norme ℓ_γ est définie par*

$$\|\boldsymbol{f}\|_\gamma = \left(\sum_m |f_m(\mathbf{x})|^\gamma \right)^{1/\gamma} .$$

Cette définition est étendue à la quasi-norme ℓ_γ pour $0 < \gamma < 1$. Lorsque $q = 0$,

$$\|\boldsymbol{f}\|_0 = \text{card}\{f_m(\mathbf{x}) \mid f_m(\mathbf{x}) \neq 0\} .$$

Enfin, lorsque $\gamma \to \infty$,

$$\|\boldsymbol{f}\|_\infty = \sup_m |f_m(\mathbf{x})| .$$

Remarque ***Normes ou quasi-normes*** — Par abus de langage, les quasi-normes ℓ_γ, où $0 < \gamma < 1$, seront également appelées normes ℓ_γ. ◇

Remarque ***Convexité et parcimonie*** — Les normes ℓ_γ, où $0 \leq \gamma \leq 1$ favorisent des solutions dites *parcimonieuses*. Lorsqu'elles sont intégrées au problème d'optimisation (2.1), elles ont pour effet de réduire l'amplitude les composantes les moins influentes, voire de les éliminer. Parmi elles, seule la norme ℓ_1 est convexe [1]. ◇

Remarque ***Algorithmes de résolution*** — La monographie de Bach et al. [HG2] présente des techniques d'optimisation permettant de résoudre efficacement les formulations de type (2.1), notamment :

(a) Les méthodes d'***homotopie*** fondées sur des ***ensemble actifs*** de composantes, proposées dans le cadre non structuré par Osborne, Presnell et Turlach [HG35], et popularisées dans le domaine de l'apprentissage par Efron et al. [HG18].

(b) Les méthodes de ***descente proximale***, également développées dans [HG36], dont font partie les algorithmes de seuillage itératif de type FISTA (pour *Fast Iterative Shrinkage and Thresholding Algorithm*) [HG5].

Ces algorithmes s'appliquent aux pénalités structurées. ◇

Nous allons à présent évoquer les structures qui peuvent être modélisées sous forme de groupes disjoints ou recouvrants et intégrées par le biais d'une norme mixte. Nous décrirons ensuite comment ces pénalités mixtes peuvent être étendues aux structures hiérarchiques. Lorsque les structures sont définies par un graphe de voisinage, nous allons voir que la norme devient fonction de la topologie (d'une factorisation) du Laplacien de ce graphe. Enfin, nous consacrerons un passage aux problèmes régularisés par des combinaisons convexes de normes.

1. Mon manuscrit de thèse [MT1] contient un chapitre dédié au comportement des normes et quasi-normes en terme de convexité et de parcimonie ainsi qu'aux méthodes associées.

2.3.2. Pénalités mixtes

Partitions $(\mathcal{P})$**.** Lorsque la structure est constituée de groupes disjoints, $P(f)$ peut être définie a l'aide d'une norme mixte $\ell_{(\gamma_0,\gamma_1)}$ sur les composantes :

$$P(f) = \|\boldsymbol{f}\|_{\gamma_0,\gamma_1}^{\gamma_0} = \sum_{g=1}^{G} w_g \left(\sum_{m\in\mathcal{G}_g} |f_m(\mathbf{x})|^{\gamma_1} \right)^{\gamma_0/\gamma_1} \tag{2.2}$$

$$= \sum_{g=1}^{G} w_g \, \|\boldsymbol{f}_{\mathcal{G}_g}\|_{\gamma_1}^{\gamma_0} , \tag{2.3}$$

où $w_g \in \mathbb{R}^+$ représente un facteur d'échelle reflétant généralement la taille du groupe $\mathcal{G}_g$ et $\boldsymbol{f}_{\mathcal{G}_g}$ représente le vecteur des M composantes du groupe $\mathcal{G}_g$, avec $f_{m\in\mathcal{G}_g}(\mathbf{x}) = f_m(\mathbf{x})$ et $f_{m\notin\mathcal{G}_g}(\mathbf{x}) = 0$.

Remarque *Pénalité pondérée* — On notera $P_{\mathbf{w}}(f)$ une pénalité dont les composantes sont pondérées par les termes d'un vecteur $\mathbf{w}$, avec $\dim(\mathbf{w}) = \dim(\boldsymbol{f})$. ◇

On constate sur l'équation (2.3) que le représentant $\boldsymbol{f}_{\mathcal{G}_g}$ du groupe est globalement pénalisé par une norme ℓ_{γ_0} tandis que sur l'équation (2.2), on voit comment chaque composante de ce groupe est pénalisée par une norme ℓ_{γ_1}. Les propriétés des normes mixtes ont été étudiées dans les travaux de Kowalski [HG30]. En particulier, les propriétés de convexité et de parcimonie sont celles des normes non structurées.

Remarque *Convexité et parcimonie des normes mixtes* — La norme mixte définie par (2.2) est *convexe* pour $1 \leq \gamma_0, \gamma_1 \leq \infty$ et favorise la *parcimonie* lorsque $0 \leq \gamma_0, \gamma_1 \leq 1$. ◇

Plusieurs combinaisons (γ_0, γ_1), induisant différents comportements de sélection sur les groupes ou les composantes, ont été étudiées dans la littérature, avec généralement $1 \leq \gamma_0, \gamma_1 \leq 2$ pour assurer la convexité du problème. Dans le cadre de problèmes linéaires parcimonieux sur les groupes, les normes mixtes convexes de type *group lasso*, proposées initialement dans la communauté par Grandvalet et Canu [HG22] et dans la thèse de Bakin [HG4] puis développées dans [HG51], ont été largement explorées. L'utilisation de ce type de norme se généralise aux composantes associées à des Espaces de Hilbert à Noyaux Reproduisants (EHNR), avec notamment les travaux sur le *Multiple Kernel Learning* [IG50, RI6]. Les problèmes structurés et parcimonieux sur les composantes ont été étudiés comme cas particulier de normes mixtes dans [HG30] ou encore dans [CI4]. Le tableau 2.1 synthétise les comportements des pénalités mixtes utilisées dans les applications de mes travaux.

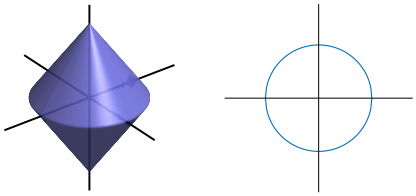
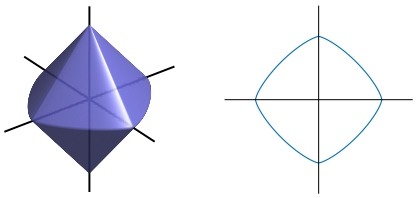
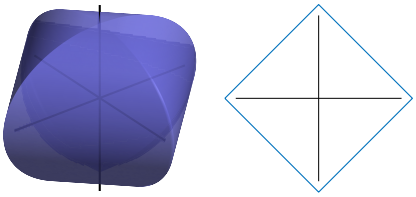
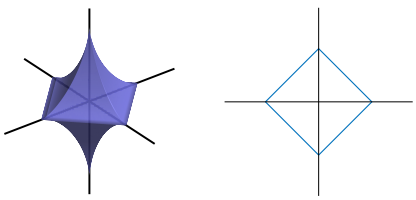
$\ell_{(1,\gamma_1)}$	Famille *group lasso* $\rightsquigarrow$ parcimonie sur les groupes	
$\ell_{(1,2)}$		[CI4]
$\ell_{(1,4/3)}$		[CI7, CI6, RI6]
$\ell_{(\gamma_0,1)}$	Famille *elitist lasso* $\rightsquigarrow$ parcimonie sur les composantes	
$\ell_{(2,1)}$		[CI4]
$\ell_{(2/3,1)}$		[CI6, RI6]

TABLE 2.1. – Boules unité associées à différentes configuration de normes mixtes. Les deux axes horizontaux des graphiques 3D représentent le plan $(f_1(\mathbf{x}), f_2(\mathbf{x}))$ associé à un premier groupe, tandis que l'axe vertical représente la composante $f_3(\mathbf{x})$ associée à un second groupe. Les singularités visibles sur ces boules dénotent un caractère parcimonieux. Le plan $(f_1(\mathbf{x}), f_2(\mathbf{x}))$ est également reproduit pour visualiser l'effet de la pénalité sur les composantes appartenant à un même groupe.

Remarque *Pénalité bridge* — Pour $R(f)$ associé au critère linéaire des moindres carrés, lorsque $\gamma_0 = \gamma_1$ et que les groupes sont réduits aux composantes individuelles, ces normes mixtes généralisent la régression *bridge* [HG19] dont font partie la régression *ridge* [HG24] avec sa pénalité ℓ_2 et le *lasso* [HG41] avec sa pénalité ℓ_1 [2]. ◇

Groupes recouvrants $(\mathcal{R})$**.** Lorsqu'une composante peut être associée à différents groupes, la résolution du problème (2.1) associé à la pénalité (2.3) conduit à des comportements plus spécifiques en terme de sélection de composantes. Jenatton, Audibert et Bach [HG28] les ont étudiés dans le cas de pénalité $\ell_{(1,2)}$ et ont montré que le support des composantes sélectionnées, $\text{supp}(f) = \{m \mid f_m(\mathbf{x}) \neq 0\}$, se trouve dans le complémentaire de l'ensemble des groupes éliminés. En notant $(u)^c$ le complémentaire de u et en considérant l'ensemble des groupes dont les composantes sont éliminées, $\mathcal{R}_0 = \{\mathcal{G}_g \in \mathcal{R} \mid f_{\mathcal{G}_g} = 0\}$, on a

$$\text{supp}\,(f) \subset \left(\bigcup_{\mathcal{G}_g \in \mathcal{R}_0} \mathcal{G}_g \right)^c .$$

Dans certains cas, cette contrainte peut s'avérer restrictive. Pour qu'une composante appartenant à différents groupes puisse être sélectionnée dans certains groupes et éliminée dans d'autres, Jacob, Obozinski et Vert [HG27] ont introduit une norme opérant sur des composantes latentes

$$P(f) \;=\; \min \left\{ \sum_{g=1}^{G} \left\| \tilde{f}_{\mathcal{G}_g} \right\|_2 \;\middle|\; \text{supp}\left(f_{\mathcal{G}_g}\right) \subseteq \mathcal{G}_g \quad ; \sum_{\mathcal{G}_g \in \mathcal{R}} \tilde{f}_{\mathcal{G}_g} = f \right\} . \tag{2.4}$$

Le problème consiste *in fine* à dupliquer les composantes appartenant à différents groupes dans une matrice d'observation $\mathbf{X}^\dagger$ de sorte que le problème (2.1) se récrit

$$\min_{f \in \mathcal{F}^\dagger} \quad R_{S^\dagger}(f) + \lambda P(f)\,,$$

où la pénalité $P(f)$ peut désormais être définie dans (2.3) à travers une partition $\mathcal{P}$ sur les composantes de $\mathbf{X}^\dagger$.

Remarque *Extension aux normes convexes* $\ell_{(1,\gamma_1)}$ — Bien que l'ensemble des résultats sur les groupes recouvrants cités ci-dessus aient initialement été présentés dans le cadre de la pénalité $\ell_{(1,2)}$, ils s'étendent aux normes mixtes $\ell_{(1,\gamma_1)}$, avec $1 < \gamma_1 \leq \infty$. ◇

Remarque *Extension aux structures de graphes* $\mathcal{N}$ — Les travaux de Jacob, Obozinski et Vert [HG27] permettent de considérer chaque paire de composantes (m, m') de l'ensemble $\mathcal{E}$ des arrêtes d'un graphe $\mathcal{N}$ comme un groupe de $\mathcal{R}$ et d'y associer une variable latente au sein de leur pénalité. ◇

2. On mentionnera tout de même les travaux concomitants de Chen, Donoho et Saunders [HG11] sur la pénalité ℓ_1 dans un cadre plus général.

2.3.3. Pénalités hiérarchiques

Graphes acycliques dirigés $(\mathcal{D})$**.** Pour les hiérarchies de composantes illustrées sur la figure 2.2a, Zhao, Rocha et Yu [HG52] ont proposé une pénalité mixte $\ell_{(\gamma_0,\gamma_m)}$ agissant sur l'ensemble des descendants de chaque composante m :

$$P(f) = \sum_{g=1}^{M} w_g \left(\sum_{m \in \mathcal{D}^g} |f_m(\mathbf{x})|^{\gamma_m} \right)^{\gamma_0/\gamma_m}, \tag{2.5}$$

avec dans leurs travaux $\gamma_m \to +\infty, \forall m$, qui permet de déterminer un chemin de régularisation sur les coefficients du modèle linéaire.

Jenatton et al. [HG29] ont étudié les propriétés de cette norme hiérarchique pour les normes mixtes $\ell_{(1,\gamma_1)}$ parcimonieuses sur les composantes identifiées comme parent d'autres composantes et avec $\gamma_1 = \{2, \infty\}$. Leur formalisme de groupes d'arbres structurés définit une relation d'ordre total sur ces groupes. Cette relation permet de réinterpréter élégamment la norme (2.5) comme une composition d'opérateurs proximaux[3] des groupes de cet ordre.

Arborescences $(\mathcal{T})$**.** Les structures arborescentes sont une autre généralisation possible des partitions $\mathcal{P}$ lorsque $H > 2$. Cette structure, définissant également une hiérarchie sous forme de DAG enraciné, diffère de $\mathcal{D}$ dans la mesure où chaque niveau de hiérarchie définit des groupes disjoints de granularités différentes sur les composantes, et non des relations de parenté.

Dans ce cas, pour $H > 2$ et en reprenant les notations décrites sur la figure 2.2b, la norme mixte est étendue sur les différents niveaux $1 \leq h \leq H$ de l'arborescence $\mathcal{T}$

$$P(f) = \sum_{g_1=1}^{G_1} w_{g_1} \left(\sum_{g_2 \in \mathcal{G}_{g_1}^H} w_{g_2} \dots \left(\sum_{g_H \in \mathcal{G}_{g_{H-1}}^H} |f_{g_H}(\mathbf{x})|^{\gamma_H} \right)^{\frac{\gamma_{H-1}}{\gamma_H}} \dots \right)^{\frac{\gamma_1}{\gamma_2}}. \tag{2.6}$$

Dans nos travaux sur l'intégration de structures, nous avons formulé des contraintes sur différents niveaux d'une arborescence de composantes au sein d'un problème régularisé et avons montré que ces contraintes correspondaient à une régularisation de type norme mixte pour une arborescence de deux niveaux (équivalente à $\mathcal{P}$), d'abord dans le cadre linéaire [CI7] puis dans le cadre de méthodes à noyaux [CI6, RI6]. Nous avons finalement étendus ces travaux à des arborescences de hauteur arbitraires dans [CI2].

3. Les opérateurs proximaux, introduits par Combettes et Pesquet [HG13], et les méthodes de résolution associées sont décrits dans la monographie de Parikh, Boyd et al. [HG36].

2.3.4. Pénalités graphiques

Pour les relations de voisinage entre les composantes illustrées sur la figure 2.3, il est possible d'intégrer la topologie d'une structure $\mathcal{N}$ à une norme ℓ_γ par le biais de la matrice d'adjacence $\mathbf{W}$ associée à la Laplacienne $\mathbf{L}$, en définissant

$$P(f) \;=\; \|f\|_{\mathbf{W}}^{\gamma} \;=\; \sum_{\substack{m\sim m'\\ m<m'}} w_{mm'}\,|f_m(\mathbf{x}) - f_{m'}(\mathbf{x})|^{\gamma}\;, \tag{2.7}$$

où $m \sim m'$ indique que $(m, m') \in \mathcal{E}$, avec généralement $\gamma = 2$. En particulier, Smola et Kondor [HG40] remettent en contexte cette régularisation au regard de l'opérateur de Laplace sur des fonctions à valeurs réelles et établissent des connexions avec les fonctions noyaux d'EHNR et la théorie spectrale des graphes.

Il est par ailleurs intéressant de constater que (2.7) intègre comme cas particulier les pénalités de type variation totale (TV pour *Total Variation*) :

$$\|f\|_{\mathbf{TV}} \;=\; \sum_{m=1}^{M-1} |f_m(\mathbf{x}) - f_{m+1}(\mathbf{x})|\;.$$

Dans ce cas,

$$\|f\|_{\mathbf{TV}} \;=\; \|f\|_{\mathbf{W}}^{1} \;=\; \sum_{\substack{m\sim m'\\ m<m'}} w_{mm'}\,|f_m(\mathbf{x}) - f_{m'}(\mathbf{x})|\;,$$

où $\mathbf{W}$ est une matrice tri-diagonale qui représente une chaîne telle que $\mathcal{E} = \{(1,2),(2,3),\ldots,(M-1,M)\}$, avec de part et d'autre de la diagonale $w_{mm+1} = w_{m+1m} = 1$ et 0 sinon.

