Analyse de données bio-médicales. Recherche reproductible et modélisation mathématique

Sébastien Li-Thiao-Té

Université Paris XIII - Sorbonne Paris Nord
École doctorale Sciences, Technologies, Santé Galilée

Habilitation à diriger des recherches

Soutenue par

Sébastien Li-Thiao-Té

Analyse de données bio-médicales Recherche reproductible et modélisation mathématique

Date de soutenance : 23 novembre 2021

Rapporteurs :

Luis Alvarez	-	Université de Las Palmas de Gran Canaria, Espagne
Antoine Chambaz	-	Université de Paris, France
Jacques Froment	-	Université de Bretagne Sud, France

Examinateurs :

Marion Darbas	-	Université Sorbonne Paris Nord, France
Agnès Desolneux	-	ENS Paris-Saclay / Centre Borelli, France
Françoise Dibos	-	Université Sorbonne Paris Nord, France
Jean-Michel Morel	-	ENS Paris-Saclay / Centre Borelli, France
Yves-Michel Frapart	-	Université de Paris, France
Hatem Zaag	-	CNRS, Université Sorbonne Paris Nord, France

Remerciements

Le travail présenté dans ce manuscrit est le résultat de mes échanges avec les nombreux chercheurs de toutes disciplines que j'ai rencontrés jusqu'à présent. Je remercie particulièrement ceux avec qui j'ai eu le plaisir d'échanger plus étroitement, autour d'un projet de recherche (CA Cuenod, Y Frapart, P Garteiser, BT Doan, D Geldwerth, P Bourdoncle, X Treton, É Ogier-Denis, Y Bouhnik) et/ou du co-encadrement d'une thèse (M Luong, JM Rocchisani, J Chaussard, H Zaag). Leurs questionnements m'ont forcé à remettre en question ma façon de faire des mathématiques appliquées et m'ont forcé à avoir les idées claires.

Je remercie tout particulièrement Françoise Dibos. De près ou de loin, elle a toujours su accompagner mes travaux depuis mon arrivée à Paris 13. Son soutien et son leadership ont beaucoup contribué à ce que je suis scientifiquement aujourd'hui.

Je remercie l'ensemble des membres du jury pour leur intérêt dans mes travaux, et en particulier les rapporteurs qui ont pris le temps de me lire. Je sais que chacune de leurs remarques me fera avancer.

Enfin, je tiens à remercier Thérèse, Fanny et Théodore, qui ont contribué chacun à leur façon à la réalisation de ce travail. J'espère que Théodore lira ces mots un jour.

Résumé

Les travaux présentés concernent le traitement mathématique du signal et des images, et l'analyse de données en général. Ils abordent la question de l'application des mathématiques dans la vie réelle, et plus particulièrement de la production rigoureuse d'énoncés scientifiques valides pour les domaines de la biologie et de la médecine.

Le problème central concerne l'utilisation des modèles en mathématiques appliquées. Bien souvent, lors de l'analyse d'un jeu de données, un certain nombre de choix et d'hypothèses sont formulées arbitrairement. Les résultats de l'analyse sont alors dépendants et affaiblis par ces choix arbitraires. Nous proposons d'envisager cette question en considérant les énoncés scientifiques comme des ensembles que l'on démontre par accumulation de données ou de modèles.

Cette démarche est soutenue par le développement du logiciel Lepton. Cet outil de recherche reproductible améliore la documentation des codes sources et rend transparente l'analyse des données effectuée dans un travail de recherche.

Cette démarche est illustrée dans trois contextes correspondant à trois encadrements de thèses de doctorat. En analyse d'images bio-médicales, nous explorons la structure de l'ensemble des images à l'aide de modèles de type patchs, et montrons l'intérêt de ces modèles pour le débruitage et la super-résolution. Ces travaux ont été réalisés dans le cadre de la thèse de Dai Viet Tran soutenue en 2018 (co-encadrement avec F. Dibos et M. Luong).

En spectroscopie par résonance paramagnétique électronique, les caractéristiques biochimiques de l'échantillon sont reliées aux paramètres d'un modèle du type fonction de Lorentz. Nous montrons comment optimiser la précision des mesures de la concentration en oxygène et du stress oxydant. Ces travaux ont été réalisés dans le cadre de la thèse de Duc-Nghia Tran soutenue en 2018 (co-encadrement avec Y. Frapart).

Pour le diagnostic des maladies inflammatoires chroniques de l'intestin, nous avons analysé les vidéos de coloscopies obtenues dans le cadre d'une collaboration avec l'hôpital Bichat-Beaujon (X Treton, E Ogier-Denis). Nous étudions la forme des lésions de la paroi intestinale à l'aide de classificateurs linéaires et nous en déduisons la répartition anatomique de ces lésions. Sur la base d'un modèle d'équations de réaction-diffusion (équation de Fisher-KPP), nous calculons un ensemble de vitesses de progression de la maladie. Ce travail a été réalisé dans le cadre de la thèse de Safaa Al Ali (co-encadrement avec H. Zaag).

Summary

The work presented here is related to signal and image processing, and data analysis in general. It concerns the application of mathematics to real life, and in particular how to construct rigorous scientific truths in biology and medecine.

The main difficulty is how to use mathematical models in science. Data analysis usually involves arbitrary choices and hypotheses. The analysis results are thus dependent on those choices and are weakened by them. We propose to consider scientific statements as mathematical sets and demonstrate them by accumulating data and/or models.

This process is supported by the development of the Lepton software. This reproducible research tool allows for better documentation of source code and improves the transparency of data analysis procedures in the context of research work.

This process is illustrated in three separate projects, each corresponding to the work of a PhD student. In the domain of biomedical image analysis, we have used patch-based models to study the structure of the set of images, and show how this improves denoising and super-resolution. This work was performed during Dai Viet Tran's PhD thesis and defended in 2018 (co-supervised with F. Dibos and M. Luong).

In electronic paramagnetic resonance spectroscopy, the biochemical properties of a sample are reflected in the parameters of the Lorentz model of the spectrum. We show how to optimize the precision of the instrument and study oxymetry and oxidative stress. This work was performed during Duc-Nghia Tran's PhD thesis and defended in 2018 (co-supervised with Y. Frapart).

For inflammatory bowel diseases, we analysed coloscopy videos obtained in collaboration with the Bichat-Beaujon Hospital (X Treton, E Ogier Denis). We study the appearance of the lesions of the gut wall through linear classifiers and deduce their anatomical position. With a reaction-diffusion model (Fisher KPP equation), we compute a set of progression speeds for the disease. This work was performed during Safaa Al Ali's PhD thesis (co-supervised with H. Zaag).

Table des matières

Chapitre 1

Contexte applicatif et motivations

Mon programme de recherche concerne l'analyse des données et l'apprentissage automatique, et plus particulièrement la façon dont on utilise des modèles pour produire des énoncés scientifiques dans les applications des mathématiques. C'est une question de méthode scientifique : la science contemporaine produit des masses de données que personne n'envisage de traiter sans l'outil informatique. Or le traitement des données repose trop souvent sur des calculs et algorithmes provenant d'hypothèses de modélisation arbitraires. On cherche donc à comprendre comment démontrer rigoureusement un énoncé scientifique par des preuves obtenues par le traitement numérique de données expérimentales.

1.1 Doute, science et mathématiques

Il n'y a pas de doute en mathématiques. Il y a des hypothèses, des théorèmes et des démonstrations. Les théorèmes d'incomplétude (par exemple Gödel [49]) ou les conjectures ne remettent pas en cause la cohérence de la théorie. Dans les autres disciplines scientifiques, le corpus des connaissances et les dogmes sont régulièrement remis en cause. En sciences physiques, la théorie de la gravitation de Newton a été complétée par la relativité générale d'Einstein. En biologie et médecine, on découvre régulièrement de nouvelles espèces (bactéries vivant sans oxygène avec un métabolisme du soufre), de nouveaux processus biologiques (analogues des bases de l'ADN en biologie synthétique), conduisant à étendre notre définition des êtres vivants. En médecine ce doute est un problème, parce qu'il y a un risque pour le patient. Mes nombreuses collaborations avec les sciences biomédicales ont certainement joué un rôle important dans ce questionnement personnel, par l'exigence de rigueur qui y est demandée et la

remise en cause des méthodes mathématiques employées.

Dans les sciences expérimentales, il y a plusieurs sources de doute. En premier lieu, il y a l'hésitation provenant de ce que la vérité est inconnue, et qu'il faut choisir entre un grand nombre de modèles ou de théories à partir de données en nombre limité. La méthode scientifique s'attache principalement à cette partie du doute. En accumulant des observations et des expériences, le scientifique s'attache à confirmer, préciser ou infirmer les énoncés étudiés.

Cependant, le développement technologique fait ressortir d'autres problèmes. Ainsi les mesures expérimentales sont de plus en plus complexes. Un premier exemple est fourni ici par la mesure de la concentration en oxygène en résonance paramagnétique électronique (oxymétrie RPE, cf chapitre 6). L'obtention du nombre passe par la mise en résonance du spin de l'électron (physique quantique), la modulation du champ magnétique (électronique), l'extraction de caractéristiques du spectre (traitement du signal) et leur interprétation par le chimiste. Un deuxième exemple est fourni en imagerie biomédicale (cf chapitres 4 et 5) par la multiplicité des modalités d'imagerie (scanner CT, PET, IRM, ultrasons, etc.), la nécessité de développer des instruments spécifiques de chaque modalité, et la volonté d'analyser ensemble toutes les données recueillies. Dans ces deux exemples, on remarque que l'exploitation des données nécessite la connaissance de nombreux domaines scientifiques et donne lieu à de riches interactions interdisciplinaires.

Indépendamment des difficultés à identifier les lois de la nature, le scientifique lui-même est parfois à l'origine du doute. De nombreux articles scientifiques publiés ne comportent pas les informations nécessaires à l'évaluation de la solidité de la démarche, que les conclusions soient elles-mêmes correctes ou non. Il manque régulièrement les données, les codes sources utilisés, la documentation des codes, la justification de la démarche utilisée, etc. Le mouvement récent autour de la recherche reproductible correspond à la prise de conscience que ces éléments sont importants pour la solidité et la confiance que l'on va accorder aux résultats, et que la complexification des données et des approches a empiré la situation.

Un dernier problème concerne les lacunes de l'analyse de données actuelle. L'arbitraire d'une hypothèse de modélisation réside parfois dans l'absence de motivation d'un choix, mais parfois aussi dans l'incapacité de motiver ce choix, ou de faire autrement. De même, le résultat d'une analyse de données peut être difficile à interpréter, par exemple quand la formulation du modèle n'est pas compacte, lisible par l'humain, ou fait intervenir les données d'apprentissage comme c'est le cas des réseaux de neurones.

1.2 Nécessité de l'outil informatique

Même si on ne comprend pas toujours bien comment, les méthodes d'intelligence artificielle ont des performances et ont eu des succès indéniables. Dans les jeux tout d'abord. En 1997 le champion d'échecs Garry Kasparov est battu par la machine DeepBlue d'IBM. En 2017, la filiale DeepMind de Google propose le programme AlphaGo qui bat Lee Sedol, champion du monde en titre de Go, dans un tournoi à cinq parties [88]. Les successeurs de ces technologies ont des performances encore meilleures, par exemple AlphaZero de DeepMind [89] dont l'apprentissage (par renforcement) a été effectué sans intervention humaine. Parmi les réussites, citons également les performances des algorithmes d'analyse d'image et de reconnaissance de visage (Imagenet [65]), ainsi que les progrès dans le diagnostic médical, par exemple la plateforme Watson d'IBM [44].

Il y a de nombreuses raisons de mettre en oeuvre l'outil informatique : c'est plus rapide (indispensable pour traiter les grandes masses de données), c'est moins cher (pas d'intervention humaine), etc. Les raisons qui nous intéressent particulièrement dans ce travail concernent

— la rigueur et la transparence : l'ordinateur applique une méthode, totalement spécifiée, selon les règles indiquées par son code source,
— la possibilité de répéter de nombreuses fois le calcul pour des simulations numériques,
— l'exploration de modèles inacessibles pour le cerveau humain.

Ces trois propriétés de l'outil informatique fournissent des moyens de lever le doute, dans des niveaux de plus en plus élaborés. Rigueur et transparence sont nécessaires pour conduire une analyse de données conforme, même si la méthode utilisée est fausse, ou sous-optimale. Ceci est important et nécessaire par rapport au cerveau humain, qui est soumis à un certain nombre de biais cognitifs qui limitent sa capacité de raisonnement. Par exemple, le cerveau humain évalue mal les probabilités des événements d'après leurs occurences. De plus, le cerveau humain accorde une importance plus importante aux informations qui confirment l'hypothèse de départ, aux informations récentes, etc. Ces défauts du raisonnement ont été étudiés par exemple par A Tversky et D Kahneman (prix Nobel d'économie 2002) [57], et conduisent à une variabilité entre experts dans le domaine médical ([71, 81] et Chapitre 7.5.1).

La puissance de calcul de l'outil informatique autorise de nouvelles approches d'analyse de données. Elle permet d'explorer l'ensemble des modèles, pour être exhaustif ou du moins avoir un échantillon représentatif. Elle permet de faire des calculs en avance, afin de calibrer et optimiser les paramètres. En faisant varier

l'aléa et les paramètres, la puissance de calcul permet d'évaluer la variabilité et la robustesse des résultats.

La possibilité d'utiliser des modèles inaccessibles pour le cerveau humain est particulièrement importante. En effet, le cerveau humain n'est pas une machine tout à fait universelle, mais plutôt une machine très efficace qui a été entraînée pour un certain type de tâches. Par exemple, le système visuel humain est très efficace pour certaines tâches basiques (reconnaissance d'alignements, de droites, de clusters, etc.) et certaines tâches élaborées (reconnaissance de visages). Cela tient certainement à son architecture particulière, qui fait encore l'objet de nombreuses recherches, par exemple [37]. Dans le même temps, on sait que le système visuel est mauvais sur certaines caractéristiques telles que la texture des images, et que ces caractéristiques peuvent comporter une information diagnostic en imagerie médicale, par exemple dans le domaine de la radiomique [66, 67].

1.3 Utilisation concrète

Si l'ordinateur est un outil indispensable pour l'analyse de données et la production de vérités scientifiques, il ne produit pas une analyse de données en complète autonomie. Même dans un apprentissage automatique tel que AlphaZero sur le jeu de Go, un opérateur humain a dû spécifier l'énoncé précis du problème à résoudre, ainsi que l'éventail des modèles considérés ou de façon équivalente l'algorithme d'apprentissage. Le modélisateur est donc à l'origine d'un certain nombre de choix ; les travaux contenus dans ce manuscrit ont pour objectif d'éclairer ces choix et de mieux comprendre la qualité des énoncés scientifiques qu'ils permettent d'obtenir.

Un premier ensemble de travaux concerne la mise en oeuvre de l'outil informatique dans une démarche objective, permettant de lever les doutes concernant les détails de l'obtention des preuves. Cette démarche, dite de *recherche reproductible*, conduit à exposer l'analyse de données sous un format adapté à la fois à l'ordinateur, ce qui permet de refaire les calculs, et adapté au lecteur humain, ce qui permet de les comprendre et de les réutiliser. Le chapitre 3 décrit le logiciel Lepton qui a été développé pour mettre en application concrètement ces idées, et inclus un exemple d'utilisation concrète. Ce logiciel a également été utilisé pour produire ce manuscrit.

Comme indiqué plus haut, les données sont devenues complexes, tant dans leur taille que dans leur multi-modalité. Leur analyse nécessite d'associer outil informatique et expertise interdisciplinaire. Une partie importante de ma recherche a été consacrée aux collaborations avec des scientifiques de toutes

disciplines, autour des problématiques de l'analyse des images biomédicales. Cette démarche transversale a donné lieu à ma participation dans le programme interdisciplinaire "Imageries du Vivant" (cf chapitre 4), mais aussi à des collaborations directes avec Jean-Marie Rocchisani (radiologue, hôpital Avicenne, cf chapitre 5), Yves Frapart (chimiste au LCBPT, Paris Descartes, cf chapitre 6) et Xavier Treton et Yoram Bouhnik (gastro-entérologues à l'hôpital Bichat-Beaujon, cf chapitre 7).

Le travail mathématique au plus près des applications m'a forcé à questionner la justification que l'on peut apporter aux hypothèses de modélisation. Les réflexions qui en découlent, présentées dans le chapitre 2, dépassent le cadre de l'analyse d'image, et sous-tendent l'ensemble des travaux qui sont présentés dans les sections ultérieures, même si elles n'ont réellement été mises en oeuvre qu'à partir de l'encadrement de la thèse de Safaa Al Ali (chapitre 7). En particulier, en fournissant un cadre théorique unifié, elles ont permis un dialogue fertile entre l'analyse des images et la modélisation mathématique par les EDP au sein de l'équipe "Mathématiques pour la Biologie et les Images" du LAGA.

Chapitre 2

Méthodologie générale

Ce chapitre présente un cadre général d'analyse des problématiques de modélisation. Je propose ici d'envisager la modélisation comme un travail sur des ensembles de modèles, dans un sens très large. Les choix ou hypothèses de modélisation sont alors vus comme des restrictions de l'ensemble des modèles, et les énoncés scientifiques sont vus comme des propriétés d'ensembles de modèles.

Afin d'illustrer les réflexions de ce chapitre, on utilisera le problème jouet suivant : "quelle analyse peut-on faire du jeu de données $(0, 0, 0, 1, 1, 1)$" ? Ce jeu de données est à la fois arbitraire et abstrait, mais il souligne l'absence de connaissances a priori communément rencontré en analyse d'images et dans les sciences bio-médicales. Les élements de ce chapitre sont également abondamment mis en oeuvre dans les travaux de modélisation de la recto-colite hémorragique présentés au chapitre 7.

2.1 Structure des données et des modèles

Tout d'abord, on postule que *les données recueillies sont potentiellement informatives et essentiellement incontestables*. Ce sont les données qui sont le point de départ de la démarche d'analyse de données ; les modèles et le modélisateur doivent s'adapter aux observations. Dans l'exemple $(0, 0, 0, 1, 1, 1)$, le modélisateur doit accepter la série de nombres qui lui a été donnée, et proposer des modèles et des méthodes appropriées, sans a priori.

Les données sont potentiellement informatives, sinon on ne pourra en tirer aucune information, uniquement des coïncidences. L'application d'une méthode d'analyse correcte et rigoureuse conduira à des résultats corrects, mais sans lien avec l'objectif initial. Par exemple, on peut analyser la structure d'un signal et y rechercher des motifs. Cependant, pour conclure que l'on a trouvé la trace

d'une communication extraterrestre dans un signal radio, il y faut supposer au préalable que le signal en question contient cette trace [1].

Les données sont essentiellement incontestables, ce qui signifie que l'on ne peut pas remettre en cause dans une démarche scientifique les valeurs observées, i.e. $(0, 0, 0, 1, 1, 1)$. Ceci ne signifie pas qu'il ne peut y avoir de problèmes dans les données : il peut y avoir des donneés manquantes, bruitées, systématiquement ou non, elles peuvent ne porter que sur une sous-population, ou dépendre de l'opérateur, etc. Cependant, on ne peut pas décider de changer arbitrairement telle ou telle valeur, ou choisir d'exclure certains individus. De plus, la démarche scientifique impose de considérer a priori que toutes les données fournissent la même quantité d'information, et qu'il n'y a pas lieu de privilégier certaines par rapport à d'autres.

L'interdiction de modifier les données résulte du fait que l'information est portée par les valeurs numériques. Ainsi, ce n'est pas la même chose d'observer $(0, 0, 0, 1, 1, 1)$ par rapport à $(0, 0, 0.1, 1, 1, 1)$. On est donc amené à comparer les observations avérées à tout ce qu'elles auraient pu être, et à fixer ou définir l'ensemble des observations potentielles. Cet ensemble sera appelé espace des données, noté $\mathcal{E}$.

L'espace des données peut être fourni sous la forme d'un espace mathématique tel que l'ensemble des nombres réels $\mathcal{E} = \mathbb{R}$, ou bien l'ensemble des fonctions continues $\mathcal{C}(E, F)$ entre deux espaces. L'espace peut également être indiqué comme en informatique par son type, `float` ou `E->F` par exemple. On ne fera pas de distinction particulière entre données et méta-données. Par conséquent, l'espace auquel appartient le jeu de données n'est pas homogène, mais plutôt "multi-modal", car il contient des informations de nature différente en général.

Afin de définir une notion de modèle, on remarque d'abord que l'apprentissage et la modélisation sont des entreprises où l'on recherche des objets similaires aux données observées. C'est en particulier le cas de l'apprentissage supervisé qui élabore des modèles prédictifs, calibrés par rapport aux observations passées d'une relation entre des "inputs" et des "outputs". Dans un apprentissage non supervisé, il s'agit souvent de modéliser la répartition des données ou bien d'en donner une version simplifiée. La similarité est donc un objectif de la modélisation, même s'il n'est pas clair que les modèles similaires soient corrects, c'est-à-dire que la proximité soit une condition suffisante pour que le modèle choisi et la réalité possèdent des propriétés communes. Dans la lignée de [28],

1. On fait référence ici au projet SETI@HOME qui analyse les signaux enregistrés par les radiotélescopes en vue d'identifier des traces d'une civilisation extraterrestres.

"all models are wrong, but some are useful".

À cause de cet objectif de similarité, les modèles sont naturellement reliés aux données, et l'ensemble des modèles est naturellement muni d'une relation d'ordre basée sur la similarité. Ils appartiennent donc à un espace noté $\mathcal{M}$ qui ressemble à celui des données. Plusieurs cas de figure sont envisageables. Tout d'abord, on peut utiliser le même ensemble, i.e. $\mathcal{M} = \mathcal{E}$. Par exemple, pour $(0, 0, 0, 1, 1, 1)$, on peut prendre $\mathcal{M} = \mathcal{E} = \{0, 1\}^6$.

On peut utiliser un ensemble de modèles $\mathcal{M}$ plus grand que l'espace des données $\mathcal{E}$ quand on considère que les données sont une observation partielle du phénomène, par exemple, $\{0, 1\}^6 \subset \mathcal{M} = \mathbb{R}^6$. En traitement du signal, l'échantillonnage, la quantification, les limites de la fenêtre d'observation conduisent à considérer que les modèles sont des fonctions $\mathbb{R} \to \mathbb{R}$ même si les observations sont une suite de nombres (entiers). Autre exemple, en analyse d'image, l'analyse harmonique des images (transformée de Fourier) conduit à travailler dans un espace $\mathcal{M}$ d'images périodisées et symétrisées afin d'éviter des effets de bord au lieu de la fenêtre d'observation. L'espace plus grand peut également inclure des variables explicatives cachées, non observées ou non observables, par exemple dans le cas où l'on utilise des modèles de type chaîne de Markov.

On peut utiliser un ensemble plus petit lorsque l'on fait des hypothèses sur les modèles, ce qui conduit à exclure certains modèles ou à les considérer comme peu vraisemblables a priori. Par exemple en traitement du signal, on restreint l'ensemble $\mathcal{M}$ par une hypothèse de régularité en ne considérant que les fonctions indéfiniment dérivables ou en appliquant un filtre passe-bas. Pour modéliser une quantité physique telle que la masse ou le volume, on peut restreindre l'ensemble $\mathcal{M}$ aux fonctions positives. Cette restriction de l'espace peut être une conséquence de l'instrument de mesure et n'est pas forcément liée à un défaut. Ainsi, observer une scène tridimensionnelle à l'aide d'une caméra applique une projection de l'espace 3D sur un plan.

De façon générale, lorsque l'espace des modèles $\mathcal{M}$ est différent de l'espace des données $\mathcal{E}$, un *modèle d'observation* indique comment l'on passe d'un modèle idéal à une observation "bruitée". La notion de problème inverse consiste à effectuer le passage dans l'autre sens. Par exemple :

- en statistique paramétrique, on donne $\mathbb{P}(y|\theta)$, où y désigne les observations et θ désigne le(s) paramètre(s) du modèle. Estimer revient à donner la valeur de θ correspondant aux observations.
- en analyse d'images, la restauration d'images consiste à retrouver une image idéale non bruitée à partir d'une ou plusieurs observations.
- en analyse numérique, on cherche à retrouver l'état initial d'après une observation du système à un instant donné ou de sa trajectoire.

L'objectif de similarité évoqué plus haut peut alors être réalisé sous différentes formes. Tout d'abord, le modèle d'observation induit une relation algébrique entre modèles de $\mathcal{M}$ et données de $\mathcal{E}$. Ainsi, on dira qu'un modèle $m \in \mathcal{M}$ et des observations $e \in \mathcal{E}$ sont compatibles si m peut générer les données e. Par exemple, considérons que les données $(0, 0, 0, 1, 1, 1)$ sont obtenues par échantillonnage d'une fonction (polynômiale) aux points $(x_1, \ldots, x_6)$ distincts. Ces données sont incompatibles avec toutes les fonctions constantes, et toutes les fonctions linéaires. Plus généralement, pour qu'une fonction polynômiale soit compatible avec $(0, 0, 0, 1, 1, 1)$, il faut et il suffit que cette fonction interpole les données, ce qui nécessite un polynôme de degré supérieur ou égal à 5. En conclusion, dans l'ensemble $\mathcal{M} = \mathbb{R}[X]$ des polynômes, les polynômes de degré inférieur ou égal à 4 sont incompatibles avec les données. Soit P un polynôme interpolateur de degré 5, l'ensemble des modèles compatibles est $\prod_i (X - x_i).\mathbb{R}[X] + P$. Un autre exemple de cette notion de compatibilité est fourni à la section 2.5.

Le modèle d'observation peut induire une "distance" entre données et modèles compatibles. Par exemple le bruit gaussien est associé à une distance L_2. En effet, étant donné un modèle m observé par échantillonnage aux points $(x_1, \ldots, x_n)$ avec un bruit gaussien additif blanc η d'écart-type σ, la probabilité d'observer un vecteur y est donnée par :

$$\mathbb{P}[y|m] = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - m(x_i))^2}{2\sigma^2}} = Ce^{-\frac{||y-m(x)||^2}{2\sigma^2}}$$

La fonction $(m, y) \mapsto \mathbb{P}[y|m]$ fournit une pseudo-distance entre les espaces $\mathcal{M}$ et $\mathcal{E}$. Suivant l'espace $\mathcal{M}$ choisi, cette fonction peut valoir 0 pour plusieurs modèles, ce qui conduit à des problèmes d'identifiabilité des modèles. Dans l'exemple utilisé ici, tous les polynômes interpolateurs sont indistinguables. En pratique, on est amené à considérer soit des modèles de degré faible au sens des moindres carrés ou bien à introduire un terme de régularisation ou pénalisation arbitraire dans la distance.

Enfin, le modèle d'observation induit une notion de distribution sur l'ensemble des modèles. Il s'agit simplement d'observer l'image réciproque du jeu de données par le modèle d'observation. Quand ce dernier est déterministe, l'image réciproque est l'ensemble des modèles compatibles. Quand le modèle d'observation est stochastique, l'image réciproque est pondérée par la probabilité de générer telle ou telle observation. En statistiques, cette notion s'appelle la vraisemblance et est donnée par la fonction $m \mapsto \mathbb{P}[y|m]$ pour des observations y fixées.

2.2 Modélisation

Il y a donc quatre choses en jeu :
— la structure des données telle que définie par l'espace $\mathcal{E}$
— la structure des modèles telle que définie par l'espace $\mathcal{M}$
— les valeurs des observations, par opposition aux autres valeurs possibles
— les valeurs des modèles, i.e. les modèles similaires par opposition aux modèles incompatibles

La similarité entre $\mathcal{M}$ et $\mathcal{E}$ apparaît implicitement dans le sous-ensemble des modèles similaires, et peut également être définie explicitement par un modèle d'observation.

La modélisation consiste donc à ajuster les différents éléments de façon à produire une analyse de données correcte, en justifiant les choix effectués à chaque étape. Si à ce stade, je suis bien incapable de donner une théorie générale pour justifier les hypothèses de modélisation, ce cadre théorique permet d'envisager un certain nombre de choses de façon claire.

Par exemple, il est à noter que en dehors de toute hypothèse de modélisation, on est obligé de considérer que les données sont ponctuelles, par opposition à la répétition d'observations d'un même phénomène. En effet, faire l'hypothèse que le jeu de données est constitué d'observations répétées revient à imposer une structure d'espace produit sur $\mathcal{E}$ et par conséquent sur $\mathcal{M}$. Cette hypothèse suggère également que l'ordre des observations n'est pas important, i.e. que les données et modèles sont invariants par permutation. La définition des espaces $\mathcal{E}$ et $\mathcal{M}$ n'est donc pas sans conséquence, et doit être justifiée soigneusement.

Certaines de ces hypothèses de modélisation sont facilement justifiables. C'est le cas en particulier des propriétés qui ont été imposées dans le modèle d'observation par l'expérimentateur, ou par l'instrument de mesure : schéma d'échantillonnage, quantification, etc. Il est important de les expliciter car ces hypothèses contraignent les modèles et leurs conclusions. Si $(0, 0, 0, 1, 1, 1)$ est issu d'un tirage aléatoire de six individus, alors les données sont invariantes par permutation, et les modèles aussi. Il est possible de travailler dans $\mathcal{E} = \mathbb{R}^6$, mais il serait plus juste de travailler dans l'espace des histogrammes.

D'autres hypothèses, notamment certaines distances, conduisent à des problèmes de modélisation triviaux. Ainsi, il est vain d'utiliser un espace de variables aléatoires comme espace de modèles $\mathcal{M}$ lorsque la similarité est une distance échantillon-à-échantillon. Plus précisément, on a le résultat suivant.

Proposition 2.2.1. *On suppose que l'on observe une loi empirique* $\mathrm{d}\mathbb{P}_Y \in \mathcal{E}$*, c'est-à-dire un élément de l'ensemble* $\mathcal{E}$ *des lois de probabilité sur un espace* $\mathcal{Y}$*. L'ensemble des modèles* $\mathcal{M}$ *est l'ensemble des variables aléatoires à valeurs dans* $\mathcal{Y}$*. On choisit la similarité entre* $\mathrm{d}\mathbb{P}_Y \in \mathcal{E}$ *et un modèle* $M \in \mathcal{M}$ *définie par l'espérance* $\mathbb{E}[d(M,Y)]$ *où* $d : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$ *est une distance. Alors*

$$\mathbb{E}[d(M,Y)] \geq \inf_{m \in \mathcal{Y}} \int d(m,y)\, \mathrm{d}\mathbb{P}_Y$$

En particulier, si le minimum est atteint, i.e. si il existe une valeur $m_0 \in \mathcal{Y}$ *telle que* $m_0 = \arg\min \int d(m,y)\, \mathrm{d}\mathbb{P}_Y$*, alors la masse de Dirac en* m_0 *est le modèle d'espérance minimale donc de similarité maximale.*

Démonstration. Ceci est une conséquence de l'indépendance nécessaire entre l'aléatoire d'un modèle et l'aléatoire du phénomène observé :

$$\begin{aligned}
\mathbb{E}[d(M,Y)] &= \int_{(m,y)\,\in\,\mathcal{Y}\times\mathcal{Y}} d(m,y)\, \mathrm{d}\mathbb{P}_{(M,Y)} && \text{(loi jointe)} \\
&= \iint d(m,y)\, \mathrm{d}\mathbb{P}_Y\, \mathrm{d}\mathbb{P}_M && \text{(par indépendance)} \\
&\geq \int \left[\inf_{m \in \mathcal{Y}} \int d(m,y)\, \mathrm{d}\mathbb{P}_Y \right] \mathrm{d}\mathbb{P}_M \\
&= \inf_{m \in \mathcal{Y}} \int d(m,y)\, \mathrm{d}\mathbb{P}_Y
\end{aligned}$$

En effet, on a minoré la fonction $m \mapsto \int d(m,y)\, \mathrm{d}\mathbb{P}_Y$ par la borne inférieure, ce qui élimine la variable m. Si cette borne inférieure est atteinte en m_0, alors

$$\mathbb{E}[d(M,Y)] \geq \int d(m_0,y)\, \mathrm{d}\mathbb{P}_Y = \iint d(m,y)\, \mathrm{d}\mathbb{P}_Y \mathrm{d}\mathbb{P}_{\delta_{m_0}} = \mathbb{E}[d(\delta_{m_0},Y)]$$

□

Plusieurs remarques dérangeantes découlent de cette proposition :

— en général, la loi empirique $\mathrm{d}\mathbb{P}_Y$ est fournie par un échantillon fini d'éléments de $\mathcal{Y}$. La proposition peut alors être démontrée sans utiliser l'indépendance. Cependant, la formulation ci-dessus est plus générale, et correspond à la tentative de modéliser une variable aléatoire Y connue par sa loi empirique.

— au sujet de l'indépendance : en pratique, c'est le générateur aléatoire de l'ordinateur qui va être utilisé comme source d'aléa pour un modèle prédictif. Quelle que soit son implémentation, il est supposé indépendant du phénomène étudié.

— dans ce cadre, la modélisation est un problème trivial car il est inutile de concevoir ou d'envisager des restrictions de l'ensemble des modèles, c'est-à-dire de faire des hypothèses de modèlisation. De la même façon, la recherche du modèle optimal est relativement simple, puisqu'il suffit d'explorer l'espace $\mathcal{Y}$ à la recherche d'un barycentre.
— sous des conditions assez générales, c'est la masse de Dirac qui est le modèle optimal, c'est-à-dire une variable aléatoire constante. C'est bien l'utilisation d'une variable *aléatoire* qui est inutile ; toute l'information est contenu dans le comportement "moyen" des données.
— on retrouverait le même type de résultat avec des paramètres ou des variables explicatives θ. Si on cherche une variable aléatoire qui minimise $\mathbb{E}[d(M, Y)|\theta]$, on retrouve un optimum qui est une variable aléatoire constante, plus précisément l'espérance conditionnelle $\mathbb{E}[Y|\theta]$ pour une distance L_2.

Appliquons la proposition ci-dessus aux données $(0, 0, 0, 1, 1, 1)$, en tant qu'échantillon de taille 6 d'une variable aléatoire à valeur dans $\mathcal{Y} = \mathbb{R}$. On utilise donc l'ensemble de modèles $\mathcal{M}$ constitué des variables aléatoires à valeurs réelles. On choisit la similarité définie par l'espérance

$$\mathbb{E}[(M - Y)^2] = \frac{3}{6}\mathbb{E}[(M - 0)^2] + \frac{3}{6}\mathbb{E}[(M - 1)^2]$$

D'après la proposition 2.2.1, la masse de Dirac $\delta_{1/2}$ réalise le minimum de cette similarité, avec $\mathbb{E}[(\delta_{1/2} - Y)^2] = \frac{3}{12}$. Par conséquent, le meilleur prédicteur au sens de la distance L_2 pour la prochaine valeur est la masse de Dirac $\delta_{1/2}$.

On s'attendait à trouver une variable de Bernoulli. Admettons que l'on s'est trompé d'espace de données et considérons maintenant des variables aléatoires à valeur dans $\mathcal{Y} = \{0, 1\}$. La même similarité s'écrit en fonction du paramètre p du modèle :

$$\begin{aligned}\mathbb{E}[(M - Y)^2] &= \frac{3}{6}\mathbb{E}[(M - 0)^2] + \frac{3}{6}\mathbb{E}[(M - 1)^2] \\ &= \frac{3}{6}\left[p.(1 - 0)^2 + (1 - p).(0)^2 + p.(0)^2 + (1 - p).(1 - 0)^2\right] \\ \mathbb{E}[(M - Y)^2] &= \frac{3}{6}\end{aligned}$$

La distance L_2 ne permet pas de distinguer les variables aléatoires de Bernoulli sur cet exemple. La proposition s'applique et les masses de Dirac δ_0 et δ_1 sont bien optimales.

Ces effets sont intimement liés au caractère échantillon-à-échantillon de la distance choisie. Si l'on définit la similarité entre $M \in \mathcal{M}$ et $\mathrm{d}\mathbb{P}_Y$ par une distance entre lois de probabilités, la distance de Kullback-Leibler par exemple,

alors c'est $M = \mathrm{d}\mathbb{P}_Y$ qui est le modèle optimal. Un autre exemple de ce phénomène est présenté dans le chapitre 7 section 7.3.1 pour la caractérisation de la couleurs des saignements dans les vidéos de coloscopie.

2.3 Énoncés scientifiques et preuves

Pour des énoncés provenant des applications des mathématiques, il n'est pas possible d'utiliser la théorie de la démonstration pour aller logiquement des axiomes aux conclusions. De fait, il n'y a pas d'axiomes, mais plutôt un ensemble de dogmes susceptibles d'être remis en cause. De plus, la présence d'incertitude sur les observations se traduit au minimum par une incertitude sur chacune des étapes de la preuve et sur sa conclusion. Par conséquent, on est obligé de se poser la question de la nature et de la force des preuves apportées par une analyse mathématique des données, et, en préalable, la question de la nature des énoncés scientifiques qu'il s'agit de démontrer.

Remarquons tout d'abord que la donnée d'un seul point, sans répétition, est suffisante pour apporter de l'information. En effet, on a vu que la similarité entre modèles et données induit une relation de compatibilité et une relation d'ordre sur l'ensemble des modèles. Il semble donc que les énoncés scientifiques que l'on peut formuler soient en général

— des énoncés d'égalité : le phénomène étudié correspond à une certaine valeur, plus ou moins précisément,
— des énoncés d'appartenance : le phénomène étudié correspond à un ensemble de valeurs cohérent, possédant les mêmes propriétés.

Il est à noter que l'égalité et l'appartenance sont parfois obtenues après application d'une fonction, qui peut être un estimateur paramétrique, ou bien une fonction indicatrice.

À ces deux types d'énoncés, il me semble que la preuve est apportée par l'accumulation d'observations concordantes. Ainsi, concernant le jeu de données $(0, 0, 0, 1, 1, 1)$, l'observation répétée de valeurs entre 0 et 1 supporte un énoncé tel que "la vraie valeur est 0,5". Ou bien, l'observation répétée de valeurs positives supporte un énoncé tel que "le phénomène est toujours positif". C'est ce type d'argument qui est utilisé au chapitre 5 section 5.6 pour justifier par l'accumulation d'exemples que "la distribution des patchs est constituée de clusters séparés".

Certains énoncés ne nécessitent pas l'utilisation de modèles, notamment parce que l'égalité ou l'appartenance peuvent être exprimées directement en fonction des observations. C'est le cas de la distribution des patchs évoquée ci-dessus, mais également du biais de répartition des lésions (cf section 7.6.2).

Dans d'autres cas, pour savoir si "le signal est croissant" par exemple, il est nécessaire de passer par un modèle pour s'affranchir notamment du bruit ou autres distortions, ou bien parce que l'énoncé lui-même a été défini par rapport à un modèle. Ainsi, on a utilisé le taux de croissance du modèle de régression exponentielle pour l'analyse du stress oxydant dans le chapitre 6.7, et pour biais de répartition des lésions dans le chapitre 7.6.3.

De la même façon, une "preuve" à un énoncé scientifique peut être apportée par accumulation de modèles concordants, à ceci près que la relation de similarité entre modèles et données est à la fois imparfaite et inconnue. Une justification de cette méthode est que le vrai modèle est supposé similaire aux données. Si tous les modèles similaires vérifient une propriété, alors le vrai modèle doit lui aussi vérifier cette propriété.

Par accumulation de modèles, considérons $(0,0,0,1,1,1)$ et tentons de démontrer l'énoncé "le signal est croissant". Il s'agit de montrer que tous les modèles similaires aux données sont des modèles croissants. On peut remarquer d'emblée que cela n'est pas possible dans n'importe quel ensemble de modèles. Ainsi, dans l'ensemble des fonctions constantes, aucun modèle ne vérifie l'énoncé. De même les polynômes de degré exactement 2 ne sont pas des fonctions croissantes. Par contre, si l'on considère l'ensemble $\mathcal{M}$ des polynômes de degré 1, c'est-à-dire les modèles linéaires, alors on peut se baser sur la pentes des modèles pour justifier cet énoncé.

On définit les modèles α-compatibles comme au chapitre 7.5.2 comme étant l'ensemble des $(a,b) \in \mathbb{R}^2$ tels que $\mathbb{E}[|aS+b-Y|^2] \leq \alpha \mathbb{E}[|\hat{a}x+\hat{b}-Y|^2]$ où $(\hat{a},\hat{b})$ correspond à la régression linéaire.

