MÉMOIRE
Pour obtenir le diplôme

HABILITATION À DIRIGER DES RECHERCHES DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et Technologies de l'Information, Informatique
Spécialité : Mathématique Appliquée
Unité de recherche : Laboratoire Jean Kuntzmann

Rencontre entre les modèles de mélange et les modèles de segmentation

Meeting of mixing and segmentation models

Présentée par :
Vincent Brault

Rapporteurs :

Arthur CHARPENTIER
Professeur des universités, Université du Québec à Montréal (UQAM)
Franck PICARD
Directeur de Recherche, CNRS
Stéphane ROBIN
Professeur des universités, Sorbonne Université

Habilitation soutenue publiquement le **10 juillet 2025**, devant le jury composé de :

Arthur CHARPENTIER Professeur des universités, Université du Québec à Montréal	Rapporteur
Sophie DONNET Directrice de Recherche, INRAE	Examinatrice
Adeline LECLERCQ SAMSON Professeure des universités, Université Grenoble Alpes	Examinatrice
Céline LÉVY-LEDUC Professeure des universités, Université Paris Cité	Examinatrice
Franck PICARD Directeur de Recherche, CNRS	Rapporteur
Stéphane ROBIN Professeur des universités, Sorbonne Université	Rapporteur

Table des matières

1 Introduction **7**
1.1 Panorama . . . 7
1.1.1 Modèle des blocs latents . . . 7
1.1.2 Modèle de mélange . . . 9
1.1.3 Bisegmentation . . . 10
1.1.4 Classification supervisée . . . 11
1.1.5 Segmentation . . . 12
1.1.6 Mélange de segmentation . . . 12
1.2 Encadrements . . . 13
1.2.1 Projets sur la dysgraphie et la science forensique . . . 13
1.2.2 Consulting . . . 14
1.2.3 Projet Gratweet . . . 14
1.2.4 Modèle des blocs latents non paramétriques . . . 14
1.2.5 Projet CoPoLangues . . . 14
1.2.6 Poolage et Covid 19 . . . 14
1.2.7 Projet sur l'utilisation de l'intelligence artificielle pour le recrutement . . . 15
1.2.8 Projet sur le Syndrome de l'Apnée du Sommeil . . . 15
1.3 Plan du manuscrit . . . 15
1.3.1 Notations générales . . . 15
1.3.2 Plan du manuscrit . . . 15

2 Modèles de mélange et des blocs latents **19**
2.1 Modèle de mélange et tests groupés pour le dépistage du SARS-CoV-2 . . . 19
2.2 Modèle des blocs latents . . . 24
2.2.1 Modèle des blocs latents avec classe de bruit . . . 24
2.2.2 Lien avec le transport optimal . . . 26
2.2.3 Procédures d'estimation et robustesse . . . 27
2.2.4 Qualité des estimations . . . 28
2.3 Comportement singulier de la vraisemblance . . . 30
2.3.1 Consistance des estimateurs . . . 31
2.3.2 Algorithme *Largest Gaps* et estimation en *Big Data* . . . 32

3 Modèles de segmentation et bisegmentation **37**
3.1 Modèle des blocs diagonaux . . . 38
3.2 Modèle de bisegmentation gaussien . . . 41
3.2.1 Estimation des emplacements de ruptures . . . 43
3.2.2 Accélération de l'estimation . . . 45
3.2.3 Consistance des estimateurs . . . 45
3.3 Modèle de bisegmentation non paramétrique . . . 46
3.3.1 Extension à K ruptures . . . 48
3.3.2 Consistance de l'estimateur des emplacements des ruptures connaissant le bon nombre de ruptures . . . 49
3.3.3 Sélection du nombre de ruptures . . . 50
3.4 Application aux véhicules autonomes . . . 50
3.5 Modèle de segmentation et croissance des plants de colza . . . 53
3.5.1 Transformation pour la méthode *LASSO* . . . 53
3.5.2 Estimation des paramètres . . . 54

3.5.3 Applications . . . 55

4 Symbiose entre les mélanges et la segmentation 57
4.1 Courbes électriques et mélange de segmentation . . . 57
4.1.1 Vraisemblance et estimation . . . 58
4.1.2 Résultats théoriques . . . 59
4.1.3 Application aux données de consommation électrique . . . 60
4.2 Protéines Tau et mélange de courbes avec sauts . . . 63
4.2.1 Modélisation . . . 63
4.2.2 Estimation . . . 63
4.2.3 Premiers résultats . . . 64
4.3 MSE^3 . . . 66
4.3.1 Présentation des données . . . 67
4.3.2 *Parsimonious Oscillatory Model of Handwriting* . . . 68
4.3.3 Modèle *POMH* et détection de la dysgraphie . . . 69
4.3.4 Modèle *POMH* et segmentation . . . 73
4.4 Syndrome d'apnée du sommeil et mélange de chaînes de Markov . . . 74

5 Perspectives 79
5.1 Projets sur l'écriture . . . 79
5.2 Projets sur l'apnée du sommeil . . . 80
5.3 Projet Taunique . . . 80
5.4 Modèle des blocs latents : robustesse et version non paramétrique . . . 81
5.5 Projets sur les addictions . . . 81
5.6 Biais dans les algorithmes de recrutement . . . 81
5.7 Développement de l'enfant . . . 82
5.8 Projet Gratweet . . . 82
5.9 Diffusion des codes . . . 82

A Revenir aux bases 83
A.1 Préambule . . . 83
A.2 Modèle de mélange . . . 84
A.2.1 Vraisemblance . . . 85
A.2.2 Identifiabilité et label switching . . . 86
A.2.3 Estimation et algorithme *Espérance Maximisation* . . . 87
A.2.4 Sélection du nombre de classes . . . 89
A.3 Modèle de segmentation . . . 90
A.3.1 Vraisemblance et identifiabilité . . . 91
A.3.2 Estimation . . . 92

B Biclustering et modèle des blocs latents 97
B.1 Biclustering, co-clustering et modèle des blocs latents . . . 97
B.1.1 Modèle des blocs latents . . . 99
B.1.2 Graphe biparti . . . 101
B.2 Estimation . . . 102
B.2.1 Énergie libre et algorithme *Variational Expectation Maximisation* . . . 103
B.2.2 Algorithme *Stochastic Expectation Maximisation* . . . 104
B.2.3 Algorithme *V-Bayes* et échantillonneur de *Gibbs* . . . 104
B.2.4 Sélection de modèle . . . 105

Remerciements

"Every little thing you said and did was right for me
I had a lot of time to think about
About the way I used to be"
Mama des *Spice Girls*

Si, pour moi, la thèse était déjà un travail d'équipe, ce manuscrit d'*habilitation à diriger des recherches* est encore plus un travail commun, un concours de circonstances ou encore une chance d'évoluer dans un environnement favorable. Bref, de mon point de vue, ce manuscrit est l'aboutissement de toutes mes rencontres. Ces remerciements vont encore être longs (désolé) et je vais essayer de n'oublier personne [1] donc, je vais essayer de les faire par ordre chronologique depuis l'obtention de ma thèse [2] avec quelques moments que je juge clef. Je vais essayer de limiter mes commentaires sur mes remerciements mais sachez que j'ai plein de raisons de vous remercier individuellement.

01/10/2014 : Merci à celles et ceux qui m'ont accueilli pour mon post-doc à AgroParisTech. Merci évidemment à Céline pour le sujet, pour l'accompagnement et pour tous ses encouragements qui m'ont permis de m'émanciper en tant que chercheur. Merci à mes co-auteur·trice·s Émilie, Julien, Laure, Maud, Sarah et Tristan pour nos nombreux échanges très enrichissants. Merci à mes co-bureaux Mounia et Paul qui ont supporté ma dactylographie bruyante [3] mais aussi Liliane, voisine de bureau qui n'a dit que tardivement qu'elle m'entendait chanter quand je pensais être seul. Merci aux doctorant·e·s Anna, Loïc, Marie entre autre pour les nombreux goûters et l'organisation d'événements festifs. Merci à Marie-Pierre et Sophie pour nos discussions les midis notamment sur les concours. Merci à Stéphane pour ses conseils précieux sur les articles et sa culture incroyable aux quiz. Et plus généralement, merci à toutes celles et tous ceux que j'ai croisé·e·s à AgroParisTech et que je ne peux pas citer en entier vue ma mémoire défaillante mais notamment Christophe, Gabriel, Irène, Isabelle, Jessica, Pierre...

Merci aussi aux membres du groupe jeunes de la SFdS avec qui nous avons pu proposer plein de nouvelles idées notamment Angelina, Baptiste, Benjamin, Charlotte, Cyrielle, Émilie, Erwan, Gaëlle, Laure, Mélanie, Rémi, Sandrine, Thomas, Valérie... et plus généralement aux membres de la SFdS avec qui j'ai pu échanger comme Adeline, Angelo, Anne, Catherine, Christian, Christine, Christophe, Gérard, Gwladys, Jean-Jacques, Jean-Michel, Margaux, Marie, Merlin, Perrine, Robin, Sophie, Servane, Vincent...

Merci aussi à celles et ceux que j'ai croisé·e·s durant les candidatures que ce soit mes compagnon·ne·s de galère ou celles et ceux qui m'ont donné des conseils précieux. Merci en particulier à Angelina, Caroline, Clément, Cyrielle, Émilie, Justine, Line, Lucie, Mélina, Mélisande, Pierre et un merci particulier à Adeline et Céline qui m'ont poussé à me surpasser.

01/09/2016 : Évidemment, un grand merci à toutes les personnes qui m'ont accueilli quand j'ai eu mon poste à Grenoble. Les collègues du LJK présent·e·s dès le début comme (et là, désolé mais je suis sûr d'en oublier) Adeline, Agnès, Anna, Arthur, Aude, Aurore, Bernard, Brigitte, Caroline, Cathy, Céline, Clémentine, Delphine, Elise, Eric, Florence, Franck, Frédéric, Frédérique, Gaëlle, Irène, Jean-Baptiste, Jean-Charles, Jean-François, Jean-Guillaume, Jérôme, Juana, Julyan, Karim, Laurence, Laurent, Marie-Jo, Modibo, Olivier, Pierre, Philippe, Rémy, Sana, Stéphane, Suzanne... ou celles et ceux qui nous ont rejoints plus tard comme Alexandre, Alice, Angélique, Aristide, Aude, Aurore, Chloé, Clovis, Charles, Christophe, Clément, Elissa, Emmanuel, Flora, Jonathan, Julien, Kevin, Lola, Luce, Magali, Manon, Margaux, Margot, Maya, Olivier, Selim, Simon, Sophie... Merci aussi aux collègues des autres instituts comme Catriona, Didier, Loren, Raphaël, Vincent...

Merci aussi aux collègues de l'IUT2 qui m'ont laissé très libre dans mes folies de cours. Merci d'abord à mes collègues de STID et SD de Grenoble Agnès, Caroline, Chantal, Delphine, Edwige, Frédérique,

1. Un formulaire est mis en page 6 si je vous ai oublié·e.
2. Il faut voir ces remerciements comme le prolongement de ceux de ma thèse. N'hésitez pas à les consulter car ils restent aussi vrais pour mon HDR.
3. Voir Bastide (2017)

Gillian, Karine, Manon, Marlène, Morgan, Myriam, Nadine, Olivier, Philippe, Sabrina, Sophie, Théa et plus généralement de SD France pour nos longues discussions sur le PN notamment Antoine, Audrey, Cédric, Chloé (en plus des Rencontre R 2024), FX, Marie... Merci aussi aux autre membre de l'IUT2 notamment à celles et ceux avec qui j'ai travaillé pour la LP Big Data Agnès, Christine, Francis, Franck, Françoise, Jérôme, Philippe, Sophie...

Merci aux ami·e·s qui m'ont accueilli en dehors du boulot et m'ont permis de m'intégrer comme Adeline, Annique, Caroline, Charlotte, Chloé, Emeline, Emilie, Franck, Ievgen, Jean-Charles, Julien, Marc-Antoine, Marion, Mélanie, Rémi, Vincent...

Merci à mes co-auteur·trice·s qui m'ont permis de ne pas me noyer dans les enseignements. En premier, merci à Charlotte qui a joué un rôle énorme. Merci à Alexandra, Amélie et Céline qui m'ont permis de terminer mes travaux liés à AgroParisTech. Dans le même style, merci à Antoine, Christine et Mahendra pour la persévérance dans l'écriture des articles un peu compliqués. Merci à toutes les personnes du projet MSE^3 comme Caroline, Etienne, Jérôme et Saifeddine, ainsi que tous les stagiaires liés à ce projet Clément, Louis, Ornella, Sacha, Sylvain et Walid ; Raphaël qui a fait une thèse avec le CEA et bien sûr Yunjiao qui a fait un post-doc avec moi sur ce projet. Il y a eu aussi les collègues de CoPoLangues Anne-Cécile, Marie-Pierre, Sylvain et les stagiaires que nous avons encadré·e·s Jules et Margaux.

C'est l'occasion de faire un merci particulier à mes filleul·le·s qui ont illuminé ma vie Gabrielle, Lubin, Mathieu et Nathanaël et leurs familles que je cite plus tard.

17/03/2020 : Merci à toutes celles et tous ceux qui m'ont permis de passer le confinement. Parmi les amis qui n'ont pas encore été citées, il y a Antoine, Arnaud, Juliette, Philomène, Randa, Victor et Yannick qui m'ont permis de passer de bonnes soirées sur Among Us et BGA. Il y a mes co-auteurs des poolages du Covid Bastien et Jean-François et ceux des articles qu'on tentait de finir comme Valérie et Yann.

Il y a eu l'après Covid aussi avec le projet IAB@R dont je remercie Alain, Christelle et Philomène ainsi que les stagiaires encadré·e·s Alice, Angélique, Florent et Shuyu. L'encadrement de stagiaire en période difficile, merci à Emeline et Julien.

C'est l'occasion de remercier mes ami·e·s de l'impro qui m'ont permis de me vider la tête et de m'aider dans les moments compliqués. Je vais commencer par les membres des Martin·e·s (et leurs proches) Anne, Baptiste, Carla, David, Laetitia, Mathieu (Eva et Inès), Matis (et Ony), Mégane (et Flo), Mélanie (Julien, Léon et Lucien), Mérédith (et Jason), Naïko, Préscilia (Alex et Nathan) et Vivek (et Chloé). Je vais tenter d'enchaîner avec celles et ceux avec qui je joue ou j'ai joué (et là encore, désolé pour les oublis) donc merci à Alain, Allison, Alice, Amandine, Anca, Anne-Cé, Annia, Antoine, Audrey, Aurélie, Aurore, Benjamin, Bertrand, Carine, Cécile, Céline, Clément, Coline, Corinne, David, Elise, Emilie, Emmanuelle, Etienne, Fabienne, Fanny, Faustine, Jean, Jérémy, Julie, Julien, Justin, Laura, Laurent, Leïla, Lolo, Lorène, Luc, Ludo, Maë, Maï-thé, Manue, Marion, Marjorie, Milène, Nathalie, Nicolas, Pascaline, Patrice, Pauline, Pierre, Polo, Rémi, Robin, Sarah, Séverine, Stéphane, Stéphanie, Sylvain, Thibauld, Thomas, Viviane, Will, Zoé...

09/03/2022 : Il est compliqué de faire des remerciements sur cette période sans évoquer la maladie et le décès de ma mère ainsi que les problèmes liés et sans remercier toutes les personnes qui m'ont soutenues (et vraiment, j'ai eu beaucoup de chances car vous avez été nombreux·ses). En plus des amis déjà cités, merci à mes amis d'enfance et d'études (et leurs enfants) Alex, Anouk, Augustin, Auxane, Basile, Bastien, Céline, Charlène, Charlotte, Corentin, David, Dorian, Elsa, Emilie, Eric, Eva, Faustine, Gabrielle, Ievgen, Jérémy, Hugues, Julia, Kenza, Laure, Léonie, Lisa, Louise, Lubin, Margriet, Mathieu, Maxime, Mélanie, Mélina, Nina, Nathanaël, Paul, Pierre, Raphaël, Rémi, Ruben, Sander, Sébastien, Simon, Solène, Solenne, Sophia, Tom, Valentin, Yasmine...

Ma famille a été très importante pour moi que ce soit à ce moment, avant et après. Merci à Adrien, Amélie, Anne, Anne-Marie, Antoine, Baptiste, Bruno, Capucine, Caroline, Denis, Dominique, Elisa, Elisabeth, Emmanuel, Eve, François, Gabriel, Gaspard, Geneviève, Guy, Isaac, Jean-Yves, Jeanne, Johan, Julie, Julia, Julius, Laetitia, Leia, Louis, Lucie, Marie, Marie-Angèle, Maurice, Maxime, Nicolas, Noé, Odile, Parrain (alias Joseph ; tu n'auras pas pu voir cette soutenance malheureusement, je n'ai pas été assez rapide mais je pense à toi), Paul, Pauline, Pierre, Pierrette, Rose, Samuel, Stéphane, Yannick (et Papatte).

Merci aux membres de SPHERE et de CIC qui m'ont accueilli durant cette période compliquée tout en me permettant de faire de la recherche intéressante. En particulier, merci à Agnès, Amélie, Anne, Aude, Bastien, Bénédicte, Bruno, Céline, Elie, Elody, Elsa, Erwan, Etienne, Floriane, François, Fred, Gaëlle,

Garance, Giulio, Jean-Benoit, Jules, Laurène, Marianne, Marie, Myriam, Pierre, Raphaële, Rouba, Solène, Sophie, Valentin, Véronique...

10/07/2025 Merci aux membres de mon jury qui ont accepté de relire ce manuscrit, d'assister à la soutenance et pour les retours intéressants que j'ai déjà eus. Merci à Adeline, Arthur, Céline, Franck, Sophie et Stéphane. J'espère que ce manuscrit et cette soutenance feront honneur au temps que vous m'avez consacré.

Je remercie également les personnes qui continuent de travailler avec moi. Merci aux membres du projet Forenscript Caroline, Fanny, Jérémy et Zoé ; aux membres du projet SAS Adeline, Elodie, Kajal et Sébastien ; le projet Gratweet Julien et Martin ; le projet Taunique avec Amélie, Emilie et Virginie et le projet Capushe avec Perrine.

Merci à mes ami·e·s qui ne sont pas entrés dans les cases précédentes comme Alana, Charles, Gaël, Gilles, Irène, Jean-François, Johanne, Lidya, Manon, Margaux, Marie, Merlin, Rodolphe...

Et des mercis plus généraux pour tous les élèves que j'ai eu·e·s (certain·e·s diraient traumatisé·e·s), les membre successifs du groupe jeunes qui font vivre le futur de la statistique (et plus généralement les personnes qui font vivre la SFdS) et les collègues de l'université que j'ai plaisir à côtoyer.

Trois mercis particuliers : Je vais conclure ce chapitre par trois mercis particuliers. Le premier est pour mon papounet ; les dernières années furent compliquées, nous avons réussi à les affronter ensemble, j'ai bien conscience de ne pas avoir toujours été le plus agréable surtout dans les moments de creux mais tu es encore là, merci à toi. Le deuxième est pour Valérie qui me supporte, m'encourage et m'aide malgré la distance depuis toutes ces années ; merci beaucoup. Et le dernier merci est pour celle qui ne pourra pas lire ce merci. Merci Maman pour toute la patience que tu as eue envers moi, tout le temps et l'amour que tu m'as consacrés et que je n'ai réalisé malheureusement qu'une fois que tu étais partie.

Et merci à toi, lectrice ou lecteur, de prendre le temps de participer à la recherche, j'espère que ce manuscrit répondra à certaines de tes questions.

Comme pour ma thèse, une de mes plus grandes craintes est d'avoir oublié quelqu'un. Si par hasard, vous êtes dans cette situation, je vous prie d'abord de m'excuser puis de m'envoyer le formulaire [4] suivant :

Je, sous signé, (nom, prénom)..... signale que dans votre manuscrit d'HDR [a] :

☐ mon prénom n'apparaît pas.

☐ mon prénom est mal orthographié.

☐ mon prénom est placé dans un groupe auquel je n'appartiens pas.

☐ autre : .....

Compléments éventuels à la demande : .....
.....
Je vous prie de corriger cette erreur assez rapidement.

a. Ne cocher qu'une seule case

Conformément au code de rédaction d'une habilitation à diriger les recherches régissant les remerciements, je ferai mon maximum pour effectuer les corrections dans un délai relativement court.

4. Existant aussi au format numérique à l'adresse suivante :
https://docs.google.com/forms/d/1liJX__BhnBqtDbNzf_Ah4lOr-3did9MCmOFjREo6c_4/viewform

Chapitre 1

Introduction

"La vie étant un éternel recommencement, seule l'acceptation de la défaite signifie la fin de tout. Tant et aussi longtemps que l'on sait recommencer, rien n'est totalement perdu."

Fleurette Levesque

Mon manuscrit d'habilitation à diriger des recherches se sépare en cinq chapitres et deux annexes. Je commence par vous proposer un état des lieux de mes travaux au moment où j'écris ces lignes et je détaille les impulsions qui ont abouti à ces recherches. Les trois chapitres suivants détaillent le contenu de mes travaux. Pour les chapitre 2 et 4, j'ai regroupé les articles par thématiques tandis que pour le chapitre 3, j'ai plutôt choisi de présenter comment chaque article répondait aux limites du précédent. Le dernier chapitre présente mes objectifs pour les années à venir avec en particulier, si j'obtiens l'habilitation, mes idées d'encadrement de thèses. Je présuppose que les lecteurs et les lectrices connaissent les modèles de mélange et les modèles de segmentation ; si ce n'est pas le cas, je conseille de lire l'annexe A qui redonne des bases. Enfin, j'aborde uniquement les travaux que j'ai initiés depuis ma thèse ; les personnes qui ne connaîtraient pas ce que j'ai fait en thèse (voir Brault (2014)) pourront lire l'annexe B pour avoir un résumé des informations importantes pour une meilleure compréhension du chapitre 2 notamment.

1.1 Panorama

Sur la figure 1.1, j'ai représenté mes articles publiés (en traits pleins), en cours de soumission (traits pointillés longs) et en fin de rédaction (traits pointillés courts) reliés aux co-autrices et aux co-auteurs regroupés par zone suivant l'une des six grandes thématiques de mes travaux de recherche. Dans la suite de cette section, je vais détailler chaque partie en expliquant comment est née la collaboration et ma participation au sein de l'article. Les détails scientifiques seront détaillés dans les chapitres suivants du manuscrit.

1.1.1 Modèle des blocs latents

Le premier sujet d'étude de ma vie de chercheur fut le **modèle des blocs latents** ou **LBM** (en rouge sur la figure 1.1) que j'étudiai en thèse dont trois articles sont sortis : Keribin et al. (2014); Brault et Lomet (2015); Brault et Mariadassou (2015). Comme l'habilitation à diriger des recherches portent principalement sur les travaux depuis mon doctorat, je ne les développerai pas ici (même si je les mentionne parfois dans le chapitre 2).

Le premier article dont j'ai été à l'initiative fut Brault et Channarond (2023) commencé durant ma première année de thèse en parallèle de mes recherches. après des débats avec mon directeur de thèse, les premières ébauches de cet article furent quand même présentées dans mon chapitre 5 de thèse (voir Brault (2014)) même si elles étaient encore jeunes et manquaient de maturité. L'idée de départ est venue après une discussion avec Stéphane Robin (AgroParisTech/INRAE à l'époque) qui encadrait Antoine Channarond (AgroParisTech à l'époque) sur le **modèle des blocs stochastiques** ou **SBM**. Ce dernier montre dans son article Channarond et al. (2012) qu'il est possible d'avoir des estimateurs consistants en ne regardant que les marginales de la matrice d'adjacence. J'ai adapté ces résultats au cas du modèle des blocs latents en tenant compte du problème de double asymptotique (voir la section 2.3.2) de notre modèle et proposé une procédure de sélection de modèles différente de leur article. À ce jour et à ma connaissance, c'est le seul critère de sélection de modèles pour le modèle des blocs latents dont il a été montré en théorie qu'il était consistant. Une fois des résultats intéressants mis en évidence, j'ai contacté

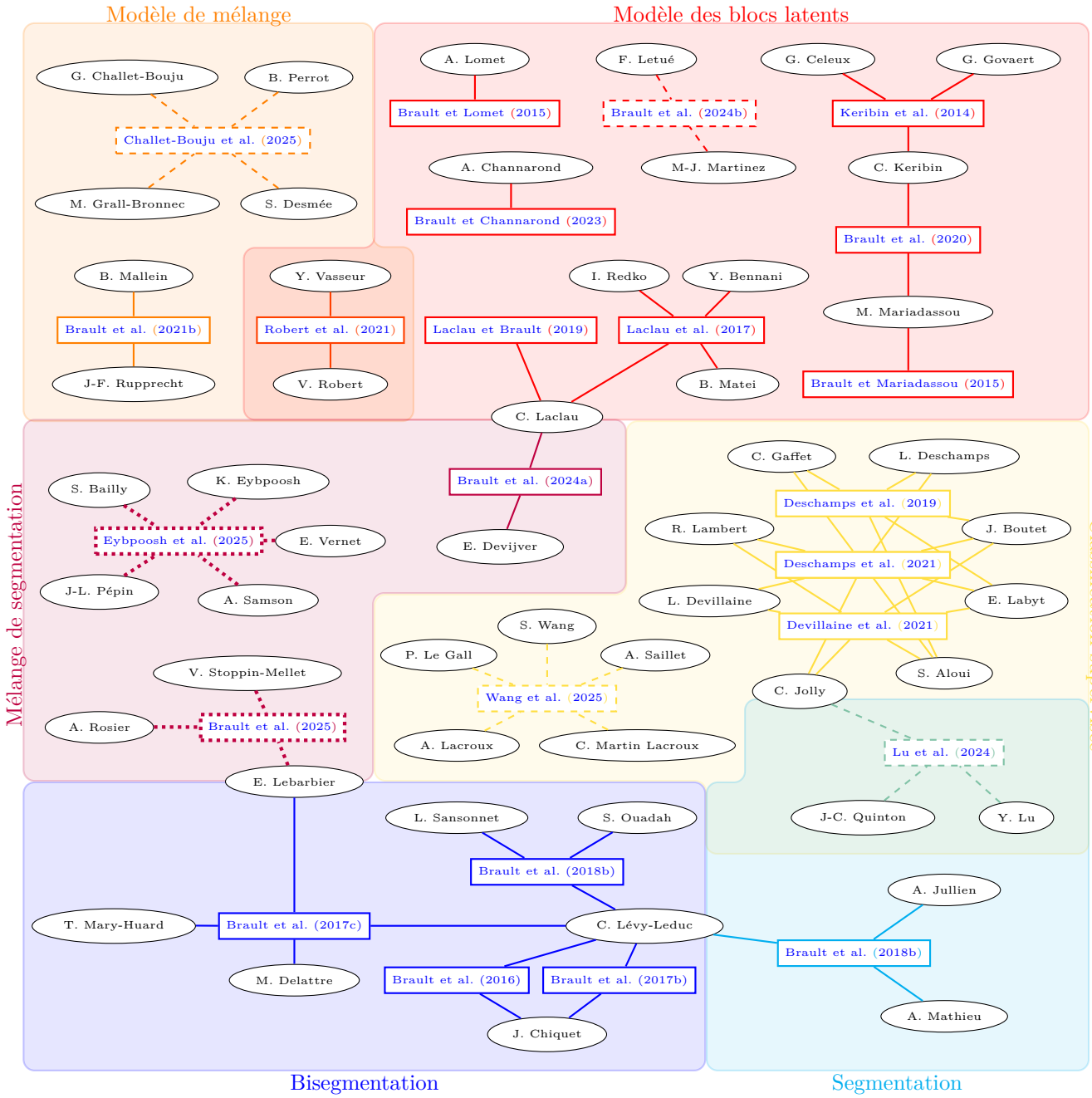

FIGURE 1.1 – Représentation des articles publiés (traits pleins), en cours de soumission (traits pointillés longs) et en fin de rédaction (traits pointillés courts) regroupés par thématiques (couleurs) : les rectangles correspondent aux articles reliés par des traits aux ovales représentant leurs co-auteur·trice·s.

Antoine durant ma première année de post-doc pour que nous écrivions l'article ensemble.

Une fois les premiers résultats de l'article Brault et Channarond (2023) obtenus durant ma thèse et profitant de ma collaboration avec Mahendra Mariadassou (INRAE de Jouy en Jossas) sur l'article Brault et Mariadassou (2015), j'ai eu envie de commencer une nouvelle collaboration avec lui sur la consistance : le principe était que je montrais dans le cas de mon article que des estimateurs basés sur les marginales étaient consistants et lui montrait dans son article Mariadassou et Matias (2015) écrit avec Catherine Matias que les lois des distributions des variables latentes basées sur le maximum a posteriori d'un paramètre *pas trop loin des vrais paramètres* tendaient vers une dirac sur la partition ayant servi à simuler les données. Pour moi, en couplant les deux, nous devions réussir à obtenir une démonstration pour la consistance des estimateurs du maximum de vraisemblance. Il fut intéressé de travailler sur cette question notamment car il avait lu les résultats, légèrement erronés, de Bickel et al. (2013) qui avaient montré la consistance pour les estimateurs dans le cadre du SBM. Comme Christine Keribin (Université Paris Saclay) était aussi intéressée pour travailler sur cette thématique après avoir montré l'identifiabilité du LBM, nous lui avons proposé de faire partie du projet. Le développement de l'article Brault et al. (2020) fut compliqué principalement à cause de la lourdeur de la démonstration : d'abord, il était compliqué de bloquer assez de temps pour nous trois afin de se mettre pleinement dans toutes les démonstrations et ensuite, une fois l'article soumis, il fut compliqué d'expliquer que cette lourdeur était importante pour adapter notre démonstration au cas du LBM tout en corrigeant l'erreur faite dans l'article de Bickel et al. (2013). Cet article est développé dans la section 2.3.1.

Une fois arrivé sur Grenoble, j'ai eu la chance de pouvoir travailler régulièrement avec Charlotte Laclau alors qu'elle était en post-doc au Laboratoire d'Informatique de Grenoble (LIG). Thésarde de Mohammed Nadif, elle avait travaillé sur les **modèles blocs diagonaux** (voir par exemple Laclau et Nadif (2015)) et souhaitait continuer en introduisant des classes *de bruits* pour séparer les individus dont le comportement serait atypique. Dans ce cadre, nous avons proposé le **modèle des blocs latents avec classe de bruit** ou **NFLBM** où nous avons adapté tous les résultats existants à ce moment là (identifiabilité, algorithme d'estimation et critère de sélection de modèle) pour le LBM dans l'article Laclau et Brault (2019) (voir la section 2.2.1).

Comme Charlotte collaborait également avec Ievgen Redko (INSA de Lyon à l'époque) qui fait du transport optimal, elle s'est aperçue qu'un pont était possible entre nos deux problématiques. Après en avoir discuté avec chacun de nous, nous avons décidé d'étudier cette intuition ce qui a abouti à l'article Laclau et al. (2017) qui montrent théoriquement, grâce aux résultats sur les modèles des blocs latents, pourquoi les algorithmes du transport optimal fonctionnent et, inversement, permet aux personnes intéressées d'utiliser les algorithmes utilisés dans le cadre du transport optimal pour estimer les partitions des modèles des blocs latents.

Enfin, le dernier article portant sur le modèle des blocs latents est issu du projet CoPoLangues en partenariat avec des collègues de langues de l'université Grenoble Alpes qui cherchaient à analyse leurs procédures de formations de groupes de niveaux à partir du test qu'ils avaient créé. Frédérique Letué (LJK) et Marie-Jo Martinez (LJK) furent contactées par les collègues des langues et comprirent que la problématique était de faire des groupes d'élèves et des groupes de questions afin d'obtenir des blocs homogènes. Elles firent le rapprochement avec mes travaux et nous entamâmes une collaboration pour proposer un outil que nous avons mis à leur disposition (voir par exemple Brault et al. (2021a)). Très vite, la mise à disposition de cet outil fit ressortir une question soulevée lors de ma soutenance de thèse par Pascal Massart sur la problématique de la robustesse des estimations : que se passe-t-il si nous enlevons quelques étudiants par exemple ? Est-ce que les groupes changent totalement ou pas ? Pour répondre à ces questions, nous proposons l'article Brault et al. (2024b) actuellement en cours de soumission.

1.1.2 Modèle de mélange

En étudiant le modèle des blocs latents, j'ai amélioré mes connaissances sur les **modèles de mélange** (en orange sur la figure 1.1). Cette thématique regroupe principalement deux articles plus un article en commun avec la thématique du modèle des bloc latent (en orange foncé sur la figure).

J'aime beaucoup l'histoire de l'article Brault et al. (2021b) qui représente pour moi un exemple de ce que pourraient faire des chercheuses et chercheurs quand on leur donne des moyens concrets. Dans

le cadre de la pandémie de Covid 19, la communauté scientifique a été appelée à trouver des solutions pour aider à sortir le plus rapidement possible de la crise. Dans ce cadre, une association de scientifiques, nommée Groupool [1], a cherché à analyser l'intérêt des tests groupés, plutôt qu'individuels, pour accélérer le dépistage et, en particulier, Bastien Mallein (Univerisité Toulouse III - Paul Sabatier) et Jean-François Rupprecht (CNRS au Centre Physique Théorique de l'Université Aix-Marseille) cherchaient à répondre aux interrogations posées par le Haut Conseil de la Santé Publique (HCSP). Dans ce cadre, ils ont proposé des modélisations de l'influence des tests groupés (où on mélange plusieurs salives pour dire si au moins une personne du groupe est malade) et ont eu besoin de nouvelles modélisations pour représenter la charge virale. En tant qu'ancien camarade de leur promo, ils ont pensé à moi et j'ai pu proposer un **modèle de mélange de gaussiennes censurées** qui a permis de montrer l'intérêt de l'utilisation de tests groupés (voir la section 2.1). Cette expérience fut, pour moi, très enrichissante car elle m'a permis de comprendre d'une part la difficulté de discuter avec les décideurs politiques (puisque nous avons eu quelques réunions avec des représentants du ministère de la santé notamment) et la nécessité d'être présent pour limiter le manque d'informations voir les contre-vérités.

Le deuxième article Challet-Bouju et al. (2025), que nous venons de soumettre, est issu de ma collaboration avec l'équipe SPHERE [2] de l'INSERM où j'ai passé un an lors de ma délégation CNRS. Durant cette année, et après, j'ai travaillé sur les questions d'addiction. Dans ce cadre, j'ai proposé une nouvelle modélisation, un peu décalée par rapport aux analyses *classiques* faites à l'INSERM, dans le but de répondre à des problématiques que mes collègues n'avaient pas pu aborder jusque là. Cette collaboration me fut enrichissante sur les pratiques en cours dans les recherches en médecine comme le fait de ne pas trop détailler les modélisations dans les articles ce qui me frustre à plusieurs moments (reproductibilité et diffusion par exemple) et m'encourage à soumettre d'autres articles en parallèle. Par exemple, dans l'article Challet-Bouju et al. (2025), j'ai proposé une modélisation et une estimation des groupes d'individus basés sur des *critères d'addiction* que j'ai faite évoluée au fur et à mesure des retours de mes collègues basées sur la compréhension biologique de l'addiction et sur la logique des observations faites. Ainsi, par exemple, je n'ai pas pu démontrer pour chaque modélisation l'identifiabilité du modèle et le comportement théorique des estimateurs en particulier pour la modélisation retenue dans la version fianle. Bien que j'ai l'impression que tout devrait fonctionner car les démonstrations que j'ai en tête sont proches de ce que j'ai déjà proposé dans d'autres articles, j'aurais bien aimé m'en assurer dans un article méthodologique proposé avant.

Enfin, le dernier article de cette section (Robert et al. (2021)), partagé avec la section du LBM, est né d'une discussion avec Valérie Robert (Université de La Réunion). Dans le cadre de sa thèse, elle travaillait sur l'adaptation du critère ARI (*Adjusted Rank Index*) pour le modèle des bloc latents et l'avait comparé aux adaptations de critères d'entropie. De mon côté, j'avais obtenu durant la rédaction de ma thèse des premiers résultats sur le comportement de l'erreur de classification étudiée notamment dans les procédures de Lomet et al. (2012). Nous avons alors décidé de mettre en commun nos résultats, de les perfectionner et de proposer une comparaison des différents critères suivant des objectifs que nous jugions importants pour comparer des estimations de partitions (voir la section 2.2.4). Notamment, cet article fut l'occasion pour moi de mettre à profit mes discussions avec Franck Iutzeler (LJK à l'époque) pour proposer une estimation de l'erreur de classification en un temps raisonnable et répondre ainsi à une question posée par Stéphane Robin lors d'un séminaire 7 ans plus tôt.

1.1.3 Bisegmentation

Les articles sur la **bisegmentation** (en bleu sur la figure 1.1) partent de l'analyse des données Hi-C et représentent ma vision la plus *classique* de mes recherches : partir d'un problème concret, proposer une modélisation, étudier les propriétés théoriques, optimiser les méthodes d'estimation, appliquer sur des données simulées et sur les données réelles, comprendre les limites et recommencer.

Le premier article était le point de départ de mon post-doc et consistait à reprendre la modélisation proposée par Lévy-Leduc et al. (2014) et de démontrer le comportement atypique de la vraisemblance. Les trois premiers mois de mon post-doc ont consisté à m'approprier le **modèle de bisegmentation par blocs diagonaux** et le principe de la segmentation en général et à proposer une démonstration quasiment définitive de l'article Brault et al. (2017c) (voir la section 3.1).

1. Voir la page https://www.groupool-covid19.org/
2. Page web : https://sphere-inserm.fr/fr

Une fois la démonstration précédente obtenue, nous avons cherché avec Céline à généraliser le modèle pour inclure des blocs en dehors de la diagonale. La première idée fut de proposer une segmentation simultanée des lignes et des colonnes mais nous perdions l'estimation par l'algorithme de programmation dynamique. Pour contourner ce problème, nous décidions d'utiliser une astuce basée sur la procédure *LASSO* (*Least Absolute Shrinkage and Selection Operator*). Comme au sein de l'institut AgroParisTech, il y avait Julien Chiquet qui était aussi un spécialiste du *LASSO*, nous avons collaboré avec lui. Cet article est un très bon exemple de la volonté d'optimiser le code car nous avons d'abord amélioré l'estimation à l'aide d'astuces mathématiques puis dans un second temps, Julien l'a codé en C (voir la section 3.2). Malheureusement, nous n'avons plus le temps de maintenir le package `BlockSeg` qui demande une mise à jour vis à vis des nouvelles règles incluant du code C et seule une ancienne version est disponible sur le site du CRAN (Brault et Chiquet (2018)).

La problématique de la précédente version était qu'il fallait supposer que les observations étaient la somme d'une moyenne et d'un bruit sous gaussien et il fallait donc normaliser les données pour permettre une application. Pour contourner ce problème, nous avons discuté avec Sarah Ouadah (AgroParisTech) qui étudiait les modélisations non paramétriques et avons proposé, avec Laure Sansonnet (AgroParisTech), une modélisation ne faisant aucune hypothèse sur les lois des blocs en dehors du fait que deux lois successives soient différentes (voir la section 3.3). Cette collaboration a permis de proposer l'article Brault et al. (2018c) et le package Brault et al. (2018a).

1.1.4 Classification supervisée

Les articles de la partie **classification supervisée** (en jaune sur la figure 1.1) représentent ma chance de travailler dans une université comme l'UGA avec une multitude de laboratoires et thèmes de recherche.

Le premier projet entamé pour lequel j'ai obtenu deux financements et dont je suis co-porteur d'un troisième est celui sur la détection de la **dysgraphie** chez les enfants en utilisant des tablettes pour évaluer la dynamique du tracé. Ce projet est en collaboration notamment avec Caroline Jolly du Laboratoire de Psychologie et NeuroCognition (LPNC) de l'UGA et Etienne Labyt puis Jérôme Boutet du CEA de Grenoble. Trois articles provenant des travaux des stagiaires pour lesquels j'ai participé au co-encadrement sont sortis. Le premier, Deschamps et al. (2019), est un *comment paper* fait sur l'article Asselborn et al. (2018) qui portait sur une base de données réalisée par Caroline Jolly avant notre collaboration et qui souffrait d'un certain nombre d'erreurs méthodologiques. Par exemple, les enfants diagnostiqués dysgraphiques provenant du CHU (Centre Hospitalo-Universitaire) de Grenoble avaient écrit sur des tablettes différentes de celles utilisées pour les enfants non diagnostiqués et n'avaient pas les mêmes calibrations de la pression. Bien que nous en ayons discuté avec les auteurs lors d'une rencontre à l'EPFL (École Polytechnique Fédérale de Lausanne), ils ont oublié cette différence et ont fait une grande partie de leur article sur la détection de la dysgraphie à partir de la pression. Dans notre *comment paper*, nous n'avons pas pu aborder *frontalement* cette erreur puisque les données n'étaient pas décrites précisément mais nous commentons la méthodologie utilisée.

Afin de remédier à ce problème dans la base, nous avons créé la nôtre avec 554 enfants et nous avons fait diagnostiquer chaque enfant par des experts. Dans l'article Deschamps et al. (2021), nous proposons une étude de différentes *features* issues à la fois de la dynamique des tracés et de paramètres sur l'écriture finale issus des critères du test *BHK* (*Brave Handwriting Kinder*) qui sert actuellement aux praticiens pour détecter la dysgraphie. La comparaison de douze méthodes de classification supervisée (comme les forêts aléatoires ou les *Support vector machines* suivant différentes configurations par exemple) permet de faire ressortir plusieurs critères pour aider à la détection ; certains étaient déjà utilisés et issus du *BHK* (comme le fait que les retours à la ligne aient des marges fixes ou variables) et d'autres sont ressortis de la dynamique (comme le temps passé avec une écriture lente).

L'inconvénient de baser la détection sur le test du *BHK* (comme fait actuellement) est que le test doit être adapté suivant les langues (en France, par exemple, il a été adapté par Charles et al. (2004)) et les enfants doivent déjà savoir écrire ; ce dernier point est contraignant car la détection de la dysgraphie sert aussi pour détecter d'autres pathologies sous-jacentes. Pour contourner ce problème, nous étudions dans l'article Devillaine et al. (2021) la détection de la dysgraphie à partir de la reproduction de formes (carrés, croix, rond, boucles...) et le fait de suivre des chemins. L'article montre une précision de 73% de bonnes classifications (contre 81% pour l'article Deschamps et al. (2021)) *juste* à l'aide de ces figures.

Enfin, le dernier sujet abordé dans la classification supervisée exclusivement est celui de l'influence de la base d'apprentissage dans la prédiction de divers algorithmes de prédictions dans le cadre du recrutement. La collaboration vient à l'origine de Christelle Martin Lacroux (CERAG) et d'Alain Lacroux (Université Paris 1 Panthéon-Sorbonne), professeurs dans la science du recrutement, qui s'intéressent à l'utilisation de l'IA (Intelligence Artificielle) dans le processus de recrutement. Leur article Lacroux et Martin-Lacroux (2022) issu de ce projet montre que les *recruteurs* vont avoir tendance à suivre les recommandations d'un algorithme (dans l'article, il était fictif) *même si cela va à l'encontre de leur intuition.* Dans ce cadre, nous proposons dans l'article Wang et al. (2025), en cours de soumission, une étude sur des algorithmes classiques de classification supervisée (modèle linéaire généralisé avec ou sans sélection de variables, *Support Vector Machine*, réseaux de neurones...) utilisés pour prédire le meilleur dossier à partir de CV simulés. Nous montrons que, si la base d'origine est biaisée, les algorithmes vont perpétuer ce biais et nous montrons aussi que l'anonymisation des dossiers fonctionne uniquement si les variables restantes sont faiblement corrélées avec les variables discriminatoires.

1.1.5 Segmentation

A la suite de mon post-doc, j'ai continué de travailler sur les **modèles de segmentation** (en cyan sur la figure 1.1) notamment avec une application sur la croissance des plants de Colza dans l'article Brault et al. (2018b). Ce travail fut le dernier en lien direct avec mon post-doc où j'ai pu améliorer mes connaissances sur l'utilisation de la méthode *LASSO* pour estimer les emplacements de ruptures notamment lorsqu'il y a une dépendance avant et après la rupture et aussi une dépendance dans le bruit (voir la section 3.5).

Le deuxième article de cette thématique est issu de mes recherches sur la dysgraphie et est donc partagé avec la rubrique classification supervisée (en vert clair sur la figure 1.1). En discutant avec Jean-Charles Quinton (LJK) de mes travaux sur la dysgraphie, il m'expliqua qu'il avait travaillé sur une modélisation de l'écriture à partir d'oscillateurs orthogonaux (le **modèle POMH** pour *Parsimonious Oscillatory Model of Handwriting*). Nous avons eu alors l'idée de voir comment se comporte cette modélisation dans le cadre d'enfants dysgraphiques c'est-à-dire d'enfant où on *pourrait penser* que leur écriture n'est pas fluide. Ainsi, j'ai obtenu un financement de post-doc auprès de l'institut Carnot pour financer Yunjiao Lu afin qu'elle puisse travailler sur cette question. Comme expliqué dans la section 4.3.3, le modèle *POMH* se base sur les moments d'annulation de la vitesse qui sont des moments de rupture ; toutefois, ces instants peuvent être compliqués à détecter à cause de l'estimation empirique de la vitesse faite à partir des positions du stylet. Dans l'article Lu et al. (2024) en soumission, l'angle choisi par Yunjiao a été de chercher à lisser les courbes à l'aide d'un filtre passe-bas dont nous cherchons à calibrer l'opérateur ; nous avons donc cherché comment optimiser les fenêtres des filtres pour détecter au mieux ces instants. De plus, nous proposons une méthode de classification de la dysgraphie comparant les reconstructions du tracé par le modèle *POMH* par rapport au tracé initial et nous obtenons une aire sous la courbe *ROC* (*Receiver Operating Characteristic*) de presque 80% pour les meilleures combinaisons de paramètres.

1.1.6 Mélange de segmentation

Enfin, la dernière partie de mes travaux porte sur les **mélanges de segmentation** (la zone violette sur la figure 1.1). Les articles présentés ici sont les plus récents et représentent différentes modélisations pour classifier des individus représentés par des processus avec des changements de comportements.

Le seul article publié à l'heure actuelle est celui sur le **mélange de segmentation** écrit avec Emilie Devijver (LIG) et Charlotte Laclau (Télécom Paris). Nous sommes partis du constat que lorsqu'il faut mélanger segmentation et classification, les auteurs classifient les données ordonnées (par exemple des données temporelles) en *espérant* que les classes soient connexes. C'est le cas par exemple de l'article Casa et al. (2021) où les données représentent des graphes sur une semaine et qu'il faut à la fois classer les nœuds et les semaines. Dans notre article Brault et al. (2024a), nous proposons de conserver l'ordre quand il y en a un et de combiner les méthodes de segmentation avec les méthodes de classification. Pour ce faire, nous appliquons ces méthodes aux données fonctionnelles (par exemple, les courbes électriques durant l'année des confinements) que nous projetons sur des bases de Haar et nous appliquons un modèle de mélange de segmentation sur les coefficients pour voir des périodes (comme les confinements justement). Dans cet article, je me suis principalement occupé du codage (avec Charlotte) et de la démonstration de consistance (avec Emilie). Il reste une coquille sur l'identifiabilité qui a passé la relecture

malheureusement et que je dois corriger. Ce travail est développé dans la section 4.1.

Dans une idée proche, j'ai entamé une collaboration avec Emilie Lebarbier (Université Paris Nanterre), Amélie Rosier (ESME Sudria Paris) et Virginie Stoppin-Mellet (Grenoble Institut Neurosciences) sur l'étude de la protéine Tau et plus précisément sur la classification de courbes issues d'expériences de détection de cette protéine. Dans l'article Brault et al. (2025) que nous sommes en train de finir d'écrire, nous classifions les courbes suivant si elles ont zéro, un ou deux sauts (dans ce dernier cas, nous devons séparer les sauts faits au même moment des sauts mais à des moments différents). Comme expliqué dans la section 4.2, la difficulté de ce travail réside dans le fait que quelque soit l'individu, la hauteur du saut est la même ce qui crée des dépendances entre les classes mais aussi dans la segmentation. En plus de la participation à la modélisation du modèle, mon travail au sein de cet article a porté sur l'optimisation des codes pour permettre une estimation dans un délai raisonnable et l'étude du comportement de l'algorithme à travers différentes simulations.

Enfin, le dernier article présenté dans la figure 1.1 est Eybpoosh et al. (2025) actuellement en cours d'écriture et porte sur la caractérisation des comportements de l'utilisation des machines dans le cadre de l'apnée du sommeil. Dans ce travail, présenté en section 4.4, nous avons des courbes d'utilisation de machine avec une particularité pour les nuits où les utilisateurs ne les utilisent pas. Dans ce cadre, il y a un changement de comportement mais qui peut n'être que pour une nuit. Pour le post-doc de Kajal Eybpoosh, plutôt que de proposer un modèle de segmentation, nous utilisons une **chaîne de Markov** afin de modéliser également les changement entre les nuits où la machine est utilisée et celles où il n'y a aucune utilisation. Dans ce travail, en plus de la co-modélisation, je me suis concentré sur l'implémentation en R des codes et je suis en train d'essayer de montrer la consistance des estimateurs en utilisant la même démonstration que pour les articles Brault et al. (2020) et Brault et al. (2024a) avec Elodie Vernet (LJK).

1.2 Encadrements

Mes projets m'ont permis le (co-)encadrement de deux post-doc, dix-huit stagiaires, une alternante et un ingénieur d'étude. Dans cette partie, je vais détailler par projet l'importance de ces encadrements.

1.2.1 Projets sur la dysgraphie et la science forensique

Dans le cadre de mon partenariat avec Caroline Jolly (LPNC) sur la détection des enfants dysgraphiques, j'ai obtenu un financement [3] qui m'a permis d'encadrer les stage de Sylvain Berthelot (M1 SSD [4] ; 2017), Sacha Peschoux (M1 SSD ; 2018) et Walid Guiza (Fin d'étude DUT STID [5] ; 2018) pour explorer la base initiale proposée par Caroline Jolly et proposer des premières modélisations. Ces explorations ont notamment permis de mettre en évidence les erreurs dans la récupération des données et ont permis d'alimenter l'article Deschamps et al. (2019).

En parallèle de ce financement, avec Caroline Jolly (LPNC) et Etienne Labyt (CEA), nous avons obtenu un financement du CEA [6] pour pouvoir récupérer de nouvelles données avec un plan d'expérience plus rigoureux et financer trois stages pour lesquels j'ai participé à l'encadrement. Les deux premiers sont Clément Gaffet (M2 SSD ; 2018) et Louis Deschamps (fin d'étude école d'ingénieur ISAE-SUPAERO [7] ; 2018) puis Louis Devillaine (fin d'étude école d'ingénieur Grenoble INP, Phelma [8] ; 2019) qui ont fait les premières analyses sur la nouvelle base de données ce qui a abouti aux articles Deschamps et al. (2021) et Devillaine et al. (2021). De mon côté, j'ai encadré Ornella Volle (L3 MIASHS [9] ; 2025) pour continuer d'explorer la base et recenser les erreurs qui se sont glissées.

Enfin, grâce à l'obtention d'un financement Carnot [10], j'ai pu encadrer Yunjiao Lu en post-doc (2022-2023) qui a exploré la modélisation *POMH* pour l'appliquer à la dysgraphie et dont l'article Lu et al. (2024) est en soumission.

3. Porteur principal du projet *MSE*^{3} pour *Modélisation Statistique pour l'Etude de l'Ecriture chez l'Enfant* financé par le LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) lui-même financé par le programme français Investissement d'avenir.
4. SSD pour *Statistique et Science des données*.
5. STID pour *Statistique et Informatique Décisionnelle* qui est une formation délivrant un DUT pour *Diplôme Universitaire Technologique* ; diplôme de deux ans obtenus au sein d'un Institut Universitaire Technologique (ou *IUT*).
6. Partenaire du projet *Ecriture CEA 17P170*.
7. ISAE-SUPAERO pour *Institut Supérieur de l'Aéronautique et de l'Espace*.
8. INP pour *Institut National Polytechnique* et Phelma pour *PHysique, ÉLectronique, MAtériaux*.
9. MIASHS pour *Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales*.
10. Porteur principal du projet Ecriture inter-Carnot.

Dans la continuité du projet sur la dysgraphie, je travaille avec Jérémy Danna (CNRS rattaché au laboratoire Cognition, Langues, Langage, Ergonomie de l'université Jean Jaurès de Toulouse), Fanny Guillet (Laboratoire Police Scientifique de Lyon) et Caroline Jolly (LPNC) pour appliquer les méthodes développées dans le cadre de la détection de la dysgraphie pour la science forensique. Dans le cadre de ce projet, Fanny a obtenu des financements de la part du ministère de l'intérieur pour l'encadrement de Zoé Riebel d'abord en stage (M1 MIASHS de Lyon ; 2024) puis en alternance (M2 MIASHS de Lyon ; 2024-2025) auquel j'ai participé pour la partie classification supervisée.

1.2.2 Consulting

Pour financer mes projets, j'ai aussi participé à des consultings pour des entreprises (Walterre puis Econeaulogis par exemple) dont deux ont abouti à des encadrements. Le premier était avec la start-up *Odit-E* qui a proposé un stage à Mouna Rhalimi (M2 MSIAM et ENSIMAG [11] ; 2018) que j'ai co-encadrée dans le cadre du consulting sur les réseaux électriques.

Le deuxième fut l'encadrement de Jules Diard (2022) en tant qu'ingénieur d'étude au sein de la société TeamCo sur la pondération des experts dans le cadre d'aide à la gestion de ressources humaines. Je n'ai encadré cet ingénieur que pendant un an car la société a choisi des orientations scientifiques qui ne me convenaient pas et j'ai préféré me retirer du projet à partir de ce moment là.

1.2.3 Projet Gratweet

A l'époque où Twitter mettait à disposition ses données, j'ai obtenu avec Julien Chevallier (LJK) un financement [12] en 2018 pour travailler sur le SBM avec processus de Hawkes sur les arrêtes. Le changement de politique de Twitter a mis un coup de frein à ces recherches mais nous sommes en train de relancer le projet en proposant un stage à Martin Combelles (M1 SSD ; 2025) pour qu'il étudie cette modélisation sur les échanges de mails.

1.2.4 Modèle des blocs latents non paramétriques

Dans le continuité du travail de l'article Brault et Channarond (2023), j'ai proposé un stage à Julien Prando (M1 Mathématiques Générales ; 2020) pour analyser la possibilité d'étendre l'algorithme *Largest Gaps* au cas d'un modèle des blocs latents où les lois ne sont pas paramétriques. Pour le moment, le projet n'a pas encore abouti et il faut que je cherche de nouveaux financements pour le continuer.

1.2.5 Projet CoPoLangues

Dans le cadre d'une collaboration avec Sylvain Coulanges (LIG et LIDILEM pour *LInguistique et DIdactique des Langues Étrangères et Maternelles*), Frédérique Létué (LJK), Marie-Pierre Jouannaud (UGA et TransCrit pour Transferts critiques anglophones), Marie-José Martinez (LJK) et Anne-Cécile Perret (CUEF pour Centre Universitaire d'Etudes Françaises), j'ai pu étudier l'application du modèle des blocs latents au test SELF [13] (voir par exemple Brault et al. (2021a)).

À l'aide de l'argent récupéré grâce aux consultings, j'ai pu encadrer le stage de Jules Trividic (M1 SSD ; 2019) sur la robustesse des méthodes d'estimation. Ce travail fut complété par le co-encadrement du stage de Margaux Leroy (M2 SSD ; 2020) grâce au financement obtenu dans le cadre du projet [14] et fut les prémices de l'article Brault et al. (2024b) en cours de soumission.

1.2.6 Poolage et Covid 19

Dans la continuité de l'article Brault et al. (2021b), j'ai encadré le stage d'Emeline Devaud (M1 Mathématiques Générales ; 2021) qui a travaillé sur l'utilisation du poolage pour aider à la détection du Covid 19 en étudiant différentes stratégies. Ce stage n'a pour le moment pas abouti à une publication.

11. MSIAM pour *Master of Science in industrial and applied mathematics* et ENSIMAG pour *École Nationale Supérieure d'Informatique et de Mathématiques Appliquées de Grenoble*.
12. Appels à projet JCJC du labex PERSYVAL-Lab qui a notamment de financer un workshop (avec le soutien de l'ARC6 aussi pour la manifestation)
13. SELF pour *Système d'Évaluation en Langues à visée Formative*.
14. Partenaire du projet *CoPoLangues* ; projet IRS financé par IDEX Université Grenoble Alpes.

1.2.7 Projet sur l'utilisation de l'intelligence artificielle pour le recrutement

Comme expliqué dans la section 1.1.4, j'ai été contacté en 2020 par Christelle Martin-Lacroux et Alain Lacroux pour travailler sur le biais dans les algorithmes de recrutement utilisant l'intelligence artificielle. Dans ce cadre, nous avons obtenu un financement [15] qui m'a permis d'encadrer quatre stages. Le premier avec Florent Magaud (M1 SSD ; 2021) sur l'étude de l'influence d'une base d'entraînement biaisée dans la prédiction des algorithmes ; ce stage fut complété par celui de Shuyu Wang (M2 SSD ; 2022) et a abouti à l'article Wang et al. (2025) en cours de soumission.

Pour compléter cette étude, Angélique Saillet (M1 SSD ; 2022) a travaillé sur des modélisations de profils de candidat·e·s et Alice Foraison (M2 WIC [16] ; 2022) sur l'analyse statistique de l'influence d'une recommandation algorithmique dans la décision de recruteurs. Ces deux stages permettront de continuer le travail du projet IAB@R.

1.2.8 Projet sur le Syndrome de l'Apnée du Sommeil

Dans le cadre d'une chaire MIAI [17], nous avons obtenu un financement pour Kajal Eybpoosh, une post-doc d'un an et trois mois, que j'ai co-encadrée sur l'étude du Syndrome de l'Apnée du Sommeil (projet SAS) et qui a abouti à l'article Eybpoosh et al. (2025) en cours de finalisation.

1.3 Plan du manuscrit

Dans ce manuscrit, je parlerai principalement de mes travaux réalisés depuis ma thèse. En particulier, je suppose que les **modèles de mélange** et les **modèles de segmentation** sont compris ; si besoin, je mets en annexe A des rappels sur ces modèles et des remarques sur les points communs et les différences. De plus, dans le chapitre 2, je vais parler du **modèle des blocs latents** et surtout de mes travaux depuis la fin de ma thèse. Toutefois, j'invite les personnes intéressées ou qui ne connaîtraient pas ces travaux à lire mon manuscrit de thèse (Brault (2014)) ou l'annexe B pour un résumé.

1.3.1 Notations générales

Avant d'expliquer la différence entre un modèle de mélange et la segmentation, j'introduis quelques notations. Pour la suite, nous supposons avoir un ensemble de données $\boldsymbol{Y}$ composé de n individus et, le plus souvent, J variables. Ainsi, les données sont le plus souvent représentées sous forme de matrice $\boldsymbol{Y} = (Y_{i,j})$ où $y_{i,j}$ représente l'observation de la variable aléatoire $Y_{i,j}$ correspondante à l'individu i pour la variable j (voir la gauche de la figure 1.2). La plupart du temps, $Y_{i,j}$ est un scalaire mais dans certains cas, il peut représenter un vecteur de valeurs : dans ce cas, nous supposons qu'il y a H valeurs et nous notons $Y_{i,j,h}$ la variable aléatoire de la coordonnée h du vecteur de la variable j pour l'individu i (voir la droite de la figure 1.2).

Notations

Pour aider à la lecture, les variables aléatoires sont souvent notées avec une lettre majuscule et leurs observations avec une lettre minuscule. De plus, une notation en gras symbolise un ensemble de valeurs (vecteur, matrice ou liste notamment) et les scalaires dans une police classique.
Dans le cas de **lois paramétriques**, nous notons φ la densité et $\boldsymbol{\theta}$ l'ensemble des paramètres de la loi. Enfin, nous notons $\boldsymbol{\Psi}$ l'ensemble des paramètres du modèle. Nous notons de façon abusive $p(x)$ la probabilité $\mathbb{P}(X = x)$ lorsque la variable X prenne la valeur x dans le cas discret et la densité $\varphi(x)$ dans le cas continue.

1.3.2 Plan du manuscrit

Mes travaux portent principalement sur l'application des modèles de mélange et des modèles de segmentation dans le cadre de tableaux. Prenons l'exemple d'un tableau de données binaires où les cases sont soient blanches, soient noires (voir la figure 1.3 (a)). Nous pouvons essayer de regrouper les lignes suivant un modèle de mélange (voir la figure 1.3 (b)) comme présenté dans la section 1.1.2 ou faire une

15. Partenaire du projet *IAB@R* pour *Intelligence Artificielle, Biais et Acceptabilité dans le Recrutement* soutenu par l'Agence Nationale de la Recherche dans le cadre du programme « Investissements d'avenir » (ANR-15-IDEX-02).
16. WIC pour *Web, Informatique et Connaissances*.
17. MIAI pour *Multidisciplinary Institute in Artificial intelligence*. Membre d'une chaire MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

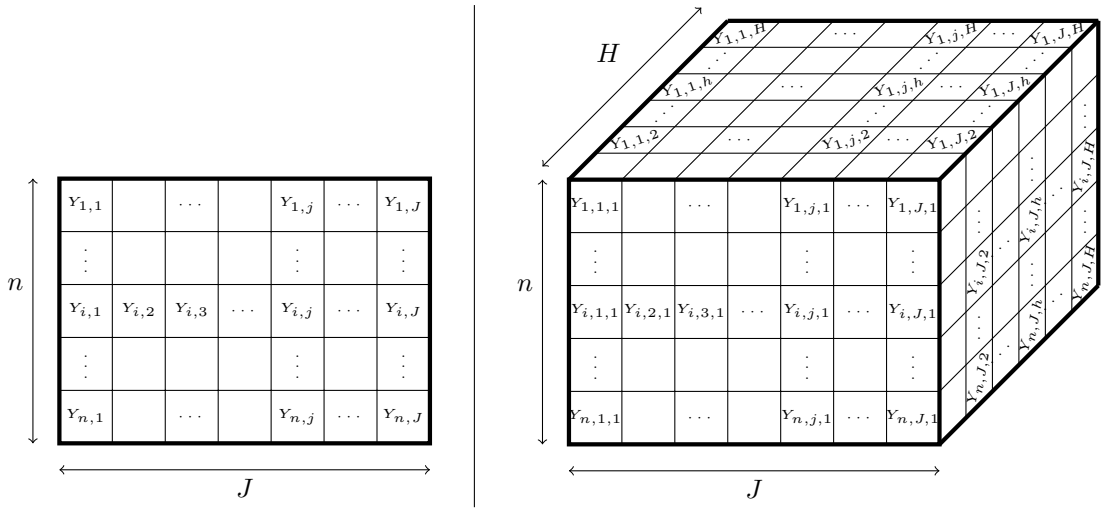

FIGURE 1.2 – Représentations schématiques des notations des données étudiées lorsque les cases $Y_{i,j}$ correspondent à des scalaires (à gauche) ou à des vecteurs (à droite).

segmentation des colonnes (voir la figure 1.3 (c)) comme pour la section 1.1.5.

Dans le chapitre 2, je m'intéresse à la classification simultanée des lignes et des colonnes d'une matrice notamment à travers l'étude du **modèles des blocs latents** (voir une application sur la gauche de la figure 1.4). De façon complémentaire, j'étudie aussi, depuis mon post-doc, la question de la segmentation simultanée des lignes et des colonnes d'une matrice ou **bisegmentation** (voir une application sur la droite de la figure 1.4) qui sera développée dans le chapitre 3.