Notons que Tibshirani et Taylor [HG42] ont également réinterprété des pénalités de type TV dans le *generalized lasso* par le biais d'une matrice $\mathbf{V} \in \mathbb{R}^{L\times M}$ structurant les composantes à considérer au sein d'une norme ℓ_1 :

$$\|f\|_{\mathbf{V}} \;=\; \sum_{l} \left| \sum_{m} v_{lm}\, f_m(\mathbf{x}) \right|\;.$$

On retrouve $\|f\|_{\mathbf{TV}}$ pour une matrice $\mathbf{V}$ bi-diagonale de taille $M-1\times M$, dont les termes généraux $v_{mm} = 1$, $v_{mm+1} = -1$ et 0 sinon, avec $\mathbf{V}^\top\mathbf{V} = \mathbf{L}$. La pénalité $\|f\|_1$ s'obtient quant-à elle pour $\mathbf{V} = \mathbf{I}_{M\times M}$.

2.3.5. Combinaisons convexes de pénalités

Au delà de la définition de pénalités structurées, il est possible d'imaginer un panel de combinaisons de pénalités très large. Lorsque les pénalités choisies définissent des ensembles convexes et que la combinaison est elle aussi convexe, les problèmes associés peuvent être résolus par exemple avec les méthodes présentées dans [HG2]. Plusieurs travaux concernant des combinaisons convexes de pénalités ont été développés pour favoriser une sélection parcimonieuse de variables corrélées.

Structures implicites. La pénalité de l'*elastic net* [HG53], dont les propriétés ont été étudiées dans les travaux de De Mol, De Vito et Rosasco [HG16], est définie par

$$P(f) \;=\; \tau\|\boldsymbol{f}\|_2^2 \,+\, (1-\tau)\|\boldsymbol{f}\|_1 \;,$$

où $\tau \in [0,1]$ module la sélection de composantes corrélées par le biais une norme ℓ_2 et l'élimination de composantes non influentes par le biais d'une norme ℓ_1.

Une alternative proposée par Bondell et Reich [HG6] est définie avec une combinaison de normes ℓ_∞ et ℓ_1 par la pénalité *OSCAR*

$$\begin{aligned} P(f) \;&=\; \tau \sum_{m<m'} \max\left\{|f_m(\mathbf{x})|, |f_{m'}(\mathbf{x})|\right\} \,+\, \|\boldsymbol{f}\|_1 \\ &=\; \sum_m \left\{\tau(m-1)+1\right\} \left|f_m^{\downarrow}(\mathbf{x})\right| \;, \end{aligned}$$

où $\tau \in \mathbb{R}^+$. Les composantes de $\boldsymbol{f}$ sont ordonnées en valeur absolue, avec $\left|f_1^{\downarrow}(\mathbf{x})\right| \leq \cdots \leq \left|f_m^{\downarrow}(\mathbf{x})\right|$, et pénalisées par un coefficient proportionnel à cet ordre.

À partir d'une combinaison ℓ_2 et ℓ_1, Argyriou, Foygel et Srebro [HG1] ont proposé une relaxation plus fine de la pénalité $\ell_2 + \ell_0$ avec la *k-support norm*. Sa norme duale s'exprime en fonction des k plus larges composantes de f en valeur absolue

$$P(f)^* \;=\; \left(\sum_{m=1}^{k} \left|f_m^{\downarrow}(\mathbf{x})\right|^2\right)^{1/2} \;.$$

La norme correspondante $P(f)$ s'écrit selon la formulation (2.4) de Jacob, Obozinski et Vert [HG27] pour les $\binom{k}{M}$ combinaisons de groupes de composantes.

Structures explicites. La pénalité du *sparse group lasso* [HG38] qui combine une norme mixte $\ell_{(1,2)}$ et une norme ℓ_1,

$$P(f) \;=\; \tau\|\boldsymbol{f}\|_{1,2} \,+\, (1-\tau)\|\boldsymbol{f}\|_1 \;,$$

où $\tau \in [0,1]$, est un cas particulier des pénalités hiérarchiques (2.5). De nombreux travaux ont été effectués avec cette pénalité [4], notamment sur des variations autour de $R(f)$, de l'algorithme permettant de trouver l'estimateur associé ou les applications. Sous un angle un peu différent, on mentionnera les travaux de Ndiaye et al. [HG34] qui ont défini pour ce problème des règles de filtrage permettant d'éliminer en amont les (groupes de) composantes ayant peu d'influence sur la solution.

Plusieurs autres variations autour de l'*elastic net* ont été proposées avec, à la place de la norme ℓ_2, une norme explicitement structurée par une matrice positive semi-définie $\mathbf{W}$ de similarités entre composantes ou encore la Laplacienne associée $\mathbf{L}$, comme par exemple dans [HG32, HG33, HG39, HG25].

4. Une recherche dans Google Scholar sur le terme exact donne environ 3500 résultats.

Pour conclure sur les combinaisons convexes de pénalités structurées, nous mentionnerons les normes coopératives parcimonieuses introduites par Chiquet, Grandvalet et Charbonnier [HG12]. La pénalité du *cooperative lasso* définie par

$$P(f) \;=\; \|\boldsymbol{f}^{+}\|_{1,2} \;+\; \|\boldsymbol{f}^{-}\|_{1,2}\;,$$

où pour tout m, $f_m^+(\mathbf{x}) = \max(0, f_m(\mathbf{x}))$ et $f_m^-(\mathbf{x}) = \max(0, -f_m(\mathbf{x}))$, $\forall m$, promeut la cohérence en terme de signe des groupes sélectionnés.

Deuxième partie

Intégration et interaction de données

Int⋆ a_tions ground

3. Intégration de sources multiples

Sommaire

3.1. Positionnement

L'intégration de données issues de sources d'information multiples et de natures différentes dans les méthodes d'apprentissage est une thématique très explorée, avec une quantité *immense* de travaux sur le sujet. Les références de cette introduction ne concernent que des articles de revue de la littérature.

3.1.1. Problématique transversale

Dans la communauté de l'apprentissage, on parlait initialement de ***sources multiples*** [IG16] mais plus encore de ***vues multiples*** : on peut notamment citer les points de vue analytique de Zhao et al. [IG75] d'une part ou spécifique au non supervisé de Fu et al. [IG24] d'autre part. Selon les disciplines, on parle aujourd'hui également de données ***multiomiques*** mais plus généralement de données ***multimodales***, avec la mise en perspective éclairante de Lahat, Adali et Jutten [IG37].

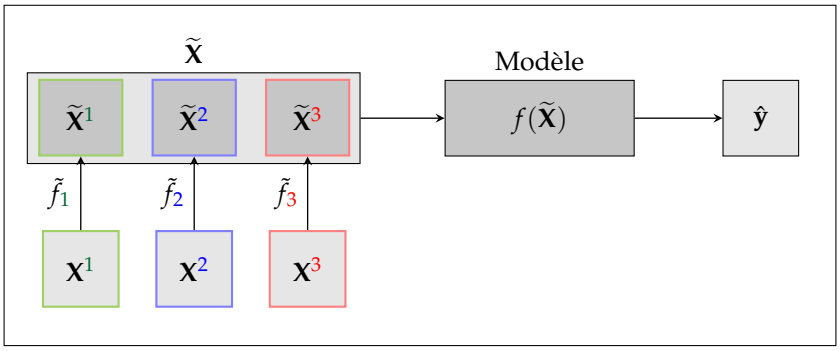
Les données ***multiomiques*** sont issues des études en sciences biologiques et médicales, avec par exemple [IG33]. On porte ici un intérêt particulier sur les problématiques dites d'***intégration***. Gomez-Cabrero et al. [IG26] discutent ces aspects en sciences du vivant. En élargissant le spectre des données à d'autres types d'information (clinique par exemple), Kline et al. [IG35] présentent les problématiques d'intégration en médecine de précision. En particulier, ils font un inventaire des méthodes associées aux schémas d'intégration ***précoce***, ***intermédiaire*** ou ***tardive***, présentés sur la figure 3.1. Ils observent que les schémas d'intégration précoce sont privilégiés tandis que les schémas d'intégration tardive sont assez étonnement dédaignés [1].

D'un point de vue plus général, les données ***multimodales*** font aussi référence aux données issues d'examens d'imagerie (scanner, échographie, IRM, etc.) dans le domaine médical mais encore aux données multimédia en vision et en traitement de la parole [IG5]. Les approches d'apprentissage profond y sont explorées de manière intensive [IG51, IG72] et les aspects autour de la ***fusion*** des données sont particulièrement scrutés, avec par exemple [IG55] dans un contexte biomédical.

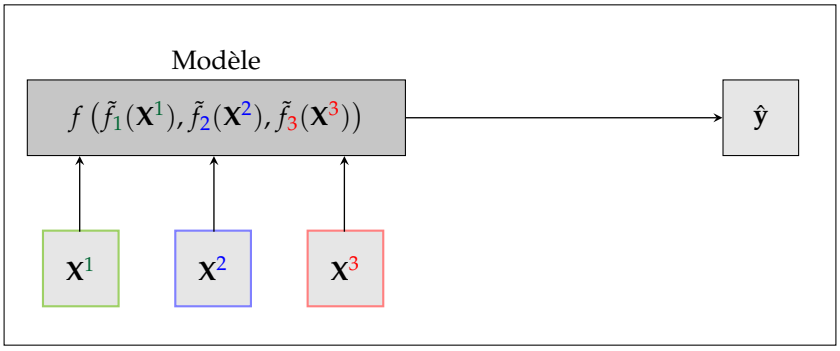
3.1.2. Typologie resserrée

Pour cette courte présentation sélective des méthodes d'***intégration intermédiaire*** (cf. figure 3.1b), nous prenons comme points de départ les typologies de Huang, Chaudhary et Garmire [IG33] présentées pour l'apprentissage supervisé ou non, ainsi que les travaux de Pierre-Jean et al. [IG49] qui ont proposé une évaluation complète de 13 méthodes non supervisées regroupées selon trois types des méthodes. On résumera encore ces méthodes en deux catégories avec celles fondées autour de la recherche d'espaces latents et celles autour de matrices positives semi-définies [2].

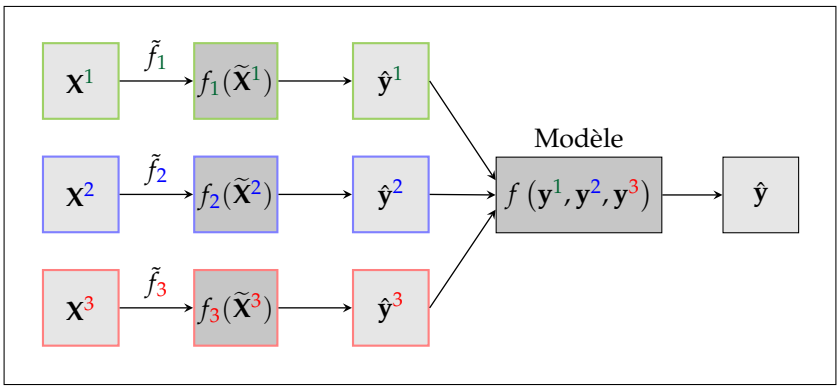
1. Nous reviendrons sur ces schémas dans les perspectives du chapitre de conclusion.
2. La description des techniques associées pourra être consultée dans [IG33], [IG24] ou [IG49].



(a) ***Schéma d'intégration précoce***. Chaque ensemble de données $\mathbf{X}^m$, éventuellement projeté indépendamment dans un espace de redescription via la fonction $\tilde{f}_m$, est combiné dans $\widetilde{\mathbf{X}}$. Les paramètres du modèle global f sont ensuite optimisés.



(b) ***Schéma d'intégration intermédiaire***. Chaque ensemble de données $\mathbf{X}^m$ est intégré dans le modèle global f. L'optimisation des paramètres des fonctions de redescription $\tilde{f}_m$ et ceux du modèle global f est conjointe.



(c) ***Schéma d'integration tardive***. Chaque ensemble de données $\mathbf{X}^m$, éventuellement projeté indépendamment dans un espace de redescription via la fonction $\tilde{f}_m$, sert à optimiser les paramètres d'un modèle indépendant f_m. L'optimisation d'un modèle global f sur l'ensemble $\{\hat{\mathbf{y}}^m\}$ des prédictions permet d'obtenir un consensus.

FIGURE 3.1. – Schémas d'intégration pour $1 \leq m \leq 3$ sources d'information.

Remarque *Matrices positives (semi-)définies* — Par abus de langage, les matrices positives semi-définies seront également qualifiées de matrices définies positives. ◇

Recherche d'espaces latents. Les méthodes fondées sur la recherche d'espaces latents consistent à extraire les caractéristiques les plus représentatives des sources d'information dont proviennent les données avec des techniques d'approximation de rang faible de matrices [HG20, IG70]. On compte parmi elles :

(a) Les méthodes dérivées d'une décomposition en valeur singulières et qui construisent des composantes orthogonales entre elles, et en particulier celles reposant sur l'analyse canonique des corrélations.

(b) Les méthodes agissant sur des matrices creuses à valeurs positives, très caractéristiques de certaines données omiques, avec comme exemple emblématique la factorisation de matrices non négatives (NMF pour *Non Negative Factorisation)*, où la contrainte d'orthogonalité est remplacée par celle de non négativité.

Ces approches peuvent être utilisées en amont d'un modèle d'apprentissage, supervisé ou non, ou y être directement intégrées.

Matrices positives semi-définies. Il existe également de nombreuses approches d'intégration fondées sur la représentation des données sous forme de matrices (symétriques) positives semi-définies, en particulier :

(a) Les méthodes où les connexions entre les observations sont représentées par autant de matrices Laplaciennes que de sources d'information, et qui reposent essentiellement sur des algorithmes de classification spectrale [IG63].

(b) Les méthodes où chaque source d'information est représentée par une similarité entre observations exprimée avec une fonction noyau appartenant généralement à un Espace de Hilbert à Noyau Reproduisant [IG53].

Ces matrices sont généralement intégrées dans la fonction de coût d'un problème d'optimisation et éventuellement associées à des contraintes supplémentaires.

3.1.3. Interprétabilité

Dans les deux familles d'approches présentées dans cette typologie, les transformations appliquées aux données masquent les spécificités des variables explicatives issues des différentes sources et les méthodes associées nécessitent des stratégies alternatives voire additionnelles pour examiner les deux aspects suivants :

1. L'identification des variables explicatives importantes issues des différentes sources dans l'estimation des classes ou la construction des partitions,
2. L'étude explicite des relations entre ces sources d'information.

Pour le premier aspect, la sélection des caractéristiques transformées est sous-jacente aux méthodes d'intégration [IG74]. Dans le cadre de méthodes à noyaux, la sélection

peut être abordée explicitement en associant un noyau à chaque variable explicative au sein d'un problème d'optimisation contraint par une norme parcimonieuse comme par exemple dans [RI6, IG71] ou plus récemment dans [IG12]. Pour les méthodes fondées sur des espaces latents, une analyse *a posteriori* des caractéristiques sélectionnées est nécessaire.

Concernant les relations entre sources d'information, les travaux de Mariette et Vialaneix [IG44] définissent des métriques mesurant la complémentarité ou au contraire l'indépendance entre les différents noyaux associés aux sources d'information. Pour des méthodes fondées sur la recherche d'espaces latents, Nguyen et Wang [IG47] déclinent les approches usuelles en ajoutant deux termes de régularisation traduisant respectivement la complémentarité ou le consensus entre les sources.

La suite de ce chapitre sera consacrée à des techniques d'***intégration intermédiaire*** de données. Dans la section 3.2, c'est sous l'angle des méthodes à noyaux supervisées, sur lesquelles une partie significative de mes travaux ont porté [CI6, RI6, CI4, CI2], qu'elles seront abordées. Dans la section 3.3, la contribution présentée, correspondant aux travaux de post-doctorat de Marie Courbariaux [RI1], s'appuiera sur un modèle de mélange d'experts axé sur la description initiale des données.

