Approches algorithmiques pour la statistique : processus déterministes par morceaux et arbres aléatoires

Romain Azaïs

▶ **To cite this version:**

Romain Azaïs. Approches algorithmiques pour la statistique : processus déterministes par morceaux et arbres aléatoires. Mathématiques [math]. ENS de Lyon, 2022. tel-03953630

HAL Id: tel-03953630
https://hal.science/tel-03953630

Submitted on 24 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Romain Azaïs

Approches algorithmiques pour la statistique : processus déterministes par morceaux et arbres aléatoires

Mémoire d'habilitation à diriger des recherches

– Décembre 2022 –

UNIVERSITÉ DE LYON

ENS DE LYON

HABILITATION À DIRIGER DES RECHERCHES DE L'UNIVERSITÉ DE LYON

Opérée par :

l'École Normale Supérieure de Lyon

Soutenue publiquement le 2 décembre 2022, par :

Romain Azaïs

Approches algorithmiques pour la statistique : processus déterministes par morceaux et arbres aléatoires

---

Devant le jury composé de :

Julien Chiquet Directeur de Recherche, INRAE	**Rapporteur**
Benoîte de Saporta Professeure, Université de Montpellier	**Rapporteuse**
Jean-Baptiste Durand Chercheur Senior, CIRAD	**Examinateur**
Anne Gégout Professeure, Université de Lorraine	**Examinatrice**
Christophe Godin Directeur de Recherche, Inria	**Garant**
Franck Picard Directeur de Recherche, CNRS	**Examinateur**
Marc Tommasi Professeur, Université de Lille	**Rapporteur**

Résumé

Ce manuscrit donne une présentation synthétique et une mise en perspective de morceaux choisis de mes travaux de recherche effectués depuis 2014 en vue de l'obtention de l'habilitation à diriger des recherches en mathématiques appliquées. Ceux-ci portent principalement sur l'analyse statistique de deux objets mathématiques : les processus déterministes par morceaux et les arbres aléatoires. Les premiers constituent un modèle dynamique sur un espace continu en temps continu quand les seconds sont statiques et intrinsèquement discrets. S'ils sont donc de nature différente, l'approche suivie pour leur étude statistique est commune : développer des méthodes algorithmiques d'extraction de l'information rigoureuses, fondées sur la théorie et vérifiées expérimentalement. C'est cette idée que les quatre chapitres de ce mémoire cherchent à illustrer.

Les deux premiers chapitres sont consacrés aux processus markoviens déterministes par morceaux. On s'intéresse d'abord au problème de l'estimation de leur taux de saut dans un cadre général et lorsqu'une unique trajectoire en temps long est observée. On se concentre dans un second temps sur l'estimation de fonctionnelles liées aux croisements, continus ou survenant lors de sauts, de ces processus.

On propose dans les deux derniers chapitres de ce manuscrit des méthodes d'analyse statistique des données arborescentes. On commence par étudier théoriquement à travers un modèle probabiliste le noyau des sous-arbres pour en proposer ensuite des généralisations. Enfin, on construit des estimateurs consistants des paramètres inconnus de la loi de naissance d'arbres de Galton-Watson conditionnés par la taille ou la hauteur.

Abstract

This manuscript, written in French, provides a summary and contextualisation of selected parts of my research work carried out since 2014 in order to obtain the habilitation to conduct research in applied mathematics. It mainly deals with the statistical analysis of two distinct mathematical objects: piecewise-deterministic processes and random trees. The first one is a dynamic model on a continuous space in continuous time, while the latter are static and intrinsically discrete. If they are therefore different in nature, the approach followed for their statistical study can be shared. It consists in the development of rigorous algorithmic methods of information extraction, based on theory and experimentally verified. The four chapters of this thesis seek to illustrate this idea.

The first two chapters are devoted to piecewise-deterministic Markov processes. We address the problem of estimating their jump rate in a general framework from the observation of a single trajectory within a long time window. We then focus on the estimation of functionals related to crossings, either continuous or occurring during jumps.

In the last two chapters of this manuscript, we propose methods for the statistical analysis of tree data. We start by studying guarantees of the subtree kernel for a probabilistic model. We take advantage of this theoretical insight to propose some generalisations of the kernel. Finally, we build consistent estimators of the unknown parameters of the birth distribution of Galton-Watson trees conditioned on size or height.

Remerciements

En premier lieu, je tiens à exprimer ma reconnaissance sincère à celle et ceux qui ont accepté de rapporter ce manuscrit, Julien Chiquet, Benoîte de Saporta et Marc Tommasi, pour leur temps, leurs efforts et leurs mots sur mon travail et mon parcours. Mes remerciements chaleureux vont également aux autres membres du jury, Anne Gégout, Jean-Baptiste Durand, Franck Picard et bien sûr Christophe Godin, pour son soutien et sa confiance dans ce projet d'habilitation. Nul doute que je me sentirai très bien entouré le jour de la soutenance.

Je pense aussi en écrivant ces lignes à mes coauteur·e·s, et plus largement à tout·e·s mes collègues chercheur·se·s ou des services d'appui, en particulier de Lyon et de Nancy, avec qui j'ai partagé des équations sur un tableau, des questionnements sérieux ou absurdes, un cours ou simplement un café. Ce manuscrit leur doit beaucoup.

Si j'ai savamment évité les risques d'un inventaire à la Prévert dans le paragraphe précédent, je tiens à remercier plus spécialement et par ordre chronologique : Alexandre Genadot, Anne Gégout, François Dufour, Christophe Godin, Benoît Henry et Florian Ingels. Leur apport scientifique et humain est incommensurable.

J'ai une pensée émue pour le reviewer anonyme, que je maudis parfois pour sa compréhension partielle et sa mauvaise foi, mais dont les rapports ont permis bien souvent d'améliorer les travaux qui constituent ce mémoire. Je veux lui dire ici que la statue érigée en son honneur à Moscou n'est pas un hommage suffisant.

À chaque période sa couleur musicale. Celle de cette habilitation a des tons de Clara Luciani et de Therapie TAXI, avec des nuances de Superblue de Kurt Elling et Charlie Watson. En osant espérer qu'un peu de leur talent ait pu s'immiscer dans mon écriture.

Enfin, pour leurs sourires et tout le reste, Marine et Iphigénie : merci.

Table des matières

Prologue

Ce manuscrit constitue une synthèse et une mise en perspective de morceaux choisis de mes travaux post-thèse, réalisés en tant que chargé de recherche à l'Inria, d'abord à Nancy puis à Lyon, en vue de l'obtention de l'habilitation à diriger des recherches en mathématiques appliquées. Les statistiques sont le principal domaine de ce document, avec le soutien des probabilités appliquées, de l'informatique théorique ou de la simulation numérique selon les objets d'intérêt et les approches choisies.

Bibliographie

Cette liste recense l'ensemble de mes travaux en langue anglaise évalués ou en cours d'évaluation par les pairs. Les deux plus récents d'entre eux sont au stade de prépublication au moment d'écrire ces lignes. Les autres ont été publiés comme articles de journaux, actes de conférences internationales ou chapitre d'ouvrage collectif.

(P26) F. Ben Naoum, C. Godin et R. Azaïs : *Characterization of random walks on space of unordered trees using efficient metric simulation.* Prépublication, 2022.

(P25) R. Azaïs et B. Henry : *Maximum likelihood estimation for spinal-structured trees.* Prépublication, 2021.

(P24) F. Ingels et R. Azaïs : *Enumeration of irredundant forests.* Theoretical Computer Science, 2022.

(P23) R. Azaïs, S. Ferrigno et M.-J. Martinez : *cvmgof: an R package for Cramér-von Mises goodness-of-fit tests in regression models.* Journal of Statistical Computing and Simulation, 2022.

(P22) A. Chaudhury, P. Hanappe, R. Azaïs, C. Godin et D. Colliaux : *Transferring PointNet++ segmentation from virtual to real plants.* CVPPA 2021.

(P21) F. Ingels et R. Azaïs : *Isomorphic unordered labeled trees up to substitution ciphering.* IWOCA 2021.

(P20) B. Leggio et R. Azaïs : *Estimation of piecewise-deterministic trajectories in a quantum optics scenario.* Statistical Topics and Stochastic Models for Dependent Data with Applications (edited by V. S. Barbu and N. Vergne), ISTE – Wiley, 2020.

(P19) C. S. Galvan-Ampudia, G. Cerutti, J. Legrand, G. Brunoud, R. Martin-Arevalillo, R. Azaïs, V. Bayle, S. Moussu, C. Wenzl, Y. Jaillais, J. U. Lohmann, C. Godin et T. Vernoux : *Temporal integration of auxin information for the regulation of patterning.* eLife, 2020.

(P18) R. AZAÏS et F. INGELS : *The weight function in the subtree kernel is decisive.* Journal of Machine Learning Research, 2020.

(P17) B. HENRY, S. R. CHOWDHURY, A. LAHMADI, R. AZAÏS, J. FRANÇOIS et R. BOUTABA : *SPONGE: software-defined traffic engineering to absorb influx of network traffic.* IEEE Local Computer Networks 2019.

(P16) R. AZAÏS, J.-B. DURAND et C. GODIN : *Approximation of trees by self-nested trees.* ALENEX 2019.

(P15) R. AZAÏS : *Nearest embedded and embedding self-nested trees.* Algorithms (Special Issue Data Compression Algorithms and their Applications), 2019.

(P14) R. AZAÏS, G. CERUTTI, D. GEMMERLÉ et F. INGELS : *treex: a Python package for manipulating rooted trees.* Journal of Open Source Software, 2019.

(P13) R. AZAÏS, A. GENADOT et B. HENRY : *Inference for conditioned Galton-Watson trees from their Harris path.* Latin American Journal of Probability and Mathematical Statistics, 2019.

(P12) R. AZAÏS et A. GENADOT : *Estimation of the average number of continuous crossings for non-stationary non-diffusion processes.* Journal of Statistical Planning and Inference, 2019.

(P11) R. AZAÏS, B. DELYON et F. PORTIER : *Integral estimation based on Markovian design.* Advances in Applied Probability, 2018.

(P10) R. AZAÏS et A. GENADOT : *A new characterization of the jump rate for piecewise-deterministic Markov processes with discrete transitions.* Communications in Statistics – Theory and Methods, 2018.

(P9) R. AZAÏS et A. MULLER-GUEUDIN : *Optimal choice among a class of nonparametric estimators of the jump rate for piecewise-deterministic Markov processes.* Electronic Journal of Statistics, 2016.

(P8) A. BEN ABDESSALEM, R. AZAÏS, M. TOUZET-CORTINA, A. GÉGOUT-PETIT et M. PUIGGALI : *Stochastic modelling and prediction of fatigue crack propagation using piecewise-deterministic Markov processes.* Journal of Risk and Reliability, 2016.

(P7) R. AZAÏS et A. GENADOT : *Semi-parametric inference for the absorption features of a growth-fragmentation model.* TEST, 2015.

(P6) R. AZAÏS, F. DUFOUR et A. GÉGOUT-PETIT : *Nonparametric estimation of the conditional distribution of the inter-jumping times for piecewise-deterministic Markov processes.* Scandinavian Journal of Statistics, 2014.

(P5) R. AZAÏS, J.-B. BARDET, A. GENADOT, N. KRELL et P.-A. ZITT : *Piecewise deterministic Markov process – recent results.* ESAIM: Proceedings and Surveys, 2014.

(P4) R. AZAÏS, R. COUDRET et G. DURRIEU : *A hidden renewal model for monitoring aquatic systems biosensors.* Environmetrics, 2014.

(P3) R. AZAÏS : *A recursive nonparametric estimator for the transition kernel of a piecewise-deterministic Markov process.* ESAIM: Probability and Statistics, 2014.

(P2) R. AZAÏS, F. DUFOUR et A. GÉGOUT-PETIT : *Nonparametric estimation of the jump rate for nonhomogeneous marked renewal processes.* Annales de l'Institut Henri Poincaré – Probabilités et Statistiques, 2013.

(P1) R. AZAÏS, A. GÉGOUT-PETIT et J. SARACCO : *Optimal quantization applied to sliced inverse regression.* Journal of Statistical Planning and Inference, 2012.

Grille de lecture

Publications issues de la thèse : (P1–6)

Les publications qui sont directement issues de ma thèse sont celles antérieures à 2014. (P1) traite de problèmes de régression semi-paramétrique abordés sous l'angle de la quantification optimale en norme $\mathbb{L}^p$. (P2) et (P4) concernent la statistique des processus de renouvellement marqués comme préliminaire au sujet central de la thèse : la statistique non-paramétrique des processus markoviens déterministes par morceaux (PDMPs pour l'anglais piecewise-deterministic Markov processes), abordée dans (P3) et (P5–6).

Processus markoviens déterministes par morceaux : (P7–10), (P12) et (P20)

Après la thèse, j'ai continué mon exploration du monde de la statistique des PDMPs jusqu'en 2017, même si le hasard d'une rencontre m'a mené à nouveau sur cette voie quelques années plus tard. Dans la poursuite de mes travaux de thèse, j'ai notamment étudié l'estimation non-paramétrique du taux de saut d'un tel processus dans (P9–10). Ces deux publications seront présentées dans le chapitre 1 avec un rappel opportun de (P2) et (P6). J'ai également considéré l'estimation de fonctionnelles liées aux croisements de ces processus dans (P7) et (P12), qui seront traitées dans le chapitre 2. Les publications restantes portent sur des problématiques de modélisation : en fiabilité en collaboration avec des mécaniciens dans (P8) (qui pourrait appartenir à la catégorie précédente puisque ce travail a débuté avant ma thèse, mais s'est aussi poursuivi bien au-delà) et en mécanique quantique dans (P20) (mon co-auteur, physicien de formation, avait étudié pendant sa thèse des modèles PDMPs de mécanique quantique).

- Chapitre 1 : (P9–10)
  Estimation du taux de saut d'un processus markovien déterministe par morceaux
- Chapitre 2 : (P7) et (P12)
  Croisements des processus markoviens déterministes par morceaux

Arbres aléatoires : (P13–16), (P18), (P21) et (P24–26)

Le post-doctorat a été l'occasion pour moi de découvrir un nouvel objet d'étude : les données arborescentes. La question initiale à laquelle je me suis intéressé dans (P15–16) puis très indirectement dans (P26) porte sur la compression avec perte des arbres non-ordonnés. J'ai fait évoluer ce sujet vers l'apprentissage statistique, notamment sous l'angle des méthodes à noyaux dans (P18) et (P24), en conservant des idées de compression (cette fois sans perte) comme outil d'énumération. Ces deux publications seront au cœur du chapitre 3. La statistique des arbres est également étudiée sous l'hypothèse de modèles probabilistes dans (P13) (arbres de Galton-Watson conditionnés par la taille) et (P25) (arbres de Galton-Watson conditionnés à survivre), contributions présentées dans le chapitre 4. Les algorithmes développés à la faveur de l'ensemble de ces travaux ont été implémentés en `Python` dans la librairie `treex`, que j'ai développée, d'abord seul puis avec l'aide d'autres contributeurs, et qui a été publiée dans (P14).

- Chapitre 3 : (P18) et (P24)
  Noyau des sous-arbres et généralisations
- Chapitre 4 : (P13) et (P25)
  Estimation de modèles de Galton-Watson conditionnés

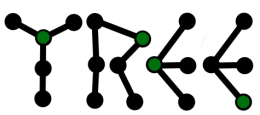
Autres thèmes : (P11), (P17), (P19) et (P22–23)

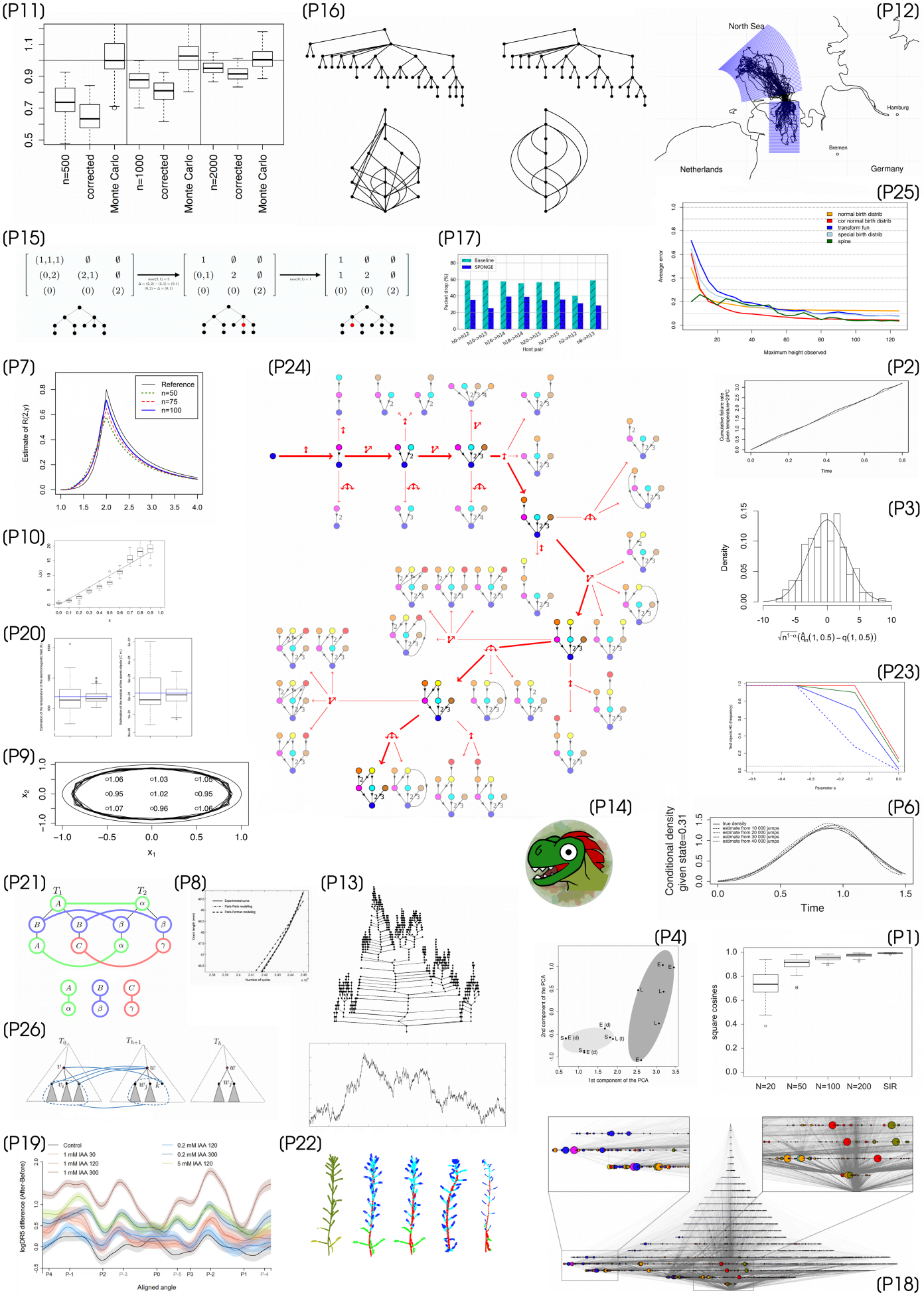
Au fil des années, j'ai eu l'occasion de m'intéresser à d'autres sujets que ceux mentionnés ci-dessus, notamment en lien avec les chaînes de Markov : (P11) porte sur des techniques d'approximation d'intégrale par échantillonnage préférentiel (importance sampling en anglais) markovien ; une collaboration avec des informaticiens des réseaux a abouti à l'introduction d'un algorithme stochastique de type MCMC (Markov chain Monte Carlo) pour l'atténuation d'un trafic trop important dans (P17). (P19) est un article de biologie auquel ma contribution se résume à un support méthodologique dans l'analyse des données. La question posée dans (P22) concerne l'acquisition de la structure arborescente de plantes via l'apprentissage par transfert d'un réseau de neurones de la littérature estimé pour la segmentation. Enfin, on se demande comment comparer numériquement différents tests d'adéquation de Cramér-von Mises pour la régression dans (P23).

Intentions

Les principaux objets mathématiques auxquels je me suis intéressé sont les PDMPs et les arbres aléatoires ; ils apparaissent naturellement dans le titre de ce document. Il s'agit de concepts très différents : le premier est un processus stochastique évoluant en temps continu sur des espaces continus alors que le second est intrinsèquement discret et ne fait a priori pas apparaitre de dynamique temporelle. Par conséquent, les outils mathématiques à mettre en œuvre pour leur étude sont eux aussi de nature différente : la convergence des chaînes de Markov joue un rôle prépondérant dans l'analyse du premier, alors que les questions algorithmiques et de complexité sont au cœur de l'étude du second. Mais finalement, peu importe la nature des objets étudiés, la démarche que j'ai suivie me semble au fond la même : la question qui m'intéresse profondément est toujours celle du développement d'une approche statistique algorithmique rigoureuse. Je veux dire par là que c'est la construction d'une méthode d'extraction de l'information à partir de données issues de modèles qui m'a préoccupé dans la plupart de mes publications. Et à chaque fois, j'ai cherché à ce que celle-ci soit fondée sur la théorie, illustrée numériquement et idéalement vérifiée expérimentalement.

La suite de ce manuscrit est constituée de 4 chapitres, chacun donnant une présentation et une mise en perspective de 2 publications : comme indiqué ci-dessus, les chapitres 1 et 2 traitent de la statistique des PDMPs alors que les chapitres 3 et 4 concernent l'analyse de données arborescentes. J'ai choisi ces 8 contributions car elles me semblent représentatives de l'ensemble de mes travaux, de mon évolution thématique, mais également de ma démarche scientifique. En plus de décrire les modèles ou les données considérés et les principaux résultats obtenus, j'ai essayé de retranscrire l'approche statistique que j'ai suivie et de montrer comment l'analyse algorithmique des données émerge de la théorie. Le contenu des chapitres est donc à mon sens très différent de celui qu'on peut attendre d'un article scientifique. En particulier, les démonstrations des résultats théoriques et leurs validations numériques ou expérimentales ne sont jamais décrites in extenso. La lecture seule des chapitres devrait néanmoins permettre d'avoir une idée assez précise du contenu de ces 8 articles, et d'aider le cas échéant à leur lecture (même si, par souci d'homogénéité des notations de ce manuscrit, celles-ci peuvent différer de l'article au chapitre afférent).




(P11)
n=500
corrected
Monte Carlo
n=1000
corrected
Monte Carlo
n=2000
corrected
Monte Carlo
(P16)
(P12)
North Sea
Hamburg
Bremen
Netherlands
Germany
(P25)
normal birth distrib
cor normal birth distrib
transform fun
special birth distrib
spine
Average error
Maximum height observed
(P15)
(P17)
Baseline
SPONGE
Packet drop (%)
Host pair
(P7)
Reference
n=50
n=75
n=100
Estimate of R(2,y)
(P24)
(P2)
Cumulative failure rate given temperature=20°C
Time
(P10)
(P3)
Density
(P20)
(P9)
(P23)
(P14)
(P6)
Conditional density given state=0.31
true density
estimate from 10 000 jumps
estimate from 20 000 jumps
estimate from 30 000 jumps
estimate from 40 000 jumps
Time
(P21)
(P8)
(P13)
(P4)
2nd component of the PCA
1st component of the PCA
(P1)
square cosines
N=20
N=50
N=100
N=200
SIR
(P26)
(P19)
Control
1 mM IAA 30
1 mM IAA 120
1 mM IAA 300
0.2 mM IAA 120
0.2 mM IAA 300
5 mM IAA 120
logDR5 difference (After-Before)
Aligned angle
(P22)
(P18)

1

Le grand saut

Estimation du taux de saut d'un processus markovien déterministe par morceaux

Ce chapitre est consacré au problème de l'estimation du taux de saut d'un processus markovien déterministe par morceaux (abrégé en PDMP pour l'anglais piecewise-deterministic Markov process), aussi général que possible, à partir d'une unique observation en temps long. Après avoir défini la dynamique de tels processus ainsi que le cadre de travail choisi dans la section 1.1, les travaux de thèse (P2) et (P6) seront rappelés en section 1.2. La stratégie suivie et les résultats obtenus dans (P10) (comme applications de (P6)) sont présentés en section 1.3. Le cœur de ce chapitre se situe en section 1.4, dédiée à l'article (P9).

1.1 Processus markoviens déterministes par morceaux

1.1.1 Définition

Un PDMP est un processus à temps continu $(X_t)_{t\in\mathbb{R}_+}$, évoluant de manière déterministe pour presque tout t selon une équation différentielle, mais subissant ponctuellement des transitions aléatoires (appelées sauts) à des instants aléatoires. Un tel processus, défini sur un ouvert E de $\mathbb{R}^d$ muni de la tribu borélienne $\mathcal{B}(E)$, est décrit par un triplet (λ, Q, Φ) où $\lambda : E \to \mathbb{R}_+$ est le taux de saut gouvernant l'apparition des instants de saut, $Q : \mathcal{B}(E) \times \overline{E} \to [0,1]$ est le noyau de transition régissant la loi des sauts, et $\Phi : \mathbb{R} \times E \to E$ est le flot auquel obéit la dynamique déterministe. Si on note T_n les instants de saut du processus (avec $T_0 = 0$), la dynamique du processus $(X_t)_{t\in\mathbb{R}_+}$ entre les instants T_n et T_{n+1}, conditionnellement à X_{T_n}, s'écrit ainsi :

- Pour tout $T_n \leq t < T_{n+1}$, $X_t = \Phi(t - T_n | X_{T_n})$;
- La position $X_{T_{n+1}}$ en T_{n+1} est sélectionnée selon le noyau Q appliqué à la position juste avant le saut $\Phi(T_{n+1} - T_n | X_{T_n})$, i.e.

$$\mathbb{E}\left[\varphi(X_{T_{n+1}}) \,|\, X_{T_n}, T_{n+1} - T_n\right] = \int_E Q(dy|\Phi(T_{n+1} - T_n|X_{T_n}))\varphi(y). \tag{1}$$

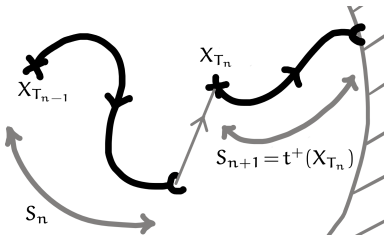


Figure 1: Trajectoire typique d'un PDMP débutant en $X_{T_{n-1}}$, évoluant de manière déterministe pendant la durée (aléatoire) S_n puis sautant vers la position (aléatoire) X_{T_n} avant de reprendre sa dynamique déterministe. Celle-ci se termine en heurtant ∂E au bout du temps $S_{n+1} = t^+(X_{T_n})$.

Il reste alors à décrire la loi des instants de saut, ce qu'on fait généralement via la fonction de survie conditionnelle des durées inter-saut $S_{n+1} = T_{n+1} - T_n$,

$$\mathbb{P}(S_{n+1} > t \,|\, X_{T_n} = x) = \exp\left(-\int_0^t \lambda(\Phi(s|x))ds\right) \mathbb{1}_{\{0 \le t < t^+(x)\}}, \tag{2}$$

où $t^+ : E \to \mathbb{R}_+ \cup \{+\infty\}$ désigne le temps (déterministe) d'atteinte de la frontière ∂E de E suivant le flot Φ,

$$t^+(x) = \inf\{t \ge 0 \ : \ \Phi(t|x) \in \partial E\}.$$

L'équation (2) se lit ainsi : conditionnellement à X_{T_n}, le $n+1^e$ saut a lieu selon le taux non-homogène $\lambda \circ \Phi(\cdot|X_{T_n})$ à moins que le flot n'ait atteint la frontière de l'espace d'état avant. La figure 1 fournit une illustration de la trajectoire d'un PDMP.

Cette classe de processus a été introduite par Davis dans les années 1980 avec l'ambition de décrire une grande variété de modèles stochastiques non-diffusifs (34, 35). Sans prétendre à l'exhaustivité, on en trouve des applications récentes en fiabilité (25, 36), en assurance (68) ou en biologie (27). D'autres exemples sont présentés dans le chapitre 2 ainsi que dans (P8) et (P20). Il existe également des extensions de ces processus en dimension infinie, notamment dans le cadre de modèles de neurosciences (22, 46).

1.1.2 Objectif : estimer λ

La question en jeu dans la suite est un problème d'inférence statistique : on suppose qu'on dispose d'observations d'un PDMP et on cherche à construire de bons estimateurs des caractéristiques inconnues, c'est-à-dire des fonctions des données dont on souhaite qu'elles s'approchent au plus près de leur cible, et surtout que leur précision s'améliore quand le nombre d'observations croit.

On peut imaginer des schémas d'observation de PDMPs variés mais celui choisi dès la thèse, i.e. dans (P3) et (P6), ainsi que dans les articles ultérieurs (P9–10), est le suivant : une seule version du processus $(X_t)_{t\in\mathbb{R}_+}$ est observée, en temps long. Cette hypothèse est équivalente à observer la chaîne de Markov $(X_{T_n}, S_{n+1})_{n\in\mathbb{N}}$ en temps long lorsqu'on connait le flot Φ. On ne s'intéressera donc pas dans la suite à l'estimation de la dynamique déterministe.

