Modèles et algorithmes pour l'apprentissage statistique des processus ponctuels avec interactions. Application : analyse et caractérisation de données cosmologiques

Nathan Gillot

UNIVERSITÉ DE LORRAINE
34 COURS LÉOPOLD — CS 25233
54052 NANCY CEDEX
TÉL : 03 54 50 54 00
CONTACT@UNIV-LORRAINE.FR
WWW.UNIV-LORRAINE.FR

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Modèles et algorithmes pour l'apprentissage statistique des processus ponctuels avec interactions. Application : analyse et caractérisation de données cosmologiques

THÈSE

présentée et soutenue publiquement le 15 décembre 2025

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention mathématiques appliquées)

par

Nathan Gillot

Composition du jury

Présidente :	Mariane Clausel	Professeure, Université de Lorraine
Rapporteurs :	Jean-François Coeurjolly	Professeur, Université Grenoble-Alpes
	Marie-Colette van Lieshout	Full Professor, CWI Amsterdam
Examinateurs :	Ed Cohen	Associate Professor, Imperial College
	Jenny Sorce	CNRS, Université de Lille
Invités :	Aila Särkkä	Full professor, Chalmers University
	Didier Gemmerlé	CNRS, Université de Lorraine
Direction de thèse :	Radu Stoica	Professeur, Université de Lorraine

Institut Élie Cartan de Lorraine - UMR 7502
Centre Inria de l'Université de Lorraine

Mis en page avec la classe thesul.

Résumé

Les relevés de décalage vers le rouge des galaxies (Redshift Survey) à grande échelle montrent que la répartition de la matière dans l'Univers n'est pas uniforme. La majorité des galaxies s'agglutinent en diverses structures telles que des murs, filaments ou clusters de galaxies et font émerger de larges zones de vide cosmique. La pluralité géométrique et topologique de la toile cosmique ainsi que la taille des jeux de données suggèrent une approche par les probabilités et les statistiques. Ce travail, motivé par la richesse de ce type de données, propose d'explorer la modélisation de la distribution des galaxies à l'aide des processus ponctuels de Gibbs. Nous supposons alors qu'une configuration de galaxies, réduite à une collection de points, est une réalisation d'un processus aléatoire régi par une densité de probabilité, permettant de générer des interactions locales entre individus. Deux types de modélisations sont alors proposées : l'une se base sur la superposition de modèles conditionnellement à des structures galactiques afin d'améliorer la connaissance sur l'agencement des galaxies autour de ces structures. L'autre s'appuie sur deux nouveaux modèles à interactions «multi-échelles» pour permettre de modéliser plus finement un jeu de donnée volumineux. La contribution clé de cette thèse traite de l'inférence en données incomplète émergeant des relevés de galaxies où certaines galaxies peuvent être obstruées par d'autres en raison de leurs positions ou de leurs trop faibles luminosités. Ce problème des données manquantes est toujours d'actualité et des solutions sont développées depuis les années 70 avec les algorithmes de type Espérance-Maximisation (EM). Mes travaux proposent de reformuler ce problème dans le cadre des statistiques Bayésiennes. À partir d'un algorithme récent construit pour l'échantillonnage *a posteriori* en données complètes (ABC Shadow), la thèse propose de nouveaux algorithmes pour échantillonner la loi jointe du paramètre et de la configuration manquante, et des algorithmes afin d'effectuer l'inférence en données incomplètes.

Mots-clés: Processus ponctuel de Gibbs, Inférence Bayésienne, Inférence en données incomplète, Modélisation, Processus ponctuels, Simulation, Cosmologie

Abstract

Large-scale redshift surveys of galaxies show that the distribution of matter in the Universe is not uniform. The majority of galaxies clump together in various structures such as walls, filaments or galaxy clusters, giving rise to large areas of cosmic emptiness. The geometric and topological plurality of the cosmic web, as well as the size of the data sets, suggest an approach based on probabilities and statistics. This work, motivated by the richness of this type of data, proposes to explore the modelling of galaxy distribution using Gibbs point processes. We assume that a configuration of galaxies, reduced to a collection of points, is a realisation of a random process governed by a probability density, allowing to generate interactions between individuals. Two types of models are proposed : one is based on the superposition of models conditionally to galactic structures in order to improve our understanding of the arrangement of galaxies around these structures. The other is based on new «multi-scale» interaction models to enable more detailed modelling of large data sets. The main contribution of this thesis deals with inference in incomplete data emerging from galaxy surveys where some galaxies may

be obscured by others due to their positions or low luminosities. This problem is still studied nowadays and solutions to the problem of missing data have been developed since the 1970s with Expectation-Maximisation (EM) algorithms. This work proposes to reformulate this problem within the framework of Bayesian statistics. Following the construction of a recent algorithm used for posterior sampling (ABC Shadow), the thesis proposes new algorithms to sample the joint distribution of the parameter and the missing configuration and algorithms to achieve inference with partly observed data.

Keywords: Gibbs point processes, Bayesian inference, Missing data inference, Modelling, Point processes, Simulation, Cosmology

Remerciements

Il y a bien ~~longtemps~~ récemment, dans une galaxie ~~lointaine~~ proche, ~~très lointaine~~ très proche... Plus précisément dans notre Voie lactée, sur Terre, à l'Université de Lorraine, France. ~~C'est une époque de guerre civile. À bord de vaisseaux spatiaux opérant à partir d'une base cachée, les Rebelles ont emporté leur première victoire sur le maléfique Empire Galactique.~~ Pardonnez-moi, je m'égare... Comme disent les plus grands, le plus important n'est pas de franchir la porte des étoiles, c'est d'en revenir ! Toutes ces petites références pour introduire les remerciements d'une thèse de mathématiques appliquées à la cosmologie. Bref, nous y voilà. Vous qui lisez ces lignes, si je vous ai oublié dans ces remerciements, je vous prie de m'excuser, ça arrive quand on est entouré de pépites au quotidien !

Je tiens à remercier chaleureusement Radu Stoica, pour avoir encadré cette thèse, pour tes conseils, relectures et discussions diverses et variées pendant ces 3 ans. J'ai appris et je continue à apprendre bien des choses à tes côtés. Pour ta confiance sans faille en mes capacités et ton soutien dans les moments de doutes. Hâte de continuer notre running gag du «vamos a vajar las notas» !

Un grand merci également à Didier Gemmerlé («Oh non ! Pas lui !») pour ton aide très précieuse dans mes combats contre le `C++` et mes e^{14562} erreurs de segmentation.. Pour ton optimisme du quotidien et toutes ces discussions dépassant largement le cadre des mathématiques.

Merci à Irène Marcovici et Julien Flamant pour avoir accepté d'être dans mon comité de suivi, vos retours et suggestions m'ont aidé à appréhender ces années plus sereinement.

Many many thanks also to Aila Särkkä, for your warm welcome in the cold city of Gothenburg. My time there was such a great experience during this PhD. Thank you for your advices, discussions and wisdom regarding our past and ongoing collaboration.

I would like to thank the members of my jury, Marie-Colette van Lieshout, Jean-François Coeurjolly, Jenny Sorce, Ed Cohen and Marianne Clausel. More specificly to the reporters for their comments and their careful reading.

L'IECL est avant tout un laboratoire où il fait bon vivre : bien qu'endurci par des températures scandinaves en hiver et tropicales en été, les rencontres que j'ai pu y faire ont toujours été chaleureuses et rafraîchissantes (suivant la saison).

Merci au personnel d'appui à la recherche sans qui ce laboratoire ne tournerait pas. Merci Cécile et Valérie M. (encore désolé pour les 45 oublis de retours de livres à la bibliothèque). Merci Jérémie, Paola, Valérie G., Virginie, Laurence et Nathalie pour votre compassion à l'égard de mes multiples bobos issus d'aventures sportives. Plus précisément, merci Valérie G. pour le plaisir de rentrer dans ton bureau en disant «j'espère que je dérange», Virginie pour ton enthousiasme par rapport à mes tenues vestimentaires, Paola pour tes petites blagues placées à la perfection, Laurence pour avoir toujours été attentive au bien-être de tes doctorants et Nathalie pour ta patience infinie pour mon usage maladroit de ~~notre cher et bien aimé~~ Notilus. Enfin, merci à Élodie, bureau de la gentillesse et de l'écoute attentive, sans qui mes années d'étude et de thèse ne se seraient pas déroulées de la même manière.

Merci à tous mes professeurs/collègues qui ont participé à l'enrichissement de mes connaissances scientifiques ou culturelles, en cours ou lors d'une pause café. Merci à Régine, Anne G.P, Pascal et Bruno pour m'avoir encouragé et conforté dans mon projet de poursuite en thèse. Merci Damien pour tes conseils avisés et ton investissement dans la médiation scientifique à travers le Club Math et les 200 autres projets que tu portes. Merci également à l'équipe Inria PASTA pour cette belle dynamique d'équipe, ces moments conviviaux, séminaires au vert et exposés. Merci

Madalina pour les sorties d'autoroute ratées à cause de discussions trop passionnantes, Pascal Moyal pour les cours vraiment géniaux en master, Sara Mazzonetto pour sa joie de vivre, Antoine Lejay pour ses anecdotes aussi surprenantes que rigolotes.

Qui dit Inria, dit également midInria. J'ai eu et je continue d'avoir le plaisir quotidien de venir me sustenter à ce restaurant dont la sympathie de l'équipe n'a d'égal que leur cuisine. Merci à Loïc de toujours nous accueillir avec tant de gentillesse malgré les oublis de personnes invitées de l'IECL dans mes mails. Merci Isabelle d'être l'échelle des personnes solaires et pour tes splendides desserts, Floriane pour tes «mon ptit loup» (je suis certain d'être le seul à qui tu dis ça !) et Caro tes «mon lapin» qui me ramènent en enfance.

Merci enfin à F.B pour son aide au dénombrement des galaxies dans l'Univers et à mes anciens professeurs de collège, lycée et prépa : Françoise Bertrand, pour l'entrain à la diffusion des mathématiques, la pomme que je suis n'est pas tombée bien loin de l'arbre. Maud Facqueur pour la profonde rigueur et le soutien pendant mes années prépa, je ne peux m'empêcher de rigoler à l'idée d'avoir fait une thèse à l'interface entre les proba et les stats, au vu de mes «lacunes» et de mon attrait pour ces dernières en terminale. Notre regretté Jacques Champagne, pour sa joie d'enseigner et son rire si communicatif, cela m'a grandement aidé à confirmer mon engouement pour les mathématiques. François Schnepf et Laurent Pietri, pour leur disponibilité et accompagnement tout au long de cette deuxième année de prépa.

Toutes ces belles années n'auraient pas non plus été les mêmes sans les nombreuses personnes qui ont accompagné mon parcours d'étude ~~quasi-linéaire~~ un peu chaotique. Je pense pouvoir dire que l'aspect fondamental de cette thèse était davantage porté sur mon cercle d'amitié et de rencontres que sur les maths (normal pour une thèse en mathématiques appliquées me direz-vous.. surtout quand on pense que les probas c'est du 50/50).

Merci à mes anciens camarades de prépa et à la team Langres/Nancy pour ces soirées, festivals, vacances et moments mémorables : Fol, Mimile, Vanane, Flo, Ghislain, Yohann, Rodrigue, Camille, Manon, Damax, Manue, Monique, Matt, Fab, Eugé, Chachou, Nico, Dondon, Jo et Bajam. Je garde un souvenir très cher de tous ces moments passés avec vous, je peux enfin vous répondre à vos «Mais Nathan, tu finis quand tes études ?» annuels.. Non, je ne reprends pas une $n-$ième formation en septembre prochain, j'ai fini ! (Bon maintenant, il va falloir répondre à «Mais Nathan, tu trouves quand un CDI ?», mais cette question sera traitée ultérieurement).

Merci aux loustics de la prépa agreg qui ont choisi la voie royale de l'enseignement dans le secondaire, pour ces moments de partages et de soutien pendant ces années covid : Geoff, JP, Etienne, Mathilde, Coralie et Florent. J'espère que l'Ellipsoïde de John Loewner ravit vos élèves et qu'ils comprennent que ça peut se recaser dans 9 leçons (en poussant «un peu»).

Merci au groupe du vendredi (quelque peu étendu depuis !) : Hugo pour tes speedrun any % du «tu vas dormir sur le paillasson !», Marie pour ton adversité dans ces parties de Canardage et ton talent particulier face à des pizzas surgelées, Samuel pour avoir été un camarade d'ego lift (hâte que tu reviennes) et t'être occupé de mes plantes, Jeanne pour toutes ces infos sur notre système de santé (ça ressemble à un reproche dit comme ça...), Nono pour tes multiples propositions de cuisine, Vincent et Valentin pour ces aprems et soirées à jouer aux cartes (n'oubliez pas, je ne suis pas la menace !). Riri pour ton enthousiasme quotidien (surtout vis-à-vis des parcs d'attractions) et tes propositions de sorties qui animent notre petit groupe (d'ailleurs, tu sais d'où vient le bois utilisé pour les jardins éphémères ?). Enfin merci à Pith pour ta présence si apaisante, j'espère que ton voyage dans les contrées scandinaves se passe bien malgré la quasi-perte des bagages.

Même si nous sommes souvent seuls face à notre sujet de thèse, je me suis rarement senti autant compris, écouté et accompagné par les camarades thésard.e.s que j'ai eu la chance de

rencontrer pendant ces 3 ans.

Il n'y a pas que la cantine qui rend la vie agréable au Loria : merci Hee-Soo, Vincent, Gabriel, Amandine L, Amandine D, Bertrand et Marie pour les aprems/soirées jeux et sports en tous genres. Merci également de participer à l'organisation du tournoi biannuel de tennis de table qui permet à l'IECL de venir rafler des médailles avec un nombre de représentants limités, il va sans dire que notre gentille rivalité reste dans les super moments de ma thèse.

Le Loria n'est pas le seul labo regorgeant de personnes géniales. Merci Thomas, Popoli, Jocelyn, Raphaël M, Benjamin, Jérémy, Nathan T (bonjour Monsieur Nathan !), Marie, Alexis, Jérémy, Magalie, Paul L, Paul B, Brieuc, Gautier, Hugo, Kilian, Louis, Fatma, Raphou, Saïd, Sophie, Bastien P, Valentin R, Vincent, Amine, Simon, Valentin C, Pierre M et Abdelkader. Plus précisément, merci Séréna pour tes imitations de Johnny, Aurélien F pour le simple fait d'exister et ces séances de course effrénées, Clara pour les ananas Victoria, Clémence pour tes tentatives de traquenard au crosstraining, Amélia pour tes goûts en anime, Mathilde pour cette chouette découverte de l'aviron (Sisyphe a convergé, aussi impossible que cela puisse paraître !), Valou pour ta vie sentimentale digne des feux de l'amour, Axel pour ton feu d'artifices chez les Budzinski, Bastien L pour ton animation du club Math et tes super exposés, Eric (de son vrai nom Pierre) pour sa présence aux manifs, David pour les parties de ping matinales, Elie pour tes «Bonjour» et tes tutos pour faire des high-kick, Virgile pour le mouvement Brownie avant le séminaire, Aurélien M pour tes notifications de présence le midi et Amine pour l'entretien des multiples micro-conflits entre Elie, Victor et toi. Many thanks also to Freja (I know you're not short !), Vidhi (I still didn't manage to translate your message in Hindi :( ), Yingtong (for your fair-play during the board games) and Juan (I know this isn't personal, you just have la dalle). Ahhhhhh et j'ai failli oublier notre Benixou national pour sa ponctualité sans égale.

Un grand «gracias» à Lili B pour les surnoms rigolos et ta connaissance fine des ordres de grandeurs. Merci à James, tu es la motivation numéro une pour se lever le dimanche et aller à la Salle (t'as vu P-drift, j'ai mis un S majuscule !).

Sweden was also a great home during my stays in Gothenburg and Smögen. Thank you to the great people I've met there : Julia, Mattias, Ruben, Konstantinos, Dave, Louise, Shekoufeh, Basil, Arnaud, Mathis, Adélie, Moritz, Marija and Ottmar.

Pierrick Zoust, P-brique, Véronique ou que sais-je l'alias... premier brise-raison choupinou et chevalier des arts et des lettres Gueminoises. Je sais que t'as rien understood au previous paragraphe because the B2 is difficile, je know. C'est why je try de put un peu de franglais dans ces few lines pour t'aider à improve ton toi-même en english. Thanks à toi d'être such un bon coach (celui là il est transparent, ça devrait aller, sauf si tu as un schtoss), coloc and ami. T'as always été de good conseils (even quand tu throw Lugia en soulink) et je miss beaucoup your pc doing la fusée ! Plus sérieusement, tu as été un pilier pendant ces années et un modèle de dépassement de soi (et ça continue !), hâte de te rejoindre en tant qu'ESAS à Villetaneuse !

Comment mentionner Patrick sans ses fidèles acolytes de basket... faiiites aussi du bruit pour Rodolphe (le Kipchoge libanais) et Mabrouk (aka Ben Jaba Durant), siuu ! Merci pour vos mails chargés d'interjections exprimant votre joie de me voir soutenir ma thèse. Je suis vraiment très heureux d'avoir pu vous passer le flambeau du séminaire des doctorants et de voir à quel point vous avez su améliorer son fonctionnement tout en gardant son essence conviviale.

Un autre occupant du 210 est évidemment à remercier : Dr Glue pour ces belles années à animer la gentille ~~toxicité~~ convivialité du bureau et de ces années d'études partagées. Merci pour ton aide pendant toutes ces années, je pense que je n'aurais pas réussi aussi bien sans toi. Merci pour ta créativité sans bornes pour pranker notre copain Sieste quand il oublie ses clés de bureau et merci d'être un Wikipédia sur pattes, ta culture ne cessera jamais de m'étonner.

Je ne peux m'empêcher d'enchaîner sur son plus bel antagoniste politique, j'ai nommé Anthony. Tu es un modèle d'abnégation et persévérance (et de répliques clairement devenues cultes dans mon répertoire). Pour n'en citer qu'une : «Là, j'avoue que j'ai abusé» après avoir cadenassé les vélos de ta résidence à 2h du matin (le contexte rendant cet acte légitime est évidemment à oublier). Merci pour tes conseils et écoutes attentives lorsque j'en avais besoin. J'espère que tu nous reviendras de Marseille avec un bel aque-cent du sud !

Merci à Victoriane pour votre joie de vivre, je suis très heureux d'avoir contribué à votre épanouissement dans la vie ! Je pense que cela constitue une de mes plus belles réussites de cette année passée (bon, quand même, peut-être après cette thèse) et que je ne pouvais pas espérer meilleur feuilleton en ce début d'année 2025. Surtout tant que notre ami Pétrisse n'y voyait que du feu et que la strat' du cheval apparaissait comme une solution «viable».

Merci Aloïs pour ces nombreuses discussions animées par ta curiosité scientifique et intellectuelle, je ne m'en lasserai probablement jamais, surtout quand tu conclus une discussion sur la thermodynamique en pensant «avoir inventé l'eau froide».

Merci Ludwig pour ces années lycée et prépa passées à tes côtés. Merci d'avoir aidé à travailler mon humilité et à entretenir une démarche critique et raisonnée sur des tas de domaines. C'est toujours un régal de passer une après-midi entière à jongler avec des sujets scientifiques, politiques et métaphysiques. Il est quand même important de mentionner notre plus belle collaboration en TP de chimie : «Ludwig... je crois qu'on a raté l'équivalence !».

Marido, je te remercie d'avoir accepté de lâcher ton poney Divino pour venir assister à ma soutenance. Merci de toujours prendre le temps de venir me voir quand tu es de passage dans le Grand Est et de venir me raconter ta vie absolument rocambolesque, à quand le best-seller ? En tout cas, j'ai hâte d'avoir quarante balais (wink wink).

Christophe, je n'aurais pu rêver mieux comme grand frère de thèse. Merci d'avoir été là pour le début de ma thèse et d'avoir contribué à ma compréhension de mon sujet. Également, dans les moments de doutes, ta présence et ta vision des choses vis-à-vis de la thèse m'ont été d'une aide vraiment précieuse. Sache que je n'oublierai jamais que «tu as sauvé le monde» avec ta thèse.

Victor, deuxième brise-raison choupinou et chevalier des arts et des lettres Gueminoises. Je te remercie d'avoir été un camarade de danse si talentueux (même quand tu fais des demandes un peu tendancieuses...), de m'avoir ouvert les yeux sur la formule déjà mentionnée et d'entrer dans mon bureau de manière incongrue sans donner la moindre explication. Sache qu'après toutes ces années, j'ai pris la décision de te pardonner pour avoir brisé mon jaune d'œuf (je ne suis pas rancunier, loin de là !). Navré pour ma prononciation douteuse du mosellan, il n'y a pas de Duolingo pour ça, mais je pense que cette appli est asmoudich de toute façon. Ah, et de rien !

Il est évident de mentionner ~~la belle~~ l'exceptionnelle triplette des occupantes (fixes ou non) du bureau 107. Merci Amélie pour ton soutien, ta présence dans le bureau et pour tes cookies, fondants et brookies, c'était vraiment «la cerise sur le bateau» lors de ces derniers mois de rédaction. Merci d'être une source de discussions si intéressantes et d'être une aussi bonne camarade de jeux de société (même si on a perdu aux petits chevaux maléfiques), hâte que tu reviennes (on croise les doigts) au Loria.

Merci à Anouk, dont la taille n'est qu'inversement proportionnelle à son talent pour les stats (surtout au tarot), la cuisine et l'escalade ! Merci de m'aider dans la différentiation des aliments de la vie de tous les jours, quelle déception cela aurait été de manger des choux romanesco en pensant prendre des brocolis... Merci pour ton éclaircissement sur mon sujet de thèse, j'ignorais que je faisais une thèse en cohomologie jusqu'à ce que tu fasses la remarque. Blague à part, tu as probablement été la personne la plus bénéfique pour ma dernière année de thèse (Yann, ne soit pas jaloux stp), partager un bureau avec quelqu'un en rédaction de thèse était une excellente idée ! Merci pour ces aprems jeux (tu pourras me laisser gagner à un moment quand même ?) et

ces pauses cafés où on refait le monde.

Coach Julia, Budz, la Ji, aka Roulia. Je pensais avoir trouvé, en la personne de P-bis, la personne la plus accro au sport que je connaisse, mais ça, c'était avant de te rencontrer. Merci pour ces (environ) 547 séances de sport pendant la thèse, j'ai rarement autant apprécié me réveiller à 6h30 du matin pour aller nager, courir, «crossfiter», pédaler ~~ou mourir d'une petite hypo~~ avec toi. Je ne me lasserai jamais de nos discussions si riches et variées sur la vie en général (surtout quand tu proposes un ptit Trampoline Park à 14h, en pleine semaine). Merci à toi et à Rodolphe pour ce petit séjour de décompression en Italie après mon dépôt de manuscrit. Enfin, merci de compléter Anouk pour l'encyclopédie culinaire que vous êtes et de contribuer à la culture de mes papilles si peu développée (même si je sais faire la différence entre des cacahuètes et des noisettes).

Ces derniers paragraphes peinent néanmoins à exprimer avec justesse les sentiments qui m'animent lorsque je rédige ces lignes tant vos amitiés sont précieuses et centrales dans ma vie. C'est encore un peu plus vrai pour mes deux amis les plus proches, Oriane et Yann. Merci Yann d'appeler la bibliothèque un samedi pour savoir si elle est ouverte pour que je puisse aller y travailler parce que j'ai horreur de téléphoner. Merci Oriane pour ces années de master avec toi, je n'aurais pu rêver meilleure binôme ~~de sudoku et de mots fléchés~~ de projets et d'agreg. Merci Yann d'être littéralement mon conseiller juridique, administratif et amoureux (même si tu n'as jamais voulu de moi, ce qui constitue probablement la plus grosse faute de goût de ta vie, après celle d'aimer jouer bleu à Magic). Merci Oriane pour tes petites attentions quasi non dénombrables tant elles sont nombreuses, pour ces balades gourmandes et toutes ces soirées à regarder des flims (ka-chow !). Merci Yann de paraître pédant en citant Montaigne et d'admettre que tu adores le capitalisme ~~lorsque tu joues à Satisfactory~~. Merci Oriane pour tes snipes de qualité, de partager toutes ces belles pépites et de prendre la peine de te lever pendant les cours en visio. Merci à vous deux de bien vouloir continuer à jouer à 7 Wonders Duel avec moi malgré vos 15 défaites d'affilée respectives (on n'a jamais dit que je ne pouvais pas être un peu vache). Et Oriane, de rien, évidemment !

Merci enfin à ma famille, vous m'avez toujours fait confiance et accompagné dans mes projets d'études. Comme quoi, je me suis un peu éloigné de mon projet professionnel initial où j'aspirais à devenir coiffeur (bon ok, j'étais en maternelle). Merci à ma mère Christine et à mon père Philippe d'avoir toujours été des modèles de compréhension, d'intégrité, de tolérance et d'ouverture d'esprit, ces valeurs contribuent à faire de moi le bon être humain que je pense être aujourd'hui. Merci à mon frère Valentin qui fut depuis ma tendre enfance une source d'inspiration et un modèle à suivre, c'est aussi grâce à toi que j'ai choisi de mener ce projet de thèse (il était hors de question que tu sois plus diplômé que moi !). Merci à vous, papi et mamie, de suivre mon avancement dans la vie avec tant de fierté et de m'avoir transmis ce goût très prononcé pour les jeux de société (j'attends encore quelques années pour vous défier au Scrabble !).

J'aimerais conclure ces remerciements avec quelques statistiques certes inutiles, mais rigolotes en lien avec cette thèse.

— Environ 650 heures passées sur la rédaction de ce manuscrit
— Environ 150 heures (estimation basse) de débeugage de code
— 834 cafés furent nécessaires à l'aboutissement de cette thèse
— Environ 413 kilomètres de course, 25 km de nage et 1100 km à vélo furent parcourus pendant les 2 dernières années de cette thèse
— Minimum 600 desserts à la cantine de l'Inria, merci Isabelle ! (en ne comptant pas les jours où j'en prenais deux...)
— Environ 60 après-midi à siester au lieu de travailler
— 75% des pauses midi du vendredi ont dépassé 14h
— 24 fichiers nommés «Manuscrit-version-finale»
— 173 articles/thèses/livres cités dans cette thèse ~~et seulement 15 lus en profondeur~~
— 47 échecs de résolution d'exercices de géométrie du Club Math suivi d'un «Demande à Damien, il saura»
— 53 pauses cafés à 10h30 en étant arrivé à 10h28 (Merci Yann)

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

État de l'art : modélisation, simulation et inférence pour les processus ponctuels avec interactions

1.1 Généralités, présentation du problème

Les processus ponctuels constituent une classe très large de modèles mathématiques permettant de décrire et d'étudier des phénomènes ayant des applications dans de nombreux domaines. Nous considérons ainsi que ce que nous observons est une réalisation d'un phénomène aléatoire régi par ces modèles. Ces derniers ont pour but de permettre la modélisation et la caractérisation de ces phénomènes aléatoires par le biais d'outils probabilistes et statistiques.

Plus spécifiquement, une réalisation de ces modèles constitue une configuration de points aléatoire. Un point peut alors représenter, suivant les applications, une bactérie, la présence d'un animal, un arbre, une occurrence de séisme ou encore une galaxie. Ces objets ne pouvant être uniquement réduits à un seul point en 2D ou 3D, il est important de pouvoir prendre en compte leurs caractéristiques (taille, poids, espèce, magnitude, etc.) afin d'affiner cette modélisation. Une telle caractéristique peut alors être associée à chaque point, de manière déterministe ou aléatoire, en leur attachant une marque. Dans ce cas, nous parlerons alors de processus ponctuels marqués.

Il peut également s'avérer utile de traduire des relations se produisant entre les individus (points) telles que la compétition inter-espèce dans un même milieu naturel ou l'attraction entre deux galaxies suffisamment proches l'une de l'autre. Pour cela, les modèles peuvent dépendre d'une relation de voisinage induite par une distance inter-points. De tels phénomènes peuvent alors être modélisés par des processus ponctuels dits de Markov. Les processus ponctuels servent également à la détection de structures particulières dans des données, telles que des clusters de galaxies ou de maladies en épidémiologie [Stoica, 2025].

Enfin, les processus ponctuels peuvent aussi servir à la modélisation de phénomènes spatio-temporels, où un temps t est associé à la position des points. Chaque point est alors vu comme un évènement apparaissant aléatoirement en temps et en espace (éruptions volcaniques, séismes, évolution de la répartition des arbres dans une forêt). Ces types particuliers de modèles ne sont ici pas abordés, nous recommandons [Daley and Vere-Jones, 2003, Daley and Vere-Jones, 2008] pour une introduction sur le sujet.

La simulation des processus ponctuels est un problème pouvant s'avérer complexe. En effet, à l'exception de certains processus, la fonction de partition du modèle n'est pas connue sous forme analytique, rendant alors les techniques de simulation directes comme l'inversion de la fonction de répartition ou des méthodes ad-hoc impossibles. Le plus souvent, des méthodes se basant sur les chaînes de Markov (méthodes MCMC) sont utilisées. Un travail théorique en amont est nécessaire pour garantir les bonnes propriétés de la chaîne de Markov (CDM) afin de s'assurer que cette dernière converge vers la loi du processus à simuler. La simulation de ces processus peut se faire de manière exacte (méthode de simulation parfaite) comme les algorithmes de couplage par le passé (CFTP), Clan des ancêtres, échantillonneur de Gibbs exact, Metropolis-Hastings exact... Cela peut être également fait en approchant la distribution d'intérêt par convergence de la CDM vers cette dernière (algorithme de Metropolis-Hastings, échantillonneur de Gibbs, algorithme à sauts réversibles). Enfin, ces méthodes peuvent être adaptées en fonction du modèle considéré [Møller et al., 1998, van Lieshout and Stoica, 2006], à noter qu'elles ne sont pas propres à la simulation des processus ponctuels.

Quant à l'estimation paramétrique des modèles, différentes solutions existent :

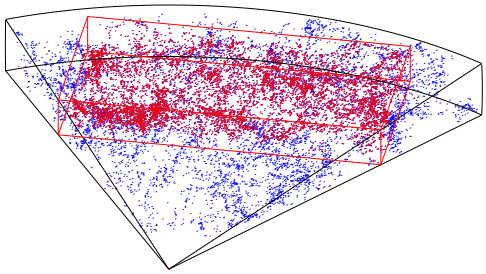
- les approches non paramétriques basées par exemple sur les statistiques descriptives [Ripley, 1976, Ripley, 1977] ou des noyaux [Diggle, 1985],
- les approches paramétriques, basées sur des modèles à densité de probabilité, comme la pseudo-vraisemblance [Besag, 1974] ou la vraisemblance.

Dans cette thèse, nous nous tournons vers les approches paramétriques pour l'inférence des processus ponctuels. Là aussi, la constante de normalisation des modèles est un frein à l'utilisation de méthodes «basiques» comme la simple dérivation et maximisation de ces fonctions. L'estimation des paramètres est donc basée sur des approximations de la fonction de partition (méthodes numériques, méthodes MCMC) ou sur des techniques ne nécessitant pas le calcul de cette dernière (pseudo-vraisemblance ou Takács-Fiksel). Il est également important de mentionner les méthodes Bayésiennes. Elles ont la particularité de considérer une distribution de probabilité sur les paramètres (loi dite *a priori*) avant de «mettre à jour» la distribution des paramètres sachant les observations (loi dite *a posteriori*) par la formule de Bayes. L'échantillonnage de la loi *a posteriori* pour les processus ponctuels est généralement difficile et nécessite des techniques adaptées (algorithmes basés sur des variables auxiliaires [Møller et al., 2006, Murray et al., 2006] ou algorithmes ABC [Stoica et al., 2017]). Les méthodes statistiques permettent également de contrôler la qualité de l'inférence (validation et choix de modèles, normalité asymptotiques, analyse résiduelle...). Pour finir, l'inférence peut devenir plus complexe lorsqu'il s'agit de traiter des données incomplètes, cette difficulté est souvent rencontrée pour des processus ponctuels (domaine privé et donc non observable en foresterie ou en épidémiologie, prédiction en séismologie, obstruction par des objets).

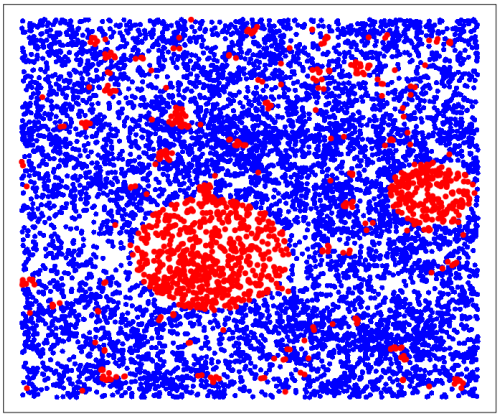
Pour des références plus générales sur les processus ponctuels, une personne souhaitant approfondir et trouver les preuves des propositions et théorèmes énoncés ci-après peut se référer aux ouvrages suivants : [van Lieshout, 2019], [van Lieshout, 2000], [Stoica, 2025],
[Møller and Waagepetersen, 2004], [Baddeley et al., 2015], [Stoyan and Stoyan, 1994], [Geyer, 1999], [Illian et al., 2008], [Daley and Vere-Jones, 2003], [Daley and Vere-Jones, 2008].

Les données cosmologiques sont des données volumineuses et particulièrement riches en structures

induites. Leurs traitement et analyse sont un travail subtil et demandent des outils rigoureux à cet effet. Elles exhibent également des obstructions dues à la luminosité des galaxies ou simplement dues à la présence d'une galaxie devant une autre par rapport à l'observateur. Les deux figures ci-dessous illustrent ces phénomènes : la quantité importante de données et les obstructions dans un jeu de données simulé.



(a) Échantillon issu du 2dF Galaxy Redshift Survey, chaque point représente une galaxie. La zone en rouge correspond à la fenêtre extraite et représente plusieurs milliers de points.



(b) Extraction d'une projection sur $\mathbb{S}^2$ d'un catalogue de galaxies simulées par des astrophysiciens. Les points bleus représentent des galaxies observées, les rouges représentent des galaxies occultées. Données fournies par Pr. Elmo Tempel.

Notre approche propose d'analyser ce type de données à travers leurs modélisation, simulation et inférence en utilisant les processus ponctuels de Gibbs. Les sections suivantes donnent un historique et un éventail des travaux connexes à cette approche dans ces trois composantes.

1.2 Modélisation

Les relevés de galaxies à grande échelle (années 1970, CfA Redshift Survey) montrent que l'organisation de la matière à grande échelle n'a pas une répartition uniforme. L'Univers s'organise comme un alliage complexe de filaments, de clusters de galaxies, et de zones de vides cosmiques. Cependant, traduire la complexité morpho-statistique de ces structures est difficile. La pluralité géométrique et topologique de la toile cosmique ainsi que la taille des jeux de données suggèrent une approche statistique où les galaxies sont réduites à des points observés dans un volume : une étude à l'aide des processus ponctuels semble donc adaptée. Pour une étude générale sur le sujet, nous recommandons [Martinez and Saar, 2001].

Les contributions des deux communautés, mathématique et astrophysique, ont développé des outils pour aboutir à la modélisation et à la détection de structures émergeant dans ce type de données. Nous résumons brièvement certains types de processus utilisés dans la littérature scientifique :

— **Processus de Scott-Neyman** : introduits par les auteurs du même nom [Neyman and Scott, 1958] pour modéliser la distribution des galaxies, cette classe de modèle se base sur un processus de Poisson homogène comme centre de clusters. Le nombre

de points de chaque cluster est ensuite tiré aléatoirement suivant une loi de probabilité discrète, leur localisation suit quant à elle une loi à densité, similaire pour chaque cluster.

— **Processus de Cox log-Gaussien** : les processus de Cox, ou «doubly stochastic Poisson process», introduits par [Cox, 1955], sont générés de la manière suivante : une intensité $\lambda(\mathbf{x})$ est choisie de manière aléatoire sur le domaine, conditionnellement à cette intensité aléatoire, le processus est un processus de Poisson inhomogène d'intensité λ. Utilisé d'abord comme champ aléatoire, le modèle log-gaussien introduit par [Coles and Jones, 1991] s'est généralisé en processus ponctuel [Møller et al., 1998] en utilisant un champ gaussien pour intensité. Ce modèle a par la suite été étendu à la sphère pour décrire la position des galaxies [Møller and Cuevas-Pacheco, 2018].

— **Processus de Cox par segment** ([Pons-Bordería et al., 1999],[Martinez and Saar, 2002]) : des segments de longueur l sont générés aléatoirement dans le domaine puis des points sont générés avec une intensité donnée autour de ces segments. Ce modèle a également été utilisé afin de quantifier la taille des filaments [Pandey, 2010]

— **Tesselations de Voronoï** [Voronoi, 1908] : elles représentent une partition de l'espace définie par un ensemble de points tiré aléatoirement (e.g. en contrôlant la corrélation spatiale). Chaque point est alors entouré par une cellule de Voronoï obtenue en traçant l'intersection des médiatrices entre ces points (dans le cas 2D). [Icke and van de Weygaert, 1987, Icke and van de Weygaert, 1991] ont utilisé ces tesselations pour décrire la distribution des galaxies dans l'Univers. Cette approche sert encore aujourd'hui à l'étude et à la détection des structures galactiques [Zaninetti, 2006, Zaninetti, 2018, Paranjape and Alam, 2020, Soares-Santos et al., 2010].

Tous ces processus constituent des modélisations différentes avec leurs avantages et inconvénients en termes de flexibilité de modélisation, simulation et inférence.

Nous allons considérer comme cadre de modélisation les processus ponctuels de Gibbs. Nous considérons également que ces processus vivent dans un compact W de $\mathbb{R}^2$. La distribution de ce processus est contrôlée par une densité de probabilité par rapport à la mesure de référence, le processus de Poisson d'intensité unité sur W. Son expression est donnée par :

$$f(\mathbf{x}) = \frac{1}{Z} \exp[-U(\mathbf{x})]$$

où U représente la fonction d'énergie de la densité et Z la constante de normalisation du modèle. Nous considérons de plus que le processus est simple, c'est-à-dire deux points dans une configuration ne peuvent pas partager la même position. Plus de détails seront donnés dans le Chapitre 2. À noter que certains résultats énoncés dans ce mémoire sont valables dans un cadre plus large et des références appropriées sont fournies le cas échéant.

— Sans interaction (Processus de Poisson) : le modèle est complètement connu et les formules analytiques sont exploitables pour des statistiques descriptives.

— Répulsion : en pénalisant l'apparition de deux points proches l'un de l'autre (Modèle de Strauss [Strauss, 1975], processus de saturation de Geyer [Geyer, 1999]) ou en contrôlant la surface occupée par les boules centrées en les points du processus (Processus Area Interaction [Baddeley and Van Lieshout, 1995]).

— Attraction : en favorisant l'apparition de deux points proches l'un de l'autre (processus de saturation de Geyer) ou à nouveau en contrôlant la surface occupée par les boules centrées en les points du processus (Processus Area Interaction).

Les processus ponctuels de Gibbs offrent une boîte à outils où il est aisé de créer de nouveaux modèles en adaptant les statistiques des modèles existants (e.g. le modèle de saturation de Geyer peut être vu comme une extension du modèle de Strauss) ou en superposant plusieurs modèles déjà existants.

C'est à partir des années 2000 que les processus ponctuels de Gibbs sont mis en avant en les appliquant directement à la détection de structures induites par la position des galaxies ou à l'analyse d'images [Stoica, 2001, Stoica et al., 2002, Stoica et al., 2004]. Le modèle Candy [Stoica, R. S. et al., 2005, Tempel et al., 2016] propose la détection des filaments galactiques à partir de l'observation de galaxies. L'idée de ce modèle se base sur l'approximation de filaments galactiques à l'aide de cylindres comme l'illustre la figure ci-dessous :

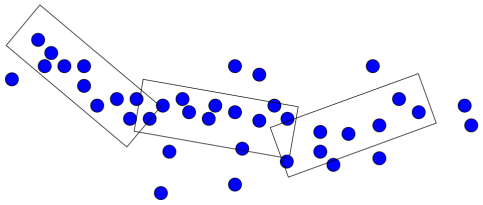


FIGURE 1.2 – Projection en 2D de l'approximation de filaments galactiques (cylindres) à partir des galaxies observées (points). Source : [Stoica, 2025]

.

Une réalisation de ce modèle spécifié par la densité d'un processus ponctuel de Gibbs marqué correspond donc à une configuration aléatoire de cylindres dans un domaine W où les marques correspondent aux caractéristiques du cylindre. Ce modèle a par la suite été généralisé à travers le modèle Bisous [Stoica et al., 2005b].

Les modèles de processus de Gibbs cités plus haut (Strauss, Area-Interaction,...) ont également fait l'objet de modélisation de données galactiques utiles à la caractérisation de ces données [Hurtado-Gil et al., 2021]. Plus récemment, des modèles de processus ponctuels de Gibbs couplés avec l'inférence Bayésienne ont été utilisés pour limiter des biais d'observations pour des catalogues de vitesses de galaxies [Sorce et al., 2023].

La section 2.1 revient sur la définition des processus ponctuels et détaille la construction des processus ponctuels de Gibbs et présente également les modèles cités dans cet état de l'art.

1.3 Simulation

Les méthodes énoncées dans cette introduction sont focalisées sur la simulation des processus ponctuels de Gibbs, ces dernières peuvent néanmoins s'appliquer à d'autres lois de probabilités qui ne sont pas des processus ponctuels. Dans ce contexte, à l'exception de certains processus,

les modèles sont définis par des densités de probabilités par rapport au processus de Poisson standard et leurs fonctions de partition ne sont pas calculables analytiquement. Considérons de plus que ces densités peuvent être écrites de la manière suivante :

$$f(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} \exp\langle t(\mathbf{x}), \theta \rangle \tag{1.1}$$

avec $Z(\theta)$ la fonction de partition, $t(\mathbf{x})$ et θ le vecteur de statistiques suffisantes et le vecteur de paramètres, respectivement.

Dans notre cas, ces algorithmes de simulation reposent sur la simulation d'une chaîne de Markov afin d'échantillonner la loi d'intérêt f, nous parlons alors de méthode MCMC (Monte Carlo Markov Chain). Dès lors que la construction d'un algorithme repose sur une chaîne de Markov, il est important de garantir la convergence de cette dernière vers la distribution d'intérêt. Ainsi, même si la simulation apparaît «simple» en première lecture des pseudo-codes, le travail théorique pour aboutir à ces derniers est difficile. L'échantillonnage peut se faire de manière exacte (algorithme de simulation parfaite) ou de manière approchée.

Pour les algorithmes de simulation parfaite, la réalisation obtenue en sortie de l'algorithme est issue de la distribution d'intérêt. Ces algorithmes reposent sur le principe de couplage par le passé (ou «coupling from the past» CFTP en anglais) [Propp and Wilson, 1996]. Ce principe réside dans le couplage de chaînes de Markov lancées simultanément à un temps $-T < 0$ possédant toutes le même mécanisme de mise à jour avec des états initiaux différents. Au moment où toutes les chaînes possèdent le même état, le phénomène de *coalescence* est atteint. Sans rentrer dans les détails théoriques, cette coalescence élimine la dépendance des états initiaux des chaînes de Markov et garantit alors que les états suivants de la chaîne sont tirés exactement suivant la loi d'intérêt. Une première application à la simulation des processus ponctuel est alors proposée par [Kendall and Møller, 2000] en créant une version «dominée» du CFTP pour simuler le processus de Strauss. Les méthodes CFTP ont par la suite suscité un vif intérêt pour la simulation des processus ponctuels de Gibbs (e.g. [van Lieshout and Stoica, 2006, Ambler and Silverman, 2009]). Un autre avantage de ces méthodes réside dans le fait que, une fois en possession d'un échantillon tiré exactement suivant la loi, les mises à jour de cet état par des algorithmes de type Metropolis-Hastings sont garanties d'être également tirées suivant la loi d'intérêt. Néanmoins, pour les processus ponctuels marqués, les performances numériques peuvent se dégrader lorsque les interactions entre les points sont trop fortes.

Les algorithmes de Metropolis-Hastings (MH) sont, quant à eux, des algorithmes convergeant vers cette loi d'intérêt [Metropolis et al., 1953, Hastings, 1970]. Hélas, il n'est pas possible de savoir quand cette convergence a lieu. Ainsi, il faudrait théoriquement itérer ces derniers une infinité de fois afin d'obtenir un échantillon tiré exactement suivant la loi cible. En pratique, seul un nombre fini d'itérations est utilisé pour obtenir un échantillon. Se pose alors la question du temps de convergence : à partir de combien d'itérations de ces algorithmes pouvons-nous considérer que l'échantillon de sortie est suffisamment proche de la distribution f ? La réponse dépend du modèle duquel nous souhaitons échantillonner et du type d'algorithme MH utilisé. Le principe général des algorithmes MH est le suivant, supposons que nous voulons échantillonner une certaine loi de probabilité π :

1. Une nouvelle valeur de la chaîne de Markov est proposée suivant une distribution de probabilité q pouvant dépendre de l'état précédent.

2. Cette nouvelle valeur est acceptée ou rejetée suivant un ratio défini par :

$$\alpha := \min\left\{1, \frac{\pi(y)}{\pi(x_{i-1})} \times \frac{q(x_{i-1}|y)}{q(y|x_{i-1})}\right\}.$$

Ce ratio a l'avantage de ne pas dépendre de la constante de normalisation de la loi à échantillonner et permet de contourner ce problème dans le cas des processus ponctuels de Gibbs. De plus, la loi q peut être choisie de manière à favoriser la convergence de la chaîne de Markov (e.g. [van Lieshout and Stoica, 2003, Taty Moukati et al., 2024, Stoica, 2025]). Nous allons dans la suite utiliser l'algorithme de Metropolis-Hastings pour l'ajout/retrait de points [van Lieshout, 2019, Møller and Waagepetersen, 2004]. Un nouvel état sera proposé de la manière suivante :

— Ajout : l'algorithme propose un point uniformément dans le domaine.
— Retrait : l'algorithme propose d'enlever un point tiré uniformément parmi les points existants.

Afin d'en comprendre les subtilités, la section 2.2 propose des rappels sur les chaînes de Markov et leurs propriétés pour ensuite se focaliser sur cet algorithme.

1.4 Inférence paramétrique

L'inférence paramétrique représente un des outils important pour l'analyse de données en utilisant les processus ponctuels. En effet, estimer les paramètres d'un modèle constitue une étape clé pour étudier les caractéristiques d'une configuration de points. Couplée avec des outils de contrôles des modèles, l'inférence paramétrique permet d'affirmer si oui ou non les configurations obtenues avec le paramètre estimé sont cohérentes avec les observations. Les valeurs estimées des paramètres permettent également de déduire les caractéristiques d'une configuration de points (e.g. répulsion, agglomération).

Cette section propose de revenir brièvement sur les diverses méthodes utilisées pour l'inférence paramétrique des processus ponctuels. Il est important de garder à l'esprit que ces méthodes ont pour la plupart été utilisées ou développées pour l'inférence paramétrique pour des classes de modèles plus générales que les processus ponctuels de Gibbs. À l'origine, l'inférence était souvent basée sur des statistiques descriptives telles que la fonction K de Ripley [Diggle, 1983, Ripley, 1977, Ripley, 1981, Ripley, 1988]. Nous nous focaliserons plutôt sur les techniques paramétriques basées sur les modèles telles que la pseudo-vraisemblance, les approximations stochastiques ou Takács-Fiksel. Les méthodes Bayésiennes seront aussi abordées et données plus en détail en section 2.3. Ces dernières années, en raison de l'amélioration des capacités de calcul des ordinateurs, les méthodes basées sur la simulation des modèles sont de plus en plus utilisées. Néanmoins, des précautions sont à prendre afin d'éviter des simulations très coûteuses.