Code chunk 1 : «modeles_alpha-compatibles»

```
def compatible_linmodels(x,y,alpha,candidates):
    # Return array containing alpha-compatible slopes among candidate slopes
    best_slope, best_intercept, _,_,_ = stats.linregress(x,y)
    best_dist = np.std(y - best_slope * x - best_intercept)
    return np.sort(np.append([ s for s in candidates
                    if np.std(y - s*x) < (1+alpha)*best_dist], best_slope))

x = np.array([0,1,2,3,4,5]).transpose()
y = np.array([0,0,0,1,1,1]).transpose()
compatible_linmodels(x,y,0.05,np.arange(0,0.5,1e-2))
```

Interpret with python2

```
array([0.22      , 0.23      , 0.24      , 0.25      , 0.25714286,
       0.26      , 0.27      , 0.28      , 0.29      , 0.3       ])
```

Ceci montre que tous les modèles 0.05-compatibles ont une pente comprise entre 0.22 et 0.3. Ce sont tous des modèles de pente positive, donc le vrai modèle aussi, ce qui justifie l'énoncé "le signal est croissant". Une notion similaire est utilisée au chapitre 7.6.4 entre les solutions de l'équation de Fisher-KPP et les données recueillies sur des patients.

Plus généralement, l'apport de preuves par accumulation de modèles consiste à justifier que l'énoncé est vérifié dans l'ensemble de modèles le plus grand possible :

— pour tous les modèles compatibles d'un ensemble de modèles $\mathcal{M}$ fixé, ici l'ensemble des modèles linéaires,
— pour tous les ensembles de modèles possibles, c'est-à-dire sous des hypothèses de modélisation les moins restrictives possibles,
— pour toutes les similarités envisageables, la vraie similarité aux données étant inconnue.
— etc.

Enfin, remarquons que la définition des ensembles $\mathcal{E}$ et $\mathcal{M}$ et de leur lien de similarité induit une restriction sur l'ensemble des énoncés scientifiques démontrables. Par exemple, si deux modèles ont la même vraisemblance par rapport aux données, ils sont indistinguables ou dit autrement, ils appartiennent à une même classe d'équivalence. Ceci rejoint la notion de "shattering" dans la théorie de la complexité des modèles de Vapnik-Chervonenkis [94], c'est-à-dire que les données ont une certaine complexité, qui peut être évaluée par le nombre de partitions différentes engendrées par un ensemble de modèles. Cette restriction opère également si l'ensemble des modèles est trop petit ou trop homogène. Ainsi, lorsque l'on utilise des équations différentielles pour modéliser un phénomène, les énoncés démontrables ne concerneront que des fonctions continues. En dehors de la question de savoir justifier une hypothèse de modélisation, il apparaît donc aussi la question de déterminer quels énoncés restent démontrables étant donné un ensemble de modèles $\mathcal{M}$.

2.4 Apprentissage et construction des énoncés scientifiques

La démonstration d'un énoncé scientifique suppose que cet énoncé est connu et défini précisément. Cependant, en pratique, beaucoup de sujets sont connus de façon imprécise et la plupart des énoncés ne sont pas bien formalisés. Par exemple, dans les travaux présentés au chapitre 7 sur la répartitition spatiale des lésions de la muqueuse intestinale, on s'est intéressé à l'accumulation de lésions dans la région du rectum. En l'absence de définition précise de la no-

tion d'accumulation, il a été choisi de traduire cela par une densité de lésions croissante en fonction de l'abscisse curviligne.

Ce travail d'élaboration des énoncés scientifiques est souvent laissé à l'expertise du modélisateur. Dans cette partie, nous envisageons deux stratégies alternatives de construction des énoncés scientifiques. Tout d'abord, les énoncés étant construits comme des propositions du type égalité à une valeur ou du type appartenance à un sous-ensemble, on peut remarquer qu'il y a un nombre limité d'énoncés "potentiellement intéressants" parmi le nombre exponentiel de sous-ensembles de l'espace des données $\mathcal{E}$ ou de l'espace des modèles $\mathcal{M}$. Par exemple, dans un espace de données discret, on peut tenter d'énumérer toutes les propositions logiques. En particulier, on peut tenter d'énumérer les implications formulables d'après un jeu de données, c'est-à-dire les potentiels liens de cause à effet.

Par exemple, si l'espace des données $\mathcal{E}$ ou l'espace des modèles $\mathcal{M}$ est l'ensemble des matrices réelles carrées, donc des éléments de $\mathbb{R}^{n\times n}$ où n est le nombre de lignes ou de colonnes, on pourra s'intéresser :

— aux ensembles de matrices de rang fixé k,
— à l'ensemble des matrices symétriques,
— à l'ensemble des matrices de transition,
— à l'ensemble des matrices triangulaires, etc.

Une première stratégie de construction automatique des énoncés scientifiques pourrait donc consister à

— énumérer les propriétés communes à toutes les répétitions du jeu de données, ou du moins les propriétés habituelles
— énumérer les modèles compatibles et leurs propriétés communes.

Ainsi la modélisation consiste à choisir les conditions ou le cadre d'analyse dans lequel l'analyse de données permet d'apporter des preuves aux questions des applications.

À l'inverse, l'apprentissage automatique effectue la construction d'un certain énoncé scientifique. Par exemple, l'apprentissage d'une classification revient à construire une partition de l'espace des données, et donc à construire des ensembles de données et de modèles possédant les mêmes propriétés. De même, l'apprentissage d'un modèle de régression revient à construire un énoncé d'égalité à une valeur. Souvent, le modèle obtenu ou les paramètres estimés sont utilisés pour construire la région de rejet d'un test statistique, c'est-à-dire le complémentaire de l'ensemble des modèles compatibles avec les données.

Considérer l'apprentissage automatique comme la construction d'un énoncé scientifique particulier permet de mettre en évidence un défaut majeur de l'ap-

prentissage : cette approche est focalisée sur un seul énoncé, et il est difficile d'envisager l'ensemble des explications possibles. Dans la suite de nos travaux, nous avons préféré la démarche consistant à déterminer l'ensemble des modèles compatibles, plutôt que d'estimer le meilleur modèle.

2.5 Exemple d'utilisation de la notion de modèle compatible

Durant ma thèse de doctorat, je me suis intéressé à un problème d'estimation paramétrique, que j'ai résolu en proposant une notion ad-hoc de modèles compatibles dans [2]. Dans cet exemple, on peut calculer explicitement l'ensemble des modèles compatibles et en déduire un estimateur optimal.

Dans un spectromètre de masse à haute résolution, on compte le nombre d'ions $X \in \mathbb{N}$ qui atteignent la surface d'un détecteur. Cependant, à cause de l'électronique utilisée et d'un facteur de gain τ inconnu, le signal observé Y est obtenu après troncature, i.e. sous la forme $Y = \lfloor \tau X \rfloor$. On souhaite retrouver la valeur de τ et la loi de X.

En l'absence de troncature, si on observe l'événement $\{X = 1\}$, on obtient $\tau = \min Y$, et on peut en déduire la variable cachée X. Dans le cas général, le problème est identifiable si et seulement si $\tau > 1$. Dans ce cas, on peut alors associer à τ le sous-ensemble $\mathcal{S}_\tau = \lfloor \tau \mathbb{N} \rfloor \subset \mathbb{N}$. Comme la relation est bijective, on peut là encore obtenir la variable cachée X à partir de τ.

On s'intéresse donc au paramètre τ et l'on considère l'ensemble de modèles $\mathcal{M} =]1; +\infty[$, ou de façon équivalente l'ensemble des grilles $\mathcal{S}_\tau$. À cause des valeurs manquantes, une observation Y est une partie de $\mathbb{N}$ incluse dans la grille $\mathcal{S}_{\tau_0}$ correspondant à la vraie valeur τ_0. L'ensemble $\mathcal{E}$ est constitué des parties de $\mathbb{N}$.

L'ensemble des modèles compatibles est constitué des grilles $\mathcal{S}_\tau$ contenant Y, i.e. $Y \subset \mathcal{S}_\tau$ ou des valeurs de τ associées. Ainsi, un modèle est compatible s'il peut "générer" les observations, ce qui correspond à la première notion de compatibilité introduite dans la section 2.1. Cette notion ne nécessite pas la définition d'une distance entre observations et modèles, et ni de notion de distribution sur l'ensemble des modèles. La figure 2.1 présente le jeu de données $Y = \{2, 3, 5, 6, 7, 11, 13\}$, ainsi quelques exemples de grilles $\mathcal{S}_\tau$ compatibles.

Dans [2], on montre que l'ensemble des modèles compatibles est formé d'une union finie d'intervalles, que ces intervalles ont une largeur minimale $\frac{1}{(\max Y)^2}$ et que l'ensemble est majoré par $\frac{\max Y + 1}{n}$ où n est le nombre d'entiers distincts

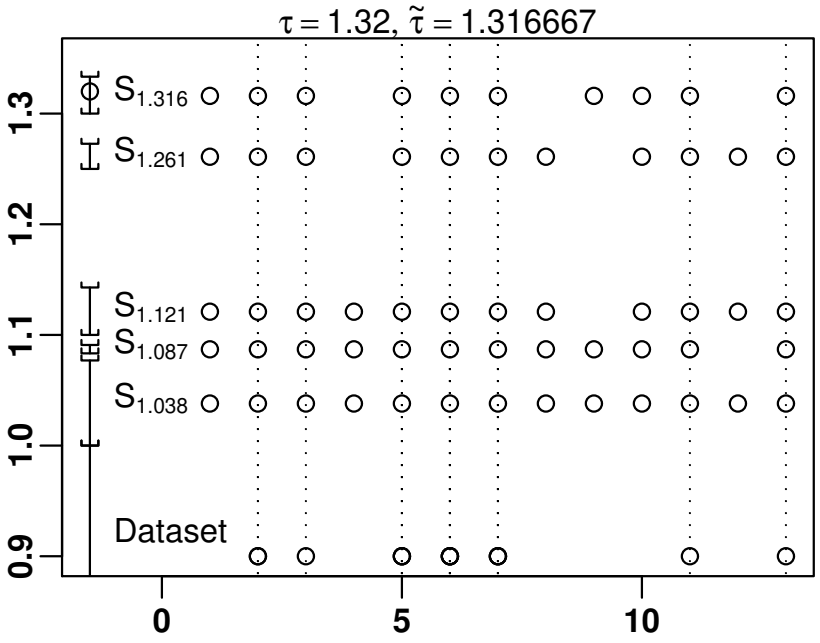


FIGURE 2.1 – Estimation du paramètre τ. Les valeurs observées Y sont indiquées par des cercles en bas de la figure. L'ensemble des valeurs compatibles est l'union des intervalles dessinés à gauche. La valeur estimée $\hat{\tau} = 1.316$ est le milieu de l'intervalle contenant les plus hautes valeurs. Pour chaque intervalle, on a dessiné la grille correspondante et les pointillés indiquent l'inclusion de Y dans chacune des grilles.

observés. De plus, la vraie valeur τ_0 appartient à l'intervalle I contenant les plus grandes valeurs. Ce dernier résultat peut être interprété comme un principe de parcimonie, car I contient les valeurs de τ minimales explicatives, c'est-à-dire les grilles $\mathcal{S}_\tau$ compatibles avec le nombre minimal d'éléments superflus.

On en déduit un algorithme d'estimation de τ_0 : on cherche la plus grande valeur compatible en commençant par la borne supérieure $\frac{\max Y+1}{n}$ et en descendant par pas de $\frac{1}{(\max Y)^2}$. La valeur obtenue permet alors de reconstruire l'ensemble des modèles compatibles, en particulier l'intervalle I, et de calculer son milieu. Dans [2], on montre que cette méthode est robuste dans le cas de données manquantes. De plus, même en l'absence de données manquantes, elle est plus précise que d'autres estimateurs basés sur la régression linéaire, ou sur la transformée de Fourier en interprétant Y comme un ensemble quasiment périodique.

Dans cet exemple, le paramètre τ n'affecte que le support des observations. L'estimation de τ permet de découpler le paramètre de "gain électronique" de la loi du processus de comptage des ions X. L'ensemble des valeurs compatibles représente donc la totalité de l'information sur τ contenue dans les données, ce qui explique la performance de l'estimateur obtenu.

De tels résultats peuvent être obtenus car l'effet de la troncature a une structure un peu spéciale. Son caractère discret permet de définir une notion de compatibilité forte entre la grille observée Y et les grilles modèles $\mathcal{S}_\tau$. Cette approche perd de son intérêt lorsque le bruit mélange les valeurs de X. Il faut alors utiliser la loi de X dans la méthode d'estimation, et passer par une notion de compatibilité faisant intervenir une distance entre modèles et observations, ou une loi de probabilité.

Comme évoqué dans la section 2.3, on peut procéder par accumulation de modèles pour justifier un énoncé scientifique. Ici, cette méthodologie est explicite. Les énoncés scientifiques étant assimilés à des parties de l'ensemble des modèles, il s'agit dans cet exemple de parties de $\mathcal{M} =]1; +\infty[$. Les énoncés justifiés par une observations Y sont donc les parties de $\mathcal{M}$ contenant l'ensemble des valeurs compatibles, i.e. l'union finie d'intervalles. En suivant le principe de parcimonie, on peut restreindre l'ensemble des valeurs compatibles à l'intervalle I. L'information étant plus précise, cela permet de justifier des énoncés scientifiques plus précis correspondant à des ensembles de modèles plus petits.

2.6 Conclusion

Ce chapitre propose un ensemble de réflexions pour guider la démarche d'analyse de données et l'apport de preuves à des énoncés scientifiques dans les applications des mathématiques. En premier lieu, il faut définir l'espace des données $\mathcal{E}$, l'espace des modèles $\mathcal{M}$ et le lien de similarité qui les relie, ce qui revient à fixer les hypothèses de modélisation et l'ensemble des énoncés scientifiques démontrables.

Ensuite, à des énoncés scientifiques précis, déjà formalisés par une définition mathématique, les preuves sont apportées par accumulation d'observations répétées, ou par l'accumulation de modèles. Ce principe a guidé les analyses présentées au chapitre 7 pour les vidéos de coloscopie obtenues dans les maladies inflammatoires chroniques de l'intestin.

Enfin, le travail du modélisateur consiste aussi à "estimer des énoncés" ou préciser des ensembles. Par exemple, dans le chapitre 5, on souhaite déterminer le sous-ensemble des patchs correspondant à des images naturelles ou médicales

parmi l'ensemble des patchs. Dans le chapitre 7.3, on souhaite déterminer le sous-ensemble des couleurs correspondant aux lésions de saignement ou d'ulcère dans les vidéos de coloscopie.

Ce cadre général permet de mieux comprendre les travaux présentés par la suite, et en particulier d'évaluer et de discuter la qualité des preuves apportées. Il facilite la discussion des résultats obtenus, et indique un certain nombre de directions naturelles dans lesquelles pourraient se poursuivre les travaux.

Chapitre 3

Recherche reproductible

Pendant le travail effectué durant ma thèse à l'Institut Pasteur et mon post-doctorat à l'INRA où j'ai travaillé en étroite collaboration avec les applications, j'ai pu constater la difficulté de réutiliser les travaux et outils que j'ai produits. En arrivant sur mon poste de maître de conférences à l'Université Paris 13, j'ai décidé de m'attaquer au problème, avec la conviction que la ré-utilisation de travaux précédents est une problématique de recherche, qui concerne à la fois la réutilisation de ses propres travaux à une date ultérieure et à la fois la réutilisation de travaux de collaborateurs, ou ceux publiés dans la littérature.

Le travail réalisé dans cette thématique a donné lieu au développement du logiciel Lepton [16], à une publication de journal [17], deux exposés en conférence internationale avec actes [8, 9], l'organisation d'un workshop [10] et plusieurs exposés en conférences nationales.

Ce travail n'a pas de lien direct avec la méthodologie générale de modélisation présentée dans le chapitre 2, mais découle de la démarche plus générale ayant pour objectif une analyse de données transparente, telle qu'évoquée dans le chapitre 1. Ce travail a également été mené dans le but de faciliter la collaboration et la ré-utilisation des travaux menés avec les étudiants que j'ai encadrés. S'il n'a pas été possible d'adopter Lepton dans les projets de thèse de Tran Dai Viet et de Tran Duc Nghia, ce logiciel est employé dans le cadre de la thèse de Safaa Al Ali et dans tous mes propres travaux, y compris ce manuscrit.

Dans le domaine de la recherche reproductible [22], on définit habituellement quatre niveaux de reproductibilité :

— **view**, c'est-à-dire la possibilité d'obtenir et de consulter le travail,
— **learn**, c'est-à-dire la disponibilité d'une documentation complète, permettant de comprendre le travail
— **execute**, c'est-à-dire la facilité de reproduire les mêmes résultats (en

exécutant les programmes),
— **modify**, c'est-à-dire une compréhension fine permettant de modifier et d'appliquer des programmes pour de nouveaux travaux.

Ce chapitre discute des différentes approches actuelles pour ces différents niveaux de reproductibilité, ainsi que des solutions logicielles correspondantes. Les fonctionnalités proposées par Lepton sont illustrées par un exemple complet (cf page 34) extrait de la documentation du logiciel.

3.1 Disponibilité de tous les éléments

La première propriété à imposer pour un travail que l'on souhaite reproductible concerne donc la disponibilité de l'ensemble des éléments nécessaires : données, codes sources, documentation, scripts, figures, etc.

Cette disponibilité est d'abord un problème d'archivage, et les premières solutions proposées sont des méthodes ou des infrastructures de stockage pour les données de la recherche. On peut penser par exemple aux grandes bases de données de génomique (GenBank, EMBL) et de protéomique (UniProt, Swiss-Prot, NCBI) qui ont été développées à partir des années 80 [1]. À une échelle plus modeste, les données de la recherche sont souvent conservées sous la forme d'archives compressées, et des services tels que Zenodo [2] ou Figshare [3] se sont développés pour faciliter l'accès à ces fichiers.

En plus de la problématique de l'archivage, on peut se poser la question de la conservation des versions successives des données. Cette question a été beaucoup travaillée dans le cas des codes sources et des logiciels, et il existe des systèmes de contrôle de version tels que CVS [4], subversion [5], ou GIT [6]. Ces outils sont adaptés pour les codes sources, mais pas pour les images ou les fichiers binaires.

Les bases de données et systèmes d'archivage fournissent des solutions techniques à la disponibilités des données de la recherche, mais ne cherchent pas à savoir quelles données doivent être mises à disposition. Une avancée récente dans cette direction est le développement de systèmes d'intégration continue

1. Une liste exhaustive des bases de données biologiques figure sur Wikipedia `https://en.wikipedia.org/wiki/List_of_biological_databases`
2. `https://zenodo.org/`
3. `https://figshare.com/`
4. `http://cvs.nongnu.org/`
5. `https://subversion.apache.org/`
6. `https://git-scm.com/`

tels que GitLab [7], qui automatisent la compilation et le déploiement d'un code source, et rendent obligatoire la présence de l'ensemble des éléments nécessaires (appelés *assets* dans ce domaine) à la production d'un package complet.

Dans le cas de Lepton, il est recommandé dans la mesure du possible de construire un fichier unique au format texte, qui rassemble le maximum d'éléments. En particulier, un fichier Lepton contiendra à la fois de la documentation et du code source, par conséquent des contenus hétérogènes écrits dans des langages différents. Lepton rend cela possible (et lisible) en implémentant l'approche "literate programming" proposée par D. Knuth [61, 78]. Ainsi, un fichier Lepton est un fichier de documentation dans un format basé texte [8], contenant des blocs de contenu arbitraire. Par exemple :

Code chunk 2 : `«exemple»`

```
documentation au format texte
<<titre1>>=
code source
@
<<titre2>>=
données
@
```

3.2 Documentation complète

Le deuxième niveau de reproductibilité nécessite un effort de documentation. En effet, c'est l'auteur humain d'un projet qui est responsable de sa présentation sous une forme compréhensible par le lecteur. L'apport d'un logiciel tel que Lepton consiste à faciliter un certain nombre de propriétés :

- des fonctionnalités pour la documentation : liens hypertexte, coloration syntaxique, formules mathématiques, etc.
- le rapprochement du code et la documentation correspondante,
- la réorganisation du code source et de la documentation en fonction du lecteur et non des contraintes du format informatique.

Le choix d'un format de documentation basé texte correspond au premier item. En particulier, tout en restant assez lisible, le format LaTeX permet de produire des documents riches. La coloration syntaxique des blocs de code source, habituellement absente de la distribution standard de LaTeX, est ici apportée

7. `https://about.gitlab.com/`
8. par exemple LaTeX, HTML, Markdown

par le package `minted`[9], qui utilise la librairie spécialisée Pygments [30] écrite en Python.

Le logiciel Lepton permet de manipuler le code source et la documentation de plusieurs façons :

— On peut découper le code source en petits blocs, et Lepton se chargera de l'assemblage, de l'écriture des fichiers sur le disque dur, de la compilation et de l'exécution des instructions.
— Un mécanisme de "référence" permet de définir un bloc à un endroit, et à l'utiliser à un autre en indiquant son `<<titre>>`.
— Ce mécanisme permet de simplifier et de donner une vision globale de l'architecture du code, en remplaçant des portions du code par leur titre.
— Ce mécanisme permet réutiliser un même bloc à plusieurs endroits du document, pour synchroniser et/ou éviter de dupliquer des portions de code.
— On peut mélanger plusieurs fichiers lorsque cela est utile, par exemple pour comparer plusieurs implémentation d'une même fonction.

Ces fonctionnalités sont illustrées dans l'exemple complet présenté page 34. En particulier, en langage C, il est nécessaire de sauvegarder le fichier avant de le compiler pour en faire un exécutable, alors que les instructions en Python et en OCaml sont directement envoyées à un interpréteur. Le code source de la fonction `fibC` n'est défini qu'une seule fois page 36, et référencé dans la fonction `main.c` page 36 et dans le programme `benchmark.c` page 39.

Pour comparaison, l'autre approche de la documentation consiste à insérer des commentaires dans des fichiers source. Un logiciel, Doxygen[10] par exemple, est alors chargé d'extraire les commentaires, et de produire la documentation. Un tel système peut fournir des fonctionnalités équivalentes pour la documentation en permettant d'écrire les commentaires en LaTeX ou en Markdown. Cependant, il est obligatoire de respecter l'architecture du code telle que définie par le langage de programmation, l'organisation en modules, l'ordre de définition des fonctions, l'indentation, etc. Il n'est pas possible de comparer des fichiers ou de réutiliser des blocs. Un système tel que Doxygen n'est donc pas adapté pour des travaux de documentation dont la portée dépasse un code source particulier, tandis que Lepton pourra être utilisé pour des rapports ou des articles de recherche.

9. `https://www.ctan.org/pkg/minted`
10. `https://www.doxygen.nl/index.html`

3.3 Exécution et reproduction des résultats

Un troisième niveau de reproductibilité concerne la possibilité de reproduire les résultats présentés, en considérant l'approche comme une boîte noire, c'est-à-dire sans forcément comprendre comment les résultats ont été obtenus. De nombreux travaux récents ont eu lieu dans ce domaine, avec le développement de solutions techniques et de logiciels [29, 45, 51, 63].

Dans un premier temps, il faut rassembler les informations sur l'environnement dans lequel les résultats ont été produits. Dans les sciences expérimentales, cela correspond à la description des produits, des machines qui ont été utilisées ainsi que leur paramétrage. Pour des approches informatiques, il s'agit des versions du système d'exploitation, des logiciels et des librairies utilisés, les scripts ainsi que tous les paramètres. Ces informations sont regroupées sous le terme de "provenance", et un certain nombre de logiciels ont été développés afin d'automatiser leur collecte [45, 51, 63]. Par exemple, l'approche VisTrails [63] va jusqu'à enregistrer la liste des appels de fonctions utilisés dans le code.

Étant donnée la description de l'environnement logiciel, il faut pouvoir reproduire cet environnement, c'est-à-dire que les librairies puissent être ré-installées et re-configurées pour relancer les calculs. Une solution consiste à enregister une copie de l'ordinateur qui a effectué le calcul sous la forme d'une machine virtuelle, qui pourra être téléchargée par un utilisateur et réutilisée pour relancer le calcul [29]. Cette solution technique est simple et complète, mais lourde car la machine virtuelle doit contenir une copie de tous les logiciels y compris ceux qui ne sont pas utilisés, le système d'exploitation par exemple. De plus, on perd un peu de puissance de calcul dans la virtualisation.

Pour faire plus léger, il faut être capable de n'installer que les logiciels nécessaires, tout en isolant ces logiciels des autres versions qui pourraient être déjà installés. Deux tendances actuelles soutiennent cette démarche. D'une part, un nombre croissant de logiciels sont disponibles en version "portable" [11], et peuvent être utilisés sans installation, c'est-à-dire sans modification du système hôte. Ils encapsulent alors l'ensemble des librairies dont ils ont besoin pour fonctionner, ainsi que l'ensemble des données de l'utilisateur.

D'autre part, notamment pour des raisons de sécurité, les systèmes d'exploitation récents séparent de plus en plus les logiciels qui s'exécutent sur une machine. Sous Linux, le système des *containers* fournit des "boîtes" dans lesquelles on peut installer et faire exécuter des logiciels sans dupliquer le système d'exploitation. Des logiciels comme Docker automatisent l'installation et le dé-

11. `https://en.wikipedia.org/wiki/List_of_portable_software`

ploiement d'un logiciel dans un container. Ce mécanisme est notamment utilisé par Gitlab pour l'intégration continue évoquée dans la section 3.1 : Gitlab déploie une image Docker contenant l'environnement, y compris le compilateur et les librairies, afin de compiler automatiquement un code source.

Enfin, il faut pouvoir exécuter toutes les commandes, et en premier lieu que les scripts utilisés soient disponibles. La difficulté réside moins dans la disponibilité des informations que dans la facilité à les employer. Dans l'idéal, il suffirait de taper une commande ou de cliquer sur un bouton. En pratique, on est souvent confronté à des situations similaires à la lecture d'un tutoriel sur une page web : il faut copier-coller toutes les commandes à la main.

Dans la mesure où les machines virtuelles sont une solution complète, Lepton prend le parti de ne traiter que l'exécution des commandes. Ainsi, les blocs contenus dans un fichier Lepton peuvent contenir des commandes shell, des instructions Python, etc. qui seront prises en charge et exécutées par Lepton. De fait, Lepton opère comme un interpréteur de commande, et peut être utilisé avec le mécanisme du `#!`. Il est donc possible de rassembler de la documentation, du code source, des données et des scripts dans un fichier Lepton `rapport.nw`, et de reproduire l'ensemble du travail avec une seule commande (cf 3.6) :

Code chunk 3 : `«exemple (part 2)»`

```
./rapport.nw
```

Un des avantages de Lepton par rapport à d'autres système est qu'il est possible de combiner des scripts correspondants à des logiciels ou langages informatiques différents, i.e. de choisir l'outil le mieux adapté pour une tâche.

En fait, l'exécution des commandes ouvre de nombreuses possibilités, et peut permettre de prendre tout en charge dans un fichier Lepton, y compris l'installation et la configuration de l'environnement de calcul :

- les fichiers nécessaires sont ré-assemblés et écrits sur le disque dur,
- l'installation des librairies peut être lancée par des commandes shell,
- la compilation des codes sources peut également être lancée par des commandes shell,
- on peut lancer les calculs, ainsi que les scripts générant des figures,
- enfin, les commandes de compilation d'un rapport (LaTeX) peuvent aussi figurer dans le même fichier.

Lepton est écrit et distribué de cette façon. Un fichier unique `lepton.nw` contient à la fois le code source, la documentation. La commande `lepton.bin lepton.nw` suffit à extraire le code source, re-compiler un exécutable et produire la documentation LaTeX/PDF.

3.4 Inspection et réutilisation

Tous les éléments évoqués plus haut sont nécessaires pour pouvoir ré-utiliser efficacement un travail. Il faut que l'ensemble des éléments du travail soient disponibles, que l'on puisse ré-exécuter les calculs et que la documentation permette de comprendre comment fonctionnent les choses. Dans les sections précédentes, nous abordé le sujet du point de vue des fonctionnalités et des implémentations logicielles. Dans cette section, on présente comment ces fonctionnalités peuvent être organisées dans une approche globale, favorisant la compréhension du travail présenté et sa réutilisation.

Tout d'abord, la compréhension du travail présenté correspond à une compréhension humaine, par opposition à la capacité d'un programme informatique à extraire des informations du texte. Le travail est donc formaté pour un lecteur humain, plutôt que pour suivre la syntaxe d'un langage de programmation. Ceci requiert un effort de la part de l'auteur, et un certain nombre de conseils ont été proposés [31, 38, 55, 84] : utiliser un système de contrôle de version (Git ou SVN), faire de la science ouverte, partager les données, noter les pré-requis, structurer le texte, etc.

Ces recommandations font de plus en plus partie du processus editorial et de la publications des travaux de recherche. Les éditeurs imposent un certain nombre de règles et de bonnes pratiques, et les relecteurs vérifient leur application. Par exemple, pour une soumission à "Biometrical Journal", les données et les codes sources sont encouragées, même si ce n'est pas obligatoire. Les instructions [12] encouragent également l'utilisation de l'approche "literate programming" et les codes soumis sont vérifiés par un éditeur spécialisé. Dans IPOL [56], il est demandé de présenter les algorithmes utilisés dans un pseudo langage de programmation en plus du code source. Le journal n'accepte que les soumissions utilisant C/C++, Python ou Octave/Matlab, mais ceci donne la possibilité de fournir des démonstrations sur le site web. Dans "Journal of Open Source Software" [13], le processus de publication a lieu à l'aide d'un serveur GitHub, ce qui permet un dialogue public entre les auteurs et les relecteurs, ainsi que la possibilité de soumettre des modifications directement dans le processus de publication.

Si la soumission des articles de recherche se fait la plupart du temps au format LaTeX, une partie importante de la communauté s'est tournée vers des solutions de type "notebook" [50, 69, 72], dont la plus connue est Jupyter [60].

12. http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1521-4036/homepage/RR_Guideline.pdf
13. https://joss.theoj.org/

Pour simplifier, cela fonctionne comme un logiciel de calcul (Matlab, Maple, …) dans une page web. Comme dans Lepton et beaucoup d'autres approches, on peut définir des blocs de documentation et des blocs de code source, et les codes sources peuvent être exécutés. D'une certaine façon, Jupyter et Lepton offrent des fonctionnalités similaires, mais Jupyter choisit une approche dynamique et interactive. De nombreux logiciels de calcul offrent maintenant des fonctionnalités de type "notebook" : Matlab, R avec rmarkdown [79], knitR [99] ou RStudio [82], ou encore Emacs avec `org-mode`[42].

Lepton prend le parti inverse d'une approche statique qui apporte une garantie de cohérence aux résultats produits. En effet, dans un système interactif, l'utilisateur peut modifier l'ordre d'exécution d'un notebook, et en changer les résultats. Dans Lepton, on a choisi de rendre l'exécution indépendante de l'utilisateur. On a donc la garantie que les éléments présentés forment un ensemble cohérent. De plus, si l'utilisateur n'intervient pas dans la production des résultats, ceux-ci devraient être reproductibles par n'importe quelle autre personne.

Un autre parti pris de Lepton est que les approches de notebook telles que Jupyter sont des projets logiciels complexes, dépendants de technologies susceptibles d'évoluer. Par conséquent, cela affaiblit la préservabilité à long terme des résultats en introduisant une dépendance supplémentaire sur le logiciel de notebook, qui s'ajoute à toutes les dépendances de l'approche elle-même. A contrario, l'approche statique implémentée dans Lepton est suffisamment simple pour être reprogrammée de façon indépendante, avec les outils du moment.

3.5 Implémentation et disponibilité

Le logiciel Lepton a été écrit dans le langage OCaml. Le code source a été publié sur Github [14].

La dernière version stable peut être téléchargée sur Zenodo [15] [87], ainsi que les précédentes versions publiées.

14. https ://github.com/slithiaote/lepton
15. https ://doi.org/10.5281/zenodo.1311587

3.6 Exemple d'application : fibonacci.nw

On reproduit ici un exemple d'utilisation de Lepton fourni avec le code source. Pour des raisons de lisibilité, le fichier source `fibonacci.nw` n'est pas inclus en entier dans ce manuscrit, on en montre simplement quelques lignes ci-dessous.

Code chunk 4 : «exemple_fibonacci»

```
head -n 33 fibonacci.nw
```

Interpret with `shell`

```
\documentclass[a4paper,10pt]{scrartcl}
% \usepackage[T1]{fontenc}
\usepackage{lmodern}
\usepackage[english]{babel}
\usepackage{graphicx}
\usepackage[bindingoffset=0cm,height=21cm]{geometry}
\usepackage[moderate]{savetrees}
\usepackage{nopageno}

\usepackage{amsmath}
\usepackage{amsfonts}
\usepackage{amssymb}
\usepackage{amsthm}
\usepackage{fancybox}
\input{lepton.sty}

\begin{document}

\title{Comparison of a few implementations\\ of the Fibonacci sequence}
\author{Li-Thiao-Té Sébastien}
\maketitle

\section{Introduction}

In this document, we use Lepton to compare a few implementations of the
computation of the Fibonacci sequence (OCaml, Python and C). This is
intended to demonstrate Lepton's features, such as

\begin{itemize}
\item embedding source code inside a document (literate programming)
\item code restructuring for better documentation
\item embedding executable instructions
\end{itemize}
```

On compile ce document avec les commandes ci-dessous. Le fichier PDF produit est inclus à la fin de ce chapitre (pages 34 à 41).

Code chunk 5 : «exemple_fibonacci (part 2)»

```
lepton fibonacci.nw -o fibonacci.tex
pdflatex -shell-escape -interaction=batchmode fibonacci.tex | fold -sw 80
pdflatex -shell-escape -interaction=batchmode fibonacci.tex | fold -sw 80
# la commande fold permet de passer à la ligne au bout de 80 caractères
```

Interpret with `shell`

```
This is the Lepton/Lex implementation.
compilation (part 1):        chunk as text,
compilation (part 2):        chunk as text,
ocaml (part 1):        chunk as ocaml, exec with ocaml, output as text,
ocaml (part 2):        chunk as ocaml, exec with ocaml, output as text,
python (part 1):        chunk as python, exec with python, output as text,
python (part 2):        chunk as python, exec with python, output as text,
fibC (part 1):        chunk as c,
main.c (part 1 write):        chunk as c,
shell (part 1):        chunk as text, exec with shell, output as text,
shell (part 2):        chunk as text, exec with shell, output as text,
ocaml (part 3):        chunk as ocaml, exec with ocaml, output as text,
ocaml (part 4):        chunk as ocaml, exec with ocaml, output as text,
shell (part 3):        chunk as text, exec with shell, output as text,
python (part 3):        chunk as python, exec with python, output as text,
shell (part 4):        chunk as text, exec with shell, output as text,
benchmark.c (part 1 write):        chunk as text,
shell (part 5):        chunk as text, exec with shell, output as text,
plot.in (part 1 write):        chunk as text,
shell (part 6):        chunk as text, exec with shell, output as text,
This is pdfTeX, Version 3.14159265-2.6-1.40.21 (TeX Live 2020/Debian)
(preloaded format=pdflatex)
 \write18 enabled.
entering extended mode
/usr/bin/pygmentize
This is pdfTeX, Version 3.14159265-2.6-1.40.21 (TeX Live 2020/Debian)
(preloaded format=pdflatex)
 \write18 enabled.
entering extended mode
/usr/bin/pygmentize
```

N.B. Il n'y a pas d'intervention manuelle dans la génération du PDF à partir de `fibonacci.nw`, ni de script caché. Les commandes nécessaires sont présentées ci-dessus, et exécutées par Lepton. Le package LaTeX `pdfpages` permet alors d'inclure le fichier PDF dans ce document.

3.7 Conclusion et perspectives

La recherche reproductible est un mouvement qui nous invite à renouveler nos pratiques, afin de faciliter la réutilisation des résultats de la recherche. Un certain nombre de principes et de bonnes pratiques ont été théorisées ces dernières années et la démarche prend de l'ampleur dans les différentes communautés utilisant l'outil informatique pour la recherche. Ces pratiques s'appuient sur des logiciels nouveaux, en développement actif, qui sont encore susceptibles d'évoluer significativement dans les prochaines années.

Lepton est un logiciel initialement développé pour la recherche reproductible qui fait un certain nombre de choix différents par rapport aux pratiques antérieures et actuelles. Il a trouvé des applications dans la quasi-totalité de mes activités, que ce soit pour la programmation logicielle, l'analyse de données ou encore l'enseignement au travers de la production de sujets et corrigés aléatoires.

Parmi les nombreuses problématiques du domaine, il en est une qui est reste inabordée. Actuellement, tout est fait pour que les résultats de recherche soient reproductibles, mais rien ne garantit que ces résultats soient corrects. Dans le cadre d'un logiciel comme Lepton, la question est de savoir s'il est possible d'introduire des propriétés garantissant que le code source reste correct vis-à-vis de sa documentation, par exemple après la mise à jour d'une librairie. Une piste de recherche serait l'utilisation de signatures cryptographiques incorporant à la fois des éléments de la documentation, du code source et des résultats produits.

Comparison of a few implementations of the Fibonacci sequence

Li-Thiao-Té Sébastien

November 12, 2021

1 Introduction

In this document, we use Lepton to compare a few implementations of the computation of the Fibonacci sequence (OCaml, Python and C). This is intended to demonstrate Lepton's features, such as

- embedding source code inside a document (literate programming)
- code restructuring for better documentation
- embedding executable instructions

From the point of view of (scientific) applications, these features make it possible to

- include and distribute the actual source code
- distribute the instructions necessary for compiling and running the code
- certify that the embedded source code is correct by executing it
- running analysis scripts and generating figures
- certify that the figures correspond to the provided source code
- embedding different programming languages in the same document / same platform
- simplifying code re-use by distributing only a single file.

1.1 Executing this document

This document is a script and is intended to be executed to produce a PDF file, as well as by-products such as extracted source code or data files. It requires :

- Lepton to produce a `.tex` LaTeX document,
- LaTeX to process the `.tex` file into PDF,
- the Pygments library for syntax highlighting and the `minted` LaTeXpackage for code beautification.

To process the document with Lepton, the user should run the command :

Code chunk 1: ≪compilation≫

```
lepton fibonacci.nw -o fibonacci.tex
```

To produce the PDF document, the user should run the following commands. LaTeX needs to run twice to process references.

Code chunk 2: ≪compilation (part 2)≫

```
# The LaTeX style file lepton.sty should be in a directory accessible to LaTeX
pdflatex fibonacci.tex
pdflatex fibonacci.tex
```

2 Problem statement

The Fibonacci sequence is defined as the sequence of integers F_n such that

$$F_0 = 0 \tag{1}$$

$$F_1 = 1 \tag{2}$$

$$F_n = F_{n-1} + F_{n-2} \tag{3}$$

The goal is to define a function that returns F_n given the integer n.

3 Implementations

The proposed implementations in this document are taken from the Rosetta Code project `https://rosettacode.org/wiki/Fibonacci_sequence`.

3.1 Recursive implementation in OCaml

We define a `fibonacci` function in OCaml in the following code chunk. The contents of this chunk are sent by Lepton to an instance of the OCaml interpreter, and the output (the type of the `fibonacci` Ocaml object) is captured below automatically.

Code chunk 3: ≪ocaml≫

```
let rec fibonacci = function
  | 0 -> 0
  | 1 -> 1
  | n -> fibonacci (n-1) + fibonacci (n-2)
;;
```

Interpret with `ocaml`

```
val fibonacci : int -> int = <fun>
```

To check that the function is correct, let us ask OCaml for the first few numbers in the sequence.

Code chunk 4: ≪ocaml (part 2)≫

```
fibonacci 0;;
fibonacci 1;;
fibonacci 2;;
fibonacci 3;;
```

Interpret with `ocaml`

```
- : int = 0
- : int = 1
- : int = 1
- : int = 2
```

3.2 Iterative implementation in Python

We define a `fibIter` function in Python in the following code chunk. The contents of this chunk are sent by Lepton to an instance of the Python interpreter. There is no output on success.

Code chunk 5: ≪python≫

```
def fibIter(n):
    if n < 2:
        return n
    fibPrev = 1
    fib = 1
    for num in xrange(2, n): fibPrev, fib = fib, fib + fibPrev
    return fib
```

Interpret with `python`

```
```

To check that the function is correct, let us ask Python for the first few numbers in the sequence.

Code chunk 6: ≪python (part 2)≫

```
for i in range(0,4): print fibIter(i),
```

Interpret with `python`

```
0 1 1 2
```

3.3 Iterative implementation in C

In a compiled language such as C, we need to define the function `fibC` first, then include it in a program to use it. Let us start with the function definition.

Code chunk 7: ≪fibC≫

```
long long int fibC(int n) {
        int fnow = 0, fnext = 1, tempf;
        while(--n>0){
                tempf = fnow + fnext;
                fnow = fnext;
                fnext = tempf;
                }
                return fnext;
}
```

We include this in a program with a `main` function. This code chunk contains a reference to the definition of the `fibC` function, and Lepton will replace the reference with the corresponding source code.