D'autres schémas d'observation sont bien sûr possibles : dans les applications décrites dans (P8) ou (P20), on s'attend plutôt à observer des versions indépendantes du processus en temps fini. Le schéma d'observation choisi ci-dessus oblige à ne tenir compte que de l'aléa markovien sans profiter d'observations i.i.d., a priori plus faciles à traiter. Il faut

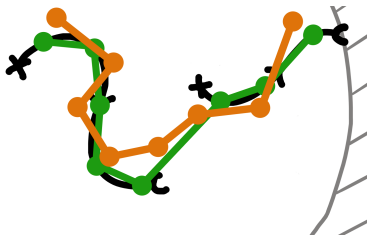


Figure 2: Trajectoire de la figure 1 observée parfaitement, i.e. sans bruit, mais sur une grille temporelle discrète en vert et observée à travers un bruit additif en orange. Dans le second cas, le saut parait beaucoup plus difficile à identifier.

noter que le contexte le plus difficile est aussi le plus réaliste : observer le PDMP (en temps long ou non) sur une grille temporelle discrète, non-alignée avec les instants de saut, et sans connaitre le flot. Si le PDMP est parfaitement observé, on peut espérer déterminer ses instants de saut (au pas de temps de la grille près) en utilisant des hypothèses de modélisation sur leur fréquence ou sur l'effet du noyau de transition, et ainsi retrouver le flot (et finalement revenir au schéma ci-dessus à l'incertitude sur le temps de saut près). Mais s'il est en plus observé à travers un bruit, comme dans (24, 26, 43), le problème d'estimation d'un tel processus partiellement observé devient plus ardu (cf. la figure 2).

Les cadres de travail choisis dans les articles (P3), (P6) et (P9–10) portant sur l'estimation des PDMPs sont très généraux : l'objectif est en effet de construire des estimateurs des quantités d'intérêt applicables à une grande variété de modèles. Par exemple, on ne souhaite pas fixer la dimension de l'espace d'état du processus (on le suppose seulement métrique séparable dans (P6), même si cette hypothèse sera abandonnée au profit de $\mathbb{R}^d$ dans les publications ultérieures), ni les formes de la dynamique déterministe, du noyau de transition ou du taux de saut. En particulier, les estimateurs construits et étudiés sont non-paramétriques.

L'espace d'état typique des PDMPs introduits notamment pour des problèmes de fiabilité est de la forme

$$E = \bigcup_{m \in M} \{m\} \times E_m,$$

où $M \subset \mathbb{N}$ et $E_m \subset \mathbb{R}^{d_m}$: l'état du processus est un couple dont la première composante (souvent appelée mode) reste constante entre deux sauts et joue le rôle de paramétrisation de la dynamique déterministe de la seconde, elle continue, et éventuellement de son espace de définition. C'est la motivation principale du choix d'un espace d'état aussi général que possible. On se contente de $\mathbb{R}^d$ dans (P3) et (P9) mais les algorithmes et résultats s'étendent sans aucune difficulté à des processus à mode discret.

Comme on l'a mentionné plus tôt, le flot Φ est supposé connu. Il nous faut donc estimer les deux quantités suivantes :

- La loi conditionnelle des durées inter-saut S_{n+1} sachant X_{T_n}, gouvernée par la composée $\lambda \circ \Phi$;
- La loi de la position X_{T_n} sachant celle juste avant le saut $\Phi(X_{T_{n-1}}, S_n)$, viz. le noyau de transition Q.

Dans les deux cas, il s'agit de lois conditionnelles. (P3), issue de la thèse, traite du problème de l'estimation de Q (via des méthodes non-paramétriques récursives) et ne sera plus considérée dans ce chapitre. La suite est dédiée à l'estimation du taux de saut λ

telle qu'envisagée dans (P10) et surtout (P9) et en regard du travail effectué au cours de la thèse dans (P6) sur l'estimation de la loi conditionnelle des durées inter-saut.

La littérature sur la statistique de PDMPs généraux observés en temps long est peu conséquente, et concerne surtout le cas unidimensionnel (44, 66, 67) pour lequel on peut profiter de la monotonie du flot pour identifier des relations entre la loi invariante et les caractéristiques du processus. Dans le cadre non-paramétrique et en temps long, mais pour des familles de processus spécifiques issues de la biologie, on peut notamment citer (39, 55, 71). Une étude de la vraisemblance est présentée dans (56, 5 Likelihood Processes). On renvoie également la lectrice ou le lecteur à toutes les références mentionnées dans (P3), (P6) et (P9–10).

1.1.3 Difficultés attendues

L'aléa d'un PDMP $(X_t)_{t\in\mathbb{R}_+}$ est contenu dans la chaîne de Markov $(X_{T_n}, S_{n+1})_{n\in\mathbb{N}}$. Si le caractère aléatoire de ces processus est donc discret, leur dynamique en temps continu ainsi que l'intrication des deux aléas (durées inter-saut et transitions) rendent le problème de leur estimation délicat, en particulier l'estimation du taux de saut, avec des difficultés originales par rapport à l'estimation des chaînes de Markov ou des processus de renouvellement. On note ci-dessous quelques particularités statistiques des PDMPs :

- Le taux de saut du processus est en fait $\lambda \circ \Phi$ (le taux λ est pris le long du flot, ce qui permet en particulier de préserver le caractère markovien du processus sans avoir des durées inter-saut exponentielles). Par conséquent, les méthodes d'estimation directe regardant S_{n+1} sachant X_{T_n} mènent à la composée $\lambda \circ \Phi$ (ou à une autre quantité caractéristique de la loi mais directement fonction de $\lambda \circ \Phi$), et non λ.
- Si $(X_t)_{t\in\mathbb{R}_+}$ est markovien, le processus à temps continu restant en X_{T_n} entre les instants T_n et T_{n+1} (et subissant les transitions décrites par l'équation (1)) n'est pas un processus de Markov mais un processus de renouvellement marqué, pour lequel la loi de la marque $X_{T_{n+1}}$ sachant le passé dépend de la marque précédente X_{T_n} et du temps passé dans celle-ci S_{n+1}.
- Le temps d'atteinte de la frontière $t^+(X_{T_n})$ joue le rôle d'une censure déterministe de la durée aléatoire S_{n+1}, fonction de la position.
- Dès que la dimension de l'espace est supérieure à 2, conditionnellement à X_{T_n}, la position juste avant le prochain saut $\Phi(S_{n+1}|X_{T_n})$ appartient presque sûrement à un espace de mesure nulle (la courbe décrite par le flot $\Phi(\cdot|X_{T_n})$), et ce même si toutes les caractéristiques du processus sont à densité.

1.1.4 Hypothèse principale : ergodicité du PDMP

Notre problème est double : il faut à la fois construire une quantité fonction des données dont on espère qu'elle sera un bon estimateur de λ et montrer ses propriétés de convergence. Évidemment, ces deux aspects sont intrinsèquement liés. L'estimation d'une loi conditionnelle $Y|X$ se fait naturellement via l'estimation de la loi du couple (X, Y) sur l'estimation de la loi du conditionnement X. Pour un PDMP, la loi du couple et la loi du conditionnement évoluent au cours du temps et il est alors tout indiqué de se passer de cette dépendance au temps en supposant une forme de stationnarité. L'hypothèse la plus forte demande à la loi de la chaîne de Markov $(X_{T_n}, S_{n+1})_{n\in\mathbb{N}}$ de ne pas dépendre du temps. Si une telle loi existe, on l'appelle loi stationnaire ou loi invariante. On considère

plutôt une hypothèse plus faible en ne requérant qu'une forme d'ergodicité assurant la convergence du processus vers sa loi invariante. L'hypothèse d'ergodicité peut être faite indifféremment sur le PDMP $(X_t)_{t\in\mathbb{R}_+}$ ou la chaîne discrète $(X_{T_n}, S_{n+1})_{n\in\mathbb{N}}$ (29). La construction des estimateurs passe donc par l'estimation de lois invariantes de chaînes de Markov et les propriétés de convergence découleront de l'hypothèse d'ergodicité.

On suppose donc qu'il existe une loi π_∞ telle que, pour toute condition initiale X_{T_0},

$$\|\pi_n - \pi_\infty\|_{VT} \to 0,$$

où π_n désigne la loi de X_{T_n} et $\|\cdot\|_{VT}$ la norme en variation totale. Cette hypothèse est vérifiée sous des conditions de minoration du noyau de la chaîne $(X_{T_n})_{n\in\mathbb{N}}$ (condition de Döblin dans (P3, Assumption A.2) ou minoration moins subtile dans (P10, Assumption 3.1)). On peut alors montrer que la chaîne de Markov $(X_{T_n})_{n\in\mathbb{N}}$ (et les chaînes associées comme $(X_{T_n}, S_{n+1})_{n\in\mathbb{N}}$) vérifient le théorème ergodique presque sûr (32, Theorem 4.3.15) et un théorème de la limite centrale (32, Theorem 4.3.16), qui nous seront utiles pour mettre en lumière la qualité de nos estimateurs.

Les autres hypothèses imposées dans les publications concernées portent sur la régularité et/ou le caractère borné des fonctions en jeu, ou sont très spécifiques à une analyse (on pense ici à (P10) où l'espace d'état de $(X_{T_n})_{n\in\mathbb{N}}$ est supposé fini). Ces hypothèses ne seront pas détaillées sauf si elles jouent un rôle majeur dans la construction de l'estimateur.

1.2 Appliquer le modèle à intensité multiplicative à tout prix

1.2.1 Estimateur de Nelson-Aalen

Éloignons nous des PDMPs un instant pour traiter de données i.i.d. $(X_n)_{1\le n\le N}$ observées sous une censure à droite C (qu'on suppose ici fixée pour ne pas s'embarrasser de détails superflus pour notre propos), i.e. on dispose seulement des $Y_n = \min(X_n, C)$ pour estimer la loi des X_n.

Le taux de risque λ des X_n, défini par $\lambda = f/G$ où f est leur densité et G leur fonction de survie, caractérise leur loi et est souvent la quantité d'intérêt dans des problèmes d'analyse de survie. Pour l'estimer, en particulier à partir de données censurées, une méthode efficace est l'estimateur de Nelson-Aalen (1). Son introduction passe par la définition du processus de comptage à temps continu $(\mathcal{N}_N(t))_{t\in[0,C)}$,

$$\mathcal{N}_N(t) = \sum_{n=1}^{N} \mathbb{1}_{\{Y_n\le t\}}.$$

Ce processus est central pour la raison suivante : le processus $(\mathcal{M}_N(t))_{t\in[0,C)}$ donné par

$$\mathcal{M}_N(t) = \mathcal{N}_N(t) - \int_0^t \lambda(s)\mathcal{Y}_N(s)ds, \tag{3}$$

avec $\mathcal{Y}_N(t) = \sum_{n=1}^{N} \mathbb{1}_{\{Y_n\ge t\}}$, fait apparaitre la quantité cible λ et est une martingale. L'estimateur de Nelson-Aalen en $t \in [0, C)$ est alors

$$\widehat{\Lambda}_N(t) = \int_0^t \mathcal{Y}_N(s)^{-1} d\mathcal{N}_N(s) = \sum_{n=1}^{N} \mathcal{Y}_N(Y_n)^{-1}\mathbb{1}_{\{Y_n\le t\}},$$

où

$$y_N(t)^{-1} = \begin{cases} 1/y_N(t) & \text{si } y_N(t) > 0, \\ 0 & \text{sinon.} \end{cases}$$

$\widehat{\Lambda}_N(t)$ fournit un bon estimateur du taux de risque cumulé $\Lambda(t) = \int_0^t \lambda(s)ds$. En effet, $\widehat{\Lambda}_N$ converge, uniformément sur tout compact, en probabilité vers sa cible (6, Theorem IV.1.1) et vérifie un résultat de normalité asymptotique (6, Theorem IV.1.2).

Pour obtenir un estimateur de λ, on applique un lissage à noyau similaire à ce qu'on fait pour estimer une densité à partir de la fonction de répartition empirique. Cet estimateur, appelé estimateur de Ramlau-Hansen (86), vérifie notamment des propriétés de convergence uniforme (sur tout compact éloigné de 0) en probabilité (6, Theorem IV.2.2).

Comme f est au signe près la dérivée de G, on a

$$G(t) = \exp\left(-\int_0^t \lambda(s)ds\right), \tag{4}$$

et on comprend, en regard de l'équation (2), que λ est le parfait analogue de la composée $\lambda \circ \Phi$ des PDMPs. En outre, la durée inter-saut S_{n+1} d'un PDMP s'écrit $S_{n+1} = \min(\check{S}_{n+1}, t^+(X_{T_n}))$ avec

$$\mathbb{P}(\check{S}_{n+1} > t|X_{T_n}) = \exp\left(-\int_0^t \check{\lambda}_{X_{T_n}}(s)ds\right),$$

où $\check{\lambda}$ doit vérifier, pour $t < t^+(x)$, $\check{\lambda}_x(t) = \lambda(\Phi(t|x))$. Il est alors clair que la stratégie de Nelson-Aalen et la méthode de lissage de Ramlau-Hansen sont de bons candidats à l'estimation du taux $\lambda \circ \Phi$ d'un PDMP.

L'estimateur de Nelson-Aalen peut en fait être défini dès lors que l'intensité de saut (l'intégrande du compensateur (terme intégral dans l'équation (3)) du processus de comptage) se met sous forme multiplicative, ce qui permet des applications de cette méthode dans des contextes bien plus variés que celui des données censurées à droite (cf. les multiples exemples présentés dans (6)). Cette hypothèse est adroitement appelée modèle à intensité multiplicative. Mais les PDMPs vérifient-ils une propriété de ce type ? Avant de s'y intéresser, on regarde le cas des processus de renouvellement marqués, dont le lien avec les processus qui nous intéresse a déjà été évoqué.

1.2.2 Application aux processus de renouvellement marqués

Dans (P2), on applique le modèle à intensité multiplicative décrit ci-dessus à des processus de renouvellement marqués dont l'espace d'état est aussi général que possible, ce qui interdit a priori le lissage en espace. On va travailler avec l'approximation discrète en espace du processus de renouvellement dont il faudra montrer que le taux est suffisamment proche de celui du processus original.

On considère donc un processus de renouvellement $(X_t)_{t\in\mathbb{R}_+}$ à valeurs dans un espace métrique séparable E. Partant de X_0, on commence par déterminer la 1re durée inter-saut S_1 dont la loi est donnée par

$$\mathbb{P}(S_1 > t|X_0) = \exp\left(-\int_0^t \lambda(s|X_0)ds\right) \mathbb{1}_{\{0\le t<t^+(X_0)\}}.$$

Entre les instants $T_0 = 0$ et $T_1 = S_1$, le processus est constant : $X_t = X_0$. En $t = T_1$, un saut a lieu selon le noyau de transition $Q(\cdot|X_0)$. Partant de X_{T_1}, la durée inter-saut

S_2 et la transition à l'instant $T_2 = T_1 + S_2$ sont déterminées de la même manière. Ainsi de suite, on obtient un processus de renouvellement marqué dont les durées inter-saut dépendent des marques, alors que les marques ne dépendent que de la précédente. On fait l'hypothèse d'ergodicité de la chaîne des marques.

La dynamique d'un tel processus est très similaire à celle d'un PDMP à la différence notable près que les marques ne dépendent pas de la durée inter-saut précédente. Il s'agit donc d'un excellent exercice d'application du modèle à intensité multiplicative.

Afin d'estimer le taux de saut $\lambda(\cdot|x)$, on considère un ensemble $A \ni x$ et on introduit le processus de comptage $(\mathcal{N}_n(t|A))_{t\in[0,t^+(A))}$ (avec $t^+(A) = \inf_A t^+$),

$$\mathcal{N}_N(t|A) = \sum_{n=0}^{N-1} \mathbb{1}_{\{X_{T_n}\in A\}}\mathbb{1}_{\{S_{n+1}\le t\}}. \tag{5}$$

On peut montrer (P2, Theorem 4.13) que son compensateur est donné par

$$\sum_{n=0}^{N-1} \mathbb{1}_{\{X_{T_n}\in A\}}\mathbb{1}_{\{S_{n+1}\ge t\}}\,\lambda(t|X_{T_n}). \tag{6}$$

Lorsque l'espace d'état est discret, en prenant A réduit à $\{x\}$, on peut éliminer la dépendance en n du taux de saut en le remplaçant par $\lambda(t|x)$ et on obtient à nouveau le modèle à intensité multiplicative (ce qui est décrit pour des processus similaires sans l'horloge déterministe t^+ dans (6, Example IV.1.9)). Dans le cas général, prendre $A = \{x\}$ ne mène à rien sinon à un processus de comptage presque sûrement nul dès que la loi des marques est à densité. Malgré tout, rien ne nous empêche de considérer le processus

$$\mathcal{Y}_N(t|A) = \sum_{n=0}^{N-1} \mathbb{1}_{\{X_{T_n}\in A\}}\mathbb{1}_{\{S_{n+1}\ge t\}}$$

ainsi que l'estimateur de Nelson-Aalen

$$\widehat{L}_N(t) = \int_0^t \mathcal{Y}_N(s|A)^{-1}\,d\mathcal{N}_N(s|A).$$

À cause de la discrétisation en espace, on peut montrer (P2, Theorem 4.17) que $\widehat{L}_N(t|A)$ estime le taux cumulé $L(t|A) = \int_0^t l(s|A)ds$ dont l'intégrande est une version moyennée sur A (sous la loi invariante π_∞) du taux de saut λ,

$$l(t|A) = \frac{\int_A f(t|z)\pi_\infty(dz)}{\int_A G(t|z)\pi_\infty(dz)},$$

où f et G sont la densité et la fonction de survie conditionnelles associées à λ vérifiant $\lambda(t|x) = f(t|x)/G(t|x)$. Dès lors, les bonnes propriétés des estimateurs peuvent être montrées en contrôlant l'erreur entre $\lambda(t|x)$ et $l(t|A)$ en fonction du diamètre de A sous des hypothèses de régularité des fonctions en jeu (P2, Lemma 4.20).

1.2.3 Cas des PDMPs

Pour les PDMPs, nous allons devoir être plus précis dans la construction des compensateurs et de leur filtration. Par souci de concision, on ne tient pas compte ici du temps déterministe d'atteinte de la frontière, i.e. $t^+ \equiv +\infty$.

On note $(\mathcal{F}_t^{n+1})_{t\in\mathbb{R}_+}$ la filtration engendrée par le processus à un saut $t \mapsto \mathbb{1}_{\{S_{n+1}\leq t\}}$. Alors, si on s'intéresse au processus $t \mapsto \mathbb{1}_{\{X_{T_n}\in A\}}\mathbb{1}_{\{S_{n+1}\leq t\}}$, son compensateur dans la filtration $(\sigma(X_{T_n}) \vee \mathcal{F}_t)_{t\in\mathbb{R}_+}$ est bien $t \mapsto \mathbb{1}_{\{X_{T_n}\in A\}}\mathbb{1}_{\{S_{n+1}\geq t\}}\lambda(\Phi(t|X_{T_n}))$. On peut tenter de reproduire la démarche mise en place pour les processus de renouvellement : on s'attend en effet à ce que le compensateur du processus à N sauts donné par l'équation (5) s'exprime comme la somme des compensateurs de manière analogue à l'équation (6), simplement en remplaçant $\lambda(t|X_{T_n})$ par $\lambda \circ \Phi(t|X_{T_n})$.

La filtration à considérer pour déterminer le compensateur de ce processus est

$$\left(\sigma(X_{T_0},\dots,X_{T_{N-1}}) \vee \bigvee_{n=0}^{N-1} \mathcal{F}_t\right)_{t\in\mathbb{R}_+}$$

comme pour les processus de renouvellement considérés ci-dessus. On fait alors face au problème suivant : la loi du processus $t \mapsto \mathbb{1}_{\{S_1\leq t\}}$, i.e. la loi de S_1, dans cette filtration est envisagée conditionnellement à la position précédente X_{T_0} mais aussi à la position qui suit X_{T_1}. Or, la connaissance de X_{T_1} change en général la loi de S_1 : le processus gouverné par $\Phi(t|x) = x + t$ et $Q(dy|x) = \delta_{x/2}(dy)$ en est un très bon exemple.

Ceci nous donne l'intuition que l'application du modèle à intensité multiplicative aux PDMPs va mener à l'estimation de la loi des durées inter-saut S_{n+1} conditionnellement aux positions courantes X_{T_n} et aux positions suivantes $X_{T_{n+1}}$. Sans supposer $t^+ \equiv +\infty$ comme dans l'énoncé des filtrations ci-dessus, on montre (P6, Proposition 3.5) que cette loi est donnée par le taux de saut

$$\widetilde{\lambda}(t|x,y) = \frac{f(t|x)Q(y|\Phi(t|x))}{H(y,t|x)}, \tag{7}$$

où

$$H(y,t|x) = \int_t^{t^+(x)} f(s|x)Q(y|\Phi(s|x))ds + G(t^+(x)|x)Q(y|\Phi(t^+(x)|x),$$

avec

$$\mathbb{P}(X_{T_1}\in B, S_1 > t|X_{T_0} = x) = \int_B dy\, H(y,t|x). \tag{8}$$

Dès lors, l'application de la stratégie décrite ci-dessus pour les processus de renouvellement va nous conduire à estimer une approximation en espace $\widetilde{l}(t|A,B)$ de $\widetilde{\lambda}(t|x,y)$, $A \times B \ni (x,y)$. Pour revenir aux caractéristiques du PDMP d'intérêt, on manipule un peu l'équation (7) afin d'obtenir la densité conditionnelle des durées inter-saut

$$f(t|x) = \int dy\, \widetilde{\lambda}(t|x,y)H(y,t|x).$$

Disposant d'un estimateur de $\widetilde{l}(t|A,B)$ et d'après l'équation (8), on considère l'approximation de $f(t|x)$

$$\sum_{B\in\mathcal{B}} \widetilde{l}(t|A,B)\mathbb{P}(X_{T_1}\in B, S_1 > t|X_{T_0}\in A),$$

où $\mathcal{B}$ est une partition de l'espace d'état, dont on peut montrer sous des hypothèses de régularité qu'elle approche aussi bien qu'on le souhaite sa cible (P6, Proposition 2.4). La probabilité conditionnelle inconnue restante sera estimée par son équivalent empirique (P6, Proposition 2.6). Le principal résultat de consistance obtenu est la convergence, uniforme sur tout compact en temps et en espace, en probabilité de l'estimateur ainsi construit de la densité conditionnelle f (P6, Theorem 2.7).

Appliquer la méthodologie de Nelson-Aalen à l'estimation du taux de saut d'un PDMP peut sembler être une bonne idée mais leur dynamique particulière induit deux défauts majeurs (qu'il était difficile de soupçonner) :

- Pour estimer $f(t|x)$ en un point x fixé, on doit estimer le taux $\widetilde{\lambda}(t|x,y)$ et la probabilité $H(y,t|x)$ sur tout l'espace d'état en y.
- Le modèle à intensité multiplicative vise à estimer un taux de risque sans passer par un estimateur quotient de type f/G. Or, on estime finalement 3 quantités ; la probabilité conditionnelle $\mathbb{P}(X_{T_1} \in B, S_1 > t|X_{T_0} \in A)$ étant elle-même estimée par un quotient.

Par ailleurs, le taux λ du processus reste inestimé à ce stade. Ces constats sont le point de départ des travaux (P9–10) présentés dans la suite.

1.3 Exploiter autrement l'équation (7)

On propose ici de construire un estimateur de λ à partir de l'équation (7) sans passer par la densité conditionnelle f. Cette section reprend succinctement l'article (P10).

1.3.1 Une expression de λ

La stratégie développée consiste à remarquer les points (certes triviaux) suivants :

- $\lambda(x) = \lambda(\Phi(0|x)) = f(0|x)/G(0|x) = f(0|x)$.
- $H(y,0|x) = R(y|x)$ où R désigne le noyau de la chaîne $(X_{T_n})_{n\in\mathbb{N}}$.

Avec l'équation (7), on obtient

$$\lambda(x) = \int dy\, \widetilde{\lambda}(0|x,y) R(y|x).$$

On pourrait réutiliser la méthode de Nelson-Aalen et le lissage de Ramlau-Hansen pour estimer $\widetilde{\lambda}(0|x,y)$ ainsi qu'un estimateur de R afin d'obtenir une estimation plug-in du taux λ. Mais cela reviendrait à se contenter de la valeur au bord de l'intervalle de temps ($t = 0$) de l'estimateur de Nelson-Aalen lissé, et on se rappelle que la convergence uniforme en probabilité est établie pour tout compact éloigné de 0. Au lieu de cela, sous l'hypothèse t^+ bornée, on propose de décomposer $\widetilde{\lambda}(\cdot t^+(x)|x,y)$ sur une base de $\mathbb{L}^2_{[0,1]}$ (indépendante du couple (x,y)), i.e., pour $s \in [0,1]$,

$$\widetilde{\lambda}(st^+(x)|x,y) = \sum_{p\in\mathbb{N}} \theta^p(x,y) B^p(s),$$

où les coefficients $\theta^p(x,y) = \langle B^p, \widetilde{\lambda}(\cdot t^+(x)|x,y)\rangle$ ont l'avantage de rendre compte du comportement de $\widetilde{\lambda}(\cdot|x,y)$ sur son intervalle de définition. On a finalement

$$\lambda(x) = \sum_{p\in\mathbb{N}} B^p(0) \int dy\, \theta^p(x,y) R(y|x). \tag{9}$$

Cette équation peut être exploitée de multiples manières, notamment dans le cas continu via des techniques de discrétisation (comme dans ce qui précède), mais la solution

proposée dans (P10) fait l'hypothèse de transitions discrètes, i.e. $Q(dy|x)$ est une mesure discrète dont le support ne dépend pas de x. Dans ce cas,

$$\lambda(x) = \sum_{p\in\mathbb{N}} B^p(0) \sum_y \theta^p(x,y) R(y|x). \tag{10}$$

Par souci de rigueur, il nous faut remarquer que dès les équations (7) et (8) on suppose que le noyau $Q(dy|x)$ est à densité par rapport à une mesure notée abusivement dy, la densité étant elle-même notée $Q(y|x)$. Le noyau $R(dy|x)$ de la chaine est par conséquent à densité, notée $R(y|x)$ par rapport à la même mesure ; ces deux quantités apparaissent dans l'expression (9) de λ. Celle-ci reste vraie dans le cas général en remplaçant $dy\, R(y|x)$ par $R(dy|x)$ et la forme (10) de λ dans le cas discret en est donc bien une application, même si, cette fois, $R(y|x)$ désigne $R(\{y\}|x)$.

1.3.2 Estimation

Commençons par remarquer le point clé suivant : comme nous sommes dans le cas discret, comme pour les processus de renouvellement (cf. (P2, 3 Discrete state space) et (6, Example IV.1.9)), le modèle à intensité multiplicative est vérifié dès lors que le compensateur fait bien intervenir le taux $\widetilde{\lambda}(t|x,y)$. Par conséquent, on peut construire l'estimateur de Nelson-Aalen $\widehat{\widetilde{\Lambda}}_N(t|x,y)$ de $\widetilde{\Lambda}(t|x,y) = \int_0^t \widetilde{\lambda}(t|x,y)dt$. Dans la suite, $\widetilde{G}$ désigne la fonction de survie conditionnelle associée, comme dans l'équation (4).

On estime les coefficients $\theta^p(x,y)$ par plug-in,

$$\widehat{\theta}^p_N(x,y) = \frac{1}{t^+(x)} \int_0^1 B^p(u) d\widehat{\widetilde{\Lambda}}_N(ut^+(x)|x,y),$$

et $R(y|x)$ par son équivalent empirique,

$$\widehat{R}_N(y|x) = \frac{\sum_{n=0}^{N-1} \mathbb{1}_{\{X_{T_n}=x, X_{T_{n+1}}=y\}}}{\sum_{n=0}^{N-1} \mathbb{1}_{\{X_{T_n}=x\}}}.$$

L'estimateur du taux de saut du PDMP est obtenu en remplaçant les inconnues apparaissant dans l'équation (10) par leur estimateur, mais également en tronquant la somme au terme τ_N qui sera défini plus tard,

$$\widehat{\lambda}_N(x) = \sum_{p=0}^{\tau_N} B^p(0) \sum_y \widehat{\theta}^p_N(x,y) \widehat{R}_N(y|x). \tag{11}$$

Il est indispensable de préciser τ_N pour rendre cet estimateur calculable. On peut se douter que sa valeur sera importante dans la preuve de la consistance de $\widehat{\lambda}_N$. On suppose dans (P10, Assumptions 3.5) que la suite $(\tau_N)_{N\in\mathbb{N}}$ vérifie une condition technique peu exploitable en pratique. Mais on montre aussi (P10, Remark 3.6) que, si on choisit la base de Fourier ou la base de Legendre, prendre $\tau_N = o(\sqrt{N})$ est suffisant pour que cette condition technique soit vérifiée. En particulier, le nombre de termes n'est pas aléatoire, mais seulement fixé par le nombre de données disponibles.

1.3.3 Résultats de convergence

Outre l'hypothèse de décomposition de $\widetilde{\lambda}$ sur une base de $\mathbb{L}^2_{[0,1]}$ (vérifiant l'hypothèse technique mentionnée ci-dessus, mais vérifiée pour les bases de Fourier et de Legendre) et le caractère borné de t^+, on suppose, comme on l'a mentionné plus haut, que

l'espace d'état $\mathcal{E}$ de la chaîne $(X_{T_n})_{n\in\mathbb{N}}$ est fini, mais également que tous les coefficients de la matrice de transition R sont strictement positifs (ce qui en particulier induit l'hypothèse d'ergodicité).

La convergence de l'estimateur $\widehat{R}_N(y|x)$ est connue (32, 4.4 Statistics of Markov chains),

$$\widehat{R}_N(y|x) \xrightarrow{\text{p.s.}} R(y|x) \qquad \text{et} \qquad \sqrt{N}\left(\widehat{R}_N(y|x) - R(y|x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_R^2(x,y)),$$

où

$$\sigma_R^2(x,y) = \frac{R(y|x)(1-R(y|x))}{\pi_\infty(x)}.$$

On montre un résultat similaire pour les $\widehat{\theta}_N^p(x,y)$, mais la convergence presque sûre est remplacée par une convergence en probabilité comme pour l'estimateur de Nelson-Aalen.