Estimation par maximum de vraisemblance

1) Estimation en données complètes

Pour des modèles spécifiés par des densités comme celle présentée ci-dessus en formule 1.1, la log −vraisemblance d'un tel modèle est donnée par

$$\log L(\theta) = \log\left(\exp\langle t(\mathbf{x}), \theta\rangle\right) - \log Z(\theta).$$

À part pour certains modèles (e.g. Poisson), l'expression de $Z(\theta)$ n'est pas connue et les dérivées partielles par rapport aux différents θ_i ne le sont également pas. L'optimisation de cette fonction paraît alors impossible de prime abord. Nous présentons ici certaines approches pour résoudre ce problème.

a) Pseudo-vraisemblance

Une première solution à ce problème a été introduite par [Besag, 1974] pour les champs de Markov et reprise par la suite pour le processus de Strauss [Besag, 1978]. Développée ensuite plus généralement pour les processus ponctuels [Møller and Jensen, 1991] et les processus de Gibbs [Särkkä, 1993] (avec interactions par paire dans ce cas) [Särkkä, 1995, Goulard et al., 1996] (cas plus général). Cette technique est encore étudiée pour d'autres modèles (e.g. Lennard-Jones) [Coeurjolly and Lavancier, 2017] ou utilisée pour estimer l'intensité d'un processus [D'Angelo et al., 2024]. La pseudo-vraisemblance se base sur l'intensité conditionnelle λ^* donnée par

$$\lambda^*(\mathbf{x}, \eta) = \frac{f(\mathbf{x} \cup \{\eta\}|\theta)}{f(\mathbf{x}|\theta)},$$

où $\mathbf{x}$ est une configuration de points (finie) de W et $\eta \in W \setminus \mathbf{x}$ avec la convention $a/0 =$ pour $a \geq 0$. Cette quantité est plus facile à calculer puisqu'elle ne dépend pas de la constante de normalisation du modèle. De plus, pour certains types de processus, cette dernière décrit entièrement le processus. Néanmoins, sa définition demande de prévenir les cas pathologiques, par exemple quand $f(\mathbf{x}|\theta) = 0$. La $\log -$pseudo-vraisemblance s'écrit alors

$$\log PL(\theta) = \sum_{\eta \in \mathbf{x}} \log \lambda^*(\mathbf{x}, \mathbf{x} \setminus \{\eta\}) - \int_W \lambda^*(\mathbf{x}, u)\mu(\mathrm{d}u)$$

pour $\mathbf{x}$ une configuration de points dans un domaine W. La pseudo-vraisemblance, pour des modèles de processus ponctuel admettant une densité sous la forme exponentielle, est concave [Møller and Jensen, 1991]. Ceci rend alors l'optimisation de l'expression ci-dessus réalisable (e.g. par descente de gradient). Pour finir, des résultats sur la consistance et la normalité asymptotiques ont été établis pour cet estimateur [Jensen and Kunsch, 1994, Mase, 2000, Billiot et al., 2008, Coeurjolly and Drouilhet, 2010].

b) Méthode de Takács-Fiksel

La méthode de Takács-Fiksel [Fiksel, 1984, Takács, 1986, Takács and Fiksel, 1986, Fiksel, 1988], repose sur l'équation de Georgii-Nguyen-Zessin :

$$\mathbb{E}\left[\sum_{\eta \in X} h(\eta, X \setminus \{\eta\})\right] = \mathbb{E}\left[\int_W h(u, X)\lambda^*(u, X)\mu(\mathrm{d}u)\right]$$

pour X un processus ponctuel stationnaire et un choix particulier de la fonction h positive et mesurable telle que les espérances considérées soient bien définies. L'idée est de choisir h et des estimations des deux termes de l'équation afin d'aboutir à des équations dont les solutions sont les paramètres du modèle [Särkkä, 1993, Diggle et al., 1994, Coeurjolly et al., 2012, Coeurjolly et al., 2016, Jansson and Cronie, 2024]. Il est par exemple possible de retrouver l'équation de la pseudo-vraisemblance en prenant h comme la dérivée par rapport au paramètre de

l'intensité conditionnelle. Pour plus de précisions sur le sujet, nous recommandons les monographies [van Lieshout, 2000, Møller and Waagepetersen, 2004, Stoica, 2025].

c) Méthodes par approximation de la vraisemblance

Les méthodes par approximation de la vraisemblance ont toutes pour but de trouver une manière d'approcher la vraisemblance afin d'en trouver le maximum. Les travaux de
[Ogata and Tanemura, 1981, Ogata and Tanemura, 1984, Ogata and Tanemura, 1985] et
[Penttinen, 1984] proposent d'approcher l'estimateur du gradient de la constante de normalisation $\nabla \log Z(\theta)$ par des approximations polynomiales puis d'intégrer ces polynômes pour retrouver une valeur approchée de $\log Z(\theta)$. Ce sont des méthodes asymptotiques avec des hypothèses très fortes, elles ne sont donc généralement pas applicables dans le cas général.

D'autres solutions sont basées sur la simulation de modèles par chaînes de Markov. Le but est alors de générer une chaîne de Markov admettant le modèle considéré pour distribution invariante afin de s'appuyer sur des réalisations du modèle pour l'inférence. [Geyer and Thompson, 1992] proposent une solution pour des modèles autologistiques et de la classe exponentielle,
[Geyer and Møller, 1993] proposent une application aux processus de Gibbs. Ces approches reposent sur l'échantillonnage préférentiel. Dans ce cas précis, ceci consiste à réécrire les ratios de fonctions de partitions présents dans le ratio de log-vraisemblance par rapport à un autre paramètre fixé ψ :

$$\log L(\theta) = \langle t(\mathbf{x}), \theta - \psi \rangle - \log \frac{Z(\theta)}{Z(\psi)},$$

comme une espérance sous la loi $f(\cdot|\psi)$:

$$\frac{Z(\theta)}{Z(\psi)} = \mathbb{E}_{\psi} \left[\exp \langle t(X), \theta - \psi \rangle \right].$$

Finalement, ce ratio est approché par sa moyenne empirique en simulant des échantillons $X_1, ..., X_n$ tirés suivant $f(\mathbf{x}|\psi)$.

Enfin, [Penttinen, 1984, Younes, 1988, Moyeed and Baddeley, 1991] proposent de combiner des méthodes MCMC et des méthodes itératives basées sur les algorithmes de Newton-Raphson et de Robbins-Monro. Ces méthodes itératives ont pour but de générer une suite de paramètres $(\theta_k)_{k \geq 0}$ qui converge vers le maximum de vraisemblance du modèle. Bien que cela ne soit pas toujours appliqué aux processus ponctuels, des travaux plus récents sur les méthodes d'approximation de la vraisemblance sont encore développés (e.g. gradient à pas optimal [Descombes et al., 1999, Stoica, 2001, van Lieshout and Stoica, 2003]).

2) Approximation de la vraisemblance en données incomplètes

Les données incomplètes émergent régulièrement dans les domaines d'applications des processus ponctuels (e.g. en cosmologie, à cause de la luminosité des galaxies ou des obstructions par des objets galactiques). Pour des données spatialisées, plusieurs mécanismes de «dégradation» des données peuvent intervenir [Lund and Rudemo, 2000, Guttorp et al., 2023] :

— L'amincissement («thinning» en anglais), où les points dans la région observée peuvent être omis dans les relevés (e.g. oublis de relevés).

— Les déplacements de points, où la position n'est pas connue assurément (e.g. incertitudes de relevés).
— La censure de points, lorsque seule une sous-région du domaine peut être observée (e.g. en foresterie : domaine public/domaine privé).
— Des ajouts involontaires de points (e.g. erreurs de relevés).

Pour l'étude menée dans ce manuscrit, nous nous focalisons sur le troisième point, où nous supposons que seule une sous-région du domaine considéré est observée, comme l'illustre la figure ci-dessous. Nous supposons néanmoins que le processus se déroule sur le domaine $W = W_Y \cup W_X$.

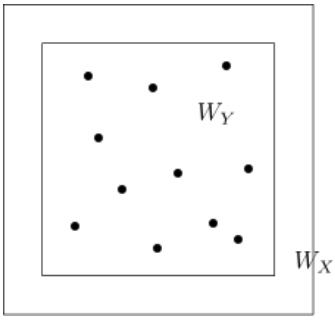


FIGURE 1.3 – Schéma explicatif de la censure de points. La configuration de points n'est observée que dans le domaine W_Y et les points dans W_X ne sont pas observés.

Pour coller à ce contexte, nous considérons la densité jointe des données

$$f(\mathbf{x}, \mathbf{y}|\theta) = \frac{\exp\langle t(\mathbf{x} \cup \mathbf{y}), \theta\rangle}{Z(\theta)}$$

où $\mathbf{y}$ est une configuration de points observée dans la région W_Y et $\mathbf{x}$ est une configuration cachée dans la région W_X.

L'inférence en données incomplètes par maximum de vraisemblance est un problème difficile. La vraisemblance dépend désormais d'intégrales qui ne sont généralement pas accessibles sous forme analytiques. En effet, puisque la zone W_X n'est pas observée, une intégrale sur toutes les configurations possibles dans cette zone intervient au numérateur de la vraisemblance. Nous résumons ici certaines approches permettant de prendre en compte les données incomplètes. Plus de détails peuvent être trouvés dans la review [Ruth, 2024].

L'algorithme Espérance-Maximisation (EM) [Dempster et al., 1977] est une méthode fondamentale pour l'analyse des données incomplètes. C'est un algorithme itératif qui peut être utilisé lorsque les données sont considérées partiellement observées (e.g. échantillons incomplets, labels des observations manquants, ...). L'algorithme EM itère deux étapes :

— L'étape E, qui consiste à calculer l'espérance conditionnelle de la vraisemblance complète sachant les données observées $\mathbf{y}$ à l'itération k :

$$Q(\theta, \theta_{k-1}) := \mathbb{E}_{\theta_{k-1}}[L(\theta, Y, X)|Y = \mathbf{y}]$$

— L'étape M, qui consiste à maximiser l'espérance conditionnelle précédemment calculée.

Bien que la convexité de la vraisemblance complète ne puisse pas être garantie, chaque itération de l'algorithme constitue toujours un «pas croissant» de la vraisemblance (la vraisemblance évaluée en θ_{k+1} est supérieure ou égale à l'évaluation en θ_k). Sous de bonnes hypothèses de régularité des modèles, il est également possible d'obtenir les différentielles successives de la vraisemblance observée, ce qui permet d'obtenir des erreurs asymptotiques pour les estimateurs des paramètres. Néanmoins, l'étape E n'est pas toujours réalisable lorsque des modèles plus complexes sont proposés.

Une manière de pallier ce problème est d'utiliser des méthodes Monte Carlo (donnant donc l'algorithme MCEM). Cette solution, proposée par [Wei and Tanner, 1990], remplace le calcul de l'espérance conditionnelle de l'étape E par une approximation Monte Carlo. Pour ce faire, des échantillons pour les données manquantes sont générés à chaque itération conditionnellement aux observations. Ensuite, l'étape M s'effectue à nouveau par une méthode d'optimisation. Les propriétés de convergence de ce type d'algorithme peuvent être trouvées dans [Fort and Moulines, 2003] ou une review plus générale [Neath, 2013].

Une autre technique repose quant à elle sur l'utilisation des méthodes d'approximations stochastiques (donnant donc l'algorithme SAEM). [Gu and Li, 1998] utilisent une estimation du score des données (le gradient de la vraisemblance observée), où les itérations de l'algorithme convergent vers un point critique de cette fonction. [Delyon et al., 1999] présentent quant à eux un algorithme SAEM appliqué au cas de la famille exponentielle où l'estimation se base directement sur l'estimation la fonction Q. D'autres approches proposant des tentatives d'optimisation globale ont aussi été abordées en perturbant la dynamique de l'algorithme EM afin d'échapper aux zones d'attractions des minimums locaux (e.g. [Celeux et al., 1995] pour une review de ces approches).

D'autres approches alliant ces deux idées existent également (e.g. [Kuhn and Lavielle, 2004]) et des approches s'affairant à échantillonner la vraisemblance jointe des données observées et manquantes sont proposées par [Gelfand and Carlin, 1993] en s'appuyant en partie sur l'échantillonnage préférentiel proposé par [Geyer and Thompson, 1992] présenté plus haut. Dans le cas des données incomplètes, la vraisemblance peut être vue comme la loi marginale suivant la variable observée $\mathbf{y}$:

$$L(\theta) = f(\mathbf{y}|\theta) = \int f(\mathbf{x}, \mathbf{y}|\theta)\mu(\mathrm{d}\mathbf{x}) = \frac{Z(\theta|\mathbf{y})}{Z(\theta)},$$

avec $Z(\theta|\mathbf{y})$ la constante de normalisation de la densité $f(\mathbf{x}|\mathbf{y}, \theta)$.

Comme précédemment présenté dans le cas des données complètes, les ratios de constantes de normalisation intervenant dans le ratio de vraisemblance sont réécrits comme des espérances et approchés par leurs espérances empiriques à l'aide d'échantillons de la loi jointe $(X_i, Y_i)_{1\leq i\leq n}$ et de la loi conditionnelle de $(X_i^*)_{1\leq i\leq n} := (X_i|Y = \mathbf{y})_{1\leq i\leq n}$.

$$\frac{Z(\theta|\mathbf{y})}{Z(\psi|\mathbf{y})} \approx \frac{1}{n}\sum_{i=1}^{n} e^{\langle t(X_i^*,\mathbf{y}),\theta\psi\rangle} \text{ et } \frac{Z(\theta)}{Z(\psi)} \approx \frac{1}{n}\sum_{i=1}^{n} e^{\langle t(X_i,Y_i),\theta-\psi\rangle}$$

Nous reviendrons sur cette approche par échantillonnage préférentiel dans la suite de ce manuscrit en section 2.3.

Les données manquantes peuvent également découler du modèle considéré. C'est le cas par exemple des «boolean models» qui sont un cas particulier des modèles «germes/grains» où les germes sont issus d'une réalisation du processus de Poisson et les grains représentent des disques, carrés ou triangles placés sur les germes. Une réalisation de ce type de processus est alors obtenue en prenant la réunion des grains. Il est alors possible qu'un germe ne soit pas observé si le disque/carré/triangle qui lui est associé est totalement recouvert par les autres germes. [van Lieshout and van Zwet, 2000, van Lieshout and van Zwet, 2001] s'intéressent à la simulation parfaite de la distribution conditionnelle des germes non observés par rapport aux observés. Ces échantillons sont alors utilisés pour obtenir un estimateur Monte Carlo du maximum de vraisemblance de l'intensité du processus puis cette estimation est comparée à celle menée via un algorithme SAEM.
[Møller and Helisová, 2010] considèrent un modèle similaire et une approche par estimation MCMC du maximum de vraisemblance pour également prendre en compte les effets de bords.

Enfin, d'autres approches basées sur l'estimation de la fonction d'intensité sont également utilisées dans le cas de processus non homogènes : [Gabriel et al., 2023] considèrent une situation similaire à celle décrite en Figure 1.3. Ils ne cherchent cependant pas directement à prédire une configuration de points dans la zone non observée. Le but est plutôt d'estimer la fonction d'intensité en tout point de cette zone conditionnellement aux observations et aux covariables définissant cette dernière.

Dans la suite du manuscrit, nous ferons l'abus de langage de parler de «problème des données incomplètes» pour faire référence à l'inférence paramétrique lorsque des mécanismes de données manquantes sont considérés.

Méthodes d'échantillonnage de la loi *a posteriori*

Jusqu'ici, nous avons considéré l'inférence paramétrique «classique» qui consiste à chercher la valeur la plus plausible par rapport aux observations. L'inférence Bayésienne, elle, considère une connaissance préalable sur la distribution des paramètres du modèle en introduisant la loi *a priori* $p(\theta)$. L'information apportée par les données $\mathbf{x}$ est alors prise en compte en considérant la loi conditionnelle des paramètres sachant $\mathbf{x}$, aussi appelé loi *a posteriori* :

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)p(\theta)}{\int f(\mathbf{x},\theta)\mu(\mathrm{d}\theta)} = \frac{f(\mathbf{x}|\theta)p(\theta)}{Z(\mathbf{x})}$$

dans notre cas, cette loi s'exprimera alors :

$$f(\theta|\mathbf{x}) = \frac{\exp\langle t(\mathbf{x}),\theta\rangle p(\theta)}{Z(\mathbf{x})Z(\theta)}$$

avec $Z(\mathbf{x})$ la constante de normalisation liée aux données et $Z(\theta)$ la constante de normalisation du modèle.

La loi *a posteriori* possède alors deux constantes de normalisation qui ne sont en général pas calculables dans le cas des processus ponctuels de Gibbs. À nouveau, des méthodes permettant de ne pas calculer ces constantes existent.

L'une de ces méthodes est l'inférence variationnelle, que nous mentionnons ici mais que nous ne détaillerons pas dans la suite. L'idée principale est également d'approcher la distribution

a posteriori $f(\theta|\mathbf{x})$. L'approximation se fait en choisissant une famille de distributions candidates q_θ dont le paramètre est par la suite ajusté afin que la distribution soit la plus proche de f en minimisant un critère (e.g. la distance en variation totale). Le problème d'échantillonnage devient alors un problème de minimisation, souvent plus commode à traiter. Pour plus de détails sur le sujet, nous recommandons l'article de review [Blei et al., 2017] et la thèse [Daudel, 2021]. Pour des travaux axés sur les processus ponctuels, le/la lecteur/rice peut se référer à [Baddeley et al., 2013a, Jahnel, 2024]. L'inférence variationnelle s'utilise également pour résoudre des problèmes de données incomplètes (e.g. pour le «spatial error model» avec différents mécanismes de censures, voir [Wijayawardhana et al., 2024] pour plus de détails) en proposant d'approcher la loi jointe du paramètre et des données manquantes ou d'utiliser des simulations MCMC pour simuler les données manquantes puis d'utiliser l'inférence variationnelle «classique», en données complètes.

D'autres méthodes bayésiennes utilisées pour échantillonner la loi *a posteriori* sont détaillées dans la suite du manuscrit et nous reviendrons plus en détail sur la méthode «ABC Shadow», utilisée dans notre cas pour mener l'inférence paramétrique des modèles.

1.5 Motivations de la thèse

Ces travaux proposent d'apporter des contributions méthodologiques pour l'analyse des données en cosmologie. Cette thèse s'inscrit dans un paysage où nous retrouvons beaucoup de travaux d'analyse descriptive par des statistiques non paramétriques pour les données représentées par des catalogues galactiques (e.g. [Totsuji and Kihara, 1969, Peebles, 1973, Snethlage et al., 2002]). Ces analyses montrent très bien l'apparition de structures à grande échelle dans la distribution des galaxies telles que des clusters, des filaments, des murs de galaxies et des zones de vide galactique. La distribution de la matière est donc tout sauf uniforme dans l'Univers.

Néanmoins, peu de travaux traitent de la modélisation de cette distribution en elle-même par les processus de Gibbs [Hurtado-Gil et al., 2021]. Les lois de la physique montrent l'importance de pouvoir modéliser des interactions entre les galaxies et de pouvoir éventuellement prendre en compte leurs masses, tailles, types ou luminosités. Ainsi, les galaxies peuvent être vues comme des objets ou des points marqués interagissant entre eux et qui forment des structures à plus grande échelle. Cette thèse propose alors deux modélisations de données galactiques : l'une, basée sur l'observation de filaments galactiques, utilise un modèle multi-interaction pour la répartition des galaxies autour des filaments. L'autre s'appuie davantage sur une modélisation multi-échelle en considérant une superposition de modèles exhibant de la répulsion et de l'agglomération afin de modéliser les phénomènes physiques observés (agglomération lorsque les galaxies sont proches entre elles et répulsion lorsqu'elles sont suffisamment éloignées).

Les données galactiques exhibent naturellement des obstructions et représentent donc un problème spatial de données manquantes. À notre connaissance, le traitement des données incomplètes dans le contexte Bayésien pour les processus ponctuel de Gibbs est encore un problème peu abordé. Une autre contribution de cette thèse est donc de proposer un algorithme pour l'échantillonnage *a posteriori* pour l'inférence paramétrique des modèles dans ce contexte où les données ne sont pas complètement observées.

Cette approche est donc largement motivée par la richesse et la complexité de ces données. Bien

que les outils existants (e.g. la librairie `R` spatstat) soient très performants et très pratiques, des limitations en termes de puissance de calculs et de flexibilité de création de modèles suggèrent d'utiliser le langage `C++`. J'ai pu ainsi contribuer au développement et au test de la librairie `C++ DRLib` (https ://gitlab.univ-lorraine.fr/labos/iecl/drlib), dédiée à la simulation et l'inférence des processus de Gibbs, en y ajoutant les modèles décrits plus loin dans ce manuscrit. Ces travaux axés sur l'informatique ont alors donné lieu à des contributions au projet européen Excosm (https ://excosm.ut.ee/) à travers un cours d'introduction à cette librairie.

1.6 Résumé des chapitres

Le chapitre 2 introduit l'intégralité des prérequis nécessaires à la définition rigoureuse et à la compréhension des contributions de cette thèse :

- La section 2.1 présente des éléments fondamentaux sur les processus ponctuels utile à la construction des processus de Gibbs que nous utilisons. Des exemples de réalisations et de mécanismes de construction de nouveaux processus sont ensuite mis en avant pour finir avec les statistiques descriptives spatiales.
- La section 2.2 aborde la simulation des processus de Gibbs par chaîne de Markov. D'abord, des rappels sur les propriétés de ces dernières sont donnés pour ensuite se focaliser sur les algorithmes de type Metropolis-Hastings pour finir avec des applications de l'algorithme utilisé.
- La section 2.3 propose d'abord de revenir sur les méthodes «Monte Carlo Maximum Likelihood Estimation» (MCMLE) pour les données complètes et incomplètes. Des méthodes bayésiennes pour échantillonner la loi *a posteriori* dans le cadre des processus de Gibbs sont ensuite présentées pour motiver l'utilisation de la méthode «approximate bayesian computation» (ABC), ABC Shadow comme outil pour mener l'inférence paramétrique des modèles. Enfin, nous présentons des outils utiles au contrôle de l'inférence et au bon paramétrage de l'algorithme ABC Shadow.

Les chapitres 3 et 4 présentent les contributions de la thèse :

- La section 3.1 présente d'abord le contexte et les données étudiées ainsi que le traitement de ces dernières pour la création du modèle. Le modèle basé sur la distance des galaxies au plus proche filament est introduit et nous discutons ensuite du traitement informatique pour sa simulation et son inférence. Enfin, nous exposons les résultats, discussions et perspectives obtenus pour ce modèle.
- La section 3.2 propose une approche dans la continuité de la section 3.1 pour un jeu de données plus volumineux. Un modèle «multi-échelle» se basant sur la superposition de modèles permettant d'obtenir un phénomène d'agglomération dans un voisinage proche des points et de répulsion à plus grande distance de ces derniers est ensuite introduit. Cette section propose enfin une autre approche motivée par la création de ce deuxième modèle reposant sur le modèle de saturation de Geyer.
- Le chapitre 4 propose une reformulation du problème des données incomplètes dans le cadre Bayésien et définit une loi *a posteriori* à échantillonner pour prendre en compte les données partiellement observées. Partant de cette distribution, nous établissons ensuite des résultats théoriques de convergence de chaînes de Markov amenant à la construction

d'un algorithme ABC pour échantillonner cette loi. Enfin, nous exhibons des stratégies numériques et les appliquons à une étude par simulation en utilisant le modèle de Strauss.

Enfin, nous donnons des conclusions et des perspectives sur ces travaux et des annexes :

— L'annexe A donne les preuves utiles à la compréhension de la construction de l'algorithme ABC Shadow.
— L'annexe B donne des détails sur la libraire `C++ DRLib`.
— L'annexe C revient sur des points clés du déroulement de la thèse.

Chapitre 2

Modélisation probabiliste, simulation et inférence pour les processus ponctuels de Gibbs

Sommaire

La section 2.1 introduit les définitions et notations utilisées pour les processus ponctuels et processus ponctuels marqués dans $\mathbb{R}^d$. Nous aborderons ensuite le processus ponctuel de Poisson et ses propriétés, exemple fondamental dans la théorie des processus ponctuels n'induisant aucune interaction entre les points. Nous verrons ensuite comment construire les processus ponctuels de Markov à partir du processus de Poisson standard, puis nous aborderons la stabilité de ces processus, garantissant leur intégrabilité et donc leur bonne définition. Pour finir, nous nous intéresserons à différents modèles utilisés dans les travaux de cette thèse ainsi que des outils pour mener l'analyse exploratoire des données. La trame suivie s'est appuyée sur les ouvrages [van Lieshout, 2000], [Møller and Waagepetersen, 2004], [van Lieshout, 2019] et [Stoica, 2025]. Un contenu plus détaillé sur les fonctions utiles à l'analyse descriptive peut être trouvé dans [Baddeley et al., 2015].

2.1 Généralités

2.1.1 Définitions

Dans la suite, nous noterons $\mathbb{R}^d$ et $\mathcal{B}(\mathbb{R}^d)$ pour l'espace euclidien de dimension d et sa tribu borélienne associée. Pour une configuration de points $\mathbf{x} \subseteq \mathbb{R}^d$, $n(\mathbf{x})$ correspondra au cardinal de cet ensemble ($+\infty$ si ce dernier n'est pas fini). Pour $B \in \mathcal{B}(\mathbb{R}^d)$ borné, $\mathbf{x}$ est dit localement fini si $n(\mathbf{x}_B) := n(\mathbf{x} \cap B) < +\infty$ et la mesure de Lebesgue sera notée ν.

Définition 2.1.1. *(Espace des configurations localement finies).*
L'espace $N^{lf} = \{\mathbf{x} \subseteq \mathbb{R}^d : n(\mathbf{x}_B) < +\infty \quad \text{pour tout borné} \quad B \subseteq \mathbb{R}^d\}$ *est appelé espace des configurations localement finies.*

Ainsi, nous noterons $\mathbf{x}, \mathbf{y}, ...$ pour des éléments de N^{lf} et $\xi, \eta, ...$ pour des points de $\mathbb{R}^d$. Les abus de notations suivants seront utilisés : $\mathbf{x} \cup \xi$ pour $\mathbf{x} \cup \{\xi\}$, $\mathbf{x} \setminus \eta$ pour $\mathbf{x} \setminus \{\eta\}$ pour $\mathbf{x} \in N^{\mathrm{lf}}$ et $\xi, \eta \in \mathbb{R}^d$. La configuration ne contenant aucun point sera notée $\emptyset$.

Munissons désormais N^{lf} de la σ-algèbre engendrée par la fonction de comptage $n(\cdot)$:

$$\mathcal{N}^{\mathrm{lf}} = \sigma(\{\mathbf{x} \in N^{\mathrm{lf}} : n(\mathbf{x}_B) = m\} : B \in \mathcal{B}_0, m \in \mathbb{N}_0).$$

Nous sommes désormais en mesure de définir un processus ponctuel :

Définition 2.1.2. *(Processus ponctuel).*
Un processus ponctuel X *sur* $\mathbb{R}^d$ *est une application mesurable définie sur un espace probabilisé* $(\Omega, \mathcal{F}, \mathcal{P})$ *à valeurs dans* $(N^{lf}, \mathcal{N}^{lf})$. *Pour* $F \in \mathcal{N}^{lf}$, *la distribution* $\mathbb{P}_X$ *de* X *est donnée par*

$$\mathbb{P}_X(F) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in F\}).$$

Nous allons parfois identifier X *à sa loi* $\mathbb{P}_X$ *et faire l'abus de langage d'appeler ces deux objets un processus ponctuel.*

Dans ce manuscrit, nous allons nous restreindre au cas des processus ponctuels localement finis et dits «simples», c'est-à-dire que deux points d'un processus ponctuel X sont distincts. Autrement dit,

i) $n(X \cap B) := n(X_B) < \infty$ dès lors que $0 < \nu(B) < \infty$.

ii) $\forall\ \xi_1, \xi_2 \in X$, $\mathbb{P}(\xi_1 \neq \xi_2) = 1$.

Dans la suite, nous supposerons donc que les processus ponctuels sont localement finis et simples.

De manière similaire, nous pouvons définir les processus ponctuels marqués en considérant Y un processus ponctuel sur $B \subseteq \mathbb{R}^d$. Soit M l'espace des marques et $m_\xi \in M$ une marque aléatoire attachée à chaque point $\xi \in Y$, alors

$$X = \{(\xi, m_\xi) : \xi \in Y\}$$

est appelé un processus de point marqué, d'espace de points B et d'espace d'états M. En d'autres termes, la loi marginale de la position des points est un processus ponctuel sur B.

Un processus ponctuel peut être caractérisé de différentes manières, cela peut se faire en considérant les distributions fini-dimensionnelles («*fidis*» en anglais) ou en étudiant la probabilité des évènements vides («*void probabilities*») :

Théorème 2.1.1. *Un processus ponctuel est entièrement caractérisé par ses distributions finidimensionnelles :*

$$\mathbb{P}(n(X_{B_1}) \leq n_1, ..., n(X_{B_m}) \leq n_m) \tag{2.1}$$

pour $B_i \in \mathcal{B}(\mathbb{R}^d)$ borné, $n_i \in \mathbb{N}$ et $m \in \mathbb{N}^*$ fini.

Théorème 2.1.2. *Un processus ponctuel est entièrement caractérisé par la probabilité des évènements vides :*

$$v(B) = \mathbb{P}(n(X_B) = 0) \tag{2.2}$$

pour $B_i \in \mathcal{B}(\mathbb{R}^d)$ borné.

2.1.2 Processus ponctuel de Poisson

Le processus ponctuel de Poisson est construit de sorte que les configurations de points obtenues n'exhibent aucune interaction entre les points. Pour cette raison, il fait office de mesure de référence lorsque l'on souhaite construire des processus plus complexes, avec interactions. Dans la suite, $W \subset \mathbb{R}^d$ représente un compact tel que $\nu(W) > 0$.

Définition 2.1.3. *(Fonction d'intensité et mesure d'intensité).*
Soit $B \subseteq W$, la fonction d'intensité d'un processus ponctuel est définie par $\rho : W \to [0, +\infty[$ telle que $\int_B \rho(\xi)d\xi < +\infty$. La mesure d'intensité μ est donnée par $\mu(B) = \int_B \rho(\xi)d\xi$. Cette mesure est également définie par $\mu(B) = \mathbb{E}[n(X_B)]$.

Nous commençons d'abord par définir un processus ponctuel utile à la définition du processus de Poisson :

Définition 2.1.4. *(Processus binomial).*
Soit f une densité sur un ensemble $B \subseteq W$ et soit $n \in \mathbb{N}$. Un processus X défini par n points i.i.d. tirés suivant la densité f est appelé un processus ponctuel binomial sur B de densité f. Nous noterons $X \sim bin(B, n, f)$.

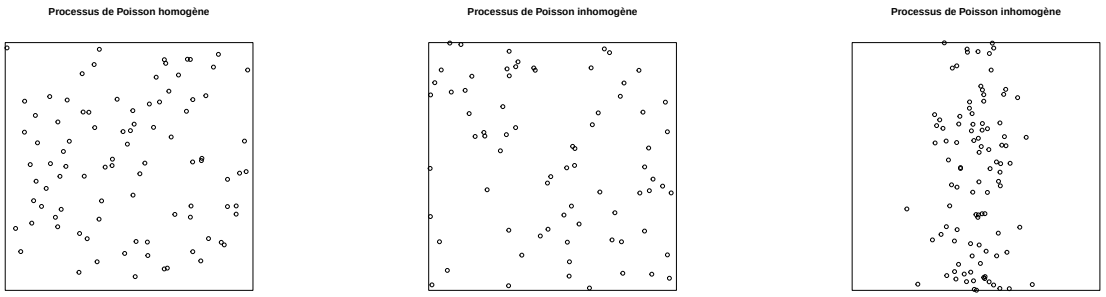
Nous pouvons alors introduire :

Définition 2.1.5. *(Processus ponctuel de Poisson).*
Un processus ponctuel X sur W est un processus ponctuel de Poisson de fonction d'intensité $\rho : W \to [0, +\infty[$ vérifiant les conditions suivantes sur μ :

i) *Pour tout $B \subseteq W$ tel que $\nu(B) > 0$, $n(X_B) \sim Poi(\mu(B))$, la loi de Poisson de moyenne $\mu(B)$. (Si $\mu(B) = 0$ alors $n(X_B) = 0$)*

ii) *Pour tous boréliens bornés $B_1, ..., B_k$ disjoints, les variables aléatoires $n(X_{B_1}), ..., n(X_{B_k})$ sont indépendantes.*
Nous noterons alors $X \sim Poisson(B, \rho)$.

Remarques :
• Si ρ est constante, le processus de Poisson(W, ρ) est appelé processus de Poisson homogène sur W d'intensité ρ. Lorsque ce n'est pas le cas, nous parlerons de processus de Poisson non homogène sur W. Enfin, Poisson$(W, 1)$ est appelé processus de Poisson standard.
• Le cas inhomogène permet par exemple de prendre en compte des structures autour desquelles se répartissent les points (e.g. des galaxies s'agglutinant autour de filaments ou dans les clusters). Plus généralement, cette inhomogénéité peut être induite par les caractéristiques du milieu dans lequel les individus (points) évoluent (e.g. altitude, humidité, température...), favorisant ou défavorisant l'apparition de ces derniers dans certaines zones de ce milieu.



(a) $\rho(x,y) = 100$ (b) $\rho(x,y) = \frac{100}{0.5\sqrt{2\pi}} \exp \frac{(x-0.5)^2}{2\times 0.5^2}$ (c) $\rho(x,y) = \frac{100}{0.1\sqrt{2\pi}} \exp \frac{(x-0.5)^2}{2\times 0.1^2}$

FIGURE 2.1 – Réalisation d'un processus de Poisson homogène d'intensité 100 (gauche) ; Réalisation du processus de Poisson inhomogène d'intensité suivant la densité d'une gaussienne $\mathcal{N}(0.5, 0.5)$ par rapport à la première variable (milieu) ; Réalisation du processus de Poisson inhomogène d'intensité suivant la densité d'une gaussienne $\mathcal{N}(0.5, 0.2)$ par rapport à la première variable (droite). Les trois réalisations sont générées dans le domaine $[0,1] \times [0,1]$.

La figure 2.1a répartit les points de manière homogène sur le domaine, la figure 2.1b favorise l'apparition des points d'abscisses autour de $x = 0.5$ sans trop défavoriser l'apparition de points d'abscisses proches des bords $x = 0$ et $x = 1$. Enfin, la figure 2.1c favorise davantage l'apparition autour $x = 0.5$ et défavorise grandement les bords $x = 0$ et $x = 1$. Puisqu'aucune dépendance par rapport à la variable y n'a été donnée dans les intensités, la répartition de l'ordonnée des points est inchangée.

Les définitions suivantes caractérisent certains processus ponctuels, ces propriétés seront utiles dans la suite pour l'étude des caractéristiques des patterns de points.

Définition 2.1.6. *Un processus ponctuel X sur $\mathbb{R}^d$ est stationnaire si sa distribution est invariante par translations : la distribution de $X + s = \{\xi + s;\ \xi \in X\}$ est la même que X pour tout $s \in \mathbb{R}^d$.*

Un processus de point est dit isotrope si sa distribution est invariante par rotations par rapport à l'origine $\mathcal{O}$ de $\mathbb{R}^d$: la loi de $RX = \{R\xi\ ;\ \xi \in X\}$ est la même que celle de X pour toute rotation autour de l'origine R de $\mathbb{R}^d$.

Exemple 2.1.1. *Le processus ponctuel de Poisson homogène d'intensité ρ Poisson$(\mathbb{R}^d, \rho)$ est stationnaire et isotrope.*

La proposition suivante permet de caractériser la distribution d'un processus de Poisson :

Proposition 2.1.3.
i) *$X \sim$ Poisson(W, ρ) si et seulement si pour tout $B \subseteq W$ avec $\mu(B) = \int_S \rho(\xi)d\xi < +\infty$ et $F \subseteq N^{lf}$,*

$$\mathbb{P}(X_B \in F) = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B ... \int_B \mathbb{1}[\{x_1, ..., x_n\} \in F] \prod_{i=1}^{n} \rho(x_i) dx_1 ... dx_n$$

où l'intégrale pour $n = 0$ est réduite à $\mathbb{1}[\emptyset \in F]$.
ii) *Si $X \sim$ Poisson(W, ρ), alors pour des fonctions $h : N^{lf} \to [0, \infty[$ et $B \subseteq W$ avec $\mu(B) < +\infty$,*

$$\mathbb{E}[h(X_B)] = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B ... \int_B h(\{x_1, ..., x_n\}) \prod_{i=1}^{n} \rho(x_i) dx_1 ... dx_n.$$

De cela, nous pouvons en déduire la «void probability» du processus de Poisson, pour $B \subseteq W$,

$$v(B) = \mathbb{P}(X_B = \emptyset) = \exp(-\mu(B)).$$

Les exemples d'applications et la complexité des données font néanmoins émerger la nécessité de créer des processus qui mettent en jeu des interactions entre les points. La section suivante traite de la construction de ces processus où le processus de Poisson standard joue un rôle de clé de voûte. Les nouveaux processus seront à densité par rapport à Poisson$(W, 1)$, la mesure de référence. Afin de construire ces processus plus complexes, nous allons nous appuyer sur les densités de probabilités des processus.

2.1.3 Densité de probabilité des processus ponctuels, intensité de Papangelou et stabilité

Énonçons tout d'abord des propriétés importantes pour la suite.

Si X_1 et X_2 sont deux processus définis sur le même espace W, alors X_1 est absolument continu par rapport à X_2 (plus précisément, la distribution de X_1 est absolument continue par rapport à celle de X_2) si et seulement si $(P(X_2 \in F) = 0) \implies (P(X_1 \in F) = 0)$ pour $F \in N^{\text{lf}}$. De manière équivalente, par le théorème de Radon-Nikodym :

Théorème 2.1.4. *(Radon-Nikodym).*
Soient μ_1 et μ_2 des mesures σ-finies de l'espace mesurable $(\Omega, \mathcal{F})$ avec μ_2 absolument continue par rapport à μ_1. Il existe une fonction mesurable $f : \Omega \to [0, +\infty[$, appelée dérivée de Radon-Nikodym, telle que pour tout $B \in \mathcal{F}$:

$$\mu_2(B) = \int_B f(\xi) d\mu_1(\xi).$$

De plus, pour deux densités f et g vérifiant ces conditions, $\mu(f \neq g) = 0$

Dans notre cas, l'espace mesuré considéré est $(N^{\text{lf}}, \mathcal{N}^{\text{lf}})$. Autrement dit, si X_1 est absolument continu par rapport à X_2, il existe une fonction $f : N^{\text{lf}} \to [0, +\infty[$ telle que

$$P(X_1 \in F) = \mathbb{E}[\mathbb{1}[X_2 \in F] f(X_2)], \quad F \subseteq N^{\text{lf}}$$

La proposition suivante permet de montrer que les processus de Poisson ne sont pas toujours absolument continus par rapport aux autres, mais qu'ils sont toujours absolument continus par rapport au processus de Poisson standard lorsque W est borné.

Proposition 2.1.5.
i) *Pour tout réel $\rho_1 > 0$ et $\rho_2 > 0$, Poisson$(\mathbb{R}^d, \rho_1)$ est absolument continu par rapport à Poisson$(\mathbb{R}^d, \rho_2)$ si et seulement si $\rho_1 = \rho_2$.*

ii) *Pour $i = 1, 2$, supposons que $\rho_i : W \to [0, +\infty[$ tels que $\mu_i(W) = \int_W \rho_i(\xi)d\xi) < +\infty$ et tels que $\rho_2(\xi) > 0$ dès lors que $\rho_1(\xi) > 0$. Alors Poisson(W, ρ_1) est absolument continu par rapport à Poisson(W, ρ_2), de densité*

$$f(\mathbf{x}) = \exp(\mu_2(W) - \mu_1(W)) \prod_{\xi \in \mathbf{x}} \rho_1(\xi)/\rho_2(\xi)$$

pour des configurations finies $\mathbf{x} \subset W$ (avec la convention $\frac{0}{0} = 0$)

Ceci donne alors l'idée d'exprimer les processus de points à partir du processus de Poisson standard. En effet, lorsque nous travaillons avec des processus à densité, il est possible d'exprimer cette densité par rapport à la mesure d'intensité du processus de Poisson standard à l'aide du théorème de Radon-Nikodym.
Considérons un processus ponctuel X sur $W \subseteq \mathbb{R}^d$ de densité f par rapport au processus de Poisson standard Poisson$(W, 1)$. Nous supposons ici que le volume de W est fini ($|W| < +\infty$), afin que cette densité soit correctement définie. La densité est donc concentrée sur l'ensemble des configurations de points finies de W, $N_{\mathrm{f}} = \{\mathbf{x} \subset W : n(\mathbf{x}) < +\infty\}$ et par la proposition 2.1.3, pour $F \subseteq N_{\mathrm{f}}$,

$$P(X \in F) = \sum_{n=0}^{\infty} \frac{\exp(-|W|)}{n!} \int_W ... \int_W \mathbb{1}[\{x_1, ..., x_n\} \in F] f(\{x_1, ...x_n\}) dx_1 ... dx_n$$

où, pour $n = 0$, nous avons $\exp(-|W|)\mathbb{1}[\emptyset \in F]f(\emptyset)$. Si $|W| = 0$ alors $P(X = \emptyset) = 1$. Dans la pratique, nous considérerons uniquement le cas $|W| > 0$.

Exemple 2.1.2. *La densité de $X \sim$ Poisson(W, ρ) avec $\mu(W) = \int_W \rho(\xi)d\xi < +\infty$, est donnée par*

$$f(\mathbf{x}) = \exp(|W| - \mu(W)) \prod_{\xi \in \mathbf{x}} \rho(\xi)$$

d'après la proposition 2.1.5.

Dans la majorité des cas, f est seulement connue proportionnellement par rapport à une fonction connue, excluant la constante de normalisation, impossible à calculer. Plus précisément, nous connaissons $h : N_{\mathrm{f}} \to [0, +\infty[$ telle que $f \propto h$. La constante de normalisation

$$c = \sum_{n=0}^{+\infty} \frac{\exp(-|W|)}{n!} \int_W ... \int_W h(\{x_1, ..., x_n\}) dx_1 ... dx_n$$

est inconnue pour les modèles que nous allons considérer dans la suite, à l'exception du processus de Poisson. En effet, l'intégrale est difficile à calculer de par la relation entre les points et la somme. Nous appellerons c la constante de normalisation ou la fonction de partition.

Ainsi, nous pouvons exprimer la densité d'un processus ponctuel par rapport au processus de Poisson standard. Ceci peut être également étendu au cas des processus ponctuels marqués. Soit X un processus ponctuel marqué sur $S \times M$ d'intensité ρ par rapport au processus de Poisson standard marqué Poisson$(S \times M, 1)$. Cette densité est également concentrée sur N_{f}. Ainsi, ce processus va d'abord répartir les points selon le processus de Poisson standard et associer une marque à chacun de ces points de manière indépendante selon une loi ν_M sur l'ensemble des

marques. Nous pouvons alors écrire comme précédemment pour $F \in N_{\mathrm{f}}$, basé sur la proposition 2.1.5 :

$$P(X \in F) = \sum_{n=0}^{\infty} \frac{\exp(-|W|)}{n!} \int_W ... \int_W \mathbb{1}[\{(x_1, w_1), ..., (x_n, w_n)\} \in F] \\ \times f(\{\{(x_1, w_1), ..., (x_n, w_n)\}\}) d\nu(w_1)...d\nu(w_n).$$

Nous allons désormais nous intéresser aux notions liées à la stabilité des processus ponctuels.

Définition 2.1.7. *(Intensité conditionnelle de Papangelou).*
L'intensité conditionnelle de Papangelou pour un processus ponctuel X de densité f est définie par

$$\lambda^*(\mathbf{x}, \xi) = f(\mathbf{x} \cup \xi)/f(\mathbf{x}), \quad \mathbf{x} \in N_f, \, \xi \in W \setminus \mathbf{x},$$

avec $a/0 = 0$ pour $a \geq 0$.

Remarques :
• L'intensité conditionnelle de Papangelou étant un ratio de densités, elle ne dépend plus de la constante de normalisation de f.
• X (ou f) est dit attractif si $\lambda^*(\mathbf{x}, \xi) \leq \lambda^*(\mathbf{y}, \xi)$ (resp. répulsif $\lambda^*(\mathbf{x}, \xi) \geq \lambda^*(\mathbf{y}, \xi)$) dès lors que $\mathbf{x} \subset \mathbf{y}$.

Exemple 2.1.3. *L'intensité conditionnelle de Papangelou pour le processus de Poisson(W, ρ) est donnée par $\lambda^*(\mathbf{x}, \xi) = \rho(\xi)$. Cette dernière ne dépend pas de $\mathbf{x}$, ce qui est cohérent avec la propriété d'indépendance du processus de Poisson.*

Heuristiquement, cette intensité conditionnelle va permettre de quantifier l'ajout d'un point à une réalisation d'un processus ponctuel X. Elle est donc étroitement liée à l'intégrabilité de la densité du processus. Dans la suite, il sera utile de considérer des processus qualifiés d'héréditaires :

Définition 2.1.8. *(Fonction héréditaire).*
Soit $h : N_f \to [0, +\infty[$, une fonction h est dite héréditaire si

$$h(\mathbf{x}) > 0 \implies h(\mathbf{y}) > 0 \text{ pour } \mathbf{y} \subset \mathbf{x}.$$

Nous dirons qu'un processus de point X est héréditaire lorsque la densité f de ce processus est héréditaire.

De cette définition, il vient :

Proposition 2.1.6. *Si X est héréditaire sur W de densité f par rapport à Poisson($W, 1$), alors*

$$f(\{\xi_1, ..., \xi_n\}) = f(\emptyset) \prod_{i=1}^{n} \lambda^*(\{\xi_1, ..., \xi_{i-1}\}, \xi_i)$$

et le produit ne dépend pas du label des points $\xi_1, ..., \xi_n$, $n \in \mathbb{N}^*$.

Autrement dit, il y a bijection entre l'intensité conditionnelle de Papangelou et la densité d'un processus ponctuel X dans le cas où X est héréditaire. L'étude de λ^* est donc directement liée à l'étude de la densité, la définition suivante et la proposition qui en découle permettent d'introduire des notions impliquant l'intégrabilité de la densité par rapport à Poisson$(W, 1)$.

Définition 2.1.9. *Soit $\phi^* : W \to [0, +\infty[$ une fonction telle que $c^* = \int_W \phi^*(\xi)d\xi$ est finie. Pour une fonction $h : N_f \to [0, +\infty[$ donnée, la stabilité locale (ou la ϕ^*-stabilité) veut dire que*

$$h(\mathbf{x} \cup \xi) \leq \phi^*(\xi)h(\mathbf{x}), \quad \forall \mathbf{x} \in N_f \text{ et } \xi \in W \setminus \mathbf{x}.$$

De plus, nous parlerons de stabilité de Ruelle ou de stabilité au sens de Ruelle lorsque $h(\mathbf{x}) \leq \alpha \prod_{\xi \in x} \phi^*(\xi)$ pour α une constante positive et tout $\mathbf{x} \in N_f$.

Proposition 2.1.7. *La stabilité au sens de Ruelle exprime que h est dominée par une densité de Poisson non normalisée. Cela implique l'intégrabilité de h par rapport à Poisson$(W, 1)$. La stabilité locale implique la stabilité au sens de Ruelle. Enfin, si $f \propto h$ est $\phi^* -$ stable, alors $\lambda^*(x, \xi) \leq \phi^*(\xi)$.*

De nombreux processus ponctuels vérifient cette condition de stabilité locale. Il existe également des exemples de processus vérifiant la stabilité au sens de Ruelle mais pas la stabilité locale (processus de Lennard-Jones). Nous verrons l'importance de la stabilité locale pour les algorithmes de simulation en section 2.2.

Nous avons donc vu, en considérant ces densités, qu'il existait des outils robustes pour montrer l'intégrabilité des processus. Dans la suite, nous allons introduire les processus de Markov. Ces processus possèdent une propriété de Markov spatiale : ceci a pour rôle de simplifier l'écriture de la densité des processus avec interactions.

2.1.4 Processus ponctuels de Markov

2.1.4.1 Généralités

Nous sommes désormais armés pour introduire de nouveaux processus ayant pour mesure de référence le processus de Poisson standard et nous assurer qu'ils soient correctement définis. Pour des raisons pédagogiques, nous allons considérer dans la suite une notion de voisinage entre les points induite par une distance sur $\mathbb{R}^d$. Soit alors $\sim$ une relation réflexive et symétrique sur $\mathbb{R}^d$ et définissons le voisinage et la frontière d'un ensemble $A \subseteq \mathbb{R}^d$ par $V_A = \bigcup_{\xi \in A} V_\xi = \{\xi \in W : \xi \sim \eta \quad \text{pour} \quad \eta \in A\}$ et $\partial_A = \{\xi \in \mathbb{R}^d \backslash A : \xi \sim a \quad \text{pour} \quad a \in A\} = V_A \backslash A$ comme illustré ci-dessous :

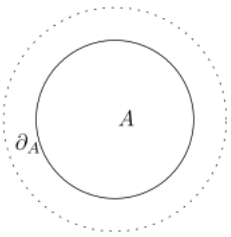


FIGURE 2.2 – Ensemble A et sa frontière ∂A.

