Détection de ruptures multiples dans des séries temporelles multivariées : application à l'inférence de réseaux de dépendance

Flore Harlé

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Signal, Image, Parole, Télécoms (SIPT)**
Arrêté ministériel : 7 août 2006

Présentée par

Flore HARLÉ

Thèse dirigée par **Sophie ACHARD**
co-encadrée par **Cédric GOUY-PAILLER**
et **Florent CHATELAIN**

préparée au sein du **GIPSA-LAB** et du **CEA-LIST**
dans **L'École Doctorale d'Électronique, d'Électrotechnique, d'Automatique et du Traitement du Signal (EEATS)**

Détection de ruptures multiples dans des séries temporelles multivariées : application à l'inférence de réseaux de dépendance

Thèse soutenue publiquement le **21 juin 2016**,
devant le jury composé de :

M. Olivier MICHEL
Professeur, Grenoble INP, Président du jury
M. Vicente ZARZOSO
Professeur, Université Nice Sophia Antipolis, Rapporteur
M. Jean-Yves TOURNERET
Professeur, Université de Toulouse, Rapporteur
M^me^ Chloé-Agathe AZENCOTT
CR, ARMINES/Mines ParisTech, Institut Curie, INSERM, Examinatrice
M^me^ Sophie ACHARD
CR, CNRS - Université de Grenoble, Directrice de thèse
M. Cédric GOUY-PAILLER
Ingénieur chercheur, CEA-LIST, Encadrant de thèse
M. Florent CHATELAIN
Maître de conférence, Grenoble INP, Encadrant de thèse, Invité

À Jonathan

Remerciements

Le moment étant venu d'apposer un point final à ce manuscrit de thèse, je tiens à remercier les nombreuses personnes qui m'ont accompagnée, de près comme de loin, dans cette aventure. J'adresse en premier lieu mes remerciements à mes encadrants de thèse, au CEA à Cédric Gouy-Pailler et au Gipsa-Lab à Florent Chatelain et à Sophie Achard, ma directrice de thèse. Travailler avec vous trois a été un véritable plaisir et je vous suis très reconnaissante pour les bonnes conditions de travail dans lesquelles s'est déroulée cette thèse, malgré la distance géographique (et les obstacles techniques qui ont pu se présenter). Je me suis sentie vraiment soutenue aux cours de ces dernières années, tant dans mon travail de recherche que sur un plan plus personnel. Je vous remercie pour m'avoir offert de travailler avec vous, de m'avoir laissé le temps d'apprendre sur des sujets nouveaux en tirant profit de vos connaissances dans de larges domaines, de m'avoir tant guidée dans les premiers mois puis laissé suivre mes pistes malgré un planning serré de fin de thèse, de m'avoir encouragée et remontée le moral à l'occasion. Merci à Florent pour les coups de fils salvateurs sur les difficultés méthodologiques et les explications très pédagogiques, à Cédric pour les conseils toujours pertinents, les relectures attentives, mes « tu aurais 5 minutes ? » intempestifs des derniers mois, et le recul que tu m'as fait prendre dans les difficultés, et merci à Sophie pour toute l'aide sur la communication, le souci du bon déroulement de mes projets, l'énergie employée sur tant de problèmes, et les grandes conversations sur des sujets très variés mais passionnants.

J'adresse également mes remerciements aux membres de mon jury de thèse, qui ont accepté d'évaluer mon travail de recherche. Je remercie particulièrement M. Vicente Zarzoso et M. Jean-Yves Tourneret, les rapporteurs, pour leur relecture attentive de mon manuscrit, qui m'a donné l'occasion d'approfondir et d'éclairer certains points de mes travaux. Je remercie encore le président du jury, M. Olivier Michel, et Mme Chloé-Agathe Azencott, examinatrice, pour leur remarques pertinentes et les perspectives qu'elles ouvrent pour la poursuite de ces recherches. Il est à la fois regrettable et encourageant de constater que les 230 pages de ce manuscrit ne permettent pas de faire le tour du sujet.

Durant plus de trois ans, j'ai eu la liberté de pouvoir me consacrer entièrement à mon sujet de recherche, le faire évoluer, tester des pistes, les poursuivre ou les abandonner, et aborder des thématiques plus éloignées. Cette expérience riche et qui sera probablement unique dans ma vie professionnelle a été rendue possible grâce aux cadres de travail propices de mes deux labos, au CEA et au Gipsa-Lab. J'adresse mes remerciements à l'ensemble des personnes de la direction qui ont fait en sorte que ma thèse se déroule favorablement,

à commencer par Anthony Larue, chef du laboratoire Ladis au CEA, qui m'a initialement fait découvrir le laboratoire et les thématiques de recherche qui m'ont amenée à me lancer dans la thèse, et son soutien au cours des années passées au CEA. Je remercie également la direction du DM2I, en particulier Laurent Disdier et Mehdi Gmar, ancien et actuel chefs du département, Olivier Gal et Annick Nguyen pour l'encadrement administratif dont les thésards bénéficient. Et que serions-nous sans le secours de Florence Chedaute, Nathalie Feiguel, et toutes les apprenties ! Du côté de Grenoble, je remercie la direction du Gipsa-lab pour avoir accepté l'organisation de cette thèse en collaboration avec le CEA, et qui s'est très bien déroulée. Je remercie notamment Jérôme Mars, Jean-Marc Thiriet, et enfin Lucia Bouffard-Tocat au secrétariat. Toujours à Grenoble, je remercie la direction de l'École Doctorale EEATS, qui s'est montrée arrangeante au vu de mes contraintes géographiques, en particulier Gang Feng pour sa confiance, le secrétariat pour son aide et les directeurs Christian Commault puis Guy Vitrant pour l'accompagnement des doctorants.

Passer toutes ces années au Ladis a été un grand plaisir, et ce en grande partie grâce à mes collègues de travail ! J'ai grandement apprécié l'ambiance qui règne au laboratoire, et qui a contribué à me motiver les jours où la thèse se passait moins bien. Je ne peux pas citer tous les collègues, mais je vous remercie sincèrement pour les bons moments passés au CEA. J'ai énormément appris à vos côtés. Commençons par remercier les participants au café du matin. J'espère que cette cérémonie se maintiendra avec tout le prestige qui lui revient, sous la houlette du Maître du Café, Antoine, et que la qualité du breuvage sera transcendée grâce au protocole de mise en œuvre de la loi du n+1, validée empiriquement par Loïs. Je garderai un excellent souvenir des échanges passionnants et éclectiques qui se sont déroulés autour d'une tasse avec Aurore, Jérôme, Marine, Bertrand, Marie-Gabrielle, Anne-Catherine et tant d'autres. Merci à Jean-Philippe le prophète du `C#` toujours souriant, Natacha, Laurence B., Laurence C., Lorène et Marine pour le réseau d'échange matériel et conseil bébé, Cédric A. et Jérôme pour la revue gastronomique, Étienne pour sa bonne humeur et ses conseils avisés, Davide pour m'avoir supporté dans le bureau (bravissima !) et pour avoir enrichi mon vocabulaire en jurons italiens, Frédéric pour ses petites blagues du matin, Aurore pour son bureau des lamentations, les collègues du LCAE et du LM2S, dont les conversations m'ont manqué à Digiteo, et de manière générale les skieurs, les grimpeurs, les coureurs du relais, les programmeurs invétérés, les dépanneurs de Linux, les littéraires, les critiques de cinéma, les finisseurs de fromage et les pourvoyeurs d'articles introuvables, « avec qui mes rapports furent aussi divers qu'enrichissants ». Une mention spéciale pour les méthodes de management de Rémi, un collègue plein de ressources pour exploiter le potentiel d'un élevage de thésards, et ses tuyaux indispensables et de bon goût pour le déroulement optimal d'une thèse.

Je salue les anciens thésards et post-docs, dont Guillaume L, Guillaume M, Seb (c'est un peu votre faute si j'ai fait une thèse), Rachel (pour les sorties aussi), Maugan (pour les bouffeurs de salade), Quentin, Cécile, Amel, Philippine et Mathieu. Un remerciement spécial aux ex-thésards du début et de la fin, avec qui j'ai partagé les hauts et les bas : Maxime, Yoann, Jérémy et ses siestes au fond de MON bureau, Emira, Lucile, Néhémy (et les raclettes), Paul (bientôt), Jana et les autres. J'espère continuer à vous croiser, malgré les chemins différents que nous emprunterons. Il n'était pas toujours aisé de s'ouvrir aux

problématiques des autres doctorants du labo tant les thématiques sont diverses, mais j'ai grandement apprécié les échanges que nous avons pu avoir à l'occasion avec l'ensemble des thésards du labo, anciens du LOAD, LIMA, et actuels du Ladis. À ces derniers, Paul, Alexis, Maxime, Cyrille, Anne, Olivier, Shivani, Harry et Lamia, j'adresse tous mes encouragements pour les années à venir, en leur souhaitant d'apprécier pleinement cette expérience, et de ne pas perdre courage, il y a une fin à tout. Du côté de Grenoble, je remercie les thésards et les post-docs du Gipsa, principalement du Gipsadoc, que j'ai eu un très grand plaisir à côtoyer, en particulier mes parentes de thèse, Aude et Céline. Un remerciement spécial au personnel de l'hôtellerie Chez Aude et Robin (chambre chez l'habitant très conviviale, petit déjeuner avec vue sur la Chartreuse). Je salue le groupe des anciens, Wei, Mathieu, Cyrille, Cindy, Manu et les autres, j'espère avoir l'occasion de partager un visionnage de La Classe Américaine, en mangeant des chips.

Il s'est passé de nombreux événements depuis le début de ma thèse. Le principal fut l'arrivée de Tiphaine en novembre 2014. Concilier le doctorat et la maternité a été possible grâce à un cadre de travail favorable que je souhaite à tous les parents, à des encadrants compréhensifs et confiants, mais aussi grâce au soutien sans faille de mes proches et de mon compagnon Jonathan. Je ne te remercierai jamais assez pour m'avoir encouragée jusqu'au bout et de t'être tant investi ces derniers mois pour que je puisse terminer ma thèse. Tu as su m'ouvrir la voie du monde de la recherche, et j'aime toujours autant les conversations enflammées que nous avons sur la science. Je remercie très sincèrement ma famille pour leur patience et leur curiosité au sujet de mon doctorat, qui je suppose reste un peu obscur. Je vous expliquerai au repas de Noël. Merci en particulier à Stella et Thierry, mes parents, à Clémence, ma sœur, à François, mon frère, à Truss, ma grand-mère, à Muriel et Thierry, mes beaux-parents, ainsi qu'à Denise, Claude et Serge, pour leurs précieux encouragements, les bons moments, les grandes discussions, et toute l'aide logistique non négligeable pour le marathon des 6 derniers mois. Il me semble toutefois que les heures passées à garder le panda roux ont été appréciées par tous. Un grand merci à ma petite Tiphaine pour les rires qui ont égayé mes journées chargées, pour le stage intensif en gestion de manque de sommeil, qui s'est avéré très utile lors de la rédaction, et pour le grand bonheur de te voir grandir et avancer plus vite que ma thèse. Un remerciement également aux amis qui nous entourent et qui ont accepté mon asociabilité de fin de thèse, et à ceux qui ont eu la charge de distraire et soulager Jonathan lors de la rédaction. Merci en particulier à Anne-Sophie et Thomas (aussi pour ta conception de la Science et du monde en général) et à Geralt qui a tenu compagnie à Jonathan pendant de longues heures.

Pour terminer, je tiens à exprimer à quel point j'ai été touchée par les encouragements puis les félicitations reçues, en particulier celles émanant des femmes de deux générations avant la mienne, qui n'ont pas eu la chance de faire les études qu'elles auraient souhaitées, et qui m'en ont exprimé le regret. Je mesure aujourd'hui ma chance d'avoir pu accomplir ce parcours, grâce d'une part à l'aide de mes parents et d'autre part au travail qui a été mené par d'autres pour que les études soient accessibles à un grand nombre.

Table des matières

Table des figures

Introduction

Contexte

La multiplication de l'instrumentation des équipements et l'augmentation des capacités de stockage des mesures au cours de ces dernières décennies a favorisé le développement d'une grande variété de solutions pour répondre à la problématique de la détection de changements dans des séries temporelles. La littérature regorge d'exemples dans divers domaines comme le contrôle qualité, la maintenance prédictive, la sécurité informatique, la biologie, le traitement de la parole, l'économétrie, l'écologie ou la géophysique. L'analyse rétrospective de séries temporelles acquises dans leur intégralité conduit à la segmentation du signal en portions où le système étudié a un comportement stationnaire. Il peut s'agir par exemple d'identifier les régimes de fonctionnement d'un moteur, ou d'estimer les propriétés physiques des milieux traversés par des ondes sismiques.

Lors de l'analyse d'un système complexe, où des dépendances régissent les relations entre les entités, les interactions entre les composantes mesurées par les capteurs doivent être prises en compte. Le formalisme des graphes permet de représenter efficacement ces systèmes : les nœuds symbolisent les variables étudiées, et les arêtes dirigées symbolisent les dépendances. Lorsque les relations intrinsèques entre les éléments du système ne varient pas au cours du temps, le graphe est statique.

Un réseau de consommation domestique d'eau et de puissance électrique est représenté par un graphe dans la figure 1a et constitue un exemple d'un tel système. Dans cette installation, plusieurs compteurs mesurent la puissance électrique instantanée consommée par certains appareils comme le four et le chauffe-eau et dans toute la maison par le compteur général, ainsi que les débits d'eau circulant dans la douche et à l'arrivée d'eau dans l'habitation. Les signaux délivrés sont tracés dans la figure 1b. L'allumage et l'extinction de différents équipements sont visibles sur les relevés de puissance sous la forme de sauts de moyenne dans les mesures issues du compteur général, et dans les signaux des équipements concernés, si ceux-ci sont mesurés. Dans cet exemple, la détection de changements peut se faire en analysant individuellement les signaux de la figure 1b, mais il est plus pertinent de tenir compte des relations qui existent entre les variables observées et d'intégrer la nature du réseau. Ici le réseau électrique est matérialisé physiquement par les câbles, tuyaux, compteurs, et débitmètres. Le traitement statistique des données est facilité par l'intégration de ces informations, par exemple via une modélisation des lois de Kirchhoff. Dans le cas où ce réseau n'est pas connu, en l'absence du plan de l'installation électrique, ce

sont les informations extraites des mesures temporelles qui permettent en partie d'estimer le graphe qui représente les relations de dépendance. Une fois la segmentation des signaux réalisée, les portions des séries temporelles sont analysées pour établir un diagnostic de consommation.

De par ses enjeux applicatifs réels, la problématique de la détection de changements a pris de plus en plus d'importance depuis les années 60. Une catégorie particulière concerne des événements abrupts, appelés ruptures. À chaque situation est associée un problème mathématique particulier, menant à une approche méthodologique donnée. Le choix d'une méthode prend en compte non seulement la nature des données et leurs propriétés, mais également les contraintes propres à l'application, telles que le traitement en ligne ou rétrospectif, la rapidité d'exécution, la complexité, les dimensions du problème, le compromis entre la puissance de détection et le taux de fausses alarmes, la nécessité ou non d'estimer les paramètres du modèle, l'importance de la localisation précise des ruptures, ou encore la robustesse à certaines perturbations dans les mesures, comme le bruit ou la perte de données.

Objectifs

Mes travaux s'inscrivent dans le cadre général de la détection d'événements dans des systèmes complexes, représentés par un graphe de dépendance. Plus précisément, les séries temporelles étudiées sont constituées par des observations mesurées de manière synchronisée sur un ensemble de variables entre lesquelles il peut exister ou non une interaction, qui demeure inchangée dans le temps. Les relations de dépendances qui en résultent sont décrites par un graphe statique. Les variables étudiées peuvent être de natures différentes, et les signaux qui en sont extraits n'ont pas forcément les mêmes caractéristiques telles que les valeurs moyennes et la nature du bruit. En l'absence de changement, les observations sont distribuées autour d'une valeur médiane. Des événements se produisent dans le système, et impactent certaines variables. Ces phénomènes prennent la forme de sauts brusques dans les médianes des signaux affectées, dont l'allure est celle d'une fonction bruitée constante par morceaux, comme dans l'exemple de la figure 1b. L'objectif poursuivi est dans un premier temps de détecter et localiser ces ruptures conjointement dans les signaux, en tenant compte des éventuelles interactions régissant les entités du système, et dans un second temps d'identifier le graphe de dépendance entre les variables.

Afin de pouvoir appliquer la méthode développée sur des séries temporelles de natures diverses sans avoir à l'adapter à chaque cas, le modèle doit être le plus généralisable possible par rapport aux distributions des données réelles. Nous nous attachons de plus à intégrer la notion de système complexe dans la résolution du problème, l'idée étant que la façon dont les changements observés affectent certaines composantes de la série temporelle et pas d'autres est intrinsèquement liée aux dépendances entre les variables. Il est donc possible d'utiliser la structure de dépendance pour améliorer la détection de changements, mais également de déduire cette structure des informations extraites de la distribution des ruptures entre les différents signaux.

Contributions

Les travaux de recherche présentés dans ce document ont conduit principalement à trois contributions. La première est un modèle nommé le *Bernoulli Detector*, qui réalise la détection de ruptures multiples dans un signal univarié. Il est construit à partir des p-valeurs d'un test statistique non paramétrique, le test de Wilcoxon-Mann-Whitney, intégrées dans un cadre bayésien. Le choix de ce test rend la méthode applicable à plusieurs distributions sans avoir à modifier le modèle. Seul un paramètre, contrôlant le risque de détecter à tort un changement, doit être choisi. La seconde contribution est une variante de la première méthode pour traiter conjointement plusieurs signaux. Un paramètre représentant les possibles liens entre les variables est introduit dans le modèle du *Bernoulli Detector*. Il en résulte une méthode flexible, pouvant tenir compte d'informations a priori sur les relations intrinsèques au système pour améliorer la détection, ou parvenant à apprendre les probabilités d'observer les mêmes ruptures sur plusieurs signaux. La méthode conduit à la segmentation de toutes les composantes de la série temporelle, incluant les ruptures communes à plusieurs signaux comme les ruptures propres à chacun. Enfin, la troisième contribution propose un moyen de retrouver et représenter en partie le modèle d'indépendance conditionnelle régissant les variables. Grâce au formalisme des réseaux bayésiens, nous vérifions que le graphe de dépendance peut être inféré à partir de l'étude des distributions des ruptures entre les signaux en adaptant une méthode d'apprentissage de la structure du graphe aux résultats du *Bernoulli Detector*.

Organisation du document

Afin de répondre à la problématique de la détection des ruptures et de la recherche du graphe de dépendance, les trois contributions sont présentées progressivement dans ce document. Nous commençons par définir dans le chapitre 1 le problème de la détection de sauts de moyenne, où il s'agit de déterminer l'existence d'une rupture entre deux portions d'un signal univarié. Le problème est ensuite complexifié, avec la recherche d'une rupture dont la position n'est pas connue, puis de plusieurs ruptures, et enfin le cas de multiples ruptures à des positions inconnues dans plusieurs séries temporelles synchronisées. Le chapitre propose une revue des principales approches rencontrées dans la considérable littérature sur le sujet. La dernière partie de ce chapitre présente le formalisme des réseaux bayésiens et les principales méthodes destinées à l'estimation d'un graphe de dépendance, en particulier lorsqu'il se prête à une interprétation causale.

Dans le chapitre 2, nous revenons sur le problème de la détection d'une rupture en un point donné entre deux segments consécutifs $\boldsymbol{s}_1$ et $\boldsymbol{s}_2$. Il se présente sous la forme d'un test d'homogénéité, où l'hypothèse nulle H_0 est qu'il n'y a pas de rupture entre $\boldsymbol{s}_1$ et $\boldsymbol{s}_2$, contre l'hypothèse alternative H_1 qu'il y a une rupture. Les tests paramétriques classiques qui permettent de détecter un saut de moyenne ou de médiane sont passés en revue. Ces derniers reposant en général sur une hypothèse de normalité des données, le principe des tests non paramétriques de rang, en particulier le test de Wilcoxon-Mann-Whitney,

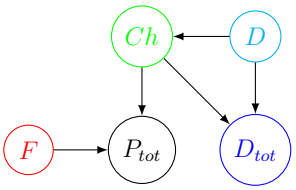
connu pour sa robustesse aux petits échantillons et au bruit non gaussien, est exposé. Nous proposons alors une solution alternative, construite sur la fonction de vraisemblance empirique. Ces différents tests sont mis en œuvre sur une série de signaux de tailles et de distributions variables, afin d'établir une comparaison. Ces expériences mettent en valeur les avantages du test de Wilcoxon-Mann-Whitney, qui est retenu pour être intégré dans notre modèle du *Bernoulli Detector*.

Ce modèle, destiné dans un premier temps à la détection de plusieurs ruptures dans un signal univarié, fait l'objet du chapitre 3. L'inférence bayésienne a été choisie afin d'estimer la segmentation du signal qui maximise la densité de probabilité a posteriori du modèle sachant les observations du signal $\boldsymbol{X}$. La segmentation est représentée par un vecteur $\boldsymbol{R}$ de variables indicatrices des ruptures. L'originalité de la méthode tient dans l'élaboration de la fonction de vraisemblance de $\boldsymbol{X}$ sachant $\boldsymbol{R}$ à partir des p-valeurs d'un test statistique appliqué en chaque point du signal, le test de Wilcoxon-Mann-Whitney ayant été retenu d'après les conclusions établies au chapitre 2. Sur le plan théorique, deux résultats sont démontrés pour le contrôle de la détection dans des cas simples. L'algorithme permettant la mise en œuvre de la méthode est décrit. Les performances de l'algorithme sont comparées expérimentalement avec deux approches de l'état de l'art, l'une paramétrique et l'autre non paramétrique, afin de mettre en avant la robustesse du *Bernoulli Detector* en présence de valeurs aberrantes dans le signal et son caractère générique.

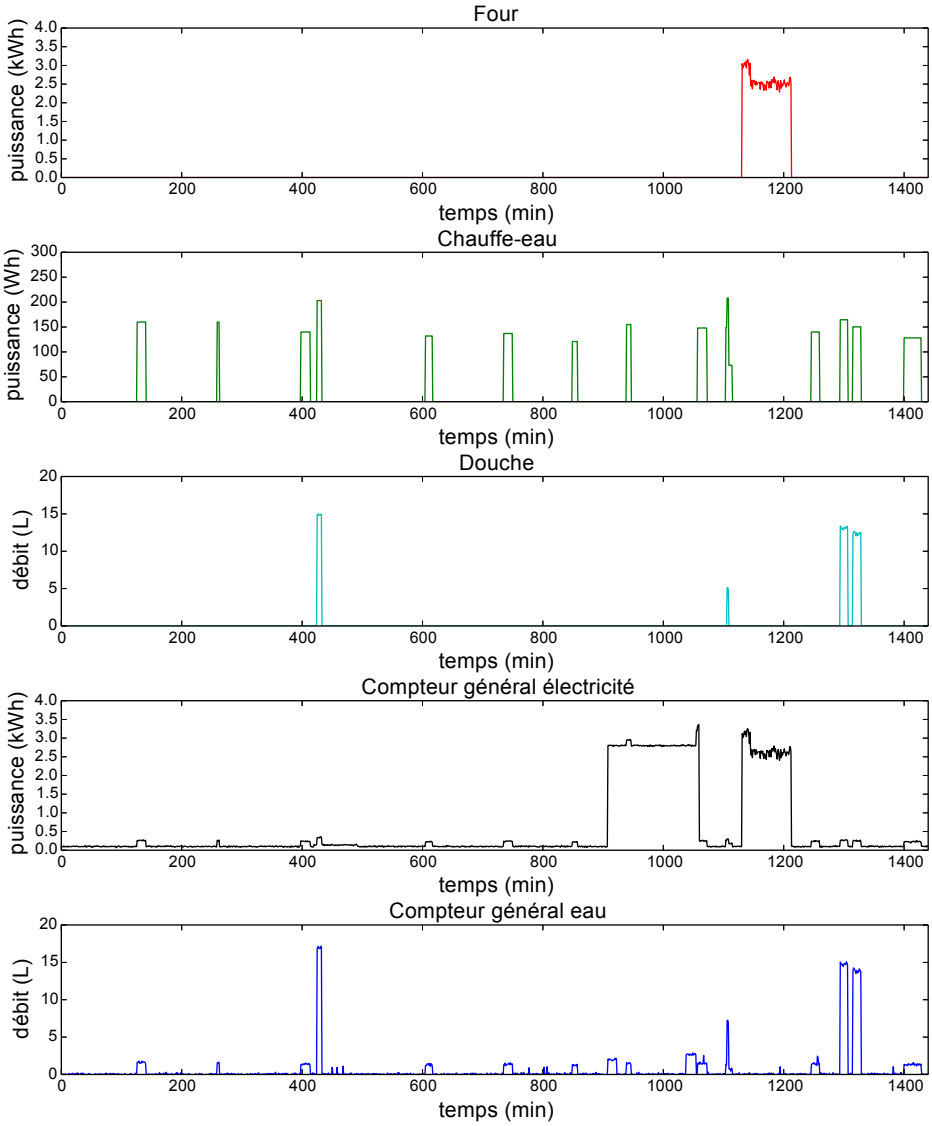
Le modèle est alors enrichi pour segmenter des séries temporelles multivariées. L'objectif du chapitre 4 est d'intégrer les relations entre les variables du système. Pour ce faire un nouveau paramètre $\boldsymbol{P}$ est introduit, représentant la probabilité que des événements soient simultanés ou non entre les signaux. L'intérêt de ce paramètre est qu'il permet d'apprendre la structure de dépendance des variables ou bien de l'utiliser lorsqu'elle est connue pour améliorer la détection et accélérer la résolution. Le *Bernoulli Detector* fournit alors une segmentation par signal, où les ruptures partagées entre plusieurs composantes sont synchronisées. L'algorithme est appliqué sur des données réelles. L'intérêt du nouveau paramètre est illustré avec le traitement de mesures de consommation électrique domestique. La seconde application est un ensemble de profils d'hybridation génomique. La segmentation de ce type de données biologiques est un problème réel classique en détection de rupture. Notre modèle conduit à un diagnostic à la fois commun et individualisé des anomalies de copies de l'ADN pour un groupe d'individus.

Enfin, dans le chapitre 5 nous proposons d'étendre l'étude à la recherche des relations d'indépendances conditionnelles sous-jacentes régissant les variables du système. Notre approche repose uniquement sur les variables indicatrices des ruptures, estimées par la méthode du *Bernoulli Detector*. Elles sont considérées comme des variables aléatoires binaires, et forment un réseau bayésien, dont le graphe acyclique dirigé représente la structure de dépendance. Nous supposons que les ruptures se produisent simultanément avec une probabilité donnée sur les signaux dont les variables indicatrices sont connectées, l'estimation des dépendances ne peut donc pas se faire par l'exploitation d'un décalage temporel dans la propagation des événements dans le réseau. Le paramètre $\boldsymbol{P}$ fournit les densités de probabilité jointes entre les variables du graphe recherché. Nous adaptons une méthode d'apprentissage du graphe dans l'espace des classes d'équivalences de Markov, basée sur

le calcul d'une fonction de score, et proposons deux critères exploitant les résultats du *Bernoulli Detector*. La méthode est testée sur des simulations puis sur les données réelles électriques, qui illustrent le prolongement de notre méthode du *Bernoulli Detector* pour la détection de rupture à l'étude plus générale d'un système complexe. Ce document se conclut par une synthèse des résultats apportés par les trois contributions principales. Les avantages et limitations des méthodes proposées y sont soulignés. Plusieurs questions soulevées au cours du travail de recherche et les pistes à explorer dans des études futures sont finalement évoquées.



(a) Représentation des relations de dépendances par un graphe entre les variables des puissances électriques (four F, chauffe-eau Ch et compteur général P_{tot}), et des débits d'eau (douche D et débit total D_{tot}) symbolisés par des cercles. Lorsque la douche fonctionne et que de l'eau chaude est requise, le chauffe-eau est actif. Ch est donc reliée aux deux compteurs et à D. Les variables inconnues comme les consommations des autres appareils électroménagers ne sont pas représentées.



(b) Mesures de consommation de puissance électrique et d'eau sur une journée.

FIGURE 1 – Suivi de consommation électrique et d'eau dans une habitation.

Chapitre 1

État de l'art : détection de ruptures et inférence de réseaux

La détection d'anomalies est un problème classique en traitement du signal, qui se décline en de nombreuses formulations. La grande variété des solutions proposées depuis les années 60 s'explique d'une part par les différentes approches méthodologiques qui ont été explorées, et d'autre part par la multitude d'applications réelles, apportant des contraintes spécifiques à chaque cas de figure. Notre problématique consiste à détecter des ruptures dans les moyennes de séries temporelles multivariées, ces dernières rassemblant les observations des éléments constitutifs d'un système complexe, régi par des interactions. Ce chapitre propose une revue des principales solutions de l'état de l'art. Il se compose de deux parties : la détection de ruptures, puis l'estimation et la représentation des dépendances du réseau.

1.1 Détection de ruptures

1.1.1 Généralités, définitions et notations

La détection de ruptures est un sujet connexe à d'autres problèmes classiques du traitement du signal, de l'information ou de la statistique, parmi lesquels on peut citer la détection d'anomalies en général, les tests d'homogénéité, l'ajustement de courbe ou encore le débruitage. La détection de ruptures elle-même correspond à plusieurs problèmes : détecter les changements dans les caractéristiques du signal et les localiser, pour pouvoir ensuite analyser individuellement les segments de la série temporelle. Les informations qui en sont extraites permettent de faire de la classification, sélectionner un modèle ou proposer un diagnostic. Les méthodes de l'état de l'art réalisent en général la détection et la localisation. Des ouvrages de référence détaillent les diverses approches existantes, le lecteur peut notamment se référer aux livres de [Basseville *et al.*, 1993] pour la détection d'une unique rupture principalement dans un contexte en ligne, [Chen et Gupta, 2011] pour un ensemble de méthodes paramétriques et leurs applications, [Brodsky et Darkhovsky, 2013]

pour les approches non paramétriques, [Csörgö et Horváth, 1997] pour le développement de résultats théoriques, [Douc *et al.*, 2014] qui est dédié à l'analyse de séries temporelles en général et [Gustafsson et Gustafsson, 2000] aux algorithmes de détection de ruptures avec des filtres adaptatifs.

L'objet de notre étude est une série temporelle, composée d'un nombre fini n d'observations successives faites à intervalles réguliers. Elle est formalisée par le vecteur $\boldsymbol{X} = (X_1, \ldots, X_n)$, dont les variables aléatoires X_i suivent une distribution de probabilité F_i donnée, $1 \leq i \leq n$. En l'absence de changement $\boldsymbol{X}$ est stationnaire. En général, on suppose que les variables $X_1, \ldots, X_n$ sont indépendantes. Une *rupture* est définie par un changement abrupt de l'une ou plusieurs des caractéristiques locales du signal. La ou les grandeurs affectées par le changement sont représentées par le paramètre θ. Il peut s'agir par exemple de la moyenne, de la variance, du moment d'ordre r, de la famille de la distribution des données, des composantes spectrales ou encore des coefficients d'autorégression. La rapidité du phénomène conduit à le qualifier de rupture : la transition de l'état antérieur à l'état modifié se fait dans un intervalle de temps inférieur ou proche de la période d'échantillonnage.

Définition 1.1.1 (Rupture)**.** *Soit un échantillon* $\boldsymbol{X} = (X_1, \ldots, X_n)$*. Pour tout* $1 \leq i \leq n$*, on associe à la variable aléatoire* X_i *un paramètre* θ_i*. Le vecteur des paramètres* θ_i *est noté* Θ*. Pour tout* $1 < i < n$*,*

$$X_i \textit{ est une } \text{rupture } \textit{si } \theta_i \neq \theta_{i+1}, \tag{1.1}$$

$$X_i \textit{ n'est pas une rupture si } \theta_i = \theta_{i+1}. \tag{1.2}$$

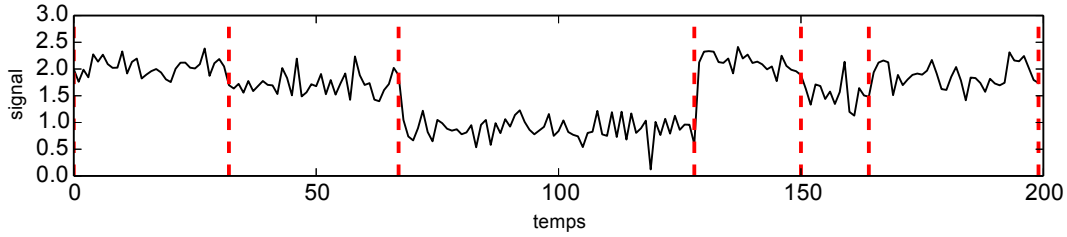
Par convention, X_1 *et* X_n *sont considérées comme des ruptures.*

La figure 1.1 montre deux exemples de séries temporelles comportant $K = 5$ ruptures (sans compter les extrémités), qui délimitent $K + 1 = 6$ segments où les observations partagent les mêmes caractéristiques. Sur le segment k, les variables X_i sont indépendantes et identiquement distribuées selon la loi normale de moyenne μ_k et de variance σ_k^2. La série temporelle de la figure 1.1a présente des ruptures dans la moyenne : les paramètres θ_i correspondent aux valeurs moyennes μ_k. Dans la figure 1.1b, les ruptures affectent la variance : les paramètres θ_i correspondent aux variances σ_k^2.

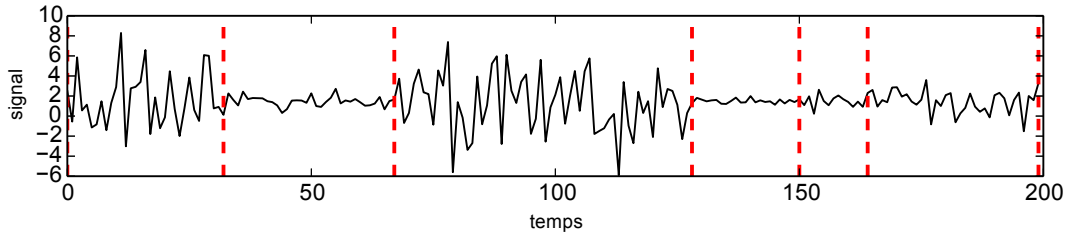
Le type de ruptures qui nous intéresse affecte la médiane de la série temporelle. Cette grandeur, égale à la moyenne pour les distributions symétriques, est plus robuste aux valeurs extrêmes que cette dernière pour donner la position de la distribution. Le signal étudié est ainsi composé de segments où la médiane est constante, comme le montre l'exemple de la figure 1.1a. En pratique, les événements se produisent à des instants inconnus qu'il faudra déterminer, leur nombre K étant lui-même inconnu dans la plupart des cas. Ces instants sont notés τ_k, $1 \leq k \leq K$, avec $\tau_1 > 1$ et $\tau_K < n$. Alors

$$\theta_1 = \cdots = \theta_{\tau_1} \neq \theta_{\tau_1+1} = \cdots = \theta_{\tau_K} \neq \theta_{\tau_K+1} = \cdots = \theta_n. \tag{1.3}$$

Le vecteur $\boldsymbol{T} = (\tau_1, \ldots, \tau_K)$ est appelé la *segmentation* de la série temporelle, il scinde le signal au niveau des ruptures en $K + 1$ segments. Notre objectif est donc d'estimer le



(a) Ruptures dans les moyennes μ_k des lois normales, variance constante.



(b) Ruptures dans les variances σ_k^2 des lois normales, moyenne constante.

FIGURE 1.1 – Exemples de séries temporelles de 200 points comportant des ruptures (en rouge) dans la moyenne (en 1.1a) et dans la variance (en 1.1b) aux instants 32, 67, 128, 150, et 164. Entre deux ruptures, sur le segment k, les observations sont indépendantes et identiquement distribuées (i.i.d) selon une loi normale de moyenne μ_k et de variance σ_k^2, $1 \leq k \leq K+1$.

vecteur $\boldsymbol{T}$ auquel est associé la relation (1.3), où θ_i est la médiane de X_i dans le segment k, pour tout $\tau_{k-1} < i \leq \tau_k$ et $1 \leq k \leq K$. Dans ce chapitre, les méthodes de l'état de l'art qui sont présentées sont en général destinées à un changement de moyenne ; sauf mention contraire, on suppose par la suite que cette grandeur est confondue avec la médiane pour les distributions des variables aléatoires X_i considérées, comme c'est le cas pour la loi normale.

L'estimation de la segmentation $\hat{\boldsymbol{T}}$ conduit à la détermination d'un signal constant par morceaux $\hat{\Theta}$. La détection de ruptures se ramène alors au contexte du débruitage de signaux. En effet, si on décompose les variables comme

$$X_i = \theta_i + \epsilon_i \tag{1.4}$$

où θ_i est la moyenne de la distribution de X_i, et où ϵ_i est une variable aléatoire de moyenne nulle et de variance finie, considérée comme du bruit, l'estimation du vecteur $\hat{\Theta}$ constant par morceaux est une approximation de $\boldsymbol{X}$, et la segmentation $\hat{\boldsymbol{T}}$ est déduite des positions des segments. Détecter les ruptures permet ensuite d'appliquer des traitements locaux sur les segments, où la série temporelle est stationnaire.

Pour estimer $\boldsymbol{T}$, le problème est formulé par un test entre l'hypothèse nulle H_0, où il n'y a pas de ruptures, et l'hypothèse alternative H_1, où des ruptures sont observées aux instants $\tau_1, \ldots, \tau_K$:

$$\begin{aligned} H_0 &: \quad \theta_1 = \theta_2 = \cdots = \theta_n, \\ H_1 &: \quad \theta_1 = \cdots = \theta_{\tau_1} \neq \theta_{\tau_1+1} = \cdots = \theta_{\tau_K} \neq \theta_{\tau_K+1} = \cdots = \theta_n. \end{aligned} \tag{1.5}$$

Il existe plusieurs hypothèses de la forme H_1, on considère pour commencer que le nombre et les positions des ruptures $\tau_1, \ldots, \tau_K$ sont connues. L'objectif est de construire une méthode qui permette de choisir entre H_0 et H_1 en fonction du vecteur $\boldsymbol{X}$. Le test statistique compare l'adéquation des observations au modèle sous H_0, où il n'y a pas de rupture, avec celle au modèle sous H_1. Une règle de décision est établie à partir d'une statistique de test Z pour trancher entre les deux hypothèses, sous la forme :

$$\begin{cases} \text{accepter } H_0 & \text{si } Z > h, \\ \text{rejeter } H_0 & \text{si } Z \leq h, \end{cases} \tag{1.6}$$

pour un seuil h approprié. En prenant une décision, le test risque de se tromper de deux manières possibles.

Définition 1.1.2 (Erreur de type I)**.** *L'erreur de type I est faite lorsqu'on détecte à tort une rupture qui n'existe pas. Elle est également appelée risque de première espèce ou niveau, et est notée α.*

Définition 1.1.3 (Erreur de type II)**.** *L'erreur de type II est faite lorsqu'on ne détecte pas une vraie rupture. Elle est également appelée risque de deuxième espèce ou puissance du test, et est notée β.*

Pour garantir l'efficacité du test, il faut limiter ces risques. Ils sont quantifiés par les probabilités de fausse alarme P_{FA} et de détection P_D :

$$\begin{aligned} P_{FA} &= P(H_0 \text{ est rejetée} | H_0 \text{ est vraie}), \\ &= \alpha, \\ P_D &= P(H_0 \text{ est rejetée} | H_0 \text{ est fausse}), \\ &= 1 - \beta. \end{aligned} \tag{1.7}$$

Lorsque ces probabilités s'expriment en fonction d'un paramètre du test, de préférence ajustable, il est possible de déterminer les conditions pour lesquelles un nombre donné d'erreurs de type I et II sont observées en probabilité. On parle alors de *contrôle de la détection*.

Plusieurs tests peuvent être construits pour un même problème. Pour choisir celui qui conduira à la meilleure détection des ruptures, il faut déterminer quel est le test le plus puissant au niveau α souhaité, c'est-à-dire celui qui maximise la probabilité de détection P_D pour une probabilité de fausse alarme P_{FA} donnée.

Définition 1.1.4 (Test uniformément plus puissant)**.** *Un test est dit uniformément le plus puissant de niveau α s'il est de niveau α et si sa puissance est supérieure à celles de tous les autres tests de niveau inférieur ou égal à α.*

Pour choisir s'il faut rejeter ou non H_0, la règle de décision est établie de manière à réaliser un compromis entre les deux risques. Elle s'applique sur la statistique de test : si cette dernière appartient à une région critique $R(\alpha)$, H_0 n'est pas rejetée, sinon on rejette H_0. La région critique est souvent définie par rapport à la probabilité P_{FA}, de telle sorte que P_D soit maximisée. Le livre de [Lehmann et Romano, 2006] rassemble un grand nombre de résultats pour la construction de tests et aborde la question de l'optimalité.

1.1.2 Enjeux

La détection de ruptures consiste à déceler la présence d'un ou plusieurs changements brutaux et à les localiser dans la série temporelle. Déterminer l'existence d'une rupture est d'autant plus difficile que cette dernière n'est pas forcément caractérisée par un décalage de grande amplitude entre θ_{τ_k} et θ_{τ_k+1} par rapport à la dispersion des observations. Un enjeu de la détection est donc d'être sensible aux faibles variations tout en garantissant une certaine robustesse au bruit. Lorsque les changements sont progressifs, c'est-à-dire qu'ils s'étalent sur plusieurs indices $\theta_{i-u} \neq \cdots \neq \theta_i \neq \cdots \neq \theta_{i+v}$, leur détection peut s'avérer problématique, soit parce que plusieurs ruptures y sont estimées, soit parce que la rupture est mal localisée voire non détectée. On notera également que si notre objectif est la détection de changements dans la médiane des segments, des ruptures peuvent affecter d'autres grandeurs, mais la méthode choisie n'y est pas forcément sensible. Dans cette partie, la démarche de la détection de ruptures est présentée progressivement, de la détection d'une simple rupture localisée à la détection de plusieurs ruptures de nombre et de positions inconnus. Les modèles sont exprimés par des fonctions de vraisemblance, afin d'illustrer l'augmentation de la difficulté lors de l'introduction de nouvelles inconnues.

Détecter une rupture unique à une position donnée

Dans le premier cas de figure, le plus simple, X_τ est l'unique rupture du signal monovarié $\boldsymbol{X}$, où $X_i \in \mathbb{R}$, pour tout $1 \leq i \leq n$. Le problème (1.5) devient

$$\begin{array}{lll} H_0: & \theta_1 = \theta_2 = & \ldots \qquad\qquad = \theta_n, \\ H_1: & \theta_1 = \theta_2 = & \cdots = \theta_\tau \neq \theta_{\tau+1} = \ldots = \theta_n. \end{array} \tag{1.8}$$

Cette formulation est équivalente à celle d'un test d'homogénéité entre les deux échantillons $\boldsymbol{X}_1 = (X_1, \ldots, X_\tau)$ et $\boldsymbol{X}_2 = (X_{\tau+1}, \ldots, X_n)$. Les tests classiques pour la détection d'un saut de moyenne, ou assimilé, comme les tests de Student, de Welch, d'Hotelling ou de Wilcoxon, sont présentés en détail au chapitre 2. Un grand nombre d'entre eux repose sur l'hypothèse que les données suivent une distribution normale.

Une notion fondamentale des méthodes paramétriques est celle de la fonction de vraisemblance. Celle-ci exprime la densité de probabilité du vecteur $\boldsymbol{X}$ en fonction du vecteur des paramètres Θ :

$$L(\boldsymbol{X}|\Theta) = f(\boldsymbol{X}|\Theta), \tag{1.9}$$

avec les notations $f(\boldsymbol{X}|\Theta) = f_\Theta(\boldsymbol{X})$ quand $\boldsymbol{X}$ est une variable aléatoire continue, et $f(\boldsymbol{X}|\Theta) = p_\Theta(\boldsymbol{X})$ quand $\boldsymbol{X}$ est une variable aléatoire discrète. Ce terme permet de déterminer le vecteur des paramètres $\hat{\Theta}$ qui décrit le mieux le vecteur des observations $\boldsymbol{X}$ et qui maximise donc la probabilité $f(\boldsymbol{X}|\Theta)$. Il est alors judicieux de l'employer dans le test d'hypothèses pour décider entre plusieurs valeurs de Θ. On suppose par la suite que les observations $X_1, \ldots, X_n$ sont indépendantes. La vraisemblance s'écrit alors comme le produit des vraisemblances marginales :

$$L(\boldsymbol{X}|\Theta) = \prod_{i=1}^{n} f(X_i|\theta_i). \tag{1.10}$$

Pour comparer les deux modèles des hypothèses H_0 et H_1 du problème (1.5), on utilise le rapport des vraisemblances :

$$\Lambda_n(\boldsymbol{X}) = \frac{\prod_{i=1}^{n} f(X_i|\theta_a)}{\prod_{i=1}^{\tau} f(X_i|\theta_a) \prod_{i=\tau+1}^{n} f(X_i|\theta_b)}, \tag{1.11}$$

où θ_a est valeur du paramètre θ_i sous H_0 et avant la rupture en τ sous H_1, et θ_b est la valeur prise sous H_1 après la rupture.

Dans le cas de figure le plus simple, les valeurs de θ_a et θ_b sont connues a priori, ainsi que les expressions des fonctions de vraisemblance ; le calcul de la statistique (1.11) est immédiat, et le seuil h de la règle de décision (1.6) peut être choisi en fonction des propriétés de la statistique de test et de la puissance α désirée.

Lorsque que les deux modèles sous H_0 et sous H_1 sont connus, le test d'hypothèses est de la forme

$$\begin{aligned} H_0 :&\quad \Theta = \Theta_0, \\ H_1 :&\quad \Theta = \Theta_1, \end{aligned} \tag{1.12}$$

où $\Theta_0 \in \Omega_0$ et $\Theta_1 \in \Omega_1$ sont connus. On parle alors de test d'hypothèses simples. Le rapport de vraisemblance s'écrit

$$\Lambda_n(\boldsymbol{X}) = \frac{L(\boldsymbol{X}|\Theta_0)}{L(\boldsymbol{X}|\Theta_1)}. \tag{1.13}$$

Pour notre problème (1.8), nous avons $\Theta_0 = (\theta_a, \ldots, \theta_a)$ et $\Theta_1 = (\theta_a, \ldots, \theta_a, \theta_b, \ldots, \theta_b)$, avec τ fois la valeur θ_a sous H_1, Ω_0 et Ω_1 étant de dimension n. Le lemme de Neyman-Pearson 1.1.1 établit un résultat fondamental pour l'établissement du test le plus puissant pour un test d'hypothèses simples [Neyman et Pearson, 1933].

Lemme 1.1.1 (Neyman-Pearson)**.** *Soient deux lois de probabilité $P_{|H_0}$ et $P_{|H_1}$, de densités de probabilité $f_{|H_0}$ et $f_{|H_1}$ respectivement, sujets à une mesure de probabilité μ.*

(i) Existence. *Pour tester $H_0 : f_{|H_0}$ contre $H_1 : f_{|H_1}$, il existe un test g et une constante λ tels que*

$$E_{|H_0}[g(\boldsymbol{X})] = \alpha \text{ sous } H_0, \tag{1.14}$$

$$g(\boldsymbol{X}) = \begin{cases} 1 & \text{quand } \frac{f_{|H_1}(\boldsymbol{X})}{f_{|H_0}(\boldsymbol{X})} \geq \lambda, \\ 0 & \text{quand } \frac{f_{|H_1}(\boldsymbol{X})}{f_{|H_0}(\boldsymbol{X})} < \lambda. \end{cases} \tag{1.15}$$

(ii) Condition suffisante pour un test le plus puissant. *Si un test satisfait* (1.14) *et* (1.15) *pour une constante λ, alors il est le plus puissant pour tester $f_{|H_0}$ contre $f_{|H_1}$ au niveau α.*

(iii) Condition nécessaire pour un test le plus puissant. *Si g est un test le plus puissant au niveau α pour tester $f_{|H_0}$ contre $f_{|H_1}$, alors il satisfait* (1.14) *presque sûrement pour un λ donné, pour la mesure μ. Il vérifie également* (1.15) *à moins qu'il n'existe un test de taille inférieure à α et de puissance 1.*

Le rapport de vraisemblance (1.13) est alors la statistique optimale pour le problème de sélection entre les modèles H_0 et H_1 (1.12) : ce test est le plus puissant pour le niveau α tel que $\Pr(\Lambda_n \geq h|H_0) = \alpha$. En partant de ce résultat, plusieurs tests basés sur le

rapport de vraisemblance ont été construits pour des variantes du problème (1.12) [Basseville *et al.*, 1993, Lehmann et Romano, 2006], mais leur optimalité n'est pas toujours conservée [Deshayes et Picard, 1986].

La résolution se complique lorsque certains paramètres sont inconnus. Dans le cadre d'un contrôle qualité, on a parfois accès à la valeur nominale θ_a du paramètre θ, et on suppose que c'est la valeur mesurée au début de la série temporelle. On cherche à déterminer si le système dérive et que θ prend une valeur $\theta_b \neq \theta_a$ à partir de l'instant τ. L'estimateur du maximum de vraisemblance est alors introduit pour remplacer θ_b dans (1.11) :

$$\Lambda_n(\boldsymbol{X}) = \frac{\prod_{i=1}^{n} f(X_i|\theta_a)}{\prod_{i=1}^{\tau} f(X_i|\theta_a) \sup_{\theta_b} \prod_{i=\tau+1}^{n} f(X_i|\theta_b)}. \tag{1.16}$$

Si de plus θ_a n'est pas connue, la statistique employée est

$$\Lambda_n(\boldsymbol{X}) = \frac{\sup_{\theta_a} \prod_{i=1}^{n} f(X_i|\theta_a)}{\sup_{\theta_a} \prod_{i=1}^{\tau} f(X_i|\theta_a) \sup_{\theta_b} \prod_{i=\tau+1}^{n} f(X_i|\theta_b)}. \tag{1.17}$$

Un résultat essentiel sur la distribution asymptotique du rapport de vraisemblance est fourni par le théorème de Wilks 1.1.1 [Wilks, 1938], quand n tend vers l'infini :

Théorème 1.1.1 (Théorème de Wilks)**.** *Soit un échantillon $\boldsymbol{X} = (X_1, \ldots, X_n)$ de variables aléatoires. On suppose que la distribution jointe $f(\boldsymbol{X})$ dépend du vecteur des paramètres inconnu $\Theta \in \Omega$ de dimension d, et que sous H_0, elle dépend du vecteur des paramètres inconnu $\Theta_0 \in \Omega_0$, de dimension d_0. Le rapport entre les vraisemblances maximales est donné par*

$$\Lambda_n(\boldsymbol{X}) = \frac{\sup\{L(\boldsymbol{X}|\Theta_0) : \Theta_0 \in \Omega_0\}}{\sup\{L(\boldsymbol{X}|\Theta) : \Theta \in \Omega\}}. \tag{1.18}$$

Alors, sous certaines conditions de régularité, lorsque H_0 est vraie la statistique $-2\log\Lambda_n(\boldsymbol{X})$ converge en distribution vers la loi du chi-deux à $(d - d_0)$ degrés de liberté, notée $\chi^2_{(d-d_0)}$, quand n tend vers l'infini.

Ce théorème permet de construire des tests d'hypothèses simples. Selon les distributions des données et les informations a priori qui sont disponibles, l'expression des fonctions de vraisemblances ou de leurs estimateurs est plus ou moins simple et le choix de la statistique de test plus ou moins immédiat.

Détecter et localiser une rupture unique

Lorsque la position τ de la rupture est inconnue, le problème (1.5) devient :

$$\begin{array}{lllll} H_0 : & & \theta_1 = & \cdots & = \theta_n \\ H_1 : & \exists \tau, 1 \leq \tau < n, & \theta_1 = & \cdots = \theta_\tau \neq \theta_{\tau+1} = \ldots & = \theta_n. \end{array} \tag{1.19}$$

τ peut être vu comme un paramètre de nuisance. Le problème a été initialement abordé dans [Page, 1957]. L'estimateur du maximum de vraisemblance (1.17) est alors exprimé en

fonction de τ pour parvenir à la statistique

$$\Lambda_\tau(\boldsymbol{X}) = \max_{1\leq\tau\leq n-1} \frac{\sup_{\theta_a} \prod_{i=1}^{n} f(X_i|\theta_a)}{\sup_{\theta_a} \prod_{i=1}^{\tau} f(X_i|\theta_a) \sup_{\theta_b} \prod_{i=\tau+1}^{n} f(X_i|\theta_b)}. \tag{1.20}$$

L'ouvrage [Chen et Gupta, 2011] rassemble des statistiques de test construites à partir du rapport de vraisemblance pour la détection d'un changement pour plusieurs types de distributions, comme la loi normale, la loi gamma ou la loi exponentielle. Les cas du saut de moyenne dans des distributions normales de variance connue et inconnue sont étudiés dans [Hinkley, 1970,Hawkins, 1977], qui présentent les distributions exactes des statistiques sous H_0. [Siegmund, 1988] calcule les intervalles de confiance de la statistique du rapport de vraisemblance pour la famille exponentielle.

En l'absence de toute information a priori sur la position de la rupture, [Deshayes et Picard, 1986] montrent qu'il n'existe pas de manière générale de statistique optimale pour construire le test le plus puissant à un niveau donné. Le test est alors défini de telle sorte qu'une approximation asymptotique de la statistique $-2\log\Lambda_\tau(\boldsymbol{X})$ garantisse un niveau de contrôle de la détection, quand n devient très grand [Csörgö et Horváth, 1997]. Dans [Deshayes et Picard, 1986], les auteurs différencient le cas asymptotique non local, où les erreurs de type I et II tendent vers 0, et le cas asymptotique local, où le niveau α est maintenu à un niveau donné et où l'erreur de type II ne tend pas vers 0. La distribution asymptotique locale sous H_0 de la statistique de test pour un saut de moyenne entre des distributions normales de variance connue est développée dans [Hinkley, 1970, Hawkins, 1977], mais il a été démontré qu'il n'existe pas de test optimal, tandis que l'optimalité peut être obtenue dans un cadre asymptotique non local [Deshayes et Picard, 1986].

Détecter et localiser plusieurs ruptures

Enfin lorsqu'il y a potentiellement plusieurs ruptures dans le signal, à des positions inconnues, le problème (1.19) devient un test d'hypothèses multiples, à appliquer en chaque point pour $1 < i < n$:

$$\begin{aligned} H_0(i) &: \quad X_i \text{ n'est pas une rupture,} \\ H_1(i) &: \quad\quad X_i \text{ est une rupture.} \end{aligned} \tag{1.21}$$

Un cas particulier est celui du problème de la rupture épidémique, où une portion de la série temporelle diffère du reste du signal sur un intervalle de temps. Le paramètre d'intérêt θ_i prend la valeur nominale θ_N par défaut, et la valeur θ_A sur le segment anormal à localiser. Ce problème, qui trouve souvent des applications médicales, est traité par exemple dans [Ramanayake et Gupta, 2003, Yao, 1993, Ning *et al.*, 2012] pour des changements de moyenne, et la distribution asymptotique d'une statistique basée sur le rapport de vraisemblance est donnée dans [Csörgö et Horváth, 1997].

Certains modèles sont construits pour détecter un nombre donné de ruptures K, à des positions inconnues, par exemple dans [Lavielle et Teyssiere, 2006,Hawkins, 2001]. [Csörgö et Horváth, 1997] présentent des résultats pour la convergence de statistiques de test basées sur la maximisation du rapport de vraisemblances. Pour un nombre donné de ruptures K, les hypothèses nulle et alternative sont données par le problème (1.5).

En supposant que les positions des ruptures sont connues, ainsi que les valeurs du paramètre θ, notées θ_a sous H_0 et sous H_1 $\theta_a, \ldots, \theta_l$ pour les segments 1 à $K+1$ respectivement, l'expression du rapport de vraisemblance est

$$\Lambda_n(\boldsymbol{X}) = \frac{\prod_{i=1}^{n} f(X_i|\theta_a)}{\prod_{i=1}^{\tau_1} f(X_i|\theta_a) \prod_{i=\tau_1+1}^{\tau_2} f(X_i|\theta_b) \ldots \prod_{i=\tau_K+1}^{n} f(X_i|\theta_l)}. \tag{1.22}$$

La complexité d'une telle approche est visible immédiatement lorsque des paramètres sont inconnus, comme c'est presque toujours le cas. En introduisant les estimateurs du maximum de vraisemblance, le terme (1.22) est alors

$$\hat{\Lambda}_\Theta(\boldsymbol{X}) = \frac{\sup_{\theta_a} \prod_{i=1}^{n} f(X_i|\theta_a)}{\sup_{\theta_a} \prod_{i=1}^{\tau_1} f(X_i|\theta_a) \ldots \sup_{\theta_l} \prod_{i=\tau_K+1}^{n} f(X_i|\theta_l)}. \tag{1.23}$$

Ce rapport doit être maximisé par rapport aux indices temporels de $\boldsymbol{T} = (\tau_1, \tau_2, \ldots, \tau_K)$ lorsque les positions des ruptures sont inconnues. La statistique devient alors

$$\hat{\Lambda}_T(\boldsymbol{X}) = \max_{\boldsymbol{T}, 1<\tau_1<\cdots<\tau_K<n} \hat{\Lambda}_\Theta(\boldsymbol{X}). \tag{1.24}$$

Cependant la maximisation directe d'une telle statistique a un coût combinatoire qui rend l'opération difficile. Enfin, quand le nombre K de ruptures n'est pas connu a priori, ce qui est en général le cas, il faut maximiser (1.24) rapport à K :

$$\widehat{\Lambda}_K(\boldsymbol{X}) = \max_{1<K<n} \hat{\Lambda}_T(\boldsymbol{X}) \tag{1.25}$$

Devant la complexité du problème (1.25), une résolution directe n'est pas toujours envisageable au-delà de quelques ruptures. Cette forte augmentation de la complexité se rencontre aussi dans les modèles autres que ceux construits sur le rapport de vraisemblance. Une solution consiste alors à introduire des approximations ; d'autres auteurs choisissent plutôt une résolution par programmation dynamique, comme [Bai et Perron, 2003, Lavielle et Teyssiere, 2006, Lung-Yut-Fong *et al.*, 2011c]. Cette méthode algorithmique a été initialement introduite dans [Bellman, 1954], elle consiste à résoudre récursivement des sous-problèmes de façon optimale. Un exemple de relations de récursion pour la détection de plusieurs ruptures est donné dans [Kay, 1998, annexe 12A] : le premier segment, de l'indice 1 à l'indice τ_1, est estimé en déterminant la position optimale de la première rupture, puis le deuxième segment est estimé à partir de l'indice $\tau_1 + 1$, jusqu'à la position optimale de la deuxième rupture, et ainsi de suite, jusqu'à l'obtention des K ruptures. Le nombre optimal de ruptures $K^\star$ peut être déduit a posteriori. Ainsi, pour résoudre un problème d'estimation des coefficients de régression linéaire multiple avec des ruptures, [Bai et Perron, 2003] proposent une procédure efficace, dont la complexité en $O(n^2)$ est indépendante du nombre de changements. Pour réduire encore la complexité, les auteurs suggèrent d'introduire une taille minimale des segments, ce qui limite le nombre de segmentations possibles.

Le problème de la détection de ruptures multiples (1.5) ne peut pas toujours être résolu par programmation dynamique, selon la méthode choisie. Une autre stratégie consiste à

détecter les ruptures une par une de manière à se rapporter à une succession de problèmes de détection d'une rupture unique (1.19). Dans les approches séquentielles par exemple, seule une portion du signal est traitée, débutant après la dernière rupture détectée où à la limite d'une fenêtre glissante, et terminant à la dernière observation acquise, lors d'un traitement en ligne, ou à la limite de la fenêtre. Une supposition communément rencontrée est que dans un intervalle restreint il ne peut se produire qu'un seul changement. Les procédures proposées dans [Niu et Zhang, 2012, Gijbels *et al.*, 1999] reposent ainsi sur un voisinage h donné autour du point i étudié, dans lequel le problème (1.19) est résolu localement avec $i = \tau$. L'approche par bissection, ou segmentation binaire, consiste à chercher les ruptures hiérarchiquement, en estimant d'abord la rupture la plus significative dans l'ensemble du signal, à la position $\tau_{(1)}$, puis une nouvelle rupture dans chacun des segments ainsi crées, $(X_1, \ldots, X_{\tau_{(1)}})$ et $(X_{\tau_{(1)}+1}, \ldots, X_n)$, en réitérant l'opération sur les sous-segments. Elle a été initialement proposée dans [Vostrikova, 1981], ses étapes sont résumées dans [Chen et Gupta, 2011, section 1.3, p.5]. Elle permet de réduire la complexité du problème en $\mathcal{O}(n \log n)$, mais ne parvient pas toujours à localiser correctement les ruptures [Fearnhead, 2006], elle ne conduit pas à la segmentation optimale si $K > 1$ [Hawkins, 2001], et nécessite l'introduction d'un critère d'arrêt, qui entraîne parfois une surestimation de K [Lavielle et Teyssiere, 2006, Matteson et James, 2014].

En se ramenant à un problème de détection d'une unique rupture, des séries temporelles de très grande dimensions, ou de taille n croissante pour les méthodes séquentielles, peuvent être traités. D'autres méthodes estiment toutes les ruptures simultanément dans l'ensemble du signal. Lorsque le nombre de ruptures inconnu est explicitement intégré au modèle, de nombreux auteurs proposent d'appliquer une méthode de détection de K ruptures pour K variant de 0 à $K_{max} < n$ [Lavielle et Teyssiere, 2006, Hawkins, 2001, Lung-Yut-Fong *et al.*, 2011c], puis de déterminer la segmentation $\boldsymbol{T}(K^*)$ optimale parmi les segmentations $\boldsymbol{T}(0), \boldsymbol{T}(1), \ldots, \boldsymbol{T}(K_{max})$. La complexité de ces procédures par sélection de modèle augmente, en raison des multiples résolutions du problème (1.24). En général, pour choisir la segmentation, un critère est appliqué, comme le critère d'information d'Akaike (AIC) [Akaike, 1998] ou le critère d'information bayésien (BIC) [Schwarz *et al.*, 1978] :

$$\text{AIC} = -2\ln(L) + 2K, \qquad \text{BIC} = -2\ln(L) + \ln(n)K, \tag{1.26}$$

où L est la fonction de vraisemblance du modèle. Un grand nombre de ruptures est ainsi pénalisé. D'autres auteurs, comme [Lavielle et Teyssiere, 2006], déterminent le nombre optimal de ruptures K^* par la localisation d'inflexion dans la courbe du critère choisi, tracé en fonction du paramètre K. Cette stratégie s'avère cependant peu précise. Dans d'autres approches, K n'apparaît pas explicitement dans le modèle comme un paramètre inconnu à estimer, il est déduit de l'estimation d'autres paramètres. Ainsi, dans la méthode BARD (pour Bayesian Abnormal Region Detector) [Fearnhead et Liu, 2007], K provient de l'estimation des longueurs et des positions des segments composant la série temporelle, et dans [Dobigeon *et al.*, 2007a], il est déterminé par l'intermédiaire de la matrice des indicatrices des positions des ruptures.

Enfin, signalons que l'établissement de résultats théoriques sur le contrôle de la détection en présence de ruptures multiples est un problème difficile, notamment lorsque les

distributions des observations ne sont pas gaussiennes. Pour déterminer quelle méthode de détection de ruptures choisir, de nombreux critères entrent en ligne de compte selon les applications, comme la dépendance à un modèle donné, la précision souhaitée dans la localisation des ruptures, la complexité ou encore le coût de calcul. La section suivante présente les principales méthodes de la littérature pour la détection de ruptures dans la moyenne d'une série temporelle.

1.1.3 Principales méthodes pour la détection de ruptures

En complément des ouvrages précédemment cités, il existe des revues de l'état de l'art concentrées sur certaines sous-parties du domaine. [Jandhyala *et al.*, 2013] présentent des méthodes pour l'analyse rétrospective de signaux affectés par des sauts de moyennes, construite sur une vraisemblance, paramétrique, non paramétrique et semi-paramétrique et des approches bayésiennes. [Khodadadi et Asgharian, 2008] recensent les articles consacrés à la détection de ruptures pour la régression de séries temporelles. Les approches statistiques peuvent être présentées de plusieurs manières. Dans son article [Lee, 2010], l'auteur propose une revue chronologique de la littérature sur la détection de cinq types de changements : de moyenne, de variance, de pente dans un processus de régression, de taux de risque, et de distribution. Dans cette partie, nous effectuons un tour d'horizon des grandes catégories de méthodes dédiées à la détection de sauts de moyennes.

Méthodes séquentielles

Les méthodes séquentielles sont dédiées au traitement de données au fur et à mesure de leur acquisition, par opposition à l'analyse rétrospective ou hors-ligne, où la taille n de la série temporelle est fixée. Elles sont généralement destinées à un cadre industriel, au contrôle qualité ou à la sécurité : l'application du test permet d'établir si le système se comporte normalement ou si un paramètre dérive, auquel cas une alarme est déclenchée. La décision est prise d'après la comparaison des dernières observations réalisées avec le début du signal, pris comme référence. Dans ce contexte en temps réel, l'enjeu principal de la détection est de décider si un changement a eu lieu, plutôt que de localiser la rupture, celle-ci pouvant même se produire progressivement. Les algorithmes sont évalués en fonction de la probabilité de détection, du délai pour la détection, qui mesure l'intervalle de temps entre la vraie rupture et sa détection, et du temps moyen entre deux fausses alarmes [Basseville *et al.*, 1993].

La première méthode de la carte de contrôle de Shewhart correspond à une application répétée du lemme de Neyman-Pearson 1.1.1 pour les hypothèses simples

$$\begin{array}{ll} H_0(i): & \theta_i = \theta_a, \\ H_1(i): & \theta_i = \theta_b, \end{array} \tag{1.27}$$

où θ_a est la valeur du paramètre θ en l'absence de changements, et θ_b est la valeur prise en cas d'anomalie dans le système [Shewhart, 1931]. La méthode repose sur le logarithme du

rapport des vraisemblances entre les variables d'indices j à k

$$S_j^k = -\sum_{i=j}^{k} \log \frac{f(X_i|\theta_a)}{f(X_i|\theta_b)}. \tag{1.28}$$

Par application du lemme 1.1.1, la statistique $S_1^m(l)$ est employée pour la détection d'une rupture dans le l^{e} intervalle de m points, permettant la détection d'une portion de signal de taille m où le paramètre d'intérêt θ prend une valeur anormale. L'instant τ où se produit la rupture est donné par le premier point sur la carte de contrôle où la statistique franchit le seuil h de la règle de décision. Dans le cas où les données sont distribuées normalement selon la loi $\mathcal{N}(\mu_a,\sigma^2)$ avant la rupture et $\mathcal{N}(\mu_b,\sigma^2)$ après, où σ^2 est supposée connue, la statistique étudiée correspond à

$$S_1^m = \frac{\mu_b - \mu_a}{\sigma^2} \sum_{i=1}^{m} \Big(X_i - \mu_a - \frac{\mu_b - \mu_a}{2} \Big). \tag{1.29}$$

Le test est alors équivalent à un test sur la moyenne empirique. Pour atténuer l'impact des observations les plus anciennes, les log-vraisemblances de (1.28) sont pondérées dans l'algorithme des moyennes géométriques glissantes (en anglais *Geometric Moving Average*, GMA) ou dans l'algorithme des moyennes glissantes finies (en anglais *Finite Moving Average*, FMA).

Dans le test séquentiel du rapport des probabilités (en anglais *Sequential Probability Ratio Test*, SPRT), le score cumulé S_1^n est calculé pour chaque nouvelle observation X_n. La règle de décision intègre deux seuils, et conduit de la sorte soit à accepter H_0, soit à rejeter H_0 et déclencher une alarme, soit à poursuivre la procédure avec une observation supplémentaire [Wald, 1973]. Le seuil ϵ en deçà duquel H_0 est acceptée est fixé à 0 dans le test de Carte de Contrôle de la Somme Cumulée (en anglais *Cumulative Sum Control Chart*, ou CUSUM) [Page, 1954]. Cet algorithme est la principale méthode statistique séquentielle pour la détection de ruptures et est largement utilisé pour le contrôle qualité. Il consiste à répéter l'algorithme SPRT tant que H_0 est acceptée, jusqu'à la détection d'une rupture. La statistique du test CUSUM est donnée par la relation récursive

$$g_k = \sup(S_{k-n_k+1}^k, 0), \tag{1.30}$$

où n_k est le nombre d'observations ajoutées depuis le dernier redémarrage du test SPRT. Le test peut être interprété comme un algorithme à fenêtre glissante, dont la taille dépend des instants où le SPRT s'arrête. Son optimalité a été démontrée dans [Lorden, 1971], lorsque les distributions des données sont connues. Lorsque la valeur de θ_b après le changement est inconnue, d'autres méthodes construites sur le rapport des vraisemblances sont disponibles, comme le test CUSUM pondéré ou le test de vraisemblance généralisé [Willsky et Jones, 1976].

Lorsque les différentes valeurs prises par le paramètre θ_i peuvent être représentées par un nombre fini d'états $S_1,\ldots,S_m$, la série temporelle peut être décrite par un modèle de Markov caché (en anglais Hidden Markov Model, HMM), dont l'estimation fournit à la fois

une segmentation et une classification des segments. La séquence des états S_k, régis par des probabilités de transition et les probabilités de l'état initial, constituent une chaîne de Markov. Le modèle de Markov caché est le processus temporel discret $\{(\theta_i,X_i), 1 \leq i \leq n\}$ où les variables aléatoires X_i, $1 \leq i \leq n$, sont supposées indépendantes, et où chaque X_i est générée conditionnellement à θ_i. Seule la séquence $\{X_i, 1 \leq i \leq n\}$ du processus est observée [Douc *et al.*, 2014, chapitre 9, p.287]. Les HMM permettent de décrire une grande variété de problèmes, par exemple pour l'analyse de données génomiques [Fridlyand *et al.*, 2004] ou le traitement de la parole [Rabiner, 1989]. L'estimation de la séquence des paramètres θ_i qui décrit le mieux les observations est déduite du modèle par maximisation de la fonction de vraisemblance. Si ce dernier n'est pas connu, les paramètres doivent également être estimés, mais il n'existe pas de méthode exacte. Les HMM peuvent être généralisées à des modèles où l'observation X_i dépend aussi de la variable précédente X_{i-1}. Des dépendances sont ainsi introduites dans le modèle. Les méthodes de filtrage particulaire permettent alors d'analyser la série temporelle [Douc *et al.*, 2014, partie 9.2, p.302].

Lorsque le modèle est linéaire, de la forme

$$\begin{cases} \theta_{i+1} = F\theta_i + GU_i + W_i \\ X_i = H\theta_i + JU_i + V_i, \end{cases} \tag{1.31}$$

où F est une matrice de transition, H une matrice d'observation, U une variable d'entrée contrôlée par J et G, et $(W_i)_i$ et $(V_i)_i$ sont des bruits blancs gaussiens [Basseville *et al.*, 1993, partie 3.2.1 p.83], le filtre de Kalman conduit à l'estimation de θ_i à partir de la connaissance de la séquence $\{X_1, \ldots, X_i\}$ [Douc *et al.*, 2014, chapitre 2] [Kalman et Bucy, 1961]. Il s'agit d'un filtre linéaire des observations, qui se met à jour pour chaque nouvelle observation X_{i+1}. Une déviation dans l'estimation de l'espace des états, après analyse des résidus, est le signe d'une rupture.

Un intérêt majeur des méthodes séquentielles est de s'appliquer à un traitement à la volée des données. En introduisant une fenêtre glissante, ou bien en ne considérant que les dernières observations depuis la détection de la rupture précédente, le problème de la détection de multiples ruptures (1.21) est rapporté à une séquence de problèmes de détection d'une rupture (1.19). Par opposition à l'approche séquentielle, on définit l'approche rétrospective, ou hors-ligne, où l'intégralité de la série temporelle $\boldsymbol{X}$, de taille fixe n, est disponible. Il est alors possible de localiser les ruptures sans le décalage temporel dû au délai de détection des méthodes séquentielles. Nous nous plaçons dans ce cadre pour la suite de notre étude.

Méthodes paramétriques

Les méthodes qualifiées de paramétriques reposent sur des hypothèses fortes sur les propriétés des distributions sous-jacentes des données, comme leur appartenance à une famille de fonctions connues. Dans [Robert, 2006, chapitre 1], l'auteur donne la définition suivante : «un modèle paramétrique statistique consiste en l'observation d'une variable aléatoire X distribuée selon $P(X|\theta)$, où seulement le paramètre θ est inconnu et appartient à un espace de dimension finie». La fonction de vraisemblance joue un rôle fondamental

dans la construction du modèle des données. La majorité des problèmes abordés dans la littérature concernent des données indépendantes et identiquement distribuées (i.i.d) sur chaque segment, suivant la loi normale $\mathcal{N}(\mu,\sigma^2)$; il s'agit en général de la détection de changements de la moyenne μ, de la variance σ^2, parfois des deux. Dans l'ouvrage [Chen et Gupta, 2011], les auteurs fournissent un grand nombre de méthodes paramétriques pour la détection de changement de moyenne, de variance, ou d'autre paramètres dans des distributions variées. Dans le cas normal avec un saut de moyenne à l'instant τ du problème (1.5), en supposant $\sigma^2 = 1$, les vraisemblances sous H_0 et H_1 s'écrivent

$$\begin{aligned} H_0 : \quad & L(\boldsymbol{X}|\mu_0) = \tfrac{1}{(\sqrt{2\pi})^n} e^{-\sum_{i=1}^{n} \frac{(X_i-\mu_0)^2}{2}}, \\ H_1 : \quad & L(\boldsymbol{X}|\mu_1,\mu_n) = \tfrac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\left(\sum_{i=1}^{\tau}(X_i-\mu_1)^2+\sum_{i=\tau+1}^{n}(X_i-\mu_n)^2\right)} \end{aligned} \tag{1.32}$$

où μ_0, μ_1 et μ_n sont les moyennes des segments sous H_0 et sous H_1, avant et après la rupture respectivement. Ces valeurs étant inconnues, on introduit les estimateurs du maximum de vraisemblance $\hat{\mu}_0$, $\hat{\mu}_1$, $\hat{\mu}_n$, qui sont les moyennes empiriques sur les segments :

$$\begin{gathered} \hat{\mu}_0 = \overline{x}_0 = \tfrac{1}{n}\sum_{i=1}^{n} X_i, \\ \hat{\mu}_1 = \overline{x}_\tau = \tfrac{1}{\tau}\sum_{i=1}^{\tau} X_i, \qquad \hat{\mu}_n = \overline{x}_{n-\tau} = \tfrac{1}{n-\tau}\sum_{i=\tau+1}^{n} X_i. \end{gathered} \tag{1.33}$$

Pour tester H_0 contre H_1 lorsque que la position τ est inconnue (problème (1.19)), [Lehmann et Romano, 2006] proposent le test du rapport vraisemblance basé sur la statistique U^2 telle que :

$$U^2 = \max_{1\leq\tau\leq n-1}(S - S_\tau) \tag{1.34}$$

avec

$$S_\tau = \sum_{i=1}^{\tau}(X_i - \overline{x}_\tau)^2 + \sum_{i=\tau+1}^{n}(X_i - \overline{x}_{n-\tau})^2 \quad \text{et} \quad S = \sum_{i=1}^{n}(X_i - \overline{x}_0)^2. \tag{1.35}$$

[Hawkins, 1977] développe les distributions exacte et asymptotique sous H_0 de $U = \max_{1\leq\tau\leq n-1}\sqrt{S - S_\tau}$, ainsi qu'une variante quand le paramètre σ n'est pas connu. La répétition de ces étapes dans la procédure par bissection de [Vostrikova, 1981] conduit à la segmentation pour plusieurs sauts de moyenne.

Certaines méthodes sont dédiées à des cas particuliers. Le test développé dans [Ramanayake et Gupta, 2003] est destiné à la détection d'une rupture épidémique dans le paramètre d'une distribution exponentielle. [Frick *et al.*, 2014] proposent une approche multi-échelle pour une distribution exponentielle à l'aide de l'algorithme SMUCE. Il est basé sur le calcul d'estimateurs locaux du maximum de vraisemblance, pour différentes échelles.

L'algorithme Screening and Ranking (SaRa), présenté dans [Niu et Zhang, 2012], repose sur le calcul d'une fonction de diagnostic locale $D(i,d)$, définie sur un voisinage d autour du point i étudié, pour détecter des sauts de moyenne dans des données normales, de variance constante. Cette fonction peut être vue comme un estimateur de la probabilité qu'il y ait une rupture dans le voisinage de X_i, et est définie comme la moyenne pondérée autour de

X_i :

$$D(i,d) = \sum_{j=1}^{n} w_j(i) x_j \text{ avec } w_j(i) = 0 \text{ quand } |j-i| > d. \tag{1.36}$$

On se ramène à la différence entre les moyennes empiriques locales avec les poids

$$w_j(i) = \begin{cases} \frac{1}{d} & \text{si } i-d+1 \leq j \leq i, \\ -\frac{1}{d} & \text{si } i+1 \leq j \leq i+d, \\ 0 & \text{sinon .} \end{cases} \tag{1.37}$$

Une autre fonction est proposée dans [Gijbels *et al.*, 1999], interprétée comme un estimateur local de la fonction dérivée du signal. Le terme $|D(i,d)|$ est maximisé lorsque X_i est une rupture. Pour détecter tous les sauts de moyennes, la fonction (1.36) est calculée en chaque point lors de l'étape dite de *screening*, et les maxima locaux sont classés lors de l'étape dite de *ranking*. Pour chaque observation X_i dans le voisinage défini par $[i-d+1,i+d]$, si la fonction $|D(i,d)|$ est localement maximale en i, i est un maximum d-local. Le vecteur $\hat{\boldsymbol{T}} = (\hat{\tau}_1,\hat{\tau}_2,\ldots,\hat{\tau}_K)$ des positions des ruptures est alors estimé par l'application d'un seuil λ tel que

$$|D(\hat{\boldsymbol{T}},d)| > \lambda. \tag{1.38}$$

Le vecteur $\boldsymbol{T}$ peut aussi être estimé par application du critère d'information bayésien (1.26). Cette méthode repose principalement la contrainte de parcimonie, c'est-à-dire sur l'hypothèse que le nombre de ruptures est petit devant le nombre de mesures et que ces événements ne sont pas trop proches. L'algorithme SaRa s'avère efficace pour les séries temporelles de grande dimension, avec une complexité en $\mathcal{O}(n)$. Des résultats théoriques sont donnés dans [Hao *et al.*, 2013], notamment sur la détection et la localisation de ruptures multiples.

D'autres statistiques de test construites à partir du rapport de vraisemblance pour les lois exponentielle, gamma, normale, de Poisson ou binomiale sont également proposées pour ces approches [Hawkins, 2001, Lehmann et Romano, 2006, Chen et Gupta, 2011]. Beaucoup de modèles paramétriques sont développés dans le cadre de l'inférence bayésienne. Les résultats théoriques obtenus pour les méthodes paramétriques ont été établis pour une famille de distributions, souvent pour la loi normale. Lorsque les hypothèses faites sur les données sont vérifiées, ces méthodes conduisent à de bons résultats. La dépendance au modèle des données, parfois mal spécifié, rend toutefois ces approches peu généralisables.

Méthodes de type LASSO

Parmi les méthodes construites sur l'hypothèse que les observations suivent la loi normale, l'approche LASSO et ses variantes sont communément rencontrées. Le principe consiste à approcher le signal par une fonction β. Dans le cas qui nous intéresse, la série temporelle $\boldsymbol{X}$ est vue comme une fonction constante par morceaux de $K+1$ segments de coefficients $\mu_1,\ldots,\mu_{K+1}$, contaminée par un bruit ϵ de moyenne nulle :

$$X_i = \mu_k + \epsilon_i, \quad \tau_{k-1} \leq i \leq \tau_k, \quad 1 \leq k \leq K+1. \tag{1.39}$$

Les coefficients de la fonction à estimer sont notés $\beta = (\beta_1, \ldots, \beta_n)$. Le problème de régression s'écrit généralement sous la forme d'une minimisation d'un critère de moindres carrés :

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} (X_i - \beta_i)^2, \tag{1.40}$$

cependant si la solution n'est pas constante par morceaux, il sera difficile de déterminer avec précision les sauts de moyenne significatifs. Afin de renforcer cette caractéristique, une pénalisation de la variation totale est ajoutée [Rudin *et al.*, 1992]. Le problème (1.40) devient un problème de régularisation :

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} (X_i - \beta_i)^2 + \lambda \sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i|. \tag{1.41}$$

La différence entre les coefficients successifs, notée $\Delta_i = \beta_{i+1} - \beta_i$, est pénalisée par la norme ℓ_1, qui permet de sélectionner les différences les plus significatives en annulant certains termes Δ_i. Cette formulation est plus adaptée que la norme ℓ_2 pour apporter une contrainte de parcimonie sur les Δ_i, et est préférée à la norme ℓ_0 pour faciliter la résolution. Le paramètre de régularisation λ contrôle la parcimonie des différences Δ_i, c'est-à-dire l'amplitude des sauts. Lorsque λ est nul, l'estimation $\hat{\beta}$ est la solution du problème des moindres carrés (1.40), et lorsqu'il est grand, le nombre de segments de $\hat{\beta}$ est faible.

Ce problème d'optimisation convexe peut être résolu efficacement par la méthode LASSO (en anglais Least Absolute Shrinkage and Selection Operator), présentée dans [Tibshirani, 1996]. L'expression (1.41) correspond au cas particulier du *fused* LASSO à une dimension [Tibshirani *et al.*, 2005]. On trouve parfois une pénalisation supplémentaire du nombre de valeurs prises par les coefficients de β, le problème est alors formulé de la façon suivante :

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} (X_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i| + \lambda_2 \sum_{i=1}^{n} |\beta_i|, \tag{1.42}$$

que l'on appelle le *sparse fused* LASSO.

Le problème (1.41) peut être exprimé par un problème dual équivalent. La résolution peut se faire par la méthode de *Least Angle Regression* (LAR) [Efron *et al.*, 2004] ou par *Alternating Direction Method of Multipliers* (ADMM) [Boyd *et al.*, 2011]. Elle fait intervenir l'opérateur de seuillage doux, qui introduit un biais dans les estimateurs de plus grande valeur. Une manière de le corriger est par exemple de pondérer les termes de la norme ℓ_1 par des poids adaptatifs [Zou, 2006]. On notera que ce biais sur l'amplitude des sauts de moyennes n'est pas gênant si l'objectif est simplement de localiser les changements.

L'application de l'algorithme LASSO pour la détection de ruptures multiples est discuté dans [Harchaoui et Lévy-Leduc, 2008], en particulier la question de l'estimation du nombre de ruptures, qui est contrôlée par le paramètre λ. Les auteurs remarquent en effet que la méthode a tendance à ajouter des sauts de moyenne à tort, bien que les vrais soient correctement estimés. L'algorithme Cachalot (CAtching CHAnge-points with LassO) est proposé pour effectuer une sélection du nombre $\hat{K}$ de ruptures a posteriori dans une procédure de programmation dynamique. La consistance de l'estimateur (1.41), dit également de

moindre carrés et variation totale, est montrée dans [Harchaoui et Lévy-Leduc, 2010] pour l'approximation du signal. En revanche on ne parvient à de tels résultats pour l'estimation des ruptures que sous certaines conditions. Ce genre de méthodes avec une pénalisation de la variation totale s'applique par exemple à la détection de ruptures et la segmentation [Harchaoui et Lévy-Leduc, 2008], le débruitage de signal ou d'images [Tibshirani, 2011], ou encore pour l'estimation de coefficients dans un processus auto-régressif [Angelosante et Giannakis, 2012].

Les résultats théoriques associés à l'algorithme LASSO ont été établis pour des erreurs ϵ_i centrées et distribuées normalement. En présence de bruit à queue lourde, qui introduit des valeurs aberrantes dans les observations, l'approche paramétrique LASSO a tendance a sur-segmenter le signal. En effet, le critère de moindres carrés dans le problème (1.41) est sensible aux fortes valeurs de $\boldsymbol{X}$. Pour que le problème soit robuste à ce genre de phénomène, on peut remplacer ce critère, équivalent à l'application de la norme ℓ_2, par la norme ℓ_1, et ainsi contraindre la solution sur la parcimonie des résidus. Dans l'article [Aravkin *et al.*, 2013], les auteurs présentent un ensemble de méthodes reposant sur des fonctions à support quadratique, dont font partie les normes ℓ_1 et ℓ_2, ainsi que leur mise en œuvre dans une série de problèmes d'optimisation. Le LASSO robuste y est introduit. Sa formulation avec la norme ℓ_1 est la suivante :

$$\min_{\beta\in\mathbb{R}^n} \sum_{i=1}^{n} |X_i - \beta_i| + \lambda \sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i|. \tag{1.43}$$

Les auteurs de [Boyd *et al.*, 2011] proposent de résoudre le problème dual par ADMM ou par la méthode du point intérieur, plus performante.

Pour approcher le signal par une fonction, une autre méthode est celle de la régression quantile, qui consiste à contraindre la solution afin que ses quantiles correspondent à ceux des données. Cette méthode est intéressante par exemple lorsqu'on ne peut pas supposer que les données suivent la distribution gaussienne. Dans notre problème de détection de rupture, on cherche à délimiter les portions du signal de moyenne ou de médiane constante. En choisissant pour quantile la médiane, on se ramène à la méthode de type LASSO robuste [Eilers et De Menezes, 2005]. La préférence pour la norme ℓ_1 est justifiée dans [Portnoy *et al.*, 1997].

La méthode paramétrique LASSO peut être interprétée d'un point de vue bayésien [Tibshirani, 1996]. En effet, en écrivant le problème (1.41) avec $AB = \beta$, $A \in \mathbb{R}^{n\times(n-1)}$ et $B \in \mathbb{R}^{n-1}$, les données sont générées selon la loi normale $\mathcal{N}(AB,I_n)$, où I_n est la matrice identité de dimension $n\times n$. On choisit de modéliser les coefficients b_i du vecteur B par la loi de Laplace de paramètre σ^2, afin que les différences $\beta_{i+1} - \beta_i$ soient fortement concentrées autour de 0. L'estimateur de (1.41) correspond alors à l'expression d'un mode de la densité de probabilité a posteriori. [Park et Casella, 2008] développent ainsi un modèle bayésien à partir de l'algorithme LASSO, où λ est un hyperparamètre. Ce type d'approche constitue un ensemble de méthodes efficaces pour l'approximation d'un signal, pouvant notamment être employées pour la détection de ruptures. Cependant, comme toute méthode paramétrique, elles sont limitées par la dépendance au modèle des données, et le paramètre de régularisation λ doit être adapté à chaque application.