Code chunk 8: ≪main.c≫

fibC

```
#include <stdlib.h>
#include <stdio.h>

<<fibC>>

int main(int argc, char **argv)
{
        int i, n;
        if (argc < 2) return 1;

        for (i = 1; i < argc; i++) {
                n = atoi(argv[i]);
                if (n < 0) {
                        printf("bad input: %s\n", argv[i]);
                        continue;
                }

                printf("%i\n", fibC(n));
        }
        return 0;
}
```

We configured the `main.c` code chunk such that its (expanded) contents are written to disk. We can now compile it with the following shell commands. Note that this implemenation returns an incorrect value for F_0.

Code chunk 9: ≪shell≫

```
gcc main.c -o a.out
./a.out 0 1 2 3
```

Interpret with `shell`

```
1
1
1
2
```

4 Comparison of running times

In this section, we compare the running times of the three proposed implementations. Let us first indicate the system configuration that is used to perform this comparison using shell commands. Note that Python writes to `stderr`, and we have to redirect its output so that it appears in the PDF document.

Code chunk 10: ≪shell (part 2)≫

```
uname -orvm
ocaml --version
python --version 2>&1
gcc --version
```

Interpret with `shell`

```
5.10.0-6-amd64 #1 SMP Debian 5.10.28-1 (2021-04-09) x86_64 GNU/Linux
The OCaml toplevel, version 4.11.1
Python 2.7.18
gcc (Debian 10.2.1-6) 10.2.1 20210110
Copyright (C) 2020 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

The time necessary for computing the number F_n depends on the algorithm, language, as well as n. We will compute the running times for several values of n, then assemble the results into a plot.

4.1 Ocaml

Let us define a function `time` to measure the time necessary in OCaml. This function uses the `Sys` module in the standard library.

Code chunk 11: ≪ocaml (part 3)≫

```
let time niter n =
  let start = Sys.time() in
  for i = 1 to niter do ignore (fibonacci n) done;
  (Sys.time() -. start) /. float_of_int niter
;;
```

Interpret with `ocaml`

```
val time : int -> int -> float = <fun>
```

Writing the results to disk.

Code chunk 12: ≪ocaml (part 4)≫

```
let oc = open_out "runtimes.ocaml" in
begin
  output_string oc "# Running times of Fibonacci sequence using Ocaml, time in seconds\n";
  for i = 1 to 9 do
    output_string oc (string_of_int i ^ "\t" ^ string_of_float (time 1000 i) ^ "\n")
  done;
  close_out oc;
end;;
```

Interpret with `ocaml`

```
- : unit = ()
```

We display below the contents of the results file.

Code chunk 13: ≪shell (part 3)≫

```
cat runtimes.ocaml
```

Interpret with `shell`

```
# Running times of Fibonacci sequence using Ocaml, time in seconds
1       1.4e-08
2       2.9e-08
3       4.5e-08
4       9.9e-08
5       1.61e-07
6       2.69e-07
7       4.4e-07
8       7.04e-07
9       1.142e-06
```

N.B. In the above example, all code chunks with the name `ocaml` share the same interpreter process. Consequently, the `fibonacci` function defined in the first chunk is available in the subsequent chunks. This is the same for the `shell` chunks and for the `python` chunks.

4.2 Python

Similarly, we define a function `timefib` to measure the time necessary in Python. There is no output on success.

Code chunk 14: ≪python (part 3)≫

```
from time import clock

def timefib(i):
  start = clock()
  for n in range(0,1000): fibIter(i)
  end = clock()
  return (end-start)/1000

file1 = open("runtimes.python","w")
file1.write("# Running times of Fibonacci sequence using Python, time in seconds\n")
for i in range(1,10):
  file1.write(str(i))
  file1.write("\t")
  file1.write(str(timefib(i)))
  file1.write("\n")

file1.close()
```

Interpret with `python`

```

```

We display below the contents of the results file.

Code chunk 15: ≪shell (part 4)≫

```
cat runtimes.python
```

Interpret with shell

```
# Running times of Fibonacci sequence using Python, time in seconds
1        6.8e-08
2        1.76e-07
3        2.04e-07
4        2.26e-07
5        2.47e-07
6        2.68e-07
7        2.9e-07
8        3.11e-07
9        3.33e-07
```

4.3 C

In C, we write a new program to run the benchmark. Note that this benchmark uses a reference to the `fibC` function defined earlier. This ensures that the same code is used in the `main.c` and `benchmark.c` programs. Lepton automatically inserts a PDF link to the definition of the `fibC` function.

Code chunk 16: ≪benchmark.c≫

fibC

```
#include <stdlib.h>
#include <stdio.h>
#include <time.h>

<<fibC>>

int main(int argc, char **argv)
{
    int i,n;
    clock_t t;

    for (n = 1; n<10; n++) {
        t = clock();
        for (i=1; i<10000; i++) fibC(n);
        t = clock() - t;
        double time_taken = ((double)t)/CLOCKS_PER_SEC / 10000; // in seconds
        printf("%i\t%e\n",n,time_taken);
    }
 }
```

We compile and run the benchmark below.

Code chunk 17: ≪shell (part 5)≫

```
gcc benchmark.c -o bench.out
echo "# Running times of Fibonacci sequence using C, time in seconds" > runtimes.c
./bench.out >> runtimes.c
head runtimes.c
```

Interpret with shell

```
# Running times of Fibonacci sequence using C, time in seconds
1        2.000000e-09
2        3.100000e-09
3        4.800000e-09
4        6.500000e-09
5        8.300000e-09
6        1.010000e-08
7        1.200000e-08
8        1.390000e-08
9        1.580000e-08
```

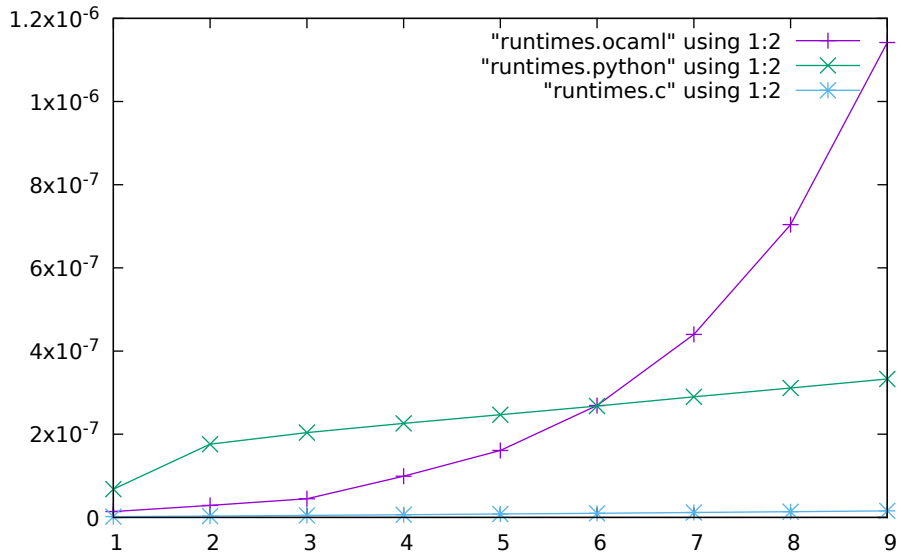


Figure 1: Running time in seconds for computing F_n as a function of n for the three proposed implementations.

4.4 Results and discussion

We use Gnuplot for making the figures. We first define the gnuplot script.

Code chunk 18: ≪plot.in≫

```
set terminal pdf
set output "runningtimes.pdf"
plot "runtimes.ocaml" using 1:2 with linespoints, \
     "runtimes.python" using 1:2 with linespoints, \
     "runtimes.c" using 1:2 with linespoints
```

We execute the script in gnuplot. This writes a PDF to disk, which is then included in a LaTeX figure.

Code chunk 19: ≪shell (part 6)≫

```
gnuplot plot.in # no output on success
```

Interpret with `shell`

```
```

In Figure 1, we plot the running times necessary to compute the Fibonacci sequence. As expected, the OCaml recursive implementation has exponential complexity in time, whereas the Python and C iterative implementations have linear time complexity. The C implementation is much faster than the Python and OCaml code, which are run inside an interactive loop whereas the C code is compiled. It would be interesting to compare with the programs produced with native code compilers of Python and OCaml or just-in-time compilation. Additionally, the recursive OCaml function runs faster than the iterative Python function for small values of n, which suggests that there is less overhead for calling functions in OCaml than in Python.

5 Conclusion

This document shows how to compute the Fibonacci sequence in three different programming languages, with one recursive implementation and two iterative implementations. Using Lepton makes it possible to

- provide everything in a single executable file that makes it easy to reproduce the results
- embed source code and executable instructions in a readable manner
- restructure the source code for easier human comprehension
- run compilation commands and analysis scripts to ensure that the figures were generated with the provided source code.

Chapitre 4

Programme Imageries du Vivant

Le LAGA est impliqué depuis 2013 dans un programme de recherche sur les Imageries du Vivant impliquant une trentaine d'équipes de recherche du PRES Sorbonne Paris Cité. Ce réseau a été créé initialement en 2008 sous la forme d'un Axe Thématique Prioritaire associant les Universités Paris Descartes et Paris Diderot sous la direction de Yves Frapart, Florence Cloppet et Charles-André Cuenod. Il a été étendu en 2013 sous la forme d'un programme interdisciplinaire financé par l'IDEX SPC à hauteur de 800 000 euros sur 5 ans, sous la direction de Charles-André Cuenod (Paris 5), Dominique Le Guludec (Paris 7) et Françoise Dibos (Paris 13).

J'ai participé au montage et à l'ensemble des activités de ce programme en tant que secrétaire et trésorier, et j'ai obtenu une décharge d'enseignement de 64h pour l'année 2016-2017 à ce titre. Initialement, il s'agissait pour moi de nouer des collaborations avec des applications en analyse d'images, et de pouvoir accéder à des données et des problématiques concrètes. Ce chapitre contient un bref exposé des activités du programme Imageries du Vivant, ainsi que des actions auxquelles j'ai directement participé.

4.1 Contexte applicatif

On s'intéresse ici à un ensemble de technologies d'imageries, c'est-à-dire de méthodes d'instrumentation capables de fournir des mesures en fonction de la position spatiale. Ce qui a motivé la constitution d'une communauté, c'est que les technologies d'imageries sont des technologies de pointe et que l'exploitation des données nécessite la collaboration entre compétences de différentes disciplines :

— en mathématiques et informatique pour l'analyse de données et le traitement du signal,

— en physique et électronique, car la plupart des méthodes exploitent les interactions entre une onde et la matière,
— en chimie, car des agents dits “de contraste” peuvent être injectés afin de sélectionner les objets biochimiques d'intérêt,
— en biologie et en médecine, pour l'interprétation correcte des images obtenues.

Le périmètre du projet a été restreint aux méthodes concourant à la production d'images pour les sciences du vivant, en allant de l'étude des cellules aux êtres humains. Il est à noter que l'on a abordé aussi bien les sujets pathologiques (e.g. maladies humaines) que les sujets sains, c'est-à-dire le fonctionnement normal d'un être vivant, et notamment les sciences cognitives. De plus, nous avons abordé un certain nombre d'enjeux éthiques et sociétaux, liés à la capacité de ces technologies à réveler des informations sur leur sujet : identité et anonymisation, maladies et risques de maladies, fonctionnement psychiques, etc.

Un enjeu important du domaine concerne la combinaison de différentes technologies afin d'en tirer des informations complémentaires.

4.2 Développement d'outils partagés

Afin de favoriser le travail collaboratif entre les différentes équipes et disciplines, le programme Imageries du Vivant a mis en place un certain nombre d'outils, et en premier lieu une plateforme de stockage et de calcul appelée Cloud-IDV afin de partager les données et les images. Une ingénieure de recherche (Leila Abidi), a été recrutée de mars 2015 à octobre 2018 afin de mettre en place cette infrastructure. Le support matériel a été fourni par la plateforme de calcul CUMULUS de la ComUE SPC.

Un service de machines virtuelles (`https://cumulus.parisdescartes.fr`) a été mis en place pour accéder et traiter les données hébergées par Cloud-IDV. Environ 150 machines virtuelles ont été déployées par les participants du programme IDV pour l'utilisation de logiciels d'analyse d'image existants ou développés dans les laboratoires. Un serveur de licences Matlab a également été mis en place, avec 10 jetons fournis par SPC et 25 jetons achetés par le programme IDV. L'infrastructure informatique déployée a été décrite dans deux chapitres de livres [34, 35]

Afin de permettre la fouille des données partagées, il s'est avéré nécessaire d'harmoniser l'annotation des jeux de données entrant sur la plateforme. En particulier, il fallait un certain nombre d'informations absentes des fichiers produits par les instruments, mais notés par les chercheurs dans leur cahier de laboratoire,

par exemple la référence de la lignée cellulaire, les conditions expérimentales, l'âge des souris, etc. Un groupe de travail a été organisé par Philippe Garteiser (CR Inserm/P7) et a rédigé une charte d'annotation. Celle-ci a été implémentée dans une interface développée en partenariat avec la société SysNCom, sous la forme d'un portail permettant d'accéder aux données et d'uploader de nouvelles images. Ce système a été déployé sur quatre sites pilotes (platforme IMAG'IC d'imagerie photonique de l'Institut Cochin, plateforme FRIM de l'Université Paris Diderot, le laboratoire PARCC de l'HEGP et sur le site des Saints-Pères de l'Université Paris Descartes).

4.3 Recherche interdisciplinaire

Le programme Imageries du Vivant a également mené des actions en faveur de la recherche interdisciplinaire en aidant au montage et au financement de projets associant des équipes de domaines différents. Nous avons ainsi financé 27 mois de postdoctorats, deux thèses de doctorat et 34 stages de M2. Ces financements ont été attribués sur appel d'offre. J'ai participé à l'organisation et aux jury d'évaluation (pour les appels auxquels je n'ai pas soumis de proposition).

Les résultats de ces projets de recherche ont été restitués lors de workshops annuels en 2014, 2015, 2016 et 2017 dont j'ai coordonné l'organisation. Ci-dessous quelques exemples des thématiques abordées :

— analyse d'image et traitement du signal : imagerie scanner (LIPADE / HEGP), imagerie IRM (L2TI / UTCBS, LIPADE / INSERM1149) microscopie optique (LAGA / Institut Cochin) et notamment le financement de la bourse de thèse de Tran Duc Nghia en spectroscopie RPE (LAGA / LCBPT)
— agents de contraste : visibilité IRM des oxydes de fer (UTCBS / LCBPT, LVTS / CSPBAT), oxydes de cérium (LMS / PCC)
— imagerie multi-modale : hypoxie (LCBPT / HEGP), PET/IRM (FRIM / LVTS) et notamment le financement de la bourse de thèse de Joao Piraquive en imagerie IRM (UTCBS / INSERM1149)
— imagerie médicale : radiomics (LIPADE / HEGP), annotation et fouille de bases de données (LIPADE / HEGP)
— théranostique : libération de médicaments (LCBPT / LPC),
— neurosciences : anatomie corticale et contrôle cognitif (LaPsyDe / CPN)

4.4 Enjeux éthiques et sociétaux

Un certain nombre de problématiques et d'enjeux sociétaux communs à l'ensemble des équipes du programme ont été identifiés : expérimentation animale, essais cliniques, secret médical, propriété intellectuelle des images, etc. Nous avons cherché des collaborations avec des laboratoires de recherche en éthique et juridique et tenté d'avancer sur ces questions.

Un verrou important du travail de recherche en imagerie médicale se trouve dans la préservation du secret médical des patients, et la nécessité d'anonymiser les données. Nous avons travaillé avec la société "Digital & Ethics" qui possède une expertise et une méthode brevetée d'anonymisation. Ce travail a conduit à une solution spécifique aux images IRM du cerveau qui est testée à l'hôpital Sainte-Anne (LaPsyDe, A. Cachia).

Un autre travail de réflexion éthique autour du contenu des images médicales a été mené en partenariat avec l'Espace Éthique Ile-de-France (E. Hirsch). En effet, les images médicales ont un contenu informatif très riche, qui peut conduire à la découverte fortuite de maladies, ou à la révélation de l'identité du sujet. Cette collaboration a donné lieu à plusieurs conférences et à l'édition d'ouvrages [52-54].

4.5 Animation d'un réseau de recherche

Afin de faire se rencontrer et soutenir le développement de projets interdisciplinaires, le programme Imageries du Vivant a organisé un certain nombre de workshops et de manifestations. En tant que trésorier, j'ai indirectement participé à l'ensemble des actions, c'est-à-dire :

- un séminaire mensuel depuis 2008 (dont comité d'organisation depuis 2016),
- 3 séminaires de travail "hors les murs" de trois jours en 2014, 2015 et 2016 (comité d'organisation des trois éditions),
- 6 workshops d'une journée, et notamment une micro-école sur l'utilisation de Cloud-IDV,
- 4 séminaires de présentation des étudiants financés par le programme IdV en 2014, 2015, 2016 et 2017 (dont comité d'organisation pour 3 éditions),
- 3 journées de rencontre Imageries du Vivant - Industries en 2015, 2016 et 2017 (dont comité d'organisation en 2016 et 2017).

4.6 Conclusion et perspectives

En tant que membre du comité exécutif et trésorier du programme, je suis intervenu dans le pilotage et dans l'ensemble des actions programme Imageries du Vivant. En plus d'une connaissance approfondie des différentes technologies et des problématiques de recherche des différentes disciplines, et d'une certaine expérience du dialogue interdisciplinaire, je suis convaincu que ma participation à ce programme a fait progresser mes réflexions sur l'analyse de données, et a contribué l'élaboration de la notion d'énoncé scientifique telle que presentée dans le chapitre 2.

De plus, le programme a apporté des financements pour mes travaux (trois bourses de stage de M2 et une bourse de thèse) et a permis d'initier la collaboration avec la plateforme de Résonance Paramagnétique Électronique de l'Université Paris Descartes (Yves Frapart, LCBPT) dont les résultats sont décrits dans le chapitre 6.

Le non-renouvellement du financement IDEX en 2016 et la clôture des crédits "biseau" fin 2018 ont conduit le programme IDV à chercher un renouvellement sous une autre forme. La direction du programme a été renouvelée le 18 mai 2018, sous la direction de Yves Frapart, avec un comité de direction composé de P. Garteiser (P7), S. Li-Thiao-Té (P13), A. Cachia (P5), B.-T. Doan (P5) et D. Geldwerth (P13), avec pour objectif de mettre en place une fédération de recherche CNRS. Cette structure serait pérenne et disposerait d'un financement autonome, non prélevé sur les dotations des équipes et laboratoires participants. À ce jour, le projet a été reçu favorablement par le CNRS et l'INSERM, et des contacts ont été pris avec les présidences de l'Université Paris 13 et de l'Université de Paris. Malheureusement, le contexte de la fusion des Universités Paris Descartes et Paris Diderot, ainsi que l'attribution d'un nouveau financement IDEX pour cette fusion ne facilitent pas l'avancée des négociations.

Chapitre 5

Modélisation basée patch en analyse des images

En parallèle des discussions pour le montage du programme interdiscriplinaire présenté au chapitre 4, j'ai souhaité mener des travaux de recherche en analyse d'image appliqués aux images bio-médicales en 2013. Le contexte y était favorable, et j'ai pu nouer une collaboration avec Marie Luong du L2TI (Paris 13) et Jean-Marie Rocchisani, radiologue à l'hôpital Avicenne. Nous avons recruté Dai Viet Tran pour un stage de M2 (encadrants Li-Thiao-Té Sébastien et Marie Luong, financé par le programme Imageries du Vivant), puis en thèse sous la direction de Françoise Dibos (PU, LAGA), co-encadrants Li-Thiao-Té Sébastien et Marie Luong, avec un financement provenant de la Fédération Math-STIC. Ce chapitre décrit les travaux qui ont été menés dans le cadre de cette thèse et comment ils s'inscrivent dans mon programme de recherche.

On s'intéresse à la modélisation des images, c'est-à-dire que l'on souhaite déterminer un ensemble de modèles ressemblant aux images observées, ou bien que l'on souhaite décrire l'ensemble des images en tant que sous-ensemble des matrices réelles ou des fonctions $\mathbb{R}^2 \to \mathbb{R}$, dans le cas de l'imagerie 2D. Dans le cas d'images tridimensionnelles, ou bien d'images multi-modales, comportant plusieurs canaux de couleurs ou de mesures physico-chimiques, il conviendra d'étendre les travaux présentés à plusieurs dimensions d'espace et à des observations dans $\mathbb{R}^k$.

Les capteurs actuels utilisés en photographie ont une résolution très élevée, et conduisent à considérer un espace des images de très grande dimension $\mathcal{E}$ du type $\mathbb{R}^n$ où n est de l'ordre de plusieurs millions pour des images pixellisées. En imagerie biomédicale, les contraintes techniques des différents instruments (scanners, IRM, microscopes, etc.) conduisent en général à des images plus réduites, de l'ordre de 512x512, suivant la technologie utilisée. Dans les deux

cas, l'espace des observations $\mathcal{E}$ est de dimension largement supérieure à la complexité effective des objets étudiés.

Par conséquent, on anticipe que l'espace des modèles $\mathcal{M}$, inconnu, sera de dimension réduite. Par exemple, les travaux de Chang et Tsao [37] estiment que le système visuel humain reconnaît les visages en les représentant dans un espace de dimension 50 environ. Ceci est à mettre en regard des approches actuelles d'apprentissage dont les réseaux de neurones opèrent parfois sur les images entières, c'est-à-dire directement dans $\mathcal{E}$.

La plupart des méthodes d'analyse d'image peuvent être envisagées au travers de la réduction de dimension. Ainsi, la théorie des ondelettes [70], le compressed sensing [33], les approches de représentation par dictionnaire correspondent à utiliser un espace de modèles $\mathcal{M}$ qui est un espace vectoriel bien choisi. Les méthodes de filtrage opèrent une régularisation, c'est-à-dire cherchent un représentant de l'image input dans un sous-ensemble de fonctions régulières.

Dans ce chapitre, on considère que l'espace des modèles $\mathcal{M}$ représentant les images est de dimension réduite, car engendré par des images élémentaires (les patches) de petite taille. Les différents travaux conduisent à justifier cette hypothèse au travers de la performance du traitement d'image qui peut être réalisé dans l'espace $\mathcal{M}$.

5.1 Notion de patch

On définit un patch ou imagette comme une image de taille $k \times k$, où k est de l'ordre de 7 pixels [21]. En général, on prendra k impair, ce qui permet de considérer le pixel central sans ambiguité. L'idée principale de ce domaine de l'analyse d'image est que la complexité des images, quelle que soit leur taille, est capturée par des relations locales entre les valeurs des pixels. On peut donc traiter les problèmes d'analyse d'image dans $\mathcal{M} \subset \mathbb{R}^{7\times 7}$.

On présente donc le modèle de génération d'une image observée de la façon suivante. Un patch observé y est obtenu à partir d'un modèle ou patch idéal $x \in \mathcal{M}$ à partir d'un opérateur d'observation H (linéaire) et d'un bruit blanc gaussien additif η :

$$y = H.x + \eta$$

La matrice H peut contenir différentes opérations telles que le flou (convolution), et en particulier un opérateur d'augmentation ou de réduction de la dimension, auquel cas H n'est pas une matrice carrée. On obtient une image observée de $\mathcal{E}$ comme l'assemblage de patchs observés, juxtaposés les uns à côté des autres, à la manière d'une mosaique, avec ou sans recouvrement.

En première intention, on peut travailler en considérant que un patch idéal x est un élément arbitraire de $\mathbb{R}^{k\times k}$. Cependant, on constate en pratique que l'espace des images est un espace de dimension plus réduite, et que x appartient à un sous-espace de $\mathbb{R}^{k\times k}$. Ceci conduit à un premier problème de modélisation consistant à bien choisir l'espace de représentation des patchs $\mathcal{M}$. On se restreindra aux sous-espaces linéaires, c'est-à-dire que l'on suppose l'existence d'une base ou d'une famille de vecteurs (souvent appelée *dictionnaire*) dont les patchs idéaux sont des combinaisons linéaires.

Le problème de reconstruction d'une image bruitée consiste à inverser l'opération d'observation, c'est-à-dire à résoudre le problème de minimisation :

$$\hat{x} = \underset{x\in\mathcal{M}}{\arg\max}\, \mathbb{P}(x|y) = \arg\min ||y - H.x||^2 - \lambda \ln\left(\mathbb{P}(x)\right)$$

En plus de la résolution du problème de minimisation, on voit apparaître un deuxième problème de modélisation dans le choix de la mesure de probabilité $\mathbb{P}(x)$. En effet, cette mesure définie sur $\mathcal{M}$ ou sur $\mathbb{R}^{k\times k}$ peut indiquer par son support ou sa distribution deux choses différentes. Elle peut caractériser l'ensemble des modèles $\mathcal{M}$ comme étant le support de $\mathbb{P}$ et/ou privilégier certains modèles a priori plus "plausibles" en leur accordant une probabilité supérieure.

5.2 Méthode SRSW

Ayant choisi un bruit gaussien additif et les déformations imposées par la structure de la matrice H, la performance de la méthode de reconstruction d'image dépend de la qualité de la modélisation de l'espace des images, c'est-à-dire du sous-ensemble de $\mathcal{M} \subset \mathbb{R}^{k\times k}$ choisi et de la mesure de probabilité $\mathbb{P}$ sur cet ensemble. Dans la thèse de Dai Viet Tran, nous sommes partis de la méthode Super-Resolution by Sparse Weight (SRSW [91, 92]) qui a été développée au L2TI durant la thèse de Dinh Hoan Trinh soutenue en 2013 sous la direction de Françoise Dibos et Marie Luong.

Dans la méthode SRSW, étant données une observation y et une base de données de patchs idéaux,

- on cherche un ensemble de patchs similaires dans la base de données suivant une distance d,
- on définit un espace de Hilbert local qui est l'ensemble des combinaisons linéaires des patchs similaires,
- on résout le problème de minimisation dans cet espace local, ce qui donne un vecteur de coefficients notés α permettant de reconstruire le patch idéal x.

Par conséquent, la méthode SRSW spécifie l'espace des patchs idéaux $\mathcal{M}$ au travers de dictionnaires locaux, extraits à partir d'une base de données d'images de bonne qualité. La mesure de probabilité $\mathbb{P}$ est choisie arbitrairement, c'est la distribution de Laplace qui est utilisée, avec une pondération calculée selon la proximité avec y.

En partant de cet algorithme, nous avons étudié les différents éléments de la méthode afin de comprendre lesquels étaient efficaces. Dans la plupart des travaux, nous considérons le cas de figure du débruitage, i.e. une matrice H carrée, mais également le cas de la super-résolution où les observations y sont plus petites que les patchs idéaux x. Dans ce cas, on utilise une base de données contenant des couples de patchs au lieu d'appliquer un zoom aux patchs y.

5.3 Travail 1 : Transformation de Anscombe

Un premier travail consiste en l'extension de la méthode SRSW pour le cas des images bruitées par un bruit suivant une loi de Poisson. On utilise la transformation de Anscombe [24] qui transforme approximativement des échantillons d'une variable aléatoire de loi de Poisson en échantillons d'une variable aléatoire de loi gaussienne. Cette transformation est appliquée à tous les patchs avant de procéder à la reconstruction habituelle.

Dans ce travail, on examine également l'effet du choix du nombre de patchs utilisés pour construire l'espace de Hilbert local et l'effet de la taille des patchs sur la qualité de la reconstruction (débruitage et super-résolution), telle que évaluée d'après les critères PSNR et SSIM. Le manuscrit [11] fait état de bonne performances, notamment pour le débruitage d'images médicales, en comparaison des méthode d'interpolation bicubique, Neighbor Embedding [36] et ScSR [101].

5.4 Travail 2 : Distance EMD

Dans la méthode SRSW, l'invariance par translation des images n'est pas utilisée. Les distances d considérées sont des distances pixel à pixel, et l'on s'appuie sur l'exhaustivité de la base de données pour retrouver des patchs similaires dans la bonne position. La base de données de patchs idéaux étant construite par extraction à partir d'images de bonne qualité, les translations d'un nombre entier de pixels sont présentes. En considérant la distance Earth Mover's Distance (EMD) ou distance de Wasserstein, notre objectif est de construire le dictionnaire local de reconstruction à partir de patchs correspondant à des translations sous-pixelliques, ou à des petites rotations.

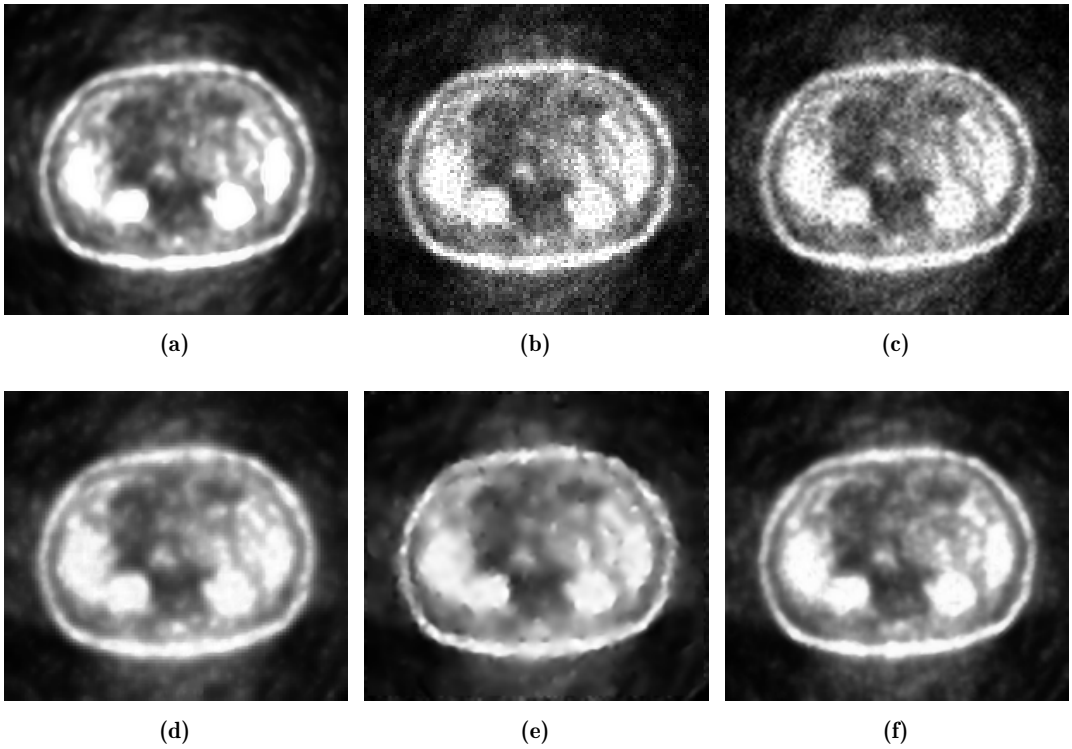


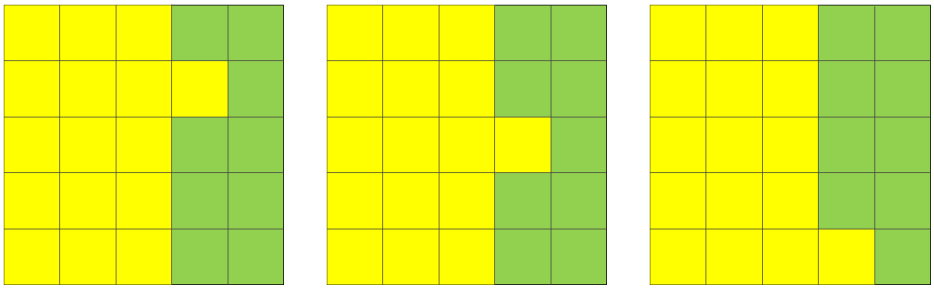
FIGURE 5.1 – Super-resolution results on PET image of abdomen. (a) Original high-resolution image. (b) The low-resolution noisy image (shown with nearest neighbor interpolation). (c) Result of bicubic interpolation (PSNR = 25.87, SSIM = 0.718). (d) Result of NE method (PSNR = 26.15, SSIM = 0.863). (e) Result of ScSR method (PSNR = 27.72, SSIM = 0.876). (f) Result of the proposed method (PSNR = 29.05, SSIM = 0.891).

On rappelle que la distance EMD [83] est définie entre deux histogrammes $h_1(i)$ et $h_2(j)$ par

$$\mathrm{EMD}(h_1, h_2) = \min_{f_{ij}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}}$$

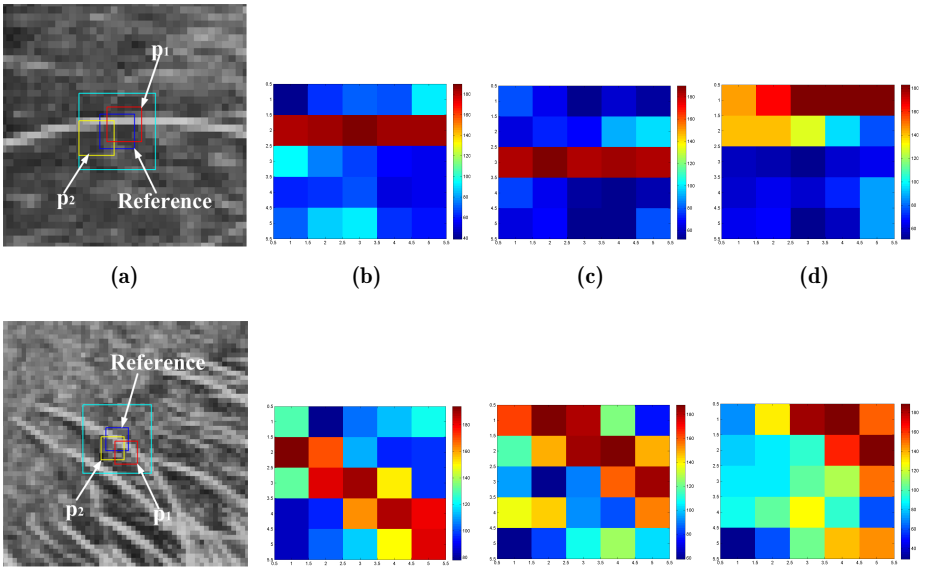
où les coefficients f_{ij} représentent la quantité de matière (earth) transportée entre les positions i et j des histogrammes. Nous avons utilisé la variante proposée par Pele [74] qui traite le cas des histogrammes non normalisés :

$$\mathrm{EMD}(h_1, h_2) = \min_{f_{ij}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}} + \left| \sum h_1(i) - \sum h_2(j) \right| \times a \max_{i,j} d_{ij}$$



(a) **(b)** **(c)**

Figure 5.2 – Comparison of the FastEMD and the Euclidean distance for patch similarity. The intensity of yellow and green pixels are 200 and 0, respectively. The Euclidean distance $L_2(a,b) = L_2(a,c) = 282.8$ whereas $\text{EMD}(a,b) = 200, \text{EMD}(a,c) = 600$. The EMD distance better reflects human perception in so far as image b is closer to a than image c.



(a) **(b)** **(c)** **(d)**

(e) **(f)** **(g)** **(h)**

Figure 5.3 – Comparison of the FastEMD and the Euclidean distance for patch similarity. (a) and (e) Two regions with one reference patch and two candidate patches. (b) Reference image patch (blue square in (a)). (c) First candidate patch p1 (red square in (a)) with $L_2 = 361.1, \text{EMD} = 812$. (d) Second candidate p2 (yellow square in (a)) with $L_2 = 296.8, \text{EMD} = 1027$. (f) Reference image patch in (e). (g) First candidate patch in (e) with $L_2 = 326.8, \text{EMD} = 1282$. (h) Second candidate in (e) with $L_2 = 326.6, \text{EMD} = 2207$

Malheureusement, le calcul de la distance EMD prend beaucoup de temps de calcul, ce qui rend difficile la sélection des patchs similaires. Nous avons proposé d'effectuer un préfiltrage basé sur la distance L_1, et calculé un seuil correspondant à des déplacements de l'image de 0.5 pixels. Cette modification permet d'avoir des temps de calcul raisonnables et de montrer qu'en pratique, la méthode basée sur la distance EMD est meilleure que la méthode SRSW (cf [12]).

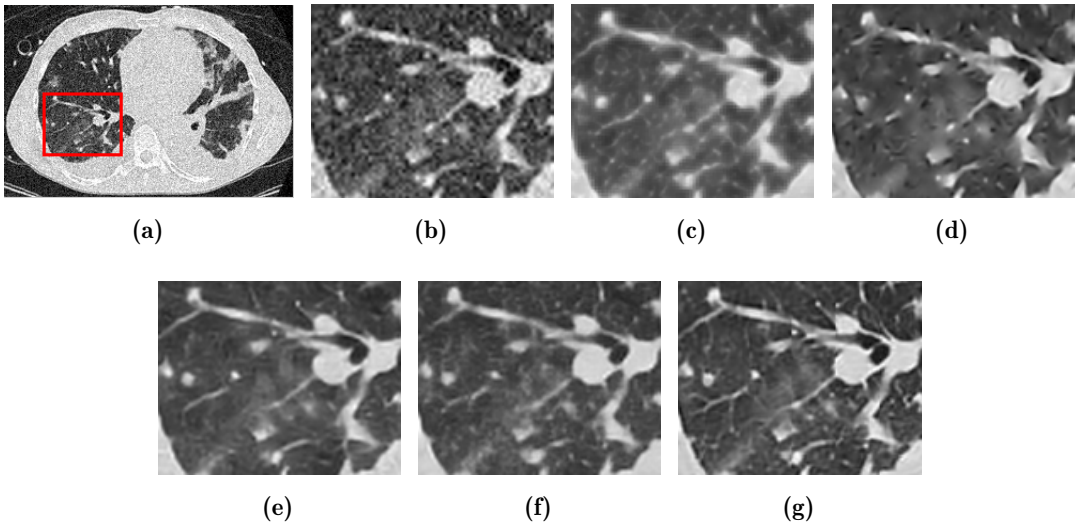


FIGURE 5.4 – Super-resolution results of a the CT image of thorax with magnification s = 2 and noise level $\sigma = 20$. (a) LR image (size 270×180) obtained by downsampling. (b)-(f) The ROI up-scaled by Bicubic interpolation, NE method (K = 4), the ScSR method ($\lambda = 0.8$), the SRSW method and the proposed SREMD method. (g) ROI in the original image.

5.5 Travail 3 : Représentation en dimension 3

Il est intéressant de remarquer que dans les précédents travaux, on utilise la base canonique de $\mathbb{R}^{k\times k}$ pour représenter l'espace des patchs, et que le fait d'utiliser des dictionnaires locaux masque la complexité réelle de l'espace des patchs. Nous nous sommes donc intéressés à la dimensionalité de l'espace des images représentées par patchs dans les deux travaux qui vont suivre.

Dans le troisième travail [14], nous avons cherché à savoir si une approche du type SRSW est applicable avec un ensemble de patchs idéaux ou de modèles de dimension 3 choisi une fois pour toute d'après la base de données. L'ensemble

des modèles $\mathcal{M}$ est un sous-espace de dimension 3 de $\mathbb{R}^{k\times k}$. On simplifie le dictionnaire, mais on travaille un peu plus sur la loi de probabilité $\mathbb{P}(x)$, qui est ici une fonction constante par morceaux estimée sur une base de données au lieu de la loi de Laplace choisie arbitrairement selon une hypothèse de parcimonie. La figure 5.5 montre que ceci permet de mieux s'adapter à la distribution des patchs, et remet en question l'hypothèse de parcimonie.

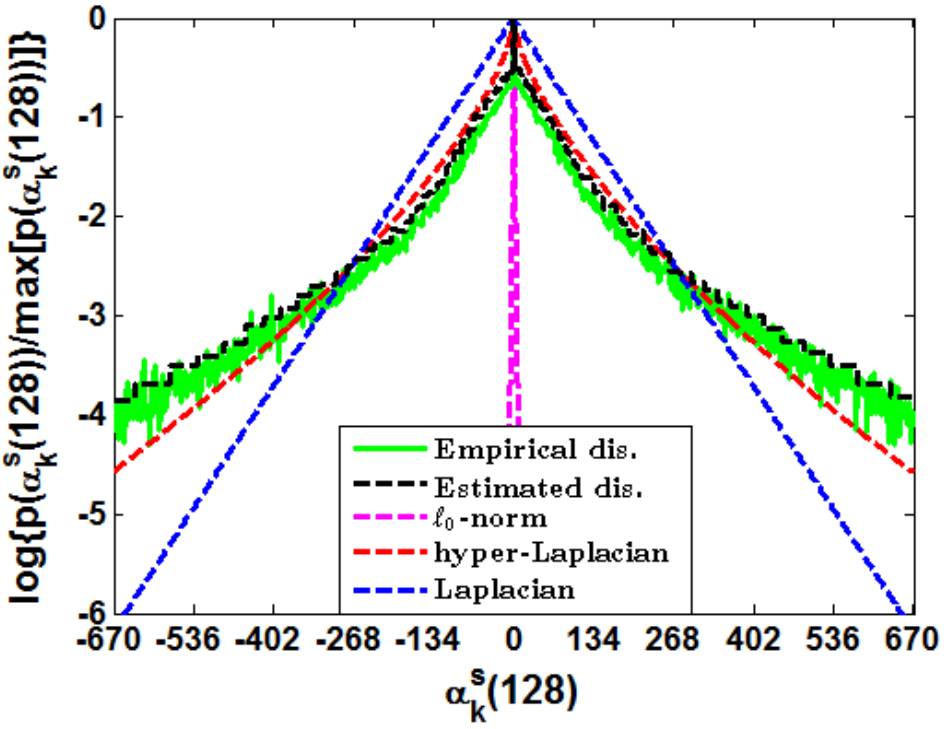


Figure 5.5 – Empirical distribution of the representation coefficients $\alpha_k^s(128)$ of image patches in the database (green curve). The fitting sparse models, including the '0-norm (in magenta dash curve), Laplacian (in red dash curve) and hyper-Laplacian (in blue dash curve). The black dash line is the histogram estimation of the real distribution of patches.

On résout donc le problème de minimisation

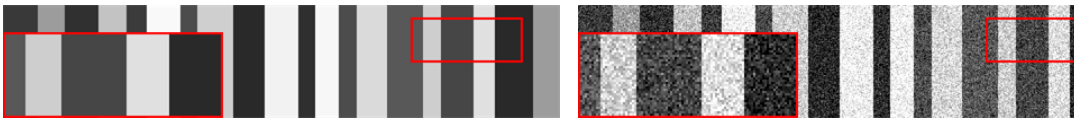
$$\hat{x} = \text{argmin}||By - BH.x||^2 - \lambda \log(\mathbb{P}(x))$$

où la matrice B réalise une projection orthogonale sur l'espace $\mathcal{M}$ et $\mathbb{P}(x)$ est constante par morceaux. On pourrait craindre que ce problème de minimisation se présente plus mal que la loi de Laplace (dont le logarithme est un polynôme). En fait, la minimisation est beaucoup plus rapide en exploitant le fait que $\mathbb{P}(x)$ est constant par morceaux.

Proposition 5.5.1. *Soit x_0 le minimum de $||By - BH.x||^2$. Si $\mathbb{P}(x)$ est constant par morceaux, alors la solution du problème de minimisation est soit x_0, soit un point situé sur la frontière des morceaux. Si de plus, les morceaux sont déterminés par une grille rectangulaire de $\mathbb{R}^k$, alors il suffit d'examiner les noeuds de la grille.*

Démonstration. Sur chaque morceau, la fonction à minimiser est quadratique, donc convexe. Ou bien x_0 est à l'intérieur du morceau, et c'est un minimum local, ou bien le minimum sur le morceau se trouve sur la frontière. □

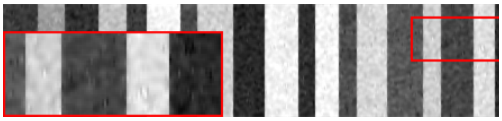
Nous avons montré que même avec un dictionnaire de dimension 3, l'estimation de $\mathbb{P}(x)$ permet d'avoir de bonnes performances de débruitage. Une première expérience consiste à se restreindre à des images qui ne comportent que des motifs verticaux avec des patchs de taille 3×3. L'espace des patchs est effectivement de dimension 3 dans ce cas, et le dictionnaire est complet. Les performances de l'algorithme sont comparées aux approches OMP [93] et LARS [104].



(a) Original image (test15)

(b) Noise image (18.56 - 0.270)

(c) OMP (28.04 - 0.646)



(d) LARS (28.43 - 0.721)

(e) ProbaEst (29.67 - 0.773)

FIGURE 5.6 – L'image (a) bruitée avec $\sigma = 30$ (b). Résultats de débruitage par les méthodes OMP (c), LARS (d) et la méthode proposée (e).

Une deuxième expérience concerne le débruitage d'images binaires, contexte dans lequel on tente de favoriser une méthode utilisant un a priori parcimonieux car on pense que l'espace des patchs est très simple. De plus, on utilise le

dictionnaire vertical. Les résultats obtenus sont très bons, notamment lorsque les images sont contaminées par un fort niveau de bruit.

(a) Original **(b)** Noise **(c)** OMP

(d) LARS **(e)** ICM **(f)** Graph Cuts

(g) ProbaEst

Figure 5.7 – Débruitage de l'image Peppers (binarisée) avec un bruit gaussien additif d'écart-type $\sigma = 30$.

Une troisième expérience a été réalisée sur des images naturelles en couleurs, afin de tester l'algorithme dans le cas où on pense que l'espace des patchs est de dimension supérieure à 3. Le résultat n'est pas parfait, mais on obtient des résultats de qualité comparable à des méthodes bien plus sophistiquées, notamment à fort niveau de bruit.

Images	$\sigma = 10$			$\sigma = 20$			$\sigma = 30$		
	OMP	**LARS**	**ProbaEst**	**OMP**	**LARS**	**ProbaEst**	**OMP**	**LARS**	**ProbaEst**
Airfield	**0.845**	0.839	0.840	0.713	0.743	**0.744**	0.607	0.655	**0.665**
Airplane	0.848	0.899	**0.901**	0.702	0.790	**0.793**	0.584	0.682	**0.704**
Baboon	**0.879**	0.841	0.821	**0.768**	0.736	0.735	0.668	0.641	**0.644**
Baby	0.856	0.902	**0.904**	0.711	**0.795**	0.792	0.596	0.692	**0.707**
Barbara	0.867	**0.874**	0.862	0.740	**0.763**	0.762	0.625	0.653	**0.666**
Boat	0.869	0.905	**0.910**	0.732	0.797	**0.801**	0.618	0.695	**0.714**
Bridge	**0.892**	0.867	0.862	**0.775**	0.761	0.767	0.678	0.681	**0.683**
Cameraman	0.849	**0.882**	0.881	0.701	0.765	**0.768**	0.592	0.662	**0.682**
Couple	0.874	**0.879**	**0.879**	0.745	0.775	**0.782**	0.640	0.686	**0.702**
Fruits	0.839	**0.876**	**0.876**	0.690	**0.768**	0.765	0.568	0.658	**0.678**
Boy	0.807	**0.791**	0.789	0.647	**0.675**	0.674	0.525	0.575	**0.583**
Hill	**0.852**	0.844	0.847	0.717	0.745	**0.750**	0.608	0.658	**0.670**
House	0.833	0.855	**0.856**	0.689	0.752	**0.754**	0.577	0.653	**0.674**
Jellybeans	0.861	**0.937**	**0.937**	0.719	**0.827**	0.823	0.601	0.711	**0.731**
Leaves	0.944	**0.961**	0.957	0.882	0.906	**0.908**	0.818	0.842	**0.858**
Lena	0.834	**0.857**	0.855	0.694	**0.757**	0.755	0.578	0.660	**0.675**
Man	0.863	**0.881**	**0.881**	0.729	0.779	**0.782**	0.618	0.685	**0.699**
Monarch	0.908	**0.936**	0.933	0.813	**0.862**	**0.862**	0.721	0.780	**0.796**
Peppers	0.813	**0.819**	0.816	0.666	**0.720**	**0.720**	0.552	0.625	**0.643**
Zelda	0.838	**0.866**	0.865	0.694	**0.759**	0.757	0.576	0.655	**0.670**
Average	0.859	**0.875**	0.874	0.726	0.774	**0.775**	0.617	0.677	**0.692**

FIGURE 5.8 – Débruitage de 20 images naturelles avec des niveaux d'erreurs croissants. La performance de débruitage est indiquée en terme de SSIM.

5.6 Travail 4 : Nombre de composantes dans un modèle de mélange gaussien

Suite aux résultats obtenus avec un dictionnaire global de dimension 3 et une estimation simple de la distribution des patchs, nous avons cherché à passer

en dimension supérieure en considérant un modèle de mélange gaussien pour $\mathbb{P}(x)$. Ceci rejoint une remarque souvent rencontrée dans la littérature [100, 102, 103] qui affirme que l'on peut se restreindre à un petit nombre de composantes du mélange, voire une seule, pour la reconstruction, et ignorer la contribution des autres composantes. Du point de vue de la modélisation de l'espace des images, ceci signifie que l'espace des patchs est formé d'un certain nombre de clusters isolés par opposition à un paysage uniforme.

Dans ce travail l'espace des modèles est donc $\mathcal{M} = \mathbb{R}^{k\times k}$ muni d'un a priori $\mathbb{P}(x)$ sous la forme d'un modèle de mélange gaussien à M composantes :

$$\mathbb{P}(x) = \sum_{m=1}^{M} \pi_m \mathcal{N}(x|\mu_m, \Sigma_m)$$

Les paramètres du mélange sont estimés en utilisant l'algorithme EM.

Le problème de reconstruction est donc

$$\hat{x} = \text{argmin}||y - H.x||^2 - \lambda \log\left(\sum_{m=1}^{M} \pi_m \mathcal{N}(x|\mu_m, \Sigma_m)\right)$$

où H contient le passage entre l'espace des modèles $\mathcal{M}$ contenant les patchs idéaux x et l'espace des observations $\mathcal{E}$.

Dans le travail publié dans [15], nous cherchons à savoir à quelle perte de performance correspond la résolution du problème simplifié

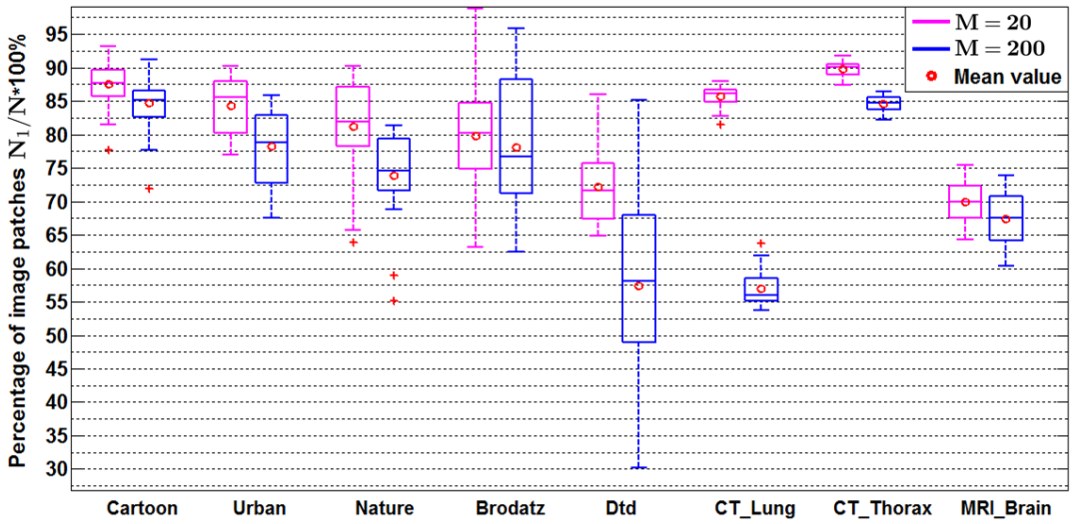
$$\hat{x} = \text{argmin}||y - H.x||^2 - \lambda \log(\pi_l \mathcal{N}(x|\mu_l, \Sigma_l))$$

où l correspond à la plus grande composante du mélange au point y.

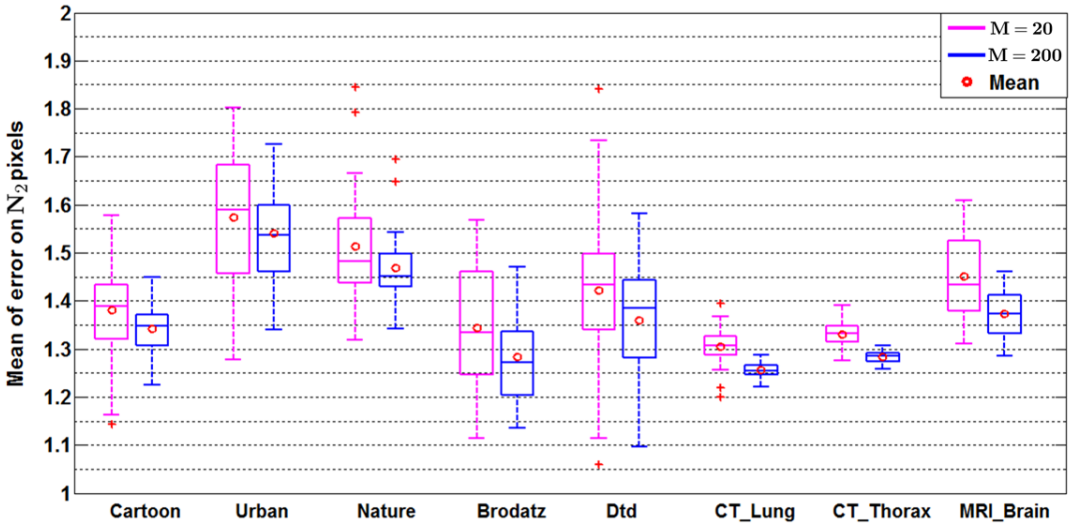
Dans une série d'expériences portant sur plusieurs jeux de données d'images naturelles et médicales (Figures 5.9, 5.10 et 5.11), nous comparons la perte en terme de PSNR et d'erreur de reconstruction L_1 suivant le nombre de composantes ($M \in \{20, 200\}$). Deux méthodes de construction du dictionnaire ont été évaluées pour des patchs de taille 8×8. Comme on est en dimension 64, nous avons utilisé la base canonique comme dictionnaire. D'autre part, nous avons construit un dictionnaire à 256 éléments selon la méthode K-SVD [19]. Ce dictionnaire est donc redondant.

On constate tout d'abord que la plupart des patchs dans les images considérées ont une composante majoritaire, c'est-à-dire que la composante de plus grande magnitude correspond à plus de 90% de la magnitude totale (Figures 5.9c et 5.10c). Ceci est observé dans la plupart des jeux de données, avec des

proportions de l'ordre de 75% à 90%, quel que soit le dictionnaire utilisé. Ce taux est plus réduit lorsque le nombre de composantes est plus élevé (elles sont plus rapprochées dans l'espace $\mathcal{M}$), et pour des images texturées (jeux de données Dtd et CT_Lung). De fait, on constate visuellement que les patchs sans composante majoritaire semblent se concentrer sur les zones texturées et sur les contours des images.



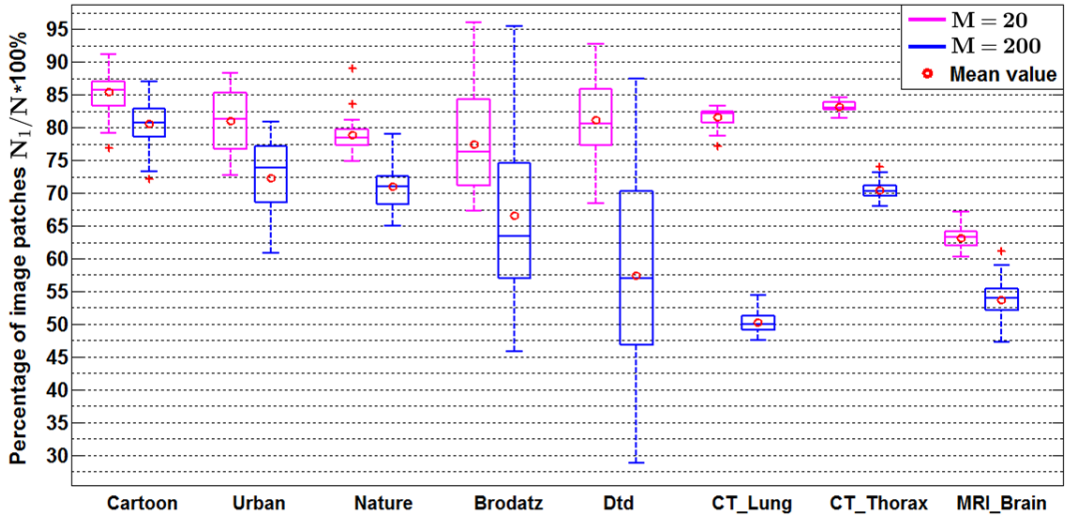
(c) Percentage of N_1 pixels in images of 8 datasets



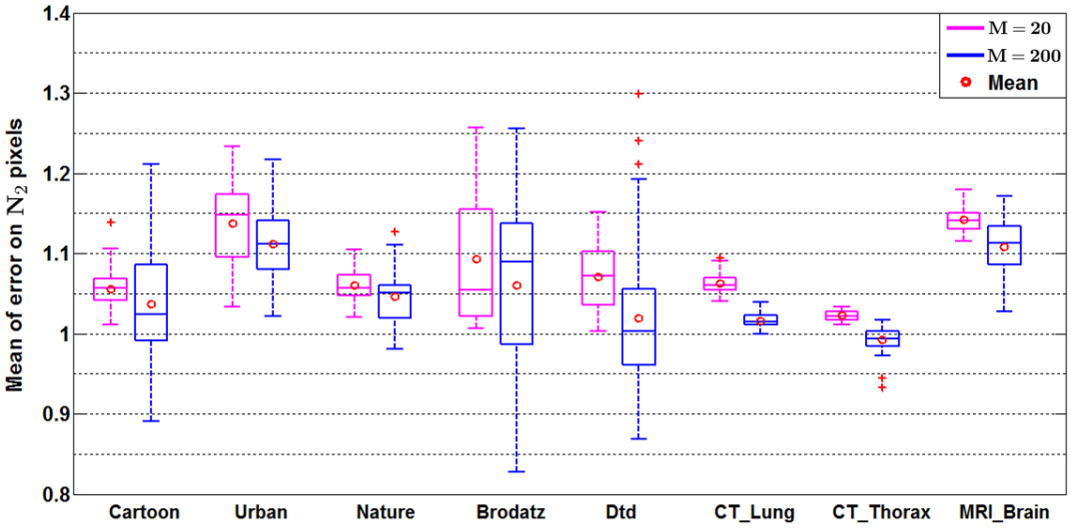
(d) Average reconstruction error $||\hat{\mathbf{X}}_{L=1} - \hat{\mathbf{X}}_{L=5}||_{L_1}$

FIGURE 5.9 – Représentation des images et débruitage en utilisant un modèle de mélange (dictionnaire base canonique).

Quel que soit le dictionnaire utilisé et le nombre de composantes, reconstruire avec une composante fait perdre peut de chose par rapport à 5 composantes (Figures 5.9d et 5.10d). Sur les pixels N_2 qui n'ont pas une composante majoritaire, on obtient des différences de 1 à 2 niveaux de gris en moyenne, et des différences de PSNR de moins de 0.2dB. Encore une fois, les pertes sont plus importantes sur les images texturées (jeux de données Dtd et CT_Lung).



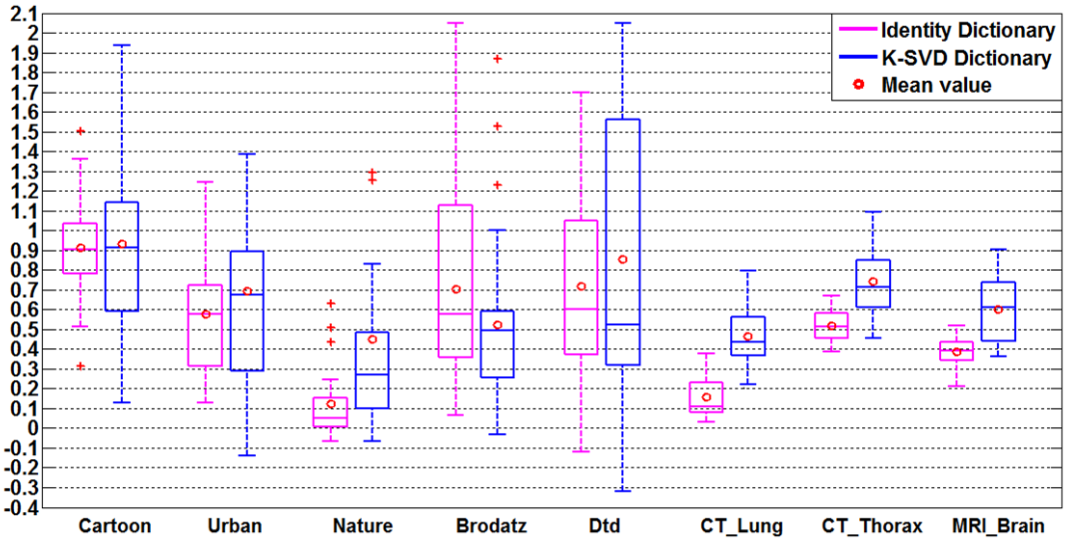
(c) Percentage of N_1 pixels in images of 8 datasets



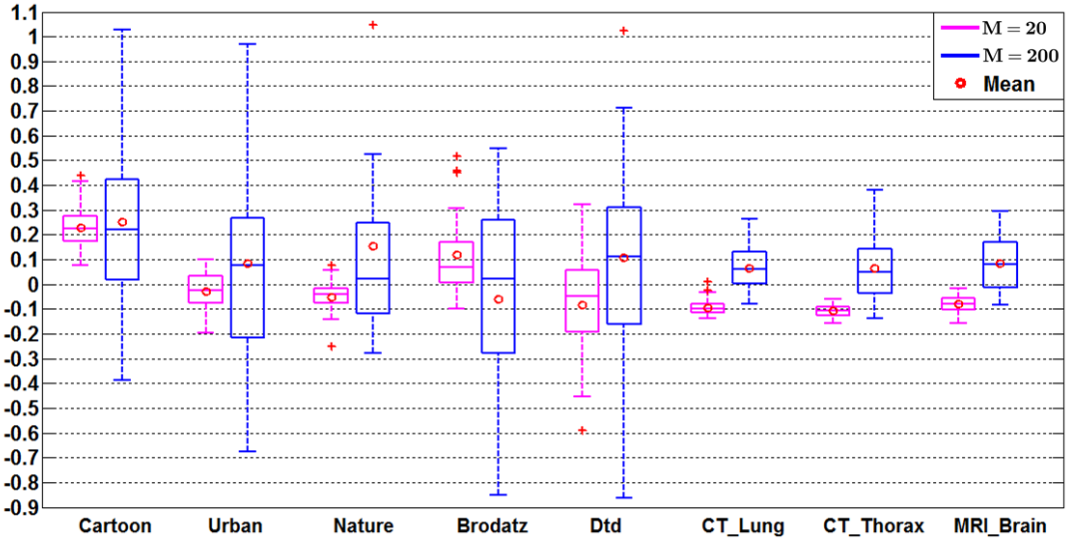
(d) Average reconstruction error $||\hat{\mathbf{X}}_{L=1} - \hat{\mathbf{X}}_{L=5}||_{L_1}$

Figure 5.10 – Représentation des images et débruitage en utilisant un modèle de mélange (dictionnaire K-SVD).

Cette série d'expérience permet également de comparer le choix de dictionnaire et le choix du nombre de composantes. On observe que l'utilisation de 200 composantes au lieu de 20 apporte un gain de l'ordre 1dB d'amélioration (Figure 5.11a), et que le dictionnaire K-SVD bénéficie un peu plus de l'augmentation de la complexité du modèle. Par contre, les différences sont marginales entre les deux méthodes de construction du dictionnaire (Figure 5.11b). Cette étude suggère qu'il y a peu d'intérêt à tenter d'adapter le dictionnaire.



(a)



(b)

FIGURE 5.11 – Reconstruction d'image avec une seule composante. Effet du nombre de composantes (a) et du choix du dictionnaire (b).

5.7 Conclusion

Les travaux réalisés dans le cadre de la thèse de Dai Viet Tran (soutenue le 26 juin 2018) en collaboration avec Françoise Dibos et Marie Luong ont porté sur les espaces de modèles $\mathcal{M}$ engendrés par des dictionnaires de patchs et leur application en analyse d'images. Dans l'ensemble, ces travaux montrent que la structure de l'ensemble des modèles est importante et que des méthodes simples de modélisation de cet espace permettent d'obtenir de bons résultats de débruitage, par opposition à des choix arbitraires.

En regard de la méthodologie générale exposée au chapitre 2, on peut constater que les travaux menés dans cette thèse ont été un peu exploratoires, et que l'on aurait pu procéder de manière plus systématique. Ainsi, il y a des hésitations sur le bruit (gaussien ou poissonien) dans le modèle d'observation des images, et sur la distance à utiliser entre les patchs idéaux et les patchs observés. Or dans les deux cas, il s'agit d'étudier le passage entre l'espace des modèles $\mathcal{M}$ et l'espace des données $\mathcal{E}$. Au lieu de faire le choix arbitraire de la distance EMD, on aurait pu s'intéresser aux caractéristiques du bruit et construire une distance adaptée aux déformations observées, un peu de la même façon que le bruit blanc gaussien additif est associé à une distance L_2.

Il est également intéressant de rapprocher le travail 4 de la section 5.6 de la méthodologie générale de démonstration des énoncés scientifiques évoquée dans la section 2.3, à savoir que la preuve est apportée par une accumulation de modèles ou d'observations. Dans le travail 4, au lieu de définir un critère de "singularité" du modèle de mélange, on accumule les images pour montrer que les choses se passent comme si le modèle ne comporte localement qu'une seule composante. Il était prévu de réaliser des simulations en variant davantage la complexité du mélange et les dictionnaires utilisés mais nous avons été limités par les temps de calcul.

Les modèles des images sont compliqués car en plus de modéliser l'intensité du phénomène (blanc ou noir) et sa position, on est amené à modéliser des couleurs, des formes, des textures, etc. Les travaux des chapitres suivants continuent à poser la question de la modélisation dans la démonstration d'énoncés scientifiques. Cependant, en se plaçant dans un espace de phénomènes de dimension 1, on espère aborder des énoncés plus concrets, avec un intérêt en biochimie dans le chapitre 6 voire un intérêt médical dans le chapitre 7.

Chapitre 6

Modèlisation du signal en spectroscopie par résonance paramagnétique électronique

6.1 Contexte

En 2013, lors des réunions de construction du programme Imageries du Vivant, j'ai rencontré Yves Frapart, chimiste au LCBPT / Paris Descartes et responsable de la plateforme de Résonance Paramagnétique Électronique (RPE). Un premier intérêt commun a été le développement de méthodes d'imagerie par RPE, et des méthodes de traitement du signal appropriées. Ce sujet était déjà exploré en collaboration entre le LCBPT et le MAP5 (Frédéric Richard, Sylvain Durand), et j'ai fait partie du jury de thèse de Maud Kerebel (directeur Sylvain Durand) qui a soutenu en 2017.

Un des problèmes non élucidés concerne l'estimation de la qualité des mesures RPE en fonction du niveau de bruit expérimental, et ses conséquences sur les objets observés dans les images et les conclusions que l'on peut en tirer. Ainsi, on cherche à évaluer le niveau de preuve apporté par les données expérimentales recueillies vis-à-vis des énoncés biologiques et chimiques. Dans le domaine de la RPE, le traitement du signal et les hypothèses de modélisation n'ont pas été suffisamment étudiés pour avoir une évaluation claire de la performance des instruments, ni des conditions expérimentales optimales.

Afin de traiter cette question, il a été décidé de revenir aux spectres RPE unidimensionnels à partir desquels sont assemblées les images. Nous avons encadré le stage de M2 de Tran Duc Nghia sur l'estimation des paramètres du spectre RPE et le calcul de leur précision. Ce stage s'est poursuivi en thèse de

doctorat sous la direction de Yves Frapart et moi-même, avec un financement apporté par le programme Imageries du Vivant. La thèse a été soutenue le 20 décembre 2018.

Ce souci de la qualité de la mesure est aussi à comprendre dans le contexte de l'installation d'un appareil de RPE clinique, et de la nécessité d'être sûr du diagnostic vis-à-vis d'une application médicale. L'instrument devait être installé au début de la thèse de Tran Duc Nghia en 2016, il a finalement été inauguré en juin 2019 à la faculté de pharmacie de Paris Descartes.

6.2 Génération du signal RPE

Un instrument de RPE mesure l'effet Zeeman, c'est-à-dire l'énergie absorbée par un électron dans la transition entre deux états quantiques. Cette énergie, égale après simplification à $g_e \beta B$, dépend de l'environnement de l'électron au travers d'un paramètre g_e appelé facteur de Landé effectif et du champ magnétique B, le paramètre β étant une constante appelée magnéton de Bohr. Pour que les électrons effectuent la transition, il faut apporter une énergie $h\nu$ sous la forme d'une onde électromagnétique, dont la fréquence ν est ici de l'ordre du GHz ; ce sont des micro-ondes. L'électron absorbe l'énergie fournie uniquement lorsqu'il y a coïncidence ou résonance :

$$\Delta E = h\nu = g_e \beta B$$

En RPE, ce sont les électrons libres qui réagissent à l'onde électromagnétique. Seules les espèces qui comportent un tel électron génèrent un signal. Il s'agit en particulier d'ions ferreux (Fe^{2+}, ...) et de radicaux libres (NO*), ce qui explique l'intérêt de cette modalité pour l'étude du stress oxydant.

Le dispositif expérimental comporte une cavité de résonance dans laquelle on place l'échantillon (cf Figure 6.1) et dans laquelle on impose une micro-onde stationnaire de fréquence fixée. On fait ensuite varier le champ magnétique B afin de mesurer les valeurs de résonance B_r, ce qui permet d'en déduire $g_e = \frac{h\nu}{\beta B_r}$, et par conséquent des informations sur les électrons et l'échantillon analysé. Ce principe fonctionnement est similaire à la résonance magnétique nucléaire (RMN) dans laquelle on observe la résonance des noyaux des molécules.

6.3 Modèles mathématiques des spectres RPE

En spectrométrie RPE, on s'intéresse à l'énergie absorbée par un échantillon $A(B)$ en fonction du champ magnétique B. L'espace des modèles $\mathcal{M}$ est

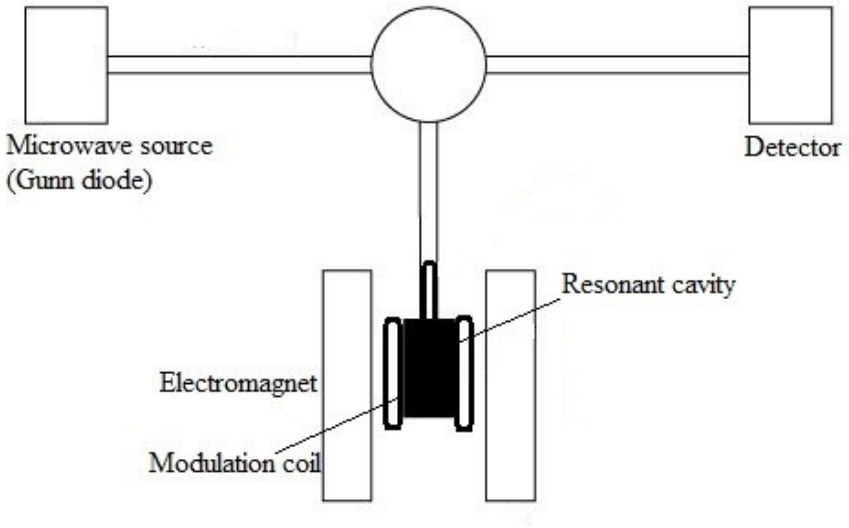


FIGURE 6.1 – Schéma simplifié d'un spectromètre RPE

donc un espace de fonctions $\mathbb{R} \to \mathbb{R}$. Le point de vue du problème direct nous indique le sous-ensemble de modèles et le processus de génération des observations correspondant à des échantillons connus. Les équations de la mécanique quantique permettent de calculer les facteurs de Landé effectifs g_e en fonction des caractéristiques du composé étudié, et d'en déduire les valeurs B_r du champ magnétique où l'on observe une résonance, c'est-à-dire la position des pics dans le spectre RPE théorique. Cependant, il n'existe pas de théorie pour la forme de l'absorption au voisinage de la résonance.

Du point de vue du problème inverse, c'est-à-dire l'identification d'un composé et de ses propriétés à partir d'un spectre RPE, la forme du signal est importante. En particulier dans cette collaboration, le signal est plus ou moins étalé autour d'un point de résonance en fonction de l'environnement chimique au voisinage de l'électron libre, et notamment de la concentration en oxygène du milieu. Ceci conduit à s'intéresser à l'écart-type ou la largeur à mi-hauteur des signaux observés, ou à des familles de lois paramétrées par la position et l'échelle. En l'absence de justification "physique quantique" des hypothèses de modélisation, nous avons choisi de traiter le cas des fonctions Lorentziennes contaminées par un bruit gaussien blanc additif [77], qui correspond à la mesure de l'oxymétrie par "spin-probe" détaillée dans la section suivante.

On considère donc que l'ensemble des modèles étudiés ici est l'ensemble $\mathcal{M}$ des fonctions $\mathbb{R} \to \mathbb{R}$ de la forme

$$A(B) = \frac{C}{\pi \mathrm{FWHM}/2(1 + (\frac{B-B_r}{\mathrm{FWHM}/2})^2)}$$

où C est l'intégrale du signal ou surface totale sous la courbe, B_r est l'unique valeur de résonance et FWHM est la largeur à mi-hauteur du signal.

Le modèle est tracé sur la figure 6.2 par les fonctions suivantes. En particulier, on vérifie la paramétrisation, C est l'intégrale de l'absorption et FWHM est bien la largeur à mi-hauteur.

Code chunk 6 : «rpe»