3.2. Combinaison de noyaux

L'apprentissage supervisé à partir d'une combinaison linéaire de noyaux (*MKL* pour *Multiple Kernel Learning*) a été initié par Lanckriet et al. [IG38] pour intégrer M sources d'information représentées par des matrices positives semi-définies, appelées noyaux. Il s'agit de résoudre un problème d'optimisation régularisé où le noyau effectif entre deux observations est défini par une combinaison linéaire de M noyaux : $\kappa(\mathbf{x}, \mathbf{x}') = \sum_m \eta_m \kappa_m(\mathbf{x}, \mathbf{x}')$, avec $\eta_m \in \mathbb{R}$, $\forall m$.

3.2.1. Noyaux reproduisants

Pour agréger M sources d'informations différentes, Bach, Lanckriet et Jordan [IG3] intègrent une combinaison linéaire de noyaux positifs dans la formulation associée à un Séparateur à Vaste Marge dans laquelle intervient le ***coût charnière*** :

$$Rc_S(f) = \sum_i \left[1 - y_i \left(\sum_m f_m(\mathbf{x}_i) + b \right) \right]_+ , \quad \text{où } [u]_+ = \max(0, u) ,$$

avec $b \in \mathbb{R}$ et où chaque fonction f_m est issue de $\mathcal{H}_m$, un Espace de Hilbert à Noyau Reproduisant (EHNR) associé au noyau reproduisant κ_m. En particulier,

$$\mathcal{H}_m = \left\{ f_m : \mathcal{X} \to \mathbb{R} \;\middle|\; f_m(\mathbf{x}) = \sum_i \alpha_i \kappa_m(\mathbf{x}, \mathbf{x}_i) , \text{ où } \forall i, \alpha_i \in \mathbb{R} \right\} .$$

Formulation du MKL. Le problème du *MKL* est défini par :

$$\min_{\{f_m\},b} \quad Rc_S(f) + \frac{\lambda}{2}\left(\sum_m \|f_m\|_{\mathcal{H}_m}\right)^2 , \tag{3.1}$$

où $\lambda \in \mathbb{R}^+$ règle le compromis entre l'attache au données $Rc_S(f)$ et la régularisation. L'élévation au carré de la norme mixte du critère (3.1) influence la force de la pénalité mais pas sa nature. La norme ℓ_1 sur les éléments f_m :

$$\begin{aligned}\sum_m \|f_m\|_{\mathcal{H}_m} &= \sum_m \langle f_m, f_m\rangle_{\mathcal{H}_m}^{1/2} \\ &= \sum_m \left|\sum_i \sum_{i'} \alpha_i \alpha_{i'} \kappa_m(\mathbf{x}_i, \mathbf{x}_{i'})\right|^{1/2} ,\end{aligned}$$

encourage quant à elle des solutions parcimonieuses sur les noyaux κ_m .

On observe cet aspect parcimonieux sous un autre angle dans la formulation du *MKL* donnée par Rakotomamonjy et al. [IG50] :

$$\begin{cases} \min\limits_{\{f_m\},\{\sigma_m\},b} & Rc_S(f) + \dfrac{\lambda}{2}\displaystyle\sum_m \frac{1}{\sigma_m}\|f_m\|^2_{\mathcal{H}_m} & & \text{(3.2a)} \\ \text{t. q.} & \displaystyle\sum_m \sigma_m \le 1 , \quad \sigma_m \ge 0 , & \forall m , & \text{(3.2b)} \end{cases}$$

où le noyau associé à ce problème est défini par la combinaison linéaire convexe

$$\kappa(\mathbf{x}, \mathbf{x}') \quad = \quad \sum_m \sigma_m \, \kappa_m(\mathbf{x}, \mathbf{x}') \, .$$

C'est la contrainte ℓ_1 (3.2b) imposée aux coefficients σ_m qui induit la parcimonie sur les éléments f_m et donc les noyaux κ_m. En ce sens, les formulations (3.1) et (3.2) sont équivalentes. Notons que Grandvalet [HG21] montrait déjà que (3.2) est une formulation variationnelle de (3.1) pour des problèmes dans lesquels $\boldsymbol{f} = \boldsymbol{\beta} \in \mathbb{R}^D$.

Modélisation proposée. Dans [CI6] et [RI6], nous avons étendu la formulation (3.2) à des *noyaux composites* structurés dans une partition $\mathcal{P}$, puis nous l'avons généralisée à des arborescences $\mathcal{T}$ dans [CI2]. En reprenant les notations de la figure 2.2b, la pénalité du *Composite Kernel Learning* définie par

$$\begin{cases} \min\limits_{\{f_m\},\{\sigma_{h,g}\},b} & Rc_S(f) + \dfrac{\lambda}{2}\displaystyle\sum_{g_1} \frac{1}{{\sigma_{1,g_1}}^{p_1}} \cdots \sum_{g_H} \frac{1}{{\sigma_{H,g_H}}^{p_H}} \|f_{g_H}\|^2_{\mathcal{H}_{g_H}} \\ \text{t. q.} & \displaystyle\sum_g \sigma_{h,g} \le 1 , \quad \sigma_{h,g} \ge 0 , \qquad 1 \le g \le G_H \quad \text{et} \quad 1 \le h \le H , \end{cases}$$

concorde avec la pénalité hiérarchique (2.6), pour $\gamma_k = 2\left(1 + \sum_{h=k}^H p_h\right)^{-1}$. Lorsque $H = 1$, $p_1 = 1$ et $1 \le g = m \le M$, on retrouve la formulation (3.2) du *MKL*.

3.2.2. Noyaux non positifs

Dans la communauté de l'apprentissage statistique, la majeure partie des contributions portent sur les méthodes à noyaux positifs. Cependant, d'autres approches ne supposent pas cette propriété. On pourra notamment se reporter aux travaux de Loosli, Canu et Ong [IG42] qui se placent dans des espaces de Kreĭn à noyaux reproduisants ainsi qu'aux références citées pour les autres approches existantes. En dehors du cadre des EHNR également, nous avons travaillé dans [CI4] sur l'intégration de M sources d'information encodées comme des proximités entre observations ne vérifiant pas nécessairement les propriétés des noyaux positifs.

Nous définissons dans cette approche, à travers une collection de M noyaux $\{\kappa_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}\}$, un espace $\mathcal{F}_S$ de recherche des solutions supporté par les N données d'un ensemble d'apprentissage S, tel que

$$\mathcal{F}_S = \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; f(\mathbf{x}) = \sum_i \sum_m \alpha_{im}\, \kappa_m(\mathbf{x}, \mathbf{x}_i) \;=\; K(\mathbf{x}, \cdot)^\top \boldsymbol{\alpha}\,. \quad \text{où } \boldsymbol{\alpha} \in \mathbb{R}^{NM} \right\},$$

avec $K(\mathbf{x}, \cdot)^\top = [\kappa_1(\mathbf{x}, \mathbf{x}_1), \ldots, \kappa_1(\mathbf{x}, \mathbf{x}_N), \ldots, \kappa_M(\mathbf{x}, \mathbf{x}_1), \ldots, \kappa_M(\mathbf{x}, \mathbf{x}_N)]$.

Contrairement aux formulations associées au *MKL* où le noyau issu de la combinaison linéaire des M noyaux s'intègre naturellement dans la formulation d'un *SVM*, le noyau effectif de notre problème s'exprime par la concaténation des M noyaux ce qui offre un cadre flexible en terme de structuration. En effet, en définissant le problème général d'optimisation par

$$\min_{f,\,b} \quad Rc_S(f)^2 \;+\; \frac{\lambda}{\gamma_0} P(f)\,, \tag{3.4}$$

avec $\frac{\lambda}{\gamma_0} \in \mathbb{R}^+$, cette modélisation, où f est définie à travers $\boldsymbol{\alpha}$, permet de configurer $P(f)$ selon deux options :

$$\begin{cases} P_d(f) \;=\; \displaystyle\sum_i \left[\sum_m |\alpha_{im}|^{\gamma_1} \right]^{\gamma_0/\gamma_1}, & \text{pour une structure sur les données,} \\ P_k(f) \;=\; \displaystyle\sum_m \left[\sum_i |\alpha_{im}|^{\gamma_1} \right]^{\gamma_0/\gamma_1}, & \text{pour une structure sur les noyaux,} \end{cases}$$

avec $\gamma_0 \in \{1, 2\}$ et $\gamma_1 \in \{1, 2\}$. On retrouve donc selon les configurations les pénalités mixtes de type *group-lasso* ou *elitist-lasso* (cf. section 2.3.2) appliquées relativement aux données ou aux noyaux, en fonction de la nature de l'application considérée.

Notons finalement que, bien que la forme du noyau effectif obtenue par la concaténation des M noyaux ne soit pas adaptée aux *SVM*, la solution du problème (3.4), faisant intervenir le coût charnière élevé au carré, peut être obtenue efficacement avec des algorithmes proximaux [HG36].

3.2.3. Aperçu des résultats

Noyaux reproduisants. La généralisation dans [CI2] des travaux associés aux *noyaux composites* a permis de confirmer les liens entre la formulation variationnelle (3.3) et une norme mixte sur une arborescence de trois niveaux (2.6). Pour $H \geq 4$, un schéma général de preuve, non publié et dont la technicité croit avec H, se dessine systématiquement pour établir l'équivalence entre les deux formulations.

Cette approche a été appliquée à des problèmes d'interfaces cerveau-machine dans une optique d'interprétation des variables sélectionnées [IG9]. Des noyaux linéaires ont été associés à chaque mesure temporelle issue d'un EEG, tandis que la structure arborescente a été établie en fonction de la répartition spatiale des électrodes (cf. figure 1.1b). Les performances en classification obtenues étaient compétitives vis-à-vis des approches concurrentes, notamment avec l'utilisation de normes non convexes et parcimonieuses sur l'ensemble de la structure spatiale. De plus, l'analyse des résultats a permis de montrer que notre approche sélectionnait des électrodes dans des zones cohérentes par rapport à l'état des connaissances sur ce type de tâches.

Noyaux non positifs. En plus du modèle et de l'algorithme proposés, des bornes de généralisation ont été établies dans [CI4] pour la norme ℓ_1 et dans certaines configurations parcimonieuses sur les noyaux, à savoir pour $P_d(f)$ lorsque $\gamma_0 = 2$ et $\gamma_1 = 1$ (*elitist-lasso* où pour chaque individu, un sous-ensemble de noyaux est sélectionné) et pour $P_k(f)$ lorsque $\gamma_0 = 1$ et $\gamma_1 = 2$ (*group-lasso* où un même sous-ensemble de noyaux est sélectionné pour l'ensemble des individus).

Bien que très flexible quand à la définition de noyaux et de structures, cette approche n'a pas été testée sur un problème réel. Des applications biologiques ou médicales, avec un nombre raisonnable (plusieurs centaines d'observations) auraient pu être envisagées. Cependant, les performances en classification, comparables à celles de *MKL* classiques utilisant des sommes pondérées de noyaux, ne compensaient pas le coût calculatoire supplémentaire nécessaire pour entraîner le modèle.

3.3. Mélange d'experts

Dans un cadre de classification partiellement supervisée, les modèles d'experts fournissent un cadre élégant pour exploiter simultanément deux types d'information. Ils permettent d'exprimer, dans des modèles classiques de mélange de K catégories, une relation conditionnelle entre un couple de variables aléatoires (Y, X), où Y est associée à un comportement observé et X à des *variables concomitantes*. Ainsi, ils permettent de comprendre, via les paramètres estimés, les effets des variables concomitantes dans la construction des catégories latentes d'un côté et dans l'explication des comportements observés de l'autre.

3.3.1. Contexte et cadre général

Ces travaux, présentés dans [RI1], ont été réalisés avec comme objectif le sous-typage en K catégories de N patients issus d'une cohorte pour laquelle nous disposions d'un suivi clinique et d'informations génétiques.

Précisions sur le contexte et les notations. Les couples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N} \in \boldsymbol{S}$ sont des réalisations indépendantes et identiquement distribuées du couple de variables aléatoires (X, Y), correspondant respectivement aux données génétiques et cliniques. Plus précisément, $\mathbf{Y}$ est une matrice multidimensionelle de taille $N \times L$ correspondant aux réponses de N patients sur L variables cliniques indicées par $1 \leq \ell \leq L$. On reprendra les conventions décrites à la section 1.3 pour la matrice $\mathbf{X}$ de taille $N \times M$ correspondant aux données génétiques et le vecteur $\mathbf{z}$ de taille N correspondant au sous-typage en K catégories recherché sur ces patients.

Mélange simple d'experts. Dans un modèle d'expert, où le k^{e} expert est représenté par la densité de probabilité $f_k(\cdot \,|\, \Theta_k(\cdot))$ et où $\forall (\mathbf{x}_i, \mathbf{y}_i) \in \boldsymbol{S}$, on définit

$$\mathbb{P}(\mathbf{y}_i \,|\, \mathbf{x}_i) = \sum_k f_k(\mathbf{y}_i \,|\, \Theta_k(\mathbf{x}_i))\, \eta_k(\mathbf{x}_i)\,, \quad \text{avec} \ \sum_k \eta_k(\mathbf{x}_i) = 1\,. \tag{3.5}$$

Les paramètres Θ_k ainsi que les poids η_k associés à cet expert sont dépendants des variables concomitantes $\mathbf{X}$. Il existe différentes configurations de mélanges d'experts [IG28]. Nous avons eu recours au mélange *simple*, illustré sur la figure 3.2, où la distribution des variables associées aux réponses dépend des variables latentes d'appartenance aux catégories, dépendant elles mêmes des variables concomitantes.

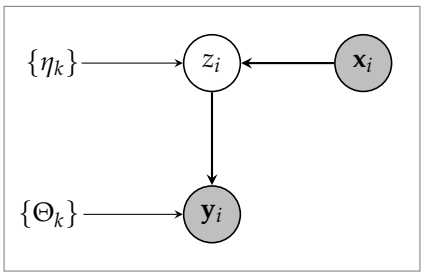


FIGURE 3.2. – Mélange d'expert simple selon Gormley et Früwirth-Schnatter [IG28], avec $\mathbb{P}(\mathbf{y}_i, z_i \,|\, \mathbf{x}_i) = \mathbb{P}(\mathbf{y}_i \,|\, z_i)\, \mathbb{P}(z_i \,|\, \mathbf{x}_i)$.

Dans cette configuration, $\mathbb{P}(\mathbf{y}_i, z_i \,|\, \mathbf{x}_i) = f_{z_i}(\mathbf{y}_i \,|\, \Theta_{z_i}(\mathbf{x}_i))\, \eta_{z_i}(\mathbf{x}_i)$, avec

$$\mathbf{y}_i \,|\, \mathbf{x}_i, z_i = k \sim f_k(\mathbf{y}_i \,|\, \Theta_k(\mathbf{x}_i))\,, \tag{3.6a}$$

$$\text{et} \ \ \mathbb{P}(z_i = k \,|\, \mathbf{x}_i) = \eta_k(\mathbf{x}_i)\,. \tag{3.6b}$$

3.3.2. Modèle longitudinal et grande dimension

Sur la base du cadre décrit ci-dessus, nous avons proposé un mélange d'experts où les symptômes cliniques des patients sont observés sur différentes visites réparties sur plusieurs années. Nous cherchons donc à décrire les catégories selon l'évolution des symptômes dans le temps, mais aussi selon les marqueurs génétiques collectés à l'inclusion des patients dans l'essai clinique.