Cadre de l'inférence bayésienne

L'inférence bayésienne consiste à proposer une description des observations en apposant un modèle probabiliste sur le vecteur des données $\boldsymbol{X} = (X_1, \ldots, X_n)$. Sa distribution est reliée au vecteur des paramètres $\Theta = (\theta_1, \ldots, \theta_m)$, de dimension finie, à travers la fonction de densité $f(\boldsymbol{X}|\Theta)$. Les paramètres inconnus sont représentés par des variables aléatoires, dont les densités de probabilités sont soit déterminées a priori à partir des informations dont on dispose, soit exprimées à l'aide de lois non informatives. Le cœur de l'inférence réside dans le théorème de Bayes, issu des travaux de Bayes et de Laplace au 18$^{\text{e}}$ siècle :

$$f(\Theta|\boldsymbol{X}) = \frac{L(\boldsymbol{X}|\Theta)f(\Theta)}{\int L(\boldsymbol{X}|\Theta)f(\Theta)d\Theta} \tag{1.44}$$

où $f(\Theta|\boldsymbol{X})$ est l'expression de la densité de probabilité a posteriori de la variable Θ, inconnue, par rapport à $\boldsymbol{X}$, $L(\boldsymbol{X}|\Theta)$ est la fonction de vraisemblance des données par rapport aux paramètres de Θ, et $f(\Theta)$ est la densité de probabilité jointe des θ_i, $1 \leq i \leq m$. Cette approche donne accès à toute la distribution des paramètres de Θ par rapport aux données, ou à une estimation de Θ avec un intervalle de crédibilité, pour le modèle choisi. Cependant l'objectif est souvent de déterminer une valeur de Θ, en général celle qui maximise la densité $f(\Theta|\boldsymbol{X})$. Dans ce cas la relation de proportionnalité

$$f(\Theta|\boldsymbol{X}) \propto L(\boldsymbol{X}|\Theta)f(\Theta) \tag{1.45}$$

est souvent employée.

De nombreuses méthodes ont une interprétation bayésienne. Pour construire un modèle, il faut définir les distributions des observations $\boldsymbol{X}$ en fonction de Θ, ainsi que la distribution a priori de Θ, notée $f(\Theta)$. Ce choix n'est pas trivial, d'une part parce qu'on ne dispose pas forcément de beaucoup d'information a priori, et d'autre part parce que les distributions choisies introduisent des hyperparamètres qu'il est nécessaire d'estimer ou de marginaliser. L'article [Carlin *et al.*, 1992] traite de la question des modèles bayésiens hiérarchiques, où plusieurs hyperparamètres ont été introduits à l'aide de densités a priori. L'échantillonneur de Gibbs est proposé pour simuler ces termes conditionnellement aux autres variables et ainsi éviter le calcul des densités de probabilités marginales. Des exemples sont fournis, pour la détection d'une rupture. Une fois l'expression de la densité a posteriori dans (1.45) obtenue, l'estimateur MAP de Θ est soit calculé analytiquement, par maximisation du terme de droite dans (1.45), soit approché numériquement. L'estimation du MAP peut se calculer de nombreuses façons (voir [Brooks *et al.*, 2011]), en échantillonnant la variable Θ jusqu'à ce que l'on considère que l'algorithme ait suffisamment convergé vers le maximum de la distribution, souvent pour un temps de calcul conséquent, mais pour une bonne performance de détection. Un exemple de procédure par méthode de Monte Carlo par Chaîne de Markov avec un échantillonneur de Gibbs est détaillé dans la partie 3.3. Pour aller plus loin, le lecteur peut se référer au livre [Robert, 2006], dédié à l'inférence bayésienne paramétrique.

De nombreuses méthodes ont été développées dans un cadre bayésien pour la détection de ruptures. La première formulation d'un modèle comportant plusieurs ruptures dans

la moyenne est attribuée à [Chernoff et Zacks, 1964]. Un modèle bayésien est construit de diverses façons, selon la nature des données $\boldsymbol{X}$, normales, exponentielles, continues ou discrètes, et selon ce que représente le paramètre Θ, comme la longueur d'un segment, sa moyenne, sa variance ou un état associé aux X_i. Ainsi, dans [Lavielle et Lebarbier, 2001, Dobigeon *et al.*, 2007a], le paramètre à estimer est le vecteur $\boldsymbol{R}$ des variables indicatrices de la présence d'une rupture, dont les coefficients sont

$$R_i = \begin{cases} 1 & \text{si } X_i \text{ est une rupture,} \\ 0 & \text{sinon,} \end{cases} \tag{1.46}$$

pour tout $1 < i < n$, et, par convention, $R_1 = R_n = 1$. Détecter les événements dans le signal $\boldsymbol{X}$ revient donc à inférer $\boldsymbol{R}$. Le modèle présenté dans [Lavielle et Lebarbier, 2001] est Bernoulli-gaussien : les données X_i suivent une loi normale et sont supposées i.i.d dans un même segment, tandis que les R_i sont i.i.d. selon la loi de Bernoulli de paramètre q :

$$f(R|q) = \prod_{i=1}^{n} q^{R_i}(1-q)^{1-R_i}. \tag{1.47}$$

Selon les applications, d'autres distributions peuvent être choisies pour modéliser la loi des variables aléatoires X_i, comme par exemple, dans un contexte multivarié, la distribution log-normale pour des mesures du vent [Dobigeon et Tourneret, 2009], ou la distribution de Poisson pour le comptage de photons de données astronomiques [Dobigeon *et al.*, 2007b].

Plutôt que d'inférer le vecteur $\boldsymbol{R}$ des positions des ruptures, une autre approche consiste à introduire une relation de récursion entre les segments du signal $\boldsymbol{X}$ grâce à laquelle on parvient à localiser les segments, et à en déduire la position des ruptures. La méthode de [Fearnhead, 2006] introduit ainsi la loi $B(t,s)$ des variables $X_t,\ldots,X_s$ $(s \geq t)$, appartenant au même segment, et la probabilité $Q(t)$ que la variable X_{t-1} soit une rupture :

$$\begin{aligned} B(t,s) &= P(X_t,\ldots,X_s|\ t \text{ et } s \text{ sont dans le même segment}) \\ Q(t) &= P(X_t,\ldots,X_n|\ X_{t-1} \text{ est une rupture}). \end{aligned} \tag{1.48}$$

Elle repose sur l'hypothèse que les paramètres θ_k des segments $1 \leq k \leq K+1$ sont indépendants les uns des autres. Le modèle fait également intervenir la distribution qui modélise la durée de l'intervalle entre deux ruptures, la loi binomiale négative est choisie a priori. Ainsi les positions des ruptures $\tau_1,\ldots,\tau_K$ sont estimées successivement et directement en partant de l'instant $i = 1$. La stratégie récursive est reprise dans [Fearnhead et Liu, 2007] pour une application en ligne, et plus récemment dans [Bardwell et Fearnhead, 2014], où l'algorithme BARD présenté permet de traiter des séries temporelles multivariées. Ces algorithmes sont adaptés selon les distributions des données, par exemple pour la loi normale et pour la loi de Student.

Les exemples précédents sont des modèles paramétriques, nous avons notamment constaté que l'efficacité de l'algorithme LASSO repose sur l'adéquation des observations avec la loi normale, si bien que quand cette hypothèse n'est pas respectée la méthode doit être adaptée. D'autres méthodes ont été développées d'un point de vue non paramétrique, y compris

dans un cadre bayésien, et permettent ainsi de s'affranchir de la dépendance au modèle. Cette alternative est intéressante en l'absence d'information a priori sur le système étudié, en particulier lorsque la normalité des données n'est pas garantie, ou bien quand le modèle doit être le plus généraliste possible, pour s'adapter à des lois de probabilités variées.

Mesure de distances

Afin de s'affranchir de la spécification d'un modèle paramétrique des observations, certains auteurs développent des approches construites à partir de mesures de distances entre les observations, en s'inspirant des méthodes de classification non supervisées. La méthode E-Divisive, présentée dans [Matteson et James, 2014], repose sur la distance euclidienne, avec comme seule condition sur les distributions que le moment d'ordre $a \in [0,2]$ existe et que les observations soient i.i.d. dans un même segment. Une mesure de divergence empirique compare les deux portions du signal $S_k = (X_{\tau_{k-1}+1}, \ldots, X_{\tau_k})$ et $S_{k+1} = (X_{\tau_k+1}, \ldots, X_{\tau_{k+1}})$ de longueurs m et n respectivement, pour tester l'existence de la rupture au point τ_k :

$$\mathcal{E}(S_k,S_{k+1};a) = \frac{2}{mn} \sum_{i=\tau_{k-1}+1}^{\tau_k} \sum_{j=\tau_k+1}^{\tau_{k+1}} |X_i - X_j|^a - \binom{n}{2}^{-1} \sum_{\tau_{k-1}+1 \leq i \leq j \leq \tau_k} |X_i - X_j|^a - \binom{m}{2}^{-1} \sum_{\tau_k+1 \leq i \leq j \leq \tau_{k+1}} |X_i - X_j|^a. \quad (1.49)$$

La statistique du test est $\mathcal{Q}(S_k,S_{k+1},a) = \frac{mn}{m+n}\mathcal{E}(S_k,S_{k+1};a)$, et sa convergence en distribution sous l'hypothèse H_0 et sous H_1 est connue. La position τ_k est estimée en maximisant la statistique locale, calculée récursivement. La stratégie de la segmentation binaire est appliquée pour détecter plusieurs ruptures. Un test, employé comme un critère d'arrêt, détermine si une rupture est significative, à partir de permutations aléatoires des segments estimés. Les auteurs montrent que les K premières ruptures sont correctement détectées, mais que des ruptures supplémentaires sont également estimées à tort.

Une approche originale a été récemment proposée dans [Chen *et al.*, 2015] : il s'agit d'utiliser la représentation des données par un graphe pour établir la statistique de test. Le graphe G est construit à partir de mesures de similarités entre les observations. La méthode est destinée à la détection d'une seule rupture dans des données multivariées, ou à une rupture épidémique, la détection de plus de deux ruptures pouvant être réalisée par bissection. Son intérêt réside notamment dans sa capacité à analyser des séries temporelles de dimension n par m, où n peut être petit devant m. Les observations composent les nœuds du graphe, l'idée étant que des variables générées selon la même distribution sont proches les unes des autres. Les arêtes sont établies par exemple par la méthode du plus proche voisin ou de l'arbre couvrant minimal, l'important étant que le graphe permette de séparer les groupes d'observations de distributions différentes. La statistique $R_G(i)$ mesure alors le nombre d'arêtes connectant une observation avant i à une observation après i. Sous l'hypothèse H_0 que X_i n'est pas une rupture, $R_G(i)$ est petit. La statistique de test Z_G est

une version standardisée de R_G, dont le maximum donne la position de la rupture τ. Ces deux méthodes reposent sur une mesure de divergence entre les observations. Elle n'est cependant pas toujours évidente à établir.

Méthodes à noyaux

Les méthodes à noyaux proposent une alternative intéressante : une transformation φ est appliquée aux données de l'espace d'entrée $\mathcal{X}$ vers un espace $\mathcal{H}$ de plus grande dimension, où une mesure de similarité entre les images des observations est calculée pour détecter le changement. Ces méthodes permettent de traiter des données de grandes dimensions. Des données structurées, par exemple des graphes ou du texte, peuvent également être analysées. L'article [Harchaoui *et al.*, 2013] présente des méthodes dédiées aux tests d'hypothèses, dans un cadre général où les données sont multivariées. Dans le cas de la détection de rupture à la position τ, le problème est formulé comme un test d'homogénéité entre deux segments $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ de dimensions n_1 et n_2, dont les lois de probabilités sont notées F_1 et F_2. Les hypothèses testées sont :

$$\begin{array}{ll} H_0 : & F_1 = F_2 \\ H_1 : & F_1 \neq F_2 \end{array} \tag{1.50}$$

Pour détecter une rupture à une position indéterminée, [Harchaoui *et al.*, 2009] applique ce test sur une fenêtre glissante. La position réelle de la rupture est déterminée en maximisant une mesure d'hétérogénéité.

L'espace image $\mathcal{H}$ de la transformation φ, où sont testées les hypothèses (1.50), est appelé espace de Hilbert à noyau reproduisant, en anglais Reproducing Kernel Hilbert Space (RKHS). On note $\langle .,. \rangle_{\mathcal{H}}$ son produit scalaire. La méthode repose sur le noyau $h(X,y) = \langle \varphi(X), \varphi(y) \rangle_{\mathcal{H}}$: pour comparer X et y on traite les projections des observations dans l'espace $\mathcal{H}$, où elles sont linéairement séparables, au lieu de traiter directement X et y dans $\mathcal{X}$. Cette opération est appelée l'*astuce du noyau*. La condition à respecter est que la matrice de Gram H, dont les coefficients sont les $h(X_i,y_j)$, est semi-définie positive. Cette condition est respectée dès lors que le noyau est symétrique et semi-défini positif. Deux exemples simples de tels noyaux sont le noyau linéaire $h(X,y) = X'y$ et le noyau gaussien $h(X,y) = \exp(-\frac{\|X-y\|^2}{\sigma^2})$.

Deux opérateurs sont employés pour représenter les lois de probabilités F dans $\mathcal{H}$: la moyenne μ_F et l'opérateur de covariance Σ_F, tels que

$$\langle \mu_F, f \rangle_{\mathcal{H}} = E(f(\boldsymbol{X})), \forall f \in \mathcal{H} \tag{1.51}$$

$$\langle f, \Sigma_F g \rangle_{\mathcal{H}} = Cov(f(\boldsymbol{X}),g(\boldsymbol{X})), \forall f,g \in \mathcal{H} \tag{1.52}$$

pour une variable $\boldsymbol{X}$ de $\mathcal{X}$. Leurs équivalents empiriques $\hat{\mu}$ et $\hat{\Sigma}$ sont donnés par

$$\langle \hat{\mu}, f \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \tag{1.53}$$

$$\langle f, \hat{\Sigma} g \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \langle \hat{\mu}, f \rangle_{\mathcal{H}})(g(X_i) - \langle \hat{\mu}, g \rangle_{\mathcal{H}}) \tag{1.54}$$

pour les variables $X_1, \ldots, X_n$ i.i.d. selon la loi F.

Une catégorie de méthodes, appelée divergence maximale moyenne (en anglais Maximum Mean Discrepancy, ou MMD) et présentée dans [Gretton *et al.*, 2006], renvoie une mesure de similarité entre les populations $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ basée sur les moyennes :

$$T^{MMD}_{n_1,n2} = (n_1 + n_2) \|\hat{\mu}_1 - \hat{\mu}_2\|^2_{\mathcal{H}} \tag{1.55}$$

Cette mesure est la déviation maximale de l'espérance d'une fonction φ, évaluée sur chaque variable aléatoire. La fonction est choisie de telle sorte que la mesure MMD soit nulle si et seulement si $F_1 = F_2$, tout en étant applicable avec une complexité raisonnable. La distribution nulle asymptotique est une somme infinie de variables aléatoires générées selon la loi du χ^2. Comme le seuil à appliquer lors du test d'homogénéité est un quantile de cette distribution, il est par conséquent difficile de déterminer la valeur de ce seuil.

Une variante du MMD est l'analyse du noyau du rapport du discriminant de Fisher (Kernel Fisher Discriminant Analysis, KFDA), introduite dans [Moulines *et al.*, 2008]. La statistique est construite à partir de l'expression (1.55), avec l'introduction du terme de covariance et d'une normalisation :

$$T^{KFDA}_{n,\tau,\gamma} = \frac{n_1 n_2}{n} \left\| (\hat{\Sigma}_W + \gamma_n I)^{1/2} (\hat{\mu}_1 - \hat{\mu}_2) \right\|^2_{\mathcal{H}} \tag{1.56}$$

où $n = n_1 + n_2$, et où $\Sigma_W = \frac{n_1}{n} \hat{\Sigma}_1 + \frac{n_2}{n} \hat{\Sigma}_2$ est l'opérateur empirique de covariance inter-classes. Pour détecter une rupture à une position inconnue, l'algorithme de [Harchaoui *et al.*, 2009] parcourt le signal avec une fenêtre glissante de taille n, et le terme (1.56) est calculé à chaque mouvement. La position du changement est celle qui maximise la statistique. Les auteurs déterminent la distribution asymptotique sous H_0 et démontrent la consistance du test sous H_1 lorsque n tend vers l'infini.

Le test du noyau du rapport des densités (en anglais Kernel Density-Ratio, KDR) repose sur le rapport des densités de probabilités, pris comme estimateur de la f-divergence [Kanamori *et al.*, 2012]. Un estimateur de ce ratio $r(X; \theta)$, selon le paramètre θ, est défini par

$$\hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i \varphi(X_i), \tag{1.57}$$

avec $\theta_1, \ldots, \theta_n \geq 0$. La statistique de test est construite à partir de $\hat{\nu}_n$

$$T^{KDR}_n = \frac{1}{n} \sum_{i=1}^{n} \log(\langle \hat{\nu}_n, \varphi(X_i) \rangle_{\mathcal{H}}) \tag{1.58}$$

et correspond dans certains cas à un estimateur de la divergence de Kullback-Leibler. Il est à noter que les deux tests précédents peuvent également être vus comme des estimateurs non paramétriques de divergences classiques, comme la divergence L^2 pour le MMD, celle du χ^2 pour le KFDA.

Une autre méthode à noyau est celle des machines à vecteurs de support (en anglais Support Vector Machine, SVM) à une classe. Elle consiste à calculer l'hyperplan qui définit la région de l'espace associée aux échantillons de $\boldsymbol{X}_1$ et celle associée aux échantillons de $\boldsymbol{X}_2$, afin d'obtenir une mesure de distance entre ces deux régions. La méthode de détection de changement par noyau (en anglais Kernel Change Detection, ou KCD) est présentée dans [Desobry *et al.*, 2005], avec un noyau gaussien et pour une application à la détection séquentielle. Pour cet algorithme le vecteur $\boldsymbol{X}_1$ est constitué des n_1 observations antérieures au point testé τ, et le vecteur $\boldsymbol{X}_2$ est constitué des n_2 observations suivantes, X_τ inclus. L'espace $\mathcal{H}$ est normalisé, de telle sorte que $\varphi(\mathcal{X})$ soit un sous ensemble de l'hypersphère unitaire S, centrée sur l'origine de $\mathcal{H}$. L'image de $\boldsymbol{X}_1$ dans $\mathcal{H}$ est le vecteur d'apprentissage : il permet de construire l'hyperplan W_1 par la résolution d'un problème d'optimisation. Cet hyperplan, paramétrisé par (w_1,ρ_1), sépare les échantillons du centre de S avec la marge ρ_1, sans tenir compte des éventuelles valeurs aberrantes. La méthode est en effet insensible à ce genre de perturbation, en fixant un seuil ν. De la même façon on obtient l'hyperplan W_2, paramétrisé par (w_2,ρ_2), qui sépare les images des observations de $\boldsymbol{X}_2$ du centre de l'hypersphère. Une différence entre F_1 et F_2 doit se traduire dans $\mathcal{H}$ par une répartition des images des observations dans des régions distinctes. La mesure de divergence suivante tient compte de l'écart entre les hyperplans ainsi que des dispersions des distributions :

$$T^{KCD}_{n_1,n_2} = \frac{\widehat{c_1c_2}}{\widehat{c_1p_1} + \widehat{c_2p_2}} \tag{1.59}$$

où c_i est le point d'intersection de S avec le vecteur prolongé w_i, et où p_i est le point d'intersection de W_i avec S dans le plan contenant les vecteurs w_1 et w_2, et inclus dans l'arc $\widehat{c_1c_2}$ (voir la figure 3 de [Desobry *et al.*, 2005]). Les longueurs sont calculées à partir de produits scalaires et de normes de l'espace $\mathcal{H}$ sur les paramètres des hyperplans, ainsi que sur la matrice du noyau et les multiplicateurs de Lagrange issus de l'étape d'estimation des hyperplans. Cette statistique est calculée à chaque translation d'une fenêtre glissante, afin d'estimer la position de la rupture. Une limite de cette méthode est l'absence de résultat asymptotique. Le seuil à appliquer pour accepter ou rejeter H_0 doit être fixé. Toutefois dans le cas gaussien, les auteurs établissent une correspondance avec la méthode KFDA, qui peut être exploitée pour décrire le comportement de (1.59).

Les méthodes à noyau sont plus généralisables que les solutions paramétriques, où les distributions des données sont définies explicitement, mais leurs performances dépendent du choix du noyau $h(.,.)$. D'autres méthodes s'inspirent de tests statistiques non paramétriques, connus pour leur robustesse aux données non normales, comme ceux présentés au chapitre 2.

Méthodes basées sur un test de rang

L'intérêt des tests non paramétriques de rangs réside principalement dans le peu d'hypothèses faites sur les données. Le livre [Lehmann et D'Abrera, 1975] rassemble les différentes méthodes basées sur ces tests et les résultats théoriques qui y sont associés. Le plus connu

d'entre eux est le test de rang de Wilcoxon-Mann-Whitney (WMW), ou test de la somme des rangs de Wilcoxon, qui établit si une population Y a tendance à avoir des valeurs plus grandes qu'une autre population Z. Il s'applique sur des distributions inconnues, contrairement au test t de Student qui suppose que les observations sont normalement distribuées, voir les parties 2.2 et 2.1.

Ce test d'homogénéité est sensible aux différences entre les rangs moyens des deux populations, ce qui revient à tester les médianes dans certains cas. Le rang R_i de l'observation X_i dans la population globale $(x_1, \ldots, x_n)$ est défini par

$$R_i = \sum_{j=1}^{n} \mathbb{1}_{\{x_j \leq X_i\}}, \tag{1.60}$$

où la fonction $\mathbb{1}_{\{.\}}$ est la fonction indicatrice

$$\mathbb{1}_{\{x\}} = \begin{cases} 1 & \text{si la condition } x \text{ est vraie,} \\ 0 & \text{sinon.} \end{cases} \tag{1.61}$$

Pour tester l'existence d'une rupture à l'instant τ, la statistique du test de WMW s'écrit

$$U_\tau = \sum_{i=1}^{\tau} \sum_{j=\tau+1}^{n} \mathbb{1}_{\{x_i \leq x_j\}}\,. \tag{1.62}$$

L'hypothèse nulle que les observations du premier et du deuxième segment, délimités par τ, suivent des distributions de même médiane est rejetée pour de grandes valeurs de U_τ. Ainsi, pour déterminer la position d'une unique rupture, [Pettitt, 1979] introduit la statistique

$$T_\tau = \max_{1 \leq \tau \leq n} |U_\tau|. \tag{1.63}$$

L'auteur développe un test applicable sur des données suivant une distribution discrète, ainsi qu'une version approchée pour les distributions continues. Le test MultiRank de [Lung-Yut-Fong *et al.*, 2011a] a également été proposé pour la détection d'un seul changement, cette fois dans des données multivariées. La signification statistique de ce test est donnée dans [Lung-Yut-Fong *et al.*, 2011c], il peut être utilisé pour la détection de plusieurs ruptures en l'appliquant sur différentes fenêtres temporelles, pourvu que les changements ne soient pas trop rapprochés.

L'algorithme dynMKW a également été élaboré pour des séries temporelles multivariées. Il s'inspire du test de Kruskal-Wallis [Kruskal et Wallis, 1952], un test de rang pour comparer plusieurs échantillons. Pour tester l'existence de ruptures aux positions $\tau_1, \ldots, \tau_K$ dans un signal univarié, la statistique de test est

$$T(\tau_1, \ldots, \tau_K) = \frac{12}{n^2} \sum_{k=0}^{K} (\tau_{k+1} - \tau_k)\Big(\overline{R}_k - \frac{n}{2}\Big)^2, \tag{1.64}$$

où $\overline{R}_k = (\tau_{k+1} - \tau_k)^{-1} \sum_{i=\tau_k+1}^{\tau_{k+1}} R_i$, avec la convention $\tau_0 = 1$ et $\tau_{K+1} = n$. Le nombre K^* optimal de changements est déterminé a posteriori, à l'aide d'une heuristique de pente sur

la valeur de la statistique en fonction du nombre de ruptures. L'intérêt de l'approche non paramétrique est illustrée par les bonnes performances obtenues sur des données contenant des valeurs aberrantes, par rapport à deux autres méthodes reposant sur l'hypothèse que les données sont gaussiennes. Outre l'absence de condition sur la nature des distributions des variables, un autre avantage de ces statistiques de rang est de pouvoir traiter des données censurées ou manquantes, en encadrant les observations par des valeurs limites lors du calcul des rangs [Gombay et Liu, 2000, Lung-Yut-Fong *et al.*, 2011a]. La statistique (1.64) permet de réaliser certaines opérations récursivement. Un algorithme de programmation dynamique est donc proposé dans [Lung-Yut-Fong *et al.*, 2011b] pour un nombre d'événements K donné, au lieu de maximiser directement la statistique par rapport à T, opération de complexité combinatoire. Une mesure de divergence n'est pas toujours évidente à établir directement à partir des observations. La complexité du problème est un point important de la détection de ruptures multiples qui peut s'avérer rapidement limitant, en particulier lorsque les données sont multivariées.

1.1.4 Séries temporelles multivariées

Dans une optique d'analyse d'un système complet, on dispose souvent de mesures réalisées de manière synchronisées par plusieurs capteurs. La série temporelle qui en est extraite est donc multivariée, on la représente par la matrice $\boldsymbol{X}$ de taille $m \times n$, dont les vecteurs ligne $\boldsymbol{X}_{j,\bullet}$, $1 \leq j \leq m$, sont les séries temporelles univariées issues de chaque capteur de n points temporels. Une approche naïve pour la détection des ruptures dans $\boldsymbol{X}$ consiste à traiter chaque signal indépendamment des autres, par l'une des méthode univariées présentées précédemment, aboutissant à l'estimation de m vecteurs de segmentation $\boldsymbol{T}_m = (\tau_{m,1}, \ldots, \tau_{m,K_m})$. Cette stratégie se justifie d'une part lorsque les grandeurs mesurées par les capteurs sont indépendantes et que les changements qu'on y observe sont causés par des processus n'affectant pas les autres signaux, auquel cas un traitement conjoint des vecteurs $\boldsymbol{X}_{j,\bullet}$ n'a pas d'intérêt. Ce cas de figure est illustré par la figure 1.2 : les ruptures (en rouge) se produisent indépendamment d'un signal à l'autre, un traitement séparé des trois signaux composant la série temporelle est approprié. D'autre part, lorsque les séries temporelles sont hétérogènes, par exemple parce que les capteurs mesurent des grandeurs de natures différentes (vitesse, pression, puissance électrique, etc.), il peut être difficile de détecter efficacement les ruptures sur tous les signaux par la même méthode.

Segmentation unique de tous les signaux

Le cas multivarié le plus simple concerne les séries temporelles où les vecteurs $\boldsymbol{X}_{j,\bullet}$ partagent les mêmes caractéristiques statistiques et où les ruptures sont localisées aux mêmes positions sur chaque vecteur, comme l'illustre la figure 1.3. Le problème consiste à estimer la segmentation $\boldsymbol{T} = (\tau_1, \ldots, \tau_K)$ commune à tous les signaux, sous forme d'un vecteur de dimension m. Pour ce type de données, les méthodes bénéficient de l'apport d'informations dû à la dimension spatiale pour améliorer la détection des ruptures.

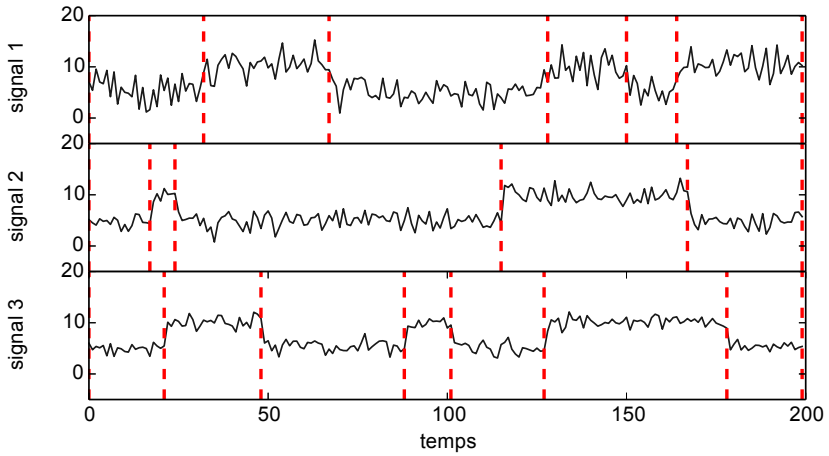


FIGURE 1.2 – Série temporelle multivariée où les ruptures (en rouge) apparaissent indépendamment les unes des autres et d'un signal à l'autre.

Le cas normal multivarié, qui étend les procédures basées sur le rapport de vraisemblance en univarié, est traité dans le chapitre 3 de [Chen et Gupta, 2011]. On suppose que les variances sont égales, et les valeurs des moyennes sont inconnues. La matrice de covariance Σ est introduite dans le modèle pour la détection d'une rupture à la position τ_k. Les estimateurs du maximum de vraisemblance du vecteur des moyennes $\boldsymbol{\mu}$ et de la matrice de covariance sont

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})', \tag{1.65}$$

où $\boldsymbol{X}_i$ est le vecteur colonne des observations à l'instant i sur tous les signaux. On définit alors le terme

$$W_{\tau_k} = \frac{1}{n-2}\Big(\sum_{i=1}^{\tau_k} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_1)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_1)' + \sum_{i=\tau_k+1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_2)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_2)'\Big) \tag{1.66}$$

où $\overline{\boldsymbol{X}}_k$ est le vecteur des moyennes empiriques des observations sur le k^{e} segment. La statistique du test T^2 de Hotelling est utilisée

$$T^2_{\tau_k} = Y'_{\tau_k} W^{-1}_{\tau_k} Y_{\tau_k} \tag{1.67}$$

où Y_{τ_k} est la différence standardisée entre les moyennes empiriques $Y_{\tau_k} = \sqrt{\frac{\tau_k(n-\tau_k)}{n}}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)$. La rupture est détectée à la position optimale $\tau_k^\star$ qui maximise la statistique (1.67), lorsque celle-ci dépasse un seuil c. Dans [Srivastava et Worsley, 1986], les auteurs proposent une approximation de la distribution nulle de la statistique

$$S_{\tau_k} = \frac{T^2_{\tau_k}}{n - 2 + T^2_{\tau_k}} \tag{1.68}$$

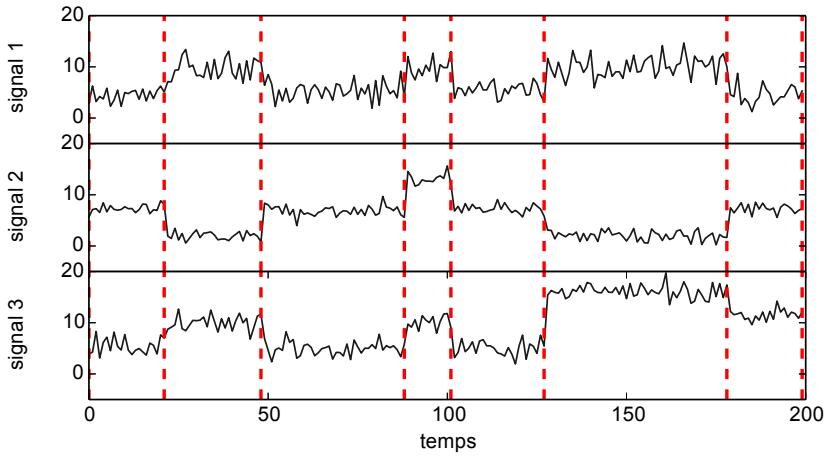


FIGURE 1.3 – Série temporelle multivariée où les ruptures (en rouge) apparaissent simultanément sur les trois signaux.

dont le maximum est atteint en $\tau_k^\star$, afin de déterminer le seuil c. Elle est obtenue à partir de l'inégalité de Bonferroni appliquée à l'événement $S_{\tau_k} > c$, et conduit à un test conservatif. Plusieurs ruptures peuvent être détectées avec la procédure de segmentation binaire de [Vostrikova, 1981]. Une procédure pour détecter des ruptures dans les moyennes et les variances d'une distribution normale multivariée est proposée dans [Chen et Gupta, 1995].

Parmi les solutions non paramétrique, les algorithmes construits sur le test de rang de WMW de [Lung-Yut-Fong *et al.*, 2011b,Lung-Yut-Fong *et al.*, 2011c,Lung-Yut-Fong *et al.*, 2011a], présentés dans la section 1.1.3, sont destinés à des séries temporelles multivariées. La statistique de test (1.64) dans le cas multivarié est donnée par

$$T(\tau_1,\ldots,\tau_K) = \frac{1}{n^2}\sum_{k=0}^{K}(\tau_{k+1}-\tau_k)\overline{R}_k'\hat{\Sigma}_n^{-1}\overline{R}_k, \tag{1.69}$$

où $\hat{\Sigma}_n$ est la matrice $m \times m$ dont l'élément (j,j') est

$$\hat{\Sigma}_{n,j,j'} = \frac{1}{n^2}\sum_{i=1}^{n}\Big(R_i^{(j)} - \frac{n}{2}\Big)\Big(R_i^{(j')} - \frac{n}{2}\Big), \tag{1.70}$$

et $\overline{R}_k = (\overline{R}_k^{(1)} - \frac{n}{2},\ldots,\overline{R}_k^{(m)} - \frac{n}{2})'$ est le vecteur colonne des sommes des rangs dans le segment k, pour chaque signal j :

$$\overline{R}_k^{(j)} = \frac{1}{\tau_{k+1}-\tau_k}\sum_{i=\tau_k+1}^{\tau_{k+1}} R_i^{(j)}. \tag{1.71}$$

Les méthodes comme celle de [Matteson et James, 2014] ou à noyau [Harchaoui *et al.*, 2013] se formulent directement dans le cas multivarié, où les distances entre les vecteurs se mesurent avec une norme. Dans ces approches, on suppose que les ruptures sont présentes

dans toutes les dimensions spatiales, et une unique segmentation $\hat{\boldsymbol{T}} = (\hat{\tau}_1, \ldots, \hat{\tau}_K)$ est estimée. Le traitement joint des m signaux est renforcé par l'augmentation de l'information et permet alors de détecter des événements de faibles amplitudes qui ne seraient pas obtenus par une analyse individuelle des vecteurs $\boldsymbol{X}_{j,\bullet}$, $1 \leq j \leq m$.

Ruptures communes à un sous-ensemble de signaux

Le traitement conjoint rend également possible la segmentation de la série temporelle multivariée aux instants où une rupture est susceptible de se produire, sans que l'événement affecte toutes les séries temporelles avec une probabilité de 1. La figure 1.4 représente une telle série temporelle, où les signaux 1 et 2 ont les mêmes ruptures, qui ont une probabilité de 0,5 de se produire dans le signal 3. L'objectif est alors l'estimation d'une segmentation de référence $\boldsymbol{T} = (\tau_1, \ldots, \tau_K)$, où des ruptures sont susceptibles de se produire. Ce type de segmentation est utile par exemple dans le traitement de données génomiques prises chez plusieurs individus (voir section 4.4.3) pour extraire les portions de l'ADN ayant une tendance sur ou sous-exprimées par rapport à un ADN de référence.

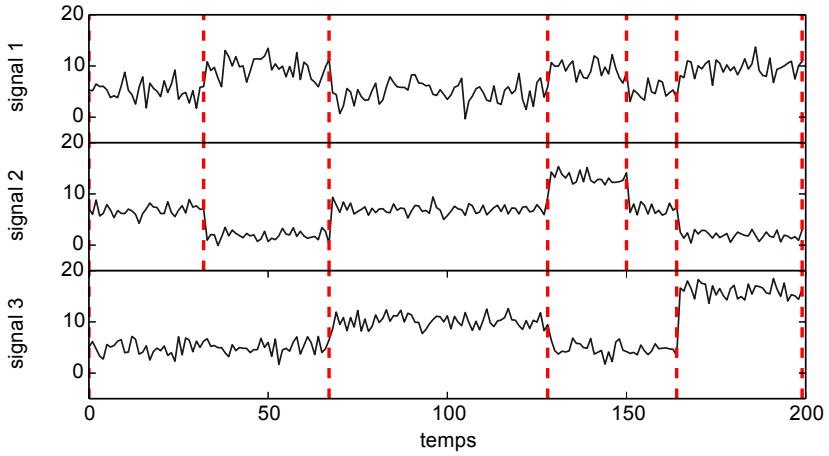


FIGURE 1.4 – Série temporelle multivariée où les ruptures (en rouge) apparaissent simultanément sur les signaux 1 et 2, et avec une probabilité de 0,5 sur le signal 3.

Une extension de la méthode du fused LASSO a été proposée dans [Vert et Bleakley, 2010], il s'agit du *group* fused LASSO. Cette fois le signal $\boldsymbol{Y} = \boldsymbol{X}'$ est approché par une fonction constante par morceaux sur les vecteurs colonnes, représentée par la matrice $\boldsymbol{B} \in \mathbb{R}^{n \times m}$. Le problème d'optimisation convexe à résoudre est maintenant

$$\min_{\boldsymbol{B} \in \mathbb{R}^{n \times m}} \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{B}\|_F^2 + \lambda \sum_{i=1}^{n-1} \frac{\|B_{i+1} - B_i\|_2}{d_i}, \tag{1.72}$$

où $\|.\|_F$ est la norme de Frobenius, B_i est la i^{e} ligne de $\boldsymbol{B}$, pour l'instant i, et d_i est un poids affecté à la position i afin de corriger des effets de bords. Comme avec le fused

LASSO de [Tibshirani *et al.*, 2005], le paramètre λ contrôle le nombre K de ruptures détectées. L'estimation du nombre optimal $K^\star$ est faite par sélection de modèle : plusieurs segmentations $\hat{\boldsymbol{T}}(\lambda)$ sont estimées pour différentes valeurs de λ, puis une somme d'erreurs quadratiques $J(K)$ est calculée entre $\boldsymbol{B}$ et $\boldsymbol{Y}$. Ce terme diminue quand $\hat{K}$ augmente, jusqu'à ce que le gain sur l'erreur ne soit plus dû qu'au bruit. $K^\star$ correspond à $J(K^\star)$ où la pente de $J(K)$ change [Bleakley et Vert, 2011]. L'estimation de la matrice $\hat{\boldsymbol{B}}$ fournit non seulement la segmentation jointe des signaux, mais aussi une estimation de la valeur des coefficients des moyennes sur chaque segment, dans chaque signal. Un algorithme inspiré de la procédure LARS est développé pour l'implémentation, permettant l'estimation de $\boldsymbol{B}$ avec K ruptures, pour une complexité en $\mathcal{O}(mnK)$. Des théorèmes sont démontrés pour la détection et la localisation d'une unique rupture. Les résultats sont asymptotiques, lorsque la dimension m tend vers l'infini. Le group fused LASSO présente ainsi la propriété intéressante d'être d'autant plus performant que le nombre de signaux est grand.

Dans un cadre bayésien, la méthode BARD de [Bardwell et Fearnhead, 2014] intègre dans son modèle la proportion de signaux impactés par une même rupture au sein de la série temporelle multivariée. Cette méthode est destinée à la détection de segments considérés comme anormaux dans un ensemble de signaux. Les portions normales et anormales se distinguent par leur moyenne. Dans ce modèle, un état caché est associé à chaque observation multidimensionnelle $\boldsymbol{X}_i = (X_{1,i},\ldots,X_{m,i})' : S_t = (C_t,B_t)$ à l'instant t, où C_t est la position de la fin du segment précédent, et B_t est le type de segment, N pour normal ou A pour anormal. Le processus des états est markovien, et les probabilités de transitions sont définies. Deux modèles de fonctions de vraisemblances sont choisis pour le cas normal et le cas anormal. Les données sont supposées i.i.d. selon la distribution P_N dans un segment normal et P_A, du paramètre Θ dans le cas anormal. En cas d'anomalie, on suppose que la dimension j est affectée avec la probabilité p_j, indépendamment des autres signaux. Le modèle pour les observations $\boldsymbol{X}_{t:s}$ entre les instants t et s est donc :

$$f(\boldsymbol{X}_{t:s}) = \begin{cases} \prod_{j=1}^{m}\prod_{i=t}^{s} P_N(X_{j,i}) & \text{si le segment est normal,} \\ \int P_A(\boldsymbol{X}_{t:s}|\Theta) f(\Theta) d\Theta & \text{si le segment est anormal,} \end{cases} \tag{1.73}$$

avec

$$f(\boldsymbol{X}_{t:s}|\Theta) = \prod_{j=1}^{m} \Big(p_j \prod_{i=t}^{s} P_A(X_{j,i}|\Theta) + (1-p_j) \prod_{i=t}^{s} P_N(X_{j,i}) \Big). \tag{1.74}$$

L'expression de la densité de probabilité a posteriori est obtenue par une relation de récursion, comme [Fearnhead, 2006]. En partant de l'extrémité du signal $S_N = (C_N,B_N)$, les limites des segments et leur type sont estimés en appliquant

$$f(C_t,B_t|C_{t+1},B_{t+1},\boldsymbol{X}) \propto f(C_t,B_t|\boldsymbol{X}_{1:t}) f(C_{t+1},B_{t+1}|C_t,B_t) \tag{1.75}$$

où $\boldsymbol{X}_{1:t}$ est la matrice des colonnes 1 à t de $\boldsymbol{X}$. L'algorithme BARD fournit une estimation de la nature des segments et leur position. La segmentation est déduite des limites des segments estimés.

Détection des ruptures propres à chaque signal

Les traitements conjoints présentés précédemment conduisent à l'estimation d'une segmentation unique pour l'ensemble des m signaux. En revanche peu de méthodes permettent d'obtenir une segmentation individuelle de chaque signal, qui inclut et synchronise les ruptures communes à d'autres dimensions. Cette approche est particulière : elle intègre à la fois les événements inter-capteurs, et les événements propres à chaque capteur. La figure 1.5 illustre ce cas : les trois signaux composant la série temporelle sont indépendants, mais des événements extérieurs entraînent l'apparition de ruptures avec une certaine probabilité. Le signaux 1 et 2 sont affectés par des ruptures totalement indépendantes, tandis que le signal 3 présente des ruptures communes avec le signal 1, et une partie des ruptures du signal 2. L'estimation d'une segmentation commune, sous forme d'un vecteur de dimension K, n'est pas adaptée, on cherche m vecteurs $\boldsymbol{T}_m = (\tau_{m,1}, \ldots, \tau_{m,K_m})$, où certaines positions $\tau_{.,k}$ sont partagées par plusieurs signaux.

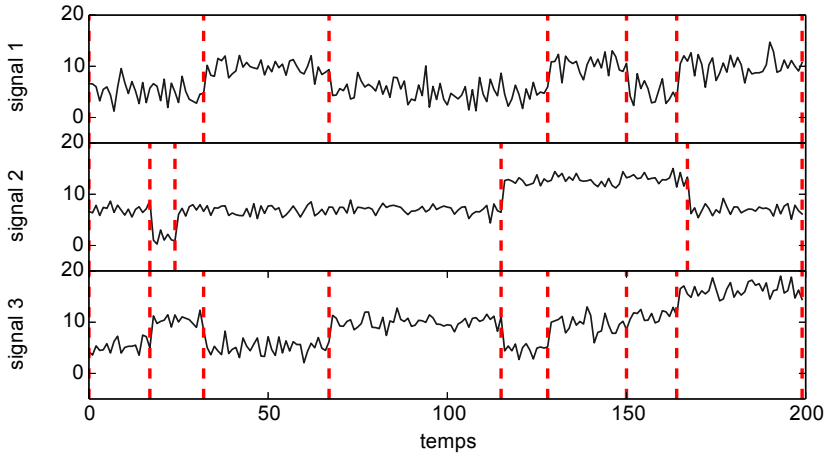


FIGURE 1.5 – Série temporelle multivariée où l'apparition de ruptures (en rouge) simultanément sur plusieurs signaux est régie par des relations de dépendance. Les positions des ruptures des signaux 1 et 2 sont indépendantes, les ruptures du signal 1 apparaissent simultanément sur le signal 3, et la moitié des ruptures du signal 2 apparaissent simultanément sur le signal 3, aux positions 17 et 115.

Dans un cadre bayésien, le paramètre à estimer est la matrice des indicatrices $\boldsymbol{R}$ associée aux observations de $\boldsymbol{X}$, telle que le coefficient du signal j et à l'instant i est défini par :

$$R_{j,i} = \begin{cases} 1 & \text{si } X_{j,i} \text{ est une rupture,} \\ 0 & \text{sinon,} \end{cases} \tag{1.76}$$

pour tout $1 \leq i \leq n$ et $1 \leq j \leq m$.

Un article récent, [Fan et Mackey, 2015], propose un modèle hiérarchique génératif consistant à partir d'une variable latente $q_i \in [0,1]$, la probabilité d'avoir une rupture à

l'instant i, puis d'obtenir la variable $R_{j,i} \in [0,1]$ à partir de q_i, et qui donne la probabilité que le signal j ait une rupture à l'instant i. Les segments étant ainsi délimités, les paramètres Θ des distributions sur chaque segment peuvent être générés comme dans les modèles bayésiens classiques. Le modèle BASIC est présenté pour des données normales, de moyennes inconnues et de variances égales à 1.

Enfin, la méthode de détection de rupture présentée dans [Dobigeon *et al.*, 2007a], intègre un paramètre supplémentaire dans le modèle : le vecteur $\boldsymbol{P}$ des probabilités d'observer simultanément des ruptures sur certains groupes de signaux. Il s'agit des probabilités d'observer certains motifs de 0 et de 1 dans les colonnes de la matrice des indicatrices $\boldsymbol{R}$, qui sont appelés configurations ϵ. Cette approche, que nous reprenons dans notre modèle multivarié, est présentée plus en détail dans le chapitre 4. La méthode est mise en œuvre sur des applications de natures diverses en modifiant le choix des distributions a priori [Dobigeon *et al.*, 2007a, Dobigeon *et al.*, 2007b, Dobigeon et Tourneret, 2009].

1.1.5 Synthèse

Les méthodes présentées dans cette partie ont été développées pour répondre au problème de la détection de sauts de moyennes, du simple cas où l'existence d'un unique changement est testée (1.5) au cas plus complexe où plusieurs ruptures se produisent à des instants inconnus (1.19), dans des séries temporelles multivariées (section 1.1.4). La complexité du problème augmente très vite lorsque le nombre de ruptures et leurs positions sont inconnues, et les propriétés théoriques du détecteur ne sont pas toujours démontrées. Toutefois des résultats sont souvent établis, garantissant les performances du détecteur, dans le cas simple d'une rupture unique.

Les méthodes paramétriques nécessitent des hypothèses fortes sur les distributions des observations, la loi normale est souvent choisie. L'inférence bayésienne permet d'introduire les lois des paramètres, considérés comme des variables aléatoires, et d'obtenir l'expression de la densité de probabilité a posteriori. De manière générale, les modèles paramétriques permettent une bonne détection des ruptures tant que les données sont conformes aux hypothèses. Quand cela ne peut pas être garanti, ou qu'une méthode plus généralisable est souhaitée, les approches non-paramétriques sont préférables. De nombreux travaux portent sur les tests de rangs, notamment le test de WMW et ses variantes, dont un des avantages est sa robustesse aux données aberrantes.

Pour les séries temporelles multivariées, l'analyse la plus répandue a pour objectif d'estimer une unique segmentation à partir du traitement conjoint de toutes les dimensions du signal. L'augmentation du nombre de canaux apporte plus d'information statistique, et favorise la détection des événements de plus faible amplitude. Certaines méthodes tiennent compte du fait que les ruptures ne sont pas toujours présentes sur l'ensemble des signaux. Pour aller plus loin, quelques algorithmes ont été mis au point pour générer une segmentation à la fois individuelle et collective de toutes les dimensions. Ainsi, les travaux de [Dobigeon *et al.*, 2007a], réalisés dans un cadre bayésien, incitent à approfondir la recherche des relations structurelles entre les signaux qui sous-tendent l'apparition ou non de ruptures communes. Ces liens entre les grandeurs mesurées dans la série temporelle se

représentent par un graphe.

1.2 Graphes de dépendance

L'analyse conjointe des signaux d'une série temporelle multivariée $\boldsymbol{X}$ prend du sens lorsqu'il existe des liens entre les signaux $\boldsymbol{X}_{j,\bullet}$. En effet, en l'absence de relation le traitement univarié de chaque composante $\boldsymbol{X}_{j,\bullet}$ est plus pertinent. Nous supposons donc que les ruptures de la série temporelle $\boldsymbol{X}$ ne se produisent pas toutes indépendamment les unes des autres. Selon ce principe, un mécanisme sous-jacent régit l'apparition des ruptures sur les différents signaux. Dans cette partie, nous présentons le formalisme des réseaux bayésiens ainsi que les principales approches rencontrées dans la littérature pour l'apprentissage de graphe. Les notions abordées ci-après forment le cadre des travaux qui sont présentés dans le chapitre 5. Pour aller plus loin, le lecteur peut se référer aux ouvrages de [Lauritzen, 1996,Naïm *et al.*, 2011,Pearl, 2014]. Le formalisme des graphes est également employé pour décrire les connections causales entre des variables, en complément de l'approche probabiliste. L'identification de relations de cause à effet entre les entités d'un système complexe à partir de leur observation est un problème central des sciences expérimentales. Les livres de [Pearl, 2000,Spirtes *et al.*, 2000] y sont consacrés et présentent un grand nombre de méthodes. Nous nous plaçons dans le cadre plus général de l'étude des relations de dépendance entre les variables.

1.2.1 Réseaux bayésiens

Pour analyser un système complexe, les éléments constitutifs sont représentés par des variables aléatoires $X_1, \ldots, X_m$, formant un ensemble fini noté U. La proposition $X_i = x_i$, où x_i est l'une des valeurs possibles du domaine de définition de X_i, est un événement [1]. Pour estimer les liens statistiques entre les variables, des tests d'indépendance et des mesures de corrélations sont effectués sur les observations. On suppose que les corrélations observées entre deux variables ont pour origine une relation de dépendance. Deux formalismes complémentaires sont adoptés pour exprimer les relations (non nécessairement causales) entre les variables aléatoires : la représentation par un graphe et la modélisation probabiliste. Un modèle graphique probabiliste combinant ces deux représentations constitue un réseau bayésien.

Représentation graphique

Une manière à la fois intuitive et efficace de dépeindre les relations de dépendance entre les variables est la notation graphique, grâce à laquelle les liens entre les éléments d'un système complexe deviennent plus facilement appréhendables. Les variables sont représentées par les sommets d'un graphe dont les arêtes symbolisent les relations de dépendance (ou

1. Ce terme est pris ici au sens probabiliste, et non pas, comme dans le contexte de la détection de rupture de la partie précédente, en tant que synonyme de changement.

de causalité), et l'absence d'arête des relations d'indépendance conditionnelle (ou l'absence de lien causal). La terminologie et les outils développés en théorie des graphes qui sont employés sont définis ci-dessous.

Définition 1.2.1 (Graphe)**.** *Un* graphe *est constitué d'un ensemble U de m nœuds ou sommets, les variables, et d'un ensemble E d'arêtes, reliant des paires de nœuds. Il est noté $G = (U,E)$. Deux sommets d'un graphe X_i et X_j sont dits* adjacents *lorsqu'il existe une arête entre X_i et X_j.*

Lorsqu'un graphe ne présente pas d'arête, il est dit *vide* (ou nul). À l'inverse, lorsque tous les sommets sont adjacents, le graphe est *complet*.

Définition 1.2.2 (Chemin et circuit)**.** *Un* chemin *(ou chaîne) est une séquence d'arêtes $((X_i,X_j), (X_j,X_k), \ldots, (X_l,X_m))$ telle que chaque arête (X_j,X_k) commence avec le sommet terminant l'arête précédente (X_i,X_j). Un* circuit *ou* cycle *est un chemin dont le sommet initial et le sommet final sont identiques. Lorsque qu'un graphe n'a pas de cycle, il est dit* acyclique.

Définition 1.2.3 (Graphe dirigé, non dirigé et squelette)**.** *Lorsque les arêtes sont orientées d'un sommet vers un autre, le graphe est dit* dirigé *ou* orienté*, et lorsqu'aucune arête n'est orientée, le graphe est* non dirigé *ou* non orienté*. Dans un graphe orienté, les arêtes sont aussi appelées* arc*. Le graphe non dirigé qui possède les même nœuds et les mêmes arêtes qu'un graphe dirigé est le* squelette *de ce graphe.*

Une catégorie de graphe en particulier nous intéresse pour modéliser une structure de dépendance : les graphes acycliques orientés.

Définition 1.2.4 (Graphes acycliques orientés)**.** *Un graphe* acyclique orienté*, en anglais directed acyclic graph, ou DAG, est un graphe orienté qui ne possède pas de cycle dirigé.*

Le champ lexical de la famille est employé dans la théorie des graphes pour décrire les relations entre des sommets : lorsqu'une arête issue de X_j pointe vers X_i, on dit que X_j est un *parent* de X_i, et X_i est un *enfant* de X_j. L'ensemble des nœuds parents de X_i est noté $\mathrm{Pa}(X_i)$. Le graphe de la figure 1.6 est un exemple de relations de dépendance entre 5 variables incarnant la saison, la pluie, le système d'arrosage, l'humidité du sol et l'aspect glissant de ce dernier [Pearl, 2000, chapitre 1.2, page 15]. Ces variables sont de natures différentes, prenant par exemple l'un des états parmi {printemps,été,automne,hiver} pour S ou {marche,arrêt} pour A. Le pendant de l'approche graphique pour la représentation des liens entre les variables est l'approche probabiliste.

Modèle probabiliste

Les relations entre les variables X_i, $1 \leq i \leq m$, s'expriment sous la forme de la probabilité jointe de toutes les variables : $\mathcal{P}(X_1,\ldots,X_m)$. Celle-ci se décompose en un produit de probabilités conditionnelles, pour un ordre donné des variables :

$$\mathcal{P}(X_1,\ldots,X_m) = \prod_{i=1}^{m} P(X_i|X_1,\ldots,X_{i-1}). \tag{1.77}$$

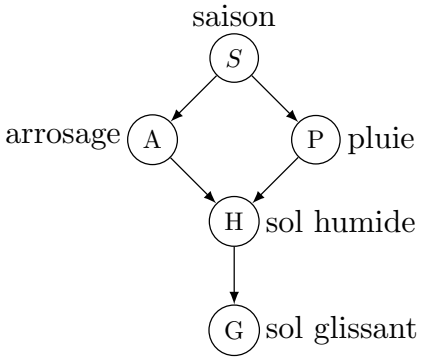


FIGURE 1.6 – Exemple de DAG entre les variables saison (S), pluie (P), système d'arrosage (A), humidité (H) et aspect glissant (G) du sol.

Une notion centrale du modèle probabiliste est l'indépendance conditionnelle.

Définition 1.2.5 (Indépendance conditionnelle)**.** *Soient X_i, X_j et X_k trois sous-ensembles disjoints de variables de l'ensemble fini U et $\mathcal{P}(.)$ la loi de probabilité jointe sur les variables de U. X_i et X_j sont dits* indépendants conditionnellement *à X_k si*

$$P(X_i|X_j,X_k) = P(X_i|X_k) \text{ si } P(X_i,X_k) > 0. \tag{1.78}$$

Cette relation est notée $X_i \perp\!\!\!\perp X_j \mid X_k$, et signifie que la connaissance apportée par la valeur de X_j ne fournit pas d'information supplémentaire sur X_i dès lors que X_k est connu. L'ensemble des triplets des sous-ensembles de U qui vérifient de telles relations constituent le *modèle d'indépendance*[2] $M(U)$. Toute loi de probabilité $\mathcal{P}$ sur un ensemble de variables U définit un modèle d'indépendance $M_{\mathcal{P}}(U)$. Pour toute variable X_i, $1 \leq i \leq m$, l'ensemble des variables de $\{X_1,\ldots,X_{i-1}\}$ qui conditionnent l'indépendance de X_i dans $M_{\mathcal{P}}(U)$ est noté $\mathrm{Pa}'(X_i)$. La probabilité jointe (1.77) devient alors :

$$\mathcal{P}(X_1,\ldots,X_m) = \prod_{i=1}^{m} P(X_i|\mathrm{Pa}'(X_i)). \tag{1.79}$$

La complexité du calcul de $\mathcal{P}(X_1,\ldots,X_m)$ est alors grandement réduite : l'introduction des relations d'indépendances conditionnelles permet d'exploiter uniquement les propriétés locales entre chaque variable et ses parents immédiats.

Pour faire le lien avec la représentation graphique, la condition de Markov globale établit que pour chaque variable X_i de U et pour chaque sous-ensemble X_j de U privé des descendants de X_i dans le DAG G, on a

$$P(X_i|\,\mathrm{Pa}(X_i),X_j) = P(X_i|\,\mathrm{Pa}(X_i)). \tag{1.80}$$

2. L'expression *modèle de dépendance* est également employée, par exemple dans [Pearl, 2000].

Suivant cette condition, les variables de $\mathrm{Pa}'(X_i)$ de l'expression (1.79) correspondent aux parents de X_i dans la représentation du système par un DAG G. La condition de Markov signifie donc que la variable X_i est séparée des autres variables sauf de ses descendants par ses parents $\mathrm{Pa}(X_i)$. L'ensemble des parents est unique si la distribution de $\mathcal{P}$ est strictement positive. G et la fonction de probabilité $\mathcal{P}$ sont dits *compatibles*, $\mathcal{P}$ est *Markov relatif* à G, et G *représente* $\mathcal{P}$. La décomposition de la probabilité jointe en (1.79) peut donc être déduite de la représentation graphique. Dans l'exemple du système d'arrosage de [Pearl, 2000, chapitre 1.2, page 15], la factorisation de la loi de probabilité jointe est aisément déduite du graphe de la figure 1.6 :

$$\mathcal{P}(S,P,A,H,G) = P(S)P(A|S)P(P|S)P(H|A,P)P(G|H). \tag{1.81}$$

Pour représenter des relations de causalité, une condition supplémentaire de complétude causale requiert que toutes les causes des variables de U soient dans U. Les probabilités conditionnelles de (1.80) sont aussi appelées *paramètres*. L'ensemble des paramètres de $\mathcal{P}$ est $\theta = P(X_i|\,\mathrm{Pa}(X_i))$ Lorsque la condition de Markov est respectée, la représentation graphique G et le modèle probabiliste $\mathcal{P}$ Markov relatif à G forment un réseau bayésien, défini par le graphe $G = (U,E)$, orienté et sans circuit, et les paramètres numériques θ de la loi $\mathcal{P}$.

Propriétés et équivalence de Markov

En associant les représentations graphiques et probabilistes, il est possible d'exploiter les propriétés de l'une ou de l'autre. Ainsi une méthode graphique, le critère de d-séparation, a été proposée pour établir les relations d'indépendances conditionnelles entre les variables à partir du DAG G [Geiger *et al.*, 1990, Pearl, 2014]. Elle consiste à identifier les nœuds de G bloquant le passage d'information entre deux variables (voir l'annexe D.1) : un nœud est d-séparé du reste du graphe par l'ensemble de ses parents, de ses enfants, et des parents de ses enfants. Un motif particulier du graphe est relevé : la V-structure (ou immoralité). Celle-ci est composée d'un nœud central vers lequel sont dirigés les arcs issus de deux sommets non adjacents entre eux. Dans le cadre d'une interprétation causale, il s'agit d'un effet commun à deux causes indépendantes. Dans la figure 1.6, les sommets A, P et H forment une V-structure, dirigée vers H.

La d-séparation permet d'établir des correspondances entre une représentation graphique G et un modèle d'indépendance $M(U)$. Ainsi, une *carte d'indépendance* d'une loi $\mathcal{P}$ est un graphe G dont toute d-séparation implique une relation d'indépendance conditionnelle dans la loi jointe. Le graphe complet est par exemple une carte d'indépendance de la loi. D'autres relations d'indépendance peuvent toutefois ne pas se retrouver dans le graphe. Pour tout graphe G' crée à partir de G en supprimant des arêtes, si G' n'est pas une carte d'indépendance de $\mathcal{P}$, alors G est une carte d'indépendance minimale, qui représente toutes les relations de dépendance du modèle $M(U)$. Lorsque G représente toutes les relations de $M(U)$, et uniquement ces relations, le graphe est une *carte parfaite*, le modèle est dit graphe-isomorphe et la loi jointe est représentable par un DAG. On dit également que G et $\mathcal{P}$ sont fidèles l'un à l'autre.

Pour une probabilité jointe donnée, plusieurs décompositions sont possibles. La loi de probabilité donnée en (1.81) Markov relative à G_1 peut par exemple aussi s'écrire

$$\mathcal{P}(S,P,A,H,G)=P(A)P(S|A)P(P|S)P(H|A,P)P(G|H) \tag{1.82}$$
$$=P(P)P(S|P)P(A|S)P(H|A,P)P(G|H). \tag{1.83}$$

Or ces deux factorisations, compatibles avec des graphes différents de G_1, nommés G_2 et G_3 respectivement, correspondent au même modèle d'indépendance que la décomposition en (1.81). Il n'y a pas d'unicité de la représentation graphique d'un modèle d'indépendance, et l'ensemble des DAG compatibles constitue la *classe d'équivalence de Markov*. La figure 1.7 illustre tous les DAG existants pour trois variables, les éléments d'une même classe d'équivalence de Markov y sont entourés en bleu.

Il a été démontré que deux DAG représentent le même modèle d'indépendance s'ils partagent le même squelette et les mêmes V-structures [Verma et Pearl, 1990, Frydenberg, 1990]. Ce critère graphique permet de construire les classes d'équivalence de Markov. Ces dernières peuvent alors être représentées par des graphes partiellement dirigés, appelés *patterns* dans [Verma et Pearl, 1990], dont les seuls arcs sont impliqués dans des V-structures. Le pattern du graphe en 1.6 est donné à la figure 1.8a. En-dehors des arcs dus aux V-structures, la direction d'autres arêtes est parfois commune à tous les DAG d'une même classe d'équivalence, comme l'arête $H \rightarrow G$ de la figure 1.8b. En effet, une orientation contraire entraînerait l'apparition d'une nouvelle V-structure. Un pattern auquel s'ajoutent ces arcs irréversibles s'appelle un graphe *essentiel*, *orienté au maximum* ou un graphe partiellement dirigé *complété* (CPDAG). Cette représentation de la classe d'équivalence de Markov, donc du modèle d'indépendance, est unique [Andersson *et al.*, 1997]. Les DAG obtenus en orientant les arêtes de direction indéterminée d'un graphe essentiel, sans introduction de nouvelles V-structures, sont des extensions de ce graphe.

1.2.2 Apprentissage du graphe

Dans un grand nombre de problèmes, la seule connaissance disponible sur un système est un ensemble de mesures indirectes prises par des capteurs, à partir desquelles on cherche à retrouver le graphe sous-jacent décrivant les liens statistiques entre les variables observées. Dans cette partie, nous présentons quelques-unes des approches de la littérature destinées à l'apprentissage de la structure.

Causalité et interventions

Quand le réseau incarne des relations de causalité entre les variables, l'arc $X_i \rightarrow X_j$ signifie que les événements de X_i causent les événements de X_j. Cette information n'est pas contenue dans la probabilité conditionnelle, puisque le théorème de Bayes permet d'écrire $P(X_i|X_j)P(X_j) = P(X_j|X_i)P(X_i)$. La symétrie induite par les relations probabilistes a pour conséquence que la seule observation des variables ne suffit pas pour établir les liens de causalité dans un système. Pearl a introduit le concept d'*intervention* (ou manipulation),

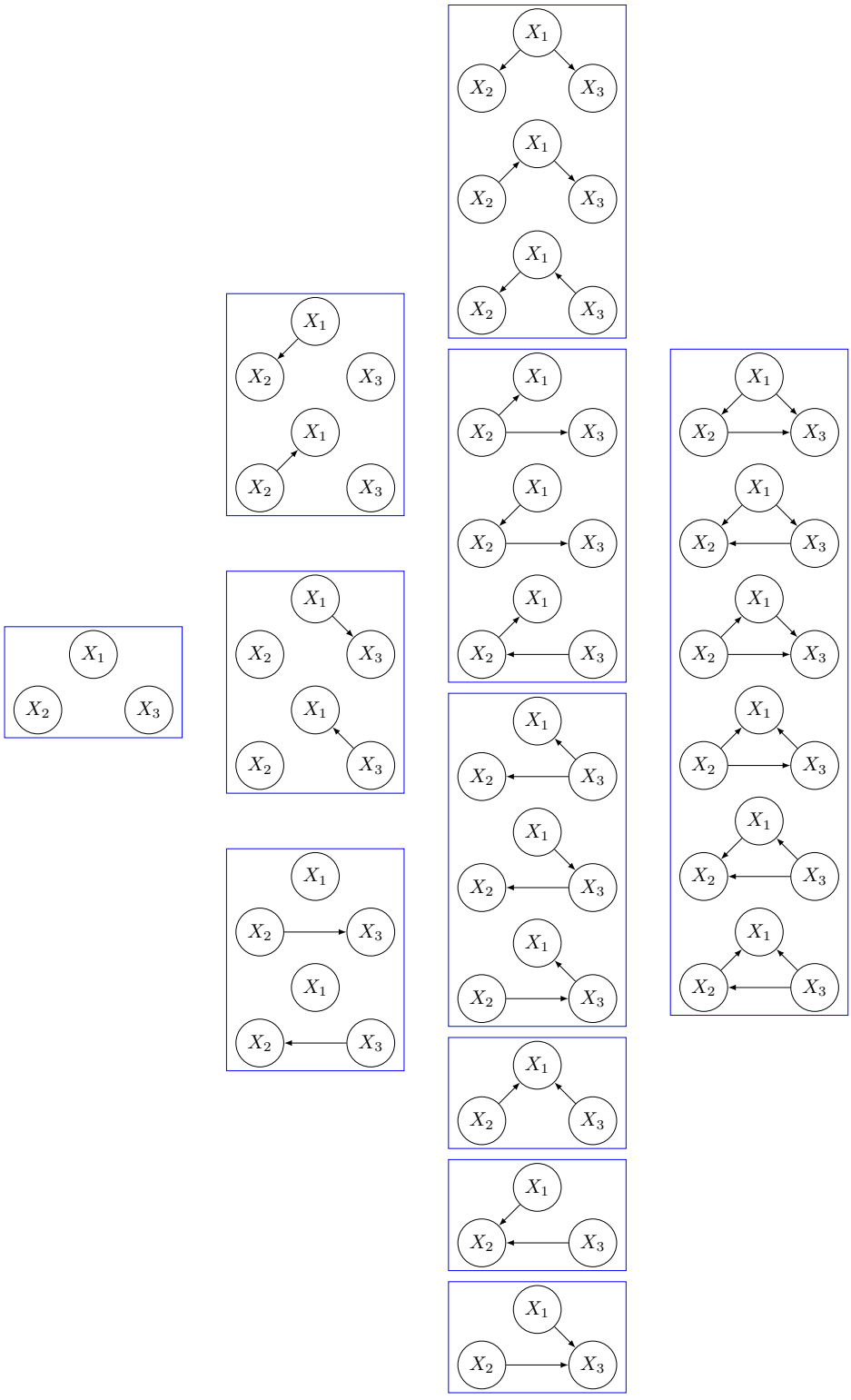


FIGURE 1.7 – Tous les DAG à 3 nœuds, de 0 à 3 arêtes de la gauche vers la droite. Les graphes d'une même classe d'équivalence sont regroupés et encadrés. On dénombre 25 DAG et 11 classes.

(a) Pattern (b) Graphe essentiel

FIGURE 1.8 – Représentation de la classe d'équivalence du graphe de la figure 1.6 par le pattern, composé du squelette et de la V-structure 1.8a, et par le graphe essentiel 1.8b.

distincte de la notion d'observation de l'événement $(X = x)$ [Pearl, 2000]. L'opérateur do(.) associé agit en imposant une valeur donnée à une variable quelles que soient les valeurs prises par ses parents, comme si les arcs provenant des parents avaient été supprimés dans le graphe. Dans l'exemple du système d'arrosage de la figure 1.6, si la variable A est mise dans l'état 1 (l'arrosage est allumé), l'arête entre A et S est supprimée, et une factorisation tronquée de la loi de probabilité jointe est alors déduite :

$$\mathcal{P}(S,P,A,H,G|\,\mathrm{do}(A=1)) = P(S)P(P|S)P(H|A=1,P)P(G|H). \tag{1.84}$$

Grâce aux interventions, il est possible d'identifier les relations de causalité entre deux variables, qui ne peuvent pas être obtenues à partir des probabilités conditionnelles, puisque le théorème de Bayes introduit une symétrie. Mais avec l'opérateur do(.), lorsque X_i est la cause de X_j, on a :

$$P(X_j = x_j|\,\mathrm{do}(X_i = x_i)) = P(X_j = x_j|X_i = x_i), \tag{1.85}$$
$$P(X_i = x_i|\,\mathrm{do}(X_j = x_j)) = P(X_i = x_i), \tag{1.86}$$

ce qui permet de déterminer la relation de causalité entre X_i et X_j, et d'orienter les arêtes correspondantes dans le graphe.

Plusieurs travaux ont été entrepris pour établir les façons d'intervenir sur un système afin d'en déduire le graphe causal complet. Une série d'expériences, où l'opérateur effectue des interventions sur certaines variables, peut être menée afin d'apprendre la structure du réseau bayésien causal, comme le proposent [Tong et Koller, 2001]. D'autres méthodes combinent les résultats d'expériences avec interventions, avec des observations sans manipulation des variables, comme [Meganck *et al.*, 2006]. Malheureusement il n'est pas toujours possible de procéder à des manipulations sur les variables, pour des raisons éthiques, de coût, de moyens, ou parce qu'on n'a pas accès au système. Les graphes causaux doivent alors être (en partie) estimés à partir de l'observation des variables.

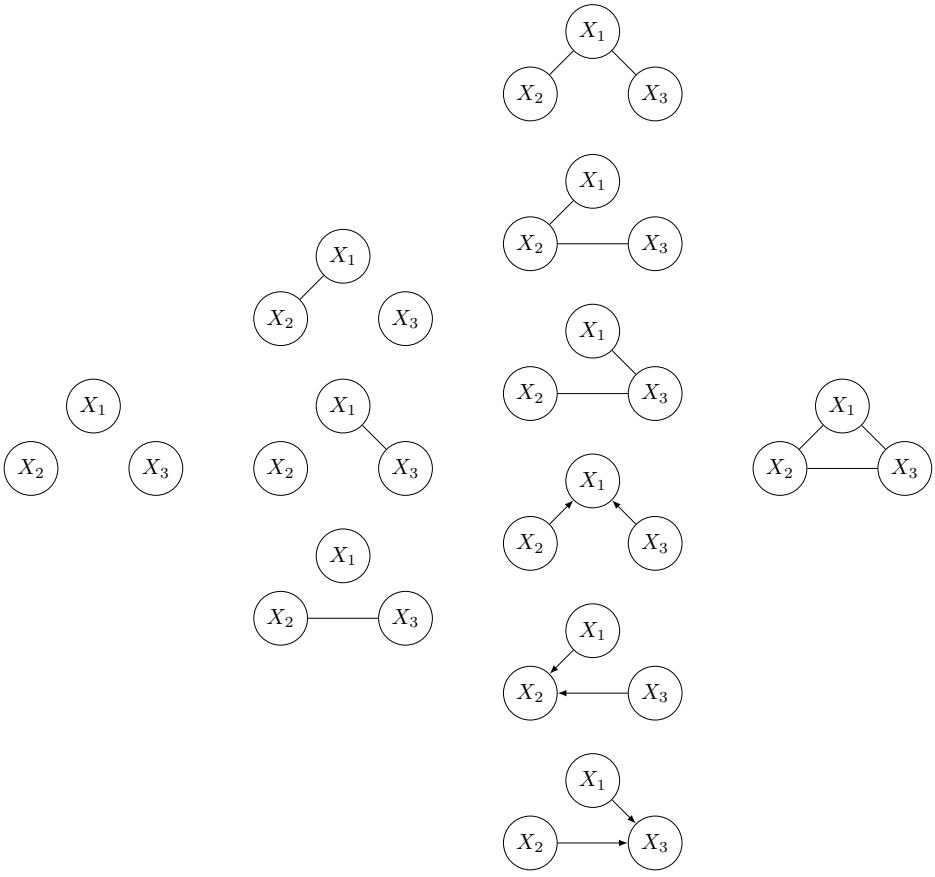


FIGURE 1.9 – Tous les graphes essentiels à 3 nœuds, représentant les classes d'équivalence de Markov des DAG de la figure 1.7.

Séries temporelles et causalité de Granger

Lorsque la dimension temporelle intervient, par exemple quand il existe un décalage dans la propagation d'un changement, il devient plus facile d'identifier une direction de causalité puisque la cause ne peut pas précéder l'effet. Dans [Granger, 1969], Granger a défini la causalité comme la relation entre deux variables X_1 et X_2 telle que si X_2 cause X_1, alors l'information apportée par le passé de X_2 permet d'améliorer la prédiction de X_1. Autrement dit, le signal $X_{1,t}$ ne cause pas le signal $X_{2,t}$ si et seulement si $P(X_{2,t}|X_{2,1:t-1},x_{1,1:t-1}) = P(X_{2,t}|X_{2,1:t-1})$. Cette notion est largement employée pour l'analyse de série temporelles stationnaires en économétrie. Le modèle causal simple à deux séries temporelles $X_{1,t}$ et $X_{2,t}$ s'écrit par exemple :

$$X_{1,t} = \sum_{i=1}^{m} a_i X_{1,t-i} + \sum_{i=1}^{m} b_i X_{2,t-i} + \epsilon_t, \tag{1.87}$$

$$X_{2,t} = \sum_{i=1}^{m} c_i X_{1,t-i} + \sum_{i=1}^{m} d_i X_{2,t-i} + \eta_t, \tag{1.88}$$

où $X_{1,t}$ et $X_{2,t}$ sont deux séries temporelles stationnaires de moyennes nulles, ϵ_t et η_t sont les coefficients d'un bruit blanc non corrélé, et où m peut être infini. $X_{2,t}$ cause $X_{1,t}$ si certains coefficients de b_i sont non nuls, et $X_{1,t}$ cause $X_{2,t}$ si certains coefficients de c_i sont non nuls. Une causalité instantanée existe également si les termes $X_{2,t}$ et $X_{1,t}$ apparaissent dans les expressions (1.87) et (1.88) respectivement. Cette formulation rencontre un grand nombre d'applications, par exemple en neurosciences pour établir un graphe de connectivité cérébral (voir par exemple [Eichler, 2005]), mais conduit à de mauvaises interprétations lorsque des variables latentes n'ont pas été prises en compte. D'autres approches, inspirées par la théorie de l'information, ou comme le critère d'indépendance spectral de [Shajarisales *et al.*, 2015], permettent de déterminer la direction de causalité dans des systèmes déterministes, sans faire intervenir la causalité de Granger. Le mécanisme causal y est décrit comme un filtre linéaire invariant.

Modèle d'équations structurelles

Les dépendances entre les variables du modèle peuvent parfois s'écrire sous forme d'équations structurelles, cette fois-ci sans notion temporelle, voir [Pearl, 2000, chapitre 5, p.133]. Un ensemble de fonctions f_i relient les variables X_i, $1 \leq i \leq m$, avec leurs parents immédiats :

$$x_i = f_i(\mathrm{Pa}(x_i),u_i), \tag{1.89}$$

où u_i représente une erreur, généralement modélisée par une loi normale. Un modèle linéaire est souvent employé pour décrire les mécanismes de dépendance entre les variables. Un diagramme causal se déduit du modèle, en reliant les parents $\mathrm{Pa}(X_i)$ à X_i par une flèche, et en ajoutant une arête bi-dirigée entre les variables dont les erreurs sont dépendantes. Le modèle linéaire peut être estimé par régression, en faisant par exemple intervenir les

coefficients de corrélation partielle. Ces termes sont nuls lorsque deux variables sont conditionnellement indépendantes. Le formalisme des équations structurelles est pratique pour étudier la propagation de perturbations, par exemple lors d'interventions.

Méthodes basées sur la matrice de précision

Lorsque les observations suivent une loi normale multivariée, de moyenne μ et de matrice de covariance Σ, définie positive, un modèle graphique gaussien peut être déduit de la matrice de concentration (ou de précision) $\Theta = \Sigma^{-1}$. En effet, l'indépendance conditionnelle entre les variables X_i et X_j se traduit par un coefficient $\theta_{i,j} = 0$ dans Θ [Lauritzen, 1996, chapitre 5.1.3, p.129]. Dans la plupart des cas, la matrice Σ est inconnue, l'estimateur empirique de la matrice de covariance S est alors employé. Estimer les paramètres du modèle revient à identifier les coefficients nuls de Θ. Une approche possible est basée sur le logarithme de la fonction de vraisemblance, à laquelle s'ajoute une pénalisation par une norme ℓ_1 [Yuan et Lin, 2007, Friedman *et al.*, 2008]. Le problème est de la forme :

$$\hat{\Theta} = \underset{\Theta \succ 0}{\operatorname{argmax}} \left\{ \log |\Theta| - tr(S\Theta) - \lambda \|\Theta\|_1 \right\}, \tag{1.90}$$

où $|A|$ est le déterminant de la matrice A, $tr(A)$ est sa trace et $\|A\|_1$ est la somme de la valeur absolue des éléments de A. La méthode du Graphical LASSO, par exemple, permet d'estimer Θ efficacement à partir du problème convexe (1.90) [Friedman *et al.*, 2008]. La résolution se fait par une méthode du point intérieur ou en estimant le problème dual pour Σ par un algorithme de descente de coordonnées par blocs. La solution dépend du paramètre de régularisation λ, et une phase de sélection de modèle est nécessaire pour obtenir la représentation graphique. Celle-ci se présente sous la forme d'un graphe non orienté. Les méthodes basées sur la matrice de concentration estiment ne permettent pas d'estimer les relations de causalité entre les variables.

Apprentissage de la structure d'un réseau bayésien

L'apprentissage de la structure du graphe d'un réseau bayésien à partir d'un ensemble d'observations des variables est un problème difficile, notamment en raison du nombre de DAG possibles qui croît superexponentiellement avec le nombre de variables. Il existe principalement deux catégories de méthodes. La première repose sur des fonctions de score, par lesquelles les modèles sont évalués et sélectionnés, au cours de l'exploration d'un espace de recherche, comme l'espace des DAG ou des arbres de recouvrement maximal [3] (MWST pour *Maximum Weight Spanning Trees* [Chow et Liu, 1968]). Différentes fonctions de score ont été proposées afin de parvenir à une carte d'indépendance du modèle tout en pénalisant les graphes les plus complexes [Buntine, 1991, Cooper et Herskovits, 1992, Bouckaert, 1993, Lam et Bacchus, 1994, Heckerman *et al.*, 1995, De Campos, 2006]. L'algorithme de *Greedy Equivalence Search* (GES) [Chickering, 2003], qui explore l'espace des classes d'équivalence

3. Arbre non orienté connectant toutes les variables, représentant tous les arbres de même squelette.

de Markov à l'aide d'opérateurs d'ajout et de suppression d'arcs, est présenté plus en détail dans le chapitre 5.

L'autre catégorie de méthodes, la recherche sous contrainte, est basée sur l'estimation de relations d'indépendances conditionnelles. Deux algorithmes principaux, IC (pour *Inductive Causation*) de [Verma et Pearl, 1992] et PC (nommé d'après ses auteurs) de [Spirtes *et al.*, 2000] en sont des exemples bien connus. Ils mettent en œuvre le principe de d-séparation, et parviennent à établir un graphe non dirigé dans lequel certaines arêtes sont orientées par l'identification des V-structures et l'application de règles de propagation des orientations. Ces méthodes reposent sur l'application de tests d'indépendance conditionnelle, comme le test du χ^2, entre deux variables X_1 et X_2, par rapport à un groupe d'autres variables X_3. L'algorithme PC commence par le graphe complètement connecté, dont les arêtes non existantes sont retirées progressivement, tandis que l'algorithme IC ajoute des arêtes à partir du graphe vide. Ces deux méthodes sont présentées plus en détail dans l'annexe D.2. Parmi les variantes qui ont été développées par la suite, nous pouvons citer l'algorithme IC*, qui intègre des variables latentes (voir [Pearl, 2000, p.52]), [Kalisch et Bühlmann, 2007] proposent une implémentation de l'algorithme PC applicable à de grandes dimensions pour des variables suivant la distribution gaussienne dans un cas parcimonieux, c'est-à-dire en présence d'un grand nombre de variables devant le nombre d'échantillons. La méthode de [Cheng *et al.*, 1997] fait intervenir des mesures d'information mutuelle.

L'inconvénient des approches basées sur un score est que la solution obtenue n'est pas toujours optimale et que le problème d'apprentissage du meilleur graphe étant NP-difficile, le nombre de variables est limité, mais leur complexité est moindre. Les méthodes de recherche sous contrainte parviennent asymptotiquement au bon graphe, mais les tests d'indépendance conditionnelle ne sont fiables qu'à partir d'un nombre important d'échantillons, et le nombre de tests augmente fortement avec le nombre de variables, nécessitant d'adapter les stratégies.

Dans ce chapitre, nous avons présenté les différentes méthodes de la littérature pour la problématique de la détection de ruptures dans des séries temporelles univariées et multivariées, puis nous avons introduit le formalisme des réseaux bayésiens et les approches destinées à l'inférence d'un graphe de dépendance. Ces deux parties ont permis de définir le cadre dans lequel nous nous plaçons pour répondre à la problématique de la détection de ruptures et la recherche d'un graphe de dépendance décrivant le système. Nous commençons par aborder la question de la comparaison de deux échantillons afin d'établir si une différence notable dans leurs distributions est présente.

Chapitre 2

Tests d'homogénéité

Ce chapitre présente les tests d'hypothèse d'homogénéité classiques, employés pour la détection d'un décalage entre les moyennes (tests t, vraisemblance empirique) ou les médianes (test de Wilcoxon-Mann-Whitney) de deux échantillons. Les hypothèses nulle et alternative sont celles du problème (1.8) de l'existence d'une rupture à une position donnée, reformulées pour comparer les paramètres m_1 et m_2 des échantillons $\boldsymbol{X}_1 = (X_1, \ldots, X_{n_1})$ et $\boldsymbol{X}_2 = (X_{n_1+1}, \ldots, X_n)$ respectivement :

$$\begin{array}{ll} H_0 : & m_1 = m_2 \\ H_1 : & m_1 \neq m_2 \end{array} \tag{2.1}$$

où, selon le test, m_1 et m_2 sont les moyennes ou les médianes des échantillons $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ respectivement. Les échantillons sont de taille n_1 pour $\boldsymbol{X}_1$ et n_2 pour $\boldsymbol{X}_2$, avec $n = n_1+n_2$, et $X_i \in \mathbb{R}$ pour tout $1 \leq i \leq n$. Dans l'échantillon $k \in \{1,2\}$, les variables X_i sont indépendantes et identiquement distribuées selon une loi dont la fonction de répartition est notée F_k. Si celle-ci admet une moyenne et une variance, elles sont notées μ_k et σ_k^2. Lorsque les paramètres des distributions ne sont pas connus, certains estimateurs sont employés, comme la moyenne empirique de l'échantillon k

$$\overline{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_i, \tag{2.2}$$

et l'estimateur non biaisé de la variance

$$S_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (X_i - \overline{X}_k)^2. \tag{2.3}$$

Les approches vont du contexte paramétrique dans la partie 2.1, avec les tests du t, au contexte non paramétrique dans la partie 2.2, avec les tests de rang. Nous présentons ensuite dans la partie 2.3 une méthode dite *semi-paramétrique* que nous avons développée à partir des travaux de Owen [Owen, 2010] sur le rapport de vraisemblance empirique, et qui a fait l'objet d'une communication dans [Harlé *et al.*, 2015]. La comparaison de ces méthodes dans la partie 2.4 conduit au choix du test qui est au cœur du modèle du *Bernoulli Detector*, introduit par la suite au chapitre 3.

2.1 Tests paramétriques du t

La famille des tests du t, présentée initialement dans [Student, 1908], repose sur une statistique suivant une distribution t de Student sous l'hypothèse nulle, i.e. lorsque les échantillons sont homogènes. Cette statistique est de la forme $T = \frac{Z}{s}$, où Z, un terme sensible aux hypothèses testées, et s, un paramètre d'échelle, sont deux fonctions des données. Comme pour un grand nombre de tests paramétriques, on suppose que les observations sont distribuées selon la loi normale. Le paramètre s suit la distribution du χ^2 à p degrés de liberté, et Z et s sont indépendants. Pour comparer les moyennes μ_1 et μ_2 des deux échantillons $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$, les hypothèses du test sont les suivantes :

$$\begin{array}{ll} H_0 : & \mu_1 = \mu_2 \\ H_1 : & \mu_1 \neq \mu_2. \end{array} \tag{2.4}$$

Sans connaissance a priori sur ces moyennes, comme de savoir si, sous H_1, $\mu_1 < \mu_2$ ou si $\mu_1 > \mu_2$, le test appliqué est bilatéral.

Le test du t de Student compare μ_1 et μ_2 sous l'hypothèse que $\boldsymbol{X}_1$ suit la loi normale $\mathcal{N}(\mu_1,\sigma_1^2)$ et $\boldsymbol{X}_2$ suit la loi normale $\mathcal{N}(\mu_2,\sigma_2^2)$, les observations étant indépendantes. Dans un premier temps, on suppose que les variances σ_1^2 et σ_2^2 sont égales, c'est l'hypothèse d'homoscédasticité. Le terme Z correspond alors à l'estimateur de l'écart entre les moyennes et le paramètre d'échelle s est l'estimateur non biaisé de la déviation standard commune, corrigé pour les dimensions des échantillons :

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S_{1,2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{2.5}$$

où

$$S_{1,2} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}. \tag{2.6}$$

Sous H_0, la distribution exacte de T est la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Le cas où les variances des deux populations ne peuvent pas être considérées comme égales (hypothèse d'hétéroscédasticité) correspond au problème dit de problème Behrens-Fisher [Lehmann et D'Abrera, 1975, Kim et Cohen, 1998, p.95]. Un test classique pour comparer les hypothèses H_0 et H_1 est alors le test du t de Welch [Welch, 1947], dont la statistique s'écrit

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{2.7}$$

Il faut cependant noter que la loi de T dépend des paramètres inconnus σ_1^2 et σ_2^2, et que le dénominateur ne correspond pas à une variance commune entre les deux échantillons. Il n'est donc pas possible d'obtenir la loi exacte de T. Le test de Welch consiste alors à

approcher la distribution réelle de T par une loi de Student de degré de liberté

$$df = \frac{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{2.8}$$

qui dépend des variances empiriques de chaque population, .