Proposition 1 *(P10, Proposition 3.3) Pour tout couple $(x,y) \in \mathcal{E}^2$, on a*

$$\widehat{\theta}_N^p(x,y) \xrightarrow{\mathbb{P}} \theta^p(x,y) \qquad \textit{et} \qquad \sqrt{N}\left(\widehat{\theta}_N^p(x,y) - \theta^p(x,y)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma_\theta^2(p,x,y)\right),$$

où la variance asymptotique $\sigma_\theta^2(p,x,y)$ est donnée par

$$\sigma_\theta^2(p,x,y) = \frac{1}{R(y|x)\pi_\infty(x)t^+(x)} \int_0^1 \frac{B^p(s)^2 \widetilde{\lambda}(t^+(x)s|x,y)}{\widetilde{G}(t^+(x)s|x,y)} ds.$$

On prouve aussi la convergence de l'estimateur (11) de λ.

Proposition 2 *(P10, Proposition 3.7) Pour tout $x \in \mathcal{E}$, on a*

$$\left|\widehat{\lambda}_N(x) - \lambda(x)\right| \xrightarrow{\mathbb{P}} 0.$$

1.3.4 Test de nullité des coefficients

On peut exploiter la normalité asymptotique de la proposition 1 afin de construire un test de nullité des coefficients ; test particulièrement utile pour ne pas tenir compte de termes inutiles dans la définition (11) de $\widehat{\lambda}_N$. Pour cela, il nous faut estimer la variance inconnue qui y apparait, ce qu'on fait à nouveau par plug-in,

$$\widehat{\sigma}_{\theta,N}^2(p,x,y) = \frac{1}{\widehat{R}_N(y|x)\widehat{\pi}_{\infty,N}(x)t^+(x)^2} \int_0^1 B^p(s)^2 d\left[\widehat{\widetilde{G}}_N(t^+(x)s|x,y)\right]^{-1},$$

où

- $\widehat{\pi}_{\infty,N}(x) = \sum_{n=0}^{N-1} \mathbb{1}_{\{X_{T_n}=x\}}/N$ estime la loi invariante $\pi_\infty(x)$;
- $\widehat{\widetilde{G}}_N(s|x,y) = \exp\left(-\widehat{\widetilde{\Lambda}}_N(s|x,y)\right)$ est l'estimateur de Fleming-Harrington de $\widetilde{G}(s|x,y)$.

La statistique de test est

$$T_N^p(x,y) = \frac{N\,\widehat{\theta}_N^p(x,y)^2}{\widehat{\sigma}_{\theta,N}^2(p,x,y)},$$

dont les comportements différents sous les hypothèses nulle $\theta_p(x,y) = 0$ et alternative $\theta_p(x,y) \neq 0$ doivent discriminer ces deux options. C'est ce qu'on montre dans le résultat ci-dessous.

Corollaire 3 *(P10, Corollary 3.4) Si B^p est dérivable à dérivée continue sur $[0,1]$, alors, sous l'hypothèse nulle $\theta^p(x,y)=0$,*

$$T_N^p(x,y) \xrightarrow{\mathcal{L}} \chi^2(1).$$

Sous l'hypothèse alternative $\theta_p(x,y)\neq 0$, la statistique de test tend vers l'infini.

1.4 Estimateur quotient de λ

Dans (P9), on cherche à estimer la loi des durées inter-saut sans passer par le modèle à intensité multiplicative comme dans tout ce qui précède, et en acceptant donc de calculer des estimateurs quotients. De plus, on souhaite, comme dans (P10), revenir à la caractéristique λ du PDMP.

Cette fois, l'espace d'état E est un ouvert de $\mathbb{R}^d$, ce qui nous autorise à utiliser des méthodes de lissage en espace. Pour permettre une présentation de la méthode d'estimation et des résultats aussi peu techniques que possible, nous allons supposer dans le manuscrit et contrairement au papier original, $t^+ \equiv +\infty$.

1.4.1 Estimation des lois jointes

On introduit les estimateurs $\widehat{\pi}_{\infty,N}$, $\widehat{\mathcal{F}}_N$ et $\widehat{\mathcal{G}}_N$,

$$\begin{aligned}
\widehat{\mathcal{F}}_N(x,t) &= \frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{v_n^d w_n}\mathbb{K}_d\left(\frac{X_{T_n}-x}{v_n}\right)\mathbb{K}_1\left(\frac{S_{n+1}-t}{w_n}\right),\\
\widehat{\mathcal{G}}_N(x,t) &= \frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{v_n^d}\mathbb{K}_d\left(\frac{X_{T_n}-x}{v_n}\right)\mathbb{1}_{\{S_{n+1}>t\}},\\
\widehat{\pi}_{\infty,N}(x) &= \frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{v_n^d}\mathbb{K}_d\left(\frac{X_{T_n}-x}{v_n}\right),
\end{aligned}$$

où $\mathbb{K}_p$ est un noyau sur $\mathbb{R}^p$, $p\in\{1,d\}$, et les fenêtres sont définies pour tout n par $v_n = v_0(n+1)^{-\alpha}$ et $w_n = w_0(n+1)^{-\beta}$. Ces trois quantités sont des estimateurs à noyau récursifs (la fenêtre de lissage w_n (en espace) ou v_n (en temps) est fonction de l'indice de la somme n et non du nombre total de données N) de lois de probabilité, tels qu'introduits dans (99). Dans notre cas, on s'attend plutôt à estimer la loi invariante des données (41, 54).

La raison pour laquelle nous introduisons ces estimateurs est la suivante :

- Le rapport $\widehat{f}_N(t|x) = \frac{\widehat{\mathcal{F}}_N(x,t)}{\widehat{\pi}_{\infty,N}(x)}$ estime la densité conditionnelle $f(t|x)$;
- Le rapport $\widehat{G}_N(t|x) = \frac{\widehat{\mathcal{G}}_N(x,t)}{\widehat{\pi}_{\infty,N}(x)}$ estime la fonction de survie conditionnelle $G(t|x)$;
- Le rapport $\widehat{\lambda\circ\Phi}_N(t|x) = \frac{\widehat{\mathcal{F}}_N(x,t)}{\widehat{\mathcal{G}}_N(x,t)}$ estime la composée $\lambda\circ\Phi(t|x)$.

Autrement dit, ce triplet nous permet de construire très simplement des estimateurs quotients des trois caractéristiques principales de la loi conditionnelle des durées inter-saut.

Afin d'obtenir leur convergence et leur normalité asymptotique, on s'intéresse à l'étude vectorielle du triplet $(\widehat{\mathcal{F}}_N(x,t),\widehat{\mathcal{G}}_N(x,t),\widehat{\pi}_{\infty,N}(x))$. Le résultat principal de (P9) établit la

convergence et la normalité asymptotique vectorielles de ce triplet d'estimateurs sous l'hypothèse d'ergodicité ainsi que des conditions de régularité des caractéristiques du PDMP (P9, Assumptions 2.3 & 2.4) et des noyaux (P9, Assumptions 3.1), et pour des paramètres de lissage α et β bien choisis, viz. $(\alpha, \beta) \in \mathcal{A}$, avec

$$\mathcal{A} = \left\{(\alpha, \beta) \in \mathbb{R}^2 \ : \ \alpha > 0, \ \beta > 0, \ \alpha d + \beta < 1, \ \alpha d + \beta + 2\min(\alpha, \beta) > 1\right\},$$

dont on peut montrer le caractère non-vide. De plus, le résultat énoncé et démontré dans (P9) intègre la possibilité d'un temps d'atteinte de la frontière fini, ce qui complique à la fois la méthode d'estimation (condition sur les fenêtres initiales) et surtout les preuves.

Théorème 4 *(P9, Theorem 3.3) Si $\pi_\infty(x)f(t|x) > 0$, alors*

$$\begin{bmatrix} \widehat{\mathcal{F}}_N(x,t) \\ \widehat{\mathcal{G}}_N(x,t) \\ \widehat{\pi}_{\infty,N}(x) \end{bmatrix} \xrightarrow{p.s.} \begin{bmatrix} \pi_\infty(x)f(t|x) \\ \pi_\infty(x)G(t|x) \\ \pi_\infty(x) \end{bmatrix}$$

et on a la normalité asymptotique,

$$N^{\frac{1-\alpha d-\beta}{2}} \left(\begin{bmatrix} \widehat{\mathcal{F}}_N(x,t) \\ \widehat{\mathcal{G}}_N(x,t) \\ \widehat{\pi}_{\infty,N}(x) \end{bmatrix} - \begin{bmatrix} \pi_\infty(x)f(t|x) \\ \pi_\infty(x)G(t|x) \\ \pi_\infty(x) \end{bmatrix} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0_3, \Sigma(x,t,\alpha,\beta)),$$

où la matrice de covariance $\Sigma(x,t,\alpha,\beta)$ est dégénérée : seul le coefficient $(1,1)$ est strictement positif et vaut

$$\Sigma(x,t,\alpha,\beta)_{1,1} = \frac{\tau_1^2 \tau_d^2 \pi_\infty(x) f(t|x)}{1 + \alpha d + \beta},$$

avec $\tau_p^2 = \int_{\mathbb{R}^p} \mathbb{K}_p^2$, $p \in \{1, d\}$.

La matrice de covariance est dégénérée car la vitesse des estimateurs $\widehat{\pi}_{\infty,N}(x)$ et $\widehat{\mathcal{G}}_N(x,t)$ est plus rapide que celle de $\widehat{\mathcal{F}}_N(x,t)$: $\widehat{\mathcal{F}}_N(x,t)$ est lissé à la fois en espace et en temps alors que $\widehat{\pi}_{\infty,N}(x)$ et $\widehat{\mathcal{G}}_N(x,t)$ sont seulement lissés en espace. Il est donc naturel de s'intéresser au couple $(\widehat{\pi}_{\infty,N}(x), \widehat{\mathcal{G}}_N(x,t))$ pour lequel on peut aussi établir un théorème de la limite centrale (P9, Theorem 3.6),

$$N^{\frac{1-\alpha d}{2}} \left(\begin{bmatrix} \widehat{\mathcal{G}}_N(x,t) \\ \widehat{\pi}_{\infty,N}(x) \end{bmatrix} - \begin{bmatrix} \pi_\infty(x)G(t|x) \\ \pi_\infty(x) \end{bmatrix} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0_2, \Sigma'(x,t,\alpha)),$$

où $\Sigma'(x,t,\alpha)$ est diagonale.

Si on admet que l'ergodicité supposée est de plus géométrique, i.e. $\|\pi_N - \pi_\infty\|_{VT} \leq b^{-n}$, $b > 1$, alors, sous la condition supplémentaire $2(\alpha d + \beta) < 1$, on peut appliquer une technique déjà utilisée dans (P4) (pour l'estimation de Q) pour établir une forme de convergence uniforme des estimateurs (P9, Proposition 3.7).

Un corollaire important du théorème 4 concerne la normalité asymptotique de $\widehat{f}_N(t|x)$, obtenue en utilisant le lemme de Slutsky, ce qui donne tout son intérêt au résultat vectoriel (P9, Corollary 3.8),

$$N^{\frac{1-\alpha d-\beta}{2}} \left(\widehat{f}_N(x,t) - f(t|x) \right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau_1^2 \, \tau_d^2 \, f(t|x)}{(1 + \alpha d + \beta)\pi_\infty(x)}\right).$$

Un résultat similaire existe pour les estimateurs de la fonction de survie conditionnelle $G(t|x)$ (P9, Corollary 3.9) et de la composée $\lambda \circ \Phi(t|x)$.

1.4.2 Estimation de λ le long du flot

Comme autre corollaire (P9, Corollary 3.10) du théorème 4, on a

$$\widehat{\lambda \circ \Phi}_N(t|x) \xrightarrow{\text{p.s.}} \lambda \circ \Phi(t|x),$$

et

$$N^{\frac{1-\alpha d-\beta}{2}} \left(\widehat{\lambda \circ \Phi}_N(t|x) - \lambda \circ \Phi(t|x) \right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau_1^2 \tau_d^2 \lambda \circ \Phi(t|x)}{(1+\alpha d+\beta)\pi_\infty(x) G(t|x)}\right).$$

Mais on souhaite avoir accès à $\lambda(x)$ et non à cette composée. Dans (P10), on écrit $\lambda(x) = \lambda \circ \Phi(0|x)$. Mais nous pouvons généraliser cette approche. Pour cela, on suppose que le flot peut être pris à rebours, ce que nous faisons depuis le début de ce chapitre sans jamais l'avoir utilisé puisque Φ est défini sur $\mathbb{R} \times E$. On note

$$\mathcal{C}_x = \{\Phi(-t|x) \ : \ 0 \leq t < t^-(x)\},$$

où $t^-(x)$ est le temps déterministe d'atteinte de la frontière partant de x en suivant le flot à rebours. Soit $\xi \in \mathcal{C}_x$. Alors par définition il existe $\tau_x(\xi)$ tel que $\Phi(\tau_x(\xi)|\xi) = x$ et donc

$$\lambda \circ \Phi(\tau_x(\xi)|\xi) = \lambda(x).$$

On peut ainsi construire une famille d'estimateurs de $\lambda(x)$ le long de $\mathcal{C}_x$,

$$\widehat{\lambda}_{\xi,N}(x) = \widehat{\lambda \circ \Phi}_N(\tau_x(\xi)|\xi).$$

De ce qui précède, chacun de ces estimateurs converge presque sûrement vers $\lambda(x)$ et est asymptotiquement gaussien de variance

$$\sigma_\xi^2 = \frac{\tau_1^2 \tau_d^2 \lambda(x)}{(1+\alpha d+\beta)\pi_\infty(\xi) G(\tau_x(\xi)|\xi)}.$$

1.4.3 Minimisation de la variance asymptotique

Parmi cette famille d'estimateurs, on aimerait choisir celui de variance asymptotique minimale, i.e. estimer $\lambda(x)$ par $\widehat{\lambda}_{\xi^\star,N}(x)$ où $\xi^\star$ maximise $\pi_\infty(\xi)G(\tau_x(\xi)|\xi)$. Ce critère est intéressant car il émerge du calcul et contredit la première intuition qui aurait été de maximiser $\pi_\infty(\xi)$ afin de profiter du plus grand nombre possible de données X_{T_n} autour de ξ. Mais ce choix ne nous aurait rien dit de la qualité de l'estimation en $(\xi, \tau_x(\xi))$ qui est le réel point d'intérêt.

La quantité $\pi_\infty(\xi)G(\tau_x(\xi)|\xi)$ est malheureusement inconnue. Nous sommes malgré tout chanceux puisqu'elle est précisément estimée par $\widehat{\mathcal{G}}_N(\xi, \tau_x(\xi))$! Finalement, on estime donc $\lambda(x)$ par

$$\widehat{\widehat{\lambda}}_N(x) = \widehat{\lambda}_{\widehat{\xi}_N^\star,N}(x), \qquad \text{où} \quad \widehat{\xi}_N^\star = \underset{\xi \in \mathcal{C}_x}{\arg\max}\, \widehat{\mathcal{G}}_N(\xi, \tau_x(\xi)).$$

1.4.4 Sélection des paramètres de lissage

Le choix des fenêtres est toujours important dans les méthodes à noyau afin d'éviter un sur ou un sous-ajustement aux données, mais il devient ici crucial : de trop fortes oscillations de $\widehat{\mathcal{G}}_N(\xi, \tau_x(\xi))$ mèneraient par exemple à un choix erroné de $\widehat{\xi}_N^\star$. Seul le paramètre de lissage α apparait dans $\widehat{\mathcal{G}}_N$ et c'est donc à lui que nous allons nous intéresser mais une

procédure similaire peut être implémentée pour sélectionner le couple (α, β) apparaissant dans $\widehat{\mathcal{F}}_N$ (P9, Proposition 3.15).

Il est classique de sélectionner le paramètre α en minimisant l'ISE (Integrated Square Error),

$$\begin{aligned} \mathrm{ISE}_N(\alpha) &= \int_{\mathcal{C}_x} \left[\pi_\infty(\xi) G(\tau_x(\xi)|\xi) - \widehat{\mathcal{G}}_N(\xi, \tau_x(\xi))\right]^2 d\xi \\ &= \int_{\mathcal{C}_x} \pi_\infty(\xi)^2 G(\tau_x(\xi)|\xi)^2 d\xi \; + \; \varepsilon_N(\alpha), \end{aligned}$$

où

$$\varepsilon_N(\alpha) = \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_N(\xi, \tau_x(\xi))^2 d\xi - 2\int_{\mathcal{C}_x} \widehat{\mathcal{G}}_N(\xi, \tau_x(\xi)) \pi_\infty(\xi) G(\tau_x(\xi)|\xi) d\xi.$$

α apparait seulement dans $\varepsilon_N(\alpha)$, c'est donc cette quantité qu'on cherche maintenant à minimiser. Certes, l'inconnue $\pi_\infty(\xi)G(\tau_x(\xi)|\xi)$ continue à y apparaitre, mais on peut contourner cette difficulté par la validation croisée : l'intégrale contre $\pi_\infty(\xi)G(\tau_x(\xi)|\xi)d\xi$ peut être estimée en évaluant l'intégrande le long d'une chaîne de Markov dont la loi est celle-ci !

L'ISE que nous considérons ici n'est pas classique car fonction de données Markov et intégrée le long du flot. Mais la difficulté principale, spécifique aux PDMPs, que nous rencontrons concerne l'étape de validation croisée : en effet, il n'y a presque sûrement aucune donnée sur $\mathcal{C}_x$ dès que $d \geq 2$. Pour traiter ce problème, nous allons considérer pour la validation croisée les données présentes dans le tube $\mathbb{T}_{x,\rho}$ de rayon ρ autour de $\mathcal{C}_x$,

$$\mathbb{T}_{x,\rho} = \bigcup_{y \in \mathbb{D}_{x,\rho}} \mathcal{C}_y,$$

où $\mathbb{D}_{x,\rho}$ est l'intersection de la boule ouverte de centre x et de rayon ρ avec l'hyperplan orthogonal à $\mathcal{C}_x$. Par abus de notation, on note $\tau_x(\xi)$ le temps déterministe d'atteinte de $\mathbb{D}_{x,\rho}$ partant de $\xi \in \mathbb{T}_{x,\rho}$. L'estimateur de $\varepsilon_N(\alpha)$ est obtenu à partir de l'observation d'un autre PDMP $(\overline{X}_t)_{t\in\mathbb{R}_+}$, indépendant du premier et de même loi,

$$\begin{aligned} \widehat{\varepsilon}_{N,\overline{N},\rho}(\alpha) &= \int_{\mathcal{C}_x} \widehat{\mathcal{G}}_N(\xi, \tau_x(\xi))^2 d\xi \\ &- \frac{2\Gamma\left(\frac{d-1}{2}+1\right)}{\overline{N}\,\pi^{\frac{d-1}{2}}\rho^{d-1}} \sum_{k=0}^{\overline{N}-1} \widehat{\mathcal{G}}_N\big(\overline{X}_{T_k}, \tau_x(\overline{X}_{T_k})\big) \mathbb{1}_{\mathbb{T}_{x,\rho}}(\overline{X}_{T_k}) \mathbb{1}_{(\tau_x(\overline{X}_{T_k}),+\infty)}(\overline{S}_{k+1}), \end{aligned}$$

où Γ désigne la fonction Gamma d'Euler. Sous des hypothèses de régularité additionnelles (P9, Assumptions 3.13), on a le résultat de convergence suivant qui nous assure la bonne marche de la méthode de validation croisée.

Proposition 5 *(P9, Proposition 3.14) Conditionnellement à l'observation de* $(X_t)_{t\in\mathbb{R}_+}$,

$$\lim_{\substack{\overline{N}\to+\infty \\ \rho\to 0}} \widehat{\varepsilon}_{N,\overline{N},\rho}(\alpha) \;=\; \varepsilon_N(\alpha) \quad \text{p.s.}$$

En pratique, on ne s'attend pas à observer un second PDMP indépendant du premier. On peut dans ce cas séparer la trajectoire observée en deux parties, l'une servant à calculer les estimateurs et l'autre à l'apprentissage des paramètres de lissage.

L'algorithme d'estimation regroupant toutes les étapes développées ici a été implémenté et testé sur des exemples d'applications, dont des données réelles de propagation de fissure, en dimensions 1 et 2 (P9, 4 Simulation study).

1.5 Conclusion et perspectives

(P9) a clos l'aventure débutée pendant la thèse autour de l'estimation du taux de saut d'un PDMP général observé en temps long. Les spécificités de ces processus l'ont rendue passionnante. Il était difficile de prévoir a priori que la stratégie de Nelson-Aalen, basée sur le modèle à intensité multiplicative, s'avèrerait si ardue alors que la proximité PDMPs / processus de renouvellement markoviens la rendait si naturelle. Ce sont les calculs qui ont permis d'aller contre l'intuition et de développer malgré tout un premier algorithme d'estimation de la loi conditionnelle des durées inter-saut. Lorsqu'on cherche à minimiser la variance asymptotique, il faut à nouveau aller contre l'intuition et mener les calculs pour comprendre qu'il faut maximiser $\pi_\infty(\xi)G(\tau_x(\xi)|\xi)$ le long du flot et non la loi invariante, ce qui a nécessité d'établir le théorème 4. Enfin, la mise en place de la validation croisée a permis de mettre en évidence une difficulté déjà rencontrée dans (P3) lors de l'estimation du noyau Q : dès la dimension 2, lorsque Q est à densité, on ne dispose pas de données le long du flot. L'estimation des PDMPs recèle bien des surprises.

Mais finalement, laquelle des deux stratégies développées dans (P6) et (P9) est-elle la plus performante ? Il est très difficile de répondre à cette question notamment car les cadres de travail ont un peu évolué au cours du temps. Dans (P6), le temps et l'espace sont traités différemment (lissage vs discrétisation) et je souhaitais unifier les deux approches, certes au prix d'un espace d'état moins général. En fait, si (P9) est écrit pour un processus défini sur $\mathbb{R}^d$, toute la démarche et les preuves s'adaptent aisément au cas très utilisé en pratique d'un processus indexé par un mode discret. On peut néanmoins remarquer que trois estimateurs ont été nécessaires à la première approche, contre deux dans la seconde, mais surtout, le retour au processus d'origine nous oblige à intégrer en espace contre la seconde variable. Devoir estimer sur tout l'espace pour un résultat ponctuel reste décevant et a grandement motivé (P9–10). Afin de répondre plus rigoureusement, il faudrait selon moi reprendre le modèle à intensité multiplicative sur $\mathbb{R}^d$ avec lissage en espace et s'attacher à montrer un théorème de la limite centrale, dont la variance serait comparée à celle obtenue dans (P9) pour $f(t|x)$.

On voit qu'il n'est pas très judicieux d'estimer $\widetilde{\lambda}(t|x,y)$ pour revenir aux caractéristiques du processus d'origine, mais on pourrait procéder autrement : approcher directement la fonction de survie associée $\widetilde{G}(t|x,y)$ par l'estimateur de Kaplan-Meier (61) (ce qui se fait comme pour l'estimateur de Nelson-Aalen sous l'hypothèse de l'intensité multiplicative), puis appliquer un lissage en temps afin d'obtenir un estimateur direct de $\widetilde{f}(t|x,y)$. En dérivant (P6, eq. (22)), on a

$$\widetilde{f}(t|x,y) \propto f(t|x)Q(y|\Phi(t|x)).$$

En remplaçant $\widetilde{f}(t|x,y)$ par l'estimateur ainsi construit dans cette équation, puis en intégrant sur y, on obtient un estimateur, certes non-normalisé, de $f(t|x)$. On souffre toujours du problème de l'estimation globale pour n'estimer qu'en un point, mais cette méthode est indubitablement plus directe. C'est elle qu'il faudrait comparer à l'approche quotient.

L'intérêt principal de (P10) est de montrer une autre application de (P6) tout en explorant une nouvelle façon de revenir au processus d'origine. Mais il présente aussi deux limites conséquentes :

- En se contentant de $\lambda(x) = \lambda(\Phi(0|x))$, on n'estime le taux de saut qu'en les points chargés par $Q(dy|x)$. Certes, il est normal en temps long de n'estimer que sur le support de la loi invariante, mais, concernant la composée $\lambda \circ \Phi$, c'est sur le support de la loi invariante de $(X_{T_n}, S_{n+1})_{n\in\mathbb{N}}$ qu'on aimerait estimer.

- L'hypothèse que le support de $Q(dy|x)$ est fini est déjà plutôt restrictive, mais supposer de plus qu'il ne dépend pas de x est vraiment très fort. Pour éviter cette hypothèse et améliorer la méthode d'estimation en la rendant applicable à des noyaux à densité, on pourrait mettre en place une approche en deux étapes : d'abord, discrétiser l'espace d'état, e.g. par des méthodes de quantification optimale en norme $\mathbb{L}^p$ (50) (techniques que j'avais étudiées pour de la régression semi-paramétrique dans (P1)), puis appliquer l'estimateur de (P10) sur cette grille. C'est l'idée à peine abordée dans la partie numérique de (P10) mais qui pourrait avoir un certain intérêt pratique.

La limite de (P9) est de se contenter de l'estimateur de variance asymptotique minimale quand on pourrait combiner la famille d'estimateurs obtenue afin de construire un meilleur estimateur (variance plus petite, meilleure vitesse...) : c'est le problème de l'agrégation d'estimateurs (88). Dans notre cadre de travail, les estimateurs sont en quantité indénombrable, indexés le long du flot, et dépendants les uns des autres à travers les données partagées dans les noyaux en temps et en espace. Si on s'inspire de la stratégie développée dans (72) pour agréger ces estimateurs, il nous faudrait étudier leurs corrélations, ce qui s'annonce pour le moins difficile mais sans aucun doute riche en rebondissements.

De mon point de vue, les algorithmes développés dans (P6) et (P9) n'ont pas vocation à être appliqué à des cas pratiques pour lesquels on a un minimum d'expertise. Tenir compte des particularités du modèle ne peut qu'améliorer l'estimation, et (44, 66, 67) en sont de très bons exemples dans un cadre non-paramétrique et en temps long. Je vois plutôt ces problèmes comme un moyen d'acquérir de la connaissance sur les PDMPs et leurs spécificités, en particulier sur ce qui les différencie des processus de renouvellement. Par ailleurs, ces études nous fournissent des clés au moment de mettre en œuvre une méthode spécifique à un problème statistique faisant intervenir des PDMPs.

2

À la croisée des chemins

Croisements des processus markoviens déterministes par morceaux

Ce chapitre est dédié à des problèmes statistiques faisant intervenir des croisements de processus markoviens déterministes par morceaux (PDMPs pour l'anglais piecewise-deterministic Markov processes), et devrait être lu après le chapitre 1 (ou du moins sa partie introductive) afin de disposer de certains prérequis. Deux publications indépendantes seront présentées : (P7) porte sur l'estimation des caractéristiques d'absorption d'un modèle de croissance-fragmentation, alors que (P12) propose un estimateur du nombre moyen de croisements dans un cadre général.

2.1 Croisements

Dans le chapitre 1, on se pose la question de l'estimation des paramètres d'un PDMP aussi général que possible, observé en temps long, et sous une hypothèse d'ergodicité. En particulier, on montre comment estimer le taux de saut d'un tel processus. Mais les paramètres de la dynamique ne sont pas nécessairement des quantités d'intérêt pour certaines applications. D'autres caractéristiques, fonctions des trajectoires, peuvent être pertinentes, comme la valeur moyenne du processus ou le nombre moyen de sauts ayant eu lieu pendant un certain intervalle de temps.

On s'intéresse ici à l'une de ces fonctionnelles, précisément aux croisements de la trajectoire d'un PDMP avec un seuil critique. Dans les applications, ce dernier correspond à un niveau à ne pas dépasser sous peine d'engendrer une situation à risque. Il peut s'agir d'un modèle décrivant la présence d'un contaminant alimentaire dont on ne souhaite pas qu'il soit en excès, comme dans (12). Cette situation apparait également en fiabilité où le PDMP en jeu décrit la taille d'une fissure (25) ou un niveau de corrosion (36) dont la valeur ne doit pas dépasser un seuil d'alerte. Dans ces trois cas, le seuil critique correspond à une valeur haute du processus, mais il peut également s'agir d'une valeur basse : lorsqu'on modélise le capital d'un ménage par un PDMP, c'est une valeur minimale (représentant un seuil de pauvreté) qu'il faut éviter de franchir (65).

Ces problèmes appliqués peuvent être appréhendés essentiellement sous deux angles différents :

- Sous l'angle statistique, on cherche à estimer les caractéristiques des croisements, comme la probabilité de croisement, la loi de l'instant de premier croisement, le nombre moyen de croisements, etc. C'est le point de vue de (12, 25), mais aussi de ce chapitre.
- Avec le regard de la théorie du contrôle, on suppose qu'on peut intervenir sur la dynamique afin d'éviter ou de retarder la situation risquée. (36) traite par exemple d'un problème d'arrêt optimal : on cherche à stopper la trajectoire avant que celle-ci n'atteigne le niveau critique de corrosion d'un matériel, mais évidemment le plus tard possible pour éviter des maintenances trop fréquentes.

Ce chapitre présente deux publications plutôt différentes, mais qui traitent toutes les deux du problème statistique de l'estimation des caractéristiques des croisements. (P7), présentée dans la section 2.2, porte sur un modèle spécifique de croissance-fragmentation issu d'une application des PDMPs en assurance (65). Le croisement d'intérêt a lieu lors d'un saut du processus et correspond à son absorption (non-souhaitée) dont on va chercher à estimer la probabilité et la loi de l'instant en fonction de la condition initiale de la trajectoire. Dans (P12), décrite dans la section 2.3, on vise cette fois à estimer le nombre moyen de croisements continus (c'est-à-dire qui ne sont pas dus à un saut) d'une hyper-surface par un PDMP général sur $\mathbb{R}^d$.