Aspect longitudinal. On note $y_{i\ell_{(v)}}$, la variable clinique ℓ observée pour le patient i à la visite v, pour $1 \leq v \leq V$. Pour prendre en compte l'aspect longitudinal du suivi clinique, nous avons adapté le modèle (3.6) ainsi

$$y_{i\ell_{(v)}} \,|\, \mathbf{x}_i, z_i = k \sim f_k(y_{i\ell_{(v)}} \,|\, \{\boldsymbol{\alpha}_{\ell k}, \sigma_{\ell k}\}) \,,$$
$$\text{et} \quad \mathbb{P}(z_i = k \,|\, \mathbf{x}_i) = \eta_k(\mathbf{x}_i \,|\, \boldsymbol{\omega}_k) \,,$$

en définissant le modèle de régression à poids logistiques suivant

$$\left(y_{i\ell_{(v)}} \,|\, z_i = k\right) = \sum_{p=0}^{P} \alpha_{\ell k p} \, (t_{iv})^{\,p} + \sigma_{\ell k} \, \varepsilon_{i\ell_{(v)}} \,, \tag{3.7a}$$

$$\text{tel que} \quad f_k(y_{i\ell_{(v)}} \,|\, \{\boldsymbol{\alpha}_{\ell k}, \sigma_{\ell k}\}) \sim \mathcal{N}\left(\sum_{p=0}^{P} \alpha_{\ell k p} \, t_{iv}^{\,p} \,,\, \sigma_{\ell k}^2\right) \,, \tag{3.7b}$$

$$\text{et} \quad \eta_k(\mathbf{x}_i \,|\, \boldsymbol{\omega}_k) = \frac{\exp\left(\omega_{k0} + \boldsymbol{\omega}_k^\top \mathbf{x}_i\right)}{\sum_{k'} \exp\left(\omega_{k'0} + \boldsymbol{\omega}_{k'}^\top \mathbf{x}_i\right)} \,, \tag{3.7c}$$

où t_{iv} est la métrique temporelle, par exemple l'âge ou le temps écoulé depuis le diagnostic pour le patient i à la visite v, P est le degré considéré dans la régression polynomiale [3], et $\{\alpha_{\ell k p}\}$, $\{\sigma_{\ell k}\}$ et $\{\boldsymbol{\omega}_k\}$ sont les (vecteurs de) paramètres à estimer, avec $\boldsymbol{\omega}_1 = \mathbf{0}_M$ pour des questions d'identifiabilité.

De plus, $\varepsilon_{i\ell_{(v)}} \underset{iid}{\sim} \mathcal{N}(0,1)$ implique certaines hypothèses d'indépendance conditionnelle entre les variables, les patients et les visites lorsque le sous-type est connu. D'une part, les variables cliniques sont choisies pour être aussi indépendantes que possible et la corrélation temporelle restante après la régression polynomiale est supposée faible. D'autre part, la corrélation entre les individus devrait essentiellement être l'expression d'un sous-type similaire de la maladie.

Enfin, si l'hypothèse gaussienne (3.7b) associée à l'expert k ne s'applique pas à la variable ℓ, une autre famille de régression, logistique ou de Poisson par exemple, pourrait être considérée à la place, pour un coût additionnel marginal.

Aspect grande dimension. Afin de tenir compte de la grande dimensionnalité des données génomiques, nous avons intégré une pénalité parcimonieuse dans l'algorithme EM utilisé pour l'inférence du modèle [IG23]. La log-vraisemblance des

3. Généralement, $P = 2$ est suffisant pour modéliser la dynamique.

données complètes (observées et latentes) pénalisée $L_{\mathbf{S}_c}(f \mid \Theta = \{\boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\omega}\}) - P(\boldsymbol{\omega})$ maximisée à chaque itération est

$$\sum_i \sum_k z_{ik} \left[\log \left[\eta_k(\mathbf{x}_i \mid \boldsymbol{\omega}_k) \right] + \sum_\ell \sum_v \log \left[f_k(y_{i\ell_{(v)}} \mid \{\boldsymbol{\alpha}_{\ell k}, \sigma_{\ell k}\}) \right] \right] - \lambda \sum_k P(\boldsymbol{\omega}_k)\,,$$

où $\lambda > 0$ contrôle le niveau de parcimonie induit par la pénalité, structurée ou non, appliquée à $\boldsymbol{\omega}$ et où $\eta_k(\cdot)$ et $f_k(\cdot \mid \Theta)$ sont définis dans (3.7).

3.3.3. Aperçu des résultats

Le modèle a été évalué sur des simulations établies à partir d'une cohorte de patients atteints de la maladie de Parkinson [IG20]. Nous avons étudié sa capacité à :

1. Retrouver le nombre de clusters simulés et les affectations des patients,
2. Identifier les variables génétiques qui influencent la construction des clusters,
3. Estimer avec précision les paramètres du modèle.

Nous avons comparé nos résultats à deux versions oracles de notre mélange d'experts, l'un avec l'ensemble des paramètres fixés avec les vraies valeurs et l'autre avec les paramètres $\{\boldsymbol{\omega}\}$, liés aux variables génétiques, fixés. Nous avons pu vérifier que notre approche se comportait de façon cohérente sur l'ensemble des aspects.

Nous avons aussi comparé notre approche globale, intégrant les informations cliniques et génétiques, à une approche en deux temps. Le premier est un clustering des patients à partir du suivi clinique uniquement, le second consiste à valider l'association génétique avec les clusters trouvés via une la régression logistique pénalisée. Les résultats ont été observés dans deux configurations pour la première étape de clustering, l'une avec les paramètres $\{\boldsymbol{\alpha}, \boldsymbol{\sigma}\}$ fixés aux valeurs réelles, l'autre avec ces paramètres estimées. Cela nous a permis d'observer la pertinence d'intégrer les informations génétiques dans un processus global de clustering.

Les résultats obtenus en conditions réellessur la cohorte mentionnée ci-dessus ont permis d'identifier quatre clusters correspondant à des symptômes cliniques caractérisant différents stades de la maladie. Nous avons identifié 15 variables génétiques ayant un potentiel impact sur le sous-typage de patients. Celle ressortant le plus significativement avait été identifiée par des précédentes études comme étant impliquée dans la maladie de Parkinson, qui semble cependant très sensible aux facteurs environnementaux [IG4]. Un repositionnement de ces travaux sur une cohorte enrichie avec des données environnementales serait certainement pertinent.

D'autres perspectives sur ces travaux pourraient concerner la précision de la modélisation, en tenant compte de la dynamique spécifique de chaque individu pour l'aspect longitudinal et en intégrant les structures associées aux variables génétiques. D'un point de vue plus général, la stratégie de sélection de variables concomitantes pourrait reposer sur des techniques de validation croisée plus efficaces, comme dans les travaux de [IG73] par exemple, ou sur la définition d'un critère de choix de modèle global intégrant également le nombre de clusters.

4. Interaction de composantes

Sommaire

4.1. Positionnement

L'étude d'interactions entre composantes d'un ensemble de données est fréquemment examinée en statistique appliquée. Dans ce positionnement, l'exemple de l'épidémiologie génétique sera utilisé pour caractériser différents types d'interactions et cibler quelques approches permettant de les analyser. Ici encore, le sujet est intensivement étudié, et les références de ce positionnement concerneront majoritairement des articles de revue de la littérature ou des thèses.

4.1.1. Illustration en épidémiologie génétique

En épidémiologie génétique, l'*interaction gène-environnement* ($G \times E$) est étudiée depuis de nombreuses décennies [IG48], avec des approches évoluant avec les capacités d'acquisition des données et en particulier avec l'avènement des études d'association pangénomique (GWAS pour *Genome Wide Association Studies*) [IG58]. Ces études permettent d'observer statistiquement le lien entre un phénotype, une maladie par exemple, et des centaines de milliers de marqueurs génétiques [IG60, IG65].

Si elles sont un outil incontournable pour analyser l'architecture complexe de processus biologiques, elles n'ont toutefois pu expliquer qu'une partie relativement faible des variations phénotypiques observées à partir d'analyses de liaison plus classiques [IG43]. Cet écart est connu sous le terme d'héritabilité manquante. Pour tenter de dissiper ce phénomène et mieux comprendre les mécanismes sous-jacents à l'expression d'un phénotype, on peut explorer d'autres axes.

Interaction $G \times E$. En particulier, on peut s'intéresser aux liens entre entre le facteur génétique et environnemental. VanderWeele et Knol [IG62] ont établi un tutoriel pour analyser ce type d'interaction, en commençant par l'analyse de mesures liées à une table de contingence croisant facteurs génétique et environnemental, puis en s'appuyant sur des modèles de régression associés à des tests statistiques. Notons qu'avec l'évolution des capacités d'acquisition et donc de la taille des données, les modèles de régression parcimonieux sont désormais largement utilisés [IG76].

Interaction $G \times G$. L'épistasie, qui consiste à analyser les liens entre gènes plutôt que les gènes indépendamment, est un cas particulier d'interaction $G \times E$. Cordell [IG19] a proposé un des premiers inventaires des méthodes permettant d'analyser cet aspect, avec comme point de départ – ici encore – l'étude de termes d'interaction issus de modèles de régression couplés à des tests d'hypothèses d'association. Un aperçu plus récent de l'étendue des approches permettant de modéliser l'épistasie peut être consulté dans le chapitre 2 de la thèse de Stanislas [IG56, figures 2.2 et 2.3].

4.1.2. Modélisation de l'interaction

Les différents états de l'art cités dans cette illustration suggèrent que l'étude d'interaction à travers des modèles – quasi-exclusivement *linéaires* – de régression associés aux tests d'hypothèse est une approche centrale autour de cette problématique.

Modélisation biologique. Notamment, cette approche est particulièrement adaptée à l'un des cinq modèles biologiques d'interaction $G \times E$ proposé par Ottman [IG48], illsutré sur la figure 4.1, où le facteur génétique et environnemental interviennent de façon conjointe dans la définition d'un phénotype P.

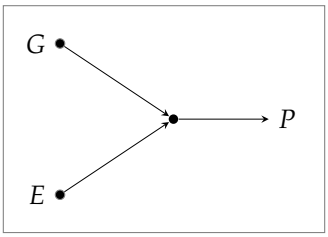


FIGURE 4.1. – Modèle d'interaction $G \times E$ selon Ottman [IG48, modèle D].

Modélisation statistique. En reprenant les notations du chapitre 1, $\mathbf{y}$ représente le vecteur de taille N associé au phénotype P, et les matrices $\mathbf{X}^G$ de taille $N \times D_G$ et $\mathbf{X}^E$ de taille $N \times D_E$ sont les observations respectivement associées aux facteurs génétique G et environnemental E. La littérature portant essentiellement sur des modèles *linéaires* d'interaction, la suite du chapitre sera présentée dans ce cadre.

Un modèle linéaire d'interaction $G \times E$ s'écrit

$$y_i = \mathbf{x}_i^G \boldsymbol{\omega}_G + \mathbf{x}_i^E \boldsymbol{\omega}_E + \mathbf{x}_i^G \boldsymbol{\Delta}_{GE} (\mathbf{x}_i^E)^\top + \epsilon_i \,, \tag{4.1}$$

où $\epsilon_i \in \mathbb{R}$ est une erreur résiduelle et où :
- les vecteurs $\boldsymbol{\omega}_G \in \mathbb{R}^{D_G}$ et $\boldsymbol{\omega}_E \in \mathbb{R}^{D_E}$ désignent les *effets simples* liés respectivement aux facteurs G et E,
- la matrice $\boldsymbol{\Delta}_{GE} \in \mathbb{R}^{D_G \times D_E}$ contient les termes d'*interaction croisée* entre toutes les paires de composantes associées aux facteurs G et E.

Remarque *Interaction $G \times G$* — Lorsqu'on s'intéresse aux interactions génétiques dans leur propre environnement, le modèle général (4.1) devient

$$y_i = \mathbf{x}_i^G \boldsymbol{\omega}_G \; + \sum_{g=1}^{D_G} \sum_{\substack{g'=1 \\ g' \neq g}}^{D_G} \delta_{gg'} \, x_{ig}^G \cdot x_{ig'}^G \; + \; \epsilon_i \,, \tag{4.2}$$

où $\{\delta_{gg'}\}$ désignent les termes d'*interaction simple* du facteur G. ◇

4.1.3. Estimation de l'interaction

Nous considérerons ici l'estimation de l'interaction de deux composantes à travers un processus en deux étapes. On peut trouver les motivations et les lignes directrices d'une telle approche pour les modèles régularisés de régression (à effets simples) dans les travaux de Bécu et al. [IG7] par exemple.

Criblage versus sélection. Les deux étapes du processus sont les suivantes.

(a) L'étape de *criblage* consiste à sélectionner un support de composantes impliquées dans la solution via des modèles régularisés de la forme

$$\min_{\boldsymbol{\omega}_G, \boldsymbol{\omega}_E, \boldsymbol{\Delta}_{GE}} \quad R_S(\boldsymbol{\omega}_G, \boldsymbol{\omega}_E, \boldsymbol{\Delta}_{GE}) + \lambda_G P_G(\boldsymbol{\omega}_G) + \lambda_E P_E(\boldsymbol{\omega}_E) + \lambda_{GE} P_{GE}(\boldsymbol{\Delta}_{GE}), \quad (4.3)$$

où les pénalités P_G, P_E et / ou P_{GE} peuvent induire des solutions parcimonieuses. Rappelons que, dans ce contexte général d'étude d'interaction, de nombreuses méthodes reposent sur ce type de mécanisme [IG76].

(b) L'étape de *sélection* à proprement parler consiste à déterminer la pertinence des composantes issues de l'étape de criblage par le biais d'une procédure de tests statistiques [IG54]. Elle permet en outre, dans un contexte médical ou biologique, de quantifier la pertinence de l'interaction, via une p-valeur par exemple.

Dépendance entre effets simples et interaction. Une interaction, simple ou croisée, peut être conditionné selon l'implication de chaque composante qui la constitue dans la construction du phénotype. Il existe deux hypothèses de dépendance.

(a) L'hypothèse *SD* (pour *strong dependency*) est la plus courante (voir [IG8] et la discussion qui s'y rapporte), et signifie qu'une interaction est potentiellement effective si et seulement si les effets simples des deux composantes font partie du support sélectionné dans l'étape de criblage.

(b) L'hypothèse *WD* (pour *weak dependency*) signifie quant-à-elle qu'une interaction peut être effective en présence de l'effet simple d'une seule composante.

Formellement, pour tout couple de composantes (m, m'), si ω_m, $\omega_{m'}$ et $\delta_{mm'}$ sont les coefficients respectivement liés aux vecteurs des effets simples $\boldsymbol{\omega}_G$, $\boldsymbol{\omega}_E$ et à la matrice d'interaction $\boldsymbol{\Delta}_{GE}$ du modèle (4.1), alors

$$\begin{array}{lllll} (SD) & \delta_{mm'} \neq 0 & \Rightarrow & \omega_m \neq 0 \quad \text{et} & \omega_{m'} \neq 0\,, \\ (WD) & \delta_{mm'} \neq 0 & \Rightarrow & \omega_m \neq 0 \quad \text{ou} & \omega_{m'} \neq 0\,. \end{array}$$

La suite de ce chapitre sera consacrée à une présentation synthétique des travaux de thèse de Florent Guinot autour d'***approches structurées pour l'interaction*** dans un contexte de très grande dimension. Ces travaux intègrent, en amont ou au sein de l'étape de criblage, la définition de ***descripteurs compressés*** obtenus à partir d'une structure arborescente $\mathcal{T}$ sur les variables initiales (cf. section 2.2 et figure 2.2b).

4.2. Interaction $G \times E$

Les travaux issus de [RI2] ont pour objectif d'observer les liens entre un facteur génétique représenté par des marqueurs génétiques $\mathbf{X}^G$ et un facteur environnemental représenté par des marqueurs métagénomiques $\mathbf{X}^E$. Dans ce contexte de très grande dimension, nous avons cherché à tirer partie des caractéristiques structurelles de chaque facteur afin de créer de nouvelles représentations compressées.

4.2.1. Principe général

Notre approche consiste à mettre en cascade différentes étapes. Dans un premier temps, une étape de construction de nouveaux descripteurs (ED) et une étape de criblage (EC) sont réalisées indépendamment pour chaque facteur. Dans un second temps, une étape de sélection (ES) s'effectue à travers le modèle d'interaction (4.1).