D'autres variantes s'appliquent au cas où les variables sont dépendantes ou appariées, et pour tester d'autres paramètres que la moyenne des populations. Le test T^2 de Hotelling est une généralisation du test t de Student dans un cas multivarié, où les variables sont distribuées selon une loi normale à d dimensions. La statistique de test T^2 suit la loi de Fisher à d degrés de liberté au numérateur et à $n_1 + n_2 - 1 - d$ degrés au dénominateur.

2.2 Tests non paramétriques de rang

Pour pallier la dépendance du test à une hypothèse forte sur la distribution des données, un test non paramétrique constitue une bonne alternative. Parmi cette catégorie, les tests de rang reposent sur une comparaison de rangs déterminés à partir des observations, comme leur position dans le vecteur ordonné des observations, ou la position de la différence entre deux observations. Le plus connu d'entre eux est le test de Wilcoxon-Mann-Whitney (WMW), également appelé test U de Mann-Whitney ou encore test de la somme des rangs de Wilcoxon. Le principe a été introduit dans [Wilcoxon, 1945] et indépendamment dans [Mann et Whitney, 1947]. Des développements théoriques sont exposés dans l'ouvrage [Lehmann et D'Abrera, 1975]. Une statistique est construite à partir de la distribution des rangs des observations dans l'échantillon global. Les rangs correspondent aux positions de toutes les observations de $\boldsymbol{X}_1$ et de $\boldsymbol{X}_2$ dans le vecteur ordonné issu de l'ensemble $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2\}$ de $n = n_1 + n_2$ valeurs. L'expression du rang R_i de la variable X_i est donnée par la formule (1.60). Les hypothèses testées s'appliquent aux distributions des populations sans avoir à les spécifier :

$$\begin{aligned} H_0 :&\quad F_1 = F_2 \\ H_1 :&\quad F_1 \neq F_2, \end{aligned} \tag{2.9}$$

pourvu que la différence entre F_1 et F_2 sous H_1 affecte la moyenne des rangs. L'intérêt du test tient dans le peu d'hypothèses faites sur les données :

- les variables de $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ sont indépendantes ;
- elles sont ordonnables ;
- pour une variable X_i de $\boldsymbol{X}_1$ et une variable X_j de $\boldsymbol{X}_2$, les probabilités $\Pr(X_i > X_j)$ et $\Pr(X_i < X_j)$ sont égales sous H_0 ;
- ces probabilités sont différentes sous H_1.

Ces hypothèses reviennent à tester les probabilités que les variables d'un échantillon soient stochastiquement plus grandes que les variables de l'autre échantillon. En revanche ce test n'est *a priori* pas sensible à un changement dans le seul paramètre de variance entre les

populations. Bien qu'il soit souvent considéré comme un test sur l'égalité des médianes, le test de WMW n'en est pas un à strictement parler [Hart, 2001]. Dans certains cas cependant, par exemple si la distribution F_2 correspond à la distribution F_1 avec un décalage dans le paramètre de localisation, ce test est équivalent à un test d'égalité des médianes. Pour simplifier, nous supposons par la suite que nous nous trouvons dans l'un de ces cas, et par raccourci nous parlons de test sur la médiane. Pour illustrer ces remarques, la figure 2.1 donne trois exemples de distributions pour lesquelles le test de WMW est approprié ou non. Les échantillons (partie inférieure), leur distribution (partie supérieure) et leur médiane (pointillés) sont représentés, en rouge pour $\boldsymbol{X}_1$ et en bleu pour $\boldsymbol{X}_2$. Dans le premier cas, en 2.1a, les distributions F_1 et F_2 ont la même forme, et sont décalées de 0,5. Le test de WMW est sensible à cette différence, et revient à tester les médianes des distributions. Le deuxième cas, en 2.1b, correspond à deux distributions de même médiane, de même moyenne, mais de variance différente. Le test de WMW n'est pas adapté à la comparaison de ces populations. Enfin, le troisième cas, en 2.1c, donne un exemple de deux distributions discrètes différentes, mais de même médiane, pour lesquelles le test de WMW est applicable. En effet, les valeurs prises par les variables de $\boldsymbol{X}_2$ sont stochastiquement plus grandes que les valeurs prises par les variables de $\boldsymbol{X}_1$.

Pour comparer $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$, la statistique employée, sous la forme introduite dans [Mann et Whitney, 1947], est

$$U = \min(U_1,U_2) \tag{2.10}$$

où

$$U_1 = S_1 - \frac{n_1(n_1+1)}{2}, \qquad U_2 = S_2 - \frac{n_2(n_2+1)}{2}, \tag{2.11}$$

et où S_k est la somme des rangs des observations de X_k :

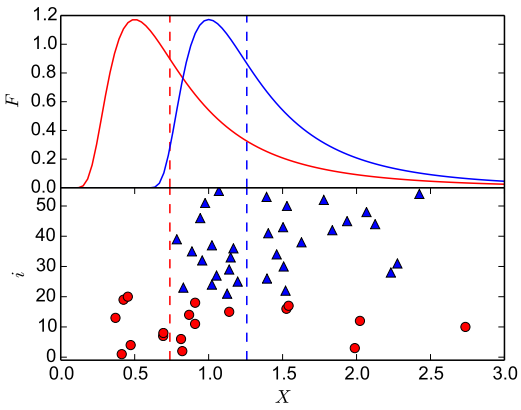
$$S_k = \sum_{j=1}^{n_k} R_j. \tag{2.12}$$

La statistique U compte le nombre de fois où une variable de $\boldsymbol{X}_k$ précède une variable de $\boldsymbol{X}_l$ dans le classement des rangs. Les rangs étant des entiers naturels, U est une variable discrète. L'hypothèse nulle est rejetée pour de grandes valeurs de U. La statistique U est généralisée pour $M \geq 2$ échantillons dans le test de Kruskal-Wallis [Kruskal et Wallis, 1952].

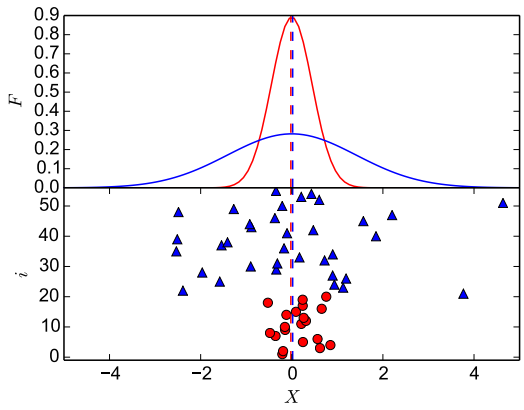
Sous l'hypothèse nulle, $\Pr(X_i > X_j) = \Pr(X_i < X_j)$, les rangs des deux échantillons sont donc mélangés aléatoirement et toutes les séquences possibles de rangs sont équiprobables. Pour les variables $X_1,\ldots,X_{n_1}$ de $\boldsymbol{X}_1$, la probabilité d'obtenir une séquence donnée $r_1,\ldots,r_{n_1}$ des rangs est donc :

$$\Pr(R_1 = r_1,\ldots,R_{n_1} = r_{n_1}|H_0) = \frac{1}{\binom{n}{n_1}}. \tag{2.13}$$

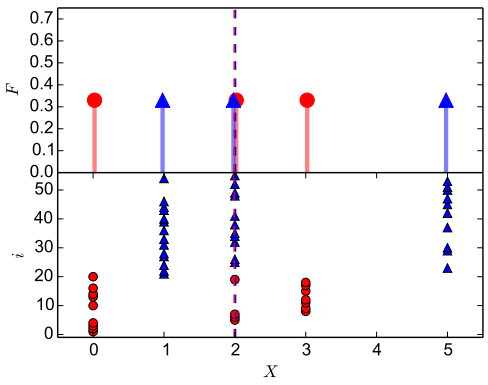
La distribution exacte de U sous H_0 peut être obtenue par une relation de récursion. Soit $P_{n_1,n_2}(U)$ la probabilité d'avoir la statistique U pour n_1 individus dans $\boldsymbol{X}_1$ et n_2 individus



(a) Distributions identiques et décalées : loi inverse-gamma de paramètres $\alpha = 3{,}0$ et $\beta = 2{,}0$. Les variables de $\boldsymbol{X}_2$ sont décalées de la valeur 0,5.



(b) Distributions normales centrées en 0 et de variances différentes : 0,2 pour $\boldsymbol{X}_1$ et 2,0 pour $\boldsymbol{X}_2$.



(c) Distributions discrètes de même médiane.

FIGURE 2.1 – Exemples d'échantillons $\boldsymbol{X}_1$ (ronds rouge), de taille $n_1 = 20$, et $\boldsymbol{X}_2$ (triangles bleus), de taille $n_2 = 35$. La partie supérieure des figures donne les fonctions de répartitions F_1 et F_2, et la partie inférieure donne une réalisation de ces échantillons. Les traits verticaux en pointillés sont les médianes.

dans $\boldsymbol{X}_2$. Alors, sous l'hypothèse nulle,

$$P_{n_1,n_2}(U) = \frac{n_1}{n} P_{n_1-1,n_2}(U-n) + \frac{n_2}{n} P_{n_1,n_2-1}(U). \tag{2.14}$$

Pour de petites valeurs de n_1 et n_2, la statistique U est facilement calculable et est tabulée. Comme le calcul exact des valeurs de U se complique quand les tailles des populations augmentent, l'approximation normale est utilisée à partir d'une dizaine d'observations (voir [Bellera *et al.*, 2010] pour une justification graphique). Sa distribution converge rapidement vers la loi normale, de moyenne

$$m_U = \frac{n_1 n_2}{2} \tag{2.15}$$

et de variance

$$\sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}. \tag{2.16}$$

Pour compenser l'approximation de la statistique U, discrète, par une loi continue, une correction de continuité peut être appliquée, afin de corriger les intervalles de confiance. Sous H_0, on a alors

$$\Pr(U \leq c) \approx \Phi\Big(\frac{c - m_U + \frac{1}{2}}{\sigma_U}\Big) \tag{2.17}$$

où $\Phi(a)$ est la mesure de l'aire sous la courbe de la densité de probabilité de la loi normale standard, à gauche de la valeur a [Lehmann et D'Abrera, 1975, chapitre 1.3. p. 14–16].

Lorsque les données sont discrètes, certaines valeurs peuvent être observées plusieurs fois dans les échantillons. Quand le nombre p de rangs attribués à plusieurs observations à la fois dans les deux échantillons devient trop important, la statistique U doit tenir compte des termes appariés. Une correction permet alors de prendre en compte ces répétitions de rangs. Comme expliqué dans [Siegel, 1956, p. 123–126], les rangs égaux sont remplacés par un rang intermédiaire moyen, et la variance de l'approximation normale devient

$$\sigma_{U,corr}^2 = \frac{n_1 n_2}{12}\Big(n + 1 - \sum_{i=1}^{p} \frac{t_i^3 - t_i}{n(n-1)}\Big), \tag{2.18}$$

où t_i est le nombre d'observations de même rang R_i. Cette correction a tendance à augmenter la valeur de la statistique, sans cette correction le test est donc plus conservatif.

Lorsque les observations sont générées selon la loi normale, le test t de Student est uniformément plus puissant que le test WMW. Ce dernier demeure toutefois intéressant : sa puissance est proche de celle du test de Student si les variances des distributions normales F_1 et F_2 sont égales, et l'écart est très faible avec des échantillons de petite taille. La puissance du test WMW est étudiée dans [Van der Vaart, 1950]. La robustesse des tests de rangs permet en outre d'analyser des distributions à queues lourdes, des échantillons avec des valeurs aberrantes ou encore des données censurées avec une bonne efficacité, ce qui en fait une solution avantageuse lorsque la distribution des données n'est pas connue.

2.3 Vraisemblance empirique

2.3.1 Définitions et propriétés

Après avoir décrit les tests bien connus de Student et de WMW, nous présentons notre test de Vraisemblance Empirique sur une Moyenne Empirique (VEME) comme une alternative aux approches paramétriques et non paramétriques. Cette partie de notre travail a été publiée dans [Harlé *et al.*, 2015]. La vraisemblance empirique a été initialement proposée par Owen dans [Owen, 1988]. Elle repose sur la fonction de vraisemblance empirique, construite entièrement sur les données, et s'affranchit donc d'une description paramétrique. Les données sont alors modélisées selon une loi multinomiale dont le nombre de paramètres

correspond à la taille de l'échantillon. Cette approche est qualifiée selon les auteurs de *non-paramétrique* ou *semi-paramétrique*, la dimension des paramètres étant asymptotiquement infinie. L'ouvrage de référence [Owen, 2010] y est consacré : cette modélisation permet de construire efficacement des tests et des intervalles de confiance. Elle s'adapte aux données, même déformées, biaisées ou incomplètes, ce qui en fait une méthode flexible. Il est de plus aisé d'introduire des contraintes spécifiques à un problème ou des informations connues a priori. On la rencontre souvent en économétrie [Kitamura, 2006], et plus récemment dans le domaine du traitement du signal [Harari-Kermadec et Pascal, 2008, Pascal *et al.*, 2010].

La vraisemblance empirique repose sur un modèle qui généralise la notion de fonction de répartition (FdR) empirique. La FdR empirique est un estimateur de la FdR qui, pour un échantillon $\boldsymbol{X} = (X_1, \ldots, X_n)$, est défini par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leqslant x}, \tag{2.19}$$

où $\mathbb{1}_A$ est la fonction qui vaut 1 lorsque la condition A est vérifiée et 0 sinon. Cet estimateur équivaut à placer une masse de probabilité $\frac{1}{n}$ en chaque X_i. La vraisemblance empirique est définie comme la probabilité d'obtenir exactement les $X_1, \ldots, X_n$ pour une FdR des variables F donnée :

$$L(F) = \prod_{i=1}^{n} (F(X_i) - F(X_{i-})), \tag{2.20}$$

où $F(x_{i-}) = \Pr(X < x_i)$ lorsque X est distribuée selon F. Cette vraisemblance est non nulle à condition de placer une masse de probabilité en chaque X_i, ces masses n'étant pas nécessairement équiprobables. En notant, pour $1 \leq i \leq n$,

$$p_i = F(X_i) - F(X_{i-}),$$

la masse de probabilité en X_i, $L(F)$ se réexprime comme

$$L(F) = \prod_{i=1}^{n} p_i, \tag{2.21}$$

ce qui correspond à modéliser les variables par une loi multinomiale dont le nombre de paramètres est le nombre de variables $X_1, \ldots, X_n$. Pour $1 \leq i \leq n$, les masses de probabilité p_i en chaque X_i sont supposées strictement positives, afin que la vraisemblance L soit non nulle, mais non nécessairement équiprobables. Cette fonction de vraisemblance L est maximisée lorsque les masses sont équiprobables, c'est-à-dire pour F_n. En effet si $F \neq F_n$, alors $L(F) < L(F_n)$ [Owen, 2010, Théorème. 2.1, p. 18].

En s'inspirant du rapport de vraisemblance paramétrique usuel, comme ceux présentés dans le chapitre 1, Owen introduit le rapport de vraisemblance empirique pour des fonctions F et F_n données :

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i. \tag{2.22}$$

La normalisation par la vraisemblance $L(F_n)$, obtenue à partir de la loi empirique, permet de calibrer le rapport quand la taille de l'échantillon n devient grande, tandis que la vraisemblance $L(F)$ tend vers 0. Une statistique de test pour un paramètre d'intérêt donné, basée sur ce rapport, peut alors être construite.

L'étude de l'échantillon $\boldsymbol{X}$ peut se faire à travers l'estimation des masses de probabilité p_i, associées aux X_i. Les poids recherchés sont ceux qui maximisent le rapport (2.22) sous des contraintes sur la distribution. Selon l'analyse à effectuer sur les données, par exemple l'estimation d'une région de confiance sur un paramètre, des contraintes sont en effet ajoutées lors de la résolution du problème de maximisation. Pour construire notre test d'homogénéité, le paramètre d'intérêt est la moyenne μ. Tester une valeur possible μ de cette moyenne revient à se restreindre aux poids tels que $\sum_{i=1}^n p_i X_i = \mu$, ce qui se traduit par une contrainte linéaire sur ces paramètres de poids. Les contraintes linéaires usuelles sur les masses d'une distribution (positivité et sommation à un des p_i) sont également ajoutées. Ceci conduit à considérer la fonction de profil de vraisemblance empirique suivante, pour une moyenne μ donnée :

$$\mathcal{R}(\mu) = \max_{\boldsymbol{p}} \left\{ \prod_{i=1}^n np_i \;\middle|\; \sum_{i=1}^n p_i X_i = \mu, p_i > 0, \sum_{i=1}^n p_i = 1 \right\}, \tag{2.23}$$

où $\boldsymbol{p}$ est le vecteur des poids. Comme cette fonction de profil est définie pour une valeur de la moyenne μ donnée, la solution où les poids sont équiprobables ($p_i = \frac{1}{n}$, pour tout $1 \leq i \leq n$) et qui permet de maximiser le rapport de vraisemblance (2.22) en vérifiant les conditions $p_i > 0$ et $\sum_{i=1}^n p_i = 1$, mais ne respecte pas la contrainte $\sum_{i=1}^n p_i X_i = \mu$ si μ n'est pas égale à la moyenne empirique. Owen a également établi le rapport pour d'autres paramètres comme la variance ou la corrélation [Owen, 1990].

Le théorème suivant, démontré dans [Owen, 1988] dans le cas univarié, puis dans [Owen, 2010] pour le cas général, est une version non paramétrique du théorème de Wilks 1.1.1 :

Théorème 2.3.1 (Vraisemblance empirique univariée)**.** *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et de même distribution. Soit μ_0 l'espérance de ces variables, et on suppose que leur variance est finie et non nulle. Alors $-2\ln(\mathcal{R}(\mu_0))$ converge en distribution quand $n \to \infty$ vers la loi du χ^2 à un degré de liberté, notée $\chi^2_{(1)}$.*

Ce résultat clef permet notamment de définir des intervalles de confiance (asymptotiques) pour le paramètre d'intérêt μ, comme avec le rapport de vraisemblances paramétrique, sous la forme

$$C = \{\mu | \mathcal{R}(\mu) \geq r_0\}. \tag{2.24}$$

Il permet aussi de formuler des tests d'hypothèses et de déterminer la loi asymptotique pour notre statistique de test. Une correction de Bartlett [DiCiccio *et al.*, 1991] peut être appliquée sur la fonction $-2\ln(\mathcal{R}(\mu))$ pour différents paramètres dont la moyenne. Ainsi l'erreur de recouvrement de la statistique, de l'ordre de $O(n^{-1})$, est réduite à $O(n^{-2})$.

La figure 2.2 illustre la façon dont les contours de la fonction de profil s'adaptent aux données. Cet exemple reprend la figure 1 de l'article [Owen, 1990], sur des données provenant de [Marx et Larsen, 2006] sur l'étude de 11 canards mâles issus d'un croisement entre

deux espèces. Les observations sont bidimensionnelles : deux indices mesurent la proximité de deux traits de caractères avec ceux typiques des espèces croisées. Les caractéristiques étudiées sont le plumage et le comportement d'un canard. On vérifie bien la nature de la vraisemblance empirique guidée par les données, en observant la déformation des contours en fonction des positions des observations dans le plan des indices de plumage et de comportement, notamment vers les mesures les plus éloignées.

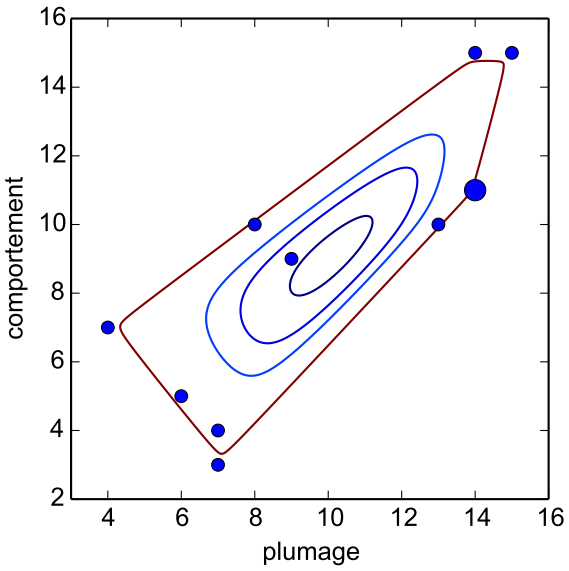


FIGURE 2.2 – Contour de la fonction de profil de vraisemblance empirique $\mathcal{R}(\mu)$ pour l'estimation du paramètre μ sur des données de plumage et de comportement de 11 canards mâles. Ces derniers sont issus d'un croisement entre des canards colverts (indices de plumage 0 et de comportement 0) et des canards pilets (indices de plumage 20 et de comportement 15). Comme il y a deux observations en (14,11), ce point est représenté par un marqueur plus gros que les autres. Les contours sont obtenus pour les niveaux de confiance 50%, 90%, 95% et 99%, dans l'ordre du contour interne (bleu foncé) au contour externe (marron).

2.3.2 Résolution

Pour une moyenne μ donnée, le problème d'optimisation sous contrainte (2.23) doit être résolu. Lorsque μ n'appartient pas à l'enveloppe convexe des données $\mathcal{H} = \{X_1, \ldots, X_n\}$, il n'existe pas d'ensemble de poids p_i qui répondent aux contraintes du problème. La valeur la plus petite de $\boldsymbol{X}$ est notée $X_{(1)}$ et la valeur la plus grande $X_{(n)}$. Par convention, Owen attribue les valeurs $\mathcal{R}(\mu) = 0$ si $\mu < X_{(1)}$ et $\mathcal{R}(\mu) = \infty$ si $\mu > X_{(n)}$. De plus, si $\mu = X_{(1)} < X_{(n)}$ ou si $\mu = X_{(n)} > X_{(1)}$, alors $\mathcal{R}(\mu) = 0$, et si $\mu = X_{(1)} = X_{(n)}$, alors $\mathcal{R}(\mu) = 1$. En revanche, dès lors que μ est un point intérieur à $\mathcal{H}$, l'ensemble des poids p_i qui satisfont les contraintes est convexe et non vide. Le logarithme de la fonction objectif

$\sum_{i=1}^{n} \ln(np_i)$ étant strictement concave sur l'ensemble convexe des poids p_i, il existe alors un unique maximum. Le problème est résolu par la méthode du multiplicateur de Lagrange.

Le lagrangien du problème dual s'écrit alors pour $\lambda \in \mathbb{R}$ et $\eta \in \mathbb{R}$ comme :

$$H(\boldsymbol{p},\lambda,\eta) = \sum_{i=1}^{n} \ln(np_i) - n\lambda \sum_{i=1}^{n} p_i(X_i - \mu) + \eta(\sum_{i=1}^{n} p_i - 1). \tag{2.25}$$

Ce terme est d'abord dérivé par rapport aux p_i. La dérivée partielle est annulée pour obtenir :

$$\frac{1}{p_i} - n\lambda(X_i - \mu) - n\gamma = 0. \tag{2.26}$$

Pour tous les poids, on a donc

$$\sum_{i=1}^{n}(1 - n\lambda p_i(X_i - \mu) - n\gamma p_i) = 0, \tag{2.27}$$

d'où, après application des contraintes,

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda(X_i - \mu)}, \tag{2.28}$$

pour $1 \leq i \leq n$, avec $\eta = -n$. Le multiplicateur de Lagrange λ est obtenu numériquement comme la solution de

$$\frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \mu}{1 + \lambda(X_i - \mu)} = 0. \tag{2.29}$$

Pour ce faire, une fonction pseudo-logarithmique est définie, qui pénalise les solutions λ où les poids p_i deviennent négatifs, tout en permettant de travailler sur une fonction convexe. Dans [Owen, 2010, chapitre 3.14], l'auteur suggère d'employer la méthode de Brent ou la méthode de Newton. La résolution s'applique de la même façon à des données multivariées $X_i \in \mathbb{R}^d$, $\lambda \in \mathbb{R}^d$ est alors le vecteur des multiplicateurs de Lagrange.

2.3.3 Méthodes de l'état de l'art

Les propriétés asymptotiques du rapport de vraisemblance empirique en font une fonction intéressante pour l'établissement d'un test à deux échantillons $\boldsymbol{X}_1 = (X_1, \ldots, X_{n_1})$ et $\boldsymbol{X}_2 = (X_{n_1+1}, \ldots, X_n)$, de moyennes respectives μ_1 et μ_2, et dont les masses de probabilités sont notées p_i pour $\boldsymbol{X}_1$, $1 \leq i \leq n_1$, et p_j pour $\boldsymbol{X}_2$, $n_1 + 1 \leq j \leq n$. Dans l'article [Jing, 1995], l'auteur cherche à établir un test pour déterminer si les échantillons $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ ont la même moyenne. Il évalue le rapport de vraisemblance empirique, défini en (2.22), pour chacun des deux échantillons de fonctions de répartition empiriques F_1 et F_2 respectivement

$$R_1(F_1) = \prod_{i=1}^{n_1} n_1 p_i \qquad \text{et} \qquad R_2(F_2) = \prod_{j=n_1+1}^{n} n_2 p_j, \tag{2.30}$$

où $\sum_{i=1}^{n_1} p_i = 1$, $\sum_{j=n_1+1}^{n} p_j = 1$ et $p_i,p_j > 0$. Une fonction de profil $\mathcal{R}(\theta)$ est proposée pour déterminer, pour un écart entre les moyennes $\theta = \mu_2 - \mu_1$ donné, l'ensemble des probabilités p_1 et p_2 qui maximisent les deux rapports de vraisemblance empiriques en (2.30). Les contraintes sur les poids sont $\sum_{i=1}^{n_1} p_i = 1$, $\sum_{j=n_1+1}^{n} p_j = 1$ et $p_i,p_j > 0$, ainsi que la condition sur l'écart entre les moyennes $\sum_{j=n_1+1}^{n} p_j X_j - \sum_{i=1}^{n_1} p_i X_i = \theta$. Jing fournit alors un théorème sur la convergence asymptotique du terme $-2\ln\mathcal{R}(\theta)$ vers la loi $\chi^2_{(1)}$ lorsque $\theta = \theta_0$, sous l'hypothèse que l'écart entre les moyennes est θ_0 et que les variances de F_1 et F_2 sont finies. Cette statistique s'avère de plus corrigeable par des coefficients de Bartlett. Les probabilités p_i pour $\boldsymbol{X}_1$ et p_j pour $\boldsymbol{X}_2$ sont obtenues par la méthode du multiplicateur de Lagrange

$$p_i = \frac{1}{n_1}\frac{1}{1 - \frac{n}{n_2}\lambda(X_i - \mu_1)}, \qquad p_j = \frac{1}{n_2}\frac{1}{1 - \frac{n}{n_1}\lambda(X_j - \mu_2)}. \tag{2.31}$$

Owen montre un résultat similaire pour la convergence vers une loi du $\chi^2_{(d)}$ dans le cas plus général d'échantillons multiples et multivariés à d dimensions [Owen, 2010, chapitre 11.4 p. 230].

L'article [Guan, 2004] présente un estimateur pour la détection d'un changement entre les deux distributions F_1 et F_2, pour le cas où F_2 est une fonction pondérée de F_1. La position de la rupture est obtenue par minimisation d'une statistique dont les valeurs asymptotiques sous H_0 et sous H_1 sont obtenues sous certaines conditions. Ces résultats sont applicables pour la détection d'un saut de moyenne. Plus récemment, la méthode présentée dans [Zou *et al.*, 2007] reprend les résultats établis dans [Jing, 1995] pour détecter une unique rupture dans un signal à une position inconnue. La contrainte d'égalité des moyennes est ajoutée ($\theta_0 = 0$) pour correspondre à l'hypothèse nulle. La statistique de test est

$$Z(k) = \max_{k/n}\{\mathcal{R}(k/n)\} \tag{2.32}$$

où k est la position de la rupture testée, c'est-à-dire la taille du premier échantillon, et la convergence de la distribution nulle asymptotique est prouvée. La position de la rupture correspond au $k^\star$ qui maximise $Z(k)$. Un inconvénient de cette statistique est qu'elle est indéfinie lorsque l'estimateur de la moyenne $\hat{\mu}$ n'est pas dans l'enveloppe convexe $\mathcal{H}_k$ des données de l'un des deux échantillons, noté $\boldsymbol{X}_k$, ce qui est d'autant plus susceptible de se produire la taille n_k est faible. En effet, quand $\hat{\mu}$ n'appartient à $\mathcal{H}_k$, il n'existe pas de masses $p_i \geq 0$ telles que $\sum_{i=1}^{n_k} p_i = 1$ et $\sum_{i=1}^{n_k} p_i X_i = \hat{\mu}$. Afin de limiter ce risque, les auteurs de [Zou *et al.*, 2007] suggèrent de restreindre le domaine de k à l'intervalle entre deux bornes suffisamment éloignées des extrémités du signal.

Le cas épidémique est étudié dans [Ning *et al.*, 2012]. La moyenne subit un décalage de la valeur μ_1 à la valeur μ_2 sur une portion de signal comprise entre les indices n'_1 et n'_2. Le premier échantillon, noté $\boldsymbol{X}'_1$, est alors formé par les variables X_i avant la première rupture et après la deuxième rupture, et le deuxième échantillon, noté $\boldsymbol{X}'_2$, est constitué des variables X_j entre les deux ruptures. On note $I = \{1,\ldots,n'_1,n'_2+1,\ldots,n\}$ et $J = \{n'_1+1,\ldots,n'_2\}$ les ensembles des indices associés à $\boldsymbol{X}'_1$ et $\boldsymbol{X}'_2$ respectivement. Le rapport

de vraisemblance empirique est alors de la forme

$$R(\mu_1,\mu_2,n'_1,n'_2) = \left(\prod_{i\in I}(n - n'_2 + n'_1)p_i\right)\left(\prod_{j\in J}(n'_2 - n'_1)p_j\right), \tag{2.33}$$

et on en déduit la fonction de profil $\mathcal{R}(\mu_1,\mu_2,n'_1,n'_2)$ en ajoutant les contraintes $\sum_{i\in I} p_i X_i = \mu_1$, $\sum_{j\in J} p_j X_j = \mu_2$ et $\sum_{i\in I} p_i = \sum_{i\in J} p_j = 1$. On teste si $\mu_1 = \mu_2$. Des fonctions de score permettent d'obtenir les estimations de la moyenne μ_1 et des positions n'_1 et n'_2 du décalage de moyenne. La consistance de la statistique de test qui en découle et sa distribution asymptotique sous H_0 sont démontrées asymptotiquement, quand les dimensions des échantillons tendent vers l'infini. Comme pour le test de [Zou *et al.*, 2007], les auteurs [Ning *et al.*, 2012] précisent que ce test n'est pas défini près des extrémités du signal, où la taille de l'un des deux échantillons devient trop petite.

Une série de tests est proposée dans [Einmahl et McKeague, 2003], pour tester la symétrie d'une distribution, le changement de distribution entre deux échantillons, l'indépendance et la distribution exponentielle. Ces tests sont dérivés d'une statistique commune faisant intervenir un rapport de vraisemblance local. Une statistique est notamment introduite pour détecter une rupture à la position n_1, qui scinde le signal en deux portions $\boldsymbol{X}_1 = (X_1,\ldots,X_{n_1})$ et $\boldsymbol{X}_2 = (X_{n_1+1},\ldots,X_n)$. Sa distribution asymptotique sous l'hypothèse nulle est donnée lorsque les distributions des échantillons sont continues. Cette approche diffère de la précédente en ce qu'elle n'implique pas de résoudre un problème de maximisation comme (2.23) pour estimer les poids p_i. Le test correspond à un test de rapport de vraisemblance, où intervient l'estimateur de la fonction de répartition empirique.

2.3.4 Test d'hypothèses VEME

Le test d'homogénéité que nous souhaitons établir à partir de la vraisemblance empirique doit être sensible à une différence entre les moyennes des fonctions de répartitions F_1 et F_2 des deux échantillons $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$. Dans le cadre de la détection de ruptures, il s'agit par exemple de déterminer si les n_1 premières variables dans une série temporelle $\boldsymbol{X}$ de $n = n_1 + n_2$ points, constituant l'échantillon $\boldsymbol{X}_1$, ont la même moyenne que les n_2 variables suivantes, qui constituent l'échantillon $\boldsymbol{X}_2$. Les hypothèses nulle et alternative sont celles données en (2.4). Les variables des deux échantillons sont supposées indépendantes et identiquement distribuées selon F_1 pour les variables de $\boldsymbol{X}_1$ et selon F_2 pour les variables de $\boldsymbol{X}_2$. On suppose également que les variances σ_1^2 et σ_2^2 sont égales. Nous proposons d'évaluer une fonction de profil de la forme (2.23) sur l'échantillon global $\boldsymbol{X}$, composé de $\boldsymbol{X}_1$ et de $\boldsymbol{X}_2$ et de taille $n = n_1 + n_2$, pour la moyenne de l'un des deux échantillons, puisque sous H_0 $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ ont la même moyenne. Comme les moyennes μ_1 et μ_2 sont inconnues, c'est l'estimateur $\overline{X}_i$ de la moyenne empirique de l'un des échantillons, par exemple $\boldsymbol{X}_1$, qui est employée. Le test VEME, pour Vraisemblance Empirique sur une Moyenne Empirique, consiste donc à estimer

$$\mathcal{R}(\overline{X}_1) = \max_{\boldsymbol{p}} \left\{ \prod_{i=1}^{n} np_i \;\middle|\; \sum_{i=1}^{n} p_i X_i = \overline{X}_1, p_i > 0, \sum_{i=1}^{n} p_i = 1 \right\}. \tag{2.34}$$

Cette statistique est obtenue pour la moyenne empirique de l'un des deux échantillons, elle est exprimée ici en fonction de $\overline{X}_1$, sans perte de généralité. L'un des avantages de ce test est que $\overline{X}_1$ appartient par construction à l'enveloppe convexe des données $\mathcal{H}$, de ce fait la statistique (2.34) est toujours définie, même pour de petits échantillons, contrairement aux tests précédents.

Rejeter l'hypothèse nulle revient maintenant à déterminer si $\mathcal{R}(\overline{X}_1) < z$ pour un seuil z donné. Afin de contrôler le risque de première espèce à un seuil de signification α donné, il est nécessaire de connaître la loi de $\mathcal{R}(\overline{X}_1)$ sous H_0. Sous l'hypothèse d'homoscédasticité, ce contrôle est assuré par le résultat asymptotique suivant :

Proposition 2.3.1. *On suppose que sous H_0 les variances des lois F_1 et F_2 existent et sont identiques. Quand $n \to \infty$, c'est-à-dire quand n_1 et n_2 tendent vers l'infini, on note $\gamma = lim_{n\to+\infty}\frac{n_2}{n_1}$, et l'on suppose que $0 < \gamma < +\infty$. Alors, sous H_0, $-2\frac{n_1}{n_2}\ln(\mathcal{R}(\overline{X}_1))$ converge en distribution quand $n \to \infty$ vers la loi du χ^2 à un degré de liberté, notée $\chi^2_{(1)}$.*

Démonstration. La preuve reprend les éléments de celle du théorème 2.3.1 donnée dans [Owen, 2010, p. 226] : une majoration du multiplicateur de Lagrange λ introduit dans (2.25) permet d'obtenir d'après l'expression des probabilités p_i donnée en (A.3) le développement suivant lorsque $n \to +\infty$:

$$-2\ln(\mathcal{R}(\overline{X}_1)) = n\frac{\left(\overline{X}-\overline{X}_1\right)^2}{S^2} + o_p(1), \tag{2.35}$$

où $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ est la moyenne empirique de l'échantillon global $\boldsymbol{X}$ et où

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_1)^2. \tag{2.36}$$

Cette partie est détaillée dans l'annexe A. Il suffit donc de montrer

$$\frac{n_1}{n_2}\frac{n\left(\overline{X}-\overline{X}_1\right)^2}{S^2} \xrightarrow[n\to+\infty]{Loi} \chi^2_{(1)}. \tag{2.37}$$

Par hypothèse, les moyennes empiriques $\overline{X}_1$ et $\overline{X}_2$ sont indépendantes, et d'après le théorème central limite elles convergent en distribution vers les lois normales $\mathcal{N}(\mu_1, \frac{\sigma_1^2}{n_1})$ et $\mathcal{N}(\mu_2, \frac{\sigma_2^2}{n_2})$ lorsque n_1 et n_2 tendent vers l'infini. On note $\sigma^2 = \text{var}(\boldsymbol{X})$ la variance des variables sous l'hypothèse nulle, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. La variable $U = \sqrt{n}\left(\overline{X} - \overline{X}_1\right) = \frac{n_2}{\sqrt{n}}\left(\overline{X}_2 - \overline{X}_1\right)$ est centrée et de variance $\text{var}\,(U) = \frac{n_2}{n_1}\sigma^2$. Par conséquent $\sqrt{\frac{n_1}{n_2}}\frac{U}{\sigma}$ converge en distribution vers la loi normale standard, et $\frac{n_1}{n_2}\frac{n\left(\overline{X}-\overline{X}_1\right)^2}{\sigma^2}$ converge donc en distribution vers la loi $\chi^2_{(1)}$. De plus S^2 est une moyenne empirique de n variables telle que

$$\begin{aligned}
E[S^2] &= \frac{1}{n}\left(E\left[\sum_{i=1}^{n_1}(X_i - \overline{X}_1)^2\right] + E\left[\sum_{i=n_1+1}^{n}(X_i - \overline{X}_1)^2\right]\right), \\
&= \frac{1}{n}\left((n_1-1)\sigma^2 + \left(n_2 + \frac{n_2}{n_1}\right)\sigma^2\right) \\
&= \sigma^2 + O\left(\tfrac{1}{n}\right).
\end{aligned}$$

Sous des conditions assez faibles de régularité, S^2 converge en probabilité vers σ^2 lorsque n tend vers l'infini. C'est par exemple une conséquence de l'inégalité de Tchebychev dès lors que $E[X_i^4] < +\infty$, pour $1 \leq i \leq n$. En effet, en posant

$$S^2 = \frac{1}{n}(n_1 S_1^2 + n_2 S_2^2) \tag{2.38}$$

avec

$$S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_i - \overline{X}_1)^2 \quad \text{et} \quad S_2^2 = \frac{1}{n_2}\sum_{j=n1+1}^{n}(X_j - \overline{X}_1)^2, \tag{2.39}$$

il vient que

$$\operatorname{var}(S^2) = \frac{1}{n^2}\left(n_1^2 \operatorname{var}(S_1^2) + n_2^2 \operatorname{var}(S_2^2) + n_1 n_2 \operatorname{cov}(S_1^2, S_2^2)\right). \tag{2.40}$$

S_1^2 est la variance empirique de l'échantillon $\boldsymbol{X}_1$. On peut montrer que sa variance s'écrit

$$\operatorname{var}(S_1^2) = \frac{(n_1-1)^2}{n_1^3}\mu_4 - \frac{(n_1-1)(n_1-3)}{n_1^3}\sigma^4, \tag{2.41}$$

où $\mu_4 = E[(X_i - \mu)^4]$ est le moment centré d'ordre 4 de la variable X_i qui est fini par hypothèse. Par conséquent,

$$\frac{n_1^2}{n^2}\operatorname{var}(S_1^2) = O\left(\frac{1}{n}\right). \tag{2.42}$$

Des calculs similaires montrent également que

$$\frac{n_2^2}{n^2}\operatorname{var}(S_2^2) = O\left(\frac{1}{n}\right), \tag{2.43}$$

$$\frac{n_1 n_2}{n^2}\operatorname{cov}(S_1^2, S_2^2) = O\left(\frac{1}{n}\right), \tag{2.44}$$

et donc $\operatorname{var}(S^2) = O\left(\frac{1}{n}\right)$ d'après (2.40). L'inégalité de Tchebychev permet d'écrire que pour tout $a > 0$,

$$\Pr(|S^2 - E[S^2]| \geq a) \leq \frac{\operatorname{var}(S^2)}{a^2}, \tag{2.45}$$

or

$$\lim_{n\to+\infty} \Pr(|S^2 - E[S^2]| \geq a) = \Pr(|S^2 - \sigma^2| \geq a) \tag{2.46}$$

et

$$\lim_{n\to+\infty} \frac{\operatorname{var}(S^2)}{a^2} = 0 \tag{2.47}$$

donc

$$\lim_{n\to+\infty} \Pr(|S^2 - E[S^2]| \geq a) = 0, \tag{2.48}$$

ce qui prouve la convergence en probabilité de S^2 vers σ^2 lorsque n tend vers l'infini. Le théorème de Slutsky assure alors la convergence écrite en (2.37), ce qui prouve le résultat énoncé. □

Ce résultat permet ainsi de calibrer la statistique de test sous H_0 afin de déterminer pour un seuil de signification α donné la valeur du seuil $z(\alpha)$ qui permet de rejeter H_0. Il faut noter qu'un résultat similaire peut également être obtenu sous l'hypothèse d'hétéroscédasticité, en injectant les variances empiriques de chaque segment dans le facteur de calibration.

La statistique de test (2.34) consiste à évaluer le profil de vraisemblance empirique (2.23) pour la moyenne empirique de l'un des deux échantillons. Cette statistique présente une asymétrie. Il convient de déterminer lequel des deux échantillons $\boldsymbol{X}_1$ ou $\boldsymbol{X}_2$ doit être retenu afin de maximiser les performances du test VEME. En effet, dans le cas de figure général où les distributions et des tailles de $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ peuvent être assez différentes, les deux statistiques $-2\frac{n_1}{n_2}\ln(\mathcal{R}(\overline{X}_1))$ et $-2\frac{n_2}{n_1}\ln(\mathcal{R}(\overline{X}_2))$ (où $\overline{X}_1$ est remplacé par $\overline{X}_2$ dans l'expression (2.34)) ne sont pas égales. L'emploi d'une statistique de test symétrique, de la forme $-2\frac{n_1}{n}\ln(\mathcal{R}(\overline{X}_1))-2\frac{n_2}{n}\ln(\mathcal{R}(\overline{X}_2))$, fournit un test intermédiaire, pénalisé par la moins bonne des deux statistiques asymétriques. Nous allons plutôt chercher à identifier lequel des deux échantillons conduit au test le plus performant avec la statistique de test asymétrique. Déterminer l'échantillon, noté $\boldsymbol{X}_1$ par défaut, qui garantit la plus grande puissance quelle que soit l'alternative est un problème théorique difficile. En pratique, plusieurs stratégies sont possibles. $\boldsymbol{X}_1$ peut être désigné par exemple comme l'échantillon remplissant l'une des conditions suivantes :

— $\mathcal{R}(\overline{X}_1) < \mathcal{R}(\overline{X}_2)$,
— $\mathcal{R}(\overline{X}_1) > \mathcal{R}(\overline{X}_2)$,
— $S_1^2 < S_2^2$,
— $S_1^2 > S_2^2$,
— $n_1 < n_2$,
— $n_1 > n_2$,

ce qui revient à comparer les statistiques, les variances empiriques ou les tailles. Ces six critères sont comparés empiriquement dans la partie 2.4.1.

Dans le cas homoscédastique, nous proposons d'opter pour l'échantillon le plus grand, $\boldsymbol{X}_1$ est tel que $n_1 \geq n_2$. Ce choix simple, discuté dans la partie 2.4, s'avère en effet puissant et robuste pour un large ensemble d'alternatives testées et est indépendant des variables des échantillons. La calibration de la proposition 2.3.1 donne donc directement la loi de $\mathcal{R}(\overline{X}_1)$ sous H_0. Notons que dans le cas d'échantillons appariés ($n_1 = n_2$), $\boldsymbol{X}_1$ est désigné de manière arbitraire ou aléatoire. Finalement, la procédure du test VEME ainsi défini se

résume par les étapes de l'algorithme 1.

Algorithme 1 : Procédure du test VEME

Données : échantillons $\boldsymbol{X}_i$ et $\boldsymbol{X}_j$, seuil de signification α
Décision : H_0 acceptée ou rejetée
Choix du vecteur $\boldsymbol{X}_1$
Calcul de la moyenne empirique des réalisations de $\boldsymbol{X}_1$: $\overline{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$
Résolution du problème dual (2.25) : les poids p_i sont obtenus d'après l'expression (A.3) et le paramètre λ est estimé numériquement
Calcul de la réalisation de la statistique de test $t = -2\frac{n_1}{n_2}\ln(\mathcal{R}(\overline{x}_1))$
Calcul du quantile $1-\alpha$ de la loi $\chi^2_{(1)}$, $z_{1-\alpha}$
si $t \geq z_{1-\alpha}$ **alors**
| H_0 est rejetée
fin
sinon
| H_0 est acceptée
fin

Ce test diffère des approches de l'état de l'art basées sur la vraisemblance empirique, présentées précédemment. Le rapport de vraisemblance empirique est obtenu pour l'échantillon global, pour l'une des deux moyennes empiriques, calculable aisément. Comme il a déjà été signalé, la statistique (2.34) présente l'avantage d'être toujours définie quelles que soient les tailles $n_1 \geq 1$ et $n_2 \geq 1$ des échantillons testés, puisque par construction $\overline{X}_1$ appartient toujours à $\mathcal{H}$. C'est une différence majeure par rapport aux tests de vraisemblance empirique proposés dans la littérature [Jing, 1995, Ning *et al.*, 2012]. Ces derniers reposent classiquement sur le profil du produit des vraisemblances empiriques des deux échantillons $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ sous la contrainte d'égalité des moyennes $\mu_1 = \mu_2 = \mu$. Comme le signalent les auteurs de [Zou *et al.*, 2007, Ning *et al.*, 2012], les statistiques de test s'avèrent indéfinies dès lors que l'estimateur de la moyenne $\hat{\mu}$ n'appartient pas à l'enveloppe convexe des données d'au moins un des deux échantillons, ce qui peut se produire quand l'échantillon est trop petit.

2.4 Comparaison

Ces exemples de tests paramétriques, semi-paramétriques et non paramétriques sont mis en œuvre dans une série de tests, afin d'établir et de comparer les performances obtenues dans différents cas de figure. L'objectif est de déterminer quelle statistique offre des propriétés intéressantes pour le modèle de détection de rupture que nous souhaitons développer. Deux critères en particulier sont mis en avant : le comportement face à des échantillons de petite taille, et la robustesse aux distributions non normales, sans a priori sur les distributions. À chaque simulation, un signal de $n = 100$ points est généré d'abord sans rupture, $\boldsymbol{x}_{H_0} = (x_1, \ldots, x_n)$, et les observations sont i.i.d. selon la distribution F_1, puis avec une rupture au point n_1, au-delà duquel les observations suivent la distribution F_2,

$\boldsymbol{x}_{H_1} = (x_1, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_n)$. Pour comparer les puissances des tests, les résultats sont présentés sous la forme de courbes de sensibilité et spécificité, ou courbes ROC, de l'anglais *Receiver Operating Characteristic*, donnant la probabilité de détection (PD) en fonction de la probabilité de fausse alarme (PFA), définies en (1.7). Les tests comparés sont :

- le test de Student, pour lequel les observations sont supposées être distribuées normalement et de même variance ;
- le test de Welch, pour lequel les observations sont supposées être distribuées normalement, avec des variances différentes ;
- notre test VEME, présenté dans la partie 2.3.4 ;
- le test sur la vraisemblance empirique à deux échantillons de [Jing, 1995] (voir la partie 2.3.3), qu'on nomme VE seg ;
- le test de rang WMW.

2.4.1 Test VEME : critère pour désigner x_1

Dans le test de VEME, le calcul de statistique $t = -2\frac{n_1}{n_2}\ln(\mathcal{R}(\overline{x}_1))$ fait intervenir la moyenne empirique de l'échantillon $\boldsymbol{x}_1$. Il convient d'établir un critère pour le choix de l'échantillon. Pour ce faire, les solutions listées dans la partie 2.3.4 sont comparées empiriquement. Dans ce paragraphe, l'indice (1) fait référence aux éléments ayant trait au premier segment du signal traité, et l'indice (2) se rapporte au deuxième segment. Pour le critère sur la comparaison des statistiques, $t_{(1)}$ et $t_{(2)}$ sont obtenues en fonction de leurs moyennes empiriques :

$$t_{(1)} = -2\frac{n_{(1)}}{n_{(2)}}\ln(\mathcal{R}(\overline{x}_{(1)})), \quad t_{(2)} = -2\frac{n_{(2)}}{n_{(1)}}\ln(\mathcal{R}(\overline{x}_{(2)})), \tag{2.49}$$

où la fonction de profil $\mathcal{R}(\overline{x}_{(i)})$, définie en (2.23), est calculée sur l'échantillon $\boldsymbol{x}_{(i)}$. Les variances sont comparées à partir de leur estimateur empirique. Dans les simulations suivantes, les tailles des échantillons sont $n_{(1)} = 70$ et $n_{(2)} = 30$. La puissance du test est illustrée par les courbes ROC, et l'adéquation de la statistique de test t avec la loi du $\chi^2_{(1)}$ est illustrée par un diagramme quantile-quantile.

La première simulation est réalisée dans le cas normal, où les échantillons ont pour moyenne $\mu_{(1)} = 0{,}0$ et $\mu_{(2)} = 1{,}0$ et sont de même variance $\sigma^2_{(1)} = \sigma^2_{(2)} = 5{,}0$. Un exemple de tels échantillons est donné dans la figure 2.3a. Les courbes ROC obtenues pour les différents critères sont tracées dans la figure 2.3b, elles sont toutes confondues. D'autres tests, non représentés ici, ont été menés en changeant les valeurs de la variance, et conduisent à des résultats similaires, où le test VEME a les mêmes performances pour tous les critères proposés. Le cas hétéroscédastique est alors abordé. Les caractéristiques des échantillons sont $\mu_{(1)} = 0{,}0$, $\sigma^2_{(1)} = 1{,}0$ et $\mu_{(2)} = 1{,}0$, $\sigma^2_{(2)} = 10{,}0$. L'échantillon $\boldsymbol{x}_{(1)}$ est le plus grand, et sa variance est la plus petite (voir la figure 2.4a). Les courbes ROC des différents critères sont données dans la figure 2.4b : les résultats sont assez similaires pour tous les critères, ceux de la plus petite variance et de la plus grande taille sont légèrement supérieurs aux autres. Les valeurs des variances sont ensuite interverties à la figure 2.5, le signal le plus grand a maintenant la plus grande variance (voir la figure 2.5a). Les résultats, donnés

à la figure 2.5b, montrent que les critères testés conduisent aux mêmes performances de détection. Ces expériences avec des distributions normales ne permettent pas de mettre en valeur un critère significativement meilleur que les autres pour le choix de la moyenne empirique de référence.

D'autres distributions sont testées. La figure 2.6 donne les courbes ROC pour des distributions inverse-gamma. Dans le cas homoscédastique, elles sont de mêmes paramètres $\alpha = 3{,}0$ et $\beta = 2{,}0$ pour $\boldsymbol{x}_{(1)}$ et $\boldsymbol{x}_{(2)}$, les observations de $\boldsymbol{x}_{(2)}$ étant décalées de 0,5 dans la figure 2.6a et de -0.5 dans 2.6b. Dans les deux cas, les meilleurs performances sont atteintes avec le critère du maximum de la variance, suivi de la taille la plus grande dans la figure 2.6a et à égalité avec la taille la plus petite dans la figure 2.6b. Dans le cas hétéroscédastique, les paramètres sont $\alpha_{(1)} = 3{,}0$, $\beta_{(1)} = 2{,}0$ pour $\boldsymbol{x}_{(1)}$ et $\alpha_{(2)} = 3{,}0$, $\beta_{(2)} = 3{,}0$ pour $\boldsymbol{x}_{(2)}$ dans la figure 2.6c, et $\alpha_{(1)} = 3{,}0$, $\beta_{(1)} = 3{,}0$ pour $\boldsymbol{x}_{(1)}$ et $\alpha_{(2)} = 3{,}0$, $\beta_{(2)} = 2{,}0$ pour $\boldsymbol{x}_{(2)}$ dans la figure 2.6d. D'après ces courbes, les meilleurs performances des tests sont obtenues avec les critères de la plus grande taille et de la plus grande variance. Les courbes de la figure 2.6d sont différentes et au-dessus des courbes de la figure 2.6c, les critères conduisant aux meilleurs résultats sont l'échantillon de plus petite taille puis la plus petite statistique et de nouveau le maximum de la variance.

Une troisième série de tests est réalisée avec des distributions exponentielles. Dans le cas homoscédastique, le paramètre de la loi est fixé à $\lambda = 1{,}0$ pour les deux échantillons. Les observations de $\boldsymbol{x}_{(2)}$ sont décalées de 0,5 dans la figure 2.7a et de $-0{,}5$ dans la figure 2.7b. Les meilleures courbes ROC sont obtenues avec le critère de l'échantillon le plus grand puis par la statistique maximale pour la figure 2.7a, et par l'échantillon le plus petit puis par la plus petite statistique pour la figure 2.7b. Dans le cas hétéroscédastique, les paramètres des lois sont $\lambda_{(1)} = 2/3$ pour $\boldsymbol{x}_{(1)}$ et $\lambda_{(2)} = 1{,}0$ pour $\boldsymbol{x}_{(2)}$. Les courbes ROC de la figure 2.7c montrent que le test VEME est le plus performant en choisissant l'échantillon le plus petit ou celui dont la statistique est la plus petite. Le choix de l'échantillon de plus grande variance donne globalement les mêmes résultats que le choix de l'échantillon de plus faible variance. Les valeurs des paramètres de la distribution sont ensuite inversés : $\lambda_{(1)} = 1{,}0$ et $\lambda_{(2)} = 2/3$. Les courbes ROC sont données dans la figure 2.7d, on peut en tirer les conclusions inverses sur le classement des critère, le meilleur étant celui de la plus grande taille.

En l'absence de formulation explicite de la statistique de test, puisque le multiplicateur de Lagrange est obtenu numériquement, les courbes ROC théoriques ne peuvent pas être établies, c'est pourquoi des simulations sont nécessaires pour identifier le meilleur critère pour le choix de $\boldsymbol{x}_1$. D'autres expériences, non présentées ici, ont été menées avec différents paramètres des lois utilisées pour générer les données. Les résultats des figures 2.4, 2.5, 2.6 et 2.7 sont des exemples représentatifs de l'impact des paramètres sur les performances du test VEME, selon le critère choisi. Ils illustrent la difficulté à établir un critère général qui conduise à la plus grande puissance possible. D'après nos conclusions, le choix du critère doit se faire en fonction des caractéristiques des données, ce qui n'est pas toujours possible en pratique.

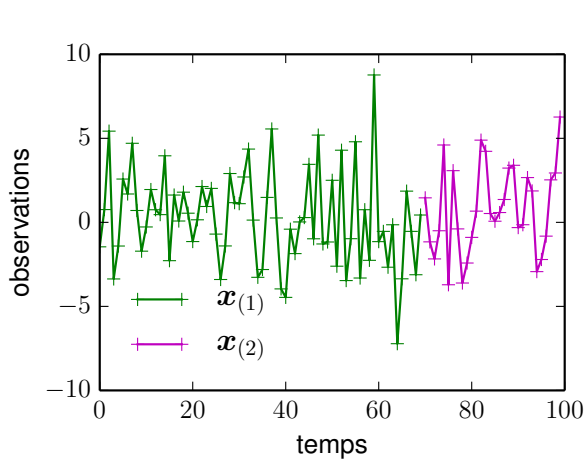
Les diagrammes quantile-quantile sont tracés pour deux critères retenus : critère du maximum de la variance (figures 2.8a, 2.9a et 2.10a), et pour le critère de l'échantillon

le plus grand (figures 2.8b, 2.9b et 2.10b). Ces courbes représentent l'adéquation de la statistique de test avec la loi asymptotique $\chi^2_{(1)}$ sous H_0. Comme les variances sont alors égales, l'application du critère de plus grande variance ne conduit pas au choix d'un des deux échantillons en particulier. Avec le critère de la plus grande taille, c'est toujours l'échantillon 1 qui est retenu. Dans le cas normal (figure 2.8), on vérifie bien qu'avec les deux critères testés, la statistique t est en adéquation avec la loi du $\chi^2_{(1)}$. Pour la distribution inverse-gamma, de paramètres $\alpha_{(1)} = 3{,}0$ et $\beta_{(1)} = 2{,}0$ pour $\boldsymbol{x}_{(1)}$ et pour $\boldsymbol{x}_{(2)}$, on obtient les diagrammes quantile-quantile de la figure 2.9. On constate dans ce cas que la statistique t s'éloigne de la distribution du $\chi^2_{(1)}$: elle est inférieure avec le critère de la variance (figure 2.9a) et supérieure avec le critère de la taille (figure 2.9b). Ce comportement pour le critère de la plus grande variance se retrouve dans le cas de la distribution exponentielle, avec $\lambda = 2/3$, comme on peut le voir à la figure 2.10a, tandis qu'avec le critère sur la taille, la statistique est bien en adéquation avec la loi du $\chi^2_{(1)}$, comme le montre la figure 2.10b.

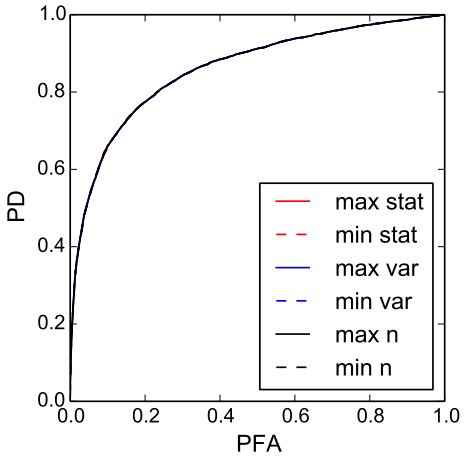
D'après ces comparaisons, nous constatons que l'approximation par la loi du $\chi^2_{(1)}$ n'est pas toujours justifiée pour le critère de la variance la plus grande, notamment pour les distributions inverse-gamma et exponentielles testées. Le choix de la plus grande taille paraît mieux garantir la convergence de t vers sa distribution asymptotique, ou au pire conduire à un test conservatif (figure 2.9b). Ce critère est donc appliqué par la suite dans le test VEME, bien qu'il n'assure pas toujours la meilleure puissance par rapport aux autres critères testés. On note que les figures sont obtenues dans le cas hétéroscédastique, tandis que la proposition 2.3.1 a été démontrée pour des variances égales.

2.4.2 Distributions normale et non normales

Les tests de Student, Welch, VEME, VEseg et WMW sont comparés pour plusieurs distributions, afin de déterminer la robustesse des statistiques aux données non nécessairement gaussiennes et d'évaluer leur généricité. Les paramètres des distributions testées sont rassemblés dans le tableau 2.1, les courbes ROC correspondantes, obtenues avec 10000 signaux, sont tracées dans la figure 2.11. La première comparaison est faite dans le cas normal où les variances sont égales, on constate que les courbes ROC de la figure 2.11a sont identiques. Dans la deuxième simulation, les variances diffèrent, les courbes ROC les meilleures de la figure 2.11b sont alors celles du test de Welch et du test de vraisemblance empirique VE seg de [Jing, 1995]. Les autres tests assurent toutefois des performances proches. Lorsque les distributions ne sont plus normales, aux figures 2.11c pour les lois inverses-gamma, 2.11d pour les lois exponentielles, et 2.11e pour les lois de Cauchy, les différences entre les tests ne sont pas très importantes, en-dehors du test de WMW. Ce dernier permet d'atteindre une bonne puissance en raison de sa robustesse aux valeurs extrêmes. Ce test est le plus efficace dans le cas particulier de la loi de Cauchy, où ni la moyenne ni la variance ne sont définies, et où tous les autres tests proposés échouent.



(a) Exemple d'échantillons $\boldsymbol{x}_{(1)}$ (vert) et $\boldsymbol{x}_{(2)}$ (magenta).

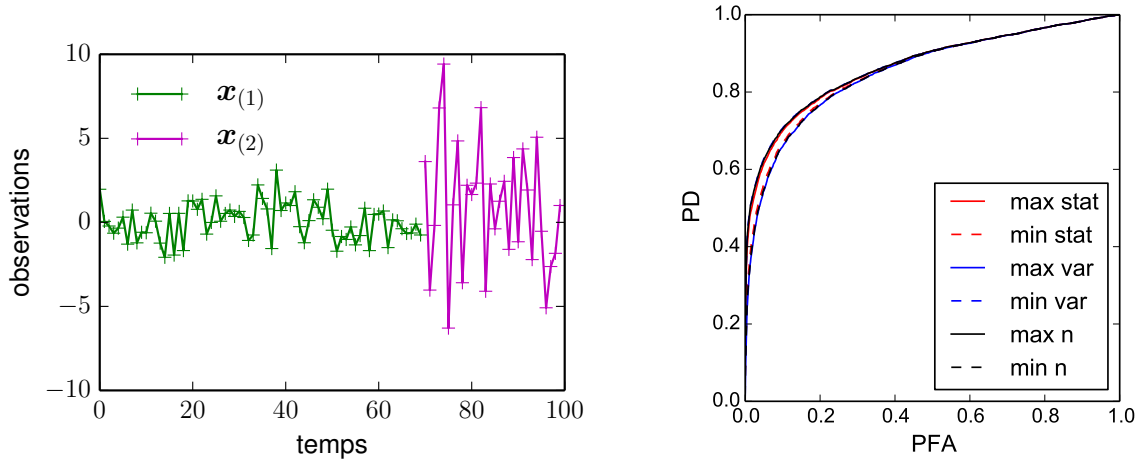


(b) Courbes ROC.

FIGURE 2.3 – Comparaison des critères pour le choix de la moyenne empirique entre les échantillons $\boldsymbol{x}_{(1)}$ (vert) de $n_{(1)} = 70$ points et $\boldsymbol{x}_{(2)}$ (magenta) de $n_{(2)} = 30$ points dans le cas homoscédastique. Les distributions sont respectivement $\mathcal{N}(0{,}0{,}10{,}0)$ et $\mathcal{N}(1{,}0{,}10{,}0)$. La figure 2.3a donne un exemple d'une telle série temporelle. Les courbes ROC en 2.3b sont tracées pour différents critères : la statistique (rouge), la variance (bleu) ou la taille (noir). Pour le choix de la valeur la plus grande, le résultat est en trait plein, et pour la valeur la plus petite le résultat est en traits pointillés. Toutes ces courbes sont confondues.

2.4.3 Petits échantillons

Enfin, on s'intéresse au comportement des statistiques lorsque les échantillons sont de petites tailles. On se place dans le cas gaussien homoscédastique, avec les paramètres $\mu_1 = 0{,}0$, $\sigma_1^2 = 1{,}0$ et $\mu_2 = 1{,}0$, $\sigma_2^2 = 1{,}0$. Dans la première simulation, les échantillons sont déséquilibrés, avec $n_1 = 95$ et $n_2 = 5$. Dans la seconde simulation ils sont de même taille $n_1 = n_2 = 5$. Les courbes ROC sont tracées dans la figure 2.12 et les diagrammes quantile-quantiles, tracés en fonction des lois asymptotiques de chaque statistique, sont donnés dans les figures 2.13 et 2.14. Le test de [Jing, 1995] n'est pas applicable dans ces cas-là, car les échantillons étant peu nombreux, le risque que la moyenne soit en-dehors de l'enveloppe des observations est grand. L'ensemble de ces résultats montre que les tests sont aussi puissants et calibrés les uns que les autres pour les petits échantillons testés (figure 2.12). Seul le test de Welch, dans le cas où les tailles sont déséquilibrées, s'avère un peu moins performant (figure 2.12a). Avec les digrammes quantiles-quantiles 2.13 et 2.14, on vérifie bien que les statistiques des tests de Student, VEME et WMW demeurent en adéquation avec leurs distributions asymptotiques, malgré la faible valeur de n.



(a) Exemple d'échantillons $\boldsymbol{x}_{(1)}$ (vert) et $\boldsymbol{x}_{(2)}$ (magenta).

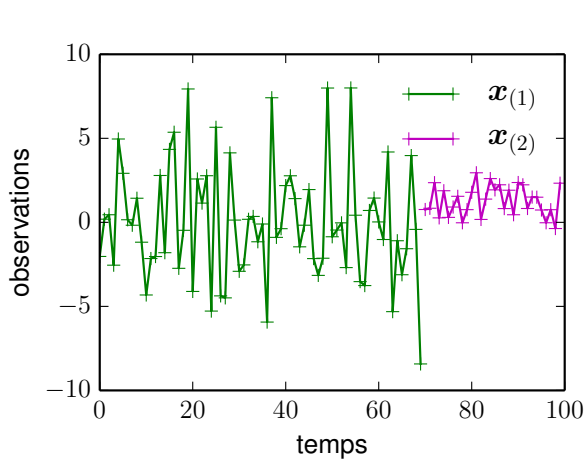
(b) Courbes ROC.

FIGURE 2.4 – Comparaison des critères pour le choix de la moyenne empirique entre les échantillons $\boldsymbol{x}_{(1)}$ (vert) de $n_{(1)} = 70$ points et $\boldsymbol{x}_{(2)}$ (magenta) de $n_{(2)} = 30$ points. Les distributions sont respectivement $\mathcal{N}(0,0,1,0)$ et $\mathcal{N}(1,0,10,0)$. La figure 2.4a donne un exemple d'une telle série temporelle. Les courbes ROC en 2.4b sont tracées pour différents critères : la statistique (rouge), la variance (bleu) ou la taille (noir). Pour le choix de la valeur la plus grande, le résultat est en trait plein, et pour la valeur la plus petite le résultat est en traits pointillés.

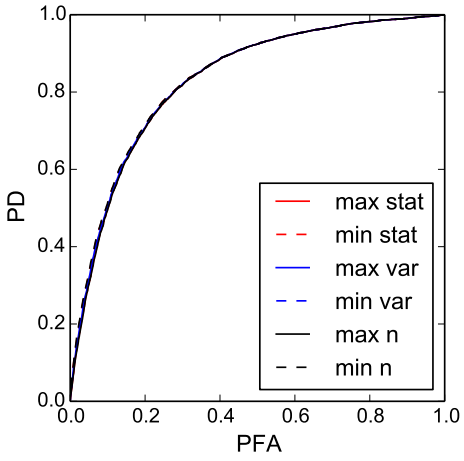
2.5 Vers la détection de rupture

Dans ce chapitre, nous avons présenté les approches classiques pour la détection d'un décalage entre les distributions F_1 et F_2 de deux populations. Les hypothèses nulle et alternatives supposent que suffisamment d'informations sur la différence entre F_1 et F_2 soient contenues dans la moyenne (test de Student, test VEME) ou dans la médiane (test de WMW). Les distributions asymptotiques des statistiques sont obtenues sous H_0, permettant ainsi d'élaborer une procédure de test pour un seuil de signification α donné. Pour notre problème de détection de ruptures dans des séries temporelles, ces tests d'homogénéité constituent un point de départ. Dans ce contexte, les échantillons $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ deviennent deux segments consécutifs du signal, et la transition entre les observations x_{n_1} et x_{n_1+1} est une rupture potentielle. Les stratégies évoquées dans la partie 1.1.2 du chapitre précédent sont employées pour passer du test d'homogénéité à la détection de ruptures multiples à des positions inconnues. Ces approches rencontrent les difficultés recensées au chapitre 1 : localisation des ruptures, détermination de leur nombre, contrôle de la détection de plusieurs événements et complexité.

Pour passer d'un test entre deux échantillons, applicable à deux segments consécutifs bien délimités, à une méthode de détection de rupture, de position inconnue, une solution classique consiste à trouver la valeur de n_1 qui maximise (ou minimise) la statistique, comme le propose l'article [Zou *et al.*, 2007]. Des exemples de statistiques obtenues en



(a) Exemple d'échantillons $\boldsymbol{x}_{(1)}$ (vert) et $\boldsymbol{x}_{(2)}$ (magenta).

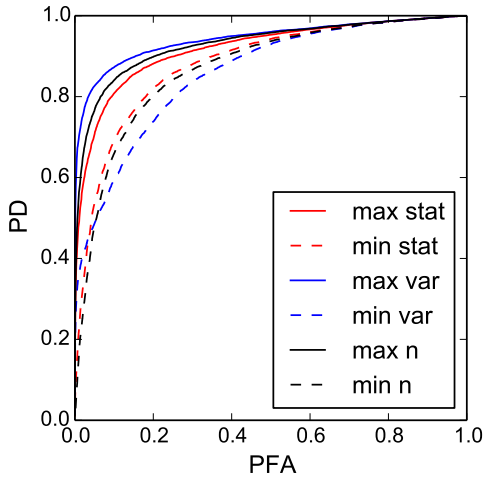


(b) Courbes ROC.

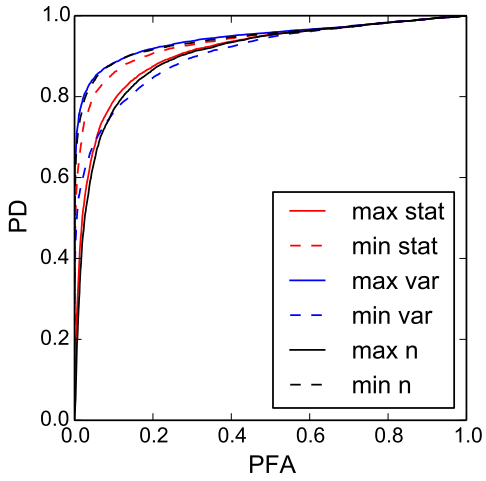
FIGURE 2.5 – Test VEME : comparaison des critères pour le choix de la moyenne empirique entre les échantillons $\boldsymbol{x}_{(1)}$ (vert) de $n_{(1)} = 70$ points et $\boldsymbol{x}_{(2)}$ (magenta) de $n_{(2)} = 30$ points. Les distributions sont respectivement $\mathcal{N}(0{,}0{,}10{,}0)$ et $\mathcal{N}(1{,}0{,}1{,}0)$. La figure 2.5a donne un exemple de telles séries temporelles. Les courbes ROC en 2.5b sont tracées pour différents critères : la statistique (rouge), la variance (bleu) ou la taille (noir). Pour le choix de la valeur la plus grande, le résultat est en trait plein, et pour la valeur la plus petite le résultat est en traits pointillés.

chaque position d'un signal sont donnés à la figure 2.15. Pour traiter des séries temporelles comportant potentiellement plusieurs ruptures, ce type de critère peut être employé localement sur une portion du signal où une seule rupture est recherchée, par exemple dans une fenêtre ou lors d'une approche par bissection. Les tests d'homogénéité peuvent également être adaptés. L'extension du test de WMW à plusieurs échantillons, le test de Kruskall-Wallis, s'applique par exemple lorsque les positions des ruptures à détecter sont connues. Comme ce cas de figure est peu courant, les méthodes présentées dans [Lung-Yut-Fong *et al.*, 2011b, Lung-Yut-Fong *et al.*, 2011a] ont été développées à partir de ce test. La détection de plusieurs événements peut aussi être formulée comme un problème d'hypothèses multiples.

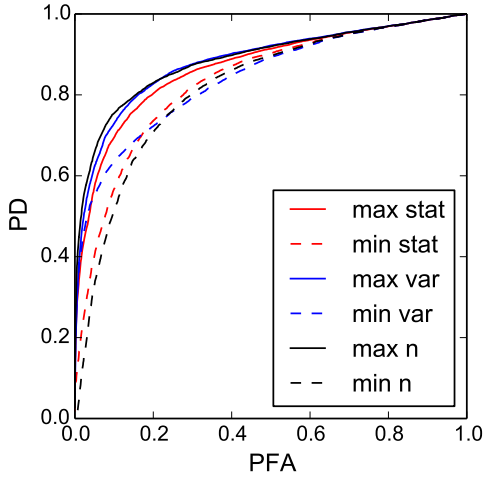
La stratégie que nous avons adoptée pour la méthode du *Bernoulli Detector*, présenté au chapitre 3, consiste à intégrer la statistique d'un test d'homogénéité comme paramètre du modèle des observations, dans un cadre bayésien. Cette statistique est calculée pour chaque observation testée x_i, entre les échantillons $\boldsymbol{x}_1 = (x_{i^-+1}, \ldots, x_i)$ et $\boldsymbol{x}_2 = (x_{i+1}, \ldots, x_{i^+})$, où i^- et i^+ sont les positions des ruptures situées avant et après x_i. D'après les conclusions tirées de la comparaison des tests dans la partie 2.4, le choix du test ne porte pas sur le test VEME, malgré sa capacité à s'adapter aux données. En effet, ce test que nous avons proposé comme alternative aux tests paramétriques ne s'avère pas toujours le plus performant, et suppose que la moyenne et la variance soient définies. Sa mise en œuvre est plus complexe que les tests paramétriques et que le test de rang. De plus, le critère pour



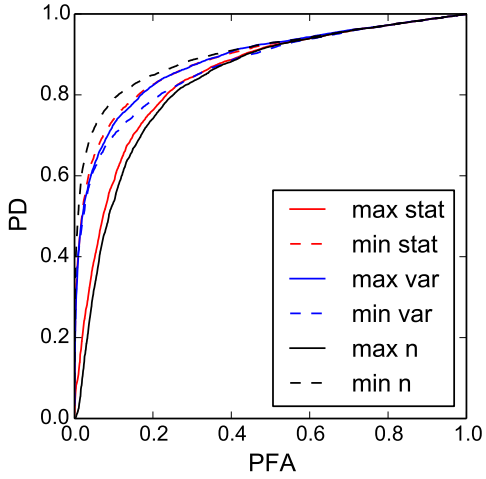
(a) Cas homoscédastique. Les paramètres des distributions sont $\mathcal{IG}(\alpha = 3{,}0, \beta = 2{,}0)$ et les observations de $\boldsymbol{x}_{(2)}$ sont décalées de 0,5.



(b) Cas homoscédastique. Les paramètres des distributions sont $\mathcal{IG}(\alpha = 3{,}0, \beta = 2{,}0)$ et les observations de $\boldsymbol{x}_{(2)}$ sont décalées de $-0{,}5$.



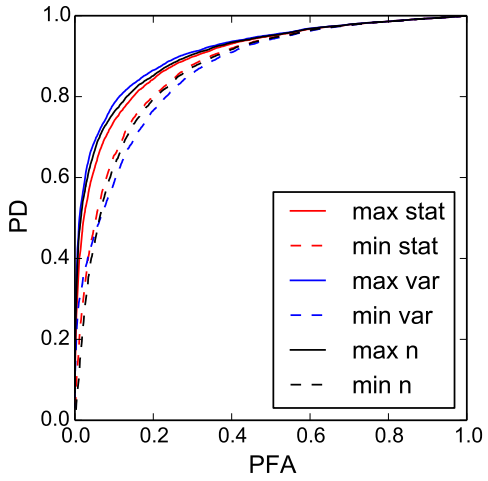
(c) Cas hétéroscédastique. Les paramètres des distributions sont $\mathcal{IG}(\alpha_{(1)} = 3{,}0, \beta_{(1)} = 2{,}0)$ pour $\boldsymbol{x}_{(1)}$ et $\mathcal{IG}(\alpha_{(2)} = 3{,}0, \beta_{(2)} = 3{,}0)$ pour $\boldsymbol{x}_{(2)}$.



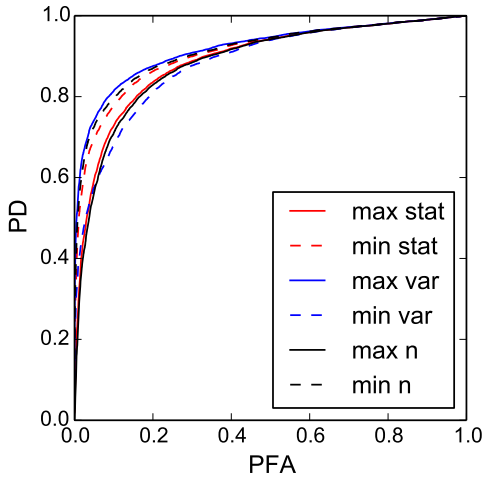
(d) Cas hétéroscédastique. Les paramètres des distributions sont $\mathcal{IG}(\alpha_{(1)} = 3{,}0, \beta_{(1)} = 3{,}0)$ pour $\boldsymbol{x}_{(1)}$ et $\mathcal{IG}(\alpha_{(2)} = 3{,}0, \beta_{(2)} = 2{,}0)$ pour $\boldsymbol{x}_{(2)}$.

FIGURE 2.6 – Test VEME : comparaison des courbes ROC des critères pour le choix de la moyenne empirique, avec des distributions inverse gamma $\mathcal{IG}(\alpha, \beta)$. Les tailles des échantillons sont $n_{(1)} = 70$ et $n_{(2)} = 30$. Les courbes ROC sont tracées pour différents critères : la statistique (rouge), la variance (bleu) ou la taille (noir). Pour le choix de la valeur la plus grande, le résultat est en trait plein, et pour la valeur la plus petite le résultat est en traits pointillés.

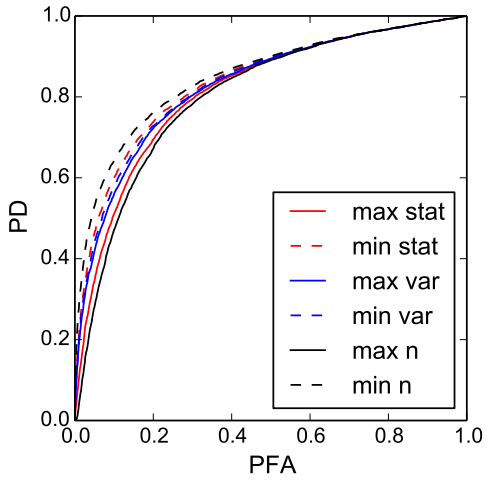
déterminer à partir de quel échantillon la moyenne empirique doit être calculée dépend des données.



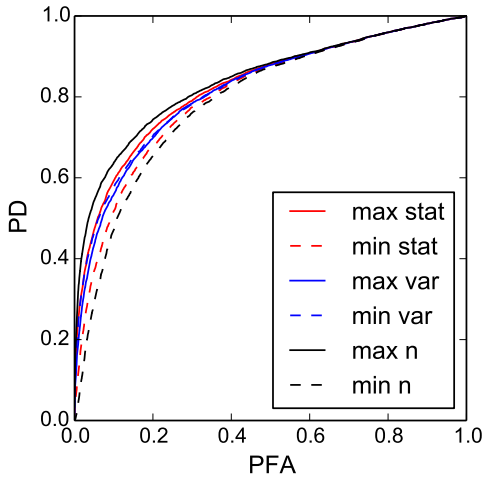
(a) Cas homoscédastique. Le paramètre de la distribution est $\lambda = 1$ pour $\boldsymbol{x}_{(1)}$ et $\boldsymbol{x}_{(2)}$ et les observations de $\boldsymbol{x}_{(2)}$ sont décalées de 0,5.



(b) Cas homoscédastique. Le paramètre de la distribution est $\lambda = 1$ pour $\boldsymbol{x}_{(1)}$ et $\boldsymbol{x}_{(2)}$ et les observations de $\boldsymbol{x}_{(2)}$ sont décalées de $-0{,}5$.



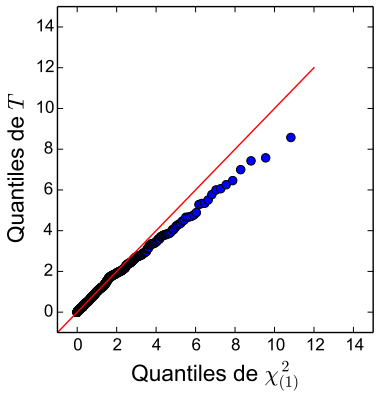
(c) Cas hétéroscédastique. Les paramètres des distributions sont $\lambda_{(1)} = 2/3$ pour $\boldsymbol{x}_{(1)}$ et $\lambda_{(2)} = 1{,}0$ pour $\boldsymbol{x}_{(2)}$.



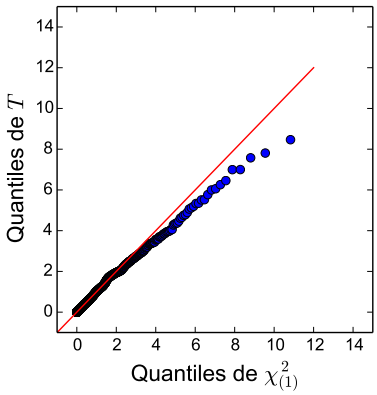
(d) Cas hétéroscédastique. Les paramètres des distributions sont $\lambda_{(1)} = 1{,}0$ pour $\boldsymbol{x}_{(1)}$ et $\lambda_{(2)} = 2/3$ pour $\boldsymbol{x}_{(2)}$.

FIGURE 2.7 – Test VEME : comparaison des courbes ROC des critères pour le choix de la moyenne empirique, avec des distributions exponentielles de paramètre λ. Les tailles des échantillons sont $n_{(1)} = 70$ et $n_{(2)} = 30$. Les courbes ROC sont tracées pour différents critères : la statistique (rouge), la variance (bleu) ou la taille (noir). Pour le choix de la valeur la plus grande, le résultat est en trait plein, et pour la valeur la plus petite le résultat est en traits pointillés.

Le test de WMW est en revanche un candidat intéressant, ce que confirment les nombreux exemples de la littérature ([Lehmann et D'Abrera, 1975, Pettitt, 1979, Gombay et

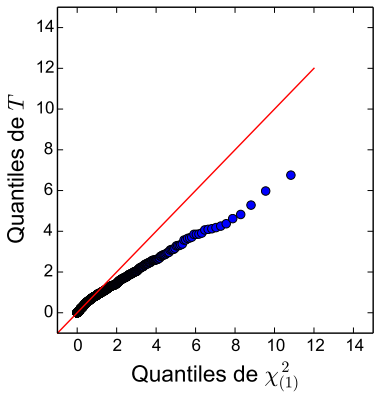


(a) Application du critère de la plus grande variance. Sous H_0, les deux variances sont égales, la moyenne choisie pour $\overline{x}_1$ est $\overline{x}_{(1)}$ ou $\overline{x}_{(2)}$.

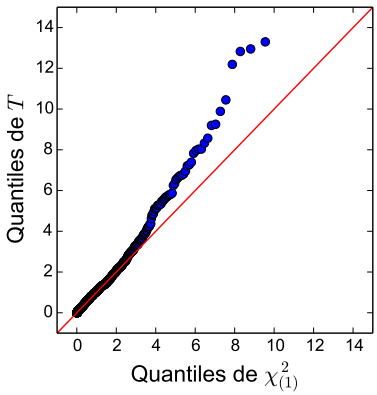


(b) Application du critère de la plus grande taille, qui conduit à choisir $\overline{x}_1 = \overline{x}_{(1)}$.

FIGURE 2.8 – Calibration du test VEME : diagramme quantile-quantile de la statique t avec la loi du $\chi^2_{(1)}$ sous H_0, avec la distribution normale $\mathcal{N}(0{,}0{,}1{,}0)$. Les tailles des échantillons sont $n_{(1)} = 70$ et $n_{(2)} = 30$. Le critère retenu pour le choix de la moyenne empirique est celui de la plus grande variance pour la figure 2.8a, et celui de la plus grande taille pour la figure 2.8b.

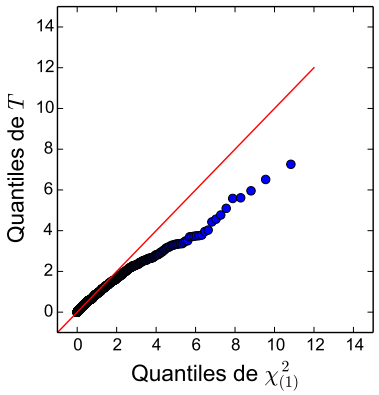


(a) Application du critère de la plus grande variance. Sous H_0, les deux variances sont égales, la moyenne choisie pour $\overline{x}_1$ est $\overline{x}_{(1)}$ ou $\overline{x}_{(2)}$.

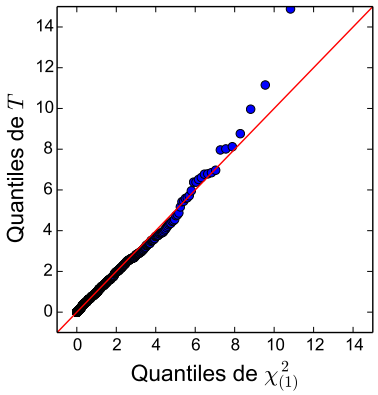


(b) Application du critère de la plus grande taille, qui conduit à choisir $\overline{x}_1 = \overline{x}_{(1)}$.

FIGURE 2.9 – Calibration du test VEME : diagramme quantile-quantile de la statique t avec la loi du $\chi^2_{(1)}$ sous H_0, avec la distribution inverse-gamma $\mathcal{IG}(\alpha_{(1)} = 3{,}0{,}\beta_{(1)} = 2{,}0)$. Les tailles des échantillons sont $n_{(1)} = 70$ et $n_{(2)} = 30$. Le critère retenu pour le choix de la moyenne empirique est celui de la plus grande variance pour la figure 2.9a, et celui de la plus grande taille pour la figure 2.9b.



(a) Application du critère de la plus grande variance. Sous H_0, les deux variances sont égales, la moyenne choisie pour $\overline{x}_1$ est $\overline{x}_{(1)}$ ou $\overline{x}_{(2)}$.



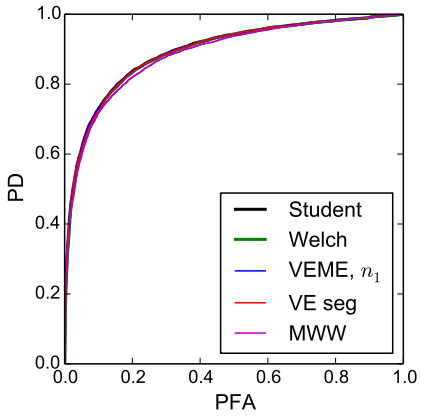
(b) Application du critère de la plus grande taille, qui conduit à choisir $\overline{x}_1 = \overline{x}_{(1)}$.

FIGURE 2.10 – Calibration du test VEME : diagramme quantile-quantile de la statique t avec la loi du $\chi^2_{(1)}$ sous H_0, avec la distribution exponentielle de paramètre $\lambda = 2/3$. Les tailles des échantillons sont $n_{(1)} = 70$ et $n_{(2)} = 30$. Le critère retenu pour le choix de la moyenne empirique est celui de la plus grande variance pour la figure 2.10a, et celui de la plus grande taille pour la figure 2.10b.