Depuis quelques années, je m'intéresse au couplage des deux notamment dans le cas d'un modèle de mélange de segmentations (voir la figure 1.5 pour un exemple d'application). Ceci sera abordé dans le chapitre 4 qui comportera surtout des projets venant d'être soumis ou en cours. L'idée générale de ce chapitre est de trouver une façon de regrouper des individus représentés par des séries temporelles possédant des ruptures dans leurs observations.

Enfin, je conclus ce manuscrit par un bilan des perspectives égrenées durant les différents chapitres et les projets que j'aimerais continuer.

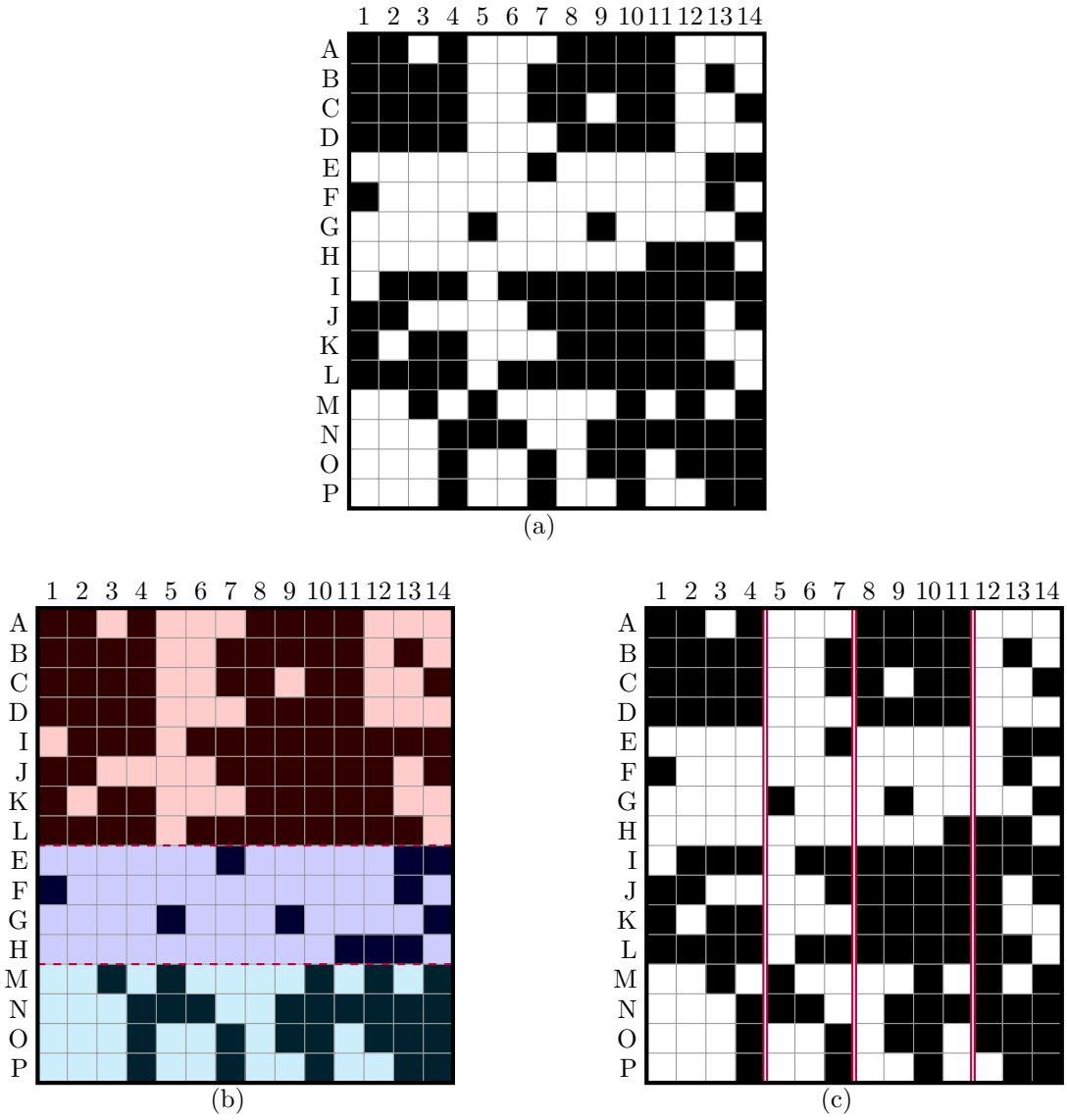

FIGURE 1.3 – Exemple d'une matrice binaire de taille 16 par 14 (a) avec un regroupement des lignes suivant un modèle de mélange (b) et une segmentation des colonnes (c).

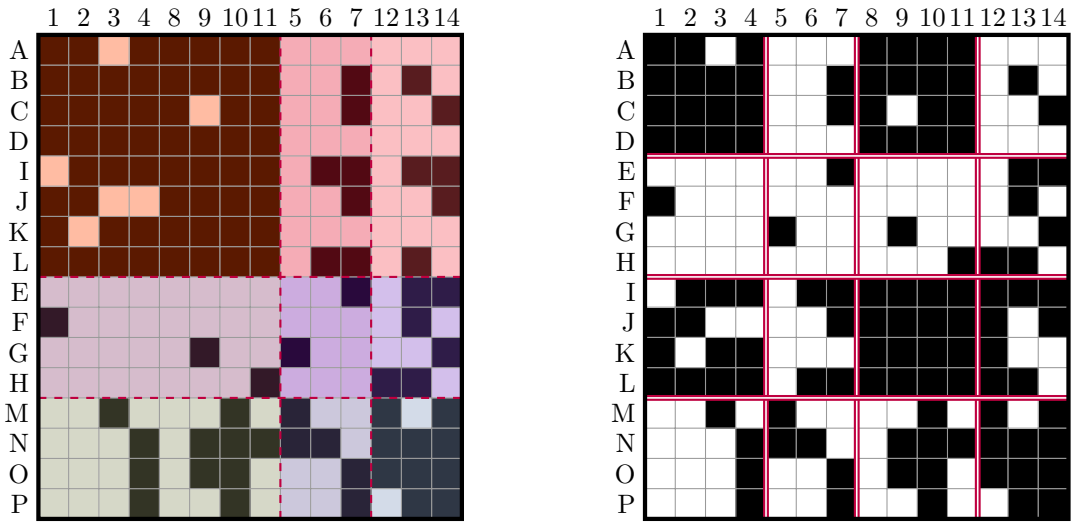

FIGURE 1.4 – Exemple d'application du modèle des blocs latents (à gauche) et de la bisegmentation (à droite) sur la matrice de la figure 1.3 (a).

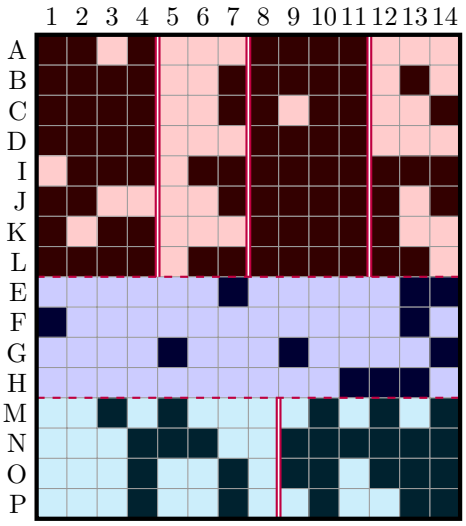

FIGURE 1.5 – Exemple d'application d'un couplage entre un modèle de mélange sur les lignes avec une segmentation des colonnes sur la matrice de la figure 1.3 (a).

Chapitre 2

Modèles de mélange et des blocs latents

"Une place pour chaque chose et chaque chose à sa place."
Aphorisme Shadok

Dans ce chapitre, je commence par parler de l'article Brault et al. (2021b) basé sur un modèle de mélange (section 2.1) puis des articles plutôt basés sur le modèle des blocs latents faits après ma thèse (pour rappel, un résumé de mes articles en lien avec ma thèse est disponible en annexe B). Pour ce faire, je commence par rappeler les hypothèses du modèle des blocs latents ainsi qu'un modèle dérivé développé dans l'article Laclau et Brault (2019) puis je présente le lien avec le transport optimal (Laclau et al., 2017) avant de parler de la méthode d'estimation des paramètres que j'ai étudié dans l'article Brault et al. (2024b) qui vient d'être soumis et je conclus la section 2.2 par l'article Robert et al. (2021) qui propose différents critères d'évaluation de la qualité des partitions. Enfin, j'ai fait le choix de faire une section particulière sur la vraisemblance du modèle des blocs latents (section 2.3.1) qui, par son comportement particulier, a permis l'obtention de deux résultats (Brault et al. (2020) et Brault et Channarond (2023)) que je juge atypiques.

2.1 Modèle de mélange et tests groupés pour le dépistage du SARS-CoV-2

Durant le premier confinement en France en 2020, j'ai travaillé avec des membres de GROUPOOL sur l'utilisation des tests groupés pour la détection du virus SARS-CoV-2. Étant donnée une population de m individus potentiellement infectés par le virus, plutôt que de faire m tests individuels, le principe des tests groupés consiste à regrouper les prélèvements de N individus et de tester ce mélange comme si c'était un seul individu. Dans le cas d'un test parfait, le résultat du groupe sera alors positif si et seulement si au moins un individu du groupe est contaminé (voir la figure 2.1 pour une représentation schématique [1]).

Une fois un groupe positif trouvé, il est possible de faire des tests individuels (voir Dorfman (1943) par exemple) ou de refaire des sous-groupes. On peut également citer des stratégies hybrides comme celle de Joly et Mallein (2022) qui testent les gens 2 ou 3 fois dans des groupes définis suivant une grille et qui permet de retrouver les individus contaminés sans avoir besoin de tests individuels s'ils ne sont pas trop nombreux.

Si nous notons p la prévalence de SARS-CoV-2 dans la population, nous voyons que la probabilité d'avoir un test positif (et donc de devoir tester à nouveau les individus du groupe) est de $1-(1-p)^N$. Par conséquent, à taille m de population fixée, plus N est grand et moins nous faisons de tests mais plus la probabilité d'avoir un test positif augmente. De plus, le choix d'un N optimal dépend de l'objectif fixé. Dans Brault et al. (2021b), nous montrons que pour une estimation de la prévalence avec un intervalle de confiance de longueur minimale, il faut choisir N_{opt} tel que le nombre moyen de tests positifs soit égal à 80% ce qui donne :

$$N_{\text{opt}} = -\frac{2 + W\left(-2e^{-2}\right)}{\log(1-p)} \approx -\frac{1.59}{\log(1-p)}$$

où W est la fonction de Lambert (introduite comme réciproque de la fonction complexe $z \mapsto ze^z$ dans

1. Crédit image : Jean-François Rupprecht, Centre de physique théorique (CNRS/Aix-Marseille Université/Université de Toulon). L'image est tirée de l'actualité CNRS faite suite à la parution de notre article Brault et al. (2021b). Pour plus d'informations, voir https://www.cnrs.fr/fr/depistage-du-covid-19-un-nouveau-modele-pour-evaluer-lefficacite-des-tests-groupes

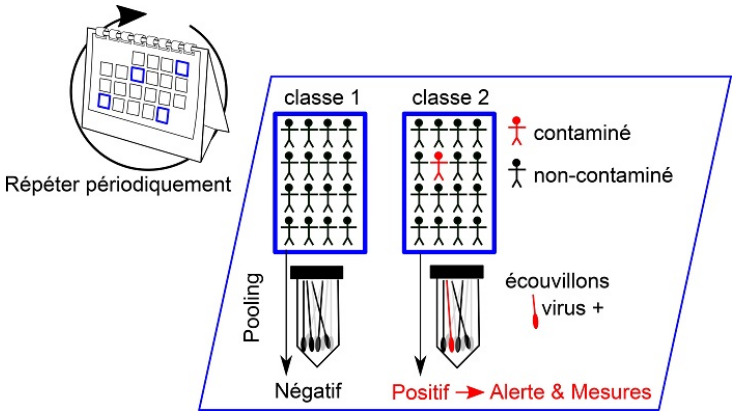

FIGURE 2.1 – Explication schématique du principe des tests groupés avec une classe 1 avec uniquement des personnes saines (donc résultat négatif pour le groupe) et une classe 2 possédant une personne contaminée (donc résultat positif pour le groupe). Crédit image : Jean-François Rupprecht, Centre de physique théorique (CNRS/Aix-Marseille Université/Université de Toulon).

Lambert (1758)). Comme la valeur du N_{opt} dépend de la prévalence, nous proposons dans l'article une méthode adaptative partant d'une première estimation arbitraire de la prévalence pour contourner ce problème.

L'un des inconvénients des tests est qu'ils ne sont pas parfaits. Il y a un risque de faux positifs (déclaration qu'une personne est contaminée alors qu'elle ne l'est pas) et de faux négatifs (déclaration qu'une personne est saine alors qu'elle est contaminée). L'une des interrogation du *Haut Conseil de la Santé Publique* (HCSP) était la perte de sensibilité due à la dilution des charges virales au sein d'un échantillon [2] pour les tests RT-PCR [3]. Pour répondre à cette interrogation, nous avons discuté avec des personnels qualifiés qui nous ont expliqué la façon de fonctionner d'un test RT-PCR. N'étant pas un spécialiste, je ne vais présenter que les éléments importants pour la modélisation notamment le fait que, après traitement, l'échantillon soit placé dans une machine PCR qui va plus ou moins doubler la quantité d'intérêt à chaque cycle et renvoyer le nombre de cycles nécessaires pour détecter cette dernière. Ainsi, plus il y a de cycles, plus la charge virale initiale était faible ou, autrement dit, le nombre de cycles Y dépend de façon logarithmique de la charge virale initiale :

$$Y = -\log_2(C) + \varepsilon$$

où ε est un bruit de mesure propre à la calibration de la machine. En pratique, si la machine tourne indéfiniment, elle risque de finir par détecter un résidu qu'elle pourrait prendre pour de la charge virale. Par conséquent, la machine est stoppée après un certain nombre de cycles, que nous noterons d_{cens}, et les échantillons qui dépassent ce nombre de cycles sont jugés sains. Plus la valeur d_{cens} est petite, plus il y a de risques d'avoir des faux négatifs et plus elle est grande, plus il y a des risques d'avoir des faux positifs. La valeur de référence dans ce cadre est autour de 38 cycles de telle sorte que, d'après les experts interrogés, il n'y ait pas de faux positifs (mais au risque d'avoir des faux négatifs). Par conséquent, l'estimation du nombre de cycles peut être caractérisée par :

$$Y = \min\left[-\log_2(C), d_{\text{cens}}\right] + \varepsilon. \tag{2.1}$$

À partir de là, nous pouvons estimer l'influence d'une dilution sur la qualité des résultats. Pour cela, nous faisons l'hypothèse que les échantillons sont mélangés en proportion égale et nous pouvons donc approcher la concentration $C^{(N)}$ de l'échantillon d'un poolage de N concentrations $(C_1, \ldots, C_N)$ par :

$$C^{(N)} \approx \frac{1}{N}\sum_{j=1}^{N} C_j.$$

Ainsi, en reprenant l'équation (2.1), nous montrons que le nombre de cycles de l'échantillon poolé est caractérisé par :

$$Y^{(N)} \approx \min\left[-\max_{j\in\{1,\ldots,N\}} \log_2\left(C_j\right), d_{\text{cens}} - \log(N)\right] + \varepsilon. \tag{2.2}$$

2. Communiqué du 10 mai 2020 : https://www.hcsp.fr/Explore.cgi/AvisRapportsDomaine?clefr=828
3. *Reverse Transcription Polymerase Chain Reaction*

Et le nombre de cycles dépend de la concentration maximale et de la taille de l'échantillon N qui vient implicitement décaler de façon logarithmique la valeur maximale du nombre de cycles par rapport à la situation où la personne avec la concentration maximale avait été testée individuellement. Pour bien répondre aux inquiétudes du HCSP, il fallait donc comprendre la distribution du nombre de cycles afin de voir la perte liée à dilution. Pour ce faire, nous avons étudié des données disponibles dans la littérature mais peu étaient en open data. Après de nombreux échanges infructueux, nous avons proposé une première étude en *mimant* les données de l'article Jones et al. (2020). Sur la figure 2.2, nous avons représenté le résultat d'une simulation *mimant* ces données (la simulation a été faite à partir d'un histogramme proposé dans leur article). Nous pouvons voir que la distribution est multimodale (au moins deux modes, peut-être trois voir plus) et nous avons choisi de faire une estimation par un modèle de mélange avec des lois gaussiennes de moyennes et variances libres :

$$Y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k f\left(Y_i; \mu_k, \sigma_k\right)$$

où f est la densité d'une loi gaussienne univariée.

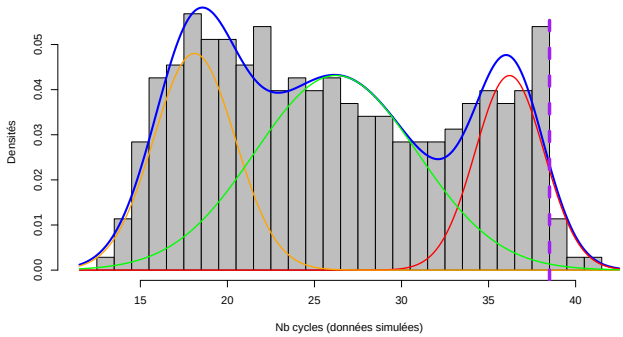

FIGURE 2.2 – Représentation sous forme d'histogramme d'une simulation mimant les données de l'article Jones et al. (2020) sur le nombre de cycles pour une populations de 3 712 patients infectés par le SARS-CoV-2 avec une estimation par un mélange de trois gaussiennes aux moyennes et variances libres : les traits pleins orange, vert et rouge correspondent aux densités des groupes et le trait plein bleu à la densité du modèle de mélange. Le trait vertical violet en pointillés correspond à ce que nous pensons être d_{cens} pour la plupart des machines.

Comme nous ne connaissons pas le nombre de groupes, nous avons utilisé le critère *BIC* (*Bayesian Information Criterion*) de Schwarz et al. (1978) qui pénalise la log-vraisemblance par le nombre de paramètres à estimer. Dans notre cas, il vaut :

$$BIC(K) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_K} \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k f\left(Y_i; \mu_k, \sigma_k\right) \right] - \frac{3K-1}{2} \log(n).$$

Comme nous avions des données simulées basées sur les histogrammes, nous avons relancé cent fois la procédure et avons obtenu trois clusters dans 95% des cas et quatre clusters dans les 5% restants. Sur la figure 2.2, nous avons représenté les estimations des densités pour un cas à trois clusters. Nous observons qu'un premier cluster (en orange) se situe autour des valeurs de 15-20 cycles (donc une forte charge virale), un deuxième (en vert) avec une variance plus élevée se positionne autour de 20-33 cycles et le dernier (en rouge) a une variance plus faible et la moyenne est très proche de la chute observée des valeurs. Or, cette chute se situe autour de la valeur 38 qui est la valeur donnée pour le nombre maximal de cycles avant de déclarer qu'un patient est sain. Nous avons représenté par un trait vertical violet en pointillés l'estimation du maximum de cycles de la plupart des machines (dépendant de l'utilisateur et des ajustements faits après l'utilisation).

En particulier, il semble que les observations correspondantes au cluster rouge sont censurées et qu'il est possible qu'il y ait plus d'individus avec une concentration virale non nulle mais qui n'apparaissent pas car jugés sains par la procédure (donc un potentiel faux négatif). Pour prendre en compte cette censure (connue), nous avons proposé d'utiliser des lois gaussiennes avec censure partielle en d_{cens}, notées

$\mathcal{CN}_{d_{\text{cens}}}(\mu, \sigma, q)$, de paramètres $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^\star$ et $q \in [0;1]$ et définies sur $\mathbb{R}$ par la densité :

$$f_{d_{\text{cens}},\mu,\sigma,q}(y) = \frac{f_{\mu,\sigma}(y)}{q + (1-q)F_{\mu,\sigma}(d_{\text{cens}})} \begin{cases} 1 & \text{si } x \leq d_{\text{cens}} \\ q & \text{sinon} \end{cases} \tag{2.3}$$

où d_{cens} est supposée fixée a priori, $f_{\mu,\sigma}$ (resp. $F_{\mu,\sigma}$) est la densité (resp. la fonction de répartition) d'une loi gaussienne $\mathcal{N}(\mu, \sigma)$ et la valeur q correspond à la probabilité d'observer le nombre de cycles Y si la concentration est plus petite que le seuil d_{cens}. Sur la figure 2.3, nous avons comparé la densité d'une loi $\mathcal{CN}_{1.5}(0, 1, 1/2)$ (en rouge) et de la loi $\mathcal{N}(0, 1)$ (en bleue transparente) correspondante à la même loi sans censure (donc $q = 1$). Nous pouvons remarquer que la densité de la loi $\mathcal{CN}_{1.5}(0, 1, 1/2)$ est plus grande avant la rupture d_{cens} et plus petite après que celle de la loi $\mathcal{N}(0, 1)$. En particulier, nous voyons sur l'équation (2.3) et la figure 2.3 que si q tend vers 1 ou si la valeur de d_{cens} tend vers l'infini, alors la densité de la loi $\mathcal{CN}_{d_{\text{cens}}}(0, 1, 1/2)$ converge simplement vers celle de la loi $\mathcal{N}(0, 1)$.

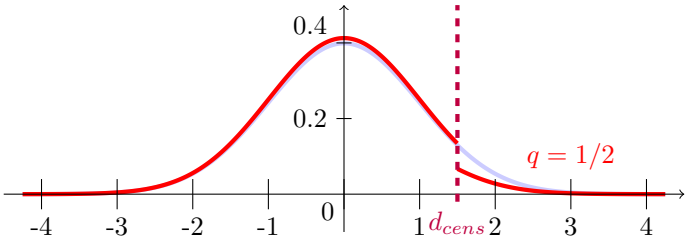

FIGURE 2.3 – Représentation de la densité d'une loi gaussienne partiellement censurée $\mathcal{CN}_{1.5}(0, 1, 1/2)$ (en rouge) et de celle d'une loi gaussienne centrée réduite $\mathcal{N}(0, 1)$ (en bleue transparente). La censure $d_{\text{cens}} = 1.5$ est symbolisée par le trait vertical en pointillés violets.

Le paramètre q est la probabilité d'observer quand même un nombre de cycles plus grand que la valeur imposée. Elle représente le fait qu'il y a une recalibration faite après l'expérience et/ou le fait que certaines fois, la machine va être programmée pour aller au delà du seuil d'autres machines. Notons que, pour ce deuxième cas, il serait préférable de proposer une probabilité q décroissante en escalier en fonction des valeurs maximales successives des machines. Pour contourner ce problème, nous avons également étudié dans Brault et al. (2021b) le cas d'une censure totale ($q = 0$) en d_{cens}, notée $\mathcal{CN}_{d_{\text{cens}}}(\mu, \sigma)$, en *oubliant* les valeurs après la censure d_{cens}.

Perspectives

Il est important de noter que, dans cette modélisation, d_{cens} est connue a priori et n'est pas remise en cause par la suite. Il aurait été possible de tester des lois où cette valeur est un paramètre ou une variable aléatoire (dépendante par exemple de la machine).

Nous montrons que, comme ces lois appartiennent à la famille exponentielle, l'estimateur du maximum de vraisemblance est fortement consistant et suit asymptotiquement une loi gaussienne (voir Barndorff-Nielsen (2014)). Par contre, il n'existe pas de forme analytique de l'estimateur du maximum de vraisemblance ; nous avons utilisé la méthode d'optimisation type Newton implémentée dans la fonction `nlm` du logiciel R (voir R Core Team (2023)) pour contourner ce problème [4]. Pour le passage au modèle de mélange, nous avons fait deux hypothèses :

- le nombre maximal théorique de cycles d_{cens} est le même pour tous les groupes.
- la probabilité q_k qu'une observation du $k^{\text{ème}}$ groupe et ayant dépassé la valeur d_{cens} soit quand même observée dépend du cluster.

Implémentations

Comme d_{cens} et q dépendent des machines et de leurs utilisations (donc de paramètres externes aux observations), il aurait été plus cohérent que tous les deux soient indépendants du cluster d'appartenance. Néanmoins, si nous faisons cette hypothèse pour q, les paramètres des clusters dépendent les uns des autres et nous sommes obligés de les maximiser en même temps. Par contre, en faisant l'hypothèse que

4. Les codes sont disponibles sur mon Github : https://github.com/Lionning/CovidPooling

les probabilités q_k dépendent des clusters, nous pouvons maximiser les paramètres de chaque cluster indépendamment (et donc en parallèle). Ce choix a été fait pour des raisons pratiques et comme les estimations des paramètres q_k sur les données réelles étaient proches, nous avons conservé cette simplification dans Brault et al. (2021b).

Pour l'estimation, nous avons utilisé l'algorithme *EM* en choisissant trois groupes, en relançant plusieurs fois avec différentes initialisations et en conservant le modèle avec la meilleure vraisemblance. Pour le choix de d_{cens}, nous avons testé différentes limites de classes en supprimant les données après la rupture quand nous testions les lois avec censure totale. Lorsque les valeurs de d_{cens} étaient différentes de 39.5 et 38.5, l'étape d'estimation pour les modèles avec censure partielle n'arrivait pas à converger. Sur la figure 2.4, nous avons représenté les estimations pour les deux types de modèles pour ces deux valeurs de d_{cens}.

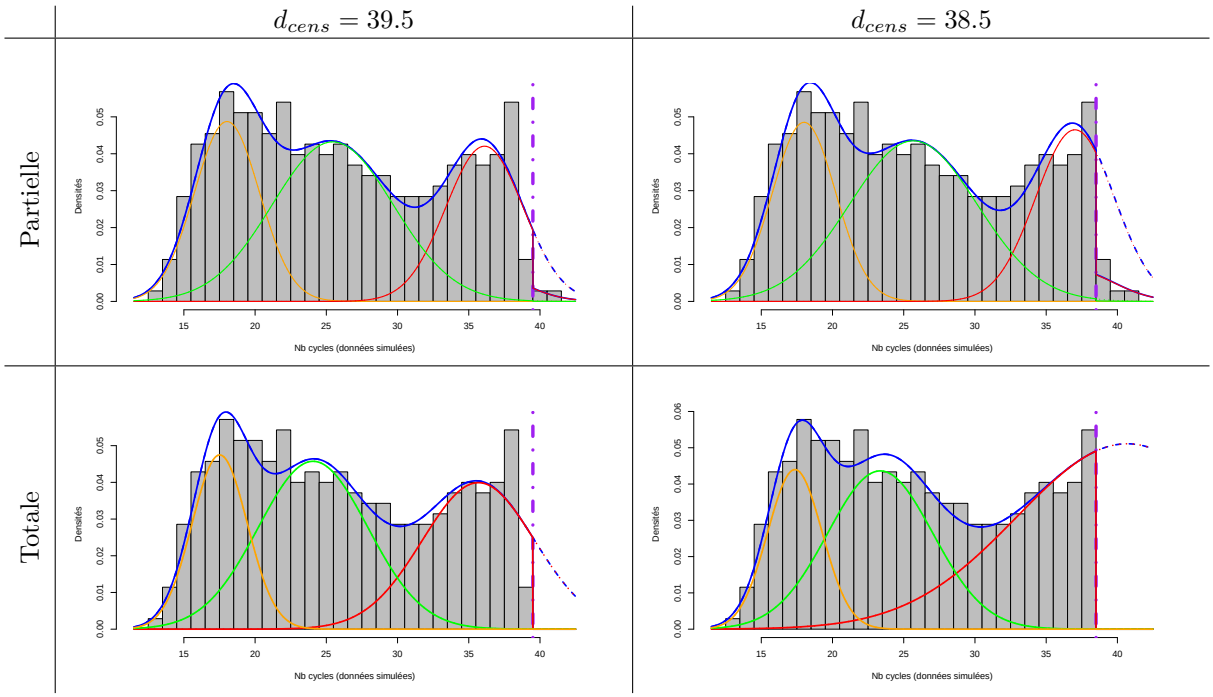

FIGURE 2.4 – Représentation des estimations par un modèle de mélange de lois gaussiennes avec censure partielle (ligne du haut) et totale (ligne du bas) suivant deux valeurs de d_{cens} (colonnes). Pour chaque graphique, nous avons représenté le nombre de cycles simulés sous forme d'histogramme en enlevant les barres à droite de la censure (trait vertical en pointillés violets) pour les censures totales ; les traits pleins orange, vert et rouge correspondent aux densités des groupes et le trait plein bleu à la densité du modèle de mélange. Les traits en pointillés bleus et rouges correspondent aux densités non tronquées.

En pointillés, nous avons prolongé les lois comme si elles n'avaient pas eu de censure afin d'avoir une idée de la proportion de faux négatifs dans la population. Attention, ce ne sont pas des densités car l'air sous la courbe ne vaut pas 1 ; pour être rigoureux, il faudrait recalculer les valeurs $\boldsymbol{\pi}$ des proportions de chaque groupe afin de mieux équilibrer. Nous pouvons voir que les courbes, notamment celle en rouge de la classe la plus proche de d_{cens} approchent mieux l'histogramme que pour le modèle de mélange classique de la figure 2.2. Pour la censure totale, nous voyons l'influence des valeurs juste avant d_{cens} : pour $d_{\text{cens}} = 38.5$, la grande barre de la classe $]37.5; 38.5]$ incite l'algorithme à proposer un très grand groupe d'individus non détectés.

À l'aide de ces estimations, nous proposons une estimation du taux de faux négatifs avant l'utilisation de poolage et après en fonction du nombre d'échantillons mélangés (voir Brault et al. (2021b)). Ce travail fut notamment l'occasion de vulgarisation auprès du grand public [5].

5. Articles de journaux tels que la gazette du laboratoire par exemple et interview pour France Bleu Isère notamment (voir https://membres-ljk.imag.fr/Vincent.Brault/)

Perspectives

Dans la continuité de ce travail, il serait intéressant de faire une étude approfondie en décalant la valeur d_{cens} dans les données réelles (donc récupérer de nouvelles données de personnes infectées) pour voir si les estimations restent cohérentes ou si on a eu un effet de bord. Une étude avec des valeurs de q en escalier afin de prendre en compte les descentes successives pourrait également être intéressante mais nécessiterait une implémentation plus subtile (notamment si nous supposons que les probabilités décroissent). Enfin, il pourrait être intéressant de généraliser l'étude à d'autres maladies utilisant ce type de test par exemple.

2.2 Modèle des blocs latents

Le **modèle des blocs latents** est un modèle probabiliste se plaçant dans le cadre de la classification croisée. Il se base sur trois hypothèses. La figure 2.5 présente un schéma des notations utilisées.

Hypothèse (Hypothèses du modèle des blocs latents)

$(\mathcal{H}_1)$: nous supposons qu'il existe une partition $\boldsymbol{z}$ des n lignes en K classes et une partition $\boldsymbol{w}$ des J colonnes en L classes de telle sorte que les variables aléatoires $Y_{i,j}$ soient indépendantes conditionnellement au couple $(\boldsymbol{z}, \boldsymbol{w})$:

$$p(\boldsymbol{y}\,|\boldsymbol{z}, \boldsymbol{w}; \boldsymbol{\Psi}) = \prod_{i=1}^{n}\prod_{j=1}^{J} p(y_{i,j}\,|z_i, w_j; \boldsymbol{\Psi}) \tag{2.4}$$

et leurs lois ne dépendent que du bloc (z_i, w_j).

$(\mathcal{H}_2)$: les variables $\boldsymbol{Z}$ et $\boldsymbol{W}$ sont indépendantes :

$$\forall(\boldsymbol{z}, \boldsymbol{w}) \in \mathcal{Z}\times\mathcal{W}, \quad p(\boldsymbol{z}, \boldsymbol{w}; \boldsymbol{\Psi}) = p(\boldsymbol{z}; \boldsymbol{\Psi})\, p(\boldsymbol{w}; \boldsymbol{\Psi}) \tag{2.5}$$

où $\mathcal{Z}$ et $\mathcal{W}$ sont les espaces de partitions.

$(\mathcal{H}_3)$: pour chaque ligne i, la loi d'affectation des labels est une loi multinomiale dont les paramètres sont les mêmes pour toutes les lignes ; de même pour les colonnes. Comme pour les modèles de mélange, $\boldsymbol{\pi}$ est la probabilité d'appartenance aux classes en lignes. Pour les colonnes, la notation sera $\boldsymbol{\rho}$.

Comme dit dans l'hypothèse $(\mathcal{H}_1)$, nous pouvons choisir la loi que nous voulons pour chaque bloc. Néanmoins, dans l'essentiel de mes travaux, je me suis intéressé au cas où chaque bloc (k, ℓ) contient des observations issues de lois paramétriques de densité φ et paramètres $\boldsymbol{\theta}_{k,\ell}$ notamment aux **lois multinomiales** et au cas particulier des **lois de Bernoulli**.

Au final, grâce aux hypothèses précédentes, nous obtenons la vraisemblance suivante du modèle :

$$p(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{(\boldsymbol{z},\boldsymbol{w})\in\mathcal{Z}\times\mathcal{W}} \underbrace{\left\{\left[\prod_{i=1}^{n}\prod_{j=1}^{J}\prod_{k=1}^{K}\prod_{\ell=1}^{L}\varphi(\boldsymbol{y}_{i,j}; \boldsymbol{\theta}_{k,\ell})^{z_{i,k}w_{j,\ell}}\right]}_{\text{Hypothèse } (\mathcal{H}_1)} \times \underbrace{\left(\prod_{k=1}^{K}\pi_k^{z_{+,k}}\right)\left(\prod_{\ell=1}^{L}\pi_\ell^{w_{+,\ell}}\right)\right\}}_{\text{Hypothèses } (\mathcal{H}_2) \text{ et } (\mathcal{H}_3)} \tag{2.6}$$

où $\mathcal{Z}\times\mathcal{W}$ sont les ensembles de quadrillages possibles de la matrice. Sur la figure 2.6, une représentation schématique des hypothèses précédentes a été proposée.

2.2.1 Modèle des blocs latents avec classe de bruit

Lors d'une collaboration avec Charlotte Laclau (post-doctorant au LIG à l'époque), nous avons étendu le modèle des blocs latents en ajoutant une nouvelle classe, appelée **classe de bruit**, sur les colonnes (voir Laclau et Brault (2019)). Dans ce cas, nous ajoutons une classe 0 à la partition $\boldsymbol{w}$ qui représente les colonnes qui *ne participent pas* à la classification des lignes ; nous supposons qu'il existe un paramètre $\phi \in [0, 1]$ représentant la probabilité pour chaque colonne d'appartenir à celles qui participeront dans la classification jointe des lignes et des colonnes. Ainsi, pour chaque colonne j, la loi de la variable W_j est une loi mutltinomiale $\mathcal{M}(1; 1-\phi, \phi\boldsymbol{\rho})$ sur $\{0, 1, \ldots, L\}$. Nous supposons que, pour chaque colonne j de

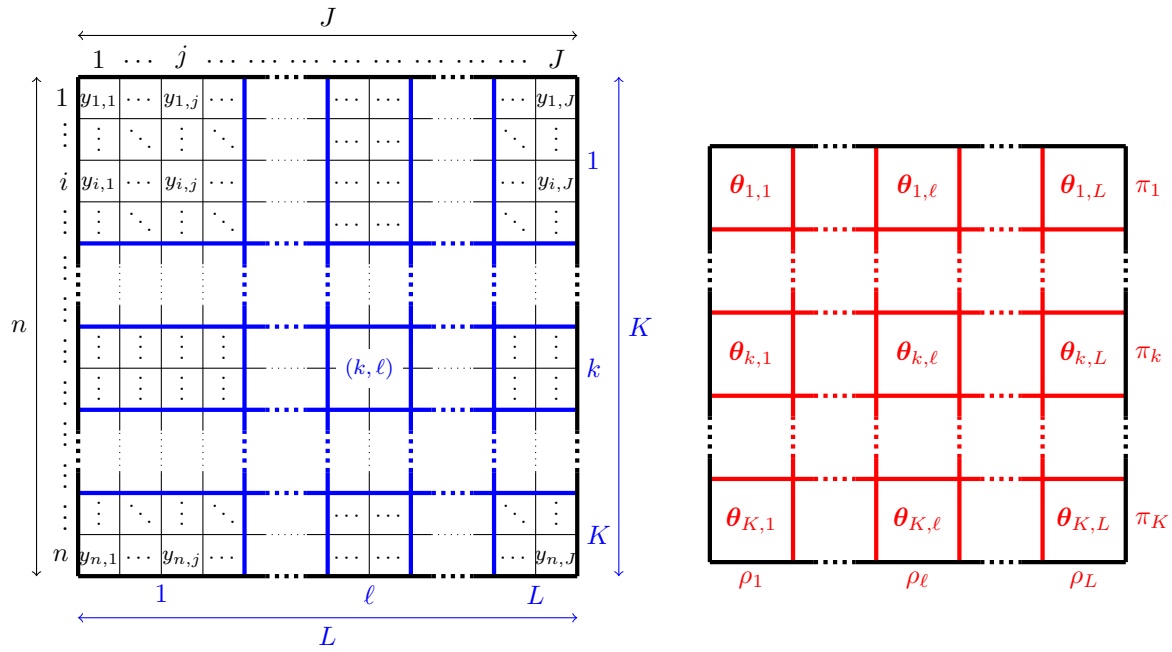

FIGURE 2.5 – Représentation schématique des notations utilisées pour le modèles des blocs latents : sur la gauche sont représentées les notations liées aux données (en noir) et aux clusters (en bleu) tandis que, sur la figure de droite, nous avons représenté les notations liées aux paramètres $\boldsymbol{\Psi}$.

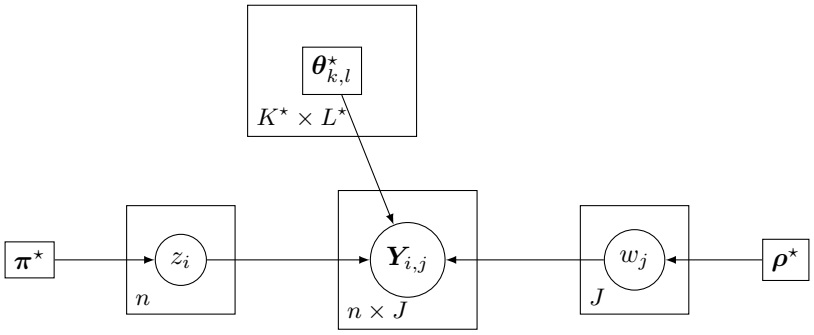

FIGURE 2.6 – Représentation graphique du modèle des blocs latents paramétrique : les carrés correspondent à des paramètres et les ronds à des variables.

la classe W_0, il existe un paramètre λ_j de telle sorte que les observations $(y_{i,j})_{1\leq i\leq n}$ sont indépendantes et de même loi dépendant du paramètre λ_j. Pour le reste de la matrice, conditionnellement au fait que les cases ne sont pas dans une colonne de la classe de bruit, les observations sont issues d'un modèle des blocs latents. Ce modèle, développé et étudié dans notre article Laclau et Brault (2019), est appelé **modèle des blocs latents avec classe de bruit** (ou *Noise-free latent block model* abrégé *NFLBM*).

Exemple

Dans le cas binaire étudié dans notre article Laclau et Brault (2019), si $w_{j,0} = 1$ alors nous avons la loi suivante :

$$y_{\cdot,j} \sim \mathcal{B}(\lambda_j)^n .$$

Dans notre article, nous comparons les deux modèles sur des données simulées et des données réelles en suivant le protocole présenté dans la section 2.2.3 avec un critère de sélection de modèle permettant de choisir entre les deux modèles. Nous constatons que le modèle *NFLBM* avait plus de chances d'être correctement sélectionné lorsque la classe de bruit était grande (donc ϕ petit), qu'il n'y avait pas trop de colonnes par rapport au nombre de lignes et que la reconnaissance des blocs dans la partie LBM n'était pas trop compliquée. A l'opposé, s'il n'y a pas de classe de bruit ou si elle est trop petite et qu'il n'y a

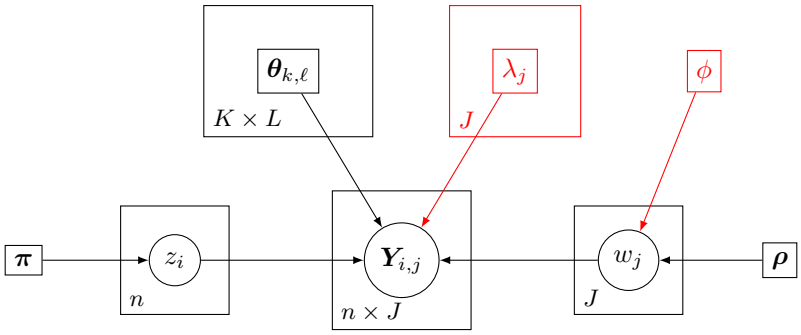

FIGURE 2.7 – Représentation graphique du modèle des blocs latents paramétriques avec classe de bruit : les carrés correspondent à des paramètres et les ronds à des variables. En rouge se trouvent les ajouts par rapport à la version du *LBM* présentée dans la figure 2.6

pas assez d'informations pour reconnaître une classe de bruit, c'est le LBM qui est sélectionné (voir la section 7.1.5 de notre article Laclau et Brault (2019)).

Nous comparons également les modèles sur des données de génétique proposées par Rosenberg et al. (2002) et, après traitements, sont représentées par une matrice de $n = 2112$ lignes et $J = 4689$ colonnes (voir la figure 2.8).

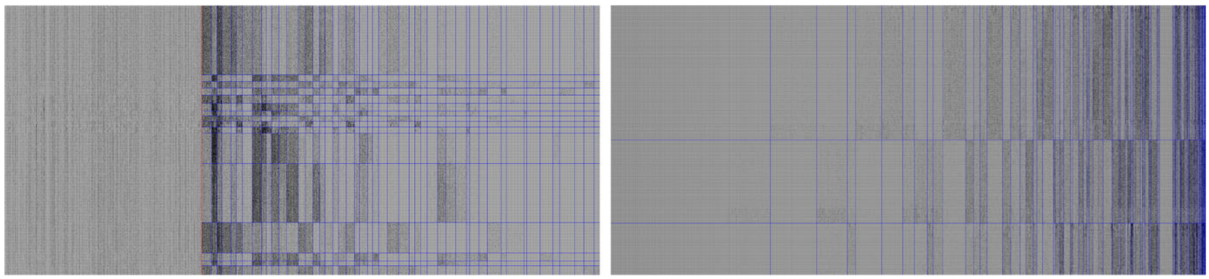

FIGURE 2.8 – Représentation des matrices obtenues par le modèle *NFLBM* (à gauche) avec 16 classes en lignes et 49 en colonnes (plus une classe de bruit) et le modèle *LBM* (à droite) avec 3 classes en lignes et 78 en colonnes. Pour le *NFLBM*, un trait rouge vertical a été tracé pour différencier la classe de bruit (à gauche) de la partie relative au *LBM* (à droite).

Le point le plus cohérent avec le fait d'introduire une classe de bruit est que le modèle des blocs latents propose 78 classes en colonnes (à droite sur la figure) contre 49 pour le *NFLBM*. En particulier, nous voyons sur la droite de la matrice et sur la gauche des colonnes assez homogènes tout le long des blocs. Toutefois, nous observons aussi que le modèle des blocs latents ne propose que 3 classes en lignes contre 16 en colonnes. Le faible nombre de classes pour le *LBM* vient peut-être du fait qu'il est plus compliqué d'être différent pour l'ensemble des blocs et le critère de sélection de modèle a choisi des blocs plus gros du coup. À l'opposé, le *NFLBM* a créé des blocs plus contrastés (sur le bord du trait rouge de la gauche de la figure ; ce trait délimitant les colonnes appartenant à la classe de bruit des autres).

Perspectives

Dans les travaux préliminaires à cet article, nous avons également testé une valeur unique de λ qui s'est avérée moins concluante. De plus, dans les projets futurs, nous souhaitons regarder avec une classe de bruit en ligne en plus ; la difficulté est de caractériser les cases appartenant à la fois aux deux classes de bruit.

2.2.2 Lien avec le transport optimal

Dans le cadre d'une collaboration avec notamment Charlotte Laclau (au LIG à ce moment là) chercheuse en machine learning et Ievgen Redko (INSA de Lyon à cette époque) en transport optimal, nous constatons que l'estimation du modèle des blocs latents à l'aide d'une approximation variationnelle peut être vue comme un problème de transport des lignes vers les colonnes. En effet, dans l'article Laclau et al.

(2017), nous regardons ce qu'il se passe si nous cherchons à *transporter* les lignes vers les colonnes d'une matrice carrée. D'un côté, nous cherchons à minimiser la distance de Wasserstein régularisée entre les mesures des probabilités empiriques de l'ensemble des lignes et de l'ensemble des colonnes. Pour résoudre ce problème, Benamou et al. (2015) démontrent qu'il existe une solution basé sur un noyau de Gibbs. Dans notre article, nous montrons que si nous cherchons à optimiser l'**énergie libre** (voir l'annexe B.2.1 si besoin) en prenant des familles de distributions jointes du couple $(\boldsymbol{Z}, \boldsymbol{W})$ dans une famille de distributions de Gibbs alors nous recherchons le même type de solutions. Ceci permet l'utilisation d'algorithmes et de méthodes de sélection du nombres de clusters développés par la communauté du transport optimal.

2.2.3 Procédures d'estimation et robustesse

Depuis la fin de ma thèse, j'ai eu l'occasion d'utiliser les estimations du modèle des blocs latents pour différents jeux de données. Je me suis donc intéressé à perfectionner mon étude des différents algorithmes proposée dans le cadre de notre article Keribin et al. (2014). Pour le moment, lorsque le nombre maximum de blocs n'est pas trop grand, je suis arrivé à une préférence pour la procédure suivante :

1. Pour chaque couple (K, L), nous faisons la procédure suivante :
   (a) Nous estimons $\widehat{\boldsymbol{\Psi}}$ maximisant l'énergie libre bayésienne en utilisant l'échantillonneur de *Gibbs* couplé à l'algorithme *V-Bayes*.
   (b) Nous cherchons la partition $\left(\widehat{\boldsymbol{Z}}_K, \widehat{\boldsymbol{W}}_L\right)$ maximisant les probabilités a posteriori obtenues avec le paramètre $\widehat{\boldsymbol{\Psi}}$.
   (c) Nous calculons la valeur $ICL(K, L)$ du critère *ICL* associé.
2. À la fin, nous sélectionnons le couple (K, L) maximisant le critère *ICL* ainsi que le paramètre $\widehat{\boldsymbol{\Psi}}$ et la partition $\left(\widehat{\boldsymbol{Z}}_K, \widehat{\boldsymbol{W}}_L\right)$ associés.

Toutefois, je n'avais pas eu le temps d'aborder toutes les études possibles de la qualité des estimations dans l'article Keribin et al. (2014) (qui comportait déjà beaucoup de cas). J'ai donc profité du projet CoPoLangues pour m'intéresser aux problèmes de **robustesse** de la procédure avec Frédérique Letué (LJK) et Marie-Jo Martinez (LJK) ce qui a aboutit à l'article Brault et al. (2024b) en cours de soumission.

Nous avons d'abord étudié la dualité entre maximiser la vraisemblance et le critère *ICL* étant donné un nombre de classes en lignes et en colonnes. Dans leur article, Biernacki et al. (2000) expliquent qu'étant donné un nombre de classes K dans un modèle de mélange, maximiser la vraisemblance et le critère *ICL* ne sont pas tout à fait les mêmes objectifs. Pour le modèle des blocs latents, nous avons étudié ceci sur des données réelles et notre stagiaire, Margaux Leroy (actuellement en thèse au LJK) a montré que c'est également le cas pour le modèle des blocs latents (voir par exemple sa présentation Leroy et al. (2021)).

De plus, comme nous appliquions nos estimations pour former des groupes d'étudiants en fonction de leurs réponses à des tests, nous proposons dans notre article Brault et al. (2024b) une étude montrant que les classes estimées ne bougent pas trop (moins de 10% des élèves ne se retrouvent pas avec les mêmes étudiants) même quand nous enlevons une partie important des élèves dans le cadre de l'examen en japonais. Toutefois, le nombre estimé de blocs varie et va être proportionnel avec le nombre d'élèves utilisés pour l'estimation des paramètres.

Littérature connexe

Cette procédure possède une complexité de l'ordre de $\mathcal{O}\left(K_{\max}^2 L_{\max}^2 n J N_{\text{Algo}}\right)$ où $K_{\max} \times L_{\max}$ sont les nombres maximums de blocs testés et N_{Algo} est le nombre d'itération des algorithmes utilisés ce qui peut prendre énormément de temps. Pour contourner ce problème, il existe des procédures comme celle implémentée dans le package `bikm1` du logiciel `R` proposée par Robert (2021) dont le principe est de trouver un chemin traversant de manière optimale la grille plutôt que de tester toutes les configurations possibles.

Perspectives

En perspective de ce projet, j'aimerais continuer à challenger la procédure dans d'autres contextes et comparer la stabilité avec la procédure `bikm1`. En parallèle de ce travail, j'aimerais pouvoir proposer un critère de *stabilité* associé aux estimations proposées par la procédure. De plus, il serait intéressant d'analyser les résultats de robustesse du modèle *NFLBM* (de l'article Laclau et Brault (2019) ; voir la section 2.2.1) afin de voir le gain, ou pas, de l'ajout de la classe de bruit.

2.2.4 Qualité des estimations

L'une des problématiques des **modèles de mélange** et a fortiori du **modèle des blocs latents** est le *label switching* qui complexifie la comparaison de deux partitions $\boldsymbol{z}$ (avec K clusters) et $\boldsymbol{z}'$ (avec K' clusters) qui peuvent être égales mais avec des permutations des labels. Dans notre article Robert et al. (2021), nous étudions voir *améliorons* trois critères de comparaison de deux partitions avant de les étendre au cas du modèle des blocs latents.

Adjusted Rand Index : Le premier critère étudié est l'*Adjusted Rand Index*, noté *ARI*,. On parle de critère *ajusté* car il est basé sur le *Rand Index* de Rand (1971) qui mesure la similarité entre les partitions $\boldsymbol{z}$ et $\boldsymbol{z}'$ en étudiant pour chaque paire d'individus si elles sont dans les mêmes clusters dans la partition $\boldsymbol{z}$ et la partition $\boldsymbol{z}'$. En procédant ainsi, le critère n'est pas gêné par le label switching et les calculs se font en $\mathcal{O}(n)$ grâce au **tableau de contingence** défini pour tout $k \in \{1, \ldots, K\}$ et $k' \in \{1, \ldots, K'\}$ par :

$$\mathbf{Cont}_{k,k'}^{(\boldsymbol{z},\boldsymbol{z}')} = \sum_{i=1}^{n} z_{i,k} z'_{i,k'}. \tag{2.7}$$

Toutefois, Fowlkes et Mallows (1983) montrent que pour des partitions $\boldsymbol{z}$ et $\boldsymbol{z}'$ tirées au hasard, l'essentiel des valeurs du critère *RI* se concentrent dans un petit intervalle proche de 1. Pour contourner ce problème, Hubert et Arabie (1985) proposent une version corrigée, l'*Adjusted Rand Index*, en s'appuyant sur la loi polyhypergéométrique dont Warrens (2008) montre qu'il peut être écrit sous la forme :

$$ARI(\boldsymbol{z}, \boldsymbol{z}') = \frac{\sum_{k=1}^{K}\sum_{k'=1}^{K'} \binom{\mathbf{Cont}_{k,k'}^{(\boldsymbol{z},\boldsymbol{z}')}}{2} - \frac{1}{\binom{n}{2}} \left[\sum_{k=1}^{K} \binom{z_{+,k}}{2}\right] \left[\sum_{k'=1}^{K'} \binom{z'_{+,k'}}{2}\right]}{\frac{1}{2}\left[\sum_{k=1}^{K} \binom{z_{+,k}}{2} + \sum_{k'=1}^{K'} \binom{z'_{+,k'}}{2}\right] - \frac{1}{\binom{n}{2}} \left[\sum_{k=1}^{K} \binom{z_{+,k}}{2}\right] \left[\sum_{k'=1}^{K'} \binom{z'_{+,k'}}{2}\right]}$$

où $z_{+,k}$ (resp. $z'_{+,k'}$) est l'effectif de la classe k de la partition $\boldsymbol{z}$ (resp. la classe k' de la partition $\boldsymbol{z}'$). Le critère vaut 1 lorsque les deux partitions sont identiques au label switching près et peut ne peut descendre en dessous d'une borne inférieure négative dont nous avons montré qu'elle tend vers 0 lorsque le nombre d'individus tend vers l'infini (voir Robert et al. (2021)).

De plus, nous étendons le critère *ARI* (*Adjusted Rand Index*) proposé par Hubert et Arabie (1985) au cas des classifications jointes appelé *CARI* (*Coclustering Adjusted Rand Index*) qui prend en entrée deux partitions $(\boldsymbol{z}_1, \boldsymbol{w}_1)$ et $(\boldsymbol{z}_2, \boldsymbol{w}_2)$ et compte pour chaque bloc de chacune des partitions le nombre de cases en commun. Nous montrons que, pour faire cela, il suffit de prendre le produit de Kronecker des tableaux de contingence $\left(\mathbf{Cont}_{k,k'}^{(\boldsymbol{z}_1,\boldsymbol{z}_2)}\right)_{1\leq k,k'\leq K}$ et $\left(\mathbf{Cont}_{\ell,\ell'}^{(\boldsymbol{w}_1,\boldsymbol{w}_2)}\right)_{1\leq \ell,\ell'\leq L}$ définis par (2.7).

L'erreur de classification : Le deuxième critère étudié dans notre article est celui de l'erreur de classification. Étant données deux partitions $\boldsymbol{z}$ et $\boldsymbol{z}'$ en K clusters, le principe est de compter le nombre moyen d'individus qui ne se retrouvent pas dans la même classe dans les deux partitions :

$$1 - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{i,k} z'_{i,k}.$$

Ainsi, ce critère vaut 0 si les partitions sont identiques et 1 si tous les individus se retrouvent chacun dans des partitions différentes. Le problème de cette formulation est que deux partitions identiques mais avec des numéros de classes complètement différents aura une valeur de 1 pour le critère. Pour contourner ce problème, Lomet et al. (2012), entre autres, calculent l'erreur en testant toutes les permutations possibles. Le critère ainsi obtenu, que nous notons *CE* (pour ***Classification Error***), s'écrit de la façon suivante :

$$CE(\boldsymbol{z}, \boldsymbol{z}') = 1 - \max_{\sigma \in \mathfrak{S}(\{1,\ldots,K\})} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{i,\sigma(k)} z'_{i,k}$$

où $\mathfrak{S}(\{1, \ldots, K\})$ est l'ensemble des permutations sur l'ensemble $\{1, \ldots, K\}$. Dans notre article Robert et al. (2021), nous montrons que ce critère possède deux inconvénients. Le premier est qu'il vaut bien 0 pour deux partitions identiques au label switching près mais la borne supérieure vaut $(K-1)/K$ ce qui est strictement plus petit que 1 et dépend de K et pose des problèmes de comparaisons de qualités lorsque les nombres de clusters sont très différents. Le second est que l'ensemble des permutations $\mathfrak{S}(\{1, \ldots, K\})$

possède un cardinal de taille $K!$ et que le calcul est numériquement impossible dès que nous avons plus de 10 clusters. Pour contourner ces problèmes, nous proposons d'une part de prendre le critère *NCE* (pour ***Normalized Classification Error***) normalisé par $(K-1)/K$:

$$NCE(\boldsymbol{z}, \boldsymbol{z}') = \frac{K \times CE(\boldsymbol{z}, \boldsymbol{z}')}{K-1} \tag{2.8}$$

et, d'autre part, de remarquer que le critère *CE* cherche une permutation des lignes et des colonnes de la matrice $\mathbf{Cont}_{k,k'}^{(\boldsymbol{z},\boldsymbol{z}')}$ définie en (2.7) afin d'obtenir la somme la plus élevée sur la diagonale. Or c'est un problème d'affectation étudié en optimisation et il existe des méthodes pour trouver des solutions en temps polynomial ; par exemple, Kuhn (1955) propose la méthode hongroise avec une complexité en $\mathcal{O}\left(K^4\right)$. Une implémentation de ce critère est disponible dans le package `bikm1` (voir Robert (2021)) du logiciel R.

Nous généralisons également le critère *CE* (pour ***Classification Error***) étudié par Lomet et al. (2012) en le normalisant par $1/K + 1/L - 1/(KL)$ afin d'obtenir le critère *NCE* (pour ***Normalized Classification Error***). De plus, nous montrons dans notre article que ce problème peut être résolu en utilisant l'algorithme proposé par Kuhn (1955) et ainsi avoir une complexité de $\mathcal{O}\left(\max\left(K^4, K'^4, n, L^4, L'^4, J\right)\right)$ (voir l'appendice A.4 de notre article) contre $\mathcal{O}\left(\max\left(K!K'n, L!L'J\right)\right)$ lorsque nous utilisons l'ensemble des permutations pour le calcul comme c'était fait avant.

Information mutuelle : Le dernier critère est celui de l'**information mutuelle** ou *MI* (voir Linfoot (1957) par exemple) dont le principe est de comparer la divergence de Kullback-Leibler entre la distribution empirique de la loi jointe et le produit des distributions empiriques. Autrement dit, pour deux partitions $\boldsymbol{z}$ et $\boldsymbol{z}'$ respectivement sur $\{1, \ldots, K\}$ et $\{1, \ldots, K'\}$, nous avons :

$$MI(\boldsymbol{z}, \boldsymbol{z}') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} \frac{\mathbf{Cont}_{k,k'}^{(\boldsymbol{z},\boldsymbol{z}')}}{n} \log\left[\frac{n\mathbf{Cont}_{k,k'}^{(\boldsymbol{z},\boldsymbol{z}')}}{z_{+,k} z'_{+,k'}}\right]$$

avec la convention que $0 \log 0$ vaut 0. Le critère vaut 0 si les deux partitions sont égales au label switching près et la valeur maximale est :

$$\max\left[-\sum_{k=1}^{K} \frac{z_{+,k}}{n} \log\left(\frac{z_{+,k}}{n}\right), -\sum_{k'=1}^{K'} \frac{z'_{+,k'}}{n} \log\left(\frac{z'_{+,k'}}{n}\right)\right].$$

Ainsi, Wyse et al. (2017) propose le critère d'**information mutuelle normalisée** ou *MI* en divisant par la borne maximale.

Pour le modèle des blocs latents, nous étudions le critère *ENMI* (pour ***Extended Normalized Mutual Information***) proposé par Wyse et al. (2017) qui fait la somme des critères *NMI* sur les partitions en lignes et en colonnes et nous proposons le critère *coNMI* (pour ***Coclustering Normalized Mutual Information***) qui porte directement sur la classification jointe plutôt que sur la somme des deux. Nous montrons dans notre article qu'au final, la différence est uniquement sur la normalisation.

Comparaisons des critères dans le cadre du modèle des blocs latents : Nous comparons ensuite ces cinq critères sur cinq objectifs (voir le tableau résumé 2.1) :

1. Le critère renvoie des résultats proportionnels au taux de cases mal classées.
2. Le résultat du critère reste identique si nous inversons les deux partitions.
3. Le critère renvoie 1 lorsque les deux classifications sont identiques et 0 lorsqu'elles jugent que c'est le pire cas.
4. Le critère renvoie le même résultat même si on permute les numéros des clusters.
5. Le critère doit être calculable en un temps raisonnable.

TABLE 2.1 – Évaluation du comportement des critères par rapport aux cinq propriétés souhaitées. ✓ signifie que le critère est validé, tandis qu'un symbole ✗ signifie que le critère n'est pas rempli (ou qu'il est le moins efficace par rapport aux autres critères).

	CARI	CE (Lomet et al., 2012)	NCE	ENMI (Wyse et al., 2017)	coNMI
1. Proportions de cellules mal classées	Modérément	✓	✓	✗	✗
2. Symétrie	✓	✓	✓	✓	✓
3.1 Limite max est 1 pour deux classifications identiques	✓	Si nous utilisons $1 - \text{CE}$	✓	✓	✓
3.2 La valeur min est 0	Asymptotiquement	✗	✓	✓	✓
4. Label switching	✓	✓	✓	✓	✓
5. Temps d'exécution	Linéaire avec n, J et le nombre de blocs	Impossible dès que $\max(K, L) > 9$	Quadratique avec le nombre de classes et linéaire avec n et J	✓	✓

Cet article permet donc de choisir le critère que veut l'utilisateur suivant ses objectifs. A titre personnel, je recommande d'utiliser *NCE* tant qu'il n'y a pas trop de classes puis *CARI* sinon car je préfère une erreur proportionnelle aux nombres de cases mal classées.

Par exemple, toujours dans notre article, nous avons comparé la dispersion des erreurs par rapport à la proportion de cases mal classées dans un cas avec deux classes en lignes et en colonnes parfaitement équilibrées et en faisant varier la proportion de classes mal classées suivant la configuration suivante (voir la figure 2.9 pour les résultats) :

$$\left(\mathbf{Cont}_{k,k'}^{(\boldsymbol{z}_1,\boldsymbol{z}_2)}\right)_{1\leq k,k'\leq K} = n\begin{pmatrix} \frac{1-x}{2} & \frac{p}{2} \\ \frac{p}{2} & \frac{1-p}{2} \end{pmatrix} \text{ et } \left(\mathbf{Cont}_{\ell,\ell'}^{(\boldsymbol{w}_1,\boldsymbol{w}_2)}\right)_{1\leq \ell,\ell'\leq L} = J\begin{pmatrix} \frac{1-p'}{2} & \frac{p'}{2} \\ \frac{p'}{2} & \frac{1-p'}{2} \end{pmatrix} \tag{2.9}$$

en faisant varier p et p' dans $[0; 1/2]$.

Nous voyons sur la figure 2.9 que pour une même valeur d'erreur (par exemple 0.5), la proportion de cases mal classées par les critères *ENMI* et *coNMI* va de 0.5 à 0.8. À l'opposé, les critères *1-CE* et *NCE* suivent une ligne car ils ont été construits dans cet objectif mais peuvent être longs à calculer. Enfin, le critère *CARI* permet un compromis entre un calcul réalisable en un temps raisonnable et une erreur proche de la proportionnalité.

Remarque

Pour la comparaison des paramètres, il faudrait également comparer toutes les permutations possibles mais, comme expliqué précédemment, cela prendrait trop de temps dès que nous avons plus de 10 classes. Une solution, que je préconise, si les paramètres sont associés à des partitions (par exemple en utilisant le maximum a posteriori) est d'utiliser la permutation obtenue dans le calcul du critère *NCE* de l'équation (2.8) et de l'appliquer pour la comparaison des paramètres.