2.2 Caractéristiques d'absorption

2.2.1 Un problème d'assurance

Le modèle en jeu dans cette partie est le processus de Markov $(X_t)_{t\in\mathbb{R}_+}$ de générateur

$$\mathcal{L}f(x) = r\max(x-1,0)f'(x) + \lambda\int_0^1 (f(zx)-f(z))G(z)dz,$$

où r et λ sont strictement positifs et G est une densité sur $[0,1]$. Sa dynamique peut être décrite ainsi :

- Partant d'une condition initiale plus grande que 1, le processus croit exponentiellement au taux r, alors que si celle-ci est plus petite que 1, il reste constant.
- Le processus subit ponctuellement des fragmentations : le processus décroit instantanément d'une certaine fraction de sa valeur, fraction tirée selon G, à des instants aléatoires distribués selon un processus de Poisson de paramètre λ.

On peut dès lors noter que l'intervalle $[0,1]$ est absorbant : lorsqu'une fragmentation amène le processus dans cet intervalle, il y reste piégé et va inexorablement plonger vers 0 au fil des fragmentations futures.

Ce modèle provient de (65) et relève en fait de problématiques d'assurance. On suppose qu'un foyer parvient à épargner à taux constant une part de ses revenus (la part au-dessus de 1, en unité arbitraire) et se constitue donc une épargne exponentiellement croissante au cours du temps. À des instants aléatoires, le foyer subit des événements indésirables qui l'amènent à puiser dans ses ressources. Lorsqu'une fragmentation fait

plonger le capital du foyer dans l'intervalle $[0, 1]$, il y reste : c'est le piège de la pauvreté (poverty trap en anglais). Les taux λ et r sont ce qu'ils sont, mais le foyer peut prendre une assurance qui l'aidera lors des événements fâcheux, avec deux conséquences :

- La part des revenus que le foyer parvient à épargner est réduite par le coût de l'assurance : il s'agit maintenant de la part au-dessus de $1 + \epsilon$.
- L'assurance permet au foyer de moins dépenser lors des fragmentations : la fraction de capital perdue sous la nouvelle loi G a tendance à être plus petite que sous la précédente loi. Mais attention, le piège de la pauvreté est aussi plus grand : $[0, 1+\epsilon]$.

Le problème qui se pose alors au foyer est de trouver une assurance G qui lui permette d'éviter d'atteindre la pauvreté (ou plutôt de l'atteindre avec une probabilité suffisamment faible ou dans un temps suffisamment long) avec un coût ϵ qui lui permette de continuer à épargner, le tout en tenant compte de son capital initial.

Les auteurs de (65) proposent notamment une technique numérique de calcul de la probabilité d'absorption, en présence ou non d'une assurance, afin de discuter la possibilité d'éviter le piège de la pauvreté. Constatant que les paramètres d'un tel processus sont a priori inconnus (sauf r qu'il suffit de demander à la banque), nous posons dans (P7) la question de l'estimation de la probabilité d'absorption, mais aussi de la loi du temps d'absorption. Ces quantités sont faciles à estimer par leur équivalent empirique lorsqu'on dispose d'observations indépendantes de plusieurs processus. Cependant, il faut noter que le triplet (λ, G, r) dépend du foyer. Il en est de même pour les caractéristiques d'absorption qui devraient donc être estimées à partir d'une seule trajectoire.

On suppose donc qu'on observe une seule trajectoire du processus $(X_t)_{t\in\mathbb{R}_+}$ en temps long. Comme le taux r est connu, ainsi qu'on l'avait remarqué dans le chapitre précédent, il est équivalent d'observer la dynamique en temps continu ou seulement lors des fragmentations, dont la suite des instants est notée $(T_n)_{n\in\mathbb{N}}$. On s'intéresse à l'estimation de la probabilité d'absorption

$$p(x) = \mathbb{P}(\exists\, t,\ X_t \in [0, 1] | X_0 = x),$$

et de la loi de l'instant d'absorption

$$t_m(x) = \mathbb{P}(X_{T_m} \in [0, 1], X_{T_{m-1}} \notin [0, 1] | X_0 = x).$$

Si elle n'est pas considérée dans (65), la loi de l'instant d'absorption parait être un excellent complément à la loi d'absorption dans le problème pratique d'assurance : si le piège de la pauvreté doit advenir mais dans un temps supérieur à la durée de vie du foyer, l'assurance parait peu utile.

2.2.2 Estimation semi-paramétrique d'un PDMP absorbant

Le processus $(X_t)_{t\in\mathbb{R}_+}$ est un PDMP (cf. le chapitre 1) sur $\mathbb{R}_+$ dont les caractéristiques locales sont les suivantes :

- Le taux est constant de valeur λ ;
- Le noyau de transition $Q(dy|x)$ est donné par

$$Q(dy|x) = \frac{1}{x} G\left(\frac{y}{x}\right) dy\ ;$$

- Le flot $\Phi(t|x)$ est donné par

$$\Phi(t|x) = \begin{cases} (x-1)\exp(rt)+1 & \text{si } x > 1, \\ x & \text{sinon.} \end{cases}$$

La loi des transitions peut sembler peu intuitive en regard de la description qualitative de la dynamique qui a été faite plus tôt. En fait, on peut montrer qu'il existe une suite i.i.d. $(Y_n)_{n\in\mathbb{N}}$ de loi G telle que

$$X_{T_n} = \Phi(T_n - T_{n-1}|X_{T_{n-1}})Y_n.$$

Les Y_n représentent les fractions de capital restant après les épisodes de fragmentation.

Dans le cas de l'observation de plusieurs trajectoires i.i.d., les caractéristiques d'absorption peuvent être estimées par leur équivalent empirique. Mais, comme dans le chapitre précédent, une seule version du processus est observée, jusqu'au N^e saut. Notre seul accès aux deux quantités d'intérêt $p(x)$ et $t_m(x)$ est donc le calcul, et celui-ci fera sans le moindre doute intervenir les inconnues λ et G. Celles-ci sont faciles à estimer : comme on vient de le voir, les durées inter-fragmentation et les fragmentations elles-mêmes sont i.i.d. Nous ne posons pas la question de leur estimation mais celle de l'estimation des caractéristiques d'absorption à partir de ces estimateurs.

Le cadre statistique est assez similaire à celui envisagé dans le chapitre précédent : on observe une seule trajectoire en temps long. On note deux différences principales :

- L'estimation des caractéristiques locales ne présente pas de difficulté particulière puisque la dynamique s'exprime comme fonction de variables i.i.d.
- Aucune hypothèse d'ergodicité ici puisque le processus est absorbant.

Nous sommes face à un problème d'estimation semi-paramétrique puisque la forme de G n'est pas spécifiée. Nous la supposons néanmoins bornée et telle que

$$\int_0^1 G(u)/u\,du < 1 + r/\lambda,$$

(cf. (P7, Assumptions 3.1) mais aussi les conditions imposées dans (P7, Theorems 3.6 & 3.8)).

De plus, on suppose qu'on dispose d'estimateurs $\widehat{\lambda}_N$ et $\widehat{G}_N$ de λ et G, construits à partir de l'observation des N premières fragmentations, qui vérifient les hypothèses suivantes :

- Il existe $0 < \lambda_* < \lambda < \lambda^*$ tels que $\widehat{\lambda}_N \in [\lambda_*, \lambda^*]$;
- $\widehat{\lambda}_N \xrightarrow{\mathbb{P}} \lambda$;
- $\|\widehat{G}_N - G\|_\infty \to 0$;
- $\int_0^1 |\widehat{G}_N(u) - G(u)|u^{-1}\,du \xrightarrow{\mathbb{P}} 0$.

Les deux premières conditions sont satisfaites par l'estimateur du maximum de vraisemblance tronqué : la tâche de déterminer λ_* et λ^* est laissée à un expert. La troisième est vérifiée si G est uniformément continue par l'estimateur de Parzen-Rosenblatt (83, 89) avec une condition additionnelle sur les fenêtres de lissage. Seule la quatrième hypothèse peut sembler baroque, mais nous montrons dans (P7, C Discussion on the condition C_2^G) qu'elle est vraie pour l'estimateur de Parzen-Rosenblatt dès lors que G est nulle sur un petit intervalle autour de 0 et suffisamment régulière.

2.2.3 Estimation du noyau de $(X_{T_n})_{n\in\mathbb{N}}$

Le noyau $R(dy|x)$ de la chaîne de Markov $(X_{T_n})_{n\in\mathbb{N}}$ est à densité $R(y|x)$ par rapport à la mesure de Lebesgue. Celle-ci s'écrit (P7, Proposition 3.2)

$$R(y|x) = \begin{cases} \frac{\lambda(x-1)^{\lambda/r}}{r} \int_0^{\min(y/x,1)} G(u) u^{\lambda/r} (y-u)^{-\lambda/r-1} du & \text{si } x > 1, \\ \frac{1}{x} G\left(\frac{y}{x}\right) & \text{sinon.} \end{cases}$$

La densité conditionnelle $R(y|x)$ est centrale pour exprimer les quantités d'intérêt $p(x)$ et $t_m(x)$. Son estimateur $\widehat{R}_N(y|x)$ est obtenu par plug-in en remplaçant les paramètres inconnus λ et G par leur estimateur vérifiant les hypothèses ci-dessus. On montre la convergence uniforme de cet estimateur.

Proposition 6 *(P7, Corollary 3.4) $\widehat{R}_N(y|x)$ converge, uniformément sur $[0,+\infty) \times [1,+\infty)$, en probabilité vers $R(y|x)$.*

Ce résultat est établi via une majoration de l'erreur uniforme entre $\widehat{R}_N(y|x)$ et $R(y|x)$ par les erreurs d'estimation de λ et de G (P7, Proposition 3.3). C'est en fait cette majoration qui est utilisée dans les démonstrations des convergences à venir.

2.2.4 Estimation de la probabilité d'absorption

L'idée ici est d'exprimer $p(x)$ comme solution d'une équation de Fredholm du second type sur $\mathbb{L}^1_{(1,+\infty)}$ (P7, Proposition 3.5),

$$p - Kp = s,$$

où l'opérateur intégral K est défini par

$$K : h \in \mathbb{L}^1_{(1,+\infty)} \mapsto \int_1^{+\infty} h(y) R(y|x) dy,$$

et $s(x) = \int_0^1 R(y|x) dy$. De manière classique, lorsque la norme subordonnée $|||K|||$ est strictement inférieure à 1, l'unique solution de cette équation peut être approchée numériquement en itérant l'opérateur K : après M itérations, p peut être approchée par

$$p_m = \sum_{m=0}^{M} K^m s.$$

On construit des estimateurs de l'opérateur K et de la fonction s en remplaçant dans leur expression $R(y|x)$ par son estimateur $\widehat{R}_N(y|x)$. On obtient l'estimateur suivant de p,

$$\widehat{p}_{M,N} = \sum_{m=0}^{M} \widehat{K}_N^m \widehat{s}_N,$$

dont la convergence est énoncée dans le résultat qui suit.

Proposition 7 *(P7, Theorem 3.6) L'équation de Fredholm $p - \widehat{K}_N p = \widehat{s}_N$ a une unique solution et $\|\widehat{p}_{M,N} - p\|_1$ tend vers 0 en probabilité lorsque M et N tendent vers l'infini.*

2.2.5 Estimation de la loi de l'instant d'absorption

On montre que la suite $(t_m)_{m\in\mathbb{N}}$ donnant la loi de l'instant d'absorption (comme fonction de la condition initiale) vérifie l'équation récursive

$$t_m = Kt_{m-1},$$

avec $t_1(x) = \int_0^1 R(y|x)dy$ (P7, Proposition 3.7). La suite des estimateurs $(\widehat{t}_{m,N}(x))_{m\in\mathbb{N}}$ est construite selon la même équation récursive dans laquelle l'opérateur K a été remplacé par son estimateur $\widehat{K}_N$, i.e. le noyau $R(y|x)$ a été remplacé par $\widehat{R}_N(y|x)$,

$$\widehat{t}_{m,N} = \widehat{K}_N\widehat{t}_{m-1,N},$$

avec $\widehat{t}_{1,N}(x) = \int_0^1 \widehat{R}_N(y|x)dy$.

On peut alors remarquer le lien existant entre $\widehat{p}_{M,N}$ et la suite $(\widehat{t}_{m,N})_{m\in\mathbb{N}}$,

$$\widehat{p}_{M,N} = \widehat{s}_N + \sum_{m=1}^{M} \widehat{t}_{m,N}.$$

Cette relation nous dit que les deux caractéristiques d'absorption peuvent être estimées en une seule passe, mais aussi que le schéma numérique mis en place pour l'estimation de p consiste à faire l'approximation suivante, pour M grand,

$$\mathbb{P}(\exists\, m, X_{T_m} \in [0,1]|X_0 = x) \simeq \mathbb{P}(\exists\, m \leq M, X_{T_m} \in [0,1]|X_0 = x). \tag{12}$$

Et on a le résultat de convergence suivant.

Proposition 8 *(P7, Theorem 3.8) Pour tout m, $\|t_m - \widehat{t}_{m,N}\|_1$ tend vers 0 en probabilité quand N tend vers l'infini.*

Outre les démonstrations des résultats de convergence, la suite de (P7) en présente des illustrations numériques.

2.3 Nombre moyen de croisements

2.3.1 Un problème de GPS

On considère ici un PDMP sur $\mathbb{R}^d$ et on se donne une barrière, i.e. une hyper-surface de l'espace d'état. On se demande combien de fois en moyenne la trajectoire va-t-elle franchir cette barrière de manière déterministe, i.e. à un autre instant que lors d'un saut, sur un horizon temporel $[0,T]$.

Dans le modèle de croissance-fragmentation précédent, le passage de la barrière 1 ne peut se produire qu'une fois, mais surtout se produit lors d'un saut et non en suivant le flot déterministe. On peut par contre se donner une barrière $x^* \gg 1$ (correspondant à un capital critique entrainant le paiement de nouveaux impôts par exemple) et se poser la question du franchissement de cette barrière, qui cette fois ne peut avoir lieu que de manière déterministe, mais peut bel et bien survenir plusieurs fois à cause des fragmentations.

On suppose maintenant que la trajectoire du processus est échantillonnée sur une grille temporelle discrète $(t_k)_{k\in\mathbb{N}}$, i.e. on dispose des $(X_{t_k})_{k\in\mathbb{N}}$. Un franchissement de l'hyper-surface S a lieu lorsque $[X_{t_k}, X_{t_{k+1}}] \cap S \neq \emptyset$, sans qu'on soit capable de distinguer les

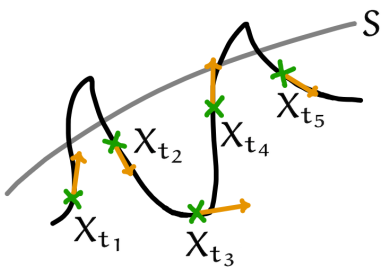


Figure 3: Trajectoire déterministe par morceaux (les sauts induisent des changements brusques de direction) franchissant la barrière S 4 fois mais échantillonnée d'un seul côté de celle-ci. Les vecteurs vitesses évalués en t_1 et en t_4 nous renseignent sur la possibilité d'un croisement entre t_1 et t_2 et entre t_4 et t_5. A contrario, le vecteur vitesse en t_3 nous indique qu'un croisement entre t_3 et t_4 est peu vraisemblable.

franchissements déterministes de ceux dus à des sauts sans hypothèses de modélisation. Bien sûr, si l'échantillonnage est suffisamment précis (par rapport à la vitesse du processus), on ne rate aucun croisement de S. Mais dans le cas contraire, on peut être amené à sous-estimer largement le nombre de croisements. Faisons alors l'hypothèse qu'on rate un croisement ayant eu lieu entre les instants t_k et t_{k+1}, i.e. X_{t_k} et $X_{t_{k+1}}$ sont du même côté de l'hyper-surface S. Si nous disposons de plus du vecteur vitesse $\dot{X}_{t_k}$, alors on a un indice exploitable sur l'éventualité du croisement raté ! C'est ce que décrit la figure 3.

Lorsqu'ils analysent le comportement d'animaux marins à partir des trajectoires qu'ils suivent, les biologistes se renseignent sur leur éloignement du nid via les franchissements de barrières virtuelles concentriques autour de celui-ci. C'est le cas dans (45) qui étudie des Goëlands bruns autour de leur nid basé sur l'île de Spiekeroog (île allemande de la mer du Nord). Les trajectoires sont enregistrées par des équipements portatifs GPS installés sur les oiseaux. Si on ajoute que certains GPS sont pourvus d'un accéléromètre fournissant une estimée du vecteur vitesse de la trajectoire, on comprend la pertinence du problème posé ci-dessus dès lors que les trajectoires peuvent être modélisées par des PDMPs.

Dans (P12), nous supposons que nous observons un PDMP et son vecteur vitesse le long d'une grille temporelle, et nous nous posons la question, sous différentes hypothèses, de l'estimation du nombre moyen de croisements continus d'une hyper-surface.

2.3.2 Processus vs hyper-surface

Nous nous intéressons donc aux croisements continus d'un PDMP $(X_t)_{t\in\mathbb{R}_+}$, défini sur $\mathbb{R}^d$, avec une hyper-surface S. Néanmoins, l'article (P12) est écrit pour la classe plus générale des processus lisses par morceaux (telle qu'introduite dans (17)), pour lesquels la distribution des sauts provient d'un processus ponctuel marqué. En particulier ces processus ne sont en général pas markoviens. Par ailleurs, on suppose que le flot $\Phi(\cdot|\zeta)$ est solution du problème de Cauchy

$$\frac{dx}{dt}(t) = r(x(t)), \qquad \text{avec} \quad x(0) = \zeta,$$

où r est dérivable à dérivée continue.

Les conditions demandées au PDMP sont classiques et peu restrictives : outre la non-explosion du processus ponctuel marqué (P12, Assumption 2.1), on suppose que, pour

tout t, X_t admet une densité p_t par rapport à la mesure de Lebesgue (P12, Assumption 2.2).

S est une hyper-surface compacte de $\mathbb{R}^d$ supposée $\mathcal{C}^1$. Dès la dimension 2, S est définie comme la frontière d'un domaine de $\mathbb{R}^d$, ce qui nous donne accès au vecteur unitaire sortant normal à la surface en x, noté $\nu(x)$. En dimension 1, $S = \{x\}$ et $\nu \equiv 1$ par convention.

On dit que le PDMP $(X_t)_{t\in\mathbb{R}_+}$ croise continûment S à l'instant τ si les deux conditions suivantes sont remplies :

- $X_{\tau^-} = X_\tau \in S$;
- Il existe $\delta > 0$ tel que $X_t \notin S$ pour tout $t \in (\tau - \delta, \tau + \delta) \setminus \{\tau\}$.

Nous supposons enfin que le flot n'atteint pas S de manière tangentielle, i.e. pour tout $x \in S$, le produit scalaire $(r(x), \nu(x))$ est non-nul (P12, Assumption 2.3). De plus, le nombre (déterministe) de croisements de S suivant le flot depuis n'importe quelle condition initiale doit rester fini à n'importe quel horizon fini.

2.3.3 État de l'art et objectif

En dimension 1, la barrière est un réel x et on a la formule suivante pour le nombre moyen de croisements $C(x, H)$ de celle-ci sur l'intervalle $[0, H]$ par un processus non-stationnaire,

$$C(x, H) = |r(x)| \int_0^H p_t(x)dt. \tag{13}$$

Ce résultat est énoncé et démontré dans (33). On en trouve une version dans le cas multi-dimensionnel mais pour des processus stationnaires dans (17). Dans ce contexte, l'objectif de (P12) est triple :

- Fournir une nouvelle démonstration de l'équation (13) (via le temps local) ;
- Étendre l'équation (13) à des processus non-stationnaires multi-dimensionnels ;
- Montrer l'intérêt d'un tel résultat pour le problème statistique cité dans l'introduction.

2.3.4 Formule de Kac-Rice

Le résultat principal de (P12) fournit le nombre moyen de croisements d'une hyper-surface S par un PDMP non-stationnaire sur une fenêtre temporelle donnée.

Théorème 9 *(P12, Theorem 2.10) Le nombre moyen $C(S, H)$ de croisements de S sur l'intervalle $[0, H]$ est donné par*

$$C(S, H) = \int_S |(r(x), \nu(x))| \int_0^H p_t(x)dt\, \mathfrak{h}_{d-1}(dx),$$

où $\mathfrak{h}_{d-1}$ est la mesure de Hausdorff de dimension $d-1$ sur $\mathbb{R}^d$.

Les formules données dans (17, 33) peuvent alors être vues comme des cas particuliers de ce résultat. De plus, on peut se contenter des croisements sortants (respectivement, des croisements entrants) en remplaçant $|(r(x), \nu(x))|$ par $\max((r(x), \nu(x)), 0)$ (respectivement, par $-\min((r(x), \nu(x)), 0)$).

2.3.5 Estimation du nombre moyen de croisements

Afin de construire un estimateur de $C(S,H)$, on envisage de procéder par plug-in en remplaçant la densité inconnue p_t apparaissant dans son expression par un estimateur. Dans le cas stationnaire $p_t = p_0$, la densité du processus peut être estimée à partir d'une seule trajectoire observée en temps long. Dans le cas non-stationnaire (et en l'absence d'hypothèses de modélisation), seule l'observation de plusieurs trajectoires nous permet d'accéder à cette quantité. Les deux cas sont considérés dans (P12) mais on se contente ici du second.

On observe N trajectoires i.i.d. $(X_t^n)_{t\in\mathbb{R}_+}$ le long d'une grille temporelle régulière $t_i = H(i-1)/(M-1)$, $1 \leq i \leq M$. Au pas de temps t_i, la densité p_{t_i} est estimée pour tout $x \in S$ par la méthode à noyau

$$\widehat{p}_{t_i,N}(x) = \frac{1}{N\sqrt{\det(B_N)}} \sum_{n=1}^{N} \mathbb{K}_d\left(B_N^{-1/2}\left[x - X_{t_i}^n\right]\right),$$

où la fenêtre de lissage B_N est une matrice symétrique définie positive. L'estimateur $\widehat{C}_{N,M}(S,H)$ de $C(S,H)$ est alors

$$\widehat{C}_{N,M}(S,H) = \int_S |(r(x),\nu(x))| \frac{H}{M-1}\left[\sum_{i=1}^{M} \widehat{p}_{t_i,N}(x)\right] \mathfrak{h}_{d-1}(dx).$$

Sous des hypothèses techniques sur le noyau $\mathbb{K}_d$ et la suite $(B_N)_{N\in\mathbb{N}}$, on a le résultat de consistance suivant.

Théorème 10 *(P12, Theorem 3.1)* $\widehat{C}_{N,M}(S,H)$ *converge presque sûrement vers* $C(S,H)$ *lorsque* N *et* M *tendent vers l'infini.*

La suite de (P12) consiste en une étude de simulations (P12, 4 Simulation study on piecewise deterministic Markov processes) visant notamment à comparer la formule de Kac-Rice estimée à l'évaluation empirique (qui compte simplement les croisements). Différents modèles sont considérés, en dimensions 1 et 2, pour des processus stationnaires et non-stationnaires continus (les sauts n'induisent que des changements de direction). Notons que dans ce cas, l'estimateur empirique ne peut que sous-estimer (au sens large) le nombre de croisements (le biais décroissant avec le pas de temps de la grille). Lorsque les sauts affectent également la position, l'évaluation empirique compte également les franchissements non-continus et peut alors (selon la fréquence des sauts et le pas de temps) sur-estimer la cible.

Une illustration sur les données de (45) (trajectoires de Goélands brun autour de leur nid) est également réalisée (P12, 5 Terrestrial and marine behaviors of a seabird species) avec la limite suivante : les données GPS ne contenant pas les vecteurs vitesses, ceux-ci ont été estimés. L'application de la formule de Kac-Rice n'utilise donc aucune information de modèle supplémentaire par rapport à l'évaluation empirique. Les résultats obtenus sont évidemment très similaires à une différence notable près. Si on regarde $C(S,H)$ comme une fonction de la distance de S au nid (qui permet typiquement de quantifier l'éloignement des oiseaux), la formule de Kac-Rice fournit une estimation plus lisse que l'évaluation empirique (P12, Fig. 14). En effet, elle fait intervenir une estimation lisse de la densité des trajectoires, qui a pour conséquence de lisser en espace le nombre de croisements.

2.4 Conclusion et perspectives

(P7) et (P12) ont permis d'aborder la question de l'estimation de fonctionnelles de PDMPs, liées aux croisements d'un seuil ou d'une hyper-surface, par des méthodes de plug-in. Même si le modèle de (P7) est très spécifique alors que celui de (P12) est assez général, la difficulté dans les deux cas est d'exprimer la cible en fonction de paramètres locaux facilement estimables puis d'étudier comment l'erreur d'estimation se propage de ceux-ci jusqu'à la cible.

Dans (P7), les caractéristiques d'absorption du PDMP $(X_t)_{t\in\mathbb{R}_+}$ sont définies comme celles de la chaîne de Markov $(X_{T_n})_{n\in\mathbb{N}}$ et sont estimées comme telles via son noyau R. Néanmoins, ce sont les caractéristiques G et λ qui sont accessibles : la difficulté du problème réside donc dans l'expression particulière de R en fonction de celles-ci. Il s'agit donc bien d'un problème de statistique des PDMPs et non des chaînes de Markov. On peut malgré tout arguer que la loi de l'instant d'absorption comme instant de fragmentation du processus ne se situe pas dans l'échelle de temps (continu) du processus d'intérêt mais dans celui (discret) de sa chaîne embarquée. C'est une limite de l'approche suivie dans (P7) : il serait plus intéressant du point de vue de l'application d'estimer la loi de $\inf\{t \geq 0 : X_t \in [0,1]\}$, directement ou via $t_m(x)$.

Au delà du potentiel intérêt applicatif, (P7) fournit un exemple d'approche hybride statistique / numérique pour les PDMPs. Certes les hypothèses sur les caractéristiques locales et les résultats de convergence sont spécifiques à la forme du noyau R mais la méthodologie est générique et peut facilement s'adapter à d'autres modèles. Par ailleurs, il serait judicieux de comparer cette approche calculatoire à celle développée dans (12) (appelée dans l'article simulation-based statistical inference) qui consiste à :

- Estimer les caractéristiques locales ;
- Simuler le modèle avec les paramètres estimés ;
- Estimer la quantité d'intérêt par Monte Carlo.

L'erreur issue de la mise en place de la méthode de Monte Carlo est très comparable à l'erreur numérique due à l'approximation (12) dans notre problème. Et dans les deux cas, on se demande comment va se propager l'erreur d'estimation à travers la méthode de calcul de la quantité d'intérêt. L'une de ces deux approches est-elle universellement meilleure que l'autre ? Si la réponse est non, peut-on identifier les familles de modèles sur lesquelles l'une est plus efficace que l'autre ?

Finalement, je vois surtout (P7) comme un premier pas vers le problème plus intéressant (du moins du point de vue de l'application) mais aussi plus difficile suivant. Le foyer a accès à un ensemble fini d'assurances $\{G_i\}_{1\leq i\leq I}$ (on peut supposer que l'une d'entre elles est égale à la loi des fragmentations de base, i.e. correspond en fait à l'absence d'assurance) et peut en changer à certains instants (e.g. lors des fragmentations pour conserver un cadre discret). Se pose alors un problème de bandit manchot : il faut choisir la meilleure assurance (au sens de la probabilité d'absorption par exemple) sans connaitre les G_i et en ayant donc à les estimer au fil de la trajectoire.

(P12) se distingue des approches statistiques développées dans le chapitre précédant ainsi que dans (P7) par le schéma d'observation sélectionné : échantillonnage le long d'une grille temporelle vs observation lors des instants de saut. Comme on l'a vu, lorsqu'on estime les caractéristiques locales du processus, le second schéma est particulièrement adaptée. Pour estimer la formule de Kac-Rice par plug-in, on a besoin de

la densité du processus au cours du temps et cette fois, c'est l'observation sur une grille temporelle qui est la plus adéquate. En particulier, les instants de saut ne sont d'aucune importance. On peut alors se poser la question de l'estimation de la formule de Kac-Rice lorsque le processus est observé lors des instants de saut, donc via l'estimation de la densité à partir des caractéristiques locales. Dans le cas non-stationnaire où plusieurs trajectoires sont observées, c'est sans intérêt : en effet, si on accède au processus lors des instants de saut et que son flot est supposé connu, on dispose d'une estimation directe par Monte Carlo du nombre de croisements. C'est lorsque le processus est stationnaire (ou seulement ergodique) et observé en temps long que la question prend tout son intérêt : il faudrait alors étudier comment l'erreur d'estimation des caractéristiques locales se propage jusqu'à la formule des croisements, mais également comprendre laquelle des stratégies d'inférence présentées dans le chapitre précédent est la plus adaptée à ce nouveau problème.

Lorsqu'on applique la formule de Kac-Rice aux trajectoires d'oiseaux observées par GPS, on semble ne faire aucune hypothèse forte de modélisation, en particulier dans le cas non-stationnaire. On suppose néanmoins que la partie déterministe évolue selon le flot d'une équation différentielle : cette condition est fausse notamment dans le cadre de trajectoires aller-retours. On peut malgré cela appliquer la formule de Kac-Rice en distinguant, au voisinage de la barrière, les trajectoires entrantes de celles sortantes (on sait qu'elles sont au même nombre car tous les oiseaux retournent au nid, dans les données utilisées du moins). A contrario, dans certaines zones (plutôt proches du nid, donc de moins d'intérêt pour étudier leur éloignement), on remarque que les oiseaux peuvent être amenés à prendre de nombreuses directions différentes : l'astuce précédente pourrait encore être appliquée mais après un clustering des trajectoires, qui reviendrait à modéliser la dynamique par un processus à mode (typiquement : éloignement vers le sud, retour depuis le sud, etc). L'application généralisée de la formule de Kac-Rice à ce type de données nécessite donc encore un travail conséquent.