Définition 2.1.10. *(Processus ponctuel de Markov).*
Un processus ponctuel défini par une densité de probabilité f par rapport au processus standard sur W est un processus ponctuel de Markov par rapport à $\sim$ si pour toute configuration de points $\mathbf{x} \in \mathbb{N}_f$ telle que $f(\mathbf{x}) > 0$,

i) *$f(\mathbf{y}) > 0$ pour tout $\mathbf{y} \subseteq \mathbf{x}$ (f est héréditaire)*

ii) *Pour tout $\xi \in W, \xi \notin \mathbf{x}$, $\lambda^*(\mathbf{x}, \xi)$ ne dépend que de ξ et de $\partial(\{\xi\}) \cap \mathbf{x}$.*

Cette propriété de Markov spatiale énoncée en **(ii)** sera utile pour décrire les densités de probabilités des processus ponctuels à l'aide des cliques pour une relation de voisinage $\sim$. Rappelons-en la définition :

Définition 2.1.11. *(Clique).*
Une configuration $\mathbf{x} \in N_f$ est une clique si tous les éléments de $\mathbf{x}$ sont voisins pour la relation $\sim$. Autrement dit, $\forall \xi_i, \xi_j \in \mathbf{x}$, $\xi_i \sim \xi_j$.

Exemple 2.1.4.
• *L'ensemble vide $\emptyset$ est une clique.*
• *Pour la relation de voisinage «être à distance plus petite qu'un rayon r», un ensemble de points $\mathbf{x}$ est une clique s'ils sont tous contenus dans une même boule de rayon r.*

Théorème 2.1.8. *(Hammersley-Clifford).*
Soit X un processus ponctuel fini sur $W \subseteq \mathbb{R}^d$ défini par une densité de probabilité par rapport au processus standard et soit $\sim$ une relation réflexive et symétrique sur W. Alors X est un processus de Markov par rapport à $\sim$ si et seulement si f peut être écrite par une décomposition en produit de cliques pour $\sim$:

$$f(\mathbf{x}) = \prod_{cliques\ \mathbf{y} \subseteq \mathbf{x}} \varphi(\mathbf{y})$$

où φ est une fonction mesurable positive définie sur N_f appelée fonction d'interaction.

Cette caractérisation sera utile dans la suite pour décrire les distributions conditionnelles des processus par rapport à une configuration déjà existante. Une autre forme pour écrire les densités markoviennes est la forme dite de Gibbs, originaire de la physique statistique.

Définition 2.1.12. *(Processus ponctuels de Gibbs).*
Soit X un processus ponctuel de densité $f : W \to \mathbb{R}^+$ par rapport au processus standard. Ce processus est appelé processus de Gibbs si f peut être mise sous la forme :

$$f(\mathbf{x}) = \frac{1}{Z} \exp[-U(\mathbf{x})]$$

avec $Z = \int_W \exp[-U(\mathbf{y})]\mu(d\mathbf{y})$ la constante de normalisation et U la fonction d'énergie.

Remarques :
• Un processus de Markov est toujours un processus de Gibbs. En effet, la densité peut se réécrire :

$$f(\mathbf{x}) = \frac{1}{Z} \exp\left[- \sum_{\substack{\text{cliques } \mathbf{y} \subseteq \mathbf{x} \\ y \neq \emptyset}} \log \varphi(\mathbf{y}) \right].$$

• La réciproque n'est pas vraie.

Exemple 2.1.5. *(Processus de Poisson).*
Ce processus ne dépendant pas d'une relation de voisinage, toutes les cliques non vides sont

alors d'ordre 1. La fonction d'interaction correspondante est donc $\varphi(\{\xi\}) = \rho(\xi)$. *La densité de Poisson*(W, ρ) *s'exprime alors sous la forme*

$$f(\mathbf{x}|\rho) \propto \prod_{\xi \in \mathbf{x}} \rho(\xi) = \exp\left[\sum_{\xi \in \mathbf{x}} \log(\rho(\xi))\right] = \exp\left[\sum_{i=1}^{n(\mathbf{x})} \log(\rho(\xi_i))\right] \tag{2.3}$$

où $n(\mathbf{x})$ *est le nombre de points de la configuration* $\mathbf{x}$. *Le processus de Poisson est donc bien un processus de Gibbs.*

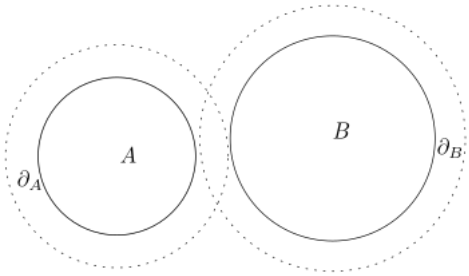
À l'aide de la propriété de Markov spatiale et de cette décomposition, il est aisé de construire la distribution conditionnelle d'un processus par rapport à une configuration de points existante dans un sous ensemble $B \subset W$. Nous avons :

Proposition 2.1.9. *Soit* X *un processus ponctuel de Markov de densité* f *et de fonction d'interaction* ϕ.

- *Si* $A, B \subset W$ *sont tels que* $A \cap v_B = \emptyset$, *alors* X_A *et* X_B *sont conditionnellement indépendants étant donné* X_C *où* $C = W \setminus (A \cup B)$
- *Pour* $B \subset W$, $X_B|X_{W\setminus B} \sim X_B|X_{\partial B}$ *et conditionnellement à* $X_{\partial B} = \mathbf{x}_{\partial B}$ *avec* $f(\mathbf{x}_{\partial B}) > 0$, *le processus ponctuel* X_B *est un processus de Markov de densité par rapport au processus de Poisson*$(B, 1)$ *:*

$$f_B(\mathbf{x}_B|\mathbf{x}_{\partial B}) \propto \prod_{\mathbf{y} \subseteq \mathbf{x}_B \cup \mathbf{x}_{\partial B} : \mathbf{y}_B \neq \emptyset} \phi(\mathbf{y}) \tag{2.4}$$

Le produit sera égal à 1 si $\mathbf{x}_B = \emptyset$ *et l'indiçage du produit porte sur toutes les sous-configurations de points de* $\mathbf{x}_B \cup \mathbf{x}_{\partial B}$ *tel que la restriction de* $\mathbf{y}$ *à* B *soit non vide.*



Cette proposition sera utile lorsque nous considérerons l'inférence en données partiellement observées dans la suite du manuscrit.

2.1.4.2 Exemples de processus de Gibbs

Dans cette sous-section, nous allons donner des exemples de processus ponctuels de Gibbs utilisés dans la suite pour la modélisation des données en astrophysique et illustrerons comment «superposer» plusieurs de ces modèles. Pour des raisons pédagogiques, nous allons considérer les fonctions d'énergies sous forme du résultat du produit scalaire entre les paramètres et les statistiques suffisantes du modèle $\langle t(\mathbf{x})|\theta\rangle$, où $t(\mathbf{x})$ représente le vecteur des statistiques suffisantes du

modèle et θ le vecteur des paramètres.

Le processus ponctuel de Strauss [Strauss, 1975, Kelly and Ripley, 1976] est un premier modèle avec interaction qui pénalise la probabilité d'avoir deux points à une distance inférieure à un rayon fixe r, l'interaction est basée sur la distance entre les points.

Définition 2.1.13. *(Processus de Strauss homogène).*
La densité du processus de Strauss homogène par rapport au processus de Poisson standard est donnée par

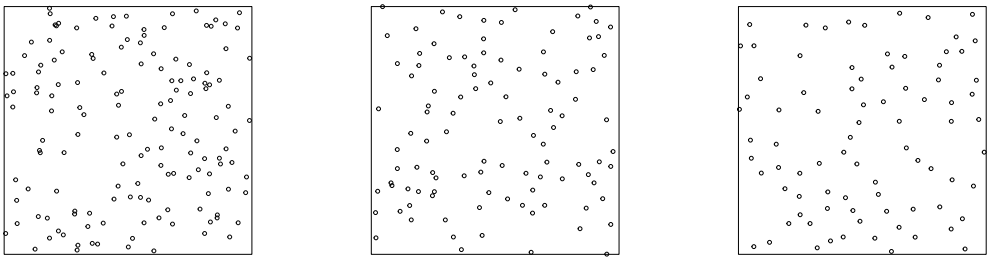
$$f(\mathbf{x}|\rho,\gamma_s) \propto \exp(n(\mathbf{x})\log(\rho) + s_r(\mathbf{x})\log(\gamma_s)) \tag{2.5}$$

où $n(\mathbf{x})$ représente le nombre de points et $s_r(\mathbf{x})$ le nombre de paires de points à une distance inférieure à r. ρ représente toujours l'intensité et $\gamma_s \in]0,1]$ est la force d'interaction. Dans ce modèle, $n(\mathbf{x})$ et $s_r(\mathbf{x})$ sont les statistiques suffisantes. Si $\gamma_s = 1$, le modèle redevient le processus de Poisson d'intensité ρ.

Ce processus est localement stable : pour tout $(\mathbf{x},\xi) \in N_{\mathrm{f}} \times W$

$$\lambda^*(\mathbf{x},\xi) = \exp[\log(\rho) + \log(\gamma_s)(s_r(\mathbf{x}\cup\xi) - s_r(x))] \leq \rho$$

car $\gamma_s \leq 1$ et $s_r(\mathbf{x}\cup\xi) - s_r(x) \geq 0$. Pour $\gamma_s < 1$ les configurations avec des points proches les uns des autres reçoivent une faible probabilité. Plus γ_s est proche de 0, plus la répulsion est forte comme l'illustrent les trois figures suivantes : Les réalisations exhibent un nombre de $r-$voisins de moins en moins importants lorsque γ_s se rapproche de 0. La figure 2.3 illustre trois configurations obtenues pour 3 valeurs de γ_s.



(a) $\gamma_s = 0.75$ (b) $\gamma_s = 0.2$ (c) $\gamma_s = 0.005$

FIGURE 2.3 – Réalisations du modèle de Strauss pour différentes valeurs de γ_s ; $\rho = 200$ et $r = 0.05$ dans $W = [0,1]\times[0,1]$.

Le processus Area Interaction est un modèle à interaction prenant en compte l'aire occupée par les boules de rayon R centrées en les points. Les interactions sont basées sur le «territoire» occupé par les points.

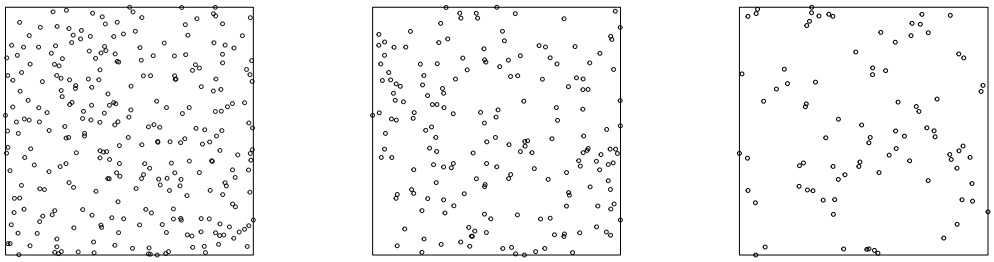
Définition 2.1.14. *(Area Interaction homogène).*
Dans le cas homogène, la densité du processus Area Interaction est la suivante

$$f(\mathbf{x}|\rho,\gamma_a) \propto \exp(n(\mathbf{x})\log(\rho) + a_R(\mathbf{x})\log(\gamma_a)) \tag{2.6}$$

où $a_R(\mathbf{x}) = -|\cup_{\xi\in\mathbf{x}} b(\xi,R)|$ représente le d-volume de la réunion des boules de rayon R centrées en les points de la configuration. $\gamma_a \geq 0$ est le paramètre du modèle. Pour ce modèle, $n(\mathbf{x})$ et

$a_R(\mathbf{x})$ *sont les statistiques suffisantes. Une fois de plus, si* $\gamma_a = 1$*, le modèle devient le processus de Poisson d'intensité* ρ*.*

Ce modèle est localement stable pour toute valeur de γ_a ($\lambda^*(\mathbf{x}, \xi) < \rho$ si $\gamma_a \geq 1$ et $\lambda^*(\mathbf{x}, \xi) < \rho e^{-R^d \times |b(0,R)|}$ sinon). Il est répulsif pour des valeurs de γ_a inférieures à 1 et agglomérant lorsque γ_a est plus grand que 1. Plus précisément, lorsque $\gamma_a < 1$, les points ont tendance à occuper l'espace disponible et exhibent donc une forme de régularité. À l'inverse, lorsque $\gamma_a > 1$, le modèle cherche à minimiser l'espace occupé par les points. La figure 2.4 ci-dessous illustrent ce phénomène :



(a) $\gamma_a = 0.005$ (b) $\gamma_a = 1$ (c) $\gamma_a = 5$

FIGURE 2.4 – Réalisations du modèle Area Interaction pour différentes valeurs de γ_a ; $\rho = 200$ et $R = 0.05$ dans $W = [0, 1] \times [0, 1]$.

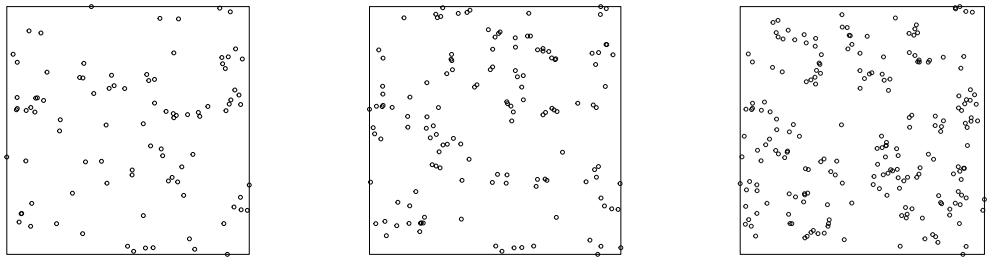
Un autre exemple de construction de modèle peut se faire à partir de statistiques déjà existantes. [Geyer, 1999] introduit une variante du modèle de Strauss en modifiant sa statistique suffisante afin d'obtenir un modèle permettant l'agglomération en contrôlant le nombre possible d'interactions pour un point.

Définition 2.1.15. *(Processus de saturation de Geyer homogène).*
La densité du processus de ce processus est donnée par :

$$f(\mathbf{x}|\rho, \gamma_g) \propto \exp(n(\mathbf{x}) \log(\rho) + \sum_{\xi \in \mathbf{x}} \min(s_r(\xi), s) \log(\gamma_g)) \tag{2.7}$$

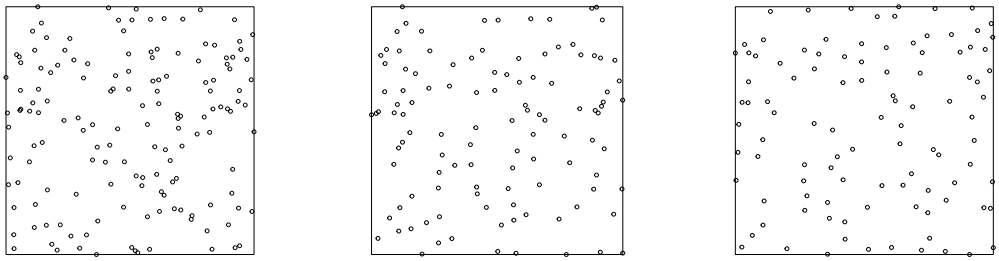
où $s_r(\mathbf{x})$ *représente le nombre* $r-$*voisins,* s *représente un seuil de saturation des paires connectées et* $\gamma_g > 0$ *est la force d'interaction. Dans ce modèle,* $n(\mathbf{x})$ *et* $\sum_{\xi \in \mathbf{x}} \min(s_r(\xi), s)$ *sont les statistiques suffisantes. À nouveau, si* $\gamma_g = 1$*, le modèle se résume au processus de Poisson d'intensité* ρ*.*

Le modèle est localement stable dans les deux cas ($\lambda^*(\mathbf{x}, \xi) \leq \rho \exp(\log(\gamma_g) \times s)$ si $\gamma_g > 1$ et $\lambda^*(\mathbf{x}, \xi) \leq \rho$ si $\gamma_g < 1$). Si $\gamma_g > 1$, les points auront tendance à se regrouper, si $\gamma_g < 1$, la tendance est inverse. Le seuil de saturation contrôle quant à lui le nombre de r-voisins, il nivelle par le haut le cas $\gamma_g > 1$ et par le bas le cas $\gamma_g < 1$. Les figures 2.5 et 2.6 montrent l'influence du seuil de saturation dans les deux cas et montrent des réalisations où nous pouvons observer les effets des différents types d'interactions.



(a) $s = 2$ (b) $s = 3$ (c) $s = 4$

FIGURE 2.5 – Réalisations du processus de saturation de Geyer pour différentes valeurs de s ;$\gamma_g = 1.5$; $\rho = 55$ et $r = 0.05$ dans $W = [0, 1] \times [0, 1]$. Nous remarquons bien une tendance au clustering dans les trois cas, le nombre de points agglomérés augmente bien lorsque le seuil s augmente.



(a) $s = 0.25$ (b) $s = 1$ (c) $s = 3$

FIGURE 2.6 – Réalisations du processus de saturation de Geyer pour différentes valeurs de s ;$\gamma_g = 0.5$; $\rho = 200$ et $r = 0.05$ dans $W = [0, 1] \times [0, 1]$. Nous remarquons bien une tendance répulsive dans les trois cas, le nombre de points agglomérés varie bien quand le seuil varie.

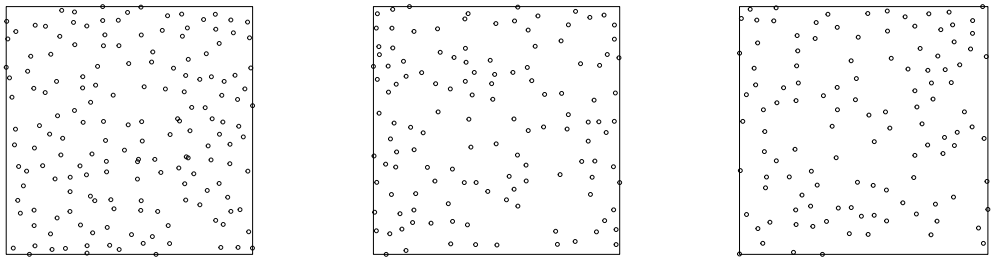
Une dernière construction de modèle s'obtient en superposant plusieurs modèles déjà existants. Ainsi, nous pouvons par exemple créer un modèle en superposant les modèles de Strauss et Area Interaction.

Définition 2.1.16. *(Superposition Strauss - Area Interaction).*
La densité qui résulte de cette construction est la suivante :

$$f(\mathbf{x}|\rho, \gamma_s, \gamma_a) \propto \exp(n(\mathbf{x}) \log(\rho) + s_r(\mathbf{x}) \log(\gamma_s) + a_R(\mathbf{x}) \log(\gamma_a)) \tag{2.8}$$

où les paramètres et les statistiques suffisantes sont les mêmes que pour les modèles précédents.

Ce modèle a été utilisé par [Stoica et al., 2007, Tempel et al., 2018], et permet à la fois de générer des patterns de points exhibant de la répulsion tout en contrôlant la surface occupée par les points.



(a) $\gamma_s = 0.5$ (b) $\gamma_s = 0.2$ (c) $\gamma_s = 0.1$

FIGURE 2.7 – Réalisations du modèle Strauss - Area Interaction pour différentes valeurs de γ_s ;$\gamma_a = 0.005$; $\rho = 200$ et $r = R = 0.05$ dans $W = [0,1] \times [0,1]$. Nous remarquons que plus le paramètre γ_s se rapproche de 0, moins le nombre de $r-$voisins est important, l'aire occupée par les boules attachées au point semble quasi-similaire dans les trois cas.

Nous allons maintenant présenter quelques statistiques descriptives spatiales. Ces fonctionnelles découlent de l'étude des moments des mesures de comptage (mesures moments, mesures moments factoriels, mesure de Palm). Les théorèmes de Campbell-Mecke et Georgii-Nguyen-Zessin, utilisés sous des hypothèses précises (stationnarité ou non, isotropie), en considérant les boréliens B particuliers (boules, cercles, segments) et en prenant les indicatrices de ces ensembles, permettent la création de ces statistiques descriptives [van Lieshout, 2000, Møller and Waagepetersen, 2004] [Stoica, 2025].

2.1.5 Statistiques descriptives spatiales

Parmi toutes les fonctions les plus connues, nous avons choisi de présenter les fonctions F (fonction d'espace vide), G (fonction de la distance au plus proche voisin), K (fonction K de Ripley) et g (fonction de corrélation par pair). Bien qu'en général, ces fonctions ne soient pas connues sous forme analytique, elles peuvent être estimées ([Baddeley et al., 2015]). Les valeurs théoriques sont connues pour le processus de Poisson et nous permettent de comparer les caractéristiques d'une réalisation d'un processus observé avec celles d'un processus de Poisson. Elles sont aussi utilisées pour l'inférence paramétrique et non paramétrique, associées aux patterns de points observés. Soit X un processus ponctuel stationnaire et isotrope, et $d(\xi, X) = \min\{||\xi - x_i||, x_i \in X\}$ la distance entre un point ξ dans $\mathbb{R}^2$ et le processus ponctuel X.

Définition 2.1.17. *(Fonction F).*
La fonction d'espace vide F est définie par :

$$F(r) = \mathbb{P}(d(\xi, X) \leq r) \text{ pour } r > 0.$$

Il s'agit de la fonction de répartition de cette distance, appelée «distance d'espace vide» ou «distribution sphérique de contact».

Définition 2.1.18. *(Fonction G).*
La fonction d'espace inter-points G est définie par

$$G(r) = \mathbb{P}(d(\xi, X \setminus \xi) \leq r | X \text{ a un point en } \xi) \text{ pour } r > 0.$$

Il s'agit de la fonction de répartition de la distance au plus proche voisin d'un point dans X.
Remarques :
• Pour un processus de Poisson homogène sur $\mathbb{R}^2$ d'intensité ρ, pour $r \geq 0$,

$$G_{pois}(r) = F_{pois} = 1 - \exp(-\rho r^2 \pi).$$

• Pour des configurations de points répulsives (resp. agrégées), pour $r \geq 0$

$$G(r) < 1 - \exp(-\rho r^2 \pi) < F(r) \text{ (resp. } F(r) < 1 - \exp(-\rho r^2 \pi) < G(r) \text{).}$$

• Ces deux fonctions sont adéquates pour analyser ce qui se passe à petite échelle (les estimateurs sont cumulatifs pour les évènements les plus proches) mais ne permettent en général pas de conclure sur la nature de phénomènes à plus grande échelle.

Définition 2.1.19. *(Fonction K de Ripley).*
Elle est définie par

$$K(r) = \frac{1}{\rho} \mathbb{E}[\text{nombre de } r\text{-voisins de } \xi | X \text{ a un point en } \xi], \text{ pour } r > 0.$$

À partir de la dérivée de cette fonction, nous pouvons alors définir

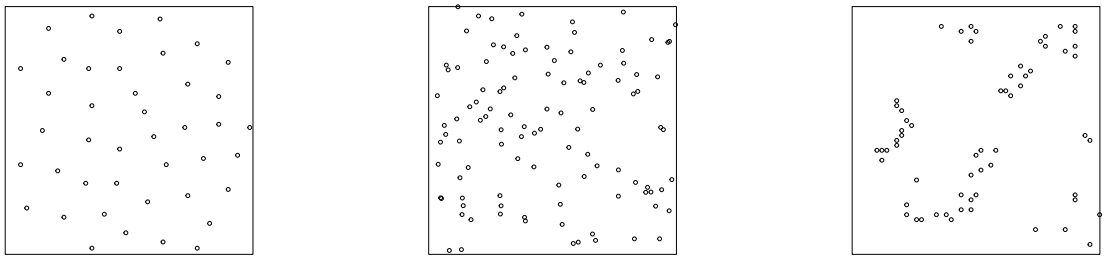
Définition 2.1.20. *(Fonction g).*
La «pair correlation function» ou fonction de corrélation par paires de points, est définie par

$$g(r) = \frac{K'(r)}{2\pi r} \text{ pour } r > 0$$

où $K'(r)$ est la dérivée de K par rapport à r.

Remarques :
• La valeur théorique pour le processus de Poisson est $g(r) = 1$.
• Une valeur $g(r) < 1$ indique que les distances entre les points égales à r sont moins fréquentes que ce qui est attendu pour un processus sans interactions. Ceci indique donc de la répulsion. À l'inverse, $g(r) > 1$ indique du clustering.
• La fonction g donne des informations sur les distances les plus et moins fréquentes entre deux points de la configuration ([Stoyan and Stoyan, 1994]).

Exemple 2.1.6. *(Illustration sur des données).*
Prenons trois types de données pour illustrer le fonctionnement de ces fonctions. Les jeux de données Cells et Redwood ont été tirés de la librairie `R spatstat` *[Baddeley et al., 2015].*

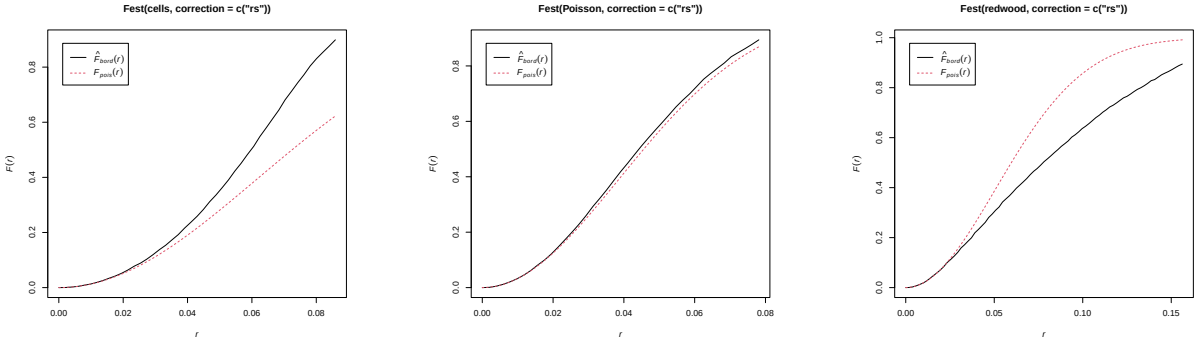


(a) Dataset Cells

(b) Réalisation d'un processus de Poisson d'intensité 100

(c) Dataset Redwood

FIGURE 2.8 – Représentations des données Cells (gauche) ; Processus de Poisson d'intensité 100 (milieu) et redwood (droite). Visuellement, nous remarquons une tendance de régularité entre les points de la première figure. Pour ce qui est du cas poissonnien, nous savons en théorie que cela représente un cas de «complete spatial randomness» : aucune tendance ni de répulsion ni de clustering n'apparaît sur la répartition des points. Enfin, une tendance de clustering s'observe sur la dataset Redwood.



(a) Dataset Cells

(b) Réalisation d'un processus de Poisson d'intensité 100

(c) Dataset Redwood

FIGURE 2.9 – En rouge : Courbe théorique poissonienne ; En noir : l'estimation de la fonction F : F_{est} pour les données Cells (gauche) ; Processus de Poisson d'intensité 100 (milieu) et redwood (droite). Figure de gauche : la courbe empirique (noire), exprime une tendance à la répulsion : la probabilité de contenir un point dans une boule de taille r est plus grande que dans le cas poissonnien. À l'inverse, sur la figure de droite, cette probabilité est plus faible : cela traduit une tendance de clustering entre les points.

Pour les fonctions G et g, la tendance est similaire à ce qui s'observe pour F.

Ces figures illustrent la différence entre la courbe analytique théorique et l'estimation des fonctions sur ces données. Afin de mener une analyse plus robuste, des enveloppes peuvent être obtenues par simulations (voir section 2.2.3.2 dans la suite de ce manuscrit) ou des tests d'hypothèses peuvent être conduits (test d'enveloppe global [Myllymäki et al., 2016]). Ces fonctions vont également jouer un rôle important dans le contrôle de l'estimation de paramètres. Néanmoins, ces statistiques ne permettent pas de caractériser les processus (e.g. comme la variance

ou la moyenne pour des processus non Gaussiens. Ou voir [Bedford and Berg, 1997] pour un contre-exemple).

Jusqu'ici, nous avons montré des réalisations de ces différents modèles sans mentionner comment ces configurations étaient obtenues. La section suivante s'intéresse à la simulation de ces processus par chaînes de Markov.

2.2 Simulation par chaîne de Markov et applications

Dans cette partie, nous rappelons des notions essentielles en lien avec les chaînes de Markov et leurs propriétés liées à leurs convergences. Nous avons suivi [Meyn and Tweedie, 1993][Geyer, 1999], [van Lieshout, 2000],[Robert and Casella, 2000],[Møller and Waagepetersen, 2004] et [Stoica, 2025]. Nous détaillons également différentes opérations sur les noyaux de transition. Pour finir, nous nous intéresserons plus en détail à l'algorithme de Metropolis-Hasting et à des applications se basant sur la simulation des modèles.

2.2.1 Définitions et propriétés des chaînes de Markov

Comme nous l'avons vu, les processus ponctuels de Gibbs sont des modèles à densités par rapport à la mesure Poissonienne. Il s'agit alors de générer une chaîne de Markov à espace d'états général. Les propriétés classiques des chaînes de Markov à espace d'états finis doivent ainsi être adaptées. Les définitions/propriétés qui suivent peuvent être vues comme des «analogues» aux propriétés connues pour les chaînes de Markov à espace d'états finis. Ces définitions s'inscrivent dans un cadre général qui n'est pas uniquement propre aux processus ponctuels. Nous commençons par rappeler les définitions de base. Soit alors un espace probabilisé $(\Omega, \mathcal{F}, \mu)$.

Définition 2.2.1. *(Noyau de transition).*
Nous appelons noyau de transition une fonction $P : \Omega \times \mathcal{F} \to [0,1]$ *vérifiant :*

i) $\forall x \in \Omega$, $P(x, \cdot)$ *est une mesure de probabilité.*

ii) $\forall A \in \mathcal{F}$, $P(\cdot, A)$ *est mesurable.*

Remarques :
• Dans le cas où Ω est discret, le noyau de transition n'est autre que la matrice de transition donnée par les probabilités de transition d'un état i à un état j.
• Si l'espace Ω est continu, le noyau de transition sera donné par

$$\mathbb{P}(X \in A|x) = \int_A P(x, \mathrm{d}x')$$

pour caractériser un pas de l'état x vers $A \in \mathcal{F}$.

Définition 2.2.2. *(Chaîne de Markov).*
Soit P *un noyau de transition. Une suite* $(X_n)_{n\in\mathbb{N}}$ *définie sur* $(\Omega, \mathcal{F}, \mu)$ *est une chaîne de Markov (CDM) si à tout instant* k*, la loi conditionnelle de* X_k *sachant les états précédents,* $x_{k-1}, x_{k-2}, ..., x_0$ *est égale à la loi conditionnelle de* X_k *sachant* x_{k-1} *:*

$$\mathbb{P}(X_k \in A|x_0, x_1, ..., x_{k-1}) = \mathbb{P}(X_k \in A|x_{k-1}) = \int_A P(x_{k-1}, dx)$$

Remarques :
Une chaîne de Markov sera dite :
• *homogène* en temps lorsque les probabilités de transitions ne dépendent pas de n.
• *réversible* par rapport à une mesure π lorsque son noyau de transition vérifie l'égalité suivante :

$$\int_{A_1} \pi(\mathrm{d}x)P(x, A_2) = \int_{A_2} \pi(\mathrm{d}x)P(x, A_1)$$

pour tout $A_1, A_2 \in \mathcal{F}$.
• *irréductible* s'il est possible de passer d'un état $x_1 \in \Omega$ à un état $x_2 \in \Omega$ en un nombre fini d'itérations.

Définition 2.2.3. *(Stationnarité).*
Une mesure de probabilité π est stationnaire ou invariante pour le noyau de transition P si $\pi P = \pi$ où la multiplication à droite par une mesure est définie par :

$$(\nu P)(A) = \int \nu(dx) P(x, A)$$

avec $A \in \mathcal{F}$ et x un état fixé.

Si cette distribution est unique, nous parlons alors de distribution d'équilibre.

Deux opérations élémentaires peuvent s'effectuer sur noyaux de transitions. Vus comme des matrices, il s'agit de la multiplication entre noyaux, que nous appellerons «composition» de noyaux, et de la combinaison linéaire de noyaux, appelée ici «mixing».

Définition 2.2.4. *(Composition).*
Soient P_1 et P_2 deux noyaux de transition. Si une mise à jour décrite par un noyau de transition P_1 est directement suivie par une autre mise à jour de noyau de transition P_2, alors la mise à jour totale peut être décrite par le noyau de transition P_1P_2 défini part

$$(P_1P_2)(x, A) = \int P_1(x, dy) P_2(y, A)$$

Nous pouvons généraliser cette notion à plus de deux noyaux.

Définition 2.2.5. *(Mixing).*
Soient $P_1, ..., P_n$ des noyaux de transitions et $p_1, \ldots, p_n$ tels que $\sum_{i=1}^{n} p_i = 1$ et $p_i > 0 \; \forall i \in \{1, ..., n\}$. Une manière de combiner les mises à jour est alors de choisir de manière aléatoire un noyau P_i avec probabilité p_i. Le noyau de transition est alors donné par $P = \sum_i p_i P_i$.

Remarques :
• Pour la composition, si π est invariante pour P_1 et P_2, il vient directement que π est stationnaire pour $P = P_1P_2$.
• De même, pour le mixing, si π est invariante pour P_i pour tout $i \in \{1, ..., n\}$, alors π est invariante pour $P = \sum_i p_i P_i$.
La notion de stationnarité est une propriété de base requise pour un échantillonneur reposant sur les méthodes MCMC. L'idée va donc être de simuler des chaînes de Markov qui vont admettre la distribution cible comme distribution stationnaire. Lorsque l'espace d'états est fini, les propriétés énoncées ci-dessus suffisent à garantir l'existence d'une loi stationnaire et la convergence vers cette dernière. Néanmoins, dans le cas où Ω a une structure plus générale, il est nécessaire d'introduire des notions plus fortes. En effet, la notion de chaîne irréductible ne fait plus sens étant donné que les probabilités de transitions d'un état fixé à un autre sont nulles et la convergence n'est alors pas garantie avec ces seules notions. Rappelons ainsi les définitions suivantes :

Définition 2.2.6. *(ϕ-irréductibilité).*
Soit ϕ une mesure non nulle sur $(\Omega, \mathcal{F})$. Une CDM $(X_n)_{n\geq 0}$ est dite ϕ-irréductible si pour tout $x \in \Omega$ et tout $A \in \mathcal{F}$ tel que $\phi(A) > 0$,

$$\mathbb{P}(\tau_A < +\infty | X_0 = x)$$

où τ_A est le temps de retour de la chaîne en $A \in \mathcal{F}$ défini par $\tau_A = \min\{n \geq 1 : X_n \in A\} \leq +\infty$.

Définition 2.2.7. *(Apériodicité).*
Soit $P(\cdot,\cdot)$ le noyau de transition d'une chaîne de Markov admettant une mesure invariante π. Supposons qu'il existe un ensemble $A \in \mathcal{F}$ et une mesure de probabilité ν tel que $\nu(A) = 1$, $\epsilon > 0$ et $n_0 \in \mathbb{N}$ tels que

$$P^{n_0}(x,\cdot) \geq \epsilon\nu(\cdot) \ \forall x \in A.$$

La CDM associée à P est alors dite apériodique si

$$pgcd\{m; \exists \epsilon_m > 0 | P^m(x,\cdot) \geq \epsilon_m \nu(\cdot)\} = 1.$$

Si une CDM vérifie ces deux définitions et admet une distribution invariante π, cette distribution est alors l'unique distribution d'équilibre de la chaîne.

Théorème 2.2.1. *Soit (X_n) une CDM π-irréductible et apériodique, où π désigne la mesure stationnaire de la chaîne. Alors il existe un sous ensemble Ω' de Ω tel que $\pi(\Omega') = 1$ vérifiant :*

$$\lim_{n\to+\infty} \sup_{A\in\mathcal{F}} |P^n(x,A) - \pi(A)| = 0 \quad \forall x \in \Omega'.$$

Ce théorème permet ainsi de garantir la convergence de la CDM lorsqu'un noyau de transition vérifie les hypothèses du théorème.

Nous introduisons désormais une condition plus forte que la ϕ-irréductibilité : la récurrence au sens de Harris. Cette notion permettra de garantir la convergence des CDM tout en se débarrassant du cas pathologique du complémentaire de Ω' dans Ω (bien que de mesure nulle, cet ensemble pourra intervenir en pratique et ainsi empêcher la convergence souhaitée.).

Définition 2.2.8. *(Harris-récurrence)*
Une chaîne de Markov est dite Harris-récurrente si elle est ϕ-irréductible et si elle vérifie la propriété suivante : $\forall x \in \Omega$, $\forall A \subseteq \mathcal{F}$ tel que $\phi(A) > 0$,

$$\mathbb{P}(X_m \in A \quad \text{pour un certain m} \quad | X_0 = x) = 1.$$

Introduisons désormais des notions permettant de quantifier la convergence : la norme en variation totale et l'ergodicité.

Définition 2.2.9. *(Norme en variation totale)*
Soient μ et ν deux mesures de probabilités. La norme en variation totale est définie par

$$||\mu - \nu||_{TV} = \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|.$$

Nous remarquons que cette norme vérifie $||\mu-\nu||_{TV} \leq 1$, $||\mu-\nu||_{TV} = 0$ si $\mu = \nu$ et $||\mu-\nu||_{TV} = 1$ si les supports des mesures sont disjoints. Ceci donne une notion de convergence qui implique la convergence en loi.

Proposition 2.2.2. *Supposons que la chaîne de Markov considérée possède une distribution stationnaire π. Alors*

1. *Indépendamment de la distribution initiale, $||P^m - \pi||_{TV}$ est décroissante en m.*
2. *Si la CDM est irréductible et apériodique, il existe $A \in \mathcal{F}$ tel que $\pi(A) = 0$ et $\forall x \notin A$,*

$$\lim_{m\to+\infty} ||P^m(x,\cdot) - \pi||_{TV} = 0.$$

3. *La CDM est Harris-récurrente et apériodique si et seulement si* $\lim_{m\to+\infty} ||P^m - \pi||_{TV} = 0$ *indépendamment de la distribution initiale.*

Nous dirons alors qu'une CDM est ergodique si elle est apériodique et Harris-récurrente. L'ergodicité implique que la limite des noyaux de transitions itérés $\lim_{m\to+\infty} P^m(x, F) = \pi(F)$, $\forall A \in \mathcal{F}$ quel que soit l'état initial x ([Meyn and Tweedie, 1993])

Définition 2.2.10. *(Ergodicité géométrique et uniforme).*
Soit $(X_n)_{n\geq 0}$ *une chaîne ergodique de distribution d'équilibre* π*. Elle sera dite géométriquement ergodique s'il existe une fonction* $M : \Omega \to \mathbb{R}^+$ *et* $\rho \in]0, 1[$ *telles que,* $\forall x \in \Omega$*,*

$$||P^n(x, \cdot) - \pi||_{TV} \leq M(x)\rho^n.$$

Nous parlerons d'ergodicité uniforme lorsqu'il existe $M > 0$ *et* $\rho \in]0, 1[$ *telles que*

$$\sup_{x\in\Omega} ||P^n(x, \cdot) - \pi||_{TV} \leq M\rho^n.$$

Ces deux notions permettent alors de quantifier la vitesse de convergence de la chaîne de Markov vers la distribution cible.

2.2.2 Algorithmes de Metropolis-Hastings

2.2.2.1 Algorithme général

Afin de comprendre l'algorithme utile pour la simulation des processus ponctuels, nous souhaitons illustrer le fonctionnement de l'algorithme de Metropolis-Hastings «classique» à travers un exemple simple pour simuler une simple loi normale. Notons à nouveau π la distribution cible , l'algorithme général est le suivant :

Algorithme 2.2.1. *(Algorithme de Metropolis-Hastings général)*

Initialisation : *Fixer x_0 l'état initial et $i = 1$.*

1) *À l'itération i :*
a) Simuler $y \sim q(y|x_{i-1})$, où $q(y|x_{i-1})$ est appelée loi de proposition.
b) Calculer la probabilité

$$\alpha = \min\left\{1, \frac{\pi(y)}{\pi(x_{i-1})} \times \frac{q(x_{i-1}|y)}{q(y|x_{i-1})}\right\}$$

c) Simuler $U \sim \mathbb{U}([0,1])$:
- *Si $U \leq \alpha$, alors $x_i = y$*
- *Sinon, $x_i = x_{i-1}$*

2) *$i = i + 1$ et répéter en partant de* ***1)*** *si nécessaire.*

3) *Retourner $x_0, x_1, x_2, ...$*

Remarques :
• Le ratio d'acceptation α ne fait pas intervenir le calcul de la constante de normalisation, comme annoncé en introduction.
• La loi de proposition q est choisie de sorte que la proposition des nouvelles valeurs soit simple à simuler.
• Les conditions sur q afin que la chaîne de Markov simulée ait les bonnes propriétés sont faciles à remplir.

L'exemple ci-dessous illustre la simplicité du choix de la loi de proposition et le fait de ne plus dépendre de la constante de normalisation du modèle. Nous prenons dans cet exemple une loi de proposition uniforme avec une condition initiale très éloignée de la moyenne de la loi à échantillonner pour montrer le peu de sensibilité par rapport à cette dernière. Des proposales bien choisies influencent les propriétés de convergence de la chaîne.

Exemple 2.2.1. *(Loi Normale centrée réduite).*
Pour simuler un échantillon gaussien $\mathcal{N}(0,1)$, choisir de proposer des nouvelles valeurs suivant la loi uniforme sur un intervalle symétrique centré en x_{i-1} d'amplitude 0.5 (valeur arbitraire) fonctionne. À l'itération i, le calcul du ratio est alors une simple évaluation du ratio $\frac{\exp(-y^2/2)}{\exp(-x_{i-1}^2/2)}$. Comprendre la dynamique de la chaîne ainsi créée est alors aisé, nous accepterons plus facilement la valeur proposée y si elle est proche de 0. L'histogramme suivant est l'histogramme de l'échantillon $(x_0 = 10, x_1, ..., x_{10000})$ des 10000 premiers pas de la chaîne avec la série temporelle

des valeurs prises par la chaîne, ces derniers illustrent la dynamique attendue.

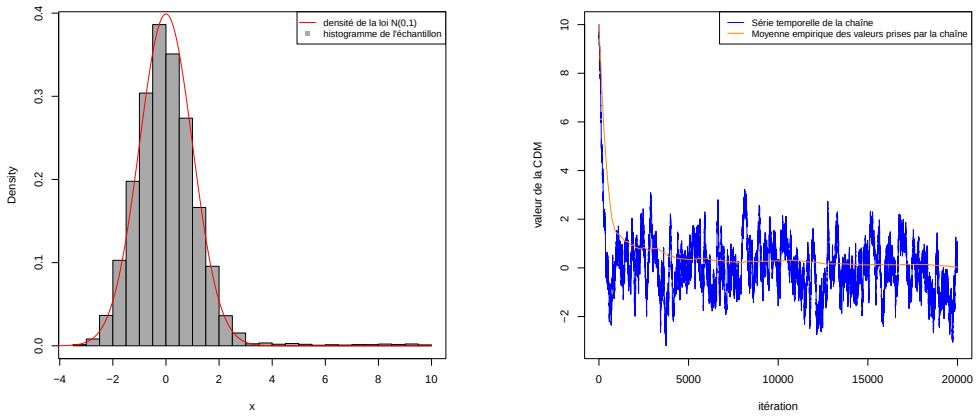


FIGURE 2.10 – Figure de gauche : Histogramme des valeurs prises par la CDM (gris) et densité de la loi $\mathcal{N}(0,1)$ (rouge). Figure de droite : Série temporelle des valeurs prises par la CDM (bleu) et moyenne empirique de ces valeurs (orange).

Nous remarquons à première vue que la loi normale centrée réduite semble correctement échantillonnée, en effet, la densité de la loi $\mathcal{N}(0,1)$ semble épouser correctement l'histogramme des valeurs. Quant à la convergence de la chaîne, nous observons que, malgré une valeur initiale $(x_0 = 10)$ assez éloignée de la moyenne de la loi, les valeurs semblent converger autour de 0 tout comme la moyenne empirique de l'échantillon obtenu.

Évidemment, ces illustrations ne constituent pas une preuve de la convergence ni du bon échantillonnage de la loi cible (un test statistique s'imposerait, e.g. Shapiro-Wilk), mais seulement une indication que notre algorithme semble correct. De plus, nous ne nous sommes pas posé la question de l'indépendance entre toutes ces valeurs obtenues, nous y reviendrons dans la suite. Nous présentons dans la section suivante un algorithme MH pour la simulation des processus ponctuels.

2.2.2.2 Algorithme MH pour la simulation des processus ponctuels

Reprenons les notations utilisées dans la partie modélisation 2.1, f désigne la densité d'un processus de Gibbs quelconque sur un compact $W \subset \mathbb{R}^d$ et sera la loi cible à échantillonner.

Comme nous l'avons vu, le principe des algorithmes de Metropolis-Hastings est d'abord de proposer une nouvelle valeur pour la chaîne de Markov puis de choisir d'accepter ou non ce nouvel état. Dans le cas des processus ponctuels, une nouvelle valeur constituera une nouvelle configuration de points. Pour l'algorithme de Metropolis-Hastings avec ajout/retrait de points, un nouvel état sera proposé de la manière suivante :

— Avec probabilité p_b, l'algorithme propose de rajouter un point η suivant la densité $b(\mathbf{x}, \eta)$.
— Avec probabilité p_d, l'algorithme propose d'enlever un point $\eta \in \mathbf{x}$ suivant la loi $d(\mathbf{x}, \eta)$.

où nous choisirons $b(\mathbf{x}, \eta) = \frac{\mathbb{1}[\eta \in W]}{\nu(W)}$, la densité de la loi uniforme sur W et $d(\mathbf{x}, \eta) = \frac{1}{n(\mathbf{x})}$, la probabilité de choisir un point au hasard dans la configuration $\mathbf{x}$.

Cette nouvelle configuration est ensuite acceptée avec probabilité $\alpha(\mathbf{x}, \mathbf{x} \cup \{\eta\})$ dans le cas d'un ajout et $\alpha(\mathbf{x}, \mathbf{x} \setminus \{\eta\})$ dans le cas où l'on a proposé d'enlever un point. Dans le cas très spécifique où la configuration courante est réduite à l'ensemble vide, la proposition d'une nouvelle configuration de point se limite à l'ajout d'un point.

Ainsi, le passage d'une configuration $\mathbf{x}$ à $A \in \mathcal{F}$ est décrit par le noyau de transition suivant, qui n'est autre qu'une combinaison linéaire de deux noyaux de transitions :

$$\begin{aligned} P(\mathbf{x}, A) =& p_b \int_W b(\mathbf{x}, \eta)\alpha(\mathbf{x}, (\mathbf{x} \cup \{\eta\}))\mathbb{1}[\mathbf{x} \cup \{\eta\} \in A]d\nu(\eta) \\ &+ p_d \sum_{\eta \in \mathbf{x}} d(\mathbf{x}, \eta)\alpha(\mathbf{x}, \mathbf{x} \setminus \{\eta\})\mathbb{1}[\mathbf{x} \setminus \{\eta\} \in A] \\ &+ \mathbb{1}[\mathbf{x} \in A]\Bigg(1 - p_b \int_W b(\mathbf{x}, \eta)\alpha(\mathbf{x}, \mathbf{x} \cup \{\eta\})d\nu(\eta) \\ &- p_d \sum_{\eta \in \mathbf{x}} d(\mathbf{x}, \eta)\alpha(\mathbf{x}, \mathbf{x} \setminus \{\eta\})\Bigg). \end{aligned} \tag{2.9}$$

Le détail des deux ratios s'obtient en écrivant la réversibilité. Moralement, il s'agit de compenser un ajout par le fait d'ôter un point et réciproquement, cela donne alors :

$$p_b \alpha(\mathbf{x}, \mathbf{x} \cup \eta) f(\mathbf{x}) b(\mathbf{x}, \eta) = p_d \alpha(\mathbf{x} \cup \eta, \mathbf{x}) f(\mathbf{x} \cup \eta) d(\mathbf{x} \cup \eta, \eta).$$

En posant $r_b(\mathbf{x}, \eta) = \frac{f(\mathbf{x} \cup \eta) p_d d(\mathbf{x} \cup \eta, \eta)}{f(\mathbf{x}) p_b b(\mathbf{x}, \eta)}$ et $r_d(\mathbf{x}, \mathbf{x} \setminus \eta) = 1/r_b(\mathbf{x} \setminus \eta, \eta)$, nous obtenons le ratio d'acceptation d'un ajout :

$$\alpha(\mathbf{x}, \mathbf{x} \cup \eta) = \min\left(1, r_b(\mathbf{x}, \eta)\right)$$

et le ratio d'acceptation d'un retrait :

$$\alpha(\mathbf{x}, \mathbf{x} \setminus \eta) = \min\left(1, r_d(\mathbf{x}, \mathbf{x} \setminus \eta)\right).$$

L'algorithme peut donc être résumé par le pseudo-code suivant :

Algorithme 2.2.2. *(Algorithme de Metropolis-Hastings ajout/retrait de points)*
Soient p_b et p_d tels que $p_b + p_d = 1$, les probabilités de choisir un ajout ou un retrait de point.
Pour un certain $m \in \mathbb{N}$, si $X_m = \mathbf{x} \in N_f$, générer X_{m+1} de la manière suivante :

1) *Tirer $U_m \sim \mathbb{U}([0,1])$ et $V_m \sim \mathbb{U}([0,1])$.*
2) *Si $U_m \leq p_b$, générer un nouveau point $\eta_m \sim b(\mathbf{x}, \eta_m)$ et mettre à jour*

$$X_{m+1} = \mathbf{x} \cup \eta_m \text{ si } V_m \leq r_b(\mathbf{x}, \eta_m) \text{ ; } \mathbf{x} \text{ sinon.}$$

3) *Si $U_m > p_b$ alors*
  a) *Si $\mathbf{x} = \emptyset$ alors $X_{m+1} = \mathbf{x}$*
  b) *Sinon générer $\eta_m \sim d(\mathbf{x}, \eta_m)$ et poser*

$$X_{m+1} = x \setminus \eta_m \text{ si } V_m \leq r_d(x, \eta_m) \text{ ; } \mathbf{x} \text{ sinon.}$$

Avec ces choix de lois pour les ajouts/retraits de points et en prenant $p_b, p_d \neq 0$, nous pouvons montrer que la chaîne de Markov générée par l'algorithme possède toutes les bonnes propriétés attendues : elle est ϕ-irréductible, géométriquement ergodique, vérifie la récurrence au sens de Harris. Elle admet pour unique loi invariante la densité f et converge vers cette dernière. [Geyer, 1999, van Lieshout, 2000, Stoica, 2025, Møller and Waagepetersen, 2004].