2.3 Comportement singulier de la vraisemblance

Nous avons vu dans la section 2.2 que le calcul de la vraisemblance (2.6) nécessite la somme de toutes les partitions possibles. Or, à cause de la double asymptotique du modèle des blocs latents (à la fois en lignes et en colonnes), nous montrons dans Brault et al. (2020) que toutes les partitions qui ne sont pas celle qui sert à simuler les données [6] sont négligeables dans le calcul de la vraisemblance. Concrètement,

6. Sauf dans de très rares cas où il existe des permutations possibles des labels

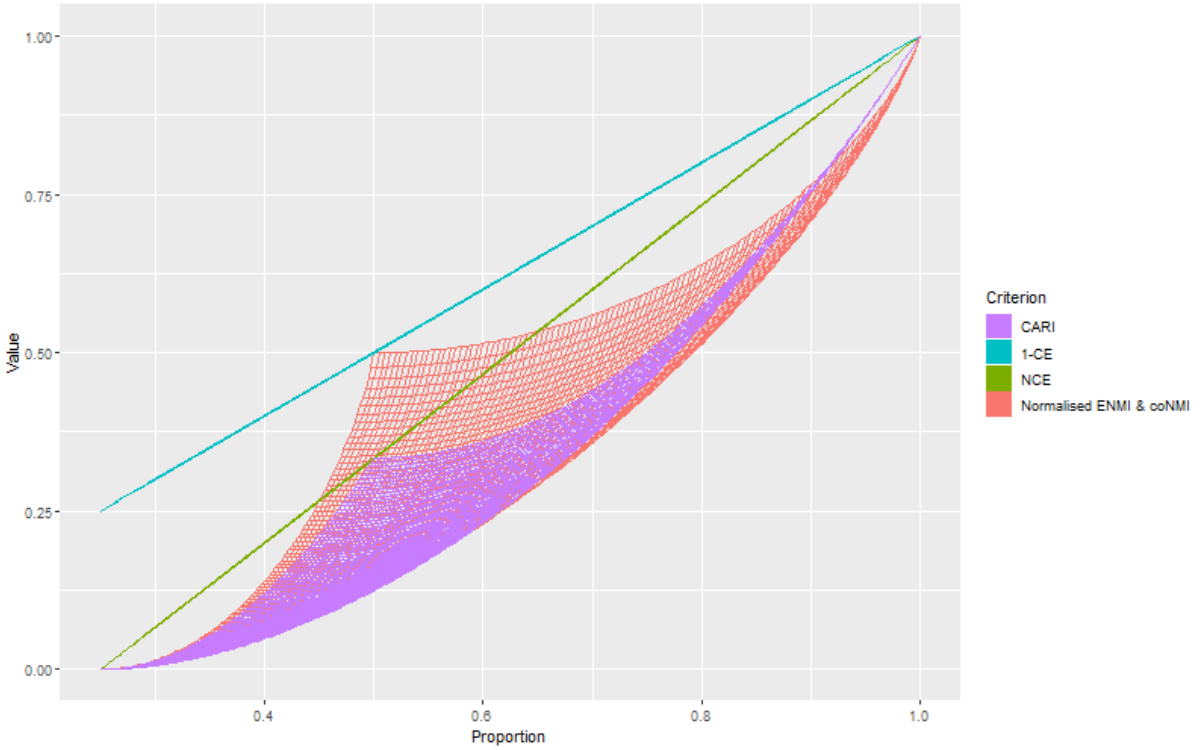

FIGURE 2.9 – Évolution de *CARI* (en violet), *1-CE* (en cyan), *NCE* (en vert) et *ENMI* divisée par 2, qui est égale à *coNMI* (en saumon) dans la configuration (2.9), en fonction de la proportion de cases bien classées pour la configuration (2.9) et pour $(n, J) = (100000, 100000)$.

nous montrons le théorème suivant dans le cas d'un modèle des blocs latents paramétriques avec une loi appartenant à la famille exponentielle univariée :

Théorème 1 (Brault et al. (2020))

Étant donnée la partition $(\boldsymbol{z}^\star, \boldsymbol{w}^\star)$ qui a servi à générer les données à l'aide du paramètre $\boldsymbol{\Psi}^\star$, sous certaines conditions de régularité du modèle généralisant les hypothèses de l'identifiabilité du modèle (voir la section 2.2) et sous la condition asymptotique suivante :

$$\frac{\log J}{n} \underset{n,J\to+\infty}{\longrightarrow} 0 \text{ et } \frac{\log n}{J} \underset{n,J\to+\infty}{\longrightarrow} 0 \qquad (\mathcal{H}^{\text{LBM}}_{\text{Asymp}})$$

alors pour tout paramètre $\boldsymbol{\Psi}$ ne possédant pas de symétries, nous avons la relation suivante :

$$\frac{p(\boldsymbol{y};\boldsymbol{\Psi})}{p(\boldsymbol{y};\boldsymbol{\Psi}^\star)} = \max_{\boldsymbol{\Psi}'\sim\boldsymbol{\Psi}} \frac{p(\boldsymbol{y},\boldsymbol{z}^\star,\boldsymbol{w}^\star;\boldsymbol{\Psi}')}{p(\boldsymbol{y},\boldsymbol{z}^\star,\boldsymbol{w}^\star;\boldsymbol{\Psi}^\star)}\left[1+o_P(1)\right]+o_P(1)$$

où $\boldsymbol{\Psi}'$ est égal à $\boldsymbol{\Psi}$ à une permutation des labels près et le o_P est uniforme sur l'ensemble des paramètres.

Perspectives

Je suis actuellement en train de montrer que ce résultat se généralise dès qu'il y a une asymptotique sur le nombre d'observations par individu. Nous retrouvons notamment ce résultat dans le cadre de l'article Brault et al. (2024a) détaillé dans la section 4.1.

2.3.1 Consistance des estimateurs

Le premier corollaire du théorème 1 est que l'estimateur du maximum de vraisemblance a un comportement similaire à l'estimateur du maximum de la vraisemblance complétée avec les partitions $(\boldsymbol{z}^\star, \boldsymbol{w}^\star)$

qui ont servi à générer les données. Or, nous montrons dans notre article la proposition suivante :

Proposition 2 (Brault et al. (2020))

Si nous notons $\widehat{\boldsymbol{\Psi}}(\boldsymbol{Y}, \boldsymbol{z}^\star, \boldsymbol{w}^\star)$ l'estimateur de $\boldsymbol{\Psi}^\star$ du maximum de vraisemblance complétée avec les partitions $(\boldsymbol{z}^\star, \boldsymbol{w}^\star)$ ayant simulé les données, $\Sigma_{\boldsymbol{\pi}^\star} = \text{Diag}(\boldsymbol{\pi}^\star) - \boldsymbol{\pi}^\star (\boldsymbol{\pi}^\star)^T$ et $\Sigma_{\boldsymbol{\rho}^\star} = \text{Diag}(\boldsymbol{\rho}^\star) - \boldsymbol{\rho}^\star (\boldsymbol{\rho}^\star)^T$ les deux matrices semis-définies positives de rangs $K-1$ et $L-1$, $[V(\boldsymbol{\theta}^\star)]_{k,\ell}$ la matrice de variance associée à la loi paramétrique du modèle et $\Sigma_{\boldsymbol{\theta}^\star} = \text{Diag}^{-1}(\boldsymbol{\pi}^\star) V(\boldsymbol{\theta}^\star) \text{Diag}^{-1}(\boldsymbol{\rho}^\star)$ alors nous avons :

$$\sqrt{n}\left[\widehat{\boldsymbol{\pi}}(\boldsymbol{z}^\star) - \boldsymbol{\pi}^\star\right] \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(\mathbb{0}_{K\times 1}, \Sigma_{\boldsymbol{\pi}^\star}), \qquad \sqrt{J}\left[\widehat{\boldsymbol{\rho}}(\boldsymbol{w}^\star) - \boldsymbol{\rho}^\star\right] \xrightarrow[J\to+\infty]{\mathcal{L}} \mathcal{N}(\mathbb{0}_{L\times 1}, \Sigma_{\boldsymbol{\rho}^\star})$$

$$\text{et } \forall (k,\ell)\ \sqrt{nJ}\left[\widehat{\theta}_{k,\ell}(\boldsymbol{Y}, \boldsymbol{z}^\star, \boldsymbol{w}^\star) - \theta^\star_{k,\ell}\right] \xrightarrow[n,J\to+\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma_{\boldsymbol{\theta}^\star,k,\ell})$$

Et le corollaire suivant :

Corollaire 3 (Consistance de l'EMV (Brault et al., 2020))

Si nous notons $\widehat{\boldsymbol{\Psi}}_{\text{EMV}}$ l'estimateur du maximum de vraisemblance alors il existe une permutation $\sigma_{\boldsymbol{z}}$ de $\{1,\dots,K\}$ et une permutation $\sigma_{\boldsymbol{w}}$ de $\{1,\dots,L\}$ telles que :

$$\widehat{\boldsymbol{\pi}}(\boldsymbol{z}^\star) - \widehat{\boldsymbol{\pi}}^{\sigma_{\boldsymbol{z}}}_{\text{EMV}} = o_P\left(n^{-1/2}\right), \qquad \widehat{\boldsymbol{\rho}}(\boldsymbol{w}^\star) - \widehat{\boldsymbol{\rho}}^{\sigma_{\boldsymbol{w}}}_{\text{EMV}} = o_P\left(J^{-1/2}\right)$$
$$\text{et } \widehat{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{z}^\star, \boldsymbol{w}^\star) - \widehat{\boldsymbol{\theta}}^{\sigma_{\boldsymbol{z}},\sigma_{\boldsymbol{w}}}_{\text{EMV}} = o_P\left[(nJ)^{-1/2}\right]$$

où $\widehat{\boldsymbol{\pi}}^{\sigma_{\boldsymbol{z}}}_{\text{EMV}}$, $\widehat{\boldsymbol{\rho}}^{\sigma_{\boldsymbol{w}}}_{\text{EMV}}$ et $\widehat{\boldsymbol{\theta}}^{\sigma_{\boldsymbol{z}},\sigma_{\boldsymbol{w}}}_{\text{EMV}}$ sont les estimateurs du maximum de vraisemblance dont les labels ont été permutés par les permutations $\sigma_{\boldsymbol{z}}$ et $\sigma_{\boldsymbol{w}}$.

De même, en adaptant légèrement la démonstration, nous montrons le même comportement pour l'estimateur du maximum de l'énergie libre (voir par exemple l'annexe B.2.1) :

Théorème 4 (Consistance de l'estimateur variationnel (Brault et al., 2020))

Si nous notons $\widehat{\boldsymbol{\Psi}}_{\text{VAR}}$ l'estimateur variationnel obtenu en maximisant l'énergie libre sur l'espace des familles ayant la contrainte en champs moyen alors il existe une permutation $\sigma_{\boldsymbol{z}}$ de $\{1,\dots,K\}$ et une permutation $\sigma_{\boldsymbol{w}}$ de $\{1,\dots,L\}$ telles que :

$$\widehat{\boldsymbol{\pi}}(\boldsymbol{z}^\star) - \widehat{\boldsymbol{\pi}}^{\sigma_{\boldsymbol{z}}}_{\text{VAR}} = o_P\left(n^{-1/2}\right), \qquad \widehat{\boldsymbol{\rho}}(\boldsymbol{w}^\star) - \widehat{\boldsymbol{\rho}}^{\sigma_{\boldsymbol{w}}}_{\text{VAR}} = o_P\left(J^{-1/2}\right)$$
$$\text{et } \widehat{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{z}^\star, \boldsymbol{w}^\star) - \widehat{\boldsymbol{\theta}}^{\sigma_{\boldsymbol{z}},\sigma_{\boldsymbol{w}}}_{\text{VAR}} = o_P\left[(nJ)^{-1/2}\right]$$

où $\widehat{\boldsymbol{\pi}}^{\sigma_{\boldsymbol{z}}}_{\text{VAR}}$, $\widehat{\boldsymbol{\rho}}^{\sigma_{\boldsymbol{w}}}_{\text{VAR}}$ et $\widehat{\boldsymbol{\theta}}^{\sigma_{\boldsymbol{z}},\sigma_{\boldsymbol{w}}}_{\text{VAR}}$ sont les estimateurs variationnels dont les labels ont été permutés par les permutations $\sigma_{\boldsymbol{z}}$ et $\sigma_{\boldsymbol{w}}$.

2.3.2 Algorithme *Largest Gaps* et estimation en *Big Data*

Une autre conséquence de ce comportement atypique est qu'à partir d'un certain nombre de données binaires (et dans certaines configurations), l'étude des marginales suffit pour estimer correctement les clusters en lignes et en colonnes. En effet, en adaptant le travail de Channarond et al. (2012) sur l'étude des marginales dans le cadre du *SBM*, nous proposons un algorithme dans Brault et Channarond (2023) qui est consistant dans l'estimation du nombre de classes, des partitions et des paramètres avec une complexité linéaire avec la taille de la matrice.

Le principe de l'algorithme *Largest Gaps* (ou *LG*) dans le cas de données binaires est le suivant (voir aussi la figure 2.10) :

1. Étant donnée une matrice de données binaire $\boldsymbol{y}$, nous prenons la moyenne des valeurs sur les colonnes (resp. les lignes) afin d'obtenir le vecteur $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$ (voir en haut à droite de la figure 2.10).

2. Si nous avons suffisamment d'observations, nous voyons les colonnes se regrouper par cluster (voir l'histogramme au milieu gauche de la figure 2.10). Du coup, nous réorganisons les valeurs du $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$ par ordre croissant (voir au milieu à droite de la figure 2.10). Nous faisons de même pour les moyennes sur les lignes.

3. Nous calculons ensuite les différences entre deux valeurs successives (voir en bas à gauche de la figure 2.10). Les valeurs des plus hauts sauts correspondent aux écarts entre la plus grande valeur d'un cluster et la plus petite du cluster suivant (définis par ordre des centres). Nous faisons pareil pour les lignes.

4. En choisissant un seuil pertinent séparant les clusters (en bas à gauche de la figure 2.10, nous pouvons prendre une valeur entre 0.02 et 0.04 par exemple), nous obtenons des clusters et pouvons réorganiser la matrice (voir en bas à droite de la figure 2.10).

Dans Brault et Channarond (2023), nous montrons que si nous notons $\equiv_{\mathcal{Z}}$ (resp. $\equiv_{\mathcal{W}}$) l'opérateur disant que deux partitions sont égales à une permutation près et si nous prenons la distance d^{∞} définie pour deux paramètres $\boldsymbol{\Psi}$ (avec (K,L) classes) et $\boldsymbol{\Psi}'$ (avec (K',L') classes) par

$$d^{\infty}\left(\boldsymbol{\Psi},\boldsymbol{\Psi}'\right)=\begin{cases}\max\left\{\left\|\boldsymbol{\pi}-\boldsymbol{\pi}'\right\|_{\infty},\left\|\boldsymbol{\rho}-\boldsymbol{\rho}'\right\|_{\infty},\left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|_{\infty}\right\} & \text{si } K=K',\ L=L'\\ +\infty \text{ sinon,}\end{cases}$$

où $\|\cdot\|_{\infty}$ est la norme infini définie pour tout $\boldsymbol{x}\in\mathbb{R}^{K}$ par $\|\boldsymbol{x}\|_{\infty}=\max_{1\leq k\leq K}|x_k|$ alors nous avons le résultat suivant :

Théorème 5 (Consistance des estimateurs de *LG* (Brault et Channarond, 2023))

Sous les conditions suffisantes d'identifiabilité proposées par Keribin et al. (2014) et rappelées en annexe B.1.1 et si nous prenons deux suites de seuils $S_K^{n,J}$ et $S_L^{n,J}$ strictement positifs de séparation des clusters vérifiant :

$$\begin{gathered}S_K^{n,J}\underset{n,J\to+\infty}{\longrightarrow}0,\quad S_L^{n,J}\underset{n,J\to+\infty}{\longrightarrow}0,\\ \varliminf_{n,J\to+\infty}S_K^{n,J}\sqrt{\frac{J}{\log n}}>\sqrt{2}\quad\text{and}\quad\varliminf_{n,J\to+\infty}S_L^{n,J}\sqrt{\frac{n}{\log J}}>\sqrt{2},\end{gathered}\tag{$\mathcal{H}_{\mathrm{LG}}$}$$

alors les estimateurs $\widehat{K}_{\mathrm{LG}}$ et $\widehat{L}_{\mathrm{LG}}$ du nombre de clusters, $\widehat{\boldsymbol{z}}_{\mathrm{LG}}$ et $\widehat{\boldsymbol{w}}_{\mathrm{LG}}$ des clusters et $\widehat{\boldsymbol{\Psi}}_{\mathrm{LG}}$ des paramètres proposés par l'algorithme *Largest Gaps* sont consistants ; c'est-à-dire que pour tout $t>0$, nous avons la convergence suivante :

$$\mathbb{P}\left(\widehat{K}_{\mathrm{LG}}\neq K^{\star}\text{ ou }\widehat{L}_{\mathrm{LG}}\neq L^{\star}\text{ ou }\widehat{\boldsymbol{z}}_{\mathrm{LG}}\not\equiv_{\mathcal{Z}}\boldsymbol{z}^{\star}\text{ ou }\widehat{\boldsymbol{w}}_{\mathrm{LG}}\not\equiv_{\mathcal{W}}\boldsymbol{w}^{\star}\text{ ou }d^{\infty}\left(\widehat{\boldsymbol{\Psi}}_{\mathrm{LG}}^{\sigma_{\mathcal{Z}},\sigma_{\mathcal{W}}},\boldsymbol{\Psi}^{\star}\right)>t\right)\underset{n,J\to+\infty}{\longrightarrow}0$$

où $\widehat{\boldsymbol{\Psi}}_{\mathrm{LG}}^{\sigma_{\mathcal{Z}},\sigma_{\mathcal{W}}}$ est la réorganisation des coordonnées du vecteur associée aux égalités $\widehat{\boldsymbol{z}}_{\mathrm{LG}}\equiv_{\mathcal{Z}}\boldsymbol{z}^{\star}$ et $\widehat{\boldsymbol{w}}_{\mathrm{LG}}\equiv_{\mathcal{W}}\boldsymbol{w}^{\star}$ c'est-à-dire la permutation telle que les clusters $\widehat{\boldsymbol{z}}_{\mathrm{LG}}$ et $\boldsymbol{z}^{\star}$ (resp. $\widehat{\boldsymbol{w}}_{\mathrm{LG}}$ et $\boldsymbol{w}^{\star}$) soient les plus ressemblants.

Notons que la condition ($\mathcal{H}_{\mathrm{LG}}$) implique la condition ($\mathcal{H}_{\mathrm{Asymp}}^{\mathrm{LBM}}$) du comportement asymptotique de la forme de la matrice nécessaire pour le théorème 1.

Remarque

Dans l'article Brault et Channarond (2023), nous montrons une inégalité de concentration pour avoir une vitesse de convergence dépendant de la distance entre deux clusters successifs et la probabilité d'appartenir au cluster avec le moins d'individus. Nous montrons des versions où les paramètres évoluent avec la taille des données.

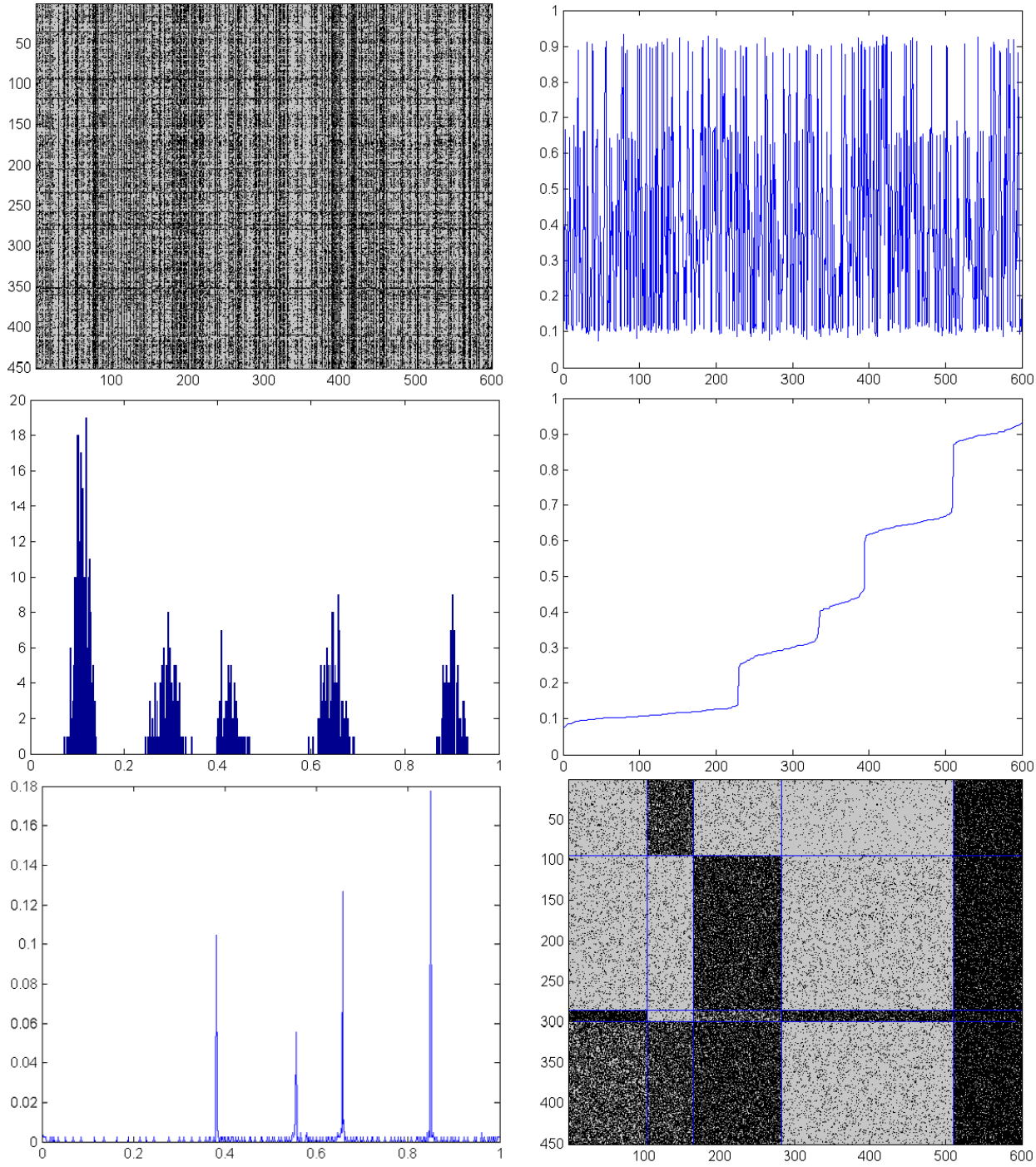

FIGURE 2.10 – En haut à gauche : matrice binaire initiale. En haut à droite : moyennes des cases $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$ valant 1 par colonnes. Milieu à gauche : l'organisation sous forme d'histogramme des moyennes $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$ pour mettre en évidence les clusters. Au milieu à droite : le vecteur réorganisé par ordre croissant des moyennes $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$. En bas à gauche : le calcul des différences entre deux valeurs successive des moyennes $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$ ordonnées par ordre croissant. En bas à droite : la réorganisation de la matrice après avoir choisi un seul judicieux pour les moyennes $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$ des colonnes et refait la même opération pour les lignes.

Perspectives

Actuellement, je suis en train de généraliser les résultats obtenus dans Brault et Channarond (2023) au cas de n'importe quelles lois ayant un moment d'ordre deux en incluant la possibilité d'avoir des données manquantes aléatoirement ce qui permettrait d'avoir un algorithme d'estimation non paramétrique. Nous proposons déjà des premiers résultats lors des 49$^{\text{èmes}}$ JdS (voir Brault et al. (2017a)).

Chapitre 3

Modèles de segmentation et bisegmentation

"Tu sais moi et les probabilités..."
Han Solo, *Star Wars, épisode V : L'Empire contre-attaque*

Le chapitre sur la bisegmentation commence avec mes travaux de mon post-doc passé à AgroParisTech en collaboration avec Céline Lévy-Leduc, Julien Chiquet, Maud Delattre, Emilie Lebarbier, Sarah Ouadah, Laure Sansonnet (tous les six à AgroParisTehc à l'époque) et Tristan Mary-Huard (INRAE/AgroParisTech). La plupart des modèles développés ici avaient pour origine de répondre aux questions des biologistes issues de la technologie *Hi-C* (voir Lieberman-Aiden et al. (2009)). Sans entrer dans les détails car je ne suis pas un spécialiste de cette technologie [1], le principe est d'essayer de comprendre la structure en trois dimensions de l'ADN afin de comprendre les domaines topologiques. L'objet de mes travaux de post-doc est la matrice fournie en sortie du protocole dont les lignes et les colonnes représentent les emplacements le long de l'ADN et une case représente les interactions entre ces deux emplacements (voir une version binarisée sur la figure 3.1) : plus une interaction est forte, plus les deux emplacements sont spatialement proches. Le but de ces études est de discrétiser les lignes (et les colonnes qui représentent la même chose) pour faire ressortir une structure et mieux comprendre les zones qui sont spatialement proches des autres.

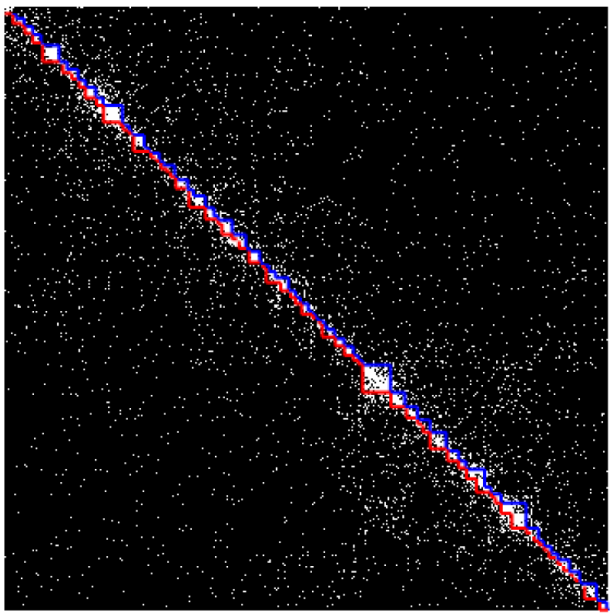

FIGURE 3.1 – Données *Hi-C* du chromosome 17 de cellules souches embryonnaires avec deux segmentations de blocs diagonaux (en bleu et en rouge). Image provenant de l'article Lévy-Leduc et al. (2014) qui fut à la base de mon post-doc.

Dans ce cadre, j'ai commencé par démontrer dans Brault et al. (2017c) le comportement particulier de sélection de modèle obtenu dans l'article Lévy-Leduc et al. (2014) d'où provient la figure 3.1 (voir la

1. Je renvoie les lectrices et lecteurs intéressé · e · s vers les sites de vulgarisation comme https://bioinfo-fr.net/hi-c-explication-des-bases

section 3.1). Ensuite, je présente et étudie dans Brault et al. (2017b) un modèle généralisant la bisegmentation proposée dans leur article (voir la section 3.2). L'inconvénient de cette méthode est qu'elle suppose que les observations sont la somme d'une moyenne et d'un bruit symétrique ce qui limite la modélisation donc j'enchaîne dans Brault et al. (2018c) avec la proposition et l'étude d'une modélisation non paramétrique (voir la section 3.3). Enfin, dans le prolongement de la modélisation 3.2, je termine ce chapitre avec l'article Brault et al. (2018b) où j'étudie la croissance des plants de Colza à l'aide d'une transformation du problème de segmentation par une méthode *LASSO*.

A nouveau, je présuppose que les bases de la segmentation (par exemple l'estimation par programmation dynamique) sont connues. J'encourage les personnes intéressée à lire l'annexe A.3 si ce n'est pas le cas.

3.1 Modèle des blocs diagonaux

Pour la première modélisation, nous supposons que la matrice $\boldsymbol{y}$ est symétrique donc nous nous intéressons qu'à la partie supérieure de la matrice. Le principe est de supposer qu'il existe $K^\star + 1$ blocs diagonaux où, au sein de chaque bloc, les observations suivent la même loi (éventuellement différente d'un bloc à l'autre) et les cases en dehors de la diagonale suivent également une même loi mais différente de chacune des lois des blocs centraux. Concrètement, cela signifie qu'il existe $K^\star$ ruptures $T_0^\star = 0 < T_1^\star < \cdots < T_{K^\star}^\star < T_{K^\star+1}^\star = n$ délimitant les blocs $D_k^\star$ (voir la figure 3.2) :

$$\forall k \in \{0, \ldots, K^\star\},\ D_k^\star = \left\{(i,j) \in \{1,\ldots,n\}^2 \,\middle|\, T_k^\star + 1 \leq i \leq j \leq T_{k+1}^\star\right\}. \tag{3.1}$$

De plus, nous notons $E_{\text{Bruit}}^\star$ la zone extérieure aux blocs diagonaux définie par l'équation (3.1) (voir la figure 3.2) :

$$E_{\text{Bruit}}^\star = \left\{(i,j) \in \{1,\ldots,n\}^2 \middle|\, 1 \leq i \leq j \leq n\right\} \setminus \left(\bigcup_{k=0}^{K^\star} D_k^\star\right).$$

Nous associons une moyenne $\theta_k^\star$ pour chaque zone $D_k^\star$ et $\theta_{\text{Bruit}}^\star$ pour la zone $E_{\text{Bruit}}^\star$ et nous supposons qu'il existe des variables $(\varepsilon_{i,j})_{1\leq i\leq j\leq n}$ indépendantes et identiquement distribuées telles que chaque observation $y_{i,j}$ est issue de la variable :

$$Y_{i,j} = \left\{ \begin{array}{ll} \theta_k^\star & \text{si } (i,j) \in D_k^\star \\ \theta_{\text{Bruit}}^\star & \text{sinon} \end{array} \right\} + \varepsilon_{i,j} \tag{3.2}$$

Dans leur article, Lévy-Leduc et al. (2014) estiment les emplacements des blocs à l'aide de l'algorithme de programmation dynamique (voir Bellman (1961)) par maximum de vraisemblance. Pour estimer les paramètres, nous définissons la fonction suivante en $(K, \boldsymbol{T}, \boldsymbol{\theta})$ en notant φ la densité de la loi $Y_{i,j}$ dépendant de la moyenne de la zone dans laquelle (i,j) se trouve (voir l'équation (3.2)) et du paramètre σ lié au bruit (voir le graphe bayésien en figure 3.3) :

$$\log p\left(\boldsymbol{Y}; K, \boldsymbol{T}, \boldsymbol{\theta}\right) = \sum_{k=0}^{K} \sum_{(i,j)\in D_k} \log \varphi\left(Y_{i,j}; \theta_k, \sigma\right) + \sum_{(i,j)\in E_{\text{Bruit}}} \log \varphi\left(Y_{i,j}; \theta_{\text{Bruit}}, \sigma\right). \tag{3.3}$$

Pour maximiser la vraisemblance, Lévy-Leduc et al. (2014) procèdent en deux étapes :

1. Étant donné un nombre $K_{\max}$ de ruptures, ils utilisent l'algorithme de programmation dynamique pour estimer les emplacements des ruptures et les paramètres associés aux régions (voir Bellman (1961)).
2. Ils choisissent le nombre de ruptures qui maximise la vraisemblance.

Remarque

En réalité, il n'est pas possible d'utiliser directement l'algorithme de programmation dynamique car la région E_{Bruit} est définie par complémentarité des unions des blocs D_k. Par conséquent, Lévy-Leduc et al. (2014) font d'abord une estimation des paramètres $\theta_{\text{Bruit}}^\star$ et σ dans la partie supérieure droite de la matrice (où la zone est censée être strictement incluse dans la région $E_{\text{Bruit}}^\star$) puis utilise l'algorithme de programmation dynamique pour les blocs. Cette précision ne change pas les résultats mais est faite pour une meilleure cohérence avec les articles Lévy-Leduc et al. (2014) et Brault et al. (2017c).

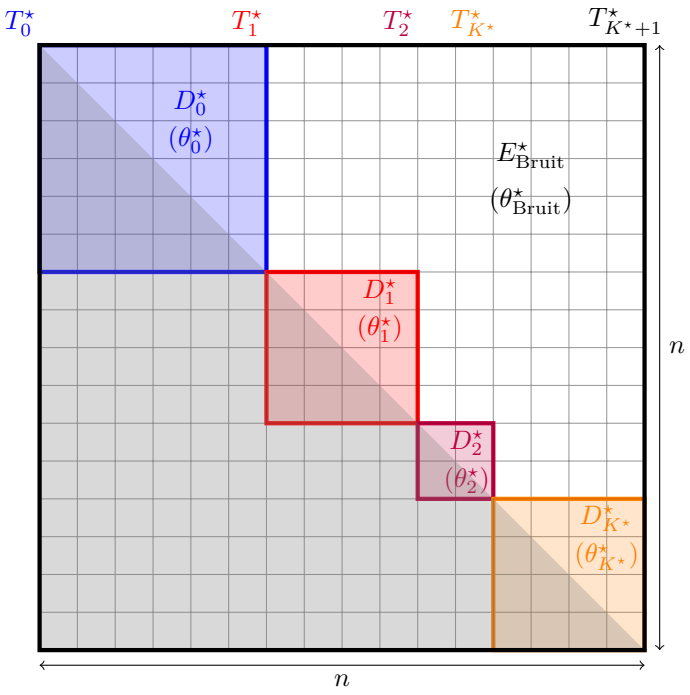

FIGURE 3.2 – Représentation schématique des notations utilisées pour le modèle (3.2) avec $n = 16$ lignes et $K^\star = 4$ blocs.

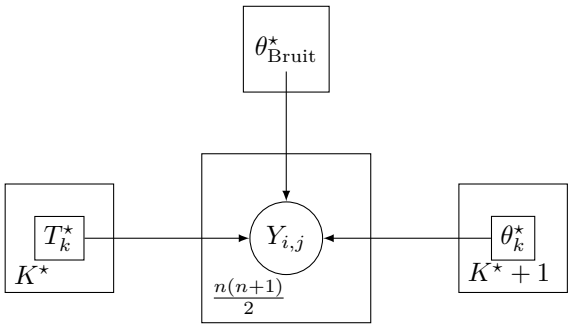

FIGURE 3.3 – Représentation graphique du modèle des blocs diagonaux : les carrés correspondent à des paramètres et les ronds à des variables.

Le choix fait par Lévy-Leduc et al. (2014) de sélectionner le nombre de ruptures par le maximum de vraisemblance peut surprendre de prime abord les personnes habituées à utiliser des critères de sélection de modèle. Nous montrons dans notre article Brault et al. (2017c) que ce choix est cohérent à cause du fait que les modèles ne sont pas imbriqués : ainsi, quand nous commençons à prendre trop de blocs, nous sommes obligés de *sacrifier* des cases qui auraient dû être présentes dans les blocs pour les mettre dans la zone E_{Bruit} ce qui n'est pas cohérent vis-à-vis de la vraisemblance (3.3) si les paramètres sont différents (voir les illustrations du *supplementary material* de notre article Brault et al. (2017c)). La suite de cette section consiste à expliquer les théorèmes présentés dans notre article Brault et al. (2017c). Pour cela, je commence par détailler les notations.

Notation (Brault et al. (2017c))

Nous notons $\boldsymbol{\tau}$ les ruptures $\boldsymbol{T}/n$ normalisées par la dernière valeur $T_{K+1} = n$; ainsi, elles se trouvent toutes dans l'intervalle $[0; 1]$. De façon similaire, nous supposons qu'il existe une suite $(\Delta_n)_{n\in\mathbb{N}^\star}$ une suite de valeurs strictement positives et nous notons $\left(\widehat{K}_n^{\Delta_n}, \widehat{\boldsymbol{\tau}}_{\widehat{K}_n^{\Delta_n}}^{\Delta_n}, \widehat{\boldsymbol{\theta}}_{\widehat{K}_n^{\Delta_n}}^{\Delta_n}\right)$ l'estimateur obtenu en maximisant la logvraisemblance (3.3) par la procédure proposée dans l'article et en admettant qu'il y ait une distance minimale de $n\Delta_n$ entre deux ruptures estimées.

Nous devons également faire quelques hypothèses :

Hypothèse (Brault et al. (2017c))

Nous commençons par supposer que le bruit $\varepsilon_{i,j}$ est sous-gaussien c'est-à-dire qu'il existe une constante $\beta > 0$ strictement positive telle que pour tout $v \in \mathbb{R}$ et tout $1 \leq i \leq j \leq n$, nous avons :

$$\mathbb{E}\left[e^{v\varepsilon_{i,j}}\right] \leq e^{\beta v^2}. \qquad (\mathcal{H}^{\text{BlocDiag}}_{\text{Sous-Gaussien}})$$

La distance minimale entre deux ruptures normalisées ayant servi à simuler les données est strictement positive :

$$0 < \Delta_{\boldsymbol{\tau}^\star} = \min_{0 \leq k \leq K^\star} \left(\tau^\star_{k+1} - \tau^\star_k\right). \qquad (\mathcal{H}^{\text{BlocDiag}}_{\Delta_{\boldsymbol{\tau}^\star}})$$

Les moyennes des blocs sont toutes différentes de la moyenne de la zone $E^\star_{\text{Bruit}}$:

$$0 < \min_{0 \leq k \leq K^\star} \left|\theta^\star_k - \theta^\star_{\text{Bruit}}\right|. \qquad (\mathcal{H}^{\text{BlocDiag}}_{\theta^\star_{\text{Bruit}}})$$

Dans notre article Brault et al. (2017c), nous obtenons des inégalités de concentration qui permettent d'obtenir le comportement des estimateurs :

Théorème 6 (Convergence de l'estimation du nombre de ruptures $\widehat{K}_n^{\Delta_n}$)

Sous l'hypothèse que nous avons une idée du nombre maximal de ruptures et que Δ_n vérifie les deux conditions suivantes à partir d'un certain moment :

$$\Delta_n < \Delta_{\boldsymbol{\tau}^\star} \text{ et } \Delta_n \frac{\sqrt{n}}{(\log n)^{1/4}} \xrightarrow[n\to+\infty]{} +\infty$$

alors sous les hypothèses $(\mathcal{H}^{\text{BlocDiag}}_{\text{Sous-Gaussien}})$, $(\mathcal{H}^{\text{BlocDiag}}_{\Delta_{\boldsymbol{\tau}^\star}})$ et $(\mathcal{H}^{\text{BlocDiag}}_{\theta^\star_{\text{Bruit}}})$, presque sûrement, nous avons :

$$\mathbb{P}\left(\widehat{K}_n^{\Delta_n} \neq K^\star\right) \xrightarrow[n\to+\infty]{} 0.$$

Remarque

En fait, les hypothèses sur la distance Δ_n ne sont nécessaires que pour le cas où l'estimation $\widehat{K}_n^{\Delta_n}$ est supérieure à $K^\star$ car la probabilité que $\widehat{K}_n^{\Delta_n}$ soit strictement plus petite que $K^\star$ tend vers 0 sans hypothèse sur Δ_n autre que d'être strictement positive ou, autrement dit, on peut prendre $\Delta_n = 1/n$.

De plus, nous avons la consistance des estimateurs des ruptures :

Théorème 7 (Consistance de l'estimateur des ruptures $\widehat{\boldsymbol{\tau}}^{\Delta_n}_{\widehat{K}_n^{\Delta_n}}$)

Sous les hypothèses du théorème 6, nous avons pour tout $\delta > 0$:

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\tau}}^{\Delta_n}_{\widehat{K}_n^{\Delta_n}} - \boldsymbol{\tau}^\star\right\|_{\text{Haus}} > \delta\right) \xrightarrow[n\to+\infty]{} 0$$

où $\|\cdot\|_{\text{Haus}}$ est la norme de Hausdorff définie pour un vecteur $\boldsymbol{\tau}$ de longueur K et un vecteur $\boldsymbol{\tau}'$ de longueur K' par :

$$\|\boldsymbol{\tau} - \boldsymbol{\tau}'\|_{\text{Haus}} = \max\left(\max_{0\leq k\leq K}\min_{0\leq k'\leq K'}\left|\tau_k - \tau'_{k'}\right|, \max_{0\leq k'\leq K'}\min_{0\leq k\leq K}\left|\tau_k - \tau'_{k'}\right|\right)$$

Remarque

De même, nous montrons que si nous connaissons le bon nombre $K^\star$ de ruptures alors l'estimateur $\widehat{\boldsymbol{\tau}}^{\Delta_n}_{K^\star,n}$ est consistant sans aucune hypothèse sur Δ_n autre que d'être strictement positive.

Littérature connexe

En travaillant sur l'analyse de l'écriture (voir la section 4.3), j'obtiens un résultat similaire de vraisemblance auto-pénalisante. L'article est en cours de rédaction.

Perspectives

En fait, à l'aide des inégalités de concentration obtenues, nous pouvons proposer des résultats en ayant des conditions moins restrictives sur les paramètres. Par exemple, $\Delta_{\boldsymbol{\tau}^\star}$ peut tendre vers 0 avec n sous condition que nous trouvions une suite $(\Delta_n)_{n\in\mathbb{N}^\star}$ qui continue de permettre à l'inégalité de concentration de tendre vers 0.

3.2 Modèle de bisegmentation gaussien

L'étude des blocs diagonaux est suffisante pour certaines données *Hi-C* mais, en discutant avec des biologistes, il peut exister des interactions à l'extérieur de la diagonale qu'il faudrait mettre en évidence car elles représentent des parties éloignées sur l'ADN qui seraient spatialement proches. Dans nos articles Brault et al. (2016, 2017b), nous proposons un modèle segmentant à la fois les n lignes et les J colonnes de la matrice $\boldsymbol{Y}$ pour faire ressortir une structure en quadrillage ; dans le cadre de ce modèle, les ruptures en lignes et en colonnes peuvent être différentes. Il existe donc deux ensembles de ruptures $\boldsymbol{T}^\star_1$ et $\boldsymbol{T}^\star_2$ tels que $0 = T^\star_{1,0} < T^\star_{1,1} < \cdots < T^\star_{1,K^\star} < T^\star_{1,K^\star+1} = n$ et $0 = T^\star_{2,0} < T^\star_{2,1} < \cdots < T^\star_{2,L^\star} < T^\star_{2,L^\star+1} = J$ et $(K^\star+1)\times(L^\star+1)$ lois paramétriques telles que les variables $Y_{i,j}$ soient indépendantes et vérifient (voir la figure 3.4 pour une représentation schématique et la figure 3.5 pour un graphe bayésien) :

$$\begin{aligned}&\forall (k,\ell)\in\{0,\dots,K^\star\}\times\{0,\cdots,L^\star\},\\ &\qquad \forall (i,j)\in\{T^\star_{1,k}+1,\dots,T^\star_{1,k+1}\}\times\{T^\star_{2,\ell}+1,\dots,T^\star_{2,\ell+1}\},\ Y_{i,j}\sim\mathcal{L}^\star_{k,\ell}\end{aligned} \tag{3.4}$$

Dans le cadre de ce modèle, les ruptures sont dépendantes les uns des autres et la programmation dynamique ne s'applique donc pas. Pour contourner ce problème, nous décidons de transformer le modèle en un problème de régression linéaire sparse afin d'utiliser des procédure *LASSO* (voir Brault et al. (2017b)). Pour cela, nous supposons que les lois sont gaussiennes de même variance et de moyenne $\theta^\star_{k,\ell}$ et le modèle (3.4) est donc équivalent à la modélisation :

$$\boldsymbol{Y} = \boldsymbol{U} + \boldsymbol{E} \tag{3.5}$$

où $\boldsymbol{U}$ est une matrice constante par blocs correspondants à chaque région du modèle (3.4) (voir la figure 3.4) et $\boldsymbol{E}$ est une matrice de variables aléatoires indépendantes et de même loi gaussienne centrée $\mathcal{N}\left(0,\sigma^2\right)$. Le premier réflexe que nous avons eu est d'utiliser différentes procédures *LASSO* (voir l'acte Brault et al. (2016)) comme le LASSO généralisé de Tibshirani (2011) implémenté dans le package `genlasso` par Arnold et Tibshirani (2022) et le *Fused LASSO Signal Approximator* (ou *FLSA*) de Hoefling (2010) implémenté dans le package `flsa` par Hoefling (2024) afin de trouver les zones des blocs. Toutefois, les frontières n'étaient pas nettes et nous proposons donc une autre approche.

Pour cela, nous remarquons que la matrice $\boldsymbol{U}$ est en bijection avec une matrice sparse $\boldsymbol{B}$ où les valeurs non nulles sont uniquement dans le coin supérieur gauche de chaque bloc par la formule suivante :

$$\boldsymbol{U} = \mathbb{T}^{(1)}_n \boldsymbol{B} {\mathbb{T}^{(1)}_J}^T \tag{3.6}$$

où $\mathbb{T}^{(1)}_n$ est la matrice triangulaire inférieure ne contenant que des 1 et nous avons la relation suivante :

$$\forall (k,\ell)\in\{0,\dots,K^\star\}\times\{0,\dots,L^\star\},\ \theta_{k,\ell} = \sum_{k'=0}^{k}\sum_{\ell'=0}^{\ell} \boldsymbol{B}_{T^\star_{1,k'}+1,T^\star_{2,\ell'}+1}. \tag{3.7}$$

Nous vectorisons l'équation (3.5) en incluant la relation (3.6) afin d'obtenir la relation suivante :

$$\mathrm{Vec}\left(\boldsymbol{Y}\right) = \mathbb{T}^{(1)}_J \otimes \mathbb{T}^{(1)}_n \underbrace{\mathrm{Vec}\left(\boldsymbol{B}\right)}_{=:\mathcal{B}} + \mathrm{Vec}\left(\boldsymbol{E}\right) \tag{3.8}$$

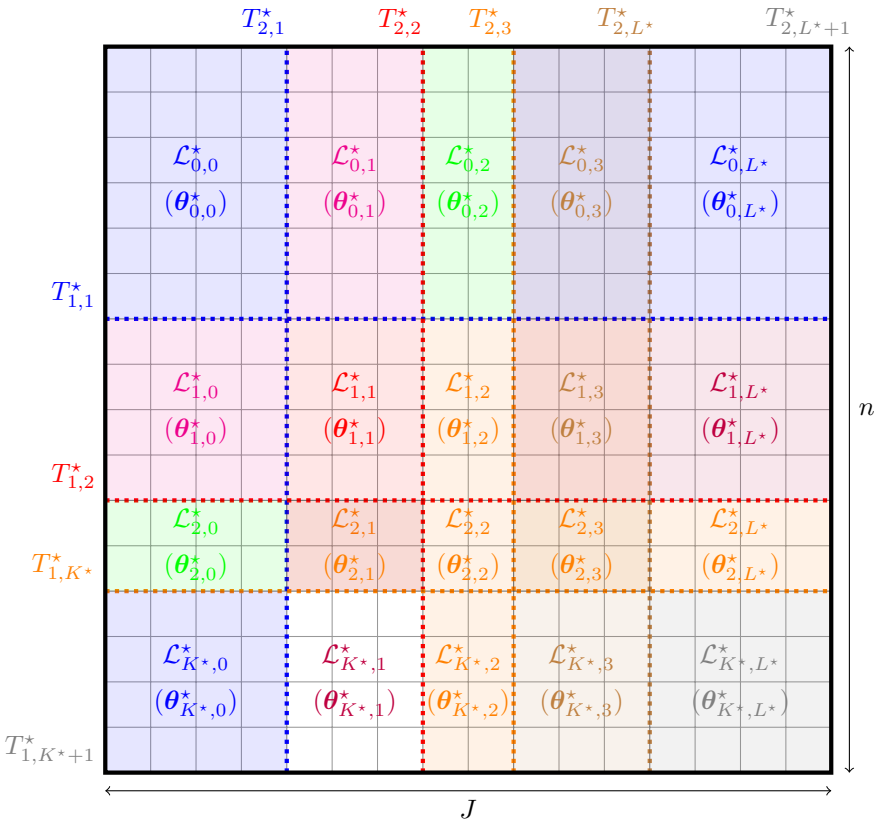

FIGURE 3.4 – Représentation schématique des notations du modèle (3.4) avec $n = 16$ lignes ayant $K^\star = 3$ ruptures et $J = 16$ colonnes avec $L^\star = 4$ ruptures. Chaque rectangle est caractérisé par une loi $\mathcal{L}^\star_{k,\ell}$ paramétrique ayant $\boldsymbol{\theta}^\star_{k,\ell}$ comme paramètre.

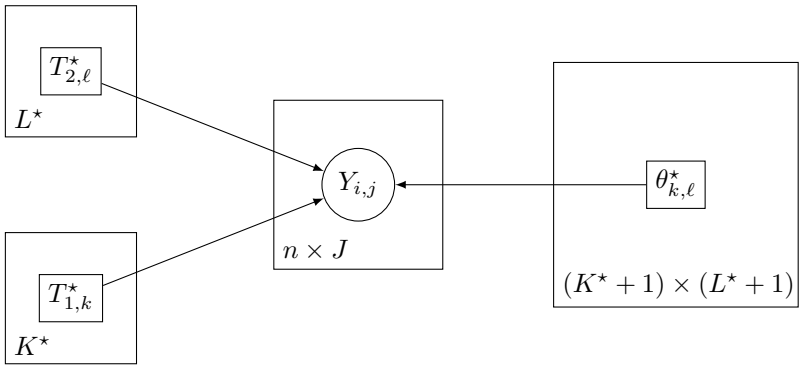

FIGURE 3.5 – Représentation graphique du modèle de bisegmentation gaussien : les carrés correspondent à des paramètres et les ronds à des variables.

avec $\otimes$ le produit de Kronecker et le vecteur $\mathcal{B}$ qui possède au plus $(K^\star+1)\times(L^\star+1)$ valeurs non nulles. Dans la suite de cette partie, nous précisons comment nous pouvons estimer les ruptures, améliorer la vitesse d'estimation et les résultats théoriques.

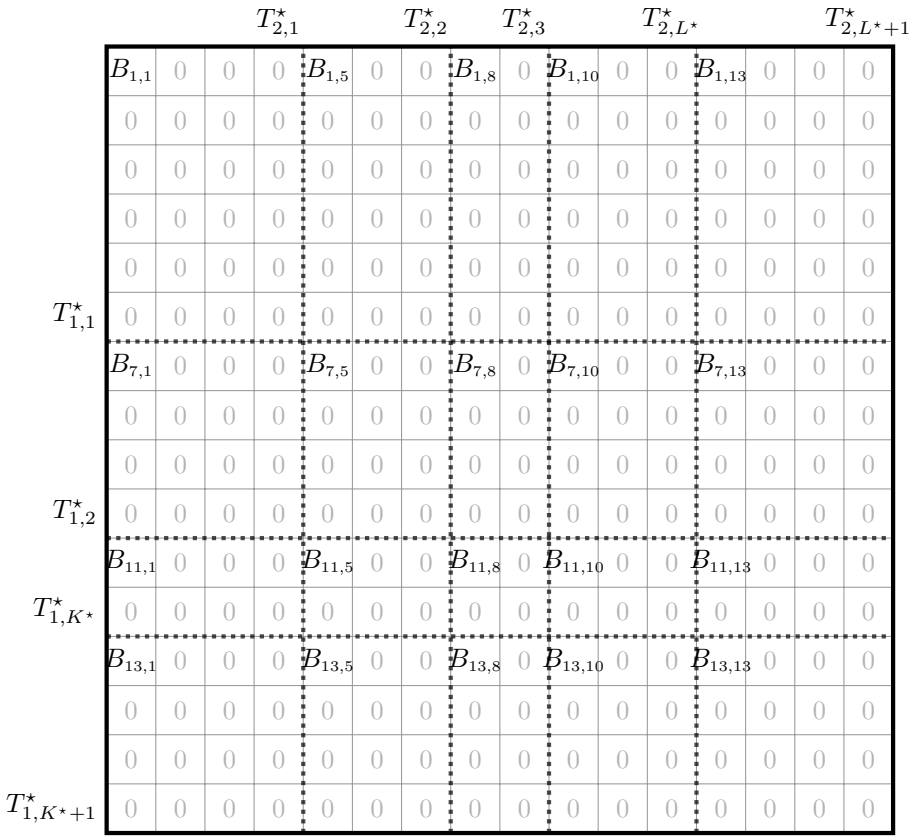

FIGURE 3.6 – Représentation de la matrice $\boldsymbol{B}$ associée à la matrice $\boldsymbol{U}$ de l'exemple de la figure 3.4 par la transformation proposée dans l'équation (3.6).

3.2.1 Estimation des emplacements de ruptures

Étant donnée une valeur $\lambda > 0$, nous notons $\widehat{\mathcal{B}}(\lambda)$ un estimateur de $\mathcal{B}$ et $\widehat{\mathcal{A}}(\lambda)$ le sous ensemble de $\{1, \dots, nJ\}$ des emplacements des valeurs non nulles de $\widehat{\mathcal{B}}(\lambda)$:

$$\widehat{\mathcal{A}}(\lambda) = \left\{ t \in \{1, \dots, nJ\} \,\middle|\, \left[\widehat{\mathcal{B}}(\lambda)\right]_t \neq 0 \right\}.$$

Or, nous savons que les emplacements des valeurs non nulles sont en haut à gauche de chaque région si nous remettons les valeurs sous forme de matrices donc pour estimer les ruptures, il faut chercher le quotient (pour les ruptures en colonnes) et le reste (pour les ruptures en lignes) de la division euclidienne des emplacements par le nombre de lignes. Autrement dit, si nous prenons un élément $a \in \widehat{\mathcal{A}}(\lambda)$ et que nous notons q_a le quotient et r_a le reste de la division euclidienne de $a-1$ par n ; c'est-à-dire que nous avons la relation $a - 1 = q_a n + r_a$. Ainsi, les estimations des emplacements des ruptures appartiennent aux ensembles suivants :

$$\widehat{\boldsymbol{T}}_1 \in \widehat{\mathcal{A}}_1(\lambda) = \left\{ r_a \,\middle|\, a \in \widehat{\mathcal{A}}(\lambda) \right\} \text{ et } \widehat{\boldsymbol{T}}_2 \in \widehat{\mathcal{A}}_2(\lambda) = \left\{ q_a \,\middle|\, a \in \widehat{\mathcal{A}}(\lambda) \right\} \tag{3.9}$$

de telle sorte que les ruptures soient strictement croissantes. Par exemple, étant donné une valeur λ, nous pouvons prendre toutes les valeurs des ensembles (3.9).

Pour sélectionner le nombre de ruptures, nous adaptons la méthode de **stabilité de la sélection** (ou *stability selection* en anglais) proposée par Meinshausen et Bühlmann (2010) consistant à faire plusieurs fois une estimation en enlevant la moitié des valeurs. Dans notre article Brault et al. (2017b), nous proposons d'exécuter plusieurs fois :

1. Enlever la moitié des lignes et la moitié des colonnes avant d'utiliser la procédure sur la matrice vectorisée.
2. Appliquer la méthode *LASSO* sur ce vecteur de taille $\left[\frac{nJ}{4}\right]$.

3. Pour chaque ligne et chaque colonne, compter combien il y a de variables actives qui sélectionnent ces lignes et ces colonnes.

À la fin, nous ne conservons que les lignes et les colonnes ayant été sélectionnées le plus grand nombre de fois.

Remarque

Dans un premier temps, nous avions essayé d'appliquer la méthode de Meinshausen et Bühlmann (2010) au pied de la lettre en ne conservant que le fait qu'une ligne ou une colonne ait été choisie comme rupture potentielle pour chaque matrice. Or, quand nous représentons les valeurs non nulles de l'estimation $\widehat{\mathcal{B}}(\lambda)$, nous pouvons voir que les faux positifs sont souvent quand même sur le bord gauche ou supérieur des rectangles même s'ils ne sont pas placés dans le coin [2]. En comptant plusieurs fois les ruptures dans l'étape 3, nous augmentons un peu les *bonnes* estimations même quand il y a des faux positifs.

Cette procédure est implémentée dans le package `blockseg` de Brault et Chiquet (2018) [3] et testée sur différents jeux de données. Dans notre article, nous simulons 100 matrices carrées de taille 500 avec des emplacements de ruptures à 100, 200, 300 et 400 et avons fait varier la variance (1, 2 et 5). Sur la figure 3.7 de l'article Brault et al. (2017b), chaque point gris correspond au nombre de valeurs actives choisies par notre procédure pour chaque ligne et chaque matrice et les points rouges les valeurs médianes pour chaque ligne.

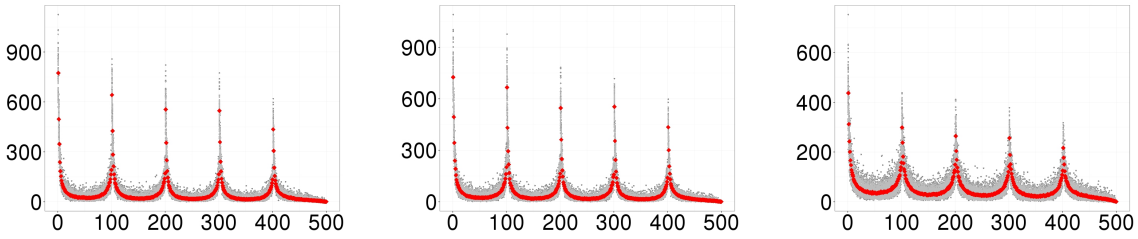

FIGURE 3.7 – Représentation en gris transparent de plusieurs simulations du nombre de valeurs actives associées à chaque ligne obtenues par la procédure adaptée de *stability selection* pour trois scénarios du plus simple à gauche ou plus compliqué à droite. Les points rouges correspondent aux valeurs médianes et les emplacements des vraies ruptures sont les multiples de 100. Image extraite de l'article Brault et al. (2017b).

Nous pouvons observer des pics très prononcés autour des valeurs correspondantes aux centaines d'autant plus prononcées que la difficulté est faible (à gauche). Dans le package `blockseg`, nous sélectionnons les ruptures les plus hautes suivant un pourcentage de la plus haute valeur (dans les simulations, 30% et 40% semblent donner les meilleurs résultats) en ne conservant qu'une seule valeur si plusieurs sont agrégées [4].

Perspectives

Quand nous regardons de plus près la qualité des estimations, nous voyons que les ruptures les moins bien estimées se situent à droite des colonnes et en bas des lignes ; je pense que c'est dû au fait qu'il y a moins de cases qui sont impactées et l'algorithme se concentre sur les ruptures en haut à gauche car elles améliorent plus le critère. Pour contourner ce problème, je travaille avec Jean-Charles Quinton (LJK) et Adeline Leclercq Samson (LJK) sur les véhicules autonomes (voir la section 3.4) et nous testons des procédures où nous décomposons la matrice en quatre blocs $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ de façon aléatoire et utilisons la procédure sur la matrice $\begin{pmatrix} D & C \\ B & A \end{pmatrix}$. En faisant cette manipulation plusieurs fois, les estimations sont alors lissées (voir Bardet et al. (2020)).

2. Voir par exemple la vidéo sur ma chaîne Youtube https://youtu.be/naT3Akl4Vqo?si=-O7X5EvfixpqFVhr
3. Le package n'est actuellement plus maintenu mais vous pouvez trouver une archive sur le CRAN : https://cran.r-project.org/src/contrib/Archive/blockseg/
4. Voir la vidéo sur ma chaîne pour une illustration de la méthode https://youtu.be/3Z97WZ5JSrk?si=g3tFwtAr00bRCuux

3.2.2 Accélération de l'estimation

Une fois que le problème est transformé dans une formulation *LASSO* comme sur l'équation (3.8), nous nous retrouvons avec la matrice carrée $\mathbb{T}_J^{(1)} \otimes \mathbb{T}_n^{(1)}$ de longueur nJ à utiliser. Or, si nous utilisons le produit matriciel ou pire l'inversion des matrices, le calcul devient très vite chronophage alors que la forme de la matrice est particulière. Nous cherchons donc à optimiser les temps de calculs par différentes astuces. Par exemple, nous avons la proposition suivante :

Proposition 8 (Calcul de $\mathbb{T}_J^{(1)} \otimes \mathbb{T}_n^{(1)} v$)
Pour tout vecteur $v \in \mathbb{R}^{nJ}$, le calcul de $\mathbb{T}_J^{(1)} \otimes \mathbb{T}_n^{(1)} v$ consiste à prendre la matrice associée à v et à faire des sommes cumulées sur les lignes et sur les colonnes. Par conséquent, nous n'avons besoin qu'au plus de $2nJ$ opérations.

Pour prouver cette proposition, il faut reprendre la forme de la matrice et faire le parallèle entre les deux calculs. De même, nous avons des calculs simplifiés pour chaque étape de la procédure *LASSO* (voir les lemmes 4, 5 et 6 de notre article Brault et al. (2017b)). Au final, nous avons la complexité suivante :

Théorème 9 (Complexité de `blockseg`)
Étant donnés A un nombre maximal de variables actives pour le vecteur $\mathcal{B}$, si nous notes m_{LASSO} le nombre d'itérations nécessaires pour obtenir ce nombre ou finir la procédure *LASSO*, l'algorithme implémenté dans `blockseg` a une complexité de $\mathcal{O}\left(\max\left(m_{LASSO} nJ, m_{LASSO} A^2\right)\right)$.

Pour comparaison, l'implémentation de la méthode *LARS* proposée par Bach et al. (2011) pour une matrice sparse quelconque possède une complexité de $\mathcal{O}\left(\max\left(m_{LASSO} n^2 J^2, m_{LASSO} A^2\right)\right)$. Nous avons d'ailleurs comparé les temps avec différents procédure dans l'article.

3.2.3 Consistance des estimateurs

Enfin, la dernière question est celle de la qualité des estimations. Pour cela, nous avons besoin de quelques notations :

Notation (Brault et al. (2017b))

Comme pour le modèle bloc diagonal de la section 3.1, nous avons besoin de connaître la distance minimale entre deux ruptures que nous définissons par :

$$\Delta_{\boldsymbol{\tau}^\star} = \min\left[\min_{0\leq k\leq K^\star}\left(\tau^\star_{1,k+1} - \tau^\star_{1,k}\right), \min_{0\leq \ell\leq L^\star}\left(\tau^\star_{2,\ell+1} - \tau^\star_{2,\ell}\right)\right].$$

Comme pour le résultat du modèle des blocs latents de la section 2.3.1, nous avons besoin de pouvoir différencier deux ruptures successives en ayant au moins un paramètre différent :

$$\Delta_{\boldsymbol{\theta}^\star} = \min\left[\min_{0\leq k\leq K^\star-1}\left(\max_{0\leq \ell\leq L^\star}\left|\theta^\star_{k,\ell} - \theta^\star_{k+1,\ell}\right|\right), \min_{0\leq \ell\leq L^\star-1}\left(\max_{0\leq k\leq K^\star}\left|\theta^\star_{k,\ell} - \theta^\star_{k,\ell+1}\right|\right)\right].$$

Autrement dit, pour chaque couple de colonnes successives (resp. lignes), nous regardons le plus grand écart entre deux valeurs qui permettrait de les différencier et nous prenons la plus petite de toutes ces valeurs pour nous guider sur la vitesse d'estimation.

Enfin, nous notons $\widehat{\boldsymbol{\tau}}_{n,J,1}$ (resp. $\widehat{\boldsymbol{\tau}}_{n,J,2}$) l'estimateur $\widehat{\boldsymbol{T}}_1$ (resp. $\widehat{\boldsymbol{T}}_2$) de la procédure `blockseg` (définie dans l'équation (3.9)) normalisé par n le nombre de lignes (resp. par J le nombre de colonnes) pour une matrice de taille $n \times J$.