(ED) Définition de descripteurs compressés $\rightsquigarrow$ indépendamment pour $\mathbf{X}^G$ et $\mathbf{X}^E$

(ED_A) Définition d'une structure de groupes arborescente sur les variables

$$\mathcal{T}_G = \{\mathcal{T}_G^h\}_{h=0}^{H_G}, \text{ avec } \mathcal{T}_G^h = \left\{\mathcal{G}_1^h, \ldots, \mathcal{G}_{G_h}^h\right\} \text{ et } \widetilde{D}_G = \sum_{h=h_g^-}^{h_g^+} G_h$$

$$\mathcal{T}_E = \{\mathcal{T}_E^h\}_{h=0}^{H_E}, \text{ avec } \mathcal{T}_E^h = \left\{\mathcal{G}_1^h, \ldots, \mathcal{G}_{E_h}^h\right\} \text{ et } \widetilde{D}_E = \sum_{h=h_e^-}^{h_e^+} E_h$$

(ED_E) Exploitation de l'arborescence pour définir des descripteurs pondérés

$$\mathbf{X}^G \xmapsto{\rho_G \in \mathbb{R}^{\widetilde{D}_G}} \widetilde{\mathbf{X}}^G, \text{ avec } \dim(\widetilde{\mathbf{X}}^G) = \widetilde{D}_G \qquad \widetilde{\mathbf{S}}_G = \left(\widetilde{\mathbf{X}}^G, \mathbf{y}\right)$$

$$\mathbf{X}^E \xmapsto{\rho_E \in \mathbb{R}^{\widetilde{D}_E}} \widetilde{\mathbf{X}}^E, \text{ avec } \dim(\widetilde{\mathbf{X}}^E) = \widetilde{D}_E \qquad \widetilde{\mathbf{S}}_E = \left(\widetilde{\mathbf{X}}^E, \mathbf{y}\right)$$

(EC) Criblage des effets simples compressés $\rightsquigarrow$ indépendamment pour $\widetilde{\mathbf{S}}_G$ et $\widetilde{\mathbf{S}}_E$

$$\boldsymbol{\beta}_G^\star = \text{argmin}_{\boldsymbol{\beta}_G} R_{\widetilde{\mathbf{S}}_G}(\boldsymbol{\beta}_G) + \lambda_G P_{\rho_G}(\boldsymbol{\beta}_G) \qquad \text{supp}(\boldsymbol{\beta}_G^\star)$$

$$\boldsymbol{\beta}_E^\star = \text{argmin}_{\boldsymbol{\beta}_E} R_{\widetilde{\mathbf{S}}_E}(\boldsymbol{\beta}_E) + \lambda_E P_{\rho_E}(\boldsymbol{\beta}_E) \qquad \text{supp}(\boldsymbol{\beta}_E^\star)$$

(ES) Sélection des termes d'interaction $\forall\, (g \in \text{supp}(\boldsymbol{\beta}_G^\star), e \in \text{supp}(\boldsymbol{\beta}_E^\star))$

$$\left(\omega_g^\star, \omega_e^\star, \delta_{ge}^\star\right) = \text{argmin}_{\omega_g, \omega_e, \delta_{ge}} R_{\widetilde{\mathbf{S}}_{GE}}(\omega_g, \omega_e, \delta_{ge}), \text{ avec } R_{\widetilde{\mathbf{S}}_{GE}} \mid (4.1)$$
$$\text{test}\left(\delta_{ge}^\star\right)$$

Hypothèses de dépendance. Par construction, cette approche respecte l'hypothèse de dépendance (SD), contrairement à un problème régularisé de type (4.3) qui, sans contrainte additionnelle, ne garanti le respect d'aucune hypothèse de dépendance.

Criblage et sélection. Les étapes (EC) et (ES), typiques des schémas en deux étapes évoqués en section 4.1, ont été adaptées à un modèle d'interaction et plusieurs optimisations rendent l'approche plus robuste. En particulier, l'étape (E_C) intègre des mécanismes d'échantillonnage pour améliorer la stabilité et l'étape (E_S) inclut des techniques de correction pour améliorer la puissance des tests statistiques.

4.2.2. Définition d'un structure de groupes arborescente (ED_A)

La spécificité de ces travaux, qui rend l'ensemble de cette approche en cascade particulièrement efficace du point de vue calculatoire, réside dans la définition des descripteurs. Nous avons travaillé sur des expansions d'arborescences $\mathcal{T}$ obtenues à partir d'une classification hiérarchique ascendante des variables, selon la méthode de Ward [IG69], avec des distances pouvant être adaptées à la nature des données.

Arborescence $\mathcal{T}_G$. Lorsque l'information génétique est disponible sous forme de polymorphismes nucléotidiques (SNP pour *Single Nucleotide Polymorphism*), l'arborescence $\mathcal{T}_G$ peut-être définie à l'aide d'une classification hiérarchique, intégrant le déséquilibre de liaison [1] en tant que mesure de la dissimilarité, afin de regrouper les SNP selon leur structure spatiale [IG22].

Plus précisément, cette classification hiérarchique modélise la matrice du déséquilibre de liaison de façon bloc diagonale de sorte que seuls des SNP contigus sur le génome puissent être regroupés. Couplé à des pré-calculs de dissimilarités dans des structures de stockage appropriés (mini-tas), l'arborescence s'obtient avec une complexité quasi-linéaire [IG1].

Arborescence $\mathcal{T}_E$. Une approche couramment utilisée lors de l'analyse de données métagénomiques consiste à regrouper les séquences en unités taxonomiques (OTU pour *Operational Taxonomic Units*) [IG10], chaque OTU représentant des espèces microbiennes regroupées selon un certain degré de similarité. Des méthodes plus récentes basées sur des techniques de débruitage ont conduit à la définition de variants de séquences d'amplicons (ASV pour *Amplicon Sequence Variant*), qui peuvent être considérés comme des versions affinées des OTU [IG14].

Une revue des méthodes statistiques et computationnelles permettant l'analyse de telles données, en fonction de différents objectifs et / ou technologies, est établie dans [IG40]. Bien que la structure des espèces microbiennes puisse être définie selon ces techniques ou selon une connaissance biologique (par exemple l'arbre phylogénétique sous-jacent), il est également pertinent d'utiliser des distances plus classiques basées sur l'abondance des OTU pour définir une hiérarchie $\mathcal{T}_E$.

1. Le déséquilibre de liaison entre deux SNP $\mathbf{x}^m$ et $\mathbf{x}^{m'}$ est défini par $1 - R^2(\mathbf{x}^m, \mathbf{x}^{m'})$.

4.2.3. Exploitation de l'arborescence (ED_E)

Cette sous-étape a pour objectif de réduire la dimension originale des variables tout en tenant compte des différentes granularité de la structure de arborescente. Plusieurs approches peuvent être envisagées pour trouver le niveau de coupe optimal (et donc le nombre optimal de groupes) dans une structure arborescente obtenue par une classification hiérarchique (voir par exemple [IG46] ou [IG27]). Quelle que soit l'approche choisie, la recherche de cette coupe optimale nécessite une exploration systématique de différents niveaux de la hiérarchie.

Pour contourner cette exploration souvent coûteuse, nous avons proposé l'alternative suivante illustrée sur la figure 4.2 :

(a) Expansion de l'arborescence avec l'ensemble des groupes contenus sur les différents niveaux $0 \leq h \leq H$.
(b) Pondération ρ de chaque groupe en fonction de la distance s entre deux niveaux successifs dans l'arborescence.
(c) Définition de descripteurs compressés $\widetilde{\mathbf{X}}$ selon ces groupes pondérés.

Expansion de l'arborescence (a). Pour réduire la dimension du problème initial, la première étape consiste à aplanir chaque arborescence $\mathcal{T}_G$ et $\mathcal{T}_E$ de sorte qu'il ne reste qu'une seule structure de groupes emboités par facteur. Chaque groupe de variables défini au niveau le plus profond est ainsi inclus dans d'autres groupes plus englobants, comme le montre la figure 4.2b.

Pondération des groupes de l'arborescence (b). Pour conserver la mémoire de l'arborescence, une mesure supplémentaire peut être incluse afin de quantifier la perte d'information entre deux niveaux successifs. Plus précisément, pour un intervalle $[h^-, h^+]$ de niveaux explorés, avec $1 \leq h^- < h^+ \leq H-1$, nous définissons s_h comme l'écart entre les hauteurs h et $h-1$. En utilisant une méthodologie similaire à celle du *multi-layer group-lasso* [IG29, IG30], nous définissons cette quantité par $\rho_h = 1/\sqrt{s_h}$. Ce processus est illustré à travers le passage de les figures 4.2a à 4.2b.

Définition de descripteurs compressés (c). Pour résumer chaque groupe pondéré de variables (voir figure 4.2c), la moyenne, la médiane ou d'autres quantiles peuvent être utilisés, tout comme des représentations plus sophistiquées basées sur une décomposition en valeurs propres, telles que le premier facteur d'une analyse en composantes principales par exemple. De plus, la dimension $\widetilde{D}$ d'une matrice compressée $\widetilde{\mathbf{X}}$ peut être limitée aux descripteurs les plus significatifs en tenant compte de la relation d'ordre sur le vecteur des pondérations $\boldsymbol{\rho}$, avec $\widetilde{D} < \text{card}(\boldsymbol{\rho})$.

Remarque *Particularités des données omiques* — Nous évoquerons dans la section 4.3 des compressions plus spécifiques aux marqueurs génétiques représentés par des SNP ainsi que des éléments permettant de calibrer l'intervalle $[h^-, h^+]$. ◇

4.2.4. Aperçu des résultats

Nous avons testé notre algorithme [CS2] dans le cadre de simulations et l'avons comparé à d'autres approches régularisées, notamment celle de Grimonprez et al. [IG30] qui utilise les descripteurs non compressés issus de la structure arborescence (cf. figure 4.2b) ainsi que celle de Bien, Taylor et Tibshirani [IG8] qui intègre des contraintes respectant les hypothèses de dépendance de variables organisées dans une structure arborescente. Nous avons observé différents aspects :

1. La capacité à identifier les interactions dans différentes configurations fonction de la taille de l'échantillon, du bruit, et du nombre d'interactions,
2. L'impact du schéma de pondération sur les performances, en comparant deux versions de notre approche : une avec les pondérations ρ et une sans,
3. L'impact du schéma de compression sur les performances,
4. Les temps de calcul nécessaires pour atteindre la convergence à mesure que la dimension d'une matrice augmente.

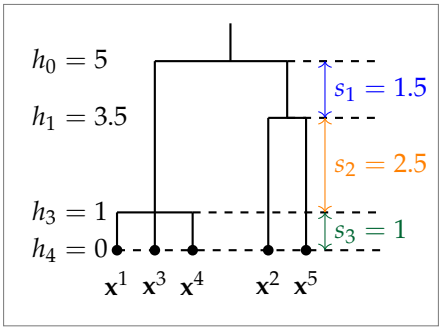
Pour les trois premiers aspects, les résultats en terme de rappel (interactions pertinentes sélectionnées) sont assez convaincants notamment vis-à-vis des autres méthodes comparées ainsi que de la version non pondérée. Les résultats en terme de précision sont par contre plus contrastés. En effet, les descripteurs sélectionnés correspondent à des regroupements situés trop haut dans l'arborescence, impliquant ainsi des interactions faussement positives englobant spatialement les vraies interactions. Un point d'amélioration pourrait être d'introduire une étape supervisée afin de mieux cibler l'échelle spatiale, comme dans l'approche présentée en section 4.3.

Pour l'aspect calculatoire, la compression des descripteurs et la mise en cascade de différentes étapes rend notre approche particulièrement performante dans un contexte de très grande dimension. En résumé, notre approche est pertinente et efficace pour identifier des régions impliquées dans des interactions mais nécessite une étude plus fine des interactions sélectionnées pour éliminer les faux positifs.

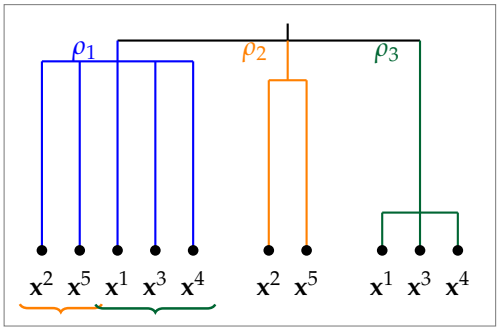
Nous avons appliqué notre approche pour détecter des interactions entre les marqueurs génomiques de la plante *Medicago truncatula* et les marqueurs métagénomiques de la communauté bactérienne issue de sa rhizosphère [2]. Certaines interactions ont fait l'objet de recoupements au regard des connaissances actuelles mais plusieurs d'entre elles nécessitent des analyses plus poussées pour être confirmées.

Notons enfin que l'algorithme [CS2] s'applique à l'interaction de jeux de données représentés par plus de deux facteurs, quelque soient leurs types, pourvu qu'une structure arborescente puisse être définie dessus.

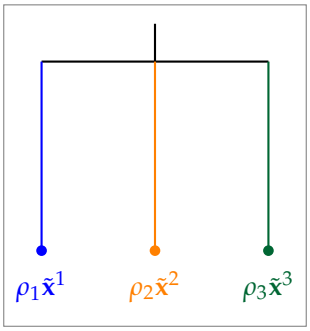
2. La partie du sol proche des racines des plantes.



(a) $\mathcal{T}$: arborescence originale.



(b) $\widetilde{\mathcal{T}}$: arborescence étendue.



(c) $\widetilde{\mathbf{X}}$: représentation compressée.

FIGURE 4.2. – Définition de descripteurs compressés. (a) Arborescence $\mathcal{T}$ originale sur un exemple de cinq variables. (b) Expansion de l'arborescence originale sur l'ensemble des groupes avec les pondérations ρ fonction des sauts hiérarchiques s. Le groupe bleu agrège les variables contenues dans les groupes orange et vert. (c) Construction de descripteurs compressés selon les groupes pondérés définis en (b).

4.3. Interaction $G \times G$

Ces travaux visent à détecter des *régions génomiques* impliquées dans des maladies génétiques dans des études d'association pangénomique. Ils ont été proposés dans [CM1] avec des modèles additifs généralisés (*GAM* pour *Generalized Additive Models* [IG45]) sur des descripteurs issus d'un regroupement spatial.

4.3.1. D'un modèle additif à un modèle d'interaction

Le cadre flexible des *GAM* permet de repositionner notre approche sur l'étude d'interaction $G \times G$. On peut trouver un aperçu des liens entre les propriétés de ces deux types de modèles dans [IG61] ou encore [IG21]. Pour conserver l'unité de la présentation, nous utiliserons le modèle additif *simple* (fonction de liaison identité).

Dans un modèle additif, on peut décrire la relation entre y_i et $\mathbf{x}_i$ par

$$y_i = \sum_m f_m(x_{im}) + \epsilon_i\,, \quad \text{avec } f_m(x_{im}) = \sum_t \beta_{mt} B_{mt}(x_{im})\,,$$

où les éléments $\{f_m\}$ représentent les composantes des variables sur une base de fonctions (polynomiales ou splines par exemple), B_{mt} est l'élément t de la base B associé à la variable m et β_{mt} le coefficient correspondant à estimer.

On peut choisir B comme étant une base supportée par les données. En particulier, pour $1 \leq t = m' \leq M$ et lorsque $B_{mm'}(x_{im}) = x_{im} \cdot x_{im'}$ et $B_{mm}(x_{im}) = x_{im}$, le modèle

$$y_i = \sum_m \sum_{m'} \beta_{mm'} B_{mm'}(x_{im}) + \epsilon_i\,, \tag{4.4}$$

devient strictement équivalent au modèle d'interaction $G \times G$ (4.2), avec $\beta_{mm} = \omega_m$ le coefficient associé à l'effet simple de la composante m et $\beta_{mm'} = \delta_{mm'}$, celui associé au terme d'interaction simple entres les composantes m et m'.

Remarque *Composantes génétiques* — On peut affiner la définition de B en tenant compte de caractéristiques structurelles, notamment en regroupant des SNP selon les régions chromosomiques auxquels ils appartiennent. Nous verrons dans la suite comment définir des composantes tenant compte de la spatialité des marqueurs génétiques. ◇

4.3.2. Principe général

Dans ces travaux reposant également sur une approche en deux étapes, la procédure de criblage est embarquée dans la construction de nouveaux descripteurs représentant des SNP agrégés selon des régions chromosomiques. Cette agrégation est fonction d'une structure arborescente $\mathcal{T}$ établie de la même façon que $\mathcal{T}_G$ pour notre approche $G \times E$ (cf. section 4.2.2). La différence réside dans le fait que, plutôt

que d'exploiter la hiérarchie sur l'ensemble des niveaux $[h^-, h^+]$, on cherche ici à établir le niveau de coupe optimal. Les étapes de cette approche peuvent être résumées ci-dessous, avec $R_S(\boldsymbol{\omega}, \boldsymbol{\Delta})$ défini selon (4.2).