2.2.3 Contrôles de la simulation et applications de la simulation

2.2.3.1 Convergence et indépendance

Comme mentionné précédemment, il s'avère important de détecter le meilleur moment pour prélever des échantillons issus de la chaîne de Markov. Si le nombre d'itérations est trop faible, la configuration de points risque de ne pas présenter les caractéristiques attendues et, à l'inverse, trop d'itérations sont à éviter d'un point de vue informatique. Un autre problème peut résider dans l'indépendance des échantillons conservés. Si le nombre d'itérations entre deux configurations de points retenues sont trop proches, les configurations risquent d'avoir des caractéristiques similaires. Le théorème suivant justifie également, dans le cas où la CDM est récurrente et ergodique, l'intérêt de minimiser la corrélation des réalisations :

Théorème 2.2.3. *Soit $\pi(\cdot)$ la distribution d'équilibre d'une chaîne récurrente et ergodique. Si la fonction g est $\pi-$intégrable, i.e*

$$s := \mathbb{E}_\pi[g(X)] = \int_x g(x)\pi(x)dx < \infty.$$

Soit alors $s_n := \frac{1}{n}\sum_{i=1}^{n} g(X_i)$ avec $X_1, X_2, \ldots, X_n$ des réalisations de la chaîne, l'application du théorème Centrale-Limite implique alors

$$\sqrt{n}(s_n - s) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

avec $\sigma^2 = \mathbb{V}ar_\pi[g(X)] + 2\sum_{k=1}^{n} Cov_\pi[g(X_0), g(X_k)]$.

Nous voyons alors que pour minimiser la variance asymptotique des réalisations retenues, il est intéressant de les prendre indépendantes.

D'autres points auraient pu être traités comme le rôle de la loi de proposition, le comportement de l'échantillonneur en fonction des autres paramètres. Nous proposons dans cette partie des stratégies pour la prise en main de l'algorithme MH pour les processus ponctuels.

Les simulations menées dans cette partie ont été obtenues à l'aide de la librairie `C++ DRLib` et les résultats affichés avec le logiciel `R`. Pour plus de détails sur cette librairie, le lecteur pourra se référer à l'annexe B. Indépendamment du modèle considéré, les paramètres de l'algorithme en `C++` sont les suivants :

- W correspond au domaine dans lequel nous simulerons (ici fixé à $[0,1] \times [0,1]$).
- N correspond au nombre de réalisations.
- p_b, p_d les probabilités d'ajout/retrait de points (ici fixé à $p_b = p_d = 0.5$).
- N_{MH} correspond au nombre d'itérations de MH.

a) Convergence

Le but principal va être de trouver un N_{MH} adapté pour que les réalisations soient proches du modèle souhaité. Un premier indicateur est la série temporelle des valeurs des statistiques suffisantes du modèle. Leur évolution permet de constater deux phases : une phase de «pré-convergence», où la variance de ces statistiques reste importante, et une phase de «convergence» où cette variance devient faible. Nous illustrons ce phénomène sur le modèle de Strauss avec $\rho = 200$, $r = 0.05$ et $\gamma_s = 0.2$, les statistiques suffisantes sont $n(\mathbf{x})$, le nombre de points, et $s_r(\mathbf{x})$, le nombre de $r-$voisins. Nous avons fixé $N = 1$ et $N_{MH} = 5000$.

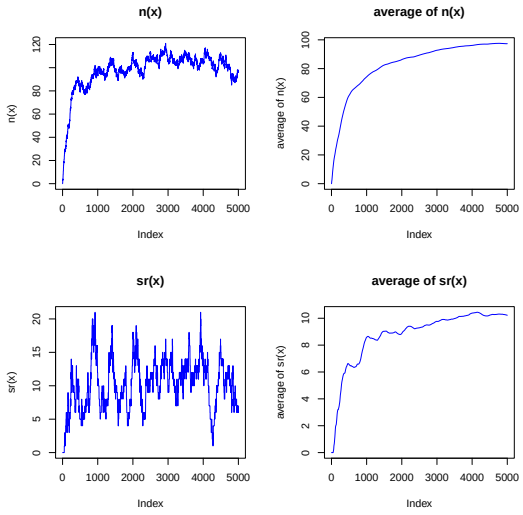


FIGURE 2.11 – Série temporelle du nombre de points (haut gauche) et moyenne cumulée du nombre de points (haut droite) ; Série temporelle du nombre de $r-$voisins (bas gauche) et moyenne cumulée associée (bas droite).

Nous remarquons sur la figure 2.11 que la série temporelle pour le nombre de points et la moyenne cumulée semblent se stabiliser après $N_{MH} = 3000$ itérations. Le même phénomène s'observe pour la moyenne de s_r. La convergence semble donc correcte à partir de cette valeur. En scindant les échantillons en deux, nous pouvons alors calculer les variances pour ces deux «phases» :

Variance pour les 3000 premières itérations	Variance pour les 2000 dernières itérations
$\mathbb{V}ar[n(\mathbf{x})] = 373.93$	$\mathbb{V}ar[n(\mathbf{x})] = 37.97$
$\mathbb{V}ar[s_r(\mathbf{x})] = 15.60$	$\mathbb{V}ar[s_r(\mathbf{x})] = 10.24$

La valeur 3000 reste néanmoins prise arbitrairement à l'œil sur les graphiques, cela permet tout de même de fixer un ordre de grandeur sur le nombre d'itérations à effectuer pour obtenir la convergence.

b) Indépendance des échantillons

Désormais, si nous supposons que la convergence est atteinte, plutôt que de repartir d'une configuration vide, il est intéressant de prendre une configuration tirée suivant le modèle d'intérêt comme état initial afin de se passer du temps de convergence. Se pose alors la question de l'indépendance entre les réalisations à conserver. À partir de combien d'itérations de mises à jour de

Metropolis-Hastings deux configurations peuvent être considérées indépendantes ? À nouveau, il est possible de s'appuyer sur les statistiques suffisantes du modèle. De plus, nous pouvons considérer qu'une configuration est indépendante d'une autre lorsque leur nombre de points en commun (deux points ayant les mêmes coordonnées) est faible.

Tout d'abord, regardons les diagrammes d'autocorrélation des statistiques suffisantes des états $X_{3000}, ..., X_{5000}$. Les diagrammes étant très similaires dans ce cas, seul le diagramme du nombre de points est illustré ci-après.

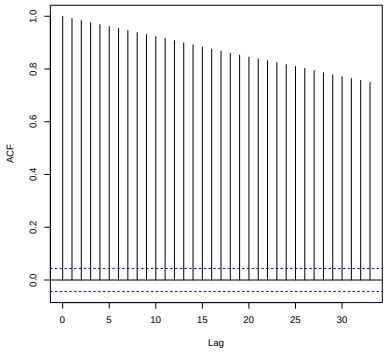
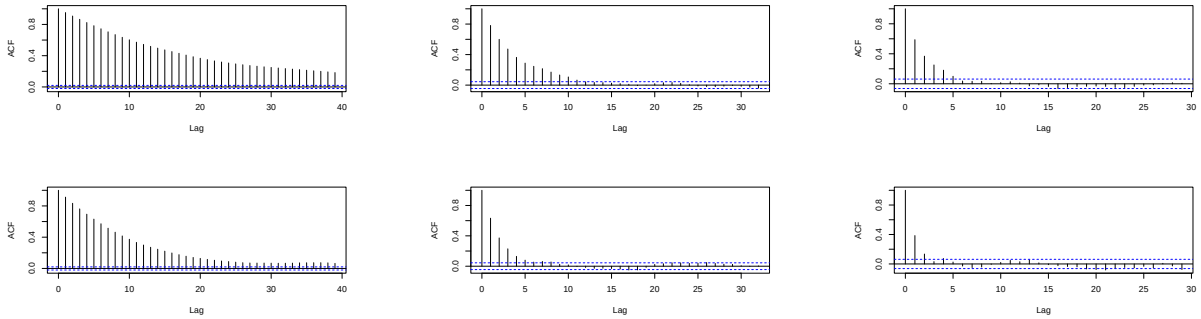
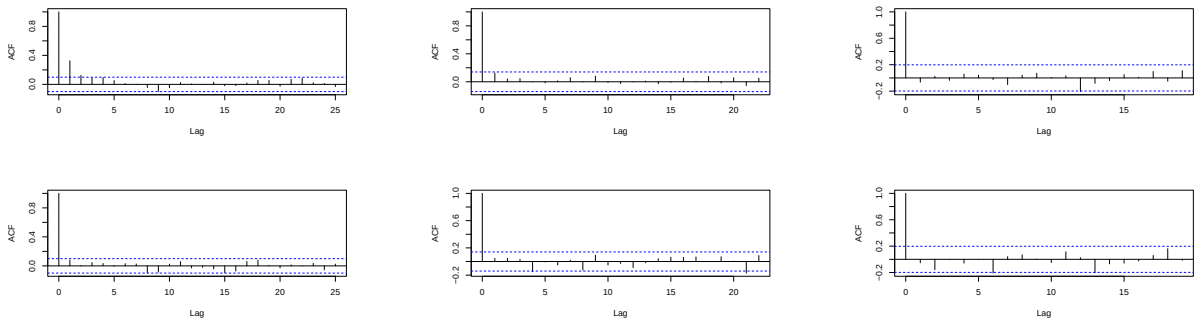


FIGURE 2.12 – Diagramme d'autocorrélation pour le nombre de points

Comme attendu, deux états successifs après mise à jour de Metropolis-Hastings sont très corrélés. Il convient alors de trouver un entier m tel que X_i et X_{i+m} soient peu corrélés.



(a) m=10 (b) m=50 (c) m=100



(d) m=250 (e) m=500 (f) m=1000

FIGURE 2.14 – Diagrammes d'autocorrélations pour différentes valeurs d'espacement m entre deux états de la chaîne. Pour chaque figure, le diagramme du haut est celui pour le nombre de points, celui du bas pour le nombre de $r-$voisins.

Nous observons que jusqu'à $m = 100$, les différentes réalisations restent encore corrélées. À partir de $m = 250$, la corrélation devient très faible entre les états de la chaîne.

Intéressons-nous désormais au renouvellement des points : combien de points, en moyenne, restent d'un état X_i à un état X_{i+m} pour les valeurs de m ci-dessus ? Nous avons alors pris 100 réalisations successives pour les différents m et nous les avons comparées successivement 2 par 2, établissant une moyenne à partir de 99 comparaisons. Le tableau ci-dessous donne la moyenne du nombre de points en commun et le nombre moyen de points des 100 configurations.

m	Moyenne du nombre de points en commun	Moyenne du nombre de points
10	102.26	105.26
50	85.58	99.36
100	74.74	100.37
250	48.40	100.54
500	23.29	99.23
1000	5.81	99.76

Ainsi, même si l'autocorrélation semble indiquer que $m = 250$ est un choix correct pour l'espacement, cette autre approche suggère plutôt de considérer $m = 1000$ pour ce modèle. En effet, près de la moitié des points demeure d'un état à un autre après 250 itérations. Ceci peut créer des caractéristiques spatiales très similaires entre deux configurations consécutives et donc rendre les configurations très dépendantes spatialement, ce qui peut nuire à la qualité des estimateurs obtenus avec ces échantillons. Dans la prochaine section, nous abordons les tests d'enveloppe MCMC.

2.2.3.2 Enveloppes MCMC

La simulation des modèles permet également de vérifier si une configuration peut être considérée comme la réalisation d'un modèle en particulier. Imaginons vouloir tester si un pattern vérifie l'hypothèse de «complete spatial randomness» à l'aide des fonctions de statistiques descriptives spatiales introduites en 2.1.5. La comparaison entre les courbes empiriques et théoriques peut être un premier indicateur. Toutefois, s'appuyer sur une seule observation ne permet pas de conclure,

ou au moins d'arriver à un consensus sur l'aspect complètement aléatoire d'une configuration. Pour pallier cela, nous allons utiliser les enveloppes MCMC. Nous détaillons ici le principe de l'utilisation de ces enveloppes :

1) Considérer une hypothèse H_0.

2) Déterminer T_0 pour $r > 0$, l'estimation de la statistique d'intérêt pour le pattern (F, G ou g par exemple).

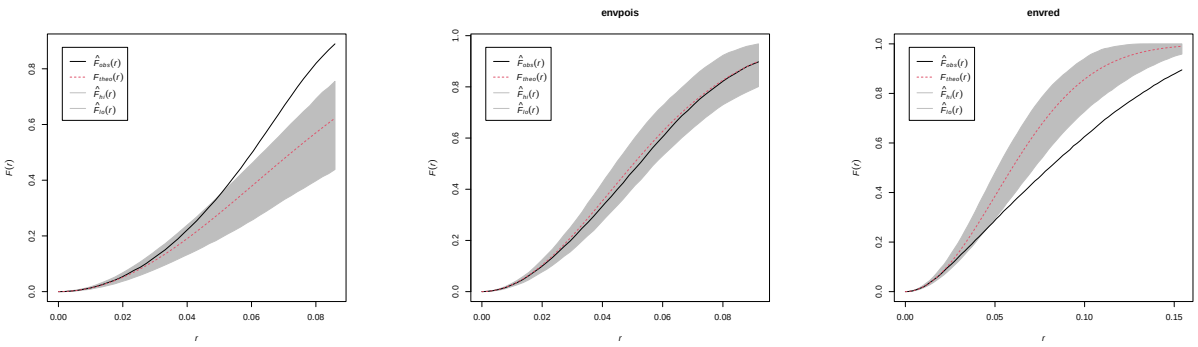
3) Simuler $X_1, ..., X_n$ un échantillon i.i.d sous H_0 et calculer les statistiques associées à chaque réalisation $T_1(r) = T(X_1, r), ..., T_n(r) = T(X_n, r)$.

4) Poser $T_{min}(r) = \min(T_1(r), ..., T_n(r))$ et $T_{max}(r) = \max(T_1(r), ..., T_n(r))$

5) Sous H_0, $\mathbb{P}(T_0(r) < T_{min}(r)) = \mathbb{P}(T_0(r) > T_{max}(r)) \leq \frac{1}{n+1}$ avec égalité si $T_0(r), ...T_n(r)$ sont presque sûrement différents.

Remarques :
• H_0 peut correspondre à beaucoup d'hypothèses différentes : CSR, inhomogénéité, le pattern suit un certain modèle, etc.
• Sous H_0 et pour $n = 39$, la statistique du pattern a 95% de chance d'être à l'intérieur de l'enveloppe.
• Le package `R spatstat` permet de tracer ces enveloppes avec diverses méthodes de corrections (effets de bords,...) pour les fonctions de statistiques descriptives présentées plus haut.
• Lors de cette procédure, plusieurs tests sont effectués en parallèles, rendant la puissance globale très faible. Nous nous contentons d'illustrer le principe des enveloppes à travers les enveloppes dites «simples». Une solution à ce problème de $p-$valeur, proposée par [Myllymäki et al., 2016], consiste à prendre des tests d'enveloppe globaux.

Exemple 2.2.2. *(Enveloppes MCMC pour la fonction F).*
Reprenant les trois exemples 2.1.6, nous souhaitons tester l'hypothèse H_0 : les patterns observés vérifient la propriété de «complete spatial randomness». Pour chaque valeur de r, la fonction «envelope» de `spatstat` *simule n réalisations i.i.d d'un processus de Poisson et détermine $T_{min}(r)$ et $T_{max}(r)$. Nous avons fixé $n = 39$.*



(a) Dataset Cells (b) Poisson d'intensité 100 (c) Dataset Redwood

FIGURE 2.15 – En rouge : Courbe théorique ; En noir : l'estimation de la fonction F : F_{est} pour les données Cells (gauche) ; Processus de Poisson d'intensité 100 (milieu) et redwood (droite). La zone grisée correspond à l'enveloppe créée par les simulations pour chaque valeur de r.

Nous remarquons en figure 2.15a que le Dataset Cells semble plus répulsif que des configurations Poissoniennes pour des rayons supérieurs à 0.05*, valeur à laquelle l'estimation de la fonction F sort de l'enveloppe. La réalisation du processus de Poisson 2.15b est, comme attendu, à l'intérieur de l'enveloppe. Enfin, la figure 2.15c illustre que le dataset Redwood semble plus aggloméré qu'un processus de Poisson.*

2.2.3.3 Étude de modèles

Cette section a pour but d'étudier plus en détail les comportements des modèles de Geyer (2.7) et de la superposition du modèle de Strauss et Area-Interaction (2.8). Nous ciblons ici l'impact du seuil de saturation s pour le premier et la valeur du paramètre γ_s pour le second. Une étude par simulation permet généralement de prendre en main les modèles et l'objectif ici est de présenter et de donner des références sur les comportements attendus pour ces deux modèles, peu décrits dans la littérature.

Le choix des paramètres de simulations est décrit ci-dessous :

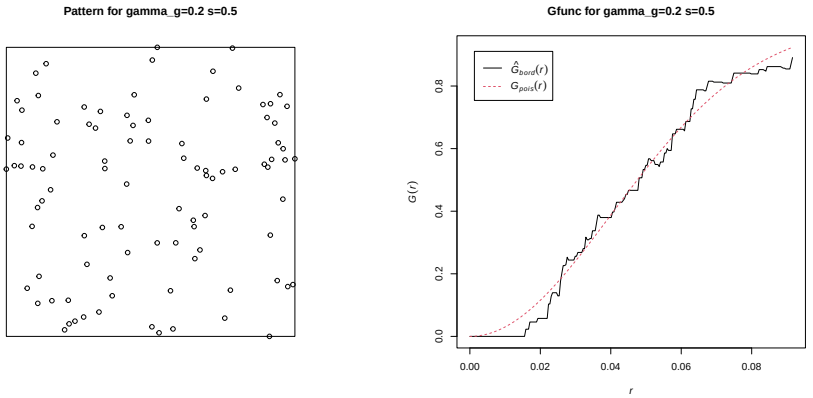
— Paramètres pour le processus de saturation de Geyer :

Variable	Description	Valeur
W	Domaine de simulation	$[0,1] \times [0,1]$
N	Nombre d'échantillons	1000
N_{MH}	Nombre d'itérations MH	10^6
m	Espacement entre les configurations	1000
p_b, p_d	Probabilité d'ajout/retrait	$0.5, 0.5$
(ρ, γ_g)	Paramètres du modèle	$(50, 1.49); (75, 1.34); (100, 1.22)$
s	Seuil de saturation	$0.5, 1, ..., 4, 4.5$
r	Rayon d'interaction	0.05

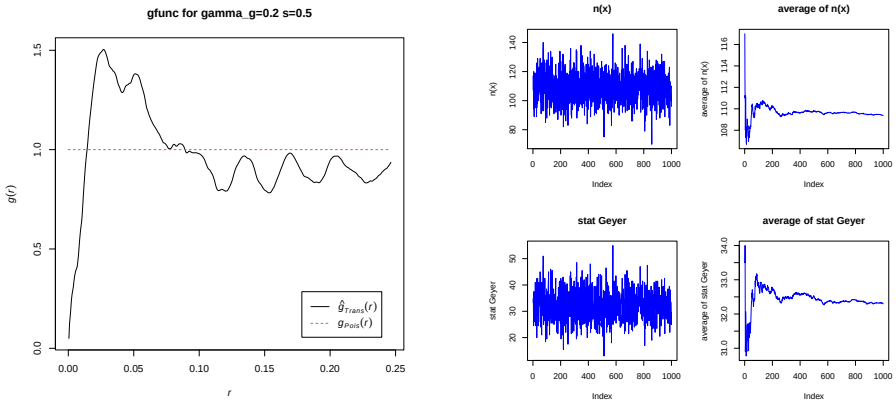
— Paramètres pour la superposition de Strauss et Area-Interaction :

Variable	Description	Valeur
W	Domaine de simulation	$[0,1] \times [0,1]$
N	Nombre d'échantillons	1000
N_{MH}	Nombre d'itérations MH	10^6
m	Espacement entre les configurations	1000
p_b, p_d	Probabilité d'ajout/retrait	$0.5, 0.5$
(ρ, γ_a)	Paramètres du modèle	$(200, 0.005); (200, 0.2); (200, 0.5)$
γ_s	Paramètre de répulsion	$0.1, 0.2, ..., 0.7$
r, R	Rayons d'interaction	$0.05, 0.05$

Pour chaque jeu de paramètres, nous traçons la série temporelle des statistiques suffisantes ainsi que la moyenne cumulée de cette série et le dernier pattern simulé avec ses estimations pour les fonctions G et g, un exemple est donné ci-dessous. Afin d'alléger le contenu de cette section, nous donnons seulement les moyennes des statistiques suffisantes sous forme de tableaux ci-après.



(a) Dernière configuration obtenue (b) Estimation pour G



(c) Estimation pour g (d) Statistiques

FIGURE 2.16 – Figures obtenues pour $\gamma_g = 0.2$ et $s = 0.5$

s	$\overline{n(\mathbf{x})}$	Moyenne de $\sum_{\xi \in \mathbf{x}} \min(s_r(\xi), s)$
0.5	63.073	14.2520
1.0	75.811	43.4510
1.5	84.811	64.7855
2.0	101.067	111.8050
2.5	114.259	153.9720
3.0	141.608	255.3110
3.5	170.912	375.2765
4.0	238.803	687.5640
4.5	334.072	1179.5205

TABLE 2.1 – Résultats pour les paramètres $(\rho, \gamma_g) = (50, 1.49)$

s	$\overline{n(\mathbf{x})}$	Moyenne de $\sum_{\xi\in\mathbf{x}} \min(s_r(\xi), s)$
0.5	85.342	22.3715
1.0	98.851	61.4530
1.5	108.970	89.4215
2.0	124.818	142.4440
2.5	136.303	183.9825
3.0	157.284	267.8580
3.5	175.129	346.0305
4.0	208.970	513.0270
4.5	245.396	708.2065

TABLE 2.2 – Résultats pour les paramètres $(\rho, \gamma_g) = (75, 1.34)$

s	$\overline{n(\mathbf{x})}$	Moyenne de $\sum_{\xi\in\mathbf{x}} \min(s_r(\xi), s)$
0.5	109.375	32.2960
1.0	120.944	79.3310
1.5	130.122	111.4160
2.0	142.022	161.3490
2.5	150.203	195.2215
3.0	161.674	248.8220
3.5	169.622	287.7805
4.0	181.344	350.3630
4.5	189.929	398.7745

TABLE 2.3 – Résultats pour les paramètres $(\rho, \gamma_g) = (100, 1.22)$

Nous remarquons que pour chaque tableau, le nombre de points évolue avec le seuil s mais à des vitesses différentes en fonction du paramètre γ_g. Le paramètre γ_g semble donc avoir plus d'impact sur la croissance du nombre de points que le paramètre ρ dès que $s \geq 1$. La statistique du modèle de saturation de Geyer évolue également avec le seuil et avec la valeur du paramètre γ_g.

γ_s	$\overline{n(\mathbf{x})}$	$\overline{s_r(\mathbf{x})}$	$\overline{a_r(\mathbf{x})}$
0.1	116.301	3.477	-115.2332
0.2	119.115	7.325	-116.7295
0.3	122.221	11.539	-118.4807
0.4	126.327	17.539	-121.1440
0.5	132.538	26.325	-125.1405
0.6	140.230	36.594	-130.3960
0.7	149.411	50.060	-136.3681

TABLE 2.4 – Résultats pour les paramètres $(\rho, \gamma_a) = (200, 0.5)$

γ_s	$\overline{n(\mathbf{x})}$	$\overline{s_r(\mathbf{x})}$	$\overline{a_r(\mathbf{x})}$
0.1	115.416	2.001	-114.3636
0.2	117.352	4.961	-115.4286
0.3	120.793	9.581	-117.6481
0.4	125.351	15.290	-120.7647
0.5	131.504	23.465	-124.7022
0.6	139.155	34.037	-129.8917
0.7	148.154	47.126	-135.8631

TABLE 2.5 – Résultats pour les paramètres $(\rho, \gamma_a) = (200, 0.2)$

γ_s	$\overline{n(\mathbf{x})}$	$\overline{s_r(\mathbf{x})}$	$\overline{a_r(\mathbf{x})}$
0.1	114.205	2.818	-115.5458
0.2	121.749	7.709	-122.2879
0.3	133.332	14.389	-132.9893
0.4	144.871	24.067	-143.0295
0.5	159.466	36.548	-155.5089
0.6	174.968	53.469	-168.1485
0.7	195.977	77.344	-185.3200

TABLE 2.6 – Résultats pour les paramètres $(\rho, \gamma_a) = (200, 0.005)$

Nous remarquons que les deux premiers tableaux exhibent des statistiques similaires pour le nombre de points, le nombre de $r-$voisins et la statistique du modèle Area-Interaction. Pour cette intensité fixée à $\rho = 200$, l'influence du paramètre γ_a varie très peu entre 0.5 et 0.2. Néanmoins, pour $\gamma_a = 0.005$, les configurations couvrant une grande surface sont favorisées, augmentant le nombre moyen de points et donc le nombre de $r-$voisins. Enfin, le paramètre γ_s permet bien de contrôler le nombre de $r-$voisins.

Dans tous les cas étudiés pour ces deux modèles, une analyse plus détaillée peut être effectuée en regardant les fonctions de statistiques descriptives spatiales et les configurations obtenues comme présenté plus haut. Pour des modèles plus complexes comme ceux que nous allons introduire plus tard dans ce manuscrit, cela permet, lorsque nous regardons les configurations obtenues, de se faire une idée plus claire sur les comportements possibles du modèle.

2.3 Inférence

Dans cette section, nous revenons plus en détail sur les notions énoncées en introduction pour l'inférence. Nous commençons par détailler une méthode MCMC pour calculer le maximum de vraisemblance aussi bien en données complètes qu'incomplètes proposée par [Geyer, 1999, Geyer and Thompson, 1992]. Ensuite, nous rappelons également des approches Bayésiennes pour échantillonner la loi *a posteriori*, loi conditionnelle des paramètres sachant les observations, plus précisément, lorsque les fonctions de partition ne sont pas accessibles sous forme analytique. Enfin, l'algorithme ABC Shadow [Stoica et al., 2017] et ses propriétés sont détaillées et illustrés avec des applications.

Dans cette partie, notons Θ l'espace des paramètres, $f(\mathbf{x}|\theta)$ une densité de probabilité que nous prendrons sous la forme

$$f(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} e^{\langle t(\mathbf{x}),\theta\rangle},$$

où $Z(\theta)$ n'est pas traçable analytiquement, $t(\mathbf{x})$ représente les statistiques suffisantes du modèle et θ le vecteur de paramètre. Les approches présentées ici s'appliquent à la famille exponentielle.

2.3.1 Maximum de vraisemblance MCMC en données complètes

Afin de réduire le nombre de simulations nécessaires pour approcher la vraisemblance, [Geyer and Thompson, 1992] proposent de considérer le ratio de log −vraisemblance par rapport à un paramètre de référence. Considérons ce ratio des log-vraisemblances observées pour un autre paramètre fixé $\psi \in \Theta$:

$$l(\theta) = \log \frac{e^{\langle t(\mathbf{x}),\theta\rangle}}{e^{\langle t(\mathbf{x}),\psi\rangle}} - \log \frac{Z(\theta)}{Z(\psi)} = \langle t(\mathbf{x}), \theta - \psi\rangle - \log \frac{Z(\theta)}{Z(\psi)}. \tag{2.10}$$

Le premier terme est facilement connu et se résume à un produit scalaire dans le cas des processus de Gibbs. Le deuxième est quant à lui plus difficile à traiter car il fait intervenir les constantes de normalisation.

Dans le cas de la famille exponentielle, nous pouvons directement réécrire le ratio $Z(\theta)/Z(\psi)$ comme une espérance sous la loi de $f(\cdot|\psi)$:

$$\frac{Z(\theta)}{Z(\psi)} = \frac{1}{Z(\psi)} \int e^{\langle t(\mathbf{x}),\theta\rangle} \mu(d\mathbf{x}) = \int \frac{e^{\langle t(\mathbf{x}),\theta\rangle}}{e^{\langle t(\mathbf{x}),\psi\rangle}} f_\psi(\mathbf{x}) \mu(d\mathbf{x}) = \mathbb{E}_\psi \left[\frac{e^{\langle t(X),\theta\rangle}}{e^{\langle t(X),\psi\rangle}} \right].$$

Si nous simulons $X_1, ..., X_n$ un échantillon sous la loi $f(\cdot|\psi)$, alors l'espérance ci-dessus peut être approximée par

$$\frac{1}{n} \sum_{i=1}^{n} \frac{e^{\langle t(X_i),\theta\rangle}}{e^{\langle t(X_i),\psi\rangle}} = \frac{1}{n} \sum_{i=1}^{n} e^{\langle t(X_i,\theta-\psi\rangle}.$$

Alors

$$l_n(\theta) = \langle t(\mathbf{x}), \theta - \psi\rangle - \log \left(\frac{1}{n} \sum_{i=1}^{n} e^{\langle t(X_i),\theta-\psi\rangle} \right) \tag{2.11}$$

est une approximation de $l(\theta)$.

Si la chaîne de Markov qui génère les échantillons $X_1, ..., X_n$ est irréductible et apériodique, alors la Loi Forte des Grands Nombres assure que $\frac{1}{n}\sum_{i=1}^{n} e^{\langle t(X_i),\theta-\psi\rangle}$ converge presque sûrement vers $\frac{Z(\theta)}{Z(\psi)}$ et implique alors que $l_n(\theta)$ converge presque sûrement vers $l(\theta)$. De plus, si $f(\mathbf{x}|\theta)$ appartient à la famille exponentielle, le maximum de $l_n(\cdot)$, $\hat{\theta_n}$, converge presque sûrement vers $\hat{\theta}$, le maximum de $l(\cdot)$ [Geyer and Thompson, 1992, Monfort, 1997, Geyer, 1999, Møller and Waagepetersen, 2004, Stoica, 2025].

Maximiser l_n en θ donne alors une approximation du maximum de vraisemblance $\hat{\theta}$, cette approximation dépend grandement du choix de ψ. Le cas de la famille exponentielle est encore une fois favorable puisque $l(\cdot)$ est convexe, étudier les dérivées de $l_n(\cdot)$ est alors intéressant.

$$\begin{aligned}\nabla l_n(\theta) = t(\mathbf{x}) - \frac{\sum_{i=1}^{n}\left(t(X_i)\frac{e^{\langle t(X_i),\theta\rangle}}{e^{\langle t(X_i),\psi\rangle}}\right)}{\sum_{i=1}^{n}\frac{e^{\langle t(X_i),\theta\rangle}}{e^{\langle t(X_i),\psi\rangle}}} &= t(\mathbf{x}) - \frac{\sum_{i=1}^{n}\left(t(X_i)e^{\langle t(X_i),\theta-\psi\rangle}\right)}{\sum_{i=1}^{n}e^{\langle t(X_i),\theta-\psi\rangle}} \\ &= t(\mathbf{x}) - \mathbb{E}_{n,\theta,\psi}[t(X)]\end{aligned}$$

où $\mathbb{E}_{n,\theta,\psi}[t(X)]$ est l'approximation MCMC de $\mathbb{E}_\theta t(X)$.

Dérivant une seconde fois, nous obtenons :

$$\nabla^2 l_n(\theta) = \mathbb{E}_{n,\theta,\psi}t(X)t(X)^T - \left[\mathbb{E}_{n,\theta,\psi}t(X)\right]\left[\mathbb{E}_{n,\theta,\psi}t(X)\right]^T = -\mathbb{V}ar_{n,\theta,\psi}t(X) \tag{2.12}$$

À partir de ces quantités, des algorithmes itératifs peuvent être mis en place pour l'optimisation locale (algorithme de Newton-Raphson [Geyer, 1999], gradient stochastique
[Moyeed and Baddeley, 1991], gradient à pas optimal [van Lieshout and Stoica, 2003]).

2.3.2 Maximum de vraisemblance MCMC en données incomplètes

Une version similaire peut-être utilisée pour le cas des données incomplètes. Rappelons le contexte donné en introduction : supposons qu'un processus ponctuel existe dans la région $W = W_X \cup W_Y$ et n'est observé que sur la région W_Y comme l'illustre la Figure 1.3 donnée dans l'état de l'art.

La densité du processus peut être écrite comme la densité jointe des données sur W :

$$f(\mathbf{x},\mathbf{y}|\theta) = \frac{\exp\langle t(\mathbf{x}\cup\mathbf{y})|\theta\rangle}{Z(\theta)}$$

où $\mathbf{y}$ est une configuration de points observée dans la région W_Y et $\mathbf{x}$ est une configuration cachée dans la région W_X.

La vraisemblance est alors la loi marginale suivant la variable $\mathbf{y}$,

$$L(\theta) = f(\mathbf{y}|\theta) = \int f(\mathbf{x},\mathbf{y}|\theta)\mu(\mathrm{d}\mathbf{x}) = \frac{Z(\theta|\mathbf{y})}{Z(\theta)},$$

avec $Z(\theta|\mathbf{y})$ la constante de normalisation de la densité $f(\mathbf{x}|\mathbf{y},\theta)$.

L'approche proposée par [Gelfand and Carlin, 1993] et également détaillée dans [Geyer, 1999] repose encore sur l'échantillonnage préférentiel. Comme présenté plus haut, le ratio de log − vraisemblance par rapport à un autre paramètre fixé ψ est considéré :

$$l(\theta) = \log \frac{Z(\theta|\mathbf{y})}{Z(\psi|\mathbf{y})} - \log \frac{Z(\theta)}{Z(\psi)}$$

et ainsi, par le même raisonnement que précédemment, nous pouvons réécrire, comme pour le cas des données complètes :

$$\frac{Z(\theta)}{Z(\psi)} = \mathbb{E}\left[\frac{e^{\langle t(X\cup Y),\theta\rangle}}{e^{\langle t(X\cup Y),\psi\rangle}}\right],$$

et en conditionnant par rapport à la configuration observée $\mathbf{y}$

$$\frac{Z(\theta|\mathbf{y})}{Z(\psi|\mathbf{y})} = \mathbb{E}\left[\frac{e^{\langle t(X\cup Y),\theta\rangle}}{e^{\langle t(X\cup Y),\psi\rangle}}\middle| Y = \mathbf{y}\right].$$

Enfin, à l'aide d'échantillons de la loi jointe $(X_i, Y_i)_{1\leq i\leq n}$ et de la loi conditionnelle de $(X_i^*)_{1\leq i\leq n} := (X_i|Y = \mathbf{y})_{1\leq i\leq n}$, les ratios de constantes de normalisation peuvent être approchés par

$$\frac{Z(\theta|\mathbf{y})}{Z(\psi|\mathbf{y})} \approx \frac{1}{n}\sum_{i=1}^{n} e^{\langle t(X_i^*,\mathbf{y}),\theta\psi\rangle} \text{ et } \frac{Z(\theta)}{Z(\psi)} \approx \frac{1}{n}\sum_{i=1}^{n} e^{\langle t(X_i\cup Y_i),\theta-\psi\rangle}$$

ce qui permet ainsi d'approcher le ratio de log −vraisemblance :

$$l_n(\theta) = \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{\langle t(X_i^*,\mathbf{y}),\theta-\psi\rangle}\right) - \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{\langle t(X_i\cup Y_i),\theta-\psi\rangle}\right).$$

Les résultats obtenus dans le cas des données complètes tiennent toujours. Néanmoins, la convexité de la vraisemblance n'est plus garantie, l'optimisation dépend alors des conditions initiales et les algorithmes itératifs risquent de converger vers le maximum local le plus proche [Younes, 1989, Gu and Li, 1998, Delyon et al., 1999]. Le gradient de $l_n(\cdot)$ est néanmoins facile à approcher par MCMC :

$$\nabla l_n(\theta) = \mathbb{E}_{n,\theta,\psi}[t(X\cup Y)|Y = \mathbf{y}] - \mathbb{E}_{n,\theta,\psi}[t(X\cup Y)].$$

2.3.3 Méthodes Bayésiennes : échantillonner la loi *a posteriori*

L'inférence Bayésienne repose sur l'emploi de la formule de Bayes afin de mettre à jour nos connaissances sur un phénomène par rapport aux observations supposées influencer ce phénomène. Formellement, considérons $p(\theta)$ une distribution de probabilité sur les paramètres (loi *a priori*), elle représente les «connaissances» que nous avons sur ces paramètres. En appliquant la formule de Bayes à la densité $f(\mathbf{x}|\theta)$ (implicitement, loi des configurations sachant les paramètres), nous obtenons alors, en décomposant la loi jointe $f(\mathbf{x},\theta) = f(\mathbf{x}|\theta)p(\theta)$,

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)p(\theta)}{\int f(\mathbf{x},\theta)\mu(\mathrm{d}\theta)} = \frac{f(\mathbf{x}|\theta)p(\theta)}{Z(\mathbf{x})}$$

dans notre cas, cette loi s'exprimera alors :

$$f(\theta|\mathbf{x}) = \frac{\exp\langle t(\mathbf{x})|\theta\rangle p(\theta)}{Z(\mathbf{x})Z(\theta)},$$

avec $Z(\mathbf{x})$ la constante de normalisation liée aux données et $Z(\theta)$ la constante de normalisation du modèle.

La loi *a posteriori* possède alors deux constantes de normalisation qui ne sont en général pas calculable dans le cas des processus ponctuels de Gibbs. Nous introduisons ici des méthodes générales afin d'échantillonner la loi *a posteriori*. Ces méthodes ne seront pas utilisées formellement dans la suite mais motivent l'introduction de l'algorithme que nous utilisons pour effectuer l'estimation des paramètres : l'algorithme ABC Shadow. Plus d'approches peuvent être trouvées dans l'article [Lu and Friel, 2024].

Remarque :
• Lorsque la loi *a priori* suit une loi uniforme sur un intervalle (ou un produit d'intervalle dans le cas où le modèle possède plusieurs paramètres), la loi *a posteriori* est une restriction de la vraisemblance à l'intervalle de définition de la loi uniforme. Ainsi, ces algorithmes permettent, pour ce choix spécifique de prior, d'approcher la vraisemblance si le support de la loi uniforme est choisi assez grand.

2.3.3.1 Échantillonnage direct par algorithme de Metropolis-Hastings

Cette approche est seulement réalisable lorsque le modèle est entièrement connu. Le fonctionnement est similaire aux algorithmes de simulations sauf que cette fois-ci la loi de proposition est une distribution sur le paramètre. Notant $q(\cdot|\cdot)$ cette loi, le fonctionnement de l'algorithme est résumé par le pseudo-code qui suit.

Algorithme 2.3.1. *(Algorithme de Metropolis-Hastings pour l'échantillonnage a posteriori)*

Initialisation : *Fixer θ_0 l'état initial et $i = 1$, $\mathbf{x}$ les données et $f(\mathbf{x}|\theta)$ le modèle considéré.*

1) *À l'itération i :*

a) Générer $\theta \sim q(\theta|\theta_{i-1})$.
b) Calculer la probabilité

$$\alpha = \min\left\{1, \frac{f(\theta|\mathbf{x})}{f(\theta_{k-1}|\mathbf{x})} \times \frac{q(\theta_{i-1}|\theta)}{q(\theta|\theta_{i-1})}\right\}$$

c) Simuler $U \sim \mathbb{U}([0,1])$:

- *Si $U \leq \alpha$, alors $\theta_i = \theta$*
- *Sinon, $\theta_i = \theta_{i-1}$*

2) *$i = i + 1$ et répéter en partant de **1)** si nécessaire.*

3) *Retourner $\theta_0, \theta_1, \theta_2, ...$*

Pour des hypothèses peu exigeantes sur $q(\cdot|\cdot)$, cet algorithme simule une chaîne de Markov conver-

gente vers la loi *a posteriori* d'intérêt et uniformément ergodique [Tierney, 1994].

Le ratio $\alpha = \min\left\{1, e^{\langle t(\mathbf{x})|\theta-\theta_{k-1}\rangle}\frac{p(\theta)}{p(\theta_{k-1})} \times \frac{Z(\mathbf{x})}{Z(\mathbf{x})} \times \frac{Z(\theta_{k-1})}{Z(\theta)} \times \frac{q(\theta_{i-1}|\theta)}{q(\theta|\theta_{i-1})}\right\}$ fait disparaître la constante liée aux données mais dépend du ratio $\frac{Z(\theta_{k-1})}{Z(\theta)}$, qui n'est pas connu sous forme analytique pour la plupart des modèles de processus ponctuels. Nous détaillons ci-après des solutions pour se passer de ce ratio.

2.3.3.2 Algorithmes basés sur une variable auxiliaire

Algorithme de Metropolis-Hastings avec variable auxiliaire

Comme énoncé ci-dessus, l'algorithme MH classique ne permet pas d'échantillonner la loi *a posteriori* dans le cas général. Les auteurs [Møller et al., 2006] ont proposé un algorithme MH reposant sur l'utilisation d'une variable auxiliaire $\mathbb{Y}$ de densité $a(\mathbf{y}|\theta, \mathbf{x})$ afin de faire disparaître le ratio de constante de normalisation exhibé ci-dessus lors du calcul de α. La densité à échantillonner devient alors $f(\theta, \mathbf{y}|\mathbf{x}) = a(\mathbf{y}|\theta, \mathbf{x})f(\theta|\mathbf{x})$. L'idée ici est donc de générer une chaîne de Markov ayant une dynamique de Metropolis-Hastings proposant des nouvelles valeurs comme étant un couple $(\theta', \mathbf{y}')$ à l'aide d'une loi de proposition bien choisie pour annuler le calcul du ratio.

Pour cela, les auteurs proposent de scinder la loi de proposition $q((\theta, \mathbf{y}) \to (\theta', \mathbf{y}'))$ en deux sous lois de proposition :

$$q((\theta, \mathbf{y}) \to (\theta', \mathbf{y}')) = q_1(\theta'|\theta, \mathbf{x}) \times q_2(\mathbf{y}'|\theta', \theta, \mathbf{y}).$$

q_1 peut être choisie indépendante de la valeur de $\mathbf{x}$ et être une loi simple (e.g. uniforme centrée en θ et d'amplitude δ). De même, q_2 peut être choisie indépendante de $(\theta, \mathbf{y})$. Le but étant d'annuler les constantes de normalisation, les auteurs proposent

$$q_2(\mathbf{y}'|\theta') = \frac{1}{Z(\theta')}e^{\langle t(\mathbf{y}'),\theta'\rangle}.$$

Le ratio devient alors

$$\alpha((\theta, \mathbf{y}) \to (\theta', \mathbf{y}')) = \frac{a(y'|\theta', \mathbf{x})p(\theta')\exp\left(\langle t(\mathbf{x}), \theta'\rangle\right)\exp\left(\langle t(\mathbf{y}), \theta\rangle\right)}{a(y|\theta', \mathbf{x})p(\theta)\exp\left(\langle t(\mathbf{x}), \theta\rangle\right)\exp\left(\langle t(\mathbf{y}'), \theta'\rangle\right)}$$

et ne dépend alors plus de la constante de normalisation.

La procédure est résumée par le pseudo-code ci-dessous avec, pour l'exemple, une densité suivant une loi de la famille exponentielle.

Algorithme 2.3.2. *(Metropolis-Hastings variable auxiliaire)*

Initialisation : *Fixer* $(\mathbf{y}, \theta)$ *comme état initial et* $\mathbf{x}$ *les données.*
1) *Générer* $\theta' \sim q_1(\theta'|\theta)$

2) *Générer* $\mathbf{y}' \sim q_2(\mathbf{y}'|\theta') = \frac{\exp(\langle t(\mathbf{y}'), \theta' \rangle)}{c(\theta')}$

3) *Calculer le ratio* $\alpha((\theta, \mathbf{y}) \to (\theta', \mathbf{y}')) = \frac{a(\mathbf{y}'|\theta', \mathbf{x}) p(\theta') \exp(\langle t(\mathbf{x}), \theta' \rangle) \exp(\langle t(\mathbf{y}), \theta \rangle)}{a(\mathbf{y}|\theta', \mathbf{x}) p(\theta) \exp(\langle t(\mathbf{x}), \theta \rangle) \exp(\langle t(\mathbf{y}'), \theta' \rangle)}$

4) *Simuler* $U \sim \mathbb{U}([0, 1])$ *:*

- *Si* $U \leq \alpha$*, le nouvel état est accepté.*
- *Sinon,* $(\theta', \mathbf{y}') = (\theta, \mathbf{y})$*.*

Pour garantir la convergence de cette chaîne de Markov vers la distribution invariante $f(\theta, \mathbf{y}|\mathbf{x})$, les échantillons issus de $q_2(\mathbf{y}'|\theta')$ nécessitent une simulation exacte, ce qui peut s'avérer difficile (e.g. quand les interactions deviennent «fortes» [van Lieshout and Stoica, 2006]). De plus, les auteurs mentionnent que le choix de la densité de la variable auxiliaire peut induire un phénomène de «poor mixing» où l'espace des paramètres n'est pas suffisamment visité, rendant l'inférence inefficace.

Exchange algorithm

Une autre approche [Murray et al., 2006] propose d'introduire une variable auxiliaire supplémentaire et une procédure de mise-à-jour basée sur un échange de paramètres. Le but est d'échantillonner à partir de la densité de probabilité $f(\theta, \mathbf{y}, \psi|\mathbf{x}) \propto f(\theta|\mathbf{x}) q(\psi|\theta) f(\mathbf{y}|\psi)$ où $f(\theta|\mathbf{x})$ est la distribution *a posteriori* à partir de laquelle nous voulons échantillonner, $q(\psi|\theta)$ la proposale sur le paramètre et $f(\mathbf{y}|\psi)$ la densité de probabilité sur la variable auxiliaire. Le principe est décrit dans les lignes qui suivent, encore une fois utilisé avec une densité sous la forme exponentielle.

Algorithme 2.3.3. *(Exchange algorithm)*

Initialisation : *Fixer* $(\theta, \mathbf{y}, \psi)$ *comme état initial et* $\mathbf{x}$ *les données.*

1) *Générer un nouveau paramètre* $\psi \sim q(\psi|\theta)$*.*

2) *Générer* $y' \sim q_2(\mathbf{y}'|\theta') = \frac{\exp(\langle t(\mathbf{y}'), \theta' \rangle)}{c(\theta')}$

3) *Calculer le «ratio d'échange» proposant* $\theta' = \psi$ *et* $\psi' = \theta$

$$\begin{aligned} \alpha((\theta', \mathbf{y}', \psi') \to (\psi, \mathbf{y}', \theta)) &= \frac{f(\theta', \mathbf{y}', \psi'|\mathbf{x})}{f(\theta, \mathbf{y}', \psi|\mathbf{x})} \\ &= \exp\left\langle t(\mathbf{x}) - t\left(\mathbf{y}'\right), \psi - \theta \right\rangle \times \frac{p(\psi) q(\theta|\psi)}{p(\theta) q(\psi|\theta)} \end{aligned}$$

4) *Simuler* $U \sim \mathbb{U}([0, 1])$ *: Le nouvel état* $(\theta', \mathbf{y}', \psi') = (\psi, \mathbf{y}', \theta)$ *est alors accepté si* $U \leq \min\{1, \alpha\}$*, sinon* $(\theta', \mathbf{y}', \psi') = (\theta, \mathbf{y}', \psi)$*.*

À nouveau, le ratio α ne dépend plus des constantes de normalisation et la convergence vers la distribution *a posteriori* est garantie par la simulation exacte de $\mathbf{y}'$. Cependant, le mécanisme d'échange empêche cette fois-ci le phénomène de «poor mixing» dans l'algorithme précédent. Plus de détails sur la convergence de cet algorithme peuvent être trouvés dans [Liang, 2010, Liang et al., 2016], où des solutions à l'échantillonnage exact sont également proposées.