À l'aide de ces notations, nous pouvons proposer les hypothèses suivantes :

Hypothèse (Brault et al. (2017b))

Nous commençons par supposer que le bruit $E_{i,j}$ est sous-gaussien c'est-à-dire qu'il existe une constante $\beta > 0$ strictement positive telle que pour tout $v \in \mathbb{R}$ et tout $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, J\}$, nous avons :

$$\mathbb{E}\left[e^{vE_{i,j}}\right] \leq e^{\beta v^2}. \qquad (\mathcal{H}^{\text{Blockseg}}_{\text{Sous-Gaussien}})$$

La suite $(\delta_{n,J})_{n\in\mathbb{N}^\star, J\in\mathbb{N}^\star}$ utilisée pour la consistance est décroissante, strictement positive et vérifie :

$$\delta_{n,J}\Delta^2_{\boldsymbol{\theta}^\star}\frac{n}{\log J} \xrightarrow[n,J\to+\infty]{} +\infty \text{ et } \delta_{n,J}\Delta^2_{\boldsymbol{\theta}^\star}\frac{J}{\log n} \xrightarrow[n,J\to+\infty]{} +\infty \qquad (\mathcal{H}^{\text{Blockseg}}_{\text{Asymp}})$$

La suite de pénalités $(\lambda_{n,J})_{n\in\mathbb{N}^\star, J\in\mathbb{N}^\star}$ utilisée pour la procédure *LASSO* est strictement positive et vérifie à partir d'un certain rang :

$$\left|\widehat{\mathcal{A}}_1(\lambda_{n,J})\right| = K^\star, \ \left|\widehat{\mathcal{A}}_2(\lambda_{n,J})\right| = L^\star, \text{ et } \frac{\delta_{n,J}\Delta_{\boldsymbol{\theta}^\star}}{\lambda_{n,J}}\min(n,J) \xrightarrow[n,J\to+\infty]{} +\infty \qquad (\mathcal{H}^{\text{Blockseg}}_{\text{LASSO}})$$

La distance minimale entre deux ruptures normalisées ayant servi à simuler les données ne peut pas tendre trop vite vers 0 :

$$\Delta_{\boldsymbol{\tau}^\star} \geq \delta_{n,J}. \qquad (\mathcal{H}^{\text{Blockseg}}_{\Delta_{\boldsymbol{\tau}^\star}})$$

Remarque

L'hypothèse $(\mathcal{H}^{\text{Blockseg}}_{\text{Asymp}})$ implique l'hypothèse $(\mathcal{H}^{\text{LBM}}_{\text{Asymp}})$ faite pour le modèle des blocs latents si les paramètres sont fixes et nous retrouvons donc le même comportement sur le rapport des longueurs.

Nous pouvons ainsi avoir la consistance des estimations des emplacements des ruptures :

Théorème 10 (Consistance des estimateurs des ruptures normalisées $\widehat{\boldsymbol{\tau}}_{n,J,1}$ et $\widehat{\boldsymbol{\tau}}_{n,J,2}$)

Sous les hypothèses $(\mathcal{H}^{\text{Blockseg}}_{\text{Sous-Gaussien}})$, $(\mathcal{H}^{\text{Blockseg}}_{\text{Asymp}})$, $(\mathcal{H}^{\text{Blockseg}}_{\text{LASSO}})$ et $(\mathcal{H}^{\text{Blockseg}}_{\Delta_{\boldsymbol{\tau}^\star}})$, nous avons la consistance des estimations des ruptures normalisées :

$$\mathbb{P}\left(\max\left(\|\widehat{\boldsymbol{\tau}}_{n,J,1} - \boldsymbol{\tau}^\star_1\|_{\text{Haus}}, \|\widehat{\boldsymbol{\tau}}_{n,J,2} - \boldsymbol{\tau}^\star_2\|_{\text{Haus}}\right) \geq \delta_{n,J}\right) \xrightarrow[n,J\to+\infty]{} 0$$

où $\|\cdot\|_{\text{Haus}}$ est la norme Hausdorff rappelée dans le théorème 7.

Dans la section 4 de notre article Brault et al. (2017b), nous testons l'algorithme `blockseg` sur différents scénarios et nous obtenons des courbes *ROC* (*Receiver Operating Characteristic*) avec des aires sous la courbe allant de 0.983 pour les cas les plus simples à 0.63 dans les cas les plus compliquées (où les méthodes concurrentes testées font moins bien que le hasard).

3.3 Modèle de bisegmentation non paramétrique

Le dernier travail que j'ai effectué sur les données *Hi-C* est parti du principe que pour utiliser le modèle de bisegmentation gaussien, il fallait avoir des données normalisées. Or, les données brutes sont des données de comptage et il est possible que nous perdions de l'information en les normalisant. Nous proposons donc dans notre article Brault et al. (2018c) un modèle **non paramétrique** qui permet d'étudier tous les types de matrices *Hi-C*.

Dans ce cadre, nous supposons que la matrice $\boldsymbol{Y}$ est symétrique et nous commençons par présenter le principe avec une seule rupture ; l'objectif étant d'adapter la statistique utilisée par Lung-Yut-Fong et al. (2015) qui étend la statistique de rang de Wilcoxon-Mann-Whitney. Nous supposons qu'il existe une rupture $T^\star_1$ en colonne séparant les lignes au même endroit et nous supposons que pour chaque ligne $i \in \{1, \dots, n\}$, les cases sont des observations issues de variables indépendantes de loi $\mathcal{L}^{(i)}_0$ avant la rupture et $\mathcal{L}^{(i)}_1$ après (voir la gauche de la figure 3.8). Nous supposons que la rupture $T^\star_1$ existe réellement

si et seulement s'il existe au moins une ligne i telle que la loi $\mathcal{L}_0^{(i)}$ est différente de la loi $\mathcal{L}_1^{(i)}$. En terme d'hypothèses de test, nous pouvons résumer ainsi :

$$\begin{array}{rcl}(\mathcal{H}_0) & : & \forall i \in \{1,\dots,n\},\ \mathcal{L}_0^{(i)} = \mathcal{L}_1^{(i)},\\ (\mathcal{H}_1) & : & \exists i \in \{1,\dots,n\},\ \mathcal{L}_0^{(i)} \neq \mathcal{L}_1^{(i)}.\end{array} \tag{3.10}$$

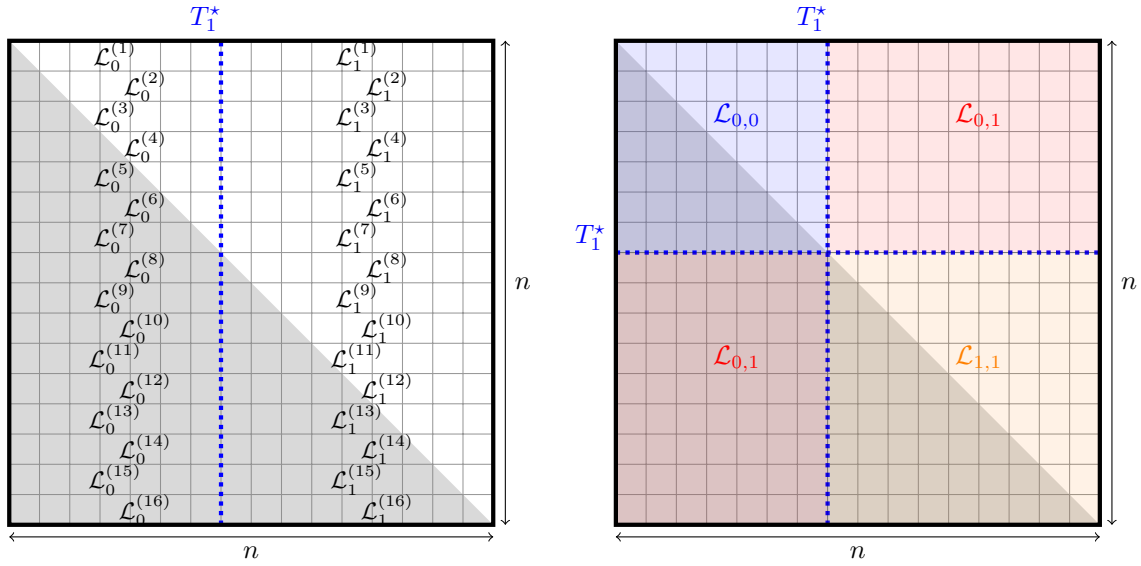

FIGURE 3.8 – Représentation schématique du modèle (3.10) avec les lois pour chaque ligne avant et après la rupture (à gauche) et les lois induisent par la symétrie (à droite) comme vues dans le modèle (3.11). Les zones grises sont mises pour rappelées que la matrice est symétrique.

Or, comme la matrice est symétrique, pour tout couple (i,j), la case $Y_{i,j}$ a la même loi que la case $Y_{j,i}$ et les hypothèses (3.10) se résument à trois lois $\mathcal{L}_{0,0}$, $\mathcal{L}_{0,1}$ et $\mathcal{L}_{1,1}$ pour les variables $Y_{i,j}$:

$$\left\{\begin{array}{ll} Y_{i,j} \sim \mathcal{L}_{0,0} & \text{si } 1 \leq i,j \leq T_1^\star,\\ Y_{i,j} \sim \mathcal{L}_{1,1} & \text{si } T_1^\star + 1 \leq i,j \leq n\\ \text{et } Y_{i,j} \sim \mathcal{L}_{0,1} & \text{sinon.}\end{array}\right. \tag{3.11}$$

Dans ce cadre, les hypothèses (3.10) peuvent être réécrites de la façon suivante :

$$\begin{array}{rcl}(\mathcal{H}_0) & : & \mathcal{L}_{0,0} = \mathcal{L}_{0,1} = \mathcal{L}_{1,1},\\ (\mathcal{H}_1) & : & \mathcal{L}_{0,0} \neq \mathcal{L}_{0,1} \text{ ou } \mathcal{L}_{0,1} \neq \mathcal{L}_{1,1}.\end{array}$$

Pour savoir s'il y a une rupture en $T_1^\star$ ou pas, nous utilisons la U **statistique** définie pour chaque ligne i par :

$$U_{n,i}\left(T_1^\star\right) = \frac{1}{\sqrt{nT_1^\star\left(n-T_1^\star\right)}} \sum_{j_0=1}^{T_1^\star} \sum_{j_1=T_1^\star+1}^{n} \left(\mathbb{1}_{\left\{Y_{i,j_0} \leq Y_{i,j_1}\right\}} - \mathbb{1}_{\left\{Y_{i,j_1} \leq Y_{i,j_0}\right\}}\right) \tag{3.12}$$

qui compare toutes les cases avant la rupture $T_1^\star$ à toutes les cases après la rupture et ajoute 1 lorsque les valeurs avant la rupture sont plus élevées et -1 sinon. Ainsi, si les lois sont identiques, la statistique $U_{n,i}\left(T_1^\star\right)$ sera proche de 0 alors que si les lois sont différentes, nous nous attendons à ce que la valeur soit loin de 0 négativement ou positivement. Afin de concaténer les informations de toutes les lignes, nous prenons la sommes des statistiques $U_{n,i}\left(T_1^\star\right)$ de l'équation (3.12) au carré :

$$S_n\left(T_1^\star\right) = \sum_{i=1}^{n} U_{n,i}^2\left(T_1^\star\right). \tag{3.13}$$

En pratique, nous ne connaissons pas l'emplacement de la *vraie* rupture $T_1^\star$ donc, dans notre article Brault et al. (2018c), nous proposons de l'estimer en maximisant la statistique $T \mapsto S_n(T)$ définie dans l'équation (3.13) :

$$\widehat{T}_1 \in \underset{1 \leq T_1 \leq n-1}{\operatorname{argmax}} S_n(T_1). \tag{3.14}$$

Dans le suite de cette partie, nous présentons l'extension au cas de K ruptures, montrons que l'estimateur obtenu est consistant lorsque nous connaissons le bon nombre $K^\star$ de ruptures et concluons par quelques pistes pour estimer le nombre de ruptures $K^\star$.

3.3.1 Extension à K ruptures

Afin de passer à l'extension à $K^\star$ ruptures quelconques (voir le graphe bayésien sur la figure 3.9), nous commençons par remarquer que pour tout $1 \leq T_1 \leq n-1$ la statistique $U_{n,i}(T_1)$ définie à l'équation (3.12) peut s'écrire de la façon suivante :

$$U_{n,i}(T_1) = \frac{2}{\sqrt{nT_1(n-T_1)}} \sum_{j_0=1}^{T_1} \left(\frac{n+1}{2} - R_{j_0}^{(i)} \right) \text{ où } R_{j_0}^{(i)} = \sum_{j'=1}^{n} \mathbb{1}_{\{Y_{i,j'} \leq Y_{i,j_0}\}} \tag{3.15}$$

c'est-à-dire en utilisant $R_{j_0}^{(i)}$ le rang de la case Y_{i,j_0} à l'intérieur de la ligne i. L'avantage de cette formulation est qu'il *suffit* de calculer les rangs de chaque case au sein de chaque ligne au lieu de comparer tous les couples deux par deux.

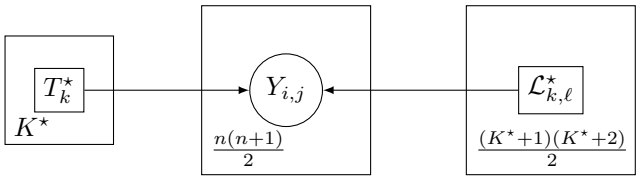

FIGURE 3.9 – Représentation graphique du modèle de bisegmentation non paramétrique : les carrés correspondent à des paramètres et les ronds à des variables.

Implémentations

En utilisant des procédures de type *quickselect* (voir par exemple Hoare (1961)), nous pouvons obtenir une complexité moyenne de $\mathcal{O}(n \log n)$ au lieu de $\mathcal{O}(n^2)$ même si cette dernière reste la complexité *dans le pire des cas*.

En adaptant les équations (3.13) et (3.15), nous obtenons pour tout ensemble de $K^\star$ ruptures $\boldsymbol{T}$ la formule suivante :

$$S_n(\boldsymbol{T}) = \frac{4}{n^2} \sum_{k=0}^{K^\star} (T_{k+1} - T_k) \sum_{i=1}^{n} \left(\overline{R}_k^{(i)} - \frac{n+1}{2} \right)^2 \text{ avec } \overline{R}_k^{(i)} = \frac{1}{T_{k+1} - T_k} \sum_{j=T_k+1}^{T_{k+1}} R_j^{(i)} \tag{3.16}$$

c'est-à-dire que $\overline{R}_k^{(i)}$ est le rang moyen des cases $Y_{i,j}$ dont les indices des colonnes sont comprises entre les ruptures T_k et T_{k+1}. Ainsi, pour estimer les emplacements des ruptures, connaissant leur nombre K, nous procédons de la même façon que pour l'équation (3.14) :

$$\widehat{\boldsymbol{T}}_n \in \underset{0=T_0<T_1<\cdots<T_K<T_{K+1}=n}{\operatorname{argmax}} S_n(\boldsymbol{T}).$$

Comme nous pouvons le voir dans l'équation (3.16), le calcul de la statistique $S_n(\boldsymbol{T})$ se décompose comme une somme de segments une fois la matrice $\left(\overline{R}_j^{(i)}\right)_{1 \leq i,j \leq n}$ des rangs moyens connue. Nous pouvons utiliser l'algorithme de programmation dynamique (voir Brault et al. (2018c) et Brault et al. (2018d)) pour estimer les ruptures ; il est implémenté dans le package `MuChPoint` du logiciel `R` (voir Brault et al. (2018a)).

Théorème 11 (Complexité de l'algorithme d'estimation implémenté dans `MuChPoint`)

L'algorithme d'estimation des emplacements de ruptures $\widehat{\boldsymbol{T}}_n$ par maximisation de la statistique $S_n(\boldsymbol{T})$ en utilisant l'algorithme de programmation dynamique implémenté dans `MuChPoint` est de complexité $\mathcal{O}(n^3)$.

L'essentiel de la complexité est dû au calcul de la matrice $\Delta(s:t)_{0\leq s<t\leq n}$ définie dans l'équation (A.5) dans le cadre de l'équation (3.16) :

$$\forall 0 \leq s < t \leq n,\ \Delta(s:t) = \sum_{i=1}^{n} \left(\overline{R}_{s:t}^{(i)} - \frac{n+1}{2} \right)^2 \text{ avec } \overline{R}_{s:t}^{(i)} = \frac{1}{t-s} \sum_{j=s+1}^{t} R_j^{(i)}.$$

Pour le moment, je n'ai pas trouvé de solutions pour descendre en dessous de $\mathcal{O}(n^3)$. En revanche, l'avantage est qu'une fois calculée, la matrice peut être stockée pour être réutilisée.

Perspectives

Pour ne pas avoir à calculer toutes les valeurs de la matrice $\Delta(s:t)_{0\leq s<t\leq n}$, je pense qu'il serait possible d'adapter le résultat obtenu par Rigaill (2015) mais je n'ai pas encore réussi à comprendre comment.

3.3.2 Consistance de l'estimateur des emplacements des ruptures connaissant le bon nombre de ruptures

Comme pour les modèles précédents, nous montrons que l'estimation proposée est consistante. Dans ce cadre là, il est intéressant de noter deux points :

- Le principe de la démonstration se rapproche de celle du modèle bloc diagonal de l'article Brault et al. (2017c).
- L'une des difficultés est le fait que les cases $Y_{i,j}$ ne sont pas toutes indépendantes car $Y_{i,j}$ et $Y_{j,i}$ ont la même loi. Néanmoins, cette dépendance reste finalement marginale dans nos démonstrations et sont négligeables lorsque n est suffisamment grand.

Pour démontrer la consistance, nous avons besoin de quelques notations :

Notation (Brault et al. (2018c))

Nous supposons que la matrice $\boldsymbol{Y}$ est symétrique, que toutes les variables $(Y_{i,j})_{1\leq i\leq j\leq n}$ sont indépendantes et qu'il existe $K^\star$ ruptures $\boldsymbol{T}^\star$ et $(K^\star+1)(K^\star+2)/2$ lois $(\mathcal{L}_{k,\ell})_{0\leq k\leq \ell\leq K^\star}$ de fonctions de répartitions $F_{k,\ell}$ telles que :

$$\forall 0 \leq k \leq \ell \leq K^\star, \forall i \in \left\{T_k^\star+1, \ldots, T_{k+1}^\star\right\}, \forall j \in \left\{\max(i, T_\ell^\star+1), \ldots, T_{\ell+1}^\star\right\},\ Y_{i,j} \sim \mathcal{L}_{k,\ell}.$$

Comme précédemment, nous notons $\widehat{\boldsymbol{\tau}}_n$ l'estimateur $\widehat{\boldsymbol{T}}_n$ renormalisé par n.

Dans le cadre du modèle `MuChPoint`, nous n'avons besoin que de deux hypothèses :

Hypothèse (Brault et al. (2018c))

Comme pour l'hypothèse ($\mathcal{H}^{\text{BlocDiag}}_{\Delta_{\boldsymbol{\tau}^\star}}$) du modèle bloc diagonal, nous avons besoin d'avoir assez de cases pour pouvoir voir qu'il y a une rupture. Nous supposons donc que la distance minimale entre deux ruptures normalisées ayant servi à simuler les données est strictement positive :

$$0 < \Delta_{\boldsymbol{\tau}^\star} = \min_{0\leq k\leq K^\star} \left(\tau_{k+1}^\star - \tau_k^\star\right). \qquad (\mathcal{H}^{\text{MuChPoint}}_{\Delta_{\boldsymbol{\tau}^\star}})$$

De plus, pour chaque rupture $T_k^\star$, il existe au moins une ligne telle que la loi de cette ligne avant la rupture soit différente de la loi après la rupture. Pour que la statistique $S_n(\boldsymbol{T})$ voit cette différence, nous avons besoin de la condition suivante :

$$\begin{array}{l} \forall k \in \{0,\ldots,K^\star-1\},\ \exists \ell_k \in \{0,\ldots,K^\star\} \text{ tels que} \\ \sum_{k'=0}^{K^\star} \left(\tau_{k'+1}^\star - \tau_{k'}^\star\right) \mathbb{E}_{X\sim\mathcal{L}_{\ell_k,k'}}\left[F_{\ell_k,k}(X) - F_{\ell_k,k+1}(X)\right] \neq 0 \end{array} \qquad (\mathcal{H}^{\text{MuChPoint}}_{F})$$

Ainsi, nous avons la consistance des estimateurs des emplacements des ruptures :

Théorème 12 (Consistance des estimateurs des emplacements des ruptures $\widehat{\boldsymbol{\tau}}_n$)

Sous les hypothèses $(\mathcal{H}^{\text{MuChPoint}}_{\Delta_{\boldsymbol{\tau}^\star}})$ et $(\mathcal{H}^{\text{MuChPoint}}_{F})$, nous avons la consistance des estimateurs $\widehat{\boldsymbol{\tau}}_n$ des emplacements des ruptures ; c'est-à-dire que pour tout $\delta > 0$, nous avons :

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}^\star\|_{\text{Haus}} > \delta\right) \xrightarrow[n\to+\infty]{} 0$$

avec $\|\cdot\|_{\text{Haus}}$ la norme de Hausdorff déjà utilisée précédemment et rappelée dans le théorème 7.

Remarque

Comme pour l'article Brault et al. (2017c), le résultat reste à mon avis vrai même si $\Delta_{\boldsymbol{\tau}^\star}$ tend vers 0 mais pas trop rapidement.

Dans notre article Brault et al. (2018c), nous testons l'estimateur sur des données gaussiennes, exponentielles et suivant une loi de Cauchy pour différentes configurations et nous nous comparons à la méthode `ecp` proposée dans Matteson et James (2014). Par exemple, sur la figure 3.10, nous représentons les erreurs de distance quadratiques moyennes entre les vraies ruptures $\boldsymbol{T}^\star$ équirépartis et les estimations $\widehat{\boldsymbol{T}}_n$ faites par la procédure MuChPoint (en gris) et la méthode `ecp` (en blanc) dans une configuration en échiquier.

Nous pouvons remarquer sur la figure 3.10 que les estimations sont d'autant meilleures pour notre méthode qu'il y a beaucoup d'observations (à droite de chaque graphique) et que le bruit est faible (en haut de chaque ligne). Bien que notre méthode préfère les lois gaussiennes et exponentielles, elle tend quand même à donner de bonnes estimations pour les lois de Cauchy contrairement à la méthode `ecp`.

3.3.3 Sélection du nombre de ruptures

Un problème encore en suspens est la question de la sélection du nombre de ruptures. Sur le jeu de données des souris étudié par Dixon et al. (2012) [5], nous utilisons l'heuristique de pente implémentée dans le package `Capushe` que je maintiens (voir Brault et al. (2023a)). Pour ce faire, nous utilisons la procédure `MuChPoint` et récupérons les valeurs de la statistique $\max_{\boldsymbol{T}_K} S_n(\boldsymbol{T}_K)$ utilisées comme contraste (voir Baudry et al. (2012)) pour des valeurs de K allant de 1 à 500. Nous utilisons le package `Capushe` (voir la figure 3.11) sur ces données par l'intermédiaire d'une implémentation présente dans notre propre package `MuChPoint`.

Lorsque l'heuristique de pente fonctionne convenablement, la droite rouge de la gauche de la figure 3.11 est censée suivre relativement bien les points ; dans notre cas, elle ne semble pas épouser la fin de la courbe de façon optimale. Cela se voit également sur la figure en haut à droite où il n'y a pas de stabilisation dans l'estimation des pentes qui est strictement décroissante. Enfin, il semblerait tout de même que nous ayons des valeurs plateaux (38, 39 et 40) mais il serait important de mieux comprendre quel contraste précisément il faudrait prendre pour estimer le nombre de ruptures.

Perspectives

Dans notre travail avec Jean-Charles Quinton et Adeline Leclercq Samson, nous sommes en train d'essayer de comprendre le comportement de la statistique $\max_{\boldsymbol{T}} S_n(\boldsymbol{T})$ afin de comprendre quel contraste nous devons fournir à l'heuristique de pente pour avoir de bonnes estimations.

3.4 Application aux véhicules autonomes

Pour conclure ce chapitre, j'aborde dans cette section un travail en cours sur une application des modèles `blockseg` et `MuChPoint` au cas des voitures autonomes. Dans le cadre d'une collaboration avec Jean-Charles Quinton et Adeline Leclercq Samson, nous avons à disposition un film d'un trajet fait par un véhicule autonome sur un parcours situé à Clermont Ferrand. Pour chaque couple d'images, nous

5. Les données étaient disponibles sur l'url http://chromosome.sdsc.edu/mouse/hi-c/download.html mais le lien semble ne plus être maintenu

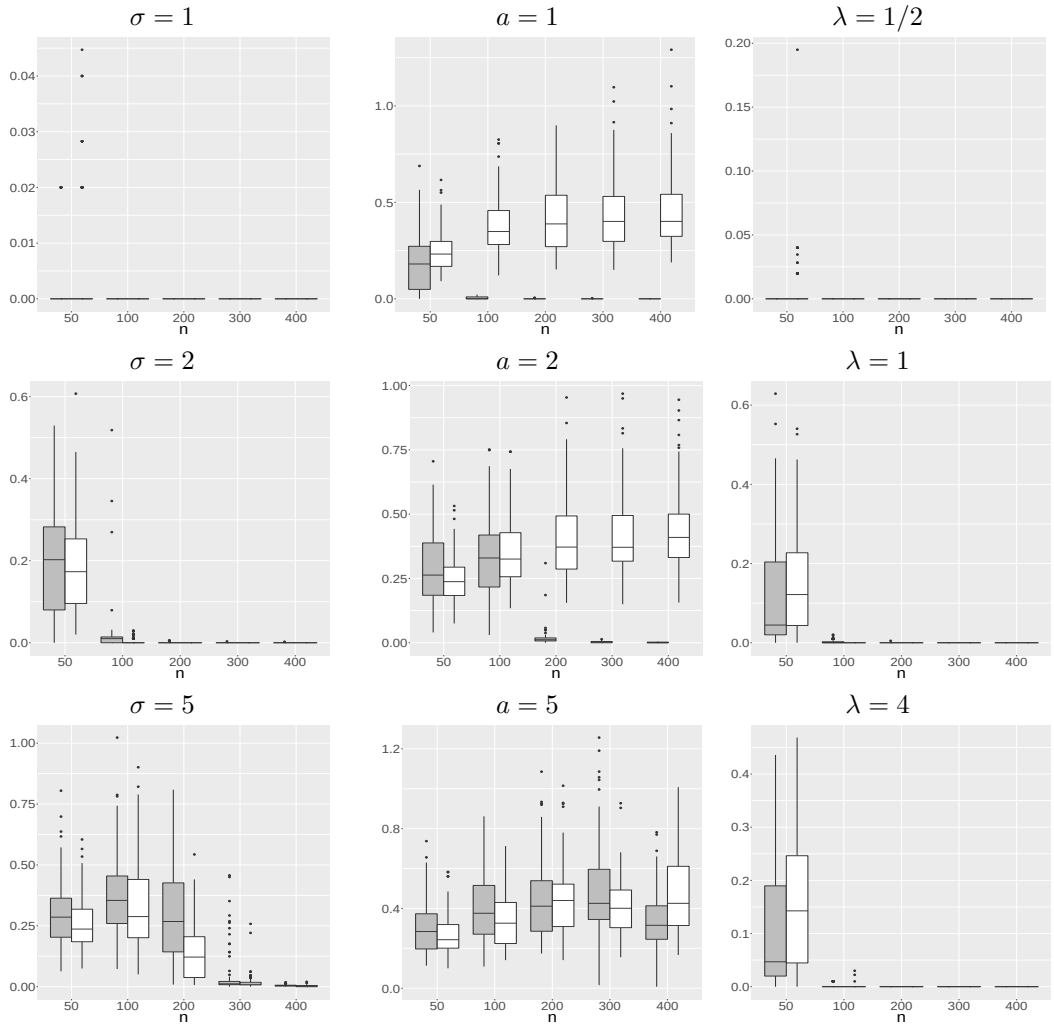

FIGURE 3.10 – Boxplots des distances quadratiques entre les emplacements des vraies ruptures $\boldsymbol{T}^\star$ équirépartis et les estimations $\widehat{\boldsymbol{T}}_n$ faites par la procédure MuChPoint (en gris) et la méthode ecp (en blanc) dans une configuration en échiquier alternant $\mathcal{L}_1$ et $\mathcal{L}_2$ suivant différentes tailles n de matrices (en abscisse). Les lois gaussiennes $\mathcal{L}_1 = \mathcal{N}(1, \sigma^2)$ et $\mathcal{L}_2 = \mathcal{N}(0, \sigma^2)$ sont à gauche, les lois de Cauchy $\mathcal{L}_1 = \mathcal{C}au(1, a)$ et $\mathcal{L}_2 = \mathcal{C}au(0, a)$ au milieu et les lois exponentielles $\mathcal{L}_1 = \mathcal{E}xp(2)$, $\mathcal{L}_2 = \mathcal{E}xp(\lambda)$ sont à droite pour différentes valeurs de σ, λ and a. Figures issues de l'article Brault et al. (2018c).

utilisons un algorithme qui calcule la similarité (ou non) entre ces deux images et on obtient une matrice de valeurs comme celle représentée sur la figure 3.12 où les lignes et les colonnes correspondent à chaque image du film et une case correspond à la similarité entre l'image en ligne et l'image en colonne (plus la couleur est rouge sur la figure, plus la similarité est forte).

Nous voyons sur cette figure 3.12 les rectangles rouges sur la diagonale correspondant souvent à des lignes droites. Dans la partie supérieure gauche de la matrice, nous voyons des carrés rouges hors de la diagonale correspondants au fait que la voiture a fait deux tours d'un rond point (voir le rectangle vert). De même, des gros motifs rouges très excentrés de la diagonale (rectangle violet) correspondent au moment où la voiture est revenue sur cette portion.

Nous appliquons les estimations des modèles présentés précédemment et regardons la cohérence des partitions obtenues (par exemple, sur la figure 3.13, nous représentons le résultat obtenu par la procédure `MuChPoint` avec 50 ruptures).

Nous voyons que des rectangles rouges ressortent notamment sur la diagonal. Par exemple, le premier rectangle très rouge en haut à gauche correspond au début du film où la voiture était arrêtée. Le rectangle rouge un peu plus à droite sur la même ligne correspond au moment où la voiture est repassée par le même endroit (c'est d'ailleurs l'interaction la plus forte en dehors de la diagonale). Les rond points ont des faibles valeurs car les images changent très vite.

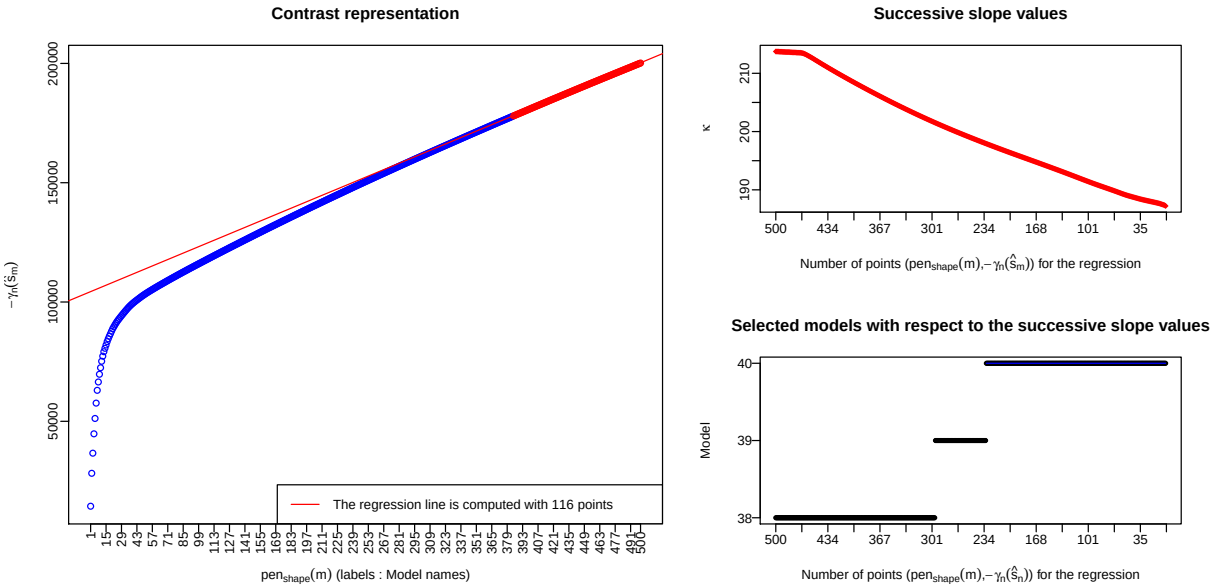

FIGURE 3.11 – Représentation de la sortie proposée par le package `Capushe` pour l'estimation du nombre de ruptures dans le cadre du modèle `MuChPoint` sur les données *Hi-C* de souris étudiées par Dixon et al. (2012). Figures issues de l'article Brault et al. (2018c).

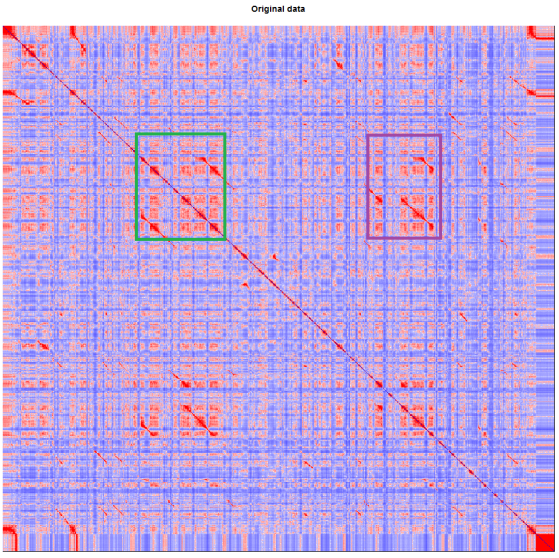

FIGURE 3.12 – Représentation de la matrice de similarité des images prises deux par deux dans un film d'un véhicule autonome : plus une case est rouge, plus les deux images mises en lignes et en colonnes sont semblables. Le rectangle vert a été ajouté pour mettre en évidence le motif correspondant à deux tours de rond point et le rectangle violet pour le moment où la voiture a refait le même trajet dans le même sens.

Le but du projet est d'améliorer les estimations afin de proposer des techniques pour permettre à des véhicules autonomes qui serviraient à des trajets réguliers de se repérer sur leur trajet sans avoir besoin d'un GPS trop précis (car très coûteux au moment où nous avons commencé le projet).

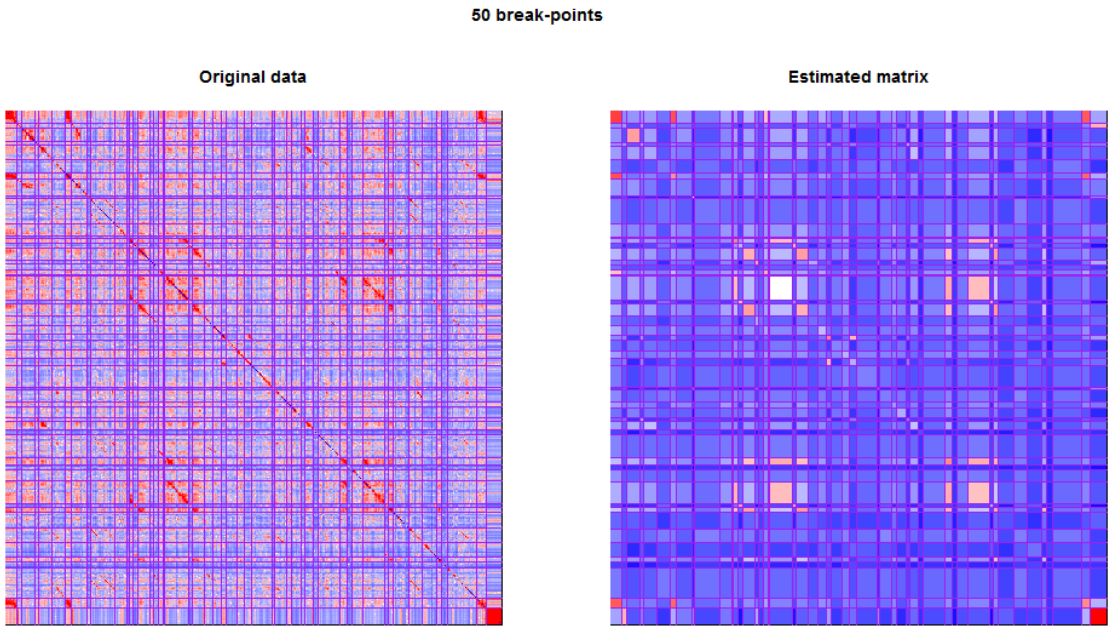

FIGURE 3.13 – Représentation des 50 ruptures obtenues sur la matrice présentée dans la figure 3.12 par l'algorithme `MuChPoint` avec la matrice originale à gauche et la matrice résumée à droite.

3.5 Modèle de segmentation et croissance des plants de colza

Dans le cadre d'une collaboration avec l'UMR ECOSYS de l'INRAE, nous avons étudié dans l'article Brault et al. (2018b) la croissance de plants de colza et plus précisément le nombre de feuilles par plantes. Cette modélisation possédait plusieurs contraintes :

- L'évolution du nombre de feuilles suit une fonction **continue** et **linéaire par morceaux** (voir par exemple Jullien et al. (2011)) en fonction du temps thermique exprimé en *degrés jour* et peut être décomposée en trois phases :
  - La première phase où le nombre de feuilles croît lentement.
  - Dans la seconde phase, une accélération de l'apparition du nombre de feuilles est observée.
  - La dernière phase est celle de la reproduction ; dans cette phase, l'apparition des fleurs remplace celle des feuilles. Nous nous attendons donc à ce que le nombre de feuilles n'évolue plus.
- L'évolution du nombre de feuilles peut varier fortement suivant la concurrence entre les plantes (voir par exemple Baey et Cournède (2011)).

Ainsi, étant donnés une plante j et un ensemble de temps thermiques $\boldsymbol{t} = (t_i)_{1\leq i\leq n}$ commun à toutes les plantes, nous supposons qu'il existe deux moments de rupture $T_{0,j} = 0 < T_{1,j} < T_{2,j} < T_{3,j} = n$ et trois paires de coefficients $(a_{k,j}, b_{k,j})_{0\leq k\leq 2}$ représentant les évolutions de croissance à chaque phase tels que :

$$\begin{cases} \forall k \in \{0,1,2\}, & \forall i \in \{T_{k,j}+1, \dots, T_{k+1,j}\},\ Y_{i,j} = a_{k,j} + b_{k,j} t_i + \varepsilon_i, \\ \forall k \in \{0,1\}, & a_{k,j} + b_{k,j}(T_{k+1,j}+1) = a_{k+1,j} + b_{k+1,j}(T_{k+1,j}+1), \end{cases} \tag{3.17}$$

où les ε_i sont indépendants d'un temps à l'autre mais vont dépendre des plantes proches (voir l'équation (3.20)). De plus, nous supposons qu'il n'y a pas de feuilles au début de l'expérience ce qui implique que les $a_{0,j}$ soient nuls.

3.5.1 Transformation pour la méthode *LASSO*

La modélisation (3.17) est équivalente à l'équation suivante :

$$\boldsymbol{Y} = \mathbb{T}_n^{(\boldsymbol{t})} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.18}$$

où $\boldsymbol{Y} \in \mathcal{M}_{n\times J}(\mathbb{R})$ est la matrice des observations des J plantes suivant les n temps thermiques, $\mathbb{T}_n^{(\boldsymbol{t})}$ est la matrice triangulaire inférieure des différences des temps thermiques définie par

$$\mathbb{T}_n^{(\boldsymbol{t})} = \left[(t_i - t_{j-1})\,\mathbb{1}_{\{i\geq j\}}\right]_{\substack{1\leq i\leq n\\ 1\leq j\leq n}} = \begin{pmatrix} t_1 & 0 & 0 & \cdots & 0 \\ t_2 & t_2 - t_1 & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ t_n & t_n - t_1 & \cdots & \cdots & t_n - t_{n-1} \end{pmatrix},$$

$\boldsymbol{\beta} \in \mathcal{M}_{n\times J}(\mathbb{R})$ est une matrice sparse où les seules valeurs non nulles correspondent aux emplacements des ruptures $T_{k,j}$ et vérifient :

$$\begin{cases} \beta_{1,j} = b_{0,j}, \\ \beta_{T_{1,j}+1,j} = b_{1,j} - b_{0,j}, \\ \beta_{T_{2,j}+1,j} = b_{2,j} - b_{1,j}, \end{cases} \tag{3.19}$$

et $\boldsymbol{\varepsilon} \in \mathcal{M}_{n\times J}(\mathbb{R})$ est une matrice aléatoire où chaque ligne suit une loi gaussienne multivariée $\mathcal{N}(\mathbb{0}_J, \boldsymbol{\Sigma}_J)$ indépendante des autres lignes et avec la matrice $\boldsymbol{\Sigma}_J = (\Sigma_{j,j'})_{1\leq j,j'\leq J}$ définie par :

$$\Sigma_{j,j'} = \sigma^2 \exp\left(\frac{\|\mathrm{Pos}_j - \mathrm{Pos}_{j'}\|_2^2}{2\ell^2}\right) \tag{3.20}$$

où Pos_j est la position de la $j^{\text{ème}}$ plante de colza dans le bac, $\|\cdot\|_2$ est la norme euclidienne et σ^2 et ℓ sont des paramètres à estimer.

Enfin, en vectorisant le modèle (3.18), nous obtenons l'équation :

$$\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E} \tag{3.21}$$

où $\mathcal{Y} = \mathrm{Vec}(Y)$, $\mathcal{X} = \mathbb{I}_K \otimes \mathbb{T}_n^{(\boldsymbol{t})}$ avec $\otimes$ représente le **produit de Kronecker**, $\mathcal{B} = \mathrm{Vec}(\boldsymbol{\beta})$ et $\mathcal{E} = \mathrm{Vec}(\boldsymbol{\varepsilon})$ et nous retrouvons la formalisation de la méthode *LASSO*.

3.5.2 Estimation des paramètres

Dans la modélisation (3.21), nous devons estimer les valeurs non nulles de la matrice $\boldsymbol{\beta}$ et les paramètres σ^2 et ℓ. Toutefois, à cause de la dépendance des observations, l'estimation des paramètres par la méthode *LASSO* peut être dégradée (voir par exemple Jia et Rohe (2015)), nous avons donc fait une estimation en quatre étapes :

1. La première étape consiste à estimer le bruit. Pour cela, nous lançons une estimation par la procédure *LASSO* en supposant le bruit indépendant et en choisissant le λ_{CV} fourni par la validation croisée. Ainsi, nous avons une estimation du bruit par les résidus :

$$\widehat{\mathcal{E}} = \mathcal{Y} - \mathcal{X}\widehat{\mathcal{B}}(\lambda_{\mathrm{CV}}) \tag{3.22}$$

   et donc une estimation de la matrice $\widehat{\boldsymbol{\varepsilon}}$.

2. La deuxième étape consiste à estimer $\widehat{\boldsymbol{\Sigma}_J}$ en estimant les paramètres de l'équation (3.20) par maximum de vraisemblance sur les données de l'équation (3.22).

3. À l'aide de l'estimation $\widehat{\boldsymbol{\Sigma}_J}$ de la matrice $\boldsymbol{\Sigma}_J$, nous procédons à un **blanchissement** de nos données en multipliant à droite par la matrice $\widehat{\boldsymbol{\Sigma}_J}^{-\frac{1}{2}}$ les deux côtés de l'équation (3.18) :

$$\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}_J}^{-\frac{1}{2}} = \mathbb{T}_n^{(\boldsymbol{t})}\boldsymbol{\beta}\widehat{\boldsymbol{\Sigma}_J}^{-\frac{1}{2}} + \boldsymbol{\varepsilon}\widehat{\boldsymbol{\Sigma}_J}^{-\frac{1}{2}}. \tag{3.23}$$

   Ainsi, si l'estimation des paramètres est correcte, nous nous attendons à ce que le bruit de l'équation (3.23) soit composé de variables indépendants et de même loi.

4. Ensuite, nous devons estimer les paramètres de la matrice $\boldsymbol{\beta}$. Pour cela, nous vectorisons le modèle (3.23) :

$$\mathcal{Y}' = \mathcal{X}'\mathcal{B} + \mathcal{E}' \tag{3.24}$$

où $\mathcal{Y}' = \text{Vec}\left(Y\widehat{\mathbf{\Sigma}_J}^{-\frac{1}{2}}\right)$, $\mathcal{X}' = \left(\widehat{\mathbf{\Sigma}_J}^{-\frac{1}{2}}\right)^T \otimes \mathbb{T}_n^{(t)}$ où A^T est la transposée de la matrice A, $\mathcal{B} = \text{Vec}(\boldsymbol{\beta})$ et $\mathcal{E}' = \text{Vec}\left(\boldsymbol{\varepsilon}\widehat{\mathbf{\Sigma}_J}^{-\frac{1}{2}}\right)$. Toutefois, comme suggéré dans Jia et Rohe (2015), nous avons utilisé une transformation de Puffer pour améliorer l'estimation des positions non nulles du vecteur $\mathcal{B}$. Pour ce faire, nous faisons une décomposition en valeurs singulières de la matrice $\mathcal{X} = UDV^T$ avec U et V des matrices orthogonales et D une matrice diagonale. À l'aide de cette transformation, nous définissons la matrice $F = UD^{-1}U^T$ et nous souhaitons appliquer la méthode *LASSO* sur une transformation du modèle (3.24) :

$$F\mathcal{Y}' = F\mathcal{X}'\mathcal{B} + F\mathcal{E}'.$$

Le problème de cette transformation est que la covariance de la loi de $F\mathcal{E}'$ est $UD^{-2}U^T$ et les valeurs trop petites de D deviennent alors gigantesques. Ainsi, nous étudions finalement la version suivante :

$$F_m\mathcal{Y}' = F_m\mathcal{X}'\mathcal{B} + F_m\mathcal{E}' \text{ avec } F_m = UD_m^{-1}U^T \tag{3.25}$$

où D_m ne contient que les m plus grandes valeurs de la matrice D. En pratique, nous avons observé empiriquement dans la section 3 de Brault et al. (2018b) qu'il fallait ne conserver que les valeurs supérieures à deux.

5. Enfin, nous estimons les paramètres du vecteur $\mathcal{B}$ de l'équation (3.25) en utilisant une procédure *LASSO* couplée à une validation croisée et les valeurs non nulles correspondent aux emplacements des ruptures. Néanmoins, comme il est assez rare que nous ayons exactement deux valeurs non nulles pour chaque plan, nous avons proposé deux méthodes pour ne conserver que deux ruptures :

   - La méthode du minimum et du maximum consistaient à prendre la plus grande et la plus petite valeur des l'estimation $\widehat{\mathcal{B}}(\lambda_{\text{CV}})$. Ceci rejoint la remarque faite dans le paragraphe précédent sur l'évolution des pentes.
   - La méthode des deux maxima consistaient à prendre pour chaque plan les deux valeurs les plus élevées parmi les valeurs absolues de l'estimation $\widehat{\mathcal{B}}(\lambda_{\text{CV}})$.

3.5.3 Applications

Enfin, nous avons fait des plans d'expérience et appliqué notre méthode sur les données réelles (voir la figure 3.14).

Sur cette figure, nous pouvons voir que la phase 2 est très courte et se produit pour tous les plants autour d'un temps thermique de 800. La méthode du maximum de vraisemblance va avoir tendance à mettre la première rupture plus tôt que les autres méthodes. La méthode du minimum et maximum (en vert) va plus souvent mettre les ruptures très proches les unes des autres. Enfin, la méthode des deux maxima peut être vu comme un compromis entre les deux.

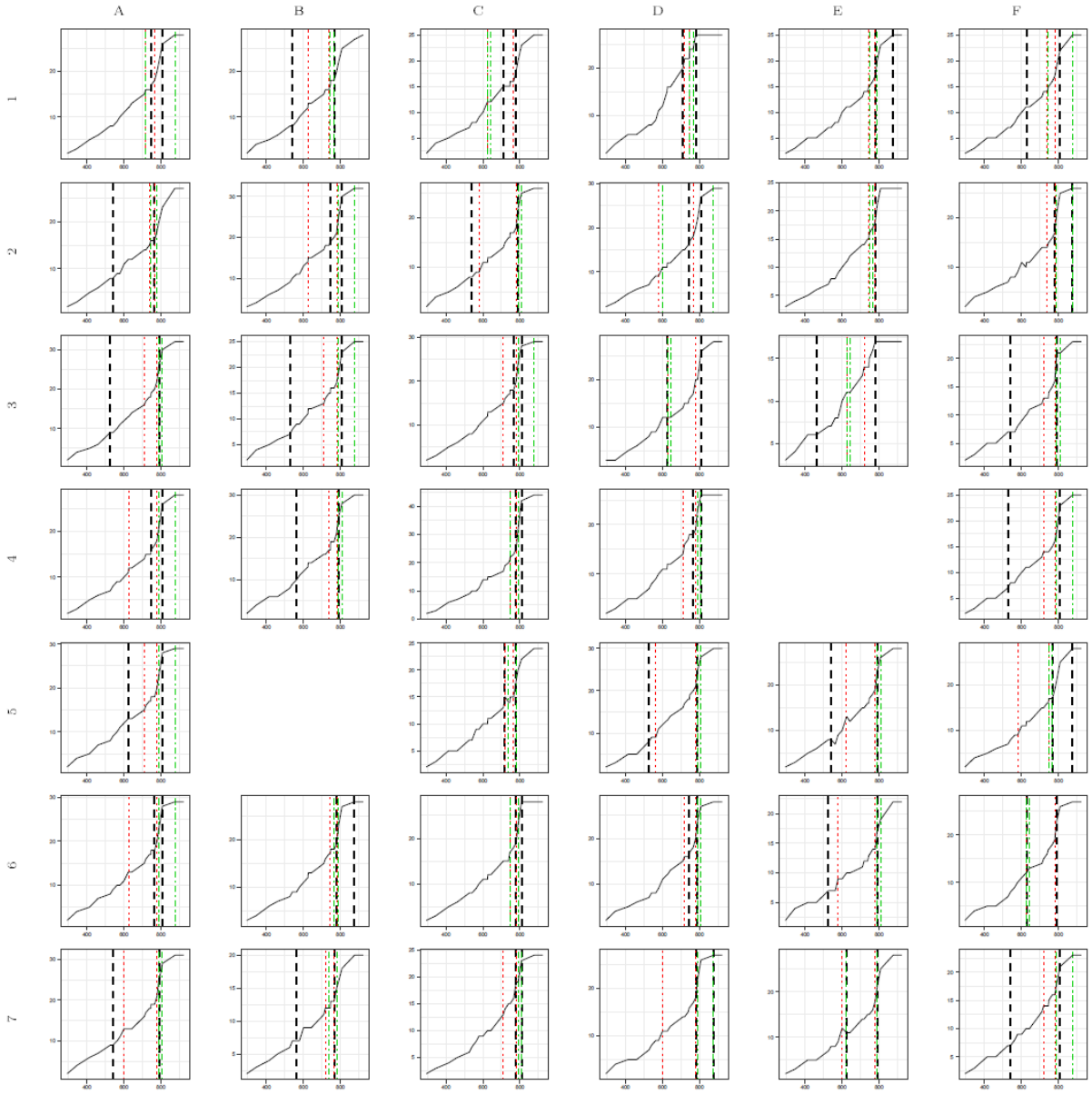

FIGURE 3.14 – Estimation des deux instants de ruptures pour chaque plant de Colza du premier container (les emplacements vides correspondent à des plants où nous n'avons pas les informations) suivant différentes méthodes : le maximum de vraisemblance (en noir), la méthode du minimum et du maximum (en vert) et la méthode des deux maxima (en rouge)

Chapitre 4

Symbiose entre les mélanges et la segmentation

"In a normal education everything is designed to suppress spontaneity, but I wanted to develop it."
Keith Johnstone, *Impro : Improvisation and the Theatre*

Depuis quelques années, j'ai été associé à différents projets analysant des données nécessitant une modélisation couplant classification et ruptures. Dans ce chapitre, je vais donc présenter la multitude des applications et des modélisations proposées. Je commence par présenter le travail le plus proche de la combinaison des deux qui est le modèle de mélange de segmentation avec une application sur les données électriques récemment publié (voir la section 4.1) et d'une modélisation sur les protéines tau qui ressemble à un mélange de segmentation mais avec des paramètres en commun entre les groupes (voir la section 4.2). Je parlerai ensuite d'un travail en cours sur l'écriture des enfants (voir la section 4.3) et enfin le travail sur l'apnée du sommeil (section 4.4). Tous ces travaux n'étant pas encore publiés, il reste encore des pistes à améliorer.

4.1 Courbes électriques et mélange de segmentation

La première combinaison est celle du modèle de **mélange de segmentations** proposée dans l'article Brault et al. (2024a). Le principe est que chaque individu $i \in \{1, \dots, n\}$ est représenté par une suite ordonnée de J variables et chaque individu appartient à l'un des $K^\star$ clusters avec probabilité $\pi_k^\star$. Pour chaque cluster $k \in \{1, \dots, K^\star\}$, nous supposons qu'il existe $L_k^\star$ ruptures $\boldsymbol{T}_k^\star = \left(T_{k,1}^\star, \dots, T_{k,L_k^\star}^\star\right)$ telles que $0 = T_{k,0}^\star < T_{k,1}^\star < \cdots < T_{k,L_k^\star}^\star < T_{k,L_k^\star+1}^\star = J$ et $L_k^\star + 1$ lois $\left(\mathcal{L}_{k,\ell}^\star\right)_{0 \leq k \leq L_k^\star}$ telles que les variables $\boldsymbol{Y}_{i,j}$ soient indépendantes vérifiant pour tout individu $i \in \{1, \dots, n\}$ et tout temps/position $j \in \{1, \dots, J\}$ (voir le graphe bayésien de la figure 4.1) :

$$\boldsymbol{Y}_{i,j} \left| \left(z_{i,k}^\star = 1, T_{k,\ell}^\star + 1 \leq j \leq T_{k,\ell+1}^\star\right) \right. \sim \mathcal{L}_{k,\ell}^\star. \tag{4.1}$$

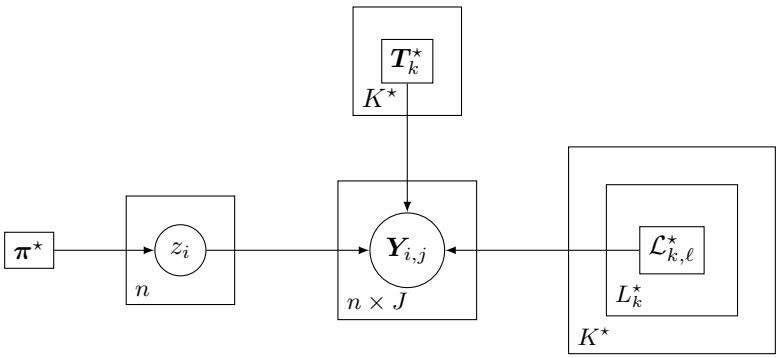

FIGURE 4.1 – Représentation graphique du modèle de mélange de segmentation : les carrés correspondent à des paramètres et les ronds à des variables.

Remarque

Par conséquent, nous supposons que chaque ligne de la matrice possède le même nombre de cases mais les groupes peuvent avoir un nombre et des emplacements de ruptures différents.

Dans notre article Brault et al. (2024a), nous supposons également que les lois sont gaussiennes de moyennes et de variances elliptiques dépendantes à la fois du cluster et de l'intervalle :

$$\forall k \in \{1, \dots, K^\star\}, \forall \ell \in \{0, \dots, L_k^\star\}, \mathcal{L}_{k,\ell} \sim \mathcal{N}_H \left[\boldsymbol{\mu}_{k,\ell}^\star, \mathbb{I}_{\boldsymbol{\sigma}_{k,\ell}^\star}\right] \tag{4.2}$$

où $\mathbb{I}_{\boldsymbol{\sigma}_{k,\ell}^\star}$ est la matrice diagonale avec les H valeurs de $\boldsymbol{\sigma}_{k,\ell}^\star = \left(\sigma_{k,\ell,1}^\star, \dots, \sigma_{k,\ell,H}^\star\right)$.

Dans la suite de cette section, je présente le modèle et l'estimation des paramètres puis les résultats théoriques avant de finir sur l'application aux courbes électriques.

4.1.1 Vraisemblance et estimation

À partir des modélisations (4.1) et (4.2), nous obtenons la vraisemblance du modèle :

$$p\left(\boldsymbol{Y}; \boldsymbol{T}^\star, \boldsymbol{\Psi}^\star\right) = \prod_{i=1}^{n} \sum_{k=1}^{K^\star} \pi_k^\star \prod_{\ell=0}^{L_k^\star} \prod_{j=T_{k,\ell}^\star+1}^{T_{k,\ell+1}^\star} \prod_{h=1}^{H} \left\{ \frac{1}{\sqrt{2\pi\sigma_{k,\ell}^{2\star}}} \exp\left[-\frac{1}{2\sigma_{k,\ell,h}^{2\star}}\left(Y_{i,j,h} - \mu_{k,\ell,h}^\star\right)^2\right] \right\}. \tag{4.3}$$

Pour estimer les paramètres $\boldsymbol{\Psi}$ et les emplacements des ruptures, nous utilisons l'algorithme *EM* couplé à l'algorithme de programmation dynamique pour l'étape *M*. En effet, l'étape *E* possède une forme explicite connaissant les paramètres $\boldsymbol{\Psi}^{(c)}$ et les ruptures $\boldsymbol{T}^{(c)}$ estimés à l'étape précédentes.

Pour l'étape *M* de maximisation, nous remarquons que si nous connaissons la matrice $s_{i,k}^{(c)}$ des probabilités conditionnelles connaissant les observations, des paramètres $\boldsymbol{\Psi}_{c-1}$ et des ruptures $\boldsymbol{T}^{(c-1)}$ alors nous avons :

$$\begin{aligned}
&\max_{\boldsymbol{\Psi}, \boldsymbol{T}} \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y}; \boldsymbol{\Psi}^{(c)}, \boldsymbol{T}^{(c)}} \left[\log p\left(\boldsymbol{Y}, \boldsymbol{Z}; \boldsymbol{\Psi}, \boldsymbol{T}\right)\right] \\
&= \quad -\frac{1}{2} \sum_{k=1}^{K^\star} \left[\min_{\boldsymbol{T}_k} \sum_{\ell=0}^{L_k^\star} \Delta_k^{(c)}\left(T_{k,\ell} : T_{k,\ell+1}\right)\right] - \frac{nJH}{2} \log\left(2\pi\right) + \max_{\boldsymbol{\pi}} \sum_{k=1}^{K^\star} s_{+,k}^{(c)} \log \pi_k \\
&\text{avec } \forall 0 \leq T_1 < T_2 \leq J \\
&\Delta_k^{(c)}\left(T_1 : T_2\right) \\
&= \quad \min_{\boldsymbol{\mu}, \boldsymbol{\sigma}} \sum_{h=1}^{H} \left[s_{+,k}^{(c)}\left(T_2 - T_1\right) \log\left(\sigma_h^2\right) + \frac{1}{\sigma_h^2} \sum_{j=T_1+1}^{T_2} \sum_{i=1}^{n} s_{i,k}^{(c)}\left(Y_{i,j,h} - \mu_h\right)^2\right].
\end{aligned}$$

La mise à jour du paramètre $\boldsymbol{\pi}$ se fait en prenant les moyennes par colonnes de la matrice $s_{i,k}^{(c)}$ et nous pouvons utiliser l'algorithme de programmation dynamique pour les autres paramètres ainsi que les instants de ruptures.

Pour optimiser le temps de calcul, nous commençons par calculer toutes les moyennes [1]

$$\overline{\boldsymbol{Y}}_{i,T_1:T_2,h} = \frac{1}{T_2 - T_1} \sum_{j=T_1+1}^{T_2}$$

et nous obtenons ainsi :

Proposition 13 (Complexité de l'algorithme d'estimation)

L'algorithme *EM* couplé avec l'algorithme de programmation dynamique possède une complexité de $\mathcal{O}\left(HnJ^2 \max\left(J, KN_{\text{algo}}\right)\right)$ où N_{algo} est le nombre d'itérations de l'algorithme *EM*.

1. Les codes sont disponibles sur https://github.com/laclauc/MixtSegmentation

4.1.2 Résultats théoriques

Le premier résultat que nous avons est l'identifiabilité du modèle, pour cela, nous avons besoin d'hypothèses :

Hypothèse (Brault et al. (2024a))

La première hypothèse consiste à supposer que les ruptures sont visibles :

$$\forall k \in \{1,\dots,K^\star\}, \ell \in \{0,\dots,L_k^\star - 1\}, \exists h \in \{1,\dots,H\}, \quad \mu^\star_{k,\ell,h} \neq \mu^\star_{k,\ell+1,h} \text{ ou } \sigma^\star_{k,\ell,h} \neq \sigma^\star_{k,\ell+1,h} \tag{$\mathcal{I}d^{\text{MixSeg}}_{\text{Ruptures}}$}$$

La deuxième hypothèse suppose que nous avons assez de colonnes pour voir toutes les ruptures :

$$J \geq \max_{k \in \{1,\dots,K^\star\}} (L_k^\star + 1) \tag{$\mathcal{I}d^{\text{MixSeg}}_{J}$}$$

La troisième hypothèse permet de différencier deux clusters :

$$\begin{aligned} &\forall k \in \{1,\dots,K^\star\}, \forall k' \in \{1,\dots,K^\star\}\setminus\{k\} \text{ tels que } L_k^\star = L_{k'}^\star \text{ alors :} \\ &\bullet\ \exists \ell \in \{0,\dots,L_k^\star\}, T^\star_{k,\ell} \neq T^\star_{k',\ell} \\ &\bullet\ \text{ou } \exists \ell \in \{0,\dots,L_k^\star\}, \exists h \in \{1,\dots,H\}, \mu^\star_{k,\ell,h} \neq \mu^\star_{k',\ell,h} \text{ ou } \sigma^\star_{k,\ell,h} \neq \sigma^\star_{k',\ell,h} \end{aligned} \tag{$\mathcal{I}d^{\text{MixSeg}}_{\text{Clusters}}$}$$

La dernière hypothèse suppose qu'il y a une possibilité d'avoir un représentant de chaque cluster :

$$\forall k \in \{1,\dots,K^\star\}, \pi_k^\star > 0 \tag{$\mathcal{I}d^{\text{MixSeg}}_{\boldsymbol{\pi}}$}$$

Remarque

Les hypothèses peuvent être classées en deux groupes : les deux premières portent sur la segmentation tandis que les deux dernières sont pour le modèle de mélange.

À l'aide de ces hypothèses, nous avons l'identifiabilité :

Théorème 14 (Identifiabilité du modèle (4.3))

Sous les hypothèses ($\mathcal{I}d^{\text{MixSeg}}_{\text{Ruptures}}$), ($\mathcal{I}d^{\text{MixSeg}}_{J}$), ($\mathcal{I}d^{\text{MixSeg}}_{\text{Clusters}}$) et ($\mathcal{I}d^{\text{MixSeg}}_{\boldsymbol{\pi}}$) le modèle (4.3) de mélange de segmentations gaussien est identifiable.

Remarque

L'une des particularités de ce modèle est qu'il n'y a pas de *label switching* possible dès que les nombre de ruptures $L_k^\star$ des clusters sont différents. Dans ce cas, nous pouvons ordonner les numéros de clusters par ordre croissant du nombre de ruptures (le cluster 1 aurait le moins de ruptures pendant que le cluster K aurait le plus grand nombre). Sinon, le problème de label switchning est réservé aux groupes de clusters qui ont le même nombre de ruptures.