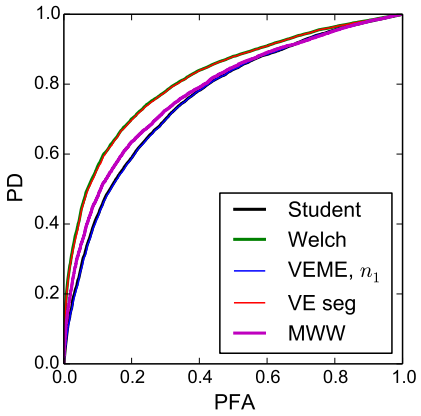
Liu, 2000, Lung-Yut-Fong *et al.*, 2011c]), pour plusieurs raisons. Tout d'abord ce test est non paramétrique, contrairement aux tests t, ce qui nous permet de nous affranchir de l'hypothèse de normalité des données, et de développer un modèle générique qui ne nécessite pas de modifications pour traiter des séries temporelles aux caractéristiques variables. Les hypothèses faites sur les observations sont celles citées dans la partie 2.2. De plus, la statistique de WMW repose sur la comparaison des rangs ; une conséquence appréciable est que les propriétés de U sont définies aussi pour des lois à queues lourdes et en présence de valeurs aberrantes, et demeure applicable à des distributions extrêmes, comme la loi de Cauchy, où les autres tests ne sont plus valables. Dans le cas gaussien, le test de WMW est presque aussi puissant que le test uniformément optimal de Student. Il reste bien calibré même pour de petits échantillons, ou pour des échantillons déséquilibrés. Ce point est important, car notre modèle ne contraint pas les dimensions des segments estimés, i^- et i^+ peuvent être proches. Enfin, concernant la complexité calculatoire, le test de WMW est plus rapide à mettre en œuvre que les méthodes de la vraisemblance empirique, qui nécessitent la résolution d'un problème d'optimisation. L'approximation normale est applicable à partir de quelques observations, et le classement des observations peut se faire une seule fois sur l'ensemble des mesures. Une stratégie de programmation dynamique est par exemple employée dans [Lung-Yut-Fong *et al.*, 2011b]. Grâce au choix de la statistique U, le modèle du *Bernoulli Detector* que l'on introduit dans la suite tire parti des propriétés inhérentes au test de WMW.

distributions	paramètres échantillon 1	paramètres échantillon 2	remarque
normale 1	$\mu_1 = 0{,}0$, $\sigma_1^2 = 1{,}0$	$\mu_2 = 0{,}5$, $\sigma_2^2 = 1{,}0$	même variance
normale 2	$\mu_1 = 0{,}0$, $\sigma_1^2 = 3{,}0$	$\mu_2 = 0{,}5$, $\sigma_2^2 = 1{,}0$	variances différentes
inverse-gamma	$\alpha_1 = 3{,}0$, $\beta_1 = 2{,}0$	$\alpha_2 = 3{,}0$, $\beta_2 = 3{,}0$	
exponentielle	$\lambda = \frac{2}{3}$	$\lambda = 1{,}0$	
Cauchy	$\alpha_1 = 1{,}0$, $x_{0,1} = 0{,}0$	$\alpha_1 = 1{,}0$, $x_{0,2} = 1{,}0$	espérances et variances indéfinies

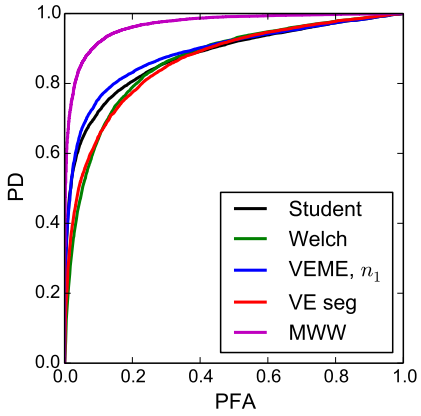
TABLE 2.1 – Distributions choisies pour comparer les tests d'homogénéité à la figure 2.11.



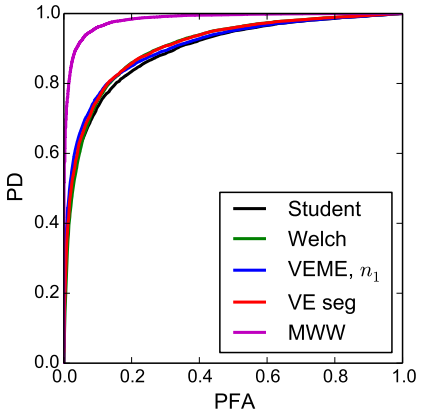
(a) Lois normales de même variance.



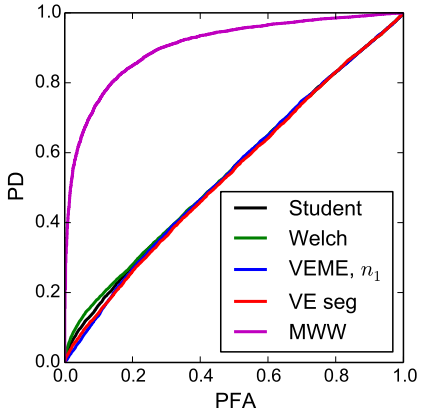
(b) Lois normales de variances différentes.



(c) Lois inverses-gamma.

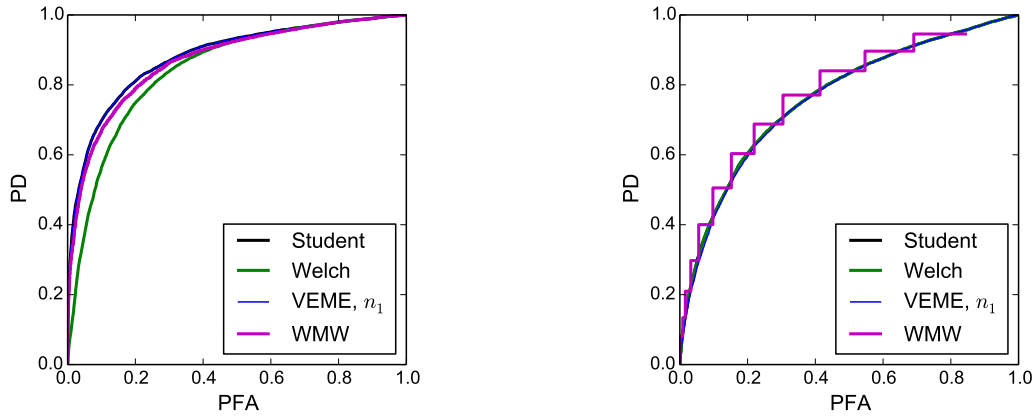


(d) Lois exponentielles



(e) Lois de Cauchy

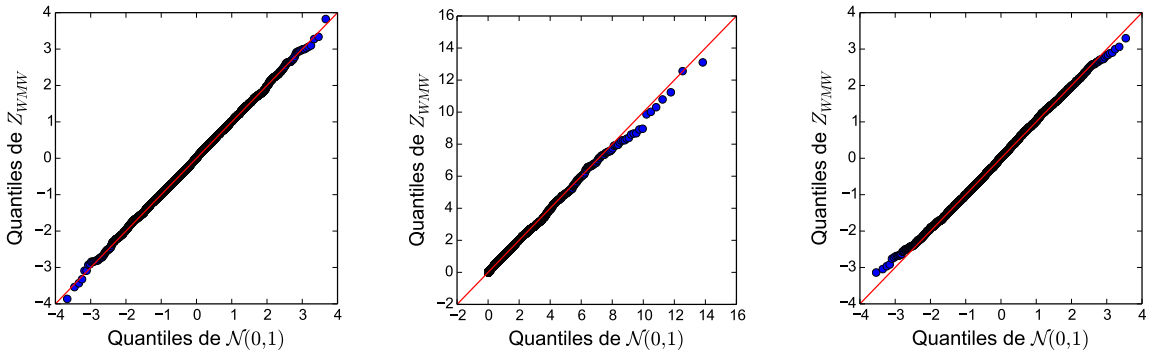
FIGURE 2.11 – Comparaison des puissances de tests t de Student (Student), de Welch (Welch), le test VEME avec le critère de l'échantillon de plus grande taille (VEME, n_1), le test de [Jing, 1995] (VE seg) et le test de rang WMW (MWW), pour différentes distributions. Les tailles des échantillons sont $n_1 = 70$ pour $\boldsymbol{x}_1$ et $n_2 = 30$ pour $\boldsymbol{x}_2$.



(a) $n_1 = 95$ et $n_2 = 5$.

(b) $n_1 = n_2 = 5$

FIGURE 2.12 – Comparaison des puissances de tests t de Student (Student), de Welch (Welch), le test VEME avec le critère de l'échantillon de plus grande taille (VEME, n_1), et le test de rang WMW (WMW), pour des petits échantillons. Les distributions sont des lois normales de même variance : $\mathcal{N}(\mu_1 = 0{,}0, \sigma_1^2 = 1{,}0)$ et $\mathcal{N}(\mu_2 = 1{,}0, \sigma_2^2 = 1{,}0)$ pour $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ respectivement.

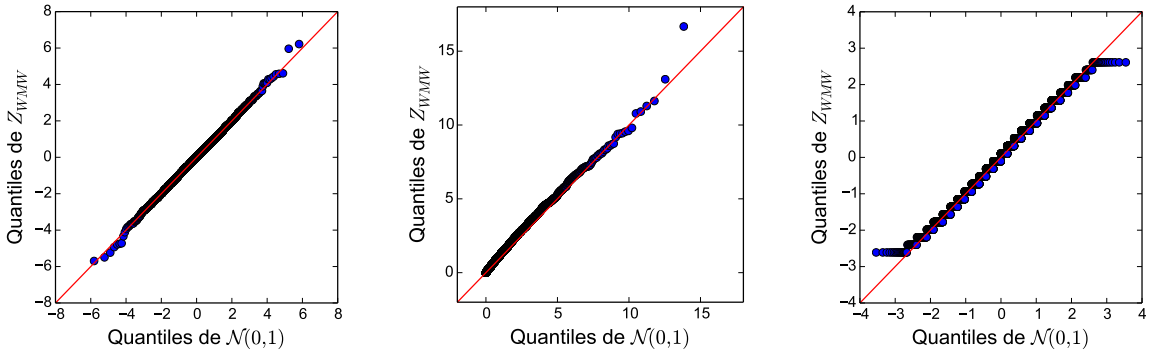


(a) Test de Student.

(b) Test VEME.

(c) Test de WMW.

FIGURE 2.13 – Calibration des tests : diagramme quantile-quantile, $n_1 = 95$ et $n_2 = 5$, sous H_0. La distribution est la loi normale standard.



(a) Test de Student. (b) Test VEME. (c) Test de WMW.

FIGURE 2.14 – Calibration des tests : diagramme quantile-quantile, $n_1 = n_2 = 5$, sous H_0. La distribution est la loi normale standard.

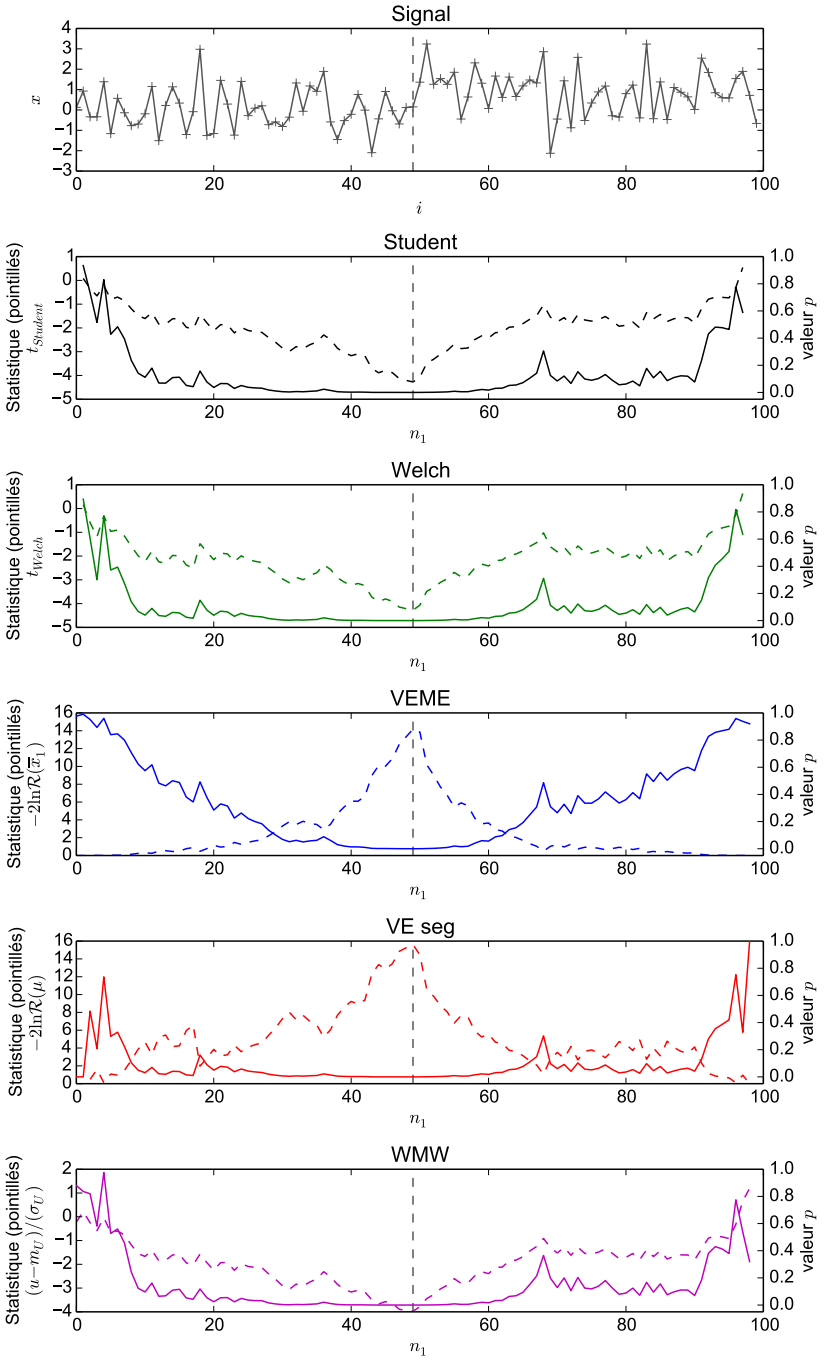


FIGURE 2.15 – Signal et statistiques (pointillés) et p-valeur (pleins) calculées pour chaque position : test de Student (noir), le test de Welch (vert), test VEME (bleu), test VE seg (rouge), et test WMW (magenta). Les observations suivent la loi normale standard, et sont décalées de la valeur 1,0 après la rupture (trait vertical pointillés). La vraie rupture est positionnée à $i = 50$.

Chapitre 3

Modèle du *Bernoulli Detector*

Dans ce chapitre, nous présentons la méthode du *Bernoulli Detector*, destinée à la détection de multiples ruptures. Dans un premier temps, elle est introduite dans un cadre univarié. Comme il est expliqué dans la partie 3.1, le modèle intègre des statistiques locales issues d'un test d'homogénéité, sous la forme de p-valeurs, dans un contexte bayésien. Le test de rang de WMW a été retenu d'après les conclusions du chapitre 2. Cette approche diffère des méthodes paramétriques présentées dans le chapitre 1 en ce qu'elle repose uniquement sur des hypothèses sur la nature des ruptures, et non sur les distributions des données. La fonction de vraisemblance est construite à partir du modèle bêta-uniforme des p-valeurs du test statistique, dépendant d'un paramètre α. Des résultats théoriques sur le contrôle de la détection où intervient le paramètre α sont exposés dans la partie 3.2. La partie 3.3 est consacrée à la résolution pratique du problème et au développement d'un algorithme. Le *Bernoulli Detector* a été appliqué à divers types de données et les résultats expérimentaux sont présentés et commentés dans la partie 3.4. Enfin, une discussion sur cette méthode, son intérêt et ses limites, en comparaison avec d'autres méthodes classiques de l'état de l'art, le fused LASSO et l'algorithme BARD, est menée dans la partie 3.5. Dans un deuxième temps, au chapitre 4, le modèle est étendu au traitement de séries temporelles multivariées. Cette méthode est présentée dans les articles [Harlé *et al.*, 2014, Harlé *et al.*, 2016].

3.1 Description du modèle

3.1.1 Objectif et notations

Notre objectif est la détection de multiples ruptures dans les valeurs autour desquelles les observations d'un signal sont distribuées, en nombre et positions inconnus. La méthode doit de plus être assez générale pour s'appliquer sans modification à une grande variété de distributions des données. On note $\boldsymbol{X}$ la série temporelle de n variables, où $x_i \in \mathbb{R}$ est l'observation à l'instant i. La variable X_i est distribuée selon une loi de probabilité dont la fonction de répartition est notée F_i. Les variables sont supposées indépendantes les unes

des autres, et identiquement distribuées entre deux ruptures. Le problème s'écrit sous la forme du test d'hypothèses

$$\begin{array}{ll} H_0 : & F_1 = F_2 = \cdots = F_n, \\ H_1 : & F_1 = F_2 = \cdots = F_{\tau_1} \neq F_{\tau_1+1} = \cdots = F_{\tau_K} \neq F_{\tau_K+1} = \cdots = F_n. \end{array} \tag{3.1}$$

où les instants des ruptures $1 < \tau_1, \ldots, \tau_K < n$ et leur nombre K sont inconnus.

Afin de formaliser le problème de la détection de ces ruptures, le vecteur $\boldsymbol{R} \in \{0,1\}^n$ des variables indicatrices de la présence ou non de ruptures est défini par

$$R_i = \begin{cases} 1 & \text{si } X_i \text{ est une rupture,} \\ 0 & \text{sinon,} \end{cases} \tag{3.2}$$

pour tout $1 < i < n$. Par convention les extrémités du signal $\boldsymbol{X}$ sont des ruptures, $R_1 = R_n = 1$, et ne sont pas comptabilisées dans le nombre $K = \sum_{i=2}^{N-1} R_i$. L'estimation du paramètre inconnu $\boldsymbol{R}$ revient à détecter et localiser les ruptures de $\boldsymbol{X}$. L'inférence de $\boldsymbol{R}$ est réalisée dans un cadre bayésien, où le théorème de Bayes permet d'écrire la relation de proportionnalité

$$f(\boldsymbol{R}|\boldsymbol{X}) \propto L(\boldsymbol{X}|\boldsymbol{R}) f(\boldsymbol{R}), \tag{3.3}$$

la densité de probabilité a posteriori $f(\boldsymbol{R}|\boldsymbol{X})$ étant, à une constante multiplicative près, la densité a priori $f(\boldsymbol{R})$ multipliée par la fonction de vraisemblance $L(\boldsymbol{X}|\boldsymbol{R})$.

La méthode du *Bernoulli Detector* propose une description originale des données : une statistique de test est introduite dans la fonction de vraisemblance $L(\boldsymbol{X}|\boldsymbol{R})$, le test étant choisi de telle sorte que la dépendance du modèle à des hypothèses fortes sur les distributions F_i des données soit limitée. La densité a posteriori introduit plus classiquement une variable de Bernoulli, comme dans les approches de [Lavielle et Lebarbier, 2001, Dobigeon *et al.*, 2007a, Fan et Mackey, 2015]. L'expression d'une fonction de vraisemblance marginale composite est présentée dans la partie 3.1.2 comme approximation de $L(\boldsymbol{X}|\boldsymbol{R})$, puis les choix effectués a priori pour $f(\boldsymbol{R})$ sont donnés dans la partie 3.1.3. La densité de probabilité a posteriori en est finalement déduite dans la partie 3.1.4.

3.1.2 Fonction de vraisemblance

Les variables X_i sont supposées indépendantes. Le problème (3.1) revient à tester pour chaque variable X_i, $1 < i < n$ les hypothèses locales suivantes :

$$\begin{array}{lllll} H_0(i) : & X_i & \text{n'est pas une rupture,} & R_i = 0 \\ H_1(i) : & X_i & \text{est une rupture,} & R_i = 1. \end{array} \tag{3.4}$$

La fonction de vraisemblance se décompose donc en un produit de vraisemblances marginales

$$L(\boldsymbol{X}|\boldsymbol{R}) = \prod_{i=2}^{n-1} f(X_i|\boldsymbol{R}), \tag{3.5}$$

où le paramètre $\boldsymbol{R}$ est équivalent à la segmentation du signal $\boldsymbol{X}$.

Pour que le *Bernoulli Detector* parvienne à détecter des ruptures dans des signaux de distribution non nécessairement gaussienne, la loi $f(X_i|\boldsymbol{R})$ de la fonction de vraisemblance (3.5) n'est pas choisie selon un modèle précis des observations, mais est construite à partir des distributions de statistiques de test $T(i)$, obtenues pour chaque $1 < i < n$. Ce test statistique, choisi judicieusement, fournit une p-valeur à la position i à partir de $\boldsymbol{X}$ et de la segmentation $\boldsymbol{R}$, notée P_i, qui est considérée comme une variable aléatoire. Le modèle bêta-uniforme est associé à ces p-valeurs, et introduit un seuil d'acceptation α. La fonction de vraisemblance (3.5) est alors exprimée en fonction de $\boldsymbol{X}$, $\boldsymbol{R}$ et α, nous verrons dans la partie 3.2 comment ce terme intervient dans le contrôle de la détection.

Valeur p

Soit un test d'hypothèse nulle H_0, de statistique T, de densité de probabilité continue sous H_0. Pour un échantillon $\boldsymbol{X}$, la p-valeur du test statistique est définie de la façon suivante :

Définition 3.1.1 (Valeur p)**.** *La p-valeur est la probabilité d'obtenir une valeur T au moins aussi extrême que le résultat du test $T = t$ sur les observations $\boldsymbol{x}$ si H_0 est vraie.*

Pour un test unilatéral à droite, on a donc $p(t) = \Pr(T \geq t|H_0)$, et dans le cas d'un test bilatéral, si la distribution du test est symétrique autour de 0, $p(t) = \Pr\left(|T| \geq |t|\middle|H_0\right)$.

La figure 3.1 donne une représentation graphique de cette notion, pour un test bilatéral : la valeur de $p(t)$ est déterminée par la somme des surfaces grisées sous la courbe de la densité de probabilité de la statistique T sous H_0, qui correspondent aux valeurs extrêmes de T, les moins probables sous l'hypothèse nulle. L'interprétation de la p-valeur dans des procédures d'analyse de données expérimentales porte parfois à confusion, notamment parce que cette notion n'est ni une mesure de la probabilité a posteriori que H_0 ou que H_1 soient vraies, ni une mesure de la probabilité de rejeter H_0 à tort, c'est pourquoi il est généralement recommandé de se ramener à la définition (3.1.1) [Sellke *et al.*, 2001].

Une caractéristique intéressante de la p-valeur est que sa distribution sous H_0 est la loi uniforme sur l'intervalle [0,1] si la statistique de test suit une loi continue, quelle que soit la distribution des données ou la taille n de l'échantillon (voir [Lehmann et Romano, 2006, lemme 3.3.1, p. 64]). En effet, par définition, la p-valeur associée à un test est le niveau de signification le plus petit pour lequel H_0 est rejetée. Ainsi, pour tout niveau $u \in [0,1]$, $p(t) \leq u \Leftrightarrow t \in \mathcal{R}_u$, où $\mathcal{R}_u$ est la région critique de niveau u. En notant P la variable aléatoire $p(T)$, quand H_0 est vraie on a $\Pr(P \leq u|H_0) = \Pr(T \in \mathcal{R}_u|H_0)$, donc $\Pr(P \leq u|H_0) = u$. On en déduit que sous H_0, P suit la loi uniforme sur [0,1]. Connaissant la distribution des p-valeurs sous H_0, nous considérons ce terme comme une variable aléatoire P, dont la fonction de répartition F_0 est celle de la loi uniforme. L'interprétation stochastique des p-valeurs est par exemple abordée dans [Sackrowitz et Samuel-Cahn, 1999].

Toutefois sous l'hypothèse alternative H_1, la distribution F_1 de P dépend des distributions des données et n'est en général pas connue. La figure 3.2 illustre les distributions

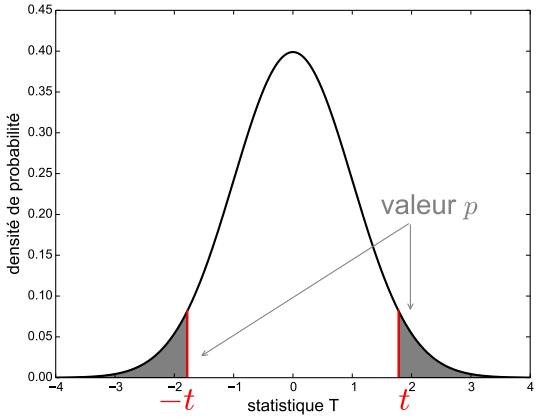
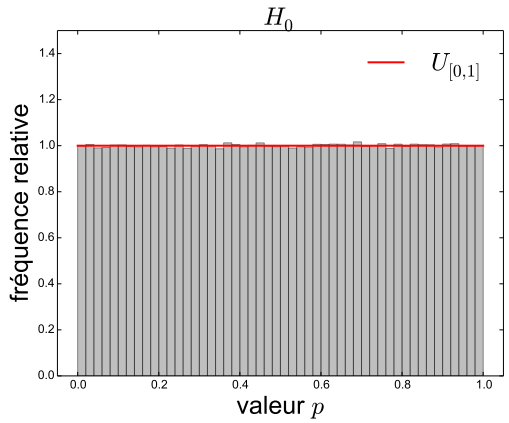


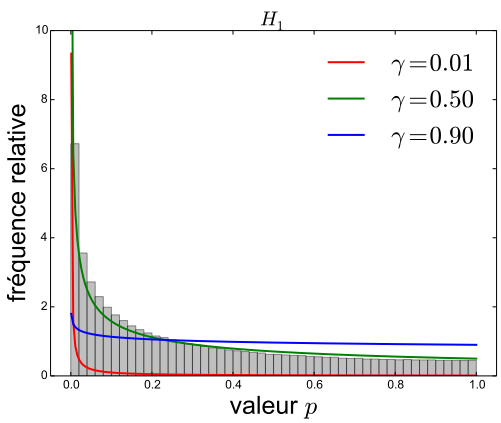
FIGURE 3.1 – Exemple de calcul de la p-valeur à partir de la densité de probabilité f_0 (en noir) de la statistique de test T sous l'hypothèse nulle, pour un test bilatéral. Pour le test de WMW par exemple, la distribution asymptotique est la loi normale standard. L'application du test sur des observations $\boldsymbol{x}$ a produit la valeur t (en rouge), on en déduit la p-valeur à partir de la somme des aires (grisées) sous la courbe T avant la valeur $-t$ et après la valeur t, qui correspondent aux observations les moins probables sous H_0.

des p-valeurs obtenues par l'application du test t de Student entre deux échantillons $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ de $n_1 = n_2 = 100$ points, distribués selon une loi normale de variance $\sigma^2 = 1{,}0$. L'histogramme 3.2a est obtenu sous H_0, lorsque les moyennes des deux échantillons sont $\mu_1 = \mu_2 = 0{,}0$, et l'histogramme 3.2b est obtenu sous H_1 lorsque $\mu_1 = 0{,}0$ et $\mu_2 = 0{,}4$. Dans [Hung *et al.*, 1997, Sellke *et al.*, 2001], les auteurs utilisent la relation $p(t) = 1 - \Phi(t)$ (ici pour un test unilatéral), entre la p-valeur et la réalisation t de la statistique de test T, dont Φ est la fonction de répartition. La fonction de répartition de P sous H_1 est exprimée en fonction des paramètres de la loi asymptotique de T sous H_0, lorsque celle-ci est approximativement normale, de la forme $T = (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)/\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. La distribution des p-valeurs sous l'hypothèse alternative F_1 ainsi obtenue dépend alors de la taille des échantillons et de la puissance du test. Dans le cas général cependant, F_1 est inconnue, mais pour un test consistant, les valeurs prises par P ont tendance à être plus faibles sous H_1 que sous H_0 [Sackrowitz et Samuel-Cahn, 1999]. En suivant la proposition de [Sellke *et al.*, 2001], le modèle de la loi bêta $\mathcal{B}e(\gamma,1)$, de paramètre $0 \leq \gamma \leq 1$, est choisi comme distribution alternative (voir les courbes rouge, verte et bleue à la figure 3.2b). Cette loi est décroissante, et la distribution uniforme est obtenue dans le cas particulier où $\gamma = 1$.

La loi bêta est une distribution alternative de Lehmann [Lehmann, 1953], où la distribution F_1 est exprimée simplement en fonction de F_0. En effet, les densités de probabilité



(a) Sous H_0 : les distributions des variables sont $\mathcal{N}(0{,}0{,}1{,}0)$ pour $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$. La distribution uniforme sur [0,1] est superposée en rouge.



(b) Sous H_1 : les distributions des variables sont $\mathcal{N}(0{,}0{,}1{,}0)$ pour $\boldsymbol{X}_1$ et $\mathcal{N}(0{,}4{,}1{,}0)$ pour $\boldsymbol{X}_2$. Les distributions bêta de paramètres $(\gamma{,}1{,}0)$ sont superposées pour $\gamma = 0{,}01$ (en rouge), $\gamma = 0{,}50$ (en vert), $\gamma = 0{,}90$ (en bleu).

FIGURE 3.2 – Exemple de distributions sous H_0 (figure 3.2a) et sous H_1 (figure 3.2b) de la p-valeur P du test t de Student bilatéral. Les histogrammes sont issus de l'application du test sur 1000000 de couples d'échantillons $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$ de 100 points chacun. Les distributions uniformes et bêta correspondant au modèle (3.12) sont ajoutées pour comparaison.

étant

$$f_0(P) = \mathbb{1}_{\{[0,1]\}}(P) \text{ sous } H_0, \tag{3.6}$$

$$f_1(P) = \frac{\Gamma(1+\gamma)}{\Gamma(\gamma)\Gamma(1)} P^{\gamma-1}\, \mathbb{1}_{\{[0,1]\}}(P) = \gamma P^{\gamma-1}\, \mathbb{1}_{\{[0,1]\}}(P) \text{ sous } H_1, \tag{3.7}$$

alors $F_1(P) = P^\gamma\, \mathbb{1}_{\{[0,1]\}}(P)$, ce qui vérifie la relation $F_1(P) = F_0(P)^\gamma$. Le paramètre γ est déterminé en fonction d'un seuil d'acceptation noté α, tel que la p-valeur P prenne la valeur α avec la même probabilité sous H_0 et sous H_1 :

$$\Pr(P = \alpha | H_0) = \Pr(P = \alpha | H_1) \tag{3.8}$$

donc

$$f_0(\alpha) = f_1(\alpha). \tag{3.9}$$

Ainsi sous H_1, γ est l'unique solution dans $[0{,}1[$ de l'équation

$$\gamma \alpha^{\gamma-1} = 1, \tag{3.10}$$

pour un seuil α donné dans $[0{,}e^{-1}]$. La valeur choisie pour α dépend des performances souhaitées pour la détection des ruptures. En effet, il est montré dans la section 3.2 que ce seuil d'acceptation permet de fixer le niveau de contrôle de la détection de la première

rupture. On notera que si $\alpha \geq e^{-1}$, alors le paramètre $\gamma = 1$ est l'unique solution de (3.10), et la distribution alternative sous H_1 se réduit à la distribution uniforme de l'hypothèse nulle H_0. Le modèle ne permet alors plus de différencier les deux hypothèses. Par la suite, on suppose donc que $\alpha \in [0,e^{-1}]$.

Choix du test

La p-valeur est une fonction de la statistique T, obtenue sur un échantillon donné. Dans le modèle du *Bernoulli Detector*, la statistique est calculée pour une position i dans le signal $\boldsymbol{X}$, et pour la segmentation $\boldsymbol{R}$. La variable représentant la p-valeur est donc notée $P_i(\boldsymbol{X})$. Le test statistique choisi pour la méthode permet d'évaluer les p-valeurs en chaque point X_i de la série temporelle, $1 < i < n$. Il s'agit d'un test d'homogénéité à deux échantillons, dont l'hypothèse nulle correspond à l'absence de rupture au point i testé. Les conclusions établies au chapitre 2 quant à la robustesse aux distributions non gaussiennes et aux petits échantillons conduisent à préférer le test non paramétrique de WMW. Le *Bernoulli Detector* est alors un détecteur de ruptures dans la médiane de la série temporelle. Les hypothèses faites sur les données sont celles du test de WMW, présentées dans la partie 2.2 : sous H_0, la probabilité qu'une variable X_i d'un échantillon soit plus grande qu'une variable X_j de l'autre échantillon est égale à la probabilité que $X_i < X_j$ et sous H_1, les variables d'un échantillon sont stochastiquement plus grandes que les variables de l'autre échantillon. Contrairement à de nombreuses méthodes de détection de ruptures, les données ne sont pas supposées appartenir à une famille de distribution précise, comme la loi normale.

La statistique du test de WMW étant discrète, les modèles des lois uniformes et bêta, qui sont continues, constituent une approximation de la loi réelle des p-valeurs. En particulier, la propriété sur la distribution uniforme des p-valeurs sous H_0 ne s'applique pas directement, la distribution de P dans le cas discret étant stochastiquement plus grande que la loi uniforme [Casella et Berger, 2002, p. 77]. Sous H_0, la distribution asymptotique de P est la loi normale de moyenne m_U (2.15) et de variance σ_U^2 (2.16). Sous cette approximation continue, la p-valeur est définie par

$$\Pr(T \geq t|H_0) \approx 2\left[1 - \Phi\Big(\frac{|t - m_U|}{\sqrt{\sigma_U^2}}\Big)\right], \tag{3.11}$$

dans le cas d'un test bilatéral, où $\Phi(a)$ est la mesure de l'aire sous la courbe de la densité de probabilité de la loi normale standard, à gauche de la valeur a (voir [Lehmann et D'Abrera, 1975, chapitre 1,5, pages 23-25]). Pour de grands échantillons, l'approximation (3.11) des p-valeurs est employée, on se ramène alors à une loi continue pour laquelle le modèle des lois uniforme et bêta peut convenir. Il existe des p-valeurs adaptées à un test d'hypothèses dans le cas discret, notamment pour un usage sur de petits échantillons [Agresti et Gottard, 2007, Agresti et Kateri, 2011]. La figure 3.3 permet de visualiser la loi de probabilité des p-valeurs sous H_0, pour 1000000 de simulations sur des échantillons de $n_1 = n_2 = 10$ points (figures 3.3a, 3.3c et 3.3e à gauche) et sur des échantillons de $n_1 = n_2 = 50$ points (figures

3.3b, 3.3d et 3.3f à droite). La première ligne donne la fonction masse des p-valeurs (figures 3.3a et 3.3b) ; on observe que le nombre de valeurs possibles augmente avec la taille des échantillons. La fonction de répartition est tracée sur la deuxième ligne (figures 3.3c et 3.3d), et la droite d'équation $y = x$ correspondant à la distribution uniforme est ajoutée. Les fonctions masses sont représentées par un histogramme de 15 canaux sur l'intervalle [0,1] dans les figures 3.3e et 3.3f. On constate bien que la loi de probabilité des p-valeurs se rapproche de la loi uniforme (en rouge) quand la taille des échantillons augmente.

Pour simplifier le modèle du *Bernoulli Detector*, la loi uniforme est appliquée sous H_0 pour les p-valeurs du test de WMW, qui conduit à un test plus conservatif en particulier pour les petits échantillons. Précisons toutefois qu'en choisissant un autre test d'hypothèses, pour lequel la distribution de T est continue, le modèle uniforme est valable. Les statistiques du test de WMW sont obtenues par l'approximation normale, bien que celle-ci ne soit adaptée qu'aux échantillons d'une certaine taille [1], l'approximation étant rapidement justifiée à partir de quelques d'observations (voir [Bellera *et al.*, 2010]).

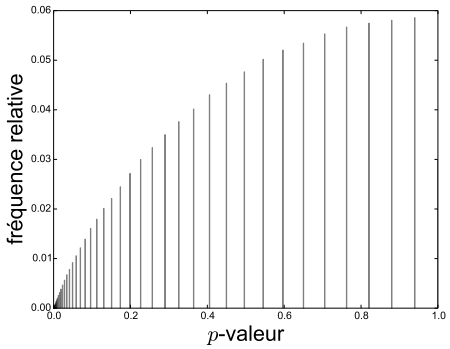
Vraisemblance marginale composite

Pour construire la fonction de vraisemblance, les p-valeurs sont déterminées pour chaque position $1 < i < n$ sur la série temporelle $\boldsymbol{X}$, dont la segmentation est donnée par $\boldsymbol{R}$. La statistique au point i est obtenue par application du test entre les deux échantillons $\boldsymbol{X}_- = (X_{i^-+1}, \ldots, X_i)$ et $\boldsymbol{X}_+ = (X_{i+1}, \ldots, X_{i^+})$, où i^- et i^+ sont les positions des ruptures situées avant et après le point testé X_i selon $\boldsymbol{R}$, elle est notée $P_i(\boldsymbol{X}_-, \boldsymbol{X}_+)$. Le schéma de la figure 3.4 représente comment les échantillons $\boldsymbol{X}_-$ et $\boldsymbol{X}_+$ sont définis par les ruptures en i^- et i^+ pour le calcul de la p-valeur au point i. Le modèle bêta-uniforme sur ces p-valeurs, considérées comme des variables aléatoires, conduit à la vraisemblance marginale suivante :

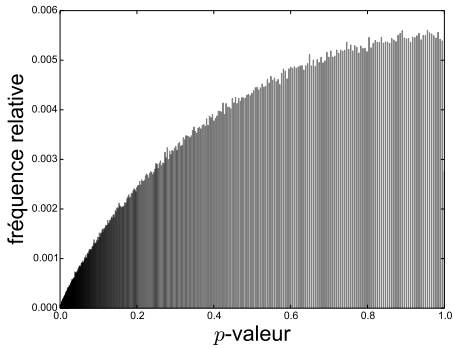
$$f(P_i(\boldsymbol{X}_-, \boldsymbol{X}_+)|\boldsymbol{R}) = \begin{cases} \mathbb{1}_{\{[0,1]\}}(P_i(\boldsymbol{X}_-, \boldsymbol{X}_+)) & \text{si } R_i = 0, \\ \gamma P_i(\boldsymbol{X}_-, \boldsymbol{X}_+)^{\gamma-1}\, \mathbb{1}_{\{[0,1]\}}(P_i(\boldsymbol{X}_-, \boldsymbol{X}_+)) & \text{si } R_i = 1, \end{cases} \tag{3.12}$$

pour tout $1 < i < n$ et pour une valeur donnée de γ, solution de l'équation (3.10). Par conséquent γ est le même pour tous les intervalles. Finalement, en s'inspirant de la relation (3.5), le terme d'attache aux données du *Bernoulli Detector* est construit à partir du

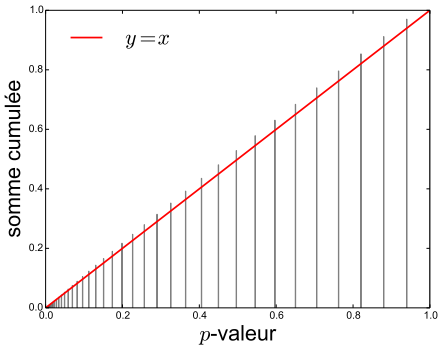
1. Les auteurs ne s'accordent pas sur une recommandation particulière au sujet les tailles limites de n_1 et n_2 au-delà desquelles l'approximation normale pour la statistique de test est justifiée. [Mann et Whitney, 1947] proposent $n_1 = n_2 = 8$, [Siegel, 1956] propose $n' > 20$, où n' est la taille de l'échantillon qui prend les plus grandes valeurs. Les documentations des différents langages de programmation ne donnent pas non plus les mêmes recommandations : la fonction `python` scipy.stats.mannwhitneyu fixe la limite à $n_1 > 20$ et $n_2 > 20$ (voir `http://docs.scipy.org/doc/scipy-0.16.0/reference/generated/scipy.stats.mannwhitneyu.html`), la fonction `R` wilcox.test suggère $n_1 > 50$ et $n_2 > 50$ (voir `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/wilcox.test.html`). De plus, ces limites varient en présence de valeurs égales dans les échantillons.



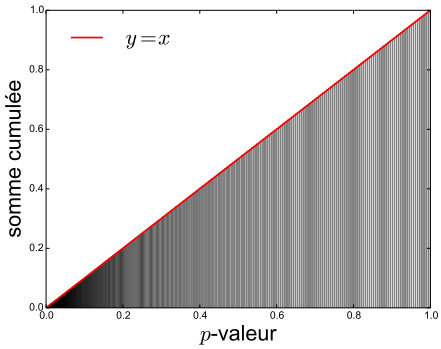
(a) Fonction masse avec $n_1 = n_2 = 10$.



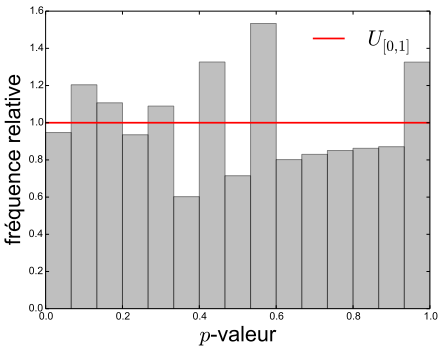
(b) Fonction masse avec $n_1 = n_2 = 50$.



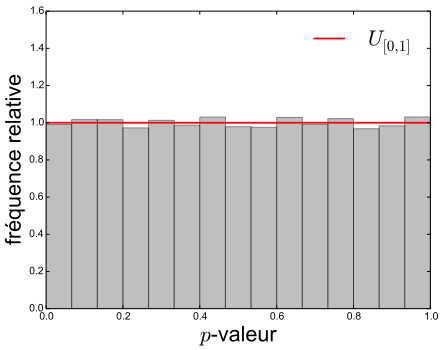
(c) Fonction de répartition avec $n_1 = n_2 = 10$.



(d) Fonction de répartition avec $n_1 = n_2 = 50$.



(e) Histogramme avec $n_1 = n_2 = 10$.



(f) Histogramme avec $n_1 = n_2 = 50$.

FIGURE 3.3 – Loi de probabilité discrète des p-valeurs sous H_0 pour le test de WMW, appliqué sur 1000000 de simulations. À gauche les échantillons sont de petites tailles ($n_1 = n_2 = 10$), et à droite les échantillons sont de grandes tailles ($n_1 = n_2 = 50$). La première ligne donne la fonction de masse, la deuxième ligne donne la fonction de répartition et la droite d'équation $y = x$ correspondant à l'approximation uniforme. La troisième ligne est une représentation des fonctions de masse par un histogramme de 15 canaux, recouvrant tout le support [0,1]. La distribution uniforme sur [0,1] est tracée en rouge.

produit des vraisemblances marginales (3.12)

$$L_*(\boldsymbol{X}|\boldsymbol{R}) = \prod_{i=2}^{n-1} f(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)|\boldsymbol{R}) \tag{3.13}$$

$$\boxed{L_*(\boldsymbol{X}|\boldsymbol{R}) = \prod_{i=2}^{n-1} \left(\gamma P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)^{\gamma-1}\right)^{R_i}.} \tag{3.14}$$

Rappelons que, par convention, les variables indicatrices des extrémités du signal R_1 et R_n sont considérées comme des ruptures.

Le terme $L_*(\boldsymbol{X}|\boldsymbol{R})$ n'est en réalité pas une fonction de vraisemblance au sens propre du terme, telle qu'elle est définie par [Monahan et Boos, 1992] afin d'obtenir des probabilités de couverture correctes. La fonction de vraisemblance théorique $L(\boldsymbol{X}|\boldsymbol{R})$ doit tenir compte des dépendances entre les p-valeurs. Il est possible d'exprimer les corrélations entre les p-valeurs voisines sous H_0, mais en présence de ruptures, il est plus difficile de modéliser ces corrélations. En effet, les tests statistiques dont sont issues les p-valeurs sont appliqués sur certaines segments, et les p-valeurs voisines $P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)$ et $P_{i^+}(\boldsymbol{X}_+,\boldsymbol{X}_{2+})$ n'ont en commun que le seul segment $\boldsymbol{X}_+$ qui les sépare (voir la figure 3.4). Afin d'éviter de devoir introduire des hypothèses fortes pour modéliser les dépendances entre les p-valeurs, la fonction de vraisemblance est remplacée par la fonction de vraisemblance marginale composite $L_*(\boldsymbol{X}|\boldsymbol{R})$ définie en (3.14).

Les fonctions de vraisemblance composite sont souvent employées pour résoudre des problèmes où l'expression analytique de la fonction de vraisemblance est difficile voire impossible à établir, ou bien lorsqu'elle est trop coûteuse à calculer. Le principe consiste à substituer la loi de probabilité conditionnelle $L(\boldsymbol{X}|\boldsymbol{R})$ par une fonction de vraisemblance composite, constituée de fonctions de vraisemblance de dimension inférieure $L_k(\boldsymbol{X}|\Theta) \propto f(\boldsymbol{X} \in A_k|\Theta)$, où Θ est le vecteur des paramètres à estimer et $\{A_1,\ldots,A_K\}$ est un ensemble d'événements marginaux ou conditionnels. La fonction de vraisemblance est donc décomposée sur deux blocs de données ou plus, paire à paire ou sur chacune des variables X_i. La nouvelle fonction est de la forme :

$$L_*(\boldsymbol{X}|\Theta) = \prod_{k=1}^{K} L_k(\boldsymbol{X}|\Theta)^{w_k}, \tag{3.15}$$

où les poids $w_k \geq 0$ doivent être choisis [Varin, 2008, Varin *et al.*, 2011]. L'estimateur du maximum du logarithme de la fonction de vraisemblance composite est noté $\hat{\Theta}_{CL}$ et peut être déduit de la fonction de score $u(\Theta;\boldsymbol{X}) = \bigtriangledown_\Theta CL(\Theta;\boldsymbol{X})$, où $CL(\Theta;\boldsymbol{X})$ est le logarithme de la fonction de vraisemblance composite (3.15). Pour remplacer la fonction de vraisemblance complète par une fonction de vraisemblance composite $L_*(\boldsymbol{X}|\Theta)$, celle-ci doit être validée. Plusieurs approches ont été proposées, et sont choisies au cas par cas. Dans un cadre bayésien, comme dans [Pauli *et al.*, 2011, Ribatet *et al.*, 2012], l'introduction de $L_*(\boldsymbol{X}|\Theta)$ doit conduire à une loi de probabilité a posteriori valable. Parmi les critères proposés pour déterminer la validité, on trouve la comparaison des intervalles de recouvrement de la loi a posteriori, obtenus par exemple par simulation [Monahan et Boos,

1992], ou l'étude du comportement asymptotique de $L_*(\boldsymbol{X}|\Theta)$ [Varin *et al.*, 2011]. Ainsi, lorsque les observations $X_1,\ldots,X_n$ sont i.i.d. et que n tend vers l'infini, le théorème central limite s'applique sous certaines conditions de régularité pour $\hat{\Theta}_{CL}$, qui suit la loi normale de dimension p :

$$\sqrt{n}(\hat{\Theta}_{CL} - \Theta) \rightarrow \mathcal{N}_p(0, G^{-1}(\Theta)), \tag{3.16}$$

où $G(\Theta)$ est la matrice d'information de Godambe pour une observation :

$$G(\Theta) = H(\Theta)J^{-1}(\Theta)H(\Theta) \tag{3.17}$$

avec

$$H(\Theta) = E_\Theta[-\bigtriangledown_\Theta u(\Theta;\boldsymbol{X})] \quad \text{et} \quad J(\Theta) = \text{var}_\Theta(u(\Theta;\boldsymbol{X})). \tag{3.18}$$

Dans notre cas, les poids w_k sont égaux et les paramètres de $L_*(\boldsymbol{X}|\Theta)$ sont les variables binaires R_i. Les théorèmes de convergence asymptotiques de l'état de l'art ne sont donc pas applicables. Les fonctions $L_k(\boldsymbol{X}|\Theta)^{w_k}$ sont les vraisemblances marginales des p-valeurs et sont supposées indépendantes. Par conséquent l'intervalle de recouvrement de la densité de probabilité a posteriori issue de $L_*(\boldsymbol{X}|\boldsymbol{R})$ diffère probablement du vrai.

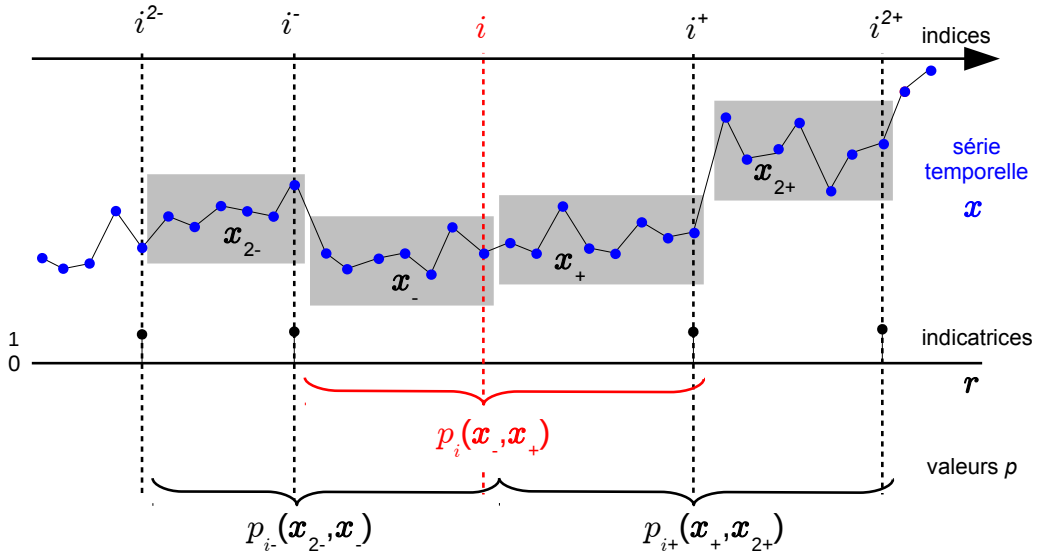


FIGURE 3.4 – Illustration du calcul de $p_i(\boldsymbol{x}_-,\boldsymbol{x}_+)$ au point d'indice i (pointillés rouge), en fonction de la série temporelle $\boldsymbol{x}$ (en bleu) et de la segmentation donnée par $\boldsymbol{r}$, où les indicatrices non nulles sont représentées par un trait vertical noir. Les échantillons $\boldsymbol{x}_-$ et $\boldsymbol{x}_+$ (grisés) sur lesquels on applique le test d'homogénéité sont définis par rapport à i et aux positions des ruptures voisines en i^- et i^+. Les accolades noires indiquent les observations sélectionnées pour le calcul des p-valeurs aux positions i^- et i^+ où se trouvent des ruptures, en fonction des ruptures en i^{2-} et en i^- pour $p_{i^-}(\boldsymbol{x}_{2-},\boldsymbol{x}_-)$ et en i^+ et i^{2+} pour $p_{i^+}(\boldsymbol{x}_+,\boldsymbol{x}_{2+})$.

3.1.3 Densité a priori

Pour chaque variable X_i, $1 < i < n$, il existe une probabilité, notée q, que X_i soit une rupture. L'hyperparamètre q est introduit dans le modèle, pour formuler la probabilité que la variable indicatrice R_i prenne la valeur 0 ou 1, selon une loi de Bernoulli :

$$f(R_i|q) = q^{R_i}(1-q)^{(1-R_i)}. \tag{3.19}$$

Comme on suppose que les indicatrices $(R_i)_{1<i<n}$ sont *a priori* indépendantes les unes des autres, alors la loi de probabilité a priori du paramètre $\boldsymbol{R}$ s'écrit :

$$f(\boldsymbol{R}|q) = \prod_{i=2}^{n-1} q^{R_i}(1-q)^{(1-R_i)}. \tag{3.20}$$

La probabilité q est considérée comme une variable aléatoire, à laquelle on choisit d'attribuer un a priori non informatif de Jeffrey : la loi bêta $\mathcal{B}e(\frac{1}{2},\frac{1}{2})$. On obtient alors l'expression de la loi a priori sur le paramètre $\boldsymbol{R}$ de la relation (3.3) :

$$f(\boldsymbol{R}) = f(\boldsymbol{R}|q)f(q) \tag{3.21}$$

$$\boxed{f(\boldsymbol{R}) = \left(\prod_{i=2}^{n-1} q^{R_i}(1-q)^{(1-R_i)}\right) \frac{1}{\pi} q^{-\frac{1}{2}}(1-q)^{-\frac{1}{2}}.} \tag{3.22}$$

3.1.4 Densité a posteriori

Finalement, l'expression de la loi de probabilité a posteriori de tous les paramètres sachant les observations $\boldsymbol{X}$ est obtenue à partir de la vraisemblance marginale composite (3.14) et de l'a priori sur $\boldsymbol{R}$ (3.22) :

$$f(\boldsymbol{R},q|\boldsymbol{X}) \propto \left(\prod_{i=1}^{n} (\gamma P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)^{\gamma-1})^{R_i}\right) \left(\prod_{i=1}^{n} q^{R_i}(1-q)^{1-R_i}\right) \frac{1}{\pi} q^{-\frac{1}{2}}(1-q)^{-\frac{1}{2}}. \tag{3.23}$$

Pour s'affranchir du paramètre de nuisance q, la loi a posteriori $f(\boldsymbol{R},q|\boldsymbol{X})$ est marginalisée par rapport à q. En effet, dans l'équation (3.23), ce paramètre suit la loi bêta $\mathcal{B}e(K+\frac{1}{2},n-K-\frac{3}{2})$, où $K = \sum_{i=2}^{n-1} R_i$ est le nombre de ruptures détectées dans le signal, en-dehors du premier et du dernier point, considérés comme des ruptures par convention. La densité de probabilité a posteriori de $\boldsymbol{R}$ sachant $\boldsymbol{X}$ est donc :

$$\boxed{f(\boldsymbol{R}|\boldsymbol{X}) \propto \Gamma\left(K+\frac{1}{2}\right)\Gamma\left(n-K-\frac{3}{2}\right)\prod_{i=2}^{n-1}\left(\gamma P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)^{\gamma-1}\right)^{R_i},} \tag{3.24}$$

et dépend de la taille du vecteur $\boldsymbol{X}$, du nombre d'événements détectés, du paramètre γ, fixé par le seuil d'acceptation α selon (3.10), et des p-valeurs obtenues grâce au test statistique appliqué sur chaque point dans les segments définis par $\boldsymbol{R}$.

Les ruptures sont localisées aux positions $\tau_1, \ldots, \tau_K$ qui maximisent la densité (3.24) pour les observations de la série temporelle $\boldsymbol{X}$. Le vecteur $\boldsymbol{R}_{MAP}$, pour Maximum A Posteriori, qui y est associé, correspond à

$$\boldsymbol{R}_{MAP} = \underset{\boldsymbol{R}\in\{0,1\}^{n-2}}{\operatorname{argmax}} f(\boldsymbol{R}|\boldsymbol{X}). \tag{3.25}$$

Comme la résolution analytique directe du problème de maximisation (3.25) n'est pas possible, on applique une méthode numérique, présentée dans la partie 3.3, pour déterminer un estimateur de $\boldsymbol{R}_{MAP}$, qui est noté $\hat{\boldsymbol{R}}_{MAP}$. Une fois la densité de probabilité a posteriori exprimée, il reste à établir les propriétés du *Bernoulli Detector* pour la détection des ruptures.

3.2 Contrôle de la détection

L'expression de la densité de probabilité a posteriori (3.24) obtenue avec le modèle du *Bernoulli Detector* dépend des $P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)$, $1 < i < n$, et de l'hyperparamètre γ, issu du choix de la distribution alternative de $P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)$, sous l'hypothèse $H_1(i)$ que X_i est une rupture. La valeur de γ est liée à celle du niveau d'acceptation α du test statistique dont résultent les p-valeurs, selon la relation (3.10). Ce paramètre α permet de définir la probabilité de détection et le risque d'erreur de type I pour une unique rupture. On se place d'abord dans le cas simple où aucune rupture n'a été détectée ($R_i = 0$ pour tout $1 < i < n$). Le lemme suivant fournit la condition à remplir sur $P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)$ pour détecter la première rupture au point i.

Lemme 3.2.1 (détection de la première rupture à la position i)**.** *Dans le cas de la détection de la première rupture à la position i, $1 < i < n$, la densité de probabilité a posteriori* (3.24) *est maximisée pour $R_i = 1$ si et seulement si :*

$$P_i(\boldsymbol{X}_-,\boldsymbol{X}_+) \leq \frac{\alpha}{(2n-1)^{\frac{1}{1-\gamma}}}, \tag{3.26}$$

où le paramètre α est le niveau d'acceptation du test statistique, introduit dans (3.10)*, avec $0 \leq \gamma < 1$.*

La proposition suivante évalue l'erreur de type I (une rupture inexistante a été détectée à tort, voir la partie 1.1.2) dans le cas où la série temporelle $\boldsymbol{X}$ ne présente aucune rupture.

Proposition 3.2.1 (contrôle global de la détection d'une unique rupture)**.** *On suppose que la série temporelle ne comporte pas de ruptures. Alors la probabilité de faire une erreur de type I dans l'estimateur du MAP $\hat{\boldsymbol{R}}_{MAP}$ contenant une seule rupture est contrôlée au niveau η tel que*

$$\eta = \frac{n-2}{(2n-5)^{\frac{1}{1-\gamma}}}\alpha. \tag{3.27}$$

Cette propriété assure notamment que $\eta \leq \frac{n-2}{2n-5}\alpha \leq \alpha$, pour tout $n > 2$.

Ces deux résultats sont démontrés ci-après, par l'introduction du facteur de Bayes au point i, puis l'application de l'inégalité de Boole. Pour le point i, le facteur de Bayes s'écrit comme le rapport des vraisemblances sous $H_0(i)$ et sous $H_1(i)$

$$B_{10,i} = \frac{f(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)|H_1(i))}{f(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)|H_0(i))} = \frac{f(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)|R_i=1)}{f(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)|R_i=0)} \tag{3.28}$$

donc d'après (3.12), dans le modèle bêta-uniforme,

$$B_{10,i} = \gamma P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)^{\gamma-1}. \tag{3.29}$$

Lorsqu'aucune rupture n'a été détectée en-dehors du point d'indice i, les échantillons de $P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)$ sont $\boldsymbol{X}_- = (X_1,\ldots,X_i)$ et $\boldsymbol{X}_+ = (X_{i+1},\ldots,X_n)$. L'expression de la densité a posteriori (3.24) employée dans cette démonstration est

$$f(\boldsymbol{R}|\boldsymbol{X}) \propto \left(\prod_{i=2}^{n-1} f(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)|R_i=0)^{1-R_i} f(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)|R_i=1)^{R_i}\right) \times \Gamma\left(K+\frac{1}{2}\right)\Gamma\left(n-K-\frac{3}{2}\right). \tag{3.30}$$

Démonstrations du lemme et de la proposition. Le rapport des probabilités a posteriori pour une potentielle unique rupture à la position i s'exprime comme

$$\begin{aligned}
&\frac{\Pr\left(R_2=0,\ldots,R_{i-1}=0,R_i=1,R_{i+1}=0,\ldots,R_{n-1}=0\right)}{\Pr\left(R_2=0,\ldots,R_{n-1}=0\right)} \\
&\quad = \frac{f(R_2=0,\ldots,R_{i-1}=0,R_i=1,R_{i+1}=0,\ldots,R_{n-1}=0|\boldsymbol{X})}{f(R_2=0,\ldots,R_{n-1}=0|\boldsymbol{X})} \qquad (3.31)\\
&\quad = \frac{f(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)|R_i=1)\prod\limits_{\substack{1<j<n\\ j\neq i}} f(P_j(\boldsymbol{X}_-,\boldsymbol{X}_+)|R_j=0)}{\prod\limits_{1<j<n} f(P_j(\boldsymbol{X}_-,\boldsymbol{X}_+)|R_j=0)} \qquad (3.32)\\
&\quad \times \frac{\Gamma\left(1+\frac{1}{2}\right)\Gamma\left(n-1-\frac{3}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(n-\frac{3}{2}\right)}. \qquad (3.33)
\end{aligned}$$

En l'absence de rupture à toutes les positions $j \neq i$, les p-valeurs $P_j(\boldsymbol{X}_-,\boldsymbol{X}_+)$ suivent la loi uniforme. On obtient alors :

$$\begin{aligned}
&\frac{\Pr\left(R_2=0,\ldots,R_{i-1}=0,R_i=1,R_{i+1}=0,\ldots,R_{n-1}=0\right)}{\Pr\left(R_2=0,\ldots,R_{n-1}=0\right)} \\
&\quad = \frac{f(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)|R_i=1)}{f(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)|R_i=0)} \times \frac{1}{2n-5}
\end{aligned} \tag{3.34}$$

$$= B_{10,i} \times O_{10}, \tag{3.35}$$

où $B_{10,i}$ est le facteur de Bayes (3.28) au point i et O_{10} est le rapport des probabilités a priori pour la détection d'une unique rupture :

$$O_{10} = \frac{1}{2n-5}. \tag{3.36}$$

Alors l'unique rupture au point i est détectée au sens du MAP si et seulement si le rapport des probabilités (3.35) est plus grand que 1, ce qui revient à la condition

$$B_{10,i} \geq O_{10}^{-1}, \tag{3.37}$$

soit, d'après (3.29) et (3.36),

$$\gamma P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)^{\gamma-1} \geq 2n-5. \tag{3.38}$$

En introduisant la paramétrisation choisie dans (3.10) pour la distribution alternative de la p-valeur $P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)$, on a $\gamma^{\frac{1}{1-\gamma}} = \alpha$ avec $\gamma \in [0,1[$. Alors, en posant $\eta = (n-2)\alpha/\left((2n-5)^{\frac{1}{1-\gamma}}\right)$, (3.38) devient

$$P_i(\boldsymbol{X}_-,\boldsymbol{X}_+) \leq \frac{\eta}{n-2}, \tag{3.39}$$

ce qui prouve le lemme 3.2.1.

D'après (3.39), la probabilité de détecter à tort une rupture dans une série temporelle qui n'en comporte pas, c'est-à-dire de faire une erreur de type I, est :

$$\Pr\left\{\bigcup_{1<i<n} P_i(\boldsymbol{X}_-,\boldsymbol{X}_+) \leq \frac{\eta}{n-2}\right\}. \tag{3.40}$$

D'après l'inégalité de Boole

$$\Pr\left\{\bigcup_{1<i<n} P_i(\boldsymbol{X}_-,\boldsymbol{X}_+) \leq \frac{\eta}{n-2}\right\} \leq \sum_{1<i<n} \Pr\left(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+) \leq \frac{\eta}{n-2}\right), \tag{3.41}$$

quelles que soient les dépendances entre les p-valeurs. Comme on suppose qu'il n'y a pas de rupture dans la série temporelle, toutes les p-valeurs $P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)$, $1 < i < n$, sont distribuées uniformément sur [0,1], donc

$$\sum_{1<i<n} \Pr\left(P_i(\boldsymbol{X}_-,\boldsymbol{X}_+) \leq \frac{\eta}{n-2}\right) = (n-2)\frac{\eta}{n-2} = \eta, \tag{3.42}$$

ce qui correspond au niveau de contrôle souhaité. On note de plus que pour toute taille $n > 2$ et $0 \leq \gamma < 1$,

$$(2n-5)^{\frac{1}{1-\gamma}} \geq 2n-5, \tag{3.43}$$

donc

$$\frac{n-2}{(2n-5)^{\frac{1}{1-\gamma}}}\alpha \leq \frac{n-2}{2n-5}\alpha, \tag{3.44}$$

et

$$\frac{n-2}{2n-5}\alpha \leq \alpha, \tag{3.45}$$

alors $\eta \leq \frac{n-2}{2n-5}\alpha \leq \alpha$, ce qui conclut la démonstration de la proposition 3.2.1. □

On note que ces résultats sont démontrés pour toutes les p-valeurs, et non pas asymptotiquement. Le contrôle de la détection de la première rupture est alors obtenu même avec des signaux de petite taille n. La proposition 3.2.1 concerne le contrôle global de la détection d'une rupture, même lorsque les tests en chaque position $1 < i < n$ sont fortement dépendants. Ceci montre que le choix des a priori de notre modèle induit implicitement une correction de Bonferroni [Dunn, 1961]. En pratique, dans la cadre des expériences présentées dans la partie 3.4, la proposition 3.2.1 n'a pas été utilisée directement pour déterminer la valeur de α pour un niveau de contrôle donné de la détection d'une seule rupture.

Des résultats théoriques pour la détection de multiples ruptures ont été démontrés pour certaines méthodes, comme [Vert et Bleakley, 2010, Harchaoui et Lévy-Leduc, 2010, Lung-Yut-Fong *et al.*, 2011c], en général asymptotiquement. Dans notre modèle du *Bernoulli Detector*, la présence de plus d'une rupture dans la série temporelle engendre des dépendances entre les p-valeurs, malheureusement impossibles à décrire. Le contrôle de la détection est donc uniquement étudié pour au plus une rupture.

3.3 Algorithme

La méthode du *Bernoulli Detector* est mise en œuvre sur la série temporelle $\boldsymbol{X}$ afin de détecter les ruptures par l'estimation du vecteur des variables indicatrices $\boldsymbol{R}$. Cette partie présente l'algorithme développé pour l'application de la méthode, basé sur une méthode de Monte Carlo par Chaînes de Markov (en anglais *Markov Chain Monte Carlo*, abrégé en MCMC) avec un échantillonneur de Gibbs, puis une approximation permettant de réduire la complexité calculatoire.

3.3.1 Méthode MCMC

La solution $\boldsymbol{R}_{MAP}$ qui maximise la densité de probabilité a posteriori $f(\boldsymbol{R}|\boldsymbol{X})$ ne peut pas être obtenue analytiquement à partir de l'expression (3.24), ni par recherche exhaustive parmi toutes les valeurs possibles de $\boldsymbol{R}$, puisque 2^{n-2} solutions seraient à tester. C'est pourquoi l'estimateur $\hat{\boldsymbol{R}}_{MAP}$ est déterminé numériquement par une méthode MCMC. Le principe est rappelé en annexe B. La méthode MCMC génère une chaîne de Markov $(\boldsymbol{R}^{(v)})_v$ dont la loi stationnaire est la distribution d'intérêt, c'est-à-dire la loi a posteriori $f(\boldsymbol{R}|\boldsymbol{X})$. L'état $\boldsymbol{R}^{(v+1)}$ est simulé à partir de l'état précédent selon la probabilité de transition $f(\boldsymbol{R}^{(v+1)}|\boldsymbol{R}^{(v)})$.

Pour un nombre suffisamment grand d'itérations V, au-delà d'une période de préchauffe (dite de *burn-in* en anglais), la chaîne résultante $\boldsymbol{R}^{(V)}$ est approximativement distribuée selon $f(\boldsymbol{R}|\boldsymbol{X})$, quelle que soit la valeur initiale $\boldsymbol{R}^{(0)}$. La figure 3.5 illustre l'évolution de la probabilité a posteriori obtenue pour chaque état de la chaîne $(\boldsymbol{R}^{(v)})_v$, avec l'algorithme 4 présenté dans la suite. On vérifie bien sur cet exemple que la période de préchauffe est courte. En pratique dans l'algorithme du *Bernoulli Detector*, le nombre d'itérations est fixé empiriquement, suffisamment grand pour atteindre la loi stationnaire, puis $\hat{\boldsymbol{R}}_{MAP}$ est mis à jour sur V simulations, tel que

$$\hat{\boldsymbol{R}}_{MAP}(V) = \underset{0 \leq v \leq V}{\operatorname{argmax}}\, f(\boldsymbol{R}^{(v)}|\boldsymbol{X}). \tag{3.46}$$

Un autre estimateur de $\boldsymbol{R}$, le Minimum Mean Square Error ou MMSE, peut également être évalué. Il est obtenu d'après la configuration moyenne de tous les $\boldsymbol{R}^{(v)}$, pour $1 \leq v \leq V$. En pratique, comme les critères d'évaluation qui sont employés sont facilement calculables à partir de $\hat{\boldsymbol{R}}_{MAP}(V)$, seuls les résultats de l'estimation du MAP sont présentés à la section 3.4.

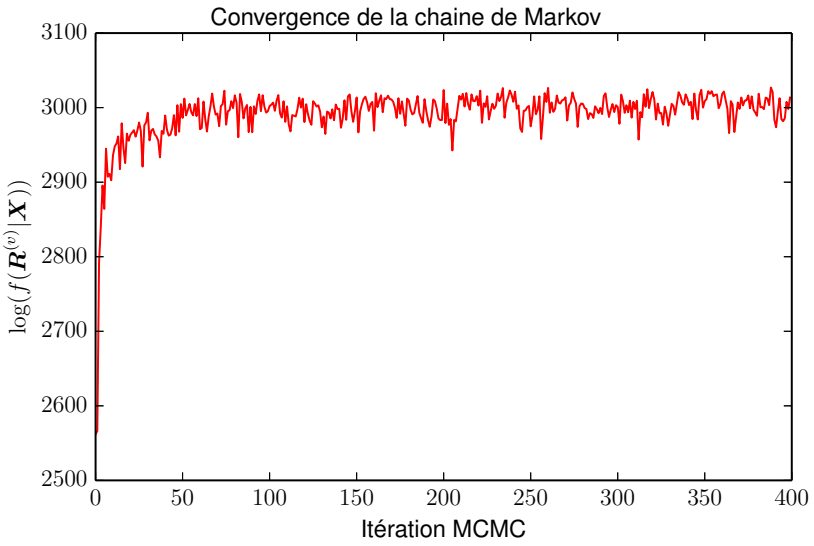


FIGURE 3.5 – Exemple d'évolution de la probabilité posteriori $f(\boldsymbol{R}^{(v)}|\boldsymbol{X})$ (à un facteur de normalisation près) obtenue avec une chaîne de Markov $(\boldsymbol{R}^{(v)})_v$, générée par l'algorithme 4, pour $V = 400$ itérations.

3.3.2 Échantillonneur de Gibbs

L'échantillonneur de Gibbs paraît indiqué dans notre cas de figure, car il permet d'éviter de simuler $\boldsymbol{R}$ (de dimension finie $n-2$) directement selon la distribution multivariée $f(\boldsymbol{R}|\boldsymbol{X})$, en effectuant l'échantillonnage successif des composantes R_i du paramètre $\boldsymbol{R}$,

$1 < i < n$, via une décomposition de la loi jointe en lois de probabilité conditionnelles complètes. On suppose que les lois conditionnelles $f(R_i|R_2,\ldots,R_{n-1};\boldsymbol{X})$ sont connues, pour tout $1 < i < n$. L'algorithme de Gibbs à l'itération MCMC v consiste à générer successivement les paramètres R_i selon les étapes données par l'algorithme 2. La loi stationnaire de chacune de ces étapes est la loi jointe $f(R_2,\ldots,R_{n-1}|\boldsymbol{X})$. L'échantillonneur de Gibbs est ainsi construit directement à partir de la loi cible.

Algorithme 2 : Échantillonneur de Gibbs

$R_2^{(v)} \sim f(R_2|R_3^{(v-1)},\ldots,R_{n-1}^{(v-1)};\boldsymbol{X})$
$R_3^{(v)} \sim f(R_3|R_2^{(v)},R_4^{(v-1)},\ldots,R_{n-1}^{(v-1)};\boldsymbol{X})$
$\ldots$
$R_i^{(v)} \sim f(R_i|R_2^{(v)},\ldots,R_{i-1}^{(v)},R_{i+1}^{(v-1)}\ldots,R_{n-1}^{(v-1)};\boldsymbol{X})$
$\ldots$
$R_{n-1}^{(v)} \sim f(R_{n-1}|R_2^{(v)},\ldots,R_{n-2}^{(v)};\boldsymbol{X})$

Les indices temporels $1 < i < n$ sont tirés aléatoirement pour déterminer dans quel ordre échantillonner les indicatrices R_i à chaque itération MCMC, ce qui permet d'accélérer la convergence. On note m la nouvelle position de l'indice i après le mélange aléatoire, $1 < m < n$. La permutation appliquée aux indices temporels est notée $\pi(.)$, elle est générée aléatoirement à chaque itération MCMC, et $\pi^{-1}(.)$ est son inverse. On a alors $m = \pi(i)$ et $i = \pi^{-1}(m)$. La distribution jointe de l'échantillonneur correspond à la distribution a posteriori donnée par l'expression (3.24). Les distributions conditionnelles de chaque R_i, $1 < i < n$, par rapport aux autres indicatrices $R_2,\ldots,R_{i-1},R_{i+1}\ldots,R_{n-1}$, en sont déduites. On note $\boldsymbol{R}_m^{(v)} = (R_2^{(v)},\ldots,R_m^{(v)},R_{m+1}^{(v-1)},\ldots,R_{n-1}^{(v-1)})$ le vecteur des variables indicatrices permutées en cours d'échantillonnage à la v^{e} itération MCMC, où les variables $(R_j^{(v)})_{1<j\leq m}$ sont dans l'état v, tandis que les variables $(R_j^{(v-1)})_{m+1\leq j<n}$ sont toujours dans l'état $v-1$. À cette étape, le vecteur $\boldsymbol{R}_{\setminus m}^{(v)}$ correspond au vecteur $\boldsymbol{R}_m^{(v)}$ privé de son m^{e} élément : $\boldsymbol{R}_{\setminus m}^{(v)} = (R_2^{(v)},\ldots,R_{m-1}^{(v)},R_{m+1}^{(v-1)},\ldots,R_{n-1}^{(v-1)})$.

Avec la notation permutée, on définit les deux probabilités jointes pour échantillonner R_m à l'itération MCMC v

$$\Pr(\boldsymbol{R}_{m,0}^{(v)}|\boldsymbol{X}) = \Pr(R_2^{(v)},\ldots,R_{m-1}^{(v)},0,R_{m+1}^{(v-1)},\ldots,R_{n-1}^{(v-1)}|\boldsymbol{X}), \tag{3.47}$$

$$\Pr(\boldsymbol{R}_{m,1}^{(v)}|\boldsymbol{X}) = \Pr(R_2^{(v)},\ldots,R_{m-1}^{(v)},1,R_{m+1}^{(v-1)},\ldots,R_{n-1}^{(v-1)}|\boldsymbol{X}). \tag{3.48}$$

Tout d'abord, les p-valeurs du test statistique sont évaluées pour les deux cas, d'après leur probabilité marginale (3.12). $\boldsymbol{R}_{i,0}$ et $\boldsymbol{R}_{i,1}$ ne différant qu'au niveau du coefficient R_i, les vecteurs des p-valeurs associés sont identiques en-dehors de la position i et aux positions des ruptures avant et après i, respectivement i^- et i^+. Les p-valeurs calculées sont donc $p_i(\boldsymbol{X}_-,\boldsymbol{X}_+)$, $p_{i^-}(\boldsymbol{X}_{2-},\boldsymbol{X}_-)$ et $p_{i^+}(\boldsymbol{X}_+,\boldsymbol{X}_{2+})$, comme représenté dans le schéma de la figure 3.4. Pour les deux segmentations $\boldsymbol{R}_{i,0}$ et $\boldsymbol{R}_{i,1}$, les probabilités a posteriori (3.47) et (3.48) sont déduites de l'expression (3.24). On obtient les probabilités conditionnelles

de l'échantillonneur de Gibbs :

$$\Pr(R_m = 0|\boldsymbol{R}^{(v)}_{\setminus m}; \boldsymbol{X}) = \frac{\Pr(\boldsymbol{R}^{(v)}_{m,0}|\boldsymbol{X})}{\Pr(\boldsymbol{R}^{(v)}_{m,0}|\boldsymbol{X}) + \Pr(\boldsymbol{R}^{(v)}_{m,1}|\boldsymbol{X})}, \tag{3.49}$$

$$\Pr(R_m = 1|\boldsymbol{R}^{(v)}_{\setminus m}; \boldsymbol{X}) = \frac{\Pr(\boldsymbol{R}^{(v)}_{m,1}|\boldsymbol{X})}{\Pr(\boldsymbol{R}^{(v)}_{m,0}|\boldsymbol{X}) + \Pr(\boldsymbol{R}^{(v)}_{m,1}|\boldsymbol{X})}. \tag{3.50}$$

Afin également d'accélérer la convergence, un échantillonnage par bloc est appliqué autour de position testée i, dans le référentiel temporel. L'échantillonneur de Gibbs présente en effet des difficultés à sortir d'une configuration locale où une rupture a été détectée, mais est légèrement mal localisée en i'. Pour aller vers la configuration où la rupture est correctement positionnée en i, il faut soit retirer la rupture en i' puis la replacer en i, soit ajouter la rupture en i puis supprimer celle en i', ces deux solutions occasionnant une diminution de la probabilité a posteriori dans la configuration sans rupture et celle avec deux ruptures en i et i' par rapport à la valeur de $f(\boldsymbol{R}|\boldsymbol{X})$ avec la rupture en i'. L'échantillonnage par bloc permet de repositionner ce type de rupture dans un voisinage proche. Cette solution est par exemple employée dans [Bourguignon et Carfantan, 2005], où une procédure est appliquée afin de déplacer une rupture à la position voisine lorsque la probabilité a posteriori est augmentée, et dans [Kail *et al.*, 2012], où les auteurs introduisent en plus une condition de distance minimale entre deux indicatrices à la valeur 1. Dans notre algorithme, lorsqu'une rupture a été détectée en i à l'itération $(v-1)$, les variables indicatrices au point i et aux points voisins $i-1$ et $i+1$ sont échantillonnées conjointement à l'itération suivante, afin de favoriser le déplacement de la rupture à une meilleure position (en $i-1$ ou en $i+1$), sa suppression, ou bien l'apparition de ruptures voisines (simultanément en $i-1$ et i, en i et $i+1$, en $i-1$ et $i+1$ ou en $i-1$, i et $i+1$). Ainsi, les indicatrices R_{i-1}, R_i et R_{i+1} sont simulées conjointement à l'itération v, pour toutes les solutions possibles, c'est-à-dire les vecteurs d'indicatrices (0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0) et (1,1,1). La figure 3.6 illustre l'évolution de la valeur de la probabilité a posteriori pour les stratégies d'échantillonnage de Gibbs sur chacune des indicatrices R_i, $1 < i < n$, et par bloc autour des ruptures. Les deux courbes sont obtenues sur une même série temporelle de 200 points, comportant 5 ruptures. On vérifie bien que l'échantillonnage par bloc permet de réduire la période de préchauffe de la méthode MCMC par rapport à l'échantillonnage individuel.

Les indicatrices R_{i-1}, R_i et R_{i+1} correspondent respectivement, sans le vecteur mélangé $\boldsymbol{R}_m$, aux coefficients $R_{m_1} = R_{\pi(i-1)}$, R_m et $R_{m_2} = R_{\pi(i+1)}$. Pour simplifier, et sans perte de généralité, on suppose que ces éléments sont consécutifs dans $\boldsymbol{R}_m$: $m_1+1 = m = m_2-1$. Le vecteur des indicatrices à mettre à jour (R_{m_1},R_m,R_{m_2}) prend la valeur $\rho \in G$ où $G = \{0,1\}^3$. Comme précédemment, on définit pour l'échantillonnage à l'itération MCMC v de ces variables le vecteur des indicatrices mélangées

$$\boldsymbol{R}^{(v)}_{m,\text{bloc}} = (R^{(v)}_2, \ldots, R^{(v)}_{m-2}, R^{(v)}_{m_1}, R^{(v)}_m, R^{(v)}_{m_2}, R^{(v-1)}_{m+2}, \ldots, R^{(v-1)}_{n-1}), \tag{3.51}$$

puis, pour le cas particulier où $(R_{m_1},R_m,R_{m_2}) = \rho$,

$$\boldsymbol{R}^{(v)}_{m,\text{bloc},\rho} = (R^{(v)}_2, \ldots, R^{(v)}_{m-2}, \rho(1), \rho(2), \rho(3), R^{(v-1)}_{m+2}, \ldots, R^{(v-1)}_{n-1}), \tag{3.52}$$

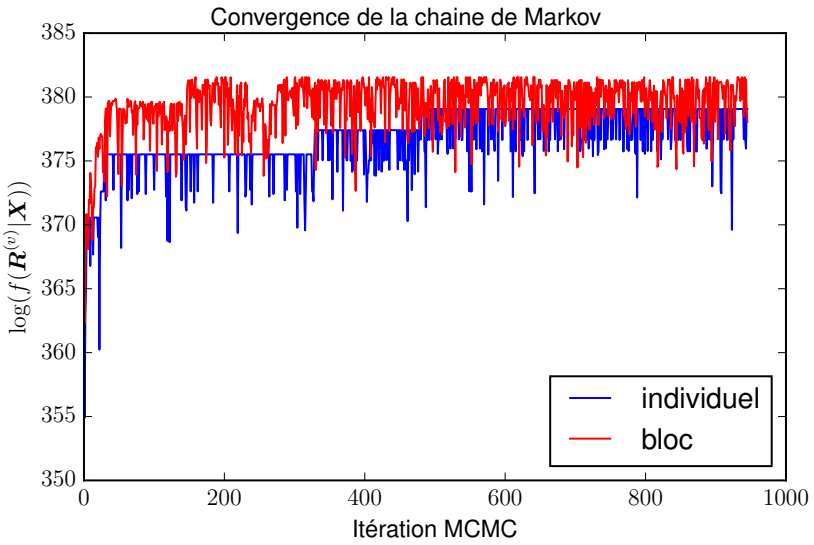


FIGURE 3.6 – Exemple d'évolution de la probabilité posteriori $f(\boldsymbol{R}^{(v)}|\boldsymbol{X})$ (à un facteur de normalisation près) lors de l'application de la méthode du *Bernoulli Detector* avec l'échantillonneur de Gibbs appliqué selon deux approches : sur chaque indicatrice individuellement (bleu) et sur un bloc autour des ruptures (rouge, algorithme 3), pour $V = 1000$ itérations. La série temporelle de cet exemple comporte $n = 200$ points et 5 ruptures.

et enfin le vecteur des indicatrices $\boldsymbol{R}_m^{(v)}$ privé des coefficients R_{m_1},R_m,R_{m_2}

$$\boldsymbol{R}_{\setminus m,\text{bloc}}^{(v)} = (R_2^{(v)},\ldots,R_{m-2}^{(v)},R_{m+2}^{(v-1)},\ldots,R_{n-1}^{(v-1)}). \tag{3.53}$$

La probabilité conditionnelle associée au bloc d'indicatrices voisines est alors

$$\Pr\left((R_{m_1},R_m,R_{m_2})^{(v)} = \rho|\boldsymbol{R}_{\setminus m,\text{bloc}}^{(v)};\boldsymbol{X}\right) = \frac{\Pr(\boldsymbol{R}_{m,\text{bloc},\rho}^{(v)}|\boldsymbol{X})}{\sum_{g\in G}\Pr(\boldsymbol{R}_{m,\text{bloc},g}^{(v)}|\boldsymbol{X})}, \tag{3.54}$$

où les probabilités a posteriori $\Pr(\boldsymbol{R}_{m,\text{bloc},g}^{(v)}|\boldsymbol{X})$ sont obtenues d'après (3.24). On remarque que dans ce cas, les p-valeurs à mettre à jour sont $p_{i^-}(\boldsymbol{X}_{2-},\boldsymbol{X}_-)^{(v)}$, $p_{i-1}(\boldsymbol{X}_-,\boldsymbol{X}_+)^{(v)}$, $p_i(\boldsymbol{X}_-,\boldsymbol{X}_+)^{(v)}$, $p_{i+1}(\boldsymbol{X}_-,\boldsymbol{X}_+)^{(v)}$ et $p_{i^+}(\boldsymbol{X}_+,\boldsymbol{X}_{2+})^{(v)}$, où $\boldsymbol{X}_-$ et $\boldsymbol{X}_+$ sont les échantillons délimités par les ruptures i^- et i^+ autour du bloc $(i-1,i,i+1)$ et par une rupture dans le bloc, et où $\boldsymbol{X}_{2-}$ et $\boldsymbol{X}_{2+}$ sont les échantillons avant et après $\boldsymbol{X}_-$ et $\boldsymbol{X}_+$ respectivement, entre les ruptures i^{2-} et i^- pour $\boldsymbol{X}_{2-}$, et entre i^+ et i^{2+} pour $\boldsymbol{X}_{2+}$, comme le montre le schéma de la figure 3.7. La procédure d'échantillonnage de Gibbs par bloc est décrite dans l'algorithme 3. On l'appelle le modèle du *Bernoulli Detector* univarié, ou UniBD. L'initialisation de $\boldsymbol{R}$ proposée correspond à la solution vide, sans rupture, où $R_i^{(0)} = 0$, pour tout $1 < i < n$.

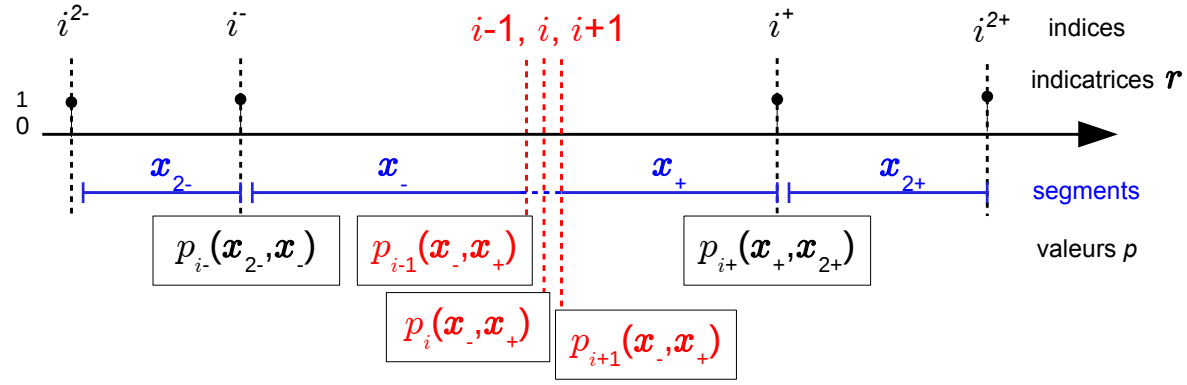


FIGURE 3.7 – Schéma des p-valeurs (encadrées) mises à jour lors de l'échantillonnage du bloc autour du point i dans l'algorithme 3. Les positions testées sont en rouge. La segmentation est représentée en bleu. Les échantillons $\boldsymbol{x}_-$ et $\boldsymbol{x}_+$ sont définis en fonction de la position de la rupture dans le bloc (r_{i-1},r_i,r_{i+1}).