2.3.3.3 Méthodes ABC («approximate Bayesian computation»)

Les algorithmes ABC sont des méthodes d'échantillonnage approché initialement conçues pour échantillonner les distributions *a posteriori* de modèles très complexes issus des sciences agricoles et environnementales. Parmi les stratégies ABC existantes [Marin et al., 2011] [Beaumont et al., 2009, Atchadé et al., 2013, Raynal et al., 2018], nous commençons par présenter le principe classique. Nous présentons ensuite l'algorithme ABC Shadow, qui s'inspire des méthodes à variables auxiliaires tout en fournissant un contrôle théorique de la méthode sans nécessiter de simulation exacte [Stoica et al., 2017].

Algorithme ABC par rejet
Le principe de cet algorithme consiste à générer d'abord un échantillon de paramètres selon la loi *a priori*, puis, dans un deuxième temps, à vérifier si les paramètres générés satisfont un critère et à rejeter ceux ne le satisfaisant pas. Par exemple, pour les modèles de la famille exponentielle présentés ci-dessus, un critère courant consisterait à contrôler la distance entre les statistiques suffisantes du modèle observé et celles simulées avec les paramètres générés.

Algorithme 2.3.4. *(Algorithme ABC par rejet)*

1) *Supposons que* $\mathbf{x}$ *soit observé, définissons un seuil de rejet* ϵ *et un nombre d'itérations* N*.*

2) *Pour* $k = 1$ *à* N *:*

- ***a)*** *Générer* θ_i *selon* $p(\theta)$*.*
- ***b)*** *Générer une réalisation* $\mathbf{y}_i$ *selon* $f(\mathbf{y}|\theta_i) = \frac{\exp(\langle t(\mathbf{y}|\theta_i)\rangle)}{c(\theta_i)}$

3) *Conserver tous les* θ_i *tels que* $d(t(\mathbf{x}), t(\mathbf{y}_i)) \leq \epsilon$

L'échantillon conservé comme résultat de cet algorithme est distribué selon $f(\theta|d(t(\mathbf{x}), t(\mathbf{y}) \leq \epsilon)$. Le choix du seuil ϵ et de la distance d doivent être effectués avec soins afin que l'échantillon soit suffisamment proche de la loi *a posteriori* et que la quantité de paramètres rejetés ne soit pas trop élevée. La section suivante présente l'algorithme «ABC Shadow» que nous avons utilisé dans cette thèse pour mener l'inférence statistique.

2.3.4 Algorithme ABC Shadow

Cet algorithme ABC combine deux idées [Stoica et al., 2017] :

- L'usage d'une variable auxiliaire comme présentée précédemment.
- La construction de deux chaînes de Markov, l'une théorique, basée sur l'algorithme MH, permettra d'obtenir la distribution *a posteriori* comme distribution invariante. Elle sera

néanmoins impossible à simuler dans la pratique. L'autre suivra la dynamique de cette première chaîne avec autant de précision que souhaité pour un nombre fini de pas et sera quant à elle simulable en pratique.

Nous détaillons la construction des deux chaînes de Markov utilisées dans cet algorithme ainsi que leurs propriétés de convergence et de contrôle. Les preuves sont données en annexe A.

Construction de la «ideal chain» : Reconsidérons l'algorithme de Metropolis-Hastings pour échantillonner la loi *a posteriori*. Supposons que le système est dans un état θ, l'algorithme choisi d'abord un nouvel état ψ suivant la proposition $q(\theta \to \psi)$. La valeur ψ est alors acceptée avec probabilité

$$\alpha_{ideal}(\theta \to \psi) = \min\left\{1, \frac{p(\psi|\mathbf{x})}{p(\theta|\mathbf{x})}\frac{q(\psi \to \theta)}{q(\theta \to \psi)}\right\}$$

Le noyau de transition de la chaîne de Markov est alors donné par

$$P_{ideal}(\theta, A) = \int_A \alpha_{ideal}(\theta \to \psi)q(\theta \to \psi)\mathbf{1}\{\psi \in A\}d\psi + \mathbf{1}\{\theta \in A\}\left[1 - \int_A \alpha_i(\theta \to \psi)q(\theta \to \psi)d\psi\right]$$

avec $A \in \mathcal{T}_\Theta$.

Le problème de l'algorithme MH présenté plus haut demeure néanmoins pour le calcul du ratio de constantes de normalisation. Comme nous l'avons vu dans les approches à variables auxiliaires, le choix d'une loi de proposition adéquate permet de faire disparaître ce ratio.

Soit $\Delta > 0$, $\nu \in \Theta$ et une réalisation $\mathbf{y}$ du modèle $f(\cdot|\nu)$. Considérons la proposition suivante :

$$q(\theta \to \psi) = q_\Delta(\theta \to \psi|\mathbf{y}) = \frac{\exp(\langle t(\mathbf{y})|\psi\rangle)/c(\psi)}{I(\theta, \Delta, \mathbf{y})}\mathbf{1}_{b(\theta,\Delta/2)}\{\psi\}$$

où $\mathbf{1}_{b(\theta,\Delta/2)}\{\cdot\}$ correspond à l'indicatrice de la boule centrée en θ et de rayon $\Delta/2$ et $I(\theta, \Delta, \mathbf{y}) = \int_{b(\theta,\Delta/2)} f(\mathbf{y}|\Phi)/c(\Phi)d\Phi$. Ce choix de proposition permet de faire converger la chaîne vers $f(\theta|\mathbf{x})$ et évite donc le calcul des constantes de normalisations apparaissant dans α_{ideal}. Néanmoins, le calcul de $I(\theta, \Delta, \mathbf{y})$ est tout aussi difficile que le calcul des constantes de normalisation. Cette «solution» permet la construction de la deuxième chaîne de Markov, la «shadow chain».

Construction de la «shadow chain» :

L'idée pour cette construction se base sur l'approximation de l'intégrale $I(\theta, \Delta, \mathbf{y})$. La solution proposée par les auteurs [Stoica et al., 2017] est de considérer la loi uniforme sur la boule $b(\theta, \Delta/2)$, notons V_Δ son volume et posons $U_\Delta(\theta \to \psi) = \frac{1}{V_\Delta}\mathbf{1}_{b(\theta,\Delta/2)}\{\psi\}$. Cette loi va être utilisée pour proposer de nouvelles valeurs du paramètre au lieu d'utiliser q_Δ comme le justifie le théorème suivant :

Théorème 2.3.1. *Soit* $\mathbf{x}$ *un point de* Ω *tel que la fonction* $f(\mathbf{x}|\phi)$ *soit strictement positive et continue en* ϕ*, on a alors :*

(i) *La distribution de probabilité donnée par la proposition* $q_\Delta(\theta \to \cdot)$ *et* $U_\Delta(\theta \to \cdot)$ *vérifient :* $\forall \theta \in \Theta$ *fixé et* $A \in \mathcal{T}_\Theta$*,*

$$\lim_{\Delta \to 0^+} \int_A |q_\Delta(\theta \to \psi) - U_\Delta(\theta \to \psi)|d\psi = 0.$$

(ii) *Pour tout $\theta \in \Theta$ fixé, les fonctions $\frac{q_\Delta(\theta\to\cdot)}{q_\Delta(\cdot\to\theta)}$ et $\frac{\frac{f(\mathbf{x}|\cdot)}{c(\cdot)}\mathbf{1}_{b(\theta,\Delta/2)}(\cdot)}{\frac{f(\mathbf{x}|\theta)}{c(\theta)}\mathbf{1}_{b(\theta,\Delta/2)}(\theta)}$ vérifient :*

$$\lim_{\Delta\to 0^+}\sup_{\psi\in\Theta}\left|\frac{q_\Delta(\theta\to\psi|\mathbf{x})}{q_\Delta(\psi\to\theta|\mathbf{x})}-\frac{\frac{f(\mathbf{x}|\psi)}{c(\psi)}\mathbf{1}_{b(\theta,\Delta/2)}(\psi)}{\frac{f(\mathbf{x}|\theta)}{c(\theta)}\mathbf{1}_{b(\psi,\Delta/2)}(\theta)}\right|=0$$

*uniformément en $\theta\in\Theta$. De plus, si $f(\mathbf{x}|\cdot)\in\mathcal{C}^1(\Theta)$, les taux de convergence pour **(i)** et **(ii)** peuvent être déterminés.*

D'après ce résultat, il apparaît naturel d'essayer d'approcher le ratio α_{ideal} par un ratio plus simple à calculer, n'utilisant plus $q_\Delta(\theta\to\psi)$ comme loi de proposition. Posons alors

$$\alpha_{shadow}(\theta\to\psi)=\min\left\{1,\frac{f(\psi|\mathbf{x})}{f(\theta|\mathbf{x})}\times\frac{f(\mathbf{y}|\theta)c(\psi)\mathbf{1}_{b(\psi,\Delta/2)}\{\theta\}}{f(\mathbf{y}|\psi)c(\theta)\mathbf{1}_{b(\theta,\Delta/2)}\{\psi\}}\right\}.$$

Remarque :
• Par construction, la «shadow chain» est irréductible et apériodique. En revanche, nous ne savons pas si cette chaîne possède une loi stationnaire.

En appliquant le théorème précédent, nous obtenons

Corollaire 2.3.2. *Les probabilités d'acceptation α_{ideal} et α_{shadow} vérifient : $\forall\theta\in\Theta$,*

$$\lim_{\Delta\to 0^+}\sup_{\psi\in\Theta}|\alpha_{ideal}(\theta\to\psi)-\alpha_{shadow}(\theta\to\psi)|=0.$$

Le ratio de la chaîne théorique peut être approché par celui de la «shadow chain». Pour s'assurer de contrôler la proximité souhaitée entre les deux chaînes, il reste à travailler sur les noyaux de transitions des deux chaînes ainsi créées. La remarque ci-dessus suggère que le comportement asymptotique de la «shadow chain» n'est pas contrôlable avec les résultats à notre disposition. Une alternative est donc de contrôler la dynamique sur un nombre fini d'itérations des noyaux P_{ideal} et P_{shadow}.

Proposition 2.3.3. *Soient P_{ideal} et P_{shadow} les noyaux de transitions respectifs des deux chaînes. Soit $\Delta>0$ et $\mathbf{x}\in\Omega$ comme dans le théorème 2.3.1. Ainsi, pour tout $\epsilon>0$ et tout $n\in\mathbb{N}$, il existe $\Delta_0=\Delta_0(\epsilon,n)>0$ tel que, pour tout $\Delta<\Delta_0$,*

$$|P^{(n)}_{ideal}(\theta,A)-P^{(n)}_{shadow}(\theta,A)|<\epsilon$$

uniformément en $\theta\in\Theta$ et $A\in\mathcal{T}_\Theta$. Si $f(\mathbf{x}|\cdot)\in\mathcal{C}^\infty(\Theta)$, il est possible d'expliciter $\Delta_0(\epsilon,n)$.

Cette proposition établit le contrôle de la dynamique entre les deux chaînes. Ainsi, pour un nombre fini de pas, il est possible d'approximer la dynamique de la chaîne théorique et ainsi échantillonner assez proche de la loi *a posteriori*.

Le fonctionnement de l'algorithme est détaillé ci-dessous.

Algorithme 2.3.5. *(Algorithme ABC Shadow)*

Initialisation : *Poser Δ le paramètre de perturbation, θ_0 comme valeur initiale du paramètre et N_{ABC} comme nombre d'itérations. Supposons que $\mathbf{x}$ soit observé.*

1) *Avec l'algorithme Metropolis-Hastings, générer $\mathbf{y}$ selon $f(\mathbf{y}|\theta_\mathbf{0})$*

2) *Pour $k = 1$ à N_{ABC} :*

a) *Générer un nouveau paramètre ψ selon la densité $U_\Delta(\theta_{k-1} \to \psi)$ définie par*

$$U_\Delta(\theta \to \psi) = \frac{1}{|b(\theta, \Delta/2)|}\mathbf{1}_{b(\theta,\Delta/2}\{\psi\}$$

.

b) *Le nouvel état $\theta_k = \psi$ est accepté avec une probabilité*

$$\alpha_s(\theta_{k-1} \to \psi) = \min\{1, \frac{f(\mathbf{x}|\theta_k)p(\theta_k)}{f(\mathbf{x}|\theta_{k-1})p(\theta_{k-1})} \times \frac{f(\mathbf{y}|\theta_{k-1})}{f(\mathbf{y}|\theta_k)}\}$$

3) *Renvoyer $\theta_{N_{ABC}}$.*

4) *Si davantage d'échantillons sont nécessaires, revenir à l'étape 1 et définir $\theta_0 = \theta_{N_{ABC}}$*

Remarque :
• Dans le cas d'une distribution appartenant à la classe exponentielle, l'observation $\mathbf{x}$ peut être résumée par les statistiques suffisantes du modèle considéré.

2.3.5 Contrôle de l'estimation

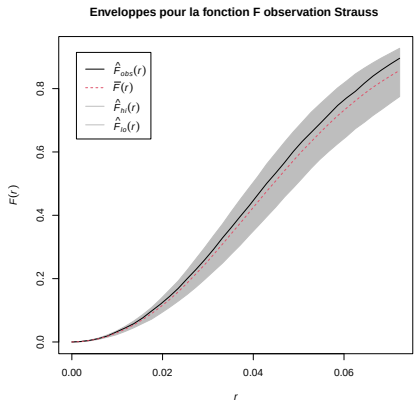
Supposons que nous ayons obtenu une estimation du maximum de vraisemblance $\hat{\theta}$ par l'une des méthodes présentées ci-dessus. Ne pouvant que s'assurer de la pertinence à première vue des valeurs obtenues (éventuellement aberrantes, tendant vers $+\infty$, indiquant une potentielle erreur dans l'exécution des algorithmes), il est important de pouvoir s'appuyer sur des outils théoriques. Pour la qualité de l'ajustement, nous re-mentionnons les tests d'enveloppe MCMC abordés en 2.2.3.2 dans le cas particulier de l'inférence. D'autres méthodes existent telles que l'analyse résiduelle ou les tests d'enveloppe globaux, nous ne les utilisons pas ici. Nous détaillons ensuite des résultats pour obtenir les erreurs asymptotiques entre le vrai maximum de vraisemblance $\hat{\theta}$ et son estimation $\hat{\theta}_n$. L'obtention de ces erreurs se base sur deux résultats de normalité asymptotique [Geyer, 1994, Monfort, 1997].

2.3.5.1 Test d'enveloppe MCMC

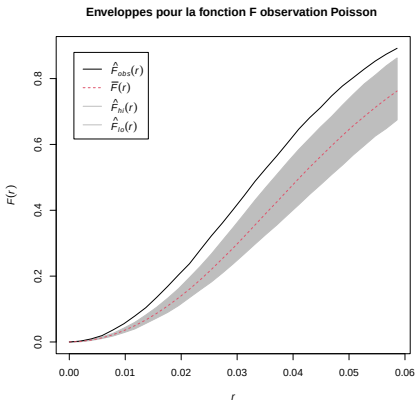
Dans ce cas précis, nous nous plaçons dans la situation suivante : nous supposons que la configuration de point observée est issue d'un processus ponctuel de Gibbs (prenons le modèle de Strauss pour l'exemple). L'inférence est alors menée de manière à trouver une valeur de $\hat{\theta}_n = (\hat{\rho}_n, \hat{\gamma}_{s_n})$.

Par rapport à l'exemple générique donné plus haut, l'hypothèse H_0 est choisie de manière à tester si le paramètre estimé $\hat{\theta}_n$ permet de re-simuler des configurations de points respectant les caractéristiques de la configuration observée (e.g. pour les fonctions de statistiques descriptives spatiales ou pour les statistiques suffisantes du modèle). Ici, nous pouvons donc poser H_0 : «La configuration observée suit un modèle de Strauss de paramètre $\hat{\theta}_n$». De la même manière que dans l'exemple 2.2.2, des réalisations du processus de Strauss de paramètre $\hat{\theta}_n$ sont simulées afin d'obtenir l'enveloppe MCMC pour la statistique d'intérêt. Si l'estimation de la statique pour la configuration observée est à l'intérieur des simulations, ceci indique que le modèle colle correctement aux observations.

Exemple 2.3.1. *Pour illustrer le phénomène, nous prenons deux configurations observées : l'une, une réalisation d'un processus de Strauss de paramètres* $(\rho, \gamma_s, r) = (200, 0.1, 0.05)$ *et l'autre une réalisation du processus de Poisson d'intensité* $\rho = 200$. *Supposons que les deux réalisations sont issues d'un modèle de Strauss de rayon* $r = 0.05$ *et que l'inférence paramétrique nous donne* $\hat{\theta}_n = (205, 0.12)$ *pour la première réalisation et* $\hat{\theta}_n = (196, 0.5)$ *pour la deuxième. (Les valeurs sont ici prises de manière arbitraire).*



(a) Enveloppe MCMC obtenue par 100 simulations du modèle de Strauss de paramètres $\hat{\theta}_n = (205, 0.12)$. Estimation de F sur l'observation (noir) ; moyenne de F sur les simulations (rouge) ; enveloppe MCMC (gris).



(b) Enveloppe MCMC obtenue par 100 simulations du modèle de Strauss de paramètres $\hat{\theta}_n = (196, 0.5)$. Estimation de F sur l'observation (noir) ; moyenne de F sur les simulations (rouge) ; enveloppe MCMC (gris).

Dans le premier cas, nous voyons que l'estimation permet de coller aux caractéristiques du pattern initial (l'estimation pour ce dernier est bien à l'intérieur de l'enveloppe générée par les simulations). Dans le deuxième cas, la courbe de l'estimation de la configuration est en dehors de l'enveloppe, indiquant que la configuration ne peut être expliquée correctement par un modèle de Strauss de paramètres $\hat{\theta}_n = (196, 0.5)$.

2.3.5.2 Erreurs asymptotiques standard et MCMC

Le cadre de travail le plus connu est peut-être celui d'un $n-$échantillon. Nous observons $x_1, x_2, \ldots, x_n$ que l'on suppose des réalisations de $X_1, X_2, \ldots, X_n$, indépendantes et identiquement distribuées selon densité de probabilité $f(x|\theta)$ avec $\theta = \theta_0$ inconnu. L'estimateur de maximum de vraisemblance de θ est noté dans ce cas $\hat{\theta}$. Le résultat suivant nous donne la possibilité de construire des erreurs contrôlant la différence entre le vrai paramètre du modèle et son esti-

mation par maximum de vraisemblance, quand la taille n de l'échantillon tend vers l'infini. [Monfort, 1997, Gourieroux and Monfort, 1989].

Théorème 2.3.4. *Supposons que*
i) *L'intérieur de* Θ *est non vide convexe de* $\mathbb{R}^d$ *et* θ_0 *appartient à l'intérieur de* Θ.
ii) *Le modèle est identifiable.*
iii) *La log-vraisemblance* $L_n(x|\theta)$ *est continue en* θ.
iv) $\mathbb{E}_{\theta_0}[\log f(X_i|\theta)]$ *existe.*
v) *La* $\log-$*vraisemblance est telle que* $(1/n)L_n(x|\theta)$ *converge presque sûrement vers* $\mathbb{E}_{\theta_0}[\log f(X_i|\theta)]$ *uniformément en* $\theta \in \Theta$.
vi) *La* $\log-$*vraisemblance est deux fois différentiable dans un voisinage ouvert de* θ_0.
vii) *La matrice* $I_1(\theta_0) = \mathbb{E}_{\theta_0}[\frac{-\partial^2 \log f(X_1|\theta_0)}{\partial\theta\partial\theta'}]$ *existe et est non singulière.*
Alors

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, I_1(\theta_0)^{-1}).$$

Dans le cadre de travail de cette thèse, une seule configuration de points est observée, donc un seul échantillon. En plus, nous devons faire appel à la simulation Monte Carlo pour approcher la fonction de vraisemblance. Le théorème suivant nous donne la possibilité de construire des erreurs contrôlant la différence entre l'estimateur par maximum de vraisemblance et son équivalent Monte Carlo construit à partir de n simulations, $\hat{\theta}_n$ [Geyer, 1994, Geyer, 1999].

Théorème 2.3.5. *Supposons que*
i) $\hat{\theta}$ *est unique et l'espace des paramètres* Θ *contient un voisinage ouvert de* $\hat{\theta}$ *dans* $\mathbb{R}^d$.
ii) $\hat{\theta}_n$ *converge en probabilité vers* $\hat{\theta}$.
iii) $\theta \mapsto Z(\theta)$ *est deux fois différentiable sous le signe intégrale.*
iv) $\sqrt{n}\nabla l_n(\hat{\theta}) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, A)$ *pour une certaine matrice de variance-covariance* A.
v) $B = -\nabla^2 l(\hat{\theta})$ *est définie positive.*
vi) *La probabilité que* $\nabla^3 l_n(\theta)$ *soit borné est égale à 1 uniformément en* θ *sur un voisinage de* $\hat{\theta}$.
Alors

$$-\nabla^2 l_n(\hat{\theta}_n) \longrightarrow B, \text{ en probabilité,}$$

et

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, B^{-1}AB^{-1}).$$

Comme les auteurs le précisent, l'hypothèse **iv)** est la plus délicate. En effet, si une méthode MCMC est utilisée pour les simulations, elle repose sur un Théorème Central limite pour la chaîne de Markov utilisée. Supposant que cette condition soit vérifiée, la variance de A ne peut pas être calculée et nécessite d'être estimée par des méthodes Monte Carlo. Ainsi, pour estimer B, nous prenons donc

$$B_n = -\nabla^2 l_n(\hat{\theta}_n)$$

donné par la formule (2.12) et pour estimer A, nous allons prendre $A_n = \frac{C_n}{d_n^2}$ avec

$$C_n = (t(\mathbf{x}) - t(X)) \exp\left[t(X)(\hat{\theta}_n - \psi)^T\right],$$

la matrice de variance covariance empirique de l'échantillon $X_1, ..., X_n$ et

$$d_n = \frac{1}{n}\sum_{i=1}^{n} e^{\langle t(X_i), \hat{\theta}_n - \psi\rangle}.$$

À partir de ce théorème, en faisant l'hypothèse que nous avons obtenu une estimation de $\hat{\theta}$ par MCMLE, des erreurs asymptotiques peuvent être obtenues [Geyer, 1994, Geyer, 1999, van Lieshout and Stoica, 2003, Stoica, 2025].

À noter également, les auteurs de [Dereudre and Lavancier, 2016] établissent la convergence de l'estimateur du maximum de vraisemblance dans le cadre des processus de Gibbs généraux lorsque le domaine d'observation est étendu vers l'infini. Les auteurs supposent que l'on peut calculer sans erreur le maximum de vraisemblance dans une fenêtre de dimensions données pour un modèle qui puisse avoir des interactions. Pour les modèles que nous considérons dans cette thèse cette étape nécessite des approximations Monte Carlo du maximum de vraisemblance comme décrit précédemment. Ceci introduit des erreurs qui doivent être contrôlées. A priori, le résultats de convergence présentés par les auteurs ne montrent pas comment gérer ce type d'erreurs. Arriver à maîtriser cet aspect pourrait peut être permettre d'évaluer la variance de ces estimateurs et ainsi ce résultat pourrait constituer une piste très intéressante lorsque qu'une observation d'un très grand jeu de donnée est fournie.

2.3.6 Réglage des paramètres et exemples d'utilisation d'ABC Shadow

Nous allons nous focaliser sur le paramétrage de l'algorithme ABC Shadow pour en comprendre le fonctionnement. Nous choisissons d'illustrer l'impact du choix des paramètres sur l'inférence et d'expliquer comment se prémunir de certains phénomènes pouvant intervenir lors de l'inférence. Le tableau ci-dessous résume les paramètres de l'algorithme décrit en 2.3.5. Puisque l'algorithme de Metropolis-Hastings est utilisé pour simuler le pattern auxiliaire, les paramètres de ce dernier se retrouvent également ici. Dans la suite, les paramètres seront donnés en échelle logarithmique, nécessaire à la stabilité numérique de l'algorithme.

Variable	Description
W	Domaine de simulation
N_θ	Nombre d'échantillons de θ
N_{ABC}	Nombre d'itérations proposition paramètre
N_{MH}	Nombre d'itérations MH
p_b, p_d	Probabilité d'ajout/retrait
$\theta_0 = (\rho_0, \gamma_{s_0})$	Paramètres initiaux
$\Delta = (\Delta_\rho, \Delta_\gamma)$	Paramètres de perturbation de θ
$[\rho_{min}, \rho_{max}]$	Support de la loi uniforme sur ρ
$[\gamma_{min}, \gamma_{max}]$	Support de la loi uniforme sur γ_s
$n(\mathbf{x})$	Nombre de points observés
$s_r(\mathbf{x})$	Nombre de $r-$voisins observés

Dans la suite, nous allons considérer une réalisation du modèle de Strauss de paramètres $\theta = (200, 0.1)$ (ce qui donne $\log\theta = (5.30, -2.30)$) et de rayon $r = 0.05$ et faire l'estimation des paramètres en supposant que ce que nous observons est une réalisation du modèle de Strauss. Nous fixons les paramètres $W = [0,1] \times [0,1]$ et $N_\theta = 10000$ de sorte à obtenir un $10000-$échantillon de la loi *a posteriori*. Cette loi est une restriction de la vraisemblance du modèle à $[\rho_{min}, \rho_{max}] \times [\gamma_{min}, \gamma_{max}]$ (en échelle logarithmique $[0,9] \times [-5,0]$). Ces intervalles sont suffisamment larges pour s'assurer que la dynamique de la chaîne de Markov permette que la vraisemblance et la loi

a posteriori soient suffisamment proches. Les paramètres de l'algorithme de Metropolis-Hastings sont fixés à $N_{MH} = 1000$ et $p_b = 0.5, p_d = 0.5$.

2.3.6.1 Choix des paramètres Δ et N_{ABC}

Ces deux paramètres ont une influence relativement similaire : Δ contrôle à quel point le nouveau paramètre candidat est «éloigné» du paramètre courant là où N_{ABC} représente le nombre de propositions à la suite pour le paramètre candidat. Ainsi, si ces deux valeurs sont trop grandes par rapport à l'échelle du paramètre, il est possible que l'échantillonnage de la loi *a posteriori* «rate» la zone d'intérêt comme l'illustre les figures en 2.18 ci-dessous. Les valeurs sont données volontairement sous formes de points pour faciliter la lisibilité.

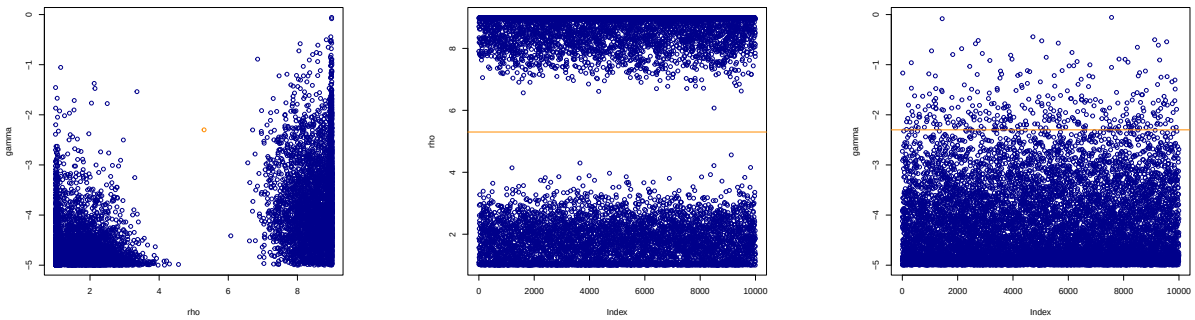


FIGURE 2.18 – La figure de gauche représente la distribution jointe des paramètres estimés $(\log \rho, \log \gamma_s)$. Les deux autres figures représentent les séries temporelles des paramètres $\log \rho$ et $\log \gamma_s$ respectivement. Ces résultats ont été obtenus pour $N_{ABC} = 100$ et $\Delta = (0.05, 0.05)$ et $\theta_0 = (4.5, -0.5)$. Les valeurs en orange représentent les vraies valeurs des paramètres $(5.30, -2.30)$.

À l'inverse, si ces paramètres N_{ABC} et Δ sont trop «petits», la dynamique de la chaîne de Markov ne semble pas échantillonner correctement l'espace des paramètres et demanderait alors un nombre d'itérations très grand pour s'assurer du bon échantillonnage comme l'illustrent les figures en 2.19.

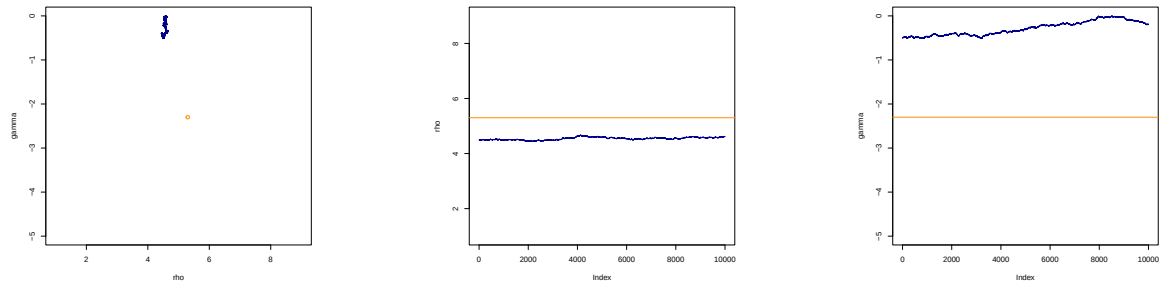


FIGURE 2.19 – La figure de gauche représente la distribution jointe des paramètres estimés $(\log \rho, \log \gamma_s)$. Les deux autres figures représentent les séries temporelles des paramètres $\log \rho$ et $\log \gamma_s$ respectivement. Ces résultats ont été obtenus pour $N_{ABC} = 10$ et $\Delta = (0.001, 0.001)$ et $\theta_0 = (4.5, -0.5)$. Les valeurs en orange représentent les vraies valeurs des paramètres $(5.30, -2.30)$.

Enfin, un phénomène «de bord» peut survenir lorsque la valeur d'un paramètre (e.g. γ_s) est trop proche de 1 (et donc trop proche de 0 en échelle logarithmique). Supposons ici seulement que la configuration observée soit issue d'un modèle de Strauss de paramètres $(200, 0.9)$ ($(5.30, -0.11)$ en échelle logarithmique) et $r = 0.05$. Si Δ est pris trop grand, l'échantillonnage semblera «rebondir» sur la borne supérieure comme l'illustre l'histogramme de gauche ci-dessous, le mode de l'histogramme est alors très proche de 0. Il convient alors de faire des tests sur la valeur de Δ pour éviter ce phénomène. La figure de droite illustre un choix de Δ permettant de pallier ce problème.

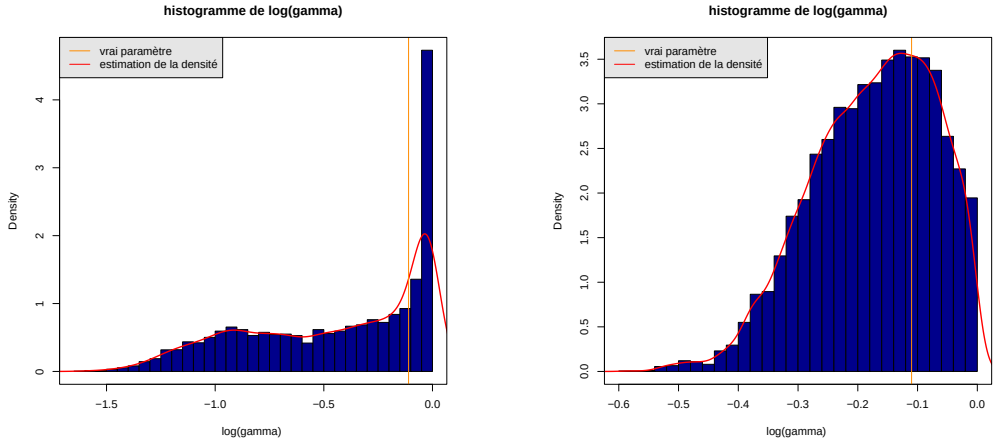


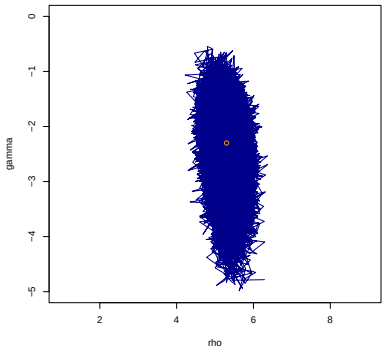
FIGURE 2.20 – Histogramme pour $\log \gamma_s$ pour $\Delta_\gamma = 0.01$ (gauche) ; Histogramme pour $\log \gamma_s$ pour $\Delta_\gamma = 0.001$ (droite) ; Les lignes orange verticales représentent les vrais paramètres $\log \gamma_s = -0.11$.

2.3.6.2 Exemple sur le modèle de Strauss

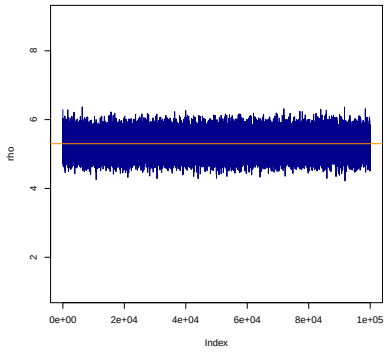
Enfin, nous donnons un exemple d'inférence pour le modèle de Strauss lorsque plusieurs réalisations sont observées. Ceci permet d'affiner les valeurs d'entrée de l'algorithme ABC Shadow qui sont alors la moyenne des statistiques suffisantes du modèle considéré sur les différentes observations. Nous choisissons également des valeurs initiales volontairement éloignées des vraies valeurs pour illustrer l'indépendance des conditions initiales. Nous résumons les paramètres utilisés dans le tableau ci-dessous

Variable	Description	Valeur
W	Domaine de simulation	$[0,1] \times [0,1]$
N_θ	Nombre d'échantillons de θ	10^5
N_{ABC}	Nombre d'itérations proposition paramètre	100
N_{MH}	Nombre d'itérations MH	10^3
p_b, p_d	Probabilité d'ajout/retrait	$0.5, 0.5$
$\theta_0 = (\rho_0, \gamma_{s_0})$	Paramètres initiaux	$(8.5, -4.5)$
$\Delta = (\Delta_\rho, \Delta_\gamma)$	Paramètres de perturbation de θ	$(0.01, 0.01)$
$[\rho_{min}, \rho_{max}]$	Support de la loi uniforme sur ρ	$[1, 9]$
$[\gamma_{min}, \gamma_{max}]$	Support de la loi uniforme sur γ_s	$[-5, 0]$
$\overline{n(\mathbf{x})}$	Moyenne du nombre de points observés	94.34
$\overline{s_r(\mathbf{x})}$	Moyenne du nombre de $r-$voisins observés	4.81

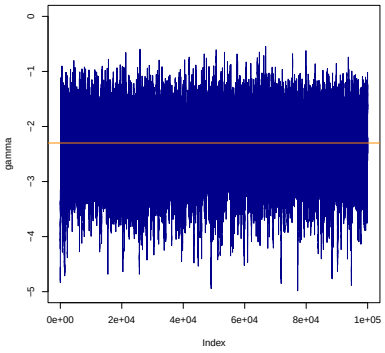
Les figures suivantes illustrent l'estimation de la loi *a posteriori* pour la loi jointe, les lois marginales et les histogrammes des paramètres.



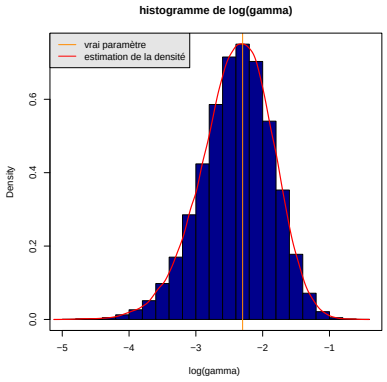
(a) Distribution jointe des paramètres



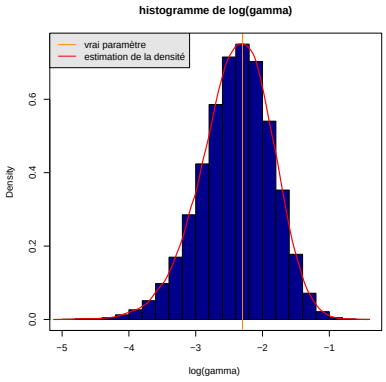
(b) Série temporelle de $\log \rho$



(c) Série temporelle de $\log \gamma_s$



(d) Histogramme de $\log \rho$



(e) Histogramme de $\log \gamma_s$

Dans tous les cas, nous observons que l'échantillonnage s'effectue correctement autour des vrais paramètres. Pour les deux histogrammes, le vrai paramètre correspond au mode de chaque histogramme. Le maximum *a posteriori* (MAP) donne alors $(5.30, -2.31)$, valeur très proche des vrais paramètres.

Nous avons désormais toutes les clés à notre disposition pour comprendre les travaux de cette thèse. Les chapitres suivants présentent ces travaux abordant la modélisation des données cosmologiques et une approche méthodologique pour l'estimation des paramètres en données partiellement observées.

Chapitre 3

Quelques modèles pour la caractérisation de la distribution des galaxies dans l'Univers

Sommaire

Contexte

Il est commun en cosmologie de tester les outils de détection, de simulation et d'inférence sur des données simulées. Ces données, bien moins coûteuses à obtenir que des relevés à grande échelle obtenus à l'aide de télescopes, sont souvent obtenues par des simulations à N corps («N-body simulation» en anglais), simulations basées sur les lois de la physique et les réalisations obtenues grâce à cette technique sont assez proches de la réalité. Les approches de [Stoica et al., 2005b, Stoica et al., 2007, Stoica, 2010] ont d'abord été appliquées à des données simulées pour ensuite être appliquées à des données réelles [Stoica et al., 2017, Hurtado-Gil et al., 2021]. Dans notre démarche, nous avons suivi une démarche similaire «du simple vers le complexe». Notre étude méthodologique est mise en place sur des données cosmologiques 2D obtenues à partir de la simulation des modèles physiques. Participant au projet européen EXCOSM, l'objectif est d'appliquer ces méthodes aux données 4MOST qui sortiront dans un futur proche. Les raisons de se

cantonner à la modélisation en 2D sont multiples :

— D'un point de vue complexité de modélisation, ces données apparaissent plus simples à étudier pour l'analyse descriptive. S'assurer du bon fonctionnement des modèles en 2D est déjà un premier jalon avant de les généraliser.
— D'un point de vue computationnel et expérimentations, les modèles utilisés, même 2D, peuvent nécessiter des temps de calculs assez longs allant de plusieurs minutes à plusieurs heures pour la procédure d'inférence par ABC Shadow.

Pour ces raisons, les modélisations proposées dans ce chapitre se restreignent à des données simulées venant de simulations physiques et de données observées en 2D.

Les données extraites de ces jeux de données restent d'une taille «raisonnable» et ne sont constituées que de quelques centaines de points représentant la position des galaxies. De plus, au vu de la pluralité morphologique de ces données (clusters, zones de vides, filaments, murs) et des différentes modélisations dont ces structures pourraient faire l'objet, il paraît judicieux de focaliser la modélisation sur des extractions de ces jeux de données plutôt que de les prendre dans leur globalité.

Ce chapitre présente des modélisations proposées dans cette thèse pour décrire la distribution des galaxies. Ces travaux ont fait l'objet de deux articles dans des proceedings de conférence.

Le travail de modélisation de cette thèse se place dans un contexte très riche. Premièrement, des généralisations des modèles connus sont à mentionner comme abordé dans [Gregori et al., 2004] ou le modèle Bisous [Stoica et al., 2005a]. Ces travaux posent un cadre très général pour la construction des modèles des processus ponctuels avec interactions complexes à partir de la superposition des modèles plus simples construits à partir de toute interaction par paires symétrique et réflexive. Cela permet d'étudier la connectivité des réseaux quand il s'agît des filaments cosmiques ou des routes [Stoica, 2025, Stoica et al., 2004, Tempel et al., 2016], la caractérisation morphologique et statistique des clusters en épidémiologie animale ou dans la distribution des galaxies [Stoica, 2025, Stoica et al., 2007, Tempel et al., 2018]. Ensuite, il faut considérer le cadre de travail présentée par les [Baddeley et al., 2013b] ou comme précédemment des superpositions des modèles d'interactions simples sont utilisées pour construire des modèles présentant des structures complexes. Notre approche de modélisation est la suite de [Hurtado-Gil et al., 2021]. Ces travaux ajustent des modèles des processus ponctuels inhomogènes avec interactions à des jeux des données provenant de la cosmologie. Une des pistes lancées par les auteurs étaient l'étude des modèles avec interactions multiples.
Une autre motivation réside dans le fait que la superposition des modèles ne s'est effectuée que sur des voisinages directs autour des points, induisant une forme de «compétition» entre les interactions proposées par les différents modèles. Pour pallier cela, nous proposons deux autres modélisations basées sur des modélisations que nous appellerons «multi-échelle» dans la suite. Ces modélisations reposent sur la modification de modèles introduits dans le chapitre 2 et permettent d'éviter ce phénomène de compétition en considérant une seule composante d'interaction (e.g. Strauss, Area-Interaction...) pour différents voisinages pris autour des points. Les constructions des modèles proposées dans cette thèse peuvent être vues comme une synthèse des cadres de modélisation mentionnés précédemment.
La section 3.1 introduit un modèle «multi-interaction» obtenu en superposant les modèles de Strauss et Area-Interaction tout en prenant en compte une inhomogénéité induite par une confi-

guration de filaments galactiques. Nous commençons par présenter les données utilisées dans l'article [Stoica et al., 2005b]. L'extraction d'une configuration et le traitement de ces données à l'aide du modèle Candy sont expliqués en section 3.1.1. La section 3.1.2 introduit le modèle proposé et détaille le traitement informatique utile à sa simulation et à son inférence. Enfin, la section 3.1.3 présente les résultats obtenus pour l'inférence menée sur ce jeu de données.
La section 3.2 présente deux modélisations obtenues en modifiant des modèles existants dans la littérature pour obtenir des modélisations «multi-échelles». Nous commençons par présenter les données et introduisons le modèle que nous intitulerons «StraussCrown» qui sera utilisé pour la modélisation présentée en 3.2.2. Ensuite, une analyse descriptive est présentée afin de déterminer les rayons des modèles utilisés puis les résultats pour cette première modélisation sont donnés et contrôlés par un test d'enveloppe MCMC. La section 3.2.5 présente un second modèle reposant sur les résultats précédents puis nous présentons les résultats obtenus pour ce deuxième modèle. Enfin, nous discutons des résultats et d'éventuelles perspectives.

3.1 Modélisation à interactions multiples conditionnellement à des structures galactiques

3.1.1 Présentation, extraction et prétraitement

Ce premier jeu de données est une réalisation en 2D obtenue par simulation de données galactiques : elle est constituée de 4247 galaxies réparties sur un domaine $[0, 128] \times [0, 128]$. De ce domaine, nous avons extrait un sous-domaine $W = [0, 30] \times [0, 30]$ représenté en rouge ci-dessous. Ce sous-domaine contient 334 points et contient plusieurs structures intéressantes à prendre en compte pour la modélisation.

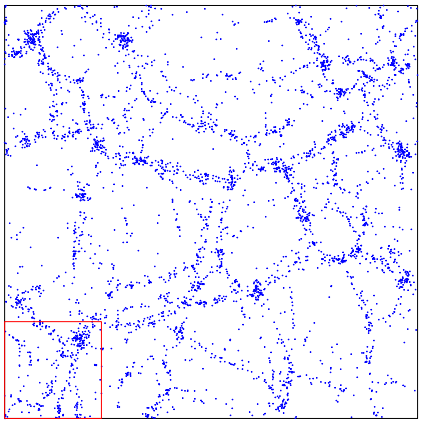
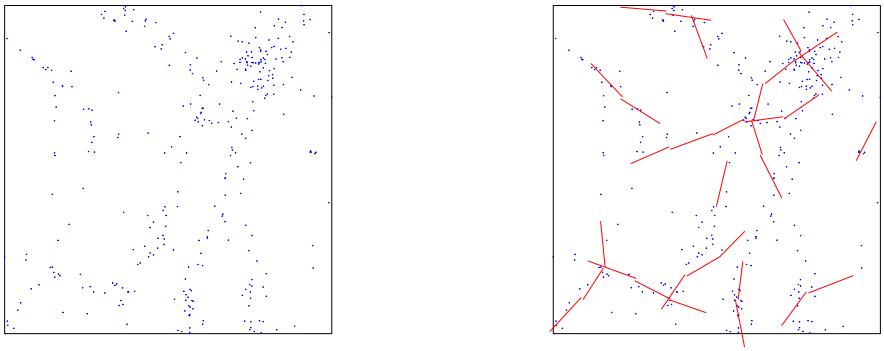


FIGURE 3.1 – Données simulées, chaque point représente une galaxie.

Ces données ont d'abord été utilisées afin de mettre en place la détection de filaments galactiques par des processus ponctuels marqués à l'aide du modèle Candy [Stoica et al., 2005b]. Une réalisation de ce processus constitue une collection de cylindres (ici en 2D, des rectangles) re-

présentés par des segments auxquels sont attachés des marques, décrivant les caractéristiques de chaque rectangle. Plus précisément, les segments s'ajustent et se connectent dans les régions où les positions de galaxies tendent à former des structures linéiques. Plus de détails sur ce modèle et le modèle Bisous sont donnés dans le chapitre 3 de [Stoica, 2025]. Notre approche se base spécifiquement sur les filaments car les travaux de [Tempel et al., 2014] mettent en avant des preuves statistiques sur le fait que les galaxies sont distribuées comme les perles d'un collier le long des filaments, montrant l'importance de ces derniers dans la répartition des galaxies.

C'est avec ce modèle que les filaments ont été détectés à partir des galaxies observées de la fenêtre extraite. La figure suivante illustre plus en détail la fenêtre extraite ainsi que les filaments détectés correspondants.



(a) Galaxies extraites. (b) Filaments détectés correspondants.

FIGURE 3.2

Les critères d'extractions, bien qu'arbitraires, sont les suivants :

- La zone contient suffisamment de points pour que les filaments détectés représentent une part substantielle du domaine.
- Les galaxies extraites sont assez représentatives de l'agencement des galaxies à plus grande échelle : nous observons un cluster de galaxies en haut à droite, des zones à moins forte densité de galaxies sur le bas du domaine ainsi que des zones où la densité est très faible.

Nous avons choisi d'extraire un sous-domaine suffisamment «riche» pour que notre modélisation ait du sens mais également assez «petit» pour que les temps de calculs soient raisonnables. D'autres extractions auraient pu être effectuées en considérant ces mêmes critères.

Pour les mêmes raisons de stabilité numériques expliquant l'utilisation de l'échelle logarithmique pour l'inférence, les données sont remises dans le domaine $W = [0,1] \times [0,1]$, sans perte de généralité. Cela permet également de travailler avec des échelles habituelles de distances d'interactions (e.g pour le choix des rayons).

Désormais, plusieurs options s'offrent à nous pour déterminer la distance au plus proche filament. Une première idée consiste à prendre, pour chaque point, le projeté orthogonal sur les différents filaments puis de prendre celui réalisant le minimum de toutes ces distances. Cette solution

peut s'avérer viable dans ce cas, où seule une vingtaine de filaments sont considérés, mais peut s'avérer lourd en terme d'opérations numériques. L'autre solution consiste à discrétiser le domaine et à déterminer la distance aux filaments pour chaque point de cette discrétisation. La fonction `distmap` de `spatstat` permet, à partir d'un processus ponctuel ou de segments (ici, les filaments), d'obtenir cette discrétisation. La figure ci-dessous illustre les distances les plus proches entre les points du domaine et les filaments.

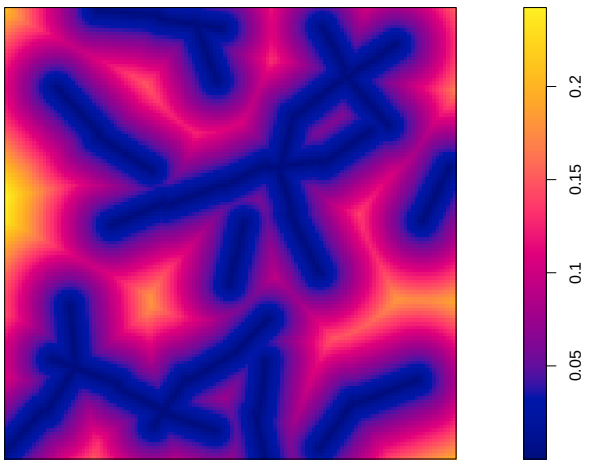


FIGURE 3.3 – Plus proche distance entre les points du domaine et les filaments. Les zones en bleu représentent les zones proches des filaments et les zones de couleurs chaudes représentent les zones éloignées des filaments.