Pour la démonstration de la consistance de l'estimateur du maximum de vraisemblance, nous faisons l'hypothèse que nous connaissons la variance et qu'elle vaut 1 pour tous les groupes et nous avons pris $H = 1$. Nous avons besoin de nouvelles hypothèses et de renforcer certaines de l'identifiabilité :

Hypothèse (Brault et al. (2024a))

D'abord, nous avons besoin que les moyennes soient dans un compact :

$$\exists M > 0, \forall k \in \{1,\dots,K^\star\}, \forall \ell \in \{0,\dots,L_k^\star\}, \mu^\star_{k,\ell} \in [-M, M]. \tag{$\mathcal{H}^{\text{MixSeg}}_{\boldsymbol{\mu}^\star}$}$$

Ensuite, la distance minimale entre deux ruptures normalisées ayant servi à simuler les données est strictement positive :

$$0 < \Delta_{\boldsymbol{\tau}^\star} = \min_{1 \leq k \leq K^\star} \min_{0 \leq \ell \leq L_k^\star} \left(\tau^\star_{k,\ell+1} - \tau^\star_{k,\ell}\right). \tag{$\mathcal{H}^{\text{MixSeg}}_{\Delta_{\boldsymbol{\tau}^\star}}$}$$

Et il faut assez d'informations pour distinguer les groupes :

$$\frac{\log(n)}{J} \xrightarrow[n,J\to+\infty]{} 0 \qquad (\mathcal{H}^{\text{MixSeg}}_{\text{Asymp}})$$

Ensuite, il faut renforcer l'hypothèse ($\mathcal{I}d^{\text{MixSeg}}_{\text{Clusters}}$) car le fait d'avoir des emplacements différents n'est plus suffisant si nous n'avons pas assez d'informations. S'il existe deux clusters $k \neq k'$ avec le même nombre de ruptures $L^\star_k = L^\star_{k'}$ alors il existe au moins $\Delta_{\boldsymbol{\tau}^\star} J$ colonnes telles que :

$$Y_{i,j} \,|z^\star_{i,k} = 1 \sim \mathcal{L} \text{ et } Y_{i,j} \,|z^\star_{i,k'} = 1 \sim \mathcal{L}' \text{ avec } \mathcal{L} \not\sim \mathcal{L}'. \qquad (\mathcal{H}^{\text{MixSeg}}_{\text{Clusters}})$$

De même, il faut un nombre minimal d'individus plutôt que juste une probabilité strictement positive qu'il y ait un individu comme dans l'hypothèse ($\mathcal{I}d^{\text{MixSeg}}_{\boldsymbol{\pi}}$) :

$$\exists \pi_{\min} > 0, \forall k \in \{1, \dots, K^\star\}, \pi^\star_k > \pi_{\min} \qquad (\mathcal{H}^{\text{MixSeg}}_{\boldsymbol{\pi}})$$

Grâce à ces hypothèses, nous avons le même résultat que pour le théorème 1 du modèle des blocs latents :

Théorème 15 (Consistance des estimateurs du maximum de vraisemblance)

Sous les hypothèses ($\mathcal{I}d^{\text{MixSeg}}_{\text{Ruptures}}$), ($\mathcal{H}^{\text{MixSeg}}_{\text{Clusters}}$), ($\mathcal{H}^{\text{MixSeg}}_{\boldsymbol{\pi}}$), ($\mathcal{H}^{\text{MixSeg}}_{\boldsymbol{\mu}^\star}$), ($\mathcal{H}^{\text{MixSeg}}_{\Delta_{\boldsymbol{\tau}^\star}}$) et ($\mathcal{H}^{\text{MixSeg}}_{\text{Asymp}}$), pour tout paramètre $\boldsymbol{\Psi}$ et pour tout ensemble de ruptures $\boldsymbol{T}$ vérifiant les conditions des hypothèses, nous avons :

$$\frac{p\left(\boldsymbol{Y};\boldsymbol{\Psi},\boldsymbol{T}\right)}{p\left(\boldsymbol{Y};\boldsymbol{\Psi}^\star,\boldsymbol{T}^\star\right)} = \max_{(\boldsymbol{\Psi}',\boldsymbol{T}')\sim(\boldsymbol{\Psi},\boldsymbol{T})} \frac{p\left(\boldsymbol{Y},\boldsymbol{z}^\star;\boldsymbol{\Psi}',\boldsymbol{T}'\right)}{p\left(\boldsymbol{Y},\boldsymbol{z}^\star;\boldsymbol{\Psi}^\star,\boldsymbol{T}^\star\right)}\left[1+o_P(1)\right]+o_P(1)$$

où les o_P sont uniformes sur l'espace des paramètres et des ruptures et $(\boldsymbol{\Psi}',\boldsymbol{T}') \sim (\boldsymbol{\Psi},\boldsymbol{T})$ signifie que le maximum est fait sur toutes les paires égales au *label switching* près.

Comme pour le modèle des blocs latents, ce théorème implique la consistance des estimateurs du maximum de vraisemblance.

4.1.3 Application aux données de consommation électrique

Dans notre article Brault et al. (2024a), nous testons l'algorithme sur différentes configurations, le comparons à d'autres méthodes et appliquons sur les données de consommation électrique en France durant l'année 2020 [2] suite à une discussion que j'ai eue avec la représentante d'Enedis dans le cadre d'un Challenge de Dataviz fait au sein du département STID où j'enseigne. Par contre, nous ne proposons pas d'utiliser directement le modèle (4.3) mais plutôt d'étudier des courbes fonctionnelles (voir la figure 4.2) dont nous projetons le comportement sur des bases d'ondelettes (cette partie est surtout le travail d'Emilie Devijver et de Charlotte Laclau).

Dans l'exemple de la figure 4.2, nous simulons $n = 100$ courbes sinusoïdales bruitées appartenant chacune à l'un des $K^\star = 3$ clusters de façon équiprobables ($\boldsymbol{\pi}^\star = (1/3, 1/3, 1/3)$) suivant la fonction suivante :

$$\forall k \in \{1,2,3\}, \forall \ell \in \{0, \dots, L^\star_k = k\}, f_{k,\ell}\left(t_j\right) = \sum_{\ell=0}^{k} (-1)^k \cos\left(\frac{2\pi}{1+\ell}\right) \mathbb{1}_{\left\{\frac{\ell J}{k}+1 \leq j \leq \frac{(\ell+1)J}{k}\right\}}$$

avec $t_j \in [0,1]$ dont nous prenons 32 valeurs équiréparties dans l'intervalle pour notre exemple et nous bruitons avec une gaussienne centrée réduite. Ainsi, nous voyons sur le haut de la figure que les courbes ont d'abord des périodes de la longueur d'une journée puis sont de plus en plus réduites ruptures après ruptures. Nous prenons ensuite les courbes et les projetons sur une base d'ondelettes (ici $H = 3$). Sur la figure 4.2, nous voyons à chaque fois les courbes par clusters et la projection sur les trois dimensions

2. Les données sont disponibles sur l'adresse suivante https://data.enedis.fr/pages/accueil/?id=init

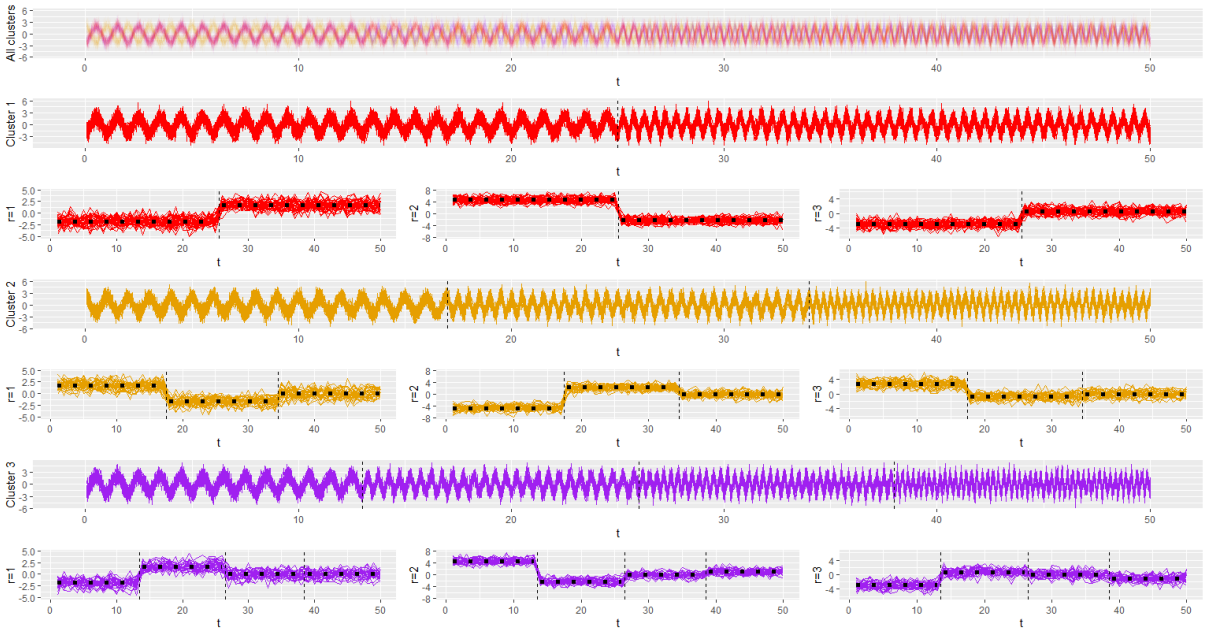

FIGURE 4.2 – Exemple illustrant l'article Brault et al. (2024a). La ligne du haut représente les données fonctionnelles initiales constituées de $n = 100$ individus (courbes) observés toutes les demi-heures pendant 50 jours. Les lignes suivantes permettent de visualiser la décomposition de la population en clusters (ici $K^\star = 3$ clusters - rouge, jaune, violet), à l'intérieur de chaque cluster la segmentation obtenue sur la dimension temporelle, la projection sur une base d'ondelettes à plusieurs niveaux $h \in \{1, 2, H = 3\}$ (dans l'article donc sur la figure, h est remplacé par r).

(comme l'image est issue de l'article, h est représenté par r sur la figure). Nous pouvons voir que, dans ce cas, les projections semblent être modélisées par le modèle (4.3).

Nous analysons les consommations électriques regroupées suivant le croisement du type de contrat, de la région et du profil utilisateurs (pour un total de 984 courbes) toutes les trente minutes durant toute l'année 2020. Après nettoyage, nous conservons $n = 889$ individus et nous projetons les courbes sur une base de Haar de telle sorte que nous ayons $J = 52$ semaines représentées chacune par $H = 42$ variables. L'un de nos objectifs est de voir s'il y a une influence des confinements sur la consommation électrique.

Pour le choix du nombre de clusters et de ruptures par cluster, nous adaptons le critère *BIC* proposé par Zhang et Siegmund (2007) dans un contexte similaire. Dans notre cas, nous axiomatisons la formule suivante :

$$\begin{aligned}\mathrm{BIC}(K, \boldsymbol{L}) \quad = \quad & \max_{\boldsymbol{\Psi}, \boldsymbol{T}} \log p\left(\boldsymbol{Y}; K, \boldsymbol{T}, \boldsymbol{\Psi}\right) - \underbrace{\frac{K-1}{2}\log(n)}_{\text{pour } \boldsymbol{\pi}} \\ & \underbrace{-\frac{1}{2}\sum_{k=1}^{K}\left[3H(L_k+1)\log(nJH) + \sum_{\ell=0}^{L_k}\log\left(\frac{\widehat{T}_{k,\ell+1} - \widehat{T}_{k\ell}}{J}nH\right)\right]}_{\text{pour } \boldsymbol{\mu} \text{ et } \boldsymbol{T}} \\ & \underbrace{-\frac{K}{2}\log(nJH)}_{\text{pour } \boldsymbol{\sigma}}.\end{aligned}$$

où $\boldsymbol{L}$ est le vecteur du nombre de ruptures. Comme il est computationnellement long de tester tous les cas possibles (puisque d'une complexité de l'ordre de $\mathcal{O}\left(L_{\max}^{K_{\max}}\right)$ pour le nombre de modèles testés à multiplier par la complexité trouvée dans la proposition 13), nous adoptons la procédure *bikm1* proposée par Robert (2021) qui explore toutes les possibilités en changeant à chaque fois une valeur entre le nombre de clusters et le nombre de ruptures et qui continue à l'étape suivante en prenant le modèle qui maximise le critère choisi (ici, le critère *BIC*). Au final, nous obtenons trois clusters avec respectivement $L_1 = 1$, $L_2 = 2$ et $L_3 = 6$ ruptures et contenant respectivement 559, 231 et 1 individus. Sur la figure 4.3, nous

TABLE 4.1 – Tableau de contingence entre les clusters et les profils *Enedis*. *ENT/PRO* correspond à des contrats de professionnels et *RES* à des contrats de particuliers (les chiffres des contrats correspondent à différents profils établis par *Enedis*).

		ENT/PRO	RES1, 3 or 4	Other RES
Cluster	1	341	36	182
	2	0	0	231
	3	0	0	1

représentons les emplacements des ruptures par groupe (lignes verticales en pointillés long) que nous comparons avec les dates de début et fin de confinement (en pointillés fins).

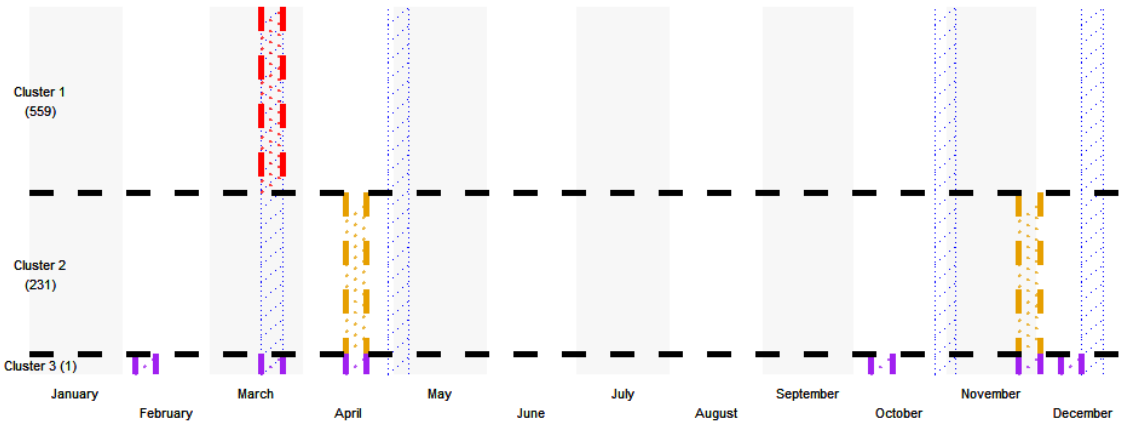

FIGURE 4.3 – Distribution des observations entre les groupes (lignes horizontales en pointillés) et localisation des moments de rupture au sein de chaque groupe (lignes verticales en pointillés long - rouge pour le cluster 1, jaune pour le 2, violet pour le 3). Les pointillés fins verticaux correspondent aux semaines de début et de fin des périodes de confinement. Figure issue de l'article Brault et al. (2024a).

Nous pouvons voir que les ruptures du groupe trois sont aux mêmes endroits que celles des groupes 1 et 2. Certaines ruptures sont en même temps que les débuts ou les fins de confinement mais cela n'est pas net et pourrait être dû à un autre effet (par exemple, la météo). Quand on regarde la distribution des contrats dans les groupes (voir le tableau 4.1), nous voyons que les clusters 2 et 3 ne contiennent que des profils de type *résidentiel*.

Enfin, nous regardons la distribution des régions mais ne trouvons rien de particulier.

Pour le moment, nous n'avons pas eu de retours de la part d'Enedis pour savoir si nos résultats leur semble cohérents ou non.

Perspectives

Dans le prolongement de ce modèle, il serait intéressant d'optimiser la procédure pour diminuer le temps de calcul. Notamment, nous avions testé sur les données journalières mais les temps de calculs étaient trop longs. À nouveau, une perspective intéressante serait de voir si nous pouvons adapter le travail de Rigaill (2015).

Perspectives

Une autre perspective serait sur la sélection de modèle. Il serait intéressant de voir si le critère *BIC* est cohérent ou, comme nous avons peu de groupes sur les données réelles, trop conservateur. Dans la même idée, j'aimerais avoir le retour d'*Enedis* pour la cohérence du modèle retenu ou voir parmi les autres modèles avec un critère *BIC* proche s'il y a des estimations plus pertinentes.

4.2 Protéines Tau et mélange de courbes avec sauts

Dans le cadre du projet *Taunique* avec Emilie Lebarbier (Université Paris Nanterre) et Virginie Stoppin-Mellet (Institut Neuroscience ; UGA) puis depuis peu Amélie Rosier (ESME Sudria Paris), nous travaillons sur une application en neuroscience. N'étant pas un spécialiste de l'application, je vais me concentrer sur les points importants pour la modélisation. La protéine Tau est présente dans le cerveau et des dysfonctionnements de sa part impliquent des pathologies comme la maladie d'Alzheimer. Virginie Stoppin-Mellet et son équipe étudient le contrôle qu'a la protéine Tau sur la dynamique des microtubules. Les données que nous étudions proviennent d'une expérience proposée par Elie et al. (2015) et représentent des profils d'intensité lumineuse au cours du temps.

En théorie, ces profils sont censés être des fonctions en escaliers décroissantes avec zéro, un ou deux sauts de même hauteur qu'importe le profil (voir la figure 4.4) ; les deux sauts peuvent être au même moment.

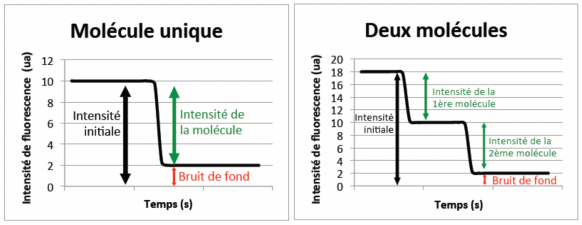

FIGURE 4.4 – Exemples de deux profils théoriques avec un saut (à gauche) et deux sauts distincts (à droite). Image issue de l'acte de conférence de Rosier et al. (2023).

L'objectif est donc de classifier les profils en quatre groupes (aucun saut, un saut, deux sauts au même moment ou à deux instants différents) en sachant que la hauteur (inconnue) d'un saut est censée être la même pour tous les profils et que la valeur initiale dépend du profil. Dans la suite de cette section, je vais expliciter le modèle que nous avons choisi (voir la section 4.2.1), la solution proposée pour estimer les paramètres (section 4.2.2) et les premiers résultats sur simulations (section 4.2.3).

4.2.1 Modélisation

Pour la modélisation (voir le graphe bayésien de la figure 4.5), nous supposons avoir n signaux de taille J_i pour le signal i appartenant à $K^\star = 4$ groupes chacun avec probabilité $\pi_k^\star$, une valeur de saut $\delta^\star < 0$ et une matrice $\boldsymbol{z}^\star$ de taille $n \times 4$ telles que :

$$\forall i \in \{1, \dots, n\},\ \boldsymbol{Y}_{i,\bullet} | z_{i,k}^\star = 1 \sim \mathcal{N}\left(\boldsymbol{\theta}_{i,k}^\star, \sigma_i^{\star,2} \mathbb{1}_{J_i}\right) \tag{4.4}$$

avec $\boldsymbol{\theta}_{i,k}^\star$ le vecteur moyenne de $J_i \times 1$ valant pour tout $j \in \{1, \dots, J_i\}$:

$$\boldsymbol{\theta}_{i,k}^\star(j) = \begin{cases} \mu_i^\star & \text{si } k = 1, \\ \mu_i^\star + \delta^\star \mathbb{1}_{\left\{j > T_{2,1}^{\star,(i)}\right\}} & \text{si } k = 2, \\ \mu_i^\star + \delta^\star \mathbb{1}_{\left\{j > T_{3,1}^{\star,(i)}\right\}} + \delta^\star \mathbb{1}_{\left\{j > T_{3,2}^{\star,(i)}\right\}} & \text{si } k = 3, \\ \mu_i^\star + 2\delta^\star \mathbb{1}_{\left\{j > T_{4,1}^{\star,(i)}\right\}} & \text{si } k = 4, \end{cases}$$

où $T_{2,1}^{\star,(i)}$ (resp. $T_{4,1}^{\star,(i)}$) est le moment du saut (resp. double) pour l'individu i s'il n'y en a qu'un seul, $T_{3,1}^{\star,(i)}$ est le premier moment de saut et $T_{3,2}^{\star,(i)} > T_{3,1}^{\star,(i)}$ le deuxième moment s'il y en a deux (voir la figure 4.6 pour des exemples de courbes théoriques).

Remarque

Comme les numéros de cluster sont attribués à une caractéristique particulière, il n'y a pas de problème de label switching dans ce modèle.

4.2.2 Estimation

Pour l'estimation, nous utilisons un algorithme *EM* où l'étape *E* s'effectue à l'aide des probabilités d'appartenance aux classes dont nous avons une formule explicite. Par contre, la difficulté est dans l'étape

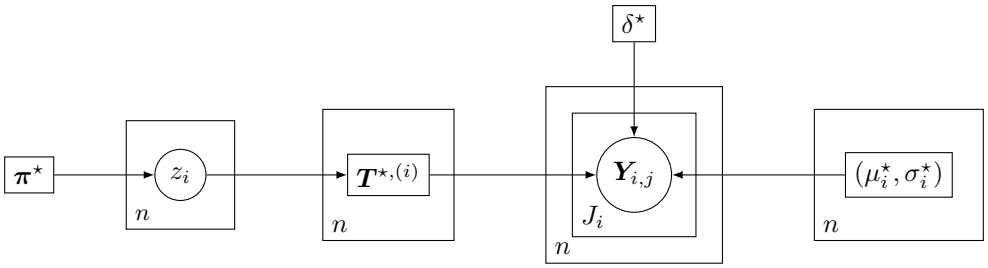

FIGURE 4.5 – Représentation graphique du modèle de mélange de courbes avec sauts : les carrés correspondent à des paramètres et les ronds à des variables.

M car la valeur de $\delta^\star$ est commune à tous les individus, mais les instants de sauts sont propres à chaque individu et les lois avant et après les ruptures ne sont pas indépendantes. Ainsi, il n'est pas possible d'utiliser l'algorithme de programmation dynamique.

Pour contourner ces problèmes, nous décidons de faire une maximisation par alternance :

- Étant connue une estimation des emplacements des ruptures, nous estimons la valeur de $\delta^\star$.
- Étant donnée une estimation de $\delta^\star$, nous estimons les emplacements des ruptures, les moyennes et la variance pour chaque individu comme s'il appartenait à chacun des clusters. Pour cela, nous testons toutes les possibilités d'emplacements de sauts et nous conservons la configuration maximisant la vraisemblance.

En pratique, nous pouvons choisir de faire un nombre fini d'itérations pour l'étape de maximisation ou utiliser un critère pour savoir quand s'arrêter. Dans tous les cas, nous obtenons la complexité suivante :

Proposition 16 (Complexité de l'algorithme *EM*)

Si nous notons $J_{\max}$ le signal le plus long, N_{algo} le nombre d'itérations faites par l'algorithme et N_M le nombre d'itérations dans l'étape de maximisation alors la complexité de l'algorithme est $\mathcal{O}\left(N_{\text{algo}} N_M n J_{\max}^2\right)$.

Dans Rosier et al. (2023), nous comparons les qualités d'estimation si nous faisons $N_M = 10$ itérations dans l'étape M avec une seule initialisation ou 10 initialisations avec une seule itération ($N_M = 1$) dans l'étape M et en conservant l'estimation maximisant la vraisemblance. Les estimations (clusters, sauts et emplacements des sauts) sont plus proches des vraies valeurs lorsque nous prenons $N_M = 1$ mais avec plus d'initialisations (voir les figures de Rosier et al. (2023)).

4.2.3 Premiers résultats

Dans Rosier et al. (2023), nous testons la qualité des estimations des paramètres, distances et classifications en prenant un bruit de 1 pour tous les individus, la même longueur $J_i = 100$ de signal et en ne changeant que la hauteur $\delta^\star \in \{-0.5, -1, -2, -5\}$ des sauts et la distance entre les deux sauts du groupe 3. De manière générale, nous remarquons que $|\delta^\star|$ est sous-estimé (le saut estimé est plus petit que le vrai saut). Les estimations sont très bonnes lorsque le vrai saut est grand mais très mauvaise pour des petits sauts. Lorsque nous regardons la répartition des clusters d'appartenances estimés en fonction des clusters d'origines (voir la figure 4.7), nous voyons que le cluster 3 tend à absorber tous les clusters.

Lorsque nous regardons les emplacements des ruptures estimées pour le cluster 3 quand les individus appartenaient au cluster 1 ou au cluster 2, nous nous apercevons que l'algorithme met le saut ou les sauts supplémentaires proche du début du signal ou au contraire à la fin ; là où les mauvaises estimations des moyennes auront moins d'impact. À l'opposé, pour les signaux issus du cluster 4 (le double saut), l'algorithme met les deux sauts proches du vrai saut.

Perspectives

Dans Brault et al. (2025), nous cherchons à comprendre de façon théorique pourquoi le cluster 3 semble absorbant afin de trouver une façon de contrer ce problème. De plus, nous regardons empiriquement l'influence d'une mauvaise estimation du paramètre $\delta^\star$ en prenant des valeurs plus petites et plus grandes

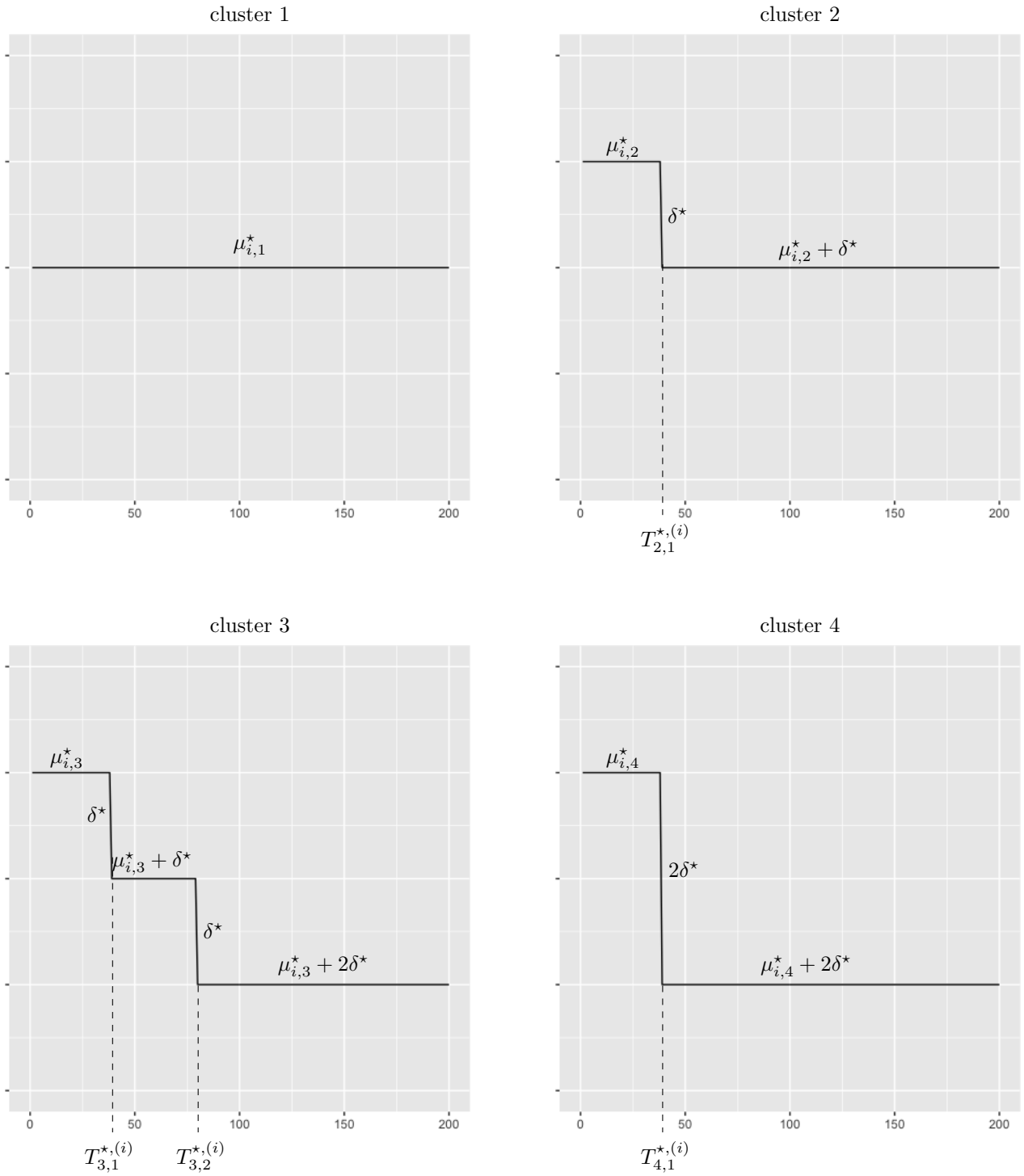

FIGURE 4.6 – Représentation des courbes moyennes d'un individu suivant son appartenance aux cluster : aucun saut dans le cluster 1 (en haut à gauche), un saut simple dans le cluster 2 (en haut à droite), deux sauts simples dans le cluster 3 (en bas à gauche) et un saut double daus le cluster 4 (en bas à droite). Images provenant de Rosier et al. (2023).

pour l'étape *M*. Nous sommes également en train de tester l'algorithme sur les données fournies par Virginie.

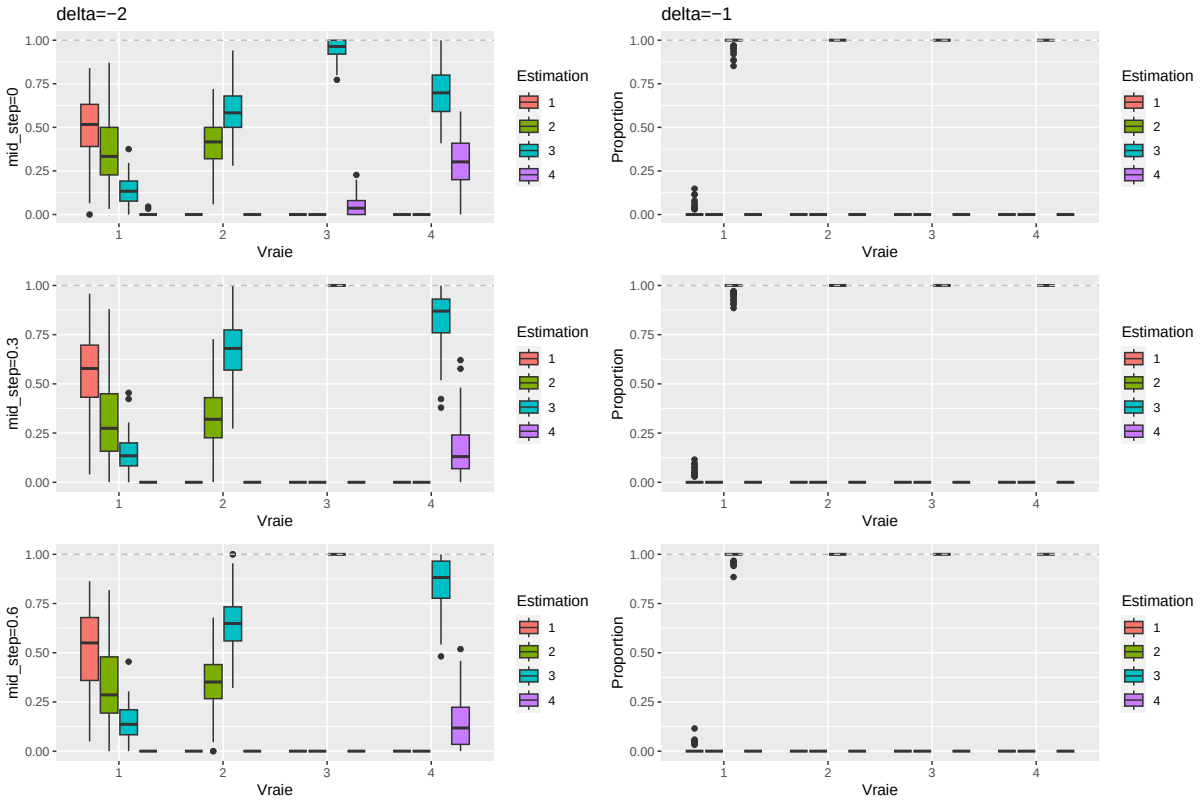

FIGURE 4.7 – Boxplot des proportions des estimations des clusters d'appartenances estimés (couleur) en fonction du cluster d'origine (abscisse), de la valeur de $\delta^\star$ (colonnes) et de la longueur du plateau central pour le cluster 3 (ligne). Images provenant de Rosier et al. (2023).

4.3 Modélisation statistique de l'évolution de l'écriture chez l'enfant

Depuis mon arrivée à Grenoble, je travaille avec Caroline Jolly (LPNC) et d'autres partenaires sur la modélisation statistique de l'évolution de l'écriture chez l'enfant (projet MSE^3). L'un des objectifs est d'essayer de proposer un outil d'aide au diagnostic pour la détection de la dysgraphie chez l'enfant. À nouveau, n'étant pas un spécialiste de la dysgraphie, je vais me contenter de présenter quelques points clefs pour comprendre ma participation à ce projet. D'abord, il est à noter que la dysgraphie est un trouble qui affecte l'écriture et son tracé [3] et qu'il existe différents types de dysgraphies comme celles portant sur la lisibilité de l'écriture ou celles portant sur la vitesse. Ensuite, en France, la détection de ces troubles se fait généralement à l'aide du test *Brave Handwriting Kinder* (ou *BHK* ; voir Hamstra-Bletz et al. (1987) et son adaptation française par Charles et al. (2004)) consistant à faire écrire des enfants pendant 5 minutes et à faire évaluer ce texte selon 13 critères par un spécialiste en psychomotricité. Notre objectif est d'aider à détecter des enfants dysgraphiques en les faisant écrire sur des tablettes et utiliser la dynamique de l'écriture plutôt que le texte final.

Au début de la collaboration, nous possédions une base de données réalisées par Caroline Jolly qui possédait quelques limites comme le fait que les enfants provenant des écoles n'avaient pas de diagnostics de poser (donc nous ne savions pas s'ils étaient dysgraphiques ou non) et surtout que les enfants diagnostiqués dysgraphiques et suivis au CHU de Grenoble avaient écrit sur des tablettes différentes (à la fois plus grande et avec des calibrations différentes sur certains paramètres comme la pression). Ces problèmes ont d'ailleurs induit des erreurs d'interprétation de la part de Asselborn et al. (2018) nous obligeant à publier un *comment paper* sur ce travail (voir Deschamps et al. (2019)).

3. Définition prise sur le site https://www.tousalecole.fr/content/dysgraphies référencé par le site de l'INSERM sur les troubles de l'apprentissage https://www.inserm.fr/dossier/troubles-specifiques-apprentissages/

Nous avons donc fait plusieurs demandes de financement (les projets MSE^3 dont je suis porteur et le projet *Ecriture CEA 17P170* et *Inter Carnot Ecriture* où je suis co-porteur) afin de créer une base de données de 545 enfants âgés entre 6,5 et 16 ans qui ont tous été diagnostiqués (66 enfants sont dysgraphiques soit environ 12% de la base de données) à qui Caroline Jolly a demandé de faire différentes activités :

- Le test *BHK* (utilisé pour le diagnostique par des professionnels).
- L'écriture individuelle des vingt six lettres et dix chiffres (dans un ordre aléatoire pour éviter l'anticipation).
- Des figures plus ou moins complexes que les enfants devaient reproduire ou des circuits qu'ils devaient suivre avec leur stylo (voir la figure 4.8).

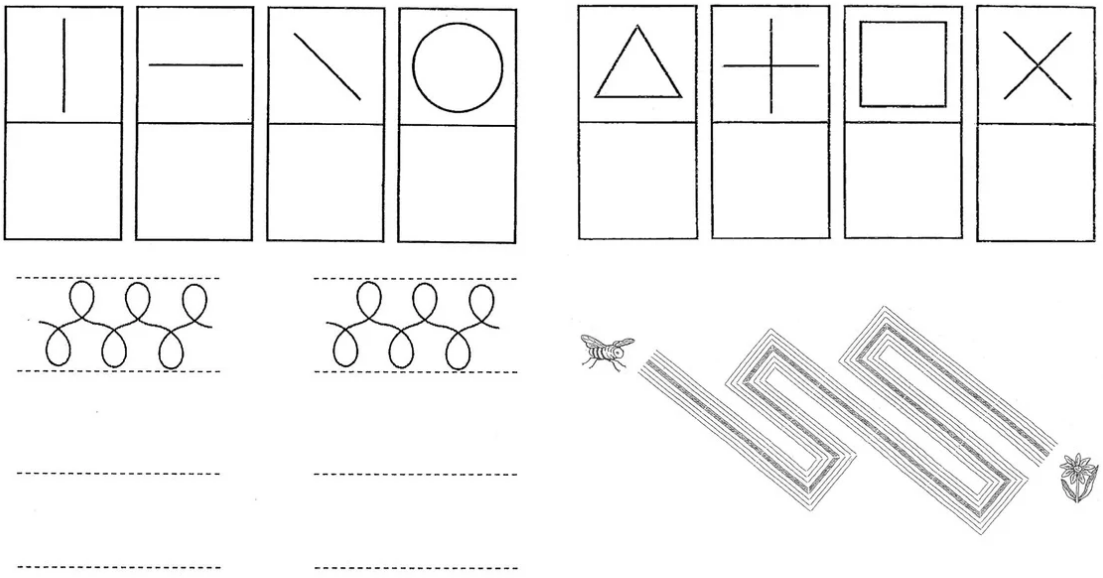

FIGURE 4.8 – Exemples de figures à reproduire et de circuits à suivre donnés aux enfants. Images provenant de l'article Devillaine et al. (2021).

Les premiers travaux relatifs à cette base (dont l'essentiel ont été faits par les partenaires du CEA) sont :

- l'application de méthodes de classification supervisée sur les textes (voir Deschamps et al. (2021)) en utilisant des caractéristiques utilisés dans le test *BHK* et des caractéristiques dynamiques de certains enfants dysgraphiques (comme le fait de se reprendre plusieurs fois pour écrire une lettre)
- l'utilisation des figures à reproduire et des circuits (voir Devillaine et al. (2021)) afin de proposer une détection qui soit indépendante de la langue et du fait de connaître les lettres ou non.

Toutefois, la partie du projet la plus intéressante pour ce manuscrit porte sur les travaux utilisant le modèle *POMH* (pour *Parsimonious Oscillatory Model of Handwriting*) entamés par Yunjiao Lu, post-doc financée grâce à l'*Inter Carnot Ecriture* que j'ai co-encadrée. Dans la suite de cette partie, je commence par présenter les données (voir la section 4.3.1) et le modèle *POMH* (section 4.3.2) puis je parlerais des premiers résultats obtenus avec Yunjiao Lu dont une publication vient d'être soumise (voir Lu et al. (2024) et la section 4.3.3) et de la version utilisant la programmation dynamique que je propose (section 4.3.4).

4.3.1 Présentation des données

Comme expliqué précédemment, chaque enfant a fait plusieurs tâches : ainsi, notre individu statistique est le croisement enfant×tâche. Pour chaque individu statistique, nous avons à chaque instant $t_j \in \{t_1, \dots, t_{J_i}\}$ l'abscisse (que nous notons $x(t_j)$) et l'ordonnée ($y(t_j)$) ainsi que l'information si le stylo touche la feuille mise sur la tablette ou pas, l'âge de l'enfant, sa classe et s'il est dysgraphique ou non.

Point de vigilance

Dans cette partie, je fais une entorse aux notations proposées durant tout ce manuscrit et je m'autorise à reprendre certaines qui ont déjà été utilisées même si elles ne représentent pas des éléments comparables le temps d'expliquer le modèle *POMH* puisque nous étudions le tracé $(x(t_j), y(t_j))_{j\in\{1,\dots,J_i\}}$.

4.3.2 *Parsimonious Oscillatory Model of Handwriting*

Le **modèle *POMH*** (pour *Parsimonious Oscillatory Model of Handwriting*) a été développé par André et al. (2014) pour modéliser les écritures fluides à l'aide de deux oscillateurs orthogonaux sur les vitesses. Dans notre cas, étant donnée une courbe $(x(t), y(t))_{t\in\{1,\dots,J_i\}}$, nous supposons qu'il existe six fonctions constantes par morceaux a_x, a_y, ω_x, ω_y , ϕ_x et ϕ_y telles que pour tout instant $t \in]0; t_{J_i}]$, nous ayons :

$$\left\{\begin{array}{rcl} \dot{x}(t) & = & a_x(t)\sin\left[\omega_x(t)\times t + \phi_x(t)\right] \\ \dot{y}(t) & = & a_y(t)\sin\left[\omega_y(t)\times t + \phi_y(t)\right] \end{array}\right. \tag{4.5}$$

où $\dot{x}$ (resp. $\dot{y}$) est la vitesse en abscisse x (resp. en ordonnée y) et les fonctions a_x, ω_x et ϕ_x (resp. a_y, ω_y et ϕ_y) changent de paliers aux mêmes instants. Nous appelons :

- a_x et a_y les **amplitudes**,
- ω_x et ω_y les **fréquences**,
- ϕ_x et ϕ_y les **déphasages**.

Sur la figure 4.9, nous représentons un exemple de fonctions a_x, ω_x, ϕ_x, a_y, ω_y et ϕ_y ainsi que les vitesses qu'elles induisent par le modèle (4.5) qui conduisent à la lettre *o* dessinée à droite.

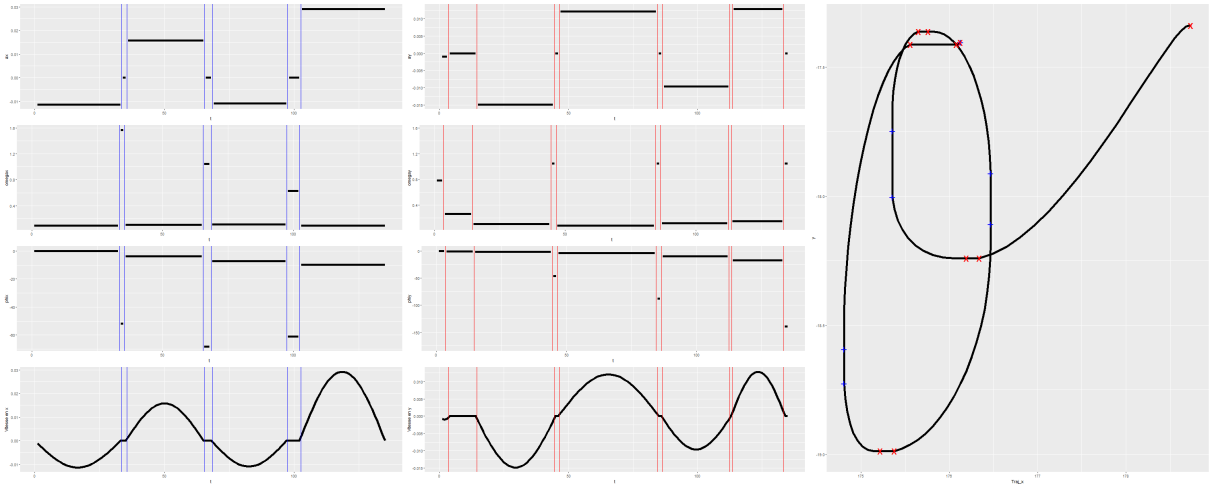

FIGURE 4.9 – À gauche et de haut en bas, exemples de fonctions a_x, ω_x et ϕ_x et les version en y au milieu ainsi que la fonction vitesse $\dot{x}$ (resp. $\dot{y}$) obtenue par l'équation (4.5) tout en bas de chaque colonne. À droite, nous avons représenté la lettre *o* formée à partir des deux fonctions vitesses ; les *plus* bleus + correspondent aux moment d'annulation de la vitesse en x et les *fois* rouges × aux moments d'annulation de la vitesse en y.

Dans leur section 2.3.2, André et al. (2014) proposent une estimation des fonctions du modèle (4.5) ; pour ce faire, nous prenons $\boldsymbol{T}_x = (T_{x,0}, T_{x,1}, \dots, T_{x,K_x+1})$ les instants où la vitesse observée s'annule (en bleu sur la figure 4.9) avec $T_{x,0} = 0$ et $T_{x,K_x+1} = J_i$ et nous estimons pour tout $k \in \{0, \dots, K_x\}$ et pour tout $t \in \left]t_{T_{x,k}}; t_{T_{x,k+1}}\right]$:

$$\left\{\begin{array}{rcl} \widehat{\omega}_x(t) & = & \frac{\pi}{t_{T_{x,k+1}} - t_{T_{x,k}}} \\ \widehat{\phi}_x(t) & = & -\frac{\pi t_{T_{x,k}}}{t_{T_{x,k+1}} - t_{T_{x,k}}} \\ \widehat{a}_x(t) & = & \frac{\pi}{2\left(t_{T_{x,k+1}} - t_{T_{x,k}} - 1\right)} \sum_{j=T_{x,k}+1}^{T_{x,k+1}} \dot{x}^{(\text{obs})}\left(t_j\right) \end{array}\right. \tag{4.6}$$

où $\dot{x}^{(\text{obs})}$ est la vitesse observée sur les données. De même, les formules pour la vitesse sur y sont les mêmes que dans l'équation (4.6) mais sur la vitesse en y.

Toutefois, la recherche des instants d'annulation de la vitesse $\boldsymbol{T}_x$ et $\boldsymbol{T}_y$ peut être compliquée car nous n'observons pas directement ces dernières. L'idée est donc de faire une estimation empirique par la formule :

$$\forall j \in \{1, \dots, J_i - 1\}, \ \frac{x(t_{j+1}) - x(t_j)}{t_{j+1} - t_j}. \tag{4.7}$$

Sur la figure 4.10, nous voyons l'estimation empirique de la vitesse pour une lettre a. Le signal semble très bruité (en haut) et savoir précisément quand la vitesse s'annule (en bas) est compliqué car nous voyons parfois plusieurs annulations dans des intervalles très petits et nous pouvons nous interroger si les valeurs non nulles sont réelles ou parfois des artefacts.

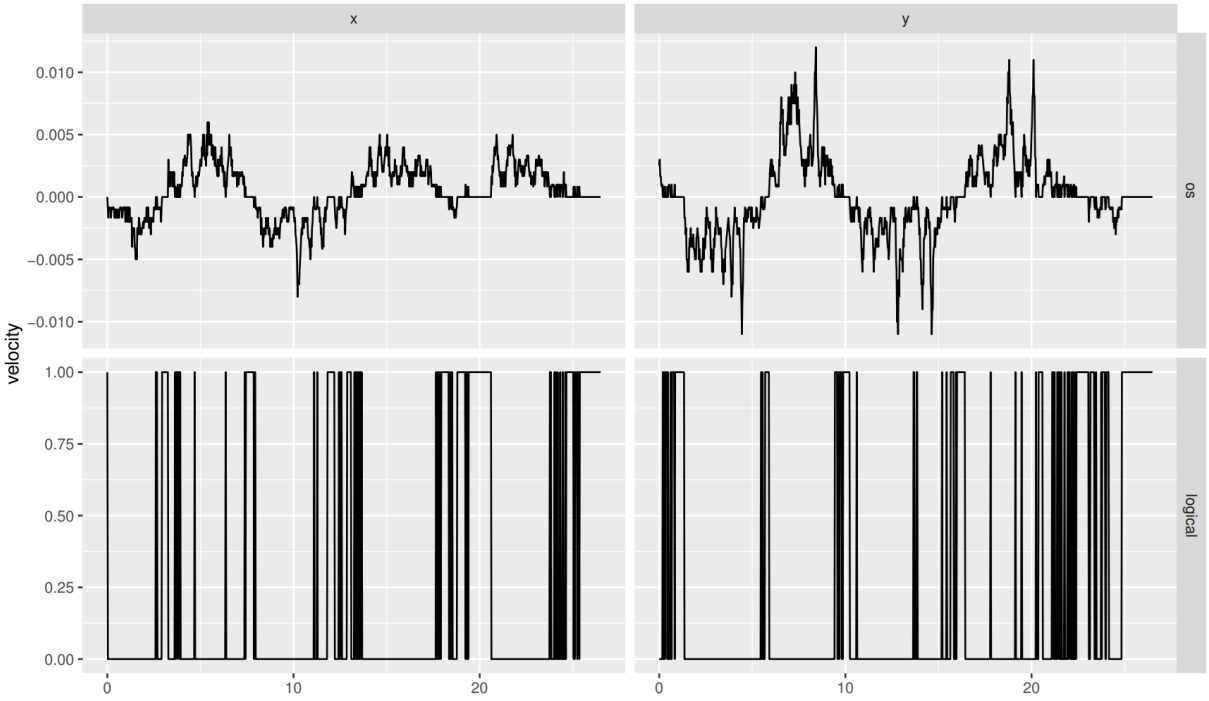

FIGURE 4.10 – Représentation en haut (*os*) d'une estimation d'une vitesse des abscisses x (à gauche) et ordonnées y (à droite) suivant la formule (4.7) d'une lettre a avec, en bas (*logical*), les instants où la vitesse empirique est nulle (`logical=TRUE`) ou non. Image issue de l'article Lu et al. (2024).

L'idée des deux sections suivantes est que si le modèle *POMH* est calibré pour représenté l'écriture *fluide*, il pourrait avoir un comportement atypique sur des écriture d'enfants *dysgraphiques*. Le problème est d'estimer de la façon la plus précise les paramètres du modèle (4.5). Dans ce cadre, nous avons obtenu le financement *Inter Carnot Ecriture* qui a permis d'avoir Yunjiao Lu en post-doc. Dans un premier temps, elle a choisi d'aborder le problème en débruitant le signal l'opérateur morphologique de fermeture (agissant comme un filtre passe-bas) (section 4.3.3). Je présenterai ensuite dans la section 4.3.4 la deuxième approche que nous lui avions proposée avec Jean-Charles Quinton et que nous explorons actuellement.

4.3.3 Modèle *POMH* et détection de la dysgraphie

Dans l'article Lu et al. (2024), nous nous concentrons uniquement sur les lettres et les chiffres (que nous appelons **symbole** par la suite). Dans cet article, nous proposons de débruiter le signal à l'aide d'un opérateur morphologique de fermeture w. Sur la figure 4.11, nous représentons le résultat pour deux opérateurs ($w = 3$ à gauche et $w = 35$ à droite). Nous voyons que plus l'opérateur est grand, plus les plages où la vitesse est nulle sont longues et donc moins il y a de zéros.

Dans un premier temps, nous calibrons donc l'opérateur $\widehat{w}$ optimal pour chaque individu statistique. Pour ce faire et étant donné un symbole $S \in \{"a", "b", \dots, "z", "0", "1", \dots, "9"\}$, nous appliquons pour

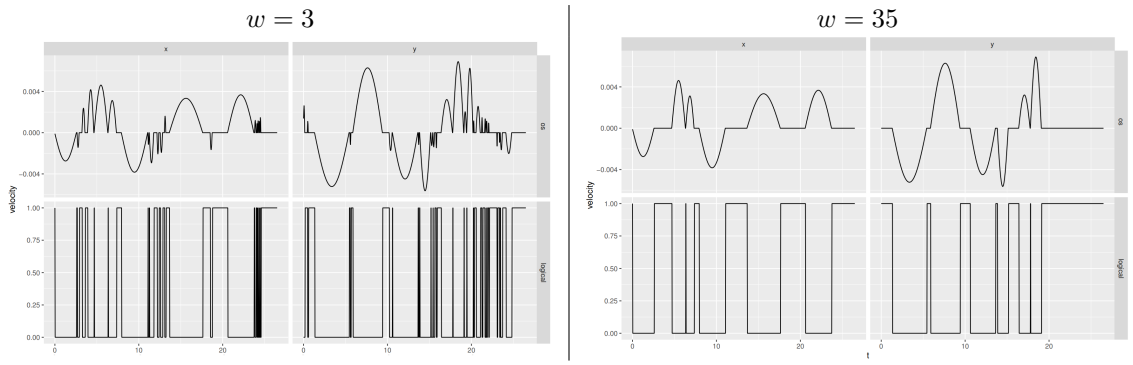

FIGURE 4.11 – Représentation de la courbe de la figure 4.10 après l'application de l'opérateur de fermeture w valant 3 (à gauche) et 35 (à droite). Image issue de l'article Lu et al. (2024).

chaque enfant les opérateurs $w \in \{3, 5, \ldots, w_{\max}\}$ (nous étudions $w_{\max} \in \{23, 25, \ldots, 39\}$) et regardons le nombre $n_{x,i,S}(w)$ d'intervalles d'annulation de la vitesse obtenus (pour les abscisses x, pour un individu i et pour un symbole S). Sur la figure 4.12, nous représentons le nombre $n_{x,i,"a"}(w)$ d'intervalles d'annulation de la vitesse en abscisse pour la lettre a.

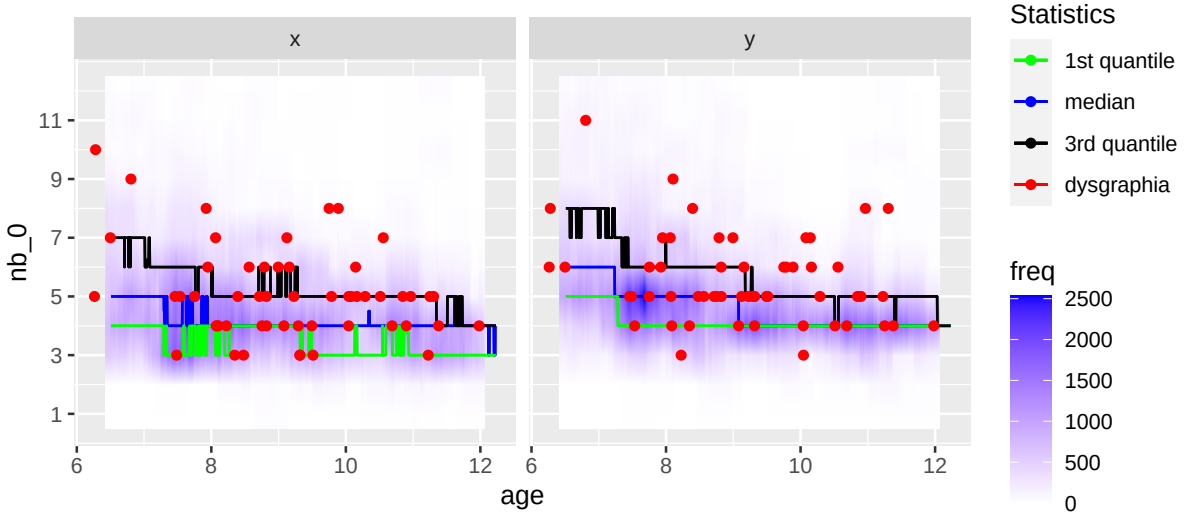

FIGURE 4.12 – Représentation sous forme de *heat map* de la distribution du nombre $n_{x,i,"a"}(w)$ d'intervalles d'annulation de la vitesse obtenus en fonction de l'âge pour des opérateurs w. Les traits correspondent aux quartiles en fonction de l'âge et les points rouges (•) aux médianes des nombres $n_{x,i,"a"}$ pour chaque enfant dysgraphique. Image issue de l'article Lu et al. (2024).

Nous voyons sur la figure 4.12 que le nombre d'intervalles d'annulation diminue avec l'âge et que l'essentiel des valeurs tout opérateur w confondu se regroupent autour de deux valeurs à partir d'un certain âge. Nous voyons également que certaines médianes des enfants dysgraphiques (symbolisées par les points rouge •) sont au dessus des enfants du même âge tandis que d'autres se retrouvent dans le premier et le troisième quartiles. Pour les points en dessous du premier quartile, les données sont parfois incomplètes (l'enfant a écrit une partie de la lettre) ; ce qui est également une information.

Afin de chercher l'opérateur *optimal* $\widehat{w}$, nous faisons le choix de prendre un opérateur w de telle sorte que le nombre d'intervalles d'annulations soit le plus proche possible de la médiane des enfants qui ont le même âge que l'enfant étudié plus ou moins six mois (voir la section 3.3 de l'article Lu et al. (2024)). Ce choix a été fait car, si un enfant a besoin de plus d'arrêts qu'un *enfant médian* alors nous nous attendons à ce que la reconstruction soit mauvaise ; nous mesurons la différence de reconstruction suivant différentes distances. De plus, nous conservons l'information de $\widehat{w}$ pour avoir une idée de si le signal initial a besoin d'être beaucoup lissé ou non et le temps mis par l'enfant pour écrire le symbole (pour essayer de détecter les enfants dysgrahpiques par la vitesse). Au final, nous obtenons les données schématisées par le tableau 4.2.

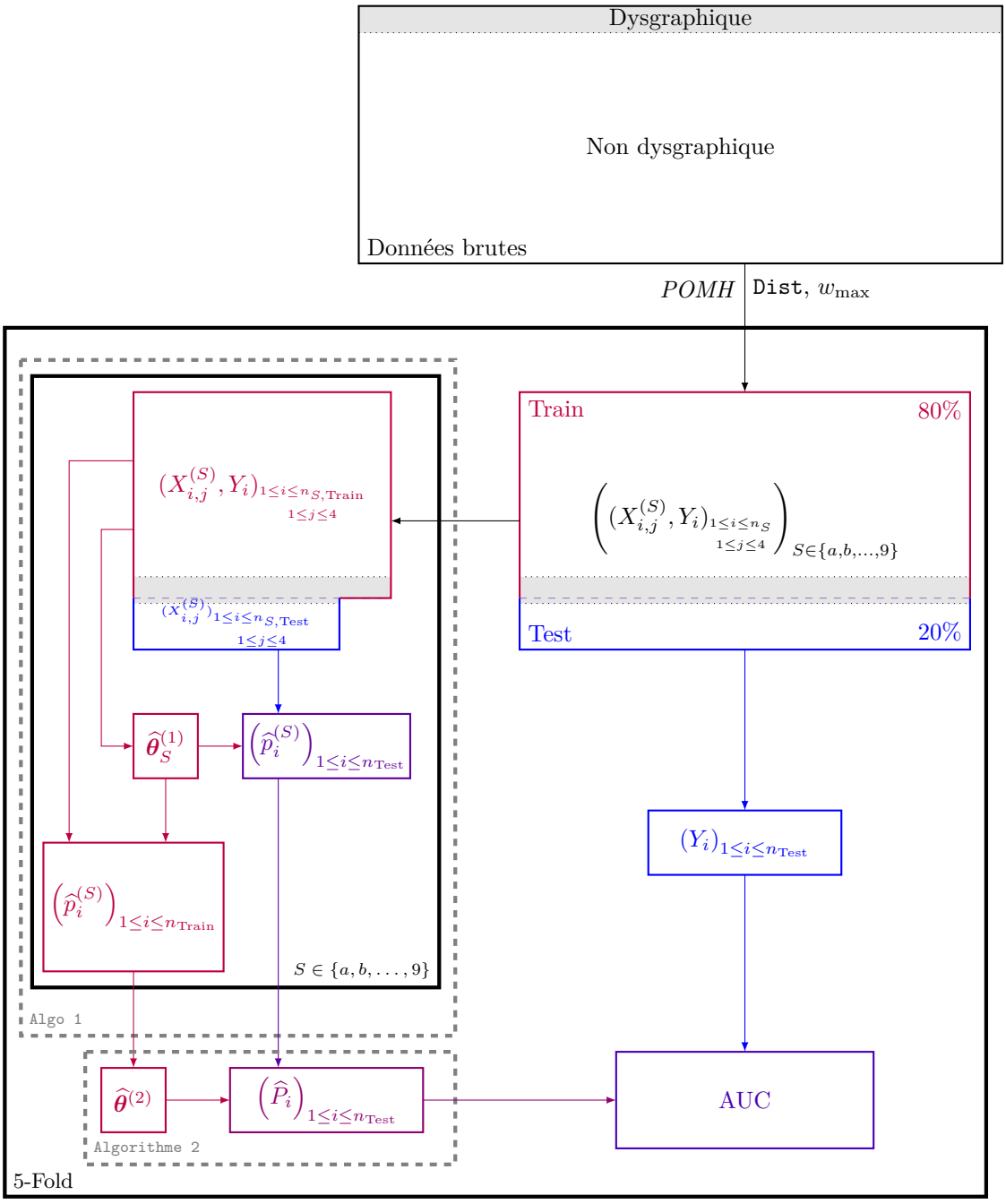

FIGURE 4.13 – Représentation schématique de la procédure utilisant deux algorithmes étant donnés une distance `Dist` et un opérateur maximal $w_{\max}$. `algorithme 1` s'applique pour chaque symbole S sur les données présentées dans la table 4.2 et est optimisé pour obtenir les paramètres $\widehat{\boldsymbol{\theta}}_S^{(1)}$ associés pour chaque symbole S qui permettent de prédire la probabilité $\widehat{p}_i^{(S)}$ que le symbole S écrit par l'enfant i provient d'un enfant dysgraphique ou non. L'`algorithme 2` s'applique sur les probabilités $\widehat{p}_i^{(S)}$ et est optimisé pour obtenir les paramètres $\widehat{\boldsymbol{\theta}}^{(2)}$ afin de prédire si un enfant est dysgraphique ou non. Enfin, les prédictions sur l'échantillon de test sont comparées aux valeurs connues afin de déterminer l'AUC c'est-à-dire l'aire sous la courbe ROC.

TABLE 4.2 – Représentation schématique des informations recueillies par enfant (en ligne) : pour chaque symbole S (groupe de colonnes), nous conservons l'opérateur $\widehat{w}$, la distance de la reconstruction et le temps mis pour écrire le symbole. A ceci, nous ajoutons la section (nous avons regroupé les classes en ordinale) et l'information de si l'enfant est dysgraphique ou non.

ID	a			b			···			9			section	dys
	w_x	dist	tt	$\widehat{w}$	dist	tt	w_x	dist	tt	w_x	dist	tt		
1	$X_{1,1}^{(a)}$	$X_{1,2}^{(a)}$	$X_{1,3}^{(a)}$	$X_{1,1}^{(b)}$	$X_{1,2}^{(b)}$	$X_{1,3}^{(b)}$		···		$X_{1,1}^{(9)}$	$X_{1,2}^{(9)}$	$X_{1,3}^{(9)}$	$X_{1,4}$	Y_1
2	$X_{2,1}^{(a)}$	$X_{2,2}^{(a)}$	$X_{2,3}^{(a)}$	$X_{2,1}^{(b)}$	$X_{2,2}^{(b)}$	$X_{2,3}^{(b)}$		···		$X_{2,1}^{(9)}$	$X_{2,2}^{(9)}$	$X_{2,3}^{(9)}$	$X_{2,4}$	Y_2
⋮		⋮			⋮						⋮		⋮	⋮
n	$X_{n,1}^{(a)}$	$X_{n,2}^{(a)}$	$X_{n,3}^{(a)}$	$X_{n,1}^{(b)}$	$X_{n,2}^{(b)}$	$X_{n,3}^{(b)}$		···		$X_{n,1}^{(9)}$	$X_{n,2}^{(9)}$	$X_{n,3}^{(9)}$	$X_{n,4}$	Y_n

À partir de ces données, nous proposons une méthode en deux temps (voir la figure 4.13) pour prédire la variable de dysgraphie :

1. Étant donnés une distance et un opérateur $w_{\max}$ maximal, nous créons le tableau de données comme présenté dans la table 4.2.
2. Nous faisons 5 fois une validation croisée c'est-à-dire que nous divisons les données en un ensemble d'apprentissage (*Train*) et un ensemble de test (*Test*) en conservant la même proportion d'enfants dysgraphiques :
   (a) Sur l'échantillon d'entraînement et pour chaque symbole $S \in \{"a", "b", \ldots, "z", "0", "1", \ldots, "9"\}$, nous optimisons un `algorithme 1` qui renvoie une probabilité $\widehat{p}_i^{(S)}$ que le symbole S de l'individu i soit fait par un enfant diagnostiqué dysgraphique ou pas.
   (b) À partir des valeurs $\left(\widehat{p}_i^{(S)}\right)_{1\leq i\leq n_{Train}\ ;\ S\in\{"a","b",\ldots,"z","0","1",\ldots,"9"\}}$, nous optimisons un `algorithme 2` pour qu'il prédise si un enfant est dysgraphique ou non.
   (c) Nous appliquons l'`algorithme 1` sur l'échantillon de test puis l'`algorithme 2` sur les sorties obtenues afin de calculer la probabilité $\widehat{P}_i$ de chaque enfant de l'ensemble de test d'être dysgraphique ou non.
   (d) Nous comparons les probabilités $\widehat{P}_i$ avec les valeurs Y_i en utilisant différents seuils afin de calculer l'aire sous la courbe *ROC* associée (pour *Receiver Operating Characteristic*) qui représente le taux de vrais positifs en fonction du taux de faux positifs.
3. Au final, nous sélectionnons le quadruplé (distance, $w_{\max}$, `algorithme 1`, `algorithme 2`) maximisant l'aire sous la courbe *ROC*.