3

La bifurcation

Noyau des sous-arbres et généralisations

Dans ce chapitre, on étudie un noyau de convolution pour les données arborescentes, appelé noyau des sous-arbres. La section 3.1 est dédiée à la définition et à la motivation de ces concepts. On discute la forme exponentielle de la fonction de poids du noyau des sous-arbres dans la section 3.2. La section 3.3 est consacrée à un nouvel algorithme de calcul de ce noyau, particulièrement efficace dans le cas d'évaluations répétées, qui permet de définir une nouvelle fonction de poids entrainée sur les données d'apprentissage. Les sections 3.2 et 3.3 sont basées sur (P18). La section 3.4 s'appuie sur (P24) et développe les outils théoriques nécessaires à l'introduction d'un nouveau noyau de convolution généralisant le noyau des sous-arbres. On conclut et présente quelques perspectives dans la section 3.5. Ce travail a été effectué dans le cadre de la thèse de Florian Ingels, de 2019 à 2022.

3.1 Arboriculture

3.1.1 Arbres : de quoi parle-t-on précisément ?

Arbres enracinés Un arbre est un graphe connexe sans cycle. Lorsqu'on désigne l'un de ses nœuds pour jouer le rôle de racine, on imprime une direction canonique sur le graphe : depuis n'importe lequel de ses nœuds, en suivant une arête, soit on s'éloigne strictement de la racine, soit on s'en rapproche strictement. Un arbre T dont un nœud a été désigné comme racine, notée $\mathcal{R}(T)$, est qualifié d'enraciné et est donc un graphe dirigé. Les notions de parent et d'enfant proviennent de cette direction et découlent donc directement du caractère enraciné :

- Depuis n'importe quel nœud v, il existe un seul chemin qui amène à se rapprocher strictement de la racine. Le premier nœud qu'on rencontre sur ce chemin est le parent de v, noté $\mathcal{P}(v)$. La racine est le seul nœud sans parent.
- Les enfants du nœud v sont tous ceux qui ont v pour parent, $\mathcal{C}(v) = \{w : \mathcal{P}(w) = v\}$. Les nœuds sans enfants sont appelés feuilles de l'arbre : leur ensemble est noté $\mathcal{L}(T)$.

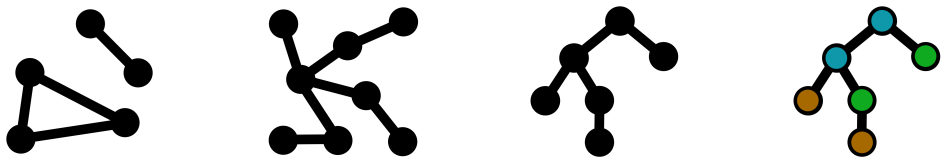


Figure 4: De gauche à droite : un graphe (non-connexe et avec un cycle) ; un graphe connexe et sans cycle, i.e. un arbre ; un arbre enraciné (la racine est en haut et imprime une direction canonique sur le graphe : de haut en bas, on part de la racine pour arriver aux feuilles) ; un arbre enraciné étiqueté (dont l'espace des étiquettes est un sous-ensemble de celui des couleurs).

La figure 4 présente le passage du concept de graphe à celui d'arbre enraciné en plus de celui d'arbre étiqueté qui sera introduit un peu plus tard.

Arbres non-ordonnés On peut estimer que l'ordre des enfants de chacun des nœuds d'un arbre enraciné n'a pas d'importance : dans ce cas l'arbre est dit non-ordonné. Formellement, on passe par la notion d'isomorphisme d'arbre : soient T et S deux arbres et $\varphi : T \to S$ une bijection ; on dit que φ est un isomorphisme d'arbre si elle préserve les relations parent-enfants, i.e. si $w \in \mathcal{C}(v) \Rightarrow \varphi(w) \in \mathcal{C}(\varphi(v))$. (Si on s'autorise à travailler avec des arbres de taille infinie, on doit ajouter $\varphi(\mathcal{R}(T)) = \mathcal{R}(S)$.) L'existence d'un isomorphisme d'arbre induit une relation d'équivalence sur l'espace des arbres enracinés : l'ensemble des arbres non-ordonnés est l'espace quotient par celle-ci.

On peut également définir la notion d'arbre ordonné pour laquelle l'ordre des enfants d'un nœud fait cette fois sens. Il faut noter en revanche qu'une part significative de la difficulté des travaux qui vont suivre provient du caractère non-ordonné des arbres considérés. Lorsqu'on sait traiter le cas non-ordonné, il suffit en général de fixer l'ordre pour passer aux arbres ordonnés. Parfois, la complexité temporelle des algorithmes reste étonnamment inchangée : déterminer l'existence d'un isomorphisme d'arbre en est un excellent exemple puisque l'algorithme est linéaire (en la taille des arbres) pour ces deux classes d'arbres. Néanmoins, si le problème est trivial pour les arbres ordonnés, il est plus subtil dans le cas non-ordonné (3, Theorem 3.3). Sauf mention contraire, les arbres de ce chapitre sont non-ordonnés.

Arbres étiquetés On peut munir chacun des nœuds d'un arbre d'une information additionnelle dont l'espace d'état reste à ce stade non spécifié, e.g. un ensemble fini ou un sous-ensemble de $\mathbb{R}^d$. Un tel arbre est qualifié d'étiqueté (cf. de nouveau la figure 4).

Quelques notions importantes Un concept clé dans ce chapitre est celui de sous-arbre : un sous-arbre $T[v]$ d'un arbre T est l'arbre composé du nœud v et de toute sa descendance dans T. Nous aurons aussi besoin de certaines caractéristiques topologiques, comme la hauteur : la hauteur $\mathcal{H}(T)$ d'un arbre T est la longueur du plus long chemin qui sépare la racine des feuilles. On peut la définir récursivement,

$$\mathcal{H}(T) = \begin{cases} 1 + \max\limits_{c \in \mathcal{C}(\mathcal{R}(T))} \mathcal{H}(T[c]) & \text{si } T \text{ n'est pas réduit à un nœud,} \\ 0 & \text{sinon.} \end{cases}$$

Autre notion topologique importante, le degré $\mathcal{D}(T)$ désigne le nombre maximal d'enfants dans l'arbre T,

$$\mathcal{D}(T) = \max_{v \in T} \#\mathcal{C}(v).$$

3.1.2 De la modélisation à des problèmes d'apprentissage

Le concept mathématique d'arbre a de nombreux intérêts applicatifs, à la fois (mais pas seulement) par son apparition spontanée dans la nature et par son utilisation ad infinitum en informatique. À propos du second point, on peut mentionner les arborescences de fichiers sur un ordinateur, la structure inhérente aux langages à balises de type `XML` (comme le `HTML`) ou encore les algorithmes de recherche fondés sur des décompositions arborescentes de l'espace, comme les octrees en computer graphics (76) ou les algorithmes CART (20) pour des problèmes de classification et de régression. Dans la nature, on peut penser à toutes les notions de généalogie (des lignées cellulaires aux arbres phylogénétiques représentant les relations de parenté entre des groupes d'êtres vivants), ainsi qu'aux plantes à différentes échelles (47) ou encore aux réseaux formés par les rivières et leurs affluents (69, pp. 130–131). Les données arborescentes sont ubiquitaires et requièrent donc une attention particulière quant à leur analyse.

Une question importante porte sur l'acquisition de ce type de données, triviale dans certains cas (l'arbre caché derrière un fichier `HTML` se construit linéairement et sans aucune source d'erreurs... sinon le développeur lui-même bien sûr), extrêmement ardue dans d'autres, en particulier en biologie (cf. par exemple (42) sur la reconstruction des lignées cellulaires en biologie des plantes). La question de l'acquisition n'est pas au cœur de mes préoccupations : seule (P22) a été l'occasion d'une excursion dans ce monde-là. Je fais plutôt l'hypothèse que les données sont disponibles et restent à analyser.

La modélisation est fortement connectée à l'acquisition : selon les applications envisagées et en fonction des informations disponibles dans les données, l'arbre sera considéré ordonné ou non, étiqueté ou non, avec le cas échéant un espace des étiquettes à identifier. L'échelle des objets étudiés est elle-même importante puisqu'elle peut faire apparaitre ou disparaitre leur caractère arborescent : les réseaux de rivières ne sont arborescents que vus de suffisamment loin.

L'étude des données arborescentes dépend évidemment du contexte applicatif, des hypothèses de modélisation qui peuvent être faites ou non, et surtout de la question posée. La formulation de celle-ci comme un problème d'apprentissage statistique est une option qui peut s'avérer pertinente. Par apprentissage statistique, on entend principalement les techniques de classification supervisée via lesquelles on souhaite apprendre, à partir de données classées, e.g. des photos de chats et de chiens, un classifieur capable de prédire avec succès si une nouvelle photo est celle d'un chat ou d'un chien. Par apprendre, on signifie deux choses :

- Le classifieur ne doit pas avoir été codé de manière explicite mais plutôt entrainé sur les données disponibles (90).
- Le classifieur devrait s'améliorer lorsqu'on augmente la base de ses données d'entrainement (78, 1.1 Well-posed learning problems).

Lorsqu'on s'intéresse à la structure des plantes par exemple, mettre en évidence et comprendre les différences entre les phénotypes de différentes espèces peut être traité par des techniques de classification supervisée, notamment via des algorithmes d'apprentissage des représentations. Ceux-ci cherchent à extraire des données d'entrainement leurs caractéristiques importantes pour le problème de classification, i.e. les caractéristiques qui comptent dans la prédiction, ce qui aide à l'interprétation du classifieur et donc à la compréhension des données. C'est précisément à ce problème-ci qu'on va s'intéresser dans ce chapitre. On peut citer (51, 93) pour des utilisations très différentes (en biologie et en informatique) mais récentes de ce genre de méthodes.

3.1.3 Approches à noyaux

L'espace des arbres n'admet pas de structure euclidienne (10), ce qui complique leur étude tout en accroissant évidemment leur intérêt. Ce constat étant fait, je distingue les grandes approches suivantes pour résoudre le problème de la classification supervisée :

- Un expert peut extraire des caractéristiques vectorielles de chacun des arbres et ce sont elles qui seront utilisées dans l'algorithme de classification. Elles peuvent décrire la topologie et/ou la distribution des étiquettes.
- On peut décrire la loi des données par un modèle probabiliste à plusieurs classes, en estimer ses paramètres, puis affecter la nouvelle donnée à prédire à la classe la plus vraisemblable selon ce modèle (52, 2.4 Statistical Decision Theory). Ce type de stratégie est employé dans (59) ; on pourra également consulter (37).
- L'espace des arbres peut être muni d'une distance (16) qui joue le rôle d'interface avec l'algorithme de classification, comme celui des k plus proches voisins (52, 13.3 k-Nearest-Neighbor Classifiers).
- Au lieu d'une distance qui augmente avec la dissimilarité des données, on peut au contraire considérer une fonction noyau qui quantifie leur similarité deux à deux. Les algorithmes SVM (30) ou les réseaux de neurones convolutifs (cf. (74) pour des problèmes non-supervisés) utilisent les données à travers le noyau sélectionné.

Sans aller jusqu'à des problèmes de classification, la question de la statistique inférentielle de modèles d'arbres aléatoires sera abordée dans le chapitre 4 dédié à (P13) et (P25). (P15–16) et (P26) traitent de sujets fortement liés à des calculs de distance entre arbres. Son titre est explicite, ce chapitre est consacré à la quatrième de ces approches, et plus précisément à (P18) et (P24). Si je ne me suis pas intéressé à la question des extractions de caractéristiques, nous allons voir qu'elle est, en un certain sens, équivalente aux approches à noyaux.

Un noyau sur un ensemble E est une fonction $\mathbf{K} : E^2 \to \mathbb{R}$ telle que pour tout N-uplet $(X_1, \ldots, X_N)$ d'éléments de E (N quelconque), la matrice de Gram $[\mathbf{K}(X_i, X_j)]_{1 \leq i,j \leq N}$ est symétrique semi-définie positive. Moralement, $\mathbf{K}(X_i, X_j)$ est un moyen d'évaluer la similarité entre X_i et X_j.

Dans la première approche mentionnée ci-dessus, viz. l'extraction de caractéristiques, on transporte les données depuis leur espace d'état E vers un espace $\mathbf{H}$ à travers une fonction $\phi : E \to \mathbf{H}$. Supposons maintenant que $\mathbf{H}$ est un espace de Hilbert, donc muni d'un produit scalaire $\langle \cdot, \cdot \rangle_{\mathbf{H}}$, et que l'algorithme de classification qu'on souhaite appliquer aux $\phi(X_i)$ utilise ce produit scalaire : c'est le cas notamment si on cherche un séparateur linéaire des classes (52, 4.5 Separating Hyperplanes). D'après le théorème de Moore-Aronszajn, se donner $\mathbf{H}$ et ϕ est équivalent à se donner une fonction noyau sur E (7) (cf. également le théorème de Mercer (30, Theorem 3.6) qui, sous des hypothèses plus fortes, a précédé (7) de près d'un demi-siècle).

Ce résultat majeur nous dit qu'un noyau $\mathbf{K}$, en plus de quantifier la similarité, est un émulateur de produit scalaire, sans qu'on ait à identifier l'espace de représentation $\mathbf{H}$, ni la transformation des données ϕ. La force des approches à noyaux vient alors du fait que des algorithmes très efficaces (on a déjà mentionné les SVMs et les réseaux de neurones convolutifs) ne font appel aux données qu'à travers leur produit scalaire. Une différence majeure entre l'extraction de caractéristiques et les approches à noyaux concerne toutefois ce qu'on modélise : dans le premier cas, on doit décrire la donnée ; dans le second, on met dans le noyau ce qu'on sait (ou suppose) de la similarité entre les données.

3.1.4 Noyaux de convolution ⊃ noyau des sous-arbres

Si la littérature regorge de propositions de noyaux adaptés à des contextes applicatifs et des problèmes de classification variés (cf. (91) en biologie computationnelle), la difficulté de notre problème est très liée au caractère combinatoire de l'espace auquel on s'intéresse. La construction par convolution (53) permet d'obtenir de grandes variétés de noyaux et s'applique non seulement aux arbres mais plus généralement aux espaces discrets comme les séquences et les graphes. Elle consiste à comparer les arbres sur la base de leurs sous-structures,

$$\mathbf{K}(T, S) = \sum_{s \in \mathcal{S}} w_s \varphi(N_T(s), N_S(s)), \tag{14}$$

où

- $\mathcal{S}$ est un ensemble de sous-structures des arbres ;
- w_s est le poids de la sous-structure s ;
- $N_T(s)$ et $N_S(s)$ comptent le nombre d'occurrences de s dans T et S ;
- φ est un noyau sur $\mathbb{N}$ ou $\mathbb{R}$. Dans la littérature, on choisit souvent $\varphi(n, m) = nm$, ce qui restreint la somme aux sous-structures communes à T et S puisque $\varphi(0, m) = 0$.

Les notions de sous-structures et de nombre d'occurrences sont volontairement vagues, mais permettent de regrouper sous une même formulation et sans technicité superflue pour ce document de nombreux noyaux. On renvoie la lectrice ou le lecteur vers (53) pour une présentation plus rigoureuse, plus technique, mais aussi plus élégante. Un exemple typique (auquel ce chapitre est consacré) est le noyau des sous-arbres (96), adapté aux arbres ordonnés ou non, étiquetés ou non. Comme son nom l'indique, l'ensemble des sous-structures $\mathcal{S}$ est l'ensemble des arbres lui-même, et le nombre d'occurrences est défini comme

$$N_T(s) = \#\{v \in T \ : \ T[v] = s\},$$

où le signe $=$ doit être compris dans l'espace quotient.

Lorsqu'on souhaite mettre en place un noyau de convolution, il faut choisir la famille de sous-structures $\mathcal{S}$ qui va permettre de quantifier au mieux la similarité entre les données de manière adaptée à l'application : c'est une question de modélisation. Mais cela doit être fait en regard de la difficulté du calcul du noyau induit. En effet, on peut être tenté de choisir une famille très riche de sous-structures mais la complexité de l'évaluation du noyau risque de devenir exponentielle, rendant le classifieur sans intérêt pratique. A contrario, choisir une famille de sous-structures pauvre induit un calcul facile du noyau mais une comparaison qui peut ne faire que peu de sens. Un bon noyau résulte d'un compromis entre la richesse des sous-structures et la complexité du calcul. Outre le noyau des sous-arbres, mentionnons le subset tree kernel (28) et le noyau des sous-chemins (64). On peut se rapporter à (75) pour un état de l'art sur les méthodes à noyaux sur des espaces d'arbres.

Le noyau des sous-arbres parait satisfaire ce compromis : les sous-arbres permettent de décrire la topologie de l'arbre dont ils sont extraits, et l'article original (96) propose, pour les arbres étiquetés, ordonnés ou non, un algorithme de calcul en temps linéaire, qui a ensuite été revisité pour les arbres ordonnés dans (4, 75). Mais dans les deux cas, la fonction de poids est supposée exponentielle en la taille des sous-structures (même si une stratégie algorithmique permettant de tenir compte de poids plus génériques est

brièvement décrite dans (96) pour le noyau des sous-chaînes), typiquement $w_s = \lambda^{\#s}$, $\lambda > 0$. Le calcul récursif du noyau exploite cette hypothèse à travers des formules du type

$$w_s = \lambda \prod_{c \in \mathcal{C}(\mathcal{R}(s))} w_{s[c]}.$$

La fonction de poids exponentielle parait avoir seulement un intérêt calculatoire. Mais on peut la justifier du point de vue de la modélisation ainsi : si un sous-arbre s est commun à T et S, alors tous ses sous-arbres le sont aussi et ont donc déjà compté dans le calcul du noyau entre T et S. Afin de compenser la multiplicité exponentielle des sous-arbres les plus petits dans le noyau, on accorde un poids décroissant exponentiellement vite aux arbres les plus gros : on suppose donc $\lambda \leq 1$.

L'histoire de ce chapitre débute ici : la fonction de poids exponentielle est-elle vraiment une bonne idée ?

3.2 Fonction de poids exponentielle : une bonne idée ?

Dans cette partie, reprise de (P18, 2 Theoretical study), on construit un modèle d'arbres aléatoires non-ordonnés ad hoc sur lequel on montre une propriété que devrait vérifier la fonction de poids du noyau des sous-arbres. Comme dans (4, 75, 96), on prend $\varphi(n, m) = nm$ et l'expression (14) du noyau est réduite aux sous-arbres communs à T et S.

3.2.1 Un modèle probabiliste arborescent à deux classes

On introduit ici un modèle probabiliste à deux classes, 0 et 1, dont les réalisations sont des arbres non-ordonnés et suffisamment simple pour montrer des résultats théoriques sur la séparabilité de ces deux classes par le noyau des sous-arbres.

Le modèle est défini à partir de deux arbres T_0 et T_1 (de même hauteur H), édités aléatoirement pour définir les données des deux classes. Précisément, les éditions aléatoires consistent à remplacer, dans T_0 ou T_1, un sous-arbre (tiré aléatoirement) par un arbre de la même hauteur issu d'une suite $(\tau_h)_{h \in \mathbb{N}}$ telle que $\mathcal{H}(\tau_h) = h$: les données de la classe i sont les arbres T_i^u obtenus comme T_i dans lequel le sous-arbre $T_i[u]$ de hauteur h_u a été remplacé par τ_{h_u}. Le nœud u est tiré uniformément au hasard parmi les sous-arbres de hauteur h dans T_i, où la hauteur h suit la loi binomiale $\mathfrak{B}_{H,\rho/H}$, de support $\{0, \dots, H\}$ et de moyenne $0 < \rho < H$.

On suppose désormais que les arbres T_0 et T_1 sont aussi différents que possible dans leur structure :

- $\forall i \in \{0,1\}, \forall u, v \in T_i \setminus \mathcal{L}(T_i)$, si $u \neq v$ alors $T_i[u] \neq T_i[v]$, i.e. deux sous-arbres de T_i ne sont jamais isomorphes, à l'exception des feuilles.
- $\forall u \in T_0 \setminus \mathcal{L}(T_0), \forall v \in T_1 \setminus \mathcal{L}(T_1)$, $T_0[u] \neq T_1[v]$, i.e. un sous-arbre de T_0 n'est jamais isomorphe à un sous-arbre de T_1, à l'exception des feuilles.
- Soient $u \in T_0$ et $v \in T_1$. On considère les arbres édités $T_0' = T_0^u$ et $T_1' = T_1^v$. Alors, $\forall u' \in T_0' \setminus (\tau_{h_u} \cup \mathcal{L}(T_0'))$, $\forall v' \in T_1' \setminus (\tau_{h_v} \cup \mathcal{L}(T_1'))$, $T_0'[u'] \neq T_1'[v']$, i.e. un sous-arbre d'un arbre de la classe 0 n'est jamais isomorphe à un sous-arbre d'un arbre de la classe 1, à l'exception des feuilles et des sous-arbres qui proviennent de la suite $(\tau_h)_{h \in \mathbb{N}}$.

L'ensemble des paires d'arbres vérifiant ces conditions est non-vide (cf. (P18, Fig. 2) pour un exemple). On peut commenter ces trois hypothèses :

- La première ne sert qu'à simplifier (ou rendre possibles) les calculs du noyau : le nombre d'apparitions de n'importe quel sous-arbre (les feuilles exceptées) dans T_0 ou dans T_1 est 1.
- La deuxième condition assure que du point de vue du noyau des sous-arbres, ces deux arbres sont les plus différents possibles. En effet, si $w_\bullet$ désigne le poids des feuilles,
$$\mathbf{K}(T_0, T_1) = w_\bullet \# \mathcal{L}(T_0) \# \mathcal{L}(T_1),$$
qui est la valeur minimale du noyau.
- La troisième condition assure que la proximité entre les arbres édités T_0' et T_1' ne provient que des éditions et non de changements collatéraux induits par elles. C'est aussi une condition calculatoire.

L'idée de ce modèle jouet est que les arbres T_0 et T_1 sont deux formats standards d'un fichier `XML` (typiquement des templates, par exemple de pages Web), aussi différents que possible. Ils sont ensuite édités avec le même contenu, modélisé par la suite $(\tau_h)_{h\in\mathbb{N}}$, pour former les données des deux classes.

- Lorsque le paramètre ρ est proche de 0, les éditions ont tendance à avoir lieu à la périphérie de l'arbre et à être peu massives. Les arbres de chacune des classes continuent de ressembler au modèle dont elles sont issues. Les deux classes restent facilement séparables sur la base du noyau des sous-arbres.
- Lorsque le paramètre ρ est grand, les éditions ont plutôt lieu proche de la racine et sont plus massives. Le modèle initial est fortement dégradé et les données des deux classes deviennent difficiles à distinguer sur la base de leurs sous-arbres.

3.2.2 Résultats théoriques

En s'appuyant sur la méthodologie développée pour les fonctions de similarité dans (11), on étudie théoriquement la pertinence du noyau des sous-arbres sur le modèle qu'on vient d'introduire pour l'algorithme de classification naïf suivant : on attribue à un nouvel arbre la classe qui maximise la moyenne des noyaux avec les données d'apprentissage de cette classe.

Pour simplifier les énoncés qui viennent, on introduit les notations suivantes,

$$\begin{aligned} C_{i,h} &= \frac{\mathbf{K}(T_i, T_i) - \max\limits_{\{u\in T_i \,:\, \mathcal{H}(T_i[u])=h\}} \mathbf{K}(T_i[u], T_i[u])}{\#\mathcal{L}(T_i)}, \\ G_\rho(h) &= 1 - \sum_{k=h+1}^{H} \mathfrak{B}_{H,\rho/H}(k). \end{aligned}$$

Proposition 11 *(P18, Corollary 3) Pour tout $0 \leq h < H$, un jeu de données d'apprentissage équilibré de taille*
$$\frac{2\max_i \mathbf{K}(T_i, T_i)^2}{\min_i C_{i,h}^2} \frac{\exp(2\rho)}{H^2} \log\left(\frac{2}{\delta}\right)$$
est suffisant pour que, avec probabilité au moins $1-\delta$, l'algorithme de classification susmentionné induise une erreur de classification d'au plus $1 - G_\rho(h) + \delta$.

Cet énoncé est équivalent (à la discrétisation de $[0, 1]$ induite par $G_\rho(h)$ près) à une erreur de classification d'au plus $1 - \epsilon + \delta$ pour tout $0 < \epsilon < 1$. À ma connaissance, c'est le premier résultat qui montre sur un modèle probabiliste (même naïf) la séparabilité des classes par le noyau des sous-arbres. Mais l'intérêt de cette étude réside surtout dans la proposition qui suit.

Proposition 12 *(P18, 2.4 Weight of leaves) Le nombre de données d'apprentissage suffisant pour induire une erreur d'au plus* $1 - G_\rho(h) + \delta$ *avec probabilité au moins* $1 - \delta$ *énoncé dans la proposition 11 est minimal lorsque le poids des feuilles dans* $\mathbf{K}$ *est nul.*

La proposition 12 montre, certes pour un modèle probabiliste ad hoc, que la forme exponentielle de la fonction de poids (4, 75, 96) est erronée puisqu'elle accorde un poids maximal aux feuilles. Un poids des feuilles non-nul augmente artificiellement la valeur du noyau au sein de chaque classe, sans faire croitre suffisamment sa valeur inter-classe, ce qui affaiblit la séparabilité. Par construction, les feuilles sont les seuls sous-arbres en commun entre les deux classes. Sans qu'on l'ait montré, il est tout à fait raisonnable de penser que, sur un modèle plus général, tout sous-arbre apparaissant systématiquement dans les deux classes devrait avoir un poids nul.

3.3 Calcul du noyau des sous-arbres

Dans cette partie, tirée également de (P18), on introduit un nouvel algorithme de calcul du noyau des sous-arbres basé sur leur énumération parcimonieuse par une technique de compression. Ceci nous permet en particulier de définir une nouvelle fonction de poids, apprise sur les données et qui tient compte du résultat théorique qu'on vient de montrer. Ces techniques sont adaptées aux arbres ordonnés ou non, étiquetés ou non, mais sont présentées préférentiellement pour les arbres non-ordonnés.

3.3.1 Compression DAG d'un arbre et d'une forêt

La compression DAG d'un arbre T consiste à construire un graphe dirigé acyclique (DAG pour l'anglais directed acyclic graph) qui en représente la structure sans les éventuelles redondances de sous-arbres (48, 95). Ce concept s'applique quelque soit le type d'isomorphisme considéré sur les arbres : ordonnés ou non-ordonnés, avec ou sans étiquettes, mais on l'explique ici dans le cas non-ordonné.

On considère le graphe dirigé $\Delta = (V, E)$ des classes d'équivalence des sous-arbres de T : V est l'ensemble des sous-arbres de T à isomorphisme près. Une flèche relie les nœuds u et v si v apparait sous la racine de u, i.e. $(u, v) \in E$ si et seulement si $v \in \mathcal{C}(\mathcal{R}(u))$ où le signe $\in$ s'entend ici à isomorphisme près. Enfin, on ajoute sur chaque flèche (u, v) une étiquette entière qui compte le nombre d'occurrences de v sous la racine de u.

Le graphe ainsi construit est un DAG qui compresse sans perte l'arbre T : il peut être reconstruit sans équivoque depuis le graphe Δ augmenté des multiplicités (cf. la figure 5 pour un exemple). À condition d'utiliser des tables de hachage, l'algorithme usuel de compression DAG des arbres non-ordonnés (48, 95) a pour complexité temporelle

$$O(\#T \times \mathcal{D}(T) \times \log \mathcal{D}(T)). \tag{15}$$

En éliminant les sous-arbres redondants, on obtient évidemment un gain de place en mémoire (19), mais on accélère aussi certains calculs (18).

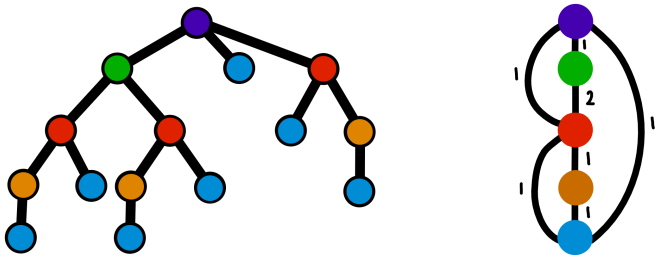


Figure 5: Un arbre non-ordonné T (à gauche) accompagné de sa compression DAG (à droite) : T n'est pas étiqueté mais les racines de ses sous-arbres ont été coloriées selon les classes d'équivalence, qu'on retrouve dans la version compressée. Dans celle-ci, on lit par exemple que le sous-arbre vert est composé de deux sous-arbres rouges, eux-mêmes composés d'un sous-arbre orange et d'un sous-arbre bleu.

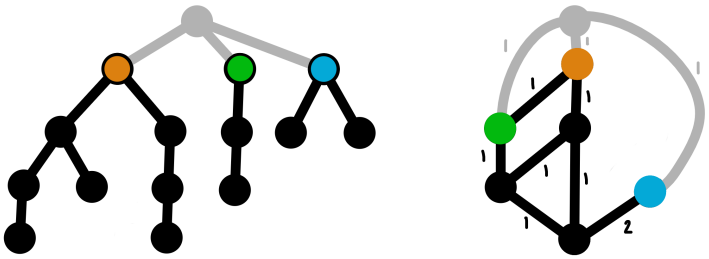


Figure 6: Une forêt (à gauche) accompagnée de sa compression DAG (à droite) : les trois arbres de la forêt ont été placés sous une racine virtuelle (grise) qu'on retrouve dans le DAG. Les flèches partant de la racine du DAG pointent vers les classes d'équivalence des arbres de la forêt (qui ont été coloriées pour faciliter l'association avec leur version non-compressée).

La compression DAG s'étend sans peine aux forêts (multi-ensembles finis d'arbres) : il suffit de placer les arbres de la forêt sous une même racine virtuelle puis de compresser l'arbre ainsi obtenu. Les racines des arbres de la forêt sont les enfants de la racine du DAG (cf. la figure 6). Dans (P18, Algorithm 1), on propose un algorithme de recompression DAG qui s'applique en particulier à la compression DAG d'une forêt à partir des compressions DAG des arbres qui la composent.