3.3.3 Approximation

Afin de réduire le coût calculatoire de l'algorithme 3, une approximation est proposée : seule la valeur $p_i(\boldsymbol{X}_-,\boldsymbol{X}_+)$ est calculée pour l'échantillonnage de la variable R_i lors du mouvement de Gibbs à l'itération v. On néglige l'effet d'une modification de la segmentation en i sur les p-valeurs du voisinage, $p_{i^-}(\boldsymbol{X}_{2-},\boldsymbol{X}_-)$ et $p_{i^+}(\boldsymbol{X}_+,\boldsymbol{X}_{2+})$ conservent leurs valeurs courantes. Ainsi

$$p_{i^-}(\boldsymbol{X}_{2-},\boldsymbol{X}_-) = \begin{cases} p_{i^-}(\boldsymbol{X}_{2-},\boldsymbol{X}_-)^{(v-1)} & \text{si } \pi(i^-) > \pi(i), \\ p_{i^-}(\boldsymbol{X}_{2-},\boldsymbol{X}_-)^{(v)} & \text{si } \pi(i^-) < \pi(i) \end{cases} \tag{3.55}$$

et

$$p_{i^+-}(\boldsymbol{X}_+,\boldsymbol{X}_{2+}) = \begin{cases} p_{i^+}(\boldsymbol{X}_+,\boldsymbol{X}_{2+})^{(v-1)} & \text{si } \pi(i^+) > \pi(i), \\ p_{i^+}(\boldsymbol{X}_+,\boldsymbol{X}_{2+})^{(v)} & \text{si } \pi(i^+) < \pi(i). \end{cases} \tag{3.56}$$

La seule p-valeur à calculer lors de l'échantillonnage de R_i est encadrée dans la figure 3.8. Les probabilités conditionnelles (3.49) et (3.50) deviennent donc

$$\Pr(R_m = 0|\boldsymbol{R}_{\setminus m}^{(v)};\boldsymbol{X}) = \frac{1}{1+\frac{K'+\frac{1}{2}}{N-K'-\frac{5}{2}}\gamma\left(p_i(\boldsymbol{X}_-,\boldsymbol{X}_+)^{(v)}\right)^{\gamma-1}}, \tag{3.57}$$

$$\Pr(R_m = 1|\boldsymbol{R}_{\setminus m}^{(v)};\boldsymbol{X}) = 1 - \Pr(R_m = 0|\boldsymbol{R}_{\setminus m}^{(v)};\boldsymbol{X}), \tag{3.58}$$

où $K' = \sum_{j=2}^{m-1} R_j^{(v)} + \sum_{j=m+1}^{n-1} R_j^{(v-1)}$ est le nombre de ruptures dans $\boldsymbol{R}_{\setminus m}^{(v)}$. Ce type de modification accélère en pratique considérablement la convergence de l'algorithme d'échantillonnage. Avec cette heuristique, une stratégie d'échantillonnage par bloc n'est plus nécessaire, et l'analyse des résultats expérimentaux montre que les ruptures estimées sont correctement positionnées. La procédure d'échantillonnage est donnée dans l'algorithme 4.

Algorithme 3 : UniBD, échantillonneur de Gibbs par bloc

Données : série temporelle $\boldsymbol{X} \in \mathbb{R}^n$
Résultat : $\hat{\boldsymbol{R}}_{MAP}$
choisir α
initialiser $\boldsymbol{R}^{(0)}$, $\hat{\boldsymbol{R}}_{MAP} = \boldsymbol{R}^{(0)}$
pour $v \leftarrow 1, V$ **faire**
  mélanger aléatoirement les indices par la fonction π
  pour $m \leftarrow 2, n-1$ **faire**
    $i = \pi^{-1}(m)$
    si $R_i^{(v-1)} = 0$ **alors**
      trouver les ruptures voisines de i en i^- et en i^+
      calculer $p_{i^-}(\boldsymbol{X}_{2-},\boldsymbol{X}_-)^{(v)}$, $p_i(\boldsymbol{X}_-,\boldsymbol{X}_+)^{(v)}$ et $p_{i^+}(\boldsymbol{X}_+,\boldsymbol{X}_{2+})^{(v)}$
      simuler $R_m^{(v)}$ d'après (3.49) et (3.50)
    fin
    sinon
      trouver les ruptures avant $i-1$ en i^- et après $i+1$ en i^+
      calculer $p_{i^-}(\boldsymbol{X}_{2-},\boldsymbol{X}_-)^{(v)}$, $p_{i-1}(\boldsymbol{X}_-,\boldsymbol{X}_+)^{(v)}$, $p_i(\boldsymbol{X}_-,\boldsymbol{X}_+)^{(v)}$,
       $p_{i+1}(\boldsymbol{X}_-,\boldsymbol{X}_+)^{(v)}$ et $p_{i^+}(\boldsymbol{X}_+,\boldsymbol{X}_{2+})^{(v)}$
      $m_1 = \pi(i-1)$, $m_2 = \pi(i+1)$
      simuler $(R_{m_1},R_m,R_{m_2})^{(v)}$ d'après (3.54)
    fin
    calculer $\Pr(\boldsymbol{R}_{m,\text{bloc}}^{(v)}|\boldsymbol{X})$ d'après (3.24)
    si $\Pr(\boldsymbol{R}_{m,bloc}^{(v)}|\boldsymbol{X}) > \Pr(\hat{\boldsymbol{R}}_{MAP}|\boldsymbol{X})$ **alors**
      $\hat{\boldsymbol{R}}_{MAP} = \boldsymbol{R}_{m,\text{bloc}}^{(v)}$
    fin
  fin
fin

3.4 Résultats expérimentaux

Après avoir présenté le modèle du *Bernoulli Detector*, permettant l'estimation du vecteur des indicatrices des ruptures $\boldsymbol{R}$ pour la série temporelle $\boldsymbol{X}$, puis déterminé le niveau de contrôle de la détection d'une unique rupture, et enfin proposé un algorithme pour une résolution par une méthode MCMC, la procédure est appliquée à des signaux simulés. Elle est comparée à d'autres approches classiques afin de mettre en valeur les avantages de la méthode, et d'illustrer les cas de figures dans lesquels son emploi présente un intérêt. Dans toute cette partie, le test statistique employé pour le calcul des p-valeurs est le test de WMW.

Les algorithmes 3 et 4 ont été implémentés en `python`. Dans les simulations, le signal $\boldsymbol{x}$ comporte $n = 100$ points en temps et une seule rupture au point $\tau = 50$. Les observations sont i.i.d. sur chacun des deux segments qui composent le signal. Plusieurs niveaux de

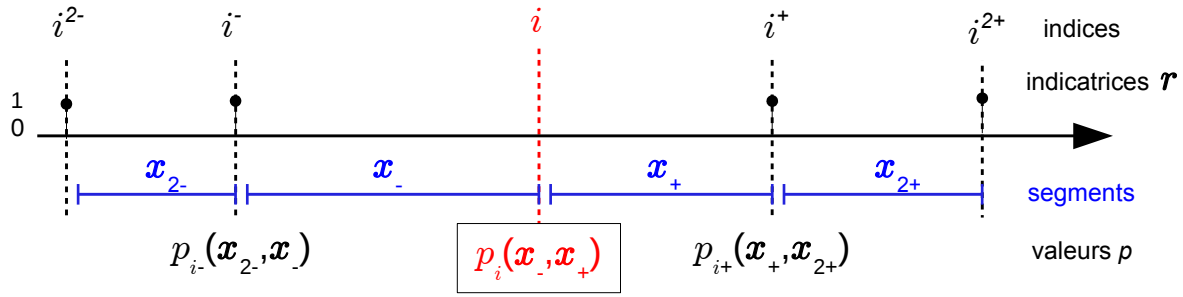


FIGURE 3.8 – Schéma de la p-valeur (encadrée) mise à jour lors de l'échantillonnage au point i (rouge) dans l'algorithme 4. La segmentation est représentée en bleu. Avec l'approximation proposée, les p-valeurs des ruptures voisines conservent leurs valeurs courantes.

bruits sont étudiés, mesurés par le rapport signal sur bruit

$$\text{SNR} = 10 \log \frac{(\mu_1 - \mu_2)^2}{\sigma^2}, \tag{3.59}$$

ici dans le cas de données normales de moyennes μ_1 et μ_2 sur les premier et le deuxième segments respectivement, et de variance σ^2. Pour chaque test, 1000 signaux sont générés. Le niveau d'acceptation choisi est $\alpha = 0{,}01$, et comme l'algorithme converge rapidement, (voir la figure 3.5), le nombre d'itérations MCMC est $V = 1000$. De la sorte on s'assure que l'espace des états a été suffisamment exploré pour avoir une estimation correcte du MAP.

3.4.1 Critères d'évaluation

Les performances de l'algorithme sont évaluées selon des critères usuels mesurant la qualité de la détection des ruptures, mais aussi en fonction des caractéristiques attendue du modèle du *Bernoulli Detector*, comme sa robustesse aux données non normales et sa flexibilité qui permet d'appliquer le modèle sans avoir à l'adapter. Les résultats de la détection sont résumés dans le tableau 3.1, où sont comparés le vrai vecteur des indicatrices et le vecteur estimé $\hat{\boldsymbol{r}}_{MAP}$. À partir de ce comptage, deux termes sont calculés, le rappel et la précision [Lung-Yut-Fong *et al.*, 2011b, Harchaoui *et al.*, 2009] :

$$\text{rappel} = \frac{VP}{VP + FN}, \qquad \text{précision} = \frac{VP}{VP + FP}. \tag{3.60}$$

Le rappel (ou sensibilité ou taux de vrais positifs) est une mesure de la proportion de vraies ruptures détectées, et la précision (ou valeur prédictive positive) mesure la proportion de vraies ruptures parmi celles qui ont été estimées.

On remarque que le comptage réalisé dans le tableau 3.1 ne tient pas compte d'un éventuel décalage dans la position d'une rupture estimée par rapport à la vraie : dans ce cas, un faux négatif et un faux positif sont comptés. Un critère complémentaire est évalué :

Algorithme 4 : UniBD, pseudo-échantillonneur de Gibbs

Données : série temporelle $\boldsymbol{X} \in \mathbb{R}^n$
Résultat : $\hat{\boldsymbol{R}}_{MAP}$
choisir α
initialiser $\boldsymbol{R}^{(0)}$, $\hat{\boldsymbol{R}}_{MAP} = \boldsymbol{R}^{(0)}$
pour $v \leftarrow 1, V$ **faire**
  mélanger aléatoirement les indices par la fonction π
  pour $m \leftarrow 2, n-1$ **faire**
    $i = \pi^{-1}(m)$
    calculer $p_i(\boldsymbol{X}_-,\boldsymbol{X}_+)^{(v)}$
    simuler $R_m^{(v)}$ d'après (3.57) et (3.58)
    calculer $\Pr(\boldsymbol{R}_m^{(v)}|\boldsymbol{X})$ d'après la loi a posteriori
    si $\Pr(\boldsymbol{R}_m^{(v)}|\boldsymbol{X}) > \Pr(\hat{\boldsymbol{R}}_{MAP}|\boldsymbol{X})$ **alors**
      $\hat{\boldsymbol{R}}_{MAP} = \boldsymbol{R}_m^{(v)}$
    fin
  fin
fin

	vraies ruptures	vrais points normaux
ruptures détectées	vrais positifs (VP)	faux positifs (FP)
point normaux estimés	faux négatifs (FN)	vrais négatifs (VN)

TABLE 3.1 – Résultats possibles de la détection : comptage des vrais et faux positifs et négatifs. Les observations $(x_i)_{1<i<n}$ sont soit des ruptures ($r_i = 1$) soit des points normaux ($r_i = 0$).

l'erreur quadratique sur la position de la rupture détectée par rapport à la position réelle i. Lorsque que plusieurs ruptures ont été estimées, seule la plus proche de i est prise en compte pour le calcul du MSE, dans une limite de plus ou moins 10 points autour de i. Les autres ruptures sont comptabilisées comme des faux positifs, et le nombre de faux positifs est ainsi obtenu. Dans les expériences présentées ci-après, le rappel, la précision, l'erreur quadratique sur la position et le nombre de faux positifs sont mesurés pour chacune des estimations, puis les valeurs moyennes sont calculées pour un grand nombre de simulations. L'erreur quadratique moyenne est notée EQM.

3.4.2 Comparaison des stratégies d'échantillonnage

Dans un premier temps, les algorithmes 3 et 4 sont comparés sur des données simulées, afin de valider expérimentalement l'approximation qui est faite dans la mise à jour des p-valeurs. Les séries temporelles sont générées suivant les distributions normales $\mathcal{N}(\mu_1,\sigma)$

sur le segment 1 et $\mathcal{N}(\mu_2,\sigma)$ sur le segment 2. Un exemple de série temporelle ainsi générée est présenté dans la figure 3.9. Plusieurs niveaux de bruit, définis par (3.59), sont simulés en faisant varier l'écart entre μ_1 et μ_2. Les résultats sont représentés dans la figure 3.10, en terme de rappel et de précision. On constate que les courbes obtenues pour les deux algorithmes sont très proches, l'approximation faite dans l'algorithme 4 est acceptable du point de vue de la détection.

Comme le rappel et la précision pénalisent fortement un éventuel décalage entre la position estimée et la vraie position d'une rupture, en comptant un faux négatif et un faux positif, le comptage est adapté afin d'inclure une tolérance dans la localisation des événements détectés. En effet, de notre point de vue, il est plus important de déceler une rupture même légèrement mal localisée, plutôt que de la manquer si elle n'est pas positionnée exactement. Dans l'exemple de la figure 3.9, la rupture en $i = 50$ est détectée malgré le mauvais rapport signal sur bruit, mais avec un décalage de 3 points. Un voisinage est donc défini autour de la vraie position i d'une rupture, de plus ou moins d points, dans lequel la détection la plus proche de i est considérée comme un vrai positif. Si une rupture est détectée dans ce voisinage, l'absence de rupture au point exact i n'est pas comptée comme un faux négatif. En incluant ces tolérances, on produit les figures 3.11 pour $d = 1$ et 3.12 pour $d = 5$. Les courbes de rappel et de précision y sont meilleures que dans la figure 3.10. Le *Bernoulli Detector* atteint même de très bonnes performances avec une tolérance de plus ou moins 5 points autour de la vraie position de la rupture.

D'après ces résultats, on déduit que la méthode, testée via les deux algorithmes 3 et 4, détecte bien la rupture réelle, avec peu de faux positifs. Lorsque le bruit est important, les ruptures sont toujours détectées, mais le modèle a plus de difficultés à les localiser précisément. L'ajout d'une fenêtre de tolérance dans la position permet d'assouplir les critères d'évaluation et ainsi de tenir compte des détections légèrement mal localisées. La comparaison des algorithmes 3 et 4 montre peu de différences dans les courbes de rappel et de précision. Le pseudo-échantillonneur de Gibbs est légèrement moins performant que l'échantillonneur par bloc, mais la stratégie par bloc rend l'algorithme 3 plus lent. Dans la perspective de traitement de séries temporelles multivariées, l'argument du coût calculatoire est décisif, l'algorithme 4 est donc retenu pour les applications ultérieures.

3.4.3 Données générées selon la loi gaussienne

Le *Bernoulli Detector* est maintenant comparé avec deux autres approches citées au chapitre 1 : la méthode BARD [2] [Bardwell et Fearnhead, 2014] et le fused LASSO [Tibshirani, 1996, Tibshirani *et al.*, 2005], tout deux pris dans un contexte univarié. La méthode BARD associe un état à chaque variable X_i : N pour l'état normal ou A pour l'état anormal. Cette approche est destinée à la détection de segments anormaux, c'est-à-dire les portions de signal où la moyenne μ_A est décalée par rapport à une moyenne de référence μ_N connue. Pour simplifier, on prend $\mu_N = 0{,}0$. Le modèle des données (1.73) entre les instants t et s

2. Les codes sont disponibles à `http://www.lancaster.ac.uk/pg/bardwell/Work.html`

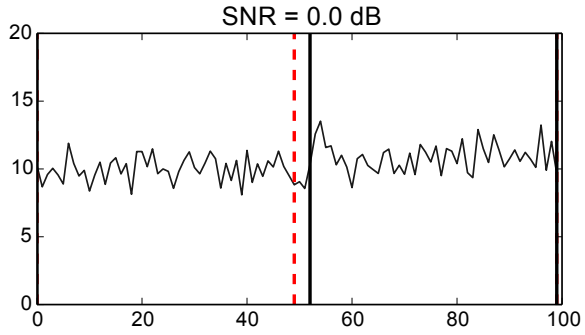
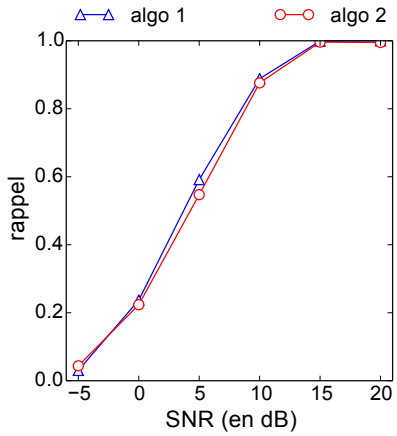
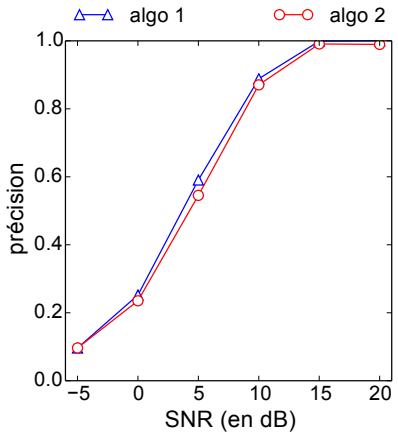


FIGURE 3.9 – Exemple de signal et de détection. $\mu_1 = 10{,}0$, $\mu_2 = 11{,}0$, $\sigma = 1{,}0$, SNR $=0{,}0$ dB. La rupture réelle est en pointillés rouges, et son estimation par $\hat{\boldsymbol{r}}_{MAP}$ est en noir.



(a) Rappel moyen en fonction du niveau de bruit.



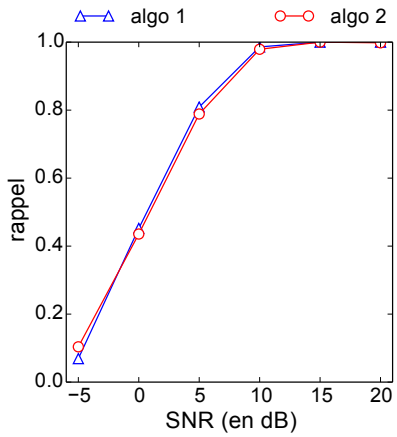
(b) Précision moyenne en fonction du niveau de bruit.

FIGURE 3.10 – Courbes de rappel et de précision en fonction du niveau de bruit, pour des données distribuées normalement. Les séries temporelles comportent $n = 100$ points, avec une rupture en $\tau = 50$. La courbe en bleu (algo 1) est obtenue avec l'algorithme 3, et la courbe en rouge (algo 2) avec l'algorithme 4.

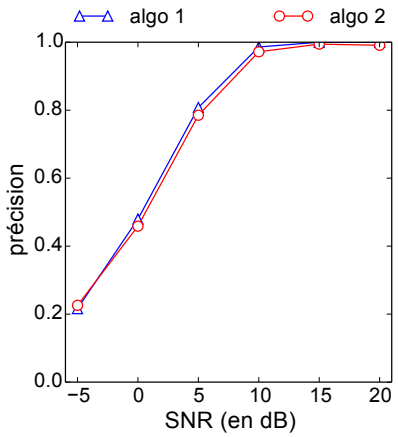
devient, pour une série temporelle univariée,

$$f(\boldsymbol{x}_{t:s}) = \begin{cases} \prod_{i=t}^{s} P_N(X_{j,i}) & \text{si le segment est normal,} \\ \int \prod_{i=t}^{s} P_A(X_i|\mu) f(\mu) d\mu & \text{si le segment est anormal,} \end{cases} \tag{3.61}$$

où $P_N(.)$ est la distribution des données dans un segment normal, $P_A(.|\mu)$ est la distribution de paramètre d'intérêt μ des données dans un segment anormal et $f(\mu)$ est la densité de probabilité a priori de la moyenne μ, dans un segment anormal. Pour les données normales, les lois a priori choisies sont $P_N = \mathcal{N}(0{,}\sigma)$ et $P_A = \mathcal{N}(\mu_A{,}\sigma)$. Le rapport des vraisemblances

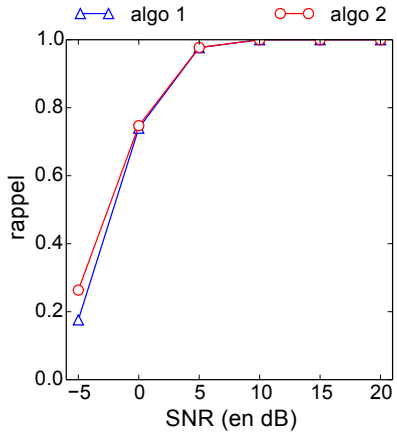


(a) Rappel moyen en fonction du niveau de bruit.

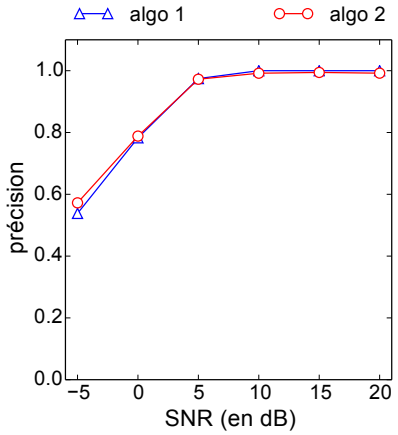


(b) Précision moyenne en fonction du niveau de bruit.

FIGURE 3.11 – Courbes de rappel et de précision en fonction du niveau de bruit, avec une tolérance sur la position des ruptures détectées, pour des données distribuées normalement. Les séries temporelles comportent $n = 100$ points, avec une rupture en $\tau = 50$, qui est détectée à $\tau \pm 1$. La courbe en bleu (algo 1) est obtenue avec l'algorithme 3, et la courbe en rouge (algo 2) avec l'algorithme 4.



(a) Rappel moyen en fonction du niveau de bruit.



(b) Précision moyenne en fonction du niveau de bruit.

FIGURE 3.12 – Courbes de rappel et de précision en fonction du niveau de bruit, avec une tolérance sur la position des ruptures détectées, pour des données distribuées normalement. Les séries temporelles comportent $n = 100$ points, avec une rupture en $\tau = 50$, qui est détectée à $\tau \pm 5$. La courbe en bleu (algo 1) est obtenue avec l'algorithme 3, et la courbe en rouge (algo 2) avec l'algorithme 4.

entre les instants t et s est alors

$$\frac{P_A(t,s|\mu)}{P_N(t,s)} = \exp\Big(\frac{\mu_A}{\sigma^2}\sum_{i=s}^{t}(X_i - \frac{\mu_A}{2})\Big). \tag{3.62}$$

Pour que ce modèle soit adapté aux données, ces dernières sont décalées de la valeur μ_1, de telle sorte que le premier segment soit considéré comme un segment normal de moyenne nulle, et σ prend la vraie valeur 1,0. En suivant les propositions de [Bardwell et Fearnhead, 2014] et d'après les vraies distributions des données, la distribution choisie pour μ dans un segment anormal est la loi uniforme sur $[\mu_2-\mu_1-\sigma,\mu_2-\mu_1+\sigma]$, et les longueurs des segments normaux S_N et anormaux S_A suivent la distribution négative binomiale de paramètres 5,0 et 0,09. Enfin, la probabilité qu'un segment anormal soit suivi d'un segment normal est fixée à $\pi_N = 0{,}9$. La détection de segments anormaux est faite récursivement, grâce à la relation (1.75). L'algorithme fourni par les auteurs de la méthode renvoie la probabilité $p(i)$ que l'observation x_i appartienne à un segment anormal. Le vecteur de ces probabilités est seuillé : les observations x_i pour lesquelles $p(i) \geq 0{,}90$ sont considérées comme anormales. Les segments étant ainsi délimités, les positions des ruptures sont déduites.

Le fused LASSO est choisi comme méthode de référence classique et non bayésienne. Le terme d'attache aux données fait intervenir la norme ℓ_2, qui rend la méthode optimale dans le cas gaussien, mais ses performances peuvent se dégrader fortement avec d'autres distributions, par exemple les distributions à queues lourdes. La série temporelle est approchée par la fonction constante par morceaux de coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$, solution du problème (1.41). Les ruptures sont détectées aux positions i, $1 < i < n$, telles que $|\beta_i - \beta_{i+1}| > 10^{-10}$. Le paquet `R genlasso` [Tibshirani et Arnold, 2014] est employé. Pour ajuster le paramètre de régularisation λ du problème (1.41), 21 valeurs [3] ont été testées dans l'intervalle [16,0; 36,0]. La solution avec $\lambda = 22.3$, donnant la courbe de rappel la plus proche du *Bernoulli Detector*, a été retenue rétrospectivement.

Comme précédemment, les données sont générées selon les lois normales $\mathcal{N}(\mu_1,\sigma)$ et $\mathcal{N}(\mu_2,\sigma)$ du test précédent, et le rapport signal sur bruit varie de $-5{,}0$ dB à 20,0 dB. Les courbes de rappel et de précision obtenues sont données à la figure 3.13. Le paramètre λ du fused LASSO a été fixé de manière à ce que la courbe de rappel corresponde à celle du *Bernoulli Detector*, on remarque qu'alors les courbes de précisions sont aussi semblables. Ce paramètre contrôle le nombre de segments dans l'estimation $\hat{\boldsymbol{\beta}}$, son choix résulte donc d'un compromis entre un bon rappel et une bonne précision. Il est en général nécessaire de sélectionner rétrospectivement la meilleure solution, par exemple selon le critère proposé dans [Bleakley et Vert, 2011]. La méthode bayésienne BARD donne des résultats moins bons que les deux autres approches. La détection des ruptures à $-5{,}0$ dB a même échoué.

Pour quantifier la distance entre la rupture estimée et la vraie rupture, ainsi que le nombre de faux positifs, l'erreur quadratique moyenne et le nombre moyen de faux positifs sont tracés dans la figure 3.14. L'EQM diminue rapidement pour les trois méthodes quand le bruit diminue, elle est proche de 0,0 à 10,0 dB. Le fused LASSO est la plus précise des méthodes, suivie du *Bernoulli Detector*. Le nombre de faux positifs est assez élevé avec le fused LASSO, comme avec la méthode BARD, alors qu'il est inférieur à 0,18 avec le *Bernoulli Detector*. Notre algorithme est donc moins bon que le fused LASSO pour localiser la rupture, mais présente très peu de faux positifs même en présence d'un bruit

3. Les valeurs testées pour le paramètre λ sont 16,0, 18,0, 20,0, 21,5, 21,6, 21,7, 21,8, 21,9, 22,0, 22,1, 22,2, 22,3, 22,4, 22,5, 24,0, 26,0, 28,0, 30,0, 32,0, 34,0 et 36,0.

méthode	temps calcul (s)
UniBD	32,6
BARD	0,4
fused LASSO	0,4

TABLE 3.2 – Temps de calcul moyen pour 100 signaux de $n = 100$ points comportant une rupture et SNR =10,0 dB.

important.

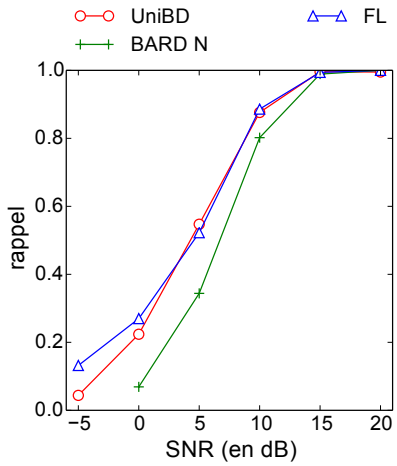
Les temps de calcul moyens pour 100 tests sont donnés dans le tableau 3.2. Le fused LASSO s'avère le plus rapide et notre algorithme UniBD est de loin le plus lent. L'implémentation de la méthode n'a toutefois pas été optimisée, et comme nous ne disposons pas de mesure théorique de la vitesse de convergence, le nombre d'itérations V a été fixé à 1000 empiriquement. En effet, la stratégie MCMC adoptée nécessite un nombre d'itérations suffisamment grand pour que l'estimateur du MAP soit assez proche de la solution. Le temps de calcul est toutefois raisonnable pour traiter des séries temporelles de $n = 100$ points.

Cette série de tests sur des données normales permet de constater que le modèle du *Bernoulli Detector* parvient à des résultats comparables à ceux du fused LASSO pour $\lambda =$ 22,3 et meilleurs que ceux de l'algorithme BARD avec un a priori normal. Notre approche a de plus l'avantage d'être applicable directement sur les séries temporelles puisqu'elle est non paramétrique, contrairement à BARD. Seul le niveau α doit être choisi, mais comme α intervient dans le contrôle de la détection d'une rupture unique (voir la partie 3.2), la proposition 3.2.1 peut guider ce choix. En revanche l'algorithme UniBD est le plus lent, mais présente l'avantage de s'appliquer quelles que soient les valeurs des moyennes des segments, tandis que la méthode BARD se réfère à une moyenne de référence μ_N fixée à 0,0, et les valeurs des hyperparamètres ont été fixées à partir de connaissances a priori sur les données. Le fused LASSO repose sur l'hypothèse de normalité des données, la méthode BARD a été adaptée à des lois a priori gaussiennes, et le modèle du *Bernoulli Detector* est non paramétrique. Il est donc intéressant de comparer ces trois approches sur des données non normales.

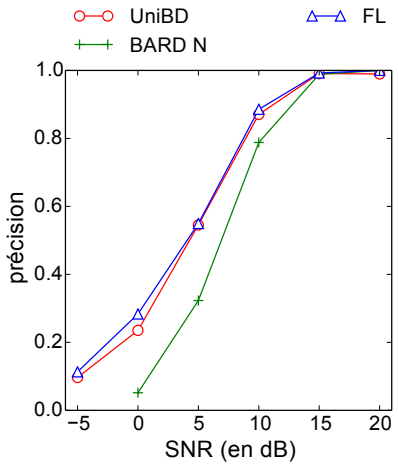
3.4.4 Données générées selon une loi à queue lourde

Pour cette nouvelle série de simulations, les distributions s'éloignent du cas gaussien : la probabilité d'observer des valeurs éloignées de la moyenne est plus grande, d'où la présence de valeurs aberrantes dans les séries temporelles. Les tests réalisés pour cette section permettent d'évaluer la robustesse des méthodes à ces valeurs extrêmes. La distribution choisie est la distribution de Student à $\nu = 3{,}0$ degrés de liberté. Elle est centrée sur μ_k au k^{e} segment, avec $\mu_1 = 10{,}0$. Pour calculer le rapport signal sur bruit (3.59), la variance est donnée par

$$\sigma^2 = \frac{\nu}{\nu - 2}. \tag{3.63}$$

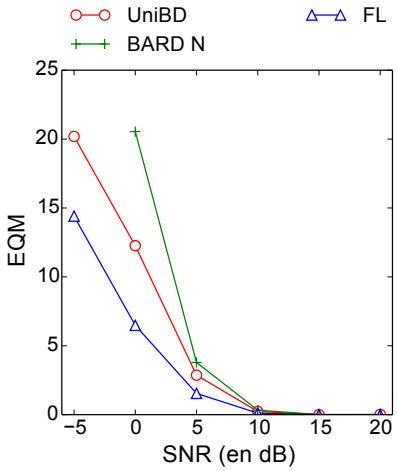


(a) Rappel moyen en fonction du niveau de bruit.

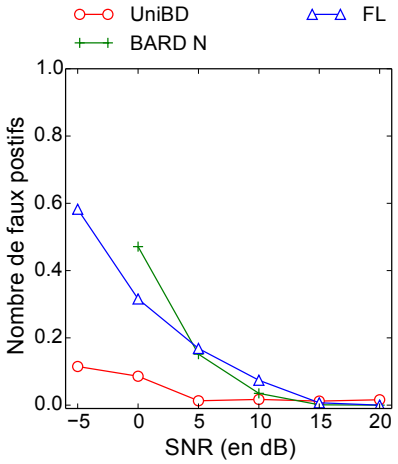


(b) Précision moyenne en fonction du niveau de bruit.

FIGURE 3.13 – Courbes de rappel et de précision en fonction du niveau de bruit, pour des données distribuées normalement. Les séries temporelles comportent $n = 100$ points, avec une rupture en $\tau = 50$. La courbe en rouge (UniBD) correspond à la méthode du *Bernoulli Detector* avec $\alpha = 0{,}01$, la courbe en vert (BARD N) à la méthode BARD et avec des lois a priori normales, et la courbe en bleu (FL) au fused LASSO avec $\lambda = 22{,}3$.



(a) EQM en fonction du niveau de bruit.



(b) Nombre moyen de faux positifs en fonction du niveau de bruit.

FIGURE 3.14 – Courbes de l'erreur quadratique moyenne et du nombre moyen de faux positifs en fonction du niveau de bruit, pour des données distribuées normalement. Les séries temporelles comportent $n = 100$ points, avec une rupture en $\tau = 50$. La courbe en rouge (UniBD) correspond à la méthode du *Bernoulli Detector* avec $\alpha = 0{,}01$, la courbe en vert (BARD N) à la méthode BARD et avec des lois a priori normales, et la courbe en bleu (FL) au fused LASSO avec $\lambda = 22{,}3$.

Comme l'hypothèse de normalité n'est plus vérifiée, les méthodes BARD et LASSO doivent être adaptées. La fonction de vraisemblance de la méthode BARD est construite à partir des distributions de Student de paramètres (ν,0) et (ν,μ_A) pour les segments normaux et anormaux respectivement, l'expression (3.62) devient alors :

$$\frac{P_A(t,s|\mu)}{P_N(t,s)} = \prod_{i=s}^{t} \left(\frac{1 + \frac{(X_i - \mu_A)^2}{\nu}}{1 + \frac{X_i^2}{\nu}} \right)^{-\frac{\nu+1}{2}}. \tag{3.64}$$

Le support de la moyenne μ_A est défini comme dans le cas normal, avec cette fois σ donné par la distribution réelle des données, selon (3.63). La série temporelle est de nouveau recentrée autour de la valeur nulle, par soustraction de μ_1. Le fused LASSO est lui aussi modifié, afin de le rendre insensible aux données aberrantes. La fonction $\boldsymbol{\beta}$ est alors la solution du problème (1.43), présenté dans [Aravkin *et al.*, 2013] [4]. Ce problème de LASSO robuste introduit une norme ℓ_1 dans le terme d'attache aux données. Il est résolu par la méthode ADMM [Boyd *et al.*, 2011], qui introduit un paramètre supplémentaire η. Comme précédemment, les valeurs de ces paramètres sont déterminées rétrospectivement, de manière à atteindre un compromis entre un bon rappel (détection de la vraie rupture) et une bonne précision (peu de fausses ruptures).

Les méthodes BARD avec l'a priori normal et fused LASSO du problème (1.41) avec la norme ℓ_2 sont également testées pour comparaison, avec les mêmes paramètres que pour les tests précédents (partie 3.4.3). La figure 3.15 donne les courbes de rappel et de précision pour toutes ces méthodes. Globalement, les courbes de rappel se ressemblent, Les méthodes testées parviennent donc toutes à détecter la rupture à la position $\tau = 50$ avec les performances comparables. En revanche les courbes de précision sont disparates. À partir de 0,0 dB, le *Bernoulli Detector* fournit les meilleurs résultats. L'algorithme BARD avec l'a priori adapté aux données (courbe BARD T) parvient également à une bonne précision quand le rapport signal sur bruit augmente, et se confond avec le *Bernoulli Detector*. Cependant il se dégrade fortement quand le bruit augmente, et n'est pas applicable sous 0,0 dB. Il est toutefois sensiblement meilleur qu'avec un a priori normal (courbe BARD N), perturbé par les valeurs aberrantes qui entraînent la détections de segments anormaux supplémentaires. Le LASSO robuste (courbe L1L1) est légèrement meilleur que les autres méthodes sous 0,0 dB, mais ne parvient pas à une aussi bonne précision que le *Bernoulli Detector* pour un SNR plus grand. Le choix des paramètres λ et η n'est pas évident, la méthode ayant tendance à introduire plusieurs ruptures successives au niveau d'un important saut de moyennes entre μ_1 et μ_2 quand la parcimonie de la solution est renforcée, ce qui créé des faux positifs autour de la vraie détection. Enfin, le fused LASSO (courbe FL) est très perturbé par les valeurs aberrantes, ce qui explique que sa courbe de précision ne dépasse pas 0,4.

L'erreur quadratique moyenne et le nombre de faux positifs sont tracés à la figure 3.16. D'après les courbes de l'erreur quadratique moyenne, les méthodes BARD produisent les moins bons résultats. Les méthodes basées sur un approche de type LASSO donnent les

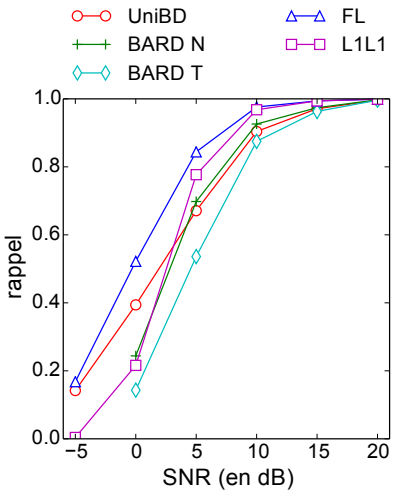
4. Les codes sont disponibles à `https://github.com/saravkin/`

distances les plus faibles entre la rupture réelle et la rupture estimée, le LASSO robuste n'ayant qu'une erreur inférieure à 6 avec un bruit de $-5{,}0\,\mathrm{dB}$. En terme de faux positifs, le LASSO est moins bon, comme nous l'avons constaté avec la courbe de précision, avec un nombre de faux positif proche de 2 dès $5{,}0\,\mathrm{dB}$. L'introduction de la norme ℓ_1 permet de réduire le nombre de fausses détections, puisque le modèle n'est plus sensible aux valeurs extrêmes. La méthode BARD avec un a priori normal n'est pas du tout adaptée à ces données, avec une forte erreur quadratique moyenne et un grand nombre de fausses détections. Les algorithmes BARD T et UniBD détectent très peu de faux positifs, ces modèles sont bien robustes aux valeurs extrêmes de la distribution.

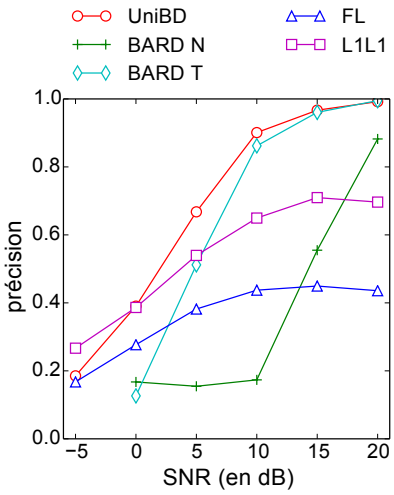
Ce test illustre l'efficacité du modèle du *Bernoulli Detector* sur des données normales comme sur des données non gaussiennes, où la moyenne est perturbée par des valeurs extrêmes. D'autre part notre approche présente l'avantage de s'adapter à des observations de nature différente sans requérir de modifications du modèle ni de prétraitement des données comme un centrage des valeurs autour de 0, son caractère généraliste est mis en valeur. D'autres méthodes peuvent être choisies pour traiter les données non normales à queues lourdes, en particulier la méthode BARD avec un a priori adéquat, mais elle nécessite des connaissances sur les données suffisantes pour construire une fonction de vraisemblance adaptée. Pour obtenir les résultats de la figure 3.15a, les vraies distributions des données ont été inclues dans le modèle, ainsi que les vrais paramètres : le degré de liberté, les moyennes μ_1 et μ_2, et la déviation standard pour fixer le support de la moyenne μ_A des segments anormaux. L'adaptation du fused LASSO, sensible aux valeurs aberrantes, en LASSO robuste introduit un paramètre supplémentaire η, qui rend la sélection de modèle plus difficile. De plus, parmi les jeux de paramètres (λ,η) testés, aucun n'a permis d'atteindre une aussi bonne précision que celle du *Bernoulli Detector*.

3.5 Discussion

Le modèle du *Bernoulli Detector*, présenté dans ce chapitre, permet de réaliser la détection de ruptures multiples dans une série temporelle en estimant pour chaque point la nature de l'observation par l'intermédiaire du paramètre $\boldsymbol{R}$. Il repose sur l'utilisation de p-valeurs, issues du test statistique non paramétrique de WMW, dont les avantages ont été montrés dans le chapitre 2. Cette approche originale équivaut à remplacer les observations X_i dans le modèle par une fonction renvoyant la p-valeur du test appliqué entre les deux segments entourant X_i, et notée $P_i(\boldsymbol{X}_-,\boldsymbol{X}_+)$. La construction de la fonction de vraisemblance se fait à partir des fonctions de vraisemblance marginales des p-valeurs, qui présentent la propriété intéressante d'être distribuées uniformément sous H_0, quelles que soient les distributions des données. Le test de WMW produit une statistique discrète, la modélisation par la loi uniforme est justifiée dans le cas asymptotique et le test est plus conservatif sur de petits échantillons. La distribution des p-valeurs sous H_1 n'étant pas connue dans le cas général, la loi bêta a été choisie en s'inspirant de [Sellke *et al.*, 2001]. Ainsi les p-valeurs suivent une loi décroissante et deviennent très petites quand la présomption contre l'hypothèse nulle est forte. Les fonctions de vraisemblance marginales



(a) Rappel moyen en fonction du niveau de bruit.



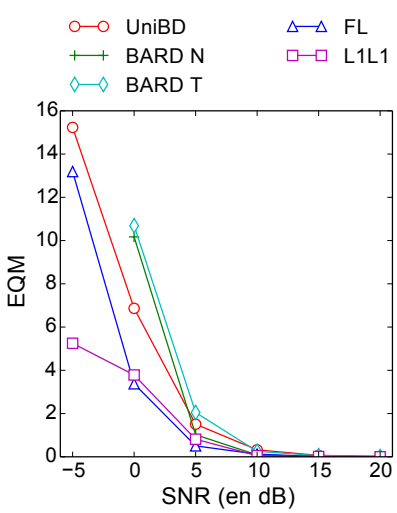
(b) Précision moyenne en fonction du niveau de bruit.

FIGURE 3.15 – Courbes de rappel et de précision en fonction du niveau de bruit, pour des données distribuées selon une loi de Student de degré de liberté $\nu = 3{,}0$, qui génère des valeurs extrêmes. Les séries temporelles comportent $n = 100$ points, avec une rupture en $\tau = 50$. La courbe en rouge (UniBD) correspond à la méthode du *Bernoulli Detector* avec $\alpha = 0{,}01$, la courbe en vert (BARD N) à la méthode BARD et avec des lois a priori normales, la courbe en cyan (BARD T) à la méthode BARD et avec des lois a priori de Student de paramètre $\nu = 3{,}0$, la courbe en bleu (FL) au fused LASSO avec $\lambda = 22{,}3$, et la courbe en magenta (L1L1) au LASSO robuste avec $\lambda = 18{,}0$ et $\eta = 4{,}0$.

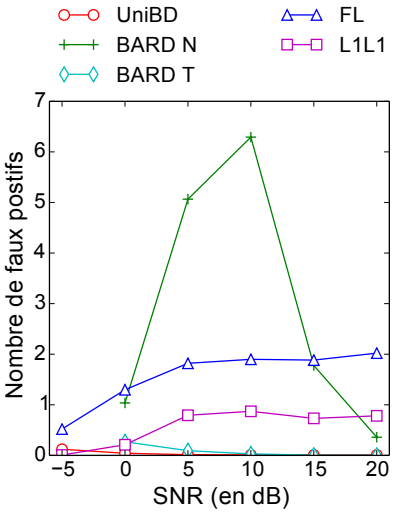
sont donc construite à partir de ce modèle bêta-uniforme.

La fonction de vraisemblance qui est obtenue par composition de ces fonctions de vraisemblance marginales ne possède pas les caractéristiques d'une fonction de vraisemblance propre. En raison des relations de dépendance entre les p-valeurs, qui ne sont pas formulables, la fonction de vraisemblance marginale composite ne peut pas être calibrée comme il est proposé dans [Pauli *et al.*, 2011, Ribatet *et al.*, 2012]. La méthode du *Bernoulli Detector* permet toutefois de calibrer le modèle à partir d'un contrôle des erreurs de type I. En effet, la modélisation par la loi bêta introduit un paramètre γ, qui est lié au paramètre α intervenant dans le contrôle de la détection, formulé par le lemme 3.26 ainsi que la proposition 3.2.1.

Ces résultats sont démontrés quelle que soit la taille n de la série temporelle, et donc valables même sur de petits échantillons, contrairement par exemple aux résultats asymptotiques de [Vert et Bleakley, 2010]. Dans le cas de ruptures multiples, la dépendance entre les p-valeurs sous H_1 apparaît, or leur expression est inconnue. La complexité du cas à multiples ruptures est relevée plusieurs fois dans la littérature [Vert et Bleakley, 2010, Harchaoui et Lévy-Leduc, 2010]. Une façon de s'affranchir des dépendances des p-valeurs serait de se placer dans le cas de ruptures suffisamment éloignées pour n'avoir qu'une faible in-



(a) EQM en fonction du niveau de bruit.



(b) Nombre moyen de faux positifs en fonction du niveau de bruit.

FIGURE 3.16 – Courbes de l'erreur quadratique moyenne et du nombre moyen de faux positifs en fonction du niveau de bruit, pour des données distribuées selon une loi de Student de degré de liberté $\nu = 3{,}0$, qui génère des valeurs extrêmes. Les séries temporelles comportent $n = 100$ points, avec une rupture en $\tau = 50$. La courbe en rouge (UniBD) correspond à la méthode du *Bernoulli Detector* avec $\alpha = 0{,}01$, la courbe en vert (BARD N) à la méthode BARD et avec des lois a priori normales, la courbe en cyan (BARD T) à la méthode BARD et avec des lois a priori de Student de paramètre $\nu = 3{,}0$, la courbe en bleu (FL) au fused LASSO avec $\lambda = 22{,}3$, et la courbe en magenta (L1L1) au LASSO robuste avec $\lambda = 18{,}0$ et $\eta = 4{,}0$.

fluence les unes sur les autres, en s'inspirant de la stratégie de [Niu et Zhang, 2012]. Le contrôle de la détection a donc été seulement montré dans le cas d'une unique rupture : la condition à remplir sur la p-valeur au point i pour détecter la rupture X_i quand aucune rupture n'a encore été détectée (lemme 3.26), et le niveau qui contrôle la probabilité de détecter à tort une rupture à une position quelconque, si aucune rupture n'existe ni n'a été détectée.

Mon approche s'inscrit dans un cadre bayésien, grâce auquel les lois a priori des paramètres $\boldsymbol{R}$ et q sont introduites pour parvenir à l'expression de la probabilité a posteriori (3.24). L'algorithme UniBD a été développé pour fournir une estimation du maximum de la densité de probabilité a posteriori, par une approche MCMC incluant un échantillonneur de Gibbs. Une approximation de cet échantillonneur a été validée empiriquement, son application est plus rapide qu'avec l'échantillonneur exact. Cette résolution par méthode MCMC implique toutefois un temps de calcul incompressible. De ce point de vue, les approches parallélisables, comme l'usage de la bissection ([Matteson et James, 2014]), ou les solutions dérivées par programmation dynamique ([Hawkins, 2001, Fearnhead, 2006, Lung-Yut-Fong *et al.*, 2011b]) sont plus compétitives.

La série de simulations qui a été réalisée, d'abord avec des distributions normales puis

avec des distributions à queues lourdes sur les données, met en valeur d'une part les bonnes performances de l'algorithme UniBD pour la détection de la rupture, avec peu de faux positifs par rapport aux méthodes du fused LASSO et BARD, et d'autre part l'intérêt du *Bernoulli Detector* pour traiter des données sans information a priori. En effet, grâce au choix du test non paramétrique de WMW, l'algorithme UniBD est robuste aux valeurs aberrantes, puisqu'il revient à tester les médianes (dans cet exemple), quand la plupart des tests construits sur les moyennes sont perturbés par les valeurs extrêmes. La méthode du *Bernoulli Detector* ne nécessite donc pas de pré-traitement, comme un filtrage des valeurs extrêmes, ni de modifications du modèle en fonction des caractéristiques de la série temporelle, comme celles qui ont été nécessaires pour appliquer la méthode BARD. L'algorithme UniBD est également une solution simple a mettre en œuvre, puisque seul le niveau α doit être fixé, tandis que le fused LASSO et sa variante robuste au bruit de Student demandent un réglage des paramètres λ et η qui est souvent établi rétrospectivement, sans toujours parvenir au meilleur compromis.

Ce chapitre a introduit le modèle du *Bernoulli Detector*, dont l'originalité tient dans l'intégration d'un test non paramétrique et dans l'introduction des p-valeurs, dont le modèle bêta-uniforme permet de contrôler l'erreur de type I. Malgré son coût de calcul et les approximations faites, ses avantages en terme de robustesse aux données non gaussiennes, notamment aux valeurs aberrantes, en font une solution relativement intéressante pour la détection de ruptures. Le cadre bayésien permet l'introduction d'un nouveau paramètre, à l'occasion du développement de la méthode au traitement de séries temporelles multivariées, qui est présenté au chapitre 4. Ce paramètre relie les composantes spatiales de la série temporelle, et incarne d'éventuelles relations de dépendances.

Chapitre 4

Extension du *Bernoulli Detector* au cas multivarié

Le modèle du *Bernoulli Detector* a été présenté au chapitre 3, nous avons vu comment l'usage de la statistique de WMW permet de construire une fonction de vraisemblance non paramétrique. La méthode résultante ne dépend pas d'hypothèses fortes sur les observations, et s'avère générale. Dans ce chapitre, je propose une extension du modèle au cas multivarié : la série temporelle à analyser est composée de m signaux de n points (aussi appelés composantes spatiales), où les observations $X_{1,i},\ldots,X_{m,i}$ sur tous les signaux sont toutes faites au même instant i. La série temporelle multivariée est notée $\boldsymbol{X}$, et Θ est la matrice des paramètres $\theta_{j,i}$ associés aux $X_{j,i}$. La notation $\boldsymbol{M}_i$ est employée pour le i^{e} vecteur colonne de la matrice $\boldsymbol{M}$, et $\boldsymbol{M}_{j,\bullet}$ représente la j^{e} ligne de $\boldsymbol{M}$. Le problème consiste maintenant à tester les hypothèses

$$
\begin{aligned}
H_0 :\quad & \theta_{1,1} = \cdots = \theta_{1,n},\\
& \cdots\\
& \theta_{m,1} = \cdots = \theta_{m,n},\\
H_1 :\quad & \theta_{1,1} = \cdots = \theta_{1,\tau_1^1} \neq \theta_{1,\tau_1^1+1} = \cdots = \theta_{1,\tau_{K_1}^1} \neq \theta_{1,\tau_{K_1}^1+1} = \cdots = \theta_{1,n},\\
& \cdots\\
& \theta_{m,1} = \cdots = \theta_{m,\tau_1^m} \neq \theta_{m,\tau_1^m+1} = \cdots = \theta_{m,\tau_{Km1}^m} \neq \theta_{m,\tau_{Km}^m+1} = \cdots = \theta_{m,n}.
\end{aligned}
\tag{4.1}
$$

où les positions des ruptures τ_k^j, $1 \leq k \leq K_j$, $1 \leq j \leq m$, sont inconnues ainsi que leurs nombres $K_1,\ldots,K_m$ dans les signaux 1 à m respectivement.

Nous nous plaçons dans le cas évoqué dans la partie 1.1.4 du chapitre d'état de l'art : les ruptures n'affectent pas tous les signaux, mais un même événement peut s'observer sur un ensemble de signaux avec une certaine probabilité. Comme précédemment, notre objectif est la détection de ces changements abrupts, mais nous cherchons à réaliser une segmentation de la série temporelle qui tienne compte des événements communs à un groupe de signaux comme des événements propres à chaque signal. Le modèle résultant est alors plus souple qu'un modèle où toutes les ruptures sont présentes sur tous les signaux,

et plus robuste que l'application indépendante sur tous les signaux d'un modèle univarié. Le paramètre à estimer est alors la matrice (1.76) de dimensions (m,n), où, par convention, les coefficients des premiers et derniers vecteurs colonnes valent 1 : $\boldsymbol{R}_1 = \boldsymbol{R}_n = (1,\ldots,1)'$.

Le paramètre $\boldsymbol{P}$, qui représente les probabilités d'observer certaines ruptures simultanément ou non sur des groupes de signaux, est introduit dans le modèle multivarié. Ces ruptures simultanées sont associées aux colonnes $\boldsymbol{R}_i$ de la matrice des indicatrices. Le modèle (3.3) devient

$$f(\boldsymbol{R}|\boldsymbol{X}) \propto \int_{\boldsymbol{P}} L_*(\boldsymbol{X}|\boldsymbol{R}) f(\boldsymbol{R}|\boldsymbol{P}) f(\boldsymbol{P}) d\boldsymbol{P}, \tag{4.2}$$

faisant apparaître une structure hiérarchique.

Dans la partie 4.1, le modèle de vraisemblance marginale composite et les lois choisies a priori pour les paramètres $\boldsymbol{R}$ et $\boldsymbol{P}$ sont détaillés, pour parvenir à expression de la densité de probabilité a posteriori selon la relation (4.2). Je propose une interprétation du paramètre $\boldsymbol{P}$ dans la partie 4.1.2. Des résultats sur le contrôle de la détection sont fournis dans la partie 4.2. Une implémentation du modèle du *Bernoulli Detector* multivarié est détaillée dans la partie 4.3. Dans la partie 4.4, l'algorithme est appliqué d'abord sur des simulations, qui soulignent l'intérêt du paramètre $\boldsymbol{P}$, puis sur deux jeux de données réelles, l'un consistant en mesures sur un réseau électrique, et l'autre en données génomiques comparatives. Enfin, ces résultats sont discutés dans la partie 4.5.

4.1 Description du modèle

4.1.1 Fonction de vraisemblance

On suppose que les observations $X_{j,i}$ sont indépendantes en temps et d'un signal à l'autre. La fonction de vraisemblance s'écrit donc comme le produit des vraisemblances sur chacun des signaux :

$$L_(\boldsymbol{X}|\boldsymbol{R}) = \prod_{j=1}^{m} L(\boldsymbol{X}_{j,\bullet}|\boldsymbol{R}_{j,\bullet}). \tag{4.3}$$

$\boldsymbol{X}_{j,\bullet}$ et $\boldsymbol{R}_{j,\bullet}$ correspondent respectivement à une série temporelle univariée et à son vecteur d'indicatrices de ruptures, nous reprenons donc le modèle univarié (3.14). La p-valeur associée à l'observation $X_{j,i}$ est calculée à partir des segments $\boldsymbol{X}_{j,-} = (X_{j,i^-+1},\ldots,X_{j,i})$ et $\boldsymbol{X}_{j,+} = (X_{j,i+1},\ldots,X_{j,i^+})$, où i^- et i^+ sont les positions des ruptures situées avant et après i dans le signal j. Elle est définie par :

$$P_{j,i}(\boldsymbol{X},\boldsymbol{R}) = \Pr\Big(|T(\boldsymbol{X}_{j,-},\boldsymbol{X}_{j,+})| \geq |t| \Big| H_0(j,i)\Big), \tag{4.4}$$

où $H_0(j,i)$ est l'hypothèse nulle locale selon laquelle $X_{j,i}$ n'est pas une rupture dans $\boldsymbol{X}_{j,\bullet}$, où T est la statistique du test choisi, et où t est la valeur obtenue pour cette statistique sur une réalisation de la série temporelle $\boldsymbol{x}$ avec la segmentation $\boldsymbol{r}$. Avec le même modèle bêta-uniforme que dans le cas univarié (3.12), nous retrouvons l'expression d'une vraisemblance

marginale composite

$$\boxed{L_*(\boldsymbol{X}|\boldsymbol{R}) = \prod_{j=1}^{m}\prod_{i=1}^{n}\left(\gamma P_{j,i}(\boldsymbol{X},\boldsymbol{R})^{\gamma-1}\right)^{R_{j,i}}.} \tag{4.5}$$

Comme dans le cas univarié, nous appliquons le test de WMW. L'indépendance des vraisemblances marginales $L(\boldsymbol{X}_{j,\bullet}|\boldsymbol{R}_{j,\bullet})$ entre elles permet d'analyser des signaux disparates. Leur dynamique et l'amplitude des ruptures peuvent ainsi être différentes, sans qu'une composante dominante ne pénalise la détection des ruptures sur les autres.

4.1.2 Densités a priori et modèle hiérarchique

Dans cette partie, le choix de la loi a priori du paramètre $\boldsymbol{R}$ est expliqué. L'hyperparamètre $\boldsymbol{P}$ est présenté, aboutissant à un modèle hiérarchique de la forme

$$f(\boldsymbol{R}) \propto f(\boldsymbol{R}|\boldsymbol{P})f(\boldsymbol{P}). \tag{4.6}$$

Le cœur du modèle du *Bernoulli Detector* multivarié réside dans l'introduction de ce terme qui permet d'intégrer des relations entre les événements des différents signaux. Il a été initialement proposé dans [Dobigeon *et al.*, 2007a], pour la segmentation conjointe de processus autorégressifs, où certaines ruptures sont corrélées. Il permet de renforcer la synchronisation des ruptures estimées entre deux signaux lorsque la probabilité d'observer les ruptures aux mêmes instants est forte et favorise la détection de ruptures propres à un signal, lorsque celui-ci présente des ruptures non communes aux autres signaux. L'interprétation que je fais de ce paramètre est une modélisation des relations de dépendances entre les ruptures, ce qui m'amène à proposer une représentation par un graphe de probabilités dans le chapitre 5.

Il est à noter que si les ruptures peuvent être dépendantes, les signaux sont indépendants conditionnellement aux ruptures. En effet, les observations $X_{j,i}$ sont i.i.d. sur les portions de signal entre deux ruptures, les distributions étant différentes de part et d'autre d'un changement. Ceux-ci peuvent toutefois se produire simultanément sur plusieurs signaux, par exemple par l'action d'une cause extérieure commune, ou par propagation d'une perturbation entre les signaux. La même rupture apparaît donc à l'instant i sur les signaux concernés $\boldsymbol{X}_{j_1,\bullet},\ldots,\boldsymbol{X}_{j_d,\bullet}$, et $R_{j_1,i} = \cdots = R_{j_d,i} = 1$.

La structure de dépendance régissant le système peut être représentée par un graphe dirigé. Selon cette modélisation, plus le graphe est connecté, plus les ruptures ont tendance à être partagées par un grand nombre de signaux, et plus l'estimation d'une segmentation unique est appropriée. À l'inverse, moins le graphe présente d'arêtes, plus les ruptures d'un signal sont propres à ce signal et plus un traitement indépendant de chaque signal paraît adapté. Une remarque importante est à faire : le mécanisme de propagation des ruptures entre les signaux n'induit pas de retard visible dans la série temporelle. Nous considérons que les événements communs sont simultanés, et donc que deux ruptures décalées ont des origines distinctes.

Le paramètre $\boldsymbol{P}$ est le vecteur des probabilités d'observer un motif donné dans les colonnes de la matrice des indicatrices $\boldsymbol{R}$. Ces motifs sont des vecteurs de 0 et de 1, de dimension m, appelés configurations. Elles sont notées ϵ et prennent leurs valeurs dans l'espace $\mathcal{E}$ de dimension L. Sans information a priori sur les motifs autorisés, toutes les combinaisons de 0 et de 1 sont possibles, et $L = 2^m$. Pour écrire la loi de probabilité de $\boldsymbol{P} = (P_{\epsilon_1}, \ldots, P_{\epsilon_L})$, nous faisons le choix d'une loi de probabilité vague, la loi de Dirichlet de paramètres $d_1, \ldots, d_L$:

$$f(\boldsymbol{P}) = \frac{1}{B} \prod_{l=1}^{L} P_{\epsilon_l}^{d_l - 1}, \tag{4.7}$$

où B est une constante de normalisation. En pratique, on choisit la même valeur déterministe pour les paramètres d_l, $1 \leq l \leq L$, fixée à 1, ce qui revient à appliquer une loi uniforme pour les probabilités P_ϵ.

La composante P_ϵ de $\boldsymbol{P}$ est la probabilité d'avoir la configuration ϵ dans les colonnes de $\boldsymbol{R}$. La loi de probabilité de $\boldsymbol{R}$ conditionnellement à $\boldsymbol{P}$ est donc :

$$f(\boldsymbol{R}|\boldsymbol{P}) = \prod_{l=1}^{L} P_{\epsilon_l}^{S_l(\boldsymbol{R})}, \tag{4.8}$$

où $S_l(\boldsymbol{R})$ est la fonction qui renvoie le nombre de fois où ϵ_l apparaît dans $\boldsymbol{R}$.

On suppose que les vecteurs $\boldsymbol{R}_i$, $1 \leq i \leq n$, sont indépendants les uns des autres. On peut donc écrire :

$$f(\boldsymbol{R}) = \prod_{i=1}^{n} f(\boldsymbol{R}_i). \tag{4.9}$$

Finalement, l'expression de l'a priori sur la matrice des indicatrices (4.6) devient :

$$\boxed{f(\boldsymbol{R}|\boldsymbol{P}) f(\boldsymbol{P}) \propto \prod_{l=1}^{L} P_{\epsilon_l}^{S_l(\boldsymbol{R})}.} \tag{4.10}$$

4.1.3 Densité a posteriori

À partir de la vraisemblance marginale composite (4.5) et des a priori sur $\boldsymbol{R}$ et $\boldsymbol{P}$ dont on a déduit l'expression (4.10), on peut écrire la loi a posteriori sur les ruptures dans (4.2) :

$$f(\boldsymbol{R},\boldsymbol{P}|\boldsymbol{X}) \propto L_*(\boldsymbol{X}|\boldsymbol{R}) f(\boldsymbol{R}|\boldsymbol{P}) f(\boldsymbol{P}), \tag{4.11}$$

$$\propto \left(\prod_{j=1}^{m} \prod_{i=1}^{n} (\gamma P_{j,i}(\boldsymbol{X},\boldsymbol{R})^{\gamma-1})^{R_{j,i}} \right) \left(\prod_{l=1}^{L} P_{\epsilon_l}^{S_l(\boldsymbol{R})} \right). \tag{4.12}$$

On marginalise la fonction (4.12) par rapport au terme $\boldsymbol{P}$, qui suit une loi de Dirichlet :

$$\boldsymbol{P} \sim \mathcal{D}_L(S_l(\boldsymbol{R}) + 1). \tag{4.13}$$

Finalement, la densité a posteriori devient :

$$\boxed{f(\boldsymbol{R}|\boldsymbol{X}) \propto \frac{\prod_{l=1}^{L}\Gamma(S_l(\boldsymbol{R})+1)}{\Gamma(n+L)}\prod_{j=1}^{m}\prod_{i=1}^{n}(\gamma P_{j,i}(\boldsymbol{X},\boldsymbol{R})^{\gamma-1})^{R_{j,i}}.} \tag{4.14}$$

La complexité de (4.14) dépend linéairement du nombre de points en temps n et du nombre de signaux m par l'intermédiaire du cardinal L de $\mathcal{E}$, c'est-à-dire du nombre de configurations L dans les colonnes de $\boldsymbol{R}$. Lorsque toutes les configurations ϵ sont possibles, $L = 2^m$, et la complexité de la loi a posteriori dépend alors exponentiellement du nombre de signaux. Certaines pistes sont proposées dans la partie 4.4.1 afin de réduire cette complexité, comme l'application d'un a priori informatif sur $\mathcal{E}$ pour éliminer certaines configurations parmi les 2^m possibles.

4.2 Contrôle de la détection

Pour l'analyse de séries temporelles multivariées, nous avons vu dans la partie 3.2 que la détection de la première rupture se fait sous une condition sur la p-valeur (lemme 3.2.1), et que le risque de détecter à tort une rupture inexistante est contrôlé à un niveau η (proposition 3.2.1). Dans cette partie, je montre que des résultats similaires peuvent être obtenus pour le modèle du *Bernoulli Detector* en multivarié, en s'appuyant sur le fait que la p-valeurs sont indépendantes d'un signal à l'autre. Le cas qui est traité est celui de la détection de la première rupture lorsqu'aucun événement n'a encore été décelé dans la série temporelle. Le lemme suivant fournit la condition à remplir pour détecter la première rupture au point i dans signal j. On suppose que $n > 2$.

Lemme 4.2.1 (détection de la première rupture à la position (j,i))**.** *Dans le cas de la détection de la première rupture de la série temporelle $\boldsymbol{X}$, dans le signal j à l'instant i, $1 < i < n$, la densité de probabilité a posteriori* (4.14) *est maximisée pour $R_{j,i} = 1$ si et seulement si :*

$$P_{j,i}(\boldsymbol{X},\boldsymbol{R}) \leq \frac{\alpha}{(n-2)^{\frac{1}{1-\gamma}}}, \tag{4.15}$$

où le paramètre α est le niveau d'acceptation du test statistique, introduit dans (3.10)*.*

À partir de ce lemme, nous établissons la proposition 4.2.1 concernant le risque de détecter à tort une rupture.

Proposition 4.2.1 (contrôle de la détection d'une unique rupture)**.** *On suppose que la série temporelle ne comporte pas de rupture. Alors la probabilité de faire une erreur de type I dans l'estimateur du MAP $\hat{\boldsymbol{R}}_{MAP}$ est contrôlée au niveau η tel que*

$$\eta = m(n-2)\frac{\alpha}{(n-2)^{\frac{1}{1-\gamma}}}. \tag{4.16}$$

Cette propriété assure notamment que $\eta \leq m\alpha$, pour tout $n > 2$.

La matrice des indicatrices sans rupture est notée $\boldsymbol{R}^0$, et la matrice contenant une unique rupture en (j,i) est notée $\boldsymbol{R}^1$. La démonstration du lemme 4.2.1 fait intervenir le facteur de Bayes défini par

$$B_{10,j,i} = \frac{L(\boldsymbol{X}|H_1(j,i))}{L(\boldsymbol{X}|H_0(j,i))} = \frac{L(\boldsymbol{X}|\boldsymbol{R}^1)}{L(\boldsymbol{X}|\boldsymbol{R}^0)} \tag{4.17}$$

donc d'après (4.5),

$$B_{10,j,i} = \gamma P_{j,i}(\boldsymbol{X},\boldsymbol{R}^1)^{\gamma-1}. \tag{4.18}$$

Démonstration du lemme. D'après l'expression (4.14), le rapport des probabilités a posteriori pour une unique rupture à la position (j,i) s'exprime comme

$$\frac{f(\boldsymbol{R}^1|\boldsymbol{X})}{f(\boldsymbol{R}^0|\boldsymbol{X})} = \gamma P_{j,i}(\boldsymbol{X},\boldsymbol{R}^1)^{\gamma-1} \times \prod_{l=1}^{L} \frac{\Gamma(S_l(\boldsymbol{R}^1)+1)}{\Gamma(S_l(\boldsymbol{R}^0)+1)}. \tag{4.19}$$

Comme les matrices $\boldsymbol{R}^0$ et $\boldsymbol{R}^1$ ne diffèrent qu'à la colonne i, la fonction de comptage des configurations S_l renvoie la valeur nulle pour toutes les configurations, sauf S_1 pour la colonne nulle et S_2 pour la colonne nulle sauf à l'indice j. De plus,

$$S_1(\boldsymbol{R}^0) = S_1(\boldsymbol{R}^1) + 1 \tag{4.20}$$

$$S_2(\boldsymbol{R}^0) + 1 = S_2(\boldsymbol{R}^1), \tag{4.21}$$

donc

$$\frac{f(\boldsymbol{R}^1|\boldsymbol{X})}{f(\boldsymbol{R}^0|\boldsymbol{X})} = \gamma P_{j,i}(\boldsymbol{X},\boldsymbol{R}^1)^{\gamma-1} \times \frac{\Gamma(S_1(\boldsymbol{R}^1)+1)\Gamma(S_2(\boldsymbol{R}^1)+1)}{\Gamma(S_1(\boldsymbol{R}^0)+1)\Gamma(S_2(\boldsymbol{R}^0)+1)} \tag{4.22}$$

$$= \gamma P_{j,i}(\boldsymbol{X},\boldsymbol{R}^1)^{\gamma-1} \times \frac{S_2(\boldsymbol{R}^0)+1}{S_1(\boldsymbol{R}^1)+1} \tag{4.23}$$

$$= B_{10,j,i} \times O_{1,0}, \tag{4.24}$$

où $B_{10,j,i}$ est le facteur de Bayes (4.17) et $O_{1,0}$ est le rapport des probabilités a priori pour la détection de la rupture en (j,i) :

$$O_{1,0} = \frac{S_2(\boldsymbol{R}^0)+1}{S_1(\boldsymbol{R}^1)+1}. \tag{4.25}$$

Alors la nouvelle rupture au point (j,i) est détectée au sens du MAP si et seulement si le rapport des probabilités (4.24) est plus grand que 1, ce qui revient à la condition

$$B_{10,j,i} \geq O_{1,0}^{-1}, \tag{4.26}$$

soit, d'après (4.18) et (4.25),

$$\gamma P_{j,i}(\boldsymbol{X},\boldsymbol{R}^1)^{\gamma-1} \geq \frac{S_1(\boldsymbol{R}^1)+1}{S_2(\boldsymbol{R}^0)+1}. \tag{4.27}$$

En introduisant la paramétrisation choisie dans (3.10) pour la distribution alternative de la p-valeur, on a $\gamma^{\frac{1}{1-\gamma}} = \alpha$ avec $\gamma \in [0,1[$. Alors

$$P_{j,i}(\boldsymbol{X},\boldsymbol{R}^1) \leq \alpha\Big(\frac{S_2(\boldsymbol{R}^0)+1}{S_1(\boldsymbol{R}^1)+1}\Big)^{\frac{1}{1-\gamma}}. \tag{4.28}$$

Or, comme

$$S_1(\boldsymbol{R}^1) = n-3 \quad \text{et} \quad S_2(\boldsymbol{R}^0) = 0, \tag{4.29}$$

alors

$$P_{j,i}(\boldsymbol{X},\boldsymbol{R}) \leq \alpha\Big(\frac{1}{n-2}\Big)^{\frac{1}{1-\gamma}}, \tag{4.30}$$

ce qui prouve le lemme 4.2.1.

On pose

$$\eta = m(n-2)\frac{\alpha}{(n-2)^{\frac{1}{1-\gamma}}}. \tag{4.31}$$

La probabilité de détecter un événement est déduite de (4.30) et conduit à l'expression de la probabilité de fausse alarme, dans la série temporelle sans rupture :

$$P_{FA} = \Pr\left\{\bigcup_{1\leq j\leq m}\bigcup_{1<i<n} P_{j,i}(\boldsymbol{X},\boldsymbol{R}) \leq \frac{\eta}{m(n-2)}\right\}. \tag{4.32}$$

D'après l'inégalité de Boole

$$\Pr\left\{\bigcup_{1\leq j\leq m}\bigcup_{1<i<n} P_{j,i}(\boldsymbol{X},\boldsymbol{R}) \leq \frac{\eta}{m(n-2)}\right\} \leq \sum_{1\leq j\leq m}\sum_{1<i<n} \Pr\left(P_{j,i}(\boldsymbol{X},\boldsymbol{R}) \leq \frac{\eta}{m(n-2)}\right), \tag{4.33}$$

quelles que soient les dépendances entre les p-valeurs. Comme sous H_0 les p-valeurs sont distribuées uniformément, et qu'on suppose qu'il n'y a pas de rupture,

$$\sum_{1\leq j\leq m}\sum_{1<i<n} \Pr\left(P_{j,i}(\boldsymbol{X},\boldsymbol{R}) \leq \frac{\eta}{m(n-2)}\right) = m(n-2)\frac{\eta}{m(n-2)} = \eta, \tag{4.34}$$

donc

$$P_{FA} \leq \eta, \tag{4.35}$$

ce qui correspond au niveau de contrôle souhaité. De plus, comme $0 \leq \gamma < 1$, $n > 2$ et $m \leq 1$,

$$\frac{1}{(n-2)^{\frac{1}{1-\gamma}}} \leq \frac{1}{n-2}, \tag{4.36}$$

donc

$$\eta \leq m(n-2)\alpha, \tag{4.37}$$

ce qui conclut la démonstration de la proposition 4.2.1.

□

Dans cette partie, j'ai énoncé des propriétés du *Bernoulli Detector* pour la détection de la première rupture. Elles portent sur les p-valeurs : le lemme 4.2.1 donne la condition pour favoriser la détection de la rupture en (j,i) et garantir de la sorte un faible risque de deuxième espèce, et la proposition 4.2.1 fournit le niveau η auquel le taux de fausses alarme est contrôlé, c'est-à-dire le risque de première espèce.De ces résultats, nous tirons les mêmes remarques et conclusions que pour le modèle univarié. Le contrôle est démontré pour toute taille $n > 2$, et non pas asymptotiquement. En cas de ruptures multiples dans le même signal j, des dépendances entre les p-valeurs $P_{j,i}(\boldsymbol{X},\boldsymbol{R})$ sont à prendre en compte, or nous avons déjà évoqué la difficulté à les exprimer. Nous soulignons toutefois que les p-valeurs sont indépendantes d'un signal à l'autre, ce qui nous ramène au problème univarié. La formulation des probabilités de détection et de fausse alarme peuvent faire l'objet d'un travail futur, j'ai néanmoins établi des conclusions à partir de résultats empiriques. Dans la partie suivante, nous examinons l'implémentation de l'algorithme du *Bernoulli Detector* pour le traitement de séries temporelles multivariées.

4.3 Algorithme

Les ruptures sont détectées via l'estimateur du MAP associé à la densité de probabilité a posteriori (4.14), noté $\widehat{\boldsymbol{R}}_{MAP}$. Comme dans le cas univarié, $\widehat{\boldsymbol{R}}_{MAP}$ est estimé par une méthode MCMC avec un échantillonneur de Gibbs. Les expériences menées dans la partie 3.4.2 ont permis de valider l'algorithme UniBD, où les p-valeurs ne sont pas mises à jour à chaque étape du mouvement de Gibbs. Étant donnée la complexité de l'expression de la densité de probabilité a posteriori (4.14), cette approximation est conservée dans la version multivariée de l'algorithme du *Bernoulli Detector*, que nous appelons MultiBD. Les notations et la stratégie détaillée dans la partie 3.3 sont adaptées.

L'échantillonneur de Gibbs permet de simuler les vecteurs des configurations, c'est-à-dire les colonnes de la matrice des indicatrices. Les indices temporels sont mélangés, ce qui revient à permuter les colonnes de $\boldsymbol{R}$ par une fonction $\pi(.)$, générée aléatoirement à chaque itération MCMC, et $p = \pi(i)$, $1 < p < n$. À l'itération MCMC v, la matrice mélangée est notée $\boldsymbol{R}^{(v)}_{mat,p} = (\boldsymbol{R}^{(v)}_2, \ldots, \boldsymbol{R}^{(v)}_p, \boldsymbol{R}^{(v-1)}_{p+1}, \ldots, \boldsymbol{R}^{(v-1)}_{n-1})$. Les colonnes $\boldsymbol{R}^{(v)}_k$, $1 < k \leq p$ sont dans l'état v et les colonnes $\boldsymbol{R}^{(v-1)}_k$, $p+1 \leq k < n$ sont toujours dans l'état $(v-1)$. La matrice $\boldsymbol{R}^{(v)}_{mat,\backslash p}$ correspond à $\boldsymbol{R}^{(v)}_{mat,p}$ sans la colonne $\boldsymbol{R}^{(v)}_p$.

Les vecteurs $\boldsymbol{R}_p$ prennent pour valeur l'une des L configurations possibles $\epsilon \in \mathcal{E}$ selon la probabilité conditionnelle :

$$\Pr(\boldsymbol{R}_p = \epsilon | \boldsymbol{R}^{(v)}_{mat,\backslash p}; \boldsymbol{X}) = \frac{\Pr(\boldsymbol{R}^{(v)}_{mat,p}(\epsilon)|\boldsymbol{X})}{\sum_{l=1}^{L} \Pr(\boldsymbol{R}^{(v)}_{mat,p}(\epsilon_l)|\boldsymbol{X})}, \tag{4.38}$$

avec

$$\Pr(\boldsymbol{R}^{(v)}_{mat,p}(\epsilon)|\boldsymbol{X}) = \Pr(\boldsymbol{R}^{(v)}_2, \ldots, \boldsymbol{R}^{(v)}_{p-1}, \epsilon, \boldsymbol{R}^{(v-1)}_{p+1}, \ldots, \boldsymbol{R}^{(v-1)}_{n-1}|\boldsymbol{X}). \tag{4.39}$$

Lorsqu'on met à jour la i^{e} colonne de $\boldsymbol{R}$, c'est-à-dire la p^{e} colonne de $\boldsymbol{R}_{mat,p}$, on ne calcule que les p-valeurs associées aux $\boldsymbol{X}_i$ selon la segmentation donnée par $\boldsymbol{R}_{mat,p}$: $P_{j,i}(\boldsymbol{X},\boldsymbol{R}_{mat,p})$,

pour tout $1 \leq j \leq m$. Soit $l^\star$ l'indice de la configuration ϵ testée dans $\mathcal{E}$ et $\epsilon_l(j)$ la valeur du j^{e} coefficient de la configuration ϵ_l. D'après la relation (4.14) et en faisant l'approximation sur les p-valeurs, la probabilité conditionnelle (4.38) devient

$$\Pr(\boldsymbol{R}_p = \epsilon | \boldsymbol{R}^{(v)}_{mat,\setminus p}; \boldsymbol{X}) \approx \frac{\prod_{j=1}^{m} \left(\gamma P_{j,i}(\boldsymbol{X}, \boldsymbol{R}^{(v)}_{mat,p})^{\gamma-1}\right)^{\epsilon(j)}}{\sum_{l=1}^{L} \left[G_{l,l^\star} \prod_{j=1}^{m} \left(\gamma P_{j,i}(\boldsymbol{X}, \boldsymbol{R}^{(v)}_{mat,p})^{\gamma-1}\right)^{\epsilon_l(j)}\right]}, \tag{4.40}$$

avec

$$G_{l,l^\star} = \frac{S_l(\boldsymbol{R}^{(v)}_{mat,\setminus p}) + 1}{S_{l^\star}(\boldsymbol{R}^{(v)}_{mat,\setminus p}) + d_{l^\star}} \tag{4.41}$$

où $S_l(\boldsymbol{R}^{(v)}_{mat,\setminus p})$ et $S_{l^\star}(\boldsymbol{R}^{(v)}_{mat,\setminus p})$ dénombrent les apparitions des configurations ϵ_l et ϵ respectivement dans les colonnes de la matrice $\boldsymbol{R}^{(v)}_{mat,\setminus p}$. La procédure MultiBD est résumée dans l'algorithme 5. Une étape optionnelle peut s'ajouter pour échantillonner le paramètre $\boldsymbol{P}$. Le détail de l'évaluation de la complexité de cet algorithme est présenté dans l'annexe C.

Le nombre de configurations possibles pour chaque colonne de $\boldsymbol{R}$ est 2^m. Lorsque le nombre de signaux est important, le nombre de configurations à tester pour simuler le vecteur des configurations devient très élevé, ce qui ralentit la résolution. Il devient alors plus judicieux d'échantillonner $\boldsymbol{R}$ coefficient par coefficient, en adaptant le nombre d'itérations MCMC à la vitesse de convergence. L'étape d'échantillonnage est réalisée en $m \times (n-2)$ opérations où seules deux valeurs sont testées. Une autre approche consiste à traiter les signaux paire à paire, puis à combiner les segmentations obtenues. Dans la suite, nous nous contentons d'un nombre limité de signaux. Les résultats de l'application de l'algorithme MultiBD sur des données simulées et des données réelles sont présentés dans la partie 4.4.

4.4 Résultats expérimentaux

Une des particularité du modèle du *Bernoulli Detector* multivarié est l'intégration des corrélations entre les ruptures des composantes spatiales, grâce au paramètre $\boldsymbol{P}$. Dans les expériences menées à la partie 3.4 du chapitre précédent, nous avons confirmé le caractère généraliste du modèle univarié, dû à l'emploi du test de WMW, et qui permet de détecter des ruptures dans des observations distribuées normalement comme en présence de valeurs extrêmes. Cette partie apporte des résultats expérimentaux sur la détection de plusieurs ruptures dans des séries temporelles multivariées par l'algorithme MultiBD 5. Les premiers, dans la section 4.4.1, sont obtenus sur des données simulées, et mettent en avant l'intérêt du paramètre $\boldsymbol{P}$. Les deux autres sections présentent les segmentations estimées sur des données réelles : un suivi de consommation d'un réseau électrique domestique, à la section 4.4.2, et un exemple de données d'hybridation génomique comparative, à la section 4.4.3.

Algorithme 5 : MultiBD, pseudo-échantillonneur de Gibbs

Données : série temporelle $\boldsymbol{X} \in \mathbb{R}^{m \times n}$
Résultat : $\hat{\boldsymbol{R}}_{MAP}$
choisir α
initialiser $\boldsymbol{R}^{(0)}$, $\hat{\boldsymbol{R}}_{MAP} = \boldsymbol{R}^{(0)}$
pour $v \leftarrow 1, V$ **faire**
  mélanger aléatoirement les indices par la fonction π
  pour $p \leftarrow 2, n-1$ **faire**
    $i = \pi^{-1}(p)$
    pour $j \leftarrow 1, m$ **faire**
      calculer $P_{j,i}(\boldsymbol{X}, \boldsymbol{R}^{(v)}_{mat,p})$
    fin
    simuler $\boldsymbol{R}^{(v)}_{p}$ d'après (4.40)
    optionnel : simuler $\boldsymbol{P}$ d'après la loi (4.13)
    calculer $\Pr(\boldsymbol{R}^{(v)}_{mat,p}|\boldsymbol{X})$ d'après la loi a posteriori (4.14)
    si $\Pr(\boldsymbol{R}^{(v)}_{mat,p}|\boldsymbol{X}) > \Pr(\hat{\boldsymbol{R}}_{MAP}|\boldsymbol{X})$ **alors**
      $\hat{\boldsymbol{R}}_{MAP} = \boldsymbol{R}^{(v)}_{mat,p}$
    fin
  fin
fin

4.4.1 Simulations : intérêt du paramètre P

Le paramètre $\boldsymbol{P}$ représente les probabilités d'observer certaines ruptures à un même instant. Il intervient dans la loi de probabilité a priori du modèle du *Bernoulli Detector* par le terme (4.10). Dans un premier temps, l'utilisation de ce paramètre pour l'apprentissage des corrélations entre les rupture et la réduction de la complexité lorsque des informations sont connues a priori est illustré pour deux types de signaux. Dans un deuxième temps, l'effet du paramètre sur des composantes très bruitées est montré. Dans chacun de ces tests, les observations $x_{j,i}$ suivent toutes une loi normale $\mathcal{N}(\mu_s,\sigma)$ dont la moyenne μ_s change d'un segment s à l'autre. Le niveau de bruit est mesuré par le SNR défini dans (3.59) entre les moyennes de deux segments successifs. Pour simplifier les notations, le vecteur d'une configuration ϵ_l, $1 \leq l \leq L$, est désigné par le nombre binaire équivalent à sa valeur, et qui correspond à $l-1$. On associe par exemple à $\epsilon_7 = (1,1,0)'$ le nombre 110, sa probabilité est donc notée P_{110}.

Pour représenter la distribution a posteriori de $\boldsymbol{P}$, on simule chaque composante P_{ϵ_l} suivant la loi (4.13), selon l'étape optionnelle de l'algorithme 5, pour tout $1 \leq l \leq L$ et pour chaque itération MCMC. Comme une rupture est un événement plutôt rare, le nombre de valeurs 1 dans $\boldsymbol{R}$ est petit devant le nombre de points en temps, la configuration vide $\epsilon_1 = (0,\ldots,0)'$ est donc très fréquente par rapport aux $L-1$ autres configurations. Afin d'améliorer la lisibilité des résultats, on calcule les probabilités a posteriori d'avoir les

configurations non vides (contenant au moins un 1) conditionnellement à l'existence d'une rupture. Par exemple, pour l'étude de trois signaux, la probabilité a posteriori d'avoir la configuration $\epsilon = (1,1,0)'$ en cas de rupture est calculée par :

$$\Pr(\boldsymbol{R}_i = (1,1,0)' | \sum_{j=1}^{m} R_{j,i} \geq 1) = \frac{P_{110}}{P_{001} + P_{010} + P_{011} + P_{100} + P_{101} + P_{110} + P_{111}}. \quad (4.42)$$

Le probabilité de la configuration ϵ en cas de rupture est notée $P_{\epsilon|\mathrm{CP}}$ (CP est l'abréviation de *change-point*, rupture en anglais). Ces probabilités conditionnelles sont calculées pour chaque estimation de P au cours des itérations MCMC, et affichées sous forme de diagramme en boîte, donnant la médiane, les quartiles supérieurs et inférieurs, et les valeurs considérées comme aberrantes [1].

A priori informatif et non informatif

La première simulation permet de comparer l'impact du choix d'un a priori informatif ou non sur les configurations : dans un premier temps toutes les configurations sont possibles et $\mathcal{E} = \{0,1\}^m$, et dans un second temps l'ensemble $\mathcal{E}$ est restreint grâce à des informations a priori sur les signaux, d'après lesquelles certaines configurations ont une probabilité nulle de se produire. La série temporelle comporte $m = 4$ signaux de $n = 1050$ points, dont le niveau de bruit entre deux segments successifs est de 0,0 dB. Dans le signal 1, les longueurs des segments sont distribuées uniformément sur l'intervalle [40,80]. Les ruptures (en dehors des premier et dernier points) se retrouvent sur les segments 2, 3 et 4 avec les probabilités 0,9, 0,5 et 0,1 respectivement. À l'exception de ces événements communs avec le signal 1, les signaux 2, 3 et 4 ne présentent pas d'autres ruptures. Les liens entre les ruptures sont représentées dans le graphe de la figure 4.1, où la direction de la flèche du nœud i vers j indique une relation de dépendance : la rupture sur le signal i induit une rupture sur le signal j avec une certaine probabilité. Les ruptures des signaux 2, 3 et 4 sont indépendantes conditionnellement aux ruptures du signal 1. La figure 4.2 donne un exemple d'une telle série temporelle (points noirs), où les positions des ruptures réelles sont indiquées en pointillés bleus.

On applique l'algorithme 5 (MultiBD) sur ces signaux, le seuil d'acceptance α étant fixé à 0,05, avec 2000 itérations MCMC et pour toutes les configurations possibles. Puis un a priori informatif est appliqué. En effet, la connaissance des relations représentées par le graphe 4.1 permet de déduire que la probabilité d'avoir une rupture sur les signaux 2, 3 ou 4 sans qu'il y en ait sur le signal 1 est nulle, en supposant qu'aucun élément extérieur ne provoque de rupture sur les signaux 2, 3 et 4. Ainsi, l'ensemble des configurations est réduit aux $L = 9$ possibilités suivantes :

$$\mathcal{E} = \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \right\}, \quad (4.43)$$

1. Les points aberrants sont ceux situés hors de l'intervalle $[Q1 - 1{,}5 * (Q3 - Q1); Q3 + 1{,}5 * (Q3 - Q1)]$ où $Q1$ et $Q3$ sont les premier et troisième quartiles.