Dans Lu et al. (2024), nous testons la méthode proposée par Zhang (2016) couplant une régression logistique généralisée et le critère *AIC* (notée `GLM`), les forêts aléatoires telles que proposées par Genuer et al. (2008) (notée `RF`), la méthode *SVM* (*Support Vector Machine*) proposée par Boser et al. (1992) avec un noyau gaussien (notée `SVM` ; uniquement pour le premier algorithme) et une méthode de comptage (notée `Comptage` ; uniquement pour le deuxième algorithme). Les combinaisons testées dans notre article Lu et al. (2024) sont résumées dans la table 4.3.

TABLE 4.3 – Résumé des combinaisons testées dans l'article Lu et al. (2024) suivant la procédure présentée dans la figure 4.13 en fonction de l'`algorithme 1` (en ligne) et de l'`algorithme 2` (en colonnes).

		algorithme 2		
		Comptage	GLM	RF
algo 1	GLM	✓	✓	
	RF	✓		✓
	SVM	✓		

Durant nos essais, cette procédure en deux couches améliore les résultats par rapport au fait de directement prédire sur la première couche. La procédure qui a obtenu la meilleure aire sous la courbe est celle utilisant la distance infinie, une valeur maximale pour l'opérateur w de $w_{\max} = 27$ et en couplant la méthode `SVM` avec un comptage avec une aire sous la courbe moyenne de 0.785. L'avantage de cette méthode est qu'elle permet de mettre en évidence les symboles plus compliqués à faire que les autres

pour les enfants. De plus, je n'ai pas abordé ce problème dans le manuscrit mais les enfants n'ont pas toujours écrit tous les symboles avec une plus grande proportion de valeurs manquantes chez les enfants dysgraphiques (voir la table 1 de notre article Lu et al. (2024)) ; ceci est pénalisé dans la méthode de comptage ce qui pourrait expliquer une plus grande réussite.

Perspectives

Dans les perspectives directes de cet article, je souhaiterais continuer de comprendre les comportements des procédures pour améliorer les qualités de détection ; en particulier, je souhaiterais voir quels sont les enfants qui ne sont pas détectés et ceux qui sont détectés à tort afin de comprendre s'il y a des pathologies associées qui sont plus faciles à détecter que d'autres. Ceci pourrait être l'occasion d'encadrements de stages voir de post-doc.

De plus, une suite normale à ces travaux serait un meta article reprenant les codes des articles Deschamps et al. (2021), Devillaine et al. (2021) et Lu et al. (2024) afin de voir si la combinaison de ces trois procédures permettrait d'améliorer l'estimation.

4.3.4 Modèle *POMH* et segmentation

Dans un travail en cours, j'aborde ce problème avec un point de vue de segmentation. Étant donné un individu statistique i et une courbe $\boldsymbol{Y}_{i,\bullet} = (Y_{i,j})_{1\leq j\leq J_i}$, je suppose qu'il existe $K_i^\star$ ruptures $\boldsymbol{T}_i^\star = \left(T_{i,0}^\star, \dots, T_{i,K_i^\star+1}^\star\right)$ avec $0 = T_{i,0}^\star < T_{i,1}^\star < \cdots < T_{i,K_i^\star+1}^\star = J_i$ et $K_i^\star + 1$ amplitudes $\boldsymbol{a}_i^\star = \left(a_{i,k}^\star\right)_{0\leq k\leq K_i^\star}$ telles que (voir la figure 4.14) :

$$\begin{aligned} &\forall k \in \{0 \leq k \leq K_i^\star\}, \forall j \in \left\{T_{i,k}^\star + 1, \dots, T_{i,k+1}^\star\right\}, \\ &\qquad Y_{i,j} = a_{i,k}^\star \sin\left(\pi \frac{t_j - t_{T_{i,k}^\star}}{t_{T_{i,k+1}^\star} - t_{T_{i,k}^\star}}\right) + \varepsilon_{i,j} \text{ où } \varepsilon_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_{i,k}^2\right) \end{aligned} \tag{4.8}$$

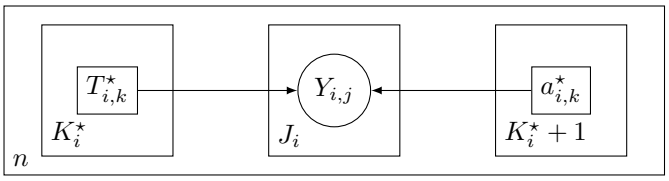

FIGURE 4.14 – Représentation graphique du modèle de rupture *POMH* avec segmentation : les carrés correspondent à des paramètres et les ronds à des variables.

Les instants de ruptures coïncident donc avec les moments d'annulation de la vitesse. De plus, l'indépendance entre les segments permet d'utiliser l'algorithme de programmation dynamique. Pour le moment, j'ai obtenu un résultat de convergence des estimateurs sans pénalisation de la vraisemblance. Pour ce faire, j'ai besoin de quatre hypothèses :

Hypothèse

La première hypothèse généralise le modèle (4.8) à n'importe quelle loi tant que j'ai la proposition suivante :

$$\forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, J_i\}, Y_{i,j} = \mathbb{E}_{\boldsymbol{T}_i^\star, \boldsymbol{a}_i^\star}\left[Y_{i,j}\right] + \varepsilon_{i,j}. \tag{$\mathcal{H}^{\text{POMH}}_{\text{Modèle}}$}$$

La deuxième hypothèse impose tout de même que le bruit est sous-gaussien c'est-à-dire qu'il existe une constante $\beta > 0$ strictement positive telle que pour tout $v \in \mathbb{R}$ et tout $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, J_i\}$:

$$\mathbb{E}\left[e^{v\varepsilon_{i,j}}\right] \leq e^{\beta v^2}. \tag{$\mathcal{H}^{\text{POMH}}_{\text{Sous-Gaussien}}$}$$

La troisième hypothèse permet de voir les ruptures :

$$0 < \boldsymbol{a}_{i,\min}^\star = \min_{0\leq k\leq K_i^\star} \left|a_{i,k}^\star\right|. \tag{$\mathcal{H}^{\text{POMH}}_{\boldsymbol{a}_i^\star}$}$$

Et enfin, si nous notons Δ_{J_i} la distance minimale entre deux instants de ruptures estimés normalisés par J_i et $\Delta_{\boldsymbol{\tau}_i^\star}$ la distance minimale entre deux *vraies* ruptures normalisées par J_i

alors la quatrième hypothèse suppose qu'à partir d'un certain rang J_i :

$$\Delta_{J_i} < \Delta_{\boldsymbol{\tau}_i^\star} \text{ et } \frac{\ln(J_i)}{\Delta_{J_i}^6 J_i} \xrightarrow[J_i \to +\infty]{} 0. \qquad (\mathcal{H}_{\Delta_{J_i}}^{\text{POMH}})$$

Alors, à l'aide de ces hypothèses, j'ai la convergence suivante :

Théorème 17 (Consistance)

Si nous notons $\widehat{K}_{i,J_i}$ et $\widehat{\boldsymbol{\tau}}_{\widehat{K}_{i,J_i}}$ les estimateurs du nombre et des emplacements renormalisés des ruptures maximisant la logvraisemblance du modèle (4.8) étant donné un nombre maximal de ruptures possibles supérieur au vrai nombre $K^\star$ de ruptures alors, sous les hypothèses $(\mathcal{H}_{\text{Modèle}}^{\text{POMH}})$, $(\mathcal{H}_{\text{Sous-Gaussien}}^{\text{POMH}})$, $(\mathcal{H}_{\boldsymbol{a}_i^\star}^{\text{POMH}})$ et $(\mathcal{H}_{\Delta_{J_i}}^{\text{POMH}})$ j'ai pour tout $\delta > 0$ le résultat de consistance suivant :

$$\mathbb{P}\left(\widehat{K}_{i,J_i} \neq K^\star \text{ ou } \left\|\widehat{\boldsymbol{\tau}}_{\widehat{K}_{i,J_i}} - \boldsymbol{\tau}_i^\star\right\|_{\text{Haus}} > \delta\right) \xrightarrow[J_i \to +\infty]{} 0$$

où $\|\cdot\|_{\text{Haus}}$ est la norme de Hausdorff définie dans le théorème 7 pour un vecteur $\boldsymbol{\tau}$ de longueur K et un vecteur $\boldsymbol{\tau}'$ de longueur K' par :

$$\|\boldsymbol{\tau} - \boldsymbol{\tau}'\|_{\text{Haus}} = \max\left(\max_{0\leq k\leq K}\min_{0\leq k'\leq K'}|\tau_k - \tau'_{k'}|, \max_{0\leq k'\leq K'}\min_{0\leq k\leq K}|\tau_k - \tau'_{k'}|\right)$$

Toutefois, nous voyons sur la figure 4.15 un exemple de simulations où le nombre de ruptures maximisant la vraisemblance est de 44 (au lieu de 5). Nous voyons que les emplacements des ruptures supplémentaires sont proches des vraies ruptures (à cause de l'arc du cosinus). Je suis en train de voir comment pénaliser le fait de mettre deux ruptures proches pour améliorer les estimations à distance finie.

Perspectives

Nous sommes en train d'appliquer cette modélisation sur les données que nous avons afin de voir si nous pouvons améliorer les résultats obtenus par Lu et al. (2024). De plus, je suis en train de chercher une pénalité en fonction de la distance minimale entre deux ruptures estimées afin d'avoir une convergence indépendante de la distance minimale Δ_{J_i}.

Perspectives

Une fois la procédure stabilisée, j'aimerais inclure la classification des individus en supposant qu'il y ait des profils semblables (par exemple, sur le nombre de ruptures ou sur les amplitudes en utilisant des lois avec effet mixtes comme dans la section 4.4).

Perspectives

Enfin, je travaille actuellement en collaboration avec Fanny Guillet de la police judiciaire de Lyon et Jérémy Danna du CNRS à Toulouse pour adapter ces modèles afin de voir s'ils pourraient répondre à certaines de leurs questions comme, par exemple, la reconnaissance de caractères particuliers dans le cadre de lettres de menace manuscrites.

4.4 Syndrome d'apnée du sommeil et mélange de chaînes de Markov

Dans le cadre d'une collaboration avec Adeline Leclercq Samson (LJK) et Sébastien Bailly (INSERM, HP2, UGA) puis avec Elodie Vernet (LJK), nous avons proposé un post-doc à Kajal Eybpoosh pour travailler sur le syndrome de l'apnée du sommeil (SAS). À nouveau, je ne suis pas un spécialiste de ce

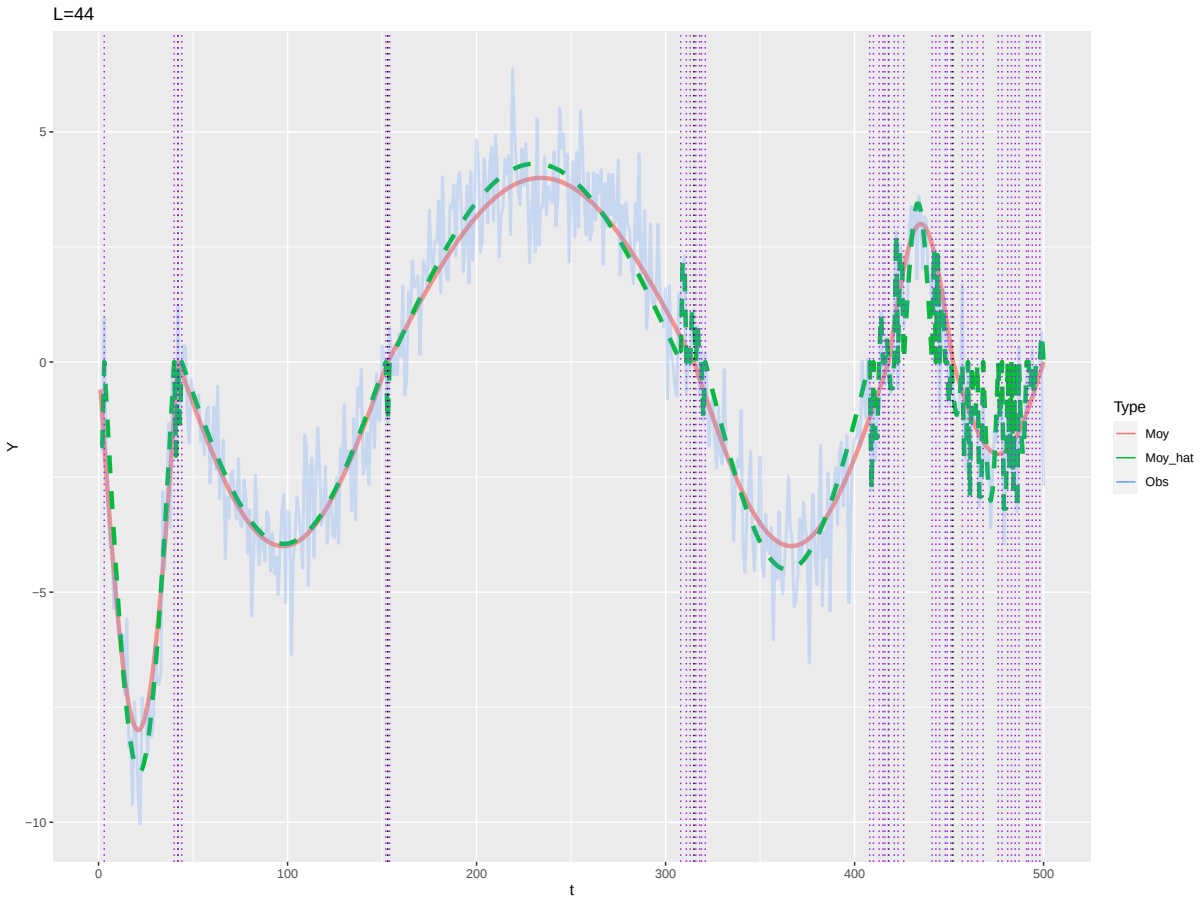

FIGURE 4.15 – Estimation des emplacements des ruptures (traits pointillées violets verticaux) pour des données simulées suivant le modèle (4.8) pour un nombre de ruptures maximisant la vraisemblance comme démontré dans le théorème 17. La courbe en saumon correspond aux paramètres de la simulation, la courbe bleue transparente aux observations et la courbe en pointillés verts aux estimations basées sur les emplacements de ruptures.

syndrome. L'information la plus importante motivant ce travail est que lorsqu'une personne est diagnostiquée, elle doit utiliser une machine la nuit qui envoie l'air dans les voies respiratoires avec une légère surpression [4]. L'inconvénient de ce traitement est qu'il peut être considéré contraignant et constitue une limite (voir Bakker et al. (2019)) et la plupart des personnes arrêtent dans les trois premiers mois (voir Kribbs et al. (1993)). Toutefois, s'ils arrivent à avoir un comportement régulier durant ces trois premiers mois, il semblerait qu'ils auront une bonne adhésion au traitement (voir Sawyer et al. (2011)). L'objectif de ce projet est donc d'étudier les comportements des utilisateurs les trois premiers mois afin de voir si nous pouvons dégager des profils à suivre plus attentivement que d'autres.

Pour ce faire, nous possédons les données de $n = 222$ personnes et l'utilisation de leurs machines sur $J = 91$ nuits (voir les exemples de six courbes sur la figure 4.16). Comme nous pouvons le voir sur la figure 4.16, les comportements peuvent varier. Déjà, nous pouvons remarquer la sur-représentation des 0 par rapport à n'importe quelle autre valeur qui correspond aux nuits où la personne n'a pas du tout utilisé son appareil. On peut notamment remarquer les deux courbes en bas à droite qui se terminent par des 0 indiquant un abandon complet du traitement. Lorsqu'une personne utilise son appareil, elle semble avoir plus ou moins la même utilisation.

Le but étant de mettre en évidence des comportements similaires, nous optons donc pour un **mélange de chaîne de Markov** c'est-à-dire que nous supposons qu'il existe une matrice $\boldsymbol{S} = (S_{i,j})$ dont les valeurs $S_{i,j} \in \{0, 1\}$ précisent si l'individu i a utilisé sa machine ($S_{i,j} = 1$) ou pas ($S_{i,j} = 0$) la nuit J. De plus, nous supposons qu'il existe K clusters, une probabilité $\boldsymbol{\pi}^\star$ pour chaque d'individu d'appartenir à l'une

4. Pour plus d'informations, voir par exemple le site suivant : https://www.ameli.fr/assure/sante/themes/apnee-du-sommeil/traitement-apnee-sommeil

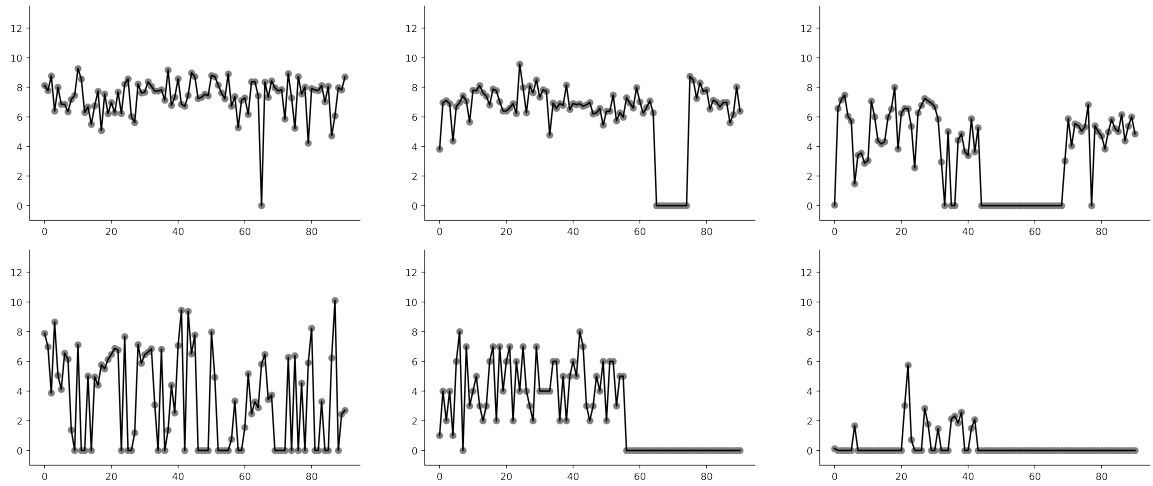

FIGURE 4.16 – Courbes d'utilisation de la machine en heures pour chacune des 91 nuits pour 6 patients de notre base de données.

ou l'autre des classes et une matrice $\boldsymbol{z}$ d'appartenance aux classes. Enfin, nous supposons que la loi $\boldsymbol{Y}_{i,\bullet}$ est une chaîne de Markov dépendante de la classe k dans laquelle l'individu se trouve. Ainsi, nous avons K probabilités $\boldsymbol{P}^\star = (P_k^\star)_{1\leq k\leq K}$ représentant pour chaque chaîne appartenant au cluster k la probabilité que le premier état $S_{i,1}$ vaille 1, K **matrices de transitions** $\boldsymbol{A}_k^\star \in \mathcal{M}_{2\times 2}([0;1])$ entre les états 0 et 1 et K triplés de paramètres $\left(a_k^\star, b_k^\star, \sigma_k^{\star 2}\right)_{1\leq k\leq K}$ définissant la loi de l'état 1 (voir le graphe bayésien de la figure 4.17) :

$$\forall i \in \{1,\dots,n\},\ \forall j \in \{1,\dots,J\},\ \forall k \in \{1,\dots,K\},\ \forall \ell \in \{0,1\} \text{ et } \forall \ell' \in \{0,1\}\ :$$
$$\begin{cases} \mathbb{P}\left(Z_{i,k}=1\right) = \pi_k^\star, \\ \mathbb{P}\left(S_{i,1}=1 \,|Z_{i,k}=1\right) = P_k^\star, \\ \mathbb{P}\left(S_{i,j}=\ell' \,|Z_{i,k}=1, S_{i,j-1}=\ell\right) = A_{k;\ell,\ell'}^\star \text{ pour } j\geq 2, \\ Y_{i,j}\,|S_{i,j}=0 \sim \delta_0, \\ Y_{i,j}\,|Z_{i,k}=1, S_{i,j}=1 \sim \mathcal{N}\left(a_k^\star + b_k^\star t_j, \sigma_k^{\star 2}\right) \end{cases} \tag{4.9}$$

où δ_0 est la **loi de Dirac** en 0 c'est-à-dire que si $S_{i,j}$ vaut 0 alors $Y_{i,j}$ est presque sûrement nulle et $(t_j)_{1\leq j\leq J}$ sont les moments où ont été prises les données (dans notre exemple, $t_j = j$ mais le modèle peut se généraliser à d'autres cas). Après discussions de cette modélisation avec les médecins, la matrice de transition les intéresse particulièrement pour le comportement des patients.

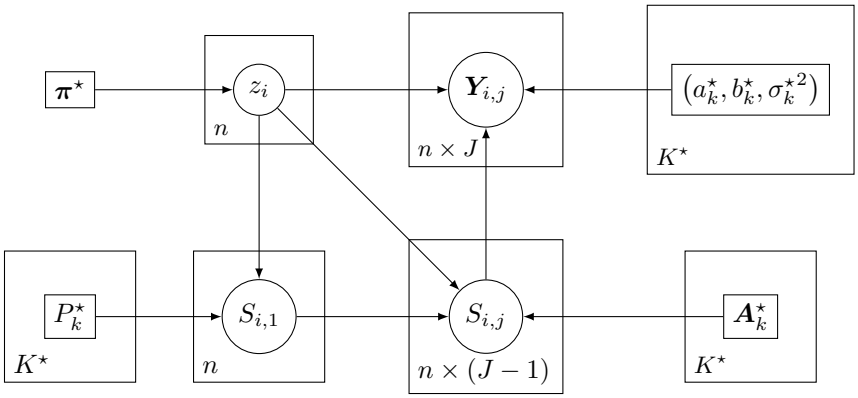

FIGURE 4.17 – Représentation graphique du modèle de mélange de chaînes de Markov : les carrés correspondent à des paramètres et les ronds à des variables.

Dans l'article en cours de rédaction, nous prouvons que la chaîne de Markov du modèle (4.9) est connue puisque nous avons une loi discrète et une loi continue, donc nous avons presque sûrement :

$$\forall i \in \{1,\dots,n\},\ \forall j \in \{1,\dots,J\},\ S_{i,j} = 0 \Leftrightarrow Y_{i,j} = 0.$$

De plus, nous pouvons utiliser l'algorithme *EM* dont les étapes se calculent et en adaptant les résultats obtenus dans les théorèmes 1 et 15, nous souhaitons montrer la convergence des estimateurs. Enfin, nous sommes actuellement en train d'étudier la forme du critère *BIC* pour les données réelles.

Perspectives

Une fois ce travail fini, nous aimerions ajouter des effets mixtes pour mieux prendre en compte le caractère individuel des personnes.

Chapitre 5

Perspectives

"Je refuse que la peur de l'échec m'empêche de faire ce qui m'importe vraiment."
Emma Watson

Pour conclure ce manuscrit, je me permets de regrouper les perspectives disséminées à travers les différentes sections en les regroupant par axes. Je termine par les projets que je n'ai pas eu le temps d'aborder car ils étaient trop récents pour une expertise suffisamment mature.

5.1 Projets sur l'écriture

Je vais commencer par présenter mes idées pour les projets sur l'écriture. Maintenant que l'article Lu et al. (2024) est soumis et que j'ai des premiers résultats sur la modélisation avec la segmentation présentée en section 4.3.4, j'aimerais explorer plusieurs pistes :

- Sur la qualité des estimations, il faut diminuer le nombre de ruptures estimé à distance finie. Pour ce faire, j'envisage de reprendre les démonstrations de convergence du théorème 17 et d'aller chercher une pénalité sur la distance entre deux ruptures successives estimées pour éviter d'avoir des ruptures trop proches.
- Sur l'implémentation, il faut accélérer la vitesse des estimations. La problématique actuelle est le calcul de la matrice $[\Delta(t,s)]_{0\leq t<s\leq J_i}$ utilisée dans la programmation dynamique (voir l'équation (A.5) en annexe si besoin) qui est très chronophage. Une idée serait de remplacer la modélisation d'un cosinus par une fonction quadratique voir peut-être polynomiale d'un degré supérieur qui, si tout se passe bien, devrait permettre de passer d'une complexité cubique à une complexité quadratique. Il faudra alors voir si et comment nous dégradons la qualité des estimations et comparons les temps d'estimation.
- Une fois ces papiers stabilisés, l'idée est alors d'appliquer sur la base de données des enfants dysgraphiques suivant deux possibilités :
  - La première, dans la continuité du travail de Yunjiao, est de se placer dans un contexte de **classification supervisée** pour sélectionner les *features* qui permettraient de détecter la dysgraphie. Dans ce cadre, après une première étude, il pourrait être intéressant de comparer les méthodes notamment avec les résultats des articles Deschamps et al. (2021) et Devillaine et al. (2021) et voir comment nous pouvons prendre le meilleur de chaque modélisation pour améliorer encore les prédictions.
  - À l'opposé, la deuxième idée serait de se placer dans un cadre de **classification non supervisée** et d'introduire des variables latentes pour classer les profils. Beaucoup de questions se posent alors :
    * Faut-il mettre les clusters directement sur les enfants ou faire symbole par symbole comme dans la procédure en deux étapes de Lu et al. (2024) ?
    * Dans le prolongement de Brault et al. (2024a), je m'attends à ce que nous ayons l'identifiabilité et la consistance mais il faudrait s'en assurer.
    * Comment prendre en compte les particularités des enfants (âge, niveau...) ? Devrons-nous introduire des **modèles mixtes** comme pour le travail Eybpoosh et al. (2025) ?

    - Comment optimiser les codes pour que tout soit calculable en un temps raisonnable ; surtout dans un contexte de réchauffement climatique où il faudra peut-être diminuer nos utilisations aux clusters ?

- Enfin, pour toutes ces méthodes et dans un but de diffusion au grand public, il faudra évaluer la robustesse des procédures aux changements de tablettes. De plus, si nous souhaitons aider les pouvoir publics, il faudra peut-être envisager de créer ou de s'associer à une start-up pour promouvoir ce travail.

Jusqu'à présent, je demandais des financements de post-doc pour ces travaux. Si j'ai mon Habilitation à Diriger des Recherches, je pense demander une bourse de thèse rapidement au moins pour les premiers points et voir, suivant l'avancée de la thésarde ou du thésard, comment continuer les points suivants.

En parallèle de cela, avec Jérémy Danna (CNRS rattaché au laboratoire Cognition, Langues, Langage, Ergonomie de l'université Jean Jaurès de Toulouse), nous demandons une bourse de thèse pour Zoé Riebel que nous co-encadrons actuellement durant son M2 alternance avec Fanny Guillet (Laboratoire Police Scientifique de Lyon) sur la prise en compte de la dynamique pour l'analyse de l'écriture. Dans ce cadre, nous avons fait déjà deux demandes auprès de l'ANR (dont j'ai été le principal investigateur la deuxième fois). Pour le moment, nous explorons les bases recueillies mais le but de la thèse est de proposer des nouvelles modélisations qui seront certainement en lien avec les modélisations proposées pour la dysgraphie.

5.2 Projets sur l'apnée du sommeil

Dans le continuité de l'article Eybpoosh et al. (2025) de la section 4.4, j'aimerais proposer une modélisation incluant des effets mixtes afin de mieux prendre en compte les individualités de chaque personne et peut-être diminuer le nombre de groupes formés au final qui semble trop grand pour le moment. Dans ce cadre, il sera plus compliqué d'obtenir les résultats théoriques et il faudra optimiser les codes. Une autre question sur cette modélisation pourrait être la prise en compte d'une chaîne de Markov avec des temps plus longs. Si nous le pouvons, j'aimerais bien obtenir un financement de post-doc pour ces analyses voir entamer une collaboration avec notre nouvelle collègue, Lola Etiévant, afin de bénéficier de ses compétences.

En parallèle de ce travail, j'aimerais proposer un article généralisant les démonstrations des articles Brault et al. (2020, 2024a); Eybpoosh et al. (2025) pour avoir l'identifiabilité et la consistance des estimateurs dans le cadre de n'importe quel modèle de mélange où les lois d'un individu peuvent avoir un nombre *illimité* d'observations. Ce travail étant plus complexe, je n'envisage pas pour le moment de le proposer pour une thèse, éventuellement une bourse de post-doc si je trouve une personne très motivée. Sinon, j'espère pouvoir le faire en continuant de collaborer avec Elodie Vernet (actuellement en détachement au LJK).

5.3 Projet Taunique

Dans le cadre du projet Taunique présenté en section 4.2, je veux aller chercher les raisons de la mauvaise classification de certaines courbes et comment pénaliser le groupe avec deux sauts qui semblent *absorber* tous les individus lorsque les sauts sont peu visibles. Avec Emilie Lebarbier (Université de Nanterre) et Amélie Rosier (ESME Sudria Paris), nous avons déjà commencé à explorer les raisons théoriques pour que les individus avec un saut double soient mis, à tort, dans la classe avec deux sauts distincts. Il faudrait continuer pour les autres groupes et la sous estimation de la valeur du saut. J'aimerais demander un financement pour une ou un post-doc voir, dans un second temps, travailler avec une ou un ingénieur·e d'étude pour mettre en place des outils à destination du GIN (Grenoble Institut Neurosciences) où travaille Virginie Stoppin-Mellet.

5.4 Modèle des blocs latents : robustesse et version non paramétrique

Dans le prolongement du chapitre 2, il y a deux pistes que j'aimerais continuer à explorer en priorité. La première est dans la continuité du projet CoPoLangues sur la robustesse des modèles. D'une part, j'aimerais comparer les classifications obtenues à l'aide du LBM et celles obtenues par le NFLBM sur les données réelles notamment car pour certaines questions, les élèves répondent quasiment tous bons. Ensuite, j'aimerais voir comment fluctuent les classifications suivant si nous avons tous les élèves ou non mais aussi dans le cas où nous nous trouvons dans un mauvais modèle. De plus, j'aimerais évaluer la possibilité de créer un coefficient de stabilité. Pour le moment, je n'ai que des idées basées sur des simulations coûteuses en temps de calcul et je réfléchis à comment proposer des idées basées sur des critères. Ceci pourrait dans un premier temps faire l'objet d'un stage ou d'un encadrement de projets avant de poursuivre potentiellement sur un financement plus important.

L'autre projet qui me tient à coeur depuis que j'ai commencé à étudier l'algorithme *Largest Gaps* est la possibilité de classifier les lignes et les colonnes sans faire d'hypothèses sur les lois autres qu'elles aient un moment d'ordre 2 fini pour appliquer le théorème de limite centrale. Cette idée a déjà été présentée lors des JdS d'Avignon (voir Brault et al. (2017a)) mais je n'ai pas encore pris le temps de concrétiser dans un article. J'aimerais proposer l'idée à un·e stagiaire ou directement avoir un financement post-doctoral pour compléter ces recherches. J'aimerais également en profiter pour réfléchir à la problématique de la sélection du nombre de clusters de façon plus efficace notamment en couplant avec des critères *BIC* sur les marginales et voir si nous pouvons remonter sur un critère de la classification croisée.

5.5 Projets sur les addictions

Durant mon année de détachement au sein de l'équipe SPHERE à l'INSERM de Tours, j'ai pu travailler sur des questionnaires avec plusieurs dimensions issus de plusieurs entretiens de personnes souffrant ou non d'addictions ce qui a abouti à l'article Challet-Bouju et al. (2025) qui vient d'être soumis. Je ne l'ai pas détaillé dans ce manuscrit car il porte principalement sur l'application du modèle développé et l'étude de ce modèle, qui aurait été plus intéressante ici, n'est pas encore finie. J'aimerais beaucoup continuer cette collaboration d'abord en finissant l'article méthodologique et en étudiant les comportements théoriques des estimateurs. Dans la continuité de ce projet, j'aimerais beaucoup profiter des connaissances de Solène Desmée (Université de Tours/SPHERE) pour inclure des effets mixtes. Suivant les premiers résultats, nous pourrions envisager le co-encadrement d'une thèse ou d'un contrat post-doctoral.

Dans un second temps, j'aimerais continuer un projet commencé avec Marianne Balem (actuellement post-doctorante à l'université de Sherbrooke) sur les courbes des personnes pariant en ligne notamment avec des ruptures dans le comportement dues aux différents confinements. L'idée que nous avons commencé à explorer est de reprendre la modélisation proposée dans l'article Brault et al. (2024a) sur les courbes électriques mais en tenant compte du fait qu'il y a un lien entre les observations et avions commencé à proposer une estimation à l'aide des méthodes LASSO (comme dans les articles Brault et al. (2016, 2017b, 2018b)). Les principaux challenges liés à cette problématique sont l'appartenance des clusters à chaque courbe notamment pour inverser les matrices dans la procédure LASSO et la gestion des paramètres de pénalisation du LASSO. Si Marianne a le temps et l'envie, j'aimerais beaucoup que nous puissions continuer ensemble voir co-encadrer un ou une post-doctorant·e sur ce projet.

5.6 Biais dans les algorithmes de recrutement

Un autre projet que je n'ai pas développé dans ce document est celui de l'article Wang et al. (2025) sur le biais dans les algorithmes de recrutement. Dans cet article, nous proposons des modélisations pour simuler des entretiens et voir comment une base biaisée peut influencer sur la qualité d'un algorithme de prédiction. Dans la suite de ce projet et du stage d'Angélique Saillet, j'aimerais proposer de nouvelles simulations encore plus proches d'entretiens de la vie de tous les jours.

D'un point de vue plus politique, une fois l'article Wang et al. (2025) publié, j'aimerais discuter avec les législateurices pour présenter le problème de laisser des algorithmes sur le marché sans contrôle. Comme

pour l'article Brault et al. (2021b) sur le poolage dans la détection du covid 19, je pense que la suite de ce travail dépendra beaucoup des questions posées et de comment répondre au mieux à leurs inquiétudes.

5.7 Analyse de l'association longitudinale entre l'exposition aux facteurs environnementaux et le développement de l'enfant

Un autre projet que je n'ai pas mentionné dans ce manuscrit car nous sommes encore au tout début est la question de l'influence de l'exposition des femmes enceintes aux facteurs environnementaux sur le développement de leur foetus en collaboration avec Johanna Lepeule (Inserm-U1209) et Adeline Samson (LJK). Pour l'instant, cette question n'est posée que pour le poids à la naissance et nous aimerions généraliser à la croissance complète durant les 9 mois à l'aide, dans un premier temps, d'un modèle de **lag distribué** (DL). Dans les questions auxquelles nous devons déjà répondre, il y a le problème des données manquantes (puisque les femmes n'ont pas fait toutes les visites ni au même moment), l'individualité des personnes ou encore le grand nombre de facteurs environnementaux et comment les prendre en compte par exemple. Dans ce cadre, nous avons commencé à encadrer un projet de M1 pour explorer les données et nous avons un financement de 3 ans de thèses ou 2 ans de post-doc pour travailler dessus. J'espère à nouveau profiter de l'arrivée de Lola Etiévant dans notre équipe pour enrichir la modélisation.

5.8 Projet Gratweet

Dans la même idée, suivant les résultats obtenus par Martin Combelles dans son stage, j'aimerais relancer le projet Gratweet pour modéliser des graphes où les arêtes représentent des processus de Hawks. Pour le moment, les principales difficultés sont sur la gestion des potentielles (très) grandes bases de données à prendre en compte. Si ces résultats sont intéressants, il y aurait de quoi proposer une bourse de thèse permettant l'analyse de ce modèle et, si besoin, de la compléter avec les travaux proposés en section 5.4. Toutefois, c'est un projet encore à l'état embryonnaire sur beaucoup d'aspects.

5.9 Diffusion des codes

Enfin, une part que je juge importante de mon travail est de proposer des codes propres et, si possible, actuels à la communauté. Dans ce cadre, je suis en train de discuter avec Perrine Lacroix (Université de Nantes) pour implémenter ses dernières avancées dans le package `Capushe` (voir Brault et al. (2023a)).

Dans une idée similaire, j'aimerais améliorer les estimations proposées dans le package `MuChPoint` et les codes issus de l'article Brault et al. (2024a) en appliquant les travaux de Rigaill (2015). Pour cela, je pense que j'aurais besoin de me former auprès de Guillem Rigaill ou une personne ayant travaillé avec lui sur ce sujet. Cela pourrait être à l'occasion d'un co-encadrement de post-doc par exemple si les personnes sont intéressées.

Annexe A

Revenir aux bases

"Soit X une variable aléatoire suivant une loi gaussienne..."
Beaucoup d'exercices

A.1 Préambule

Lançons une pièce mille fois afin de déterminer si elle est biaisée ou non, l'unique loi que nous pourrons utiliser pour la modélisation est une loi de Bernoulli. Lançons ensuite un dé à six faces, une loi mutinomiale, éventuellement contrainte, semble la plus adaptée. Maintenant, modélisons la distribution des tailles dans une population et prenons pour cela le jeu de données provenant de l'enquête sur les comportements alimentaires de 226 personnes âgées de la région de Bordeaux en 2000 (voir de Micheaux et al. (2013)). Si nous représentons les données à l'aide d'un histogramme (voir la gauche de la figure A.1) et d'une estimation de la densité (trait plein rose), nous observons une distribution qui semble unimodale et qui pourrait être approximée par une loi gaussienne (trait pointillé rouge).

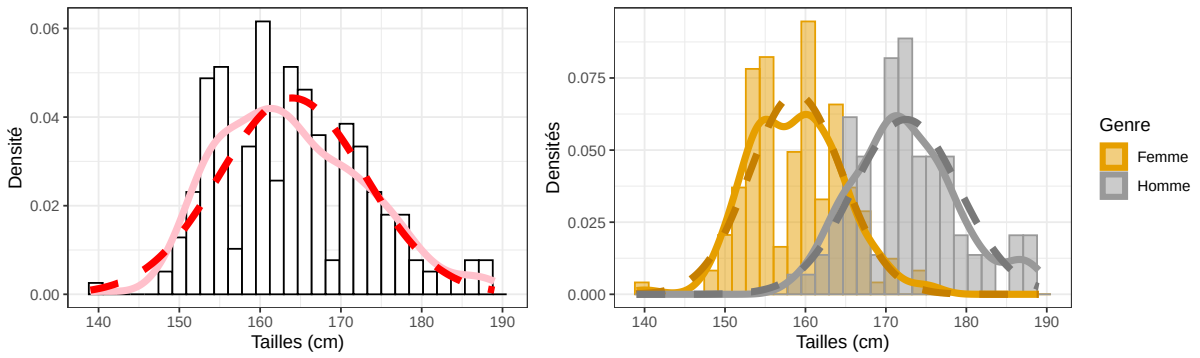

FIGURE A.1 – Représentation des tailles des personnes âgées ayant répondu à l'enquête sur leurs comportements alimentaires sous forme d'histogramme avec la population entière (à gauche) et subdivisée en fonction de leur genre (à droite). Les traits pleins correspondent à une estimation de la densité et les pointillés à la densité d'une loi gaussienne basée sur les moyennes et écart-types des populations correspondantes.

Toutefois, si nous prenons en compte le genre de chaque individu soient 141 femmes (en jaune sur la droite de la figure A.1) et 85 hommes (en gris), nous observons que les observations des tailles de chaque groupe sont mieux approchées par une loi gaussienne propre à chacun. Ainsi, l'observation globale des tailles semble être la combinaison de deux lois d'émission : les tailles des femmes et celles des hommes.

Dans ce cas particulier, les observations de chaque sous groupe étaient distribuées de telle sorte que les observations réunies pouvaient être approchées par une loi gaussienne. Toutefois, nous ignorons si ce phénomène est dépendant de l'échantillon ou resterait pertinent quelque soit le tirage. De plus, dans ce cas précis, le genre des individus semble permettre de bien expliquer les deux sous distributions mais, dans d'autres cas, rien ne garantit que les lois d'émissions peuvent être résumées par des co-variables ou des combinaisons de co-variables.

Ce constat de déviation a déjà été fait par Pearson (1894) au 19ème siècle où il proposait une modélisation pour deux groupes car, comme nous le verrons dans la section A.2.3, l'estimation analytique est déjà complexe.

L'objectif de ce manuscrit est de proposer un panorama de différents travaux sur l'étude de données issues de plusieurs lois d'émission.

A.2 Modèle de mélange

Les **modèles de mélange** supposent qu'il existe K **classes** ou **clusters** et que chaque individu i appartient à une et une seule classe (ainsi, l'ensemble des classes forme une partition de l'ensemble $\{1,\dots,n\}$). Pour la suite, nous notons $\boldsymbol{z}$ la partition des individus dans les classes et $(z_{i,k})$ avec i allant de 1 à n et k allant de 1 à K la **matrice binaire** d'appartenance des individus aux classes (voir un exemple sur la gauche de la figure A.2). Comme un individu appartient à une et une seule classe, la matrice binaire possède exactement un seul un sur chacune de ses lignes.

Notation

De façon équivalente, la matrice binaire $(z_{i,k})$ est parfois notée sous forme de vecteur de longueur K où chaque case z_i appartient à $\{1,\dots,K\}$ (voir un exemple sur la droite de la figure A.2) :

$$\forall i \in \{1,\dots,n\},\ \forall k \in \{1,\dots,K\},\ z_{i,k} = 1 \Leftrightarrow z_i = k.$$

Ces notations étant équivalentes et si aucune précision n'est nécessaire, $\boldsymbol{z}$ symbolise dans ce document les deux puisque la notion qu'elle représente est avant tout la partition.

$$(z_{i,k})_{\substack{i\in\{1,\dots,n\}\\ k\in\{1,\dots,K\}}} = \begin{pmatrix} 1 & 0 & 0\\ 1 & 0 & 0\\ 0 & 1 & 0\\ 1 & 0 & 0\\ 0 & 0 & 1\\ 0 & 0 & 1\\ 0 & 1 & 0\\ 0 & 0 & 1\\ 1 & 0 & 0 \end{pmatrix} \Leftrightarrow (z_i)_{i\in\{1,\dots,n\}} = \begin{pmatrix} 1\\ 1\\ 2\\ 1\\ 3\\ 3\\ 2\\ 3\\ 1 \end{pmatrix}$$

FIGURE A.2 – Exemple de représentations d'une partition de 9 individus dans 3 classes sous forme binaire (à gauche) et vectorielle (à droite).

Dans le cas des modèles de mélange, la partition $\boldsymbol{z}$ est dite **latente** c'est-à-dire qu'elle est *inconnue* et il faudra donc l'estimer. Les modèles de mélange appartiennent ainsi au domaine de la **classification non supervisée** par opposition à ceux de la **classification supervisée** (où la partition $\boldsymbol{z}$ est connue) et **semi-supervisée** (où une partie de la partition $\boldsymbol{z}$ est connue). En particulier, dans le cas de l'exemple de la section A.1, le sexe est une variable connue ; si nous cherchons à le retrouver à partir des tailles, nous sommes dans un cas de classification supervisée.

De plus, les modèles de mélange supposent qu'à chaque classe k est associée une loi d'émission $\mathcal{L}_k$ de telle sorte que les observations du vecteur $\boldsymbol{Y}_{i,\bullet}$, connaissant la classe d'appartenance k de l'individu i, sont issues de la loi $\mathcal{L}_k$:

$$\forall i \in \{1,\dots,n\},\ \forall k \in \{1,\dots,K\},\ \boldsymbol{Y}_{i,\bullet} | z_{i,k} = 1 \sim \mathcal{L}_k$$

et les variables aléatoires $(\boldsymbol{Y}_{i,\bullet})_{i\in\{1,\dots,n\}}$ sont indépendantes. Enfin, les modèles de mélange supposent que l'appartenance à chaque classe est choisie aléatoirement et de façon indépendante par une loi multinomiale de paramètre $\boldsymbol{\pi} = (\pi_1,\dots,\pi_K)$:

$$\forall i \in \{1,\dots,n\},\ Z_i \sim \mathcal{M}(1;\boldsymbol{\pi}).$$

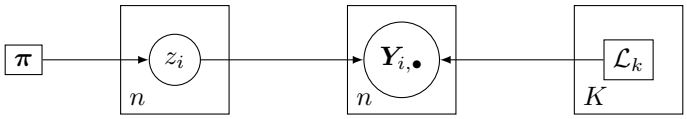

FIGURE A.3 – Représentation graphique du modèle de mélange générique : les carrés correspondent à des paramètres et les ronds à des variables.

Une autre façon de le voir est de dire que les variable Z_i sont indépendantes et identiquement distribuées suivant une loi pondérée par $\boldsymbol{\pi}$ sur l'ensemble $\{1, \dots, K\}$ (voir la représentation schématique en figure A.3).

Aucune contrainte n'est imposée sur les lois $\mathcal{L}_k$; en particulier, rien n'empêche que les lois soient de types différents (une loi gaussienne pour le premier cluster et une loi de Cauchy pour le second par exemple). Toutefois, la définition de base (ainsi appelée dans McLachlan et al. (2019)) suppose que les lois appartiennent à la même famille et, dans le cas de **lois paramétriques** de densité φ, la loi $\mathcal{L}_k$ du cluster k est totalement définie par ses paramètres $\boldsymbol{\theta}_k$.

Remarque

Comme la notation $\boldsymbol{\Psi}$ est utilisée pour symboliser *l'ensemble des paramètres du modèle* et la notation $\boldsymbol{\theta}_k$ est l'ensemble des paramètres du cluster k alors, dans le cas des modèles de mélange, la notation $\boldsymbol{\Psi}$ englobe à la fois $\boldsymbol{\pi}$ et les $\boldsymbol{\theta}_k$.

Exemple fil rouge

En annexe de ma thèse (voir annexe A de Brault (2014)), je présente le cas fictif issu de simulations de deux tireurs qui visent chacun une cible (parmi trois possibles) et dont nous ne voyons que les impacts de balles sans savoir à quel tireur chacun de ces derniers appartient (voir la gauche de la figure A.4). Nous supposons que chaque impact est l'observation d'un mélange gaussien à deux composantes dont la probabilité d'appartenir au premier tireur est π et celle d'appartenir au deuxième est $1-\pi$. Étant donné que l'impact appartient au tireur $k \in \{1, 2\}$, nous supposons qu'il est le résultat d'une variable gaussienne sphérique dont la moyenne et la variance sont inconnues :

$$\forall i \in \{1, \dots, n\},\ \forall k \in \{1, \dots, K\},\ Y_i | z_{i,k} = 1 \sim \mathcal{N}\left(\boldsymbol{\mu}_k, \sigma_k \mathbb{I}_2\right)$$

où $\boldsymbol{\mu}_k \in \mathbb{R}^2$, $\sigma_k \in \mathbb{R}_+^\star$ et $\mathbb{I}_n$ est la matrice identité de taille n. Dans ce cas, nous avons $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \sigma_k)$.

La probabilité π donne le nombre moyen attendu d'impacts par tireur (si π vaut 50%, les tireurs devraient tirer autant de balles l'un que l'autre), $\boldsymbol{\mu}_k$ représente l'endroit visé par le tireur k (si $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$ sont égaux alors les tireurs visent au même endroit) et σ_k la qualité du tireur (plus cette valeur est petite, plus le tireur a une bonne précision).

La droite de la figure A.4 montre les estimations (voir la section A.2.3) obtenues. On peut remarquer que le tireur de gauche (en bleu sur la figure) est moins précis que le tireur de droite. De plus, 58 impacts sont associés au tireur de droite (en or) contre 42 à celui de gauche (en réalité, 3 impacts de balles sont associés à tort au tireur de gauche ; en croix cerclée sur la figure). Enfin, les moyennes ne sont pas pile au centre des cibles (alors que les tireurs avaient visé les centres) : cela est dû d'une part à la variance des estimations par la moyenne et à l'influence des impacts associés aux autres tireurs.

Dans la suite de cette partie, nous présentons quelques particularités des modèles de mélange notamment le calcul de la vraisemblance, l'identifiabilité et la notion de *label switching*, la difficulté à obtenir une forme analytique et un algorithme *classique* d'estimation, l'évaluation des estimations puis nous concluons par l'exemple proposé dans l'article Brault et al. (2021b) pour modéliser la distribution de la charge virale des individus ayant eu le SARS-CoV-2.

A.2.1 Vraisemblance

Pour le reste de cette partie, nous supposons que les lois sont paramétriques de densité φ et de paramètre $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Étant données que les variables sont indépendantes et en utilisant la formule des probabilités totales, la (log)vraisemblance d'un modèle de mélange paramétrique est :

$$p(\boldsymbol{Y}; \boldsymbol{\Psi}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \varphi\left(\boldsymbol{Y}_{i,\bullet}; \boldsymbol{\theta}_k\right) \text{ et } \log p(\boldsymbol{Y}; \boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \varphi\left(\boldsymbol{Y}_{i,\bullet}; \boldsymbol{\theta}_k\right) \right]. \tag{A.1}$$

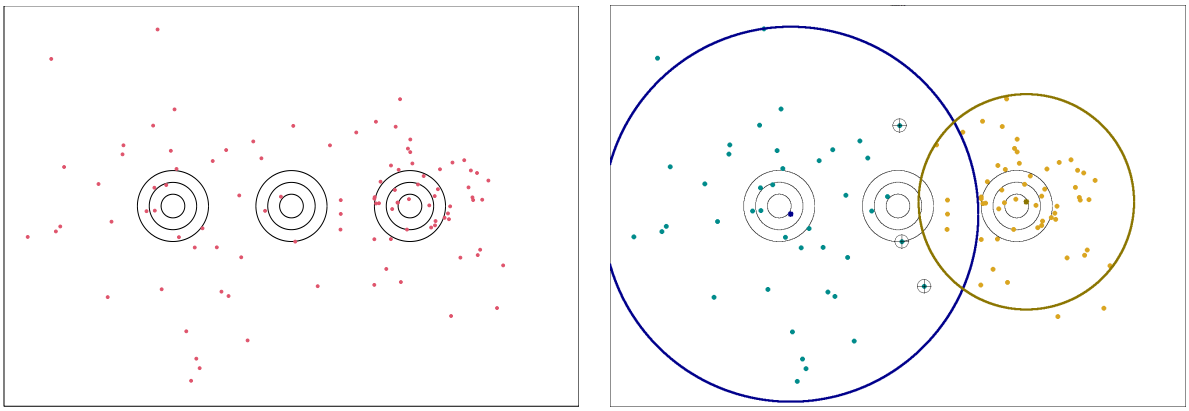

FIGURE A.4 – Exemple proposé dans Brault (2014) de deux tireurs qui ont chacun visé une cible (gauche et droite). À gauche, nous représentons les impacts de balles sans savoir quels impacts correspondent à quel tireur et à droite, nous représentons une estimation des paramètres et des appartenances : chaque impact est colorié suivant l'appartenance à un tireur (bleu à gauche et or à droite), les points plus foncés correspondent aux moyennes estimées et les cercles à région de confiance à 95% des impacts. Les croix cerclées correspondent aux points jugés à tort d'appartenir au tireur bleu.

Comme nous pouvons le voir, la (log)vraisemblance se calcule avec une complexité $\mathcal{O}(nKC_\varphi)$ où C_φ est le temps de calcul de $\varphi\left(\boldsymbol{Y}_{i,\bullet};\boldsymbol{\theta}_k\right)$; par exemple, elle est en $\mathcal{O}(J)$ dans le cas d'une loi gaussienne multivariée sur $\mathbb{R}^J$.

Implémentations

Bien que calculable en un temps raisonnable, il arrive qu'il y ait des problèmes numériques à l'intérieur de la somme sur k si la dimension du vecteur $\boldsymbol{Y}_{i,\bullet}$ est trop grande ou si le nombre K de classes est trop important par exemple. Un cas rencontré est que les termes soient en dessous de la valeur numérique et que nous obtenons au final 0 après sommation. Dans ce cas, il est souvent nécessaire de passer par le calcul de $\log\left[\pi_k\varphi\left(\boldsymbol{Y}_{i,\bullet};\boldsymbol{\theta}_k\right)\right]$ pour tout $k \in \{1,\dots,K\}$ et soit de repasser à l'exponentielle (ce cas s'utilise si la dimension du vecteur $\boldsymbol{Y}_{i,\bullet}$ est trop grande), soit de s'assurer que la plus grande des valeurs soient au moins prise en compte (pour éviter d'avoir une somme nulle ensuite).

A.2.2 Identifiabilité et label switching

L'une des difficultés des modèles de mélange vient du fait que l'ordre des classes est arbitraire si aucune contrainte n'est donnée. Par exemple, dans le cas de mélange gaussien réduit unidimensionnel, les deux jeux de paramètres $\boldsymbol{\Psi}$ et $\boldsymbol{\Psi}'$ suivants :

$$\begin{aligned} \boldsymbol{\pi} = (0.2, 0.8) \quad &\text{et} \quad \boldsymbol{\mu} = (0, 1) \\ \text{et } \boldsymbol{\pi}' = (0.8, 0.2) \quad &\text{et} \quad \boldsymbol{\mu}' = (1, 0) \end{aligned}$$

donnent les mêmes densités puisque la somme est commutative. Ce phénomène est appelé **label switching** et pose des difficultés sur l'analyse des résultats ou l'identifiabilité notamment. Des solutions peuvent limiter l'influence (voir par exemple Stephens (2000); Brault et al. (2024a)) en mettant, par exemple, un ordre sur les paramètres comme dans le cas d'un modèle de mélange de lois de Poissons. Au final, nous parlerons d'identifiabilité au label switching près.

Un autre problème pour l'identifiablité est le fait qu'une probabilité d'appartenir à une classe soit nulle. En effet, si une probabilité π_k est nulle alors le cluster k n'a pas d'influence sur la densité et nous pouvons prendre n'importe quel paramètre $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}$. Par conséquent, une condition nécessaire est que toutes les probabilités soient strictement positives.

Point de vigilance

Il ne faut pas confondre la nécessité que les probabilités soient strictement positives et le problème des **classes vides**. Nous parlons de classes vides lorsque, étant donné un modèle de mélange avec K clusters, il n'y a aucune observation qui provient d'une des classes. Une autre façon de le formuler est de dire que la matrice binaire $(z_{i,k})$ possède au moins une colonne ne contenant que des zéros. En particulier, si une probabilité π_k est infiniment petite, il faudra un très grand nombre d'observations pour avoir au moins un représentant avec une forte probabilité (puisque nous sommes dans le cas d'une loi géométrique).

Exemple fil rouge

Dans le cas de l'exemple fil rouge sur les deux tireurs, le label switching consisterait à inverser les couleurs (or à gauche et bleu à droite sur la droite de la figure A.4) et, dans ce cas, si nous comparons les partitions sans en tenir compte, nous n'aurions que trois impacts correctement affectés (les trois impacts avec une croix cerclée). De même, nous parlerions de classe vide s'il y avait eu un troisième tireur (au moins) qui n'avait pas eu le temps de tirer au moment où nous avions noté les impacts de balles ; par exemple, dans le cas où nous aurions tiré au sort à chaque fois quel tireur avait le droit d'utiliser son arme.

Ces deux notions étant posées, nous pouvons remarquer que l'identifiabilité d'un modèle de mélange paramétrique avec K clusters nécessite au moins les trois conditions suivantes :

1. **Probabilité d'appartenance à la classe k :** Pour chaque classe k, la probabilité d'appartenance doit être strictement positive. Autrement dit, nous avons π_k strictement positif pour tout cluster k.
2. **Densité injective :** l'application qui à un ensemble de paramètre $\boldsymbol{\theta}$ associe la densité $\varphi(\cdot;\boldsymbol{\theta})$ est injective.
3. **Paramètres différents :** étant donnés deux clusters k et k' différents ($k \neq k'$) alors les ensembles de paramètres $\boldsymbol{\theta}_k$ et $\boldsymbol{\theta}_{k'}$ sont différents.

La deuxième condition vient de l'identifiabilité d'une loi paramétrique. Pour la troisième condition, si elle n'est pas présente, il n'est pas possible de distinguer le cluster k du cluster k' et il suffit de prendre n'importe quel couple de proportion $(\pi'_k, \pi'_{k'})$ tel que $\pi_k + \pi_{k'} = \pi'_k + \pi'_{k'}$ pour obtenir la même densité.

Point de vigilance

Si ces conditions sont nécessaires, elles ne sont pas suffisantes pour autant. En particulier, le modèle de mélange avec des lois de Bernoulli multivariées n'est pas identifiable (voir Gyllenberg et al. (1994)). Dans leur livre, Droesbeke et al. (2013) propose des conditions suffisantes pour obtenir l'identifiabilité.

A.2.3 Estimation et algorithme *Espérance Maximisation*

Comme l'avait déjà remarqué Pearson (1894) en son temps, le fait que la vraisemblance présentée en (A.1) soit un produit de sommes rend impossible l'obtention d'une forme analytique des estimateurs du maximum de vraisemblance à chaque fois. Pour contourner ce problème, Dempster et al. (1977) propose l'algorithme *Espérance Maximisation* (*Expectation Maximisation* en anglais) ou *EM* dont le principe est de prendre un ensemble de paramètres $\boldsymbol{\Psi}^{(c)}$ et de remarquer que pour tout ensemble de paramètres $\boldsymbol{\Psi}$, nous avons la décomposition suivante :

$$\log p(\boldsymbol{Y};\boldsymbol{\Psi}) = \underbrace{\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}}\left[\log p(\boldsymbol{Y},\boldsymbol{Z};\boldsymbol{\Psi})\right]}_{=:Q\left(\boldsymbol{\Psi}\middle|\boldsymbol{\Psi}^{(c)}\right)} - \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}}\left[\log p(\boldsymbol{Y}|\boldsymbol{Z};\boldsymbol{\Psi})\right]. \tag{A.2}$$

Or, grâce à l'inégalité de Jensen (voir par exemple la proposition A.3.1 de ma thèse Brault (2014)), nous remarquons que pour tout ensemble de paramètres $\boldsymbol{\Psi}$ et $\boldsymbol{\Psi}^{(c)}$, nous avons :

$$\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}}\left[\log p(\boldsymbol{Y}|\boldsymbol{Z};\boldsymbol{\Psi})\right] \leq \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}}\left[\log p\left(\boldsymbol{Y}\middle|\boldsymbol{Z};\boldsymbol{\Psi}^{(c)}\right)\right].$$

Le principe de l'algorithme *EM* est donc de choisir $\boldsymbol{\Psi}^{(c+1)}$ à chaque étape de façon à améliorer la fonction $\boldsymbol{\Psi} \mapsto Q\left(\boldsymbol{\Psi}\middle|\boldsymbol{\Psi}^{(c)}\right)$ afin d'augmenter la valeur de la logvraisemblance. Il se divise ainsi en deux étapes (voir l'algorithme 1) où la première consiste à calculer la fonction $Q\left(\cdot\middle|\boldsymbol{\Psi}^{(c)}\right)$ ce qui revient à calculer pour chaque individu i sa probabilité $s_{i,k}^{(c)}$ d'appartenance à chaque classe k étant donnés les observations $\boldsymbol{Y}$ et le paramètre $\boldsymbol{\Psi}^{(c)}$ puis à maximiser la fonction.

Input: Données $\boldsymbol{Y}$, nombre de clusters K et un ensemble initial de paramètre $\boldsymbol{\Psi}^{(0)}$

while la convergence n'est pas atteinte **do**
  // Étape E (Espérance)
  for $i \in \{1, \dots, n\}$ **do**
    for $k \in \{1, \dots, K\}$ **do**
      Calcul de $s_{i,k}^{(c)} = \mathbb{P}\left(Z_{i,k} = 1 \middle| \boldsymbol{Y}; \boldsymbol{\Psi}^{(c)}\right)$
  // Étape M (Maximisation)
  Calcul de $\boldsymbol{\Psi}^{(c+1)} \in \operatorname{argmax}_{\boldsymbol{\Psi}} Q\left(\boldsymbol{\Psi} \middle| \boldsymbol{\Psi}^{(c)}\right)$. Notamment :

$$\forall k \in \{1, \dots, K\},\ \pi_k^{(c+1)} = \frac{s_{+,k}^{(c)}}{n}$$

  où $s_{+,k}^{(c)} = \sum_{i=1}^{n} s_{i,k}^{(c)}$.

Output: Estimation des paramètres $\boldsymbol{\Psi}^{(\infty)}$ et des probabilités d'appartenance aux classes $\left(s_{i,k}^{(\infty)}\right)$

Algorithm 1: Principe de l'algorithme *EM*.

Pour définir la convergence, plusieurs critères peuvent être utilisés comme le fait que la vraisemblance n'évolue presque plus ou que les paramètres estimés ne bougent presque plus. Le second impliquant le premier (sous condition de correctement choisir la notion de *proche*), je recommande plutôt ce dernier en faisant attention de bien définir les distances.

L'algorithme *EM* converge vers un maximum local et il est donc conseillé de relancer plusieurs fois l'algorithme et de choisir l'estimation de $\boldsymbol{\Psi}$ maximisant la vraisemblance. Afin de proposer une classification des individus dans une classe, l'estimation $\widehat{\boldsymbol{z}}$ de la partition $\boldsymbol{z}$ est obtenue en appliquant le maximum a posteriori :

$$\forall i \in \{1, \dots, n\},\ \widehat{z}_i \in \operatorname*{argmax}_{k \in \{1,\dots,K\}} s_{i,k}^{(\infty)}$$

où $s_{i,k}^{(\infty)}$ sont les estimations pour chaque individu i d'appartenir à la classe k obtenues à la fin de l'algorithme *EM*.

Pour accélérer la vitesse de convergence, Celeux et Govaert (1992) proposent d'ajouter une étape de *classification* entre l'étape *E* et l'étape *M* (on parle alors d'algorithme *Classification Espérance Maximisation* ou *CEM*) : une fois la matrice $\left(s_{i,k}^{(c)}\right)$ des probabilités d'appartenance aux classes pour chaque individu calculée, elle est remplacée par l'estimation $\widehat{\boldsymbol{z}}$ faite par le maximum a posteriori. En discrétisant ainsi l'espace des estimations des paramètres, la convergence est plus rapide mais l'algorithme est encore plus sensible aux initialisations. De plus, l'algorithme *CEM* proposera de mauvaises estimations dans le cas de lois très rapprochées (comme des mélanges gaussiens avec des moyennes relativement proches compte-tenu des variances respectives) car il choisira une seule classe pour les individus *à cheval* entre deux groupes. Par contre, si les groupes sont très séparés, l'algorithme *CEM* donnera de bonnes estimations.

Pour contourner la dépendance aux initialisations, Celeux et al. (1995) proposent de remplacer l'étape de classification de l'algorithme *CEM* par une simulation de l'affectation de chaque individu à l'une des classes par une loi multinomiale de paramètres $\boldsymbol{s}_{i,\bullet}^{(c)}$. L'algorithme *Stochastique Espérance Maximisation* ou *SEM* simule ainsi une chaîne de Markov dont la loi stationnaire permet une estimation des paramètres et des classes d'appartenance (voir Nielsen (2000)).

Implémentations

Dans un travail non publié que j'avais fait en marge de ma thèse mais présent dans l'annexe 2.A (voir Brault (2014)), pour bien fonctionner, l'algorithme *SEM* doit éviter de tomber dans certains états absorbants :

- si l'algorithme vide une classe alors il ne pourra plus jamais la remplir puisque la probabilité $\widehat{\pi_k}$ estimée sera nulle.
- si l'algorithme estime une loi dégénérée (comme pour une gaussienne multivariée avec des points alignés et donc avec une variance non inversible par exemple ou une classe d'une loi de Poisson

avec uniquement des 0), il ne lui sera plus possible d'affecter les points qui ne respectent pas la loi dégénérée. Ce cas peut se voir, par exemple, lorsque des données continues (par exemple des tailles) sont *trop* discrétisées.