Cette approche permet de stocker ces distances sous forme d'une matrice et de s'y référer pour chaque point du processus à partir de ses coordonnées. La distance au filament est alors approchée par le point de coordonnées (i, j) le plus proche du point du processus dans le domaine.

3.1.2 Présentation du modèle

En se basant sur ces distances, le modèle obtenu en superposant des modèles existants est décrit ci-après par la densité suivante :

$$f(\mathbf{x}|\rho, \gamma_s, \gamma_a) \propto \exp\left(w(\mathbf{x}) \log \rho + s_{r_s}(\mathbf{x}) \log(\gamma_s) + a_{r_a}(\mathbf{x}) \log(\gamma_a)\right) \tag{3.1}$$

Nous décomposons la superposition des différents modèles en «composantes» pour faciliter la compréhension de son introduction :

— Composante Poisson : inhomogénéité qui prend en compte $d(\xi, F)$, la distance la plus proche entre un point $\xi \in W$ et le réseau de filaments détecté. Cette distance est donc représentée par la figure 3.3 ci-dessus. La statistique suffisante associée à cette composante

est :

$$w(\mathbf{x}) = \sum_{i=1}^{n(\mathbf{x})} \mathbf{1}_{d(\xi_i,F)\leq 0{,}05}(\xi) \times \frac{1}{1+d(\xi_i,F)}.$$

— Composante Strauss : le modèle de Strauss est considéré sur le r_s-voisinage des points.
— Composante Area-Interaction : le modèle Area-Interaction est considéré en prenant les boules centrées en les points de rayon r_a.

Le premier terme de l'inhomogénéité assure que les points soient obligatoirement proches des filaments. Ceci est garanti par la fonction indicatrice $\mathbf{1}_{d(\xi_i,F)\leq 0{,}05}(\cdot)$, quoiqu'un peu «brutal», ce choix est à nouveau motivé par les aspects numériques de cette modélisation. Le choix de la valeur $r_a = 0.05$ permet de limiter les temps de calculs de la statistique du modèle Area-Interaction. Ce choix est aussi cohérent par rapport au nombre de galaxies vérifiant cette relation sur la configuration donnée en Figure 3.2b, 255 des 334 galaxies sont à une distance inférieure à 0.05 d'un filament. Le deuxième terme est quant à lui choisi décroissant en $d(\xi_i, F)$, comme le suggèrent les changements de pentes de la courbe ci-dessous, obtenue en traçant le nombre de galaxies vérifiant $d(\xi, F) \leq r$ pour r compris entre 0 et 0.15, avec un pas de 0.005. Enfin, pour calculer la statistique liée à la composante de Poisson, nous utilisons la discrétisation induite par la matrice des distances donnée plus haut. Cette matrice est de taille 128×128, ainsi, pour un point de coordonnées (ξ_1, ξ_2), nous trouvons le point associé à la matrice en prenant la partie entière de $\xi_1 \times 128 + 1$ et $\xi_2 \times 128 + 1$.

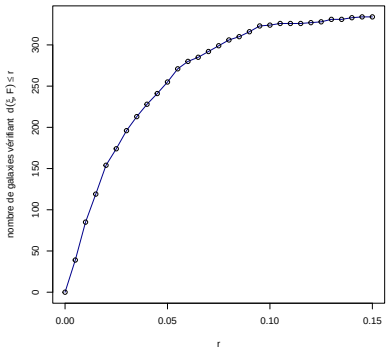


FIGURE 3.4 – Fonction du nombre de galaxies à distance inférieure ou égale à r en fonction de r. Les pentes entre deux rayons successifs indiquent une fonction de moins en moins croissante.

Ce choix permet alors de s'assurer que les galaxies sont distribuées proches des filaments et qu'elles ont une tendance à favoriser la proximité avec ces derniers.

Le choix de superposition des modèles de Strauss et Area-Interaction est motivé par la détection de potentielle «sous-interactions» présentes dans le phénomène d'agglomération autour des filaments. Plus précisément, l'idée est de contrôler le phénomène d'agglomération à l'aide du processus Area-Interaction et de contrôler la répartition de ces points à l'intérieur de ces clusters à l'aide du processus de Strauss. Les valeurs estimées pour ces paramètres permettent alors d'indiquer un potentiel phénomène d'agglomération ou de répulsion entre les galaxies distribuées autour des filaments. Les choix des rayons r_s et r_a ne sont néanmoins pas clairs, nous avons donc considéré 9 cas pour le couple (r_s, r_a) : toutes les combinaisons pour $r_s \in \{0.01, 0.03, 0.05\}$ et $r_a \in \{0.01, 0.03, 0.05\}$. Le tableau ci-dessous donne les statistiques observées qui seront utilisées

comme observations en utilisant l'algorithme ABC Shadow.

r_s, r_a	0.01	0.03	0.05
$n(\mathbf{x})$	334	334	334
$s_{r_s}(\mathbf{x})$	71	539	1268
$-a_{r_a}(\mathbf{x})$	272	143	83

Nous exposons dans la section suivante les résultats de l'échantillonnage de la loi *a posteriori*, pour p une loi uniforme détaillée dans le tableau ci-dessous.

$$f(\rho, \gamma_s, \gamma_a|\mathbf{x}) \propto \exp\left(w(\mathbf{x}) \log \rho + s_{r_s}(\mathbf{x}) \log(\gamma_s) + a_{r_a}(\mathbf{x}) \log(\gamma_a)\right) p(\rho, \gamma_s, \gamma_a) \tag{3.2}$$

3.1.3 Résultats

Le tableau ci-dessous résume le réglage des paramètres pour l'algorithme ABC Shadow, pour chaque couple de rayons (r_s, r_a) parmi ceux indiqués ci-dessus, l'algorithme ABC Shadow a été initialisé avec les statistiques suffisantes du modèle observé. Δ a d'abord été fixé à $(0.01, 0.01, 0.01)$, nous avons rencontré le phénomène «de bord» illustré en Figure 2.20 pour l'échantillonnage de $\log(\gamma_s)$. Pour cette raison, le paramètre de perturbation de γ_s est plus petit que celui des autres paramètres.

Variable	Description	Valeur
W	Domaine de simulation	$[0,1] \times [0,1]$
N_θ	Nombre d'échantillons de θ	10^5
N_{ABC}	Nombre d'itérations proposition paramètre	100
N_{MH}	Nombre d'itérations MH	10^3
p_b, p_d	Probabilité d'ajout/retrait	$0.5, 0.5$
$\theta_0 = (\rho_0, \gamma_{s_0}, \gamma_{a_0})$	Paramètres initiaux	Aléatoire
$\Delta = (\Delta_\rho, \Delta_{\gamma_s}, \Delta_{\gamma_a})$	Paramètres de perturbation des paramètres	$(0.01, 0.001, 0.01)$
$[\rho_{min}, \rho_{max}]$	Support de la loi uniforme sur ρ	$[0, 10]$
$[\gamma_{s,min}, \gamma_{s,max}]$	Support de la loi uniforme sur γ_s	$[-10, 0]$
$[\gamma_{a,min}, \gamma_{a,max}]$	Support de la loi uniforme sur γ_a	$[-10, 10]$

Ainsi, pour chacun des 9 choix de couples de rayons, nous obtenons un 10^5-échantillon de la loi *a posteriori* des paramètres. Les figures ci-dessous illustrent cette inférence pour $(r_s, r_a) = (0.03, 0.03)$ et nous résumons les résultats sous la forme de box-plot pour différentes valeurs de r_s et d'un tableau avec l'erreur MCMC de chaque paramètre estimé obtenue pour 10^3 simulations à partir des paramètres estimés.

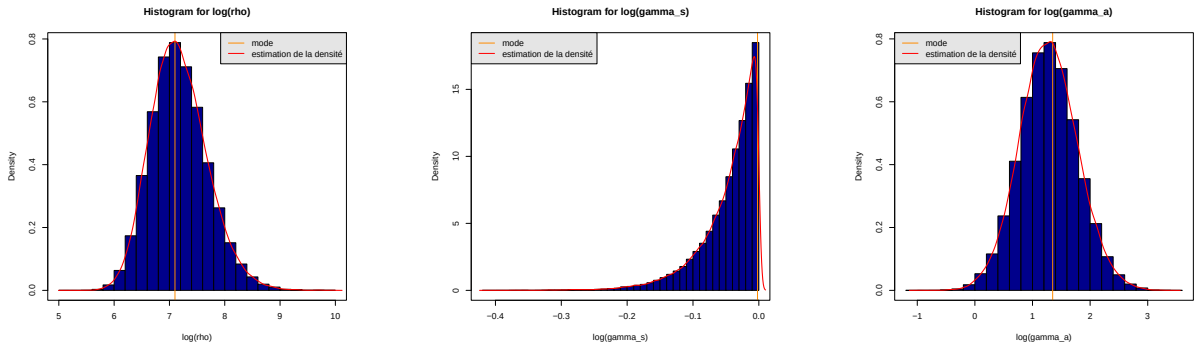


FIGURE 3.5 – Histogrammes de la distribution *a posteriori* des marginales de $\log(\rho)$ (gauche), $\log(\gamma_s)$ (milieu) et $\log(\gamma_a)$ (droite) pour $(r_s, r_a) = (0.01, 0.03)$. En rouge l'estimation de la densité obtenue par la fonction `density` de `R` et en orange le mode de l'histogramme.

Pour $(r_s, r_a) = (0.01, 0.03)$, le paramètre estimé pour $\log \gamma_s$ est très proche de 0, indiquant que le phénomène de répulsion est probablement négligeable entre les galaxies autour des filaments. Ce phénomène se répète pour tous les choix de couples de rayons comme l'indiquent les box-plots qui suivent, à l'exception de $(r_s, r_a) = (0.01, 0.01)$.

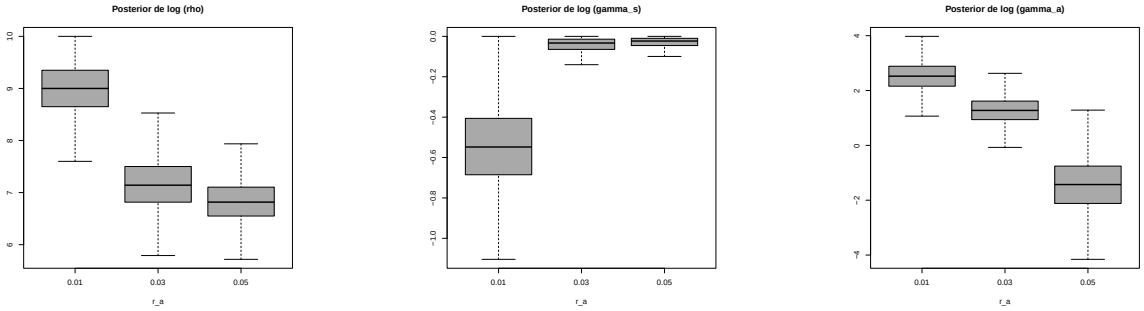


FIGURE 3.6 – Box-plot de la distribution *a posteriori* des marginales de $\log(\rho)$ (gauche), $\log(\gamma_s)$ (milieu) et $\log(\gamma_a)$ (droite) pour $r_s = 0.01$ en fonction de r_a. Pour chaque triplet de box-plots : $r_a = 0.01$ (gauche), $r_a = 0.03$ (milieu) et $r_a = 0.05$ (droite).

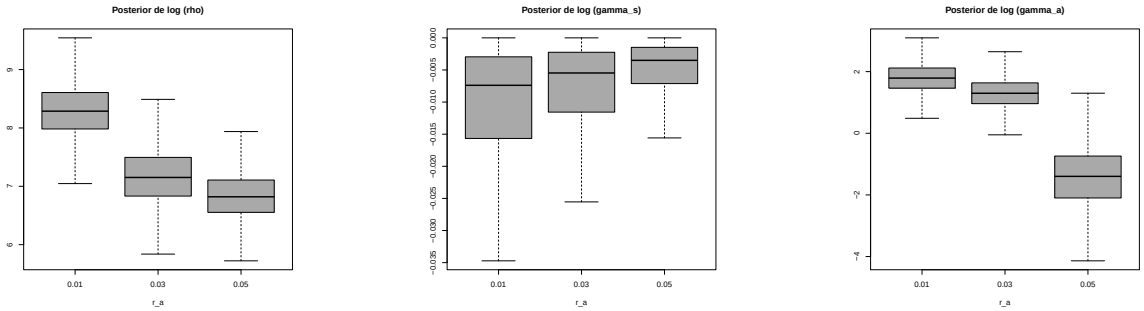


FIGURE 3.7 – Box-plot de la distribution *a posteriori* des marginales de $\log(\rho)$ (gauche), $\log(\gamma_s)$ (milieu) et $\log(\gamma_a)$ (droite) pour $r_s = 0.03$ en fonction de r_a. Pour chaque triplet de box-plots : $r_a = 0.01$ (gauche), $r_a = 0.03$ (milieu) et $r_a = 0.05$ (droite).

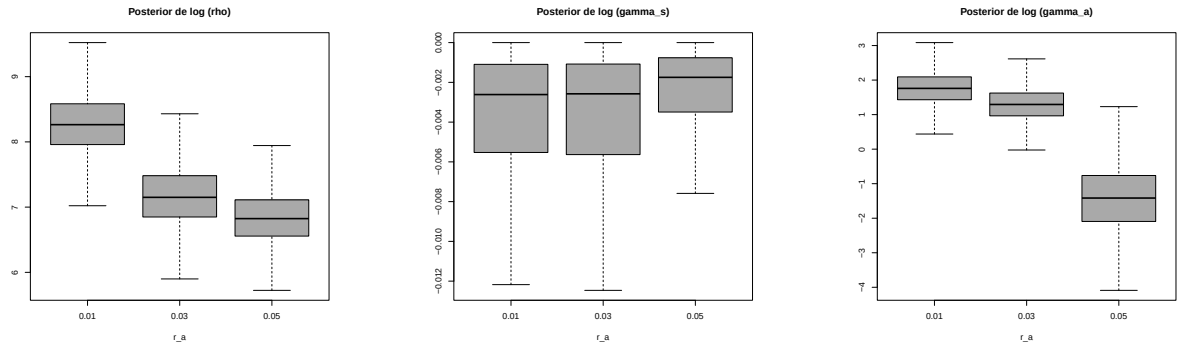


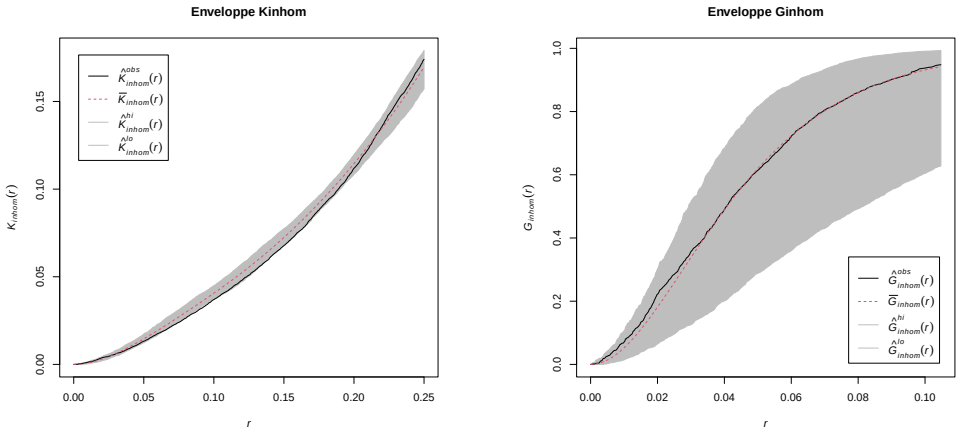
FIGURE 3.8 – Box-plot de la distribution *a posteriori* des marginales de $\log(\rho)$ (gauche), $\log(\gamma_s)$ (milieu) et $\log(\gamma_a)$ (droite) pour $r_s = 0.05$ en fonction de r_a. Pour chaque triplet de box-plots : $r_a = 0.01$ (gauche), $r_a = 0.03$ (milieu) et $r_a = 0.05$ (droite).

Le tableau suivant présente les valeurs obtenues en prenant le mode des histogrammes des lois marginales des paramètres. La loi *a priori* étant la loi uniforme, l'échantillon obtenu pour la loi *a posteriori* $f(\theta|\mathbf{x})$ est un échantillon de la vraisemblance restreinte au support de la loi uniforme. Pour cette raison, nous considérons donc que le maximum *a posteriori* (MAP) correspond au maximum de vraisemblance et que l'échantillon obtenu permet d'approcher ce MAP. Nous pouvons alors obtenir les erreurs asymptotiques standard (la différence entre le maximum de vraisemblance et le vrai paramètre) et MCMC (la différence entre l'approximation du maximum de vraisemblance et le maximum de vraisemblance) comme présenté en section 2.3.5.2. Nous avons choisi de mettre ici les erreurs MCMC pour s'assurer que la convergence vers le maximum de vraisemblance est bonne.

Rayons	Estimations de $\log(\rho)$, $\log(\gamma_S)$ et $\log(\gamma_A)$		
(r_s, r_a)	$\log(\widehat{\rho})$	$\log(\widehat{\gamma_s})$	$\log(\widehat{\gamma_a})$
$(0.01, 0.01)$	8.98 ± 0.0246	-0.5500 ± 0.0166	2.52 ± 0.0292
$(0.01, 0.03)$	7.10 ± 0.0086	-0.0020 ± 0.0132	1.35 ± 0.0324
$(0.01, 0.05)$	6.74 ± 0.0082	-0.0015 ± 0.0175	-1.26 ± 0.0876
$(0.03, 0.01)$	8.18 ± 0.0244	-0.0005 ± 0.0036	1.82 ± 0.0263
$(0.03, 0.03)$	6.99 ± 0.0222	-0.0005 ± 0.0066	1.17 ± 0.0433
$(0.03, 0.05)$	6.71 ± 0.0144	-0.0005 ± 0.0058	-1.14 ± 0.0914
$(0.01, 0.05)$	8.19 ± 0.0192	-0.0005 ± 0.0016	1.70 ± 0.0208
$(0.03, 0.05)$	7.10 ± 0.0237	-0.0005 ± 0.0029	1.40 ± 0.0429
$(0.05, 0.05)$	6.73 ± 0.0218	-0.0005 ± 0.0040	-1.42 ± 0.1048

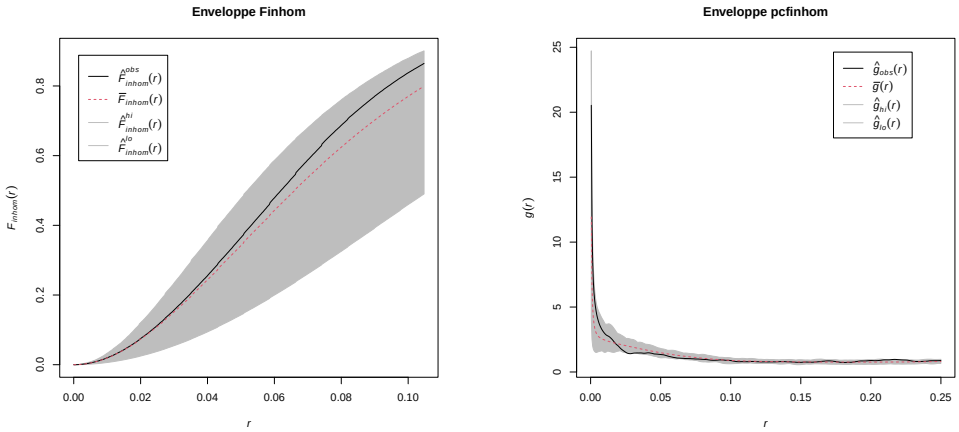
TABLE 3.1 – Estimations des paramètres pour différentes valeurs du couple (r_s, r_a) et leurs erreurs asymptotiques MCMC.

Nous contrôlons l'inférence ainsi menée à travers des tests d'enveloppes MCMC en utilisant la méthodologie expliquée en Exemple 2.3.1. Ici, nous illustrons le contrôle de l'inférence pour $(r_s, r_a) = (0.01, 0.03)$ via les enveloppes «pointwise» pour les fonctions F, G, K, g à l'aide de 1000 simulations. H_0 est ici prise comme : "La configuration de points observée suit le modèle considéré de paramètre $(\log\rho, \log\gamma_s, \log\gamma_a) = (7.10, -0.002, 1.35)$". Les autres choix de (r_s, r_a) montrent des résultats similaires. Les enveloppes ont été tracées à l'aide des versions inhomogènes des fonctions F, G, K et g de `spatstat` et les simulations menées à l'aide de la librairie `DRLib`.



(a) Enveloppe MCMC pour K

(b) Enveloppe MCMC pour G



(c) Enveloppe MCMC pour F

(d) Enveloppe MCMC pour g

FIGURE 3.9 – Enveloppes MCMC pour les fonctions K (haut gauche), G (haut droite), F (bas gauche) et g (bas droite). Les courbes rouges correspondent à la moyenne des simulations, les courbes noires correspondent à l'estimation des fonctions sur les données observées, l'enveloppe MCMC apparaît en gris.

L'estimation semble plutôt correcte en regardant les fonctions F et G. Pour les fonctions K et g, il arrive que la courbe noire soit légèrement en dehors de l'enveloppe, indiquant de légers défauts dans l'ajustement du modèle.

3.2 Modélisations multi-échelle avec interactions d'un grand jeu de données

3.2.1 Contexte, présentation et extraction des données

Par rapport aux travaux que nous venons de présenter, plusieurs aspects peuvent être pris en compte pour «aller plus loin» dans la modélisation. Le modèle, basé sur la détection des filaments, ne permet pas de modéliser correctement des zones à faible densité là où les filaments sont absents. De plus, le modèle considéré, bien qu'ajustant les données correctement, ne donne pas une interprétation claire quant aux résultats obtenus par rapport aux rayons choisis et ne permet pas de détecter un phénomène de répulsion entre les galaxies, pourtant observé par les astrophysiciens. C'est en essayant de prendre en compte ces remarques que nous proposons un deuxième modèle présenté dans cette partie 3.2. Cette modélisation est un premier pas vers l'analyse de données massives et constitue une étude du comportement spatial des interactions entre galaxies à travers l'espace. Nous introduisons dans la section suivante un modèle à interactions multiples qui considère plusieurs échelles d'interactions.

Le jeu de données considéré est ici sensiblement plus grand que le précédent, il représente une configuration 2D de 36047 galaxies extrait d'un catalogue de données réelles de galaxies 3D utilisé dans les articles [Stoica et al., 2015, Stoica et al., 2017].

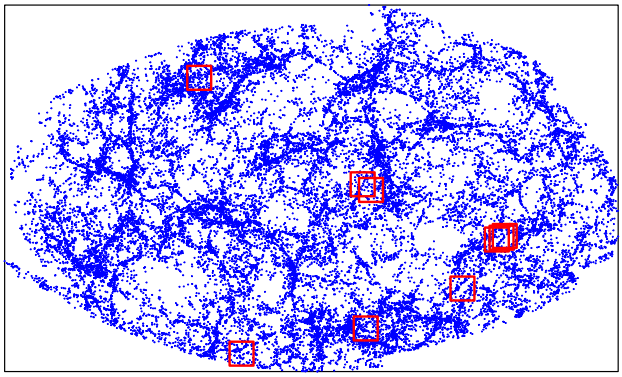
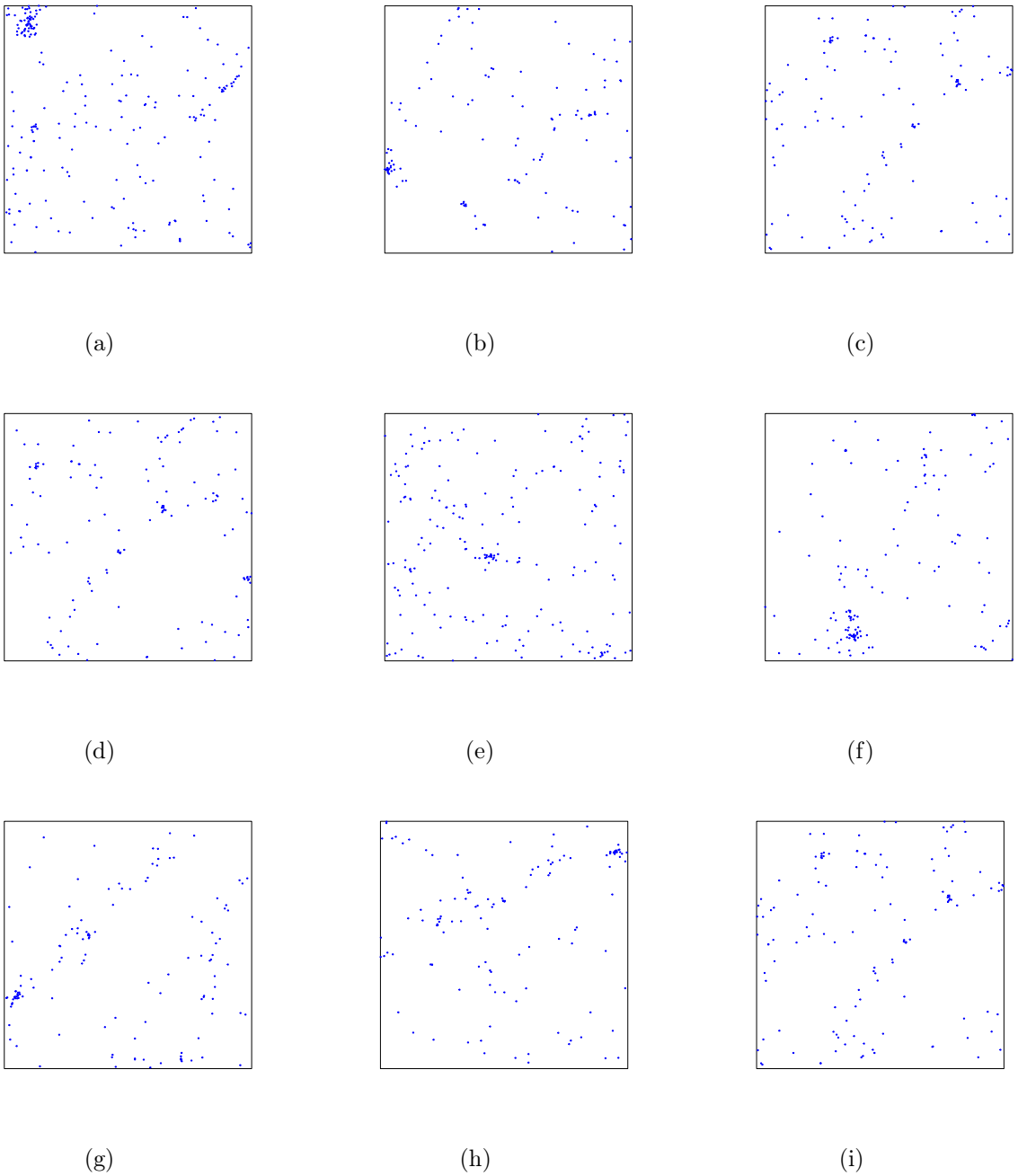


FIGURE 3.10 – Catalogue de position des galaxies et configurations extraites.

À première vue, ces données exhibent le comportement observé par les astrophysiciens : de grandes zones de «quasi» vide cosmique (à l'exception de quelques galaxies présentes) et des zones à très fortes concentrations de galaxies. Notre modèle cherche à modéliser ce phénomène. Pour les mêmes raisons que précédemment, nous nous restreignons à des extractions locales de ce jeu de données. Cette fois-ci, nous considérons plusieurs sous-domaines et non un seul afin de montrer que le modèle proposé est plus flexible que le précédent. Les extractions sont illustrées par les carrés rouges sur la figure ci-dessous. Le critère pour garder ces configurations extraites de manière aléatoires dans le catalogue est de contenir une ou plusieurs zones à très fortes densités de galaxies. Les configurations sont présentées ci-après :



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

FIGURE 3.11 – Configurations extraites.

L'analyse descriptive de ces configurations est illustrée dans la suite à travers l'analyse détaillée de la configuration 3.11i ci-dessus. Nous présentons d'abord une première pierre à l'édifice de la construction du modèle afin d'obtenir le comportement observé, nous avons appelé ce modèle le modèle «StraussCrown».

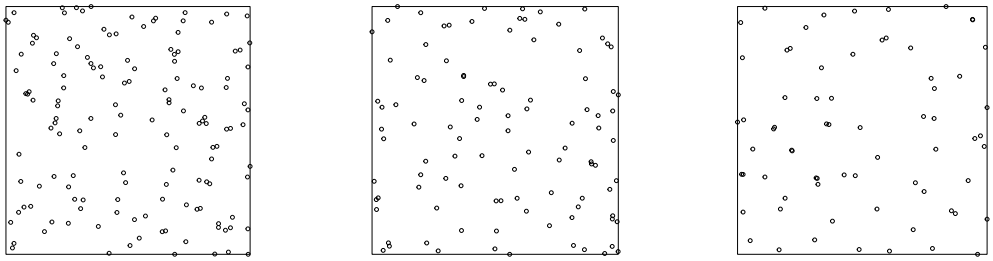
3.2.2 Une modélisation basée sur le modèle «StraussCrown»

Motivée par l'idée de devoir introduire un phénomène d'agglomération dans un voisinage proche des points et de répulsion dans un voisinage plus éloigné, une légère modification du modèle de Strauss permet d'aboutir à un modèle répulsif entre les points qui sont situé dans un voisinage en couronne autour des points. Ce modèle s'obtient simplement en modifiant la statistique du modèle de Strauss. Le nombre de $r-$voisins devient le nombre de voisins d'un point dans une couronne centrée en ce point. La densité de ce modèle par rapport au processus standard est donnée par

$$f(\mathbf{x}|\rho,\gamma_{sc}) = \exp(n(\mathbf{x})\log\rho + s_{r_1,r_2}(\mathbf{x})\log\gamma_{sc}),$$

avec $s_{r_1,r_2}(\mathbf{x}) = \frac{1}{2}\sum_{\xi\in\mathbf{x}}\sum_{\eta\in\mathbf{x},\eta\neq\xi}\mathbf{1}_{[r_1,r_2]}(d(\xi,\eta))$.

La stabilité locale et le caractère Markovien du modèle découlent directement de ceux du modèle de Strauss. De plus, si nous fixons $r_1 = 0$, le modèle est un modèle de Strauss de rayon r_2. Nous illustrons le comportement de ce modèle à l'aide de trois configurations, obtenues pour $(r_1, r_2) = (0.05, 0.10)$ et différents choix de paramètres.



(a) $\gamma_{sc} = 0.75$ (b) $\gamma_{sc} = 0.2$ (c) $\gamma_{sc} = 0.05$

FIGURE 3.12 – Réalisations du modèle StraussCrown pour différentes valeurs de γ_{sc} ; $\rho = 300$ et $(r_1, r_2) = (0.05, 0.10)$ dans $W = [0,1]\times[0,1]$.

Le modèle, lorsque γ_{sc} est très proche de 0 semble correctement inhiber le voisinage en couronne considéré. Le voisinage en cercle autour des points semble ne pas être affecté par la répulsion. Nous observons quelques paires/triplets de points très proches les uns des autres et montrent qu'à faibles portées les points peuvent se rapprocher. Néanmoins, ce modèle seul ne parvient pas à modéliser des grands amas de points comme ceux observés dans les données cosmologiques, où la concentration en galaxie est beaucoup plus forte. Le modèle que nous allons introduire dans la prochaine section permet de modéliser ces phénomènes.
En combinant l'idée de superposition et le comportement de répulsion du modèle StraussCrown, nous introduisons un nouveau modèle permettant l'agglomération à petite échelle et la répulsion à grande échelle. La densité non normalisée du modèle est donnée par :

$$f(\mathbf{x}|\rho,\gamma_{sc1},\gamma_{sc2},\gamma_a) \propto \exp(n(\mathbf{x})\log(\rho) + s_{r_1r_2}(\mathbf{x})\log(\gamma_{sc1}) + s_{r_2r_3}(\mathbf{x})\log(\gamma_{sc2}) + a_{r_1}(\mathbf{x})\log(\gamma_a))$$

où $n(\mathbf{x})$ et $a_{r_1}(\mathbf{x})$ sont les statistiques du processus de Poisson et Area-Interaction respectivement. Les statistiques s_{r_1,r_2} (resp. s_{r_2,r_3}) représentent le nombre de paires de points dans une couronne de rayons r_1 et r_2 (resp. r_2 et r_3) autour des points. Le voisinage d'un point est résumé dans le schéma suivant :

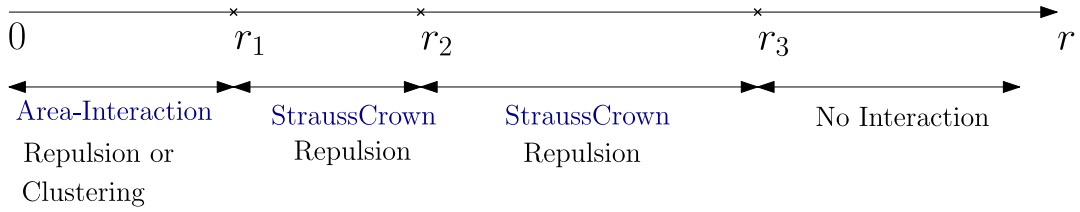
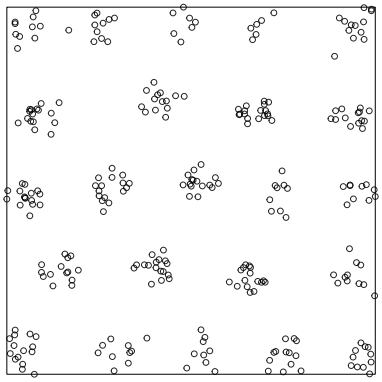


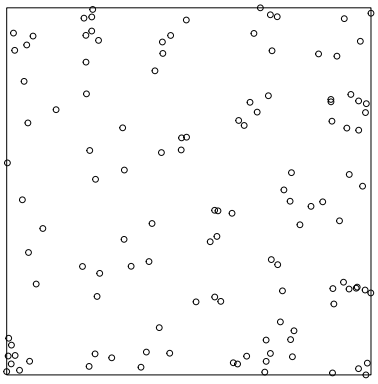
FIGURE 3.13 – Différentes interactions autour des points en fonction du rayon.

Entre les distances 0 et r_1, le modèle se comporte comme un modèle Area-Interaction, permettant la répulsion ou l'agglomération. Ensuite, entre les distances r_1 et r_3, le modèle agit comme le modèle de Strauss, permettant la répulsion entre les points. Cependant, la répulsion peut avoir deux échelles différentes, une répulsion à petite échelle dans $]r_1, r_2]$ et une répulsion à grande échelle dans $]r_2, r_3]$. Il peut aussi ne pas y avoir d'interaction entre les points après r_1, c'est-à-dire que les paramètres liés aux composantes Strauss peuvent être fixés à 0. La figure 3.15 présente quelques configurations simulées du modèle pour différentes valeurs de paramètres. Une grande variété de comportements différents peut être construite en modifiant les valeurs des paramètres. Les paramètres γ_{sc1} et γ_{sc2} pénalisent les configurations de points avec des paires de points situés respectivement dans les couronnes définies par (r_1, r_2) et (r_2, r_3). Le paramètre ρ contrôle le nombre de points du modèle. Le paramètre γ_a contrôle la surface occupée par les disques de rayons r_1 et centrés autour de chaque point. Les différents rayons choisis jouent également un rôle très importants, ils permettent d'influencer la taille des clusters formés et leurs concentrations en points.

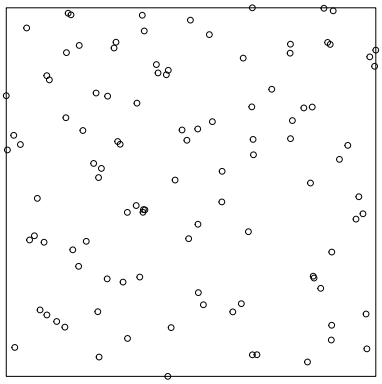
Les configurations ci-dessous illustrent des comportements différents qu'il est possible d'obtenir à l'aide de valeurs spécifiques des paramètres. La configuration (a) en haut à gauche est obtenue en forçant les points à occuper le plus de surface possible ($r_1 = 0.01$, $\gamma_a = 0.005$, $\gamma_{sc1} = 1$) tout en pénalisant la présence des points autours des clusters ($\gamma_{sc2} = 0.05$ et $(r_2, r_3) = (0.10, 0.15)$. D'autres configurations traduisant davantage la répartition des galaxies dans des zones à faibles densités peuvent être obtenues comme l'illustrent les figures (b), (c) et (d). Ce modèle semble alors prometteur pour modéliser des données cosmologiques.



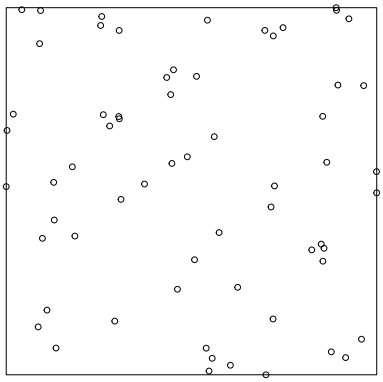
(a) Configuration simulée pour $r_1 = 0.01$; $r_2 = 0.10$; $r_3 = 0.15$; $\rho = 300$; $\gamma_{sc1} = 1$; $\gamma_{sc2} = 0.05$; $\gamma_a = 0.005$



(b) Configuration simulée pour $r_1 = 0.01$; $r_2 = 0.10$; $r_3 = 0.15$; $\rho = 300$; $\gamma_{sc1} = 1$; $\gamma_{sc2} = 0.2$; $\gamma_a = 0.4$



(c) Configuration simulée pour $r_1 = 0.05$; $r_2 = 0.05$; $r_3 = 0.08$; $\rho = 300$: $\gamma_{sc1} = \gamma_{sc2} = 0.05$; $\gamma_A = 0.4$



(d) Configuration simulée pour $r_1 = 0.05$; $r_2 = 0.10$; $r_3 = 0.15$; $\rho = 300$: $\gamma_{sc1} = 0.37$; $\gamma_{sc2} = 0.05$; $\gamma_A = 0.1$

FIGURE 3.15 – Configurations simulées.

3.2.3 Analyse descriptive des données extraites, choix des rayons

Dans cette section, les statistiques suffisantes du modèle StraussCrown Area-Interaction pour la configuration 3.11i ci-dessus sont comparées aux statistiques correspondantes calculées pour plusieurs réalisations du processus de Poisson avec le même nombre de points. Différentes valeurs de r_1 sont prises en compte pour mieux comprendre le comportement à petite échelle de la configuration observée. Les choix de r_2 et r_3 sont expliqués dans la suite.

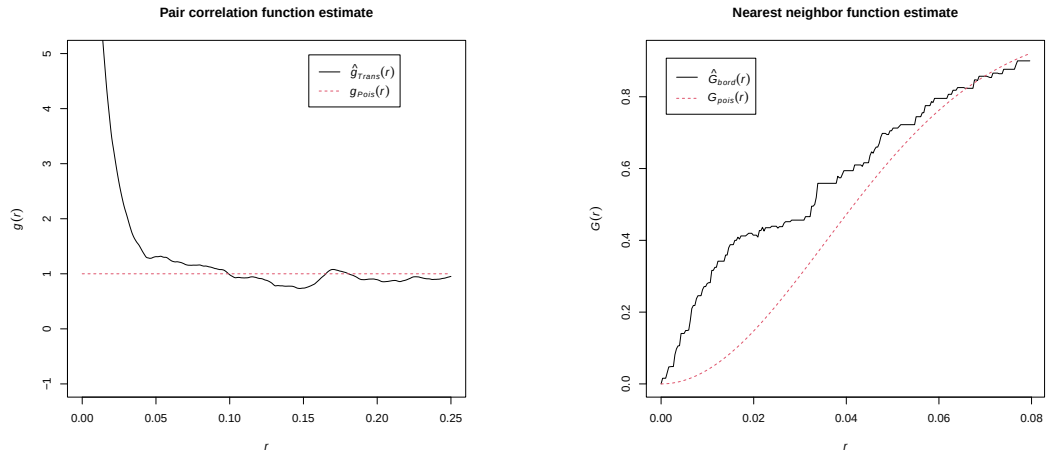
3.2.3.1 Comparaison des statistiques suffisantes

Dans le tableau 3.2, nous voyons que les statistiques suffisantes pour le modèle sélectionné sont toujours inférieures à celles du modèle de Poisson. Autour d'un point choisi au hasard, cela indique une tendance à la répulsion entre r_2 et r_3. Pour chaque r_1 choisi, la statistique d'Area-Interaction est toujours plus grande pour le modèle sélectionné par rapport au modèle de Poisson, ce qui est cohérent avec l'agglomération. Cependant, lorsque $r_1 = r_2$, la différence entre les deux statistiques est très faible, ce qui suggère de choisir une valeur r_1 différente de r_2 afin que les statistiques ne soient pas trop proches d'un processus de Poisson. Enfin, les points ont tendance à occuper une plus grande surface pour le modèle de Poisson que dans le modèle sélectionné. Ce tableau indique qu'une modélisation par un simple processus de Poisson inhomogène est inadéquat, nous allons ainsi introduire notre modèle en tenant compte de ces comparaisons.

TABLE 3.2 – Comparaison des statistiques suffisantes entre des configurations Poissoniennes et la configuration 3.11i

Statistique suffisante	Configuration 3.11i	Moyenne Poisson
$n(\mathbf{x})$	127	127
$s_{r_2,r_3}(\mathbf{x})$ for $(r_2, r_3) = (0.13, 0.15)$	90	117.9
$s_{r_1,r_2}(\mathbf{x})$ for $(r_1, r_2) = (0.01, 0.13)$	457	383.6
$s_{r_1,r_2}(\mathbf{x})$ for $(r_1, r_2) = (0.02, 0.13)$	422	376.1
$s_{r_1,r_2}(\mathbf{x})$ for $(r_1, r_2) = (0.03, 0.13)$	390	363.3
$s_{r_1,r_2}(\mathbf{x})$ for $(r_1, r_2) = (0.05, 0.13)$	331	324.65
$s_{r_1,r_2}(\mathbf{x})$ for $(r_1, r_2) = (0.065, 0.13)$	278	281.6
$a_{r_1}(\mathbf{x})$ for $r_1 = 0.01$	-106.47	-122.88
$a_{r_1}(\mathbf{x})$ for $r_1 = 0.02$	-93.34	-115.30
$a_{r_1}(\mathbf{x})$ for $r_1 = 0.03$	-82.25	-104.41
$a_{r_1}(\mathbf{x})$ for $r_1 = 0.05$	-60.46	-78.60
$a_{r_1}(\mathbf{x})$ for $r_1 = r_2/2$	-46.76	-60.06
$a_{r_1}(\mathbf{x})$ for $r_1 = r_2$	-17.26	-18.69

3.2.3.2 Statistiques descriptives spatiales



(a) Fonction g pour la configuration observée 3.11i

(b) Fonction G pour la configuration observée 3.11i

FIGURE 3.16 – Fonctions g et G pour la configuration observée Figure 3.11i

Nous pouvons voir dans la figure 3.16a que le pattern est attractif jusqu'à $r = 0.10$. En effet, la fonction g pour les données (ligne noire continue) est supérieure à la courbe théorique du cas CSR. (ligne rouge). De plus, nous observons une tendance à la répulsion entre $r = 0.13$ et $r = 0.15$ (la courbe estimée est inférieure à la courbe théorique). La courbe estimée pour G des données en figure 3.16b montre également une tendance à l'attraction à courte portée pour les rayons proches de 0.

La fonction de corrélation par paires peut être utilisée pour essayer de trouver des valeurs appropriées pour r_2 et r_3 : ces deux rayons représentent les valeurs où la fonction g cesse de diminuer

(en r_2) et commence à augmenter (en r_3). Cependant, le choix de r_1 est plus difficile, car la taille des clusters et le nombre de points à l'intérieur ne sont pas les mêmes pour tous. Pour cette raison, nous envisagerons différents choix pour r_1 lors de l'inférence.

3.2.4 Application aux données

Pour chaque configuration extraite et pour chaque r_1 parmi $\{0.01, 0.02, 0.03, 0.05, r_2/2\}$, l'algorithme ABC Shadow a été initialisé avec les statistiques suffisantes calculées à partir du pattern observé. Cette procédure conduit à 5 ensembles différents de rayons (r_1, r_2, r_3) pour chaque extraction, ce qui donne 45 estimations de paramètres. La densité *a priori* $p(\theta)$ a été choisie comme étant la distribution uniforme sur l'intervalle $[0, 50] \times [-50, 0] \times [-50, 0] \times [-50, 50]$. À chaque étape, le motif auxiliaire a été généré avec 500 itérations de l'algorithme de Metropolis-Hastings. Le paramètre de perturbation Δ a été fixé à $(0.01, 0.0025, 0.0025, 0.01)$ pour les quatre paramètres ρ, γ_{sc1}, γ_{sc2} et γ_a. Le choix de 0.0025 pour les deux composantes «StraussCrown» est encore fait de manière à éviter certaines erreurs d'estimation lorsque les estimations des paramètres γ_{sc1} et γ_{sc2} sont trop proches de 0. Ces données sont résumées dans le tableau ci-dessous :

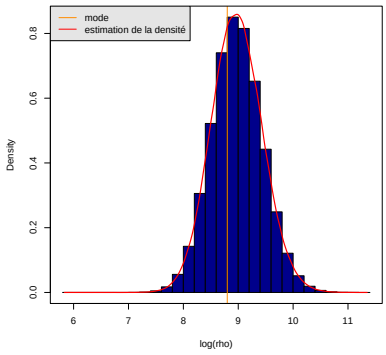
Variable	Description	Valeur
W	Domaine de simulation	$[0,1] \times [0,1]$
N_θ	Nombre d'échantillons de θ	10^6
N_{ABC}	Nombre d'itérations proposition paramètre	100
N_{MH}	Nombre d'itérations MH	500
p_b, p_d	Probabilité d'ajout/retrait	$0.5, 0.5$
θ_0	Paramètres initiaux	Aléatoire
$\Delta = (\Delta_\rho, \Delta_{\gamma_a})$	Paramètres de perturbation des paramètres	$(0.01, 0.01)$
$\Delta = (\Delta_{\gamma_{sc1}}, \Delta_{\gamma_{sc2}})$	Paramètres de perturbation des paramètres	$(0.0025, 0.0025)$
$[\rho_{min}, \rho_{max}]$	Support de la loi uniforme sur ρ	$[0, 10]$
$[\gamma_{sc1,min}, \gamma_{sc1,max}]$	Support de la loi uniforme sur γ_{sc1}	$[-50, 0]$
$[\gamma_{sc2,min}, \gamma_{sc2,max}]$	Support de la loi uniforme sur γ_{sc2}	$[-50, 0]$
$[\gamma_{a,min}, \gamma_{a,max}]$	Support de la loi uniforme sur γ_a	$[-50, 50]$

Dans 27 des 45 cas, la deuxième composante StraussCrown a été estimée comme étant différente de zéro. Tout d'abord, nous résumons les résultats en utilisant les cinq valeurs différentes de r_1 ainsi que les erreurs standard asymptotiques dans le tableau 3.3 pour les données de la figure 3.11i. Ensuite, nous présentons les histogrammes et les box-plots pour les paramètres estimés, ainsi qu'un modèle simulé avec les paramètres estimés. Enfin, nous discutons de la qualité de l'ajustement du modèle en nous appuyant sur des tests d'enveloppe MCMC réalisés avec les fonctions G et g.