Proposition 13 *(P18, Proposition 7) Soient* $(D_1, \ldots, D_N)$ *les versions compressées des arbres non-ordonnés d'une forêt. La compression DAG de celle-ci se calcule à partir des* D_i *en temps*

$$O\left(\sum_{i=1}^{N} \#D_i \times \max_{1\leq i\leq N} \mathcal{D}(D_i) \times \left[\log \max_{1\leq i\leq N} \mathcal{D}(D_i) + \max_{1\leq i\leq N} \mathcal{H}(D_i)\right]\right),$$

où les notions de degré $\mathcal{D}$ *et de hauteur* $\mathcal{H}$ *prennent sur les DAGs le sens qu'elles ont sur les arbres (cf. la sous-section 3.1.1).*

3.3.2 Évaluation du noyau via la compression DAG des données

Par construction, la compression DAG d'une forêt permet l'énumération parcimonieuse de tous les sous-arbres qui apparaissent dans celle-ci : seuls les sous-arbres de la forêt

sont représentés par un nœud du DAG et chacun d'eux l'est exactement une fois. Par conséquent, le support du noyau des sous-arbres (14) est précisément le DAG de la forêt des données : les nœuds du DAG forment le plus petit ensemble tel que, quelle que soit la paire d'arbres considérés parmi les données, le noyau s'exprime comme une somme sur celui-ci.

Dans (P18, Proposition 12), on montre que le noyau des sous-arbres peut être évalué via la compression DAG de la forêt des données. Si la complexité spatiale est sans commune mesure avec celle de l'algorithme récursif originel (96) (car on doit stocker le DAG), la complexité temporelle peut elle s'avérer intéressante.

Que les données arborescentes $(T_1, \dots, T_N)$ soient accessibles dans leur forme compressée $(D_1, \dots, D_N)$ ou non, on calcule la compression DAG Δ de la forêt qu'elles composent (les complexités temporelles sont données par l'équation (15) si on dispose des arbres et dans la proposition 13 si on dispose des DAGs). En une seule traversée de Δ, on peut évaluer, pour chaque paire d'arbres (T_i, T_j), le support du noyau entre eux $\mathbf{K}(T_i, T_j)$ qui se calcule alors en $O(\min(\#D_i, \#D_j))$, ce qui est bien plus rapide que le $O(\#T_i + \#T_j)$ de (96). Autrement dit, une fois le prix de la compression DAG payé, le noyau peut être évalué très rapidement, ce qui est particulièrement adapté à des situations qui nécessitent des calculs répétés, comme le test de plusieurs fonctions de poids.

3.3.3 Fonction de poids

Dans la section 3.2, on regarde un modèle à deux classes pour lequel on montre que les feuilles, seules sous-structures communes aux deux classes, devraient avoir un poids nul. On suppose que dans des modèles plus généraux, toutes les sous-structures communes aux deux classes devraient avoir un poids nul. Dans un problème à $K \geq 2$ classes, on étend ce principe de la façon suivante : un sous-arbre qui apparait dans plusieurs classes devrait avoir un poids nul... sauf s'il apparait dans toutes sauf une puisque dans ce cas, il permet de la discriminer.

En pratique on ne s'attend pas à ce qu'une sous-structure apparaisse systématiquement dans tous les arbres d'une ou plusieurs classes ; on procède donc de la façon suivante. On note $p_k(s)$ la proportion des arbres de la classe k qui contiennent le sous-arbre s. Le vecteur $p(s) = (p_1(s), \dots, p_K(s))$ est un élément de l'hypercube unité de dimension K. Les cas favorables où s aide à discriminer une classe apparaissent lorsque ce vecteur est proche d'un des sommets

$$e_k = (0, \dots, 0, \underset{\uparrow}{1}_{k}, 0, \dots, 0) \qquad \text{ou} \qquad \bar{e}_k = (1, \dots, 1, \underset{\uparrow}{0}_{k}, 1, \dots, 1).$$

La fonction de poids qu'on propose est donc de la forme

$$w_s = f\left(\min_{k=1}^{K} \min(|p(s) - e_k|, |p(s) - \bar{e}_k|)\right), \tag{16}$$

où f est une fonction positive décroissante telle que $f(0) = 1$ (s discrimine, positivement ou négativement, une classe) et $f(1) = 0$ (l'argument de f dans l'équation (16) est supérieur à 1 dès lors que s apparait systématiquement dans deux classes tout en étant absent de deux autres, $K \geq 4$).

Pour définir cette fonction de poids, il est nécessaire de calculer le vecteur $p(s)$ pour chaque sous-arbre s susceptible d'apparaitre dans le calcul du noyau. La compression

DAG de la forêt des données est donc particulièrement adaptée à la construction de cette fonction de poids, qui se fait en temps $O(\#\Delta(N + K^2))$ (P18, Remark 15). Notons qu'en pratique, pour éviter tout phénomène de sur-apprentissage, on scinde le jeu de données d'apprentissage en deux parties : la première sert à construire la fonction de poids, alors que la seconde est utilisée pour apprendre le classifieur.

3.3.4 Résultats sur des jeux de données réelles

La méthodologie décrite ci-dessus, couplée à un classifieur de type SVM, a été déployée avec succès sur 8 jeux de données réelles, contenant des arbres ordonnés ou non, avec ou sans étiquettes (P18, Tab. 1), avec des résultats au moins aussi bons qu'une fonction de poids exponentielle dans 7 cas sur 8 et une amélioration relative du classifieur d'au moins 90% dans 3 cas sur 8 (P18, Fig. 18). Ce résultat est d'autant plus remarquable que la moitié des données d'apprentissage ont servi à apprendre la fonction de poids.

L'exemple phare auquel nous avons appliqué notre méthode est le suivant : on essaie de prédire la langue d'un article Wikipédia à partir de sa topologie non-ordonnée (formée par l'emboîtement arborescent de ses paragraphes, titres, listes numérotées ou non...) sans tenir compte du moindre contenu textuel. (J'invite la lectrice ou le lecteur à consulter (P18, Fig. 6) pour un exemple illustratif de ce qu'on entend par topologie du fichier `HMTL` associé à une page Web.) Il s'agit bien sûr d'un exemple jouet, mais notons que la question de la classification de pages Web à partir de leur topologie trouve certaines applications pratiques, en particulier dans la détection de sites frauduleux (93). Ajoutons que si l'application n'est pas des plus pertinentes, le problème de classification parait très difficile (et c'est la raison pour laquelle il a été choisi !).

Dans cette étude, on se restreint à des articles Wikipédia écrits dans l'une des 4 langues suivantes : anglais, allemand, espagnol et français ; et ceux composant les bases de données d'apprentissage et de test ont été tirés aléatoirement parmi les centaines de milliers, voire millions, de pages disponibles dans chaque langue. Peu importent le nombre de données d'apprentissage, la valeur du paramètre de décroissance et la mesure utilisée pour évaluer la classification, le taux de succès des SVMs en utilisant une fonction de poids exponentielle est autour de 50%. En utilisant la fonction de poids apprise sur les données, ce taux est systématiquement au-dessus de 90% et frôle parfois les 100% (P18, Fig. 8) !

La figure 7 reproduit le DAG d'un ensemble d'articles issus de Wikipédia (P18, Fig. 10). Les sous-structures aidant à discriminer les langues, par leur présence ou par leur absence, sont peu nombreuses devant le grand nombre de sous-arbres d'une telle forêt. Le poids (16) permet de les différencier et de leur donner une importance significative dans le calcul du noyau. En plus de la classification précise que cette fonction de poids induit, elle permet d'extraire les sous-structures d'intérêt et donc d'interpréter le classifieur. Par exemple, l'un des sous-arbres permettant de discriminer par leur présence les pages écrites en français correspond au copyright spécifique à cette langue. Mentionnons que depuis le format standard des articles Wikipédia a quelque peu évolué et les propriétés découvertes en 2019–2020 peuvent ne plus être valides aujourd'hui.

Enfin, si le poids (16) a de bonnes propriétés à la fois en classification et en interprétation, la forme exponentielle ne doit pas nécessairement être abandonnée : on a constaté sur les données Wikipédia que lorsqu'on calcule le poids (16) moyen par hauteur, on retrouve assez précisément une décroissance exponentielle (sauf pour les feuilles qui doivent avoir un poids nul comme prédit par l'étude théorique) (P18, Fig. 7).

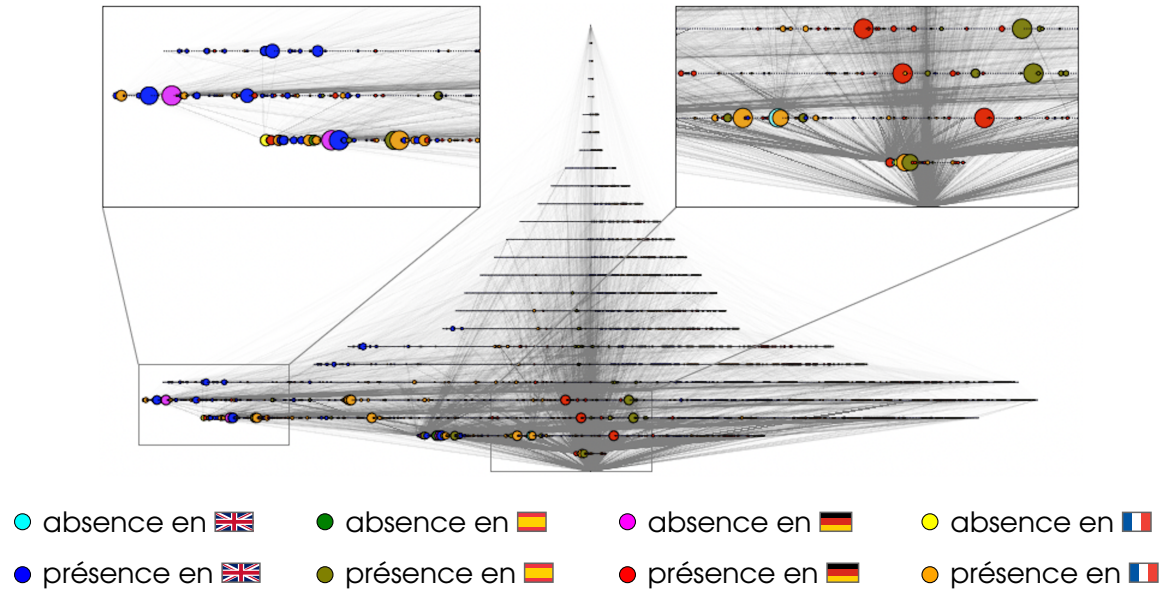


Figure 7: DAG d'une base de 240 articles dans 4 langues issus de Wikipédia. La taille d'un nœud correspond à son poids (16) et sa couleur au sommet de l'hypercube dont il est le plus proche. Par exemple, les nœuds bleus permettent de discriminer les articles écrits en anglais par leur présence, alors que les nœuds jaunes ont tendance à être présents dans toutes les langues sauf en français. Figure reproduite de (P18, Fig. 10).

3.4 Vers le noyau des sous-forêts

3.4.1 Limite du noyau des sous-arbres

Les noyaux de convolution (14) sont déterminés par la famille de sous-structures choisie pour comparer les données. On a déjà mentionné qu'un bon choix de sous-structures résulte d'un compromis entre leur richesse et la difficulté de calcul du noyau qu'elles induisent. Mais on sait désormais que la fonction de poids joue un rôle déterminant et que celle-ci peut être apprise sur les données dès lors qu'on sait énumérer le support du noyau.

Le noyau des sous-arbres est inopérant sur l'exemple de la figure 8. Pourtant, les sous-arbres sont la bonne famille de sous-structures à considérer, même si ce sont ici leurs paires qui font sens pour séparer les données. Plus généralement, il peut s'agir de N-uplets d'arbres (au lieu d'arbres seuls) qui permettent de discriminer des classes : les sous-structures à considérer dans la convolution sont alors les sous-forêts. Pour définir le noyau correspondant, avec l'estimation de la fonction de poids proposée précédemment, il nous suffit d'énumérer les sous-forêts de la forêt des données : c'est, pour les arbres non-ordonnés, la question qu'on se pose dans cette partie.

3.4.2 Recherche inversée

Par énumération, on entend ici la construction parcimonieuse de toutes les sous-forêts d'un jeu de données. Pour résoudre ce problème, on passe par la question plus difficile de l'énumération des forêts. Bien sûr, cet ensemble est infini : on ne s'attend donc pas

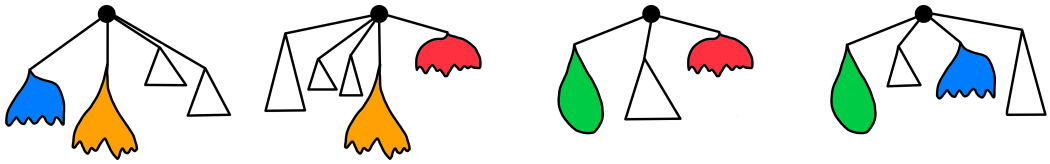


Figure 8: Aucun des sous-arbres colorés ne permet de différencier, seul, ces 4 arbres, que ce soit par sa présence ou par son absence. Mais les paires de sous-arbres le peuvent : l'arbre de gauche par exemple est le seul à avoir dans sa structure le sous-arbre bleu et le sous-arbre orange.

à construire explicitement toutes les forêts, mais plutôt à une stratégie de parcours de l'espace de sorte à atteindre chaque forêt une et une seule fois. Un tel parcours peut être décrit sous la forme d'un arbre d'énumération dont on attend (en plus de la couverture parcimonieuse de l'espace) les bonnes propriétés suivantes :

- Le nombre de successeurs d'une forêt (dans l'arbre d'énumération) doit être linéaire en la taille de la forêt ;
- Leur construction doit être polynomiale en la taille de la forêt ;
- La construction des forêts de profondeur $K+1$, i.e. qui nécessitent $K+1$ pas dans l'arbre d'énumération, à partir des forêts de profondeur K, doit se faire en temps polynomial.

En d'autres termes, un bon arbre d'énumération ne doit pas croitre trop vite.

Comme lorsqu'on construit un estimateur du taux de saut d'un processus déterministe par morceaux dans le chapitre 1, on poursuit deux objectifs intrinsèquement liés : on doit construire un arbre d'énumération des forêts tout en garantissant ses propriétés souhaitées. Pour traiter ce problème, on procède par recherche inversée (9) (cf. également (82, pp. 45–51)), technique utilisée avec succès pour l'énumération des arbres ordonnés (80) puis non-ordonnés (81).

Plus formellement, l'espace qu'on cherche à parcourir est un ensemble partiellement ordonné $(E, \leq)$ admettant un plus petit élément $\emptyset$. L'énumération par recherche inversée passe par la construction d'une règle de réduction $f : E \setminus \{\emptyset\} \to E$ qui vérifie les deux propriétés suivantes :

- Pour tout $x \in E \setminus \{\emptyset\}$, $f(x) \leq x$;
- Pour tout $x \in E \setminus \{\emptyset\}$, il existe un entier k tel que $f^k(x) = \emptyset$.

Il s'agit alors d'inverser f (au sens de la préimage, i.e. $f^{-1}(y) = \{x \in E : f(x) = y\}$) pour munir E d'un arbre d'énumération enraciné en son plus petit élément $\emptyset$.

Si on se donne de plus une propriété $g : E \to \{\top, \bot\}$ anti-monotone (si $x \leq y$ et $g(y)$ alors $g(x)$, i.e. une propriété vérifiée sur un élément l'est aussi sur les éléments plus petits), alors on peut filtrer l'arbre d'énumération pour ne parcourir que les éléments qui vérifient g. C'est exactement ce qu'on fera pour ne retenir de l'arbre d'énumération que les sous-forêts de la forêt des données : la propriété « être une sous-forêt des données » est bien anti-monotone.

3.4.3 Énumération des forêts non-redondantes

Par souci de parcimonie, on cherche à n'énumérer que les forêts non-redondantes, i.e. les forêts $(T_1, \dots, T_N)$ telles qu'aucun des arbres non-ordonnés T_i n'apparait comme sous-arbre d'un des T_j. (On montre dans (P24, 4.1 Extension to forests with repetitions) qu'on peut a posteriori ajouter de la redondance.) La difficulté provient du caractère non-ordonné des arbres qui composent les forêts qu'on souhaite énumérer, de leur propre caractère non-ordonné (on ne souhaite pas énumérer la forêt (T_1, T_2) puis plus tard la forêt (T_2, T_1)), et enfin de la condition de non-répétition.

Lorsque se posent des problèmes d'ordre dans des tâches d'énumération, on peut passer par la définition d'une forme canonique des objets d'intérêt, qui devrait fixer tous leurs degrés de liberté. On énumère alors les formes canoniques sans se soucier d'éventuelles répétitions. C'est notamment l'option choisie dans (81) où la forme canonique d'un arbre non-ordonné est sa version ordonnée dont la masse est à gauche.

On se propose d'énumérer les forêts non-redondantes directement dans leur forme compressée, ce qui permet de traiter plus facilement la condition de non-répétition et d'être plus efficace, algorithmiquement et numériquement, dans l'utilisation des résultats. On doit donc identifier une forme canonique pour les DAGs compressant des forêts (qui forment un sous-ensemble des DAGs), ce qu'on propose de faire via la notion d'ordre topologique.

Un ordre topologique sur un DAG Δ est une bijection $\psi : \Delta \to \{0, \dots, \#\Delta - 1\}$ telle que $\psi(x) > \psi(y)$ s'il existe une flèche de x vers y. On sait que les DAGs admettent un ordre topologique (en fait, un graphe dirigé est un DAG si et seulement s'il admet un ordre topologique (60)) mais il n'y a en général pas unicité : on peut même s'intéresser à l'énumération par recherche inversée des ordres topologiques d'un DAG (9) ! Pour déterminer une forme canonique, nous allons donc devoir être plus subtils. Le résultat qui suit définit une forme canonique sur les DAGs compressant des forêts.

Théorème 14 *(P24, Theorem 2.4) Un DAG compresse une forêt non-redondante si et seulement s'il admet un unique ordre topologique ψ qui vérifie :*

- *Si* $\mathcal{H}(x) > \mathcal{H}(y)$ *alors* $\psi(x) > \psi(y)$ *;*
- *Si* $\mathcal{H}(x) = \mathcal{H}(y)$ *et* $\mathcal{C}_\psi(x) > \mathcal{C}_\psi(y)$ *alors* $\psi(x) > \psi(y)$*,*
  $\uparrow$ *ordre lexicographique*

où $\mathcal{C}_\psi(x)$ *désigne le* N*-uplet* $(\psi(u) : u \in \mathcal{C}(x))$ *rangé dans l'ordre décroissant, et où les notions de hauteur* $\mathcal{H}$ *et d'enfants* $\mathcal{C}$ *ont le même sens sur les DAGs que sur les arbres.*

Ce résultat étant établi, on construit une règle de réduction (P24, Theorem 3.12) dont la préimage se décompose selon les 3 règles suivantes (qui au passage préservent la canonicité (P24, Proposition 2.10)) :

- Branchement (P24, Definition 2.6) : on ajoute dans Δ une flèche entre v (nœud le plus grand de Δ selon l'ordre topologique ψ du théorème 14) et un nœud de hauteur strictement plus petite de telle sorte que $\mathcal{C}_\psi(v)$ reste décroissant (au sens de l'ordre lexicographique) ;
- Élongation (P24, Definition 2.7) : on ajoute à Δ un nœud de hauteur $\mathcal{H}(\Delta) + 1$ dont l'unique enfant se situe à la hauteur $\mathcal{H}(\Delta)$;

- Élargissement (P24, Definition 2.8) : on ajoute à Δ un nœud w de hauteur $\mathcal{H}(\Delta)$ dont les enfants vérifient $\mathcal{C}_\psi(w) > \mathcal{C}_\psi(v)$ (où v était, avant l'ajout de w, le nœud le plus grand de Δ selon l'ordre topologique ψ).

On définit ainsi un arbre d'énumération des forêts non-redondantes directement dans leur forme compressée (cf. (P24, Fig. 7) pour en voir une toute petite portion), et on contrôle finement sa croissance. Dans le résultat qui suit, on note E_K (respectivement $E_{\leq K}$) l'ensemble des DAGs compressant des forêts à profondeur K (respectivement à profondeur au plus K) dans l'arbre d'énumération.

Théorème 15 *L'arbre d'énumération vérifie les propriétés suivantes :*

- *(P24, Theorem 3.3) Le nombre de successeurs d'un DAG compressant une forêt Δ est $\Theta(\#\Delta)$;*
- *(P24, Proposition 3.5) Ils se construisent incrémentalement en temps $O(\#\Delta \times \mathcal{D}(\Delta))$;*
- *(P24, Theorem 3.6) Énumérer $E_{\leq K+1}$ se fait en temps $O(K^2 \# E_{\leq K})$;*
- *(P24, Theorem 3.1) et (21, 58) Quand K tend vers l'infini,*

$$E_K = K! \left(\frac{12}{\pi^2}\right)^K \left(\frac{6\sqrt{2}}{\pi^2} \exp\left(\frac{\pi^2}{24}\right) + O\left(\frac{1}{K}\right)\right).$$

Le dernier point provient d'une équivalence inattendue entre les matrices d'adjacence (à quelques transformations près) des DAGs compressant des forêts et les matrices de Fishburn lignes (P24, A A bijection between FDAGs and row-Fishburn matrices).

Dans le problème de calcul du noyau, on souhaite n'énumérer que les sous-arbres d'une forêt de données, ce qu'on fait en contraignant l'arbre d'énumération qu'on vient de construire à rester dans celle-ci (P24, 5 Enumeration of forests of subtrees). Si malgré tout il s'avère que le nombre de sous-forêts est trop grand pour l'application considérée, on montre, en s'inspirant de (8) pour les sous-arbres non-ordonnés, comment on peut ne générer que les sous-forêts suffisamment fréquentes (P24, 5 Enumeration of forests of subtrees).

3.5 Conclusion et perspectives

La forme exponentielle de la fonction de poids du noyau des sous-arbres ne semble pas être un choix judicieux en classification supervisée, à la fois d'un point de vue théorique mais aussi sur des jeux de données réelles. L'alternative proposée dans (P18), finalement très empirique même si fondée sur des résultats probabilistes pour un modèle spécifique, montre qu'on peut améliorer significativement les taux de prédiction sans changer la famille de sous-structures dans la convolution et en conservant le même algorithme de classification. Toutefois, elle demande l'énumeration explicite de l'ensemble des sous-structures présentes dans les données, ce qu'on fait via un algorithme de compression pour le noyau des sous-arbres. On a poursuivi cette idée dans (P24) aux ordres ultérieurs en mettant en place tous les outils théoriques et algorithmiques nécessaires à l'introduction du noyau des sous-forêts par des méthodes d'énumération par recherche inversée.

L'histoire ne peut s'arrêter ainsi : il nous faut maintenant mettre en place toute la machinerie décrite dans (P18) pour ce nouveau noyau (à ma connaissance, on ne le trouve nulle part dans la littérature). Deux questions importantes sont en suspens :

- La méthode d'énumération par recherche inversée développée dans (P24) permet-elle d'évaluer sa matrice de Gram en un temps raisonnable sur des jeux de données réelles comme ceux utilisés dans (P18) ?
- Permet-elle d'améliorer encore les taux de classification et/ou de mieux comprendre ce qui caractérise les données ? Idéalement, on aimerait trouver des paires de sous-arbres typiques de certaines classes comme dans l'exemple de la figure 8.

Dans le cas où l'énumération exhaustive des sous-forêts ne permet pas un calcul suffisamment rapide du noyau, plusieurs solutions sont envisageables. La première, déjà mentionnée, consiste à n'énumérer que les sous-forêts suffisamment fréquentes, ce qui nécessite l'introduction d'un seuil dont il faudra fixer la valeur intelligemment. Mais si se restreindre aux sous-structures les plus nombreuses est une solution au problème numérique, elle peut aussi aller à l'encontre du message principal de ce chapitre : les sous-structures les plus fréquentes apparaissent a priori dans toutes les classes et ne permettent donc pas de les discriminer. On pourra peut-être résoudre cette apparente contradiction en s'inspirant de (87) où des approximations de noyaux de convolution pour les arbres sont construites en restreignant, par des techniques de programmation dynamique, leur support aux sous-structures les plus pertinentes pour le problème considéré.

Dans le contexte des méthodes à noyaux, la méthode de Nyström consiste à approcher la matrice de Gram $[\mathbf{K}(T_i, T_j)]_{1 \leq i,j \leq N}$ par une matrice de rang plus faible (98), via la sélection (aléatoire ou basée sur leurs propriétés (79)) de certaines données. Elle est utilisée dans des cas où la matrice de Gram ne peut être stockée en mémoire ou pour réduire (ou rendre envisageables) les temps de calcul du classifieur. On pourrait considérer la méthode de Nyström comme un moyen de compenser le coût numérique de l'énumération des sous-forêts. Par ailleurs, on peut s'interroger sur le lien existant entre un algorithme comme celui de (79) (qui commence par trier les données selon leur pertinence pour le problème de classification avant d'en sélectionner un petit nombre) et notre approche qui procède finalement de manière assez similaire sur les sous-structures.

La définition de la fonction de poids apprise sur les données est inspirée par le calcul établi pour le noyau des sous-arbres dans le cadre d'un modèle probabiliste à deux classes ad hoc. On montre que le poids des feuilles, seules sous-structures communes aux deux classes, devrait être nul, ce qu'on généralise en trois étapes :

- Toute sous-structure commune aux deux classes devrait avoir un poids nul ;
- Dans un problème multi-classes, toute sous-structure qui apparait dans plusieurs classes ou est absente de plusieurs classes devrait avoir un poids nul ;
- Dans la pratique, on considère la version continue de ce principe : le poids d'une sous-structure est fonction de sa distance à ces deux archétypes.

Si les résultats obtenus sur les différents jeux de données semblent donner raison à ces généralisations successives, force est de constater que rien de tout ça n'a été montré théoriquement. Il serait de mon point de vue pertinent de s'y intéresser sans spécifier les sous-structures en jeu dans la convolution. En plus de la forme du noyau, ajoutons que le calcul utilise fortement l'hypothèse $\varphi(n, m) = nm$ et qu'il serait judicieux d'envisager d'autres noyaux φ.

La question des étiquettes a été très peu évoquée tout au long de ce chapitre. Néanmoins, les méthodes de (P18) s'étendent aisément aux arbres (ordonnés ou non) portant des étiquettes discrètes car la compression DAG leur est adaptée. Mais concernant l'énumération des forêts et des sous-forêts, le passage à ces objets plus compliqués devrait nécessiter un peu plus de travail. Dans (P21), on développe de nouvelles techniques algorithmiques qui prennent en compte les étiquettes à réécriture près. Typiquement, on aimerait que deux arbres isomorphes dont les étiquettes de l'un sont toutes obtenues en appliquant une transformation bijective (inconnue) à leur équivalent dans l'autre soient considérés comme un même motif. On espère construire sur cette base de nouveaux noyaux pertinents, qui généraliseront quoi qu'il en soit les méthodes de (P18). Enfin, les étiquettes continues restent un problème difficile pour lequel on pourra s'inspirer des algorithmes introduits dans (31) pour le noyau des sous-arbres.

Dans ce chapitre, les données arborescentes sont systématiquement comparées sur la base de leurs sous-arbres. Dans certaines applications, de telles comparaisons ne semblent pas pertinentes, e.g. lorsqu'on observe des généalogies tronquées comme dans (P25). Dans d'autres, les sous-arbres sont la bonne sous-structure à considérer du point de vue de la modélisation sans que les résultats de classification ne le montrent. En effet, l'étape d'acquisition peut induire des erreurs dans la topologie des données (e.g. causées par une sur-segmentation) qui rendent la comparaison par les sous-arbres erronée : il est facile de se convaincre qu'un bruit topologique de très faible ampleur (ajout ou suppression de quelques nœuds à proximité des feuilles) peut rendre deux arbres isomorphes très dissemblables pour le noyau des sous-arbres. Dans ce cas, on peut commencer par éliminer le bruit afin de reconstruire les topologies cachées des données pour lesquelles le noyau des sous-arbres sera efficace. Une version de ce problème est la détermination de l'arbre auto-emboîté le plus proche d'un arbre au sens de la distance d'édition (qui évalue le nombre d'opérations élémentaires nécessaires pour transformer un arbre en un autre) : un arbre est dit auto-emboîté lorsque tous ses sous-arbres d'une même hauteur sont isomorphes. Un algorithme polynomial est présenté dans (48) pour un bruit soustractif. Dans (P15), j'ai proposé une solution exacte à ce problème lorsque le bruit est additif ainsi qu'un nouvel algorithme dans le cas soustractif réalisant une meilleure complexité que (48). On met au point une heuristique adaptée à des bruits quelconques dans (P16). Dans (P26), on construit un algorithme incrémental de calcul de la distance d'édition le long de marches aléatoires sur l'espace des arbres non-ordonnés, comme préalable à la mise en place d'un algorithme de recuit simulé pour la résolution approchée de ce problème d'optimisation.

Enfin, même si le contexte, les objets et les outils utilisés sont un peu différents, je souhaite mentionner ici une importante source d'inspiration et de réflexions : le modèle de réseau de neurones `word2vec` (77). En supposant que le nombre de mots d'une langue est certes grand mais fini (disons N), on peut les voir comme des vecteurs en utilisant le one-hot encoding : le n^{e} mot du dictionnaire est encodé par un vecteur de taille N composé d'un unique 1 en position n entouré de zéros. Cette représentation passe ensuite à travers les couches d'un réseau de neurones entrainé pour apprendre des propriétés statistiques d'une langue. Les résultats impressionnants qui ont été obtenus montrent qu'une représentation vectorielle naïve de très grande dimension peut être pertinente malgré la combinatoire tant que le modèle d'analyse qui suit est suffisamment riche et bien estimé. Dans (70), des arbres partiellement ordonnés étiquetés (représentant des équations) sont considérés comme entrées et sorties d'un réseau de neurones du type `seq2seq` entrainé pour calculer des primitives, montrant au passage que la combinatoire des arbres peut être absorbée par ce type de modèle. Mes tentatives de construction d'un réseau `tree2vec` restent malgré tout vaines pour l'instant.