(EDC) Définition et criblage de descripteurs compressés

À partir d'une arborescence $\mathcal{T} = \{\mathcal{T}^h\}_{h=0}^{H}$ $\qquad \forall h \mid h^- \leq h \leq h^+$

(a) Compression des SNP par groupe $\qquad \widetilde{\mathbf{S}}_h = \left(\widetilde{\mathbf{X}}^h, \mathbf{y}\right)$

$$\mathbf{X} \xmapsto{\mathcal{T}^h = \left\{\mathcal{G}_1^h, \dots, \mathcal{G}_{G_h}^h\right\}} \widetilde{\mathbf{X}}^h, \text{ avec } \dim(\widetilde{\mathbf{X}}^h) = G_h$$

(b) Estimation des paramètres d'un modèle d'interaction

Criblage des SNP compressés [3] $\qquad \operatorname{supp}\left(\boldsymbol{\omega}^h\right) \leftrightsquigarrow \operatorname{supp}\left(\boldsymbol{\Delta}^h\right)$

$$(\boldsymbol{\omega}^h, \boldsymbol{\Delta}^h) \;=\; \operatorname{argmin}_{\boldsymbol{\omega}, \boldsymbol{\Delta}} \; R_{\widetilde{\mathbf{S}}_h}(\boldsymbol{\omega}, \boldsymbol{\Delta}) \;+\; \lambda_h \, P(\boldsymbol{\omega}, \boldsymbol{\Delta})$$

Calcul d'une erreur de prédiction [4] $\qquad E_h\left(\widetilde{\mathbf{S}}_h, \boldsymbol{\omega}^h, \boldsymbol{\Delta}^h\right)$

Estimation du niveau de coupe optimal dans $\mathcal{T}$ $\qquad h^\star = \operatorname{argmin}_h E_h$

(ES) Sélection des termes d'interaction $\qquad \forall \left(m \in \operatorname{supp}\left(\boldsymbol{\omega}^{h^\star}\right), m' \in \operatorname{supp}\left(\boldsymbol{\omega}^{h^\star}\right)\right)$

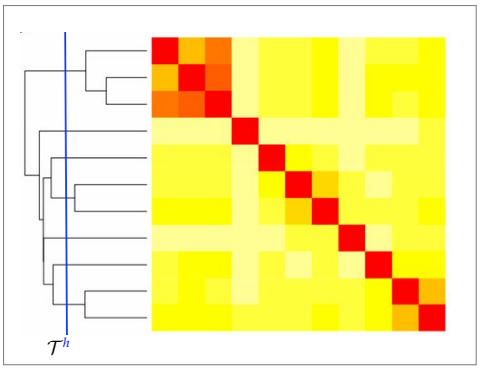
$$(\omega_m^\star, \delta_{mm'}^\star) \;=\; \operatorname{argmin}_{\omega_m, \delta_{mm'}} \; R_{\widetilde{\mathbf{S}}_{h^\star}}(\omega_m, \delta_{mm'})$$
$$\text{test}\left(\delta_{mm'}^\star\right)$$

Adaptation de la pénalité aux hypothèses de dépendance. Notre approche n'étant initialement pas définie dans une optique d'étude d'interaction classique, l'étape de criblage doit être adaptée pour respecter les hypothèses de dépendance vis à vis des effets simples. Pour cela, on peut se tourner vers des pénalités dédiées introduisant des contraintes satisfaisant ces hypothèses, comme par exemple dans les travaux de Bien, Taylor et Tibshirani [IG8]. Bien que le recours à ce type de pénalités ne consiste qu'en une adaptation marginale dans notre algorithme (changement de pénalité), il nécessite toutefois un temps de calcul important (cf. [RI2]).

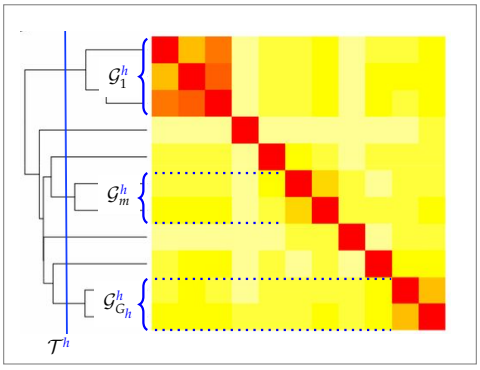
Adaptation de la procédure aux hypothèses de dépendance. Une séquence plus efficace consisterait, en s'inspirant de notre approche présentée pour l'interaction $G \times E$, à utiliser une pénalité parcimonieuse sur le modèle à effets simples puis à calculer l'erreur de prédiction sur le modèle d'interaction. La consistance d'une telle procédure dans l'estimation du niveau de coupe optimal dans l'arborescence mériterait cependant d'être étudiée plus précisément.

3. Avec $R_{\widetilde{\mathbf{S}}_h}$ | (4.2) et P garantissant le respect de l'hypothèse de dépendance choisie.

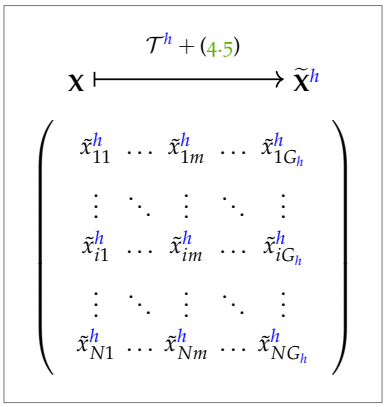
4. La mesure E_h est généralement liée à $R_{\widetilde{\mathbf{S}}_h}$ lorsque $y \in \mathbb{R}$ ou est fonction des taux de vrais et faux positifs et négatifs lorsque $y \in \{\pm 1\}$.



(a1) $\mathcal{T}$: arborescence originale.



(a2) $\mathcal{T}^h$: groupes au niveau h.



(a3) Matrice compressée.

FIGURE 4.3. – Vue schématique de l'étape (a) permettant d'aboutir à la matrice compressée $\widetilde{\mathbf{X}}^h$: (a1) Définition de l'arborescence $\mathcal{T}$; (a2) Définition de $\mathcal{T}^h = \left\{\mathcal{G}_1^h, \ldots, \mathcal{G}_{G_h}^h\right\}$, l'ensemble des groupes à la hauteur h ; (a3) Définition de la matrice $\widetilde{\mathbf{X}}^h$ issue des groupes de $\mathcal{T}^h$.

4.3.3. Définition et criblage de descripteurs compressés (EDC)

Cette étape a pour objectif de réduire la dimension originale à travers la recherche d'un regroupement spatial optimal en explorant les différents niveau d'une structure arborescente sur des marqueurs génétiques représentés par des SNP. On se concentrera ici sur la description des deux aspects suivants :

(a) Compression des marqueurs génétiques selon une structure de groupe $\mathcal{T}^h$.

(b) Procédure d'estimation du niveau de coupe optimal $h^\star$ dans l'arborescence $\mathcal{T}$.

Compression des SNP par groupe (a). Comme évoqué dans la section 4.2, et à condition de disposer de données brutes sur le génotype et éventuellement sur le phénotype, d'autres compressions peuvent être envisagées pour représenter des groupes de variables préalablement identifiés, allant de résumés classiques tels qu'un percentile pertinent pour le problème considéré au(x) premier(s) facteur(s) d'une analyse en composante principale [IG64].

D'autres méthodes plus sophistiquées font appel aux coefficients des fonctions d'une analyse de Fourier [IG68], aux pondérations liées au déséquilibre de liaison associé aux SNP d'une région spécifique [IG41] ou encore à des résumés issus de regroupements fondés sur des similarités génétiques [IG13].

Dans notre approche, pour une hauteur $h \in [h^-, h^+]$ de l'arborescence $\mathcal{T}$ (construite de façon à tenir compte du déséquilibre de liaison), nous associons à chaque groupe de $\mathcal{T}^h$ une nouvelle composante. Nous définissons ces $G_h \ll D$ nouvelles composantes comme le nombre d'allèles mineurs associés à chaque composante. Plus précisément, pour chaque groupe $1 \leq m \leq G_h$, nous définissons, pour l'individu i, la nouvelle composante m par :

$$\tilde{x}_{im}^h = \sum_{j \in \mathcal{G}_g^h} x_{ij} \,. \tag{4.5}$$

Cette représentation (4.5) est similaire à celle utilisée dans les *burden tests* [IG2, IG39], qui visent à détecter l'effet de groupes de variants rares sur un phénotype. Pour que les valeurs dans les différentes composantes soient comparables, nous éliminons l'effet de la taille du groupe en centrant et réduisant la matrice $\widetilde{\mathbf{X}}^h$.

Le processus général de la compression des SNP est illustré sur la figure 4.3.

Remarque *Calibration de l'intervalle de recherche* — Pour accélérer le processus de recherche de $h^\star$, on peut restreindre les niveaux explorés à un intervalle $[h^-, h^+]$ que l'on pourra circonscrire en fonction de connaissances *a priori*, par exemple en regardant si la taille des groupes définis à certains niveaux de l'arborescence correspond à des tailles de régions génomiques pertinentes pour le problème considéré. ◇

Procédure supervisée d'estimation du niveau de coupe (b). L'estimation de $h^\star$ dans une hiérarchie consiste à rechercher le nombre optimal de groupes $G_{h^\star}$ à

sélectionner. Cette étape fondamentale conditionne la pertinence de l'analyse. Du point de vue applicatif, bien que le génome soit structuré par blocs d'haplotypes [5], avec peu (voire pas) de recombinaison à l'intérieur, il n'est pas facile de déterminer comment ces blocs sont répartis dans le génome pour un ensemble donné de SNP. Du point de vue mathématique, plusieurs méthodes permettent de déterminer le nombre optimal de groupes dans une classification hiérarchique (voir par exemple [IG36, IG59]).

De notre côté, nous avons proposé d'utiliser une approche supervisée pour atteindre cet objectif. En effet, dans la mesure où une finalité des GWAS consiste à évaluer la probabilité d'un phénotype à partir de (l'interaction de) marqueurs génétiques, nous pouvons tirer partie de cette information. Pour trouver cet optimum, nous avons scindé notre jeu de données en deux sous-ensembles étiquetés. Un ensemble $\mathcal{S}_1 = (\mathbf{X}_{S_1}, \mathbf{y}_{S_1})$ permet, par une procédure de validation, de décider du niveau de coupe. Une fois $h^\star$ déterminé, on peut définir, à partir d'un ensemble $\mathcal{S}_2 = (\mathbf{X}_{S_2}, \mathbf{y}_{S_2})$, de $\mathcal{T}^{h^\star}$ et de la fonction d'agrégation (4.5), la matrice compressée optimale $\widetilde{\mathbf{X}}_{S_2}$ qui sera utilisée dans l'étape de sélection (ES).

4.3.4. Aperçu des résultats

La méthode développée dans [CM1] a été proposée dans le cadre d'étude d'association pangénomique. Dans ce contexte, la prise en compte de la structure de déséquilibre de liaison pour construire des descripteurs de SNP fortement corrélés est une alternative intéressante à l'analyse de marqueur standard dans la mesure où elle a pour effet de réduire significativement la dimension du problème initial. De plus, elle a permis une amélioration notable en termes de rappel et de précision notamment lorsqu'elle est couplée à un modèle de régression additif généralisé, ou lorsqu'elle est combinée à des tests d'association spécifiques (cf. [RI3]).

La transposition effective de cette approche à l'interaction $G \times G$ nécessiterait cependant des ajustements supplémentaires pour intégrer une pénalité adaptée, comme celle proposée par [IG8], et des simulations plus poussées pour étudier son comportement en terme de sélection d'interaction. Il serait également intéressant d'observer si l'approche supervisée pour choisir le niveau de coupe optimal permet de diminuer le nombre de faux positifs par rapport à l'approche globale de [RI2] consistant à explorer l'ensemble des niveaux de l'arborescence.

5. Un haplotype est un groupement physique de variants qui ont tendance à être hérités ensemble.

Perspectives

Higher ground

Vers des schémas d'intégration tardive

Mes travaux sur l'intégration de données issues de plusieurs sources d'information ont été abordés avec des approches globales autour de schémas d'***intégration intermédiaire*** de données. Dans ce paradigme, les données sont transformées, structurées, voire les deux, dans un modèle mathématique commun où les paramètres associés aux différentes sources sont optimisés conjointement, comme illustré au chapitre 3 sur la figure 3.1b.

Un objectif important des méthodes présentées dans ce document consiste à identifier les sources et les descripteurs pertinents. Ce type d'approches nécessite toutefois des analyses supplémentaires pour confirmer statistiquement la pertinence des éléments sélectionnés. En outre, peu intègrent des mécanismes pour facilement étudier les relations explicites entre les différents éléments, une des principales raisons étant probablement la combinatoire associée à l'estimation des paramètres traduisant une interaction dans un modèle global.

Pour étudier de telles interactions, nous avons privilégié des méthodes mettant en cascade une étape de criblage de descripteurs structurés avec des modèles de régression, puis une étape de sélection avec des tests statistiques d'association. Une fois les tests d'association effectués, on peut observer *a posteriori* les sources présentant des complémentarités ou des redondances, voire définir des métriques traduisant ces aspects. Cependant, on pourrait dans certains problèmes vouloir privilégier l'une de ces caractéristiques pour guider cette recherche d'interaction.

Observations

En présence de sources d'information multiples, en particulier lorsqu'elles sont hétérogènes, il apparait finalement pertinent de :

1. Tirer partie des spécificités de chaque type de données avec des modèles ou des algorithmes dédiés,
2. Pourvoir recombiner l'ensemble des résultats dans une perspective d'interprétation globale du problème.

En effet, chaque source de données est généralement liée à un format caractéristique. L'utilisation de modèles ou algorithmes spécifiques à une source permet d'optimiser la qualité des estimations et des prédictions en tenant compte des particularités de données, à travers une fonction de redescription, une fonction de coût ou encore une pénalité, en réduisant les variabilités induites par l'hétérogénéité des sources de différentes natures. De plus, il est souvent plus efficace de considérer des processus flexibles, avec des traitements dédiés en amont (comme les étapes de criblage vues dans le chapitre 4), notamment pour les problèmes de grande dimension.

Le paradigme d'***intégration tardive*** synthétisé sur la figure 3.1c, où les prédictions de S modèles indépendants sont combinées dans un modèle global, semble finalement bien adapté pour à la fois tenir compte efficacement des spécificités des données et pour interpréter de façon globale le phénomène étudié. Mon projet de recherche sur l'intégration et l'interaction de données issues de sources d'information hétérogènes se situe donc désormais autour de ce paradigme.

Positionnement

En classification non supervisée, on parle de ***consensus clustering*** ou d'***ensemble clustering*** lorsque qu'un modèle est appliqué plusieurs fois à un même jeu de données, avec différents paramètres ou différentes initialisations par exemple, puis que l'on cherche à agréger les résultats correspondants [IG25].

Le terme de ***meta clustering*** désigne un repositionnement de cette problématique autour de l'intégration tardive de sources multiples. Nous nous intéresserons ici à quelques familles d'approches utilisant une représentation construite sur la proximité entre observations comme entrée.

Représentation des données. En particulier, que le cadre d'apprentissage soit supervisé ou non, la représentation de l'appartenance des observations aux classes peut s'effectuer avec une matrice d'adjacence :

$$(\mathbf{A})_{ii'} = a_{ii'} = \begin{cases} 1, \text{ si les individus } i \text{ et } i' \text{ appartiennent à la même classe,} \\ 0, \text{ sinon.} \end{cases}$$

Remarque *Permutation aléatoire des étiquettes (label shift)* — Une classe étiquetée k par un algorithme de classification non supervisé peut être étiquetée $k' \neq k$ avec ce même algorithme, lorsque les conditions d'initialisation sont différentes par exemple. La représentation des classes par matrice d'adjacence possède l'atout majeur d'être invariante relativement à ces permutations, en particulier dans une optique de *consensus* ou de *meta clustering* où plusieurs partitions sont disponibles. ◇

Remarque *Matrices de voisinage* — Plus généralement, les approches que nous allons décrire peuvent utiliser comme entrée des graphes de voisinage ou les matrices associées, comme illustré sur la figure 2.3 du chapitre 2. ◇

Quelques familles d'approches. À partir de ce type de représentation, différentes approches permettent d'aborder les problématiques d'intégration tardive de sources d'information multiples.

Les ***méthodes à noyaux*** ont largement été explorées dans une perspective d'intégration tardive autour du *Multiple Kernel Clustering* (voir [IG67, IG66] par exemple). Il s'agit dans ce cas de déterminer une matrice représentant un consensus entre les partitions, le plus souvent avec des techniques d'optimisation qui impliquent des décompositions en valeurs singulières. La pondération des différentes sources d'information est quasiment toujours modélisée dans le problème d'optimisation, contrairement à leur complémentarité ou à leur façon d'interagir.

Les ***approches tensorielles*** se fondent sur des décompositions algébriques multilinéaires (décomposition de Tucker ou décomposition de rang du tenseur) pour identifier des relations entre plusieurs matrices de voisinage. On pourra notamment citer les travaux de Jing et al. [IG34] qui intègrent l'estimation d'une partition des observations par source d'information mais également une partition des sources.