```
mol=list(Br=40, fwhm=0.4, C=0.1)                # paramètres de l'échantillon
mac=list(B=seq(38,42,by=0.001),ma=0.3,v=0.002) # paramètres de l'instrument
rabsorption = function(mac,mol)
  list(y=mol$C / (pi*(mol$fwhm/2) * (1 + ((mac$B-mol$Br)/(mol$fwhm/2) )^2)),
       B=mac$B,ma=mac$ma,v=mac$v)

y_abs = rabsorption(mac,mol)
pdf("lorentzAbsorption.pdf",width=8,height=6)
    plot(mac$B,y_abs$y,typ="l",xlab="B (mT)",ylab="arbitrary unit")
    ymax = max(y_abs$y)
    arrows(mol$Br-mol$fwhm/2,ymax/2,mol$Br+mol$fwhm/2,ymax/2,
           lwd=3,code=3,length=0.1)
    text(mol$Br+mol$fwhm/2,ymax/2,"FWHM",pos=4)
    legend("topr",
           c("Absorption",paste("mol$C = ",mol$C),
             paste("Intégrale = ",signif(sum(y_abs$y)*mean(diff(mac$B))))),
           col=c(1,0,0),lty=c(1,0,0),lwd=c(1,0,0))
i = dev.off()
```

Interpret with `R`

L'instrument ne fournit pas directement une mesure de cette absorption $A(B)$, mais procède par différences finies. On observe donc un signal avec un bruit gaussien additif :

$$Y(B) = A(B + \mathrm{MA}/2) - A(B - \mathrm{MA}/2) + \mathcal{N}(0, \nu)$$

où le paramètre MA est appelée modulation d'amplitude.

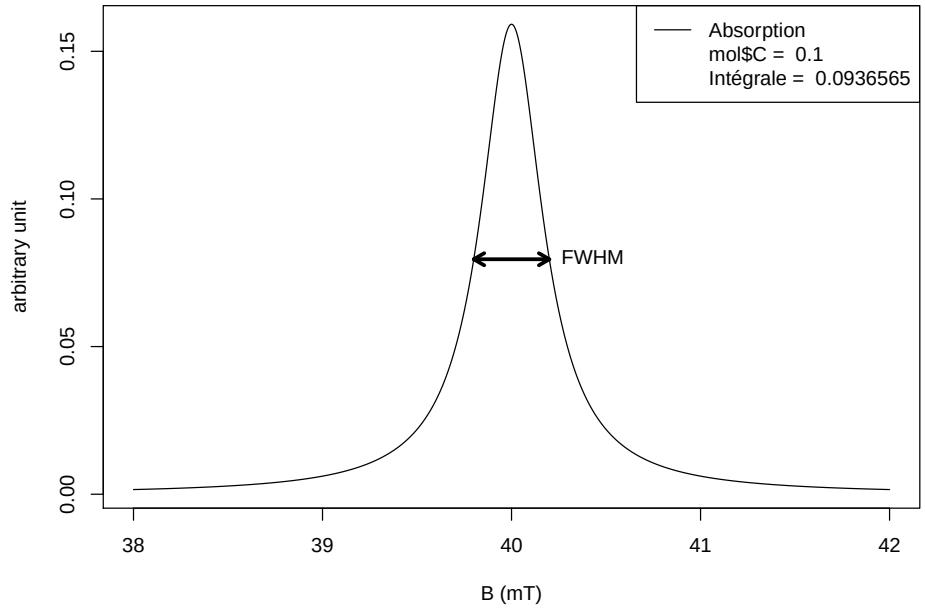


FIGURE 6.2 – Spectre d'aborption théorique selon de modèle de Lorentz

Code chunk 7 : «rpe (part 2)»

```
rsignal = function(mac,mol,ftype=dcauchy)
  # reads machine and molecular parameters, and outputs observations
  list(y=mol$C*ftype(mac$B+mac$ma/2,mol$Br,mol$fwhm/2)
         - mol$C*ftype(mac$B-mac$ma/2,mol$Br,mol$fwhm/2)
         + rnorm(length(mac$B),0,mac$v),
         B=mac$B,ma=mac$ma,v=mac$v)
```

Interpret with R

```
```

Dans le domaine de la RPE, on considère habituellement le cas MA proche de 0, appelé cas sous-modulé, ce qui correspond à MA $<$ FWHM/10 en pratique [20, 43]. On peut alors approximer $Y(B)$ par la dérivée de la fonction de Lorentz, $Y(B) \sim \text{MA}.A'(B)$, et l'on peut mesurer les paramètres C et FWHM par proportionnalité au travers de l'amplitude pic-à-pic I et de la largeur pic-à-pic P2P (cf Figure 6.3).

Code chunk 8 : «rpe (part 3)»

```
y_epr1  = rsignal(c(ma=mol$fwhm/10, mac),mol)
pdf("lorentzRPE1.pdf",width=8,height=6)
plot (mac$B,y_epr1$y,typ="l",xlab="B (mT)",ylab="arbitrary unit")
ymax = max(y_epr1$y); ymin = min(y_epr1$y);
p2p2 = mol$fwhm/sqrt(3)/2; h = 0.03
arrows(mol$Br-p2p2,h,mol$Br+p2p2,h,lwd=2,code=3,length=8,angle=90,lty=2,col=2)
arrows(38,ymax,38,ymin,lwd=2,code=3,length=3.3,angle=90,lty=2,col=3)
legend("topr",
    c("Spectre RPE", "Intensité pic-à-pic", paste("ma = fwhm/10 =",y_epr1$ma),
    paste("P2P ~ fwhm/sqrt(3) = ",signif(mol$fwhm/sqrt(3)))),
    col=c(1,3,0,2),lwd=c(1,2,0,2),lty=c(1,2,0,2))
i = dev.off()
```

Interpret with `R`

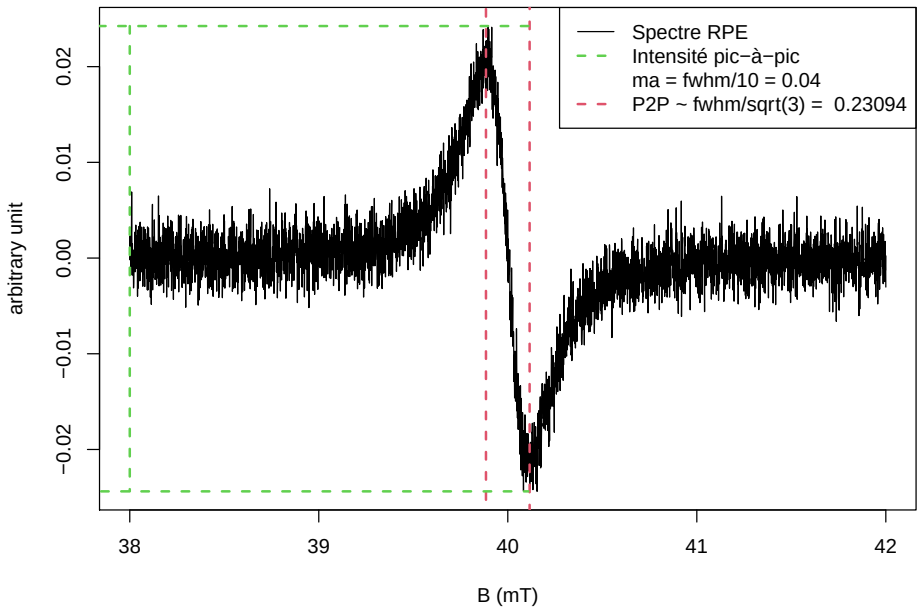


Figure 6.3 – Spectre RPE simulé dans le cas sous-modulé.

On comprend immédiatement que l'intensité pic-à-pic observée I est à peu près proportionnelle à la modulation d'amplitude MA, et que l'on a intérêt à s'écarter de 0 pour augmenter le rapport signal sur bruit. Effectivement, un certain nombre d'expériences sont réalisées avec MA $\sim$ FWHM/3, notamment in vitro et in vivo pour compenser le niveau de bruit élevé, sur la plateforme de RPE de Paris Descartes et dans d'autres laboratoires.

En contrepartie de l'augmentation du rapport signal-sur-bruit, on s'écarte de la forme idéale de la dérivée de la fonction de Lorentz, et les indicateurs simples (intensité pic-à-pic et largeur pic-à-pic) ne sont plus proportionnels à C et FWHM. Or, la plupart du temps, ces effets sont négligés par l'utilisateur.

Code chunk 9 : «rpe (part 4)»