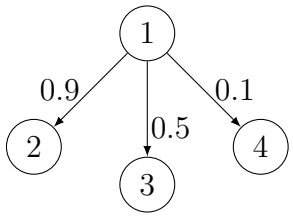
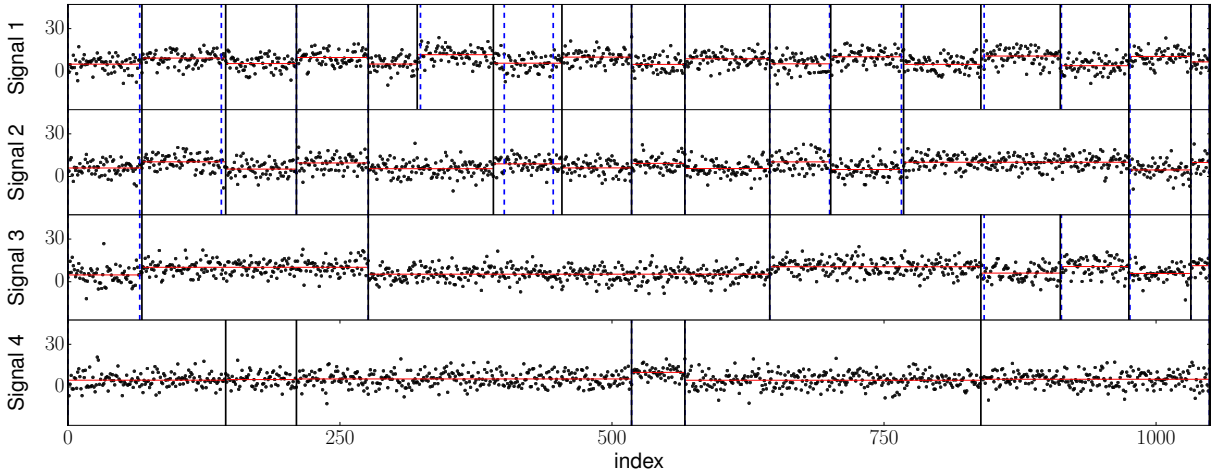


FIGURE 4.1 – Graphe des dépendances entre les ruptures des signaux. Les nœuds représentent l'indice du signal où se produit la rupture, et les arêtes indiquent l'existence d'un lien entre les ruptures. Les arêtes sont dirigées du nœud 1 vers les autres, ce qui signifie que les ruptures du signal 1 entraînent l'apparition simultanée d'une rupture sur les autres signaux, avec une probabilité 0,9 pour le signal 2, 0,5 pour le signal 3 et 0,1 pour le signal 4.

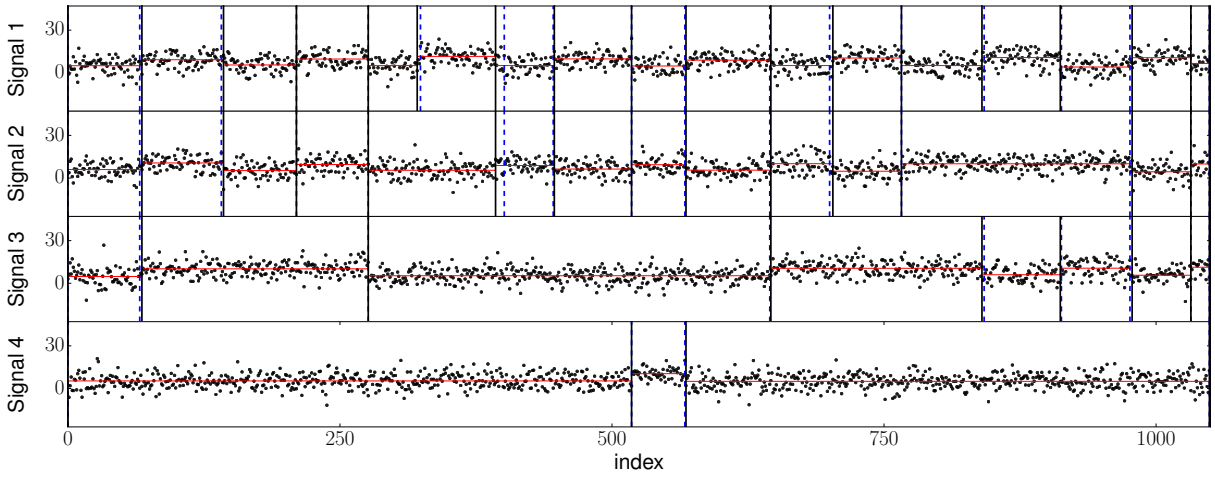
au lieu des 16 valeurs du cas non informatif.

Les estimations de $\boldsymbol{R}_{MAP}$ sont illustrées dans la figure 4.2a pour le cas non informatif et dans la figure 4.2b pour le cas informatif, les ruptures détectées étant indiquées par les traits verticaux noirs. Sur cet exemple, les deux résolutions ont permis de détecter conjointement les ruptures sur les différents signaux, et de synchroniser les événements simultanés. Le modèle du *Bernoulli Detector* parvient donc à apprendre les probabilités jointes des ruptures sur les signaux. Les segmentations obtenues sont à la fois globales, car les ruptures communes à plusieurs signaux sont bien estimées à la même position, et locales, car les particularités de chaque signal sont retrouvées. L'introduction d'informations a priori réduit notablement la complexité du problème, ce qui conduit à un gain de temps de calcul significatif dans l'application de l'algorithme MultiBD, passant de $4{,}7 \times 10^3$s à $3{,}2 \times 10^3$s. Comme certaines ruptures ne sont pas positionnées à la position exacte de la rupture réelle, par exemple vers $i = 270$, la qualité des détections selon les deux a priori testés est évaluée par rapport à la position des ruptures estimées. Les critères de rappel et de précision sont calculés en fonction de la tolérance d sur les positions des ruptures issues de $\boldsymbol{R}_{MAP}$ par rapport aux positions exactes. Les courbes, obtenues pour 20 simulations avec 2000 itérations MCMC, sont données dans la figure 4.3. On constate que les deux cas de figure génèrent des valeurs très proches. Les critères augmentent rapidement avec l'introduction de la tolérance, et dépassent la valeur 0,5 dès un écart de plus ou moins un point.

On représente dans la figure 4.4 les valeurs des probabilités a posteriori $\hat{P}_{\epsilon|\mathrm{CP}}$, pour chaque $\epsilon \in \mathcal{E}$ conditionnellement à l'existence d'une rupture, obtenues sur la série temporelle de la figure 4.2. Les deux probabilités ayant les plus grandes valeurs médianes sont obtenues pour les configurations 1110 (en bleu) et 1100 (en rouge). Le *Bernoulli Detector* a permis d'estimer que les ruptures apparaissent simultanément le plus souvent sur les signaux 1, 2 et 3, et sur les signaux 1 et 2, en accord avec la série temporelle de la figure 4.2 et du graphe 4.1. Le cas informatif estime des distributions a posteriori centrées sur des valeurs plus grandes que dans le cas non informatif.



(a) La segmentation est obtenue avec un a priori non informatif, où $L = 16$.

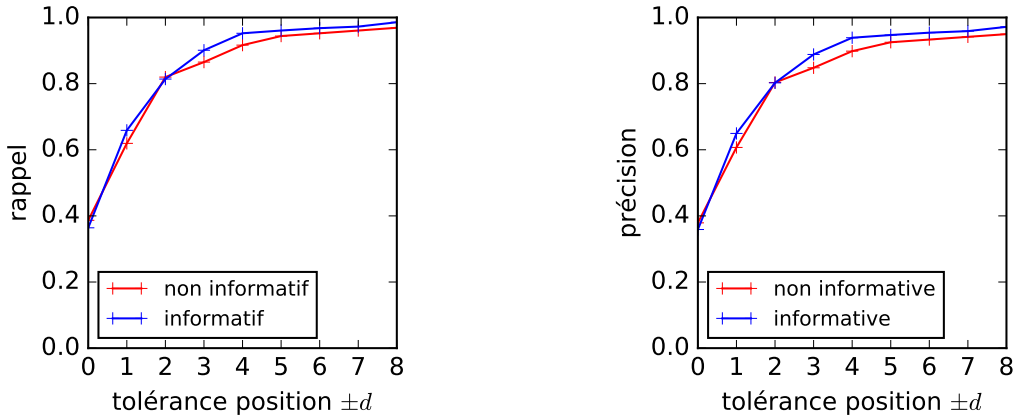


(b) La segmentation est obtenue avec un a priori informatif, où $L = 9$, d'après la connaissance des relations de dépendances entre les ruptures des signaux, donnée par le graphe 4.1. Les seules configurations possibles sont données par (4.43).

FIGURE 4.2 – Segmentation d'une série temporelle multivariée par l'estimateur $\boldsymbol{R}_{MAP}$ (noir) de l'algorithme MultiBD avec $\alpha = 0{,}05$ et 2000 itérations MCMC. La série temporelle comporte $n = 1052$ points, $m = 4$ signaux, et les observations suivent des lois normales de moyennes 5,00 et 10,00 alternativement d'un segment sur l'autre. Le rapport signal sur bruit entre deux segments est 0,0 dB. Les relations de dépendance entre les ruptures des segments sont données par le graphe 4.1. Les ruptures estimées sont en trait vertical noir, les vraies sont en pointillés bleus. Les segments rouges sont les moyennes empiriques calculées sur chaque segment estimé.

Détection d'événements dans un signal très bruité

La seconde simulation permet de tester la capacité du *Bernoulli Detector* à segmenter un signal très bruité, en apprenant la structure de dépendance entre les ruptures pour



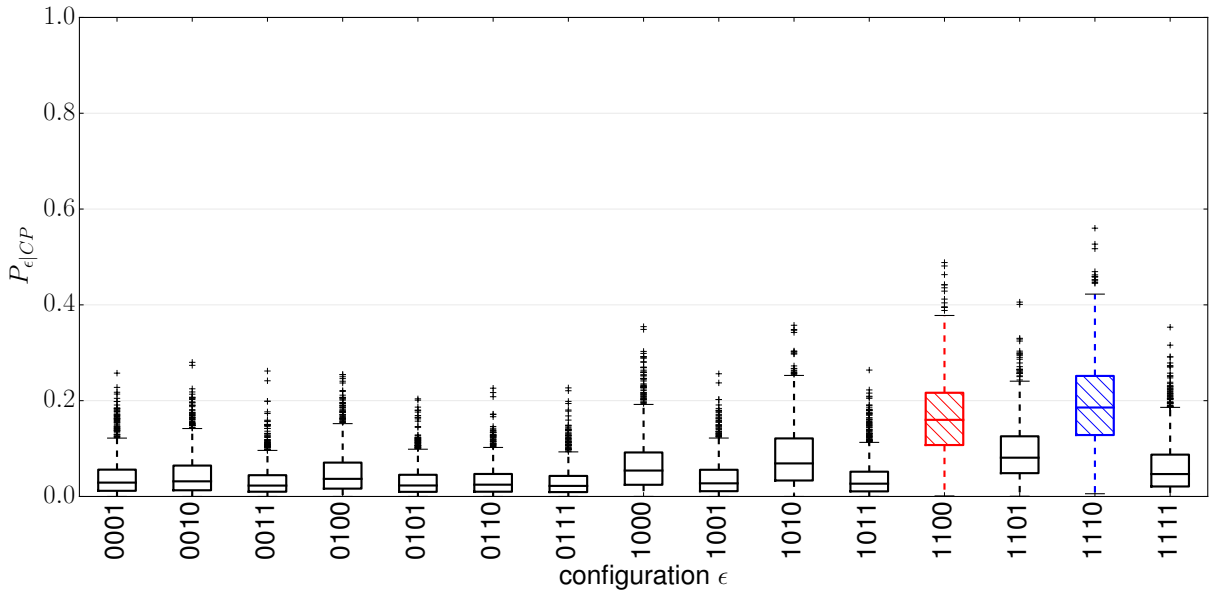
(a) Courbe de rappel.

(b) Courbe de précision.

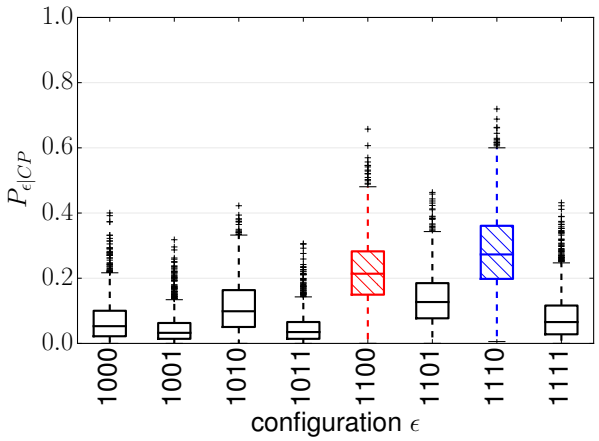
FIGURE 4.3 – Courbes de rappel 4.3a et de précision 4.3b en fonction de la tolérance d autorisée sur les positions des ruptures estimées. Les segmentations $\boldsymbol{R}_{MAP}$ sont estimées par l'algorithme MultiBD avec $\alpha = 0{,}05$ sur 20 séries temporelles du même type que la figure 4.2, pour 2000 itérations MCMC. La courbe en rouge est obtenue avec a priori non informatif, où $L = 16$, et la courbe en bleu est obtenue avec a priori informatif où $L = 9$.

favoriser la détection malgré le niveau de bruit. Trois séries temporelles sont générées, où le rapport signal sur bruit entre deux segments successifs est de 5,0 dB dans les signaux 1 et 2, et de $-2{,}5$ dB dans le signal 3. Les trois signaux sont segmentés de façon indépendantes, avec des longueurs de segments générés uniformément sur les intervalles [60,100], [40,80] et [100,130] pour les signaux 1 à 3 respectivement. Dans le dernier signal, des événements supplémentaires s'ajoutent : chaque rupture des signaux 1 et 2 se produit simultanément dans le signal 3 avec la probabilité 0,5, comme l'explique le schéma de la figure 4.5.

Ces séries temporelles sont traitées conjointement par l'algorithme MultiBD avec un a priori non informatif sur les probabilités $\boldsymbol{P}$. Pour évaluer l'effet de ce paramètre, les signaux sont également analysés un par un avec l'algorithme UniBD, ce qui revient à supposer que les événements observés se produisent indépendamment les uns des autres. Dans chaque cas, on effectue 2000 itérations MCMC. La figure 4.6 donne un exemple de segmentation, obtenue avec un traitement conjoint par MultiBD (figure4.6a) et par un traitement signal par signal par UniBD (figure 4.6b). Ces résultats montrent que le traitement conjoint a permis de détecter plus de ruptures dans le signal 3, le plus bruité, tandis que certaines ruptures ne sont pas du tout détectées par le traitement indépendant (par exemple à $i = 393$), ou sont mal positionnées (à $i = 246$). Ce signal est difficile à segmenter, en raison du mauvais rapport signal sur bruit, mais aussi de la longueur des segments entre deux ruptures qui peuvent être petits, comme autour de 250. En effet, ce signal présente des ruptures propres et des ruptures provoquées indépendamment par les signaux 1 et 2. Comme les sauts de moyennes se font seulement entre les deux valeurs 5,0 et 10,0, les observations d'un segment trop petit ont tendance à être perçues comme des valeurs aberrantes, auxquelles le test de WMW est robuste. Les événements des signaux



(a) Application d'un a priori non informatif, où $L = 16$.



(b) Application d'un a priori informatif, où $L = 9$, d'après la connaissance des relations de dépendances entre les ruptures des signaux, donnée par le graphe 4.1.

FIGURE 4.4 – Estimation des probabilités $P_{\epsilon|CP}$ d'obtenir chaque configuration ϵ conditionnellement à l'existence d'une rupture, obtenues par l'algorithme MultiBD, sur la série temporelle de la figure 4.2, pour 1500 itérations MCMC, après 500 itérations de préchauffe. Les distributions de plus grande valeur médiane sont affichées en bleu et en rouge. Il s'agit des configurations les plus probables 1110 et 1100.

1 et 2, moins bruités, sont détectés par les deux traitements. On constate toutefois que la méthode multivariée force la synchronisation des ruptures très proches mais indépendantes entre les signaux 1 et 2, comme à $i = 804$ et $i = 808$, et conduit à l'estimation d'un faux positif à $i = 126$ sur le signal 2, alors que cette rupture n'affecte que le signal 1.

Les valeurs moyennes de rappel et de précision pour le signal 3, le plus bruité, sont calculées sur les segmentations de 20 séries temporelles générées comme celle de la figure 4.6. Les courbes sont tracées à la figure 4.7 en fonction de la tolérance d appliquée sur la position des ruptures estimées. Le rappel est inférieur à 0,4, ce qui montre que les ruptures sont difficiles à détecter, et la précision est inférieure à 0,7 pour $d = 5$, ce qui indique que les faux positifs sont assez nombreux dans les détections. Pour une faible tolérance, inférieure à 5, l'algorithme MultiBD est plus performant que le traitement indépendant des signaux par UniBD. Le traitement conjoint permet donc de mieux positionner les ruptures et de synchroniser les événements simultanés de $\boldsymbol{R}$, malgré le niveau de bruit élevé du signal 3. Un inconvénient est que l'approche MultiBD a tendance aussi à associer des ruptures indépendantes entre les signaux 1 et 2. Lorsqu'on subodore l'existence d'une forme de dépendance entre les ruptures, l'introduction du paramètre $\boldsymbol{P}$ présente un intérêt pour la détection, notamment si l'une des composantes est fortement bruitée, et ce malgré l'absence d'informations sur la structure de dépendance entre les ruptures. Ces propriétés du *Bernoulli Detector* multivarié ayant été établies, le modèle est appliqué sur des données réelles.

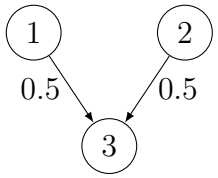
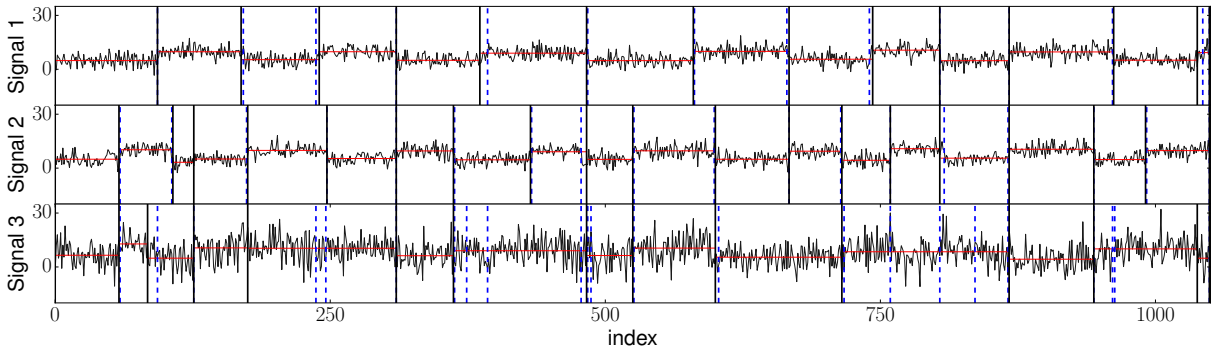


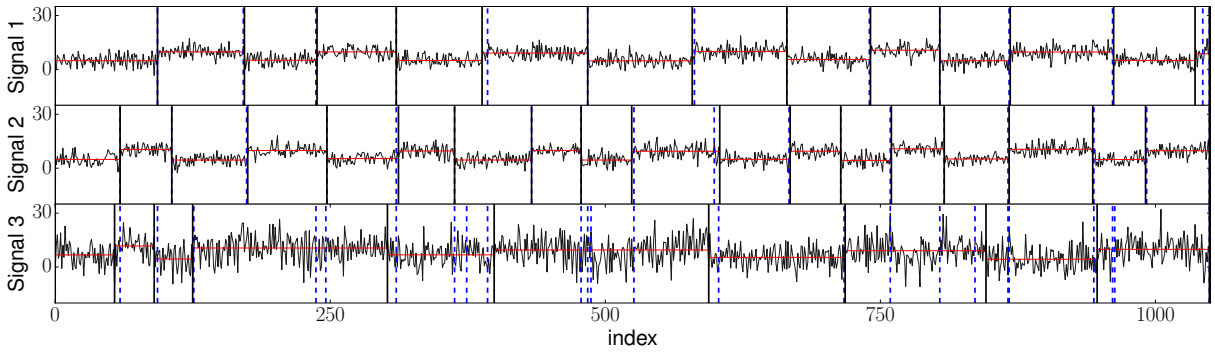
FIGURE 4.5 – Graphe des dépendances entre les ruptures des signaux. Les nœuds représentent l'indice du signal où se produit la rupture, et les arêtes indiquent l'existence d'un lien entre les ruptures. Les ruptures des signaux 1 et 2 se produisent indépendamment les unes des autres, et se retrouvent dans le signal 3 avec une probabilité de 0,5.

4.4.2 Application réelle sur des données électriques

La première application du *Bernoulli Detector* sur des données réelles consiste à analyser des mesures de consommation électrique domestique. La série temporelle provient de la base de données [Lichman, 2013]. L'expérience réalisée a consisté à enregistrer plusieurs compteurs de puissance électrique dans un réseau domestique, avec une mesure toutes les minutes pendant une durée de 4 ans. Les signaux que nous intéressent correspondent au suivi du compteur général (signal 1) et de compteurs secondaires, couvrant la cuisine (signal 2), la buanderie (signal 3), et la partie alimentant à la fois le chauffe-eau et la climatisation (signal 4). Les autres parties du réseau électrique ne sont pas disponibles, la consommation électrique de certains éléments inconnus dans l'habitation n'est donc décelable qu'au niveau du compteur général. Un des intérêts de ce jeu de données est la simplicité du système, dont la structure de dépendance est aisément identifiable. En effet, l'allumage ou l'extinction d'un équipement consommant du courant électrique induit une variation immédiate



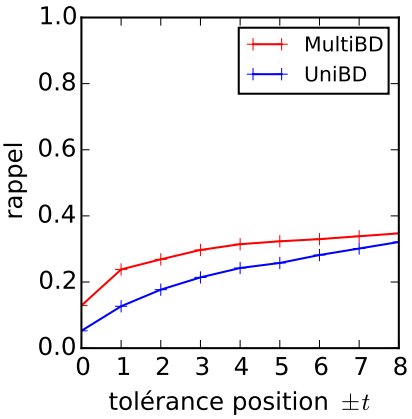
(a) Segmentation obtenue avec l'algorithme MultiBD.



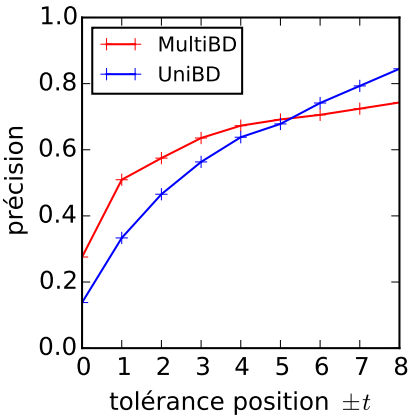
(b) Segmentations obtenues avec l'algorithme UniBD.

FIGURE 4.6 – Segmentation d'une série temporelle multivariée (noir) par l'algorithme MultiBD et l'algorithme UniBD appliqué sur chaque signal. 2000 itérations MCMC sont effectuées avec $\alpha = 0{,}05$. Les dimensions de la série temporelle sont $n = 1052$ et $m = 3$, les observations suivent des lois normales. Le rapport signal sur bruit entre deux segments est de 5,0 dB pour les signaux 1 et 2, et de $-2{,}5$ dB pour le signal 3. Les relations de dépendance entre les ruptures des segments sont données par le graphe 4.5. Les ruptures estimées sont en traits pleins noirs, les vraies sont en pointillés bleus. Les segments rouges sont les moyennes empiriques calculées sur chaque segment estimé.

de la puissance mesurée par le compteur associé à l'équipement, ainsi que de la puissance mesurée par le compteur général. Les relations entre les ruptures des signaux 1 à 4 sont ainsi représentées par le graphe de la figure 4.8. Une variable latente, notée $5^\star$, représente les autres éléments du système qui ne sont pas mesurés, et qui est à l'origine de ruptures indépendantes de celles des signaux 2, 3 et 4, dans le signal 1. Les ruptures des compteurs 2, 3 et 4 sont causées par des processus distincts et sont donc indépendantes. Signalons que pour cette application, le graphe 4.8 décrit également les relations de dépendances entre les variables mesurées par les compteurs. En effet, la puissance du compteur général est la somme des puissances consommées par chaque appareil, mais aussi des puissances consommées par le reste des équipements électriques de l'habitation, et dont nous n'avons aucune mesure intermédiaire. Pour intégrer ces éléments dans le modèle, un a priori infor-



(a) Courbe de rappel.



(b) Courbe de précision.

FIGURE 4.7 – Courbes de rappel et de précision en fonction de la tolérance d autorisée sur les positions des ruptures estimées, pour le signal 3 de 20 séries temporelles similaires à celle de la figure 4.6. Les courbes en bleu (UniBD) sont obtenues avec l'algorithme UniBD 4 pour un traitement indépendant de chaque signal, et les courbes en rouge (MultiBD) sont obtenues avec l'algorithme MultiBD 5 pour un traitement conjoint des trois signaux.

matif a été introduit, et seules les configurations ϵ_l où les ruptures des signaux 2, 3 et 4 apparaissent aussi dans le signal 1 ont été sélectionnées, ainsi que la configuration sans rupture et la configuration 1000 avec uniquement une rupture dans le signal 1, due au signal inconnu 5*. La fonction de vraisemblance (4.5) du *Bernoulli Detector* n'a pas été modifiée, puisqu'elle repose sur les p-valeurs et pas directement sur les mesures de puissance $\boldsymbol{X}_{j,\bullet}$. Le *Bernoulli Detector* est appliqué sur les signaux de $\boldsymbol{X}$, considérés comme indépendants, mais auxquels nous attachons la structure de dépendance des ruptures.

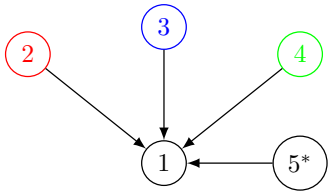


FIGURE 4.8 – Représentation des relations de dépendances du réseau électrique par un graphe. Le nœud j est une rupture dans la mesure de la puissance du compteur j. Le compteur 1 est le compteur principal sur l'ensemble de l'habitation, le 2 est celui de la cuisine, le 3 celui de la buanderie et le 4 celui du chauffe-eau et de la climatisation. Le nœud 5 marqué d'une étoile fait référence au reste de l'équipement électrique de l'habitation pour lequel nous n'avons pas de mesure.

Nous disposons d'assez d'informations pour pouvoir réduire le nombre de configurations possibles : celles où une rupture dans l'un des signaux 2, 3 ou 4 ne s'observe pas dans

le signal 1 sont éliminées, en partant du principe que les ruptures des sous-parties du réseau sont décelables au niveau du compteur général. Nous retrouvons l'ensemble $\mathcal{E}$ défini en (4.43), comportant $L = 9$ configurations. L'algorithme MultiBD est appliqué sur un intervalle temporel de 400 points, avec 2000 itérations MCMC. La série temporelle est tracée dans la figure 4.9, et les ruptures estimées par $\hat{\boldsymbol{R}}_{MAP}$ sont représentées par des traits verticaux. On remarque que de nombreuses portions, surtout sur les signaux 2, 3 et 4, sont peu bruitées, et les mêmes valeurs se répètent souvent dans les mesures. En raison de l'emploi du test de WMW sur les rangs, le détecteur est sensible aux faibles variations au sein des segments, ce qui explique l'apparition de ruptures inattendues, par exemple sur le signal 3. On constate en revanche que les petites portions de quelques points ne sont pas segmentées (vers 100 dans le signal 2 ou vers 60 ou 360 minutes dans le signal 4), probablement en raison du nombre insuffisant d'observations. Le signal 1 présente de nombreuses variations qui lui sont propres, mais dont le comportement s'éloigne d'une fonction constante par morceaux. Il paraît difficile d'évaluer la qualité de la détection sur ces parties, et nous ne disposons pas d'un relevé des événements réels. La segmentation de la figure 4.9 montre cependant que le modèle du *Bernoulli Detector* synchronise les ruptures, en accord avec les relations de dépendance connues.

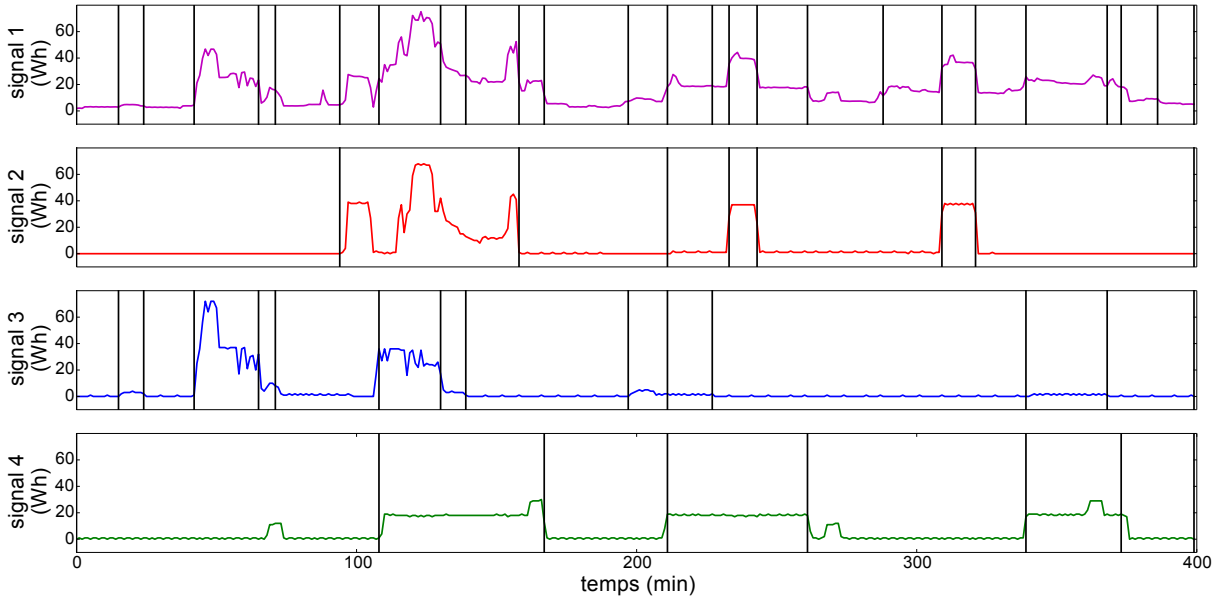


FIGURE 4.9 – Segmentation des mesures de consommation électrique domestique par la méthode MultiBD, obtenue pour $\alpha = 0{,}01$, avec un a priori informatif, après 2000 itérations MCMC. L'estimation du MAP est représentée par les traits verticaux noirs.

Les distributions des probabilités $P_{\epsilon|CP}$ sont tracées à la figure 4.10. Les configurations qui apparaissent le plus souvent en cas de rupture sont, dans l'ordre décroissant des médianes, 1010, 1100, 1001 et 1000 (en couleurs). Le *Bernoulli Detector* a donc appris que les événements simultanés sur les couples de signaux (1,3), (1,2) et (1,4) sont les plus pro-

bables, en accord avec le modèle de dépendance 4.8. Il apparaît de plus que de nombreuses ruptures ne se produisent que dans le signal 1, elles sont dues aux ruptures de la variable latente $5^\star$, et sont indépendantes des signaux 2, 3 et 4 étudiés.

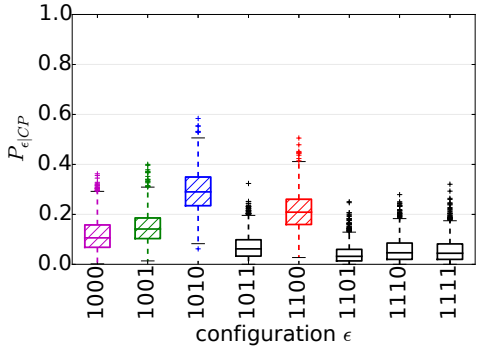


FIGURE 4.10 – Probabilités a posteriori $P_{\epsilon|\mathrm{CP}}$ obtenues par la méthode multiBD sur les données électriques, avec un a priori informatif. Les distributions avec les plus grandes médianes sont en couleurs.

Finalement, l'application du *Bernoulli Detector* sur ces données de consommation électrique met en évidence l'intérêt du paramètre $\boldsymbol{P}$, grâce auquel les ruptures communes à plusieurs signaux sont détectées au même instant. La loi choisie pour $\boldsymbol{P}$ correspond à un a priori informatif, où les relations entre les signaux électriques sont prises en compte. Il s'agit de la loi de Dirichlet d'ordre $L = 9$, de paramètres $d_l = 1$, pour tout $1 \leq l \leq L$. Ces indices l correspondent aux configurations ϵ_l qui ont une probabilité non nulle d'apparaître dans $\boldsymbol{R}$: la configuration sans rupture 0000, celle avec seulement une rupture sur le signal 1, 1000, et celles où les ruptures des signaux 2, 3 et 4 sont observées simultanément dans le signal 1, comme 1010, 1100, 1001 ou 1011. L'existence des relations de dépendance entre les ruptures, qui conduit à l'introduction d'un a priori informatif dans le modèle, permet une détection plus précise et surtout plus rapide des événements. La simple connaissance des arêtes du graphe 4.8 ou de leur absence est suffisante, les valeurs des probabilités jointes entre les ruptures ne sont pas nécessaires. Les données étant peu bruitées, la segmentation par notre modèle est sensible aux petites variations, mais ne parvient pas toujours à détecter les petits segments. En traitant un extrait de quelques centaines de points de ce jeu de données, on parvient toutefois à estimer partiellement la structure qui relie les signaux entre eux, en faisant ressortir les configurations les plus probables pour la portion de signal étudié.

4.4.3 Application réelle sur des données génomiques

La deuxième application choisie pour l'étude du modèle du *Bernoulli Detector* est la détection d'anomalies dans des données d'hybridation génomique comparative. Elle est fréquemment présentée dans des articles traitant de la détection de ruptures multiples,

notamment dans un contexte multivarié [Picard *et al.*, 2011, Shah *et al.*, 2007, Bleakley et Vert, 2011], et constitue un champ de recherche important dans le domaine de la bioinformatique. Il s'agit de la comparaison du nombre de copies de portions d'ADN dans les cellules prélevées chez un patient avec les quantités de référence du génome d'une cellule saine. En effet, à l'origine de certaines pathologies, en particulier pour des tumeurs, il arrive que certains segments de l'ADN des cellules mutées se retrouvent dupliqués (on parle d'amplification) ou soient manquants (on parle de délétion). En décelant des dérégulations dans le nombre de copies de l'ADN du patient, et sur quelles portions du génome elles se produisent, cette technique permet d'identifier les gènes défectueux, et d'élaborer un diagnostic sur l'origine d'une pathologie.

Le dispositif expérimental emploie des puces à ADN qui sont des micro matrices sur lesquelles les fragments d'ADN viennent s'hybrider, d'où le nom *array Comparative Genomic Hybridization* ou aCGH. Le principe de l'hybridation génomique comparative est expliqué dans [Pollack *et al.*, 1999], et est illustré dans la figure 4.11. L'ADN est extrait des cellules et découpé en petites portions de quelques milliers de paires de bases. Celui de la cellule malade est marqué par des fluorochromes de couleur rouge, tandis que l'ADN de référence est marqué par des fluorochromes de couleur verte. Les segments sont alors envoyés sur la puce à ADN, dont la surface est recouverte de sondes auxquelles les fragments d'ADN sain et défectueux viennent s'hybrider. L'emplacement de chaque sonde correspond à une portion précise de l'ADN. Grâce aux marqueurs fluorescents, une analyse optique de la puce permet de déterminer les endroits où l'ADN est en excès, là où la puce prend la couleur rouge, en défaut, là où la couleur est verte, et là où les quantités d'ADN sont normales, où la couleur est jaune. Un dispositif d'imagerie mesure le ratio entre le nombre de copies des deux ADN pour tous des fragments, et renvoie une séquence du logarithme en base 2 du ratio. Cette séquence, appelée profil, est assimilée à une série temporelle, où la dimension temporelle correspond aux indices des portions d'ADN, donc à leur position dans un génome de référence. Comme les variations du nombre de copies couvrent plusieurs fragments successifs d'ADN, le profil prend la forme d'un signal constant par morceaux, où le logarithme en base 2 du ratio de l'ADN en quantité normale est distribué autour de 0, tandis qu'en présence d'anomalies on observe des segments de moyenne non nulle, positive pour une amplification et négative pour une délétion. De telles séquences sont visibles à la figure 4.12. Les données aCGH mesurent les variations du nombre de copies de l'ADN, mais il existe d'autres types de mutations, comme les inversions de portions de code, que cette technique ne révèle pas. Un des exemples d'utilisation de ce genre de données en recherche médicale peut être trouvé dans [Blaveri *et al.*, 2005], où les auteurs parviennent à localiser les fragments de chromosomes ayant souvent tendance à s'altérer dans le cas de cancers, et à classifier des patients selon leurs profils.

De nombreuses méthodes ont été proposées pour traiter ces données, par HMM [Fridlyand *et al.*, 2004,Shah *et al.*, 2006], où les états (normal, en excès et en défaut) sont estimés avec la segmentation, par le fused LASSO [Tibshirani et Wang, 2008] ou une régression quantile [Eilers et De Menezes, 2005], qui fournissent également la fonction d'approximation, par la méthode E-divisive présentée au chapitre 1 [Matteson et James, 2014] ou encore par l'algorithme SaRa de [Niu et Zhang, 2012]. Les données aCGH présentent des valeurs

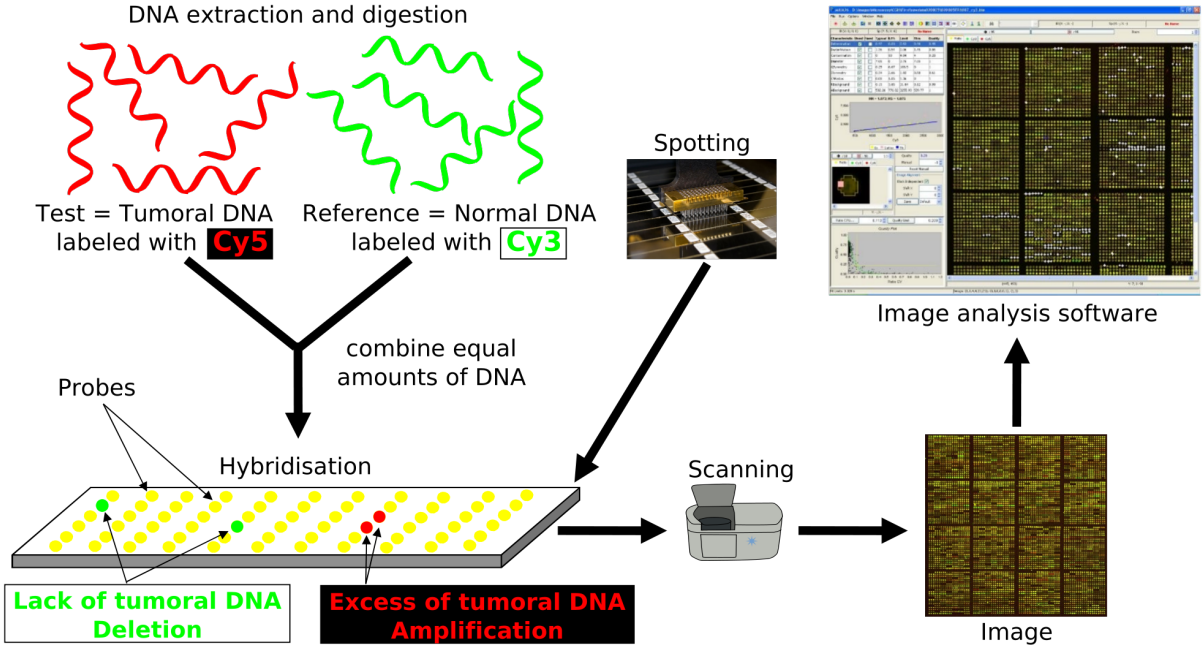


FIGURE 4.11 – Principe de l'hybridation génomique comparative sur une puce à ADN [Wikimedia, 2012].

aberrantes, comme on peut de voir à la figure 4.12. Les profils sont alors filtrés, afin de supprimer ces valeurs extrêmes, ou bien une méthode de segmentation robuste à cette forme de bruit est choisie. Ainsi la régression quantile de [Eilers et De Menezes, 2005] est appliquée à la médiane du signal. Le développement de nouvelles technologies donne accès depuis quelques années à des mesures de profils de plusieurs patients, réalisées en parallèle, nécessitant des méthodes de segmentation multivariées. L'intérêt d'un traitement conjoint de plusieurs profils réside dans le renforcement statistique de certaines caractéristiques du signal. En supposant que les patients ont des anomalies génétiques communes, ces méthodes les mettent mieux en valeur. Suivant ce raisonnement, la méthode du *group* fused LASSO, présentée dans la section 1.1.4, a été proposée dans [Bleakley et Vert, 2011]. La méthode BARD est également appliquée aux données aCGH multivariées [Bardwell et Fearnhead, 2014]. L'approche de [Fan et Mackey, 2015] est illustrée sur ce type de données. La complexité du problème devenant rapidement problématique, les auteurs de [Picard *et al.*, 2011] développent une procédure de programmation dynamique sur un modèle gaussien.

Notre modèle du *Bernoulli Detector* réalisant une segmentation à la fois commune et spécifique à chaque signal d'une série temporelle, son application aux données aCGH est intéressante, notamment en regard de la plupart des autres approches qui estiment une segmentation générale pour tous les profils. Le jeu de données que nous traitons provient du paquet `R ecp` [James et Matteson, 2013], et a été présenté dans la publication médicale [Stransky *et al.*, 2006]. Il est constitué de 43 profils de 2215 fragments d'ADN, extraits de cellules tumorales de vessies chez 43 patients. Cet ensemble de mesures et considéré

comme une série temporelle multivariée, formée des profils d'hybridation de plusieurs patients et dont la composante temporelle correspond à la position des fragments d'ADN dans l'ADN complet issu des 23 chromosomes mis bout-à-bout. Comme la complexité de l'algorithme MultiBD dépend du nombre L de configurations, et que nous ne disposons d'aucune information a priori sur les corrélations entre les ruptures qui permettrait de réduire L, seuls quelques profils sont traités conjointement : 8, 9, 19, 21, 48, 49 et 53. Ces derniers, donnés dans la figure 4.12, ont été choisis de manière à confronter des signaux avec apparemment peu d'anomalies, comme le profil 8, avec des profils plus morcelés, comme le profil 53. Le nombre de configurations est donc $L = 2^7 = 128$. La valeur du seuil d'acceptation α est fixée à $1{,}0 \times 10^{-4}$ a posteriori, évitant ainsi la sur-segmentation des profils avec peu d'anomalies, tout en détectant un maximum de ruptures dans le signal 53.

Le résultat de la détection pour 3000 itérations MCMC est indiqué par les traits verticaux de la figure 4.12. On observe des ruptures communes, délimitant les mêmes portions d'ADN anormalement copié sur plusieurs profils. La plupart des signaux possèdent des ruptures qui leur sont propres. Certaines ruptures paraissent manquer sur le profil 53 autour de l'indice 900. À l'inverse, certaines ruptures dans les longues portions des profils 48 et 19 semblent douteuses, aux indices 1400, 1550 ou 1700. Comme l'interprétation des résultats est limitée en l'absence de vérité terrain ou d'un avis d'expert médical, nous nous comparons à d'autres méthodes.

L'extension multivariée du fused LASSO, le *group* fused LASSO, est appliquée. L'approximation de chaque profil par une fonction constante par morceaux est améliorée par l'augmentation du nombre de profils. L'algorithme efficace est fourni par le paquet `MATLAB GFLseg` [Bleakley et Vert, 2011]. Le résultat est donné par la figure 4.13. La segmentation résultante est estimée pour tous les profils, en revanche les coefficient des approximations constantes par morceaux sont adaptés à chaque patient. La troisième méthode testée est la méthode BARD de [Bardwell et Fearnhead, 2014]. Cette approche permet de classer les segments en deux catégories : normaux et anormaux. La distribution normale est choisie a priori pour la fonction de vraisemblance, et le support de la moyenne μ_A des segments anormaux est $[-0{,}70, -0{,}15] \cup [0{,}15, -0{,}70]$. Les longueurs des segments sont distribuées selon des lois négatives binomiales de paramètres (2,0,0,02) pour les segments normaux et (3,0,0,08) pour les segments anormaux. La probabilité de transition d'un segment anormal vers un normal est fixée à $\pi_N = 0{,}9$. La proportion de profils simultanément affectés par une anomalie p_j de l'expression (1.74) vaut 3/7 pour tout $1 \leq j \leq 7$. Ces nombreuses valeurs ont été évaluées empiriquement, d'après l'allure générale des données. La probabilité que les observations soient dans un segment anormal est tracée en bleu à la figure 4.14, en superposition du profil 21. Les ruptures sont placées au seuil 0,17, au-delà duquel une anomalie est détectée.

Les trois méthodes employées aboutissent à des segmentations assez différentes les unes des autres. Il est essentiel de souligner que les démarches diffèrent dans leur objectifs, et sont complémentaires. Le *group* fused LASSO et l'algorithme BARD estiment une segmentation unique pour tous les signaux. L'implémentation de l'approche de type LASSO est très efficace, et ne demande pas de réglage de paramètres, la valeur du paramètre de régularisation λ étant déterminée a posteriori par application d'un critère de moindres carrés sur

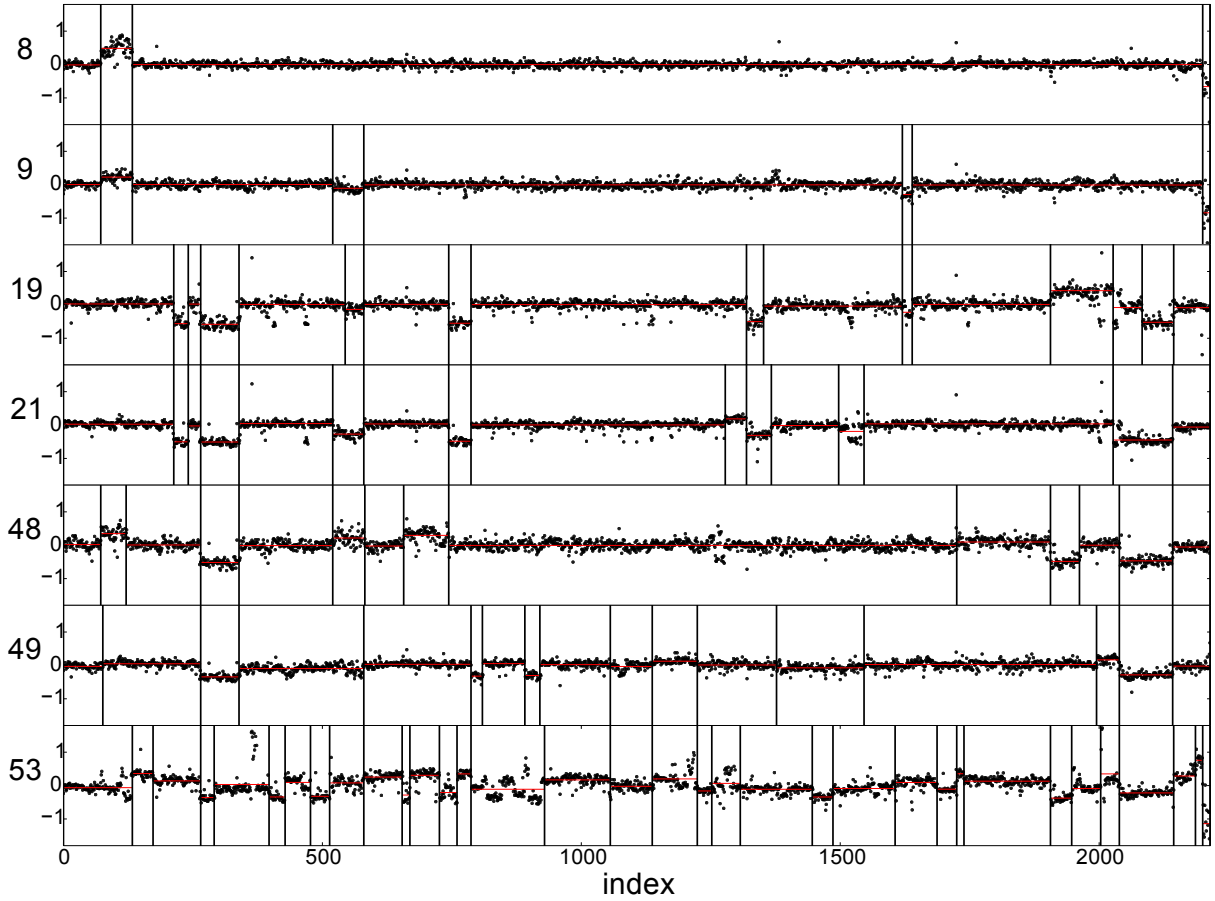


FIGURE 4.12 – Détection conjointe des ruptures sur les données aCGH par l'algorithme MultiBD (lignes verticales), avec α =1,0 $\times 10^{-4}$ et 3000 itérations MCMC. Les moyennes empiriques des segments estimés sont tracées en rouge.

les différentes segmentations obtenues. La méthode estime l'approximation constante par morceaux de chaque profil, en accord avec la segmentation générale. La méthode BARD fournit la probabilité que les observations soient anormales, la segmentation dépend donc du seuil choisi pour la règle de décision. L'inconvénient majeur de la méthode est la forte dépendance au modèle des données. On peut toutefois supposer que pour une application bien connue comme l'hybridation génomique comparative, les choix des distributions des variables sont guidés par des a priori validés empiriquement. Un avantage de l'algorithme BARD est de fournir une classification des segments, normaux ou anormaux. Enfin, le *Bernoulli Detector* estime une segmentation par profil, où les ruptures partagées par plusieurs signaux sont synchronisées. Dans cette application, les positions des ruptures, données par $\hat{\boldsymbol{R}}_{MAP}$, importent plus que les distributions des probabilités des configurations de $\boldsymbol{P}$, puisque l'objectif est l'identification des portions d'ADN dont le nombre de copies est anormal. Contrairement à l'algorithme BARD, l'étiquetage des segments n'est pas réalisé,

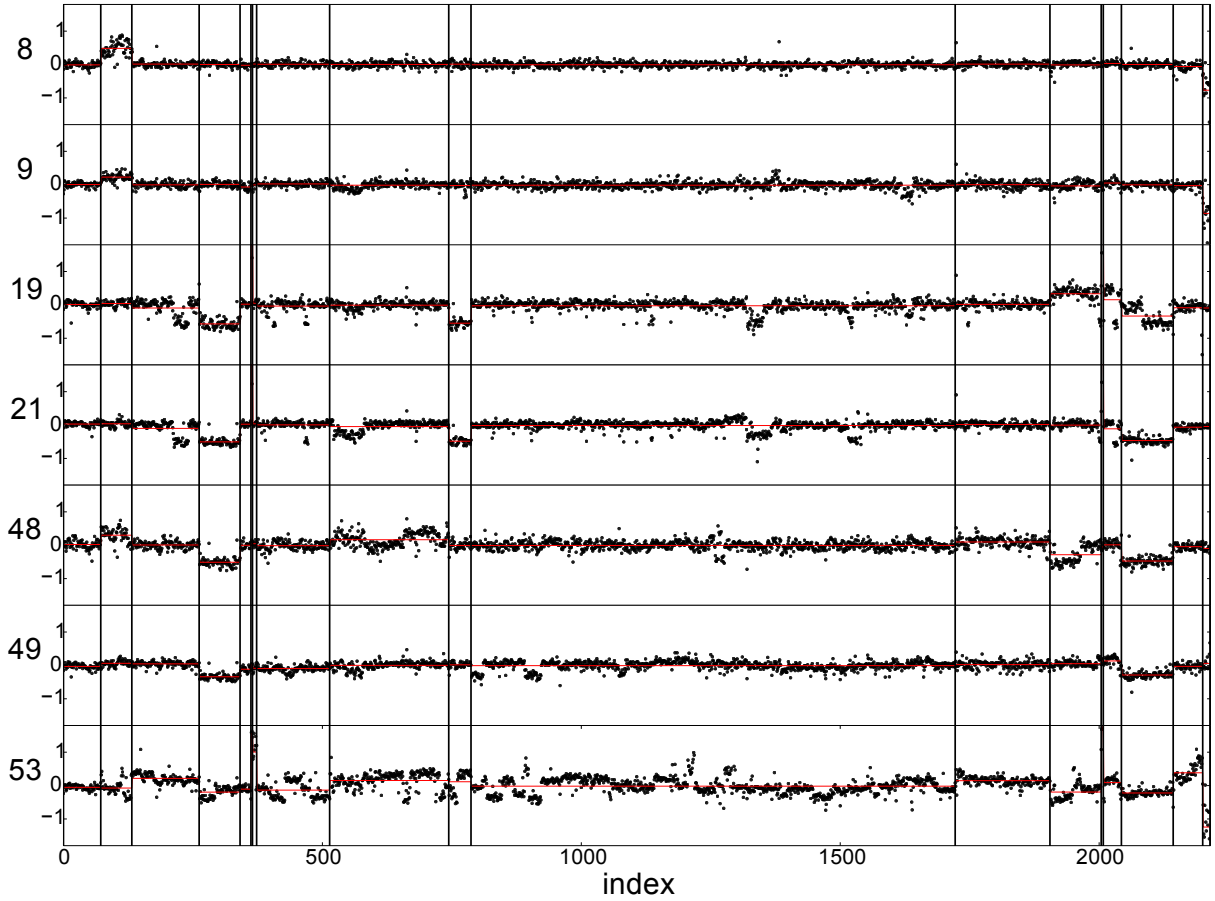


FIGURE 4.13 – Approximation constante par morceaux (en rouge) des données aCGH par la méthode du *group* fused LASSO. Les ruptures de la segmentation globale qui en est déduite sont représentées par des lignes verticales.

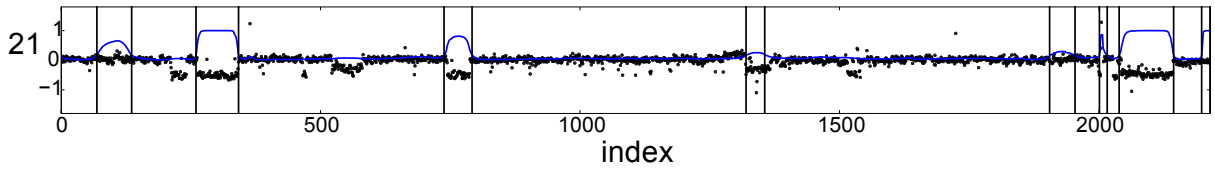


FIGURE 4.14 – Détection de segments anormaux dans des données aCGH, par la méthode BARD. La probabilité que les observations soient anormales est tracée en bleu, sur le profil 21, et les ruptures, détectées avec le seuil 0,17, sont matérialisées par des traits noirs verticaux.

et doit se faire ultérieurement. Le principal inconvénient de l'algorithme MultiBD est sa complexité. Pour traiter ces données aCGH, nous avons dû nous restreindre à groupe de 7 patients sur les 43 dont nous disposions, pour que le temps de calcul, de près de 70 heures,

demeure raisonnable. Il est cependant bien plus long qu'avec les autres méthodes. Cette approche est donc limitée en nombre de signaux m de la série temporelle. Elle conduit toutefois à des segmentations précises, globales et spécifiques des profils, insensibles aux valeurs aberrantes, sans avoir à choisir d'autre paramètre que le seuil α.

4.5 Discussion

En m'inspirant des travaux de [Dobigeon *et al.*, 2007a], j'ai proposé une extension du modèle du *Bernoulli Detector* pour le traitement de séries temporelles multivariées. Le paramètre $\boldsymbol{P}$ qui est introduit prend en compte les corrélations entre les ruptures. Des résultats sur le contrôle de la détection ont été montrés, valables pour le cas d'une unique rupture, quelle que soit la longueur de la série temporelle $n > 2$. L'étude du cas de ruptures multiples doit intégrer les dépendances entre les p-valeurs d'un signal, qui sont difficiles à exprimer. Nous avons néanmoins montré empiriquement que le *Bernoulli Detector* détecte plusieurs ruptures dans un même signal. L'algorithme MultiBD a été développé à partir de sa version univariée, UniBD, et effectue une approximation dans la mise à jour des p-valeurs au cours du mouvement de Gibbs. De nombreux résultats ont été présentés sur des données simulées et des données réelles.

Le *Bernoulli Detector* multivarié réalise l'estimation des ruptures communes à un groupe de signaux comme des ruptures spécifiques à chaque signal. La segmentation obtenue est alors adaptée à toutes les composantes spatiales, tout en tenant compte des événements impactant plusieurs signaux. Cette approche diffère de la plupart des méthodes de détection de ruptures dans un contexte multivarié. Dans les simulations et l'application sur la consommation électrique, nous avons proposé un modèle d'indépendance entre les ruptures, qui est représenté par un graphe. L'algorithme MultiBD estime $\boldsymbol{P}$ en accord avec cette représentation, cette dernière permettant aussi d'éliminer des configurations impossibles dans le système, réduisant alors la complexité de la résolution. L'apprentissage de la structure permet également de mieux détecter les ruptures communes dans un signal très bruité, par rapport à un traitement indépendant.

La limite majeure de MultiBD est son temps de calcul : la densité a posteriori (4.14) dépend linéairement de n et L, donc exponentiellement de m avec un a priori non informatif. Avec l'échantillonneur de Gibbs qui a été proposé, la complexité dépend exponentiellement du nombre de signaux m, dans le cas général non informatif. L'algorithme MultiBD conduit à de bons résultats tant que la dimension m de la série temporelle reste faible, par exemple moins de 7 signaux. D'autres pistes pour la résolution quand m devient grand sont toutefois envisageables. L'échantillonneur de Gibbs de l'algorithme 5 permet d'obtenir les colonnes de $\boldsymbol{R}$, mais nécessite de tester les L configurations. En échantillonnant les indicatrices $R_{j,i}$ individuellement, conditionnellement aux autres, le nombre d'opérations à réaliser à chaque étape est fortement réduit. Il faut toutefois s'assurer de la convergence de l'algorithme, qui sera probablement ralentie, et donc réaliser plus d'étapes MCMC. Il convient d'évaluer la stratégie la plus rapide selon les dimensions des données. Le traitement peut également être appliqué paire à paire sur les signaux. L'algorithme peut aussi être modifié en parallélisant

certaines étapes lors de l'échantillonnage d'une colonne de $\boldsymbol{R}$, ou bien la résolution peut être accélérée par une approche de type bayésien variationnel.

L'application du *Bernoulli Detector* sur les données aCGH a mis en avant l'aspect complémentaire de la méthode avec d'autres approches comme le *group* fused LASSO et la méthode BARD. Notre modèle fournit ainsi des segmentations à partir desquelles un expert peut établir un diagnostic individuel pour chaque patient, mais aussi analyser de manière globale les portions de l'ADN fréquemment en nombre anormal. L'étude du paramètre $\boldsymbol{P}$ peut permettre d'associer un patient à un groupe de patients de référence, partageant des anomalies. Ce vecteur des probabilités des configurations de ruptures apporte de l'information sur les liens entre les variables indicatrices d'un signal à l'autre. Dans les exemples de ce chapitre, nous sommes partis d'un graphe de dépendance entre les ruptures pour générer ou expliquer les séries temporelles à segmenter. La question se pose de savoir si cette structure de dépendance sous-jacente peut être reconstruite à partir des résultats du *Bernoulli Detector*. Dans le chapitre suivant, je propose un moyen de répondre à cette interrogation.

Chapitre 5

Estimation du graphe de dépendance

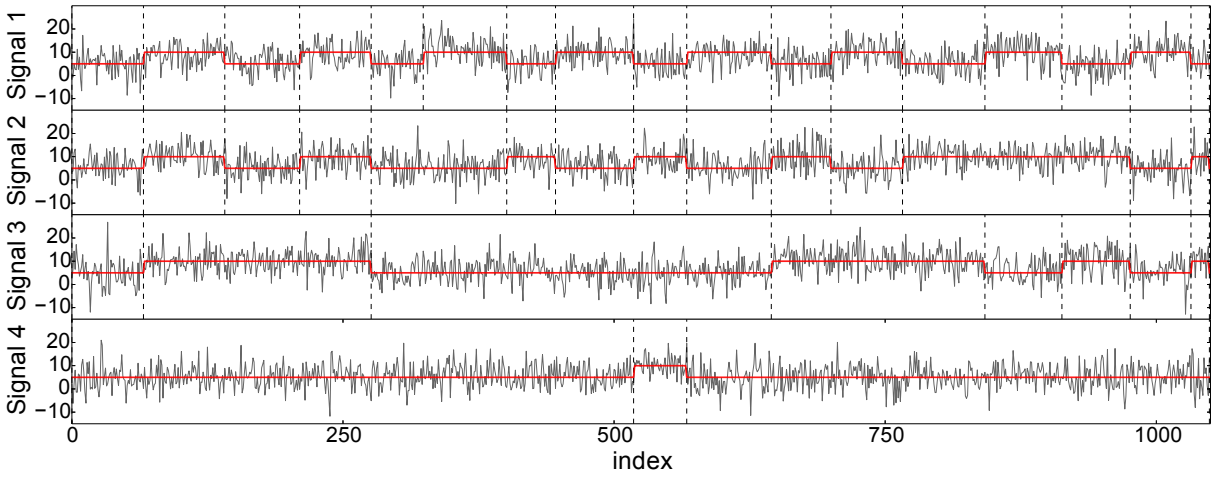
L'estimation conjointe des ruptures dans la série temporelle multivariée $\boldsymbol{X}$ est réalisée par le *Bernoulli Detector*, qui fait intervenir le paramètre $\boldsymbol{P}$. Ce terme correspond aux probabilités d'observer certains motifs dans les colonnes de $\boldsymbol{R}$; son introduction dans le modèle présente un intérêt quand il existe des relations de dépendance entre les ruptures d'un signal à l'autre. En effet, en l'absence de liens, le traitement indépendant des m signaux est plus pertinent que la segmentation conjointe. À l'inverse, lorsque les mêmes ruptures se retrouvent toutes dans l'ensemble des signaux, la recherche d'une segmentation unique est la plus adaptée, et la configuration la plus probable est $\epsilon_L = (1, \ldots, 1)'$, les autres étant négligeables. Pour les cas de figure intermédiaires, le modèle du *Bernoulli Detector* est approprié. Dans ce chapitre, j'expose comment le modèle d'indépendance entre les ruptures peut être en partie estimé, grâce à l'algorithme *Greedy Equivalent Search* (GES) qui réalise l'apprentissage de la structure de dépendance d'un réseau bayésien en explorant l'espace des classes d'équivalence de Markov [Meek, 1997, Auvray et Wehenkel, 2002, Chickering, 2003]. Cette approche exploite le formalisme des graphes présenté dans la partie 1.2 de l'état de l'art. Le problème est présenté dans la partie 5.1, où le réseau bayésien construit sur les variables indicatrices est présenté. Le principe de l'algorithme GES est exposé dans la partie 5.2, puis le lien avec les résultats issus de l'algorithme MultiBD est établi sous la forme de deux fonctions de score dans la partie 5.3. Dans la partie 5.4, la méthode est appliquée sur des exemples issus du traitement de séries temporelles par l'algorithme MultiBD et sur les données réelles du réseau électrique. Une discussion sur l'interprétation des résultats en 5.5 conclut ce chapitre.

5.1 Ruptures et modèle d'indépendance

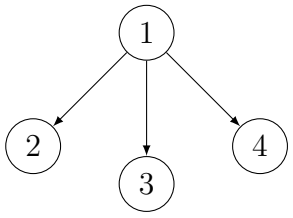
5.1.1 Définition du problème

Pour déterminer la structure de dépendance sous-jacente de la série temporelle $\boldsymbol{X}$, l'approche élémentaire consiste à évaluer les corrélations entre les signaux $\boldsymbol{X}_{1,\bullet}, \ldots \boldsymbol{X}_{m,\bullet}$ qui composent $\boldsymbol{X}$ et d'en déduire un graphe dont les sommets sont les variables X_j, $1 \leq j \leq m$

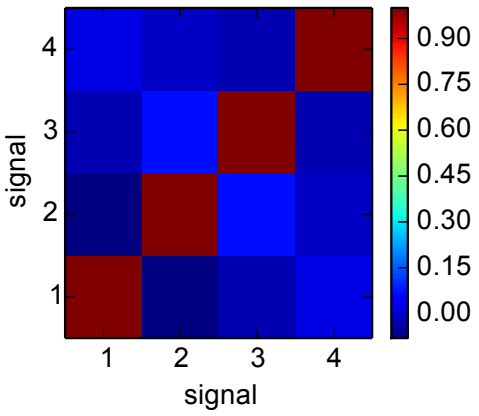
des différents signaux, et où les arêtes symbolisent un lien fort entre deux signaux. Or, si nous reprenons l'exemple de la série temporelle de la partie 4.4.1, tracée dans la figure 5.1a, et où les valeurs moyennes des segments sont indiquées en rouge, il est aisé de constater que la comparaison des valeurs prises par les variables X_j entre elles, pour $j \in \{1,2,3,4\}$, ne permet pas de retrouver la structure de dépendance de la figure 5.1b. Le calcul de la matrice de corrélation, donnée à la figure 5.1c, le confirme. De fait, les observations $x_{j,i}$ ont été générées indépendamment d'un signal à l'autre, conditionnellement aux ruptures (en pointillés noirs) qui délimitent les segments. Ces dernières se produisent sur les différents signaux conformément aux relations de dépendances décrites par le graphe 5.1b.



(a) Signaux de la série temporelle (en gris), emplacement des ruptures (en pointillés) et valeurs moyennes théoriques de chaque segment (en rouge).



(b) Graphe de dépendance



(c) Matrice de corrélation entre les signaux.

FIGURE 5.1 – Corrélations en 5.1c entre les signaux de la série temporelle 5.1a, dont les ruptures sont régies par le graphe 5.1b.

De manière générale, dans les séries temporelles que nous segmentons par le *Bernoulli*

Detector, nous supposons que les variables aléatoires $X_{j,i}$ sont i.i.d. sur chaque segment, et ainsi qu'elles sont indépendantes des variables des autres signaux, conditionnellement aux positions des ruptures. Par conséquent la mesure de corrélations entre les signaux n'est pas pertinente. En revanche des relations de dépendance existent bel et bien entre les ruptures, désignées par les variables aléatoires indicatrices $R_{j,i}$, $2 \leq i \leq n-1$, $1 \leq j \leq m$, prenant la valeur 1 lorsqu'une rupture se produit dans le signal j à l'instant i. En faisant l'hypothèse, comme pour construire le modèle du *Bernoulli Detector* multivarié, que si le signal j dépend du signal i, une rupture dans $\boldsymbol{X}_{i,\bullet}$ a une forte probabilité de se produire simultanément dans $\boldsymbol{X}_{j,\bullet}$, alors l'observation des valeurs prises par l'ensemble des variables indicatrices révèle des éléments de la structure de dépendance sous-jacente régissant le système. Les variables aléatoires indicatrices, binaires, sont notées $R_1, \ldots, R_m$ pour les signaux 1 à m, en faisant l'hypothèse qu'elles sont indépendantes en temps.

Notre objectif est d'inférer le graphe acyclique dirigé G représentant les relations de dépendance entre les variables aléatoires $R_1, \ldots, R_m$, à partir des résultats du *Bernoulli Detector*. Pour ce faire, nous adaptons la méthode GES de [Chickering, 2003] pour l'apprentissage de la structure d'un réseau bayésien. Celui-ci est défini par l'ensemble U des variables $R_1, \ldots, R_m$, dont l'ensemble $\mathcal{D}$ rassemble un nombre N d'observations. La méthode GES est basée sur une fonction de score et permet d'apprendre le modèle d'indépendance $M(U)$ qui corresponde le mieux aux données de $\mathcal{D}$. Cette procédure est réalisée efficacement sous certaines conditions, comme il est souligné dans [François, 2006, chapitre 9]. Notons que dans certains cas une structure de dépendance peut également régir les variables des signaux $\boldsymbol{X}_{j,\bullet}$, mais notre approche propose de ne s'intéresser qu'aux ruptures. Lorsque la matrice réelle des indicatrices $\boldsymbol{R}$ est connue, l'ensemble $\mathcal{D}$ est constitué des vecteurs colonnes de $\boldsymbol{R}$, et le graphe G décrivant les dépendances entre les indicatrices est appris directement à partir de ces données. En revanche, quand les positions réelles des ruptures sont inconnues, $\mathcal{D}$ doit être estimé. Deux moyens d'y parvenir grâce aux résultats de l'algorithme MultiBD sont présentés dans ce chapitre, dans la partie 5.3. La complexité du problème d'estimation du graphe à partir des ruptures est moindre que dans le cas de l'étude des variables $X_1, \ldots, X_m$ puisque les R_j sont des variables binaires. De plus cette stratégie est applicable quelle que soit la nature des signaux, à partir du moment où les hypothèses du *Bernoulli Detector* sont respectées et que l'algorithme MultiBD détecte bien les ruptures.

Les échantillons $(R_1, \ldots, R_m)$ de $\mathcal{D}$ sont obtenus à partir de la matrice $\boldsymbol{R}$ des indicatrices des ruptures dans $\boldsymbol{X}$. Dans ce chapitre, nous ne tenons plus compte de l'indice temporel i qui positionne les ruptures dans la série $\boldsymbol{X}$, et considérons donc que les variables aléatoires R_j, $1 \leq j \leq m$, sont indépendantes en temps. Cette approximation permet de s'affranchir des dépendances temporelles entre les variables, qui nécessiteraient de prendre en compte les positions des ruptures les unes par rapport aux autres dans $\boldsymbol{X}$, rendant le problème de l'estimation du modèle d'indépendance trop complexe. L'ensemble $\mathcal{D}$ est donc simplement constitué des différentes configurations prises par les variables binaires $R_1, \ldots, R_m$, indépendamment de leur position, où R_j est associée au signal j, et dont $r_{j,i}$ est une réalisation à l'instant i.

Le principe de l'apprentissage de la structure d'un réseau bayésien à partir de l'en-

semble d'observations $\mathcal{D}$ repose sur deux hypothèses. La première, la *fidélité*, suppose que le modèle d'indépendance conditionnelle des données puisse être décrit par un graphe G, autrement dit qu'il existe une représentation graphique qui représente toutes les relations d'indépendances du modèle, et uniquement celles-ci. Ce graphe est une carte parfaite du modèle (voir la partie 1.2.1). La seconde hypothèse est la *suffisance causale*, et suppose que l'ensemble des variables U suffit pour représenter toutes les relations d'indépendances conditionnelles observées dans les données $\mathcal{D}$. Ceci implique notamment, dans un contexte de relations causales, que toute cause commune à plusieurs variables appartient à U, ou bien est figée dans le même état pour toutes les observations de $\mathcal{D}$.

5.1.2 Espace de recherche

L'objectif de l'apprentissage de la structure de dépendance d'un réseau bayésien est de déterminer la structure de dépendance qui décrive le mieux possible les données de l'ensemble $\mathcal{D}$. Ce problème est NP-difficile. L'espace $\mathcal{B}$ des graphes orientés sans circuit a une taille superexponentielle en fonction du nombre de sommets. [Robinson, 1977] a démontré que le nombre de graphes à m nœuds peut être obtenu par la formule récursive suivante :

$$N_G(m) = \sum_{i=1}^{m} (-1)^{i+1} \binom{m}{i} 2^{i(m-1)} N_G(m-i) \quad \text{pour } m > 1. \tag{5.1}$$

Le nombre de DAG de plus petites tailles sont $N_G(2) = 3$, $N_G(3) = 25$, $N_G(4) = 543$, $N_G(5) = 29281$, $N_G(6) = 3781503$, et ainsi de suite. Une approche exhaustive consistant à tester tous les DAG possibles de l'espace $\mathcal{B}$ s'avère donc rapidement impraticable. Des heuristiques ont été proposées afin de réduire l'espace de recherche, par exemple aux arbres [Chow et Liu, 1968].

L'algorithme GES, pour *Greedy Equivalent Search* ou Recherche Gloutonne d'Équivalences, réalise lui l'apprentissage de la structure d'un réseau bayésien à partir des données en explorant l'espace connecté $\mathcal{C}$ des classes d'équivalences de Markov [Chickering, 2003, Auvray et Wehenkel, 2002]. Comme il a été expliqué dans le chapitre 1, il n'y a pas forcément unicité du graphe G représentant la factorisation récursive de la loi de probabilité $\mathcal{P}$, et plusieurs orientations conviennent à certaines arêtes du DAG décrivant les données $\mathcal{D}$. L'ensemble de ces graphes compatibles avec $\mathcal{P}$ constitue la classe d'équivalence de Markov de $\mathcal{P}$, et est représentée par un unique graphe essentiel, partiellement orienté, dont les arêtes dirigées sont communes à tous les graphes de la classe. L'espace de recherche $\mathcal{C}$ de ces graphes essentiels est donc de plus petite taille que $\mathcal{B}$, et la méthode GES évite de tester des modèles équivalents.

Bien qu'il n'existe pas d'expression générale pour déterminer le nombre exact de classes d'équivalences de Markov dans $\mathcal{C}$ en fonction du nombre de sommets, il est établi que celui-ci augmente de façon superexponentielle avec le nombre de variables, comme $\mathcal{B}$. L'exploration exhaustive de l'espace est donc irréalisable en pratique [1]. Plusieurs procédures ont été

1. Pour 8 variables, il existe plus de 700 milliards de DAG et plus de 200 milliards de classes d'équivalence.

proposées pour estimer le nombre de classes. [Gillispie et Perlman, 2001] présentent un algorithme énumérant toutes les classes jusqu'à 10 nœuds, et mettent en avant le fait que nombre de classes ne contiennent qu'un seul élément. D'après leurs résultats, le rapport du nombre de classes d'équivalences sur le nombre de DAG tend vers une asymptote autour de 0,267 quand le nombre de sommets augmente. Plus récemment, une approche par MCMC a été proposée dans [Sonntag *et al.*, 2015] afin d'estimer le ratio pour des graphes de plus grandes tailles. Le gain en complexité en passant du parcours de l'espace $\mathcal{B}$ au parcours de l'espace $\mathcal{C}$ s'avère limité, en-dehors des graphes contenant moins d'une dizaine de sommets : les nombres des classes jusqu'à 5 nœuds sont $N_C(2) = 2$, $N_C(3) = 11$, $N_C(4) = 185$, $N_C(5) = 8782$. Dans notre contexte, l'emploi du *Bernoulli Detector* pour estimer les variables indicatrices contraint au traitement d'un nombre raisonnable de signaux dans la série temporelle initiale $\boldsymbol{X}$. Le nombre de sommets de la structure de dépendance est alors faible, et l'exploitation de l'espace $\mathcal{C}$ des classes d'équivalences est justifié dans le cadre de notre application.

La méthode GES ne réalise pas une exploration exhaustive systématique de tout l'espace des classes d'équivalences de Markov. Grâce à l'emploi d'une fonction de score, seul un voisinage limité autour de la meilleure classe locale est exploré à chaque étape, et il est raisonnable de supposer que l'algorithme s'arrête avant d'atteindre la classe des graphes complets, les modèles étudiés étant en général simples, c'est-à-dire avec relativement peu d'arêtes. Ce type d'approche est d'autant plus efficace qu'il y a peu de connexions entre les états. Comme il est signalé dans [Chickering et Meek, 2002], le nombre de modèles adjacents dans $\mathcal{C}$ est relativement faible pour des modèles à peu de variables.

L'estimation des dépendances sous-jacentes dans la série temporelle $\boldsymbol{X}$ est réalisée sur les variables indicatrices des ruptures $R_1, \ldots, R_m$, dont l'ensemble $\mathcal{D}$ rassemble N observations. Ces dépendances sont représentées par le graphe G d'un réseau bayésien, appris à partir de $\mathcal{D}$. Comme plusieurs DAG sont compatibles avec la factorisation d'une loi de probabilité jointe $\mathcal{P}$, il n'est pas toujours possible de déterminer l'orientation de toutes les arêtes de G. Le graphe résultant est alors partiellement orienté et représente la classe d'équivalence de Markov associée à G. L'apprentissage est effectué par l'algorithme GES, qui réalise l'exploration de l'espace $\mathcal{C}$ des classes d'équivalences de Markov. Le principe de cette méthode est présenté dans la partie suivante.

5.2 Apprentissage du modèle par l'algorithme GES

L'apprentissage par l'algorithme GES [Meek, 1997, Chickering, 2003] est une méthode graphique qui fait partie de la catégorie des algorithmes basés sur une fonction de score. Elle consiste à explorer une partie de l'espace $\mathcal{C}$ grâce à deux opérateurs et à déterminer localement puis globalement la meilleure classe parmi celles testées grâce à une fonction de score. Chickering montre dans [Chickering, 2003] que la structure optimale peut être atteinte par une recherche gloutonne dans l'espace $\mathcal{C}$ consistant en une phase d'ajout d'arc suivie d'une phase de suppression. Le principe repose sur la conjoncture de Meek [Meek, 1997], démontrée dans [Chickering, 2003], dont la formulation est donnée par le théorème

5.2.1. D'après cette conjoncture, il est possible de passer d'un DAG G_2 à un graphe G_1 en un nombre fini d'opérations élémentaires, lorsque toutes les relations d'indépendance qui peuvent être déduites de la représentation graphique G_2 sont également présentes dans le modèle d'indépendance représenté par G_1 [2]. On dit alors que G_1 est une carte d'indépendance de G_2 (voir le chapitre 1, page 41). Lors des modifications apportées à G_2, G_1 demeure une carte d'indépendance de G_2. Un exemple d'application de la conjoncture de Meek est montré dans la figure 5.2.

Théorème 5.2.1 (Conjoncture de Meek)**.** *Soient deux graphes G_1 et G_2 tels que G_1 est une carte d'indépendance de G_2. Alors il existe une séquence finie d'ajouts et de retournements d'arcs appliqués à G_2, tels qu'après chaque changement G_2 demeure un DAG et G_1 une carte d'indépendance de G_2, et qu'après ces changements $G_1 = G_2$.*

Dans [Chickering et Meek, 2002], les auteurs montrent que cette séquence est composée d'au plus $r + 2a$ opérations de retournements et d'ajouts d'arcs, où r est le nombre d'arcs de G_1 d'orientation contraire dans G_2 et où a est le nombre d'arêtes de G_1 qui n'existent pas dans G_2. Ainsi, dans l'exemple de la figure 5.2, la transition de G_2 vers G_1 est réalisée en 2 retournements et 1 ajout d'arc, et le nombre de ces étapes et bien inférieur ou égal à $r + 2a = 4$, avec $r = 2$ arcs d'orientations contraires et $a = 1$ arête manquante. L'article fournit des garanties sur l'optimalité de l'algorithme appliqué à de grands échantillons pour estimer les relations d'indépendances conditionnelles. Ces conditions concernent le choix d'un score consistant asymptotiquement, c'est-à-dire favorisant le modèle le plus simple qui est en adéquation avec les données.

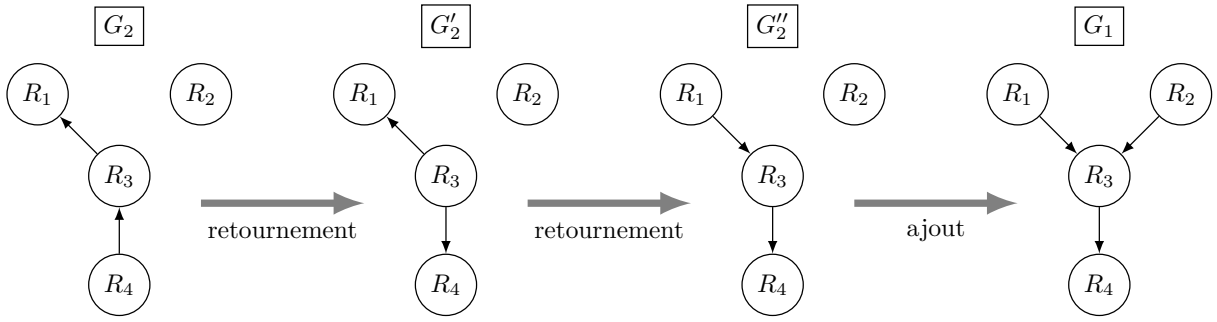


FIGURE 5.2 – Exemple d'application de la conjoncture de Meek pour passer du DAG G_2 au DAG G_1, par deux retournements d'arêtes et un ajout. Les relations d'indépendance conditionnelles associées à G_1 sont vérifiées par le DAG G_2 et les DAG intermédiaires G'_2 et G''_2 : $R_1 \perp\!\!\!\perp R_2$, $R_1 \perp\!\!\!\perp R_4 | R_3$ et $R_2 \perp\!\!\!\perp R_4 | R_3$.

Le principe général de la méthode GES, détaillé dans cette partie, consiste à explorer l'espace $\mathcal{C}$ des classes d'équivalences de Markov à la recherche du graphe partiellement orienté H, pour lequel l'un des DAG de la classe d'équivalence maximise une fonction de score S calculée sur les données $\mathcal{D}$. Elle repose donc sur les transitions entre le graphe

2. Une relation d'indépendance conditionnelle dans la loi jointe $\mathcal{P}$ est équivalente à une d-séparation dans le graphe G compatible avec $\mathcal{P}$.

partiellement orienté H et le DAG G, sur les étapes d'exploration de l'espace $\mathcal{C}$, par ajout puis par suppression d'arc, et sur une fonction de score. Ces trois éléments sont présentés dans cette partie.

5.2.1 DAG et CPDAG

La méthode GES parcourt l'espace $\mathcal{C}$, par conséquent la solution du problème d'apprentissage de la structure de dépendance pour les variables $R_1, \ldots, R_m$ n'est pas le graphe dirigé sans circuit G, mais la classe d'équivalence de Markov à laquelle il appartient, notée C_G. Elle est représentée par un graphe essentiel H_G (ou CPDAG), dont les arcs sont non réversibles et communs à tous les DAG de la classe. Le graphe essentiel possède le même squelette que le DAG G. Ses seules arêtes orientées sont celles des V-structures, c'est-à-dire les arêtes dirigées vers un même nœud, et dont les nœuds d'origine ne sont pas adjacents (d'où la forme d'un V entre 3 nœuds), ainsi que les arêtes pour lesquelles une orientation contraire entraîne l'apparition d'un cycle ou d'une V-structure absente du DAG original. Les arêtes non dirigées de H_G remplissent elles la condition suivante : il existe au moins une orientation qui ne crée ni de circuit, ni de nouvelle V-structure. Le graphe recherché G comme solution du problème d'apprentissage de la structure de dépendance est donc l'une des extensions consistantes de H_G en un graphe complètement orienté sans circuit.

Lors de la procédure GES, les différents graphes acycliques partiellement dirigés (PDAG) constituant le voisinage d'un CPDAG, augmenté ou diminué d'un arc, sont comparés. Comme nous l'expliquons dans la partie 5.2.4, l'algorithme s'appuie sur l'application d'une fonction de score mesurant l'adéquation d'un graphe G avec les données. Pour calculer le score, il est donc nécessaire de sélectionner un DAG complètement dirigé, appartenant à la même classe d'équivalence de Markov que le graphe partiellement dirigé, noté H. Si le PDAG H est extensible, c'est-à-dire si on peut en déduire un DAG G en proposant une orientation pour les arêtes non dirigées sans introduire de nouvelle V-structure ni de cycle orienté, alors la fonction de score est calculée sur G, pour la solution H. En revanche si H n'admet pas d'extension, cette solution est éliminée. La figure 5.3 donne un exemple d'un CPDAG H, comportant la V-structure $R_2 \rightarrow R_3 \leftarrow R_4$, qui est extensible. Les deux extensions possibles, les DAG G_1 et G_2, sont représentées dans la figure 5.4. Ces deux solutions correspondent aux deux orientations possibles de l'arête $R_1 - R_4$ non dirigée dans H. Comme H est extensible, la fonction de score peut être calculée sur l'une des deux extensions G_1 ou G_2. La figure 5.5 donne un exemple de deux PDAG qui ne sont pas extensibles. En effet, en orientant les arêtes indéterminées, on observe soit l'apparition d'un cycle, et le graphe résultant n'est alors pas un DAG, soit une nouvelle V-structure qui n'est pas présente dans H, et que ne peut donc pas appartenir à la même classe d'équivalence de Markov que le PDAG d'origine. Ce type de solution est éliminé au cours de l'algorithme GES.

En pratique, pour déterminer pour chaque PDAG H extensible un DAG sur lequel calculer le score, et pour éliminer les PDAG non extensibles, la méthode de [Dor et Tarsi, 1992] est employée. La procédure d'extension de H, réalisée en temps polynomial, est détaillée dans l'algorithme 6. Elle oriente récursivement les arêtes de sous-graphes en les

dirigeant vers les nœuds R_i dont les propriétés sont les suivantes :

1. R_i est un puits, c'est-à-dire qu'aucune arête n'est dirigée de R_i vers l'un des sommets adjacents composant l'ensemble A_i ;
2. les sommets voisins pour lesquels l'arête avec R_i n'est pas orientée sont adjacents à tous les voisins de R_i.

Algorithme 6 : $CPDAGtoDAG(H)$, extension du PDAG H de [Dor et Tarsi, 1992].

Données : PDAG $H = (U_H, E_H)$
Résultat : DAG G s'il existe, *Faux* sinon.
initialiser le sous-graphe $G' = H$ et le graphe $G = H$
tant que $E_{G'} \neq \varnothing$ **faire**
  sélectionner un sommet R_i qui possède les propriétés 1 et 2
  si *un tel sommet n'existe pas* **alors**
    fin de l'algorithme, renvoyer *Faux*
  sinon
    pour $R_j \in A_i$ *tel que* $R_i - R_j$ **faire**
      orienter $R_j \rightarrow R_i$ dans G
      retirer (R_i, R_j) de $E_{G'}$
    fin
  fin
fin

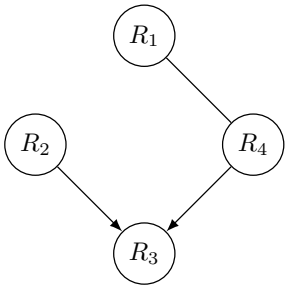
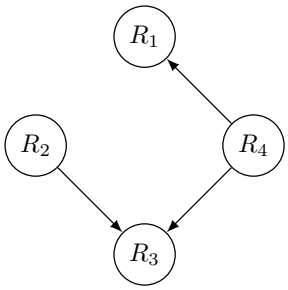


FIGURE 5.3 – Exemple de CPDAG H comportant une V-structure

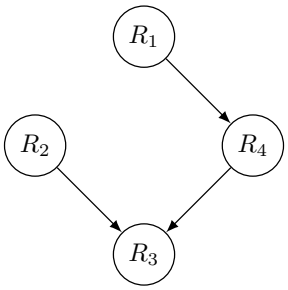
Après l'extension des PDAG extensibles en DAG consistants, le meilleur modèle $G^\star$ qui maximise la fonction de score est sélectionné. Le CPDAG $H_G^\star$ représentant la classe d'équivalence de $G^\star$ est déduit du DAG. Pour ce faire, [Chickering, 2003] propose un algorithme $DAGtoCPDAG$ qui étiquette les arêtes de $G^\star$ en «forcée» ou «réversible», indiquant si l'orientation est commune à tous les éléments de la classe $H_G^\star$ ou non.