Enfin, les ***modèles à blocs stochastiques ou latents*** à partir de *réseaux multicouches* [cf. IG17, chapitre 1] permettent également d'envisager des approches d'intégration tardive. En particulier, Boutalbi, Labiod et Nadif [IG11] définissent un modèle de Poisson à blocs latents, par le biais d'une représentation tensorielle, permettant de partitionner l'espace à travers les dimensions du tenseur. Stanley et al. [IG57] ou encore Rebafka [IG52] étendent les modèles à blocs stochastiques en considérant que les différents réseaux sont issus d'un modèle de mélange.

Approches envisagées

Les approches tardives fondées sur des modèles permettent de spécifier explicitement la redondance ou la complémentarité entre sources d'information, guidant ainsi le *meta clustering* simultanément effectué à partir des partitions disponibles.

Cependant, dans la plupart des travaux mentionnés ci-dessus, soit la partition des observations est globale mais l'interaction entre les sources des données n'est pas modélisée, soit elle l'est mais la partition des observations est relative à chaque source, et la partition globale doit être déduite *a posteriori*.

Modèles à blocs stochastiques. Les travaux de thèse de Kylliann De Santiago [6] se placent dans ce contexte. Nous travaillons sur ces modèles à partir de sources décrites par S matrices d'adjacence $\mathbf{A}^s$ de terme général $a^s_{ii'}$, donnant lieu à une représentation tensorielle $\mathbf{A} \in \{0,1\}^{N\times N\times S}$.

Dans le modèle proposé, la partition des observations recherchée est partagée à travers les sources : on recherche $\mathbf{Z} \in \{0,1\}^{N\times K}$, une partition commune des observations, plutôt que S partitions. La recherche de cette partition sur les observations

6. Kylliann a débuté en 2022 une thèse CIFRE avec l'entreprise Sensorion, sous la direction de Christophe Ambroise et en co-encadrée avec Guillaume Andéol (IRBA) et moi-même.

est conjointe à celle d'une partition $\mathbf{W} \in \{0,1\}^{S \times Q}$ des S sources d'information en Q composantes. Sous l'hypothèse d'indépendance des observations et des sources d'information, $\mathbf{Z}$ et $\mathbf{W}$ suivent respectivement des lois multinomiales de paramètres $\{\pi_k\}$ avec $1 \leq k \leq K$ et $\{\rho_q\}$, avec $1 \leq q \leq Q$.

La probabilité que les observations i et i' soient globalement liées dans $\mathbf{A}$, conditionnellement aux paramètres $\Theta = \left\{\{\pi_k\}, \{\rho_q\}\right\}$ du modèle, s'écrit

$$\mathbb{P}\left(\mathbf{A} \mid \mathbf{Z}, \mathbf{W}, \Theta\right) = \prod_{\substack{i=1 \\ i'<i}}^{N} \prod_{\substack{k=1 \\ k'=1}}^{K} \prod_{s=1}^{S} \prod_{q=1}^{Q} \left(\mathbb{P}(a_{ii'}^{s} \mid z_{ik} = 1, z_{i'k'} = 1, w_{sq} = 1, \Theta\right)^{z_{ik}\, z_{i'k'}\, w_{sq}}.$$

Ce type de problème, une fois la log-vraisemblance calculée, est classiquement résolu avec un algorithme EM. Nous avons privilégié une approche Bayesienne variationnelle afin de bénéficier d'un critère de sélection de modèle pour choisir K et Q. Nos premières simulations donnent des résultats encourageants. Il reste néanmoins de nombreux points autour de ces travaux à considérer à court ou moyen termes.

(a) D'un point de vue théorique, nous souhaitons étudier l'identifiabilité du modèle ainsi que la convergence de l'algorithme. Le point de départ concerne les travaux de Celisse, Daudin et Pierre [IG15] étendus aux réseaux mulicouches par Barbillon et al. [IG6] ou Chabert-Liddell et al. [IG18] pour des partitions $\mathbf{Z}^s$ spécifiques à chaque source. La particularité de notre approche réside dans le fait que la partition $\mathbf{Z}$ est traversante sur l'ensemble des couches.

(b) D'autre part, l'application de ce modèle est prévue sur une cohorte de patients collectée par l'IRBA. L'objectif consiste à stratifier[7] des patients atteints de traumatismes sonores aigus, pour lesquels différentes mesures audiologiques sont collectées, en parallèle de données génomiques et protéomiques. Les données étant confidentielles, nous souhaitons trouver une alternative pour consolider la présentation de notre modèle dans une optique de publication.

(c) Enfin, la levée de l'hypothèse d'indépendance entre les sources fait également partie des perspectives à moyen terme dans le cadre applicatif évoqué ci-dessus, où différents types de mesures audiologiques sont considérées d'une part, et des données de nature omiques d'autre part. D'un point de vue méthodologique, une telle modélisation pourrait permettre de concilier les objectifs du *consensus clustering* et du *meta clustering*, en considérant pour une source les partitions obtenues avec différentes paramétrisations d'un algorithme.

Mélange d'experts. À plus long terme, je souhaiterais poursuivre des travaux autour des modèles de mélange d'experts, sous l'angle de l'intégration tardive cette fois. Également inspiré par les problématiques autour du projet PATRIOT, l'objectif consisterait à déterminer, à partir des S matrices d'adjacence $\mathbf{A}^s \in \{0,1\}^{N \times N}$ issues des sources audiologiques et omiques, la probabilité y_i de récupération du patient i.

7. La stratification consiste ici à déterminer un sous-typage (ou une partition) des observations par le biais d'une approche non supervisée.

En notant $\mathbf{A}_i \in \{0,1\}^{N\times S}$, la sous-matrice d'adjacence de taille entre l'observation i et les autres observations sur l'ensemble des sources, on peut envisager de retrouver un sous-typage global en K catégories pour expliquer y_i en transformant le modèle général (3.5) de sorte que

$$\mathbb{P}(y_i \,|\, \mathbf{A}_i) = \sum_{s,k} \eta_k(\mathbf{a}_i^s)\, f_k(y_i|\Theta_k(\mathbf{a}_i^s))\,.$$

Pour spécifier plus précisément la modélisation, il serait intéressant d'intégrer des aspects supplémentaires par rapport à ceux évoqués précédemment.

(a) Les différents types de contributions permettant d'expliquer le phénomème observé, par exemple la contribution des sources (ou de leurs composantes) à la partition latente globale $\mathbf{Z}$ ou aux étiquettes associées à $\mathbf{y}$, pourraient être traduits via les paramètres, les poids des experts ou une combinaison des deux.

(b) La variabilité sous-jacente aux différents centres dans lesquels ont été collectées les informations audiologiques (biais de mesure) pourrait être prise en compte dans la densité f_k, au travers de modèles à effets mixtes par exemple.

Vers une recherche (mieux) finalisée

Depuis que j'ai intégré une équipe de statistiques appliquées aux problèmes de santé et de biologie, la dimension applicative de mes travaux a pris plus de place. Je souhaite continuer à travailler dans ce contexte de façon plus complète.

L'interdisciplinarité en conditions réelles

Si les questions associées à des problématiques concrètes sont des inspirations particulièrement motivantes pour la définition de nouveaux modèles ou algorithmes, il est essentiel de trouver un équilibre entre les aspects méthodologiques et les aspects expérimentaux dans ce type de collaborations : les mathématiciens ou les informaticiens seront attentifs à l'originalité d'un modèle ou d'un algorithme tandis les biologistes ou les cliniciens seront attachés aux réponses qu'apportent les méthodes. Cela nécessite sans surprise un dialogue important en amont entre les différents acteurs, pour la compréhension globale du problème évidemment mais aussi pour une prise en main des données aussi rapide et efficace que possible.

Il faut encore un investissement significatif pour s'approprier ces données et voir quelles sont les atouts et les limites des méthodes de référence existantes. Cette partie préliminaire, bien que chronophage, est essentielle pour la définition de méthodes plus originales. Elle est pourtant souvent mise de côté car moins facile à valoriser, le temps (et les financements) étant compté(s). Par ailleurs, si la mise à disposition des codes associés aux méthodes développées semble être aujourd'hui entrée dans les mœurs dans la communauté de l'apprentissage machine, l'accès aux

données réelles sur lesquelles les méthodes sont testées reste très sensible et entrave la reproductibilité des résultats. Cela arrive fréquemment lorsqu'il s'agit de données liées par exemple à l'identification de patients ou de données utilisées en vue de la valorisation industrielle d'un médicament ou d'un dispositif médical.

Plus généralement, il est complexe sur des collaborations inscrites dans des projets financés, dépassant rarement la durée d'une thèse, de pouvoir couvrir l'ensemble des aspects et d'intégrer les allers-retours nécessaires entre méthodologie et application pouvant mener à des retombées concrètes en terme d'explication du phénomène étudié. L'expérimentation de ces différentes facettes a fini de me convaincre de la nécessité de la pratique d'une recherche sur un temps plus long [HG47]. Cela demande d'inscrire les collaborations avec biologistes ou médecins dans la durée, en gardant contact après la fin des projets et la publication des premiers travaux, et peut être aussi d'être plus parcimonieux sur d'autres opportunités de collaborations.

Comparaisons de méthodes et protocoles d'évaluation

Je souhaiterais à l'avenir mieux intégrer les aspects autour des protocoles d'évaluation des méthodes de référence sur des données réalistes.

Concernant la mise à disposition des données évoquée précédemment, si les informaticiens ou statisticiens n'ont concrètement que peu (voire pas) de prise dessus, il n'est plus souhaitable de travailler en dehors d'une optique de reproductibilité. Une façon de traiter cela, en partie du moins, serait de travailler sur la simulation de données aussi réalistes que possible à partir d'échantillons réels perturbés dont l'incertitude est modélisée, ou du moins contrôlée, et de fournir ces données conjointement aux codes permettant de tester les approches de référence.

En particulier, une fois que l'on dispose de données exploitables, ce type de travaux parait adapté pour donner à des étudiants de niveau L3 ou M1 un aperçu du monde de la recherche et de certaines démarches scientifiques. Notamment, les comparaisons d'algorithmes nécessitent des protocoles rigoureux pour que l'évaluation soit équitable [IG32]. C'est aussi une occasion de prendre en compte les performances des algorithmes relativement à leur impact énergétique [IG31]. En fonction du contexte, on peut se demander, de façon caricaturale ici, s'il est nécessaire d'utiliser une méthode qui dépense dix fois plus d'énergie qu'une autre pour améliorer un taux de classification de quelques (dixièmes de) points.

De manière globale, on peut se demander dans quelles conditions il est pertinent de définir des modèles mathématiques sophistiqués mais souvent coûteux à optimiser lorsqu'une mise en cascade efficace de méthodes de référence permet de répondre globalement à la problématique. En complément de résultats empiriques bien menés, la connaissance de garanties théoriques sur (l'association de) plusieurs méthodes de référence est un levier supplémentaire pour orienter son choix.

Bibliographie

Hinsight Ground

[HG1] A. Argyriou, R. Foygel et N. Srebro. Sparse Prediction with the k-Support Norm. *Advances in Neural Information Processing Systems*, tome 25. Curran Associates, Inc., 2012.
(Cf. page 25)

[HG2] F. Bach, R. Jenatton, J. Mairal et G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1) :1-106, 2012.
(Cf. pages 18, 19, 24)

[HG3] F. Bach, R. Jenatton, J. Mairal et G. Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4) :450-468, 2012.
(Cf. pages 18, 19)

[HG4] S. Bakin. *Adaptive regression and model selection in data mining problems*. Thèse de doctorat, The Australian National University, 1999.
(Cf. page 20)

[HG5] A. Beck et M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1) :183-202, 2009.
(Cf. page 19)

[HG6] H. D. Bondell et B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1) :115-123, 2008.
(Cf. page 25)

[HG7] B. E. Boser, I. M. Guyon et V. N. Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144-152, 1992.
(Cf. page iii)

[HG8] L. Breiman. Statistical modeling : The two cultures. *Statistical science*, 16(3) :199-231, 2001.
(Cf. page iii)

[HG9] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh et E. R. Dougherty. Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40(3) :807-824, 2007.
(Cf. page 10)

[HG10] G. Celeux, S. Frühwirth-Schnatter et C. P. Robert. Model selection for mixture models – perspectives and strategies, *Handbook of mixture analysis*, pages 117-154. Chapman et Hall/CRC, 2019.
(Cf. page 10)

[HG11] S. S. Chen, D. L. Donoho et M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1) :129-159, 2001.
(Cf. page 22)

[HG12] J. Chiquet, Y. Grandvalet et C. Charbonnier. Sparsity with sign-coherent groups of variables via the cooperative-lasso. *The Annals of Applied Statistics*, 6(2) :795-830, 2012.
(Cf. page 26)

[HG13] P. L. Combettes et J.-C. Pesquet. Proximal splitting methods in signal processing, *Fixed-point algorithms for inverse problems in science and engineering*, pages 185-212. Springer, 2011.
(Cf. page 23)

[HG14] A. Cornuéjols, L. Miclet et V. Barra. *Apprentissage artificiel : Deep learning, concepts et algorithmes*. Eyrolles, 2018.
(Cf. page 8)

[HG15] C. Cortes et V. Vapnik. Support-vector networks. *Machine learning*, 20 :273-297, 1995.
(Cf. page iii)

[HG16] C. De Mol, E. De Vito et L. Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2) :201-230, 2009.
(Cf. page 25)

[HG17] A. Dehman, C. Ambroise et P. Neuvial. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16(1) :148-161, 2015.
(Cf. page 13)

[HG18] B. Efron, T. Hastie, I. Johnstone et R. Tibshirani. Least Angle Regression. *The Annals of Statistics*, 32(2) :407-499, 2004.
(Cf. page 19)

[HG19] L. E. Frank et J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2) :109-135, 1993.
(Cf. page 22)

[HG20] G. Govaert. *Data analysis*. John Wiley & Sons, 2013.
(Cf. page 31)

[HG21] Y. Grandvalet. Least absolute shrinkage is equivalent to quadratic penalization. *International Conference on Artificial Neural Networks*, 1998.
(Cf. page 33)

[HG22] Y. Grandvalet et S. Canu. Outcomes of the Equivalence of Adaptive Ridge with Least Absolute Shrinkage. *Advances in Neural Information Processing Systems*, tome 11, 1998.
(Cf. page 20)

[HG23] N. Hirtt. L'approche par compétences : une mystification pédagogique. *L'école démocratique*, 39(1) :1-34, 2009.
(Cf. page ii)

[HG24] A. E. Hoerl et R. W. Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55-67, 1970.
(Cf. page 22)

[HG25] J. Huang, S. Ma, H. Li et C.-H. Zhang. The sparse Laplacian shrinkage estimator for high-dimensional regression. *Annals of Statistics*, 39(4) :2021, 2011.
(Cf. page 25)

[HG26] L. Hubert et P. Arabie. Comparing partitions. *Journal of Classification*, 2(1) :193-218, 1985.
(Cf. page 10)

[HG27] L. Jacob, G. Obozinski et J.-P. Vert. Group lasso with overlap and graph lasso. *Proceedings of the 26th annual International Conference on Machine Learning*, 2009.
(Cf. pages 22, 25)

[HG28] R. Jenatton, J.-Y. Audibert et F. Bach. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12 :2777-2824, 2011.
(Cf. page 22)

[HG29] R. Jenatton, J. Mairal, G. Obozinski et F. Bach. Proximal methods for hierarchical sparse coding. *The Journal of Machine Learning Research*, 12 :2297-2334, 2011.
(Cf. pages 16, 23)

[HG30] M. Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3) :303-324, 2009.
(Cf. page 20)

[HG31] H. Kuroda et D. Kitahara. Block-sparse recovery with optimal block partition. *IEEE Transactions on Signal Processing*, 70 :1506-1520, 2022.
(Cf. page 14)

[HG32] C. Li et H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9) :1175-1182, 2008.
(Cf. page 25)

[HG33] A. Lorbert, D. Eis, V. Kostina, D. Blei et P. Ramadge. Exploiting covariate similarity in sparse regression via the pairwise elastic net. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop et Conference Proceedings, 2010.
(Cf. page 25)

[HG34] E. Ndiaye, O. Fercoq, A. Gramfort et J. Salmon. GAP Safe Screening Rules for Sparse-Group Lasso. *Advances in Neural Information Processing Systems*, tome 29, 2016.
(Cf. page 25)