```
y_epr2  = rsignal(c(ma=2*mol$fwhm, mac),mol)
pdf("lorentzRPE2.pdf",width=8,height=6)
plot (mac$B,y_epr2$y,typ="l",xlab="B (mT)",ylab="arbitrary unit")
ymax = max(y_epr2$y); ymin = min(y_epr2$y);
p2p2 = y_epr2$ma/2; h = 0.3
arrows(mol$Br-p2p2,h,mol$Br+p2p2,h,lwd=2,code=3,length=8,angle=90,lty=2,col=2)
arrows(38,ymax,38,ymin,  lwd=2,code=3,length=3.8,angle=90,lty=2,col=3)
legend("topr",c("Spectre RPE", "Intensité pic-à-pic",
       paste("ma = 2fwhm =",y_epr2$ma), paste("P2P ~ ma = ",y_epr2$ma)),
       col=c(1,3,0,2),lwd=c(1,2,0,2),lty=c(1,2,0,2))
i = dev.off()
```

Interpret with `R`

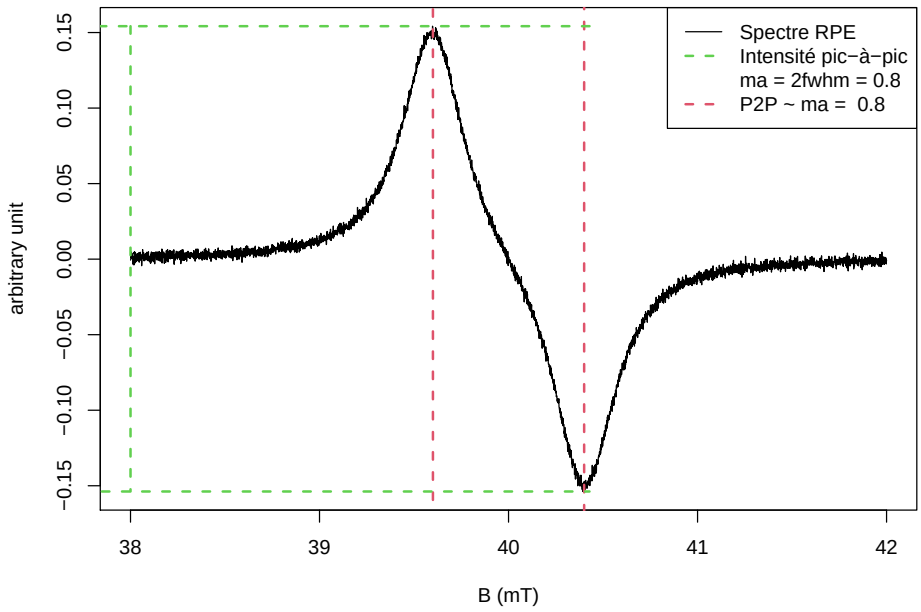


FIGURE 6.4 – Spectre RPE simulé dans le cas MA = 2FWHM.

D'après [73, 96], le rapport signal-sur-bruit optimal est obtenu pour des valeurs MA = FWHM, mais la justification est incomplète et ceci ne reflète pas nécessairement l'estimateur paramétrique de précision maximale. Les travaux menés ci-après visent à améliorer l'estimation des paramètres, C et FWHM en particulier, au travers du traitement du signal et du choix des paramètres expérimentaux, et notamment la modulation d'amplitude MA.

Il est naturel d'utiliser la fonction gaussienne pour modéliser la forme du signal, mais ce n'est pas le modèle le plus utilisé dans le domaine. En fait, les travaux de Robinson [40, 80] indiquent que le signal est à l'origine une fonction de Lorentz à valeur complexe, et que l'absorption A correspond à sa partie imaginaire :

$$S(B) = \frac{-1}{\pi} \frac{1}{B - B_r + i\mathrm{FWHM}/2}$$
$$A(B) = \frac{-1}{\pi} \frac{\mathrm{FWHM}/2}{(B - B_r)^2 + (\mathrm{FWHM}/2)^2}$$

Le modèle de Robinson inclut des termes de déformation en fonction du rapport entre la fréquence de la modulation d'amplitude et du rapport gyromagnétique de l'électron qui ont été négligés dans ce travail.

En pratique, on observe souvent des déformations de ce modèle, qui ont été attribuées à la superposition de signaux avec des points de résonance B_r proches [75]. Ceci a conduit aux modèles dits de Voigt, qui correspondent à la convolution d'une fonction de Lorentz par un noyau gaussien [25, 97]. Ces fonctions ont une forme intermédiaire entre les fonctions de Lorentz et les fonctions gaussiennes. Dans nos données, on observe peu de déformation et l'on a décidé de se restreindre au modèle de Lorentz "pur".

6.4 Oxymétrie et stress oxydant

Dans les travaux menés avec Yves Frapart, on s'est intéressé en particulier aux propriétés du tri-aryl-méthyl (TAM), molécule qui fait partie d'une famille de composés développés en collaboration avec SANOFI. Le signal RPE observé est une raie simple bien contrastée qui facilite la mesure de la concentration en oxygène du milieu et la mesure du stress oxydant[23, 41, 68].

Pour la mesure de la concentration en oxygène, on utilise la dépendance linéaire entre la largeur de raie et la teneur en oxygène (pO2) qui a été observée

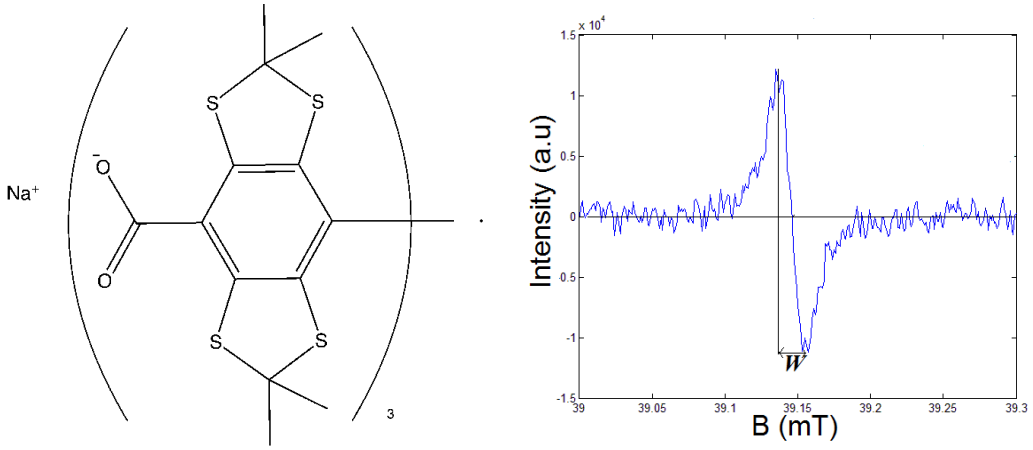


FIGURE 6.5 – Structure de la molécule TAM, et signal RPE correspondant.

et calibrée dans des travaux précédents [27]. Ceci a été démontré principalement pour la largeur pic-à-pic P2P, dans le cas sous-modulé. Or les conditions expérimentales in vivo nécessitent d'améliorer le rapport signal-sur-bruit, et d'utiliser une modulation d'amplitude MA non négligeable. Nous souhaitons par exemple montrer que l'on peut mesurer la teneur en oxygène d'après FWHM, qui est indépendante de la modulation d'amplitude, plutôt que la largeur pic-à-pic P2P.

Pour la mesure du stress oxydant, on utilise l'amplitude C du signal RPE provenant de la molécule TAM. En l'absence de stress oxydant, cette molécule est stable donc l'amplitude du signal RPE est constante au cours du temps. Il a été montré dans les travaux de [39] que le signal baisse uniquement en présence de superoxide, et que les processus biochimiques impliqués suivent une cinétique d'ordre 1. On peut donc conclure à la présence de stress oxydant quand on observe une décroisance du signal au cours du temps. Cette propriété est utilisée pour le travail 3, Section 6.7.

6.5 Travail 1 : Cas sous-modulé

Dans le cas sous-modulé, on considère que le signal $Y(B)$ provient de la dérivée de la fonction de Lorentz :

$$Y(B) = A'(B) + \mathcal{N}(0, \nu)$$

où $\mathcal{N}$ est un bruit additif gaussien d'écart-type ν. On a donc un problème d'estimation paramétrique à 4 variables $(B_r, C, \mathrm{FWHM}, \nu)$.

On calcule la vraisemblance $L(B_r, C, \mathrm{FWHM}, \nu) = \mathbb{P}(y|B_r, C, \mathrm{FWHM}, \nu)$ et

on implémente en Matlab l'estimateur du maximum de vraisemblance.

$$\{\hat{C}, \hat{B}_r, \hat{\text{FWHM}}, \hat{v}\} = \arg\max \ln L$$
$$\ln L = -\frac{n}{2}\ln(2\pi\nu^2) - \frac{1}{2\nu^2}\sum_{k=1}^{n}\left(Y(B_k) - A'(B_k)\right)^2$$

La borne de Cramer-Rao [98] fournit une borne inférieure de la variance de l'estimateur, et l'on se sert de cette borne comme d'une estimation. Pour justifier l'utilisation pratique de cette approximation, on vérifie que l'on est bien en régime asymptotique pour la spectroscopie RPE, d'abord sur des spectres simulés sous le modèle de Lorentz, puis en réalisant des séries de spectres correspondant au même échantillon (Figures 6.6 et 6.7).

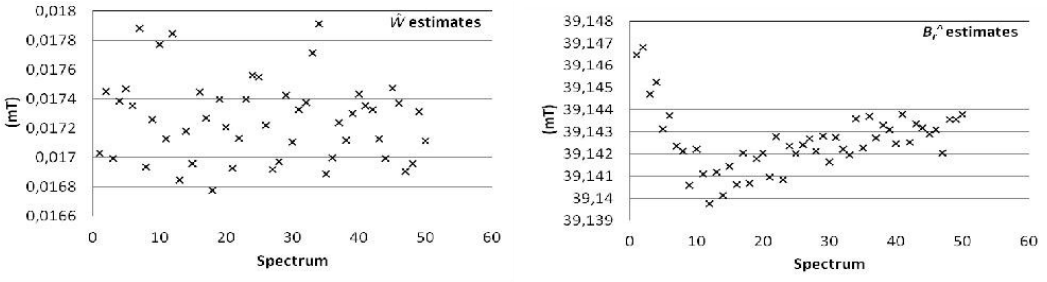


Figure 6.6 – Valeurs $\hat{\text{FWHM}}$ et $\hat{B}_r$ estimées pour 50 spectres RPE du même échantillon de TAM avec un rapport SNR de 14dB environ.

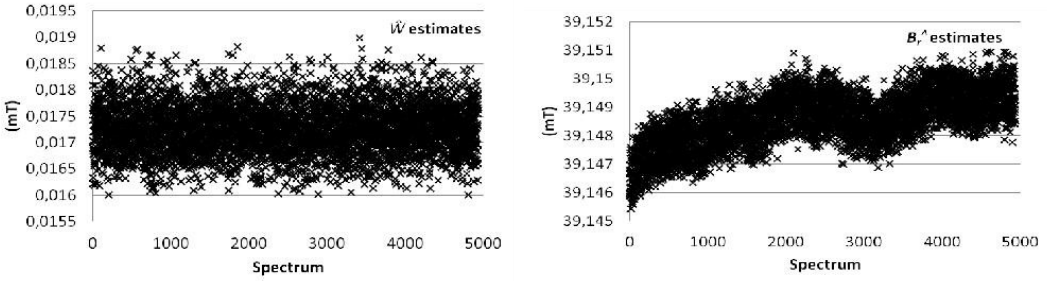


Figure 6.7 – Valeurs $\hat{\text{FWHM}}$ et $\hat{B}_r$ estimées pour 5000 spectres RPE du même échantillon de TAM avec un rapport SNR de 7dB environ.

Les résultats montrent une bonne adéquation de l'erreur prédite d'après la borne de Cramer-Rao et le calcul d'erreur effectué sur une série de spectres du même échantillon (cf Figure 6.8). En particulier pour l'application en chimie, il est donc possible d'estimer la précision des paramètres à partir d'un seul spectre, sans réaliser une série d'expériences. Ces résultats ont été publiés dans [13]. On remarque également une instabilité de la valeur $\hat{B}_r$. Celle-ci est attribuable à l'instrument car il s'agit de la valeur de résonance de la molécule, mais le phénomène n'a pas été davantage investigué.

W (mT)	***SNR (dB)***	***Predicted standard derivation of W (mT)***	***Observed standard derivation of W estimation (mT)***
0.01726	14	0.00021(1.2443%)	0.00027(1.5643%)
0.01727	7	0.00052(3.0110%)	0.00046(2.6636%)

FIGURE 6.8 – Comparaison entre l'écart-type prédit par la borne de Cramer-Rao et la valeur observée sur la série de spectres RPE.

6.6 Travail 2 : Cas sur-modulé

On reproduit la même démarche que dans la section précédente pour le cas sur-modulé. On considère que le signal $Y(B)$ provient d'une différence finie de la fonction d'absorption :

$$Y(B) = A(B + \mathrm{MA}/2) - A(B - \mathrm{MA}/2) + \mathcal{N}(0, \nu)$$

On a donc un problème d'estimation paramétrique portant sur les 5 variables $(B_r, C, \mathrm{FWHM}, \mathrm{MA}, \nu)$.

On calcule la vraisemblance et on implémente en Matlab l'estimateur du maximum de vraisemblance.

$$\{\hat{\mathrm{MA}}, \hat{C}, \hat{B}_r, \hat{\mathrm{FWHM}}, \hat{v}\} = \arg\max \ln L$$

$$\ln L = -\frac{n}{2}\ln(2\pi\nu^2) - \frac{1}{2\nu^2}\sum_{k=1}^{n}\left(Y(B_k) - (A(B_k + \mathrm{MA}/2) - A(B_k - \mathrm{MA}/2))\right)^2$$

On se sert de la même façon de la borne de Cramer-Rao comme approximation de la variance des estimateurs. Pour justifier cette approximation en pratique, nous avons de nouveau réalisé des séries de spectres RPE simulés (cf Table 6.1 et Table 6.2) et des séries de spectres RPE in vitro (cf Table 6.3). En particulier, on retrouve que le rapport signal-sur-bruit (ratio intensité pic-à-pic sur v) est maximal pour MA = FWHM, mais que la variance des estimateurs est minimale autour de MA = 2.FWHM (cf Table 6.3). Ces travaux ont été publiés dans [18].

MA (mT)	FWHM (mT)	Predicted error (%)	Observed error (%)
0.01	0.01	0.15	0.16
0.02	**0.01**	**0.12**	**0.12**
0.04	0.01	0.13	0.13
0.08	0.01	0.15	0.15
0.01	0.02	0.84	0.84
0.02	0.02	0.44	0.43
0.04	**0.02**	**0.35**	**0.34**
0.08	0.02	0.36	0.38
0.01	0.04	8.32	8.56
0.02	0.04	2.39	2.33
0.04	0.04	1.24	1.24
0.08	**0.04**	**0.99**	**1.00**
0.01	0.08	264.68	76.20
0.02	0.08	25.62	34.56
0.04	0.08	6.72	6.80
0.08	0.08	3.51	3.62

Table 6.1 – Erreurs d'estimation sur $\hat{C}$. Erreur prédite et erreur observée sur une série de spectres RPE simulés. Les valeurs correspondant à l'erreur minimale ont été surlignées.

MA (mT)	FWHM (mT)	Predicted error (%)	Observed error (%)
0.01	0.01	0.086	0.087
0.02	**0.01**	**0.082**	**0.083**
0.04	0.01	0.094	0.094
0.08	0.01	0.118	0.118
0.01	0.02	0.321	0.322
0.02	0.02	0.243	0.238
0.04	**0.02**	**0.234**	**0.231**
0.08	0.02	0.267	0.283
0.01	0.04	1.532	1.552
0.02	0.04	0.921	0.896
0.04	0.04	0.691	0.687
0.08	**0.04**	**0.663**	**0.672**
0.01	0.08	8.166	6.632
0.02	0.08	4.327	4.533
0.04	0.08	2.602	2.599
0.08	0.08	1.953	2.017

TABLE 6.2 – Erreurs d'estimation sur $\hat{\text{FWHM}}$. Erreur prédite et erreur observée sur une série de spectres RPE simulés. Les valeurs correspondant à l'erreur minimale ont été surlignées.

MA	0.02 mT	0.04 mT	0.08 mT
Estimated $\hat{C}$	0.00220	0.00224	0.00245
Predicted precision	0.0000285	0.0000221	0.0000267
Observed precision	0.0000335	0.0000236	0.0000181
Estimated $\hat{\text{FWHM}}$ (mT)	0.0252	0.0254	0.0275
Predicted precision	0.000161	0.000157	0.000210
Observed precision	0.000172	0.000154	0.000131
Estimated $\hat{\text{MA}}$ (mT)	0.0190	0.0375	0.0696
Predicted precision	0.0000600	0.0000482	0.0000678
Observed precision	0.0001152	0.0000994	0.0000844

TABLE 6.3 – Estimation de $\hat{C}$, $\hat{\text{FWHM}}$, $\hat{\text{MA}}$ en fonction de la valeur MA paramétrée dans l'instrument pour un échantillon de TAM. Il y a une bonne adéquation entre les précisions observées et la précision prédite.

6.7 Travail 3 : Mesure du stress oxydant dans l'oeil du rat

Il n'existe pas à l'heure actuelle de méthode permettant de mesurer directement le stress oxydant. Or ce phénomène biologique est associé à un certain nombre de maladies inflammatoires telles que le diabète ou certaines rétinopathies. Dans cette section, on utilise la sonde moléculaire TAM afin d'en déduire le niveau de stress oxydant dans l'oeil du rat. Pour cela, la molécule est injectée dans l'oeil du rat anesthésié, et l'on place l'animal entier dans la cavité de résonance de l'instrument de RPE.

On dispose de trois groupes principaux de rats : un groupe contrôle, un groupe DMSO et un groupe roténone :

- Le groupe contrôle ne reçoit que l'injection de la molécule TAM (et l'anesthésiant) juste avant la mesure des spectres RPE.
- Le groupe roténone reçoit une molécule, la roténone, dont l'effet est supposé induire du stress oxydant sur plusieurs jours. Certains rats sont mesurés deux jours après l'injection (rats R2), d'autres après huit jours (rats R8). La roténone étant sous forme solide, il est nécessaire de la solubiliser dans le solvant DMSO avant injection. Le groupe de rats roténone reçoit donc les deux produits, DMSO et roténone.
- Le groupe DMSO sert de contrôle et permet de distinguer l'effet du DMSO de celui de la roténone. Là aussi, certains rats sont mesurés deux jours après l'injections (rats D2) et d'autres après huit jours (rats D8).

Comme expliqué dans la section 6.4, la molécule TAM est stable, donc l'amplitude C du signal correspondant doit être constante au cours du temps. Si on observe une variation, cela doit être attribué à l'évacuation de la molécule TAM, à sa disparition selon le mécanisme indiqué dans [39] ou bien à la présence d'une autre espèce chimique. Dans ces expériences, l'oeil du rat étant un milieu fermé, il n'y a pas d'évacuation de la sonde moléculaire, et le signal provient essentiellement du TAM injecté. On s'attend donc à une diminution du signal liée au stress oxydant avec une échelle de temps de l'ordre de la demi-heure.

Compte tenu du mécanisme de la réaction, nous avons proposé de modéliser la décroissance du signal dans le spectre RPE par un modèle d'ordre un, c'est-à-dire une régression exponentielle. On étudie en particulier la valeur du paramètre α qui indique si le modèle est décroissant $\alpha > 0$ ou non.

$$I(t) = C.\exp^{-\alpha t} + \eta$$

où $I(t)$ représente l'amplitude ou l'intensité pic-à-pic du spectre RPE mesuré à

l'instant t. Les estimateurs $\hat{C}$ et $\hat{\alpha}$ sont obtenus en considérant le logarithme :

$$\ln I(t) = \ln C - \alpha t + \nu$$

Ceci conduit donc à un modèle linéaire avec ν un bruit gaussien additif.

Dans un premier temps, nous avons appliqué ce modèle à l'amplitude $C(t)$ ou aire sous la courbe. La figure 6.9 présente les valeurs de $\hat{\alpha}$ obtenues en fonction du groupe de rats. Le modèle linéaire permet de calculer les incertitudes sur $\hat{\alpha}$. Dans ces données, elles sont négligeables par rapport à la variance inter-individus. Le graphique indique que l'amplitude du signal est à peu près stable dans le groupe contrôle, et qu'elle est décroissante pour les autres groupes. Ces résultats suggèrent que l'effet est produit par le DMSO, indépendamment de la présence de roténone. Le DMSO provoquerait ainsi un stress oxydant, ce qui n'est pas l'effet attendu.

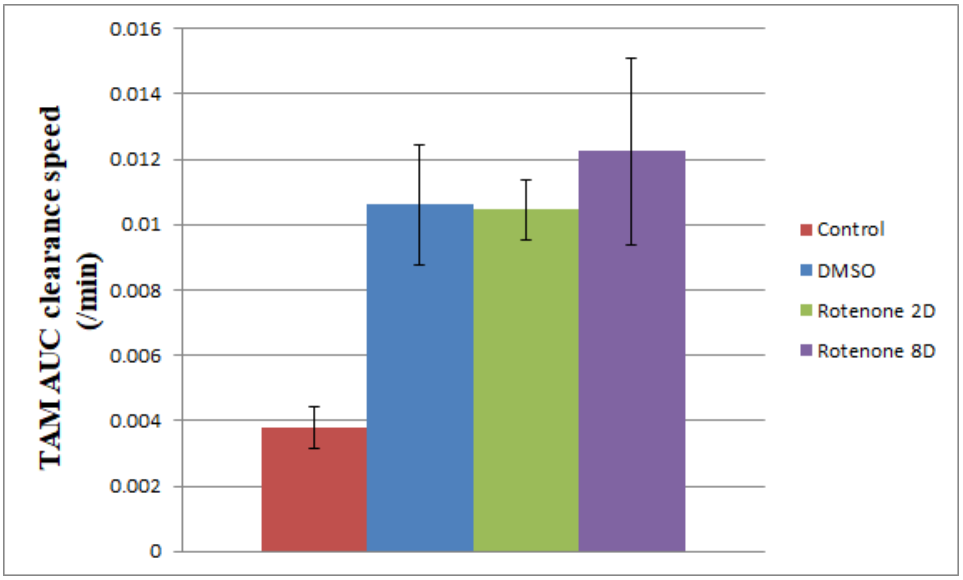


FIGURE 6.9 – Valeur de $\hat{\alpha}$ en fonction du groupe de rats. On observe que le signal décroît dans tous les groupes, et que cette décroissance est plus rapide et similaire entre les groupes DMSO et roténone.

Nous avons également appliqué ce modèle à l'intensité pic-à-pic. La figure 6.10 présente les valeurs de $\hat{\alpha}$ obtenues en fonction du groupe de rats. Là aussi, les incertitudes sont négligeables par rapport à la variabilité inter-individus. Les résultats présentent des valeurs de $\hat{\alpha}$ négatives, c'est-à-dire une intensité pic-à-pic croissante au cours du temps, ce qui n'est pas attendu, sauf si la largeur de raie FWHM diminue en même temps. Cette hypothèse n'a pas pu être explorée

durant la thèse, les estimateurs décrits dans le travail 2 n'étant pas encore finalisés au moment de l'exploitation des données expérimentales.

D'autre part, le modèle de régression exponentielle sur l'intensité pic-à-pic fait une différence entre les groupes "roténone" et les groupes contrôle et DMSO. Ceci pourrait suggérer que la roténone a bien un effet, mais qui serait plutôt sur la largeur de raie (donc sur l'oxygénation de l'oeil) plutôt que sur le stress oxydant. Il n'a pas été possible de creuser davantage cette direction pour le moment.

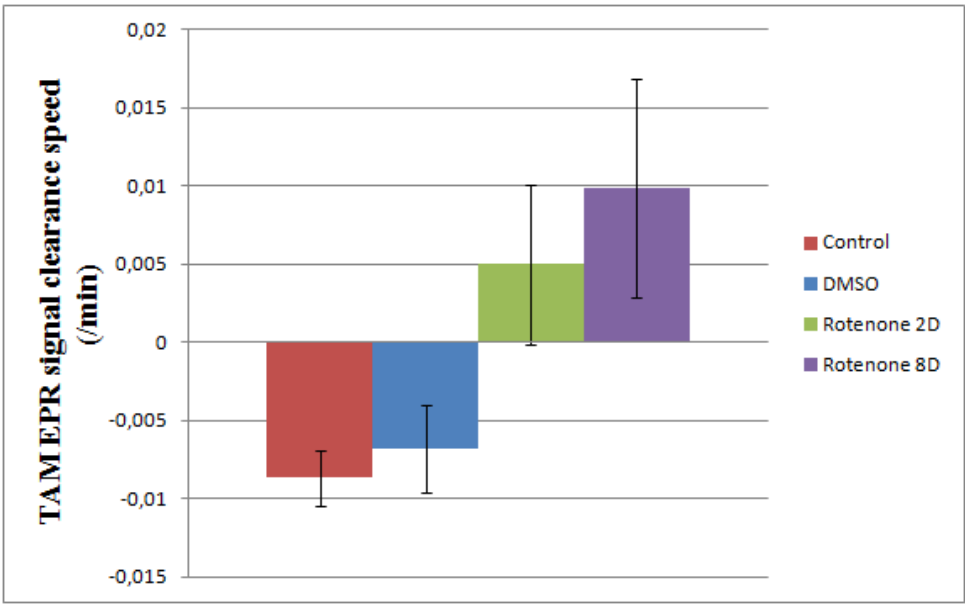


Figure 6.10 – Comparaison entre l'écart-type prédit par la borne de Cramer-Rao et la valeur observée sur la série de spectres RPE.

6.8 Conclusion et perspectives

Ce chapitre présente un certain nombre de travaux en traitement du signal pour la spectroscopie RPE. Dans ce domaine, l'analyse des spectres RPE obtenus nécessite à la fois la modélisation des signaux et du bruit observés.

Dans la thèse de Nghia Tran Duc, nous avons principalement travaillé sur la modélisation du signal, qui présente l'originalité d'être obtenu par un processus de différence contrôlable (modulation d'amplitude). En utilisant la borne de Cramer-Rao comme une estimation de la variance, nous montrons que l'on peut obtenir des valeurs approchées de la précision des estimateurs, et que

ces valeurs approchées fonctionnent en pratique. Nous envisageons de montrer l'intérêt de cette approche à l'intention de la communauté RPE en comparant la précision des différentes méthodes existantes sur plusieurs appareils. Par la suite, on pourra mettre en place des stratégies plus poussées d'acquisition, en tirant parti des possibilités de programmation du nouvel instrument de RPE clinique. Ainsi, l'espace des signaux $\mathcal{M}$ est de dimension très réduite, et il serait souhaitable de proposer des méthodes parcimonieuses pour exploiter au mieux cette structure.

Pour ce qui est de la modélisation du bruit, le cadre général présenté dans le chapitre 2 indique les grandes pistes suivantes :

— par accumulation d'observations, on peut justifier que la distribution du bruit ressemble à une certaine loi (par exemple une loi gaussienne). On estime ainsi la distribution observée du bruit dans l'ensemble des lois possibles.
— pour démontrer qu'il s'agit bien d'une loi gaussienne, il faut pouvoir séparer les distributions gaussiennes des autres types de distributions dans l'espace des lois de probabilités. Comme dans beaucoup d'autres situations, cela sera difficile.
— cependant, je pense que certaines propriétés du bruit sont accessibles à une preuve empirique. Par exemple, on a supposé dans ce chapitre que le bruit gaussien s'applique au spectre RPE Y et non à l'absorption A. Ces deux cas sont distinguables si on observe le comportement du bruit en fonction du paramètre MA, et plus particulièrement à quelle échelle se font les corrélations dans le bruit.

Le travail d'analyse des données expérimentales sur le stress oxydant a été présenté lors de la soutenance de thèse de Tran Duc Nghia, avant soumission pour publication, et a soulevé un certain nombre de questions non résolues. En particulier, la nouvelle définition de la largeur de raie en terme de FWHM nécessite des justifications supplémentaire avant son adoption dans le domaine de la chimie et il est important que l'on arrive à comprendre les différences entre la valeur indiquée par l'instrument et celle estimée sur le spectre par le modèle.

D'autre part, on peut observer des distortions non prédites par le modèle lorsque l'on utilise une modulation d'amplitude élevée dans ces expériences in vivo. Nous n'avons pas encore élucidé s'il s'agit d'une conséquence de la biochimie des échantillons, de la mesure par l'instrument de RPE, ou bien si le modèle que nous avons choisi doit être étendu.

Chapitre 7

Modélisation appliquée aux maladies inflammatoires de l'intestin

Le LAGA est impliqué depuis 2011 dans le LabEx Inflamex, en grande partie sous l'impulsion de Hatem Zaag. Cette structure regroupe des équipes de recherche de toutes les disciplines autour des maladies inflammatoires. Dans ce cadre, une collaboration plus particulière a été initiée autour des maladies inflammatoires de l'intestin, entre l'équipe de gastro-entérologie du groupement hospitalier Bichat-Beaujon (Y. Bouhnik, X. Treton, E. Ogier-Denis) et l'axe math-bio du LAGA (H. Zaag, I. Morilla, V. Milisic).

En 2016, les gastro-entérologues ont proposé aux mathématiciens de l'équipe image (S. Li-Thiao-Te, J. Chaussard, W. Jiaping, F. Dibos) de se rencontrer car les examens médicaux dans les maladies inflammatoires de l'intestin sont basés sur une vidéo de la paroi du colon. J'ai représenté le laboratoire dans les différentes demandes de financement qui ont été effectuées autour de cette thématique, et en particulier les dossiers d'appel à projet RHU (Réseaux Hospitalo-Universitaires) en 2017, 2018 et H2020. Ces dépôts de projet pour des financement de 15 à 30 M€, dont l'équivalent de 3 bourses de thèse et 6 ans de postdoctorat pour le LAGA, ont représenté une part importante de mon temps de recherche à ce moment là. Même si les financements n'ont pas été accordés, et même si la crise sanitaire a freiné un certain nombre de choses, le réseau de recherche reste dynamique et continue à développer de nouveaux projets.

Dans le cadre de cette collaboration, nous avons obtenu les financements suivants :

— la gratification du stage de M2 de Mlle Chenyu Zha, effectué en 2016 sous la direction de John Chaussard, accordée par le programme interdisciplinaire Imageries du Vivant,
— 15000 euros pour le projet "IBDimage : Elaboration d'un nouvel algorithme de mesure automatique des lésions muqueuses de la rectocolite hémorragique durant la coloscopie", accordés par l'Association François Aupetit (association de malades), et qui a permis de financer la gratification de stage de Safaa Al Ali en 2018,
— la bourse de thèse de Safaa Al Ali, accordée par le DIM MathInnov de la région Ile-de-France, et gérée par Fondation des Sciences Mathématiques de Paris. La thèse a débuté le 1er octobre 2018 sous la direction de Hatem Zaag, avec pour co-encadrants John Chaussard et Sébastien Li-Thiao-Té.

Dans ce chapitre, nous donnons quelques éléments de contexte sur les maladies inflammatoires de l'intestin, et plus particulièrement sur la rectocolite hémorragique, avant de présenter les questions de modélisation qui sont abordées dans la thèse de Safaa Al Ali.

7.1 Maladies inflammatoires de l'intestin

La rectocolite hémorragique (RCH) est une maladie inflammatoire chronique de l'intestin, résultant d'une réponse excessive des défenses naturelles du système immunitaire digestif, qui touche environ 100 000 personnes en France et 7000 nouveaux cas par an [59]. Elle se manifeste par des lésions telles que saignements et ulcères au niveau du rectum et du côlon. C'est une maladie incurable à l'heure actuelle, caractérisée par des poussées inflammatoires d'intensité variable, entrecoupées de périodes de rémission (périodes sans symptômes). Elle expose le patient à un risque de cancer du côlon plus élevé que dans la population générale et à l'ablation de l'organe (colectomie). Les traitements proposés actuellement visent à contrôler les douleurs et à réduire la fréquence et la durée des poussées, et ainsi soulager les symptômes[58].

Le patient malade est décrit par un certain nombre d'indicateurs biologiques et cliniques :

— l'interrogatoire et l'examen du patient par le médecin,
— le dosage de la calprotectine dans les selles et de la protéine CRP, qui sont des marqueurs de l'inflammation,
— la vidéo de l'examen de coloscopie,
— les images histologiques correspondant aux biopsies prélevées durant

l'examen de coloscopie,
— les analyses ADN et ARN du microbiote intestinal,
— les analyses des ARN exprimés par l'hôte dans l'intestin, etc.

Ces indicateurs permettent au gastro-entérologue d'évaluer l'état du patient (sévérité de la maladie), d'envisager son évolution probable (pronostic), de choisir le traitement le mieux adapté, et d'évaluer la réponse au traitement proposé. Dans ce chapitre, l'espace général des modèles considérés est donc un ensemble de fonctions du type : patient $\rightarrow$ objectif, l'objectif pouvant être décrit dans différents espaces suivant la situation : score numérique de sévérité (cf Section 7.5), vitesse de propagation de l'inflammation (cf Section 7.6), choix d'un traitement parmi n, etc.

Il est à noter qu'en l'état actuel des connaissances, aucun de ces indicateurs n'est spatial, c'est-à-dire que les informations ne sont jamais localisées à une position anatomique précise. De la même façon, l'étendue des lésions est actuellement ignorée en pratique médicale, faute de méthode validée permettant d'analyser cette information. Les acteurs du domaine sont bien conscients de l'intérêt de ces informations, par exemple de leur association avec la survenue de complications telles que le cancer du côlon [26]. Il manque cependant des méthodes d'analyse validées. Les travaux de la thèse de Safaa Al Ali ont donc pour objectif de proposer des méthodes pour l'analyse de la répartition et l'étendue des lésions dont on peut envisager une validation clinique à moyen terme.

Pour les différents appels à projets auxquels nous avons participé depuis 2017, il a été envisagé à chaque fois de réaliser l'ensemble des analyses biologiques et d'analyser conjointement l'ensemble des indicateurs observés. Dans le cadre de la thèse de Safaa, on s'est restreint aux vidéos de coloscopie. Le patient est donc décrit par une vidéo [1], c'est-à-dire par une suite d'images indexée par le temps. Cette description étant trop riche, on se restreindra à $\mathbb{R}$ ou $\mathbb{R}^k$ en fonction de la position dans la vidéo en calculant un ou plusieurs indicateurs.

7.2 Vidéos endoscopiques

Pour les patients atteints de la RCH, la coloscopie est l'examen de référence permettant d'évaluer la sévérité de la maladie et la gravité des lésions du colon. C'est un examen de routine, pratiqué à intervalles réguliers durant le suivi du patient. Il consiste à introduire une caméra fixée au bout d'un tube articulé

1. Nous ne disposons que d'une seule coloscopie par patient. Il n'y a donc qu'une seule observation temporelle.

dans le colon (Figure 7.1). Le gastro-entérologue peut contrôler la progression de la caméra, et prélever une biopsie si nécessaire.

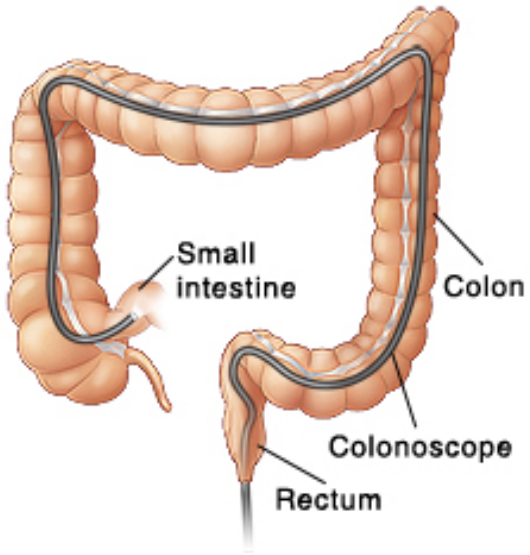


Figure 7.1 – Schéma d'une coloscopie. La caméra fixée sur un tube flexible est introduite dans le colon.

Les lésions observées dans les vidéos concernent principalement
— la perte de visibilité de la trame vasculaire, c'est-à-dire la disparition des vaisseaux sanguins associée à la formation de tissus fibreux empêchant l'absorption des nutriments,
— l'inflammation et les saignements, qui correspondent à des zones rouges de la paroi intestinale,
— les ulcères, des creusements de la paroi qui apparaissent en blanc.
Ces différentes lésions sont présentées sur la Figure 7.2.

Dans le cadre de la collaboration avec le service de gastro-entérologie de l'hôpital Bichat-Beaujon, nous nous sommes concentrés sur la RCH et les vidéos de coloscopie correspondantes. Une base de données comportant 37 patients et une interface permettant l'annotation de ces vidéos a été développée par John Chaussard, sur la base du logiciel Vatic [95].

À partir de ces données, nous avons considéré trois types de modèles :
— des modèles de l'aspect des lésions dans le travail 1 (cf Section 7.3),
— des modèles de la sévérité de la RCH dans le travail 3 (cf Section 7.5),
— des modèles de la répartition des lésions pour un patient dans le travail 4 (cf Section 7.6).

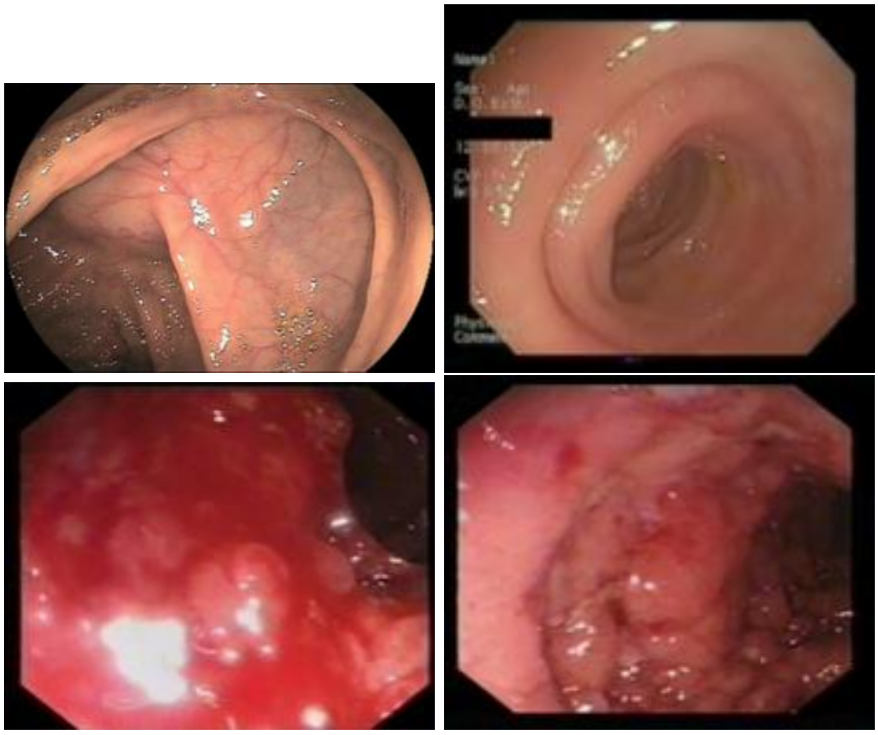


FIGURE 7.2 – Aspect de la paroi instestinale lors d'une coloscopie (paroi saine, perte de la trame vasculaire ou fibrose, inflammation et saignements, ulcères).

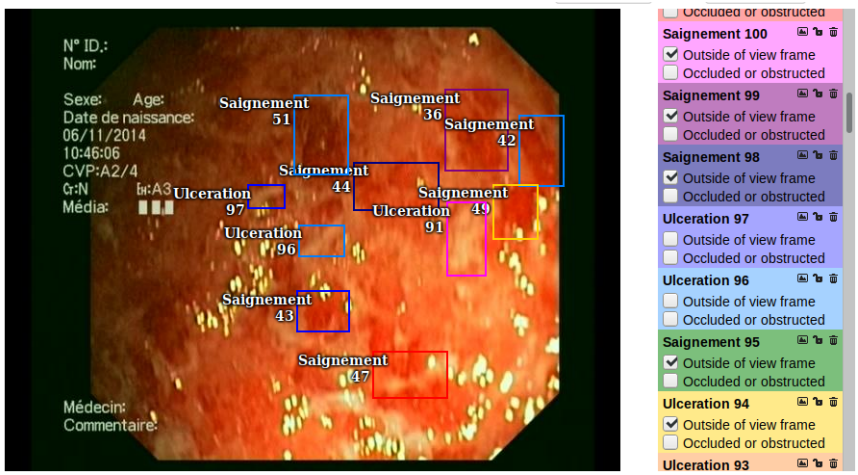


FIGURE 7.3 – Annotation d'une vidéo de coloscopie par le gastro-entérologue avec le logiciel Vatic.

Le travail 2 présenté dans la section 7.4 permet d'introduire la représentation que nous avons proposé aux gastro-entérologues, ainsi que le système de coordonnées utilisé par la suite. Il est à noter que le travail d'analyse d'images visant à détecter automatiquement les lésions dans les vidéos, en plus de son intérêt théorique, vise à soulager le travail d'annotation des gastro-entérologues, et a rendre les diagnostics plus fiables, ou du moins indépendants de la variabilité de l'expertise médicale.

7.3 Travail 1 : Détection automatique des saignements et des ulcères

Dans cette section, on considère des modèles de l'aspect des lésions (saignements et ulcères). Il s'agit donc de décrire l'ensemble des images correspondant à des saignements ainsi que l'ensemble des images correspondant à la paroi intestinale normale. Nous avons tenté de le faire de deux façons :

— en décrivant une loi de probabilité sur l'espace des images au travers de la variable aléatoire associée dans la section 7.3.1, ce qui a conduit à la proposition 2.2.1 du chapitre 2,
— en décrivant un ensemble d'images au travers de la partition engendrée par un classificateur binaire dans la section 7.3.2.

7.3.1 Couleur rouge

Durant le stage de M2 de Safaa Al Ali, nous avons commencé par considérer un problème jouet dans lequel l'état de la paroi intestinale est décrit uniquement par sa couleur, et plus spécifiquement par la composante rouge dans l'espace RGB. Dans ce travail, on est confronté à plusieurs types d'objets emboîtés :

— une frame de la vidéo de coloscopie, que l'on représente par la couleur rouge modale pour simplifier,
— un patient ou une vidéo qui est un ensemble de frames, et donc un ensemble de valeurs de la couleur rouge,
— la RCH, dont les saignements sont observés sur un ensemble de patients.

Suivant l'objectif de modélisation, on va prendre pour espace des données $\mathcal{E}$ l'un ou l'autre de ces objets. Comme l'on souhaitait modéliser les saignements de la RCH, nous avons considéré que l'on dispose d'un ensemble de patients résumés par leur couleur rouge (moyenne), et que la RCH est un processus qui conduit à observer certaines valeurs de rouge plutôt que d'autres. On ne dispose que d'un seul "point", car il n'y a qu'une seule population. Il n'y a pas de répétition au sens du chapitre 2.3. En effet, on ne compare pas les saignements produits par des maladies différentes, ou des groupes de patients différents.

Comme on ne dispose pas d'information supplémentaire sur les patients, ceux-ci sont considérés comme interchangeables et l'ensemble des données est invariant par permutation. Le saignement dans la RCH est donc décrit par la distribution de la couleur de chaque patient sur le segment $[0, 255]$. L'espace des données $\mathcal{E}$ est formé des distributions ou histogrammes sur $[0, 255]$, et l'on observe la loi empirique, comme dans la proposition 2.2.1. Les saignements apparaissent habituellement comme rouge sombre par rapport à une paroi intestinale rose (donc lumineuse), ils sont associés à des valeurs faibles dans l'espace RGB.

Code chunk 10 : «r»