Pour contrer ce problème, des vérifications régulières sont nécessaires. De plus, théoriquement, si la longueur de la chaîne tend vers l'infini, les paramètres subiront le problème du label switching en étant les mêmes qu'à un autre moment de la chaîne à une permutation des labels près. Il est donc important d'en tenir compte au moment de l'estimation des paramètres sans influer durant la simulation car cela biaiserait la chaîne de Markov associée (voir par exemple Celeux et al. (2000)).

Exemple fil rouge

Dans l'exemple des cibles, nous avons utilisé l'algorithme *EM* pour estimer les paramètres [1]. La classe de chaque impact a été choisie par maximum a posteriori mais les impacts proches des deux centres ont des probabilités plus faibles que ceux situés sur les extrêmes gauches et droites.

Littérature connexe

Il existe d'autres types d'algorithmes d'estimation comme, par exemple, l'algorithme *Expectation Propagation* qui reprend la décomposition de la vraisemblance (vue dans l'équation (A.2)) mais en minimisant un majorant cette fois (voir le chapitre 10, section 7 de Bishop et Nasrabadi, 2006). De même, il peut parfois être intéressant de se placer dans un cadre bayésien et ainsi utiliser un algorithme *EM variationnel* (voir par exemple Keribin (2009)) ou encore un échantillonneur de Gibbs (voir le livre de Mengersen et al. (2011)).

A.2.4 Sélection du nombre de classes

Une autre question que j'ai étudiée est celle de la sélection du nombre de classes. Principalement, je me suis intéressé à deux critères suivant les objectifs :

- Le **critère *BIC*** (pour *Bayesian Information Criterion* en anglais) de Schwarz et al. (1978) pénalise la logvraisemblance par le nombre de paramètres à estimer divisé par deux multiplié par le logarithme du nombre d'observations. Ce critère est réputé comme consistant sous certaines conditions (voir Keribin (2000)) c'est-à-dire que si les données ont été générées suivant le modèle de mélange étudié alors le critère donnera asymptotiquement le bon nombre de classes.
- Le **critère *ICL*** (pour *Integrated Complete Likelihood*)) cherche à maximiser la log-vraisemblance complète intégrée (voir Biernacki et al. (2000)) dans le but de favoriser les clusters minimisant l'entropie (voir par exemple Baudry et al. (2010)). Par conséquent, il faut estimer une partition cohérente $\widehat{\boldsymbol{z}}_K$, calculer la valeur $ICL(K) = \log p(\boldsymbol{y}, \widehat{\boldsymbol{z}})$ et prendre le nombre de clusters qui maximisent ce critère. Pour cela, Biernacki et al. (2000) proposent de prendre la partition associée au maximum de vraisemblance tandis que Wyse et al. (2017) présentent un algorithme glouton pour trouver le meilleur critère *ICL*. Notons que ces deux techniques ne donnent pas tout à fait les mêmes valeurs optimales ; nous en discutons dans le cadre du modèle des blocs latents dans la section B.2.4 et dans l'article Brault et al. (2024b).

Généralement, le critère *BIC* choisit des modèles avec plus de classes que le critère *ICL*.

Il existe d'autres façons d'estimer le nombre de clusters. Durant l'un de mes stages de fin d'étude, j'ai pu programmer et continue de maintenir le package `Capushe` (Brault et al., 2023a) qui implémente l'heuristique de pente (voir Baudry et al. (2012)) et dont l'exemple proposé est un mélange de vingt et une gaussiennes sphériques.

Perspectives

Actuellement, je discute avec Perrine Lacroix pour proposer une nouvelle version du package afin d'implémenter les nouvelles approches qu'elle propose dans sa thèse Lacroix (2022).

1. Des vidéos montrant le comportement sur l'exemple fil rouge sont disponibles sur ma chaîne : https://www.youtube.com/@vincentbrault236/videos

A.3 Modèle de segmentation

Dans les **modèles de segmentation**, l'ordre des observations a un intérêt. Dans ce cadre, l'observation Y_2 se situant entre les observations Y_1 et Y_3 et donc, si Y_1 et Y_3 appartiennent au même groupe alors Y_2 appartient aussi à ce groupe. Les modèles de segmentation supposent donc qu'il existe K **ruptures** $T_1 < T_2 < \cdots < T_K$ telles que l'ensemble des intervalles $(]T_k; T_{k+1}])_{0\leq k\leq K}$ forme une partition de l'ensemble $\{1, \ldots, n\}$ avec les conventions $T_0 = 0$ et $T_{K+1} = n$. Ainsi, un individu i appartient au segment $K \in \{0, \ldots, K\}$ si et seulement si $T_k < i \leq T_{k+1}$ ou, comme les valeurs sont discrètes, si et seulement si $T_k + 1 \leq i \leq T_{k+1}$.

Notation

Dans ce manuscrit, nous noterons D_k l'ensemble $\{i \in \{1, \ldots, n\} \,|T_k + 1 \leq i \leq T_{k+1}\}$ correspondant aux valeurs discrètes de chaque segment $]T_k; T_{k+1}]$.

Remarque

Si nous considérons que les segments sont des classes comme pour le modèle de mélange de la section A.2, la matrice $\boldsymbol{z}$ associée serait une matrice diagonale par blocs avec pour chaque bloc un vecteur rempli de 1 de longueur $(T_{k+1} - T_k)_{0\leq k\leq K}$.

Contrairement aux variables latentes $\boldsymbol{z}$ du modèle de mélange, les ruptures $\boldsymbol{T} = (T_k)_{1\leq k\leq K}$ sont considérées comme des **paramètres** du modèle. En revanche, comme pour les modèles de mélange, nous supposons qu'il existe $K + 1$ lois $(\mathcal{L}_k)_{0\leq k\leq K}$ chacune associée à un ensemble D_k. Ainsi, nous avons :

$$\forall i \in \{1, \ldots, n\}, \forall k \in \{0, \ldots, K\}, \boldsymbol{Y}_{i,\bullet} | i \in D_k \sim \mathcal{L}_k$$

et les variables aléatoires $(\boldsymbol{Y}_{i,\bullet})_{i\in\{1,\ldots,n\}}$ sont indépendantes. De même, nous proposons une représentation schématique en figure A.5.

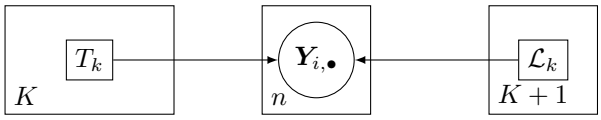

FIGURE A.5 – Représentation graphique du modèle de rupture générique : les carrés correspondent à des paramètres et les ronds à des variables.

Comme pour les modèles de mélange, les lois $\mathcal{L}_k$ peuvent être de types différents mais il est d'usage qu'elles appartiennent à la même famille et, à nouveau dans le cas de **lois paramétriques** de densité φ, la loi $\mathcal{L}_k$ du segment D_k est totalement définie par ses paramètres $\boldsymbol{\theta}_k$.

Point de vigilance

Bien que les ruptures $\boldsymbol{T} = (T_k)_{0\leq k\leq K}$ soient considérées comme des paramètres, il est d'usage qu'elles soient mentionnées à part de l'ensemble $\boldsymbol{\Psi}$ des autres paramètres (contrairement aux modèles de mélange où les proportions $\boldsymbol{\pi}$ d'appartenance à chaque classe sont incluses). Ainsi, nous noterons, par exemple, la vraisemblance par $p(\boldsymbol{Y}; \boldsymbol{T}, \boldsymbol{\Psi})$.

Exemple fil rouge

Dans l'exemple fictif de la figure A.6, nous avons simulé un jeu de données de 29 observations avec des temps de ruptures aux instants 7, 11 et 20. Pour chaque segment k, nous supposons que les observations sont issues de lois $\mathcal{N}(\mu_k; \sigma^2)$ indépendantes. Dans ce cas, nous avons l'équation suivante :

$$\forall k \in \{0, \ldots, 3\}, \forall i \in \{T_k + 1, \ldots, T_{k+1}\}, Y_i \sim \mathcal{N}\left(\mu_k, \sigma^2\right) \tag{A.3}$$

et toutes les observations sont indépendantes.

Dans la suite de cette partie, nous présenterons la vraisemblance du modèle puis l'identifiabilité. Nous présenterons ensuite l'algorithme de programmation dynamique pour estimer les instants de ruptures lorsque les segments sont indépendants et une façon de transformer le problème en modèle linéaire sparse,

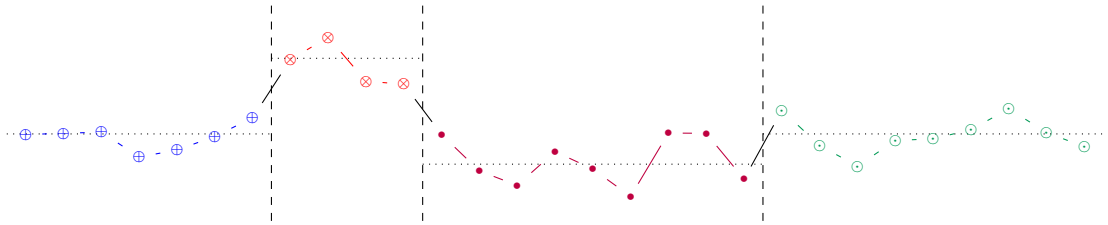

FIGURE A.6 – Représentation d'une simulation d'un modèle de segmentation où les individus d'un même segment sont issus d'une loi $\mathcal{N}(\mu_k; \sigma^2)$. Les traits horizontaux en pointillés correspondent aux moyennes des segments, ceux verticaux aux changements entre deux segments, une couleur et un symbole différents ont été associés à chaque segment pour représenter les points et les traits pleins noirs correspondent à des liaisons entre points appartenant à des segments différents.

si les lois le permettent, afin d'utiliser des méthodes *LASSO* (*Least Absolute Shrinkage and Selection Operator*) pour estimer les paramètres et les ruptures. Nous conclurons par l'exemple proposé dans l'article Brault et al. (2018b).

Littérature connexe

Dans tous les articles présentés dans ce manuscrit, nous supposerons que nous avons l'ensemble des observations ; nous parlons alors de **modèle de ruptures *offline*** (voir par exemple Truong et al. (2020)). À l'opposé, nous pourrions supposer que les observations arrivent progressivement et de vouloir détecter le plus précocement possible l'arrivée d'une rupture ; nous parlons alors de **modèle de ruptures *online*** (voir par exemple Aminikhanghahi et Cook (2017)). Ce dernier cas ne sera pas traité dans le manuscrit.

A.3.1 Vraisemblance et identifiabilité

Comme pour le modèle de mélange, nous supposons que les lois sont paramétriques de densité φ et de paramètre $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Connaissant les emplacements des ruptures $\boldsymbol{T}$, les observations sont indépendantes et leur loi ne dépend que du segment dans lequel l'observation se situe. Ainsi, la (log)vraisemblance d'un modèle de segmentation paramétrique est :

$$p(\boldsymbol{Y}; \boldsymbol{T}, \boldsymbol{\Psi}) = \prod_{k=0}^{K} \prod_{i=T_k+1}^{T_{k+1}} \varphi\left(\boldsymbol{Y}_{i,\bullet}; \boldsymbol{\theta}_k\right) \text{ et } \log p(\boldsymbol{Y}; \boldsymbol{T}, \boldsymbol{\Psi}) = \sum_{k=0}^{K} \sum_{i=T_k+1}^{T_{k+1}} \log\left[\varphi\left(\boldsymbol{Y}_{i,\bullet}; \boldsymbol{\theta}_k\right)\right]. \tag{A.4}$$

Cette fois, la (log)vraisemblance se calcule avec une complexité $\mathcal{O}(nC_\varphi)$ où C_φ est le temps de calcul de $\varphi\left(\boldsymbol{Y}_{i,\bullet}; \boldsymbol{\theta}_k\right)$. En particulier, le temps est indépendant du nombre de ruptures et de leurs positions puisque nous parcourons une seule fois les observations.

Comme les ruptures sont ordonnées, les segments D_k le sont également ; il n'y a donc pas de problème de label switching.

Ainsi, pour avoir l'identifiabilité, il faut que deux lois successives soient différentiables et qu'il y ait au moins une observation par segment (voir par exemple Brault et al. (2024a)). Ainsi, nous devons avoir les deux conditions suivantes :

- Il faut que n soit plus grand ou égal à $K + 1$
- Deux lois successives $\mathcal{L}_k$ et $\mathcal{L}_{k+1}$ doivent être différentes. Dans le cas de lois paramétriques identifiables, cela veut dire que :

$$\forall k \in \{0, \ldots, K\},\ \boldsymbol{\theta}_k \neq \boldsymbol{\theta}_{k+1}.$$

Exemple fil rouge

Dans l'exemple fil rouge, ce sont à chaque fois les mêmes variances mais avec deux moyennes successives différentes. Notons que dans l'exemple, les moyennes du premier et du dernier segment sont égales ; cela n'est pas un problème puisque les comparaisons doivent être faites entre deux segments successifs (contrairement aux modèles de mélange où il faut comparer tous les groupes deux à deux).

A.3.2 Estimation

Il peut être possible d'estimer les paramètres $(\boldsymbol{\Psi}, \boldsymbol{T})$ d'un modèle de ruptures en fixant les ruptures et en maximisant les paramètres et de tester ainsi toutes les combinaisons possibles (voir par exemple Brault et al. (2023b)). Si cela est plus petit que de tester toutes les partitions d'un modèle de mélange (dont l'ensemble des possible a une taille de K^n), la complexité est tout de même de l'ordre de $\mathcal{O}\left(n^K\right)$ ce qui est computationnellement compliqué dès que K est plus grand que 2. Pour contourner ce problème, nous présenterons deux méthodes utilisées dans mes travaux à savoir l'**algorithme de programmation dynamique** (voir par exemple Bellman (1961)) et les méthodes basées sur les procédures *LASSO* (*Least Absolute Shrinkage and Selection Operator*).

Algorithme de Programmation Dynamique

Cette partie est adaptée de la section 3.3 de l'acte de conférence Brault et al. (2018d). Pour pouvoir utiliser l'algorithme de programmation dynamique, il faut que la maximisation de la vraisemblance, étant donné un ensemble de ruptures $\boldsymbol{T}$, puisse se faire en maximisant chaque segment D_k indépendamment. Autrement dit, nous avons :

$$\max_{\boldsymbol{\Psi}\in\boldsymbol{\Theta}^{K+1}} \log p\left(\boldsymbol{Y};\boldsymbol{T},\boldsymbol{\Psi}\right) = \sum_{k=0}^{K} \underbrace{\max_{\boldsymbol{\theta}_k\in\boldsymbol{\Theta}} \sum_{i=T_k+1}^{T_{k+1}} \log\left[\varphi\left(\boldsymbol{Y}_{i,\bullet};\boldsymbol{\theta}_k\right)\right]}_{=:\Delta(T_k:T_{k+1})} \tag{A.5}$$

où les $\Delta\left(T_k : T_{k+1}\right)$ ne dépendent que des bornes du segment D_k. Le but est donc de trouver les ruptures $(T_1, \ldots, T_K)$ telles que la somme des $\Delta\left(T_k : T_{k+1}\right)$ de l'équation (A.5) soit maximale (voir la figure A.7 (a)).

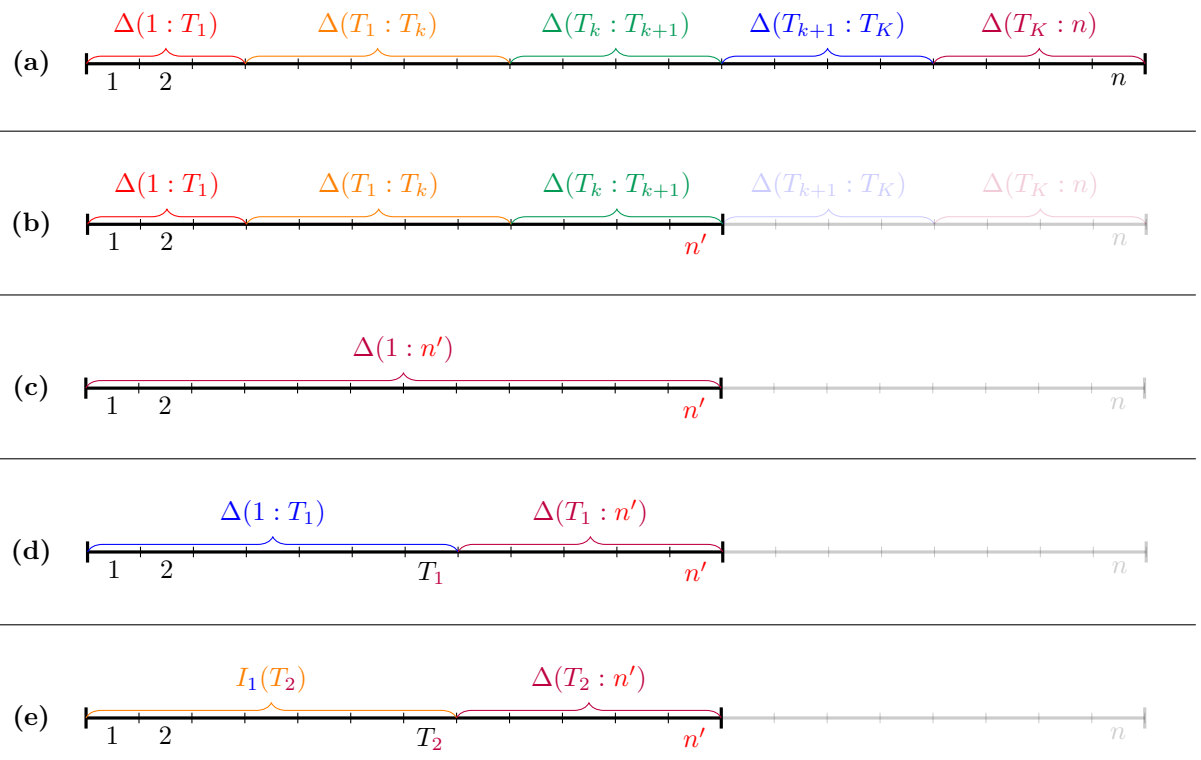

FIGURE A.7 – Représentation schématique des maximisations suivant le nombre de ruptures : principe global (a), introduction de la fonction $I_{K'}(n')$ (b), cas avec zéro rupture (c), cas avec une rupture (d) et principe de récurrence avec deux ruptures (e).

Pour ce faire, nous introduisons pour tout $K' \in \{0, \ldots, K\}$ et tout $n' \in \{K'+1, \ldots, n\}$ la fonction $I_{K'}(n')$ représentant la valeur maximale de la logvraisemblance $\log p\left[(\boldsymbol{Y}_{i,\bullet})_{1\leq i\leq n'}; (T_1, \ldots, T_{K'}), (\boldsymbol{\theta}_k)_{1\leq k\leq K'}\right]$

s'il n'y avait que K' ruptures et n' observations (voir la figure A.7 (b)) :

$$I_{K'}(n') = \max_{1<n_1<\cdots<n_{K'}<n_{K'+1}=n'} \sum_{k=0}^{K'} \Delta(T_k : T_{k+1}).$$

Le maximum de la logvraisemblance pour K ruptures vaut donc $I_K(n)$. Le but de cette décomposition est d'observer que nous avons une relation de récurrence qui apparaît sur K' :

(K'=0) : Le cas $K' = 0$ correspond au cas où nous cherchons la valeur maximale pour aucune rupture. Nous avons donc directement (voir la figure A.7 (c)) :

$$I_0(n') = \max_{1<T_1=n'} \Delta(1 : T_1) = \Delta(1 : n').$$

(K'=1) : Étant fixé $n' > 1$, le cas $K' = 1$ peut être résolu linéairement en regardant toutes les ruptures T_1 possibles (voir la figure A.7 (d)) :

$$I_1(n') = \max_{1<T_1<T_2=n'} \{\Delta(1 : T_1) + \Delta(T_1 : n')\}$$

ce qui fait une complexité totale de $\mathcal{O}\left(n^2\right)$.

(K'=2) : Pour le cas $K' = 2$, nous commençons par remarquer que le maximum peut être fait en deux fois :

$$\begin{aligned} I_2(n') &= \max_{1<T_1<T_2<T_3=n'} \{\Delta(1 : T_1) + \Delta(T_1 : T_2) + \Delta(T_2 : n')\} \\ &= \max_{2<T_2<n'} \left[\max_{1<T_1<T_2} \{\Delta(1 : T_1) + \Delta(T_1 + 1 : T_2)\} + \Delta(T_2 : n') \right]. \end{aligned}$$

Donc, si nous connaissons la rupture T_2 optimale, la maximisation consiste à trouver la rupture T_1 optimale. Or, nous connaissons déjà la valeur optimale par l'étape précédente puisqu'elle vaut $I_1(T_2)$ (voir la figure A.7 (e)) et la maximisation ne dépend plus que de T_2 :

$$I_2(n') = \max_{2<T_2<n'} \left[I_1(T_2) + \Delta(T_2 : n') \right].$$

ce qui est fait avec une complexité totale de $\mathcal{O}\left(n^2\right)$ à nouveau.

L'algorithme de programmation dynamique est détaillé dans l'algorithme 2.

Input: Données $\boldsymbol{Y}$ et nombre de ruptures K

// Calcul de la matrice $[\Delta(n_1 : n_2)]_{0 \leq n_1 < n_2 \leq n}$

for n_1 allant de 0 à $n-1$ **do**
  for n_2 de $n+1$ à n **do**
    Calcul de $\Delta(n_1 : n_2)$

// Calcul de la matrice $[I_{K'}(n')]_{0 \leq K' < n' \leq n}$

Calcul de la première ligne $[I_0(n')]_{1 \leq n' \leq n}$ en prenant la première ligne $[\Delta(1 : n')]_{1 \leq n' \leq n}$ de la matrice $[\Delta(n_1 : n_2)]_{0 \leq n_1 < n_2 \leq n}$

for K' allant de 1 à K **do**
  for n' allant de $K'+1$ à n **do**
    Calcul de $I_{K'}(n')$:

$$I_{K'}(n') = \max_{K'<T<n'} \left[I_{K'-1}(T) + \Delta(T : n') \right]$$

Output: Estimation de la vraisemblance $I_K(n)$ et des instants de ruptures $\boldsymbol{T}$

Algorithm 2: Principe de l'algorithme de programmation dynamique.

Implémentations

Ainsi, nous pouvons remarquer qu'étant donnée la matrice $[\Delta(n_1 : n_2)]_{0\leq n_1<n_2\leq n}$, la complexité de l'algorithme de programmation dynamique est $\mathcal{O}(Kn^2)$ donc la complexité finale est le maximum entre le calcul de la matrice $[\Delta(n_1 : n_2)]_{0\leq n_1<n_2\leq n}$ et $\mathcal{O}(Kn^2)$.

Littérature connexe

Dans son article, Rigaill (2015) montre que, pour certains cas, il n'est pas nécessaire d'explorer toutes les cases de la matrice $[I_{K'}(n')]_{0\leq K'<n'\leq n}$ et qu'il est ainsi possible de parfois baisser la complexité jusqu'à $\mathcal{O}(Kn\log(n))$. Ceci est d'autant plus intéressant si le calcul de la matrice $[\Delta(n_1 : n_2)]_{0\leq n_1<n_2\leq n}$ est déjà fait ou de complexité minime.

Point de vigilance

La condition de maximisation séparée des segments de l'équation (A.5) peut être contraignante. Par exemple, il n'est pas possible d'avoir des liens entre deux moyennes successives ou, comme dans l'exemple fil rouge, il est nécessaire de séparer la maximisation de la variance. Il existe des parades pour contourner certains cas mais cela se fait au cas par cas.

Littérature connexe

Dans leur article, Chakar et al. (2017) étudient des processus *ARMA* avec des changements de paramètres entre deux ruptures. Comme les observations d'un processus *ARMA* sont corrélées avec leurs voisines, il n'est pas possible d'utiliser l'algorithme de programmation dynamique sur les données de base. L'idée proposée dans l'article est d'étudier des transformations des données (par exemple, sur la différence de deux observations dans le cas d'un processus *AR(1)*) afin d'avoir des observations décorrélées.

Utilisation des procédures *Least Absolute Shrinkage and Selection Operator* (*LASSO*)

Dans certains cas, estimer les paramètres et les ruptures est analogue à estimer les paramètres d'un **modèle linéaire sparse** (voir par exemple Harchaoui et Lévy-Leduc (2007); Vert et Bleakley (2010)). Dans le cas de l'exemple de la figure A.3, nous avons vu que la modélisation était :

$$\forall k \in \{0,\dots,3\},\ \forall i \in \{T_k+1,\dots,T_{k+1}\},\ Y_i \sim \mathcal{N}(\mu_k,\sigma^2).$$

Or, nous pouvons remarquer que cela revient à estimer les paramètres du modèle :

$$\boldsymbol{Y} = \mathbb{T}_n^{(1)}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{A.6}$$

avec $\boldsymbol{Y}$ le vecteur des observations, $\boldsymbol{\beta}$ le vecteur ne contenant que des 0 sauf pour les valeurs $(\beta_{T_k+1})_{0\leq k\leq K}$ où nous avons :

$$\beta_1 = \mu_0 \text{ et pour tout } k \in \{1,\dots,K\},\ \beta_{T_k+1} = \mu_k - \mu_{k-1}, \tag{A.7}$$

$\mathbb{T}_n^{(1)}$ est la matrice triangulaire inférieure de taille $n\times n$ ne contenant que des 1 et $\boldsymbol{\varepsilon}$ le vecteur aléatoire de longueur n et de loi $\mathcal{N}(\mathbb{0}_n,\sigma^2\mathbb{I}_n)$ où $\mathbb{0}_n$ est le vecteur nul de longueur n et $\mathbb{I}_n$ est la matrice identité de taille $n\times n$. Ainsi, grâce aux équations (A.6) et (A.7), nous voyons que $\mathbb{T}_n^{(1)}\boldsymbol{\beta}$ vaut μ_0 pour les T_1 premières valeurs puis $\mu_1-\mu_0+\mu_0=\mu_1$ pour les T_2-T_1 suivantes et ainsi de suite. De plus, il y a bien une bijection entre les deux modélisations et les emplacements des ruptures peuvent être retrouvés à l'aide des positions des valeurs non nulles dans le vecteur $\boldsymbol{\beta}$.

D'après les critères d'identifiabilité vus dans la section A.3.1, le vecteur $\boldsymbol{\beta}$ possède $K+\mathbb{1}_{\{\mu_0\neq 0\}}$ valeurs non nulles où $\mathbb{1}_{\{.\}}$ est la fonction indicatrice ce qui implique que le vecteur $\boldsymbol{\beta}$ peut être considéré comme sparse ; c'est-à-dire essentiellement composé de 0. Pour résoudre ce problème, il est donc possible d'utiliser la procédure *Least Absolute Shrinkage and Selection Operator* (***LASSO***) introduite par Tibshirani (1996). Le principe est de chercher l'estimateur $\widehat{\boldsymbol{\beta}}$ tel que $\mathbb{T}_n^{(1)}\widehat{\boldsymbol{\beta}}$ soit le plus proche possible des observations tout en essayant d'avoir un vecteur $\widehat{\boldsymbol{\beta}}$ le plus sparse possible. Pour ce faire, étant donnée une valeur λ strictement positive, l'estimateur *LASSO* minimise l'équation suivante :

$$\widehat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\operatorname{argmin}} \left[\left\| \boldsymbol{Y} - \mathbb{T}_n^{(1)}\boldsymbol{\beta} \right\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right] \tag{A.8}$$

avec $\|\cdot\|_2$ la **norme euclidienne** et $\|\cdot\|_1$ la **norme absolue**.

Dans l'équation (A.8), nous voyons que le paramètre λ joue un rôle de **pénalité** : plus sa valeur est élevée, plus il y aura de 0 dans l'estimateur $\widehat{\boldsymbol{\beta}}_\lambda$. À l'extrême, lorsque λ est suffisamment grand, la seule valeur possible pour $\widehat{\boldsymbol{\beta}}_\lambda$ est le vecteur nul $\mathbb{0}_n$. À l'opposé, si λ tend vers 0, il devient de moins en moins coûteux de prendre des valeurs non nulles et le nombre de zéros dans l'estimateur $\widehat{\boldsymbol{\beta}}_\lambda$ tend vers 0 ; nous faisons alors du surapprentissage.

L'estimation des emplacements des ruptures par cette procédure possède deux problèmes principaux :

- La première, inhérente aux méthodes *LASSO*, est le choix du *bon* paramètre λ pour avoir une bonne estimation du nombre de zéros (voir par exemple Meinshausen et Bühlmann (2010)).
- La seconde est que l'estimateur $\widehat{\boldsymbol{\beta}}_\lambda$ a tendance, empiriquement, à ne pas avoir une seule valeur non nulle mais plutôt une zone autour des valeurs non nulles.

Point de vigilance

Il est important de rappeler que nous ne pouvons utiliser cette méthode que lorsqu'il est possible de transformer le problème en un modèle linéaire sparse vérifiant l'équation (A.6).

Annexe B

Biclustering et modèle des blocs latents

"Si tu donnes un poisson à un homme, il mangera un jour. Si tu lui apprends à pêcher, il mangera toujours."
Lao Tseu

Dans ce chapitre, nous développons les travaux en lien avec le **modèle des blocs latents** que j'ai étudié durant ma thèse Brault (2014). En particulier, dans les articles Brault et Mariadassou (2015) et Brault et Lomet (2015), je rappelle notamment la place de ce modèle au sein de la famille du **biclustering** et les différences avec l'utilisation des modèles de mélange sur les lignes et sur les colonnes de façon indépendante (section B.1). Dans un second temps, je parle des algorithmes d'estimations, de leurs évaluations et du choix du nombre de blocs (section B.2).

B.1 Biclustering, co-clustering et modèle des blocs latents

Une façon d'étendre les modèles de mélange aux tableaux est de rechercher des **blocs atypiques** par rapport aux restes des données ; nous parlons alors de **biclustering**. La question se pose alors des contraintes que nous mettons pour rechercher ces blocs. Comme nous sommes dans le cas d'un modèle de mélange, l'ordre des lignes et des colonnes n'a pas d'importance et, étant donnée une matrice quelconque (voir en haut à gauche de la figure B.1), nous cherchons à permuter les lignes et les colonnes pour faire ressortir une structure particulière (voir un exemple en haut à droite de la figure B.1). Dans notre article Brault et Lomet (2015) et dans le premier chapitre de ma thèse (voir Brault (2014)), nous proposons une review non exhaustive mais plus détaillée de différentes méthodes.

Dans ce manuscrit, je présente trois exemples (voir le bas de la figure B.1) :

- La méthode la moins restrictive consiste à rechercher un ou plusieurs blocs avec des cases ayant une *structure cohérente* voir vérifiant un objectif particulier. En bas à gauche de la figure B.1, nous mettons en évidence trois blocs vérifiant à la fois une forte densité de cases noires tout en étant les plus grands possibles. Nous pouvons faire plusieurs remarques :
  - Le bloc 3 n'est composé que de cases noires ; si nous rajoutons la ligne H, nous augmentons la taille mais faisons diminuer la proportion de cases noires. Suivant les critères, cela peut être intéressant ou pas. Aucune autre permutation des lignes et les colonnes ne permet d'ajouter des cases noires à ce bloc car les autres cases des colonnes 3, 5 et 7 et des lignes B, F et J contiennent quasiment que des cases blanches.
  - Les blocs 1 et 2 se *chevauchent* c'est-à-dire que certaines cases se retrouvent dans les deux blocs. En l'occurrence, ce sont celles qui sont à la fois dans les colonnes 2 et 6 et les lignes A, C et H.
  - Les blocs 1 et 2 sont moins denses que le bloc 3 mais plus gros.
  - Il s'avère que dans cet exemple jouet, il est possible de représenter tous les blocs en une seule permutation des lignes et des colonnes. Toutefois, cela n'a pas de raisons d'être le cas à chaque fois puisque la recherche d'un nouveau bloc est indépendante des blocs précédents trouvés.
- Une méthode plus restrictive consiste à former des blocs de façon **hiérarchique** : nous cherchons d'abord un ou plusieurs blocs puis nous cherchons les blocs suivants au sein des blocs de l'étape

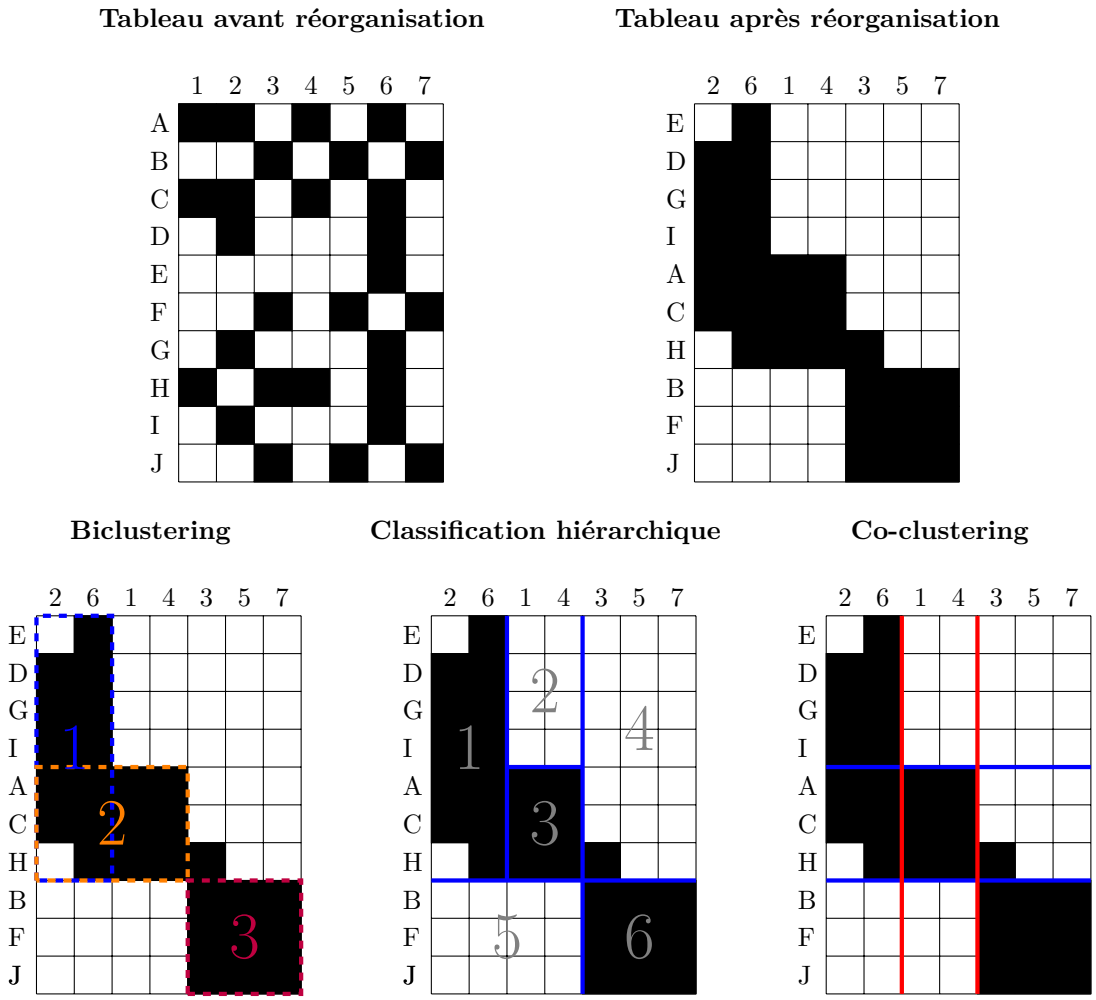

FIGURE B.1 – Exemple de matrices avec des données binaires à classifier (en haut à gauche) avec une réorganisation possible (en haut à droite). En bas à gauche, nous avons un exemple possible de *biclustering* avec trois blocs dont deux qui se chevauchent. Au centre en bas, nous avons une recherche de blocs par une procédure hiérarchique mettant en évidence six blocs. En bas à droite, un exemple de co-classification mettant en avant une structure en quadrillage. Exemple adapté de celui proposé dans l'article de Govaert et Nadif (2008).

précédente comme s'ils étaient maintenant la matrice initiale et ainsi de suite jusqu'à ce qu'il ne soit plus possible de cherche un nouveau bloc (par exemple si un bloc ne contient plus qu'une seule case). Dans l'exemple en bas, au centre de la figure B.1, nous séparons d'abord les blocs $\{1, 2, 3\}$ et le bloc $\{6\}$. Au sein du bloc $\{1, 2, 3\}$, nous différencions ensuite les blocs $\{1\}$ et $\{2, 3\}$ et enfin, nous isolons le bloc $\{3\}$. Suivant le critère, il peut aussi être intéressant de créer un sous bloc au sein du bloc $\{1\}$ en écartant les lignes E et H voir, dans ce cas, récupérer le bloc composé des cases de la colonne 6 et des lignes E et H. Pour les blocs $\{3\}$ et $\{6\}$, comme ils sont composés uniquement de cases noires, nous n'avons pas d'intérêt à les subdiviser.

- Enfin, le modèle étudié durant ma thèse appartient aux méthodes de **coclustering** ou **co-classification** ou encore **classification croisée** en français. Dans ce cadre, nous faisons une classification des lignes et une classification sur les colonnes de telle sorte que le produit cartésien des classifications met en évidence une structure particulière. Par exemple, en bas à droite de la figure B.1, la classification des lignes (représentée par des lignes horizontales bleues) et la classification des colonnes (lignes verticales rouges) mettent en évidence une structure de neuf blocs composés chacun et essentiellement de cases de même couleur (voir la figure 1.11 de ma thèse Brault (2014)).

Littérature connexe

Comme nous faisons une classification à la fois sur les lignes et les colonnes, il faut souligner que les variables doivent représenter *la même chose* (comme nous le faisons déjà pour les individus dans le cas de modèle de mélange). Par exemple, si nous classifions les réponses à des questions, la réponse à "Êtes-vous en couple ?" ne sera pas classée de la même façon que celle de la question "Êtes-vous célibataire ?". Pour explorer ce problème, j'encourage les personnes intéressées à lire l'article de Biernacki et Lourme (2019).

B.1.1 Modèle des blocs latents

Le **modèle des blocs latents** est un modèle probabiliste se plaçant dans le cadre de la classification croisée. Il se base sur trois hypothèses (voir la section 1.3 de ma thèse Brault (2014) pour des réflexions plus approfondies sur ces hypothèses). La figure B.2 présente un schéma des notations utilisées.

Hypothèse (Hypothèses du modèle des blocs latents)

$(\mathcal{H}_1)$: nous supposons qu'il existe une partition $\boldsymbol{z}$ des n lignes en K classes et une partition $\boldsymbol{w}$ des J colonnes en L classes de telle sorte que les variables aléatoires $Y_{i,j}$ soient indépendantes conditionnellement au couple $(\boldsymbol{z}, \boldsymbol{w})$:

$$p\left(\boldsymbol{y} \,|\boldsymbol{z}, \boldsymbol{w}; \boldsymbol{\Psi}\right) = \prod_{i=1}^{n}\prod_{j=1}^{J} p\left(y_{i,j} \,|z_i, w_j; \boldsymbol{\Psi}\right) \tag{B.1}$$

et leurs lois ne dépendent que du bloc (z_i, w_j).

$(\mathcal{H}_2)$: les variables $\boldsymbol{Z}$ et $\boldsymbol{W}$ sont indépendantes :

$$\forall (\boldsymbol{z}, \boldsymbol{w}) \in \mathcal{Z} \times \mathcal{W}, \quad p\left(\boldsymbol{z}, \boldsymbol{w}; \boldsymbol{\Psi}\right) = p\left(\boldsymbol{z}; \boldsymbol{\Psi}\right) p\left(\boldsymbol{w}; \boldsymbol{\Psi}\right) \tag{B.2}$$

où $\mathcal{Z}$ et $\mathcal{W}$ sont les espaces de partitions.

$(\mathcal{H}_3)$: pour chaque ligne i, la loi d'affectation des labels est une loi multinomiale dont les paramètres sont les mêmes pour toutes les lignes ; de même pour les colonnes. Comme pour les modèles de mélange, $\boldsymbol{\pi}$ est la probabilité d'appartenance aux classes en lignes. Pour les colonnes, la notation sera $\boldsymbol{\rho}$.

Comme dit dans l'hypothèse $(\mathcal{H}_1)$, nous pouvons choisir la loi que nous voulons pour chaque bloc. Néanmoins, dans l'essentiel de mes travaux, je me suis intéressé au cas où chaque bloc (k, ℓ) contient des observations issues de lois paramétriques de densité φ et paramètres $\boldsymbol{\theta}_{k,\ell}$ notamment aux **lois multinomiales** et au cas particulier des **lois de Bernoulli**.

Au final, grâce aux hypothèses précédentes, nous obtenons la vraisemblance suivante du modèle :

$$p\left(\boldsymbol{y}; \boldsymbol{\Psi}\right) = \sum_{(\boldsymbol{z},\boldsymbol{w})\in\mathcal{Z}\times\mathcal{W}} \underbrace{\left\{\left[\prod_{i=1}^{n}\prod_{j=1}^{J}\prod_{k=1}^{K}\prod_{\ell=1}^{L} \varphi\left(\boldsymbol{y}_{i,j}; \boldsymbol{\theta}_{k,\ell}\right)^{z_{i,k}w_{j,\ell}}\right]\right.}_{\text{Hypothèse } (\mathcal{H}_1)} \times \underbrace{\left.\left(\prod_{k=1}^{K} \pi_k^{z_{+,k}}\right)\left(\prod_{\ell=1}^{L} \pi_\ell^{w_{+,\ell}}\right)\right\}}_{\text{Hypothèses } (\mathcal{H}_2) \text{ et } (\mathcal{H}_3)} \tag{B.3}$$

où $\mathcal{Z} \times \mathcal{W}$ sont les ensembles de quadrillages possibles de la matrice. Sur la figure B.3, une représentation schématique des hypothèses précédentes a été proposée.

Remarque

Notons que, contrairement aux modèles de mélange, la factorisation proposée dans la section A.2.1 pour la vraisemblance n'est pas possible. Pour avoir le calcul exact de la vraisemblance, nous devons faire une somme sur la totalité des quadrillage possibles soient $K^n \times L^J$ possibilités. Ceci n'est pas possible dans un temps raisonnable tant que nous n'aurons pas accès à un ordinateur quantique. Néanmoins, nous voyons dans les sections B.2 et 2.3.1 des façons de contourner les problèmes qui en découlent.

Enfin, dans notre article Keribin et al. (2014), Christine Keribin a montré des conditions suffisantes pour avoir l'identifiabiltié du modèle dans le cas binaire :

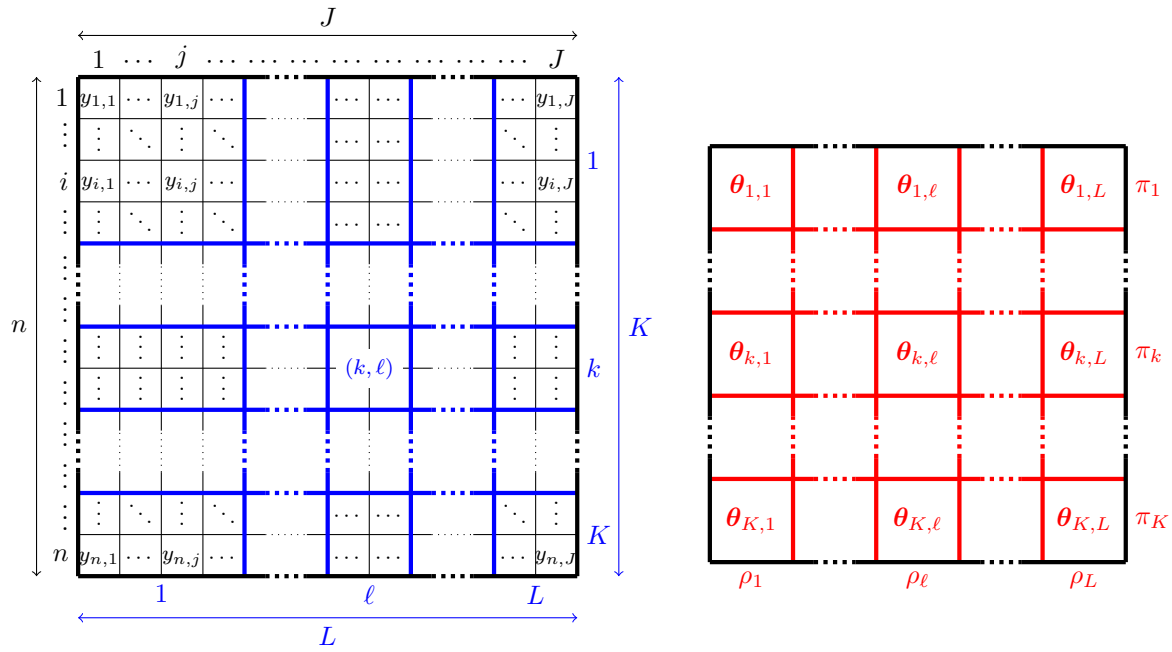

FIGURE B.2 – Représentation schématique des notations utilisées pour le modèles des blocs latents : sur la gauche sont représentées les notations liées aux données (en noir) et aux clusters (en bleu) tandis que, sur la figure de droite, nous avons représenté les notations liées aux paramètres $\boldsymbol{\Psi}$.

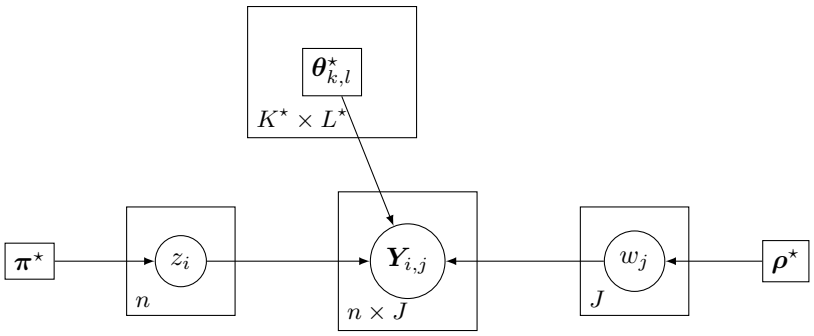

FIGURE B.3 – Représentation graphique du modèle des blocs latents paramétrique : les carrés correspondent à des paramètres et les ronds à des variables.

Hypothèse (Conditions suffisantes d'identifiabilité pour le cas binaire)

Dans le cas de données binaires où $\boldsymbol{\theta} = (\theta_{k,\ell})_{K\times L} \in [0,1]^{K\times L}$ est la matrice $K \times L$ des paramètres des lois de Bernoulli. Les trois conditions supposent que :

- $\left(\mathcal{I}d^{\text{LBM}}_{\boldsymbol{\pi}}\right)$: pour tout $1 \leq k \leq K$, $\pi_k > 0$ et les coordonnées du vecteur $\boldsymbol{\theta\rho}$ sont distinctes,
- $\left(\mathcal{I}d^{\text{LBM}}_{\boldsymbol{\rho}}\right)$: pour tout $1 \leq \ell \leq L$, $\pi_\ell > 0$ et les coordonnées du vecteur $\boldsymbol{\pi\theta}^T$ sont distinctes,
- $\left(\mathcal{I}d^{\text{LBM}}_{n,J}\right)$: $n \geq 2L-1$ et $J \geq 2K-1$.

Théorème 18 (Keribin et al. (2014))

Sous les conditions $\left(\mathcal{I}d^{\text{LBM}}_{\boldsymbol{\pi}}\right)$, $\left(\mathcal{I}d^{\text{LBM}}_{\boldsymbol{\rho}}\right)$ et $\left(\mathcal{I}d^{\text{LBM}}_{n,J}\right)$, le modèle des blocs latents binaire est identifiable.
Sous la condition $\left(\mathcal{I}d^{\text{LBM}}_{n,J}\right)$, l'ensemble des paramètres ne vérifiant pas les conditions $\left(\mathcal{I}d^{\text{LBM}}_{\boldsymbol{\pi}}\right)$

et $\left(\mathcal{I}d_{\boldsymbol{\rho}}^{\mathrm{LBM}}\right)$ étant de mesure de Lebesgue nulle, le modèle des blocs latents binaire est généralement identifiable. Enfin, dans le cas de deux classes en ligne et en colonne, Christine a montré que le modèle pouvait rester identifiable même sans vérifier ces conditions (mais en respectant les conditions de l'article Brault et al. (2020)).

Ces hypothèses sont réutilisées dans les section 2.3.1 et 2.3.2.

Remarque

Comme nous sommes dans le cas d'une classification des lignes et des colonnes, nous pouvons nous dire qu'il suffirait de faire une classification alternée des unes puis des autres. Or, dans l'article de Govaert et Nadif (2008) et dans notre article Brault et Mariadassou (2015), il a été montré qu'il y a un réel apport à faire une classification simultanée dans la plupart des cas.

B.1.2 Graphe biparti

Il existe un lien entre les **graphes bipartis** et le modèle des blocs latents (voir la figure B.4) dont les matrices sont dites **d'adjacence**. En effet, si nous supposons que les lignes représentent les noeuds d'un graphe qui interagissent avec les noeuds d'un autre graphe représentés par les colonnes (voir à gauche de la figure B.4) alors les cases représentent les liens entre ces noeuds. Dans ce cadre, le modèle des blocs latents permet de mettre en évidence des clusters de chaque graphe ayant des noeuds aux interactions similaires avec les clusters de l'autre graphe. Sur l'exemple, les noeuds D, E, G et I ont une forte interaction avec les noeuds 2 et 6 (en violet) uniquement. Les noeuds A, C et H ont des fortes interactions à la fois avec les noeuds 2 et 6 (en orange) mais aussi avec les noeuds 1 et 4 (en jaune). Enfin, les noeuds B, F et J et les noeuds 3, 5 et 7 interagissent presqu'exclusivement ensemble (en vert).

Le modèle des blocs latents permet aussi une représentation simplifiée de la matrice (voir le centre de la figure B.4) où nous ne représentons que les clusters et colorions les cases et arrête avec une couleur d'autant plus foncée qu'il y a en moyenne beaucoup d'interactions entre les noeuds des groupes. Dans ce cadre, nous voyons que quatre des neufs blocs sont assez foncés (voir en haut au centre de la figure B.4) et, en dehors des quatre blocs blancs, un seul est très clair ; il correspond au fait que le groupe $\{A, C, H\}$ ne possède qu'une seule interaction avec le groupe $\{3, 5, 7\}$.

Enfin, nous pouvons faire une représentation résumée en seuillant le tableau précédent et en représentant uniquement les interactions plus grandes que le seuil, nous obtenons un tableau et un graphe résumé (à droite de la figure B.4). Dans ce cadre, nous pouvons faire évoluer le seuil suivant l'intérêt recherché. Sur la droite de la figure B.4, tous les seuils entre 1/9 et 5/6 donnent la représentation proposée. Dans ce cadre, nous voyons deux clusters qui n'ont aucun lien l'un vers l'autre apparaître.

Point de vigilance

La structure en quadrillage du modèle des blocs latents peut être contraignante. En particulier, elle oblige à ce que toutes les lignes et les colonnes soient classées (ce qui nous a amenés à proposer un autre modèle développé dans la section 2.2.1) et empêche les structures moins rigides comme les *escaliers* et les *spirales* par exemple (voir l'introduction de notre article Brault et Mariadassou (2015)).

Littérature connexe

Dans le cas particulier où les lignes et les colonnes représentent les mêmes noeuds, nous sommes dans le cadre du **modèle à blocs stochastiques** (ou *Stochastic Block Model*). Les matrices d'adjacence sont carrées et la diagonale joue un rôle particulier puisque correspond au fait que les noeuds du graphe peuvent interagir (ou pas) avec eux mêmes. Les arrêtes du graphe peuvent être orientées ou non ; dans ce second cas, la matrice d'adjacence est symétrique. Dans son article, Keribin (2021) montre que dans le cas d'arêtes orientées, il est intéressant de ne plus supposer que les lignes et les colonnes représentent les mêmes noeuds et d'utiliser les modèles des blocs latents. Je fis notamment cette hypothèse dans le cadre d'une collaboration avec Yann Vasseur sur l'étude des interactions entre les 1937 facteurs de transcriptions de la plante *Arabidopsis thaliana* (voir le chapitre 6 de ma thèse Brault (2014)).

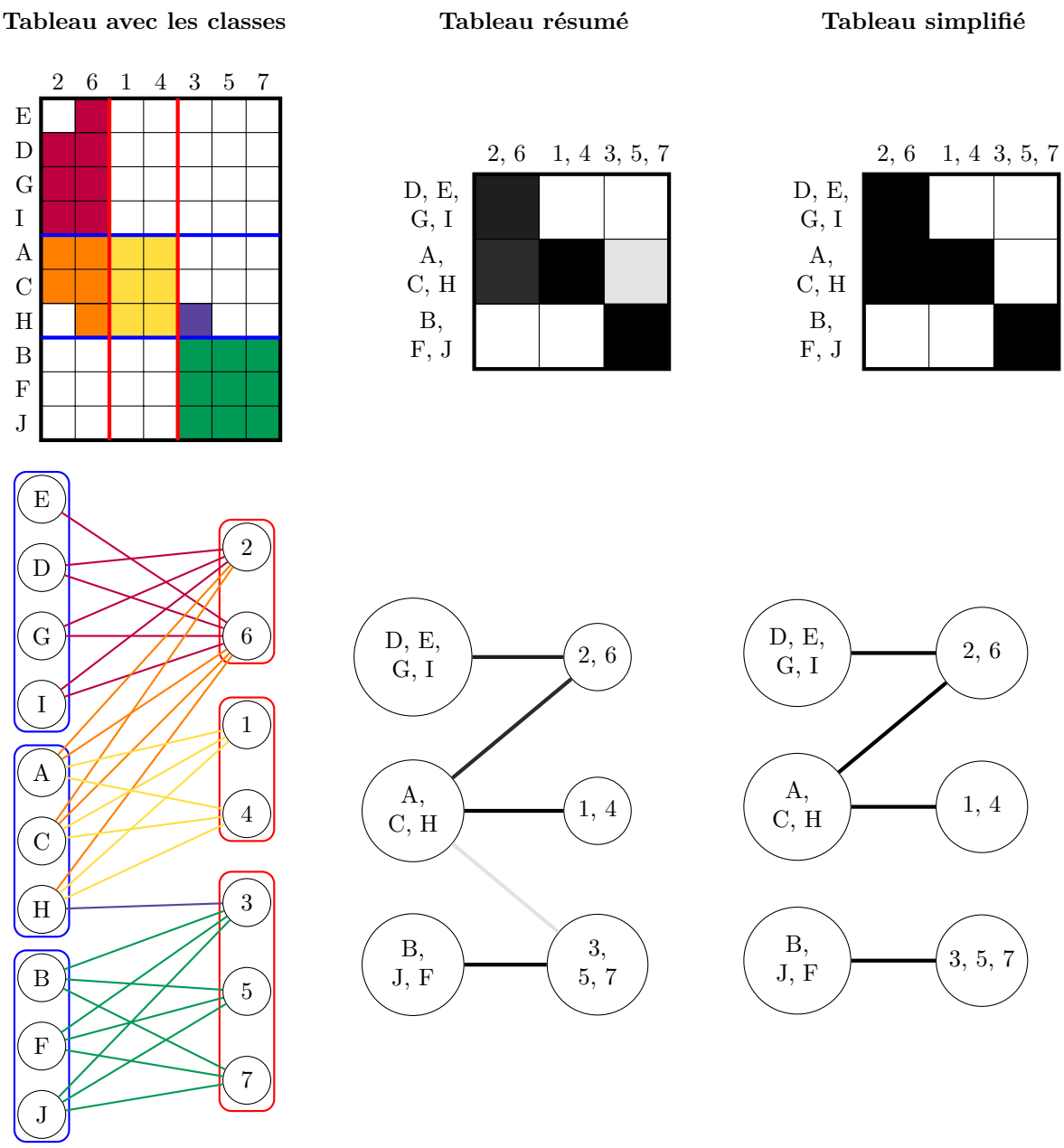

FIGURE B.4 – Représentation de la matrice de l'exemple de la figure B.1 (en haut) avec la représentation sous forme de graphe biparti (en bas). À gauche, la réorganisation suivant un modèle de co-clustering de la matrice avec des couleurs pour les blocs et la représentation en graphe biparti avec les liens entre les noeuds. Au centre, la version simplifiée en coloriant plus ou moins noir suivant l'intensité entre les blocs. À droite, la version discrétisée où seules les probabilités moyennes plus grandes que 0.5 sont conservées.

B.2 Estimation

Comme nous avons vu dans la section A.2.3, une méthode pour estimer les paramètres du modèle de mélange est d'utiliser l'algorithme *EM*. Dans le cas du modèle des blocs latents, l'espérance de la

vraisemblance complète comme vue dans l'équation (A.2) est la suivante :

$$\begin{aligned} Q\left(\boldsymbol{\Psi}\middle|\boldsymbol{\Psi}^{(c)}\right) &= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{W}|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}}\left[\log p\left(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\Psi}\right)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{\ell=1}^{L}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{W}|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}}\left[Z_{i,k}W_{j,\ell}\right]\log\varphi\left(\boldsymbol{Y}_{i,j};\boldsymbol{\theta}_{k,\ell}\right) \\ &\quad + \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{W}|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}}\left[Z_{i,k}\right]\log\pi_k + \sum_{\ell=1}^{L}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{W}|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}}\left[W_{j,\ell}\right]\log\pi_\ell \\ &= \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{\ell=1}^{L}\mathbb{P}\left(Z_{i,k}W_{j,\ell}=1\middle|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}\right)\log\varphi\left(\boldsymbol{Y}_{i,j};\boldsymbol{\theta}_{k,\ell}\right)}_{\text{Partie sur les } \boldsymbol{Y}_{i,j}} \\ &\quad + \underbrace{\sum_{k=1}^{K}\mathbb{P}\left(Z_{i,k}=1\middle|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}\right)\log\pi_k + \sum_{\ell=1}^{L}\mathbb{P}\left(W_{j,\ell}=1\middle|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}\right)\log\pi_\ell}_{\text{Partie sur les } (\boldsymbol{Z},\boldsymbol{W})}. \end{aligned} \tag{B.4}$$

Or, s'il existe une formule analytique pour calculer la loi de $\boldsymbol{Z}$ (ou $\boldsymbol{W}$) connaissant les observations, il faut par contre faire une somme sur toutes les partitions vérifiant à la fois que $z_{i,k}=1$ et $w_{j,\ell}=1$ ce qui est quasiment la même taille que pour l'ensemble $\mathcal{Z}\times\mathcal{W}$. Pour les mêmes raisons que pour le calcul de la vraisemblance, l'algorithme *EM* n'est pas utilisable en un temps raisonnable pour le moment.

B.2.1 Énergie libre et algorithme *Variational Expectation Maximisation*

Pour contourner ce problème, Govaert et Nadif (2003) proposent d'utiliser une approximation variationnelle. Le principe est d'introduire une distribution quelconque $q_{\boldsymbol{Z},\boldsymbol{W}}$ du couple $(\boldsymbol{Z},\boldsymbol{W})$ et de décomposer la logvraisemblance de la manière suivante (voir par exemple Keribin (2009)) :

$$\log p\left(\boldsymbol{y};\boldsymbol{\Psi}\right) = \underbrace{\mathbb{E}_{(\boldsymbol{Z},\boldsymbol{W})\sim q_{\boldsymbol{Z},\boldsymbol{W}}}\left[\log\left(\frac{p\left(\boldsymbol{y},\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\Psi}\right)}{q_{\boldsymbol{Z},\boldsymbol{W}}(\boldsymbol{Z},\boldsymbol{W})}\right)\right]}_{=:\mathcal{F}(q_{\boldsymbol{Z},\boldsymbol{W}};\boldsymbol{\Psi})} + \mathrm{D}_{\mathrm{KL}}\left(q_{\boldsymbol{Z},\boldsymbol{W}}\,||\,p\left(\cdot,\cdot\,|\boldsymbol{y};\boldsymbol{\Psi}\right)\right) \tag{B.5}$$

où $\mathcal{F}$, appelée **l'énergie libre**, est une fonction sur les distributions possibles $q_{\boldsymbol{Z},\boldsymbol{W}}$ et $\mathrm{D}_{\mathrm{KL}}\left(\cdot\,||\,\cdot\right)$ est la **divergence de Kullback-Leiber** entre deux distributions. Le principe est que si nous prenons comme distribution $q_{\boldsymbol{Z},\boldsymbol{W}}=p\left(\cdot,\cdot\,|\boldsymbol{y};\boldsymbol{\Psi}\right)$ alors maximiser la logvraisemblance et l'énergie libre revient au même objectif puisque la divergence de Kullback-Leibler sera nulle.

L'idée est de prendre une famille de distributions suffisamment restreinte pour que les calculs soient possibles mais aussi suffisamment grande pour que la divergence de Kullback-Leibler soit la plus petite possible. La proposition de Govaert et Nadif (2003) est de faire une **approximation en champs moyen** c'est-à-dire de ne prendre que les distributions où les variables $\boldsymbol{Z}$ et $\boldsymbol{W}$ sont supposées indépendantes :

$$q_{\boldsymbol{Z},\boldsymbol{W}}(\boldsymbol{Z},\boldsymbol{W}) = q_{\boldsymbol{Z}}(\boldsymbol{Z})q_{\boldsymbol{W}}(\boldsymbol{W}) \tag{B.6}$$

ce qui revient à faire à peu près l'hypothèse suivante dans la formule (B.4) :

$$\mathbb{P}\left(Z_{i,k}W_{j,\ell}=1\middle|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}\right) \approx \mathbb{P}\left(Z_{i,k}=1\middle|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}\right)\times\mathbb{P}\left(W_{j,\ell}=1\middle|\boldsymbol{Y};\boldsymbol{\Psi}^{(c)}\right)$$

et les calculs sont réalisables avec une complexité linéaire avec le nombre de cases et de classes du modèle. Cet algorithme est appelé ***Variationnal Expectation Maximisation*** et l'étape E se fait par maximisation alternée des distributions sur $\boldsymbol{Z}$ et $\boldsymbol{W}$ (en pratique Govaert et Nadif (2008) montrent qu'il est plus efficace en terme de temps et de qualité d'estimation de ne faire qu'une maximisation partielle avant de passer à l'étape M).

Dans notre article Keribin et al. (2014), nous montrons empiriquement que l'algorithme *VEM* a tendance à vider des classes et nous proposons une version bayésienne détaillée dans la section B.2.3. De plus, dans notre article Brault et al. (2020), nous montrons que l'estimateur du maximum de l'énergie libre est consistant (voir la section 2.3.1).

B.2.2 Algorithme *Stochastic Expectation Maximisation*

Comme expliqué dans la section A.2.3, l'algorithme *EM* et a fortiori l'algorithme *VEM* sont sensibles aux initialisations. Pour limiter ce problème, Keribin et al. (2012) proposent d'utiliser une version stochastique en simulant les couples $(\boldsymbol{Z}, \boldsymbol{W})$ à la sortie de l'étape E au lieu de prendre les probabilités d'appartenance ; c'est l'algorithme ***Stochastic Expectation Maximisation*** ou *SEM*.