5.2.2 Phase d'insertion d'arc

Le point de départ de l'algorithme est la classe C_0 du graphe sans arêtes, où toutes les variables sont indépendantes. L'intérêt de commencer avec C_0 est que le nombre de

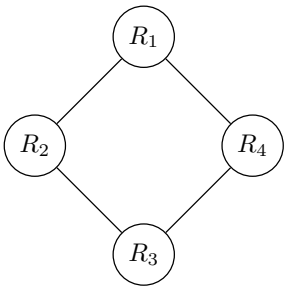


(a) DAG G_1

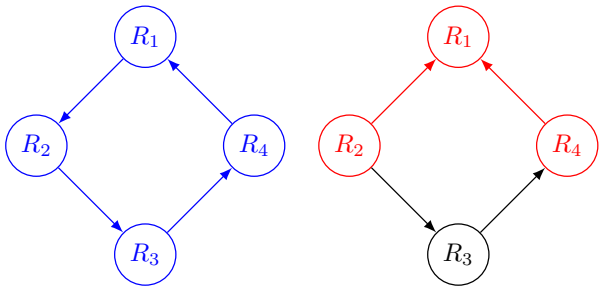


(b) DAG G_2

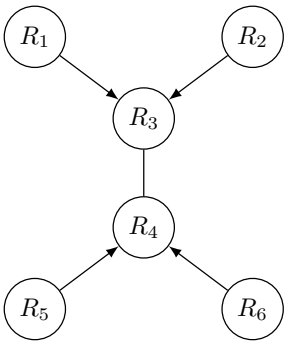
FIGURE 5.4 – Les deux extensions possibles du CPDAG H de la figure 5.3.



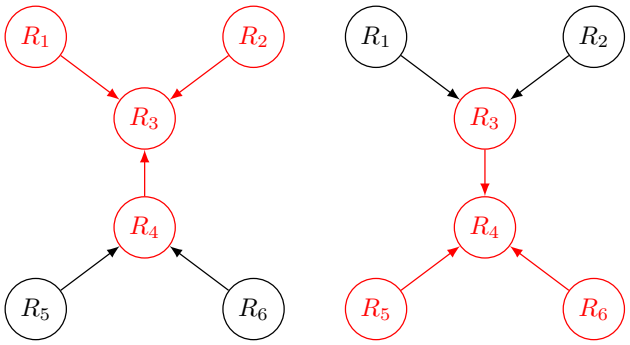
(a) PDAG H_1



(b) Graphes complètement orientés issus de H_1.



(c) PDAG H_2



(d) Graphes complètement orientés issus de H_2.

FIGURE 5.5 – Exemple de PDAG non extensibles 5.5a et 5.5c et de graphes complètement orientés qui en sont issus, respectivement en 5.5b et en 5.5d. Quelles que soient les orientations choisies, les graphes complètement orientés résultants comportent soit une V-structure qui n'existe pas dans le PDAG d'origine (rouge), soit ou un cycle dirigé (bleu).

paramètres est minimal, et que les modèles les plus simples sont testés en premier. En supposant que le modèle optimal comporte peu de paramètres, cette phase s'interrompt avant d'atteindre la classe des graphes complets, limitant ainsi le coût calculatoire. Les arêtes sont ajoutées une par une jusqu'à maximisation de la fonction de score par l'action d'un opérateur permettant de passer du CPDAG H_k contenant k arêtes à sa limite d'inclusion supérieure $\mathcal{V}^+(H_k)$, constituée des PDAG de $k+1$ arêtes qui ne diffèrent de H_k que par l'ajout d'un arc et éventuellement une orientation d'un arc de H_k [Auvray et Wehenkel,

2002]. L'opérateur d'insertion, défini en 5.2.1, s'applique directement sur le CPDAG H_k pour générer $\mathcal{V}^+(H_k)$. La classe d'équivalence obtenue à l'issue de cette phase contient la distribution générative $\mathcal{P}$ [Chickering, 2003].

Définition 5.2.1 (Opérateur d'insertion d'arc $Insert(H,R_i,R_j,A_j)$)**.** *Pour deux nœuds R_i et R_j non adjacents dans le CPDAG essentiel H, et pour tout sous-ensemble A_j des voisins de R_j non adjacents à R_i, l'opérateur $Insert(H,R_i,R_j,A_j)$ modifie H de deux façons :*

1. *il insère l'arc $R_i \rightarrow R_j$;*
2. *pour tout $R_k \in A_j$, il oriente l'arête non dirigée avec R_j en $R_k \rightarrow R_j$.*

5.2.3 Phase de suppression d'arc

Lors de la seconde phase, une arête est supprimée, et l'opération répétée jusqu'à obtention du score maximum, pour lequel le graphe essentiel $H^\star$ correspond à la solution optimale. Ainsi, lorsque le nombre d'échantillons est suffisamment grand, la classe d'équivalence de $H^\star$ est une carte de dépendance minimale et parfaite de la distribution générative $\mathcal{P}$. À chaque étape, les éléments de la limite d'inclusion inférieure $\mathcal{V}^-(H_k)$ sont testés. Dans ce voisinage, la carte d'indépendance est plus petite. L'opérateur de suppression d'arête est défini en 5.2.2.

Définition 5.2.2 (Opérateur de suppression d'arc $Delete(H,R_i,R_j,A_j)$)**.** *Pour deux nœuds R_i et R_j adjacents dans le CPDAG essentiel H par $R_i - R_j$ ou $R_i \rightarrow R_j$, et pour tout sous-ensemble $A_{j,i}$ des voisins de R_j qui sont adjacents à R_i, l'opérateur $Delete(H,R_i,R_j,A_j)$ modifie H en supprimant l'arête entre R_i et R_j, et, pour tout $R_k \in A_{j,i}$:*

1. *en orientant l'arête entre R_j et R_k en $R_j \rightarrow R_k$;*
2. *en orientant tout arête non dirigée entre R_i et R_k en $R_i \rightarrow R_k$.*

Les graphes résultants sont évalués par une fonction de score.

5.2.4 Fonction de score

La fonction de score mesure l'adéquation du modèle testé avec les données disponibles, et permet de guider la recherche de la structure vers une classe d'équivalence contenant $\mathcal{P}$ lors de la phase d'insertion, puis vers le modèle optimal lors de la phase de suppression. Le score obtenu sur les données $\mathcal{D}$ pour le graphe G est noté $S(G,\mathcal{D})$. Un grand nombre de formulations ont été proposées par différents auteurs. La fonction permet de choisir le meilleur modèle pour un nombre d'arêtes donné ainsi que le nombre d'arêtes du graphe. Dans la phase d'insertion, le score augmente lorsqu'une arête est ajoutée si celle-ci élimine une contrainte d'indépendance conditionnelle qui n'est pas compatible avec $\mathcal{P}$ et lors de la phase de suppression, le score augmente lorsque l'arête retirée introduit une contrainte d'indépendance compatible avec $\mathcal{P}$.

Pour construire la fonction de score, une règle générale est de favoriser les modèles les plus simples qui représentent la distribution générative avec le moins de paramètres, selon

le principe du rasoir d'Occam. Dans [Chickering et Meek, 2002], les auteurs montrent que la solution de la méthode GES est un modèle optimal si la fonction de score appliquée parvient au score maximal pour le modèle représentant la distribution générative des données contenant le moins de paramètres, lorsque le nombre d'échantillons devient grand.

L'algorithme parcourt donc l'espace $\mathcal{C}$ à la recherche de maxima locaux de la fonction de score, et conserve la meilleure solution. Pour limiter le nombre de calculs nécessaires, il est judicieux d'employer un score *décomposable*, s'écrivant comme la somme de termes dépendant uniquement d'un nœud R_i et de ses parents $\mathrm{Pa}(R_i)^G$ dans le DAG G :

$$S(G,\mathcal{D}) = \sum_{i=1}^{m} s(R_i, \mathrm{Pa}(R_i)^G). \tag{5.2}$$

L'avantage de ces scores est qu'ils ne nécessitent qu'une modification partielle de certains termes $s(R_i, \mathrm{Pa}(R_i)^G)$ pour passer d'un graphe G à un graphe du voisinage de la classe de G. Le score est ainsi mis à jour simplement lors de l'application des opérateurs d'insertion ou de suppression.

Une autre catégorie de scores intéressants est celle des scores *équivalents*, qui prennent la même valeur pour des modèles équivalents [Chickering, 1995]. Dans ce cas, pour deux DAG G_1 et G_2 de la même classe d'équivalence de Markov, $S(G_1,\mathcal{D}) = S(G_2,\mathcal{D})$. La formulation générale intègre deux termes : l'un qui mesure la vraisemblance du modèle testé G avec les données $\mathcal{D}$ et l'autre qui représente la complexité du modèle.

L'un des scores décomposables et équivalents proposé pour la méthode GES dans la littérature est le score BIC, de [Schwarz *et al.*, 1978]. Pour évaluer le DAG G par rapport aux observations $\mathcal{D}$, sa formule générale s'écrit

$$S_{BIC}(G,\mathcal{D}) = \log L(\mathcal{D}|G,\theta^{MV}) - \frac{1}{2}\dim(G)\log N, \tag{5.3}$$

où θ^{MV} est l'ensemble des *paramètres* du modèle G qui maximisent la vraisemblance, $\dim(G)$ est la *dimension* du graphe, et N est le nombre d'échantillons de $\mathcal{D}$, sur lesquels est estimée la vraisemblance. La dimension du graphe est donnée par

$$\dim(G) = \sum_{i=1}^{m} \dim(R_i,G) = \sum_{i=1}^{m} (s_i - 1)q_i, \tag{5.4}$$

où s_i est le nombre d'états possibles que peut prendre la variable R_i et où q_i est le nombre de configurations possibles que peut prendre l'ensemble des parents de R_i. Les paramètres, notés θ, correspondent aux différentes probabilités conditionnelles entre les variables, utilisées dans le graphe. Pour le modèle G et les échantillons 1 à N, la vraisemblance s'écrit :

$$L(\mathcal{D}|\theta) = \prod_{h=1}^{N} P(R_1 = r_{k_1}^{(h)}, \ldots, R_m = r_{k_m}^{(h)}|\theta). \tag{5.5}$$

La loi de probabilité jointe sur les variables $R_1,\ldots,R_m$ Markov relative à G se décompose en un produit des probabilités conditionnelles de chaque variable R_i sachant ses parents

immédiats dans G, $\mathrm{Pa}(R_i)$, et des probabilités marginales des variables sans parent dans G. L'expression de la vraisemblance devient donc :

$$L(\mathcal{D}|\theta) = \prod_{h=1}^{N} \prod_{i=1}^{m} P(R_i = r_{k_i}^{(h)} | \, \mathrm{Pa}(R_i) = r_j^{(h)}, \theta) \tag{5.6}$$

$$= \prod_{i=1}^{m} \prod_{h=1}^{N} \theta_{i,j(h),k(h)} \tag{5.7}$$

où i est l'indice de la variable R_i, prenant la valeur r_k dans l'état k, où les parents $\mathrm{Pa}(R_i)$ de R_i prennent la valeur r_j dans l'état j, et où le paramètre θ_{ijk} est la probabilité d'observer $R_i = r_k$ conditionnellement à $\mathrm{Pa}(R_i) = r_j$. Finalement, en introduisant N_{ijk} le nombre de fois où cette configuration est observée dans $\mathcal{D}$,

$$L(\mathcal{D}|\theta) = \prod_{i=1}^{m} \prod_{j=1}^{q_i} \prod_{k=1}^{s_i} \theta_{ijk}^{N_{ijk}} \tag{5.8}$$

où s_i est le nombre d'états possibles que peut prendre R_i et où q_i est le nombre de configurations possibles pour ses parents. Comme nous travaillons avec des données binaires, $s_i = 2$ pour tout $1 \leq i \leq m$ et $q_i = 2^{n(\mathrm{Pa}(R_i))}$ avec $n(\mathrm{Pa}(R_i))$ le nombre de parents de R_i dans G. Les paramètres $\theta_{i,j,k}$ étant inconnus, ils sont remplacés dans (5.8) par leurs estimateurs du maximum de vraisemblance (voir [Naïm *et al.*, 2011, chapitre 6.1, p.118]) :

$$\hat{\theta}_{ijk}^{MV} = \frac{N_{ijk}}{\sum_{k=1}^{s_i} N_{ijk}}. \tag{5.9}$$

Pour un nombre raisonnable de variables, et comme celles-ci sont binaires, les probabilités conditionnelles empiriques sont aisément calculables pour chaque nœud et ses parents. Il existe d'autres fonctions de score que le score BIC, comme le critère *Bayesian Dirichlet* et ses variantes [Buntine, 1991, Cooper et Herskovits, 1992, Heckerman *et al.*, 1995], ou la longueur de description minimale [Bouckaert, 1993, Lam et Bacchus, 1994, De Campos, 2006], dont une revue est réalisée dans [Naïm *et al.*, 2011, chapitre 6.2.5 p.144].

5.2.5 Algorithme

Les principales étapes de la méthode d'apprentissage de la structure du graphe à partir des données du *Bernoulli Detector* sont résumées dans les algorithmes 7 et 8 pour les phases d'insertion et de suppression d'arc respectivement, quelle que soit la fonction de score choisie. Lorsque le score est décomposable, le calcul de $S(G_{k-1}^{+},\mathcal{D})$ est simplement une mise à jour de S_{k-1} pour l'insertion d'une arête $R_i - R_j$ lors de la phase d'insertion, et le calcul de $S(G_{k+1}^{-},\mathcal{D})$ est une mise à jour de S_{k+1} pour la suppression d'une arête $R_i - R_j$ lors de la phase de suppression.

Un exemple de déroulement de la méthode GES sur 4 variables R_1, R_2, R_3 et R_4 est illustré par les figures 5.6 à 5.13, correspondant aux étapes d'insertion puis de suppression d'arêtes par les algorithmes 7 et 8. Dans ces figures, la colonne de gauche rassemble les

Algorithme 7 : GES sur les résultats de MultiBD, insertion d'arc

Données : Résultats du *Bernoulli Detector*
Résultat : Classe d'équivalence de Markov H
générer le jeu de données $\mathcal{D}$ à partir des échantillons $(R_1,\ldots,R_m)$
classe d'équivalence sans arête : H_0
DAG sans arêtes : G_0
ensemble des arêtes non orientées $E = \varnothing$
score : $S_{old} = S(G_0,\mathcal{D})$
$S = S_{old}$, $k = 0$
tant que *(*$S \geq S_{old}$*) et (*$k+1 \leq \frac{m(m-1)}{2}$*)* **faire**
  $k = k+1$, $S_{max} = -\infty$
  pour $(R_i,R_j) \notin E$ **faire**
    $A_j = \{R_l | (R_l,R_j) \in E, (R_l,R_i) \notin E\}$
    insertion : $H^+_{k-1} = Insert(H_{k-1},R_i,R_j,A_j)$, défini en 5.2.1
    extension : $CPDAGtoDAG(H^+_{k-1})$, algorithme 6
    si $CPDAGtoDAG(H^+_{k-1}) = G^+_{k-1}$ **alors**
      score : $S(G^+_{k-1},\mathcal{D})$
      si $S(G^+_{k-1},\mathcal{D}) > S_{max}$ **alors**
        $S_{max} = S(G^+_{k-1},\mathcal{D})$
        $H_k = DAGtoCPDAG(G^+_{k-1})$ (voir [Chickering, 2003, Annexe C])
      fin
    fin
  fin
  $S_{old} = S$, $S = S_{max}$
  si $S \geq S_{old}$ **alors**
    $H = H_k$
    ajout de la nouvelle arête (R_i,R_j) de H dans E
    $k_{max} = k$
  fin
fin

CPDAG des classes d'équivalence explorés à cette étape, et la colonne de droite liste les DAG sur lesquels sont calculés les scores, chaque DAG correspondant à un élément de la classe d'équivalence de Markov représentée à sa gauche. Ces DAG sont obtenus par extension du CPDAG à leur gauche par application de l'algorithme 6. Pour chaque étape à un nombre d'arêtes donné, le DAG qui maximise localement la fonction de score est tracé en rouge. Pour réaliser l'apprentissage de la structure du réseau bayésien, on dispose d'un ensemble $\mathcal{D}$ de N observations.

L'algorithme GES débute par la phase d'insertion d'arc. La première étape consiste à tester l'unique classe d'équivalence sans arêtes. Le CPDAG correspondant, H_0, est représenté à la figure 5.6a, ainsi que l'unique graphe pour cette classe d'équivalence de Markov, G_0, à la figure 5.6b. Le score initial $S(G_0,\mathcal{D})$ est calculé sur ce DAG, par exemple avec la

Algorithme 8 : GES sur les résultats de MultiBD (suite), suppression d'arc

$S = S_{old}$, $k = k_{max}$
tant que *($S \geq S_{old}$) et ($k \geq 0$)* **faire**
  $k = k - 1$, $S_{max} = -\infty$
  pour $(R_i,R_j) \in E$ **faire**
    $A_j = \{R_l | (R_l,R_j) \in E, (R_l,R_i) \in E\}$
    suppression : $H^-_{k+1} = Delete(H_{k+1},R_i,R_j,A_j)$, défini en 5.2.2
    extension : $CPDAGtoDAG(H^-_{k+1})$, algorithme 6
    si $CPDAGtoDAG(H^-_{k+1}) = G^-_{k+1}$ **alors**
      score : $S(G^-_{k+1},\mathcal{D})$
      si $S(G^-_{k+1},\mathcal{D}) > S_{max}$ **alors**
        $S_{max} = S(G^-_{k+1},\mathcal{D})$
        $H_k = DAGtoCPDAG(G^-_{k+1})$ (voir [Chickering, 2003, Annexe C])
      fin
    fin
  fin
  $S_{old} = S$, $S = S_{max}$
  si $S \geq S_{old}$ **alors**
    $H = H_k$
    suppression de l'arête (R_i,R_j) manquante dans H_k de E
  fin
fin

fonction de score BIC définie en (5.3). Après l'initialisation, les CPDAG à une arête sont évalués. Ils sont représentés dans la colonne de gauche de la figure 5.7. La fonction de score est appliquée sur chacun des DAG de la colonne de droite, et est maximisée sur le DAG $G_{1,3}$ de la figure 5.7f. Le CPDAG correspondant au score maximal $S(G_{1,3},\mathcal{D})$, noté $H_{1,3}$, est donc pris comme graphe de départ pour l'insertion d'une nouvelle arête. Les CPDAG et les DAG associés de l'étape suivante, avec 2 arêtes, sont représentés dans les figures 5.8 et 5.9. Le DAG qui maximise le score est $G_{2,1}$ (figure 5.8b). Le score $S(G_{2,1},\mathcal{D})$ obtenu avec 2 arêtes étant plus élevé que le score $S(G_{1,3},\mathcal{D})$ obtenu avec 1 arête, l'insertion d'arcs se poursuit. Une arête supplémentaire est insérée à partir du CPDAG $H_{2,1}$ (figure 5.8b), pour parvenir aux CPDAG représentés dans les colonnes de gauche des figures 5.10 et 5.11. Le score maximal est alors obtenu pour le DAG $G_{3,5}$ de la figure 5.10j, et est de nouveau supérieur au score de l'étape précédente, avec 2 arêtes. Une nouvelle arête est donc ajoutée au CPDAG de la figure 5.10j. Les CPDAG résultants sont représentés dans la figure 5.12. Le score le plus grand $S(G_{4,3},\mathcal{D})$ est obtenu pour le DAG G_3 de la figure 5.12f, mais est inférieur à $S(G_{3,5},\mathcal{D})$. La phase d'insertion d'arc s'interrompt alors, et le CPDAG qui en est issu est $H_{3,5}$.

La phase de suppression d'arc démarre sur la solution de la phase d'insertion, le graphe essentiel $H_{3,5}$ avec 3 arêtes. Un ensemble de CPDAG sont obtenus par application de

l'algorithme 8 sur $H_{3,5}$ (figure 5.10i), et sont représentés dans la colonne de gauche de la figure 5.13. Le score maximal local est obtenu pour le DAG $G_{2,10}$ (figure 5.13a), mais comme $S(G_{2,10},\mathcal{D}) < S(G_{3,5},\mathcal{D})$, la phase de suppression d'arc s'interrompt et l'algorithme GES prend fin. Dans cet exemple, l'apprentissage des relations de dépendances entre les variables R_1, R_2, R_3 et R_4 à partir des observations de $\mathcal{D}$ a finalement conduit au graphe partiellement orienté $H_{3,5}$, comportant 3 arêtes et la V-structure $R_1 \rightarrow R_2 \leftarrow R_3$.

Pour un DAG G, le calcul de la fonction de score $S(G,\mathcal{D})$ repose sur l'ensemble des observations $\mathcal{D}$ des vecteurs des variables indicatrices $(R_1,\ldots,R_m)'$. Pour constituer cet ensemble, le modèle du *Bernoulli Detector* est appliqué sur la série temporelle $\boldsymbol{X}$, et conduit à l'introduction de deux nouveaux scores, différant des fonctions de l'état de l'art.

(a) CPDAG H_0 sans arête. (b) DAG G_0 sans arête.

FIGURE 5.6 – Initialisation de l'algorithme GES sur le CPDAG à 4 variables, sans arête, en 5.6a. L'unique DAG de cette classe d'équivalence de Markov, sur lequel est calculé le score $S(G_0,\mathcal{D})$, est représenté en 5.6b. Ce DAG maximise la fonction de score, et est représenté en rouge.

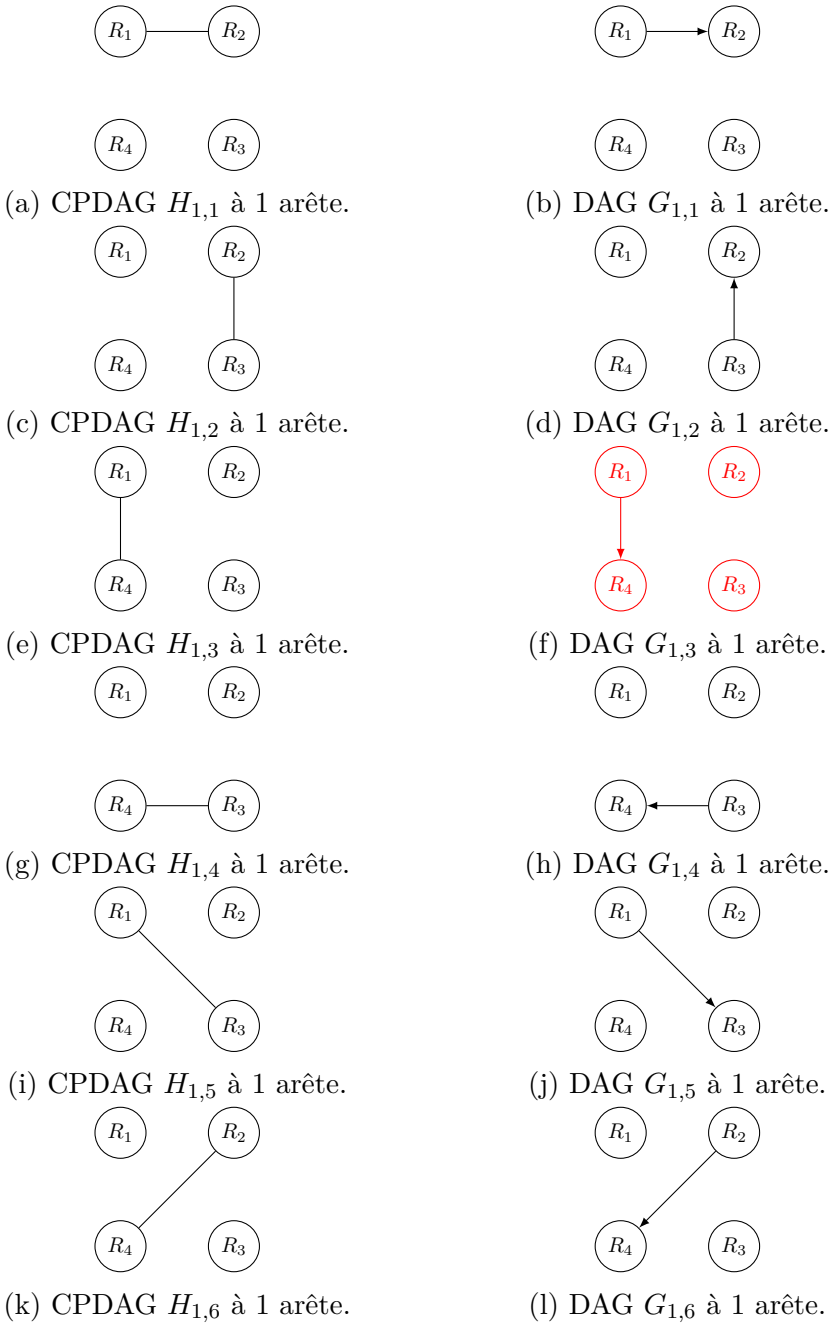
5.3 Exploitation des résultats du *Bernoulli Detector*

En l'absence de connaissance sur les positions réelles des ruptures qui permettraient de constituer des observations des variables indicatrices, l'algorithme MultiBD, présenté au chapitre 4, fournit plusieurs moyens de construire l'ensemble $\mathcal{D}$. À chacune des itérations MCMC v [3], la matrice $\boldsymbol{R}$ des indicatrices des ruptures dans $\boldsymbol{X}$, et $\boldsymbol{P}$, le vecteur des probabilités d'observer chacune des configurations ϵ_l, $1 \leq l \leq L$, sont échantillonnés. En m'inspirant du score BIC de l'état de l'art, je propose un nouveau score BIC construit sur les estimations de $\boldsymbol{R}$ issues de l'algorithme MultiBD, ainsi qu'un deuxième score basé sur la mesure de la divergence de Kullback-Leibler, ou intervient l'estimation de $\boldsymbol{P}$.

5.3.1 Indicatrices des ruptures et score BIC

À l'issue de V itérations, l'estimateur $\hat{\boldsymbol{R}}_{MAP}$ est obtenu. Les $(n-2)$ vecteurs colonnes de $\hat{\boldsymbol{R}}_{MAP}$, extrémités exclues, peuvent donc être employés comme échantillons de $(R_1,\ldots,R_m)$. L'ensemble ainsi constitué est noté $\mathcal{D}_{MAP}$. Une erreur dans la position temporelle d'une rupture n'est pas pénalisante, puisque les variables indicatrices sont supposées indépendantes en temps. L'ensemble $\mathcal{D}_{MAP}$ est constitué des observations de $(R_1,\ldots,R_m)$ qui maximisent

3. Nous supposons que la période de chauffe de la chaîne de Markov est terminée et que cette dernière est proche de la stationnarité.



(a) CPDAG $H_{1,1}$ à 1 arête. (b) DAG $G_{1,1}$ à 1 arête.

(c) CPDAG $H_{1,2}$ à 1 arête. (d) DAG $G_{1,2}$ à 1 arête.

(e) CPDAG $H_{1,3}$ à 1 arête. (f) DAG $G_{1,3}$ à 1 arête.

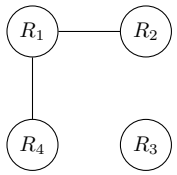
(g) CPDAG $H_{1,4}$ à 1 arête. (h) DAG $G_{1,4}$ à 1 arête.

(i) CPDAG $H_{1,5}$ à 1 arête. (j) DAG $G_{1,5}$ à 1 arête.

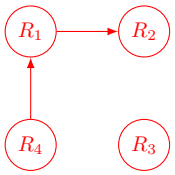
(k) CPDAG $H_{1,6}$ à 1 arête. (l) DAG $G_{1,6}$ à 1 arête.

FIGURE 5.7 – Première insertion d'arête par l'algorithme GES à partir du CPDAG H_0 de la figure 5.6a. Tous les CPDAG sont représentés dans la colonne de gauche, et à droite de chacun de ces CPDAG se trouve l'un des DAG de la classe d'équivalence de Markov représentée par le CPDAG. Le score avec 1 arête est maximal pour le DAG $G_{1,3}$ de la figure 5.7f, en rouge. Comme $S(G_{1,3},\mathcal{D}) > S(G_0,\mathcal{D})$, la phase d'insertion d'arc se poursuit.

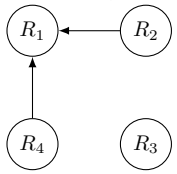
la densité de probabilité a posteriori en (4.14). Le comptage des différentes configurations N_{ijk} dans l'expression de la vraisemblance (5.8) du score BIC est donc réalisé sur $\mathcal{D}_{MAP}$.



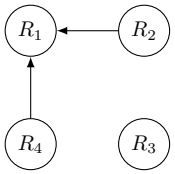
(a) CPDAG $H_{2,1}$ à 2 arêtes.



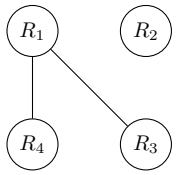
(b) DAG $G_{2,1}$ à 2 arêtes.



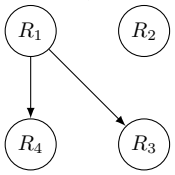
(c) CPDAG $H_{2,2}$ à 2 arêtes.



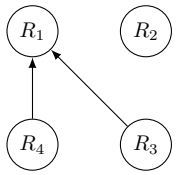
(d) DAG $G_{2,2}$ à 2 arêtes.



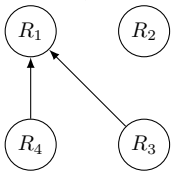
(e) CPDAG $H_{2,3}$ à 2 arêtes.



(f) DAG $G_{2,3}$ à 2 arêtes.



(g) CPDAG $H_{2,4}$ à 2 arêtes.



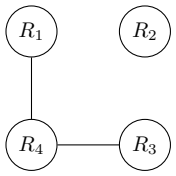
(h) DAG $G_{2,4}$ à 2 arêtes.

FIGURE 5.8 – Deuxième insertion d'arête par l'algorithme GES à partir du CPDAG $H_{1,3}$ de la figure 5.7e, première partie de la figure. Les CPDAG sont représentés dans la colonne de gauche, et à droite de chacun se trouve l'un des DAG de la classe d'équivalence de Markov. Le score calculé sur ces DAG avec 2 arêtes et ceux de la figure 5.9 est maximal pour le DAG $G_{2,1}$ (figure 5.8b), en rouge. Comme $S(G_{2,1},\mathcal{D}) > S(G_{1,3},\mathcal{D})$, la phase d'insertion d'arc se poursuit.

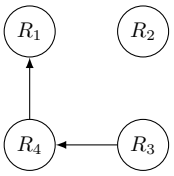
Pour estimer les paramètres $\hat{\theta}_{ijk}^{MV}$, l'ensemble de la chaîne de Markov $(\boldsymbol{R}^{(v)})_v$ est exploitée. Les $(n-2)*V$ échantillons, issus des V itérations MCMC, forment l'ensemble $\mathcal{D}_{moy}$, et comptabilisent la totalité des configurations obtenues avec MultiBD. Ainsi, dans le score BIC, les estimations des paramètres $\hat{\theta}_{ijk}^{MV}$ sont déterminés d'après le comptage moyen des configurations dans $\boldsymbol{R}$ au cours des itérations MCMC, c'est-à-dire de $\mathcal{D}_{moy}$. Pour différencier les deux origines des données, les comptages des configurations où $R_i = r_k$ et $\text{Pa}(R_i) = r_j$ sont notées $N_{ijk,moy}$ et $N_{ijk,MAP}$, pour $\mathcal{D}_{moy}$ et $\mathcal{D}_{MAP}$ respectivement. La vraisemblance (5.8) est donc approximée par

$$L(\mathcal{D}|\theta) = \prod_{i=1}^{m}\prod_{j=1}^{q_i}\prod_{k=1}^{s_i}\left(\frac{N_{ijk,moy}}{\sum_{k=1}^{s_i} N_{ijk,moy}}\right)^{N_{ijk,MAP}}. \tag{5.10}$$

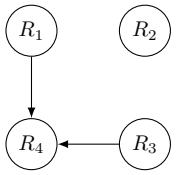
Le deuxième terme de l'expression (5.3) qui intègre la complexité du modèle est construit à partir du nombre d'échantillons intervenant dans le calcul de la vraisemblance, c'est-à-



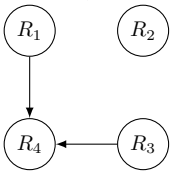
(a) CPDAG $H_{2,5}$ à 2 arêtes.



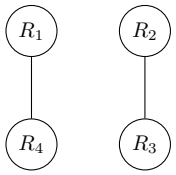
(b) DAG $G_{2,5}$ à 2 arêtes.



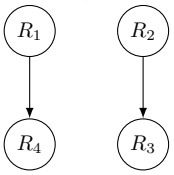
(c) CPDAG $H_{2,6}$ à 2 arêtes.



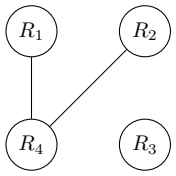
(d) DAG $G_{2,6}$ à 2 arêtes.



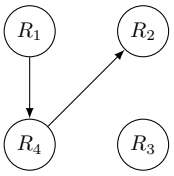
(e) CPDAG $H_{2,7}$ à 2 arêtes.



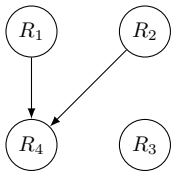
(f) DAG $G_{2,7}$ à 2 arêtes.



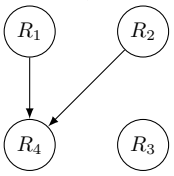
(g) CPDAG $H_{2,8}$ à 2 arêtes.



(h) DAG $G_{2,8}$ à 2 arêtes.



(i) CPDAG $H_{2,9}$ à 2 arêtes.

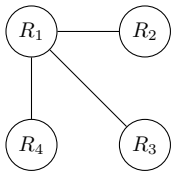


(j) DAG $G_{2,9}$ à 2 arêtes.

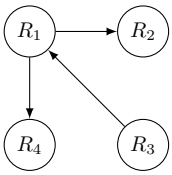
FIGURE 5.9 – Deuxième insertion d'arête par l'algorithme GES à partir du CPDAG $H_{1,3}$ de la figure 5.7e, suite de la figure 5.8. Les CPDAG sont représentés dans la colonne de gauche, et à droite de chacun se trouve l'un des DAG de la classe d'équivalence de Markov.

dire du nombre d'éléments de $\mathcal{D}_{MAP}$, soit $(n-2)$. Finalement, le score BIC, décomposé sur chaque nœud i du réseau, $1 \leq i \leq m$, est obtenu d'après les résultats de l'application du *Bernoulli Detector* selon :

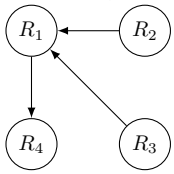
$$S_{BIC}(G,\mathcal{D}_{MAP},\mathcal{D}_{moy}) = \sum_{i=1}^{m} \left(\sum_{j=1}^{q_i} \sum_{k=1}^{2} N_{ijk,MAP} \log \frac{N_{ijk,moy}}{\sum_{k=1}^{2} N_{ijk,moy}} - \frac{1}{2} 2^{n(\mathrm{Pa}(R_i))} \log(n-2) \right). \tag{5.11}$$



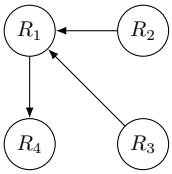
(a) CPDAG $H_{3,1}$ à 3 arêtes.



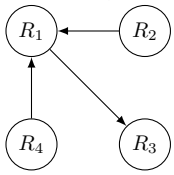
(b) DAG $G_{3,1}$ à 3 arêtes.



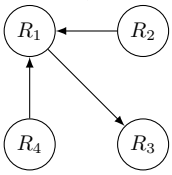
(c) CPDAG $H_{3,2}$ à 3 arêtes.



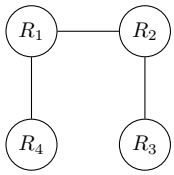
(d) DAG $G_{3,2}$ à 3 arêtes.



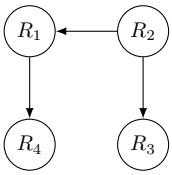
(e) CPDAG $H_{3,3}$ à 3 arêtes.



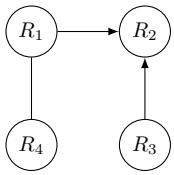
(f) DAG $G_{3,3}$ à 3 arêtes.



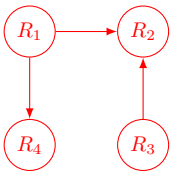
(g) CPDAG $H_{3,4}$ à 3 arêtes.



(h) DAG $G_{3,4}$ à 3 arêtes.



(i) CPDAG $H_{3,5}$ à 3 arêtes.

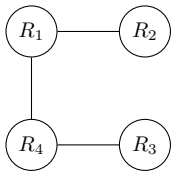


(j) DAG $G_{3,5}$ à 3 arêtes.

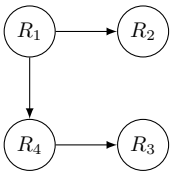
FIGURE 5.10 – Troisième insertion d'arête par l'algorithme GES à partir du CPDAG $H_{2,1}$ de la figure 5.8a, première partie de la figure. Les CPDAG sont représentés dans la colonne de gauche, et à droite de chacun se trouve l'un des DAG de la classe d'équivalence de Markov. Le score calculé sur ces DAG avec 3 arêtes et ceux de la figure 5.11 est maximal pour le DAG $G_{3,5}$ (figure 5.10j), en rouge. Comme $S(G_{3,5},\mathcal{D}) > S(G_{2,1},\mathcal{D})$, la phase d'insertion d'arc se poursuit.

5.3.2 Probabilités des configurations et divergence de Kullback-Leibler

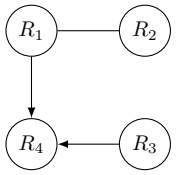
Le vecteur des variables indicatrices de U peut prendre $L = 2^m$ valeurs possibles, qui sont les configurations ϵ_l, pour $1 \leq l \leq L$. Le modèle du *Bernoulli Detector* introduit le paramètre $\boldsymbol{P} = (P_{\epsilon_1}, \ldots, P_{\epsilon_L})$, qui est le vecteur des probabilités d'obtenir ces configurations. Par exemple, pour les trois variables discrètes R_1, R_2 et R_3 et avec les notations



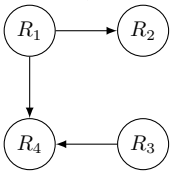
(a) CPDAG $H_{3,6}$ à 3 arêtes.



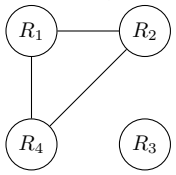
(b) DAG $G_{3,6}$ à 3 arêtes.



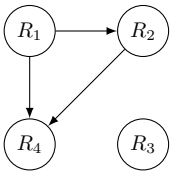
(c) CPDAG $H_{3,7}$ à 3 arêtes.



(d) DAG $G_{3,7}$ à 3 arêtes.



(e) CPDAG $H_{3,8}$ à 3 arêtes.



(f) DAG $G_{3,8}$ à 3 arêtes.

FIGURE 5.11 – Troisième insertion d'arête par l'algorithme GES à partir du CPDAG $H_{2,1}$ de la figure 5.8a, suite de la figure 5.10. Les CPDAG sont représentés dans la colonne de gauche, et à droite de chacun se trouve l'un des DAG de la classe d'équivalence de Markov.

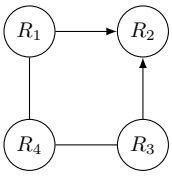
employées pour le *Bernoulli Detector* multivarié, ces probabilités sont :

$$\begin{aligned}
&\Pr(R_1 = 0, R_2 = 0, R_3 = 0) = P_{000} = P_{\epsilon_1} \\
&\Pr(R_1 = 0, R_2 = 0, R_3 = 1) = P_{001} = P_{\epsilon_2} \\
&\Pr(R_1 = 0, R_2 = 1, R_3 = 0) = P_{010} = P_{\epsilon_3} \\
&\Pr(R_1 = 0, R_2 = 1, R_3 = 1) = P_{011} = P_{\epsilon_4} \\
&\Pr(R_1 = 1, R_2 = 0, R_3 = 0) = P_{100} = P_{\epsilon_5} \\
&\Pr(R_1 = 1, R_2 = 0, R_3 = 1) = P_{101} = P_{\epsilon_6} \\
&\Pr(R_1 = 1, R_2 = 1, R_3 = 0) = P_{110} = P_{\epsilon_7} \\
&\Pr(R_1 = 1, R_2 = 1, R_3 = 1) = P_{111} = P_{\epsilon_8}.
\end{aligned} \tag{5.12}$$

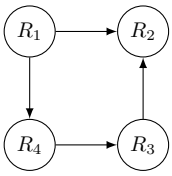
La loi jointe $\mathcal{P}$ du modèle $M(U)$ correspond donc à la loi de probabilité du vecteur $\boldsymbol{P}$, dont l'algorithme MultiBD procure la distribution empirique. Dans notre modèle du *Bernoulli Detector*, le paramètre est généré selon la loi de Dirichlet (4.13) pour une matrice des indicatrices $\boldsymbol{R}$ donnée. Les probabilités marginales P_{ϵ_l} suivent donc les lois bêta :

$$P_{\epsilon_l} | \boldsymbol{R} \sim Beta\Big(S_l(\boldsymbol{R}) + 1, \sum_{\substack{l'=1 \\ l' \neq l}}^{L} (S_{l'}(\boldsymbol{R}) + 1)\Big), \tag{5.13}$$

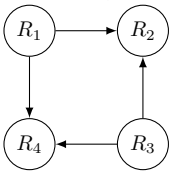
où $S_l(\boldsymbol{R})$ est le nombre de fois où la configuration ϵ_l apparaît dans $\boldsymbol{R}$, en-dehors des premiers et derniers points temporels, auxquels la configuration ϵ_L est attribuée par convention, et



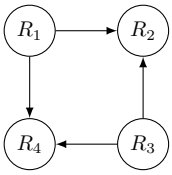
(a) CPDAG $H_{4,1}$ à 4 arêtes.



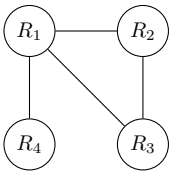
(b) DAG $G_{4,1}$ à 4 arêtes.



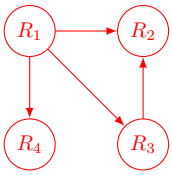
(c) CPDAG $H_{4,2}$ à 4 arêtes.



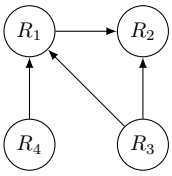
(d) DAG $G_{4,2}$ à 4 arêtes.



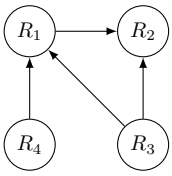
(e) CPDAG $H_{4,3}$ à 4 arêtes.



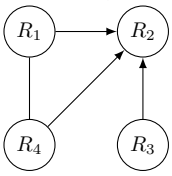
(f) DAG $G_{4,3}$ à 4 arêtes.



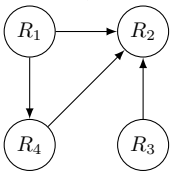
(g) CPDAG $H_{4,4}$ à 4 arêtes.



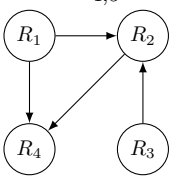
(h) DAG $G_{4,4}$ à 4 arêtes.



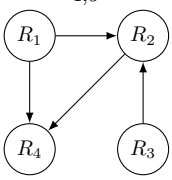
(i) CPDAG $H_{4,5}$ à 4 arêtes.



(j) DAG $G_{4,5}$ à 4 arêtes.



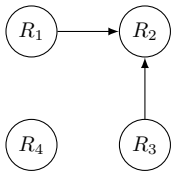
(k) CPDAG $H_{4,6}$ à 4 arêtes.



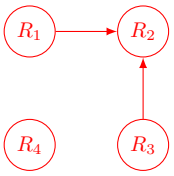
(l) DAG $G_{4,6}$ à 4 arêtes.

FIGURE 5.12 – Quatrième insertion d'arête par l'algorithme GES à partir du CPDAG $H_{3,5}$ de la figure 5.10i. Tous les CPDAG sont représentés dans la colonne de gauche, et à droite de chacun de ces CPDAG se trouve l'un des DAG de la classe d'équivalence de Markov représentée par le CPDAG. Le score avec 4 arêtes est maximal pour le DAG $G_{4,3}$ de la figure 5.12f, en rouge, mais comme $S(G_{4,3},\mathcal{D}) < S(G_{3,5},\mathcal{D})$, la phase d'insertion d'arête prend fin.

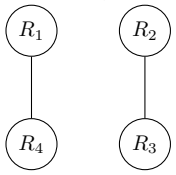
qui ne sont pas pris en compte dans ce chapitre. Pour les $(n-2)$ colonnes de $\boldsymbol{R}$ et les L



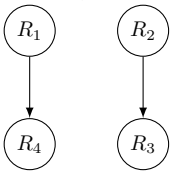
(a) CPDAG $H_{2,10}$ à 2 arêtes.



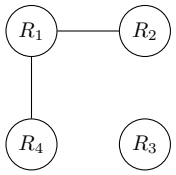
(b) DAG $G_{2,10}$ à 2 arêtes.



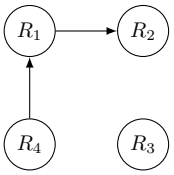
(c) CPDAG $H_{2,7}$ à 2 arêtes.



(d) DAG $G_{2,7}$ à 2 arêtes.



(e) CPDAG $H_{2,11}$ à 2 arêtes.



(f) DAG $G_{2,1}$ à 2 arêtes.

FIGURE 5.13 – Première suppression d'arête par l'algorithme GES, à partir du CPDAG $H_{3,5}$ de la figure 5.10i. Les CPDAG sont représentés dans la colonne de gauche, et à droite de chacun se trouve l'un des DAG de la classe d'équivalence de Markov. Le score calculé sur ces DAG avec 2 arêtes est maximal pour le DAG $G_{2,10}$ (figure 5.13b), en rouge. Comme $S(G_{2,10},\mathcal{D}) < S(G_{3,5},\mathcal{D})$, la phase de suppression d'arc prend fin.

configurations, la loi (5.13) s'écrit :

$$P_{\epsilon_l}|\boldsymbol{R} \sim Beta\Big(S_l(\boldsymbol{R}) + 1, L + n - S_l(\boldsymbol{R}) - 3\Big), \tag{5.14}$$

Contrairement au cadre habituel de l'apprentissage de la structure d'un réseau, le *Bernoulli Detector* nous permet donc d'exploiter directement la loi de probabilité jointe via $\boldsymbol{P}$. La distribution empirique formée des V échantillons de $\boldsymbol{P}$, générés au cours des itérations MCMC de l'algorithme MultiBD, est notée $\mathcal{P}_{BD}$.

Pour compléter les résultats obtenus avec la fonction de score BIC, la divergence de Kullback-Leibler [Kullback et Leibler, 1951] est également employée. Ce critère est calculé entre la loi jointe issue des données du *Bernoulli Detector* et sa factorisation pour le graphe G. La loi jointe empirique est donnée directement par $\mathcal{P}_{BD}$, et les distributions des paramètres $\theta_{i,j,k}$ pour un graphe G se déduisent de $\mathcal{P}_{BD}$. Nous pouvons déterminer les termes de la factorisation de la loi jointe selon le modèle G testé, pour parvenir à la distribution décomposée :

$$\mathcal{P}_G = P_G(R_1,\ldots,R_m) = \prod_{i=1}^{m} P(R_i|\,\mathrm{Pa}(R_i),\theta). \tag{5.15}$$

Par exemple dans le cas du modèle représenté dans la figure 5.14, la factorisation

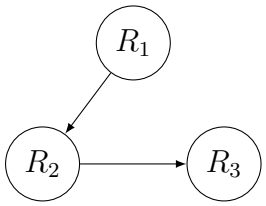


FIGURE 5.14 – Exemple de DAG G testé pour trois variables et deux arêtes.

compatible avec G est

$$\mathcal{P}_G = P(R_1)P(R_2|R_1)P(R_3|R_2). \tag{5.16}$$

Les probabilités marginales et conditionnelles de (5.16) sont alors déterminées à partir des probabilités jointes (5.13), issues du *Bernoulli Detector* :

$$\begin{aligned}
\Pr(R_1 = 0) &= P_{000} + P_{001} + P_{010} + P_{011},\\
\Pr(R_1 = 1) &= P_{100} + P_{101} + P_{110} + P_{111},\\
\Pr(R_2 = 0|R_1 = 0) &= \frac{P_{000} + P_{001}}{P_{000} + P_{001} + P_{010} + P_{011}},\\
\Pr(R_2 = 1|R_1 = 0) &= \frac{P_{010} + P_{011}}{P_{000} + P_{001} + P_{010} + P_{011}},\\
\Pr(R_2 = 0|R_1 = 1) &= \frac{P_{100} + P_{101}}{P_{100} + P_{101} + P_{110} + P_{111}},\\
\Pr(R_2 = 1|R_1 = 1) &= \frac{P_{110} + P_{111}}{P_{100} + P_{101} + P_{110} + P_{111}},\\
\Pr(R_3 = 0|R_1 = 0) &= \frac{P_{000} + P_{010}}{P_{000} + P_{001} + P_{010} + P_{011}},\\
\Pr(R_3 = 1|R_1 = 0) &= \frac{P_{001} + P_{011}}{P_{000} + P_{001} + P_{010} + P_{011}},\\
\Pr(R_3 = 0|R_1 = 1) &= \frac{P_{100} + P_{110}}{P_{100} + P_{101} + P_{110} + P_{111}},\\
\Pr(R_3 = 1|R_1 = 1) &= \frac{P_{101} + P_{111}}{P_{100} + P_{101} + P_{110} + P_{111}}.
\end{aligned} \tag{5.17}$$

Pour comparer les distributions $\mathcal{P}_G$ et de $\mathcal{P}_{BD}$, la divergence de Kullback-Leibler [Kullback et Leibler, 1951] est un critère approprié. Elle est définie par :

$$d_{KL}(\mathcal{P}_{BD},\mathcal{P}_G) = \sum_{r_1,\ldots,r_m} \mathcal{P}_{BD}(R_1 = r_1,\ldots,R_m = r_m) \log \frac{\mathcal{P}_{BD}(R_1 = r_1,\ldots,R_m = r_m)}{\mathcal{P}_G(R_1 = r_1,\ldots,R_m = r_m)} \tag{5.18}$$

et mesure la différence entre la distribution du modèle associé à G avec la distribution des données. Pour intégrer toutes les configurations possibles des variables $R_1,\ldots,R_m$ sur tous

les échantillons issus des itérations MCMC de la méthode MultiBD, la divergence moyenne, décomposée sur les configurations, est donnée par :

$$d_{KL,V}(\mathcal{P}_{BD},\mathcal{P}_G) = \frac{1}{V}\sum_{v=1}^{V}\sum_{L=1}^{l}\mathcal{P}_{BD}^{(v)}((R_1,\ldots,R_m)=\epsilon_l)\log\frac{\mathcal{P}_{BD}^{(v)}((R_1,\ldots,R_m)=\epsilon_l)}{\mathcal{P}_G^{(v)}((R_1,\ldots,R_m)=\epsilon_l)}. \quad (5.19)$$

Le modèle optimal est celui qui minimise la divergence.

Cependant comme ce critère ne prend pas en compte la complexité du modèle, la fonction de score définie par $S_{KL}(G,\mathcal{P}_{BD}) = -d_{KL,V}(\mathcal{P}_{BD},\mathcal{P}_G)$ est monotone et ne permet pas de parvenir à un maximum. La phase d'insertion d'arc se déroule jusqu'à la classe des graphes complets, aucune suppression d'arc ne se produisant par la suite. En revanche on observe que la courbe du score en fonction du nombre d'arêtes se décompose en deux parties. Lorsque le nombre d'arêtes est inférieur au nombre réel N_A, des relations de dépendances conditionnelles du modèle manquent dans la représentation graphique, et l'ajout d'une des arêtes manquantes améliore nettement le score. Lorsque le nombre d'arêtes atteint ou dépasse N_A, les graphes sont des cartes d'indépendance du modèle. L'ajout de ces arêtes jusqu'au graphe complet n'augmente pas sensiblement le score, et une rupture de pente dans la courbe de $S_{KL}(G,\mathcal{P}_{BD})$ en fonction du nombre d'arcs permet de localiser la carte d'indépendance minimale.

Pour employer la divergence de Kullback-Leibler avec la méthode GES, une heuristique de pente est introduite avec la fonction de score, à la fin des étapes d'insertion et de suppression d'arc. La phase d'insertion se déroule donc jusqu'au graphe complet. La séquence des meilleurs scores pour un nombre d'arêtes croissant est scindée en deux parties, de 0 à e arêtes et de e à e_{max} arêtes, avec $e_{max} = m(m-1)/2$. Une régression linéaire par moindres carrés sur les deux parties de la courbe est calculée. La position de e varie entre 1 et $e_{max} - 1$. Le nombre optimal d'arêtes est estimé à la position $e^\star$ qui minimise la somme des carrés des résidus de la régression. La phase de suppression démarre ensuite du graphe $G(e^\star)$, et s'effectue tant que le score augmente. Le graphe final est déterminé de la même façon après application de l'heuristique sur les scores de la phase de suppression. Cette approche basée sur la divergence de Kullback-Leibler, appelée critère KL, est mise en œuvre sur des données estimées par l'algorithme MultiBD, et est comparée avec la fonction de score BIC.

5.4 Résultats expérimentaux

Pour mettre en œuvre la méthode GES sur les résultats issus de l'algorithme MultiBD, avec les fonctions de score présentées aux parties 5.3.1 et 5.3.2, nous avons repris et adapté les codes de la `Bayes Net Toolbox` pour `Matlab` de [Murphy *et al.*, 2001], rassemblant des méthodes d'apprentissage de réseaux bayésiens, et du `Structure Learning Package` de [Leray et François, 2004] qui fournit une implémentation efficace de la méthode GES. Plusieurs expériences ont été menées à partir de séries temporelles simulées, puis les informations extraites du réseau électrique par le *Bernoulli Detector* dans la partie 4.4.2, ont été traitées.

5.4.1 Simulations

Simulation 1 : 4 variables et ruptures propres à chaque signal

Pour la première simulation, une structure de dépendance simple est choisie, comprenant $m = 4$ variables et une V-structure. Le DAG original et le graphe essentiel associé sont représentés à la figure 5.15. Chaque signal j présente des ruptures qui lui sont propres, c'est-à-dire qui n'ont pas pour origine un autre signal, et des ruptures communes à d'autre signaux avec lesquels il existe des relations de dépendance, et dont l'origine est un nœud parent de R_j dans le graphe G. Les ruptures propres aux signaux sont générées en échantillonnant des longueurs de segments selon une loi uniforme sur les intervalles [40,80] pour le signal 1, [100,120] pour le 2, [30,70] pour le 3 et enfin [90,100] pour le 4. Les ruptures ainsi simulées sont propagées selon le graphe de dépendance G de la figure 5.15a. Chaque rupture du signal 1 se produit donc au même instant dans le signal 2 avec la probabilité $p_{1\to2} = 0{,}8$, et dans le signal 4 avec $p_{1\to4} = 0{,}8$. De la même façon, les ruptures du signal 3 se produisent aussi dans le signal 2 avec la probabilité $p_{3\to2} = 0{,}5$.

(a) DAG (b) CPDAG H_G, désigné par 20552 [4]

FIGURE 5.15 – DAG (5.15a) et CPDAG (5.15b) de la classe d'équivalence pour la première simulation. Le poids de l'arête orientée de R_j vers R_i indique la probabilité $p_{j\to i}$ que l'événement $R_j = 1$ cause l'événement $R_i = 1$.

La matrice réelle des indicatrices étant créée, la série temporelle $\boldsymbol{r}$ est alors générée en simulant les observations x_k sur le segment k selon une loi normale de moyenne μ variant entre 5,0 et 10,0 d'un segment à l'autre, et telle que le SNR défini dans (3.59) vaut 5,0 dB. L'algorithme MultiBD est appliqué à la série temporelle résultante, avec $\alpha = 0{,}01$ et pour 1000 itérations MCMC. Les 500 premières itérations ne sont pas prises en compte pour s'assurer que la chaîne de Markov soit proche de l'état stationnaire. Ainsi, $V = 500$. À l'issue du *Bernoulli Detector*, les ensembles de données $\mathcal{D}_{MAP}$, $\mathcal{D}_{moy}$ et $\mathcal{P}_{BD}$ sont constitués.

Dans un premier temps, l'algorithme d'apprentissage GES est appliqué sur la matrice des indicatrices réelles, dont les colonnes forment l'ensemble d'observations noté $\mathcal{D}_{reel}$. La figure 5.16 montre l'évolution de la proportion de graphes essentiels correctement estimés par l'algorithme GES avec le score BIC de l'état de l'art, donné en (5.3) en fonction du nombre de points n de la série temporelle, sur les observations $\mathcal{D}_{reel}$ et pour 1000 séries temporelles. Nous pouvons constater qu'à partir de 600 observations, la classe d'équivalence

4. Le nom de la classe d'équivalence est obtenu en convertissant en un entier décimal le nombre binaire formé à partir des vecteurs ligne de la matrice d'adjacence de H_G mis bout-à-bout.

est bien obtenue dans la majorité des cas, et que le modèle d'indépendance peut bien être déduit de l'observation des ruptures. Lorsque la longueur de la série temporelle augmente, les lois de probabilités des différentes configurations des ruptures sont mieux estimées, et les échantillons de l'ensemble fini des données $\mathcal{D}_{reel}$ sont plus représentatifs du comportement réel du système. Avec $n = 1000$, plus de 80% des graphes essentiels réels sont retrouvés. Par la suite, nous générons des séries temporelles de 1000 points. Soulignons que pour l'ensemble des simulations réalisées, la phase de suppression d'arc n'augmente pas le score, le maximum issu de la phase d'insertion donne directement la solution.

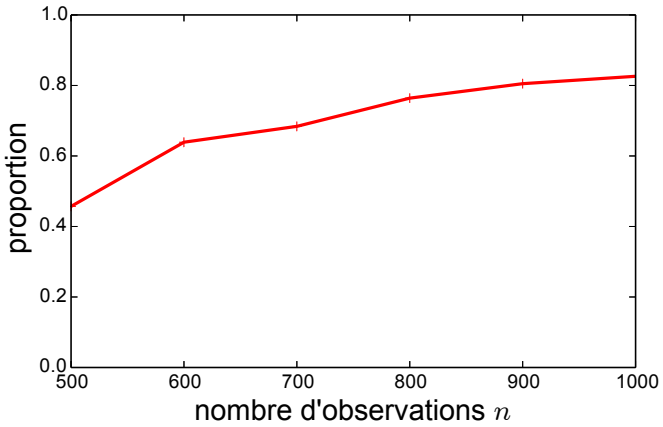
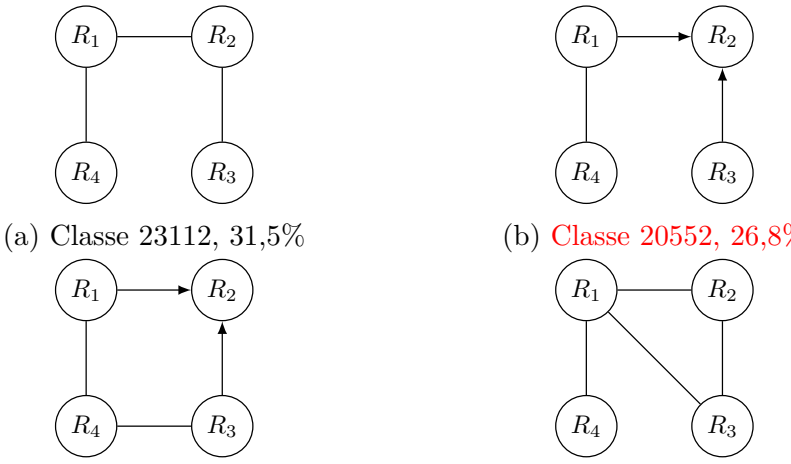


FIGURE 5.16 – Proportion de classes d'équivalences correctement apprises pour la simulation 1, en fonction du nombre d'observations et à partir des vraies positions des ruptures sur 1000 tests, pour la méthode d'apprentissage GES. La structure de dépendance réelle est celle de la figure 5.15a.

Nous appliquons ensuite la méthode GES avec le score BIC sur 200 séries temporelles. Les CPDAG obtenus le plus souvent et la proportion des ces solutions sont représentés dans la figure 5.17. Le graphe essentiel le plus estimé, dans 31,5% des cas, possède le même squelette que le graphe réel, et seule la V-structure est manquante. Le graphe réel est obtenu pour un quart des séries temporelles. Les deux autres solutions les plus fréquentes correspondent au bon graphe, auquel une arête est ajoutée. La méthode ne parvient donc pas toujours à retrouver la V-structure et a légèrement tendance à sur-apprendre. L'écart important entre ces résultats et les 80% de bons graphes obtenus avec les vraies positions des ruptures s'explique par la difficulté à segmenter les séries temporelles. En effet, certaines ruptures se produisent dans un intervalle de temps très court sur un même signal, et comme les moyennes μ des observations $x_{j,i}$ alternent entre deux valeurs, il arrive que les ruptures des segments de quelques points ne soient pas détectées.

Le critère KL est également employé sur les résultats du *Bernoulli Detector*, cette fois sur l'ensemble des probabilités des configurations $\mathcal{P}_{BD}$. L'évolution du score moyen sur les 200 séries temporelles lors de la phase d'insertion est tracée à la figure 5.18. Une légère rupture de pente est visible au niveau du vrai nombre d'arêtes, 3. Les graphes essentiels



(a) Classe 23112, 31,5%

(b) Classe 20552, 26,8%

(c) Classe 20570, 8,9%

(d) Classe 31432, 8,9%

FIGURE 5.17 – CPDAG des classes d'équivalence obtenues le plus souvent avec le score BIC pour la simulation 1, sur 200 séries temporelles. Les proportions sont indiquées dans la légende. Le graphe essentiel réel (en rouge) est en 5.17b.

déduits de l'application de l'heuristique de pente sur la fonction de score sont représentés dans la figure 5.19. Dans un tiers des cas, la classe d'équivalence 22536 est apprise. Elle ne comporte que deux des arêtes réelles, l'arc de R_3 vers R_2 étant manquant. L'estimation du nombre d'arêtes par la comparaison des pentes sur la divergence de Kullback-Leibler n'est pas assez précise, et à tendance a sous-estimer la valeur réelle. Le bon nombre d'arêtes a toutefois été retrouvé pour 60% des tests. Le critère KL conduit à la bonne classe dans 28,6% des cas, ce qui est comparable au taux issu du score BIC. Le troisième CPDAG très représenté est le graphe non orienté de même squelette que le vrai graphe, sans la V-structure.

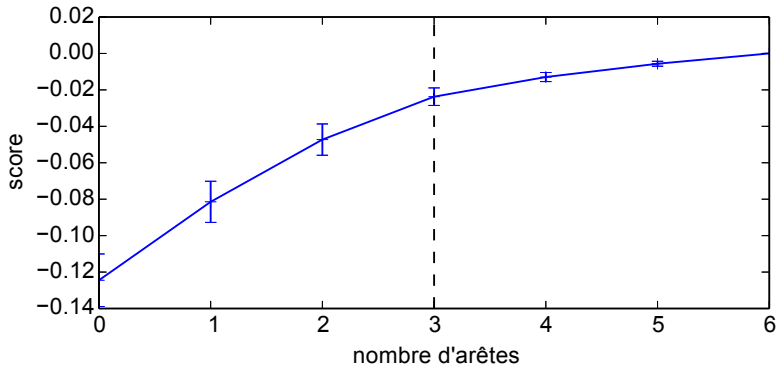
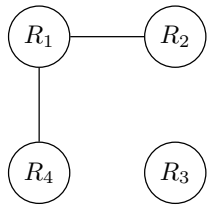
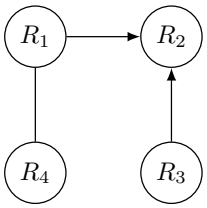


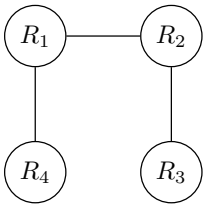
FIGURE 5.18 – Évolution du score moyen lors de la phase d'insertion d'arc de la méthode GES avec le critère KL, pour 200 séries temporelles de la simulation 1. Le nombre réel d'arêtes est indiqué en pointillés.



(a) Classe 22536, 34,3%



(b) Classe 20552, 28,6%



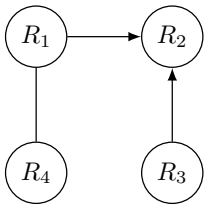
(c) Classe 23112, 26,8%

FIGURE 5.19 – CPDAG des classes d'équivalence obtenues le plus souvent avec le score KL pour la simulation 1, sur 200 séries temporelles. Les proportions sont indiquées dans la légende. Le graphe essentiel réel (en rouge) est en 5.19b.

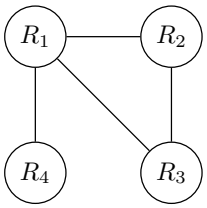
Simulation 2 : 4 variables et pas de rupture propre au signal 2

Les séries temporelles de la simulation précédente comportent des ruptures spécifiques à chaque signal. Nous analysons maintenant des séries temporelles de même graphe de dépendance 5.15a, où toutes les ruptures du signal 2 ont pour origine une rupture soit dans le signal 1, soit dans le signal 3. Ainsi, l'événement $(R_1 = 0 \cup R_2 = 1 \cup R_3 = 0)$ n'est jamais observé, ce qui correspond aux configurations $\epsilon_5 = (0,1,0,0)'$ et $\epsilon_6 = (0,1,0,1)'$. 200 séries temporelles de $n = 1000$ observations sont générées comme précédemment. Les principaux graphes essentiels appris avec le score BIC et leurs proportions sont indiqués dans la figure 5.20. Il s'agit des mêmes classes d'équivalence que pour la simulation 1. Le graphe réel est obtenu le plus souvent et dans un plus grand nombre de fois que dans la figure 5.17 : 35,0%, contre 26,8%. Lorsqu'on applique la méthode GES sur les données $\mathcal{D}_{reel}$, avec les vraies ruptures, la solution oracle parvient à la bonne classe d'équivalence pour quasiment tous les tests, contre seulement 86% dans la simulation précédente. L'absence d'événements spécifiques au signal 2 favorise donc l'apprentissage du modèle d'indépendance conditionnel.

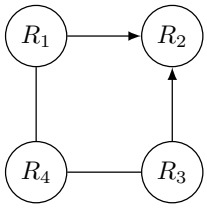
Le critère KL est également appliqué, les graphes essentiels obtenus sont donnés dans la figure 5.21. Il s'agit des mêmes classes d'équivalence que lors de la simulation 1. Le graphe réel est cette fois-ci appris dans 61,4% des cas, ce qui est bien plus important que dans la simulation précédente et qu'avec la fonction de score BIC. Le bon nombre d'arêtes est estimé pour 83% des tests ; l'heuristique sur la pente de la fonction de score donne de bons résultats.



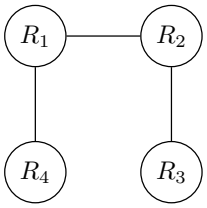
(a) Classe 20552, 35,0%



(b) Classe 31432, 24,1%

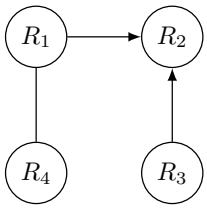


(c) Classe 20570, 19,5%

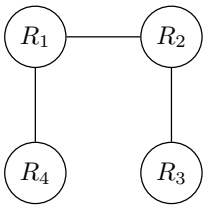


(d) Classe 23112, 9,5%

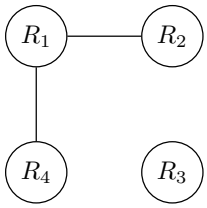
FIGURE 5.20 – CPDAG des classes d'équivalence obtenues le plus souvent avec le score BIC pour la simulation 2, sans ruptures propres au signal 2, pour 200 séries temporelles. Les proportions sont indiquées dans la légende. Le graphe essentiel réel (en rouge) est en 5.20a.



(a) Classe 20552, 61,4%



(b) Classe 23112, 19,1%



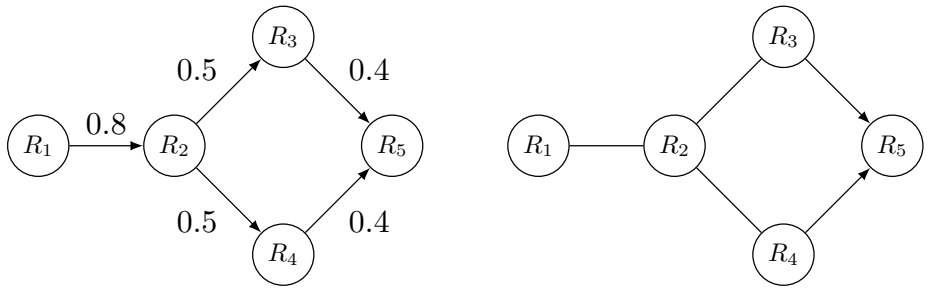
(c) Classe 22536, 14,1%

FIGURE 5.21 – CPDAG des classes d'équivalence obtenues le plus souvent avec le score KL pour la simulation 2, sans ruptures propres au signal 2, pour 200 séries temporelles. Les proportions sont indiquées dans la légende. Le graphe essentiel réel (en rouge) est en 5.21a.

Simulation 3 : 5 variables

La troisième simulation fait intervenir une structure un peu plus complexe à $m = 5$ variables. Le DAG de la structure de dépendance ainsi que la classe d'équivalence associée sont représentés dans la figure 5.22. Les longueurs des segments composant les signaux de la série temporelle sont simulés selon une loi uniforme sur les intervalles [40,80] pour le signal 1, [100,120] pour le 2, [30,70] pour le 3, [90,100] pour le 4 et [105,140] pour le 5. Ces ruptures initialement propres à chacun des signaux se propagent entre les composantes connectées

dans le graphe G de la figure 5.22a, avec les probabilités $p_{1\to2} = 0{,}8$, $p_{2\to3} = 0{,}5$, $p_{2\to4} = 0{,}5$, $p_{3\to5} = 0{,}4$ et $p_{4\to5} = 0{,}4$. 200 séries temporelles de $n = 1000$ points sont générées. Les observations sont i.i.d. et suivent une loi normale de moyenne μ valant successivement 5,0, 10,0 et 15,0 selon les segments, avec SNR =5,0 dB. L'algorithme MultiBD est appliqué avec $\alpha = 0{,}01$ pour détecter les ruptures. 1000 itérations MCMC sont effectuées, dont les 500 dernières sont exploitées pour constituer les ensembles de données $\mathcal{D}_{moy}$ et $\mathcal{P}_{BD}$.



(a) DAG (b) CPDAG, désigné par 9119008

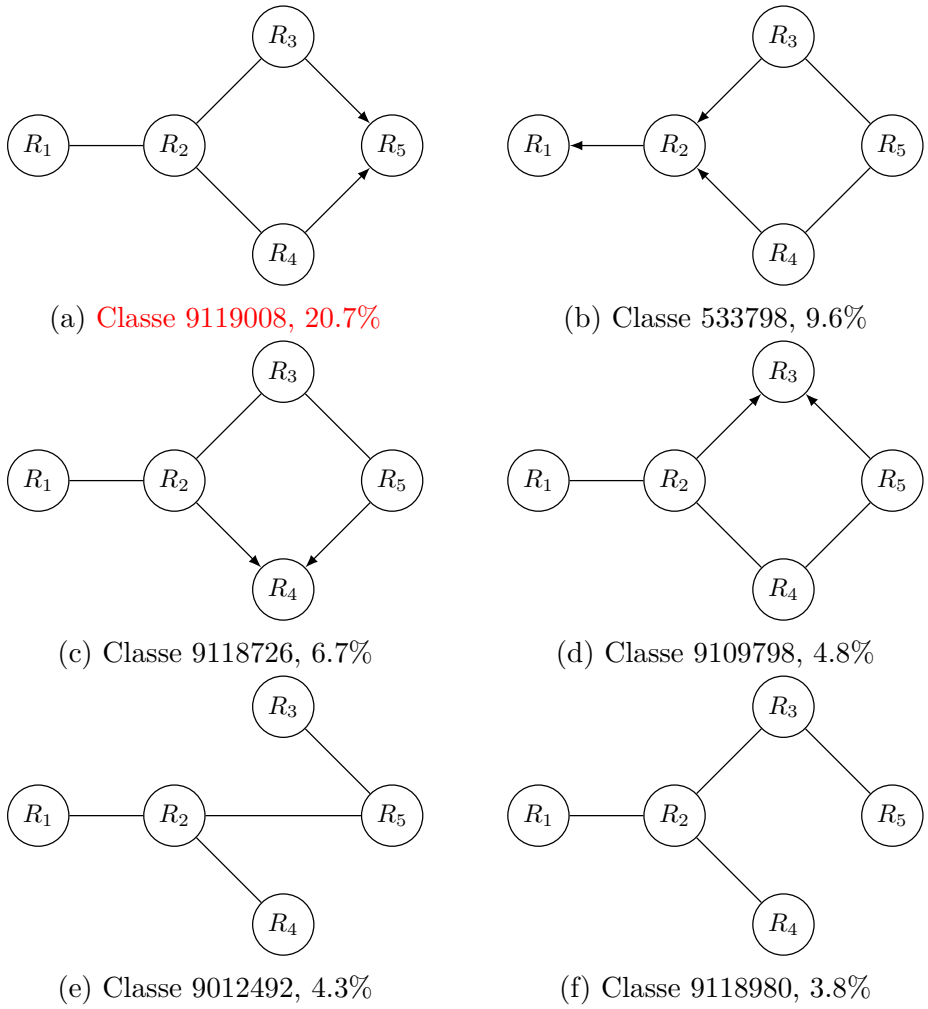
FIGURE 5.22 – DAG (5.22a) et CPDAG (5.22b) de la classe d'équivalence pour la troisième simulation.

Les graphes essentiels les plus retrouvés sont tracés dans la figure 5.23. La classe apprise majoritairement (20.7% des tests) est la vraie classe d'équivalence de Markov. Dans plus de 40% des cas, le bon squelette est obtenu. Dans les CPDAG 9118980 en 5.23f et 9012492 en 5.23b, il manque une ou deux arêtes au squelette, et une arête superflue apparaît en 9012492.

Les CPDAG obtenus grâce au critère KL sont donnés dans la figure 5.19. Le squelette de la classe réelle est partiellement estimé avec ces graphes (en-dehors de celui en 5.24d, avec une arête non existante), mais elle n'y est pas représentée. Pour 92% des tests, seules 4 arêtes ont été estimées, contre 6% pour le nombre réel, 5. L'heuristique sur la pente du score a sous-estimé le nombre d'arêtes. En revanche, lorsque le nombre réel est connu, le taux de bonnes classes d'équivalence est de 35.6%, ce qui est meilleur qu'avec le score BIC.

5.4.2 Application aux données électriques

De part la nature des données et des informations disponibles sur les mesures de la consommation électrique domestique du chapitre 4.4.2, la structure de dépendance sous-jacente entre les puissances relevées sur les différents compteurs a pu être estimée, et représentée sous la forme d'un DAG à la figure 4.8. Nous pouvons raisonnablement supposer que les dépendances entre les ruptures suivent la même structure. Le CPDAG correspondant est noté G_h, et est figuré en 5.25, sans l'hypothétique variable latente du schéma 4.8. Nous proposons d'appliquer la méthode GES sur la segmentation des signaux pour retrouver le modèle d'indépendance entre les variables indicatrices, puis nous comparons la structure obtenue avec G_h.



(a) Classe 9119008, 20.7%

(b) Classe 533798, 9.6%

(c) Classe 9118726, 6.7%

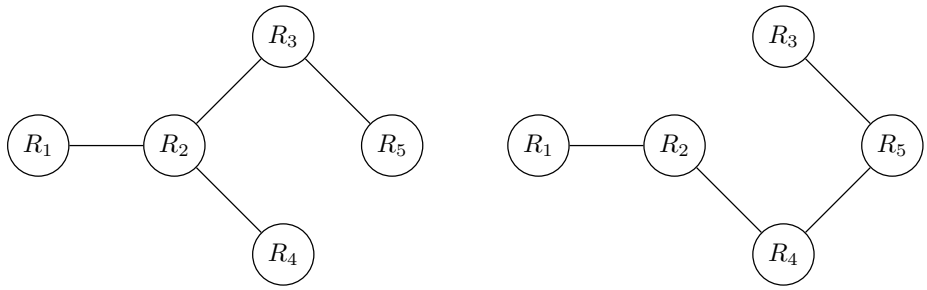
(d) Classe 9109798, 4.8%

(e) Classe 9012492, 4.3%

(f) Classe 9118980, 3.8%

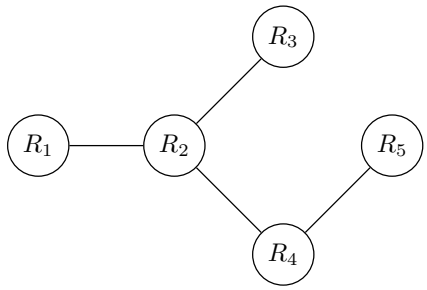
FIGURE 5.23 – CPDAG des classes d'équivalence obtenues le plus souvent avec le score BIC pour la simulation 3, pour 200 séries temporelles. Les proportions sont indiquées dans la légende. Le graphe essentiel réel (en rouge) est en 5.23a.

Afin de disposer d'un grand nombre d'échantillons, la série temporelle analysée comporte $n = 4320$ points, et couvre une période de 3 jours. Dans un premier temps, l'algorithme MultiBD est appliqué, avec $\alpha = 0{,}001$ et 2000 itérations MCMC. 5 résolutions indépendantes ont été menées. Une fois le *Bernoulli Detector* appliqué, les jeux de données $\mathcal{D}_{MAP}$, $\mathcal{D}_{moy}$ et $\mathcal{P}_{BD}$ sont construits sur les 1000 dernières itérations, puis l'algorithme GES est employé. Les solutions avec le score BIC et le score KL sont représentées dans les figures 5.26 et 5.27 respectivement. Le critère KL a permis d'estimer le bon nombre d'arêtes. Sur les 5 tests, le DAG G_h, dont le numéro de classe est le 2184, n'est jamais retrouvé. En revanche les modèles obtenus comportent le bon squelette, seule la V-structure des sommets 2, 3 et 4 vers 1 n'est pas apprise correctement. Une manière de valider l'un de ces modèles consisterait à appliquer l'algorithme MultiBD avec chacun des a priori informatif correspondant aux différentes solutions, puis d'analyser les segmentations résultantes. En

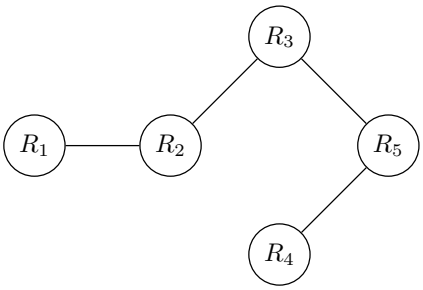


(a) Classe 9118980, 33.2%

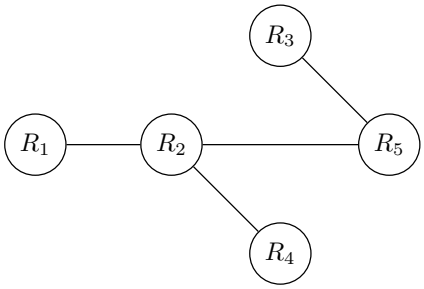
(b) Classe 8979750, 20.7%



(c) Classe 9117986, 11.1%



(d) Classe 9053222, 7.2%



(e) Classe 9012492, 6.7%

FIGURE 5.24 – CPDAG des classes d'équivalence obtenues le plus souvent avec le score KL pour la simulation 3, pour 200 séries temporelles. Les proportions sont indiquées dans la légende. Le graphe essentiel réel n'apparaît pas.

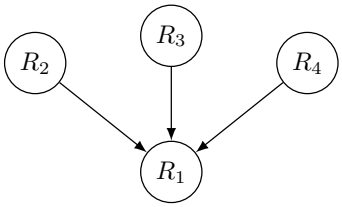
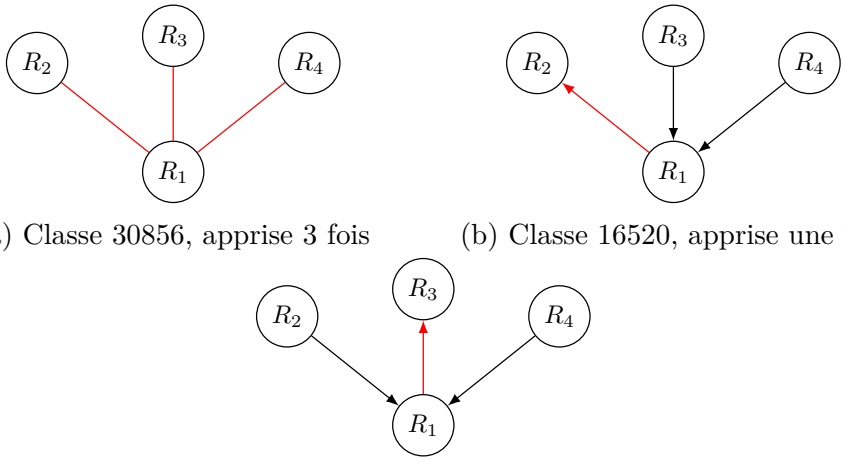


FIGURE 5.25 – DAG (et CPDAG) G_h supposant régir les interactions entre les variables indicatrices des ruptures pour les données électriques.

l'absence de vérité terrain, il toutefois est difficile d'évaluer la segmentation des signaux. Pour ces 5 résolutions, la méthode a échoué à retrouver G_h. Ce résultat négatif est toutefois

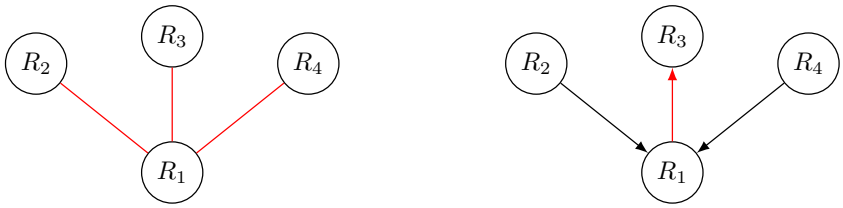
à nuancer en soulignant que le squelette est correctement obtenu, et que la V-structure est souvent apprise en partie, une erreur se produisant sur l'orientation de l'arc $R_2 \to R_1$ ou de l'arc $R_3 \to R_1$. Elle peut s'expliquer par la difficulté de la méthode MultiBD à synchroniser les ruptures entre les signaux 1 et 2, et 1 et 3.



(a) Classe 30856, apprise 3 fois (b) Classe 16520, apprise une fois

(c) Classe 6272, apprise une fois

FIGURE 5.26 – CPDAG des principales classes d'équivalence obtenues pour les données électriques avec la méthode GES et le score BIC. Les arêtes dont l'orientation diffère par rapport à G_h sont en rouge.



(a) Classe 30856, apprise 4 fois (b) Classe 6272, apprise une fois

FIGURE 5.27 – CPDAG des principales classes d'équivalence obtenues pour les données électriques avec la méthode GES et le score KL. Les arêtes dont l'orientation diffère par rapport à G_h sont en rouge.

5.5 Discussion

5.5.1 Conclusions

Dans ce dernier chapitre, j'ai proposé d'exploiter les résultats de l'algorithme MultiBD pour la détection de ruptures dans une série temporelle multivariée en vue de l'apprentissage de la structure de dépendance qui relie les variables indicatrices des ruptures entre

elles. La représentation de ces relations par un réseau bayésien ainsi que la pertinence de ce formalisme appliqué aux variables du *Bernoulli Detector* ont été détaillées. Nous faisons la distinction entre les relations entre les variables $X_{j,i}$ constituant la série temporelle et les variables indicatrices des ruptures $R_{j,i}$. Dans notre modèle, les $X_{j,i}$ sont supposés i.i.d. sur un segment, conditionnellement aux positions des ruptures. Notre étude porte sur les variables $R_{j,i}$. L'un des avantages est que les variables étant binaires et en nombre raisonnable, les probabilités conditionnelles peuvent être calculées. De plus, le fait de ne traiter que les informations portant sur la segmentation permet de s'affranchir des particularités des signaux, et des séries temporelles de natures très différentes sont analysées de la même manière, pourvu que les hypothèses du modèle du *Bernoulli Detector* soient respectées.

La méthode qui a été choisie pour l'apprentissage de la structure est la méthode GES. L'intérêt de l'espace des classes d'équivalence de Markov $\mathcal{C}$ comme espace de recherche est de réduire le nombre d'états à explorer. Bien que le nombre de classes augmente superexponentiellement avec le nombre de variables, le gain en complexité par rapport à l'espace des DAG demeure intéressant pour notre application. En effet le *Bernoulli Detector* est plutôt destiné au traitement de séries temporelles comportant un nombre m raisonnable de signaux afin de limiter le temps de calcul. Pour des valeurs de m inférieures à la dizaine, l'espace de recherche $\mathcal{C}$ est approprié. De plus, la méthode GES ne réalise pas une exploration exhaustive de tout l'espace.

Deux fonctions de score sont présentées : le BIC et le KL. Ces critères sont calculés à partir des données. Dans le cadre de l'exploitation des résultats de MultiBD, deux types de jeux de données sont disponibles : le premier, $\mathcal{D}_{MAP}$, est issu de l'estimateur du MAP de la matrice des indicatrices, de mêmes dimensions que la série temporelle, et le deuxième, $\mathcal{D}_{moy}$, correspond à un ensemble d'échantillons i.i.d. des colonnes de la matrice $\boldsymbol{R}$ obtenus empiriquement. J'ai modifié le score BIC afin d'intégrer les échantillons de $\mathcal{D}_{MAP}$ et de $\mathcal{D}_{moy}$. Le score KL repose lui sur la mesure de la divergence de Kullback-Leibler entre la probabilité jointe $\mathcal{P}_{BD}$ estimée par le *Bernoulli Detector* et la factorisation des paramètres du modèle testé, déduits de $\mathcal{P}_{BD}$. La divergence n'intégrant pas de pénalisation par rapport à la complexité du modèle, une heuristique sur la pente de la fonction de score est nécessaire pour déterminer la meilleure solution globale parmi toutes les solutions locales à un nombre d'arêtes donné.