4

Une dernière excursion

Estimation de modèles de Galton-Watson conditionnés

Dans ce chapitre, on se concentre sur des problèmes d'inférence statistique pour des modèles d'arbres aléatoires obtenus comme des processus de Galton-Watson conditionnés. Deux articles sont présentés ici : (P13) concerne le conditionnement par la taille, alors que (P25) traite du conditionnement à la survie. Même si le cadre théorique et les problèmes posés sont de nature différente, on réutilise de nombreuses définitions et notations posées dans le chapitre 3.

4.1 Conditionnement des arbres de Galton-Watson

4.1.1 Modèle de Galton-Watson

Le modèle de Galton-Watson est une loi de probabilité sur les arbres ordonnés définie à partir d'une loi μ sur $\mathbb{N}$, dite loi de naissance. Le processus récursif de définition d'un arbre sous ce modèle est générationnel. L'unique nœud de génération 0 est la racine de l'arbre. Récursivement, les nœuds de la génération $n+1$ sont obtenus à partir de ceux de la génération précédente : à chaque nœud de la génération n, on ajoute un nombre aléatoire d'enfants distribué selon μ. Dans ce contexte, la génération d'un nœud v est aussi sa distance à la racine, qu'on appelle en général profondeur et qu'on notera dans la suite $\mathfrak{d}(v)$.

Il y a beaucoup à dire sur ce modèle, introduit indépendamment par Bienaymé (15) et Galton et Watson (97) dans la seconde moitié du XIX^{e} siècle pour étudier la disparition des noms de famille. On se contentera ici de caractériser les grands comportements émergents.

Si $\mu(0) = 0$, les arbres issus de ce modèle sont tous infinis. A contrario, si $\mu(0) > 0$, ils peuvent n'être composés que d'un seul nœud. On distingue en fait les trois régimes suivants selon la valeur moyenne $\overline{\mu}$ de la loi de naissance :

- Régime sous-critique $\overline{\mu} < 1$: le nombre moyen de nœuds d'un arbre de Galton-Watson est fini.

- Régime critique $\overline{\mu} = 1$ (et $\mu(0) > 0$) : le nombre de nœuds est fini presque sûrement mais infini en moyenne.
- Régime sur-critique $\overline{\mu} > 1$: le nombre de nœuds est infini avec probabilité strictement positive.

4.1.2 Aparté : processus de codage

Tout arbre ordonné T peut être décrit de manière univoque par une suite finie d'entiers, qu'on appelle processus de codage. Un exemple important de processus de codage est le processus de contour $\mathfrak{C}$ qui renvoie la suite des profondeurs des nœuds lorsque l'arbre est visité en le contournant par la gauche comme sur la figure 9. Le nombre de profondeurs reportées selon ce parcours est $2\#T - 1$; le processus de contour est en fait l'interpolation linéaire de cette suite d'entiers sur l'intervalle $[0, 2(\#T - 1)]$.

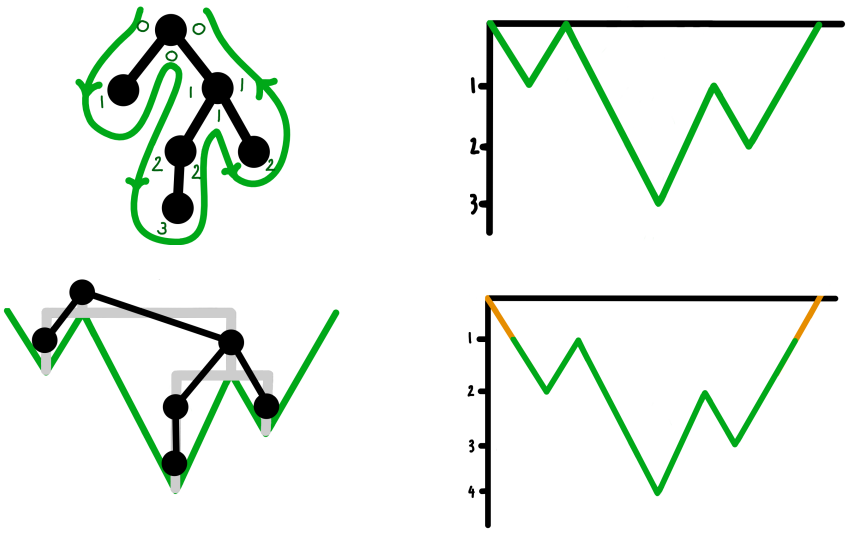


Figure 9: Arbre (en haut à gauche) contourné par la gauche : les profondeurs des nœuds sont reportées pour constituer le processus de contour $\mathfrak{C}$ (en haut à droite), à partir duquel l'arbre peut être retrouvé (en bas à gauche). Le chemin de Harris $\mathfrak{H}$ (en bas à droite) est défini comme le processus de contour augmenté de 1 et rattaché à droite et à gauche en 0.

Il existe d'autres processus de codage, comme la marche des hauteurs, la marche de Łukasiewicz ou le chemin de Harris qui va beaucoup nous intéresser par la suite. Le processus de contour a le mauvais goût de revenir en 0 à chaque fois qu'on termine l'exploration d'une branche émanant de la racine, qu'il s'agisse ou non de la dernière. Le chemin de Harris $\mathfrak{H}$ corrige cette observation : pour tout entier $1 \leq n \leq 2\#T - 1$, $\mathfrak{H}(n) = \mathfrak{C}(n-1) + 1$ et $\mathfrak{H}(0) = \mathfrak{H}(2\#T) = 0$. Comme le processus de contour, le chemin de Harris est l'interpolation linéaire de cette suite, cette fois sur $[0, 2\#T]$.

Grâce à ce décalage, et contrairement au processus de contour, le chemin de Harris peut être vu comme une excursion d'une marche aléatoire, i.e. une marche qui retourne à son point de départ. Ce point est clé dans le conditionnement par la taille des modèles de Galton-Watson.

4.1.3 Conditionnement par la taille

Le modèle de Galton-Watson conditionné par la taille est la loi sur les arbres ordonnés décrite ci-dessus mais regardée conditionnellement au nombre de nœuds de l'arbre généré. Bien sûr, le conditionnement a un effet global sur le processus génératif et est non-trivial. (38) présente une technique de simulation de cet objet probabiliste qui évite la méthode de rejet directe (simuler jusqu'à l'obtention d'un arbre à n nœuds) et donne une idée de l'effet du conditionnement sur sa loi.

De manière plutôt étonnante, plusieurs classes d'arbres aléatoires peuvent être vues comme des modèles de Galton-Watson conditionnés par la taille (38, 57). C'est par exemple le cas de la loi uniforme sur les arbres ordonnés à n nœuds, obtenue à partir de la loi de naissance géométrique de paramètre $1/2$.

Autre fait remarquable, la loi d'un arbre de Galton-Watson de loi de naissance μ conditionné par la taille est la même que pour n'importe quelle loi de naissance μ' de la forme $\mu'(i) \propto \theta^i \times \mu(i)$, $\theta > 0$ (tant que la constante de normalisation est finie) (84, 6.3 Brownian asymptotics for conditioned Galton-Watson trees). Cela signifie que pour les arbres de Galton-Watson conditionnés, la moyenne n'est pas identifiable et on peut supposer sans perte de généralité qu'on est dans le régime critique $\overline{\mu} = 1$.

On note désormais $\mathfrak{H}_n$ la marche de Harris d'un arbre de Galton-Watson conditionné à avoir n nœuds. Comme on l'a précédemment remarqué, elle est définie sur l'intervalle $[0, 2n]$. On peut alors considérer sa version normalisée $t \in [0,1] \mapsto \mathfrak{H}_n(2nt)$. Le modèle de Galton-Watson conditionné par la taille voit sa loi asymptotique caractérisée de la façon suivante (5, Theorem 23),

$$\left(\frac{\mathfrak{H}_n(2nt)}{\sqrt{n}}\right)_{t\in[0,1]} \xrightarrow{\mathcal{L}} \left(\frac{2}{\sigma}\mathbb{e}_t\right)_{t\in[0,1]}, \tag{17}$$

où $\mathbb{e}$ désigne l'excursion brownienne (un mouvement brownien conditionné à ne revenir en 0 qu'à l'instant 1) et σ l'écart-type de la loi de naissance μ : à la limite, seule la variance de la loi de naissance subsiste.

La question statistique qui émane de ce résultat est la suivante : la loi de naissance d'un arbre de Galton-Watson conditionné par la taille n'est pas pleinement identifiable, mais peut-on estimer au moins sa variance, à partir d'une ou plusieurs observations, et obtenir des garanties théoriques sur cet estimateur ? Cette question est traitée dans (P13) dont les résultats sont décrits en section 4.2.

4.1.4 Conditionnement par la hauteur

On regarde maintenant un tout autre type de conditionnement des arbres de Galton-Watson : le conditionnement par la hauteur. On considère la loi des arbres de Galton-Watson induite par une loi de naissance sous-critique $\overline{\mu} < 1$ lorsqu'ils sont conditionnés à être de hauteur h ou plus, i.e. à survivre jusqu'à la génération h si on poursuit l'utilisation d'un vocabulaire issu de la généalogie. On sait qu'un arbre sous-critique est fini presque sûrement ; lorsque h est grand, le conditionnement va donc à l'encontre de ce phénomène. À la limite en loi, c'est un arbre infini qui apparait, appelé arbre de Kesten (2, 62).

L'arbre de Kesten est un modèle de Galton-Watson à deux types : les nœuds normaux dont la loi de naissance est μ et les nœuds spéciaux dont la loi de naissance $\mathcal{B}\mu$ est

biaisée, $\mathcal{B}\mu(i) \propto i \times \mu(i)$. Le processus générationnel est similaire aux arbres de Galton-Watson classiques à ceci près qu'il faut distinguer les nœuds spéciaux : la racine de l'arbre est spéciale et chaque nœud spécial donne naissance à exactement un nœud spécial, tiré au sort parmi les enfants ; les autres sont normaux. Ceci fait sens puisque le nombre d'enfants d'un nœud spécial est au moins 1 (en effet, $\mathcal{B}\mu(0) = 0$).

Un arbre de Kesten est donc constitué d'une unique lignée de nœuds spéciaux, infinie, et qui constitue une épine dorsale à laquelle sont attachés des arbres de Galton-Watson sous-critiques (de loi de naissance μ).

Lorsqu'un arbre de Galton-Watson modélise un phénomène générationnel, typiquement une généalogie, plusieurs schémas d'observation sont envisageables : l'un d'entre eux consiste à supposer qu'on observe le processus jusqu'à la génération h. D'un point de vue statistique, on s'attend alors à estimer de mieux en mieux la loi de naissance lorsque h grandit. Si la lignée qu'on observe n'est pas éteinte à la génération h, cela peut être dû au hasard, ou à un biais dans la sélection des données, en particulier dans le cas d'une population sous-critique. Dans ce cas, l'observation n'est pas distribuée selon un modèle de Galton-Watson mais suit asymptotiquement la loi de Kesten. Mais comment décider si on ne dispose que d'un seul arbre ? C'est ce problème qui est abordé dans (P25) et qui sera traité dans la section 4.3.

4.2 Conditionnement par la taille : estimation de la variance

Dans cette partie reprise de (P13) on cherche à estimer le paramètre de variance d'un modèle de Galton-Watson conditionné par la taille.

4.2.1 Généalogie

Lorsque j'ai découvert la convergence, donnée par l'équation (17), des arbres de Galton-Watson conditionnés vers l'excursion brownienne (l'environnement probabiliste nancéien n'y est pas pour rien), je m'interrogeais sur les différentes techniques disponibles pour analyser des données non-euclidiennes comme les arbres (cf. la sous-section 3.1.3 du chapitre précédent sur les arbres non-ordonnés). Les processus de codage fournissent un point de vue fonctionnel sur les arbres ordonnés et permettent donc de les analyser comme des courbes : l'approche semblait prometteuse.

Outre le fait que (92) utilisait déjà ces concepts pour l'analyse de données biomédicales, ce n'est en fait pas si simple. Premièrement, le support des processus de codage dépend de la taille des arbres. Si on souhaite comparer des données de différentes tailles, on peut vouloir appliquer une renormalisation selon l'axe des abscisses. Mais ensuite, selon quelle norme ? Pour se rendre compte du problème auquel on fait face, on peut s'intéresser à l'exemple suivant : on cherche à comparer deux arbres ordonnés T et S où S est obtenu à partir de T en ajoutant un unique nœud sous sa racine à gauche. Si T est suffisamment gros, T et S peuvent être considérés comme très similaires. Pourtant la différence des processus de contour subit fortement l'existence du premier enfant de la racine de S. On comprend alors qu'on voudrait plutôt recaler les contours pour mettre en évidence la similarité de T et S. Mais le phénomène mis en exergue sous la racine peut exister à toutes les échelles : on voudrait alors s'autoriser à scinder les contours afin de les recaler par morceaux. On est en train de redécouvrir la distance d'édition pour les arbres ordonnés (16). Toutes mes tentatives autour de ces questions sont restées infructueuses.

Malgré tout subsiste la question statistique déjà posée de l'estimation du seul paramètre identifiable des arbres de Galton-Watson conditionnés par la taille, viz. la variance de la loi de naissance σ^2. C'est la question qu'on traite dans (P13), mais je vois surtout cet article comme un modeste apport à la statistique des arbres via leur processus de codage.

Au moment de soumettre pour la première fois (P13) dans un journal, on découvre l'article très récent (13) qui traite du même problème (estimation de la variance de la loi de naissance d'arbres de Galton-Watson conditionnés par la taille), avec les mêmes outils (leur convergence vers l'excursion brownienne). D'un point de vue théorique, (13) suppose directement l'observation de données issues du modèle asymptotique (ce qui ne fait pas sens en pratique), et donc n'étudie pas la convergence des estimateurs lorsque la taille des arbres tend vers l'infini, alors que c'est la question principale de (P13). Mais au delà de ce point important, la publication (13) rate un certain nombre de phénomènes qu'elle nous a pourtant permis de pointer lorsqu'on a réécrit notre article (on y reviendra dans la sous-section 4.2.4).

4.2.2 Moindres carrés

On observe un arbre de Galton-Watson conditionné à avoir n nœuds et on cherche à estimer l'inverse de l'écart-type de sa loi de naissance σ^{-1} en se basant sur la convergence en loi, décrite par l'équation (17), de sa marche de Harris $\mathfrak{H}_n$ vers l'excursion brownienne. Pour cela, on utilise une approche de moindres carrés.

Cette convergence en loi a également lieu en moyenne (40, Theorem 1),

$$\mathbb{E}\left[\frac{\mathfrak{H}_n(2nt)}{\sqrt{n}}\right] \to \frac{2}{\sigma}E_t,$$

où $E_t = \mathbb{E}[e_t] = 4(2\pi)^{-1/2}\sqrt{t(1-t)}$. Une quantité qui semble alors pertinente pour approcher σ^{-1} est

$$\begin{aligned}\widehat{\lambda}_n &= \operatorname*{arg\,min}_{\lambda>0}\left\|\frac{\mathfrak{H}_n(2n\cdot)}{\sqrt{n}} - 2\lambda E\right\|_2^2 \\ &= \frac{\langle \mathfrak{H}(2n\cdot), E\rangle}{2\sqrt{n}\|E\|_2^2}.\end{aligned}$$

Malheureusement, d'une convergence faible on ne peut rien espérer d'autre qu'une convergence faible,

$$\widehat{\lambda}_n \xrightarrow{\mathcal{L}} \sigma^{-1}\Lambda_\infty,$$

où la variable aléatoire $\Lambda_\infty = \frac{\langle e, E\rangle}{\|E\|_2^2}$ est caractérisée dans le résultat qui suit.

Théorème 16 *(P13, Proposition 3.4) Λ_∞ est à densité par rapport à la mesure de Lebesgue avec*

$$\mathbb{E}[\Lambda_\infty] = 1 \qquad \textit{et} \qquad \mathbb{V}[\Lambda_\infty] = \frac{1}{\|E\|_2^4}\int_0^1\int_0^1 g(t,s)E_tE_s\,dt\,ds - 1,$$

où, lorsque $0 \le t \le s \le 1$,

$$g(t,s) = \frac{2}{\pi}\left[3\sqrt{t(s-t)(1-s)} + (2t(1-s)+s(1-t))\arcsin\left(\frac{t(1-s)}{s(1-t)}\right)\right],$$

et $g(t,s) = g(s,t)$ pour $s < t$.

$\widehat{\lambda}_n$ n'est donc pas un estimateur de σ^{-1} (une erreur aléatoire intrinsèque persiste même à la limite), mais si on le considérait comme tel, il serait asymptotiquement sans biais. Si on veut construire un véritable estimateur de σ^{-1} à partir de la convergence (17), il nous faut observer une forêt. Mais ne nous y trompons pas : la partie difficile de ce théorème est en fait l'absolue continuité de Λ_∞ par rapport à la mesure de Lebesgue.

4.2.3 Estimation à partir d'une forêt

On suppose maintenant qu'on observe une forêt d'arbres de Galton-Watson conditionnés indépendants et de même variance de la loi de naissance. On note N la taille de la forêt et $\mathbf{n} = (n_1, \dots, n_N)$ le vecteur des tailles des N arbres. Si on réitère l'approche des moindres carrés en concaténant les contours des N arbres de la forêt, on obtient simplement la moyenne empirique des $\widehat{\lambda}^i_{n_i}$ définis pour chacun des arbres de la forêt,

$$\overline{\lambda}_{\mathbf{n}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\lambda}^i_{n_i}.$$

Au lieu de mesurer l'adéquation des N marches de Harris au contour moyen théorique, on peut tâcher d'exploiter notre connaissance de leur distribution autour de celui-ci. En effet, on sait qu'à la limite, les $\widehat{\lambda}^i_{n_i}$ sont distribués comme $\sigma^{-1}\Lambda_\infty$. On peut donc estimer σ^{-1} par

$$\overline{\lambda}^W_{\mathbf{n}} = \operatorname*{arg\,min}_{\lambda>0} d_W \left(\frac{1}{N} \sum_{i=1}^{N} \delta_{\widehat{\lambda}^i_{n_i}} , \lambda\Lambda_\infty \right),$$

où d_W désigne la distance de Wasserstein. Un peu de calcul montre qu'on construit ainsi une moyenne pondérée des $\widehat{\lambda}^i_{n_i}$,

$$\overline{\lambda}^W_{\mathbf{n}} = \frac{1}{\|F^{-1}_{\Lambda_\infty}\|_2^2} \sum_{i=1}^{N} \widehat{\lambda}^i_{n_i} \int_{(i-1)/N}^{i/N} F^{-1}_{\Lambda_\infty}(s)\,ds,$$

où $F^{-1}_{\Lambda_\infty}$ désigne la fonction de répartition inverse de Λ_∞. On tient ainsi compte de l'asymétrie de la loi limite ; (P13, Fig. 3.4) en donne une représentation graphique.

Outre des résultats sur le biais (cf. (P13, Proposition 4.1) pour $\overline{\lambda}_{\mathbf{n}}$ et (P13, Proposition 4.5) pour $\overline{\lambda}^W_{\mathbf{n}}$), on montre le résultat de consistance suivant valide pour les deux estimateurs qu'on vient de construire.

Proposition 17 *(P13, Propositions 4.2 & 4.8) Pour tout $\epsilon > 0$, il existe n tel que, pour tout $\mathbf{n}$ vérifiant $\min(\mathbf{n}) \geq n$,*

$$\mathbb{P}\left(\limsup_{N\to+\infty} \left| \overline{\lambda}^*_{\mathbf{n}} - \sigma^{-1} \right| < \epsilon \right) = 1,$$

où $* \in \{\emptyset, W\}$.

4.2.4 Comparaison avec (13)

L'approche développée dans (13) exploite la convergence en loi vers l'excursion brownienne des arbres de Galton-Watson conditionnés d'une manière différente. Si on tire un nœud au hasard dans un tel arbre, alors, comme corollaire de l'équation (17), sa profondeur normalisée par $\sqrt{n}$, notée δ_n, converge en loi vers une variable aléatoire

$\sigma^{-1}\Delta_\infty$, où Δ_∞ suit la loi de Rayleigh de paramètre d'échelle 1 (13, Proposition 4). La méthode d'estimation présentée dans (13) est basée sur l'utilisation de δ_n, qui joue en quelque sorte un rôle équivalent à celui de $\widehat{\lambda}_n$ ci-dessus.

Le premier inconvénient à cette approche est qu'elle génère un biais : $\mathbb{E}[\Delta_\infty] = (\pi/2)^{1/2}$. On peut le corriger aisément en considérant à la place $\widehat{\delta}_n = (2/\pi)^{1/2}\delta_n$, mais on fait alors face à un second défaut, du moins quand on compare $\widehat{\delta}_n$ à $\widehat{\lambda}_n$: la variance asymptotique. En effet, en ne tenant compte que d'un seul nœud au lieu de tout le contour pour estimer σ^{-1}, on observe une variance intrinsèque plus grande,

$$\mathbb{V}(\Lambda_\infty) \simeq 0.0690785 \qquad \text{vs} \qquad \mathbb{V}\left(\sqrt{\frac{2}{\pi}}\Delta_\infty\right) \simeq 0.2732395.$$

À la fois biaisé et de variance asymptotique plus grande, l'estimateur construit dans (13) est donc moins performant que le nôtre. Mais les différences entre les deux approches ne s'arrêtent pas là.

On a constaté dans les simulations numériques que, du moins pour les lois de naissance binaires, la convergence (17) se fait par en dessous. Pour des arbres finis, les estimateurs de σ^{-1} construits sur la convergence vers l'excursion brownienne, qu'il s'agisse de $\widehat{\lambda}_n$ ou $\widehat{\delta}_n$, présentent donc un biais d'autant plus grand que les arbres sont petits. Dans (P13, 5 Simulation study), nous estimons cette erreur par des simulations numériques afin de débiaiser les estimateurs.

Mais le véritable point important, à la fois d'un point de vue théorique et pratique, à côté duquel passe (13) est sans aucun doute celui qui suit.

4.2.5 Et la variance empirique ?

En exploitant la convergence faible (17), on ne peut espérer estimer la variance de la loi de naissance qu'à partir de l'observation d'une forêt. Mais en se focalisant sur l'analyse fonctionnelle des processus de codage, ne passe-t-on pas à côté d'un estimateur construit sur un unique arbre ? S'il y a bien un candidat naturel, c'est la variance empirique des nombres d'enfants. On ne l'a pas considérée jusqu'ici pour deux raisons :

- La moyenne empirique M_n des nombres d'enfants d'un arbre de Galton-Watson conditionné T_n vérifie
$$M_n = \frac{1}{n}\sum_{v\in T_n} \#\mathcal{C}(v) = 1 - \frac{1}{n}.$$
Ce résultat est purement déterministe et reste donc vrai quel que soit le modèle sous-jacent. Si la moyenne empirique n'estime pas la moyenne de la loi de naissance (en présence ou non d'un conditionnement), que peut-on espérer de la variance ?
- On s'attend à ce que le conditionnement par la taille bouleverse les réalisations de la loi de naissance dans tout l'arbre.

Et pourtant... On note V_n la variance empirique des nombres d'enfants,

$$V_n = \frac{1}{n}\sum_{v\in T_n} (\#\mathcal{C}(v) - M_n)^2.$$

Théorème 18 *(P13, Theorem 2.6) Quand n tend vers l'infini,*

$$\mathbb{E}\left[\left|V_n - \sigma^2\right|\right] \to 0.$$

Ce résultat, dont la preuve technique utilise la représentation développée dans (38) pour la simulation ainsi que (100), montre qu'on peut estimer le seul paramètre identifiable de la loi de naissance à partir d'une unique observation d'un arbre de Galton-Watson conditionné par la taille, et enlève donc beaucoup de l'intérêt de (13) ainsi que de notre étude des marches de Harris.

On montre néanmoins dans (P13, 5.3 Missing or noisy data) comment l'approche asymptotique peut être appliquée lorsqu'on dispose de données manquantes ou bruitées qui ne permettent pas de calculer la variance empirique. Mais au delà du modèle particulier des arbres de Galton-Watson conditionnés, je pense qu'il faut surtout voir ces deux approches basées sur les marches de Harris comme des contributions à la statistique des données arborescentes d'un point de vue fonctionnel.

En plus des méthodes et des résultats déjà présentés, une étude numérique des différents estimateurs considérés et une application à des données réelles complètent le contenu de (P13).

4.3 Conditionnement à survivre : estimation des lois de naissance

On revient maintenant sur les arbres de Galton-Watson conditionnés par la hauteur. Dans (P25), on étudie l'estimation par maximum de vraisemblance des lois de naissance d'arbres à épine dorsale qui généralisent les modèles de Galton-Watson conditionnés ou non à survivre.

4.3.1 Arbres de Galton-Watson à épine dorsale

Si un arbre de Galton-Watson de loi de naissance sous-critique μ n'a pas été conditionné à survivre, tous les nombres d'enfants qu'il contient sont distribués selon μ. Si au contraire il est observé conditionnellement à la survie, il s'agit d'un arbre de Kesten dont une unique branche suit la loi biaisée $\mathcal{B}\mu$, où $\mathcal{B}\mu(i) \propto i \times \mu(i)$, alors que tous les autres nombres d'enfants suivent la loi μ. Afin de distinguer statistiquement ces deux cas, on les voit comme des instances particulières d'un modèle plus général qu'on appelle arbre à épine dorsale.

Un arbre à épine dorsale est un modèle de Galton-Watson à deux types : les nœuds normaux dont les nombres d'enfants suivent la loi de naissance μ et les nœuds spéciaux dont les nombres d'enfants sont distribués selon ν où $\nu(i) \propto f(i) \times \mu(i)$. La racine est spéciale et chaque nœud spécial donne naissance à un unique enfant spécial pris au hasard parmi les enfants (comme pour l'arbre de Kesten).

Si on sait estimer le biais f d'un tel modèle observé pendant un grand nombre de générations, on peut décider de la présence ou non du conditionnement à la survie : si $f \equiv 1$, on retrouve un arbre de Galton-Watson usuel, alors que si $f(i) = i$, il s'agit d'un arbre de Kesten.

Dans (P25), on s'intéresse à l'estimation des paramètres μ et f de ce modèle, à partir d'une unique observation de l'arbre sur un grand nombre de générations. La difficulté principale vient du fait que les types des nœuds (normaux ou spéciaux) ne sont évidemment pas observés.

4.3.2 Vraisemblance

Dans toute la suite, on suppose $f(0) = 0$, i.e. les arbres qu'on considère sont constitués d'une épine dorsale infinie notée $\mathcal{S}$ (pour l'anglais spine). On fait également l'hypothèse que le support de μ est fini, $\mu(i) = 0$ pour tout $i \geq N+1$. Enfin, il est important de remarquer que le paramètre f n'est pas identifiable puisque tout autre biais αf induit la même loi de naissance spéciale ν : on suppose sans perte de généralité que f vérifie

$$\sum_{i=0}^{N} f(i) \times \mu(i) = 1.$$

On observe un arbre T issu de ce modèle jusqu'à la génération h. Précisément, les nœuds v observés, c'est-à-dire les nœuds pour lesquels $\#\mathcal{C}(v)$ est connu, sont ceux de profondeur $\mathfrak{d}(v) < h$ (on note leur ensemble T_h). Les sous-arbres de T ne sont donc pas nécessairement observés jusqu'aux feuilles. Dans ce contexte, la hauteur observée d'un sous-arbre $T[v]$ est définie comme

$$\mathcal{H}_o(T[v]) = \min(\mathcal{H}(T[v]), \ell - \mathfrak{d}(v)),$$

où ℓ est la longueur du chemin descendant de v aux nœuds non-observés : $\ell = +\infty$ si v n'a pas de descendants non-observés et $\ell = h - \mathfrak{d}(v)$ sinon.

Pour estimer μ et f, on propose de procéder par maximum de vraisemblance. En reproduisant le raisonnement mené dans (14) pour les processus de Galton-Watson, on peut voir que la vraisemblance du modèle s'écrit

$$\mathcal{L}_h(\mu, f) = \sum_{\mathfrak{d}(v)<h} \log\, \mu(\#\mathcal{C}(v)) + \sum_{v\in\mathcal{S},\mathfrak{d}(v)<h} \log\, f(\#\mathcal{C}(v)) - h \log \sum_{l=0}^{N} f(l)\mu(l).$$

Si $\mathcal{L}_h$ fait intervenir l'ensemble de nœuds $\mathcal{S}$ supposé inconnu, son maximum en μ lui n'en dépend pas, ce qui nous fournit un premier estimateur de la loi de naissance,

$$\begin{aligned} \widehat{\mu}_h &= \underset{\mu \text{ t.q. } \sum \mu = 1}{\arg\max}\ \mathcal{L}_h(\mu, f) \\ &= \left(\frac{1}{\#T_h} \sum_{\mathfrak{d}(v)<h} \mathbb{1}_{\{\#\mathcal{C}(v)=i\}} \right)_{i\in\{0,\dots,N\}}. \end{aligned} \tag{18}$$

Le maximum en f dépend de l'épine dorsale $\mathcal{S}$ et n'est donc d'aucune utilité à ce stade.

4.3.3 Le vilain petit canard

Une partie de l'épine dorsale $\mathcal{S}$ de l'arbre, ainsi que certains nœuds normaux, peuvent être identifiés de manière purement algorithmique. C'est l'objet de la proposition qui suit.