```
patients = c(4, 20, 11, 9, 16, 13, 25, 20, 10, 56, 7, 13, 20, 11, 14, 18,
             12, 23, 19, 15, 22, 20, 19, 20, 17, 20, 8, 35, 15, 21, 15, 16,
             15, 14, 16, 35)
```

Interpret with `R`

```
```

Durant le stage, nous avons fait le choix de modéliser la RCH, i.e. sa distribution sur $[0, 255]$ par une variable aléatoire notée Y, et d'évaluer la similarité entre Y et un modèle M par la performance prédictive du modèle :

$$D(M, Y) = \sum_{i=1}^{37} d(m_i, y_i)$$

où y_i est la couleur rouge observée pour le patient i et $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ est une distance de de type L_1 ou L_2.

En appliquant la méthodologie décrite dans le chapitre 2, la modélisation de l'aspect du saignement consiste à déterminer un ensemble de modèles $\mathcal{M}$, ici un sous-ensemble de variables aléatoires à valeur dans $[0, 255]$, ainsi que les éléments de $M \in \mathcal{M}$ qui sont compatibles avec les observations. Safaa a donc tenté de trouver le modèle de distance minimale dans plusieurs classes de variables aléatoires. On s'attendait à ce que la loi empirique, fournie par les observations, soit optimale quel que soit le choix de la distance, et que la restriction de l'espace des modèles à une sous-classe (une hypothèse de modélisation) conduise à une version dégradée ou une projection de cette loi empirique.

Nous avons d'abord considéré l'ensemble des modèles restreint aux lois de Dirac δ_a pour $a \in \mathbb{R}$, c'est-à-dire aux variables aléatoires constantes. L'optimisation revient à calculer la couleur moyenne de l'ensemble des patients pour une distance L_2, et revient à calculer la couleur médiane pour une distance L_1. Dans la base de données Vatic, on trouve respectivement 17.88889 et 16.

Dans un deuxième temps, nous avons considéré l'ensemble des modèles constitué des lois M dont le support n'a que deux atomes i.e. $M = \alpha\delta_a + (1-\alpha)\delta_b$ pour $a, b \in \mathbb{R}$ et $\alpha \in [0,1]$. Contrairement à nos attentes, on obtient une masse de Dirac seule, correspondant à la couleur moyenne pour la distance L_2 et à la couleur médiane pour la distance L_1. Le même phénomène de concentration se reproduit dans la classe des lois gaussiennes de moyenne μ et d'écart-type σ où l'on obtient une gaussienne dégénérée $\sigma = 0$, concentrée sur la moyenne ou la médiane.

Ceci est un exemple concret du phénomène décrit par la proposition 2.2.1 dans le chapitre 2. Comme la similarité entre modèles et observations est donnée par l'espérance d'une distance sur $[0, 255]$, l'optimum est atteint pour une masse de Dirac en $m_0 = \arg\min_{m\in[0,255]} \int d(m, y)\, \mathrm{d}\mathbb{P}_Y$. Cet optimum étant facile à calculer et global dans l'ensemble $\mathcal{M}$ des variables aléatoires à valeur dans $[0, 255]$, il n'y a pas lieu d'étudier différentes hypothèses de modélisation.

Dans cet "échec de modélisation", interviennent deux ingrédients essentiels. En premier lieu, l'indépendance entre les modèles et les observations est inévitable en pratique. En effet, la maladie de chaque patient est une variable aléatoire indépendante des autres patients, de même loi commune inconnue, voisine de la loi empirique $\mathrm{d}\mathbb{P}_Y$. Si on utilise un modèle M pour prédire l'état d'un nouveau patient, on est donc dans la situation de comparer une nouvelle instance indépendante de Y à notre modèle M. Le même phénomène se produit si l'on dispose d'informations supplémentaires sur les patients et que l'on cherche à effectuer une régression sur ces co-variables : on obtient une masse de Dirac conditionnelle à la valeur des co-variables.

Ensuite, cet effet est spécifique du type de similarité entre l'espace des données $\mathcal{E}$ et l'espace des modèles $\mathcal{M}$. Il n'aurait pas eu lieu pour une distance entre distributions de probabilité du type Kullback-Leibler. En effet, supposons que la similarité entre $\mathcal{E}$ et $\mathcal{M}$ est donnée par une distance qui compare la probabilité d'un événement sous le modèle $M \in \mathcal{M}$ et cette même probabilité d'après les observations. Cette similarité compare donc le modèle M à la loi empirique $\mathrm{d}\mathbb{P}_Y$. Soit $Y' \in \mathcal{M}$ une variable aléatoire de loi $\mathrm{d}\mathbb{P}_Y$. La distance entre Y' et Y est nulle car les probabilités des événements sont identiques. On a donc un modèle Y' non dégénéré optimal pour la similarité entre $\mathcal{E}$ et $\mathcal{M}$, et le phénomène de concentration sur une masse de Dirac n'a pas lieu. On est forcé d'admettre que certaines similarités conduisent à des problèmes de modélisation dégénérés, sans bien savoir pourquoi une distance entre distributions est bonne alors qu'une distance entre variables aléatoires ne l'est pas.

7.3.2 Classificateurs linéaires dans l'espace des couleurs

Dans un deuxième temps, afin d'obtenir une description objective et reproductible de la paroi intestinale permettant de travailler sur la répartition des lésions, nous avons proposé une méthode de détection des saignements et des ulcères dans les images de coloscopie.

Un certain nombre de travaux récents ont proposé des méthodes de détection des saignements et des ulcères. Il s'agit principalement de travaux appliquant une méthode d'apprentissage standard (SVM[76], réseaux de neurones, K-Nearest Neighbors [46], etc.) à un ensemble de descripteurs bien choisis (moments statistiques des histogrammes[47], matrice de co-occurence[76], etc.) et ajustant les paramètres en vue d'obtenir une performance maximale. On a du mal à analyser les propriétés de l'aspect des saignements ou des ulcères d'après ces travaux. On retient cependant que de nombreux travaux détectent les saignements et les ulcères uniquement d'après la couleur des pixels. L'espace couleur RGB est adapté pour la détection des saignements, et en particulier les composantes R et G[48], tandis que l'espace YCbCr est adapté pour la détection des ulcères, et en particulier les composantes Y et Cb[64]. La forme de ces lésions, très complexe, est ignorée dans ces travaux. La texture est prise en compte dans les travaux récents, à partir de descripteurs génériques des textures (matrice de co-occurence [76], paramètres statistiques après application de filtres [64]).

La littérature suggère donc que l'on peut considérer le problème de la détection en dimension 2, c'est-à-dire considérer les classificateurs binaires dans les espaces (R, G) et (Y, Cb), donc les fonctions $\mathcal{M} : \mathbb{R}^2 \to \{0, 1\}$ qui attribuent une classe d'après les deux couleurs. De façon équivalente, le problème de classification correspond à trouver un ensemble de couleurs regroupant les couleurs observées pour les saignements, et son complémentaire regroupant les couleurs observées pour la paroi intestinale saine. Ceci peut donc être déduit des histogrammes (présentés sur la figure 7.4). Nous avons choisi de commencer par explorer les classificateurs linéaires, qui correspondent à des droites dans la figure 7.4.

Nous avons choisi d'évaluer la performance des classificateurs dans l'espace "ROC", c'est-à-dire d'après leur sensibilité et spécificité. Les résultats ont été décevants. En fait, les lésions visibles dans la vidéo de coloscopie sont de formes et tailles arbitraires ce qui rend leur délimitation manuelle difficile et imprécise. Dans notre base de données, les médecins ont utilisé des rectangles pour délimiter les saignements et les ulcères (cf Figure 7.3). Par conséquent, la base de données contient de nombreuses erreurs, correspondant aux pixels dont l'annotation est incorrecte, notamment dans les coins des rectangles.

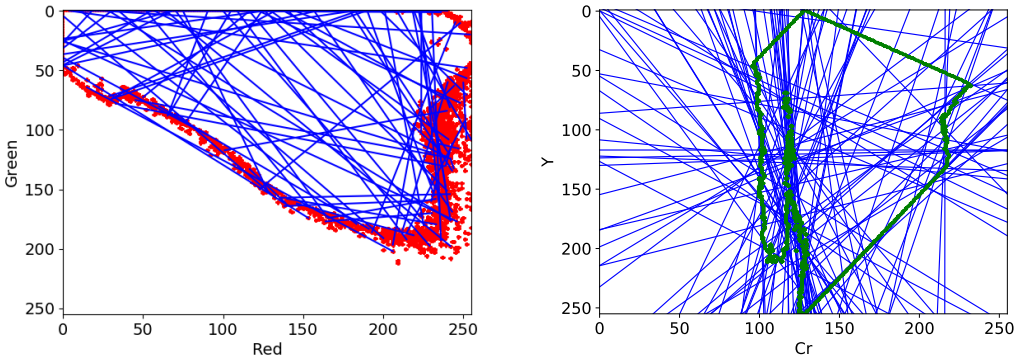


Figure 7.4 – Contours des histogrammes des pixels sains dans l'espace RG (gauche) et dans l'espace YCr (droite). On réalise un tirage aléatoire de 100 classificateurs linéaires (droites bleues) en prenant des couples de points appartenant au contour.

Pour remédier ce problème, nous avons proposé de modifier la définition de la sensibilité ordinaire de la façon suivante :

$$\text{Sensibilité} = \frac{\sum TP}{\sum TP + \sum FN} = \frac{\text{Pixels détectés corrects}}{\text{Pixels annotés}}$$

$$\text{Sensibilité}^N = \frac{\text{Annotations détectées correctes (nombre)}}{\text{Annotations (nombre)}}$$

$$\text{Sensibilité}^A = \frac{\text{Annotations détectées correctes (surface)}}{\text{Annotations (surface)}}$$

Ainsi, dès que l'on détecte un pixel dans une annotation, on compte l'annotation entière comme détectée. Dès lors, on compte le taux de lésions ou annotations détectées au travers de leur nombre (Sensitibilité^N) ou au travers de leur surface (Sensibilité^A), au lieu de compter le nombre de pixels détectés. Ceci permet de ne pas pénaliser les erreurs sur les bords, en ne les comptant pas comme des faux négatifs. Nous avons privilégié le calcul par surface, qui tient compte de l'importance visuelle des lésions.

La spécificité n'est pas modifiée. Le classificateur idéal est donc celui qui maximise le nombre de lésions détectées correctement d'après les annotations des médecins, tout en détectant peu de pixels en dehors des annotations. Chaque classificateur est représenté par un point dans l'espace ROC, et le détecteur idéal est celui qui se rapproche le plus du point $(0, 1)$. On prendra la distance à la diagonale comme critère de performance d'un classificateur.

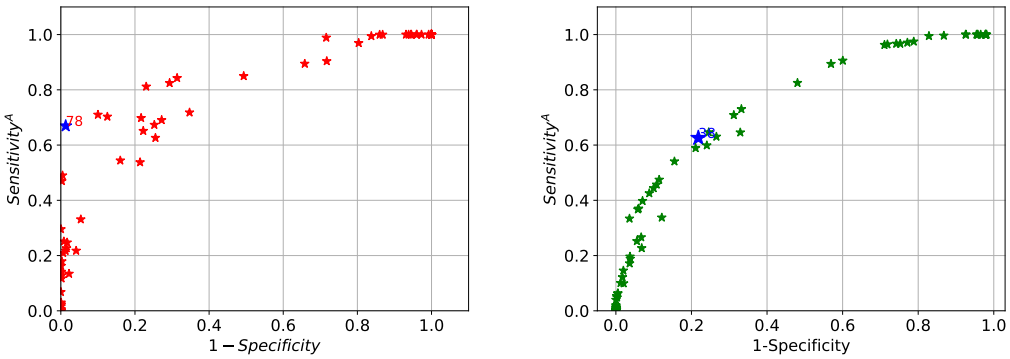


FIGURE 7.5 – Performance dans l'espace ROC modifié des classificateurs linéaires pour les saignements (gauche) et les ulcères (droite). Les meilleurs modèles trouvés sont indiqués par une étoile bleue.

Résoudre le problème de classification revient donc à explorer l'espace $\mathcal{M}$, ici l'ensemble des droites, en identifiant celle dont les performances sont optimales. Cette exploration peut se faire à partir d'un échantillon aléatoire obtenu d'après une loi de probabilité, d'une marche aléatoire dans la direction du gradient par exemple, ou bien d'un schéma explicite. Ce qui a retenu notre attention dans cette application est que l'espace $\mathcal{M}$ contient de nombreux modèles triviaux, correspondant à des classificateurs dont la réponse est toujours la même. Par exemple, la droite $R > G$ attribue la classe "saignement" à tous les exemples de la base de données. Sa sensibilité est de 1, car toutes les annotations sont retrouvées. Sa spécificité est de 0 car elle attribue la classe "saignement" à tous les pixels sains, il n'y a que des faux positifs et pas de vrais négatifs. Elle correspond au point $(1, 1)$ en haut à droite dans l'espace ROC.

Un classificateur optimal au sens de la distance à la diagonale réalise un compromis entre sensibilité et spécificité. Il faut donc éviter les droites triviales, pour lesquels l'un des histogrammes est entièrement d'un côté de la droite. Nous en déduisons qu'une droite non triviale doit "couper l'histogramme", c'est-à-dire avoir une intersection avec son contour. Nous avons donc proposé d'échantillonner l'espace des droites $\mathcal{M}$ en tirant deux points au hasard parmi le contour de l'histogramme (des pixels "saignement"). Les résultats d'un tirage de 100 modèles aléatoires sont présentés sur la figure 7.5.

Les figures 7.6 et 7.7 présentent quelques exemples de détections obtenues par les meilleurs modèles identifiés. Pour la détection des saignements, le modèle $G < 0.193R - 0.758$ obtient Sensibilité$^A = 66.9\%$ et Specificité $= 98.7\%$. Pour la détection des ulcères, le modèle $0.611Cr - 2.947 < Y$ obtient Specificité $= 78.18\%$ et Sensibilité$^A = 62.58\%$.

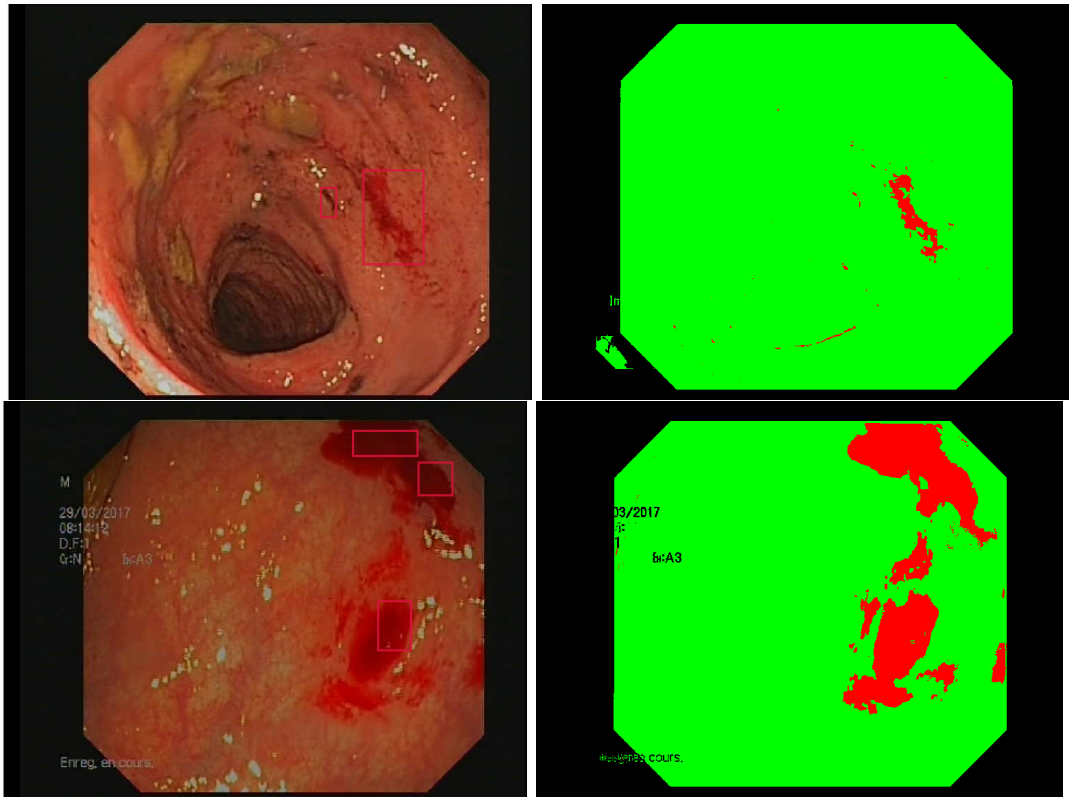


Figure 7.6 – Exemples de détections de saignements par le meilleur classificateur linéaire $G < 0.193R - 0.758$. (Annotations médicales à gauche, saignements détectés à droite en rouge).

Ce schéma d'échantillonnage, appuyé sur les contours de l'histogramme est intéressant car il est extensible à d'autres types de modèles. Par exemple, il est extensible aux coniques et aux courbes polynômiales qui seront définies par plusieurs points du contour. De la même façon, le schéma est extensible en dimension supérieure, ce qui fournit une extension naturelle de ces travaux s'il s'avère que le classificateur linéaire ne puisse pas être validé pour la pratique médicale.

Il est à noter que l'on a procédé par échantillonnage. La validation croisée utilise des jeux de données aléatoire et conduit à évaluer un générateur de modèles aléatoire. Elle évalue donc la performance du schéma d'échantillonnage à sélectionner un bon classificateur, et non la performance d'un modèle donné. Par conséquent, nous avons préféré évaluer la variabilité de l'estimation de performance pour un modèle donné, afin de savoir s'il est utilisable en pratique médicale, plutôt que la qualité de la méthode d'apprentissage.

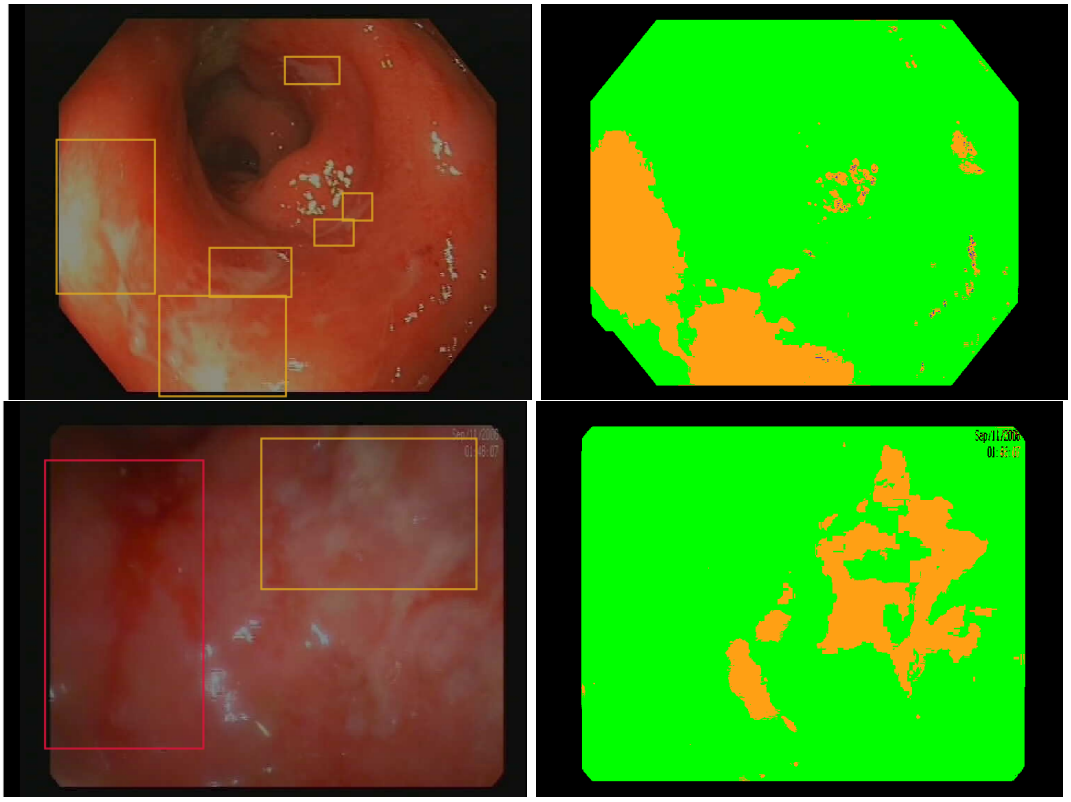


FIGURE 7.7 – Exemples de détections d'ulcères par le meilleur classificateur linéaire $0.611Cr - 2.947 < Y$. (Annotations médicales à gauche, ulcères détectés à droite en orange).

La figure 7.8 présente la performance et la variabilité des trois meilleurs modèles en évaluant la spécificité et Sensibilité^A, c'est-à-dire la position dans l'espace ROC modifié. Chaque modèle est évalué 20 fois sur chacun des 10 patients, sur un jeu de données aléatoire obtenu en sélectionnant 10% des frames de la coloscopie. Ces 20 points de mesure sont représentés par une ellipse, et trois points sont indiqués par des symboles.

On remarque que la performance est correctement estimée car les ellipses sont petites, et que l'essentiel de la variabilité correspond aux différences inter-patients. Ceci signifie que les modèles obtenus ne sont pas universels, et que les vidéos ont des caractéristiques très différentes. Il faut donc un modèle spécifique pour chaque nouveau patient, ou bien améliorer le pré-traitement et la standardisation des coloscopies. Ce travail a été soumis à la conférence CAIP 2021.

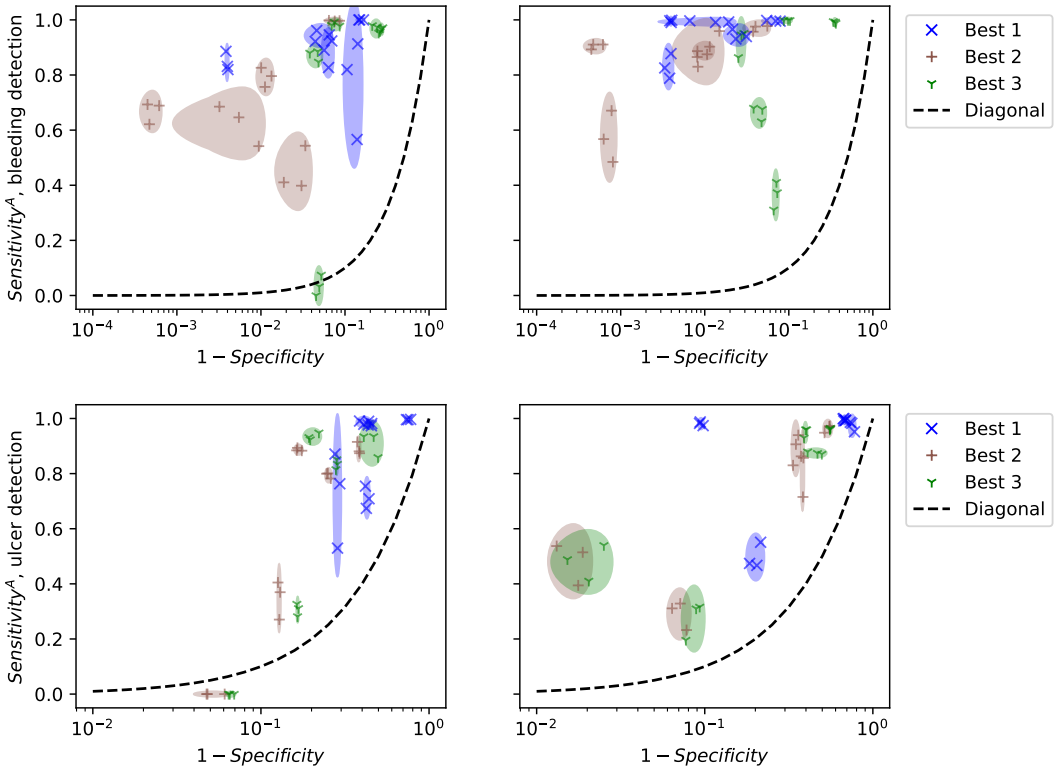


Figure 7.8 – Performance des 3 meilleurs modèles sur 10 patients différents : 5 vidéos d'apprentissage (gauche) and 5 autres vidéos (droite). La performance a été évaluée 20 fois pour chaque vidéo, sur 10% du nombre de frames tirées aléatoirement. Les ellipses représentent la variabilité de la performance d'après les 20 tirages, mais seulement 3 points sont indiqués par les symboles.

7.4 Travail 2 : Visualisation de la répartition des lésions

Comme indiqué précédemment, les méthodes actuelles d'évaluation de la sévérité de la RCH analysent la vidéo endoscopique en isolant chacune des images, comme dans le travail 1 (section 7.3). Or il est reconnu que l'extension et la distribution spatiale des lésions sont des indicateurs de la sévérité de la maladie et de sa progression. Dans cette section, nous présentons la méthode utilisée pour faire le lien entre le numéro de frame dans la vidéo, et le paramètre spatial. Elle sera utilisée dans la suite pour étudier des modèles de la sévérité (Travail 3, Section 7.5) et des modèles de répartition des lésions (Travail 4, Section 7.6).

Le colon est un organe mou, tubulaire. On observe la surface intérieure du tube, mais seule la position sur la longueur du tube a un intérêt. Nous avons donc choisi d'utiliser un seul paramètre $s \in [0, 1]$ qui représente l'abscisse curviligne le long du colon. Au cours d'une coloscopie, l'endoscope est introduit jusqu'au niveau de l'appendice (repère A sur la figure 7.9), puis retiré du colon. Nous faisons ensuite l'hypothèse que la vitesse de retrait est constante, donc que l'abscisse curviligne est reliée linéairement au numéro de frame.

Afin de faciliter la lecture des informations par le médecin, il a été proposé de visualiser les lésions sur un schéma du colon (cf figure 7.9). L'abscisse s est reportée sur un chemin passant approximativement par l'intérieur du colon sur le schéma. Les saignements sont présentés en rouge, tandis que les ulcères sont présentés en vert. À l'heure actuelle, la hauteur des segments représente le nombre de lésions présentes à une abscisse donnée. On pourrait également utiliser la surface des lésions, ou le rapport entre cette surface et la taille de l'image. Cependant, ce schéma est la première tentative de représentation spatiale des lésions et l'on ne sait pas quelle information est la plus pertinente.

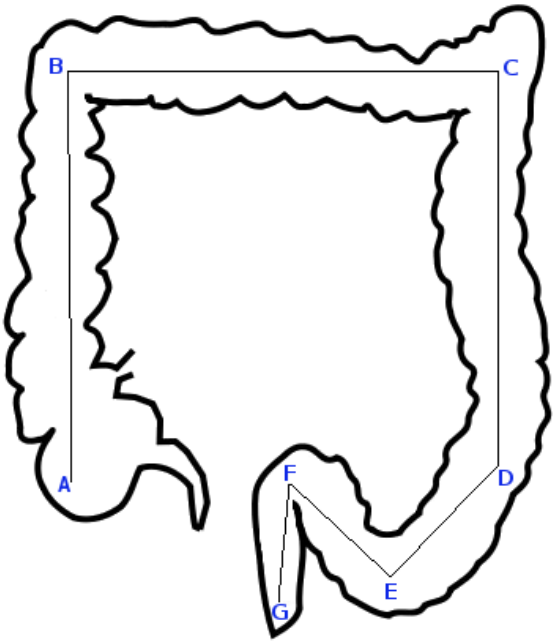


FIGURE 7.9 – Schéma du colon utilisé pour positionner les saignements et ulcères. L'appendice est au niveau du point A, le rectum correspond au segment FG. Les points B, C et E correspondent à des "coudes" repérables par le gastro-entérologue lors de la coloscopie.

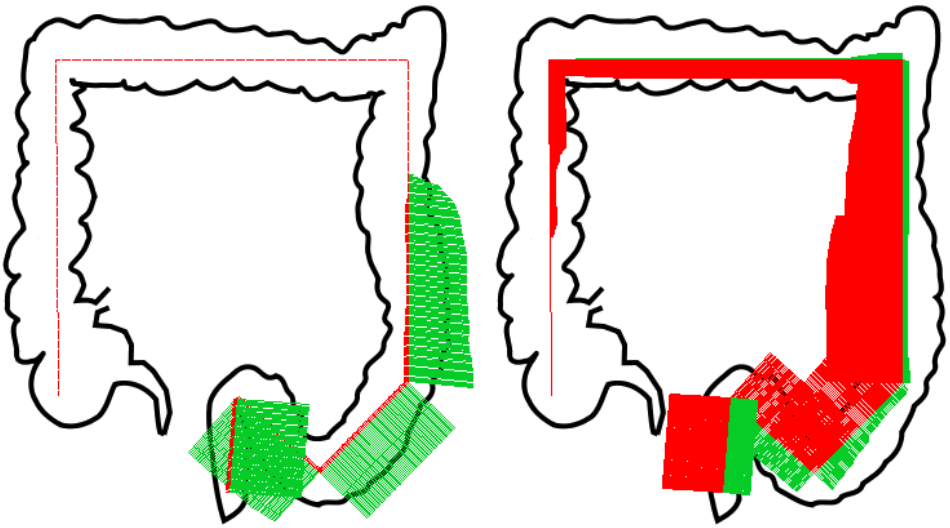


Figure 7.10 – Répartition des lésions de la RCH chez deux patients ayant un score UCEIS égal à 5 (annotations médicales). Le schéma montre les zones du colon touchées par les saignements (rouge) et les ulcères (vert).

La figure (7.10) permet de visualiser les lésions de la RCH chez deux patients de même score UCEIS (cf Section 7.5.1) égal à 5. On observe que les deux patients présentent des atteintes distinctes, le patient 1 présentant de nombreux ulcères, alors que le patient 2 présente une muqueuse très fragile et hémorragique. Cette nouvelle représentation présente clairement ces deux types d'atteinte de la muqueuse, ainsi que les différences de répartition des lésions. Elle suggère l'existence de plusieurs profils de patients, pour lesquels le traitement pourrait être adapté. Cette visualisation peut être appliquée pour les annotations fournies par le gastro-entérologue, ou bien pour les détections automatiques obtenues d'après le travail précédent (cf Figure 7.11).

Il est clair que l'hypothèse d'une relation linéaire entre l'axe temporel de la vidéo et l'abscisse curviligne est très simplificatrice, et qu'elle nécessite d'être validée par rapport aux observations cliniques. Tout d'abord, nous avons demandé à ce que le retrait de l'endoscope se fasse autant que possible à vitesse constante. De plus, nous avons proposé aux gastroentérologues de se repérer par rapport à la graduation (en cm) indiquée sur l'endoscope, mais ceci complique grandement la conduite de l'examen médical. Une possibilité en cours d'investigation est de repérer la position des points B, C et E indiqués sur la Figure 7.9. Ceux-ci correspondent à des coudes du colon facilement repérables par le médecin. Des expériences étaient prévues pour vérifier cette propriété de linéarité, mais elles n'ont pas encore pu être réalisées.

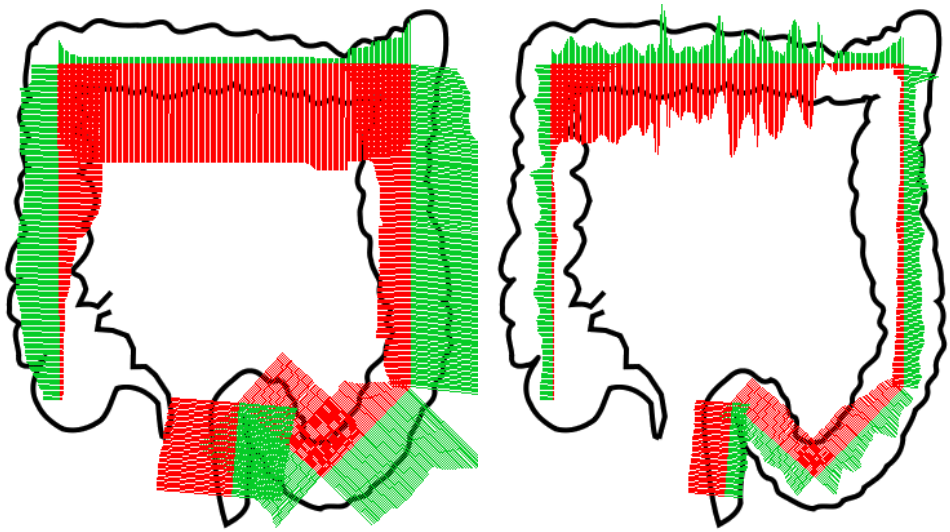


FIGURE 7.11 – Répartition des lésions de la RCH chez le même patient. Annotations médicales à gauche, détections automatiques (Travail 1) à droite.

En l'absence de données, nous nous sommes contentés de relations linéaires. Il est intéressant de remarquer que le cadre général s'applique aussi dans ce problème de visualisation, plus précisément dans le recalage entre la vidéo et l'abscisse curviligne. En effet, l'ensemble général des modèles de recalage est l'ensemble des fonctions croissantes $\mathbb{N} \to [0, 1]$. En faisant l'hypothèse de linéarité, on s'est restreint au sous-ensemble constitué par les droites croissantes. Or il existe des algorithmes de recalage tels que l'approche Dynamic Time Warping permettant de recaler selon des fonctions croissantes arbitraires, si l'on dispose de données suffisantes. Ceci pourrait être appliqué pour recaler des vidéos du même patient à deux instants différents, au cours du suivi du traitement par exemple.

7.5 Travail 3 : Modèles de la sévérité de la RCH

Dans cette section, on étudie l'évaluation numérique de la sévérité du patient. On suppose donc que le patient est décrit par un certain nombre de quantités observables aussi appelées marqueurs biologiques, tandis que sa sévérité est décrite par un nombre réel. Une observation de $\mathcal{E}$ est donc un couple formé des données patient et de sa sévérité. L'espace des modèles est formé des fonctions $\mathcal{M} : \{\text{patient}\} \to \mathbb{R}$.

7.5.1 Scores endoscopiques

Parmi les fonctions de $\mathcal{M} : \{\text{patient}\} \to \mathbb{R}$, les plus pertinentes et les plus utilisées dans la pratique clinique à l'heure actuelle sont le score MAYO [85] et le score UCEIS (Ulcerative Colitis Endoscopic Index Score, [90]). Le score MAYO évalue à la fois des variables endoscopiques et des observations cliniques comme la fréquence des selles, les saignements, l'activité inflammatoire de la sigmoïdoscopie, l'évaluation globale du médecin et les activités quotidiennes du patient [86]. Lorsque l'on ne s'intéresse qu'aux observations endoscopiques, il est habituel de se limiter au sous-score MAYO composé des variables endoscopiques.

Le score UCEIS est le seul indicateur de sévérité considéré par les gastro-entérologues comme "validé cliniquement". Il comporte trois éléments :
- la visibilité de la trame vasculaire (score entre 0 et 2),
- la présence de saignements (score entre 0 et 3),
- la présence d'ulcères (score entre 0 et 3)

On ne retient que la lésion la plus grave pour chacun des éléments. Les scores associés sont ensuite additionnés pour former un score UCEIS entre 0 et 8.

Le score UCEIS, comme le sous-score MAYO et les autres scores de sévérité proposés à l'heure actuelle, a un certain nombre de défauts :
- il est subjectif, et dépend l'expertise du gastro-entérologue,
- il manque de granularité, et ne permet pas de suivre finement l'évolution de la maladie et l'effet d'un traitement,
- il évalue uniquement la lésion la plus grave, sans tenir compte du nombre ou de l'extension des lésions,
- il ne tient pas compte de la position et de la répartition des lésions,
- il n'a pas été construit en fonction d'un indicateur clinique objectif tel que l'intervalle entre deux crises ou la durée avant chirurgie.

7.5.2 Modélisations de la sévérité

Une grande part des défauts du score UCEIS sont la conséquence de la réduction de données qui s'opère dans la simplification de la vidéo endoscopique et sa restriction aux trois lésions les plus graves. Cette opération réduit l'espace $\mathcal{M}$ à un sous-ensemble de fonctions de $\mathbb{R}^3 \to \mathbb{R}$. Si on les envisage de plus comme croissantes, avec un nombre limité de grades pour chaque lésion, le nombre de modèles possibles devient très limité.

Comme nous disposons de vidéos endoscopiques annotées et/ou d'annotatons automatiques de ces vidéos, nous avons envisagé un certain nombre de modèles de granularité plus fine et permettant de tenir compte de l'extension des lésions. Ce travail a été réalisé sur 14 vidéos comportant à la fois des sai-

gnements et des ulcères, et ayant été notés par trois gastro-entérologues :

Code chunk 11 : «severite»

```
UCEIS_XT = [8,7,2,1,3,1,5,5,5,5,7,5,4,7]
UCEIS_CS = [8,6,2,1,2,4,5,5,6,3,6,4,4,7]
UCEIS_YB = [7,8,3,1,1,0,4,5,3,2,3,3,2,4]

MAYO_XT  = [3,3,1,1,1,1,3,3,3,3,3,3,2,3]
MAYO_CS  = [3,3,1,1,1,2,3,3,3,1,3,3,2,3]
MAYO_YB  = [3,3,2,1,1,0,3,2,2,1,2,2,1,2]
```

Interpret with `python2`