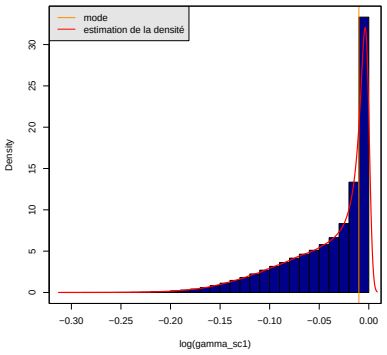
TABLE 3.3 – Paramètres estimés et erreurs standard associées

Rayon	Estimations de $\log(\rho)$, $\log(\gamma_{sc1})$, $\log(\gamma_{sc2})$ et $\log(\gamma_a)$			
(r_1, r_2, r_3)	$\log(\rho)$	$\log(\gamma_{sc1})$	$\log(\gamma_{sc2})$	$\log(\gamma_a)$
$(0.01, 0.13, 0.15)$	48 ± 1.26	-0.05 ± 0.33	-0.05 ± 0.75	44 ± 1.28
$(0.02, 0.13, 0.15)$	8.8 ± 0.20	-0.02 ± 0.014	-0.15 ± 0.04	4.5 ± 0.25
$(0.03, 0.13, 0.15)$	7.4 ± 0.19	-0.02 ± 0.017	-0.1 ± 0.05	3.5 ± 0.27
$(0.05, 0.13, 0.15)$	6.6 ± 0.20	-0.02 ± 0.02	-0.15 ± 0.05	3 ± 0.40
$(0.065, 0.13, 0.15)$	6.4 ± 0.21	-0.12 ± 0.05	-0.15 ± 0.09	4 ± 0.48

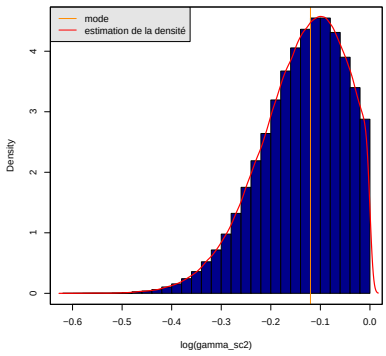
À l'exception du cas $r_1 = 0.01$, les erreurs asymptotiques sont plutôt faibles, ce qui indique une estimation assez bonne. Pour r_1 dans $\{0.02, 0.03, 0.05\}$, la première composante de «Strauss-Crown» est très proche de zéro, ce qui est cohérent avec la fonction de corrélation par paires pour ce pattern. En revanche, pour r_1 supérieur à 0.02, nous observons un comportement répulsif d'une couronne de rayons $r_2 = 0.13$ et $r_3 = 0.15$ autour des points. À l'exception du cas $r_1 = 0.01$, les valeurs estimées de γ_a semblent assez proches les unes des autres, ce qui indique un comportement assez similaire pour chaque r_1 considéré. Les histogrammes et les box-plots ci-dessous décrivent l'échantillon de paramètres obtenu par l'algorithme ABC Shadow.



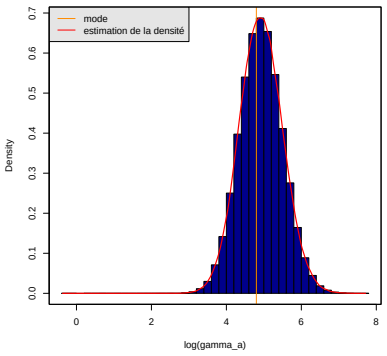
(a) Histogramme de $\log(\rho)$



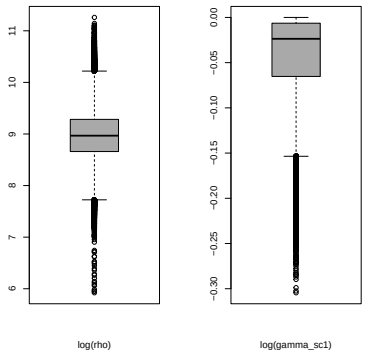
(b) Histogramme de $\log(\gamma_{sc_1})$



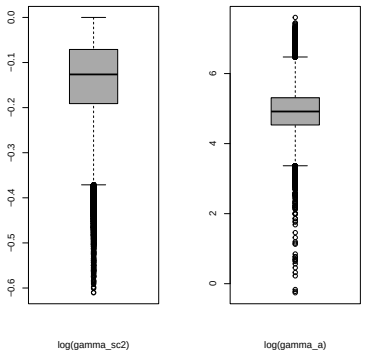
(c) Histogramme de $\log(\gamma_{sc_2})$



(d) Histogramme de $\log(\gamma_a)$



(e) Box-plots pour $\log(\rho)$ (gauche) et $\log(\gamma_{sc_1})$ (droite)



(f) Box-plots pour $\log(\gamma_{sc_1})$ (gauche) et $\log(\gamma_a)$ (droite)

FIGURE 3.18 – Histogrammes et Box-plots pour l'estimation réalisée avec $(r_1, r_2, r_3) = (0.02, 0.13, 0.15)$

Les figures ci-dessous montrent une simulation obtenue pour $(\log(\rho), \log(\gamma_{sc1}), \log(\gamma_{sc2}), \log(\gamma_a)) = (8.8, -0.02, -0.15, 4.5)$ (gauche) et le pattern extrait 3.11i. (droite)

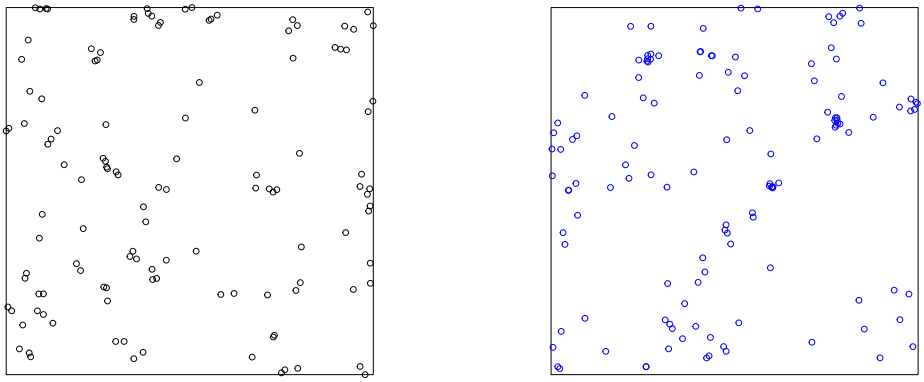
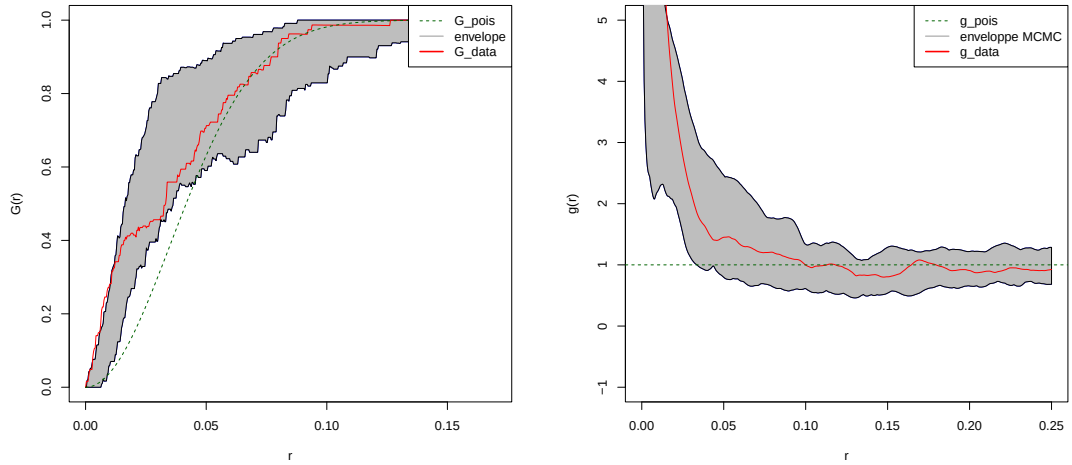


FIGURE 3.19 – configuration simulée (gauche) et configuration observée 3.11i (droite)

À première vue, la taille des clusters et les zones vides semblent assez similaires dans les deux configurations. Dans la section suivante, nous présentons une comparaison plus détaillée entre 200 configurations simulées et la configuration extraite ci-dessus à l'aide de tests d'enveloppe MCMC pour les fonctions g et G.

Les enveloppes MCMC suivantes (en gris) ont été obtenues en traçant les fonctions G (3.20a) et g (3.20b) pour 200 configurations simulées avec les paramètres $(\log(\rho), \log(\gamma_{sc1}), \log(\gamma_{sc2}), \log(\gamma_a)) = (8.8, -0.02, -0.15, 4.5)$. La courbe rouge représente les fonctions G ou g estimées à partir du pattern observé 3.11i. Enfin, la courbe verte représente les fonctions théoriques de Poisson pour G et g.

À gauche de la figure 3.20a, nous observons que les modèles simulés sont attractifs jusqu'à $r = 0.05$. Cependant, la fonction G du modèle observé se trouve légèrement en dehors de l'enveloppe créée par les simulations pour les petites valeurs de r, même si elle reste très proche de celle-ci. Cela indique un éventuel décalage entre le modèle et le modèle à très petite échelle. Dans la figure 3.20b, nous observons que la fonction g du pattern observé se trouve toujours à l'intérieur de l'enveloppe MCMC. Cela indique que notre modèle correspond aux caractéristiques de la fonction de corrélation par paires de l'observation.



(a) Enveloppe MCMC pour G (b) Enveloppe MCMC pour g

FIGURE 3.20 – Enveloppes MCMC pour les fonction G (gauche) et g (droite)

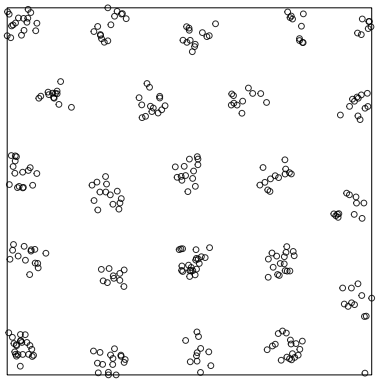
3.2.5 Une modélisation basée sur le modèle «GeyerCrown»

Ce modèle est un second modèle basé sur cette première modélisation. Après avoir implémenté le modèle de saturation de Geyer dans la librairie `DRLib`, il était aisé de l'adapter en créant, tout comme le modèle «StraussCrown», un modèle «GeyerCrown». À nouveau, la stabilité et le caractère markovien du modèle découlent de celles du modèle de saturation de Geyer. La statistique de ce modèle est calculée sur une couronne de rayons r_1 et r_2 autour des points. L'avantage du modèle de saturation de Geyer, c'est qu'il est défini à la fois pour de la répulsion et de l'agglomération. Nous avons choisi de retraiter les mêmes données avec les rayons retenus mais en utilisant cette fois-ci uniquement le modèle de saturation de Geyer et sa variante «GeyerCrown». La densité du modèle est la suivante :

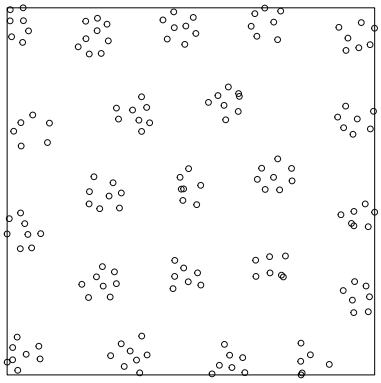
$$f(\mathbf{x}|\rho,\gamma_{g_1},\gamma_{g_2},\gamma_{g_3}) \propto \exp\left(n(\mathbf{x})\log(\rho) + \sum_{\xi\in\mathbf{x}} \min(s_{0,r_1}\ (\xi), s_1)\log(\gamma_{g_1}) + \sum_{\xi\in\mathbf{x}} \min(s_{r_1,r_2}(\xi), s_2)\log(\gamma_{g_2}) + \sum_{\xi\in\mathbf{x}} \min(s_{r_2,r_3}(\xi), s_3)\log(\gamma_{g_3})\right) \quad (3.3)$$

Ce modèle offre plus de flexibilité par rapport au modèle introduit précédemment mais il demande également un réglage plus fin : les seuils de saturations sont à fixer pour chaque composante,

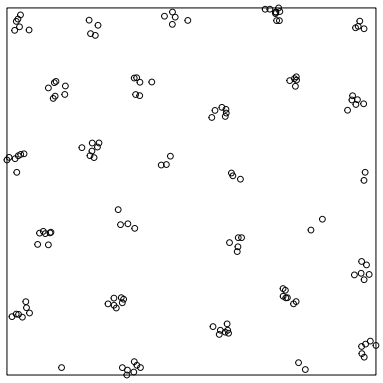
choix non évident, même après une analyse descriptive des données. Nous illustrons à nouveau les comportements possibles de ce modèle à travers 4 réalisations obtenues en fixant tous les seuils de saturation s_1, s_2 et s_3 à 4.5.



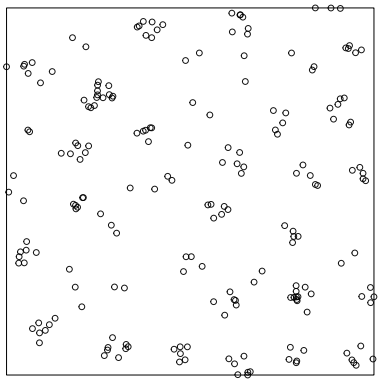
(a) Configuration simulée pour $r_1 = 0.05$; $r_2 = 0.10$; $r_3 = 0.15$; $\rho = 300$; $\gamma_{g_1} = 1.5$; $\gamma_{g_2} = 1$; $\gamma_{g_3} = 0.05$



(b) Configuration simulée pour $r_1 = 0.03$; $r_2 = 0.10$; $r_3 = 0.15$; $\rho = 300$; $\gamma_{g_1} = 0.05$; $\gamma_{g_2} = 2.72$; $\gamma_{g_3} = 0.2$



(c) Configuration simulée pour $r_1 = 0.05$; $r_2 = 0.10$; $r_3 = 0.15$; $\rho = 300$: $\gamma_{g_1} = 1.35$; $\gamma_{g_2} = 0.37$; $\gamma_{g_3} = 0.05$



(d) Configuration simulée pour $r_1 = 0.05$; $r_2 = 0.08$; $r_3 = 0.10$; $\rho = 300$; $\gamma_{g_1} = 1.22$; $\gamma_{g_2} = 1$; $\gamma_{g_3} = 0.05$

FIGURE 3.21 – Configurations simulées.

La réalisation en haut à gauche ressemble très fortement à celle obtenue en 3.14a et illustre bien la possibilité d'obtenir des configurations très agglomérées. La réalisation en haut à droite illustre que le «remplissage» de ces clusters peut être contrôlé en réglant le premier rayon et en prenant le paramètre γ_{g_1} proche de 0. La réalisation en bas à gauche montre que la taille de ces clusters peut être ajustée en jouant sur la force d'agglomération, γ_{g_1} est pris légèrement plus petit que

pour la configuration 3.21a et $\gamma_{g_2} < 1$ pour obtenir de la répulsion dès $r_2 = 0.10$ autour des points. Enfin, la dernière réalisation montre qu'il est possible «d'assouplir» le clustering obtenu par rapport à la figure 3.21c.

L'analyse descriptive menée précédemment est réutilisée dans la suite pour effectuer l'inférence à partir de ce nouveau modèle. Nous présentons les résultats pour la même configuration étudiée plus haut.

3.2.6 Application aux données

Les paramétrages de l'algorithme ABC Shadow sont expliqués à nouveau par le tableau ci-dessous. Les seuils de saturations n'étant pas directement interprétables en termes de données réelles, plusieurs réglages pour différentes valeurs ont été menés. Nous présentons l'inférence pour les mêmes choix de rayons que précédemment et pour $s_1 = s_2 = s_3 = 4.5$.

Variable	Description	Valeur
W	Domaine de simulation	$[0,1] \times [0,1]$
N_θ	Nombre d'échantillons de θ	10^5
N_{ABC}	Nombre d'itérations proposition paramètre	100
N_{MH}	Nombre d'itérations MH	10^3
p_b, p_d	Probabilité d'ajout/retrait	$0.5, 0.5$
θ_0	Paramètres initiaux	Aléatoire
$\Delta = (\Delta_\rho, \Delta_{\gamma_{g_1}})$	Paramètres de perturbation des paramètres	$(0.001, 0.001)$
$\Delta = (\Delta_{\gamma_{g_2}}, \Delta_{\gamma_{g_3}})$	Paramètres de perturbation des paramètres	$(0.001, 0.001)$
$[\rho_{min}, \rho_{max}]$	Support de la loi uniforme sur ρ	$[0, 10]$
$[\gamma_{g_i,min}, \gamma_{sg_i,max}]$	Support de la loi uniforme sur γ_{g_i}	$[-5, 2]$

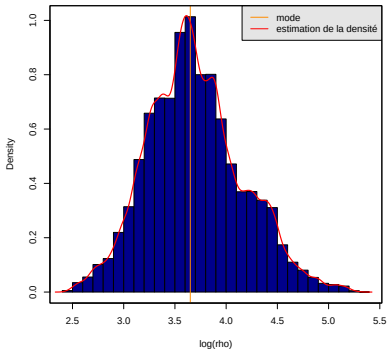
Le modèle de saturation de Geyer étant sensible aux perturbations des paramètres (les statistiques varient beaucoup d'un paramètre à un autre, même «proches»), le choix des paramètres de perturbation est choisi à 0.001. Aucun phénomène de bord n'ayant été rencontré, le choix initial de prendre les lois uniformes décrites ci-dessus est resté le même et n'a pas dû être adapté au cours de l'inférence. Le tableau ci-dessous résume l'inférence menée pour tous les rayons et nous illustrons à nouveau par des histogrammes et des box-plots celle menée pour $(r_1, r_2, r_3) = (0.02, 0.13, 0.15)$.

TABLE 3.4 – Paramètres estimés et erreurs standard associées

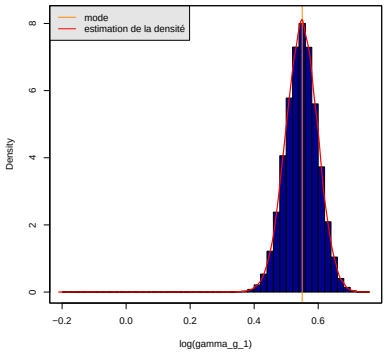
Rayon	Estimations de $\log(\rho)$, $\log(\gamma_{g_1})$, $\log(\gamma_{g_2})$ et $\log(\gamma_{g_3})$			
(r_1, r_2, r_3)	$\log(\rho)$	$\log(\gamma_{g_1})$	$\log(\gamma_{g_2})$	$\log(\gamma_{g_3})$
$(0.01, 0.13, 0.15)$	4.05 ± 0.62	0.75 ± 0.05	0.175 ± 0.11	-0.09 ± 0.05
$(0.02, 0.13, 0.15)$	3.65 ± 0.51	0.55 ± 0.06	0.175 ± 0.09	-0.07 ± 0.05
$(0.03, 0.13, 0.15)$	3.85 ± 0.61	0.425 ± 0.03	0.175 ± 0.11	-0.07 ± 0.03
$(0.05, 0.13, 0.15)$	4.05 ± 0.38	0.275 ± 0.04	0.075 ± 0.06	-0.09 ± 0.04
$(0.065, 0.13, 0.15)$	4.05 ± 0.30	0.270 ± 0.05	0.010 ± 0.04	-0.11 ± 0.05

Cette fois-ci, l'inférence pour le rayon $r_1 = 0.01$ reste proche des autres estimations et nous n'observons pas de phénomène d'explosion comme précédemment où les estimations pour $\log(\rho)$ et $\log(\gamma_a)$ étaient très élevées. Les erreurs asymptotiques pour $\log \rho$ sont plutôt grandes mais

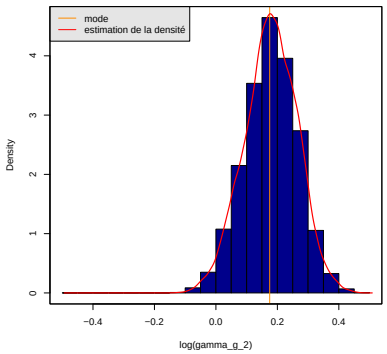
celles pour les $\log \gamma_{g_i}$ sont très faibles. Ces résultats indiquent bien une tendance à la répulsion entre les rayons $r_2 = 0.13$ et $r_3 = 0.15$ et le phénomène d'agglomération à petite échelle s'exprime bien par l'estimation de $\log \gamma_{g_1}$ supérieur à 0 dans les 5 cas considérés. Ces résultats vont dans le sens de ceux trouvés précédemment résumés dans le tableau 3.3. Néanmoins, nous remarquons que l'estimation pour la deuxième composante de GeyerCrown, associée à $\log(\gamma_{g_2})$, est cohérente avec de l'attraction entre les points. Précédemment, cette échelle (entre r_1 et r_2) était associée à la première composante de StraussCrown qui était alors estimée très proche de 0 ($\log(\gamma_{sc1})$. Il est donc possible qu'un phénomène d'attraction, bien que faible, soit finalement présent sur cette échelle.



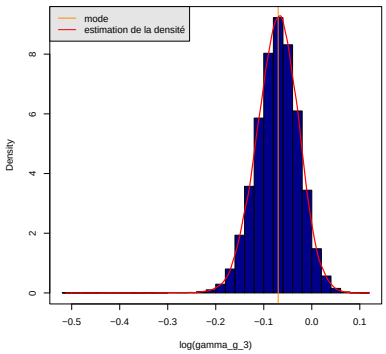
(a) Histogramme de $\log(\rho)$



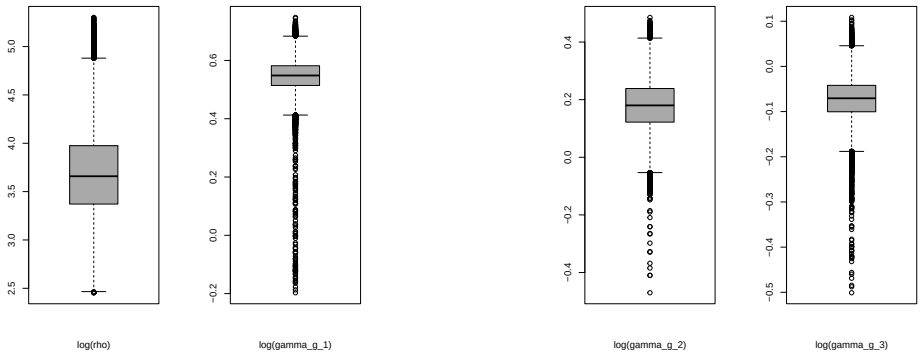
(b) Histogramme de $\log(\gamma_{g_1})$



(c) Histogramme de $\log(\gamma_{g_2})$



(d) Histogramme de $\log(\gamma_{g_3})$



(a) Box-plots pour $\log(\rho)$ (gauche) et $\log(\gamma_{sc_1})$ (droite)

(b) Box-plots pour $\log(\gamma_{sc_1})$ (gauche) et $\log(\gamma_a)$ (droite)

FIGURE 3.23 – Histogrammes et Box-plots pour l'estimation réalisée avec $(r_1, r_2, r_3) = (0.02, 0.13, 0.15)$

Les figures ci-dessous montrent une simulation obtenue pour $(\log(\rho), \log(\gamma_{g_1}), \log(\gamma_{g_2}), \log(\gamma_{g_3})) = (3.65, 0.55, 0.175, -0.07)$ (gauche) et le pattern extrait 3.11i. (droite)

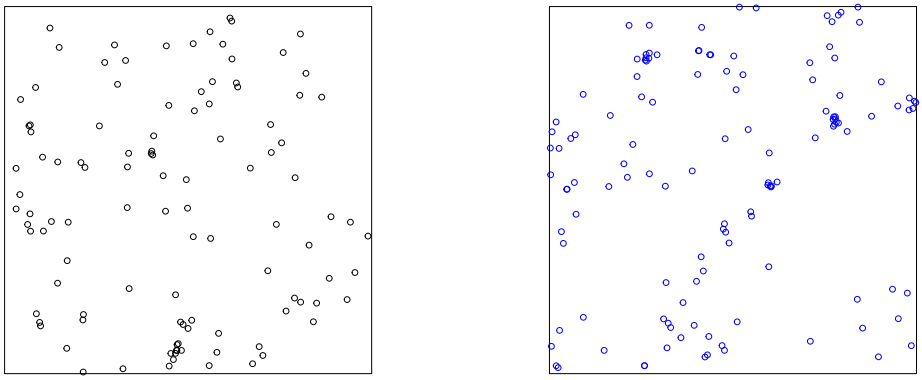
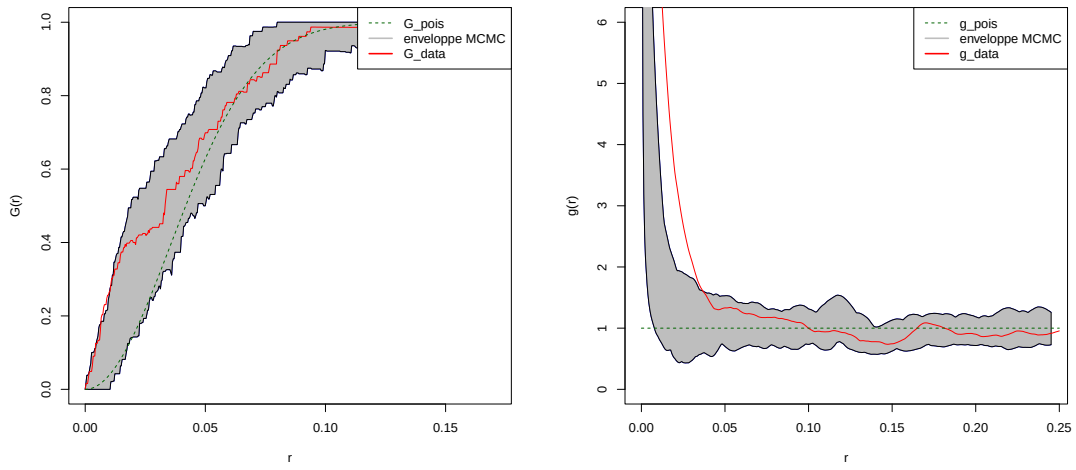


FIGURE 3.24 – Configuration simulée (gauche) et configuration observée 3.11i (droite)

À nouveau, la taille des clusters et les zones vides semblent assez similaires dans les deux configurations, cette fois-ci, moins de petits clusters s'observent dans la configuration simulée et l'espacement entre les galaxies semble être respecté sur cette configuration. Pour contrôler cela, nous présentons comme précédemment une comparaison plus détaillée entre 200 configurations simulées et la configuration extraite ci-dessus à l'aide de tests d'enveloppe MCMC pour les fonctions g et G.

Les enveloppes MCMC suivantes (en gris) ont été obtenues en traçant les fonctions G (3.25a) et g (3.25b) pour 200 configurations simulées avec les paramètres $(\log(\rho), \log(\gamma_{g_1}), \log(\gamma_{g_2}), \log(\gamma_{g_3})) = (3.65, 0.55, 0.175, -0.07)$. La courbe rouge représente les fonctions G et g estimées à partir du

pattern observé 3.11i. Enfin, la courbe verte représente les fonctions théoriques de Poisson pour G et g.
Similairement à ce qui est observé pour les premiers tests d'enveloppes MCMC, à gauche de la figure 3.25a, nous observons que les modèles simulés sont attractifs jusqu'à $r = 0.05$. Cependant, la fonction G du modèle observé se trouve légèrement en dehors de l'enveloppe créée par les simulations pour les petites valeurs de r, même si elle reste très proche de celle-ci. Cela indique un éventuel décalage entre le modèle et le modèle à très petite échelle. Dans la figure 3.25b, nous observons également ce phénomène pour des rayons jusqu'à $r = 0.04$, bien que la fonction g soit davantage adaptée à l'analyse du comportement pour des rayons plus grands, cela indique tout de même un potentiel souci d'ajustement à très petite échelle autour des points. Néanmoins, dans les deux cas ci-dessous, la courbe estimée pour les données se trouve bien dans l'enveloppe pour des rayons plus grands.



(a) Enveloppe MCMC pour G (b) Enveloppe MCMC pour g

FIGURE 3.25 – Enveloppes MCMC pour les fonctions G (gauche) et g (droite)

Chapitre 4

Reformulation Bayésienne pour l'inférence en données incomplètes

Sommaire

Contexte et hypothèses

Comme nous l'avons présenté dans l'état de l'art, l'émergence des données incomplètes est fréquente en statistique et plusieurs mécanismes de données manquantes sont généralement considérés dans la littérature pour les données spatialisées. Cette étude se focalise sur le mécanisme de censure où seule une sous-partie W_Y du domaine complet W est observée. Dans cette situation, la vraisemblance du processus ponctuel considéré ne peut être déterminée à l'aide des observations disponibles. Une solution est d'utiliser des données simulées produites via des procédures de Markov Chain Monte Carlo (MCMC) dans la région W_X, où aucune observation directe n'est disponible. Cette opération augmente le coût de calcul général, tandis que la convexité de la vraisemblance ne peut être garantie. Dans le contexte des données cosmologiques, la position des galaxies est observée dans l'Univers mais certaines régions sont obstruées en raison

de corps célestes ou à cause de la luminosité de certaines galaxies. Ainsi le cadre où le domaine $W = W_Y \cup W_X$ est divisé en une zone observée W_Y et une zone non-observée W_X est adéquat à ce type de données partiellement observées. Les algorithmes de type EM sont principalement utilisés pour effectuer l'inférence paramétrique pour ce type de données, nous proposons une méthodologie adaptée au contexte Bayésien pour l'échantillonnage *a posteriori* en données incomplètes. Dans la suite, nous proposons une nouvelle méthode où le domaine complet est considéré rectangulaire et le domaine W_X est pris comme une section de ce rectangle comme l'illustre la figure ci-dessous. Cela permet de limiter la zone d'interaction entre la zone W_X et W_Y lors de la simulation conditionnelle. La surface de cette zone est prise plus petite que la zone observée pour permettre aux statistiques observées d'être «proches» des statistiques non-observées. Cette décision ne change pas la généralité des développements proposés dans la suite, elle rend la mise en œuvre et le contrôle des programmes plus aisés.

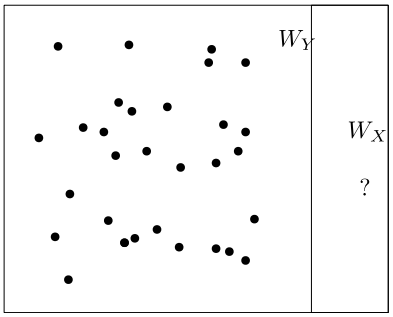


FIGURE 4.1 – Schéma du problème des données manquantes.

Ce chapitre est structuré de la manière suivante : nous rappelons d'abord plusieurs approches et illustrons les limites d'une tentative «naïve» basée sur la vraisemblance en données incomplètes. Nous introduisons ensuite deux lois *a posteriori* à échantillonner pour effectuer l'inférence en données incomplètes et proposons une construction, similaire à celle présentée dans la section 2.3.5 pour l'algorithme ABC Shadow, pour échantillonner les lois introduites. Nous proposons plusieurs algorithmes en pseudo-code pour échantillonner cette loi et dérivons, à partir de ces derniers, des stratégies numériques. Enfin, une étude par simulation illustre ces stratégies numériques en considérant le modèle de Strauss pour différentes valeurs du paramètre et différentes configurations d'observations.

4.1 Position du problème

Dans ce chapitre, nous reprenons les notations présentées en partie 2.3.2. Pour $\theta \in \Theta$, notons $f(\cdot, \cdot|\theta)$ la loi jointe des données complètes $(\mathbf{x}, \mathbf{y})$. Nous supposons ici que $\mathbf{y} \in W_Y$ est observé et que $\mathbf{x} \in W_X$ est une configuration cachée. La densité de la loi jointe des configurations sur W_Y et sur W_X est donnée par

$$f(\mathbf{x}, \mathbf{y}|\theta) = \frac{\exp\langle t(\mathbf{x} \cup \mathbf{y})|\theta\rangle}{Z(\theta)}$$

avec $Z(\theta) = \int\int \exp\langle t(\mathbf{x} \cup \mathbf{y})|\theta\rangle \mu_{W_X}(\mathrm{d}\mathbf{x}) \mu_{W_Y}(\mathrm{d}\mathbf{y})$ la constante de normalisation, μ_{W_Y} et μ_{W_X} les restrictions de la mesure de référence Poissonienne à W_Y et W_X.

La difficulté apparaît lorsque nous considérons la vraisemblance. Étant donné que $\mathbf{x}$ n'est pas observée, une intégrale sur toutes les configurations possibles dans la fenêtre W_X apparaît au numérateur de cette dernière :

$$L(\theta) = \frac{\int \exp\langle t(\mathbf{x} \cup \mathbf{y})|\theta\rangle \mu_{W_X}(\mathrm{d}\mathbf{x})}{Z(\theta)}.$$

Comme nous l'avons vu en partie 2.3.2, l'approche proposée par [Gelfand and Carlin, 1993, Geyer, 1999] offre une solution pour trouver le maximum de vraisemblance en considérant le ratio des log-vraisemblances à partir d'un paramètre auxiliaire. Dans le cas des données incomplètes, le paramètre auxiliaire est utilisé pour simuler à la fois des réalisations de la loi jointe et de la loi conditionnelle des données incomplètes sachant les observations. La qualité de l'estimation dépend grandement du choix du paramètre auxiliaire et de sa proximité avec le vrai paramètre θ.

Motivée par l'adaptation de la méthode ABC Shadow au cadre des données incomplètes, notre approche se base donc à nouveau sur l'échantillonnage de la loi *a posteriori* des paramètres :

$$f(\theta|\mathbf{y}) = \int \frac{e^{\langle t(\mathbf{x}\cup\mathbf{y})|\theta\rangle}}{Z(\theta)} \mu_{W_X}(\mathrm{d}\mathbf{x}) \times \frac{p(\theta)}{Z(\mathbf{y})}$$

Les approches proposées par les algorithmes basés sur les variables auxiliaires permettraient de faire disparaître la constante de normalisation des données et du modèle, mais pas de se dédouaner de l'intégrale toujours présente au numérateur. Pour pallier ce problème, nous choisissons d'échantillonner la loi *a posteriori* de $(\theta, \mathbf{x}|\mathbf{y})$, la loi jointe des paramètres et de la configuration non observée, avec l'idée d'échantillonner les lois marginales de ces données et des paramètres. La section suivante est dédiée à cette distribution et à la construction de deux chaînes de Markov pour échantillonner, en théorie et en pratique, cette dernière. Par soucis de «complétude» de ce manuscrit, nous présentons également une autre approche, bien qu'elle ne soit pas au cœur de nos projets de recherche actuels, elle motive également la création des algorithmes. La démarche qui suit peut être adaptée à cette autre distribution.

4.2 Proposition de lois *a posteriori* pour l'inférence en données incomplètes

Notations : Dans cette section, θ_Y désignera le paramètre du modèle obtenu à partir de l'observation $\mathbf{y}$ et θ_W désignera le paramètre sur W tout entier.

4.2.1 Loi jointe du paramètre et de la configuration manquante sachant l'observation

Nous allons ici nous intéresser à la distribution jointe des paramètres et de la configuration non observée sachant l'observation $\mathbf{y}$:

$$f(\theta_W, \mathbf{x}|\mathbf{y}) = f(\mathbf{x}|\theta_W, \mathbf{y}) p(\theta_W|\mathbf{y}) \tag{4.1}$$

$$\propto f(\mathbf{x}, \mathbf{y}|\theta_W) p(\theta_W) \tag{4.2}$$

où $p(\theta_W|\mathbf{y})$ et $p(\theta_W)$ sont des lois *a priori*.

L'heuristique derrière le choix de cette loi est la suivante : bien que l'espace produit des paramètres et des configurations sur W_X soit très grand, échantillonner cette loi permet de générer à la fois des configurations manquantes et des paramètres pour le modèle complet à partir des observations. Les travaux de cette thèse se concentrent davantage sur l'échantillonnage de cette loi plutôt que sur l'échantillonnage de la loi basé sur le modèle hiérarchique présenté précédemment. Bien que cette piste n'ait pas encore été abordée en pratique, la création des deux chaînes présentées dans les sous-sections suivantes est adaptable au modèle hiérarchique précédemment introduit. Ce modèle hiérarchique constitue ainsi une potentielle approche intéressante pour la suite de ces travaux mais ne sera pas abordé dans la suite. Les sections suivantes traitent de la construction de deux chaînes de Markov pour échantillonner la loi $f(\theta_W, \mathbf{x}|\mathbf{y})$ en s'inspirant de la construction des chaînes «idéale» et «shadow» introduite en partie 2.3.5.

4.2.2 Construction d'une chaîne de Markov admettant la loi *a posteriori* comme distribution invariante

4.2.2.1 Une construction basée sur l'algorithme de Metropolis-Hastings

Considérons l'algorithme de Metropolis-Hastings général : l'algorithme propose d'abord une nouvelle valeur pour l'état courant : $(\theta_W, \mathbf{x}) \to (\psi_W, \mathbf{x}')$. Cette nouvelle valeur est ensuite acceptée avec probabilité

$$\begin{aligned}
\alpha_i :&= \alpha_i\left((\theta_W, \mathbf{x}) \to (\psi_W, \mathbf{x}')\right) \\
&= \min\left\{1, \frac{f(\psi_W, \mathbf{x}'|\mathbf{y})}{f(\theta_W, \mathbf{x}|\mathbf{y})} \times \frac{q((\psi_W, \mathbf{x}') \to (\theta_W, \mathbf{x}))}{q((\theta_W, \mathbf{x}) \to (\psi_W, \mathbf{x}'))}\right\} \\
&= \min\left\{1, \frac{Z(\theta_W)p(\mathbf{y})}{Z(\psi_W)p(\mathbf{y})} \times \frac{e^{\langle t(\mathbf{y}\cup\mathbf{x}')|\psi_W\rangle}}{e^{\langle t(\mathbf{y}\cup\mathbf{x})|\theta_W\rangle}} \times \frac{p(\psi_W)}{p(\theta_W)} \times \frac{q((\psi_W, \mathbf{x}') \to (\theta_W, \mathbf{x}))}{q((\theta_W, \mathbf{x}) \to (\psi_W, \mathbf{x}'))}\right\}.
\end{aligned}$$

La convergence de cette chaîne de Markov va dépendre du choix de la loi de proposition $q(\cdot \to \cdot)$. Nous allons choisir pour cela une loi similaire à celle introduite pour la convergence de la chaîne idéale de l'ABC Shadow et une loi sur la configuration manquante.

4.2.2.2 Choix des lois de propositions

En supposant que les lois de propositions sont indépendantes entre elles, nous pouvons écrire :

$$q((\theta_W, \mathbf{x}) \to (\psi_W, \mathbf{x}')) = q(\theta_W \to \psi_W) \times q(\mathbf{x} \to \mathbf{x}').$$

En reprenant la même loi de proposition que dans [Stoica et al., 2017] pour proposer un nouveau paramètre : pour une valeur fixée de Δ et une réalisation $\mathbf{z}$ (sur W) du modèle $f(\cdot|\widetilde{\nu})$ pour une valeur de paramètre fixée de $\widetilde{\nu}$.

— $q(\theta_W \to \psi_W) := q_\Delta(\theta_W \to \psi_W|\mathbf{z}) = \frac{f(\mathbf{z}|\psi_W)/Z(\psi_W)}{I(\theta_W, \Delta, \mathbf{z})} \mathbb{1}_{b(\theta_W, \Delta/2)}(\psi_W)$

avec $I(\theta, \Delta, \mathbf{z}) = \int_{b(\theta,\Delta/2)} f(\mathbf{z}|\phi)/Z(\phi)\mathrm{d}\phi$ et $b(\theta, \Delta/2)$ la boule de centre Δ et de rayon $\Delta/2$.

— $q(\mathbf{x} \to \mathbf{x}')$: un Metropolis Hastings ajout/retrait de points. La proposition d'une nouvelle configuration constituera un ajout d'un point à $\mathbf{x}$ uniformément dans W_X ou une suppression d'un point parmi les points de $\mathbf{x}$.

Le ratio devient alors :

$$\alpha_i = \min\left\{1, \frac{Z(\theta_W)p(\mathbf{y})}{Z(\psi_W)p(\mathbf{y})} \times \frac{e^{\langle t(\mathbf{y}\cup\mathbf{x}')|\psi_W\rangle}}{e^{\langle t(\mathbf{y}\cup\mathbf{x})|\theta_W\rangle}} \times \frac{p(\psi_W)}{p(\theta_W)} \times \frac{Z(\psi_W)}{Z(\theta_W)} \times \frac{f(\mathbf{z}|\theta_W)}{f(\mathbf{z}|\psi_W)} \times \frac{q(\mathbf{x}' \to \mathbf{x})}{q(\mathbf{x} \to \mathbf{x}')} \times \frac{I(\theta_W, \Delta/2, \mathbf{z})}{I(\psi_W, \Delta/2, \mathbf{z})} \times \frac{\mathbb{1}_{b(\psi_W,\Delta_2)}(\theta_W)}{\mathbb{1}_{b(\theta_W,\Delta_2)}(\psi_W)}\right\} \tag{4.3}$$

Cette nouvelle expression ne dépend plus des constantes de normalisation du modèle. Néanmoins, le ratio d'intégrales $\frac{I(\theta_W,\Delta/2,\mathbf{z})}{I(\psi_W,\Delta/2,\mathbf{z})}$ est aussi difficile à calculer que les constantes de normalisation.

Pour échantillonner cette loi, une approche consiste à utiliser une combinaison entre le choix d'un nouveau paramètre, l'ajout d'un nouveau point et la suppression d'un point. Avec une probabilité p_p, un nouveau paramètre sera proposé selon $q_\Delta(\theta_W \to \cdot)$. Le noyau de transition est décrit par l'égalité suivante, pour $(A_W, F) \in \mathcal{T}_\Theta \times \mathcal{F}$:

$$\begin{aligned}
&P_{ideal}((\theta_W, \mathbf{x}) \to (A_W, F)) \\
&= p_b \int_{W_X} b(\mathbf{x}, \eta)\alpha_i\left((\theta_W, \mathbf{x}) \to (\theta_W, \mathbf{x} \cup \{\eta\})\right) \mathbb{1}[\mathbf{x} \cup \{\eta\} \in F]\mathbb{1}[\theta_W \in A_W]d\nu(\eta) \\
&+ p_d \sum_{\eta\in\mathbf{x}} d(\mathbf{x}, \eta)\alpha_i((\theta, \mathbf{x}) \to (\theta, \mathbf{x} \setminus \{\eta\}))\mathbb{1}[\mathbf{x} \setminus \{\eta\} \in F]\mathbb{1}[\theta_W \in A_W] \\
&+ p_p \int_{A_W} \alpha_i((\theta_W, \mathbf{x}) \to (\psi_W, \mathbf{x}))q_\Delta(\theta_W \to \psi_W)\mathbb{1}[\mathbf{x} \in F]\mathbb{1}[\psi_W \in A_W]d\psi_W \\
&+ \mathbb{1}[\mathbf{x} \in F]\mathbb{1}[\theta_W \in A_W]\Bigg(1- \\
&\quad p_b \int_{W_X} b(\mathbf{x}, \eta)\alpha_i\left((\theta_W, \mathbf{x}) \to (\theta_W, \mathbf{x} \cup \{\eta\})\right) d\nu(\eta) \\
&\quad - p_d \sum_{\eta\in\mathbf{x}} d(\mathbf{x}, \eta)\alpha_i((\theta, \mathbf{x}) \to (\theta, \mathbf{x} \setminus \{\eta\})) \\
&\quad - p_p \int_{A_W} \alpha_i((\theta_W, \mathbf{x}) \to (\psi_W, \mathbf{x}))q_\Delta(\theta_W \to \psi_W)d\psi_W\Bigg)
\end{aligned}$$

Proposition 4.2.1. *La chaîne admettant comme noyau P_{ideal} est apériodique et $\pi-$irréductible, où π est la loi a posteriori d'intérêt.*

Démonstration. Par construction, π est la loi invariante de la chaîne. Puisque la $\phi-$irréductibilité implique la $\pi-$irréductibilité ([Møller and Waagepetersen, 2004] Proposition 7.2), montrons la $\phi-$irréductibilité. Nous supposons que les choix de p_b, p_d et $b(\cdot, \eta), d(\cdot, \eta)$ permettent de vérifier les hypothèses posées dans la Proposition 7.13 de [Møller and Waagepetersen, 2004]. Par les mêmes arguments que ceux donnés dans la démonstration de cette proposition, en posant $\phi(A_W, F) = \mathbf{1}[(0, \emptyset) \in F], (A_W, F) \in \mathcal{T}_\Theta \times \mathcal{F}$. Puisque, individuellement, ces deux chaînes sont

$\phi-$irréductibles : la probabilité, à paramètre fixé, de revenir en $\emptyset$ depuis une certaine configuration $\mathbf{x}$ est strictement positive. Nous avons alors, si $m = n(\mathbf{x}) \geq 1$, $P^m_{ideal}((0,\mathbf{x}) \to (\{0\},\{\emptyset\})) > 0$. Ainsi, si $\phi(A_W, F) > 0$, alors $P^m_{ideal}((0,\mathbf{x}) \to (\{0\}, F)) > 0$ pour $m \geq 1$. De même, à configuration fixée, par $\phi-$irréductibilité par rapport à la première variable, la probabilité de revenir en 0 depuis un paramètre fixé en un nombre fini de pas m' est strictement positive. Ainsi, P_{ideal} est $\phi-$irrédutible.

Pour l'apériodicité, un critère vérifiable facilement est l'existence d'un état pour lequel la probabilité de rester est strictement positive. Nous obtenons alors, pour $\theta_W \in \Theta$, $P_{ideal}((\theta_W, \emptyset) \to (\{\theta_W\}, \{\emptyset\}) = p_d$ puisqu'il s'agit de la probabilité de rester dans l'état $\emptyset$. La chaîne est donc apériodique.

□

La chaîne ainsi créée converge vers la loi *a posteriori* d'intérêt. Néanmoins, cette chaîne ne peut pas être utilisée dans la pratique à cause du ratio d'intégrales $\frac{I(\theta_W, \Delta/2, \mathbf{z})}{I(\psi_W, \Delta/2, \mathbf{z})}$. Pour résoudre cette difficulté, la section suivante donne des détails sur la construction de la «chaîne shadow», une chaîne de Markov dont la dynamique va suivre celle de la «chaîne idéale» aussi proche que souhaité, pour un nombre fixé de pas.

Une autre manière de mettre à jour la CDM serait d'accepter ou non le nouvel état formé du nouveau paramètre et de la nouvelle configuration.

4.2.3 Construction d'une deuxième chaîne de Markov, la chaîne «shadow», suivant la dynamique de la première

Soient V_Δ le volume de la boule $b(\theta_W, \Delta/2)$ et $U_\Delta(\theta_W \to \psi_W) = \frac{1}{V_\Delta}\mathbb{1}_{b(\theta_W,\Delta/2)}(\psi_W)$ la densité de la loi uniforme sur $b(\theta_W, \Delta/2)$.

Théorème 4.2.2. *Soit* $\mathbf{z}$ *une réalisation dans* Ω *telle que les fonctions* $f(\mathbf{z}|\phi)$ *soient strictement positives et continues en* ϕ *et* Δ*, des réels positifs, alors :*

(i) *Les distributions de* $q_\Delta(\theta \to \cdot)$ *et* $U_\Delta(\theta \to \cdot)$ *vérifient :* $\forall \theta \in \Theta$ *fixe et* $A \in \mathcal{T}_\Theta$*,*

$$\lim_{\Delta \to 0^+} \int_A |q_\Delta(\theta \to \psi) - U_\Delta(\theta \to \psi)| d\psi = 0$$

(ii) *Pour tout* $\theta_W \in \Theta$ *fixé, la fonction* $\frac{q_\Delta(\theta_W \to \cdot)}{q_\Delta(\cdot \to \theta_W)}$ *et* $\frac{\frac{f(\mathbf{z}|\cdot)}{Z(\cdot)}\mathbf{1}_{b(\theta_W,\Delta/2)}(\cdot)}{\frac{f(\mathbf{z}|\theta_W)}{Z(\theta_W)}\mathbf{1}_{b(\cdot,\Delta/2)}(\theta_W)}$ *vérifient :*

$$\lim_{\Delta \to 0} \sup_{\psi_W \in \Theta} \left| \frac{q_\Delta(\theta_W \to \psi_W|\mathbf{z})}{q_\Delta(\psi_W \to \theta_W|\mathbf{z})} - \frac{\frac{f(\mathbf{z}|\psi_W)}{Z(\psi_W)}\mathbf{1}_{b(\theta_W,\Delta/2)}(\psi_W)}{\frac{f(\mathbf{z}|\theta_W)}{Z(\theta_W)}\mathbf{1}_{b(\psi_W,\Delta/2)}(\theta_W)} \right| = 0$$

uniformément en $\theta_W \in \Theta$*. De plus, si* $f(\mathbf{z}|\cdot)$ *est* $\mathcal{C}^1(\Theta)$*, les vitesses de convergence dans* ***(i)*** *et* ***(ii)*** *peuvent être explicitées.*

Démonstration. Ce théorème est démontré dans l'article [Stoica et al., 2017] et est donné en annexe A du manuscrit. □

Ce théorème justifie l'usage d'une seconde chaîne de Markov qui va jumeler sa dynamique avec celle de la chaîne «idéale» pour Δ proche de 0 pour un nombre de pas fixé. Cette nouvelle chaîne a la dynamique suivante : supposons la chaîne dans l'état $(\theta_W, \mathbf{x})$, une nouvelle valeur pour ψ_W est proposée uniformément dans la boule $b(\theta_W, \Delta_2/2)$. Le nouveau pattern est proposé par une proposition de type Metropolis-Hastings avec ajout/retrait de points. Le nouvel état est alors accepté avec probabilité :

$$\begin{aligned}\alpha_s &:= \alpha_s((\theta_W, \mathbf{x}) \to (\psi_W, \mathbf{x}')) \\ &= \min\left\{1, \frac{Z(\theta_W)p(\mathbf{y})}{Z(\psi_W)p(\mathbf{y})} \times \frac{e^{\langle t(\mathbf{y}\cup\mathbf{x}')|\psi_W\rangle}}{e^{\langle t(\mathbf{y}\cup\mathbf{x})|\theta_W\rangle}} \times \frac{p(\psi_W)}{p(\theta_W)} \times \frac{Z(\psi_W)}{Z(\theta_W)}\right. \\ &\qquad \left. \times \frac{q(\mathbf{x}' \to \mathbf{x})}{q(\mathbf{x} \to \mathbf{x}')} \times \frac{f(\mathbf{z}|\theta_W)}{f(\mathbf{z}|\psi_W)} \times \frac{\mathbb{1}_{b(\psi_W,\Delta_2/2)}(\theta_W)}{\mathbb{1}_{b(\theta_W,\Delta_2/2)}(\psi_W)}\right\} \end{aligned} \tag{4.4}$$

Du théorème 4.2.2, nous pouvons en déduire un corollaire sur les ratios d'acceptations donnés par les ratios α_i et α_s :

Corollaire 4.2.3. *Les probabilités d'acceptation définies des chaînes «idéales» et «shadow», définis par (4.3) et (4.4), sont uniformément aussi proches que souhaité lorsque* Δ *tend vers* 0.