[HG35] M. R. Osborne, B. Presnell et B. A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3) :389-403, 2000.
(Cf. page 19)

[HG36] N. Parikh, S. Boyd et al. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3) :127-239, 2014.
(Cf. pages 19, 23, 34)

[HG37] L.-B. Qiao, B.-F. Zhang, J.-S. Su et X.-C. Lu. A systematic review of structured sparse learning. *Frontiers of Information Technology & Electronic Engineering*, 18(4) :445-463, 2017.
(Cf. pages 13, 18)

[HG38] N. Simon, J. Friedman, T. Hastie et R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2) :231-245, 2013.
(Cf. page 25)

[HG39] M. Slawski, W. **zu** Castell et G. Tutz. Feature selection guided by structural information. *The Annals of Applied Statistics*, 4(2) :1056-1080, 2010.
(Cf. page 25)

[HG40] A. J. Smola et R. Kondor. Kernels and regularization on graphs, *Learning theory and kernel machines*, pages 144-158. Springer, 2003.
(Cf. page 24)

[HG41] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*, 58(1) :267-288, 1996.
(Cf. page 22)

[HG42] R. J. Tibshirani et J. Taylor. The solution path of the generalized lasso. *The annals of statistics*, 39(3) :1335-1371, 2011.
(Cf. page 24)

[HG43] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, 1995.
(Cf. pages iii, 9)

[HG44] U. Von Luxburg et al. Clustering stability : an overview. *Foundations and Trends in Machine Learning*, 2(3) :235-274, 2010.
(Cf. page 10)

[HG45] J. Wang et H. Jia. Metagenome-wide association studies : fine-mining the microbiome. *Nature Reviews Microbiology*, 14(8) :508-522, 2016.
(Cf. page 4)

[HG46] W. Y. Wang, B. J. Barratt, D. G. Clayton et J. A. Todd. Genome-wide association studies : theoretical and practical concerns. *Nature Reviews Genetics*, 6(2) :109-118, 2005.
(Cf. page 4)

[HG47] Wikipedia. Slow science. Ainsi que les références contenues.
(Cf. page 58)

[HG48] H. Xiong et Z. Li. Clustering validation measures, *Data Clustering*, pages 571-606. Chapman et Hall/CRC, 2018.
(Cf. page 10)

[HG49] Xkcd. Fields arranged by purity. https://xkcd.com/435/.
(Cf. page iii)

[HG50] Xkcd. Machine Learning. https://xkcd.com/1838/.
(Cf. page iii)

[HG51] M. Yuan et Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49-67, 2006.
(Cf. page 20)

[HG52] P. Zhao, G. Rocha et B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A) :3468-3497, 2009.
(Cf. page 23)

[HG53] H. Zou et T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : series B (statistical methodology)*, 67(2) :301-320, 2005.
(Cf. pages 14, 25)

Bibliographie

[IG1] C. Ambroise, A. Dehman, P. Neuvial, G. Rigaill et N. Vialaneix. Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. *Algorithms for Molecular Biology*, 14(1) :22, 2019.
(Cf. page 44)

[IG2] J. L. Asimit, A. G. Day-Williams, A. P. Morris et E. Zeggini. ARIEL and AMELIA : testing for an accumulation of rare variants using next-generation sequencing data. *Human Heredity*, 73(2) :84-94, 2012.
(Cf. page 51)

[IG3] F. R. Bach, G. R. Lanckriet et M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. *Proceedings of the twenty-first International Conference on Machine Learning*, 2004.
(Cf. page 32)

[IG4] N. Ball, W.-P. Teo, S. Chandra et J. Chapman. Parkinson's disease and the environment. *Frontiers in neurology* :218, 2019.
(Cf. page 38)

[IG5] T. Baltruaitis, C. Ahuja et L.-P. Morency. Multimodal machine learning : A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2) :423-443, 2018.
(Cf. page 29)

[IG6] P. Barbillon, S. Donnet, E. Lazega et A. Bar-Hen. Stochastic block models for multiplex networks : an application to a multilevel network of researchers. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* :295-314, 2017.
(Cf. page 56)

[IG7] J.-M. Bécu, Y. Grandvalet, C. Ambroise et C. Dalmasso. Beyond support in two-stage variable selection. *Statistics and Computing*, 27 :169-179, 2017.
(Cf. page 42)

[IG8] J. Bien, J. Taylor et R. Tibshirani. A Lasso for hierarchical interactions. *Annals of statistics*, 41(3) :1111, 2013.
(Cf. pages 42, 46, 49, 52)

[IG9] B. Blankertz et al. The BCI competition 2003 : progress and perspectives in detection and discrimination of EEG single trials. *IEEE Transactions on Biomedical Engineering*, 51(6) :1044-1051, 2004.
(Cf. page 35)

[IG10] M. Blaxter, J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd et E. Abebe. Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 360(1462) :1935-1943, 2005.
(Cf. page 44)

[IG11] R. Boutalbi, L. Labiod et M. Nadif. Implicit consensus clustering from multiple graphs. *Data Mining and Knowledge Discovery*, 35 :2313-2340, 2021.
(Cf. page 55)

[IG12] C. Brouard, J. Mariette, R. Flamary et N. Vialaneix. Feature selection for kernel methods in systems biology. *NAR Genomics and Bioinformatics*, 4(1) :1-17, 2022.
(Cf. page 32)

[IG13] A. Buil, A. Martinez-Perez, A. Perera-Lluna, L. Rib, P. Caminal et J. M. Soria. A new gene-based association test for genome-wide association studies. *BMC proceedings*, tome 3, S130. BioMed Central, 2009.
(Cf. page 51)

[IG14] B. J. Callahan, P. J. McMurdie et S. P. Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12) :2639, 2017.
(Cf. page 44)

[IG15] A. Celisse, J.-J. Daudin et L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model, 2012.
(Cf. page 56)

[IG16] N. Cesa-Bianchi, D. R. Hardoon et G. Leen. Guest Editorial : Learning from multiple sources. *Machine Learning*, 79(1-2) :1, 2010.
(Cf. page 29)

[IG17] S.-C. Chabert-Liddell. *Statistical learning of collections of networks with applications in ecology and sociology*. Thèse de doctorat, Université Paris-Saclay, 2022.
(Cf. page 55)

[IG18] S.-C. Chabert-Liddell, P. Barbillon, S. Donnet et E. Lazega. A stochastic block model approach for the analysis of multilevel networks : An application to the sociology of organizations. *Computational Statistics & Data Analysis*, 158 :107179, 2021.
(Cf. page 56)

[IG19] H. J. Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6) :392-404, 2009.
(Cf. page 40)

[IG20] J.-C. Corvol, F. Artaud, F. Cormier-Dequaire, O. Rascol, F. Durif, P. Derkinderen, A.-R. Marques, F. Bourdain, J.-P. Brandel, F. Pico et al. Longitudinal analysis of impulse control disorders in Parkinson disease. *Neurology*, 91(3) :e189-e201, 2018.
(Cf. page 38)

[IG21] B. A. Coull, D. Ruppert et M. Wand. Simple incorporation of interactions into additive models. *Biometrics*, 57(2) :539-545, 2001.
(Cf. page 48)

[IG22] A. Dehman, C. Ambroise et P. Neuvial. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16(1) :148, 2015.
(Cf. page 44)

[IG23] A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* :1-38, 1977.
(Cf. page 37)

[IG24] L. Fu, P. Lin, A. V. Vasilakos et S. Wang. An overview of recent multi-view clustering. *Neurocomputing*, 402 :148-161, 2020.
(Cf. page 29)

[IG25] K. Golalipour, E. Akbari, S. S. Hamidi, M. Lee et R. Enayatifar. From clustering to clustering ensemble selection : A review. *Engineering Applications of Artificial Intelligence*, 104 :104388, 2021.
(Cf. page 54)

[IG26] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa et J. Tegnér. Data integration in the era of omics : current and future challenges. *BMC Systems Biology*, 8(2) :1-10, 2014.
(Cf. page 29)

[IG27] A. D. Gordon. *Classification*. Monographs on statistics and applied probability. Chapman & Hall, CRC Press, Boca Raton, Florida, United-States of America, 1999.
(Cf. page 45)

[IG28] I. C. Gormley et S. Früwirth-Schnatter. Mixture of Experts Models. S. Frühwirth-Schnatter, G. Celeux et C. P. Robert, éditeurs, *Handbook of mixture analysis*, chapitre 12, pages 271-308. CRC Press, 2019.
(Cf. page 36)

[IG29] Q. Grimonprez. *Sélection de groupes de variables corrélées en grande dimension*. Thèse de doctorat, Université de Lille, 2016.
(Cf. page 45)

[IG30] Q. Grimonprez, S. Blanck, A. Celisse et G. Marot. MLGL : an R package implementing correlated variable selection by hierarchical clustering and group-lasso. *Journal of Statistical Software*, 106 :1-33, 2023.
(Cf. pages 45, 46)

[IG31] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky et J. Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *The Journal of Machine Learning Research*, 21(1) :10039-10081, 2020.
(Cf. page 58)

[IG32] S. Hoffmann, F. Schönbrodt, R. Elsas, R. Wilson, U. Strasser et A.-L. Boulesteix. The multiplicity of analysis strategies jeopardizes replicability : lessons learned across disciplines. *Royal Society Open Science*, 8(4) :201925, 2021.
(Cf. page 58)

[IG33] S. Huang, K. Chaudhary et L. X. Garmire. More is better : recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 8 :84, 2017.
(Cf. page 29)

[IG34] B.-Y. Jing, T. Li, Z. Lyu et D. Xia. Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics*, 49(6) :3181-3205, 2021.
(Cf. page 55)

[IG35] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng et Y. Luo. Multimodal machine learning in precision health : A scoping review. *NPJ Digital Medicine*, 5(1) :171, 2022.
(Cf. page 29)

[IG36] W. J. Krzanowski et Y. T. Lai. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44(1) :23-34, 1988.
(Cf. page 52)

[IG37] D. Lahat, T. Adali et C. Jutten. Multimodal data fusion : an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9) :1449-1477, 2015.
(Cf. page 29)

[IG38] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui et M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5(Jan) :27-72, 2004.
(Cf. page 32)

[IG39] B. Li et S. M. Leal. Methods for detecting associations with rare variants for common diseases : application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3) :311-321, 2008.
(Cf. page 51)

[IG40] H. Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2 :73-94, 2015.
(Cf. page 44)

[IG41] M. Li, K. Wang, S. F. A. Grant, H. Hakonarson et C. Li. ATOM : a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics*, 25(4) :497-503, 2008.
(Cf. page 51)

[IG42] G. Loosli, S. Canu et C. S. Ong. Learning SVM in Kren Spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(6) :1204-1216, 2016.
(Cf. page 34)

[IG43] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265) :747-753, 2009.
(Cf. page 40)

[IG44] J. Mariette et N. Vialaneix. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34(6) :1009-1015, 2018.
(Cf. page 32)

[IG45] L. Meier, S. v. d. Geer et P. Buhlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37 :3779-3821, 2009.
(Cf. page 48)

[IG46] G. W. Milligan et M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2) :159-179, 1985.
(Cf. page 45)

[IG47] N. D. Nguyen et D. Wang. Multiview learning for understanding functional multiomics. *PLoS computational biology*, 16(4) :e1007677, 2020.
(Cf. page 32)

[IG48] R. Ottman. Gene–environment interaction : definitions and study design. *Preventive medicine*, 25(6) :764-770, 1996.
(Cf. pages 40, 41)

[IG49] M. Pierre-Jean, J.-F. Deleuze, E. Le Floch et F. Mauger. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Briefings in Bioinformatics*, 21(6) :2011-2030, 2020.
(Cf. page 29)

[IG50] A. Rakotomamonjy, F. Bach, S. Canu et al. SimpleMKL. *Journal of Machine Learning Research*, 9 :2491-2521, 2008.
(Cf. pages 20, 33)

[IG51] D. Ramachandram et G. W. Taylor. Deep multimodal learning : A survey on recent advances and trends. *IEEE Signal Processing magazine*, 34(6) :96-108, 2017.
(Cf. page 29)

[IG52] T. Rebafka. Model-based graph clustering of a collection of networks using an agglomerative algorithm. *arXiv preprint*, 2022.
(Cf. page 55)

[IG53] B. Schölkopf, K. Tsuda et J.-P. Vert. *Kernel methods in computational biology*. MIT press, 2004.
(Cf. page 31)

[IG54] P. C. Sham et S. M. Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5) :335-346, 2014.
(Cf. page 42)

[IG55] S. R. Stahlschmidt, B. Ulfenborg et J. Synnergren. Multimodal deep learning for biomedical data fusion : a review. *Briefings in Bioinformatics*, 23(2) :bbab569, 2022.
(Cf. page 29)

[IG56] V. Stanislas. *Statistical approaches to detect epistasis in Genome Wide Association Studies*. Thèse de doctorat, Université Paris-Saclay (ComUE), 2017.
(Cf. page 40)

[IG57] N. Stanley, S. Shai, D. Taylor et P. J. Mucha. Clustering network layers with the strata multilayer stochastic block model. *IEEE Transactions on Network Science and Engineering*, 3(2) :95-105, 2016.
(Cf. page 55)

[IG58] D. Thomas. Gene–environment-wide association studies : emerging approaches. *Nature Reviews Genetics*, 11(4) :259-272, 2010.
(Cf. page 40)

[IG59] R. Tibshirani, G. Walther et T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society : Series B*, 63(2) :411-423, 2001.
(Cf. page 52)

[IG60] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. De Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen et D. Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1) :59, 2021.
(Cf. page 40)

[IG61] F. A. Van Eeuwijk. Multiplicative interaction in generalized linear models. *Biometrics* :1017-1032, 1995.
(Cf. page 48)

[IG62] T. J. VanderWeele et M. J. Knol. A tutorial on interaction. *Epidemiologic methods*, 3(1) :33-72, 2014.
(Cf. page 40)

[IG63] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4) :395-416, 2007.
(Cf. page 31)

[IG64] K. Wang et D. Abbott. A principal components regression approach to multilocus genetic association studies. *Genetic Epidemiology : The Official Publication of the International Genetic Epidemiology Society*, 32(2) :108-118, 2008.
(Cf. page 51)

[IG65] M. H. Wang, H. J. Cordell et K. Van Steen. Statistical methods for genome-wide association studies. *Seminars in cancer biology*, tome 55, pages 53-60. Elsevier, 2019.
(Cf. page 40)

[IG66] S. Wang, X. Liu, L. Liu, S. Zhou et E. Zhu. Late fusion multiple kernel clustering with proxy graph refinement. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
(Cf. page 55)

[IG67] S. Wang, X. Liu, E. Zhu, C. Tang, J. Liu, J. Hu, J. Xia et J. Yin. Multi-view Clustering via Late Fusion Alignment Maximization. *IJCAI*, pages 3778-3784, 2019.
(Cf. page 55)

[IG68] T. Wang et R. C. Elston. Improved power by use of a weighted score test for linkage disequilibrium mapping. *The American Journal of Human Genetics*, 80(2) :353-360, 2007.
(Cf. page 51)

[IG69] J. H. J. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236-244, 1963.
(Cf. page 44)

[IG70] Wikistat. Exploration multivariée, 2016.
(Cf. page 31)

[IG71] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing et M. Sugiyama. High-dimensional feature selection by feature-wise kernelized lasso. *Neural computation*, 26(1) :185-207, 2014.
(Cf. page 32)

[IG72] X. Yan, S. Hu, Y. Mao, Y. Ye et H. Yu. Deep multi-view learning methods : A review. *Neurocomputing*, 448 :106-129, 2021.
(Cf. page 29)

[IG73] X. Yi et C. Caramanis. Regularized EM algorithms : a unified framework and statistical guarantees. *Advances in Neural Information Processing Systems*, tome 28, pages 1-9, 2015.
(Cf. page 38)

[IG74] R. Zhang, F. Nie, X. Li et X. Wei. Feature selection with multi-view data : A survey. *Information Fusion*, 50 :158-167, 2019.
(Cf. page 31)

[IG75] J. Zhao, X. Xie, X. Xu et S. Sun. Multi-view learning overview : Recent progress and new challenges. *Information Fusion*, 38 :43-54, 2017.
(Cf. page 29)

[IG76] F. Zhou, J. Ren, X. Lu, S. Ma et C. Wu. Gene–environment interaction : A variable selection perspective. *Epistasis : Methods and Protocols* :191-223, 2021.
(Cf. pages 40, 42)