La méthode GES parvient bien, sous certaines conditions, à retrouver la structure initiale du réseau bayésien régissant les ruptures : les tests avec la vraie segmentation ont montré que le modèle de dépendance peut être déduit des indicatrices des ruptures. Comme elles ne sont pas connues, les jeux de données sont constitués à partir des résultats de la segmentation de $\boldsymbol{X}$ par MultiBD. Plusieurs expérimentations ont été menées. La phase de suppression d'arêtes s'est avérée inutile dans nos exemples. Le nombre de réussites augmente quand le nombre de points de la série temporelle augmente, ce qui est en accord avec les propriétés asymptotiques connues des fonctions de score optimales. Même quand la classe d'équivalence réelle n'est pas retrouvée, elle est en partie estimée. Souvent le squelette est le bon, mais les V-structures sont plus délicates à identifier. Sous réserve que le bon nombre d'arêtes est correctement évalué par l'heuristique de pente, le score KL présente des performances équivalentes voire meilleures que le score BIC. Malheureusement pour

certaines simulations ce critère a tendance à sous-estimer la dimension du graphe.

La méthode GES a également été appliquée sur la segmentation des données de consommation électrique, pour lesquelles la structure de dépendance réelle est définie à partir de notre connaissance du système. Une longue série temporelle de plus de 4000 points a été segmentée par le *Bernoulli Detector*. Les 5 segmentations obtenues indépendamment par MultiBD n'aboutissent pas à un modèle unique ni ne conduisent à la structure attendue. Ce résultat négatif est à nuancer par le fait que le squelette est en général bien retrouvé, et que la V-structure est souvent apprise partiellement. La difficulté de la méthode à retrouver la V-structure complète trouve probablement son origine dans la vérification de l'hypothèse de fidélité.

5.5.2 Perspectives

L'autre hypothèse faite sur les données est la suffisance causale. Celle-ci établit que les causes communes à plusieurs variables de l'ensemble U appartiennent à U ou demeurent constantes. Lorsque cette hypothèse n'est pas vérifiée, le modèle doit être adapté. Des variables latentes, dont l'existence est connue mais qui ne sont pas observées, y sont alors intégrées. L'apprentissage de la structure de dépendance des variables indicatrices du *Bernoulli Detector* à partir de données incomplètes peut être réalisé en adaptant des méthodes dédiées de l'état de l'art [Rubin, 1976, François, 2006]. Les questions des variables de sélection, figées dans un état constant, sont aussi à considérer.

Concernant les dimensions de la série temporelle, nous avons signalé à plusieurs reprises le coût calculatoire prohibitif de l'application de MultiBD sur des séries temporelles contenant un grand nombre de signaux. Cet algorithme pouvant être adapté pour traiter un nombre m de composantes un peu plus important, l'algorithme GES aurait alors à explorer l'espace des classes d'équivalences de Markov à m sommets. Pour simplifier la résolution, il est possible d'éliminer certaines relations d'indépendances par l'application de tests d'indépendances paire à paire sur les vecteurs lignes de $\boldsymbol{R}$. L'apprentissage de la structure reviendrait alors dans certains cas particuliers à retrouver des sous-graphes.

La dimension temporelle des signaux a tendance à favoriser l'obtention du bon modèle, sous réserve qu'un grand nombre d'observations dans $\boldsymbol{X}$ enrichisse la base des différentes configurations des ruptures. Pour détecter ces dernières, le modèle du *Bernoulli Detector* dépend d'un test statistique dont les performances diminuent quand les segments sont de plus en plus petits. Le nombre de ruptures est donc nécessairement faible devant le nombre d'observations n, et la configuration vide $(R_1 = 0, \ldots, R_m = 0)$ est largement surreprésentée. Il serait intéressant d'évaluer la proportion de ruptures nécessaires dans les signaux par rapport à la dimension temporelle et au nombre de configurations vides pour garantir une estimation la plus complète possible du modèle de dépendance. L'augmentation du nombre de points temporels ne dégrade pas le temps de calcul de la méthode GES avec le score BIC, car les jeux de données $\mathcal{D}_{MAP}$ et $\mathcal{D}_{moy}$ peuvent être représentés par le compteur des différentes configurations.

Dans l'exemple des données électriques du chapitre 4.4.1, un a priori informatif a été introduit, de telle sorte que les configurations impossibles, identifiées par notre connaissance

du système, ne sont pas intégrées au modèle, et leurs probabilités P_{ϵ_l} sont nulles. Or pour les réseaux bayésiens, la représentation graphique associée à la distribution jointe est unique si cette dernière est strictement positive (voir [Pearl, 2000, p.15]). Pour appliquer la méthode GES, toutes les configurations, même fortement improbables, ont une probabilité non nulle, ce qui est en contradiction avec l'a priori informatif. Les connaissances sur le système peuvent toutefois être mises à profit, par exemple pour éliminer des relations entre les variables qui sont impossibles et restreindre ainsi l'espace de recherche.

5.5.3 Interprétation causale

La structure de dépendance qui fait l'objet de cette étude régit les interactions entre les variables indicatrices des ruptures. Les données ont été générées à partir de formulations telles que «une rupture sur le signal i se produit simultanément sur le signal j avec la probabilité $p_{i\to j}$», faisant implicitement référence à une relation de cause à effet. Certaines séries temporelles traitées par le *Bernoulli Detector* se prêtent à une interprétation causale, par exemple pour les données des compteurs électriques. À l'inverse l'application sur les données d'hybridation génomique ne permet pas l'élaboration directe de liens de causalité, à moins d'introduire des variables supplémentaires. Lorsque le système étudié est interprétable dans ce contexte, plusieurs remarques peuvent être faites.

La structure apprise par la méthode GES est un CPDAG, dont toutes les arêtes ne sont pas orientées. Des avis d'experts peuvent alors être intégrés pour déterminer les orientations manquantes, de la même façon que nous avons proposé le DAG G_h de la figure 5.25 pour les données électriques. Dans notre étude, les variables sont des indicatrices de changements. La causalité s'exprime donc par l'intermédiaire de ces perturbations, qui se propagent ou non dans plusieurs signaux. Notre modèle spécifie que les changements communs à deux signaux ou plus sont simultanés, par exemple parce que la fréquence d'échantillonnage est insuffisante pour mesurer les décalages dûs à la propagation de la perturbation dans un milieu. Les approches comme celles basées sur la causalité de Granger ne sont donc pas valables.

Dans l'article [Tian et Pearl, 2001], les auteurs se placent dans un cadre similaire au nôtre : des changements spontanés, dû à des phénomènes sous-jacents, se produisent sur plusieurs variables connectées. Ils sont considérés comme des interventions non contrôlées dans un environnement dynamique, et apportent des informations. Une méthode est proposée pour retrouver la structure sous-jacente et tenter de déterminer les directions de causalité. Elle repose sur le fait qu'un changement dans une variable n'affecte que les probabilités marginales de ses descendants. Des variables indicatrices sont alors introduites pour représenter le fait qu'une variable change ou non. Contrairement à notre modèle, ils supposent que tous les changements se propagent parmi les variables descendantes, et parviennent ainsi à orienter certaines arêtes du graphe final. La dépendance à un test de détection de changement est soulignée.

Ce chapitre présente une extension des travaux menés pour la détection de ruptures dans des séries temporelles multivariées, afin d'estimer au moins en partie le graphe des relations d'indépendances conditionnelles entre les variables indicatrices des ruptures. Pour

cela, la méthode GES a été mise en œuvre non pas sur les signaux mais sur les estimations des matrices des indicatrices. Deux critères adaptés aux résultats du *Bernoulli Detector* ont été proposés. Les expériences réalisées ont permis de conclure que l'apprentissage de la structure est possible. Dans certains cas, une interprétation causale du réseau bayésien obtenu peut être menée.

Conclusion et perspectives

Détection robuste de multiples ruptures

Les travaux présentés dans ce document s'inscrivent dans le cadre général de l'étude d'un système complexe à partir d'une série temporelle multivariée. La problématique à laquelle j'ai tâché de répondre se décompose en deux parties : la première traite de la détection de ruptures multiples dans l'ensemble des signaux, et la seconde de l'estimation du graphe dirigé qui décrit les relations de dépendance et d'indépendance conditionnelle entre les entités composant le système. Trois contributions principales ont été proposées. Nous avons d'abord défini le problème de la détection de ruptures, unique puis multiples, de positions connues et inconnues, dans un signal simple puis dans une série temporelle multivariée. Les principales approches de l'état de l'art sont présentées. Le formalisme des réseaux bayésiens, s'appliquant à certaines séries temporelles, est également introduit, ainsi que certaines des méthodes permettant d'estimer le modèle d'indépendance entre les variables du système.

À la suite de cette présentation de l'état de l'art, nous nous sommes intéressés aux tests d'homogénéité comme moyens de déceler une rupture entre deux échantillons, c'est-à-dire de tester l'existence d'une unique rupture à une position donnée dans un signal. Après avoir présenté les tests classiques de Student, de Welch et le test de rang de Wilcoxon-Mann-Whitney, j'ai proposé la méthode VEME reposant sur la fonction de vraisemblance empirique comme une solution intermédiaire aux tests paramétriques et non paramétriques. Les essais réalisés avec des distributions non gaussiennes et avec des échantillons de tailles déséquilibrées ont montré que le test de WMW est le plus robuste, et demeure applicable sur des distributions où ni la moyenne ni la variance ne sont définies. Ce test est de plus rapide à mettre en œuvre, en particulier en l'absence de répétitions dans les rangs et avec l'approximation normale. Le test VEME est quant à lui applicable sur de petits échantillons, contrairement à d'autres approches issues de la vraisemblance empirique, mais dépend d'un critère pour désigner la moyenne empirique de référence, et demande la résolution d'un problème d'optimisation. Le test de WMW est donc finalement retenu pour la détection d'une rupture.

L'intégration de la statistique de test dans un modèle bayésien est le cœur de notre méthode. Les p-valeurs du test de WMW sont évaluées localement sur chaque portion de signal, délimitée par les ruptures, et sont considérées comme des variables aléatoires représentant les observations de la série temporelle. La fonction de vraisemblance est obtenue

en introduisant les distributions des p-valeurs sous H_0 et sous H_1, grâce à un modèle bêta-uniforme. Le vecteur des indicatrices des ruptures R est estimé en maximisant la densité de probabilité a posteriori de R sachant le signal X. La première contribution au problème de la détection de ruptures multiples est ce modèle du *Bernoulli Detector*, développé dans un premier temps pour un signal univarié. Des résultats théoriques garantissent la détection de la première rupture à une position donnée et le contrôle du risque de faire une détection à tort d'une rupture dans un signal sans changement, en fonction du paramètre α, grâce à partir duquel le paramètre γ de la loi bêta est obtenu. Ces résultats sont non asymptotiques, et donc valables y compris sur des signaux de petite taille. Le cas de multiples ruptures n'a pas été résolu, étant donné la complexité du problème et l'impossibilité d'intégrer les dépendances entre les p-valeurs des ruptures voisines.

L'algorithme UniBD permet d'obtenir l'estimateur du maximum a posteriori de R grâce à une approche par MCMC. L'échantillonneur de Gibbs a été modifié pour diminuer la complexité du calcul en négligeant la mise à jour de certaines p-valeurs. Nous avons validé expérimentalement cette approximation. L'algorithme UniBD a été comparé aux méthodes BARD et LASSO de l'état de l'art sur une série de simulations, d'abord sur des données normales puis sur des données présentant des valeurs aberrantes. Les résultats ont été quantifiés en termes de rappel et de précision. Ces tests ont mis en valeur la robustesse du *Bernoulli Detector* conférée par le test de WMW aux données non gaussiennes, suivant une loi de Student. Les méthodes BARD et LASSO peuvent traiter ce type de données, mais nécessitent des modifications, et reposent sur une connaissance préalable de la distribution des observations. L'algorithme UniBD ne renvoie que peu de faux positifs, en revanche les ruptures ne sont pas toujours localisées aussi précisément qu'avec l'algorithme LASSO.

Segmentation de séries temporelles multivariées et intégration du graphe de dépendance

Le modèle a été étendu au traitement de séries temporelles multivariées $\boldsymbol{X}$. Pour ce faire, le paramètre $\boldsymbol{P}$, représentant le vecteur des probabilités d'observer chacune des configurations possibles des variables indicatrices binaires de tous les signaux à un instant donné, est ajouté dans l'expression de la densité a posteriori de $\boldsymbol{R}$ sachant $\boldsymbol{X}$. Nous avons choisi une loi vague a priori pour ce terme. Le modèle du *Bernoulli Detector* en multivarié constitue la deuxième contribution de mes recherches, et répond à la première partie de la problématique. Il conduit à l'estimation la matrice $\boldsymbol{R}$ et fournit de la sorte une segmentation par signal de la série temporelle, où les ruptures communes à plusieurs signaux sont synchronisées tandis que les ruptures propres à une composante n'apparaissent que dans le signal concerné. Le contrôle de la détection de la première rupture sur tout le signal multivarié est démontré pour un niveau dépendant de α. L'algorithme MultiBD, inspiré de UniBD, est fourni. La stratégie d'échantillonnage de Gibbs permet de simuler les vecteurs colonnes $(R_1, \ldots, R_m)'$ de $\boldsymbol{R}$. Le temps de calcul inhérent à cette procédure est élevé, en raison du nombre important de configurations possibles pour chaque colonne, 2^m pour m signaux,

et impose de ne traiter qu'un nombre raisonnable de signaux, par exemple moins d'une dizaine pour 1000 points temporels.

La méthode est appliquée sur des données simulées, qui nous permettent d'aborder la question de la structure de dépendance sous-jacente régissant les éléments du système. Nous la présentons sous la forme d'un graphe orienté reliant les variables indicatrices des ruptures de chaque signal. Dans ce graphe, les arêtes dirigées d'un nœud parent R_i vers un nœud fils R_j s'interprètent comme une relation de dépendance de R_j par rapport à R_i, induisant une probabilité significativement non nulle qu'une rupture sur le signal i s'observe également dans le signal j, exactement au même instant. Nous avons montré expérimentalement comment l'introduction de la connaissance de ce graphe permet d'améliorer la qualité de la détection tout en réduisant le temps de calcul, mais également qu'en l'absence de ces informations l'algorithme MultiBD parvient à attribuer de plus grandes probabilités aux configurations favorisées par le système, même si elles sont sous-estimées. Le modèle s'adapte donc à la série temporelle et nous avons de plus vérifié que la méthode MultiBD bénéficie de l'apprentissage de la structure de dépendance pour améliorer la détection de ruptures dans un signal très bruité par rapport au traitement indépendant par UniBD de chaque signal.

Enfin la méthode a été mise en œuvre sur des données réelles. Le premier jeu de données est un suivi de la consommation électrique d'une habitation, dont nous proposons le graphe de dépendance, qui nous permet d'introduire un a priori informatif dans MultiBD. Le deuxième jeu de données est un ensemble de profils d'hybridation génomique comparative, où des anomalies de nombre de copies de portions d'ADN sont recherchées chez plusieurs patients. Cet exemple met en valeur l'intérêt du *Bernoulli Detector* pour le traitement conjoint de signaux partageant des caractéristiques tout en présentant des segments qui leur sont propres. Pour établir une comparaison, les méthodes du *group* fused LASSO et BARD sont également appliquées. Elles conduisent à l'estimation d'une segmentation unique pour l'ensemble des profils, où apparaissent les ruptures les plus partagées par les signaux, et bénéficient de l'augmentation du nombre de composantes pour améliorer la détection. Plusieurs différences ressortent. Les méthodes *group* fused LASSO et BARD proposent une évaluation globale des possibles défauts dans l'ADN d'un groupe de personnes, tandis que le *Bernoulli Detector* permet d'établir un diagnostic à la fois général et individualisé pour chaque patient, mais au prix d'un très important temps de calcul, ce qui limite le traitement à quelques profils. Le *group* fused LASSO dépend du choix d'un paramètre de régularisation et est sensible aux valeurs aberrantes, mais estime les valeurs moyennes de chaque segment. La méthode BARD classe les segments estimés dans les catégories «normal» et «anormal». La réussite de cette méthode repose sur le choix a priori d'un modèle de vraisemblance, ce qui requiert un minimum de connaissances sur les données. À l'inverse seul le paramètre α est à déterminer pour le MultiBD et la méthode est robuste à certains bruits de nature non gaussienne.

Estimation du modèle d'indépendances conditionnelles

Dans la dernière partie, nous avons entrepris d'approfondir la recherche des relations entre les variables du système complexe. Le contexte diffère des exemples habituels. D'une part les événements se propageant instantanément entre les signaux connectés, par rapport à l'intervalle de temps entre deux observations, aucune information temporelle ne peut donc être exploitée pour retrouver des relations causales. D'autre part les signaux sont considérés comme indépendants conditionnellement aux ruptures ; le modèle d'indépendance du système régit les variables indicatrices des ruptures $R_1, \ldots, R_m$, prises comme des variables aléatoires, et non les variables des séries temporelles. Le système est décrit sous la forme d'un réseau bayésien, dont nous cherchons une représentation par un graphe acyclique dirigé. Pour simplifier la résolution nous appliquons une approximation consistant à considérer les variables indicatrices indépendantes en temps les unes des autres. Comme la matrice réelle des variables indicatrices n'est pas connue, nous utilisons la méthode MultiBD pour générer les échantillons $(R_1, \ldots, R_m)$ sur lesquels le modèle est appris.

Pour répondre à la deuxième partie de la problématique, j'ai repris le principe de la méthode GES, qui explore l'espace des classes d'équivalence de Markov à l'aide d'un opérateur d'ajout d'arc puis d'un opérateur de suppression, jusqu'à la maximisation d'une fonction de score. La troisième contribution réside dans la proposition de deux critères évaluant les résultats du *Bernoulli Detector* pour l'apprentissage du graphe équivalent compatible avec le modèle d'indépendance des données. Le premier est une variante du score BIC, où la fonction de vraisemblance est exprimée selon le comptage des différentes configurations dans la matrice $\hat{\boldsymbol{R}}_{MAP}$ estimées par MultiBD mais aussi dans l'ensemble des matrices $\boldsymbol{R}^{(v)}$ échantillonnées à chaque étape MCMC de l'algorithme lorsque la distribution stationnaire est atteinte. Le second critère est la mesure de la divergence de Kullback-Leibler, qui évalue l'adéquation de la loi jointe $\mathcal{P}_{BD}$, estimée grâce aux échantillons du paramètre $\boldsymbol{P}$, avec la factorisation en paramètres correspondant à un modèle donné, déduite de $\boldsymbol{P}$. Ce critère étant monotone, une heuristique sur la pente de la fonction de score est appliquée pour sélectionner la carte d'indépendance minimale.

Les deux critères BIC et KL ont été testés sur des données simulées, pour plusieurs graphes de dépendance. Les résultats montrent que le squelette est dans la plupart des cas correctement appris, et qu'il est possible de retrouver la classe d'équivalence du graphe réel. Le critère KL a tendance à sous-estimer le nombre d'arêtes, l'heuristique n'étant pas très précise. La mise en œuvre de cette approche sur les données réelles de consommation électrique n'a pas conduit au graphe de dépendance que nous attendions, même si le squelette est le bon, les arêtes de la V-structure étant partiellement ou mal orientées. Cette application ne répond donc pas de manière totalement satisfaisante à la problématique de l'estimation du graphe de dépendance du système, mais la réussite de l'algorithme GES dépend principalement des données, donc d'une part de la capacité de l'algorithme MultiBD à détecter correctement les ruptures pour estimer les échantillons $(R_1, \ldots, R_m)$, et d'autre part du respect des hypothèses de fidélité et de suffisance causale de l'ensemble des échantillons. Sous réserve que ces conditions soient remplies, le graphe partiellement orienté représentant la classe d'équivalence du graphe réel est obtenu.

Perspectives

Certaines interrogations soulevées au long de la présentation des méthodes et les possibilités que ces dernières offrent ouvrent des perspectives intéressantes pour des travaux futurs. Pour commencer, le modèle du *Bernoulli Detector* repose sur une approximation : les dépendances entre les p-valeurs n'ont pas été prises en compte dans la fonction de vraisemblance marginale composite, malheureusement nous ne savons pas les exprimer. Pour parvenir à établir théoriquement les conditions pour le contrôle de la détection de plusieurs ruptures dans un signal, ces dépendances doivent également être inclues. Une piste permettant de simplifier le problème serait de considérer des ruptures suffisamment éloignées pour que leurs influences réciproques soient limitées.

Dans sa version multivariée, l'inconvénient majeur de la méthode du *Bernoulli Detector* est son temps de calcul. En effet, comme la matrice des indicatrices est échantillonnée colonne par colonne dans l'algorithme MultiBD, le nombre de configurations à tester peut aller jusqu'à 2^m dans un contexte non informatif, avec m signaux, auxquelles s'ajoutent les V itérations MCMC. Différentes façons de pallier ce défaut sont envisageables. Tout d'abord il est parfois possible d'introduire des connaissances sur le système impliquant la suppression de certaines configurations, comme nous l'avons fait avec les données électriques. L'échantillonnage de Gibbs peut également s'appliquer individuellement sur chacun des coefficients de la matrice $\boldsymbol{R}$, en ajustant le nombre d'itérations MCMC nécessaires, ou encore en traitant conjointement des sous-groupes de signaux avant de fusionner les segmentations. Les observations temporelles voisines pourraient aussi être groupées dans un bloc à partir duquel un seul terme est estimé, mais la précision dans la position des ruptures en pâtirait. D'autres solutions peuvent s'inspirer des méthodes de résolution bayesiennes variationnelles ou des algorithmes de type *message-passing*.

Le modèle du *Bernoulli Detector* est construit sur la combinaison d'une statistique de test avec une approche bayésienne. Pour détecter des sauts de médianes et être généralisable à d'autres distributions des données que le cas gaussien, nous avons choisi le test de WMW. D'autres tests conviendraient à certains types de séries temporelles, et permettraient de détecter des changements de nature différente, comme des ruptures de variance ou de pente. Cette extension s'obtient relativement facilement, puisque seule l'étape de calcul des p-valeurs est à modifier dans la méthode.

Les séries temporelles de notre étude sont acquises dans leur intégralité et segmentées par le *Bernoulli Detector* rétrospectivement. Il serait intéressant d'examiner une variante de la méthode pour le traitement à la volée des observations, par exemple en n'appliquant l'algorithme que sur les segments les plus récents. Le paramètre $\boldsymbol{P}$ serait alors appris progressivement, et favoriserait la détection simultanée des ruptures communes aux signaux fortement connectés au fur et à mesure de l'arrivée de nouvelles valeurs. Le signal pourrait être représenté par un compteur des configurations estimées, réduisant le stockage des données à un nombre entier par configuration. La méthode GES avec le score BIC ou le critère KL peut s'appliquer sur ces résultats. Cette idée n'est valable que si le graphe des dépendances est statique, et à condition que la résolution par l'algorithme MultiBD, éventuellement modifié selon les propositions précédentes, soit plus rapide que la période

d'échantillonnage de la série temporelle.

La méthode GES a été mise en œuvre dans le dernier chapitre pour l'estimation du modèle d'indépendance, mais il existe d'autres approches, comme les algorithmes IC et PC, qui peuvent être employées. La comparaison des résultats de toutes ces méthodes permettrait d'identifier la stratégie la plus efficace pour exploiter les segmentations issues du *Bernoulli Detector*. Enfin, au sujet de l'application génomique, la recherche d'un graphe non orienté représentant des associations entre les variables indicatrices de tous les patients prendrait peut-être plus de sens que la représentation par un graphe acyclique dirigé, et mérite une attention plus détaillée.

Bibliographie

[Agresti et Gottard, 2007] Agresti, A. et Gottard, A. (2007). Nonconservative exact small-sample inference for discrete data. *Computational statistics & Data analysis*, 51(12):6447–6458.

[Agresti et Kateri, 2011] Agresti, A. et Kateri, M. (2011). *Categorical data analysis*. Springer.

[Akaike, 1998] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. *In Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

[Andersson *et al.*, 1997] Andersson, S. A., Madigan, D. et Perlman, M. D. (1997). A characterization of markov equivalence classes for acyclic digraphs. *Ann. Statist.*, 25(2):505–541.

[Angelosante et Giannakis, 2012] Angelosante, D. et Giannakis, G. B. (2012). Group lassoing change-points in piecewise-constant AR processes. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–16.

[Aravkin *et al.*, 2013] Aravkin, A. Y., Burke, J. V. et Pillonetto, G. (2013). Sparse/-robust estimation and kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory. *The Journal of Machine Learning Research*, 14(1):2689–2728.

[Auvray et Wehenkel, 2002] Auvray, V. et Wehenkel, L. (2002). On the construction of the inclusion boundary neighbourhood for markov equivalence classes of bayesian network structures. *In Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 26–35. Morgan Kaufmann Publishers Inc.

[Bai et Perron, 2003] Bai, J. et Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18(1):1–22.

[Bardwell et Fearnhead, 2014] Bardwell, L. et Fearnhead, P. (2014). Bayesian detection of abnormal segments in multiple time series. *arXiv preprint arXiv:1412.5565*.

[Basseville *et al.*, 1993] Basseville, M., Nikiforov, I. V. *et al.* (1993). *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs.

[Bellera *et al.*, 2010] Bellera, C. A., Julien, M., Hanley, J. A. *et al.* (2010). Normal approximations to the distributions of the Wilcoxon statistics: accurate to what n ? Graphical insights. *Journal of Statistics Education*, 18(2):1–17.

[Bellman, 1954] Bellman, R. (1954). The theory of dynamic programming. Rapport technique, DTIC Document.

[Blaveri *et al.*, 2005] Blaveri, E., Brewer, J. L., Roydasgupta, R., Fridlyand, J., DeVries, S., Koppie, T., Pejavar, S., Mehta, K., Carroll, P., Simko, J. P. *et al.* (2005). Bladder cancer stage and outcome by array-based comparative genomic hybridization. *Clinical cancer research*, 11(19):7012–7022.

[Bleakley et Vert, 2011] Bleakley, K. et Vert, J.-P. (2011). The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*.

[Bouckaert, 1993] Bouckaert, R. R. (1993). Probabilistic network construction using the minimum description length principle. *In Symbolic and quantitative approaches to reasoning and uncertainty*, pages 41–48. Springer.

[Bourguignon et Carfantan, 2005] Bourguignon, S. et Carfantan, H. (2005). Bernoulli-gaussian spectral analysis of unevenly spaced astrophysical data. *In Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*, pages 811–816. IEEE.

[Boyd *et al.*, 2011] Boyd, S., Parikh, N., Chu, E., Peleato, B. et Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

[Brodsky et Darkhovsky, 2013] Brodsky, B. E. et Darkhovsky, B. S. (2013). *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media.

[Brooks *et al.*, 2011] Brooks, S., Gelman, A., Jones, G. et Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis.

[Buntine, 1991] Buntine, W. (1991). Theory refinement on bayesian networks. *In Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*, pages 52–60. Morgan Kaufmann Publishers Inc.

[Carlin *et al.*, 1992] Carlin, B. P., Gelfand, A. E. et Smith, A. F. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied statistics*, pages 389–405.

[Casella et Berger, 2002] Casella, G. et Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

[Chen *et al.*, 2015] Chen, H., Zhang, N. *et al.* (2015). Graph-based change-point detection. *The Annals of Statistics*, 43(1):139–176.

[Chen et Gupta, 1995] Chen, J. et Gupta, A. (1995). Likelihood procedure for testing change point hypothesis for multivariate Gaussian model. *Random Operators and Stochastic Equations*, 3(3):235–244.

[Chen et Gupta, 2011] Chen, J. et Gupta, A. K. (2011). *Parametric statistical change point analysis: With applications to genetics, medicine, and finance*. Springer Science & Business Media.

[Cheng *et al.*, 1997] Cheng, J., Bell, D. A. et Liu, W. (1997). An algorithm for bayesian belief network construction from data. *In proceedings of AI & STAT'97*, pages 83–90. Citeseer.

[Chernoff et Zacks, 1964] Chernoff, H. et Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, pages 999–1018.

[Chickering, 1995] Chickering, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. *In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 87–98. Morgan Kaufmann Publishers Inc.

[Chickering, 2003] Chickering, D. M. (2003). Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554.

[Chickering et Meek, 2002] Chickering, D. M. et Meek, C. (2002). Finding optimal bayesian networks. *In Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 94–102. Morgan Kaufmann Publishers Inc.

[Chow et Liu, 1968] Chow, C. et Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3): 462–467.

[Cooper et Herskovits, 1992] Cooper, G. F. et Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347.

[Csörgö et Horváth, 1997] Csörgö, M. et Horváth, L. (1997). *Limit theorems in change-point analysis*, volume 18. John Wiley & Sons Inc.

[De Campos, 2006] De Campos, L. M. (2006). A scoring function for learning bayesian networks based on mutual information and conditional independence tests. *The Journal of Machine Learning Research*, 7:2149–2187.

[Deshayes et Picard, 1986] Deshayes, J. et Picard, D. (1986). Off-line statistical analysis of change-point models using non parametric and likelihood methods. *In Detection of Abrupt Changes in Signals and Dynamical Systems*, pages 103–168. Springer.

[Desobry *et al.*, 2005] Desobry, F., Davy, M. et Doncarli, C. (2005). An online kernel change detection algorithm. *Signal Processing, IEEE Transactions on*, 53(8):2961–2974.

[DiCiccio *et al.*, 1991] DiCiccio, T., Hall, P. et Romano, J. (1991). Empirical likelihood is bartlett-correctable. *The Annals of Statistics*, 19(2):1053–1061.

[Dobigeon et Tourneret, 2009] Dobigeon, N. et Tourneret, J.-Y. (2009). MCMC sampling for joint segmentation of wind speed and direction. *In Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. DSP/SPE 2009. IEEE 13th*, pages 250–255. IEEE.

[Dobigeon *et al.*, 2007a] Dobigeon, N., Tourneret, J.-Y. et Davy, M. (2007a). Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach. *Signal Processing, IEEE Transactions on*, 55(4): 1251–1263.

[Dobigeon *et al.*, 2007b] Dobigeon, N., Tourneret, J.-Y. et Scargle, J. D. (2007b). Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model. *Signal Processing, IEEE Transactions on*, 55(2):414–423.

[Dor et Tarsi, 1992] Dor, D. et Tarsi, M. (1992). A simple algorithm to construct a consistent extension of a partially oriented graph. *Technicial Report R-185, Cognitive Systems Laboratory, UCLA*.

[Douc *et al.*, 2014] Douc, R., Moulines, E. et Stoffer, D. (2014). *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. CRC Press.

[Dunn, 1961] Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

[Efron *et al.*, 2004] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. *et al.* (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

[Eichler, 2005] Eichler, M. (2005). A graphical approach for evaluating effective connectivity in neural systems. *Philosophical Transactions of the Royal Society of London B : Biological Sciences*, 360(1457):953–967.

[Eilers et De Menezes, 2005] Eilers, P. H. et De Menezes, R. X. (2005). Quantile smoothing of array CGH data. *Bioinformatics*, 21(7):1146–1153.

[Einmahl et McKeague, 2003] Einmahl, J. et McKeague, I. (2003). Empirical likelihood based hypothesis testing. *Bernoulli*, 9(2):267–290.

[Fan et Mackey, 2015] Fan, Z. et Mackey, L. (2015). An Empirical Bayesian Analysis of Simultaneous Changepoints in Multiple Data Sequences. *arXiv preprint arXiv:1508.01280*.

[Fearnhead, 2006] Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2):203–213.

[Fearnhead et Liu, 2007] Fearnhead, P. et Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605.

[François, 2006] François, O. (2006). *De l'identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplètes.* Thèse de doctorat, INSA de Rouen.

[Frick *et al.*, 2014] Frick, K., Munk, A. et Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580.

[Fridlyand *et al.*, 2004] Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G. et Jain, A. N. (2004). Hidden Markov models approach to the analysis of array CGH data. *Journal of multivariate analysis*, 90(1):132–153.

[Friedman *et al.*, 2008] Friedman, J., Hastie, T. et Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

[Frydenberg, 1990] Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, pages 333–353.

[Geiger *et al.*, 1990] Geiger, D., Verma, T. et Pearl, J. (1990). d-separation : From theorems to algorithms. *In Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence*, pages 139–148. North-Holland Publishing Co.

[Gelfand et Smith, 1990] Gelfand, A. E. et Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

[Geman et Geman, 1984] Geman, S. et Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):721–741.

[Gijbels *et al.*, 1999] Gijbels, I., Hall, P. et Kneip, A. (1999). On the estimation of jump points in smooth curves. *Annals of the Institute of Statistical Mathematics*, 51(2):231–251.

[Gilks *et al.*, 1996] Gilks, W., Richardson, S. et Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. Interdisciplinary Statistics Series. Chapman & Hall.

[Gillispie et Perlman, 2001] Gillispie, S. B. et Perlman, M. D. (2001). Enumerating markov equivalence classes of acyclic digraph dels. *In Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 171–177. Morgan Kaufmann Publishers Inc.

[Gombay et Liu, 2000] Gombay, E. et Liu, S. (2000). A nonparametric test for change in randomly censored data. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 113–121.

[Granger, 1969] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica : Journal of the Econometric Society*, pages 424–438.

[Gretton *et al.*, 2006] Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B. et Smola, A. J. (2006). A kernel method for the two-sample-problem. *In Advances in neural information processing systems*, pages 513–520.

[Guan, 2004] Guan, Z. (2004). A semiparametric changepoint model. *Biometrika*, 91(4): 849–862.

[Gustafsson et Gustafsson, 2000] Gustafsson, F. et Gustafsson, F. (2000). *Adaptive filtering and change detection*, volume 1. Wiley New York.

[Hao *et al.*, 2013] Hao, N., Niu, Y. S. et Zhang, H. (2013). Multiple change-point detection via a screening and ranking algorithm. *Statistica Sinica*, 23(4):1553.

[Harari-Kermadec et Pascal, 2008] Harari-Kermadec, H. et Pascal, F. (2008). On the use of empirical likelihood for non-gaussian clutter covariance matrix estimation. *In Proc. of the IEEE-RADAR-08*, Roma, Italy.

[Harchaoui *et al.*, 2013] Harchaoui, Z., Bach, F., Cappé, O. et Moulines, E. (2013). Kernel-based methods for hypothesis testing: a unified view. *IEEE Signal Processing Magazine*, 30(4):87–97.

[Harchaoui et Lévy-Leduc, 2008] Harchaoui, Z. et Lévy-Leduc, C. (2008). Catching change-points with lasso. *In Advances in Neural Information Processing Systems*, pages 617–624.

[Harchaoui et Lévy-Leduc, 2010] Harchaoui, Z. et Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492).

[Harchaoui *et al.*, 2009] Harchaoui, Z., Moulines, E. et Bach, F. R. (2009). Kernel change-point analysis. *In Advances in Neural Information Processing Systems*, pages 609–616.

[Hart, 2001] Hart, A. (2001). Mann-Whitney test is not just a test of medians: differences in spread can be important. *BMJ : British Medical Journal*, 323(7309):391.

[Hastings, 1970] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

[Hawkins, 1977] Hawkins, D. M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, 72(357):180–186.

[Hawkins, 2001] Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Computational Statistics & Data Analysis*, 37(3):323–341.

[Heckerman *et al.*, 1995] Heckerman, D., Geiger, D. et Chickering, D. M. (1995). Learning bayesian networks : The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243.

[Hinkley, 1970] Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17.

[Hung *et al.*, 1997] Hung, H. J., O'Neill, R. T., Bauer, P. et Kohne, K. (1997). The behavior of the P-value when the alternative hypothesis is true. *Biometrics*, pages 11–22.

[James et Matteson, 2013] James, N. A. et Matteson, D. S. (2013). ecp: An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data. Rapport technique, Cornell University.

[Jandhyala *et al.*, 2013] Jandhyala, V., Fotopoulos, S., MacNeill, I. et Liu, P. (2013). Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*.

[Jing, 1995] Jing, B.-Y. (1995). Two-sample empirical likelihood method. *Statistics & probability letters*, 24(4):315–319.

[Kail *et al.*, 2012] Kail, G., Tourneret, J.-Y., Hlawatsch, F. et Dobigeon, N. (2012). Blind deconvolution of sparse pulse sequences under a minimum distance constraint : A partially collapsed gibbs sampler method. *Signal Processing, IEEE Transactions on*, 60(6):2727–2743.

[Kalisch et Bühlmann, 2007] Kalisch, M. et Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 8:613–636.

[Kalman et Bucy, 1961] Kalman, R. E. et Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of basic engineering*, 83(1):95–108.

[Kanamori *et al.*, 2012] Kanamori, T., Suzuki, T. et Sugiyama, M. (2012). f-Divergence Estimation and Two-Sample Homogeneity Test Under Semiparametric Density-Ratio Models. *Information Theory, IEEE Transactions on*, 58(2):708–720.

[Kay, 1998] Kay, S. (1998). *Fundamentals of Statistical Signal Processing: Detection theory*. Fundamentals of Statistical Signal Processing. PTR Prentice-Hall.

[Khodadadi et Asgharian, 2008] Khodadadi, A. et Asgharian, M. (2008). Change-point problem and regression: an annotated bibliography. *COBRA Preprint Series, Paper*, 44.

[Kim et Cohen, 1998] Kim, S.-H. et Cohen, A. S. (1998). On the Behrens-Fisher problem: a review. *Journal of Educational and Behavioral Statistics*, 23(4):356–377.

[Kitamura, 2006] Kitamura, Y. (2006). Empirical Likelihood Methods in Econometrics: Theory and Practice. Cowles Foundation Discussion Papers 1569, Cowles Foundation for Research in Economics, Yale University.

[Kruskal et Wallis, 1952] Kruskal, W. H. et Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

[Kullback et Leibler, 1951] Kullback, S. et Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

[Lam et Bacchus, 1994] Lam, W. et Bacchus, F. (1994). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational intelligence*, 10(3):269–293.

[Lauritzen, 1996] Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press.

[Lavielle et Lebarbier, 2001] Lavielle, M. et Lebarbier, E. (2001). An application of MCMC methods for the multiple change-points problem. *Signal Processing*, 81(1):39–53.

[Lavielle et Teyssiere, 2006] Lavielle, M. et Teyssiere, G. (2006). Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306.

[Lee, 2010] Lee, T.-S. (2010). Change-point problems: bibliography and review. *Journal of Statistical Theory and Practice*, 4(4):643–662.

[Lehmann, 1953] Lehmann, E. L. (1953). The power of rank tests. *The Annals of Mathematical Statistics*, pages 23–43.

[Lehmann et D'Abrera, 1975] Lehmann, E. L. et D'Abrera, H. J. (1975). *Nonparametrics: statistical methods based on ranks*. Holden-Day.

[Lehmann et Romano, 2006] Lehmann, E. L. et Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

[Leray et François, 2004] Leray, P. et François, O. (2004). BNT Structure Learning Package: Documentation and Experiments. *Laboratoire PSI, Tech. Rep.*

[Lichman, 2013] Lichman, M. (2013). UCI machine learning repository. `http://archive.ics.uci.edu/ml`. University of California, Irvine, School of Information and Computer Sciences.

[Lorden, 1971] Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, pages 1897–1908.

[Lung-Yut-Fong *et al.*, 2011a] Lung-Yut-Fong, A., Lévy-Leduc, C. et Cappé, O. (2011a). Robust changepoint detection based on multivariate rank statistics. *In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 3608–3611. IEEE.

[Lung-Yut-Fong *et al.*, 2011b] Lung-Yut-Fong, A., Levy-Leduc, C. et Cappe, O. (2011b). Robust retrospective multiple change-point estimation for multivariate data. *In Statistical Signal Processing Workshop (SSP), 2011 IEEE*, pages 405–408.

[Lung-Yut-Fong *et al.*, 2011c] Lung-Yut-Fong, A., Lévy-Leduc, C. et Cappé, O. (2011c). Homogeneity and change-point detection tests for multivariate data using rank statistics. *arXiv preprint arXiv:1107.1971*.

[Mann et Whitney, 1947] Mann, H. B. et Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

[Marx et Larsen, 2006] Marx, M. L. et Larsen, R. J. (2006). *Introduction to mathematical statistics and its applications*, volume 31. Pearson/Prentice Hall Upper Saddle River, NJ, USA.

[Matteson et James, 2014] Matteson, D. S. et James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

[Meek, 1995] Meek, C. (1995). Causal inference and causal explanation with background knowledge. *In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410. Morgan Kaufmann Publishers Inc.

[Meek, 1997] Meek, C. (1997). *Graphical models : selecting causal and statistical models*. Thèse de doctorat, Carnegie Mellon University.

[Meganck *et al.*, 2006] Meganck, S., Maes, S., Leray, P., Manderick, B., Rouen, I. et St Etienne du Rouvray, F. (2006). Learning semi-markovian causal models using experiments. *In Probabilistic Graphical Models*, pages 195–206. Citeseer.

[Metropolis *et al.*, 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. et Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

[Monahan et Boos, 1992] Monahan, J. F. et Boos, D. D. (1992). Proper likelihoods for bayesian analysis. *Biometrika*, 79(2):271–278.

[Moulines *et al.*, 2008] Moulines, E., Harchaoui, Z. et Bach, F. (2008). Testing for homogeneity with kernel Fisher discriminant analysis. *Advances in Neural Information Processing Systems*.

[Murphy *et al.*, 2001] Murphy, K. *et al.* (2001). The Bayes Net Toolbox for Matlab. *Computing science and statistics*, 33(2):1024–1034.

[Naïm *et al.*, 2011] Naïm, P., Wuillemin, P.-H., Leray, P., Pourret, O. et Becker, A. (2011). *Réseaux bayésiens*. Editions Eyrolles.

[Neyman et Pearson, 1933] Neyman, J. et Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 231(694-706):289–337.

[Ning *et al.*, 2012] Ning, W., Pailden, J. et Gupta, A. (2012). Empirical likelihood ratio test for the epidemic change model. *Journal of Data Science*, 10:107–127.

[Niu et Zhang, 2012] Niu, Y. S. et Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *The annals of applied statistics*, 6(3):1306.

[Owen, 1990] Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120.

[Owen, 2010] Owen, A. (2010). *Empirical Likelihood*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

[Owen, 1988] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

[Page, 1954] Page, E. (1954). Continuous inspection schemes. *Biometrika*, pages 100–115.

[Page, 1957] Page, E. (1957). Estimating the point of change in a continuous process. *Biometrika*, 44(2):248–252.

[Park et Casella, 2008] Park, T. et Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

[Pascal *et al.*, 2010] Pascal, F., Harari-Kermadec, H. et Larzabal, P. (2010). The empirical likelihood method applied to covariance matrix estimation. *Signal Processing*, 90(2):566–578.

[Pauli *et al.*, 2011] Pauli, F., Racugno, W. et Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica*, 21(1):149–164.

[Pearl, 2000] Pearl, J. (2000). *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press.

[Pearl, 2014] Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann. version 2014 de Pearl1988b.

[Pettitt, 1979] Pettitt, A. (1979). A non-parametric approach to the change-point problem. *Applied statistics*, pages 126–135.

[Picard *et al.*, 2011] Picard, F., Lebarbier, E., Hoebeke, M., Rigaill, G., Thiam, B. et Robin, S. (2011). Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics*, 12(3).

[Pollack *et al.*, 1999] Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D. et Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature genetics*, 23(1):41–46.

[Portnoy *et al.*, 1997] Portnoy, S., Koenker, R. *et al.* (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Ramanayake et Gupta, 2003] Ramanayake, A. et Gupta, A. K. (2003). Tests for an epidemic change in a sequence of exponentially distributed random variables. *Biometrical journal*, 45(8):946–958.

[Ribatet *et al.*, 2012] Ribatet, M., Cooley, D. et Davison, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22(2):813–845.

[Robert, 2006] Robert, C. P. (2006). *Le choix bayésien*. Springer.

[Robinson, 1977] Robinson, R. W. (1977). Counting unlabeled acyclic digraphs. *In Combinatorial mathematics V*, pages 28–43. Springer.

[Rubin, 1976] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

[Rudin *et al.*, 1992] Rudin, L. I., Osher, S. et Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268.

[Sackrowitz et Samuel-Cahn, 1999] Sackrowitz, H. et Samuel-Cahn, E. (1999). P values as random variables—expected p values. *The American Statistician*, 53(4):326–331.

[Schwarz *et al.*, 1978] Schwarz, G. *et al.* (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Sellke *et al.*, 2001] Sellke, T., Bayarri, M. et Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71.

[Shah *et al.*, 2007] Shah, S. P., Lam, W. L., Ng, R. T. et Murphy, K. P. (2007). Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics*, 23(13):i450–i458.

[Shah *et al.*, 2006] Shah, S. P., Xuan, X., DeLeeuw, R. J., Khojasteh, M., Lam, W. L., Ng, R. et Murphy, K. P. (2006). Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, 22(14):e431–e439.

[Shajarisales *et al.*, 2015] Shajarisales, N., Janzing, D., Schoelkopf, B. et Besserve, M. (2015). Telling cause from effect in deterministic linear dynamical systems.

In Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pages 285–294.

[Shewhart, 1931] Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. ASQ Quality Press.

[Siegel, 1956] Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. McGraw-hill.

[Siegmund, 1988] Siegmund, D. (1988). Confidence sets in change-point problems. *International Statistical Review/Revue Internationale de Statistique*, pages 31–48.

[Sonntag *et al.*, 2015] Sonntag, D., Peña, J. M. et Gómez-Olmedo, M. (2015). Approximate Counting of Graphical Models Via MCMC Revisited. *International Journal of Intelligent Systems*, 30(3):384–420.

[Spirtes *et al.*, 2000] Spirtes, P., Glymour, C. N. et Scheines, R. (2000). *Causation, prediction, and search*, volume 81. MIT press.

[Srivastava et Worsley, 1986] Srivastava, M. et Worsley, K. J. (1986). Likelihood ratio tests for a change in the multivariate normal mean. *Journal of the American Statistical Association*, 81(393):199–204.

[Stransky *et al.*, 2006] Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., de Medina, S. G. D., Segraves, R., de Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C. *et al.* (2006). Regional copy number–independent deregulation of transcription in cancer. *Nature genetics*, 38(12):1386–1396.

[Student, 1908] Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.

[Tian et Pearl, 2001] Tian, J. et Pearl, J. (2001). Causal discovery from changes. *In Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 512–521. Morgan Kaufmann Publishers Inc.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[Tibshirani *et al.*, 2005] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. et Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

[Tibshirani et Wang, 2008] Tibshirani, R. et Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18–29.

[Tibshirani, 2011] Tibshirani, R. J. (2011). *The solution path of the generalized lasso*. Stanford University.

[Tibshirani et Arnold, 2014] Tibshirani, R. J. et Arnold, T. B. (2014). *genlasso: Path algorithm for generalized lasso problems*. R package version 1.2.

[Tong et Koller, 2001] Tong, S. et Koller, D. (2001). Active learning for structure in bayesian networks. *In Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, pages 863–869, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Van der Vaart, 1950] Van der Vaart, H. (1950). Some remarks on the power function of Wilcoxon's test for the problem of two samples. *In Nederl. Akad. Wetensch., Proc., Ser. A*, volume 53, pages 494–520.

[Varin, 2008] Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1):1–28.

[Varin *et al.*, 2011] Varin, C., Reid, N. et Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

[Verma et Pearl, 1990] Verma, T. et Pearl, J. (1990). Equivalence and synthesis of causal models. *In Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270. Elsevier Science Inc.

[Verma et Pearl, 1992] Verma, T. et Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. *In Proceedings of the Eighth international conference on uncertainty in artificial intelligence*, pages 323–330. Morgan Kaufmann Publishers Inc.

[Vert et Bleakley, 2010] Vert, J.-P. et Bleakley, K. (2010). Fast detection of multiple change-points shared by many signals using group LARS. *In Advances in Neural Information Processing Systems*, pages 2343–2351.

[Vostrikova, 1981] Vostrikova, L. (1981). Detection of the disorder in multidimensional random-processes. *Doklady Akademii Nauk SSSR*, 259(2):270–274.

[Wald, 1973] Wald, A. (1973). *Sequential analysis*. Courier Corporation.

[Welch, 1947] Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika*, pages 28–35.

[Wikimedia, 2012] Wikimedia (2012). Array-CGH protocol. [Online ; accessed 15-February-2016].

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83.

[Wilks, 1938] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

[Willsky et Jones, 1976] Willsky, A. S. et Jones, H. L. (1976). A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *Automatic Control, IEEE Transactions on*, 21(1):108–112.

[Yao, 1993] Yao, Q. (1993). Tests for change-points with epidemic alternatives. *Biometrika*, 80(1):179–191.

[Yuan et Lin, 2007] Yuan, M. et Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

[Zou *et al.*, 2007] Zou, C., Liu, Y., Qin, P. et Wang, Z. (2007). Empirical likelihood ratio test for the change-point problem. *Statistics & probability letters*, 77(4):374–382.

[Zou, 2006] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Publications

[Harlé *et al.*, 2014] Harlé, F., Chatelain, F., Gouy-Pailler, C. et Achard, S. (2014). Rank-based multiple change-point detection in multivariate time series. *In Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE.

[Harlé *et al.*, 2015] Harlé, F., Chatelain, F., Gouy-Pailler, C. et Achard, S. (2015). Utilisation de la vraisemblance empirique pour un test d'homogénéité. *In GRETSI 2015*.

[Harlé *et al.*, 2016] Harlé, F., Chatelain, F., Gouy-Pailler, C. et Achard, S. (2016). Bayesian Model for Multiple Change-points Detection in Multivariate Time Series. *Signal Processing, IEEE Transactions on*.

Annexe A

Démonstration de la proposition 2.3.1

La démonstration de la proposition 2.3.1 s'inspire de celle du théorème 2.3.1, donnée dans [Owen, 2010, p. 226]. La première partie (équation (2.35)) consiste à montrer que

$$-2\ln(\mathcal{R}(\overline{X}_1)) = n\frac{\left(\overline{X} - \overline{X}_1\right)^2}{S^2} + o_p(1), \tag{A.1}$$

où

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_1)^2. \tag{A.2}$$

Pour y parvenir, une majoration du multiplicateur de Lagrange $\lambda \in \mathbb{R}$ (voir l'équation (2.25))permet de conclure que $\lambda = O_p(n^{-1/2})$. Ce terme est ensuite exprimé sous la forme $\lambda = \frac{\overline{X}-\overline{X}_1}{S^2} + o_p(n^{-1/2})$, puis cette expression est introduite dans la fonction de profil $-2\ln(\mathcal{R}(\overline{X}_1))$. Un développement de Taylor permet alors de parvenir à l'équation (A.38).

Comme $\overline{X}_1$ est la moyenne empirique de l'échantillon X_1, alors $\overline{X}_1$ est un point intérieur de l'enveloppe convexe des variables X_i. Il existe alors un ensemble unique de masses de probabilités $p_i > 0$ qui vérifient les contraintes $\sum_{i=1}^n p_i X_i = \overline{X}_1$, $p_i > 0$, et $\sum_{i=1}^n p_i = 1$ et qui maximisent le produit $\prod_{i=1}^n np_i$. Ils peuvent s'écrire comme

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda(X_i - \overline{X}_1)}. \tag{A.3}$$

Le multiplicateur de Lagrange satisfait l'équation $g(\lambda) = 0$ où $g(\lambda)$ est donné par

$$g(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \overline{X}_1}{1 + \lambda(X_i - \overline{X}_1)}. \tag{A.4}$$

On pose

$$Y_i = \lambda(X_i - \overline{X}_1). \tag{A.5}$$

Dans $g(\lambda)$, on a donc

$$g(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \overline{X}_1}{1 + Y_i}. \tag{A.6}$$

Comme

$$\frac{1}{1+Y_i} = 1 - \frac{Y_i}{1+Y_i}, \tag{A.7}$$

on parvient à

$$g(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_1 - \frac{X_i - \overline{X}_1}{1+Y_i}\right) \tag{A.8}$$

donc

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_1) = \lambda\frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \overline{X}_1)^2}{1+Y_i}, \tag{A.9}$$

$$\overline{X} - \overline{X}_1 = \lambda\frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \overline{X}_1)^2}{1+Y_i}. \tag{A.10}$$

Comme toutes les masses p_i sont strictement positives, $1+Y_i > 0$, donc pour tout $1 \leq i \leq n$,

$$(X_i - \overline{X}_1)^2 \leq (X_i - \overline{X}_1)^2\frac{1 + \max_{1\leq i\leq n}|Y_i|}{1+Y_i}, \tag{A.11}$$

$$\sum_{i=1}^{n}(X_i - \overline{X}_1)^2 \leq \sum_{i=1}^{n}\frac{(X_i - \overline{X}_1)^2}{1+Y_i}(1 + \max_{1\leq i\leq n}|Y_i|). \tag{A.12}$$

En introduisant le terme

$$Z_n^\star = \max_{1\leq i\leq n}|X_i - \overline{X}_1|, \tag{A.13}$$

et S^2, défini en (A.2), l'inégalité (A.12) devient :

$$|\lambda|S^2 \leq |\lambda|\frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \overline{X}_1)^2}{1+Y_i}(1 + |\lambda|Z_n^\star). \tag{A.14}$$

D'après (A.10), on obtient

$$|\lambda|S^2 \leq |\lambda||\overline{X} - \overline{X}_1|(1 + |\lambda|Z_n^\star), \tag{A.15}$$

d'où

$$|\lambda|(S^2 - Z_n^\star|\lambda||\overline{X} - \overline{X}_1|) \leq |\lambda||\overline{X} - \overline{X}_1|. \tag{A.16}$$

En appliquant le théorème central limite à $\overline{X} - \overline{X}_1$, on en déduit que $|\lambda||\overline{X} - \overline{X}_1| = O_p(n^{-1/2})$. La vitesse de convergence de $Z_n^\star$ est déduite du lemme suivant [Owen, 2010, Lemme 11.2, page 225] :

Lemme A.1. *Soient les variables aléatoires indépendantes Y_i de même distribution et telles que $E[Y_i^2] < \infty$. Soit $Z_n = \max_{1\leq i\leq n}|Y_i|$. Alors $Z_n = o(n^{1/2})$.*

D'après ce lemme, $Z_n^\star = o_p(n^{1/2})$. On obtient :

$$|\lambda|(S^2 + o_p(1)) = O_p(n^{-1/2}). \tag{A.17}$$

Comme les variances des lois F_1 et F_2 existent et sont finies, et que S^2 est une moyenne empirique, alors $S^2 = O_p(1)$ et

$$|\lambda| = O_p(n^{-1/2}). \tag{A.18}$$

D'après le lemme A.1 et (A.18) :

$$\max_{1\leq i\leq n} |Y_i| = O_p(n^{-1/2})o(n^{1/2}) = o_p(1). \tag{A.19}$$

En introduisant deux fois la relation (A.7) dans (A.6), et comme $g(\lambda) = 0$, on obtient :

$$0 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_1)(1 - Y_i + \frac{Y_i^2}{1+Y_i}) \tag{A.20}$$

$$= \overline{X} - \overline{X}_1 - S^2\lambda + \frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \overline{X}_1)Y_i^2}{1+Y_i} \tag{A.21}$$

$$\lambda = \frac{\overline{X} - \overline{X}_1}{S^2} + \frac{1}{S^2 n}\sum_{i=1}^{n}\frac{(X_i - \overline{X}_1)Y_i^2}{1+Y_i} \tag{A.22}$$

$$\tag{A.23}$$

Le dernier terme dans (A.22) est borné par :

$$\frac{1}{S^2 n}\sum_{i=1}^{n}\frac{(X_i - \overline{X}_1)Y_i^2}{1-Y_i} \leq \frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}_1|^3\lambda^2|1-Y_i|^{-1}. \tag{A.24}$$

Le lemme [Owen, 2010, Lemme 11.3, page 225], donné ci-après, est une conséquence du lemme A.1.

Lemme A.2. *Soient les variables aléatoires indépendantes Y_i de même distribution et telles que $E[Y_i^2] < \infty$. Alors $\frac{1}{n}\sum_{i=1}^{n}|Y_i|^3 = o(n^{1/2})$.*

On en déduit pour (A.24) :

$$\frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}_1|^3\lambda^2|1-Y_i|^{-1} = o(n^{1/2})O_p(n^{-1})O_p(1) = o_p(n^{-1/2}), \tag{A.25}$$

λ s'écrit donc sous la forme :

$$\lambda = \frac{\overline{X} - \overline{X}_1}{S^2} + \beta \tag{A.26}$$

où

$$\beta = o_p(n^{-1/2}). \tag{A.27}$$

Un développement de Taylor permet d'écrire :

$$\ln(1+Y_i) = Y_i - \frac{1}{2}Y_i^2 + \eta_i, \tag{A.28}$$

où, pour une constante $B > 0$ finie et pour tout $1 \leq i \leq n$,

$$\lim_{n\to\infty}\Pr(|\eta_i| \leq B|Y_i|^3) = 1. \tag{A.29}$$

La fonction de profil du test VEME peut donc s'écrire :

$$-2\ln(\mathcal{R}(\overline{X}_1)) = -2\sum_{i=1}^{n}\ln(np_i) \tag{A.30}$$

$$= 2\sum_{i=1}^{n}\ln(1+Y_i) \tag{A.31}$$

$$= 2\sum_{i=1}^{n}Y_i - \sum_{i=1}^{n}Y_i^2 + 2\sum_{i=1}^{n}\eta_i \tag{A.32}$$

$$= 2\sum_{i=1}^{n}\lambda(X_i-\overline{X}_1) - \sum_{i=1}^{n}\lambda^2(X_i-\overline{X}_1)^2 + 2\sum_{i=1}^{n}\eta_i \tag{A.33}$$

$$= 2n\lambda(\overline{X}-\overline{X}_1) - n\lambda^2S^2 + 2\sum_{i=1}^{n}\eta_i. \tag{A.34}$$

D'après (A.26) :

$$-2\ln(\mathcal{R}(\overline{X}_1)) = n\frac{(\overline{X}-\overline{X}_1)^2}{S^2} - n\beta^2S^2 + 2\sum_{i=1}^{n}\eta_i. \tag{A.35}$$

Comme

$$n\beta^2S^2 = o(n)o_p(n^{-1})O_p(1) = o(1), \tag{A.36}$$

et

$$|\sum_{i=1}^{n}\eta_i| \leq B|\lambda|^3\sum_{i=1}^{n}(X_i-\overline{X}_1)^3 = O_p(n^{-3/2})o_p(n^{3/2}) = o_p(1), \tag{A.37}$$

alors

$$-2\ln(\mathcal{R}(\overline{X}_1)) = n\frac{\left(\overline{X}-\overline{X}_1\right)^2}{S^2} + o_p(1). \tag{A.38}$$

La deuxième partie de la démonstration, à la section 2.3.4, consiste à montrer que S^2 converge en probabilité vers la variance σ^2, puis par application du théorème de Slutsky on parvient au résultat

$$\frac{n_1}{n_2}\frac{n\left(\overline{X}-\overline{X}_1\right)^2}{S^2} \xrightarrow[n\to+\infty]{Loi} \chi^2_{(1)}. \tag{A.39}$$

Annexe B

Méthodes MCMC

Les méthodes MCMC sont apparues dans les années 50, accompagnant les progrès informatiques en puissance de calcul et la plus grande accessibilité aux ordinateurs. Elles sont couramment employées dans le cadre de l'inférence bayésienne, mais également dans des applications en statistique, en physique (dont est issue une partie du vocabulaire) ou en chimie. Le principe consiste à générer des chaînes de Markov $(\boldsymbol{R}^{(v)})_v$ dont la loi stationnaire est la distribution d'intérêt, c'est-à-dire, dans notre contexte bayésien, la loi a posteriori $f(\boldsymbol{R}|\boldsymbol{X})$; l'idée étant que pour un nombre suffisamment grand d'itérations u, la chaîne résultante $\boldsymbol{R}^{(u)}$ est approximativement distribuée selon $f(\boldsymbol{R}|\boldsymbol{X})$, quelle que soit la valeur initiale $\boldsymbol{R}^{(0)}$. $\hat{\boldsymbol{R}}_{MAP}$ correspond alors à la solution qui maximise $f(\boldsymbol{R}|\boldsymbol{X})$ parmi toutes les solutions simulées dans la chaîne de Markov. Plusieurs ouvrages, comme [Gilks *et al.*, 1996, Brooks *et al.*, 2011], présentent ce type de méthodes et ses applications, dans un contexte plus large que l'inférence bayésienne.

La chaîne de Markov $(\boldsymbol{R}^{(v)})_v$ est composée de la séquence des variables aléatoires $\{\boldsymbol{R}^{(0)},\boldsymbol{R}^{(1)},\ldots\}$, où l'état $\boldsymbol{R}^{(v+1)}$ est généré à partir de l'état précédent par la probabilité de transition $f(\boldsymbol{R}^{(v+1)}|\boldsymbol{R}^{(v)})$. L'état futur $\boldsymbol{R}^{(v+1)}$ est donc indépendant des états antérieurs $\{\boldsymbol{R}^{(0)},\ldots,\boldsymbol{R}^{(v-1)}\}$ conditionnellement à l'état présent $\boldsymbol{R}^{(v)}$. L'ensemble dans lequel $\boldsymbol{R}^{(v)}$ prend ses valeurs est appelé l'espace de états. Si la chaîne remplit les conditions suivantes, alors elle converge vers une distribution stationnaire $f(.)$, qui ne dépend ni de v, ni de $\boldsymbol{R}^{(0)}$ (voir [Gilks *et al.*, 1996, définition 3.1, p. 46]) :

- elle est *irréductible*, tous les états peuvent être atteints en un certain nombre d'itérations, quelle que soit l'initialisation $\boldsymbol{R}^{(0)}$;
- elle est *apériodique*, et n'oscille donc pas entre les mêmes états,
- elle est *récurrente positive*, ce qui est équivalent à dire que si l'état initial $\boldsymbol{R}^{(0)}$ est généré à partir de la distribution stationnaire $f(.)$, alors les états suivants sont distribués selon $f(.)$.

Dans une méthode MCMC, la loi stationnaire est $f(\boldsymbol{R}|\boldsymbol{X})$. La procédure consiste donc à initialiser $\boldsymbol{R}^{(0)}$, puis à générer la chaîne $(\boldsymbol{R}^{(v)})_v$ en simulant l'état futur $\boldsymbol{R}^{(v+1)}$ à partir de l'état précédent selon la probabilité de transition $f(\boldsymbol{R}^{(v+1)}|\boldsymbol{R}^{(v)})$. Au-delà de l'itération u, correspondant à une période de préchauffe, la loi stationnaire $f(\boldsymbol{R}|\boldsymbol{X})$ est atteinte, et les $\boldsymbol{R}^{(v)}$, $v \geq u$, sont alors échantillonnés d'après l'approximation de la densité de probabilité a

posteriori. L'estimateur du MAP peut donc être déterminé comme la solution qui maximise la distribution à l'équilibre.

Une difficulté réside dans l'évaluation du nombre d'itérations u nécessaires pour atteindre la loi stationnaire, et pour lesquelles on ne tiendra pas compte de résultats. Dans certains cas la vitesse de convergence, qui mesure le taux de décroissance de la distance entre la loi de $\boldsymbol{R}^{(v)}$ et sa limite, peut être déterminée théoriquement, mais elle dépend en général de $\boldsymbol{R}^{(0)}$. Cette convergence peut être très rapide, pour peu que l'initialisation se fasse dans un état proche du centre de la distribution à l'équilibre. En pratique dans notre algorithme on se contente de fixer empiriquement un nombre d'itérations V assez grand, puis on réalise les V simulations, en mettant à jour $\hat{\boldsymbol{R}}_{MAP}$ tel que

$$\hat{\boldsymbol{R}}_{MAP}(V) = \underset{0 \leq v \leq V}{\operatorname{argmax}} f(\boldsymbol{R}^{(v)}|\boldsymbol{X}). \tag{B.1}$$

On s'assure de ne terminer le processus que lorsque la valeur de $\hat{\boldsymbol{R}}_{MAP}$ n'est plus modifiée assez fréquemment.

La méthode MCMC permet d'approcher la densité de probabilité a posteriori $f(\boldsymbol{R}|\boldsymbol{X})$ par la chaîne de Markov $(\boldsymbol{R}^{(v)})_v$, de loi stationnaire $f(\boldsymbol{R}|\boldsymbol{X})$. Pour générer de telles chaînes, plusieurs stratégies sont possibles. La première qui a été développée est celle de Metropolis-Hastings, initialement introduite dans [Metropolis *et al.*, 1953] en physique statistique puis généralisée dans [Hastings, 1970]. Cet algorithme repose sur l'emploi de ratios de densités de probabilité, ce qui permet de s'affranchir des constantes de normalisation des lois. Ainsi, la distribution $f(\boldsymbol{R}^{(v)}|\boldsymbol{X})$ est remplacée par le membre de droite de l'expression (3.24). Cette densité non normalisée, notée $f(\boldsymbol{R}^{(v)}|\boldsymbol{X})$, est strictement positive et constitue la loi cible. Pour effectuer un mouvement de l'état courant $\boldsymbol{R}^{(v)}$ vers l'état $\boldsymbol{R}'$, une densité de probabilité conditionnelle $f(.|\boldsymbol{R}^{(v)};\boldsymbol{X})$ intervient, appelée la loi de proposition. Le ratio de Hastings est défini comme

$$\rho(\boldsymbol{R}^{(v)},\boldsymbol{R}') = \frac{f(\boldsymbol{R}'|\boldsymbol{X}) f(\boldsymbol{R}'|\boldsymbol{R}^{(v)};\boldsymbol{X})}{f(\boldsymbol{R}^{(v)}|\boldsymbol{X}) f(\boldsymbol{R}^{(v)}|\boldsymbol{R}';\boldsymbol{X})}. \tag{B.2}$$

Le mécanisme de mise à jour de $\boldsymbol{R}$ consiste à accepter le mouvement $\boldsymbol{R}'$ avec la probabilité

$$a(\boldsymbol{R}^{(v)},\boldsymbol{R}') = \min\left(1, \rho(\boldsymbol{R}^{(v)},\boldsymbol{R}')\right). \tag{B.3}$$

Selon la valeur du ratio de Hastings, l'état final $\boldsymbol{R}^{(v+1)}$ est donc soit l'état proposé $\boldsymbol{R}'$, soit l'état initial $\boldsymbol{R}^{(v)}$. La distribution stationnaire de la chaîne de Markov $(\boldsymbol{R}^{(v)})_v$ générée de la sorte est la loi cible si la chaîne est irréductible. On s'assure ainsi que tout le support de $f(\boldsymbol{R}^{(v)}|\boldsymbol{X})$ est exploré. Des conditions sur les supports des lois cible et de proposition sont suffisantes pour garantir l'irréductibilité de la chaîne. Il existe plusieurs façon de proposer le mouvement, le plus souvent une marche aléatoire est effectuée, de façon à explorer les alentours de la valeur courante $\boldsymbol{R}^{(v)}$. Une variante assez répandue de cet algorithme est connue sous le nom d'échantillonneur de Gibbs [Geman et Geman, 1984, Gelfand et Smith, 1990], et ne nécessite pas de mettre en place une fonction d'exploration. C'est celle que nous avons retenue pour la mise en œuvre du modèle du *Bernoulli Detector*.

Annexe C

Étude de la complexité calculatoire du *Bernoulli Detector*

L'expression de la loi de probabilité a posteriori du modèle du *Bernoulli Detector* pour des séries temporelles multivariées est la suivante :

$$f(\boldsymbol{R}|\boldsymbol{X}) \propto \frac{\prod_{l=1}^{L}\Gamma(S_l(\boldsymbol{R})+1)}{\Gamma(n+L)}\prod_{j=1}^{m}\prod_{i=1}^{n}(\gamma P_{j,i}(\boldsymbol{X},\boldsymbol{R})^{\gamma-1})^{R_{j,i}}. \tag{C.1}$$

La maximisation de cette expression est réalisée par une méthode MCMC, en suivant la stratégie de l'échantillonneur de Gibbs, auquel une approximation est appliquée. L'algorithme MultiBD, présenté en 5, est rappelé par le pseudo-code 9. La complexité calculatoire de cet algorithme est évaluée à partir des complexités de chaque étape, détaillées dans le tableau C.1. Le test statistique appliqué est celui de WMW. L'étape optionnelle où le paramètre $\boldsymbol{P}$ est simulé d'après la loi (4.13) n'est pas inclue.

Ligne	Étape	Complexité
1	Choix du seuil d'acceptance α	$O(1)$
2	Initialisation de $\boldsymbol{R}$ et $\hat{\boldsymbol{R}}_{MAP}$	$O(2mn)$
4	Permutation aléatoire des indices temporels	$O(n)$
6	Conversion d'indice	$O(1)$
8	Calcul de la p-valeur d'après le test de WMW	$O(n\log n)$
10	Simuler le i^{e} vecteur colonne de $\boldsymbol{R}$ d'après (4.40)	$O(2Lm)$
11	Calculer la probabilité a posteriori de $\boldsymbol{R}$ après avoir mis à jour la i^{e} colonne	$O(2mn\log n)$
12	Comparaison de la probabilité a posteriori avec la plus grande des valeurs précédentes	$O(1)$
13	Mise à jour de l'estimateur du MAP	$O(nm)$

Table C.1 – Complexité calculatoire de chaque étape de l'algorithme MultiBD 9.

Les p-valeurs sont issues du test de WMW : en (j,i), $P_i(\boldsymbol{X}_{j,-},\boldsymbol{X}_{j,+})$ est obtenue en comparant les rangs des coefficients des segments $\boldsymbol{X}_{j,-}$ et $\boldsymbol{X}_{j,+}$ autour du point i dans le j^{e} signal. Pour obtenir les rangs des observations, un algorithme de tri est appliqué sur le segment total, rassemblant les coefficients de $\boldsymbol{X}_{j,-}$ et $\boldsymbol{X}_{j,+}$. Dans le pire des cas, le signal ne présente pas de rupture, alors le tri des coefficients a une complexité en $O(n \log n)$. Le coût du calcul de la p-valeur à partir de la somme des rangs est alors négligeable devant le coût du tri.

La simulation du vecteur colonne $\boldsymbol{R}_p$ à l'itération v requiert au préalable de calculer la probabilité a posteriori de chaque configuration possible pour $\boldsymbol{R}_p$, puis à simuler $\boldsymbol{R}_p$ selon ces probabilités. Avec l'approximation de l'échantillonneur de Gibbs, proposée dans la partie 3.3.3, seules les p-valeurs d'indice temporel $i = \pi^{-1}(p)$ sont évaluées. Elles sont calculées à l'étape 8 de l'algorithme. La probabilité a posteriori pour une configuration donnée est donc approchée en mettant à jour les vraisemblances de la colonne i et le compteur des configurations, par rapport à la configuration prise par la colonne $\boldsymbol{R}_{\bullet,i}$ à l'itération $v-1$. L'étape de simulation 10 nécessite le calcul des L probabilités a posteriori associées aux L configurations possibles. Ce calcul consiste à mettre à jour la valeur courante du logarithme de $f(\boldsymbol{R}|\boldsymbol{X})$, en réalisant m opérations de modification des fonctions de vraisemblance marginales aux points d'indice i, par soustraction de leur valeur précédente et addition de leur nouvelle valeur. Le comptage des configurations $S_l(\boldsymbol{R})$ est également mis à jour : seule une colonne de $\boldsymbol{R}$ est modifiée, le comptage de deux configurations seulement est ajusté de plus ou moins 1, occasionnant 4 changements. Le vecteur $\boldsymbol{R}_{\bullet,i}$ est ensuite échantillonné selon les probabilités a posteriori, normalisées. Cette étape de simulation du i^{e} vecteur colonne de $\boldsymbol{R}$ a finalement une complexité en $O(2Lm)$. La valeur exacte de la probabilité a posteriori est ensuite calculée pour la nouvelle valeur courante du vecteur colonne $\boldsymbol{R}_{\bullet,i}$ à l'étape 11. Pour cela, les p-valeurs aux ruptures précédent et suivant la position i dans les m signaux sont mises à jour, et la probabilité a posteriori est également mise à jour pour ces nouvelles p-valeurs. Le coût de cette opération est dominé par la complexité du calcul des p-valeurs, en $O(2mn \log n)$.

En tenant compte des boucles **pour** des lignes 7 à 8 (facteur m), des lignes 5 à 13 (facteur $n-2$) et des lignes 3 à 13 (facteur V), et en supposant que la condition de mise à jour de l'estimateur du MAP, à la ligne 12, est rencontrée à chaque itération MCMC, on parvient à une complexité de la forme : $O(Vn^2(3mn\log(n) + mn + 2Lm))$. La dimension temporelle n est plusieurs ordres de grandeurs plus grande que le nombre de signaux m, et dans le cas non informatif, le nombre de configurations est $L = 2^m$. Dans la parenthèse, le terme mn est négligeable devant les deux autres. Selon les dimensions m et n, l'un des deux termes $3mn\log(n)$ et $2Lm$ l'emporte sur l'autre, comme le montrent les courbes de la figure C.1.

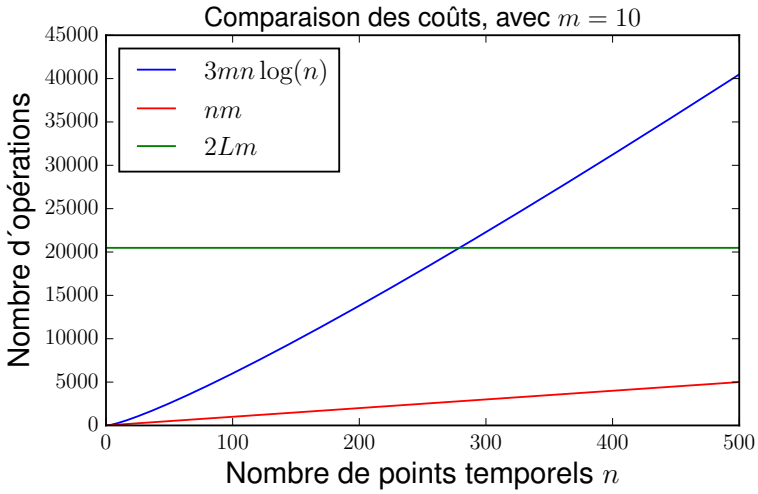


FIGURE C.1 – Comparaison des coûts calculatoires en fonction du nombre de points temporels pour l'étude de la complexité de l'algorithme MultiBD, avec $m = 10$ et $V = 1024$.

Algorithme 9 : MultiBD, pseudo-échantillonneur de Gibbs

Données : série temporelle $\boldsymbol{X} \in \mathbb{R}^{m \times n}$
Résultat : $\hat{\boldsymbol{R}}_{MAP}$

choisir α
initialiser $\boldsymbol{R}^{(0)}$, $\hat{\boldsymbol{R}}_{MAP} = \boldsymbol{R}^{(0)}$
pour $v \leftarrow 1, V$ **faire**
  mélanger aléatoirement les indices par la fonction π
  pour $p \leftarrow 2, n-1$ **faire**
    $i = \pi^{-1}(p)$
    pour $j \leftarrow 1, m$ **faire**
      calculer $P_{j,i}(\boldsymbol{X}, \boldsymbol{R}^{(v)}_{mat,p})$
    fin
    simuler $\boldsymbol{R}^{(v)}_{p}$ d'après (4.40)
    calculer $\Pr(\boldsymbol{R}^{(v)}_{mat,p}|\boldsymbol{X})$ d'après la loi a posteriori (C.1)
    si $\Pr(\boldsymbol{R}^{(v)}_{mat,p}|\boldsymbol{X}) > \Pr(\hat{\boldsymbol{R}}_{MAP}|\boldsymbol{X})$ **alors**
      $\hat{\boldsymbol{R}}_{MAP} = \boldsymbol{R}^{(v)}_{mat,p}$
    fin
  fin
fin

Annexe D

Compléments sur les réseaux bayésiens

D.1 Principe de *d*-séparation

Le principe de *d*-séparation décrit la façon dont l'information circule entre deux nœuds d'un graphe acyclique dirigé, et en particulier dans quelles conditions elle reste bloquée. La propagation d'information est à distinguer de la relation de causalité, orientée de la cause vers l'effet. Ainsi, en présence d'une relation causale de X vers Y, symbolisée par l'arête $X \to Y$, la connaissance de la valeur prise par la cause X apporte de l'information sur l'effet Y, mais inversement la connaissance de l'effet Y peut modifier la connaissance qu'on a de la cause X.

Définition D.1.1 (*d*-séparation [Geiger *et al.*, 1990, Pearl, 2014])**.** *Soit un DAG G. Un chemin C est dit* d-séparé *(ou bloqué) par l'ensemble de nœuds Z si et seulement si au moins l'une des deux conditions suivantes est vérifiée :*

- *C contient une chaîne $X_1 \to X_2 \to X_3$ ou la structure $X_1 \leftarrow X_2 \to X_3$ telle que X_2 appartient à l'ensemble Z ;*
- *C contient une V-structure $X_1 \to X_2 \leftarrow X_3$ telle que ni X_2 ni aucun de ses descendants n'appartiennent à l'ensemble Z.*

Les ensembles disjoints X et Y sont dits d-séparés *par l'ensemble Y si et seulement si tous les chemins d'un nœud de X vers un nœud de Y sont bloqués par Z.*

Lorsque deux nœuds X et Y sont d-séparés par l'ensemble Z, la propriété de fidélité garantit que X et Y sont conditionnellement indépendants par rapport à Z, et réciproquement.

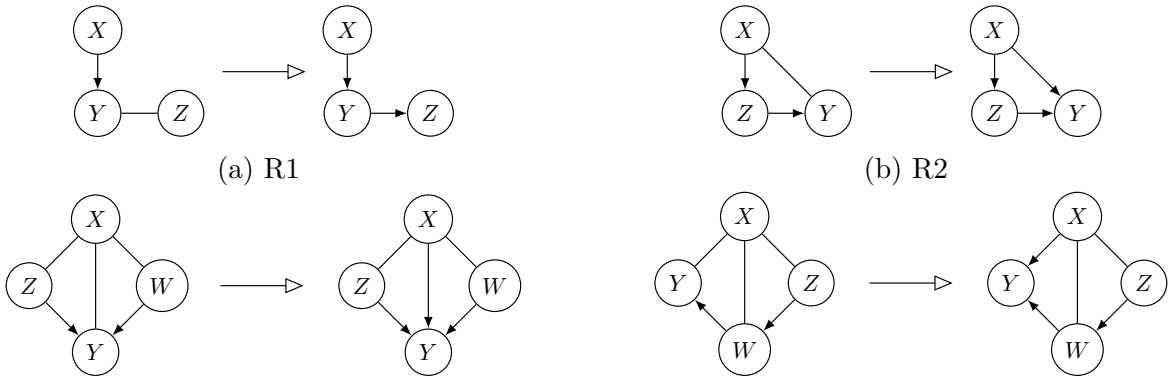
D.2 Algorithmes IC et PC

Pour estimer le modèle causal à partir des échantillons de la probabilité jointe, [Verma et Pearl, 1992] proposent l'algorithme d'Inductive Causation (voir [Pearl, 2000, p. 50]),

qui permet d'obtenir le graphe essentiel H qui décrit les influences causales entre les variables pour une distribution $\mathcal{P}$, en supposant que le modèle d'indépendance conditionnelle puisse être décrit par un DAG. Le graphe H est construit à partir du graphe vide, complété progressivement par les arêtes déduites d'un ensemble de relations d'indépendances conditionnelles. Ces relations sont établies par application de tests d'indépendance sur un ensemble d'observations des variables de U. Les trois principales étapes sont résumées dans l'algorithme 10. La première génère un graphe non orienté, la deuxième introduit les V-structures, et la troisième oriente le plus possible d'arêtes sans introduire de nouvelle V-structure ni de cycle. L'étape d'extension du graphe partiellement se fait par l'application des quatre règles suivantes :

— R1 : orienter $Y - Z$ en $Y \to Z$ s'il existe un arc $X \to Y$ tel que X et Z ne sont pas adjacents ;
— R2 : orienter $X - Y$ en $X \to Y$ s'il existe une chaîne $X \to Z \to Y$;
— R3 : orienter $X - Y$ en $X \to Y$ s'il existe deux chaînes $X - Z \to Y$ et $X - W \to Y$ telles que Z et W ne sont pas adjacents ;
— R4 : orienter $X - Y$ en $X \to Y$ s'il existe deux chaînes $X - Z \to W$ et $Z \to W \to Y$ telles que Z et Y ne sont pas adjacents.

[Meek, 1995] a montré que ces règles sont suffisantes pour parvenir au graphe essentiel. Elles sont schématisées dans la figure D.1. L'algorithme 6 de [Dor et Tarsi, 1992] peut également être employé pour tester si un graphe essentiel admet une extension, en un temps polynomial.



(a) R1 (b) R2

(c) R3 (d) R4

FIGURE D.1 – Règles pour l'orientation des arêtes [Verma et Pearl, 1992, Meek, 1995].

Une variante de l'algorithme IC, présentée dans [Spirtes *et al.*, 2000], est l'algorithme PC, nommé d'après ses auteurs. Le graphe initial est le graphe complet. L'application de tests d'indépendance permet de retirer certaines arêtes. Comme avec l'algorithme IC, les V-structures sont ensuite identifiées, puis les orientations sont propagées dans le reste du graphe quand une telle opération est possible. Ces étapes sont résumées dans l'algorithme 11, où Card(S) est le cardinal de l'ensemble S et Adj(H,X,Y) est l'ensemble des nœuds adjacents à X et Y dans le graphe H. L'article [Kalisch et Bühlmann, 2007] s'intéresse

Algorithme 10 : IC

Données : Ensemble de variables U, distribution $\mathcal{P}$
Résultat : Graphe essentiel H
1. **pour** *chaque paire* X *et* Y *de* U **faire**
  chercher l'ensemble S_{XY} tel que $X \perp\!\!\!\perp Y | S_{XY}$
  construire un graphe non orienté H tel que X et Y soient connectés si et seulement si S_{XY} est vide
fin
2. **pour** *chaque paire* X *et* Y *non adjacents dans* H *ayant pour voisin commun* Z **faire**
  si $X \in S_{XY}$ **alors**
    poursuivre
  sinon
    orienter les arcs : $X \rightarrow Z \leftarrow Y$
  fin
fin
3. orienter autant d'arêtes que possible dans H par application répétée des règles R1, R2, R3 et R4.

aux propriétés de l'algorithme PC dans le cas de graphes parcimonieux dont les variables suivent des distributions gaussiennes.

Algorithme 11 : PC

Données : Ensemble de variables U, graphe complet non orienté H
Résultat : Graphe essentiel H
initialisation : graphe complet non orienté H, $i = 0$
1. **tant que** *le nombre de paires X,Y de nœuds adjacents dans H est supérieur ou égal à i* **faire**
  pour *tout $S \subset Adj(H,X,Y)$ tel que* $\text{Card}(S) = i$, **faire**
    si $X \perp\!\!\!\perp Y|S$ **alors**
      supprimer l'arête $X - Y$ dans H
      ajouter S à l'ensemble S_{XY}
      ajouter S à l'ensemble S_{YX}
    fin
  fin
  i = i+1
fin
2. **pour** *chaque paire X et Y non adjacents dans H ayant pour voisin commun Z* **faire**
  si $X \notin S_{XY}$ **alors**
    orienter les arcs : $X \rightarrow Z \leftarrow Y$
  fin
fin
3. **tant que** *des arêtes peuvent être orientées dans H* **faire**
  si *il existe une arête entre X et Y et un chemin dirigé de X vers y* **alors**
    ajouter $X \rightarrow Y$
  fin
  si *X et Y ne sont pas adjacents et il existe Z tel que $X \rightarrow Z$ et $Z - Y$* **alors**
    ajouter $Z \rightarrow Y$
  fin
fin

Résumé : Cette thèse présente une méthode pour la détection hors-ligne de multiples ruptures dans des séries temporelles multivariées, et propose d'en exploiter les résultats pour estimer les relations de dépendance entre les variables du système. L'originalité du modèle, dit du *Bernoulli Detector*, réside dans la combinaison de statistiques locales issues d'un test robuste, comparant les rangs des observations, avec une approche bayésienne. Ce modèle non paramétrique ne requiert pas d'hypothèse forte sur les distributions des données. Il est applicable sans ajustement à la loi gaussienne comme sur des données corrompues par des valeurs aberrantes. Le contrôle de la détection d'une rupture est prouvé y compris pour de petits échantillons. Pour traiter des séries temporelles multivariées, un terme est introduit afin de modéliser les dépendances entre les ruptures, en supposant que si deux entités du système étudié sont connectées, les événements affectant l'une s'observent instantanément sur l'autre avec une forte probabilité. Ainsi, le modèle s'adapte aux données et la segmentation tient compte des événements communs à plusieurs signaux comme des événements isolés. La méthode est comparée avec d'autres solutions de l'état de l'art, notamment sur des données réelles de consommation électrique et génomiques. Ces expériences mettent en valeur l'intérêt du modèle pour la détection de ruptures entre des signaux indépendants, conditionnellement indépendants ou complètement connectés. Enfin, l'idée d'exploiter les synchronisations entre les ruptures pour l'estimation des relations régissant les entités du système est développée, grâce au formalisme des réseaux bayésiens. En adaptant la fonction de score d'une méthode d'apprentissage de la structure, il est vérifié que le modèle d'indépendance du système peut être en partie retrouvé grâce à l'information apportée par les ruptures, estimées par le modèle du *Bernoulli Detector*.

Mots-clés : détection de ruptures, inférence bayésienne, statistiques de rang, séries temporelles multivariées, réseaux bayésiens, classes d'équivalence de Markov

Abstract: This thesis presents a method for the multiple change-points detection in multivariate time series, and exploits the results to estimate the relationships between the components of the system. The originality of the model, called the *Bernoulli Detector*, relies on the combination of a local statistics from a robust test, based on the computation of ranks, with a global Bayesian framework. This non parametric model does not require strong hypothesis on the distribution of the observations. It is applicable without modification on gaussian data as well as data corrupted by outliers. The detection of a single change-point is controlled even for small samples. In a multivariate context, a term is introduced to model the dependencies between the changes, assuming that if two components are connected, the events occurring in the first one tend to affect the second one instantaneously. Thanks to this flexible model, the segmentation is sensitive to common changes shared by several signals but also to isolated changes occurring in a single signal. The method is compared with other solutions of the literature, especially on real datasets of electrical household consumption and genomic measurements. These experiments enhance the interest of the model for the detection of change-points in independent, conditionally independent or fully connected signals. The synchronization of the change-points within the time series is finally exploited in order to estimate the relationships between the variables, with the Bayesian network formalism. By adapting the score function of a structure learning method, it is checked that the independency model that describes the system can be partly retrieved through the information given by the change-points, estimated by the *Bernoulli Detector*.

Keywords: change-point detection, Bayesian inference, rank statistics, multivariate time series, Bayesian networks, Markov equivalence classes