```
```

Travail 1 Un premier travail consiste à envisager un modèle linéaire de la relation entre les lésions et le score endoscopique. On représente les données du patient dans $\mathbb{R}^2$ en considérant uniquement le nombre total de saignements et le nombre total d'ulcères. La sévérité est donc une fonction à trouver parmi les fonctions de $\mathbb{R}^2 \to \mathbb{R}$, et l'on restreint l'espace des modèles aux modèles linéaires, c'est-à-dire aux fonctions $f(S,U) = a\,S + b\,U + c, (a,b,c) \in \mathbb{R}^3$. Dans cet espace de fonctions, le meilleur modèle linéaire obtenu est la fonction $f(S,U) = 0.00039\,S + 0.00114\,U + 3.23$.

Code chunk 12 : «severite (part 2)»

```
# Nombres d'annotations
n_bleeding = [1946, 1101, 67, 2728, 1026, 674, 157, 673, 1081, 202, 8294,
              766, 0, 1574]
n_ulcer    = [2332, 3888, 20, 0, 0, 92, 956, 905, 904, 0, 50, 1292, 0, 0]
nb_annot = np.array([n_bleeding, n_ulcer]).transpose()

m = LinearRegression().fit(nb_annot,UCEIS_XT)
print 'f(S,U)_XT =', round(m.coef_[0],6), 'S + ', \
                    round(m.coef_[1],6), 'U + ', round(m.intercept_,3)
```

Interpret with `python2`

```
f(S,U)_XT = 0.000391 S +  0.00114 U +  3.226
```

Le même travail peut être effectué en découpant le colon en six segments, et en considérant le nombre de saignements et d'ulcères sur chacun de ces segments. On explore alors l'espace des fonctions $f : \mathbb{R}^{14} \to \mathbb{R}$, et on obtient les coefficients suivants.

Code chunk 13 : «severite_copie»

```
VbetaUCEISXT_p6= [[ 2.34114508e+00][ -1.60093247e-02][ 4.83337464e-03]
[ 2.44090421e-02][ -4.37072393e-02][ 4.99162039e-02]
[ -1.16436451e-03][ 8.78151257e-02][ -3.45440864e-02]
[ -7.25386297e-03][ -1.31653361e-02][ 4.43814090e-02]
[ -1.10326955e-01]]
```

Travail 2 Un deuxième travail consiste à modéliser la notion de sévérité pour chaque médecin au travers des scores UCEIS ou Mayo qu'il attribue aux vidéos endoscopiques. Dans la base de données Vatic, nous disposons en effet des scores de trois médecins. On obtient donc trois modèles linéaires :

Code chunk 14 : «severite (part 3)»

```
models_uceis = [LinearRegression().fit(nb_annot,g)
                for g in [UCEIS_XT, UCEIS_CS, UCEIS_YB]]
for (m,g) in zip(models_uceis,['XT','CS','YB']):
    print 'f(S,U)_',g, '=', round(m.coef_[0],6), 'S + ', \
                           round(m.coef_[1],6), 'U + ', round(m.intercept_,3)
```

Interpret with `python2`

```
f(S,U)_ XT = 0.000391 S +  0.00114 U +  3.226
f(S,U)_ CS = 0.000286 S +  0.00097 U +  3.363
f(S,U)_ YB = 0.000104 S +  0.001682 U +  1.88
```

Et de même pour les scores MAYO :

Code chunk 15 : «severite (part 4)»

```
models_mayo = [LinearRegression().fit(nb_annot,g)
               for g in [MAYO_XT, MAYO_CS, MAYO_YB]]
for (m,g) in zip(models_mayo,['XT','CS','YB']):
    print 'f(S,U)_',g, '=', round(m.coef_[0],6), 'S + ', \
                           round(m.coef_[1],6), 'U + ', round(m.intercept_,3)
```

Interpret with `python2`

```
f(S,U)_ XT = 9.1e-05 S +  0.000398 U +  1.928
f(S,U)_ CS = 0.000124 S +  0.000453 U +  1.769
f(S,U)_ YB = 6.8e-05 S +  0.000555 U +  1.273
```

On constate que les coefficients du nombre d'ulcères sont à peu près identiques pour tous les modèles, ce qui signifie que les gastro-entérologues accordent à peu près la même importance (subjectivement) aux ulcères. Par contre, les coefficients pour les saignements sont très différents. Ceci permet d'illustrer la variabilité inter-experts du processus de scoring.

Ces deux travaux sont limités par la simplicité des modèles considérés, mais aussi et surtout par la difficulté à se positionner dans le colon. Le découpage en six segments proposé est ainsi très artificiel, et l'utilisation d'un résumé tel que le nombre total de lésions, ou la lésion la plus grave, permet de s'affranchir de ce problème de recalage, au prix d'une simplification excessive des données.

Travail 3 Une troisième façon de s'intéresser à la sévérité consiste à chercher à en justifier des propriétés (mathématiques). Nous avons par exemple étudié la question : "est-ce que la sévérité est une fonction croissante du nombre total de saignements ?". Cet énoncé correspond au sous-ensemble des modèles croissants parmi l'ensemble des fonctions du type `n_bleeding` $\to \mathbb{R}$. Cet énoncé porte sur la sévérité, donc sur le comportement global d'une population de patients. On ne dispose donc pas de répétition, à moins de disposer de données sur plusieurs populations de patients considérées comme équivalentes : des données dans plusieurs pays, à des dates différentes, etc.

En l'absence de répétition, nous proposons de vérifier cet énoncé par accumulation de modèles compatibles, c'est-à-dire que l'on cherche à savoir si tous les modèles compatibles de la sévérité sont des fonctions croissantes. Pour cela, si restreint l'ensemble des modèles aux fonctions linéaires $\mathcal{M} = \{f(S) = aS+b\}$, les fonctions croissantes sont exactement les fonctions de pente a positive.

Une observation de la sévérité de la RCH est donnée sous la forme d'un ensemble de patients et de leurs scores numériques, ici les saignements des 14 vidéos de coloscopie et les scores UCEIS et MAYO. On définit les modèles compatibles comme étant les modèles proches à 5% de la performance du modèle optimal. L'ensemble des modèles compatibles est donc l'ensemble des $(a, b) \in \mathbb{R}^2$ tels que $\mathbb{E}[|a\,S + b - Y|^2] \leq 1.05\,\mathbb{E}[|\hat{a}\,S + \hat{b} - Y|^2]$ où le couple $(\hat{a}, \hat{b})$ correspond à la régression linéaire, et Y désigne le score UCEIS ou le score MAYO.

Code chunk 16 : «severite (part 5)»

```
def compatible_linmodels(x,y,alpha,candidates):
    # Return array containing alpha-compatible slopes among candidate slopes
    best_slope, best_intercept, _,_,_ = stats.linregress(x,y)
    best_dist = np.std(y - best_slope * x - best_intercept)
    return np.sort(np.append([ s for s in candidates
                    if np.std(y - s*x) < (1+alpha)*best_dist], best_slope))
```

Interpret with `python2`

```
```

Code chunk 17 : «severite (part 6)»

```
x = np.array(n_bleeding).transpose()
fig,ax = plt.subplots(figsize=(6,4))
for (i,uc,l) in zip(range(3),[UCEIS_XT,UCEIS_CS,UCEIS_YB],['XT','CS','YB']):
    s = compatible_linmodels(x,uc,0.05,np.arange(-0.02,0.02,3e-5))
    _ = ax.plot(s,s*0+5-i,'.-',label='UCEIS_'+l)

for (i,mayo,l) in zip(range(3),[MAYO_XT,MAYO_CS,MAYO_YB],['XT','CS','YB']):
    s = compatible_linmodels(x,mayo,0.05,np.arange(-0.02,0.02,3e-5))
    _ = ax.plot(s,s*0+2-i,'.-',label='MAYO_'+l)

_ = ax.plot([0,0],[0,5])
_ = ax.legend(loc=4); _ = ax.set_xlabel('Slope')
fig.savefig('severite.pdf')
```

Interpret with `python2`

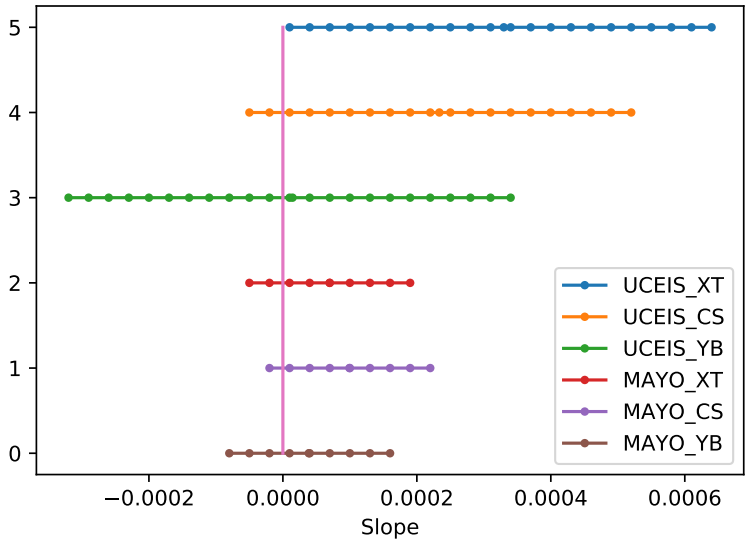


Figure 7.12 – Pentes 5% compatibles pour les scores UCEIS et MAYO attribués par les trois gastro-entérologues. Les modèles compatibles ont plutôt une pente positive, excepté dans le cas de UCEIS/YB.

La figure 7.12 montre les pentes compatibles avec les scores UCEIS et MAYO attribués par les trois gastro-entérologues. Excepté pour les scores UCEIS/YB, les modèles compatibles ont une pente positive, ce qui signifie que le score attribué est croissant en fonction du nombre de saignements pour tous les modèles

compatibles. Nous avons donc justifié l'énoncé par accumulation de modèles, et pour tous les médecins. Dans le cas des scores UCEIS/YB, les modèles compatibles ne présentent pas de biais dans un sens ou dans l'autre, et l'on ne peut pas conclure.

Conclusion Dans cette partie, nous avons présenté quelques travaux préliminaires de modélisation de la sévérité de la RCH à l'aide de modèles linéaires. Un premier type de travail consiste à estimer, c'est-à-dire à représenter la sévérité par un modèle linéaire. En plus de la capacité de représentation des modèles linéaires, ces travaux sont limités par la qualité de l'information contenue dans les données. Ainsi, les modèles obtenus à partir des scores UCEIS reflètent nécessairement les scores UCEIS. Ils sont différents selon les experts, et probablement différents de la vérité de la maladie. De la même façon, la difficulté de repérage des lésions dans le colon rend difficile la modélisation spatiale de la sévérité.

Un deuxième type de travail consiste à justifier des énoncés scientifiques. La notion d'ensemble de modèles compatibles permet ici de comparer un énoncé tel que "sévérité croissante" et le niveau de preuve apporté par les données. On est cependant confronté à la difficulté de définir un énoncé scientifique utile pour le médecin, et à l'arbitraire de la définition de la notion de compatibilité.

7.6 Travail 4 : Modèles de la répartition des lésions

Les gastro-entérologues font l'hypothèse que la répartition des lésions contient une information médicale, mais on ne sait pas bien quelle information. En effet, en l'absence d'outils automatiques de répérage des lésions, il n'existe pas d'étude indiquant si cela est relié à la sévérité de la maladie, ni d'étude permettant de caractériser des sous-populations de malades par exemple. Ainsi, les investigations décrites dans cette partie, on été menées sans a priori médical.

7.6.1 Modèles de répartition et énoncés scientifiques

Les annotations des médecins et les méthodes de détection automatique des lésions fournissent la position et le nombre de saignements et d'ulcères en fonction de la position dans le colon, i.e. d'une abscisse curviligne $s \in [0, 1]$. Les données vivent donc dans l'espace des fonctions $\mathcal{E} = [0, 1] \to \mathbb{R}^k$. Nous avons commencé par le cas $k = 1$, ce qui correspond à traiter saignements et ulcères séparément. Safaa Al Ali a également souhaité envisager le nombre de lésions en fonction de la position, une lésion étant définie comme un saignement ou un ulcère.

L'espace des fonctions $[0,1] \to \mathbb{R}$ étant un espace de dimension infinie, nous sommes amenés à considérer des sous-espaces de modèles :

— fonctions affines, i.e. les fonctions de la forme $\{y = a\,s + b, a, b \in \mathbb{R}\}$,
— espace des solutions d'une équation différentielle du premier ordre,
— espace des modèles de réaction-diffusion.

Vu comme un espace de fonctions de $[0,1] \to \mathbb{R}$, l'espace des fonctions affines est un espace $\mathcal{M} = \{y = a\,s+b, a, b \in \mathbb{R}\}$ de dimension 2, et peut être représenté dans le plan $\mathbb{R}^2 = \{(a,b)\}$. En particulier, on peut représenter les fonctions affines compatibles avec les observations comme des sous-ensembles de $\mathbb{R}^2$, et les énoncés scientifiques ou les propriétés mathématiques correspondent à des régions du plan. Par conséquent, démontrer un énoncé scientifique correspond à montrer que toutes les répétitions des données appartiennent à la même région du plan, ou bien que tous les modèles compatibles y appartiennent.

L'espace des solutions d'une équation différentielle du premier ordre, i.e. $y' = ay$ est constitué des fonctions exponentielles $y = Ce^{as}$. C'est aussi un espace à deux paramètres que l'on pourra représenter dans le plan $\mathbb{R}^2$. Dans cet espace de modèles, on peut interpréter le paramètre C comme un nombre global de lésions, et le paramètre a comme un indicateur de répartition. Si $a > 0$ alors la fonction est croissante, et les lésions sont plus abondantes au niveau du rectum que au début du colon. Une valeur de a proche de 0 indique que les lésions sont réparties de façon uniforme.

Parmi les modèles de réaction-diffusion, nous avons choisi l'équation de Fisher-KPP [62], avec l'intuition que la propagation des lésions dans le colon peut être modélisée par la propagation d'un front d'onde. Une application médicale du modèle consiste alors à estimer la vitesse de propagation et d'en déduire un pronostic pour le patient. Ceci revient à caractériser l'ensembles des paramètres et conditions initiales conduisant à l'apparition d'un front d'onde, c'est-à-dire à tracer un "portrait de phase" pour l'équation de Fisher-KPP. Il s'agit d'étudier les partitions de l'ensemble des modèles $\mathcal{M}$, ici les solutions de l'équation de Fisher-KPP, et de situer le patient observé dans ces partitions.

7.6.2 Biais de répartition (cas binaire)

Le travail présenté dans cette partie a pour point de départ l'énoncé suivant formulé par les gastro-entérologues : "les lésions de la RCH partent toujours du rectum", et pour lequel nous avons tenté d'apporter des "preuves mathématiques". Conformément à la démarche de modélisation présentée dans le chapitre 2, cela consiste à définir une partition de l'espace des données et à justifier l'énoncé par accumulation d'observations ou accumulation de modèles.

Un premier problème est que l'énoncé médical est trop vague, et qu'il conduit à différentes partitions selon l'interprétation que l'on en fait. Nous avons considéré les deux énoncés suivants définis pour les patients (donc sur l'espace de données $\mathcal{E} = [0,1] \to \mathbb{R}$) :

— **E1** les lésions sont localisées préférentiellement dans la partie basse du colon,
— **E2** les lésions sont plus abondantes lorsque l'on se rapproche du rectum.

L'énoncé E1 est défini précisément de la façon suivante :

Definition 7.6.1. Soit $p \in \mathcal{E}$ la fonction décrivant l'état du colon d'un patient. On dira que p est localisée dans la partie basse du colon si

$$\int_0^{0.5} p(s)\mathrm{d}s < \int_{0.5}^{1} p(s)\mathrm{d}s$$

Cet énoncé décrit bien une partition de l'espace des fonctions $[0,1] \to \mathbb{R}$. Comme il est calculable directement d'après l'observation $p \in \mathcal{E}$, il ne nécessite pas d'accumulation de données ou de modèles pour être justifié. Le code Python suivant permet d'effectuer le calcul pour les 37 patients de notre base de données.

Code chunk 18 : «biais_repartition»

```
biais_b, biais_u = np.full(37,True), np.full(37,True)
for i in range(0,36):
  p = Data_perc_b[i]
  L,L2 = len(p), (int) (len(p)/2)
  biais_b[i] = sum(p[0:L2])/L < sum(p[(L2+1):L])/L
  p = Data_perc_u[i]
  biais_u[i] = sum(p[0:L2])/L < sum(p[(L2+1):L])/L

print sum(biais_b), "patients sur", len(biais_b),"vérifient E1 " \
  "(gauche < droite) pour les saignements."
print sum(biais_u), "patients sur", len(biais_u),"vérifient E1 " \
  "(gauche < droite) pour les ulcères."
```

Interpret with `python2`

```
28 patients sur 37 vérifient E1 (gauche < droite) pour les saignements.
21 patients sur 37 vérifient E1 (gauche < droite) pour les ulcères.
```

Comme l'énoncé E1 porte sur le patient $p : [0,1] \to \mathbb{R}$, on a en fait 37 énoncés $\mathrm{E1}_p$, que l'on a envie de rassembler en un énoncé global sur la RCH. Ainsi, si les 37 patients vérifient chacun l'énoncé E1, on aura démontré par accumulation d'observations l'énoncé $\mathrm{E1}_{\mathrm{RCH}}$: "la RCH produit des lésions localisées préférentiellement dans la partie basse du colon".

En pratique, il est peu probable que les patients vérifient tous l'énoncé E1. On est donc amené à évaluer le niveau de preuve apporté par le nombre de patients calculé ci-dessus. Pour cela, on cherche à construire un test statistique d'appartenance à la région E1 contre l'absence de biais.

Sous l'hypothèse d'absence de biais, les lésions sont réparties uniformément dans le colon, entre 0 et 1, car toutes les positions sont équiprobables. Par conséquent, l'appartenance à E1 pour un patient est un événement de probabilité 0.5 (par symétrie), et le nombre de patients vérifiant E1 suit une loi binômiale. Le test statistique présenté ci-dessous nous amène à rejeter l'hypothèse $\lambda = 0.5$ pour les saignements.

Code chunk 19 : «biais_repartition (part 2)»

```
print "Proba[B >= ",sum(biais_b), "sur", len(biais_b),"] =", \
   1-binom.cdf(sum(biais_b),len(biais_b),0.5)
```

Interpret with `python2`

```
Proba[B >=  28 sur 37 ] = 0.0003764485300052911
```

En fait, on peut faire une analyse presque complète dans ce cas. La RCH désigne un ensemble de patients présentant des symptômes similaires, caractéristiques de la maladie. Au travers de l'énoncé E1, on réalise une projection de cet ensemble de patients sur un ensemble de nombres booléens. L'espace des observations pour un énoncé sur la RCH est donc $\mathcal{E} = \{0,1\}^{37}$. On peut faire les hypothèses suivantes par "construction" :

- indépendance : la RCH n'est pas une maladie transmissible, jusqu'à preuve du contraire les lésions sont spécifiques pour chaque patient,
- même distribution : on a considéré par hypothèse qu'il s'agit de patients indistinguables.

Chaque patient peut donc être considéré comme la réalisation d'une variable aléatoire de Bernoulli indépendante identiquement distribuée. L'ensemble des modèles de la RCH est l'ensemble des variables de Bernoulli de paramètre $\lambda \in [0,1]$, ou de façon équivalente le segment $\mathcal{M} = [0,1]$. La vraisemblance $\mathbb{P}(\text{données}|\lambda)$ indique naturellement si les valeurs observées sont plausibles pour un modèle, ou réciproquement si un modèle est compatible avec les données.

Un énoncé scientifique portant sur la RCH correspond à une partition de l'ensemble des modèles $\mathcal{M} = [0,1]$. L'énoncé E1_{RCH} correspond aux variables de Bernoulli de paramètre supérieur à 0.5, i.e. aux modèles $]0.5;1]$. Cet énoncé est donc démontré par accumulation de modèles si l'ensemble des modèles compatibles est contenu dans $]0.5;1]$.

Pour la Figure 7.13, on définit arbitrairement la notion de modèle compatible par $\ln(\mathbb{P}(\text{données}|\lambda)) > -5$), et l'on trace l'ensemble des modèles λ compatibles avec les patients de notre base de données.

Code chunk 20 : «biais_repartition (part 3)»

```
def plot_compat(figname,n,k,p):
    logp = binom.logpmf(k,n,p)
    p_compat = p[logp>-5]
    fig,ax = plt.subplots(figsize=(6,3.5))
    _ = ax.margins(y=0.12)
    ax.plot(p,logp)
    ax.plot(p_compat,p_compat*0-5)
    ax.plot([0.5, 0.5],[-15, 0])
    ax.text(float(k)/n,np.max(logp),'Max vraisemblance',
            va='bottom',ha='center')
    ax.text(np.mean(p_compat),-5,'Modeles compatibles',
            va='top',ha='center')
    ax.set_xlabel('Parametre lambda')
    ax.set_ylabel('Vraisemblance (log)')
    fig.savefig(figname)

plot_compat('biais_compat_b.pdf',
    len(biais_b),sum(biais_b),np.arange(0.4,0.98,0.01))
plot_compat('biais_compat_u.pdf',
    len(biais_u),sum(biais_u),np.arange(0.2,0.9,0.01))
```

Interpret with `python2`

```
```

La figure indique que les modèles $\lambda < 0.55$ sont incompatibles avec le biais de répartition des saignements. Comme l'ensemble des modèles compatibles pour les saignements est inclus dans $]0.5, 1]$, on en conclut que l'énoncé E1 est bien démontré par accumulutation de modèles. Par contre, les modèles $\lambda > 0.75$ sont incompatibles avec le biais de répartition des ulcères. Comme l'ensemble des modèles compatibles avec le biais des ulcères n'est ni inclus dans E1 ni dans son complémentaire, il n'est pas possible de conclure pour les ulcères.

Remarque Cette analyse, qui passe par un ensemble de modèles de Bernoulli $\mathcal{M} = \{\lambda \in [0, 1]\}$ est plus complète que la précédente, qui ne passe que par les données des patients. L'utilisation des modèles permet ici de localiser ou de quantifier le biais de répartition, au lieu de simplement mesurer l'adhésion à l'énoncé E1.

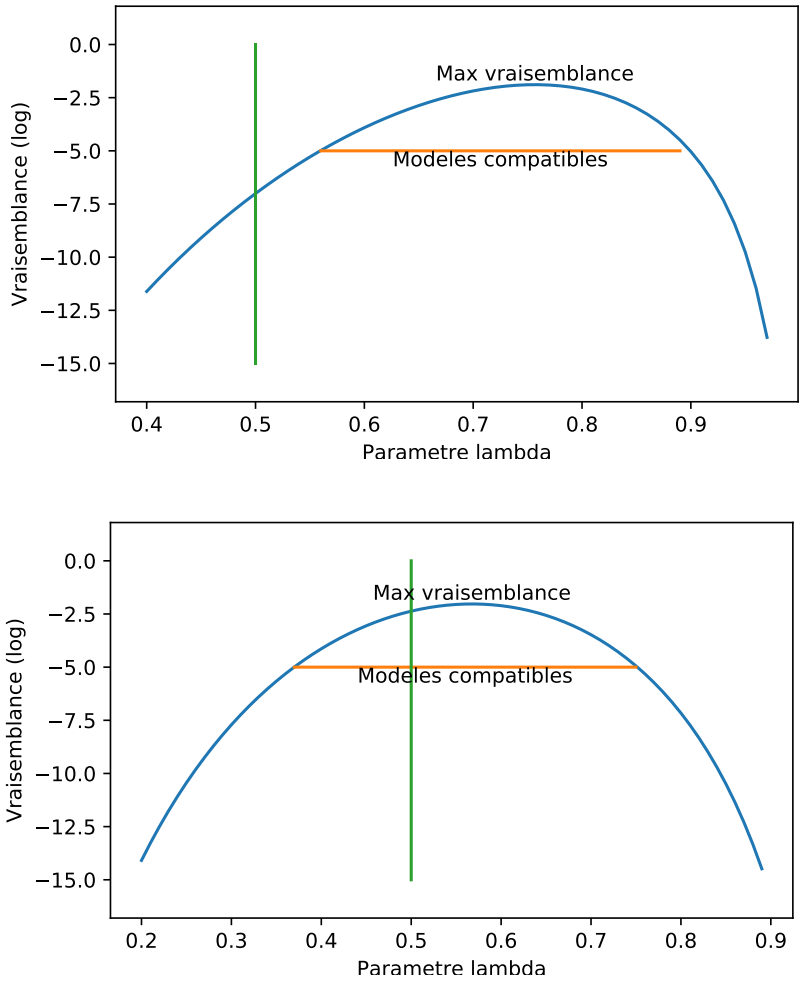


Figure 7.13 – Modèles de Bernoulli compatibles avec le biais de répartition observé. Tous les modèles compatibles appartiennent à $]0.5, 1]$ dans le cas des saignements (haut), mais la situation est indéterminée pour les ulcères (bas).

7.6.3 Biais de répartition (nombres réels)

Pour l'énoncé E2, on considère que les lésions sont plus abondantes lorsque l'on se rapproche du rectum si la fonction p est croissante pour le patient correspondant. Cependant, comme on compte un nombre de lésions dans chaque frame, les données sont très bruitées, et l'on n'observe jamais de fonction croissante ou décroissante directement dans les observations. Pour l'énoncé E2, on va donc passer par les modèles : on va justifier que la plupart des modèles compatibles vérifient l'énoncé E2. De plus, on se place dans des classes de modèles de fonctions monotones.

Pour cela on définit arbitrairement la notion de compatibilité suivante :

Definition 7.6.2. (Modèles compatibles) Étant donné une notion de similarité d entre un modèle m et des données y (éventuellement sous la forme d'une distance), on appelle $d^* = \min_{m \in \mathcal{M}} d(m, y)$ la performance du modèle optimal dans la classe $\mathcal{M}$. Les modèles α-compatibles sont les modèles m tels que $d(m, y) < (1 + \alpha)d^*$.

Si on considère l'espace $\mathcal{M}$ des modèles linéaires, l'énoncé E2 correspond aux modèles de pente strictement positive. Pour chaque patient, on cherche donc la régression linéaire, et l'ensemble des modèles α-compatibles. Ici, on s'intéresse uniquement aux pentes des modèles compatibles. On utilise la fonction `compatible_linmodels` définie précédemment à la section 7.5 chunk 16.

Code chunk 21 : «`biais_repartition (part 4)`»

```
def plot_linear(figname,datap,alpha):
    plt.figure(figsize=(6,4))
    _ = plt.plot([0,0],[0,36])
    _ = plt.xlabel('Slope')
    _ = plt.ylabel('Patient number')
    for i in range(0,37):
        p = datap[i]
        x = np.arange(0,len(p),1.0)
        ac = compatible_linmodels(x,p,alpha,np.arange(-0.0005,0.001,2e-5))
        _ = plt.plot(ac,ac*0+i,'.-')

    plt.savefig(figname)
    plt.close()

plot_linear('biais_linearb.pdf',Data_perc_b,0.1) # Saignements
plot_linear('biais_linearu.pdf',Data_perc_u,0.1) # Ulcères
```

Interpret with `python2`

```
```

La figure 7.14 présente les pentes des modèles linéaires 0.1-compatibles pour chaque patient. Dans le cas des saignements (en haut), on observe qu'une majorité des patients présente un biais de répartition car les modèles compatibles ont une pente positive. Cependant, ce biais est peu marqué, car la valeur des pentes est faible et que plusieurs patients ont un comportement inverse ou indéterminé. La situation est plus claire dans le cas des ulcères (en bas), car tous les patients ont des modèles compatibles de pente positive, avec un écart plus marqué avec la pente nulle.

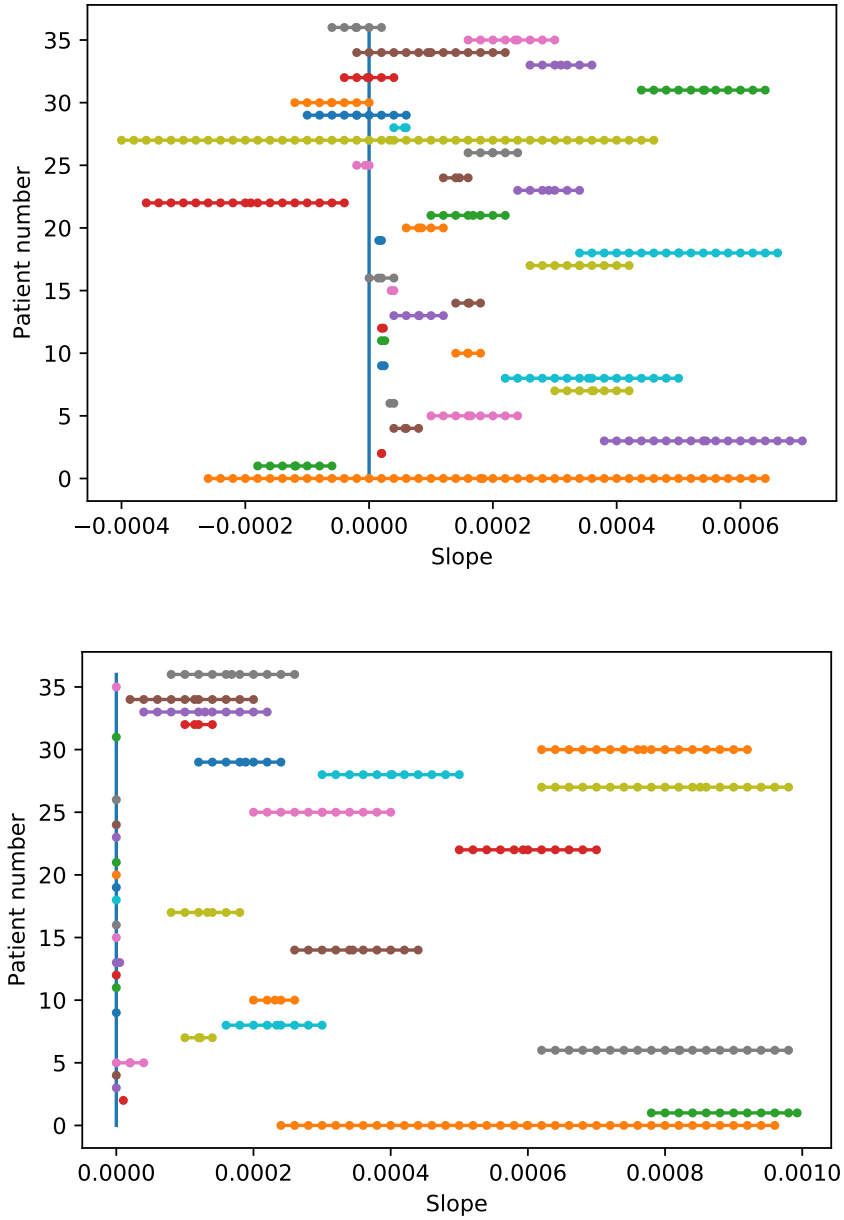


Figure 7.14 – Pentes des modèles linéaires compatibles à 0.1 près. Saignements (haut) et Ulcères (bas). On observe un léger biais pour la répartition des saignements, mais le biais est plus significatif dans le cas des ulcères.

Dans l'espace $\mathcal{M}$ des modèles exponentiels $y = Ce^{as}$, l'énoncé E2 correspond aux modèles de paramètre a strictement positif. Pour chaque patient, on cherche donc la régression exponentielle, et l'ensemble des modèles α-compatibles. Pour simplifier, on appliquera simplement une transformation logarithmique.

Code chunk 22 : «biais_repartition (part 5)»

```
def plot_exp(figname,datap,alpha):
    plt.figure(figsize=(6,4))
    _ = plt.plot([0,0],[0,36])
    _ = plt.xlabel('a')
    _ = plt.ylabel('Patient number')
    for i in range(0,37):
        logp = np.log(datap[i]+0.0001)
        x = np.arange(0,len(logp),1.0)
        ac = compatible_linmodels(x,logp,alpha,np.arange(-0.008,0.025,4e-4))
        _ = plt.plot(ac,ac*0+i,'.-')

    plt.savefig(figname)
    plt.close()

plot_exp('biais_expb.pdf',Data_perc_b,0.1) # Saignements
plot_exp('biais_expu.pdf',Data_perc_u,0.1) # Ulcères
```

Interpret with `python2`

```
```

La figure 7.15 présente les pentes des modèles exponentiels 0.1-compatibles pour chaque patient. Dans le cas des saignements (en haut), on observe qu'une majorité des patients présente un biais de répartition car les modèles compatibles ont un coefficient positif. Cependant, ce biais est peu marqué, de valeur plus faible que dans le cas des ulcères. Pour un certain nombre de patients, le coefficient est nul dans le cas des ulcères, ce qui signifie qu'il n'apparaît pas de biais de répartition.

Il apparaît donc que pour la plupart des patients et pour les modèles de régression linéaire et exponentielle, l'énoncé E2 est vérifié car les pentes des modèles compatibles sont positives.

Remarque L'enoncé E2 représente chaque patient par un nombre réel, ou un ensemble de pentes compatibles, alors que l'énoncé E1 représente un patient par un nombre binaire. L'analyse des propriétés de la maladie est plus difficile, car il faut considérer des variables aléatoires à valeurs réelles. C'est un ensemble plus grand que l'ensemble des variables aléatoires à valeur binaires, qui est restreint aux lois de Bernoulli.

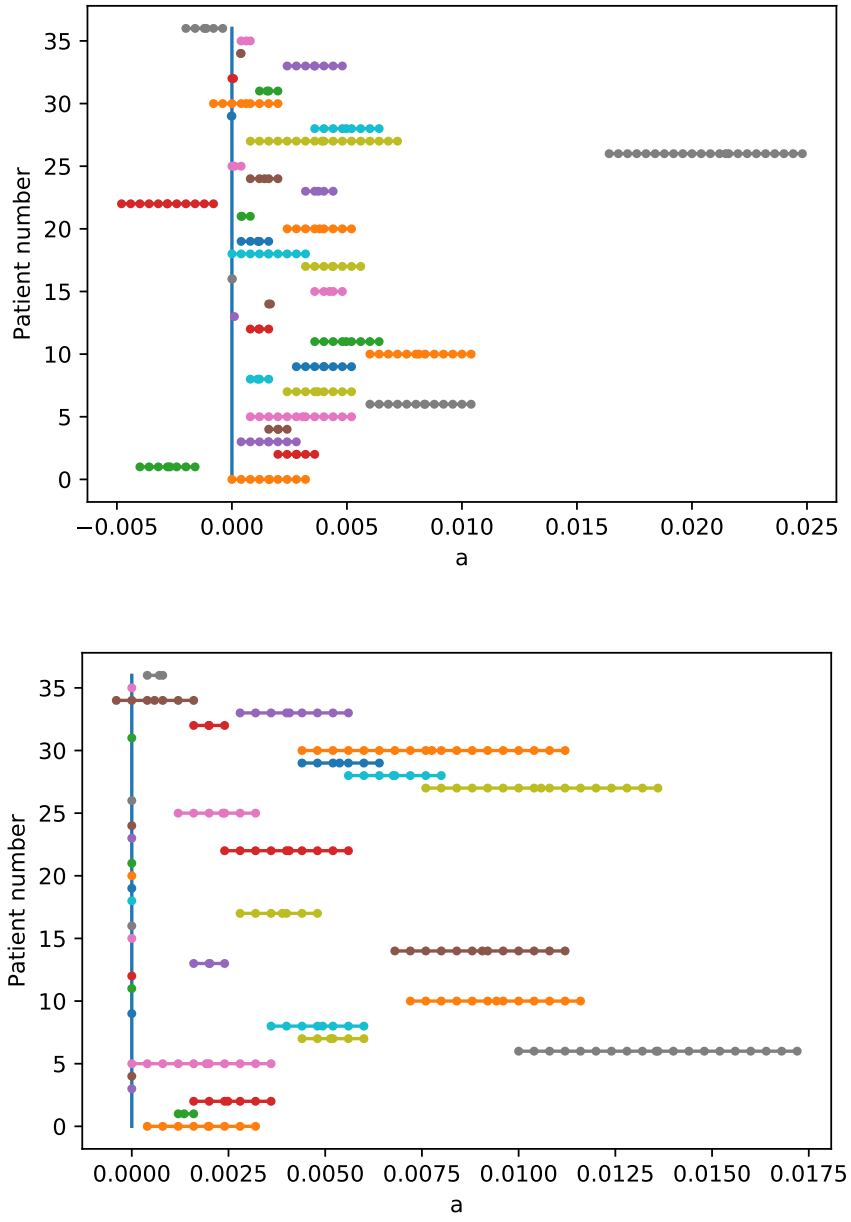


Figure 7.15 – Pentes des modèles exponentiels compatibles à 0.1 près. Saignements (haut) et Ulcères (bas). On observe un léger biais pour la répartition des saignements. Dans le cas des ulcères, certains patients ont un biais plus marqué, alors que d'autres apparaissent sans biais.

7.6.4 Fronts d'onde dans l'équation de Fisher-KPP

Au lieu de travailler sur des caractéristiques globales de la maladie, nous avons également souhaité travailler sur des modèles spatiaux de la répartition des lésions. Nous avons envisagé les modèles de réaction-diffusion, qui sont couramment utilisés pour modéliser la propagation d'un phénomène qui pourrait être ici l'inflammation ou les lésions.

La forme générale des équations de réaction-diffusion est décrite par l'équation suivante :

$$\frac{\partial}{\partial t}u - D\frac{\partial^2}{\partial x^2}u = f(u, x, t), \quad t \geq 0,\ x \in \mathbb{R}^n \tag{7.1}$$

où $u(x,t)$ est une fonction de $\mathbb{R}\times\mathbb{R}^k \to \mathbb{R}$. Les termes D et f sont appelés respectivement matrice de diffusion et terme de réaction. Quand le terme de réaction $f(u)$ est nul, il s'agit de l'équation de la chaleur. Quand le terme $f(u) = u(1-u)$, on parle de l'équation de Fisher Kolmogorov-Petrovski-Puskinov (Fisher-KPP) initialement proposée pour décrire la propagation de populations[62].

Une équation de réaction-diffusion, i.e. la donnée de la matrice de diffusion D et du terme de réaction f, caractérise un ensemble de fonctions u qui sont les solutions de cette équation pour différents choix de la condition initiale $u(x,0)$. On peut ainsi caractériser un ensemble $\mathcal{M}$ de modèles par une équation de réaction-diffusion.

Dans ce travail, nous proposons d'utiliser les solutions de l'équation de Fisher Kolmogorov-Petrovski-Puskinov comme ensemble de modèles $\mathcal{M}$, et de prédire l'évolution de la RCH chez un patient en fonction des modèles compatibles avec l'examen de coloscopie observé. En particulier, nous souhaitons évaluer la possibilité d'estimer une vitesse de propagation de la RCH.

On considère donc a priori l'ensemble des modèles $\mathcal{M}$ constitué des solutions de l'équation :

$$\frac{\partial}{\partial t}u - D\frac{\partial^2}{\partial s^2}u = u(1-u), \quad t \geq 0,\ s \in \mathbb{R} \tag{7.2}$$

où s désigne l'abscisse curviligne dans le colon et u désigne une mesure de sévérité locale de la maladie, telle que le nombre de lésion ou la surface détéctée dans le travail 1. L'espace $\mathcal{M}$ est paramétré par les couples (D, u_0) où D est le paramètre de diffusion et u_0 la condition initiale.

Remarque On a ici formulé l'équation de Fisher-KPP sur le domaine non borné $\mathbb{R}$ alors que le côlon est paramétré par l'abscisse curviligne sur $[0, 1]$. Ceci est une approximation raisonnable car la propagation des lésions est un phénomène local par rapport à l'ensemble du colon (cf Figure 7.16). Comme les lésions se forment au niveau du rectum avant de se propager dans le colon d'après les médecins (cf 7.6.2), on prendra comme conditions au bord $u(1, t) = 1$ et $u(0, t) = 0$ pour tout t.

Nous utilisons le code suivant afin de simuler l'évolution de l'équation de Fisher-KPP pour un patient donné. Les résultats pour le patient 23 sont présentés sur la figure 7.16. Le travail pour les autres patients est en cours de réalisation.

Code chunk 23 : «fkpp»

```
L=len(Data_perc_b[23])
x0, xL, delta_x = 0, L-1, 1
t0, tF, delta_t = 0, 50, 0.003
Nx, Nt = int((xL-x0)/delta_x+1), int((tF-t0)/delta_t+1)
X, T   = np.linspace(x0,xL,Nx),  np.linspace(t0,tF,Nt)

# initial condition
U1=np.zeros((Nx,Nt))
U1[0,:] = 0.0           # at position x=0 for all times
U1[L-1,:] = 1.0         # at position x=L-1 for all times
U1[:,0]= Data_perc_b[23] / max(Data_perc_b[23]) # at t=0