Proposition 19 *(P25, Proposition 1) Soit* T *un arbre à épine dorsale observé jusqu'à la génération* h *et* v *un nœud observé de* T *de profondeur* $\mathfrak{d}(v)$*.*

- *Si* $\mathcal{H}_o(T[v]) + \mathfrak{d}(v) < h$*, alors* v *est normal.*
- *Si* v *est spécial, si ses enfants sont observés, et s'il a un unique enfant* c *tel que* $\mathcal{H}_o(T[c]) + \mathfrak{d}(c) \geq h$*, alors* c *est spécial.*

Figure 10: Arbre à épine dorsale (en vert alors que les nœuds normaux sont en orange) observé jusqu'à la ligne horizontale grise. En appliquant la proposition 19, une portion de l'épine dorsale peut être identifiée (surlignée en vert). Les nœuds dont la descendance s'éteint dans l'intervalle d'observation sont identifiés comme normaux (surlignés en orange). Ni identifiées comme normales ou spéciales, les lignées candidates $\mathfrak{S}_h$ sont surlignées en bleu.

Ce résultat est illustré par l'exemple de la figure 10. En l'appliquant (récursivement depuis la racine pour le second point) à l'arbre observé, on classe ses nœuds en trois catégories : les nœuds identifiés comme normaux, ceux identifiés comme spéciaux, et ceux pour lesquels on n'a pu conclure et qui forment les épines dorsales candidates (à la suite de la portion spéciale détectée) dont on note l'ensemble $\mathfrak{S}_h$.

Au moment de choisir laquelle des lignées candidates est la vraie épine dorsale de l'arbre, on peut être tenté de maximiser la vraisemblance $\mathcal{L}_h$ en $\mathcal{S}$. On fait alors le constat suivant : l'optimum en $\mathcal{S}$ dépend de f et réciproquement. Cette approche mène à un algorithme itératif d'optimisation composante par composante. On préfère ici un algorithme en une seule étape et pour lequel on parviendra à montrer un résultat de convergence.

La clé de l'algorithme d'estimation de $\mathcal{S}$ est la suivante. Toutes les lignées candidates qu'on observe ont été conditionnées à survivre jusqu'à la génération h (leur loi de naissance est donc biaisée). Toutes sauf une : la véritable épine dorsale de l'arbre ! On la détecte donc comme le vilain petit canard de l'ensemble $\mathfrak{S}_h$ (où la loi de naissance normale conditionnée à la survie $\mathcal{B}\mu$ est inconnue et donc estimée par $\mathcal{B}\widehat{\mu}_h$),

$$\widehat{\mathcal{S}}_h = \operatorname*{arg\,max}_{s \in \mathfrak{S}_h} d_{KL}(\tilde{s}, \mathcal{B}\widehat{\mu}_h), \tag{19}$$

où $\tilde{s}$ est la loi empirique des nombres d'enfants le long de s et d_{KL} désigne la divergence de Kullback-Leibler. Bien sûr, on évite un algorithme itératif car le maximum (18) de $\mathcal{L}_h$ en μ ne dépend ni de f, ni de $\mathcal{S}$.

4.3.4 Correction de $\widehat{\mu}_h$ et estimation de f

Maintenant qu'on a estimé l'épine dorsale de l'arbre, on peut corriger l'estimateur $\widehat{\mu}_h$ en ôtant les nombres d'enfants de $\widehat{\mathcal{S}}_h$ de la moyenne empirique,

$$\widehat{\mu}^\star_h(i) = \frac{1}{\#T_h - h} \sum_{v \in T_h \setminus \widehat{\mathcal{S}}_h} \mathbb{1}_{\{\#c(v)=i\}}. \tag{20}$$

Et on peut estimer f en maximisant $\mathcal{L}_h$ où $\mathcal{S}$ a été remplacée par son estimateur $\widehat{\mathcal{S}}_h$,

$$\widehat{f}_h(i) = \frac{1}{\widehat{\mu}_h^\star(i)h} \sum_{v \in \widehat{\mathcal{S}}_h} \mathbb{1}_{\{\#\mathcal{C}(v)=i\}}.$$

4.3.5 Convergence des estimateurs

Nous allons énoncer un unique résultat de convergence mais qui regroupe deux cas de figure en fait très différents. En effet, nous nous sommes jusqu'ici placés dans un cadre où la loi de naissance normale μ est sous-critique. On montre que dans ce cas l'ensemble $\mathfrak{S}_h$ est essentiellement réduit à la vraie épine dorsale $\mathcal{S}$ plus de petites perturbations, qui ne sont pas d'ampleur suffisante pour fausser l'estimation de μ et f (P25, 4 Proof of Theorem 4 in the subcritical case). L'algorithme du vilain petit canard (19) est de peu d'utilité ici : on peut même sélectionner $\widehat{\mathcal{S}}_h$ au hasard parmi $\mathfrak{S}_h$!

Le cas véritablement intéressant apparait lorsque μ est critique ou sur-critique : $\mathfrak{S}_h$ est alors de taille plus conséquente et peut contenir des grandes déviations de la loi de naissance normale telles que

$$d_{KL}(\tilde{s}, \mathcal{B}\mu) \gg d_{KL}(\tilde{\mathcal{S}}, \mathcal{B}\mu).$$

En pareille situation, on ne peut malheureusement pas distinguer la vraie épine dorsale de l'arbre. On peut en fait espérer estimer $\mathcal{S}$ lorsque la loi spéciale ν et la loi biaisée $\mathcal{B}\mu$ sont suffisamment différentes devant la vitesse de croissance de la population normale gouvernée par $\overline{\mu}$:

- Si la croissance de la population est raisonnable, $\mathfrak{S}_h$ n'est pas trop gros et on observe peu de déviations à $\mathcal{B}\mu$. $\mathcal{S}$ peut être estimée même si ν et $\mathcal{B}\mu$ sont proches (sans être égales bien sûr).
- Si la population croit très vite $\overline{\mu} \gg 1$, on ne peut pas estimer $\mathcal{S}$ à moins que ν et $\mathcal{B}\mu$ ne soient très différentes.

D'un point de vue théorique, la similarité entre ν et $\mathcal{B}\mu$ est mesurée par une divergence ad hoc $\mathfrak{D}$ (P25, eq. (12)).

Théorème 20 *(P25, Theorem 4) Si $\mathfrak{D}(\nu, \mathcal{B}\mu) > \log \overline{\mu}$, alors*

$$\widehat{\mu}_h^\star \xrightarrow{\text{p.s.}} \mu, \qquad \widehat{f}_h \xrightarrow{\text{p.s.}} f \qquad \textit{et} \qquad \frac{\#\widehat{\mathcal{S}}_h \cap \mathcal{S}}{h} \xrightarrow{\text{p.s.}} 1.$$

La démonstration dans les cas critique et sur-critique de ce résultat étudie les grandes déviations de la population $\mathfrak{S}_h$ (P25, 5 On the rate function of large deviations in sample selection). Une courte étude de simulations finalise cet article (P25, 7 Numerical illustration).

4.4 Conclusion et perspectives

Dans ce chapitre on s'est intéressé à des problèmes d'inférence statistique de deux modèles d'arbres aléatoires finalement très différents même si tous deux sont obtenus via le conditionnement de processus de Galton-Watson. Dans les deux cas, on a construit des estimateurs des paramètres du modèle et montré leur convergence.

Lorsqu'on conditionne un arbre de Galton-Watson à avoir une taille fixée à l'avance, l'unique quantité à estimer est la variance de la loi de naissance. On a prouvé que celle-ci peut être estimée à partir de l'observation d'un unique arbre par sa version empirique, alors même que la moyenne empirique est en général un piètre estimateur de la moyenne de la loi de naissance. D'un point de vue théorique, il reste à traiter la question de la normalité asymptotique de cet estimateur, notamment dans le but de construire un test d'égalité de la variance, qui pourrait permettre de comparer les lois de deux arbres indépendants.

Les bonnes propriétés de la variance empirique enlèvent de l'intérêt à l'approche basée sur la convergence de la marche de Harris vers l'excursion brownienne. On a exhibé dans (P13, 5.3 Missing or noisy data) des exemples jouets de bruits qui biaisent fortement le calcul de de la variance empirique alors que la forme du contour est conservée, mais les applications réelles font toujours défaut. Malgré cela, je crois toujours pertinentes les approches fonctionnelles pour les données arborescentes : (92) en est un très bon exemple dans un cadre appliqué. Mais leur étude théorique passera sans doute par d'autres modèles que les processus de Galton-Watson conditionnés.

Dans (P13), on a aussi montré que l'approche des moindres carrés (faisant intervenir à la limite la variable aléatoire $\sigma^{-1}\Lambda_\infty$) est de variance plus faible que lorsqu'on regarde la marche de Harris en un nœud pris au hasard (qui converge vers $\sigma^{-1}\Delta_\infty$) (13). Dans ce contexte, on peut se demander quelle est la variable aléatoire sans biais de variance asymptotique minimale qu'on peut construire à partir de l'observation de la marche de Harris. À propos du biais observé pour des arbres finis causé par la convergence par en dessous vers l'excursion brownienne, il serait intéressant d'avoir des résultats théoriques afin de corriger les estimateurs de manière plus rigoureuse, mais la littérature semble avare sur ce point.

Enfin, rien ne distingue d'un point de vue théorique les estimateurs $\overline{\lambda}_{\mathbf{n}}$ et $\overline{\lambda}^{W}_{\mathbf{n}}$ construits à partir de l'observation d'une forêt : tous deux vérifient la proposition 17. On observe toutefois dans les simulations qu'ils ne sont pas sujets aux données aberrantes de la même façon (P13, 5.3.1 Estimation with outliers) : par exemple, l'estimateur minimisant la distance de Wasserstein est moins sensible que son homologue des moindres carrés aux données aberrantes plus petites que la cible σ^{-1}, ce qu'on peut expliquer par les faibles poids qui leur sont accordés (de la forme $\int_{(i-1)/N}^{i/N} F^{-1}_{\Lambda_\infty}(s)ds$ avec i proche de 1). On pourrait profiter de cette observation pour sélectionner l'estimateur en fonction de la distribution des données.

Dans le cas des arbres de Galton-Watson à épine dorsale (P25), on a estimé la loi de naissance μ et le biais f alors que les types des nœuds ne sont pas observés. On a prouvé que dans le cas sous-critique, la détection algorithmique de l'épine dorsale est suffisante pour construire de bons estimateurs. Dans le cas plus difficile (et finalement plus intéressant) où la loi de naissance normale est critique ou sur-critique, on peut détecter l'épine dorsale lorsque la loi spéciale ν et la loi biaisée $\mathcal{B}\mu$ sont suffisamment différentes devant le taux de croissance de la population normale $\overline{\mu}$.

Après avoir estimé l'épine dorsale $\mathcal{S}$, on a corrigé l'estimateur $\widehat{\mu}_h$ en ôtant de la moyenne empirique les nœuds identifiés comme spéciaux : il s'agit de l'estimateur $\widehat{\mu}^\star_h$ donné par l'équation (20). Ceci n'a d'intérêt que dans le cas sous-critique où le nombre de nœuds normaux n'est pas grand devant h ; ce phénomène est d'ailleurs observé dans les simulations (P25, 7 Numerical illustration). Mais alors, on pourrait proposer de réestimer l'épine dorsale $\mathcal{S}$, ce qui permettrait à la fois de corriger l'estimateur de f et construire un nouvel estimateur de μ... et ainsi de suite. L'algorithme itératif qu'on obtient ainsi

n'a pas d'intérêt du point de vue de la convergence établie dans le théorème 20 (car ce qui compte dans le cas sous-critique est que les lignées candidates sont à peu de choses près réduites à $\mathcal{S}$), mais pourrait améliorer sensiblement les résultats numériques, en particulier pour des observations de profondeur faible.

Lorsqu'on cherche à appliquer un tel résultat de convergence, il est important de pouvoir attester de sa validité sur les données observées. Dans notre cas, cela revient à identifier le signe de $\log \overline{\mu} - \mathfrak{D}(\nu, \mathcal{B}\mu)$, ce qui ne peut être fait sans connaitre les paramètres μ et f. Vérifier à partir des données le critère de convergence est une question ouverte importante pour la mise en place de tests statistiques. Rappelons que la question qui a motivé ce travail est de savoir si un arbre observé sur un grand nombre de générations a été conditionné à survivre ou est issu d'un autre modèle. Les premiers outils théoriques nécessaires au développement de ce test sont maintenant à notre disposition.

Les arbres à épine dorsale ne sont en définitive qu'une instance particulière de processus de Galton-Watson multitypes, pour lesquels les problèmes d'inférence ont été traités de multiples manières dans la littérature, mais la plupart du temps en supposant l'observation des types (23, 63, 73). Seulement quelques travaux (49, 85, 94) ne font pas cette hypothèse et aucun ne traite des aspects théoriques. L'étude des arbres à épine dorsale réalisée dans (P25) souligne les difficultés inhérentes à l'estimation des modèles à types cachés et peut être vue comme un premier pas dans cette direction.

Bibliographie exogène

(1) O. O. AALEN : Nonparametric inference for a family of counting processes. *Ann. Statist.*, 6(4):701–726, 1978.

(2) R. ABRAHAM et J.-F. DELMAS : Local limits of conditioned Galton-Watson trees: the infinite spine case. *Electronic Journal of Probability*, 19:1–19, 2014.

(3) A. V. AHO, J. E. HOPCROFT et J. D. ULLMAN : *The Design and Analysis of Computer Algorithms*. Addison-Wesley series in computer science and information processing. Addison-Wesley Publishing Company, 1974.

(4) F. AIOLLI, G. DA SAN MARTINO, A. SPERDUTI et A. MOSCHITTI : Fast on-line kernel learning for trees. *In Sixth International Conference on Data Mining*, pages 787–791. IEEE, 2006.

(5) D. ALDOUS : The Continuum Random Tree III. *The Annals of Probability*, 21(1):248–289, 1993.

(6) P. K. ANDERSEN, O. BORGAN, R. D. GILL et N. KEIDING : *Statistical models based on counting processes*. Springer Series in Statistics. Springer New York, 1996.

(7) N. ARONSZAJN : Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

(8) T. ASAI, H. ARIMURA, T. UNO et S.-I. NAKANO : Discovering frequent substructures in large unordered trees. *In International Conference on Discovery Science*, pages 47–61. Springer, 2003.

(9) D. AVIS et K. FUKUDA : Reverse search for enumeration. *Discrete Applied Mathematics*, 65(1-3):21–46, 1996.

(10) R. BABILON, J. MATOUŠEK, J. MAXOVÁ et P. VALTR : Low-distortion embeddings of trees. *In Graph Drawing*, pages 343–351. Springer Berlin Heidelberg, 2002.

(11) M.-F. BALCAN, A. BLUM et N. SREBRO : A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008.

(12) P. BERTAIL, S. CLÉMENÇON et J. TRESSOU : A storage model with random release rate for modeling exposure to food contaminants. *Mathematical Biosciences and Engineering*, 5(1):35–60, 2008.

(13) K. BHARATH, P. KAMBADUR, D. K. DEY, A. RAO et V. BALADANDAYUTHAPANI : Statistical tests for large tree-structured data. *Journal of the American Statistical Association*, 112(520):1733–1743, 2017.

(14) B. R. BHAT et S. R. ADKE : Maximum likelihood estimation for branching processes with immigration. *Advances in Applied Probability*, 13(3):498–509, 1981.

(15) I.-J. BIENAYMÉ : De la loi de multiplication et de la durée des familles. *Soc. Philomath. Paris*, pages 37–39, 1845.

(16) P. BILLE : A survey on tree edit distance and related problems. *Theoretical Computer Science*, 337:217–239, 2005.

(17) K. BOROVKOV et G. LAST : On Rice's formula for stationary multivariate piecewise smooth processes. *Journal of Applied Probability*, 49(2):351–363, 2012.

(18) S. BÖTTCHER, R. HARTEL et J. RABE : Efficient XML keyword search based on DAG-compression. *In International Conference on Database and Expert Systems Applications*, pages 122–137, 2013.

(19) M. BOUSQUET-MÉLOU, M. LOHREY, S. MANETH et E. NOETH : XML compression via directed acyclic graphs. *Theory of Computing Systems*, 57(4):1322–1371, 2015.

(20) L. BREIMAN, J. FRIEDMAN, C. J. STONE et R. A. OLSHEN : *Classification and Regression Trees*. Taylor & Francis, 1984.

(21) K. BRINGMANN, Y. LI et R. C. RHOADES : Asymptotics for the number of row-Fishburn matrices. *European Journal of Combinatorics*, 41:183–196, 2014.

(22) E. BUCKWAR et M. RIEDLER : An exact stochastic hybrid model of excitable membranes including spatio-temporal evolution. *Journal of mathematical biology*, 63:1051–93, 12 2011.

(23) M. L. CARVALHO : A joint estimator for the eigenvalues of the reproduction mean matrix of a multitype Galton-Watson process. *Linear algebra and its applications*, 264:189–203, 1997.

(24) J. CHIQUET, M. EID et N. LIMNIOS : Modelling and estimating the reliability of stochastic dynamical systems with markovian switching. *Reliability Engineering & System Safety*, 93(12):1801–1808, 2008. 17th European Safety and Reliability Conference.

(25) J. CHIQUET, N. LIMNIOS et M. EID : Piecewise deterministic Markov processes applied to fatigue crack growth modelling. *Journal of Statistical Planning and Inference*, 139(5):1657–1667, 2009. Special Issue on Degradation, Damage, Fatigue and Accelerated Life Models in Reliability Testing.

(26) A. CLEYNEN et B. DE SAPORTA : Change-point detection for piecewise deterministic Markov processes. *Automatica*, 97:234–247, 2018.

(27) B. CLOEZ, R. DESSALLES, A. GENADOT, F. MALRIEU, A. MARGUET et R. YVINEC : Probabilistic and piecewise deterministic models in biology. *ESAIM: Procs*, 60:225–245, 2017.

(28) M. COLLINS et N. DUFFY : Convolution kernels for natural language. *In Advances in neural information processing systems*, pages 625–632, 2002.

(29) O. L. V. COSTA et F. DUFOUR : Stability and ergodicity of piecewise deterministic Markov processes. *SIAM Journal on Control and Optimization*, 47(2):1053–1077, 2008.

(30) N. CRISTIANINI et J. SHAWE-TAYLOR : *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

(31) G. DA SAN MARTINO, N. NAVARIN et A. SPERDUTI : Tree-based kernel for graphs with continuous attributes. *IEEE Transactions on Neural Networks and Learning Systems*, 2015.

(32) D. DACUNHA-CASTELLE et M. DUFLO : *Probability and Statistics*, volume 2. Springer New York, 2012.

(33) F. DALMAO et E. MORDECKI : Rice formula for processes with jumps and applications. *Extremes*, 18:15–35, 2015.

(34) M. H. A. DAVIS : Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B*, 46(3): 353–388, 1984.

(35) M. H. A. DAVIS : *Markov models and optimization*, volume 49 de *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1993.

(36) B. de SAPORTA, F. DUFOUR, H. ZHANG et C. ELEGBEDE : Optimal stopping for the predictive maintenance of a structure subject to corrosion. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 226(2):169–181, 2012.

(37) F. DENIS, A. HABRARD, R. GILLERON, M. TOMMASI et É. GILBERT : On probability distributions for trees: representations, inference and learning. *In NIPS Workshop on Representations and Inference on Probability Distributions*, Whistler, Canada, 2007.

(38) L. DEVROYE : Simulating size-constrained Galton–Watson trees. *SIAM Journal on Computing*, 41(1):1–11, 2012.

(39) M. DOUMIC, M. HOFFMANN, N. KRELL et L. ROBERT : Statistical estimation of a growth-fragmentation model observed on a genealogical tree. *Bernoulli*, 21(3):1760–1799, 2015.

(40) M. DRMOTA et J.-F. MARCKERT : Reinforced weak convergence of stochastic processes. *Statistics & Probability Letters*, 71(3):283–294, 2005.

(41) M. DUFLO : *Random iterative models*. Applications of Mathematics. Springer-Verlag, Berlin, 1997.

(42) R. FERNANDEZ, P. DAS, V. MIRABET, E. MOSCARDI, J. TRAAS, J.-L. VERDEIL, G. MALANDAIN et C. GODIN : Imaging plant growth in 4-d: robust tissue reconstruction and lineaging at cell resolution. *Nature Methods*, 7(7):547–553, 2010.

(43) A. FINKE, A. JOHANSEN et D. SPANÒ : Static-parameter estimation in piecewise deterministic processes using particle Gibbs samplers. *Annals of the Institute of Statistical Mathematics*, 66(3):577–609, June 2014.

(44) T. FUJII : Nonparametric estimation for a class of piecewise-deterministic Markov processes. *Journal of Applied Probability*, 50(4):931–942, 2013.

(45) S. GARTHE, P. SCHWEMMER, V. H. PAIVA, A.-M. CORMAN, H. O. FOCK, C. C. VOIGT et S. ADLER : Terrestrial and marine foraging strategies of an opportunistic seabird species breeding in the Wadden sea. *PLOS ONE*, 11(8):1–19, 08 2016.

(46) A. GENADOT et M. THIEULLEN : Averaging for a fully coupled piecewise-deterministic Markov process in infinite dimensions. *Advances in Applied Probability*, 44(3):749–773, 2012.

(47) C. GODIN : Representing and encoding plant architecture: A review. *Annals of Forest Science*, 57(5):413–438, 2000.

(48) C. GODIN et P. FERRARO : Quantifying the degree of self-nestedness of trees: application to the structural analysis of plants. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(4):688–703, 2010.

(49) M. GONZÁLEZ, J. MARTÍN, R. MARTÍNEZ et M. MOTA : Non-parametric bayesian estimation for multitype branching processes through simulation-based methods. *Computational statistics & data analysis*, 52(3):1281–1291, 2008.

(50) S. GRAPH et H. LUSCHGY : *Foundations of quantization for random vectors*, volume 1730 de *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.

(51) L. GUIGNARD, U.-M. FIUZA, B. LEGGIO, J. LAUSSU, E. FAURE, G. MICHELIN, K. BIASUZ, L. HUFNAGEL, G. MALANDAIN, C. GODIN et P. LEMAIRE : Contact area-dependent cell communication and the morphological invariance of ascidian embryogenesis. *Science*, 369(6500):158, juillet 2020.

(52) T. HASTIE, R. TIBSHIRANI et J.H. FRIEDMAN : *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009.

(53) D. HAUSSLER : Convolution kernels on discrete structures. Rapport technique, Department of Computer Science, University of California, 1999.

(54) O. HERNÁNDEZ-LERMA, S. O. ESPARZA et B. S. DURAN : Recursive nonparametric estimation of nonstationary Markov processes. *Boletín de la Sociedad Matemática Mexicana. Segunda Serie*, 33(2):57–69, 1988.

(55) P. HODARA, N. KRELL et E. LÖCHERBACH : Non-parametric estimation of the spiking rate in systems of interacting neurons. *Statistical Inference for Stochastic Processes*, 21, 04 2018.

(56) M. JACOBSEN : *Point process theory and applications: marked point and piecewise deterministic processes*. Probability and Its Applications. Birkhäuser Boston, 2006.

(57) S. JANSON : Simply generated trees, conditioned Galton–Watson trees, random allocations and condensation. *Probability Surveys*, 9:103–252, 2012.

(58) V. JELÍNEK : Counting general and self-dual interval orders. *Journal of Combinatorial Theory, Series A*, 119(3):599–614, 2012.

(59) F. JOUSSE, R. GILLERON, I. TELLIER et M. TOMMASI : Conditional random fields for XML trees. *In Workshop on Mining and Learning in Graphs*, Berlin, Germany, 2006.

(60) A. B. KAHN : Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562, 1962.

(61) E. L. KAPLAN et P. MEIER : Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

(62) H. KESTEN : Subdiffusive behavior of random walk on a random cluster. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 22(4):425–487, 1986.

(63) R. H. KHAĮRULLIN : On estimating parameters of a multitype Galton-Watson process by ϕ-branching processes. *Siberian Mathematical Journal*, 33(4):703–713, 1992.

(64) D. KIMURA, T. KUBOYAMA, T. SHIBUYA et H. KASHIMA : A subpath kernel for rooted unordered trees. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 62–74. Springer, 2011.

(65) R. M. KOVACEVIC et G. Ch. PFLUG : Does insurance help to escape the poverty trap?—A ruin theoretic approach. *The Journal of Risk and Insurance*, 78(4):1003–1027, 2011.

(66) N. KRELL : Statistical estimation of jump rates for a specific class of Piecewise Deterministic Markov Processes. *ESAIM: Probability and Statistics*, 20:196–216, 2016.

(67) N. KRELL et E. SCHMISSER : Nonparametric estimation of jump rates for a specific class of piecewise deterministic Markov processes. *Bernoulli*, 27(4):2362–2388, 2021.

(68) P. KRITZER, G. LEOBACHER, M. SZÖLGYENYI et S. THONHAUSER : Approximation methods for piecewise deterministic Markov processes and their costs. *Scandinavian Actuarial Journal*, 2019(4):308–335, 2019.

(69) D. LAMBERT et DIAGRAM GROUP : *The Field Guide to Geology*. Facts on File, 1998.

(70) G. LAMPLE et F. CHARTON : Deep learning for symbolic mathematics. *International Conference on Learning Representations*, 2020.

(71) F. LAVANCIER et R. LE GUÉVEL : Spatial birth-death-move processes : basic properties and estimation of their intensity functions. *Journal of the Royal Statistical Society: Series B*, 2021.

(72) F. LAVANCIER et P. ROCHET : A general procedure to combine estimators. *Computational Statistics & Data Analysis*, 94:175–192, 2016.

(73) F. MAAOUIA et A. TOUATI : Identification of multitype branching processes. *The Annals of Statistics*, 33(6):2655–2694, 2005.

(74) J. MAIRAL, P. KONIUSZ, Z. HARCHAOUI et C. SCHMID : Convolutional kernel networks. *In Advances in Neural Information Processing Systems*, volume 27, 2014.

(75) G. Da San MARTINO : *Kernel Methods for Tree Structured Data*. Thèse de doctorat, University of Bologna, Italy, 2009.

(76) D. MEAGHER : Geometric modeling using octree-encoding. *Computer Graphics and Image Processing*, 19:129–147, 06 1982.

(77) T. MIKOLOV, K. CHEN, G. CORRADO et J. DEAN : Efficient estimation of word representations in vector space. *International Conference on Learning Representations*, 2013.

(78) T. M. MITCHELL : *Machine Learning*. McGraw-Hill, Inc., USA, 1ère édition, 1997.

(79) Ca. MUSCO et Ch. MUSCO : Recursive sampling for the Nyström method. *In Advances in Neural Information Processing Systems*, volume 30, 2017.

(80) S.-I. NAKANO : Efficient generation of plane trees. *Information Processing Letters*, 84(3):167–172, 2002.

(81) S.-I. NAKANO et T. UNO : Efficient generation of rooted trees. *National Institute for Informatics (Japan), Tech. Rep. NII-2003-005E*, 8, 2003.

(82) S. NOWOZIN : *Learning with structured data: applications to computer vision.* Thèse de doctorat, Berlin Institute of Technology, 2009.

(83) E. PARZEN : On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

(84) J. PITMAN : *Combinatorial stochastic processes*, volume 1875 de *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006.

(85) J. QI, J. WANG et K. SUN : Efficient estimation of component interactions for cascading failure analysis by EM algorithm. *IEEE Transactions on Power Systems*, 33(3): 3153–3161, 2017.

(86) H. RAMLAU-HANSEN : Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, 11(2):453–466, 1983.

(87) K. RIECK, T. KRUEGER, U. BREFELD et K.-R. MÜLLER : Approximate tree kernels. *Journal of Machine Learning Research*, 11(16):555–580, 2010.

(88) Ph. RIGOLLET et A. B. TSYBAKOV : Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16:260–280, 2017.

(89) M. ROSENBLATT : A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences*, 42(1):43–47, 1956.

(90) A. L. SAMUEL : Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):535–554, 1959.

(91) B. SCHÖLKOPF, K. TSUDA et J.-P. VERT : *Kernel Methods in Computational Biology*. The MIT Press, 07 2004.

(92) D. SHEN, H. SHEN, S. BHAMIDI, Y. MUNOZ MALDONADO, Y. KIM et J. S. MARRON : Functional data analysis of tree data objects. *Journal of Computational and Graphical Statistics*, 23, 04 2014.

(93) K. SHIN, T. ISHIKAWA, Y.-L. LIU et D. L. SHEPARD : Learning DOM trees of web pages by subpath kernel and detecting fake e-commerce sites. *Machine Learning and Knowledge Extraction*, 3(1):95–122, 2021.

(94) A. STANEVA et V. STOIMENOVA : EM algorithm for statistical estimation of two-type branching processes – a focus on the multinomial offspring distribution. *AIP Conference Proceedings*, 2302(1):030003, 2020.

(95) G. VALIENTE : *Algorithms on Trees and Graphs*. Springer-Verlag, Berlin, Heidelberg, 2002.

(96) S. V. N. VISHWANATHAN et A. J SMOLA : Fast kernels on strings and trees. *Advances on Neural Information Proccessing Systems*, 14, 2002.

(97) H. W. WATSON et F. GALTON : On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 4:138–144, 1875.

(98) C. WILLIAMS et M. SEEGER : Using the Nyström method to speed up kernel machines. *In Advances in Neural Information Processing Systems*, volume 13, 2000.

(99) C. T. WOLVERTON et T. J. WAGNER : Recursive estimates of probability densities. *IEEE Transactions on Systems Science and Cybernetics*, 5(3):246–247, 1969.

(100) S. L. ZABELL : A limit theorem for expectations conditional on a sum. *Journal of Theoretical Probability*, 6:267–283, 1993.