Cet algorithme souffre également du problème de classes qui se vident : d'abord dans le cas général, Celeux et al. (1995) montrent qu'une fois qu'une classe se vidait dans un *SEM*, elle le resterait jusqu'à la fin puisque la probabilité d'être a priori dans la classe était nulle. Dans leurs simulations, Keribin et al. (2012) remarquent empiriquement que lorsque les valeurs des Bernoulli tendaient vers 0 ou 1, cela implique souvent un cas dégénéré et proposaient une alternative en remettant arbitrairement la valeur du paramètres à 0.1 ou 0.9. Dans la section 2.A de mon manuscrit de thèse (voir Brault (2014)), je montre qu'il existe du coup une autre possibilité pour le modèle des blocs latents binaires de dégénérer : lorsqu'un bloc est composé d'une même valeur, nous sommes dans un état absorbant de la chaîne et nous ne pouvons jamais visiter toutes les possibilités. Ceci est dû au fait que la loi estimée dans l'étape M est une *loi dégénérée* qui n'accepte qu'une seule valeur possible pour tout le bloc.

B.2.3 Algorithme *V-Bayes* et échantillonneur de *Gibbs*

Pour contourner ces problèmes de classes qui se vident, nous proposons dans notre article Keribin et al. (2014) d'adopter un point de vue bayésien pour le cas binaire. Pour cela, nous ajoutons des lois de Dirichlet sur les paramètres (voir le schéma de la figure B.5) :

$$\boldsymbol{\pi} \sim \mathcal{D}ir\left(a, \ldots, a\right), \quad \boldsymbol{\rho} \sim \mathcal{D}ir\left(a, \ldots, a\right) \quad \text{et} \quad \forall (k,\ell),\, \theta_{k,\ell} \sim \mathcal{B}e\left(b, b\right) \tag{B.7}$$

avec a et b des hyperparamètres strictement positifs. L'idée est basée sur l'article de Frühwirth-Schnatter (2011) qui s'est aperçue que, contrairement au cas classique où on prend des valeurs pour a et b égales à 1/2 (loi de Jeffrey) ou 1 (loi uniforme), si nous prenons des valeurs de a plus grandes que 1 (dans son article, elle préconise 4 ou 16), les échantillonneurs de Gibbs qu'elle utilise pour des modèles de mélange vident moins les classes.

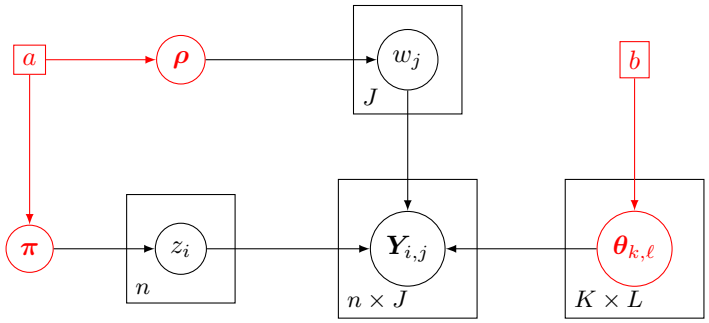

FIGURE B.5 – Représentation graphique du modèle des blocs latents paramétrique bayésien : les carrés correspondent à des hyperparamètres et les ronds à des variables. En rouge se trouvent les ajouts par rapport à la version fréquentiste du *LBM* présentée dans la figure B.3

Dans notre article Keribin et al. (2014), nous proposons deux extensions :

- L'algorithme *V-Bayes* a la même structure que l'algorithme *VEM* mais en tenant compte de la loi a priori. Ainsi, nous montrons que l'hyperparamètre a permet d'avoir une valeur minimale pour les probabilités des classes dès que a est strictement plus grand que 1 et ainsi, elles ne sont jamais estimées à 0 même quand la classe est vide.
- L'échantillonneur de *Gibbs* généralise l'algorithme *SEM* en simulant les variables $(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\theta})$ en plus des partitions $(\boldsymbol{Z}, \boldsymbol{W})$. L'avantage est que lorsqu'une classe se vide, l'échantillonneur continue de simuler des paramètres aux classes correspondantes ce qui laisse une chance de remplir de nouveau cette classe [1].

1. Une vidéo d'un exemple de chaîne où une classe se vide puis se remplit est disponible sur ma chaîne Youtube : https://youtu.be/-r_VaR0E6ZA?si=JCnK_MGs9RIkNRMb

Nous testons empiriquement les comportements de ces algorithmes et comparons les versions fréquentistes et bayésiennes (voir notre article Keribin et al. (2014)) :

- Lorsque la valeur de a est strictement plus grande que 1, l'algorithme *V-Bayes* a tendance à moins vider les classes que l'algorithme *VEM*.
- À l'opposé, lorsque la valeur de b est strictement plus grande que 1, l'algorithme *V-Bayes* et l'échantillonneur de *Gibbs* ont tendance à avoir des classes vides à la fin lors de l'affectation des individus aux classes.

B.2.4 Sélection de modèle

La dernière question est celle de la sélection du nombre de classes en lignes et en colonnes. Comme vue dans la section A.2.4, il existe pour les modèles de mélange des critères comme le critère *BIC* (*Bayesian Information Criterion*) de Schwarz et al. (1978) qui pénalise la logvraisemblance. Or, nous avons vu que la logvraisemblance n'est pas calculable en un temps raisonnable. Une première façon de procéder est de remplacer la logvraisemblance par l'énergie libre (B.5) dans le cadre d'une approximation en champs moyen (B.6). L'inconvénient est que si nous montrons dans notre article Brault et al. (2020) que les emplacements des maximums de la logvraisemblance et de l'énergie libre sont asymptotiquement proches (voir la section 2.3.1) lorsque nous avons le bon nombre de classes, nous ne connaissons pas la valeur de la divergence de Kullback-Leibler ni le comportement lorsque ce n'est pas le cas. Sur les simulations de notre article Keribin et al. (2014), nous remarquons empiriquement que le critère *BIC* a tendance à sous-estimer le nombre de classes.

Partant de ce constat et comme nous proposons des adaptations bayésiennes des algorithmes usuels, nous étudions, dans le même article, le critère *ICL* (*Integrated Complete Likelihood*)) qui, étant donné un nombre de clusters (K, L), l'estimateur du maximum de vraisemblance et une partition $\left(\widehat{\boldsymbol{Z}}_K, \widehat{\boldsymbol{W}}_L\right)$ estimée par maximum a posteriori, nous conservons le modèle qui maximise la logvraisemblance complète intégrée $ICL\,(K,L) = \log p\left(\boldsymbol{y}, \widehat{\boldsymbol{Z}}_K, \widehat{\boldsymbol{W}}_L\right)$ (voir par exemple Biernacki et al. (2000)). Ce critère a l'avantage de posséder une forme exacte calculable (étant donnés une partition et des hyperparamètres) et nous montrons empiriquement dans notre article (voir Keribin et al. (2014)) qu'il est plus consistant que le critère *BIC*.

Perspectives

La conjecture que nous proposons dans notre article Keribin et al. (2014) est qu'asymptotiquement, les critères *BIC* et *ICL* ont le même comportement. Comme expliqué dans la section 2.3, si on peut admettre que la logvraisemblance complète et la logvraisemblance ont le même comportement pour le bon nombre de clusters, nous ne connaissons pas encore leurs comportements lorsque nous nous plaçons dans un cadre différent.

Bibliographie

S. Aminikhanghahi et D. J. Cook. A survey of methods for time series change point detection. Knowledge and information systems, 51(2) :339–367, 2017.

G. André, V. Kostrubiec, J.-C. Buisson, J.-M. Albaret, et P.-G. Zanone. A parsimonious oscillatory model of handwriting. Biological cybernetics, 108(3) :321–336, 2014.

T. B. Arnold et R. J. Tibshirani. genlasso : Path Algorithm for Generalized Lasso Problems, 2022. URL https://CRAN.R-project.org/package=genlasso. R package version 1.6.1.

T. Asselborn, T. Gargot, L. Kidziński, W. Johal, D. Cohen, C. Jolly, et P. Dillenbourg. Automated human-level diagnosis of dysgraphia using a consumer tablet. NPJ digital medicine, 1(1) :42, 2018.

F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Convex optimization with sparsity-inducing norms. Optimization for Machine Learning, pages 19–53, 2011.

C. Baey et P.-H. Cournède. Using a hierarchical segmented model to assess the dynamics of leaf appearance in plant populations. Dans 14th Applied Stochastic Models and Data Analysis International Conference (ASMDA 2011), 2011.

J. Bakker, T. Weaver, S. Parthasarathy, et M. Aloia. Adherence to cpap : what should we be aiming for, and how can we get there ? Chest, 155 :1272–87, 2019.

J.-M. Bardet, V. Brault, S. Dachian, F. Enikeeva, et B. Saussereau. Change-point detection, segmentation, and related topics. ESAIM : Proceedings and Surveys, 68 :97–122, 2020.

O. Barndorff-Nielsen. Information and exponential families : in statistical theory. John Wiley & Sons, 2014.

P. Bastide. Shifted stochastic processes evolving on trees : application to models of adaptive evolution on phylogenies. PhD thesis, Université Paris Saclay (COmUE), 2017.

J.-P. Baudry, A. E. Raftery, G. Celeux, K. Lo, et R. Gottardo. Combining mixture components for clustering. Journal of computational and graphical statistics, 19(2) :332–353, 2010.

J.-P. Baudry, C. Maugis, et B. Michel. Slope heuristics : overview and implementation. Statistics and Computing, 22(2) :455–470, 2012.

R. Bellman. On the approximation of curves by line segments using dynamic programming. Communications of the ACM, 4(6) :284, 1961.

J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, et G. Peyré. Iterative bregman projections for regularized transportation problems. SIAM Journal on Scientific Computing, 37(2) :A1111–A1138, 2015.

P. Bickel, D. Choi, X. Chang, et H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. The Annals of Statistics, pages 1922–1943, 2013.

C. Biernacki et A. Lourme. Unifying data units and models in (co-) clustering. Advances in Data Analysis and Classification, 13 :7–31, 2019.

C. Biernacki, G. Celeux, et G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(7) :719–725, 2000.

C. M. Bishop et N. M. Nasrabadi. Pattern recognition and machine learning, volume 4. Springer, 2006.

B. E. Boser, I. M. Guyon, et V. N. Vapnik. A training algorithm for optimal margin classifiers. Dans Proceedings of the fifth annual workshop on Computational learning theory, pages 144–152, 1992.

V. Brault. Estimation et sélection de modèle pour le modèle des blocs latents. PhD thesis, Paris 11, 2014.

V. Brault et A. Channarond. Fast and consistent algorithm for the latent block model. Computational Statistics, 39(3) :1621–1657, 2023.

V. Brault et J. Chiquet. blockseg : Two dimensional change-points detection. https://cran.r-project.org/web/packages/blockseg/index.html, 2018.

V. Brault et A. Lomet. Revue des méthodes pour la classification jointe des lignes et des colonnes d'un tableau. Journal de la Société Française de Statistique, 156(3) :27–51, 2015.

V. Brault et M. Mariadassou. Co-clustering through latent bloc model : A review. Journal de la Société Française de Statistique, 156(3) :120–139, 2015.

V. Brault, J. Chiquet, et C. Lévy-Leduc. Fast Detection of Block Boundaries in Block-Wise Constant Matrices, pages 214–228. Springer International Publishing, Cham, 2016. ISBN 978-3-319-41920-6. doi : 10.1007/978-3-319-41920-6_16. URL http://dx.doi.org/10.1007/978-3-319-41920-6_16.

V. Brault, A. Channarond, et V. Robert. Généralisation de l'algorithme largest gaps pour le modèle des blocs latents non-paramétrique. Dans 49èmes Journées de Statistique, 2017a.

V. Brault, J. Chiquet, et C. Lévy-Leduc. Efficient block boundaries estimation in block-wise constant matrices : An application to hic data. Electron. J. Statist., 11(1) :1570–1599, 2017b. ISSN 1935-7524. doi : 10.1214/17-EJS1270.

V. Brault, M. Delattre, E. Lebarbier, T. Mary-Huard, et C. Lévy-Leduc. Estimating the number of block boundaries from diagonal blockwise matrices without penalization. Scandinavian Journal of Statistics, pages n/a–n/a, 2017c. ISSN 1467-9469. doi : 10.1111/sjos.12266. URL http://dx.doi.org/10.1111/sjos.12266.

V. Brault, G. Cougoulat, S. Ouadah, et L. Sansonnet. Muchpoint : Multiple change point. https://CRAN.R-project.org/package=MuChPoint, 2018a.

V. Brault, C. Lévy-Leduc, A. Mathieu, et A. Jullien. Change-point estimation in the multivariate model taking into account the dependence : Application to the vegetative development of oilseed rape. Journal of Agricultural, Biological and Environmental Statistics, pages 1–16, 2018b.

V. Brault, S. Ouadah, L. Sansonnet, et C. Lévy-Leduc. Nonparametric multiple change-point estimation for analyzing large hi-c data matrices. Journal of Multivariate Analysis, 165 :143–165, 2018c.

V. Brault, A. Samson, et J.-C. Quinton. Modélisation statistique pour détecter des séquences vidéos similaires : application aux véhicules autonomes. Dans SMF 2018 : Congrès de la Société Mathématique de France, pages 375–390, 2018d. https://smf.emath.fr/publications/smf-2018-congres-de-la-societe-mathematique-de-france.

V. Brault, C. Keribin, et M. Mariadassou. Consistency and asymptotic normality of latent block model estimators. Electron. J. Statist., 14(1) :1234–1268, 2020. ISSN 1935-7524. doi : 10.1214/20-EJS1695.

V. Brault, S. Coulange, F. Letué, M.-P. Jouannaud, M.-J. Martinez, et A.-C. Perret. Comment former des groupes d'étudiants homogènes à partir des résultats de self ? présentation d'un outil d'aide à la décision pour la création de groupes. Mediazioni. Rivista online du studi interdisciplinari su lingue e culture, 32 :185–203, 2021a.

V. Brault, B. Mallein, et J.-F. Rupprecht. Group testing as a strategy for covid-19 epidemiological monitoring and community surveillance. PLOS Computational Biology, 17(3) :e1008726, 2021b.

V. Brault, S. Arlot, J.-P. Baudry, C. Maugis, et B. Michel. capushe : CAlibrating Penalities Using Slope HEuristics, 2023a. URL https://CRAN.R-project.org/package=capushe. R package version 1.1.2.

V. Brault, E. Lebrarbier, A. Rosier, et V. Stoppin-Mellet. Classification contrainte de signaux, application à l'étude de la protéine neuronale tau. Dans 54es Journées de Statistique de Société Française de Statistique, 2023b.

V. Brault, E. Devijver, et C. Laclau. Mixture of segmentation for heterogeneous functional data. Electron. J. Statist., 2024a.

V. Brault, F. Letué, et M.-J. Martinez. Examining the robustness of a model selection procedure in the binary latent block model through a language placement test data set. 2024b.

V. Brault, E. Lebarbier, A. Rosier, et V. Stoppin-Mellet. Classification of multiple segmented profiles, application to the study of the neuronal protein tau. En cours de rédaction, 2025.

A. Casa, C. Bouveyron, E. Erosheva, et G. Menardi. Co-clustering of time-dependent data via the shape invariant model. Journal of Classification, 38 :626–649, 2021.

G. Celeux et G. Govaert. A classification em algorithm for clustering and two stochastic versions. Computational statistics & Data analysis, 14(3) :315–332, 1992.

G. Celeux, D. Chauveau, et J. Diebolt. On stochastic versions of the EM algorithm. PhD thesis, INRIA, 1995.

G. Celeux, M. Hurn, et C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association, 95(451) :957–970, 2000.

S. Chakar, E. Lebarbier, C. Lévy-Leduc, et S. Robin. A robust approach for estimating change-points in the mean of an ar(1) process. 2017.

G. Challet-Bouju, V. Brault, B. Perrot, . JEU-GROUP, D. Solène, , et M. Grall-Bronnec. A clustering based on the dynamics of dsm-5 criteria for gambling disorder : a 5-year follow-up of gamblers with and without gambling disorder. En cours de rédaction, 2025.

A. Channarond, J.-J. Daudin, S. Robin, et al. Classification and estimation in the stochastic blockmodel based on the empirical degrees. Electronic Journal of Statistics, 6 :2574–2601, 2012.

M. Charles, R. Soppelsa, et J.-M. Albaret. Bhk : échelle d'évaluation rapide de l'écriture chez l'enfant. Ecpa, 2004.

P. L. de Micheaux, R. Drouilhet, et B. Liquet. The R software. Springer, 2013.

A. P. Dempster, N. M. Laird, et D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society : Series B (Methodological), 39(1) :1–22, 1977.

L. Deschamps, C. Gaffet, S. Aloui, J. Boutet, V. Brault, et E. Labyt. Methodological issues in the creation of a diagnosis tool for dysgraphia. npj Digital Medicine, 2(1) :36, 2019.

L. Deschamps, L. Devillaine, C. Gaffet, R. Lambert, S. Aloui, J. Boutet, V. Brault, E. Labyt, et C. Jolly. Development of a pre-diagnosis tool based on machine learning algorithms on the bhk test to improve the diagnosis of dysgraphia. Advances in Artificial Intelligence and Machine Learning, pages https–www, 2021.

L. Devillaine, R. Lambert, J. Boutet, S. Aloui, V. Brault, C. Jolly, et E. Labyt. Analysis of graphomotor tests with machine learning algorithms for an early and universal pre-diagnosis of dysgraphia. Sensors, 21(21) :7026, 2021.

J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, et B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485(7398) :376–380, 2012.

R. Dorfman. The detection of defective members of large populations. The Annals of Mathematical Statistics, 14(4) :436–440, 1943.

J.-J. Droesbeke, G. Saporta, et C. Thomas-Agnan. Modèles à variables latentes et modèles de mélange. Editions TECHNIP, 2013.

A. Elie, E. Prezel, C. Guérin, E. Denarier, S. Ramirez-Rios, L. Serre, A. Andrieux, A. Fourest-Lieuvin, L. Blanchoin, et I. Arnal. Tau co-organizes dynamic microtubule and actin networks. Scientific reports, 5(1) :9964, 2015.

K. Eybpoosh, E. Vernet, S. Bailly, J.-L. Pépin, A. Samson, et V. Brault. Clustering sleep apnea compliance data with a regression mixture model based on a markov chain. En cours de rédaction, 2025.

E. B. Fowlkes et C. L. Mallows. A method for comparing two hierarchical clusterings. Journal of the American statistical association, 78(383) :553–569, 1983.

S. Frühwirth-Schnatter. Dealing with label switching under model uncertainty. Dans Mixtures : Estimation and Applications, chapter 10, pages 213–239. Wiley, 2011.

R. Genuer, J.-M. Poggi, et C. Tuleau. Random forests : some methodological insights. arXiv preprint arXiv :0811.3619, 2008.

G. Govaert et M. Nadif. Clustering with block mixture models. Pattern Recognition, 36(2) :463–473, 2003.

G. Govaert et M. Nadif. Block clustering with bernoulli mixture models : Comparison of different approaches. Computational Statistics & Data Analysis, 52(6) :3233–3245, 2008.

M. Gyllenberg, T. Koski, E. Reilink, et M. Verlaan. Non-uniqueness in probabilistic numerical identification of bacteria. Journal of Applied Probability, 31(2) :542–548, 1994.

L. Hamstra-Bletz, J. DeBie, B. Den Brinker, et al. Concise evaluation scale for children's handwriting. Lisse : Swets, 1 :623–662, 1987.

Z. Harchaoui et C. Lévy-Leduc. Catching change-points with lasso. Dans NIPS, volume 617, page 624, 2007.

C. A. Hoare. Algorithm 65 : find. Communications of the ACM, 4(7) :321–322, 1961.

H. Hoefling. A path algorithm for the fused lasso signal approximator. Journal of Computational and Graphical Statistics, 19(4) :984–1006, 2010.

H. Hoefling. flsa : Path Algorithm for the General Fused Lasso Signal Approximator, 2024. URL https://CRAN.R-project.org/package=flsa. R package version 1.5.5.

L. Hubert et P. Arabie. Comparing partitions. Journal of classification, 2 :193–218, 1985.

J. Jia et K. Rohe. Preconditioning the lasso for sign consistency. Electronic Journal of Statistics, 9 : 1150–1172, 2015.

K. Johnstone. Impro : Improvisation and the theatre. Routledge, 2012.

E. Joly et B. Mallein. A tractable non-adaptative group testing method for non-binary measurements. ESAIM : Probability and Statistics, 26 :283–303, 2022.

T. C. Jones, B. Mühlemann, T. Veith, G. Biele, M. Zuchowski, J. Hofmann, A. Stein, A. Edelmann, V. M. Corman, et C. Drosten. An analysis of sars-cov-2 viral load by patient age. MedRxiv, pages 2020–06, 2020.

A. Jullien, A. Mathieu, J.-M. Allirand, A. Pinet, P. De Reffye, P.-H. Cournède, et B. Ney. Characterization of the interactions between architecture and source–sink relationships in winter oilseed rape (brassica napus) using the greenlab model. Annals of botany, 107(5) :765–779, 2011.

C. Keribin. Consistent estimation of the order of mixture models. Sankhyā : The Indian Journal of Statistics, Series A, pages 49–66, 2000.

C. Keribin. Les méthodes bayésiennes variationnelles et leur application en neuroimagerie : une étude de l'existant. PhD thesis, INRIA, 2009.

C. Keribin. Cluster or co-cluster the nodes of oriented graphs ? Journal de la société française de statistique, 162(1) :46–69, 2021.

C. Keribin, V. Brault, G. Celeux, G. Govaert, et al. Model selection for the binary latent block model. Dans Proceedings of COMPSTAT, volume 2012, 2012.

C. Keribin, V. Brault, G. Celeux, et G. Govaert. Estimation and selection for the latent block model on categorical data. Statistics and Computing, pages 1–16, 2014. ISSN 0960-3174. doi : 10.1007/s11222-014-9472-2. URL http://dx.doi.org/10.1007/s11222-014-9472-2.

N. Kribbs, A. Pack, L. Kline, et et al. Objective measurement of patterns of nasal cpap use by patients with obstructive sleep apnea. Am Rev Respir Dis, 147 :887–895, 1993.

H. W. Kuhn. The hungarian method for the assignment problem. Naval research logistics quarterly, 2 (1-2) :83–97, 1955.

C. Laclau et V. Brault. Noise-free latent block model for high dimensional data. Data Mining and Knowledge Discovery, 2 :446–473, 2019.

C. Laclau et M. Nadif. Diagonal co-clustering algorithm for document-word partitioning. Dans International Symposium on Intelligent Data Analysis, pages 170–180. Springer, 2015.

C. Laclau, I. Redko, B. Matei, Y. Bennani, et V. Brault. Co-clustering through optimal transport. Dans International Conference on Machine Learning, pages 1955–1964. PMLR, 2017.

P. Lacroix. Contributions to variable selection in high-dimension and its uses in biology. PhD thesis, Université Paris-Saclay, 2022.

A. Lacroux et C. Martin-Lacroux. Should i trust the artificial intelligence to recruit ? recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening. Frontiers in Psychology, 13 :895997, 2022.

J. H. Lambert. Observationes variae in mathesin puram. Acta Helvetica, 3(1) :128–168, 1758.

M. Leroy, V. Brault, S. Coulange, M. Jouannaud, F. Letué, M.-J. Martinez, et A.-C. Perret. Robustesse dans le modèle des blocs latents : application au test de positionnement en langues self. Dans 52 èmes journées de la statistique de la SFdS, 2021.

C. Lévy-Leduc, M. Delattre, T. Mary-Huard, et S. Robin. Two-dimensional segmentation for analyzing hic data. Bioinformatics, 30(17) :386–392, 2014.

E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science, 326(5950) :289–293, 2009.

E. H. Linfoot. An informational measure of correlation. Information and control, 1(1) :85–89, 1957.

A. Lomet, G. Govaert, et Y. Grandvalet. Design of artificial data tables for co-clustering analysis. Universit de Technologie de Compigne, France, 2012.

Y. Lu, J.-C. Quinton, C. Jolly, et V. Brault. A statistical procedure to assist dysgraphia detection through dynamic modelling of handwriting. 2024.

A. Lung-Yut-Fong, C. Lévy-Leduc, et O. Cappé. Homogeneity and change-point detection tests for multivariate data using rank statistics. Journal de la Société Française de Statistique, 156(4) :133–162, 2015.

M. Mariadassou et C. Matias. Convergence of the groups posterior distribution in latent or stochastic block models. Bernoulli, 21(1) :537–573, 2015.

D. S. Matteson et N. A. James. A nonparametric approach for multiple change point analysis of multivariate data. Journal of the American Statistical Association, 109(505) :334–345, 2014.

G. J. McLachlan, S. X. Lee, et S. I. Rathnayake. Finite mixture models. Annual review of statistics and its application, 6 :355–378, 2019.

N. Meinshausen et P. Bühlmann. Stability selection. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 72(4) :417–473, 2010.

K. L. Mengersen, C. Robert, et M. Titterington. Mixtures : estimation and applications. John Wiley & Sons, 2011.

S. F. Nielsen. The stochastic em algorithm : estimation and asymptotic results. Bernoulli, pages 457–489, 2000.

K. Pearson. Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London. A, 185 :71–110, 1894.

R Core Team. R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

W. M. Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336) :846–850, 1971.

G. Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to k_max change-points. Journal de la Société Française de Statistique, 156(4) :180–205, 2015.

V. Robert. bikm1 : Co-Clustering Adjusted Rand Index and Bikm1 Procedure for Contingency and Binary Data-Sets, 2021. URL https://CRAN.R-project.org/package=bikm1. R package version 1.1.0.

V. Robert, Y. Vasseur, et V. Brault. Comparing high-dimensional partitions with the co-clustering adjusted rand index. Journal of Classification, 38(1) :158–186, 2021.

N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, et M. W. Feldman. Genetic structure of human populations. Science, 298(5602) :2381–2385, 2002.

A. Rosier, E. Lebarbier, V. Brault, et V. Stoppin-Mellet. Classification contrainte de signaux, application à l'étude de la protéine neuronale tau. Dans 54ème journées de statistique de la SFdS, 2023.

A. Sawyer, N. Gooneratne, C. Marcus, D. Ofer, K. Richards, T. Weaver, et al. systematic review of cpap adherence across age groups : clinical and empiric insights for developing cpap adherence interventions. Sleep Med Rev., 15 :343–56, 2011.

G. Schwarz et al. Estimating the dimension of a model. The annals of statistics, 6(2) :461–464, 1978.

M. Stephens. Dealing with label switching in mixture models. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 62(4) :795–809, 2000.

R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B : Statistical Methodology, 58(1) :267–288, 1996.

R. J. Tibshirani. The solution path of the generalized lasso. Stanford University, 2011.

C. Truong, L. Oudre, et N. Vayatis. Selective review of offline change point detection methods. Signal Processing, 167 :107299, 2020.

J.-P. Vert et K. Bleakley. Fast detection of multiple change-points shared by many signals using group lars. Dans Advances in neural information processing systems, pages 2343–2351, 2010.

S. Wang, A. Saillet, P. Le Gall, A. Lacroux, C. Martin-Lacroux, et V. Brault. Study of the influence of a biased database on the prediction of standard algorithms for selecting the best candidate for an interview. En cours de soumission, 2025.

M. J. Warrens. On the equivalence of cohen's kappa and the hubert-arabie adjusted rand index. Journal of classification, 25 :177–183, 2008.

J. Wyse, N. Friel, et P. Latouche. Inferring structure in bipartite networks using the latent blockmodel and exact icl. Network Science, 5(1) :45–69, 2017.

N. R. Zhang et D. O. Siegmund. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics, 63(1) :22–32, 2007.

Z. Zhang. Variable selection with stepwise and best subset approaches. Annals of translational medicine, 4(7), 2016.

Table des figures

1.1 Représentation des articles publiés (traits pleins), en cours de soumission (traits pointillés longs) et en fin de rédaction (traits pointillés courts) regroupés par thématiques (couleurs) : les rectangles correspondent aux articles reliés par des traits aux ovales représentant leurs co-auteur·trice·s. 8
1.2 Représentations schématiques des notations des données étudiées lorsque les cases $Y_{i,j}$ correspondent à des scalaires (à gauche) ou à des vecteurs (à droite). 16
1.3 Exemple d'une matrice binaire de taille 16 par 14 (a) avec un regroupement des lignes suivant un modèle de mélange (b) et une segmentation des colonnes (c). 17
1.4 Exemple d'application du modèle des blocs latents (à gauche) et de la bisegmentation (à droite) sur la matrice de la figure 1.3 (a). 18
1.5 Exemple d'application d'un couplage entre un modèle de mélange sur les lignes avec une segmentation des colonnes sur la matrice de la figure 1.3 (a). 18

2.1 Explication schématique du principe des tests groupés avec une classe 1 avec uniquement des personnes saines (donc résultat négatif pour le groupe) et une classe 2 possédant une personne contaminée (donc résultat positif pour le groupe). Crédit image : Jean-François Rupprecht, Centre de physique théorique (CNRS/Aix-Marseille Université/Université de Toulon). 20
2.2 Représentation sous forme d'histogramme d'une simulation mimant les données de l'article Jones et al. (2020) sur le nombre de cycles pour une populations de 3 712 patients infectés par le SARS-CoV-2 avec une estimation par un mélange de trois gaussiennes aux moyennes et variances libres : les traits pleins orange, vert et rouge correspondent aux densités des groupes et le trait plein bleu à la densité du modèle de mélange. Le trait vertical violet en pointillés correspond à ce que nous pensons être d_{cens} pour la plupart des machines. . . . 21
2.3 Représentation de la densité d'une loi gaussienne partiellement censurée $\mathcal{CN}_{1.5}(0, 1, 1/2)$ (en rouge) et de celle d'une loi gaussienne centrée réduite $\mathcal{N}(0, 1)$ (en bleue transparente). La censure $d_{\text{cens}} = 1.5$ est symbolisée par le trait vertical en pointillés violets. 22
2.4 Représentation des estimations par un modèle de mélange de lois gaussiennes avec censure partielle (ligne du haut) et totale (ligne du bas) suivant deux valeurs de d_{cens} (colonnes). Pour chaque graphique, nous avons représenté le nombre de cycles simulés sous forme d'histogramme en enlevant les barres à droite de la censure (trait vertical en pointillés violets) pour les censures totales ; les traits pleins orange, vert et rouge correspondent aux densités des groupes et le trait plein bleu à la densité du modèle de mélange. Les traits en pointillés bleus et rouges correspondent aux densités non tronquées. 23
2.5 Représentation schématique des notations utilisées pour le modèles des blocs latents : sur la gauche sont représentées les notations liées aux données (en noir) et aux clusters (en bleu) tandis que, sur la figure de droite, nous avons représenté les notations liées aux paramètres $\boldsymbol{\Psi}$. 25
2.6 Représentation graphique du modèle des blocs latents paramétrique : les carrés correspondent à des paramètres et les ronds à des variables. 25
2.7 Représentation graphique du modèle des blocs latents paramétriques avec classe de bruit : les carrés correspondent à des paramètres et les ronds à des variables. En rouge se trouvent les ajouts par rapport à la version du *LBM* présentée dans la figure 2.6 26
2.8 Représentation des matrices obtenues par le modèle *NFLBM* (à gauche) avec 16 classes en lignes et 49 en colonnes (plus une classe de bruit) et le modèle *LBM* (à droite) avec 3 classes en lignes et 78 en colonnes. Pour le *NFLBM*, un trait rouge vertical a été tracé pour différencier la classe de bruit (à gauche) de la partie relative au *LBM* (à droite). . . 26

2.9 Évolution de *CARI* (en violet), *1-CE* (en cyan), *NCE* (en vert) et *ENMI* divisée par 2, qui est égale à *coNMI* (en saumon) dans la configuration (2.9), en fonction de la proportion de cases bien classées pour la configuration (2.9) et pour $(n, J) = (100000, 100000)$. 31
2.10 En haut à gauche : matrice binaire initiale. En haut à droite : moyennes des cases $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$ valant 1 par colonnes. Milieu à gauche : l'organisation sous forme d'histogramme des moyennes $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$ pour mettre en évidence les clusters. Au milieu à droite : le vecteur réorganisé par ordre croissant des moyennes $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$. En bas à gauche : le calcul des différences entre deux valeurs successive des moyennes $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$ ordonnées par ordre croissant. En bas à droite : la réorganisation de la matrice après avoir choisi un seul judicieux pour les moyennes $\left(\overline{y}_{\cdot,j}\right)_{1\leq j\leq J}$ des colonnes et refait la même opération pour les lignes. 34

3.1 Données *Hi-C* du chromosome 17 de cellules souches embryonnaires avec deux segmentations de blocs diagonaux (en bleu et en rouge). Image provenant de l'article Lévy-Leduc et al. (2014) qui fut à la base de mon post-doc. 37
3.2 Représentation schématique des notations utilisées pour le modèle (3.2) avec $n = 16$ lignes et $K^\star = 4$ blocs. 39
3.3 Représentation graphique du modèle des blocs diagonaux : les carrés correspondent à des paramètres et les ronds à des variables. 39
3.4 Représentation schématique des notations du modèle (3.4) avec $n = 16$ lignes ayant $K^\star = 3$ ruptures et $J = 16$ colonnes avec $L^\star = 4$ ruptures. Chaque rectangle est caractérisé par une loi $\mathcal{L}^\star_{k,\ell}$ paramétrique ayant $\boldsymbol{\theta}^\star_{k,\ell}$ comme paramètre. 42
3.5 Représentation graphique du modèle de bisegmentation gaussien : les carrés correspondent à des paramètres et les ronds à des variables. 42
3.6 Représentation de la matrice $\boldsymbol{B}$ associée à la matrice $\boldsymbol{U}$ de l'exemple de la figure 3.4 par la transformation proposée dans l'équation (3.6). 43
3.7 Représentation en gris transparent de plusieurs simulations du nombre de valeurs actives associées à chaque ligne obtenues par la procédure adaptée de *stability selection* pour trois scénarios du plus simple à gauche ou plus compliqué à droite. Les points rouges correspondent aux valeurs médianes et les emplacements des vraies ruptures sont les multiples de 100. Image extraite de l'article Brault et al. (2017b). 44
3.8 Représentation schématique du modèle (3.10) avec les lois pour chaque ligne avant et après la rupture (à gauche) et les lois induisent par la symétrie (à droite) comme vues dans le modèle (3.11). Les zones grises sont mises pour rappelées que la matrice est symétrique. . 47
3.9 Représentation graphique du modèle de bisegmentation non paramétrique : les carrés correspondent à des paramètres et les ronds à des variables. 48
3.10 Boxplots des distances quadratiques entre les emplacements des vraies ruptures $\boldsymbol{T}^\star$ équirépartis et les estimations $\widehat{\boldsymbol{T}}_n$ faites par la procédure MuChPoint (en gris) et la méthode ecp (en blanc) dans une configuration en échiquier alternant $\mathcal{L}_1$ et $\mathcal{L}_2$ suivant différentes tailles n de matrices (en abscisse). Les lois gaussiennes $\mathcal{L}_1 = \mathcal{N}(1, \sigma^2)$ et $\mathcal{L}_2 = \mathcal{N}(0, \sigma^2)$ sont à gauche, les lois de Cauchy $\mathcal{L}_1 = \mathcal{C}au(1, a)$ et $\mathcal{L}_2 = \mathcal{C}au(0, a)$ au milieu et les lois exponentielles $\mathcal{L}_1 = \mathcal{E}xp(2)$, $\mathcal{L}_2 = \mathcal{E}xp(\lambda)$ sont à droite pour différentes valeurs de σ, λ and a. Figures issues de l'article Brault et al. (2018c). 51
3.11 Représentation de la sortie proposée par le package `Capushe` pour l'estimation du nombre de ruptures dans le cadre du modèle `MuChPoint` sur les données *Hi-C* de souris étudiées par Dixon et al. (2012). Figures issues de l'article Brault et al. (2018c). 52
3.12 Représentation de la matrice de similarité des images prises deux par deux dans un film d'un véhicule autonome : plus une case est rouge, plus les deux images mises en lignes et en colonnes sont semblables. Le rectangle vert a été ajouté pour mettre en évidence le motif correspondant à deux tours de rond point et le rectangle violet pour le moment où la voiture a refait le même trajet dans le même sens. 52
3.13 Représentation des 50 ruptures obtenues sur la matrice présentée dans la figure 3.12 par l'algorithme `MuChPoint` avec la matrice originale à gauche et la matrice résumée à droite. 53
3.14 Estimation des deux instants de ruptures pour chaque plant de Colza du premier container (les emplacements vides correspondent à des plants où nous n'avons pas les informations) suivant différentes méthodes : le maximum de vraisemblance (en noir), la méthode du minimum et du maximum (en vert) et la méthode des deux maxima (en rouge) 56

4.1 Représentation graphique du modèle de mélange de segmentation : les carrés correspondent à des paramètres et les ronds à des variables. 57
4.2 Exemple illustrant l'article Brault et al. (2024a). La ligne du haut représente les données fonctionnelles initiales constituées de $n = 100$ individus (courbes) observés toutes les demi-heures pendant 50 jours. Les lignes suivantes permettent de visualiser la décomposition de la population en clusters (ici $K^\star = 3$ clusters - rouge, jaune, violet), à l'intérieur de chaque cluster la segmentation obtenue sur la dimension temporelle, la projection sur une base d'ondelettes à plusieurs niveaux $h \in \{1, 2, H = 3\}$ (dans l'article donc sur la figure, h est remplacé par r). 61
4.3 Distribution des observations entre les groupes (lignes horizontales en pointillés) et localisation des moments de rupture au sein de chaque groupe (lignes verticales en pointillés long - rouge pour le cluster 1, jaune pour le 2, violet pour le 3). Les pointillés fins verticaux correspondent aux semaines de début et de fin des périodes de confinement. Figure issue de l'article Brault et al. (2024a). 62
4.4 Exemples de deux profils théoriques avec un saut (à gauche) et deux sauts distincts (à droite). Image issue de l'acte de conférence de Rosier et al. (2023). 63
4.5 Représentation graphique du modèle de mélange de courbes avec sauts : les carrés correspondent à des paramètres et les ronds à des variables. 64
4.6 Représentation des courbes moyennes d'un individu suivant son appartenance aux cluster : aucun saut dans le cluster 1 (en haut à gauche), un saut simple dans le cluster 2 (en haut à droite), deux sauts simples dans le cluster 3 (en bas à gauche) et un saut double daus le cluster 4 (en bas à droite). Images provenant de Rosier et al. (2023). 65
4.7 Boxplot des proportions des estimations des clusters d'appartenances estimés (couleur) en fonction du cluster d'origine (abscisse), de la valeur de $\delta^\star$ (colonnes) et de la longueur du plateau central pour le cluster 3 (ligne). Images provenant de Rosier et al. (2023). 66
4.8 Exemples de figures à reproduire et de circuits à suivre donnés aux enfants. Images provenant de l'article Devillaine et al. (2021). 67
4.9 À gauche et de haut en bas, exemples de fonctions a_x, ω_x et ϕ_x et les version en y au milieu ainsi que la fonction vitesse $\dot{x}$ (resp. $\dot{y}$) obtenue par l'équation (4.5) tout en bas de chaque colonne. À droite, nous avons représenté la lettre o formée à partir des deux fonctions vitesses ; les *plus* bleus $+$ correspondent aux moment d'annulation de la vitesse en x et les *fois* rouges $\times$ aux moments d'annulation de la vitesse en y. 68
4.10 Représentation en haut (*os*) d'une estimation d'une vitesse des abscisses x (à gauche) et ordonnées y (à droite) suivant la formule (4.7) d'une lettre a avec, en bas (*logical*), les instants où la vitesse empirique est nulle (`logical=TRUE`) ou non. Image issue de l'article Lu et al. (2024). 69
4.11 Représentation de la courbe de la figure 4.10 après l'application de l'opérateur de fermeture w valant 3 (à gauche) et 35 (à droite). Image issue de l'article Lu et al. (2024). 70
4.12 Représentation sous forme de *heat map* de la distribution du nombre $n_{x,i,\text{"}a\text{"}}(w)$ d'intervalles d'annulation de la vitesse obtenus en fonction de l'âge pour des opérateurs w. Les traits correspondent aux quartiles en fonction de l'âge et les points rouges (•) aux médianes des nombres $n_{x,i,\text{"}a\text{"}}$ pour chaque enfant dysgraphique. Image issue de l'article Lu et al. (2024). 70
4.13 Représentation schématique de la procédure utilisant deux algorithmes étant donnés une distance `Dist` et un opérateur maximal $w_{\max}$. `algorithme 1` s'applique pour chaque symbole S sur les données présentées dans la table 4.2 et est optimisé pour obtenir les paramètres $\widehat{\boldsymbol{\theta}}_S^{(1)}$ associés pour chaque symbole S qui permettent de prédire la probabilité $\widehat{p}_i^{(S)}$ que le symbole S écrit par l'enfant i provient d'un enfant dysgraphique ou non. L'`algorithme 2` s'applique sur les probabilités $\widehat{p}_i^{(S)}$ et est optimisé pour obtenir les paramètres $\widehat{\boldsymbol{\theta}}^{(2)}$ afin de prédire si un enfant est dysgraphique ou non. Enfin, les prédictions sur l'échantillon de test sont comparées aux valeurs connues afin de déterminer l'*AUC* c'est-à-dire l'aire sous la courbe *ROC*. 71
4.14 Représentation graphique du modèle de rupture *POMH* avec segmentation : les carrés correspondent à des paramètres et les ronds à des variables. 73

4.15 Estimation des emplacements des ruptures (traits pointillées violets verticaux) pour des données simulées suivant le modèle (4.8) pour un nombre de ruptures maximisant la vraisemblance comme démontré dans le théorème 17. La courbe en saumon correspond aux paramètres de la simulation, la courbe bleue transparente aux observations et la courbe en pointillés verts aux estimations basées sur les emplacements de ruptures. 75
4.16 Courbes d'utilisation de la machine en heures pour chacune des 91 nuits pour 6 patients de notre base de données. 76
4.17 Représentation graphique du modèle de mélange de chaînes de Markov : les carrés correspondent à des paramètres et les ronds à des variables. 76

A.1 Représentation des tailles des personnes âgées ayant répondu à l'enquête sur leurs comportements alimentaires sous forme d'histogramme avec la population entière (à gauche) et subdivisée en fonction de leur genre (à droite). Les traits pleins correspondent à une estimation de la densité et les pointillés à la densité d'une loi gaussienne basée sur les moyennes et écart-types des populations correspondantes. 83
A.2 Exemple de représentations d'une partition de 9 individus dans 3 classes sous forme binaire (à gauche) et vectorielle (à droite). 84
A.3 Représentation graphique du modèle de mélange générique : les carrés correspondent à des paramètres et les ronds à des variables. 85
A.4 Exemple proposé dans Brault (2014) de deux tireurs qui ont chacun visé une cible (gauche et droite). À gauche, nous représentons les impacts de balles sans savoir quels impacts correspondent à quel tireur et à droite, nous représentons une estimation des paramètres et des appartenances : chaque impact est colorié suivant l'appartenance à un tireur (bleu à gauche et or à droite), les points plus foncés correspondent aux moyennes estimées et les cercles à région de confiance à 95% des impacts. Les croix cerclées correspondent aux points jugés à tort d'appartenir au tireur bleu. 86
A.5 Représentation graphique du modèle de rupture générique : les carrés correspondent à des paramètres et les ronds à des variables. 90
A.6 Représentation d'une simulation d'un modèle de segmentation où les individus d'un même segment sont issus d'une loi $\mathcal{N}(\mu_k; \sigma^2)$. Les traits horizontaux en pointillés correspondent aux moyennes des segments, ceux verticaux aux changements entre deux segments, une couleur et un symbole différents ont été associés à chaque segment pour représenter les points et les traits pleins noirs correspondent à des liaisons entre points appartenant à des segments différents. 91
A.7 Représentation schématique des maximisations suivant le nombre de ruptures : principe global (a), introduction de la fonction $I_{K'}(n')$ (b), cas avec zéro rupture (c), cas avec une rupture (d) et principe de récurrence avec deux ruptures (e). 92

B.1 Exemple de matrices avec des données binaires à classifier (en haut à gauche) avec une réorganisation possible (en haut à droite). En bas à gauche, nous avons un exemple possible de *biclustering* avec trois blocs dont deux qui se chevauchent. Au centre en bas, nous avons une recherche de blocs par une procédure hiérarchique mettant en évidence six blocs. En bas à droite, un exemple de co-classification mettant en avant une structure en quadrillage. Exemple adapté de celui proposé dans l'article de Govaert et Nadif (2008). 98
B.2 Représentation schématique des notations utilisées pour le modèles des blocs latents : sur la gauche sont représentées les notations liées aux données (en noir) et aux clusters (en bleu) tandis que, sur la figure de droite, nous avons représenté les notations liées aux paramètres $\boldsymbol{\Psi}$. 100
B.3 Représentation graphique du modèle des blocs latents paramétrique : les carrés correspondent à des paramètres et les ronds à des variables. 100
B.4 Représentation de la matrice de l'exemple de la figure B.1 (en haut) avec la représentation sous forme de graphe biparti (en bas). À gauche, la réorganisation suivant un modèle de co-clustering de la matrice avec des couleurs pour les blocs et la représentation en graphe biparti avec les liens entre les noeuds. Au centre, la version simplifiée en coloriant plus ou moins noir suivant l'intensité entre les blocs. À droite, la version discrétisée où seules les probabilités moyennes plus grandes que 0.5 sont conservées. 102
B.5 Représentation graphique du modèle des blocs latents paramétrique bayésien : les carrés correspondent à des hyperparamètres et les ronds à des variables. En rouge se trouvent les ajouts par rapport à la version fréquentiste du *LBM* présentée dans la figure B.3 104

Liste des tableaux

2.1 Évaluation du comportement des critères par rapport aux cinq propriétés souhaitées. ✓ signifie que le critère est validé, tandis qu'un symbole ✗ signifie que le critère n'est pas rempli (ou qu'il est le moins efficace par rapport aux autres critères). 30

4.1 Tableau de contingence entre les clusters et les profils *Enedis*. *ENT/PRO* correspond à des contrats de professionnels et *RES* à des contrats de particuliers (les chiffres des contrats correspondent à différents profils établis par *Enedis*). 62

4.2 Représentation schématique des informations recueillies par enfant (en ligne) : pour chaque symbole S (groupe de colonnes), nous conservons l'opérateur $\widehat{w}$, la distance de la reconstruction et le temps mis pour écrire le symbole. A ceci, nous ajoutons la section (nous avons regroupé les classes en ordinale) et l'information de si l'enfant est dysgraphique ou non. 72

4.3 Résumé des combinaisons testées dans l'article Lu et al. (2024) suivant la procédure présentée dans la figure 4.13 en fonction de l'`algorithme 1` (en ligne) et de l'`algorithme 2` (en colonnes). 72

Index

Absolu
  Norme absolue, 95
Adjacence
  Matrice, 101
Algorithme
  CEM, 88
  EM stochastique, 104
  EM variationnel, 89, 103
  EM, 58, 87
  EP, 89
  SEM, 88
  V-Bayes, 104
  de programmation dynamique, 58, 92
  Largest Gaps, 33
Amplitude, 68
Approximation
  en champs moyen, 103
ARI
  Adjusted Rand Index, 28

Bayesien
  Algorithme *V-Bayes*, 104
Bayésien
  Bayesian Information Criterion, 21, 61, 89, 105
Biclustering, 97
  hiérarchique, 97
Biparti
  Graphe, 101
Bisegmentation, 10, 16
  Modèle de bisegmentation par blocs diagonaux, 10
Blanchissement, 54
Bloc
  Modèle blocs diagonaux, 9
  Modèle de bisegmentation par blocs diagonaux, 10
  Modèle des blocs latents, 7, 15, 16, 24, 28, 97, 99
  Modèle des blocs latents avec classe de bruit, 9, 25
  Modèle des blocs stochastiques, 7
  Modèle à blocs stochastiques, 101
Brave
  Handwriting Kinder (*BHK*), 66
Bruit
  Classe de bruit, 24
  Modèle des blocs latents avec classe de bruit, 9, 25

Censure
  Modèle de mélange de gaussiennes censurées, 10
Chaîne
  de Markov, 13
  Mélange de chaîne de Markov, 75
Champs
  Approximation en champs moyen, 103
Chevauchement, 97
Classe, 84
  de bruit, 24
  latente, 84
  Modèle des blocs latents avec classe de bruit, 9
  vide, 87
Classification
  Algorithme *CEM*, 88
  croisée, 98
  Error (CE), 28, 29
  non supervisée, 84
  Normlized Classification Error (NCE), 29
  semi supervisée, 84
  supervisée, 11, 84
Cluster, 84
Coclassification, 98
Coclustering, 98
  Normalized Mutual Information (coNMI), 29
Complete
  Integrated Complete Likelihood, 89, 105
Contingence
  Tableau, 28
Courbe
  ROC, 72
Critère
  Bayesian Information Criterion, 21, 61, 89, 105
  Adjusted Rand Index (ARI), 28
  Classification Error (CE), 28, 29
  d'information mutuelle (MI), 29
  d'information mutuelle normalisée (NMI), 29
  Normalized Classification Error (NCE), 29
Croisement
  Classification croisée, 98

Déphasage, 68
Diagonal
  Modèle blocs diagonaux, 9
  Modèle de bisegmentation par blocs diagonaux, 10
Dirac
  Loi, 76

Divergence
de Kullback-Leibler, 29, 103
Dynamique
Algorithme de programmation dynamique, 58, 92
Dysgraphie, 11

Échantillonneur
de Gibbs, 89, 104
Energie
libre, 27, 103
Erreur
Classification Error (CE), 28, 29
Normalized Classification Error (NCE), 29
Espérance
Algorithme *EM*, 58, 87
Euclide
Norme euclidienne, 95
Expectation
Algorithme *CEM*, 88
Algorithme *EM stochastique*, 104
Algorithme *EM variationnel*, 89, 103
Algorithme *EP*, 89
Algorithme *SEM*, 88
Extended
Normalized Mutual Information (ENMI), 29

Fréquence, 68

Gaps
Largest Gaps, 33
Gaussien
Modèle de mélange de gaussiennes censurées, 10
Sous-gaussien, 40, 46, 73
Gibbs
Échantillonneur, 89, 104
Graphe
biparti, 101

Handwriting
Brave Handwriting Kinder (*BHK*), 66
Parsimonious Oscillatory Model of Handwriting (*POMH*), 67, 68
Hiérarchie, 97

Identifiabilité, 91, 100
conditions suffisantes, 100
d'un modèle de mélange, 87
Information
Bayesian Information Criterion, 21, 61, 89, 105
Coclustering Normalized Mutual Information (coNMI), 29
Extended Normalized Mutual Information (ENMI), 29
mutuelle (MI), 29
mutuelle normalisée (NMI), 29
Integrated
Complete Likelihood, 89, 105

Kinder
Brave Handwriting Kinder (*BHK*), 66
Kullback
Divergence de Kullback-Leibler, 29, 103

Label
switching, 28, 86, 87, 89, 91
Largest
Gaps, 33
Latent
Classes latentes, 84
Modèle des blocs latents, 7, 15, 16, 24, 28, 97, 99
Modèle des blocs latents avec classe de bruit, 9, 25
Leibler
Divergence de Kullback-Leibler, 29, 103
Libre
Energie, 27, 103
Likelihood
Integrated Complete Likelihood, 89, 105
Linéaire
Modèle linéaire sparse, 94
Logvraisemblance
Modèle de mélange, 85
Modèle de segmentation, 91
Loi
de Dirac, 76
paramétrique, 15, 85

Markov
Chaîne de Markov, 13
Mélange de chaîne de Markov, 75
Matrice
binaire d'appartenance des individus aux classes, 84
d'ajacence, 101
de transitions, 76
Maximisation
Algorithme *CEM*, 88
Algorithme *EM stochastique*, 104
Algorithme *EM variationnel*, 89, 103
Algorithme *EM*, 58, 87
Algorithme *SEM*, 88
Mélange
de chaîne de Markov, 75
de segmentation, 12
Modèle, 9, 28, 84
Modèle de mélange de gaussiennes censurées, 10
Modèle de mélange de segmentations, 57
Modèle de segmentation, 15
Modèle
blocs diagonaux, 9
de bisegmentation par blocs diagonaux, 10
de mélange, 9, 15, 28, 84
de mélange de gaussiennes censurées, 10
de mélange de segmentations, 57
de ruptures *offline*, 91

de ruptures *online*, 91
de segmentation, 12, 15, 90
des blocs latents, 7, 15, 16, 24, 28, 97, 99
des blocs latents avec classe de bruit, 9, 25
des blocs stochastiques, 7
linéaire sparse, 94
Parsimonious Oscillatory Model of Handwriting (*POMH*), 67, 68
à blocs stochastiques, 101
Moyen
Approximation en champs moyen, 103
Mutual
Coclustering Normalized Mutual Information (coNMI), 29
Extended Normalized Mutual Information, 29
Mutuel
Information (MI), 29
Information mutuelle normalisée (NMI), 29

Normalized
Coclustering Normalized Mutual Information (coNMI), 29
Extended Normalized Mutual Information, 29
Normalized Classification Error (NCE), 29
Norme
absolue, 95
euclidienne, 95

Offline
Modèle de ruptures *offline*, 91
Online
Modèle de ruptures *online*, 91
Optimal
Transport, 26
Oscillatory
Parsimonious Oscillatory Model of Handwriting (*POMH*), 67, 68

Paramétrique
Loi, 15, 85
Parsimonious
Oscillatory Model of Handwriting (*POMH*), 67, 68
Programmation
Algorithme de programmation dynamique, 58, 92
Propagation
Algorithme *EP*, 89

robustesse, 27
Rupture, 90
Modèle de ruptures *offline*, 91
Modèle de ruptures *online*, 91

Segmentation
Modèle, 90
Modèle de mélange de segmentations, 57
Modèle de segmentation, 12, 15
Mélange, 12
Sélection
Stabilité de la sélection, 43
Sous
gaussien, 40, 46, 73
Sparse
Modèle linéaire sparse, 94
Stable
Stabilité de la sélection, 43
Statistique
U statistique, 47
Stocastique
Modèle des blocs stochastiques, 7
Stochastique
Algorithme *EP*, 89
Algorithme *SEM*, 88
Modèle à blocs stochastiques, 101
Supervisé
Classification supervisée, 11
Switching
Label, 28, 86, 87, 89, 91
Symbole, 69

Tableau
de contingence, 28
Transition
Matrice de transitions, 76
Transport
optimal, 26

Variationnel
Algorithme *EM stochastique*, 104
Algorithme *EM variationnel*, 89, 103
Algorithme *V-Bayes*, 104
Vide
Classe, 87
Vraisemblance
Modèle de mélange, 85
Modèle de segmentation, 91

Acronymes
AIC : *Akaike Information Criterion*, 72
ARI : *Adjusted Rand Index*, 28
BHK : *Brave Handwriting Kinder*, 11, 66, 67
BIC : *Bayesian Information Criterion*, 21, 61, 89, 105
CARI : Coclustering Adjusted Rand Index, 28
CEM : Classification Espérance Maximisation ou *Classification Expectation Maximisation*, 88
CE : *Classification Error*, 28, 29
coNMI : *Coclustering Normalized Mutual Information*, 29
CUEF : Centre Universitaire d'Etudes Françaises, 14
DUT : Diplôme Universitaire Technologique, 13
EM : Espérance Maximisation ou *Expectation Maximisation*, 58, 87

ENMI : *Extended Normalized Mutual Information*, 29
ENSIMAG : École Nationale Supérieure d'Informatique et de Mathématiques Appliquées de Grenoble, 14
EPFL : École Polytechnique Fédérale de Lausanne, 11
EP : *Expectation Propagation*, 89
FLSA : *Fused LASSO Signal Approximator*, 41
HCSP : Haut Conseil de la Santé Publique, 10, 20
R : Intelligence Artificielle, Biais et Acceptabilité dans le Recrutement, 15
IA : Intelligence Artificielle, 12
ICL : *Integrated Complete Likelihood*, 89, 105
INP : Institut National Polytechnique, 13
ISAE-SUPAERO : Institut Supérieur de l'Aéronautique et de l'Espace, 13
IUT : Institut Universitaire Technologique, 13
LASSO : *Least Absolute Shrinkage and Selection Operator*, 11, 91, 92, 94
LBM : Modèle des Blocs Latents, 7
LG : algorithme *Largest Gaps*, 33
LIDILEM : LInguistique et DIdactique des Langues Étrangères et Maternelles, 14
LIG : Laboratoire d'Informatique de Grenoble, 9
MIAI : *Multidisciplinary Institute in Artificial intelligence*, 15
MIASHS : Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales, 13
MI : *Mutual Information*, 29
MSE^3 : Modélisation Statistique pour l'Etude de l'Ecriture chez l'Enfant, 13
MSIAM : *Master of Science in industrial and applied mathematics*, 14
NCE : *Normalized Classification Error*, 29
NFLBM : *Noise-free latent block model* ou modèle des blocs latents avec classe de bruit, 25
NFLBM : Modèle des Blocs Latents avec Classe de Bruit, 9
NMI : *Normalised Mutual Information*, 29
Phelma : PHysique, ÉLectronique, MAtériaux, 13
POMH : *Parsimonious Oscillatory Model of Handwriting*, 12, 67, 68
RI : *Rand Index*, 28
ROC : *Receiver Operating Characteristic*, 12, 72
SAS : Syndrome de l'Apnée du Sommeil, 15, 74
SBM : Modèle des Blocs Stochastiques, 7
SELF : Système d'Évaluation en Langues à visée Formative, 14
SEM : Espérance Maximisation stochastique ou *Stochastic Expectation Maximisation*, 88, 104
SSD : Statistique et Science des données, 13
STID : Statistique et Informatique Décisionnelle, 13
SVM : *Support Vector Machine*, 72
TransCrit : Transferts critiques anglophones, 14
V-Bayes : algorithme *V-Bayes*, 104
VEM : Espérance Maximisation Variationnel ou *Variationnal Expectation Maximisation*, 103
WIC : Web, Informatique et Connaissances, 15

Notations

$(z_{i,k})$: matrice binaire d'appartenance des individus aux classes, 84
$F_{k,\ell}$: fonction de répartition associée à la loi $\mathcal{L}_{k,\ell}$, 49
H : dimension des données dans la case $Y_{i,j}$, 15
$I_{K'}(n')$: valeur maximale de la fonction à maximiser dans un algorithme de programmation dynamique en ayant pris uniquement K' ruptures et n' observations, 93
K : nombre de classes ou clusters, 84
L : nombre de classes ou clusters en colonnes, 24, 99
$R_j^{(i)}$: rang de la case $Y_{i,j}$ au sein de la ligne i, 48
Δ : fonction à deux variables utilisée pour la maximisation dans l'algorithme de programmation dynamique, 92
$\Delta_{\boldsymbol{\tau}^\star}$: distance minimale entre deux estimations ruptures normalisés $\boldsymbol{\tau}^\star$ ayant servi à simuler les données, 40, 45, 49
$\Delta_{\boldsymbol{\theta}^\star}$: distance minimale permettant de différencier deux colonnes (resp. deux lignes) successives., 45
Δ_n : distance minimale entre deux estimations ruptures normalisés $\widehat{\boldsymbol{\tau}}_n$, 39
$\|\cdot\|_{\text{Haus}}$: norme de Hausdorff, 40, 74
$\mathbb{I}_n$: matrice identité de taille n, 85
J : nombre de variables (si identique), 15
$\widehat{K}_n^{\Delta_n}$: estimateur du nombre de ruptures en admettant qu'il y ait une distance minimale de Δ_n entre deux ruptures, 39
$\overline{R}_k^{(i)}$: rang moyen des cases $(Y_{i,j})_{T_k+1\leq j\leq T_{k+1}}$, 48
$S_n(\cdot)$: statistique utilisée pour estimer les emplacements des ruptures dans le modèle `MuChPoint`, 47, 48
T_k : rupture dans le modèle de segmentation, 90
$\mathbb{T}_n^{(1)}$: matrice triangulaire inférieure de taille $n \times n$ ne contenant que des 1, 94
$\mathbb{T}_n^{(\boldsymbol{t})}$: matrice des différences des temps thermiques, 54
$U_{n,i}(T_1^\star)$: U statistique utilisée pour le modèle `MuChPoint`, 47
$Y_{i,j}$: variable aléatoire correspondante à l'in-

dividu i et à la variable j, 15
$Y_{i,j,h}$: variable aléatoire de la coordonnée h du vecteur de la variable j pour l'individu i, 15
$\boldsymbol{\Psi}$: ensemble des paramètres du modèle, 15
$\boldsymbol{\Theta}$: ensemble des paramètres $\boldsymbol{\theta}$, 85, 91
$\boldsymbol{Y}$: ensemble de données, 15
$\boldsymbol{\pi}$: paramètre d'appartenance aux classes, 84
$\boldsymbol{\tau}$: emplacement des ruptures $\boldsymbol{T}$ normalisés par $T^{\star}_{K^{\star}+1} = n$, 39
$\widehat{\boldsymbol{\tau}}^{\Delta_n}_{\widehat{K}^{\Delta_n}_n}$: estimateur des ruptures normalisées par n en admettant qu'il y ait une distance minimale de Δ_n entre deux ruptures, 39
$\boldsymbol{\theta}$: ensemble des paramètres d'une loi paramétrique, 15, 85
$\widehat{\boldsymbol{\theta}}^{\Delta_n}_{\widehat{K}^{\Delta_n}_n}$: estimateur des paramètres dans un modèle de ruptures en admettant qu'il y ait une distance minimale de Δ_n entre deux ruptures, 39
$\boldsymbol{\theta}_k$: paramètres du cluster k ou du segment D_k dans le cas d'une loi paramétrique, 85, 90
$\boldsymbol{\theta}_{k,\ell}$: paramètres du bloc (k, ℓ) dans le cas de familles paramétriques, 24, 99
$\boldsymbol{w}$: partition des colonnes d'une matrice, 24, 99
$\boldsymbol{z}$: partition des individus dans les classes, 24, 84, 99
$\widehat{\boldsymbol{z}}$: estimation de la matrice $\boldsymbol{z}$, 88
$\mathbf{Cont}^{(\boldsymbol{z},\boldsymbol{z}')}$: tableau de contingence entre les partitions $\boldsymbol{z}$ et $\boldsymbol{z}'$, 28
d_{cens} : seuil de censure (connu) utilisé dans l'article Brault et al. (2021b), 20
$\mathbb{I}_n$: matrice identité de taille $n \times n$, 94
$\mathbb{1}_{\{\cdot\}}$: fonction indicatrice, 94
λ_{CV} : paramètre *LASSO* obtenu par validation croisée, 54
$\mathcal{CN}_{d_{\text{cens}}}(\mu, \sigma, q)$: lois gaussiennes avec censure partielle en d_{cens} de paramètres $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^{\star}_{+}$ et $q \in [0; 1]$, 22
$\mathcal{CN}_{d_{\text{cens}}}(\mu, \sigma)$: lois gaussiennes avec censure totale en d_{cens} de paramètres $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}^{\star}_{+}$, 22
$\mathfrak{S}(\{1, \ldots, K\})$: ensemble des permutations sur l'ensemble $\{1, \ldots, K\}$, 28
$\mathcal{W}$: espace des partitions $\boldsymbol{w}$, 24, 99
$\mathcal{Z}$: espace des partitions $\boldsymbol{z}$, 24, 99
n : nombre d'individus, 15
$\otimes$: produit de Kronecker, 42, 54
π_k : probabilité d'appartenance à la classe k, 84
$s^{(c)}_{i,k}$: probabilité pour l'individu i d'appartenir à la classe k étant données les observations $\boldsymbol{Y}$ et le paramètre $\boldsymbol{\Psi}^{(c)}$, 87
$s^{(c)}_{i,k}$: probabilité pour l'individu i d'appartenir à la classe k étant donnés les observations $\boldsymbol{Y}$ et le paramètre $\boldsymbol{\Psi}^{(c)}$, 58
φ : densité d'une loi paramétrique, 15, 85, 90
$y_{i,j}$: observation de la variable $Y_{i,j}$, 15
$\mathbb{0}_n$: vecteur nul de longueur n, 94