# simulation code
D=100  # Diffusion parameter
CFL=D*delta_t/delta_x**2
for t in range(Nt-1):
    U1[1:-1,t+1] = CFL*U1[2:,t] + CFL*U1[:-2,t] + \
                   (-2*CFL + 1 + delta_t)* U1[1:-1,t] - delta_t*U1[1:-1,t]**2

# figure code
_ = plt.figure()
_ = plt.xlabel('Frame number'); _ = plt.ylabel('Bleeding')
ts = np.floor(np.linspace(0,Nt-1,10))
for t in ts.astype(int):
  roundedt = round(T[t],1)
  _ = plt.plot(X,U1[:,t], label='t=%s'%roundedt)

_ = plt.legend()
plt.savefig('test-patient-23.pdf')
plt.close()
```

Interpret with python2

```
```

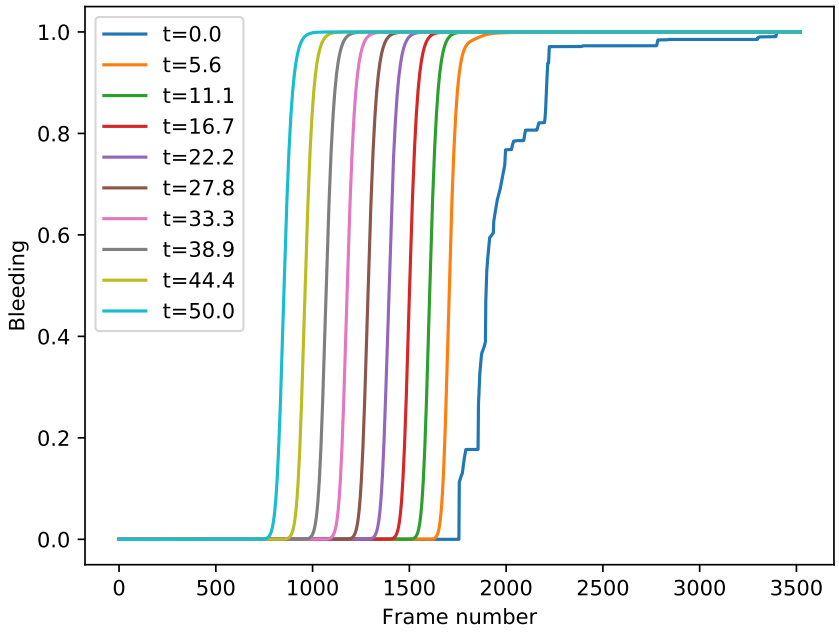


FIGURE 7.16 – Simulation de l'évolution des saignements selon le modèle de Fisher-KPP pour le patient 23 avec un coefficient de diffusion $D = 100$. On observe l'apparition d'un front d'onde se déplaçant à la vitesse $2\sqrt{D} = 20$.

Un patient $p(s)$ est observé à un instant t qui n'est pas l'instant initial. En effet, les lois de la biologie modélisées par l'équation de Fisher-KPP s'appliquent à tout moment de la vie du patient. Le moment où l'examen est réalisé n'est pas l'instant où la maladie débute ; celle-ci n'est pas "inoculée". Par conséquent, l'inflammation a évolué pendant un certain temps inconnu avant que le patient fasse des examens. D'autre part, considérer que le patient est observé à $t = 0$ conduirait à un problème mal posé. Comme il n'y a pas de restriction sur la condition initiale et que le coefficient de diffusion D est un paramètre libre, on serait obligé de conclure $u_0(s) = p(s)$ et D inconnu.

On est donc amenés à chercher le modèle (D, u_0) tel qu'il existe un instant t pour lequel la solution $u(s,t)$ de l'équation de Fisher-KPP coincide avec l'observation $p(s)$ du patient. On considérera ici que cet instant t inconnu est grand, c'est-à-dire que le patient est observé en régime asymptotique. Pour l'équation de Fisher-KPP, avec les conditions initiales $u = 0$ à gauche et $u = 1$ à droite, ce régime asymptotique est nécessairement un front d'onde progressant vers $-\infty$ [62].

Plus précisément, pour les modèles $(D = 1, u_0)$ où u_0 décroît suffisamment vite, c'est-à-dire $u(s,0) << s^\alpha e^{-s\sqrt{f'(0)}}$ pour un certain $\alpha < -2$, et en particulier pour $u(s,0) = 0$ à partir d'une certaine valeur, le comportement asymptotique est un front d'onde W_c indépendant de u_0, de vitesse $c = 2\sqrt{f'(0)} = 2$. La forme du front d'onde est solution de l'équation différentielle $W_c'' + cW_c' + f(W_c) = 0$, mais sa forme explicite n'est pas connue dans le cas général [32]. Pour les modèles (D, u_0), le comportement asymptotique et la forme du front d'onde s'obtiennent par changement de variable, donc par dilatation de W_c d'un facteur $\sqrt{D}$.

On a donc réduit l'ensemble des modèles $\mathcal{M}$ à un ensemble de formes de fronts d'onde $\sqrt{D}.W_c$, qui peut être paramétré par le coefficient de diffusion D. Le problème inverse consiste à trouver les coefficients D compatibles, c'est-à-dire les fronts d'onde compatibles avec la fonction $p(s)$ à translation près. L'ensemble des vitesses compatibles est obtenu avec la formule $v = 2\sqrt{D}$.

La fonction `fkpp_interp` réalise le recalage du front d'onde $\sqrt{D}.W_c$ par translation. On a remarqué qu'il suffit d'égaliser l'intégrale du signal pour obtenir une bonne approximation de la translation.

Code chunk 24 : «`fkpp_recalage (part 2)`»

```
def fkpp_interp(p,D,figname=None):
    Xp = np.linspace(0,1,num=len(p))
    sqrtD = np.sqrt(D)
    Xi = sqrtD*X \
         + (np.trapz(U1,dx=delta_x*sqrtD) -np.trapz(p,dx=Xp[1]-Xp[0]) \
         - (xL*sqrtD-1))
    Ui = np.interp(Xp,Xi,U1)
    distD = np.sqrt(np.trapz( (p-Ui)**2,dx=Xp[1]-Xp[0] ))
    if figname != None:
       fig,ax = plt.subplots()
       ax.plot(Xi,U1)
       ax.plot(Xp,p)
       ax.set_title("sqrtD = " + str(round(sqrtD,3)) + \
                    "  distance = " + str(round(distD,3))  )
       ax.set_xlabel("Abscisse curviligne s")
       fig.savefig(figname)
    return distD

fkpp_interp(p,0.014**2,'fkpp-recalage.pdf')
```

Interpret with `python2`

```
0.03539466016294326
```

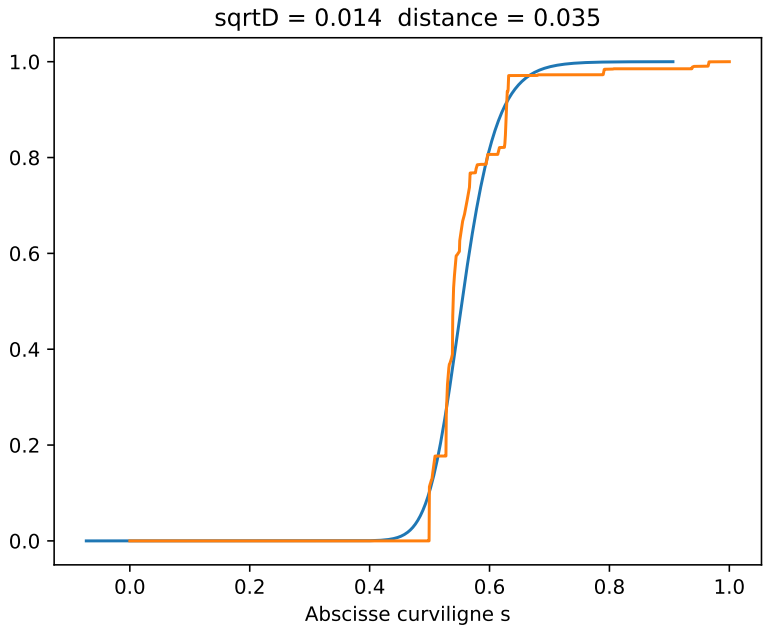


FIGURE 7.17 – Recalage du front d'onde $\sqrt{D}.W_c$ sur le patient 23 avec un coefficient de diffusion $D = 0.014^2$.

On explore l'ensemble des modèles en échantillonnant les valeurs de $\sqrt{D}$, et on en déduit les vitesses 0.05-compatibles :

Code chunk 25 : «fkpp_recalage (part 3)»

```
D = np.arange(0.01,0.02,0.0005) ** 2
dists = [fkpp_interp(p,d) for d in D]
alpha, min_dist = 0.05, np.amin(dists)
scompat = 2 * np.sqrt(np.extract(dists<(1+alpha)*min_dist,D))
print 'Vitesses compatibles', scompat
Xp = np.linspace(0,1,num=len(p))
print 'Temps d\'invasion',np.round((1-np.trapz(p,dx=Xp[1]-Xp[0]))/scompat,2)
```

Interpret with `python2`

```
Vitesses compatibles [0.025 0.026 0.027 0.028 0.029 0.03  0.031]
Temps d'invasion [22.39 21.53 20.73 19.99 19.3  18.66 18.06]
```

Comme présenté ci-dessus, on obtient des vitesses de propagation du front d'onde pour un patient, que l'on peut convertir en une date d'invasion à laquelle l'inflammation aurait envahi l'ensemble du colon, i.e. le segment $[0, 1]$. Cependant, l'unité de temps n'a pas été précisée, et il faudrait la déterminer en utilisant des coloscopies à des dates différentes.

7.7 Conclusion et perspectives

Les travaux présentés dans ce chapitre mélangent les thématiques de l'analyse d'image et de la modélisation mathématique. Le cadre formel présenté au chapitre 2 permet de les traiter de la même façon, c'est-à-dire sous la forme de la recherche de modèles compatibles dans un ensemble de modèles $\mathcal{M}$. De même, ce chapitre illustre sur des exemples concrets comment définir des énoncés scientifiques et comment y apporter des éléments de preuves avec les outils et la rigueur des mathématiques.

Dans ces travaux, on voit apparaître une difficulté scientifique majeure qui réside dans la formulation des énoncés scientifiques. La plupart du temps, les hypothèses scientifiques que l'on explore sont vagues et mal définies. Une partie du travail mathématique ou de modélisation va consister à explorer les différentes définitions, à la fois en calculant la validité des différentes définitions possibles, comme pour les énoncés E1 et E2, mais aussi en proposant de nouveaux outils et modèles mathématiques correspondant à de "nouvelles définitions". Il est intéressant de remarquer qu'il s'agit de définir et de décrire les caractéristiques de sous-ensembles particuliers de modèles.

Plusieurs types de modèles sont utilisés dans ce chapitre suivant l'objet étudié, et les données sont vues comme des exemples informatifs :

- un modèle de l'aspect visuel d'une lésion est une fonction du type image $\to \{0, 1\}$ qui indique quelles formes ou images sont associées à ce type de lésion (saignement ou ulcère). On envisage donc le problème de classification ou de détection des lésions comme l'estimation de l'ensemble des images correspondant à chaque classe. Les annotations des médecins fournissent des exemples à partir desquels on estime une partition de l'espace des modèles, ici de l'espace des couleurs.
- un modèle de la sévérité de la RCH est une fonction du type patient $\to \mathbb{R}$ qui représente la liaison entre les marqueurs biologiques et/ou endoscopiques et la gravité de la maladie. Chaque couple (coloscopie, score UCEIS) fournit un exemple que le modèle de sévérité doit tenter de reproduire.
- un modèle de la répartition des lésions le long du colon est une fonction du type $[0, 1] \to \mathbb{R}$ qui représente l'abondance des lésions en fonction de la position dans le colon. Un seul exemple est fourni pour chaque patient dans notre base de données actuelle.

Enfin, l'énumération des modèles compatibles est une forme d'analyse des données qui n'est pas un test statistique, car on ne formule pas d'hypothèse à rejeter, et qui n'est pas non plus du domaine des statistiques inférentielles, car

on ne quantifie pas l'incertitude sur un "vrai paramètre", mais l'adéquation des modèles eux-mêmes. Cette analyse s'écarte aussi des approches fréquentistes et bayésiennes, car il n'y a ni estimation ni calcul par rapport à une distribution a priori. Cette approche ouvre la possibilité d'envisager la structure mathématique ou topologique des énoncés scientifiques compatibles avec des données, et, par exemple, de les énumérer systématiquement.

Chapitre 8

Conclusion générale

Dans ce document, on aborde le problème de l'analyse de données, et des modèles utilisés pour conduire cette analyse. La question principale ici est de savoir comment conduire une analyse rigoureuse alors que des hypothèses de modélisation plus ou moins arbitraires sont nécessaires.

Au travers de mes recherches, j'ai été confronté à plusieurs domaines applicatifs, avec lesquels l'interaction a été riche, et pour lesquels il m'a été nécessaire de construire un cadre unificateur pour l'analyse de données. Dans le cadre de la thèse de Tran Dai Viet, on s'est intéressé aux modèles basés patchs des images naturelles et médicales. Dans le cadre de la thèse de Tran Duc Nghia, nous avons exploré les modèles de signaux de résonance paramagnétique électronique pour la biochimie. Dans la thèse de Safaa Al Ali, nous avons étudié les modèles de la sévérité et de la répartition des lésions de la rectocolite hémorragique dans le colon.

C'est l'interaction avec des scientifiques de toutes ces disciplines qui a permis l'élaboration d'un cadre général. Sans eux, ma réflexion aurait été cantonnée à l'un ou l'autre des outils mathématiques, à l'analyse d'images, au traitement du signal, aux statistiques bayésiennes ou aux équations différentielles. Je pense aujourd'hui pouvoir comprendre un problème d'analyse de données indépendamment d'un domaine ou d'un ensemble d'outils mathématiques. Je pense avoir la méthodologie pour choisir et combiner ces différents outils mathématiques.

Au coeur de mon approche, il y a la notion de structure des ensembles des données $\mathcal{E}$ et des modèles $\mathcal{M}$, et la similarité qui les relie. Ainsi, les hypothèses de modélisation sont traduites dans les propriétés des ensembles ou sous-ensembles choisis, et transportées d'un côté ou de l'autre par la similarité. Dans ce cadre, on remarque que les énoncés scientifiques sont aussi des sous-ensembles (de $\mathcal{E}$ ou de $\mathcal{M}$), et que leur démonstration passe par l'accumulation d'observations,

ou de modèles similaires aux observations. Enfin, la construction de ces énoncés scientifiques, souvent confiée à l'expertise du modélisateur, peut être obtenue par une méthode d'apprentissage automatique ou par énumération.

Ce cadre général permet de mieux comprendre la nature et la portée des preuves contenues dans un travail d'analyse de données. Il fournit également quelques grands axes naturels de travail tels que

— **la zoologie des espaces et hypothèses de modélisation**. Le cadre présenté permet d'envisager une modélisation mathématique transverse aux sous-domaines tels que la théorie des probabilités, les méthodes variationnelles, la morphologie mathématique, etc. Il s'agit de rassembler les différentes approches, non pas en fonction des outils mathématiques employés, mais plutôt en fonction du type de données auquel les modèles se réfèrent : ensembles de cardinal fini, nombres entiers, nombres réels, chaînes de caractères, signaux, images, etc.

— **les stratégies d'accumulation de modèles**. Certaines hypothèses de modélisation sont difficiles à justifier, la nature du bruit (gaussien, poissonien, etc.) dans un modèle d'observation par exemple. Dans ce cas de figure, le cadre présenté ici indique que les énoncés scientifiques valides sont ceux qui sont supportés par tous les types de bruits. La question est alors de savoir comment construire une analyse de données qui explore tous les types de bruits possibles. Une première possibilité consiste à considérer des suites d'espaces de modèles de complexité croissante : modèles hiérarchiques, modèles de mélange, etc. Une autre possibilité revient à s'intéresser à la notion de flexibilité des modèles en tant que "complétude", c'est-à-dire à chercher un sous-ensemble de modèles dense dans l'espace $\mathcal{M}$.

— **le choix d'une méthode d'apprentissage**. Un des grands problèmes ouverts de l'apprentissage par réseaux convolutionnels consiste à comprendre comment la structure du réseau influence ses performances. Le cadre présenté ici permet d'envisager cette question comme le calcul de l'espace des modèles engendré par un réseau donné, et d'envisager la vitesse d'apprentissage comme la vitesse de convergence d'un algorithme d'approximation en fonction du nombre de données. Beaucoup de méthodes de validation actuelles (validation croisée, bootstrap, etc.) consistent à explorer le voisinage des données d'apprentissage. Le cadre présenté ici souligne la faiblesse des conclusions basées sur un seul modèle et permet d'envisager des méthodes d'exploration du voisinage dans l'espace des modèles.

Publications

[1] I. Abraham et al. « Significant edges in the case of non-stationary Gaussian noise ». In : *Pattern Recognition* 40.11 (2007), p. 3277-3291.

[2] S. Li-Thiao-Té. « Semiparametric Estimation of the Gain Parameter with Quantization Errors ». In : (2008). URL : http://hal.archives-ouvertes.fr/hal-00390325/fr/.

[3] M. Vandenbogaert et al. « Alignment of LC-MS images, with applications to biomarker discovery and protein identification ». In : *Proteomics* 8 (fév. 2008), p. 650-672.

[4] Sebastien Li-Thiao-Te et al. « Modèles de mélange tronqués pour l'écologie microbienne. Estimation du nombre d'espèces manquantes. » Français. In : *42èmes Journées de Statistique*. Marseille, France, France, 2010. URL : http://hal.inria.fr/inria-00494719.

[5] Sébastien Li-Thiao-Te, Jean-Jacques Daudin et Stephane Robin. « Mixture models of truncated data for estimating the number of species ». In : *19. International Conference on Computational Statistics*. Short paper : Biostatistics. Paris, France, août 2010, np. URL : https://hal.archives-ouvertes.fr/hal-01197572.

[6] J.J. Daudin, S. Li-Thiao-Te et S. Robin. « Bayesian model averaging for estimating the number of classes. Applications to the total number of species in metagenomics ». In : *Journal of Applied Statistics* (2012).

[7] S. Li-Thiao-Té et B. Schwikowski. « Feature Detection with the M-N rule in Liquid Chromatography-Mass Spectrometry Images ». In : *Journal of Computational Biology* (2012).

[8] Sébastien Li-Thiao-Té. « Literate Program Execution for Reproducible Research and Executable Papers ». In : *Procedia Computer Science* 9.0 (2012). Proceedings of the International Conference on Computational Science, ICCS 2012, p. 439-448. ISSN : 1877-0509. DOI : 10.1016/j.procs.2012.04.047. URL : http://www.sciencedirect.com/science/article/pii/S1877050912001688.

[9] Sébastien Li-Thiao-Té. « Literate Program Execution for Teaching Computational Science ». In : *Procedia Computer Science* 9.0 (2012). Proceedings of the International Conference on Computational Science, ICCS 2012, p. 1723-1732. ISSN : 1877-0509. DOI : 10.1016/j.procs.2012.04.190. URL : http://www.sciencedirect.com/science/article/pii/S1877050912003110.

[10] Sébastien Li-Thiao-Té. « Using Lepton for documenting source code : a guided example in computer vision ». In : *ESAIM : ProcS* 45 (2014), p. 239-246. DOI : 10.1051/proc/201445024. URL : https://doi.org/10.1051/proc/201445024.

[11] Dai-Viet Tran et al. « Super-resolution for medical images corrupted by heavy noise ». In : *Medical Imaging 2015 : Image Processing, Orlando, Florida, USA, February 24-26, 2015*. SPIE, 2015, 94130E. DOI : 10.1117/12.2082314. URL : http://dx.doi.org/10.1117/12.2082314.

[12] Dai-Viet Tran et al. « Example-Based Super-Resolution for Enhancing Spatial Resolution of Medical Images ». In : *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Orlando, Florida, USA, August, 2016*. IEEE, 2016.

[13] N. Duc-Tran, Y. M. Frapart et S. Li-Thiao-Té. « Estimation of spectrum parameters for Quantitative EPR in the derivative limit ». In : *2017 International Conference on Advanced Technologies for Communications (ATC)*. Oct. 2017, p. 214-219. DOI : 10.1109/ATC.2017.8167620.

[14] D. V. Tran et al. « Patch-based image denoising : Probability distribution estimation vs. sparsity prior ». In : *2017 25th European Signal Processing Conference (EUSIPCO)*. Août 2017, p. 1490-1494. DOI : 10.23919/EUSIPCO.2017.8081457.

[15] Dai-Viet Tran et al. « Number of Useful Components in Gaussian Mixture Models for Patch-Based Image Denoising ». In : *Image and Signal Processing - 8th International Conference, ICISP 2018, Cherbourg, France, July 2-4, 2018, Proceedings*. 2018, p. 108-116. DOI : 10.1007/978-3-319-94211-7_13. URL : https://doi.org/10.1007/978-3-319-94211-7%5C_13.

[16] Sébastien Li-Thiao-Té. *Lepton : An Automaton for Literate Executable Papers*. Version v1.2. Oct. 2019. DOI : 10.5281/zenodo.3492221. URL : https://doi.org/10.5281/zenodo.3492221.

[17] Sébastien Li-Thiao-Té. « Lepton : An automaton for Literate Executable Papers ». In : *Journal of Open Source Software* 4.42 (oct. 2019), p. 1005. DOI : 10.21105/joss.01005. URL : https://doi.org/10.21105%2Fjoss.01005.

[18] N. Duc-Tran, Y. M. Frapart et S. Li-Thiao-Té. « Parameter Estimation for Quantitative EPR Spectroscopy ». In : *IEEE Transactions on Instrumentation and Measurement* (2021). DOI : 10.1109/TIM.2021.3084289.

Références

[19] M. Aharon, M. Elad et A. Bruckstein. « K-SVD : An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation ». In : *IEEE Transactions on Signal Processing* 54.11 (nov. 2006), p. 4311-4322.

[20] Rizwan Ahmad et Periannan Kuppusamy. « Theory, instrumentation, and applications of electron paramagnetic resonance oximetry ». In : *Chemical reviews* 110.5 (2010), p. 3212-3236.

[21] Monagi H Alkinani et Mahmoud R El-Sakka. « Patch-based models and algorithms for image denoising : a comparative review between patch-based images denoising methods for additive noise reduction ». In : *EURASIP Journal on Image and Video Processing* 2017.1 (2017), p. 1-27.

[22] David B. Allison, Richard M. Shiffrin et Victoria Stodden. « Reproducibility of research : Issues and proposed remedies ». In : *Proceedings of the National Academy of Sciences* 115.11 (2018), p. 2561-2562. issn : 0027-8424. doi : 10.1073/pnas.1802324115. eprint : https://www.pnas.org/content/115/11/2561.full.pdf. url : https://www.pnas.org/content/115/11/2561.

[23] Sven Andersson et al. *Free radicals*. US Patent 5,530,140. Juin 1996.

[24] F. J. ANSCOMBE. « THE TRANSFORMATION OF POISSON, BINOMIAL AND NEGATIVE-BINOMIAL DATA ». In : *Biometrika* 35.3-4 (déc. 1948), p. 246-254. issn : 0006-3444. doi : 10.1093/biomet/35.3-4.246. eprint : https://academic.oup.com/biomet/article-pdf/35/3-4/246/785684/35-3-4-246.pdf. url : https://doi.org/10.1093/biomet/35.3-4.246.

[25] BH Armstrong. « Spectrum line profiles : the Voigt function ». In : *Journal of Quantitative Spectroscopy and Radiative Transfer* 7.1 (1967), p. 61-88.

[26] Anita Bálint et al. « How disease extent can be included in the endoscopic activity index of ulcerative colitis : the panMayo score, a promising scoring system ». In : *BMC gastroenterology* 18.1 (2018), p. 7.

[27] A. Boutier-Pischon et al. « EPR and electrochemical quantification of oxygen using newly synthesized para-silylated triarylmethyl radicals ». In : *Free Radical Research* 49.3 (2015), p. 236-243. doi : 10.3109/10715762.2014.995183. eprint : https://doi.org/10.3109/10715762.2014.995183. url : https://doi.org/10.3109/10715762.2014.995183.

[28] George E. P. Box. « Science and Statistics ». In : *Journal of the American Statistical Association* 71.356 (1976), p. 791-799. doi : 10.1080/01621459.1976.10480949. eprint : https://www.tandfonline.com/doi/pdf/10.1080/01621459.1976.10480949. url : https://www.tandfonline.com/doi/abs/10.1080/01621459.1976.10480949.

[29] Grant R. Brammer et al. « Paper Mâché : Creating Dynamic Reproducible Science ». In : *Procedia Computer Science* 4.0 (2011). Proceedings of the International Conference on Computational Science, ICCS 2011, p. 658-667. issn : 1877-0509. doi : 10.1016/j.procs.2011.04.069. url : http://www.sciencedirect.com/science/article/pii/S187705091100127X.

[30] Georg Brandl, Tim Hatch et Armin Ronacher. *Pygments*. url : http://pygments.org/.

[31] Karl Broman et al. « Recommendations to funding agencies for supporting reproducible research ». In : *American statistical association*. T. 2. 2017.

[32] Eric Brunet et Bernard Derrida. « An Exactly Solvable Travelling Wave Equation in the Fisher–KPP Class ». In : *Journal of Statistical Physics* (2015), p. 1-20. doi : 10.1007/s10955-015-1350-6. url : https://hal.sorbonne-universite.fr/hal-01196767.

[33] Emmanuel J Candès et al. « Compressive sampling ». In : *Proceedings of the international congress of mathematicians*. T. 3. Madrid, Spain. 2006, p. 1433-1452.

[34] Christophe Cérin et al. *Approches contemporaines en hébergement et gestion de données*. 2017.

[35] Christophe Cérin et al. *Utilizing Big Data Paradigms for Business Intelligence*. Sous la dir. de Jérôme Darmont et Sabine Loudcher. IGI Global, 2018. Chap. 1.

[36] H. Chang, D.-Y. Yeung et Y. Xiong. « Super-resolution through neighbor embedding ». In : *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*. T. 1. Juin 2004, p. 1275-1282.

[37] Le Chang et Doris Y. Tsao. « The Code for Facial Identity in the Primate Brain ». In : *Cell* 169.6 (juin 2017), 1013-1028.e14. issn : 0092-8674. doi : 10.1016/j.cell.2017.05.011. url : https://doi.org/10.1016/j.cell.2017.05.011.

[38] Natalie Cooper et al. « A guide to reproducible code in ecology and evolution ». In : (2017).

[39] Christophe Decroos et al. « Oxidative and Reductive Metabolism of Tris(p-carboxyltetrathiaaryl)methyl Radicals by Liver Microsomes ». In : *Chemical Research in Toxicology* 22.7 (2009). PMID : 19545126, p. 1342-1350. doi : 10.1021/tx9001379. eprint : https://doi.org/10.1021/tx9001379. url : https://doi.org/10.1021/tx9001379.

[40] Yuanmu Deng et al. « Application of magnetic field over-modulation for improved EPR linewidth measurements using probes with Lorentzian lineshape ». In : *Journal of Magnetic Resonance* 181.2 (2006), p. 254-261.

[41] Ilirian Dhimitruka et al. « Large-scale synthesis of a persistent trityl radical for use in biomedical EPR applications and imaging ». In : *Bioorganic & medicinal chemistry letters* 17.24 (2007), p. 6801-6805.

[42] Carsten Dominik. *The Org-Mode 7 Reference Manual : Organize Your Life with GNU Emacs*. with contributions by David O'Toole, Bastien Guerry, Philip Rooke, Dan Davison, Eric Schulte, and Thomas Dye. UK : Network Theory, 2010.

[43] Gareth R Eaton et al. *Quantitative Epr*. Springer Science & Business Media, 2010.

[44] David Ferrucci et al. « Watson : Beyond Jeopardy ! » In : *Artificial Intelligence* 199-200 (2013), p. 93-105. issn : 0004-3702. doi : https://doi.org/10.1016/j.artint.2012.06.009. url : https://www.sciencedirect.com/science/article/pii/S0004370212000872.

[45] Matan Gavish et David Donoho. « A Universal Identifier for Computational Results ». In : *Procedia Computer Science* 4.0 (2011). Proceedings of the International Conference on Computational Science, ICCS 2011, p. 637-647. issn : 1877-0509. doi : 10.1016/j.procs.2011.04.067. url : http://www.sciencedirect.com/science/article/pii/S1877050911001256.

[46] Tonmoy Ghosh, Antara Das et Rosni Sayed. « Automatic small intestinal ulcer detection in capsule endoscopy images ». In : *International Journal of Scientific and Engineering Research* 7.10 (2016), p. 737-741.

[47] Tonmoy Ghosh, Shaikh Anowarul Fattah et Khan A Wahid. « CHOBS : color histogram of block statistics for automatic bleeding detection in wireless capsule endoscopy video ». In : *IEEE journal of translational engineering in health and medicine* 6 (2018), p. 1-12.

[48] Tonmoy Ghosh et al. « Cluster based statistical feature extraction method for automatic bleeding detection in wireless capsule endoscopy video ». In : *Computers in biology and medicine* 94 (2018), p. 41-54.

[49] Kurt Gödel. « Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I ». In : *Monatshefte für mathematik und physik* 38.1 (1931), p. 173-198.

[50] Pieter Van Gorp et Steffen Mazanek. « SHARE : a web portal for creating and sharing executable research papers ». In : *Procedia Computer Science* 4.0 (2011). Proceedings of the International Conference on Computational Science, ICCS 2011, p. 589-597. issn : 1877-0509. doi : 10.1016/j.procs.2011.04.062. url : http://www.sciencedirect.com/science/article/pii/S1877050911001207.

[51] Konrad Hinsen. « A data and code model for reproducible research and executable papers ». In : *Procedia Computer Science* 4.0 (2011). Proceedings of the International Conference on Computational Science, ICCS 2011, p. 579-588. issn : 1877-0509. doi : 10.1016/j.procs.2011.04.061. url : http://www.sciencedirect.com/science/article/pii/S1877050911001190.

[52] Emmanuel Hirsch. *Imagerie cérébrale : enjeux épistémologiques, éthiques et politiques*. Espace éthique Région Ile-de-France : Les cahiers de l'espace éthique, 2018.

[53] Emmanuel Hirsch, Léo Coutellec et Paul-Loup Weil-Dubuc. *Big data et pratiques biomédicales, Implications éthiques et sociétales dans la recherche, les traitements et le soin*. Espace éthique Région Ile-de-France : Les cahiers de l'espace éthique, 2015.

[54] Emmanuel Hirsch et François Hirsch. *Traité de bioéthique IV : Les nouveaux territoires de la bioéthique*. Eres, 2018.

[55] Benjamin Hofner, Matthias Schmid et Lutz Edler. « Reproducible research in statistics : A review and guidelines for the Biometrical Journal ». In : *Biometrical journal* 58.2 (2016), p. 416-427.

[56] *Image Processing On Line*. doi : 10.5201/ipol. url : http://www.ipol.im/.

[57] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[58] Jeffry A Katz, Gil Melmed, Bruce E Sands et al. « The facts about inflammatory bowel diseases ». In : *Crohn's & Colitis Foundation of America, New York* (2011).

[59] J Kirchgesner et al. « Therapeutic management of inflammatory bowel disease in real-life practice in the current era of anti-TNF agents : analysis of the French administrative health databases 2009–2014 ». In : *Alimentary pharmacology & therapeutics* 45.1 (2017), p. 37-49.

[60] Thomas Kluyver et al. « Jupyter Notebooks – a publishing format for reproducible computational workflows ». In : *Positioning and Power in Academic Publishing : Players, Agents and Agendas*. Sous la dir. de F. Loizides et B. Schmidt. IOS Press. 2016, p. 87-90.

[61] Donald E. Knuth. « Literate programming ». In : *THE COMPUTER JOURNAL* 27 (1984), p. 97-111.

[62] Andrei N Kolmogorov. « Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique ». In : *Bull. Univ. Moskow, Ser. Internat., Sec. A* 1 (1937), p. 1-25.

[63] David Koop et al. « A Provenance-Based Infrastructure to Support the Life Cycle of Executable Papers ». In : *Procedia Computer Science* 4.0 (2011). Proceedings of the International Conference on Computational Science, ICCS 2011, p. 648-657. issn : 1877-0509. doi : 10.1016/j.procs.2011.04.068. url : http://www.sciencedirect.com/science/article/pii/S1877050911001268.

[64] Nimisha Elsa Koshy et Varun P Gopi. « A new method for ulcer detection in endoscopic images ». In : *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*. IEEE. 2015, p. 1725-1729.

[65] Alex Krizhevsky, Ilya Sutskever et Geoffrey E Hinton. « Imagenet classification with deep convolutional neural networks ». In : *Advances in neural information processing systems* 25 (2012), p. 1097-1105.

[66] Virendra Kumar et al. « Radiomics : the process and the challenges ». In : *Magnetic Resonance Imaging* 30.9 (2012). Quantitative Imaging in Cancer, p. 1234-1248. issn : 0730-725X. doi : https://doi.org/10.1016/j.mri.2012.06.010. url : https://www.sciencedirect.com/science/article/pii/S0730725X12002202.

[67] Philippe Lambin et al. « Radiomics : Extracting more information from medical images using advanced feature analysis ». In : *European Journal of Cancer* 48.4 (mars 2012), p. 441-446. issn : 0959-8049. doi : 10.1016/j.ejca.2011.11.036. url : https://doi.org/10.1016/j.ejca.2011.11.036.

[68] Ib Leunbach et Jan Henrik Ardenkjaer-Larsen. *Method for determining oxygen concentration using magnetic resonance imaging*. US Patent 5,765,562. Juin 1998.

[69] R.C. Machado, L. Rittner et R.A. Lotufo. « Adessowiki Collaborative platform for writing executable papers ». In : *Procedia Computer Science* 4 (2011), p. 759-767. URL : https://www.sciencedirect.com/science/article/pii/S1877050911001384.

[70] Stéphane Mallat. *A Wavelet Tour of Signal Processing (The Sparse Way)*. Third Edition. Academic Press, 2009. ISBN : 978-0-12-374370-1. DOI : https://doi.org/10.1016/B978-0-12-374370-1.50001-9. URL : http://www.sciencedirect.com/science/article/pii/B9780123743701500019.

[71] Bjoern H Menze et al. « The multimodal brain tumor image segmentation benchmark (BRATS) ». In : *IEEE transactions on medical imaging* 34.10 (2014), p. 1993-2024.

[72] Piotr Nowakowski et al. « The Collage Authoring Environment ». In : *Procedia Computer Science* 4.0 (2011). Proceedings of the International Conference on Computational Science, ICCS 2011, p. 608-617. ISSN : 1877-0509. DOI : 10.1016/j.procs.2011.04.064. URL : http://www.sciencedirect.com/science/article/pii/S1877050911001220.

[73] J Palmer, LC Potter et R Ahmad. « Optimization of magnetic field sweep and field modulation amplitude for continuous-wave EPR oximetry ». In : *Journal of Magnetic Resonance* 209.2 (2011), p. 337-340.

[74] Ofir Pele et Michael Werman. « Fast and robust earth mover's distances ». In : *2009 IEEE 12th International Conference on Computer Vision*. IEEE. 2009, p. 460-467.

[75] JR Pilbrow. « Principles of computer simulation of EPR spectra ». In : *Applied Magnetic Resonance* 10.1-3 (1996), p. 45-53.

[76] Konstantin Pogorelov et al. « Bleeding detection in wireless capsule endoscopy videos—Color versus texture features ». In : *Journal of applied clinical medical physics* 20.8 (2019), p. 141-154.

[77] Charles P Poole. *Electron spin resonance : a comprehensive treatise on experimental techniques*. Courier Corporation, 1996.

[78] N. Ramsey. « Literate programming simplified ». In : *Software, IEEE* 11.5 (1994), p. 97-105.

[79] *rmarkdown : Dynamic Documents for R*. R package version 2.5.1. 2020. URL : https://github.com/rstudio/rmarkdown.

[80] BH Robinson, C Mailer et AW Reese. « Linewidth analysis of spin labels in liquids : I. Theory and data analysis ». In : *Journal of Magnetic Resonance* 138.2 (1999), p. 199-209.

[81] PJ Robinson. « Radiology's Achilles' heel : error and variation in the interpretation of the Röntgen image. » In : *The British Journal of Radiology* 70.839 (1997), p. 1085-1098.

[82] RStudio Team. *RStudio : Integrated Development Environment for R*. RStudio, PBC. Boston, MA, 2020. URL : http://www.rstudio.com/.

[83] Y. Rubner, C. Tomasi et L. J. Guibas. « A metric for distributions with applications to image databases ». In : *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*. 1998, p. 59-66. DOI : 10.1109/ICCV.1998.710701.

[84] A Rule et al. « Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks ». In : *PLoS Comput Biol* 15.7 (2019), e1007007.

[85] Kenneth W Schroeder, William J Tremaine et Duane M Ilstrup. « Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis ». In : *New England Journal of Medicine* 317.26 (1987), p. 1625-1629.

[86] Kenneth W Schroeder, William J Tremaine et Duane M Ilstrup. « Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis ». In : *New England Journal of Medicine* 317.26 (1987), p. 1625-1629.

[87] Li-Thiao-Té Sébastien. « lepton : v1.0 ». In : (juill. 2018). DOI : 10.5281/zenodo.1311588. URL : https://doi.org/10.5281/zenodo.1311588.

[88] David Silver et al. « Mastering the game of Go with deep neural networks and tree search ». In : *Nature* 529.7587 (jan. 2016), p. 484-489. ISSN : 1476-4687. DOI : 10.1038/nature16961. URL : https://doi.org/10.1038/nature16961.

[89] David Silver et al. « Mastering the game of Go without human knowledge ». In : *Nature* 550.7676 (oct. 2017), p. 354-359. DOI : 10.1038/nature24270.

[90] Simon PL Travis et al. « Developing an instrument to assess the endoscopic severity of ulcerative colitis : the Ulcerative Colitis Endoscopic Index of Severity (UCEIS) ». In : *Gut* 61.4 (2012), p. 535-542.

[91] D. H. Trinh et al. « An effective example-based learning method for denoising of medical images corrupted by heavy Gaussian noise and poisson noise ». In : *2014 IEEE International Conference on Image Processing*. Oct. 2014, p. 823-827.

[92] D. H. Trinh et al. « Novel Example-Based Method for Super-Resolution and Denoising of Medical Images ». In : *IEEE Transactions on Image Processing* 23.4 (avr. 2014), p. 1882-1895.

[93] Joel A Tropp et Anna C Gilbert. « Signal recovery from random measurements via orthogonal matching pursuit ». In : *IEEE Transactions on information theory* 53.12 (2007), p. 4655-4666.

[94] V. N. Vapnik et A. Ya. Chervonenkis. « On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities ». In : *Measures of Complexity : Festschrift for Alexey Chervonenkis*. Sous la dir. de Vladimir Vovk, Harris Papadopoulos et Alexander Gammerman. Cham : Springer International Publishing, 2015, p. 11-30. isbn : 978-3-319-21852-6. doi : 10.1007/978-3-319-21852-6_3. url : https://doi.org/10.1007/978-3-319-21852-6_3.

[95] Carl Vondrick, Donald Patterson et Deva Ramanan. « Efficiently scaling up crowdsourced video annotation ». In : *International journal of computer vision* 101.1 (2013), p. 184-204.

[96] Hugo Wahlquist. « Modulation broadening of unsaturated Lorentzian lines ». In : *The Journal of Chemical Physics* 35.5 (1961), p. 1708-1710.

[97] John A Weil et James R Bolton. *Electron paramagnetic resonance : elementary theory and practical applications*. John Wiley & Sons, 2007.

[98] Samuel S Wilks. « Mathematical statistics ». In : *New York, John Wileyand Sons* (1962).

[99] Yihui Xie. *knitr : A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.30. 2020. url : https://yihui.org/knitr/.

[100] Jun Xu et al. « Patch group based nonlocal self-similarity prior learning for image denoising ». In : *Proceedings of the IEEE international conference on computer vision*. 2015, p. 244-252.

[101] J. Yang et al. « Image Super-Resolution Via Sparse Representation ». In : *IEEE Transactions on Image Processing* 19.11 (nov. 2010), p. 2861-2873.

[102] G. Yu, G. Sapiro et S. Mallat. « Solving Inverse Problems With Piecewise Linear Estimators : From Gaussian Mixture Models to Structured Sparsity ». In : *IEEE Transactions on Image Processing* 21.5 (2012), p. 2481-2499. doi : 10.1109/TIP.2011.2176743.

[103] Daniel Zoran et Yair Weiss. « Natural images, Gaussian mixtures and dead leaves ». In : *Advances in Neural Information Processing Systems*. 2012, p. 1736-1744.

[104] H. Zou et T. Hastie. « Regularization and variable selection via the Elastic Net ». In : *Journal of the Royal Statistical Society, Series B* 67 (2005), p. 301-320.