Proposition 4.2.4. *Soient* $P_i := P_{ideal}$ *et* $P_s := P_{shadow}$ *les noyaux de transition proposés respectivement pour la chaîne «idéale» et la chaîne «shadow». Comme dans le théorème 4.2.2, nous supposons que* $\Delta > 0$ *et que* $\mathbf{z} \in \Omega$ *est une configuration sur* W.
Pour tout $\epsilon > 0$ *et tout* $n \in \mathbb{N}$, *il existe* $\Delta_i = \Delta_i(\epsilon, n) > 0$ *tel que pour tout* $\Delta \leq \Delta_i$, *alors* $|P_i^{(n)}((\theta_W, \mathbf{x}) \to (A_W, F)) - P_s^{(n)}((\theta_W, \mathbf{x}) \to (A_W, F))| < \epsilon$ *uniformément en* $\theta_W \in \Theta$ *et* $A_W \in \mathcal{T}_\Theta$

Démonstration. La preuve est similaire à celle donnée dans [Stoica et al., 2017], les termes impliquant l'ajout et le retrait disparaissant lors du calcul de la différence de $|P_i - P_s|$, laissant le mécanisme de mise à jour des paramètres à contrôler comme dans l'article.

□

Maintenant que la chaîne est construite et justifiée théoriquement, nous pouvons introduire l'algorithme pour simuler cette chaîne. Les résultats précédents donnent la construction d'un algorithme ABC, nous résumons son fonctionnement en pseudo-code :

Algorithme 4.2.1. *ABC Shadow données manquantes*

ABC Shadow données manquantes

Posons Δ un paramètre de perturbation, N_{ABC} un nombre d'itérations, p_b, p_d et p_p tels que $p_b + p_d + p_p = 1$, les probabilités de proposer respectivement d'ajouter un point, de supprimer un point et de proposer une nouvelle valeur de paramètre. Supposons que la configuration observée soit $\mathbf{y}$ et que l'état actuel soit $(\theta_W^0, \mathbf{x}^0)$

1) Générer une configuration $\mathbf{z}$ sur W suivant $f(\mathbf{z}|\theta_W^0)$.

2) Pour $k = 1$ à n :

- Avec probabilité p_b, proposer d'ajouter un point η à $\mathbf{x}$ selon $b(\mathbf{x}, \eta)$. Accepter cette modification avec probabilité $\alpha_s((\theta_W, \mathbf{x}) \to (\theta_W, \mathbf{x} \cup \{\eta\}))$ et définir $(\theta_W^{k+1}, \mathbf{x}^{k+1}) = (\theta_W, \mathbf{x} \cup \{\eta\})$
- Avec probabilité p_d, proposer de supprimer un point $\eta \in \mathbf{x}$ selon $d(\mathbf{x}, \eta)$. Accepter cette suppression avec probabilité $\alpha_s((\theta_W, \mathbf{x}) \to (\theta_W, \mathbf{x} \setminus \{\eta\}))$ et définir $(\theta_W^{k+1}, \mathbf{x}^{k+1}) = (\theta_W, \mathbf{x} \setminus \{\eta\})$
- Avec probabilité p_p, proposer une nouvelle valeur ψ_W selon $q_\Delta(\mathbf{x}, \eta)$. Accepter cette nouvelle valeur avec probabilité $\alpha_s((\theta_W, \mathbf{x}) \to (\psi_W, \mathbf{x}))$ et définir $(\theta_W^{k+1}, \mathbf{x}^{k+1}) = (\psi_W, \mathbf{x})$

3) Renvoyer $(\theta_W^n, \mathbf{x}^n)$.

4) Aller en **1)** si d'autres échantillons sont nécessaires avec l'état initial $(\theta_W^n, \mathbf{x}^n)$.

4.2.4 Loi *a posteriori* basée sur un modèle hiérarchique

Une autre idée de loi *a posteriori* utile à l'inférence en donnée incomplète émerge en considérant un deuxième paramètre, θ_Y. Cette idée amène à considérer un modèle hiérarchique décrit par la distribution de $f(\theta_W, \theta_Y, \mathbf{x}|\mathbf{y})$. Une idée naturelle serait de restructurer cette distribution en fonction des sous-distributions conditionnelles afin d'obtenir des distributions que nous savons contrôler.

$$f(\theta_W, \theta_Y, \mathbf{x}|\mathbf{y}) = f(\mathbf{x}|\theta_W, \theta_Y, \mathbf{y}) p(\theta_W|\theta_Y, \mathbf{y}) p(\theta_Y|\mathbf{y})$$

Supposons maintenant que la distribution de $f(\mathbf{x}|\theta_W, \theta_Y, \mathbf{y})$ ne dépend que de θ_W et $\mathbf{y}$ et posons $p(\theta_W|\theta_Y, \mathbf{y}) = p(\theta_W|\theta_Y)$, nous pouvons écrire :

$$f(\theta_W, \theta_Y, \mathbf{x}|\mathbf{y}) = f(\mathbf{x}|\theta_W, \mathbf{y}) p(\theta_W|\theta_Y) p(\theta_Y|\mathbf{y}) \tag{4.5}$$

où $f(\mathbf{x}|\theta_W, \mathbf{y})$ est la distribution conditionnelle des configurations sur W_X sachant $\mathbf{y}$, $p(\theta_W|\theta_Y)$ une loi *a priori* à préciser et $p(\theta_Y|\mathbf{y})$ la loi *a posteriori* de θ_Y sachant $\mathbf{y}$.

Cette approche repose sur l'idée que les deux paramètres, bien que différents, devraient être relativement proches l'un de l'autre. Si ce n'était pas le cas, cela voudrait dire que la configuration non-observée est complètement différente de celle observée et que les hypothèses que nous faisons ne sont pas vérifiées.

Nous pouvons maintenant discuter des choix pour la distribution de $p(\theta_W|\theta_Y)$. Tant que nous partons du principe que θ_W doit être «proche» de θ_Y, la distribution $p(\theta_W|\theta_Y)$ peut être soit une loi uniforme sur l'intervalle $[\theta_Y - \delta; \theta_Y + \delta]$ avec $\delta > 0$, soit une distribution gaussienne avec une moyenne θ_Y et une variance σ^2, $\mathcal{N}(\theta_Y, \sigma^2)$, avec $\sigma > 0$. Pour $p(\theta_Y)$, notre choix se portera sur la loi *a posteriori* $p(\theta_Y|\mathbf{y})$ avec un hyper-prior uniforme afin que cette distribution soit proportionnelle à la vraisemblance pour les données complètes $\mathbf{y}$. Ce choix est motivé par l'algorithme ABC Shadow que nous pouvons appliquer à la configuration observée.

Des pseudo-codes similaires à celui exposé plus haut peuvent être donnés pour échantillonner la loi $f(\theta_W, \theta_Y, \mathbf{x}|\mathbf{y})$. À partir de ces idées, nous allons tester des stratégies numériques afin d'échantillonner la loi $f(\theta_W, \mathbf{x}|\mathbf{y})$. Ces stratégies sont plus faciles à réaliser numériquement en utilisant l'état actuel de la `DRLib`, des travaux plus approfondis sur le sujet sont en cours.

4.3 Implémentation des algorithmes pour l'inférence en données incomplètes

À partir de ce que nous avons présenté plus haut, nous proposons différents algorithmes d'échantillonnage dérivés de la construction ci-avant. Gardant à l'esprit que la loi clé à échantillonner est la loi *a posteriori* $f(\theta_W|\mathbf{y})$, nous avons introduit des distributions permettant d'approcher cette distribution en la considérant comme une loi marginale des lois introduites en 4.5 et 4.1. La loi introduite en 4.1 suggère alors l'utilisation de l'ABC Shadow sur les observations pour obtenir un échantillon de $(\theta_Y|\mathbf{y})$ et l'utilisation de ces valeurs pour la loi $p(\theta_W|\theta_Y)$. La loi introduite en 4.5 et l'algorithme destiné à l'échantillonner suggèrent quant à eux de simuler des configurations sur W_X conditionnellement à $\mathbf{y}$ à l'aide de valeurs candidates pour θ_W. Les solutions introduites par [Gelfand and Carlin, 1993] et [Geyer, 1999] suggèrent également l'utilisation d'un paramètre auxiliaire, de simulations issues de la loi conditionnelle des données non-observées sachant les observations et de la loi jointe des observations et des données manquantes. Le paramètre estimé sur les données observées, $\widehat{\theta_Y}$, ne devrait pas être trop éloigné du paramètre pour le modèle complet θ_W. Enfin, l'avantage de l'ABC Shadow en données complètes est que nous avons la possibilité d'obtenir un échantillon pour θ_Y basé sur les données observées. Les stratégies proposées se basent sur l'estimation obtenue à l'aide de l'ABC Shadow sur les données observées $\mathbf{y}$ et sur des simulations sur les zones non observées pour les données manquantes conditionnellement à cette observation.

L'idée repose sur les étapes suivantes :

1. Échantillonnage du paramètre $\widehat{\theta_Y}$ en se basant seulement sur l'observation $\mathbf{y}$ en utilisant l'algorithme ABC Shadow sur W_Y pour obtenir un échantillon de $f(\theta_Y|Y = \mathbf{y})$, la loi *a posteriori* en données complètes.
2. Utiliser $\widehat{\theta_Y}$ pour simuler des configurations sur W_X conditionnellement à l'observation $\mathbf{y}$. Pour ce faire, nous pouvons utiliser différentes valeurs pour θ_Y comme le mode, la moyenne ou des valeurs aléatoires de l'échantillon obtenu précédemment.

3. Utiliser les configurations simulées sur W_X et les observations pour former une vraisemblance. À partir de là, nous pouvons supposer que nous travaillons en données complètes avec des données composées des observations et des simulations.
4. Enfin, l'algorithme ABC Shadow est utilisé pour obtenir un échantillon de $f(\theta_W|\mathbf{y}, \widetilde{\mathbf{x}})$ où $\widetilde{\mathbf{x}}$ fait référence à l'échantillon simulé. Plus précisément, l'algorithme ABC Shadow prenant en entrées les statistiques suffisantes du modèle, $\overline{t(\mathbf{y} \cup \widetilde{\mathbf{x}})}$ sera utilisé.

À partir d'estimations $\widehat{\theta_Y}$, le but sera alors d'estimer $\mathbb{E}_{\widehat{\theta_Y}}[t(X \cup Y)|Y = \mathbf{y}]$, l'espérance des statistiques suffisantes sur W_X en sachant l'observation $\mathbf{y}$. Pour cela, nous pouvons approcher cette espérance par son espérance empirique

$$\widehat{\mathbb{E}}_{\widehat{\theta_Y}} t(X) = \frac{1}{n} \sum_{i=1}^{n} t(X_i^*)$$

où $X_1^*, ..., X_n^*$ sont des réalisations de la distribution conditionnelle de X sachant $Y = \mathbf{y}$. Puisque plusieurs choix de $\widehat{\theta_Y}$ sont possibles, nous pouvons également prendre plusieurs valeurs de $\widehat{\theta_Y}$ tirées de $f(\theta_Y|\mathbf{y})$ et estimer une moyenne pour chacune de ces valeurs et ainsi obtenir une estimation de :

$$\frac{1}{n_{\widehat{\theta_Y}}} \sum_{i=1}^{n_{\widehat{\theta_Y}}} \mathbb{E}_{\widehat{\theta_Y}}[t(X)|Y = \mathbf{y}]$$

avec $n_{\widehat{\theta_Y}}$ un nombre de paramètres extrait fixé.

À partir de cette démarche, nous pouvons utiliser différentes valeurs de $\widehat{\theta_Y}$. Dans la suite, lors de l'étude par simulations sur les modèles de Strauss, nous utiliserons :

Algorithme 1 :

1. Générer un échantillon $(\theta_{Y_1}, ...\theta_{Y_n})$ tiré suivant $f(\theta_Y|\mathbf{y})$ avec l'algorithme ABC Shadow.
2. Extraire n_{θ_Y} valeurs aléatoires de cet échantillon et simuler pour chaque θ_Y extrait N_c configurations sur W_X conditionnellement à $\mathbf{y}$.
3. Calculer les moyennes empiriques des statistiques suffisantes $\left(\widehat{\mathbb{E}}_{\widehat{\theta_{Y_i}}} t(X)\right)_{1 \leq i \leq n_{\theta_Y}}$.
4. Pour chacune de ces moyennes, l'algorithme ABC Shadow est utilisé avec les moyennes obtenues.
5. Les échantillons obtenus par l'algorithme ABC Shadow sont alors concaténés pour obtenir un échantillon de θ_W.

Ce premier algorithme se rapproche des algorithmes proposés par [Gelfand and Carlin, 1993] et [Geyer, 1999] de par l'utilisation de plusieurs paramètres $\widehat{\theta_Y}$ et des simulations conditionnelles, un lien reste à établir entre cette proposition d'algorithme et l'algorithme EM. L'utilisation de plusieurs valeurs aléatoires de paramètres pour simuler des configurations sur W_X permet de mimer la dynamique décrite dans le pseudo-code 4.2.3. Les algorithmes qui suivent permettent de simplifier les coûts numériques tout en s'inspirant de ce premier algorithme.

Algorithme 2 :

1. Générer un échantillon $(\theta_{Y_1}, ...\theta_{Y_n})$ tiré suivant $f(\theta_Y|\mathbf{y})$ avec l'algorithme ABC Shadow.
2. Extraire n_{θ_Y} valeurs aléatoires de cet échantillon et simuler pour chaque θ_Y extrait N_c configurations sur W_X conditionnellement à $\mathbf{y}$.
3. Calculer les moyennes empiriques des statistiques suffisantes $\left(\widehat{\mathbb{E}}_{\widehat{\theta_{Y_i}}} t(X)\right)_{1\leq i\leq n_{\theta_Y}}$.
4. Prenant cette fois-ci la moyenne de ces moyennes, l'algorithme ABC Shadow est utilisé avec cette approximation comme entrée pour obtenir un échantillon de $(\theta_W|\overline{t(\mathbf{y}\cup\widetilde{\mathbf{x}})})$.

Algorithme 3 :

1. Générer un échantillon $(\theta_{Y_1}, ...\theta_{Y_n})$ tiré suivant $f(\theta_Y|\mathbf{y})$ avec l'algorithme ABC Shadow.
2. Prendre $\widehat{\theta_Y}$ comme le mode de l'histogramme de l'échantillon et générer N_c configurations sur W_X conditionnellement à $\mathbf{y}$.
3. Calculer la moyenne empirique des statistiques suffisantes $\widehat{\mathbb{E}}_{\widehat{\theta_Y}} t(X)$.
4. À partir de cette moyenne empirique, utiliser l'algorithme ABC Shadow pour obtenir un échantillon de $(\theta_W|\overline{t(\mathbf{y}\cup\widetilde{\mathbf{x}})})$

4.4 Étude par simulations

Dans cette section, nous proposons de comparer les algorithmes exposés ci-dessus pour différentes configurations observées et pour différentes valeurs du paramètre γ_s du modèle de Strauss.

4.4.1 Modèles considérés, scénarios d'observation

4.4.1.1 Modèles de Strauss

Pour rappel, le modèle de Strauss est défini par la densité de probabilité

$$f(\mathbf{y}|\rho,\gamma_s) \propto \exp(n(\mathbf{y})\log(\rho) + s_r(\mathbf{y})\log(\gamma_s)).$$

Le modèle de Strauss pouvant contrôler la répulsion entre les points, nous souhaitons prendre différentes valeurs du paramètre afin de regarder l'impact de la force d'interaction sur l'inférence. Le paramètre θ dans la section précédente fait ici référence au vecteur (que nous prendrons en échelle log) de paramètres $(\log\rho, \log\gamma_s)$. Dans cette étude, nous considérons 4 valeurs pour ces paramètres : $(5.7, -0.11)$; $(5.7, -0.29)$; $(5.7, -0.69)$; $(5.7, -2.30)$. ($\ln(300) = 5.7$ et $\gamma_s \in \{0.9, 0.75, 0.5, 0.1\}$).

4.4.1.2 Fenêtre observée et domaine complet

Le processus ponctuel est défini sur la fenêtre $W = [0, 1.2] \times [0, 1]$. Nous considérons 3 scénarios pour la fenêtre observée : $Y = [0, 1] \times [0, 1]$; $Y = [0, 0.8] \times [0, 1]$ et $Y = [0, 0.6] \times [0, 1]$ afin de mesurer l'impact de la surface de la zone manquante sur l'inférence. Pour chaque ensemble de paramètres présenté ci-dessus, une simulation via l'algorithme de Metropolis-Hastings est

effectuée afin d'obtenir ce que nous supposons être les données complètes sur $[0, 1.2] \times [0, 1]$ (voir figure 4.3 ci-dessous). Nous extrayons ensuite les différentes fenêtres observées évoquées précédemment, ce qui donne 4 modèles observés sur chaque fenêtre observée $W_Y = [0, 1] \times [0, 1]$; $W_Y = [0, 0.8] \times [0, 1]$ et $W_Y = [0, 0.6] \times [0, 1]$ illustrées par les différentes couleurs dans la figure 4.2. Cela donne alors 12 cas de figures à traiter au total.

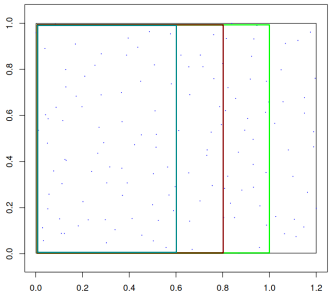
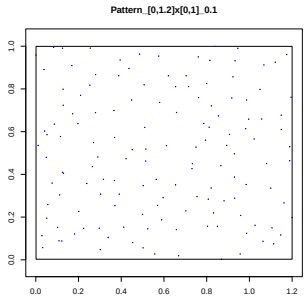
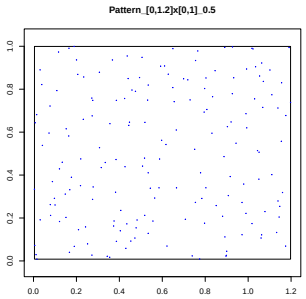


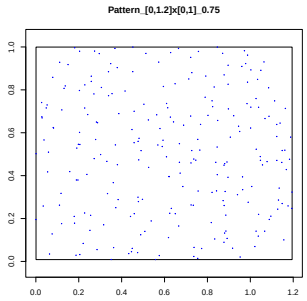
FIGURE 4.2 – Configuration obtenue pour $(\ln(\rho), \ln(\gamma_s)) = (5.7, -2.30)$. La fenêtre verte représente $W_Y = [0, 1]$; la fenêtre rouge représente $W_Y = [0, 0.8]$ et la bleue représente $W_Y = [0, 0.6]$.



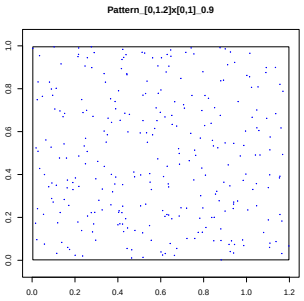
(a) Configuration obtenue pour $(\ln(\rho), \ln(\gamma_s)) = (5.7, -2.30)$



(b) Configuration obtenue pour $(\ln(\rho), \ln(\gamma_s)) = (5.7, -0.69)$



(c) Configuration obtenue pour $(\ln(\rho), \ln(\gamma_s)) = (5.7, -0.29)$



(d) Configuration obtenue pour $(\ln(\rho), \ln(\gamma_s)) = (5.7, -0.11)$

FIGURE 4.3 – 4 réalisations du processus de Strauss simulées sur $[0, 1.2] \times [0, 1]$ via l'algorithme MH. Chacune de ces réalisations sera divisée en W_Y et W_X comme illustré ci-dessus.

4.4.2 Applications et résultats

4.4.2.1 Statistiques observées

Puisque l'algorithme ABC Shadow repose sur les statistiques suffisantes du modèle, nous résumons sous forme d'un tableau les statistiques suffisantes observées. La deuxième colonne ne sera, en pratique, pas connue mais il est intéressant, puisque nous sommes dans une étude par simulation, de garder la trace des «vraies» données.

	$(n(\mathbf{y}), s_r(\mathbf{y}))$ on			
$(\ln(\rho), \ln(\gamma_s))$	$[0, 1.2] \times [0, 1]$	$[0, 1] \times [0, 1]$	$[0, 0.8] \times [0, 1]$	$[0, 0.6] \times [0, 1]$
$(5.7, -0.11)$	$(298, 262)$	$(259, 243)$	$(209, 200)$	$(153, 146)$
$(5.7, -0.29)$	$(241, 142)$	$(197, 105)$	$(151, 76)$	$(110, 54)$
$(5.7, -0.69)$	$(177, 57)$	$(147, 48)$	$(118, 41)$	$(94, 34)$
$(5.7, -2.30)$	$(140, 9)$	$(120, 9)$	$(90, 6)$	$(72, 5)$

TABLE 4.1 – Valeurs de $(n(\mathbf{y}), s_r(\mathbf{y}))$ pour chaque fenêtre observée. La colonne $[0, 1.2] \times [0, 1]$ n'est pas connue en pratique pour une étude en données incomplètes.

La décroissance des statistiques suffisantes n'est pas linéaire par rapport à la surface non-observée, ce qui fera émerger des valeurs différentes pour les estimations basées seulement sur l'observation.

4.4.2.2 Estimation sur W, données complètes

Afin d'obtenir une vraie référence pour une comparaison avec l'inférence que nous allons mener dans la suite et puisque nous avons accès à toutes les données, nous avons exécuté l'algorithme ABC Shadow pour chaque cas. En entrée de l'algorithme, les statistiques observées sur $W = [0, 1.2] \times [0, 1]$ décrites en deuxième colonne du tableau 4.1 ont été utilisées. Nous avons paramétré l'algorithme de manière à obtenir un échantillon de taille 10^6, Δ a été fixé à $(0.01, 0.01)$ lorsque $\ln(\gamma) = -0.69$ ou $\ln(\gamma) = -2.30$ et à $(0.01, 0.001)$ lorsque $\ln(\gamma) = -0.11$ ou $\ln(\gamma) = -0.29$ en raison de la proximité de la limite supérieure 0. Le tableau ci-dessous résume les paramètres de la procédure :

Variable	Description	Valeur
W	Domaine de simulation	$[0, 1.2] \times [0, 1]$
N_θ	Nombre d'échantillons de θ	10^6
N_{ABC}	Nombre d'itérations proposition paramètre	100
N_{MH}	Nombre d'itérations MH	500
p_b, p_d	Probabilité d'ajout/retrait	$0.5, 0.5$
θ_0	Paramètres initiaux	Aléatoire
$\Delta = (\Delta_\rho, \Delta_{\gamma_s})$	Paramètres de perturbation des paramètres	$(0.01, 0.01/0.001)$
$[\rho_{min}, \rho_{max}]$	Support de la loi uniforme sur ρ	$[0, 10]$
$[\gamma_{s,min}, \gamma_{s,max}]$	Support de la loi uniforme sur γ_s	$[-5, 0]$

Les figures ci-dessous illustrent les box-plots obtenus pour $\ln(\rho)$ (à gauche) et $\ln(\gamma_s)$ (à droite). La ligne rouge représente le paramètre réel (5.7 et $\gamma = 0.1$).

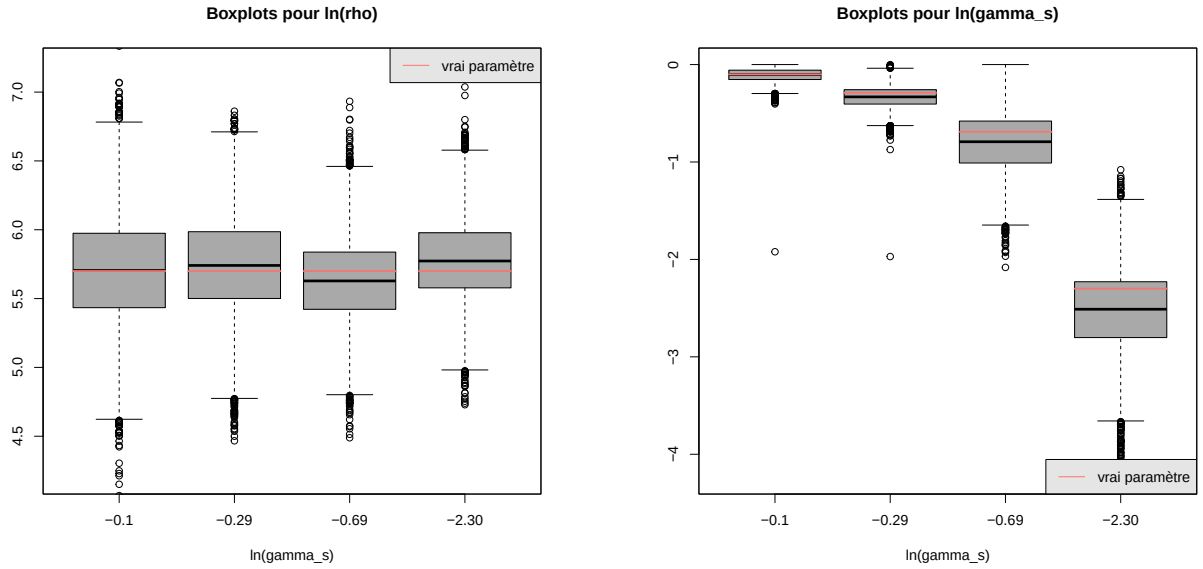


FIGURE 4.4 – Box-plots pour l'échantillon de $\log(\rho)$ (gauche) et $\log(\gamma_s)$ (droite). Les lignes rouges représentent les vrais paramètres. Pour chaque figure en partant de la gauche, les box-plots représentent l'échantillon obtenu pour $\gamma_s = 0.9$, $\gamma_s = 0.75$, $\gamma_s = 0.5$, $\gamma_s = 0.1$ respectivement.

Les vrais paramètres sont toujours contenus dans les box-plots, indiquant une bonne estimation. Le tableau ci-dessous résume les valeurs estimées pour les paramètres en prenant le mode des histogrammes des échantillons.

$(\ln(\rho), \ln(\gamma_s))$	Estimation de $(\ln(\rho), \ln(\gamma_s))$
$(5.7, -0.11)$	(5.70,-0.1)
$(5.7, -0.29)$	(5.75,-0.34)
$(5.7, -0.69)$	$(5.60, -0.78)$
$(5.7, -2.30)$	(5.75,-2.5)

TABLE 4.2 – Estimation de $(\ln \rho, \ln \gamma_s)$ sur W pour chaque couple de paramètres.

Une légère différence s'observe pour l'estimation par rapport aux paramètres réels, ceci est dû au fait que l'observation n'est constituée que d'une seule réalisation.

4.4.2.3 Estimation sur W_Y

Tous les algorithmes proposés s'appuient sur un échantillon de $(\theta_Y|\mathbf{y})$. Comme dans la sous-section précédente, nous utilisons l'algorithme ABC Shadow pour cela. Les paramètres seront les mêmes que ceux décrits dans le tableau 4.4.2.2 à l'exception que le domaine dans lequel l'inférence se conduit sera W_Y pour les différents cas étudiés ($[0,1] \times [0,1]; [0,0.8] \times [0,1]; [0,0.6] \times [0,1]$). Les box-plots qui suivent décrivent les échantillons obtenus pour $\log \rho$ (gauche) et $\log \gamma_s$ (droite) en fonction de la fenêtre observée.

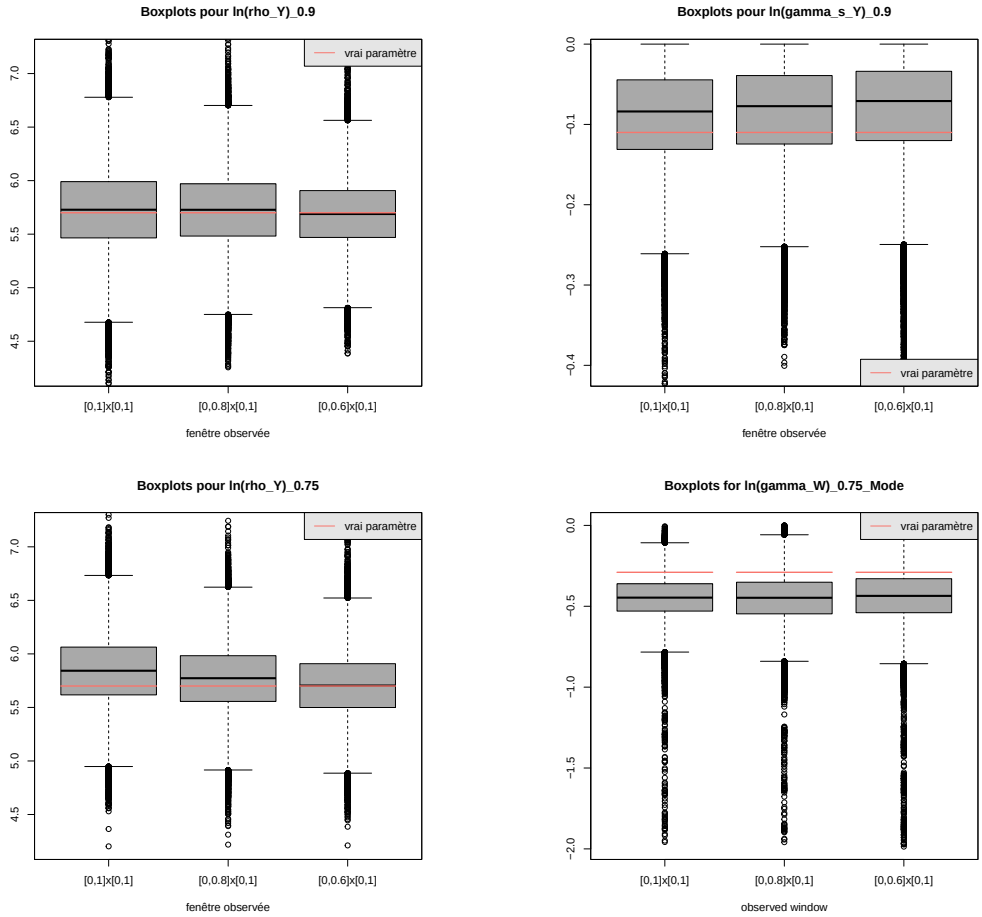


FIGURE 4.5 – Box-plots obtenus pour $\log \rho_Y$ (haut-gauche) et $\log \gamma_{s_Y}$ (haut-droite) dans le cas où les vrais paramètres sont fixés à $\log \rho = 5.7$ et $\log \gamma_s = -0.11$.
Box-plots obtenus pour $\log \rho_Y$ (bas gauche) et $\log \gamma_{s_Y}$ (bas droite) dans le cas où les vrais paramètres sont fixés à $\log \rho = 5.7$ et $\log \gamma_s = -0.29$. Pour chaque figure, de gauche à droite : box-plot pour $W_Y = [0,1] \times [0,1]$, $Y = [0, 0.8] \times [0,1]$, $Y = [0, 0.6] \times [0,1]$ respectivement.

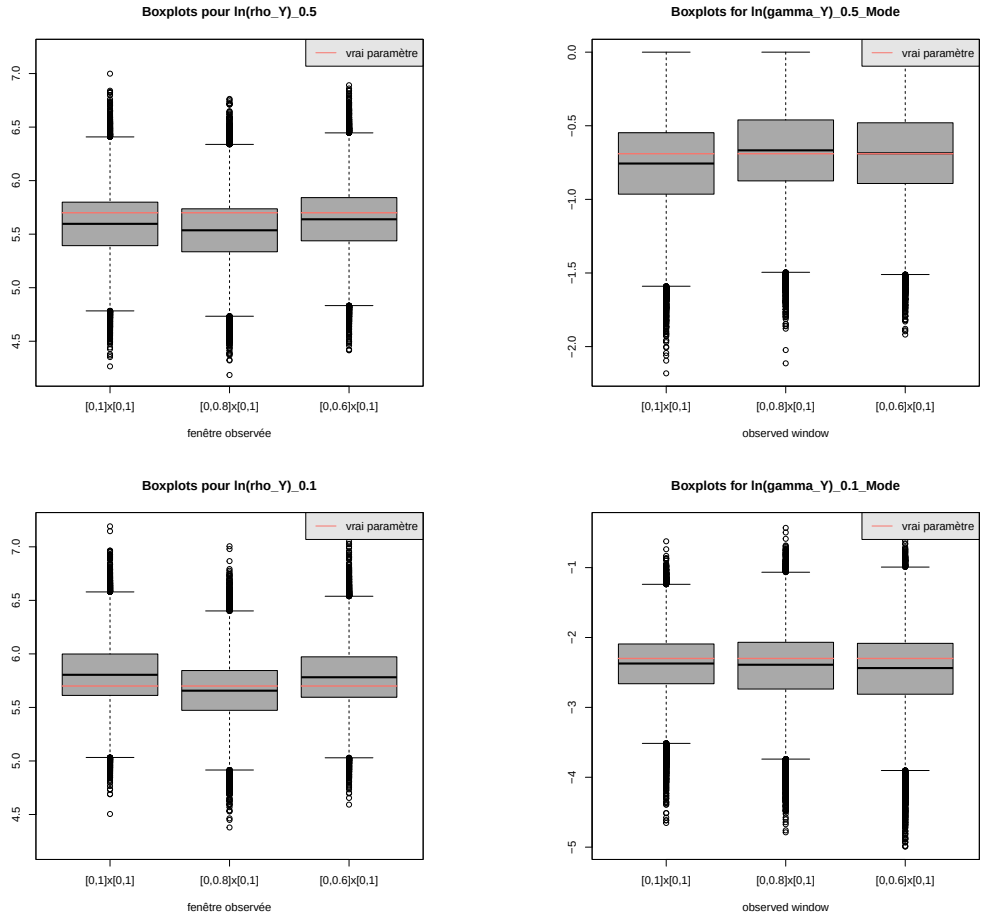


FIGURE 4.6 – Box-plots obtenus pour $\log \rho_Y$ (haut gauche) et $\log \gamma_{s_Y}$ (haut droite) dans le cas où les vrais paramètres sont fixés à $\log \rho = 5.7$ et $\log \gamma_s = -0.69$.
Box-plots obtenus pour $\log \rho_Y$ (bas gauche) et $\log \gamma_{s_Y}$ (bas droite) dans le cas où les vrais paramètres sont fixés à $\log \rho = 5.7$ et $\log \gamma_s = -2.30$. Pour chaque figure, de gauche à droite : box-plot pour $W_Y = [0,1] \times [0,1]$, $Y = [0, 0.8] \times [0,1]$, $Y = [0, 0.6] \times [0,1]$ respectivement.

	$(\ln(\rho_Y), \ln(\gamma_{s_Y}))$ estimation pour $W_Y =$		
Paramètres réels $(\ln(\rho), \ln(\gamma_s))$	$[0,1] \times [0,1]$	$[0, 0.8] \times [0,1]$	$[0, 0.6] \times [0,1]$
$(5.7, -0.11)$	$(5.70, -0.08)$	$(5.75, -0.06)$	$(5.65, -0.02)$
$(5.7, -0.29)$	$(5.85, -0.46)$	$(5.75, -0.44)$	$(5.65, -0.46)$
$(5.7, -0.69)$	$(5.58, -0.76)$	$(5.52, -0.74)$	$(5.64, -0.70)$
$(5.7, -2.30)$	$(5.78, -2.35)$	$(5.66, -2.40)$	$(5.74, -2.40)$

TABLE 4.3 – Estimation de $(\ln(\rho_Y), \ln(\gamma_{s_Y}))$ sur Y pour chaque couple de paramètre et fenêtre observée

Nous constatons un écart par rapport à l'estimation résumée dans le tableau 4.2 et par rapport au vrai paramètre, ce qui est cohérent avec un phénomène d'effet de bord dû à la perte de points observés dans W_X. Nous allons utiliser ces valeurs pour effectuer les simulations sur W_X conditionnellement à l'observation $\mathbf{y}$, les parties qui suivent montrent les résultats obtenus pour

les trois algorithmes.

4.4.2.4 Résultats pour l'Algorithme 1

Cette procédure a été effectuée en répétant les étapes suivantes pour chaque θ_Y extrait de l'échantillon $(\theta_{Y_1}, ...\theta_{Y_n})$ tiré suivant $f(\theta_Y|\mathbf{y})$:

1. Générer $N_c = 1000$ configurations sur W_X conditionnellement à $\mathbf{y}$ en utilisant θ_Y.
2. Générer un échantillon de taille $n_{\theta_W} = 10^4$ de θ_W à l'aide de l'ABC Shadow en utilisant les configurations simulées et l'observation.

Comme précédemment, le tableau suivant décrit les réglages de l'algorithme ABC Shadow.

Variable	Description	Valeur
W_X	Domaine de simulation	En fonction du cas
N_θ	Nombre d'échantillons de θ	10^6
N_{ABC}	Nombre d'itérations proposition paramètre	100
N_{MH}	Nombre d'itérations MH	500
p_b, p_d	Probabilité d'ajout/retrait	$0.5, 0.5$
θ_0	Paramètres initiaux	Aléatoire
$\Delta = (\Delta_\rho, \Delta_{\gamma_s})$	Paramètres de perturbation des paramètres	$(0.01, 0.01/0.001)$
$[\rho_{min}, \rho_{max}]$	Support de la loi uniforme sur ρ	$[0, 10]$
$[\gamma_{s,min}, \gamma_{s,max}]$	Support de la loi uniforme sur γ_s	$[-5, 0]$

Nous obtenons ainsi un échantillon de taille 10^6 de θ_W. Nous illustrons à nouveau cette procédure pour $W_X = [1, 1.2] \times [0, 1]$ et $\ln(\gamma_s) = -2.30$. La ligne rouge représente le paramètre réel, la ligne jaune représente le mode de l'histogramme, la ligne verte représente l'estimation effectuée sur les données observées $\mathbf{y}$ et la ligne bleue représente l'estimation complète des données effectuée en 4.2. Nous remarquons une estimation très proche pour les paramètres ρ et γ_s. La figure 4.8 donne la dynamique de l'échantillonnage de la distribution jointe des paramètres et les séries temporelles des paramètres.

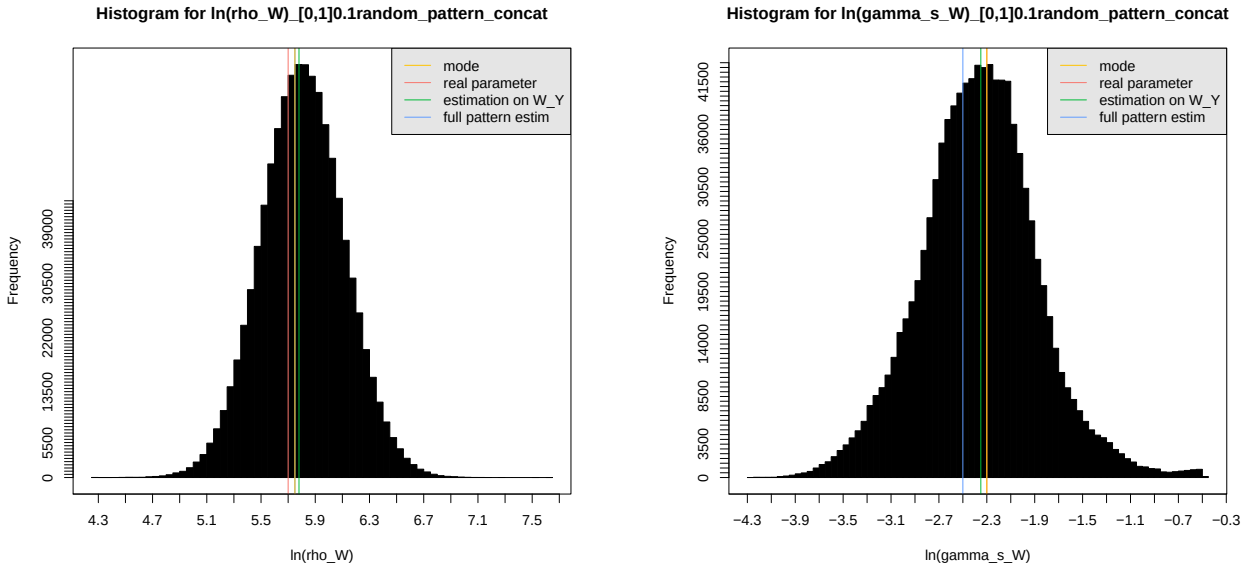


FIGURE 4.7 – Histogrammes pour les échantillons de $\log(\rho_W)$ (gauche) et $\log(\gamma_{s_W})$ (droite). La ligne rouge représente le paramètre réel, la ligne jaune représente le mode de l'histogramme, la ligne verte représente l'estimation effectuée sur les données observées $\mathbf{y}$ et la ligne bleue représente l'estimation complète des données résumée dans le tableau 4.2.

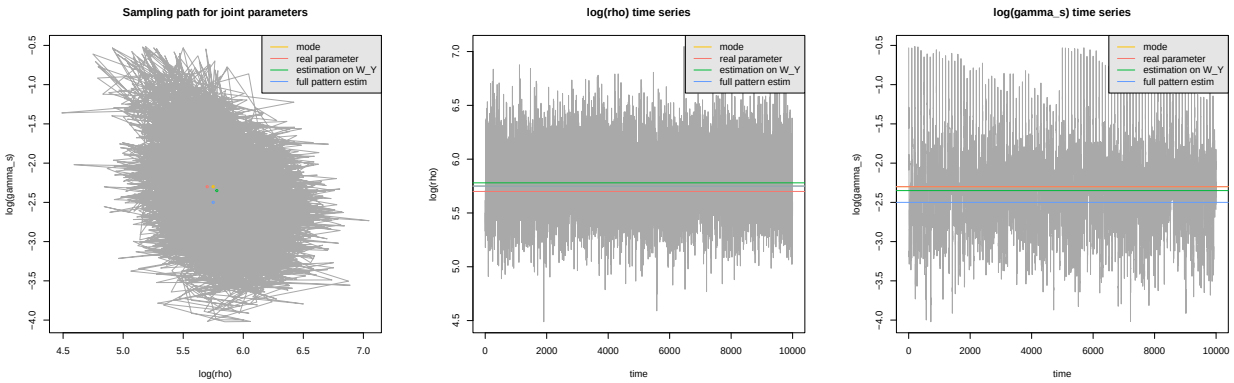


FIGURE 4.8 – Exemple de dynamique pour la loi jointe des deux paramètres (gauche) ; série chronologique de $\ln(\rho_W)$ (milieu) ; série chronologique de $\ln(\gamma_{s_W})$ (droite).

Nous constatons que l'inférence ainsi réalisée à l'aide des simulations a permis d'obtenir une nouvelle valeur pour l'estimation des paramètres. Cette fois-ci, l'écart entre l'estimation obtenue sur W_Y et celle obtenue avec la configuration complète est légèrement plus petit que celui entre l'estimation avec la configuration complète et l'estimation basée sur l'observation et les simulations. Les box-plots et le tableau ci-après résument les estimations obtenues.

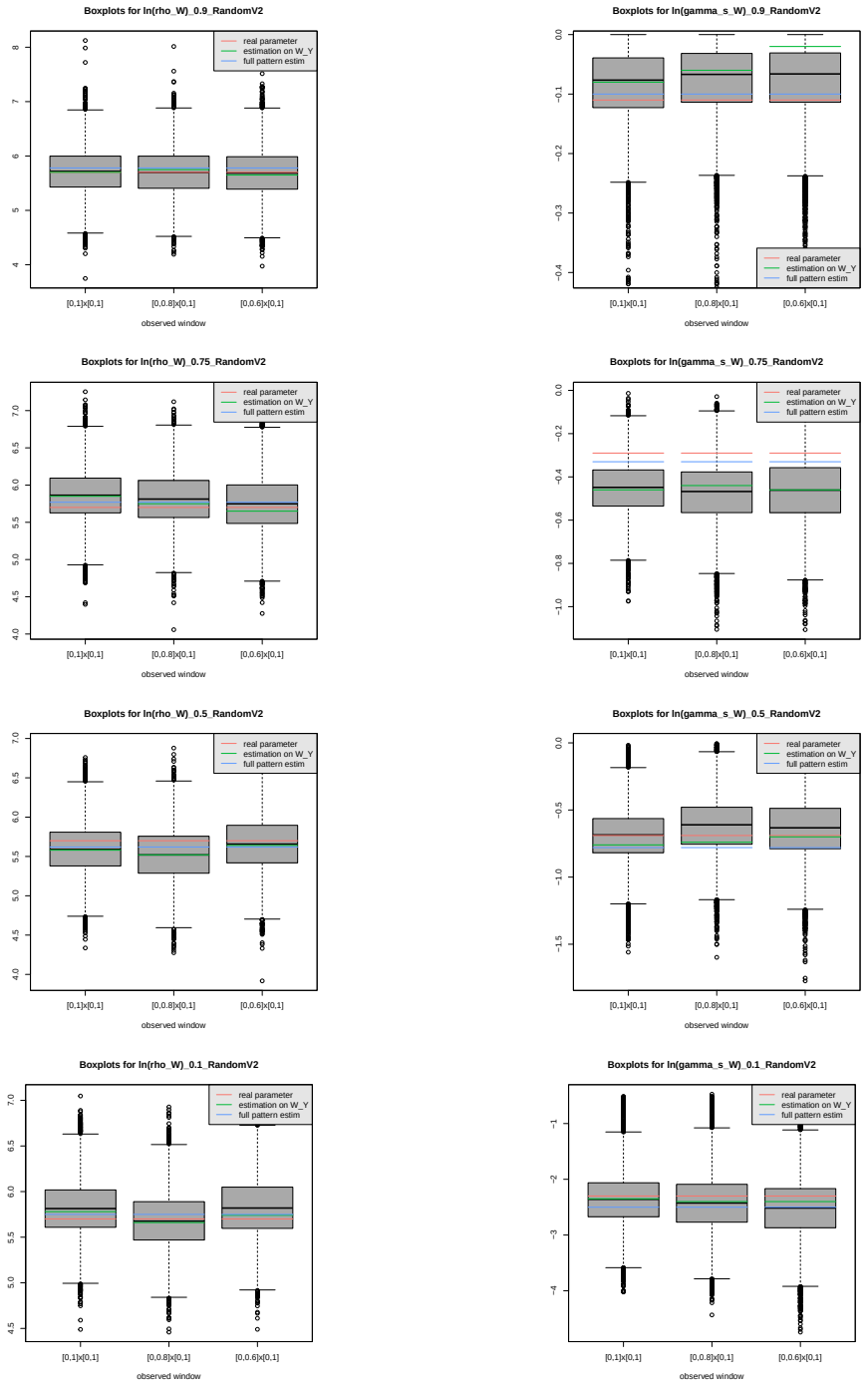


FIGURE 4.9 – Box-plots pour l'estimation de $\ln(\rho_W)$ (colonne gauche) et $\ln(\gamma_{s_W})$ (colonne droite) pour chaque couple de paramètres (du haut vers le bas : $(5.7, -0.11)$; $(5.7, -0.29)$; $(5.7, -0.69)$; $(5.7, -2.30)$). Chaque figure montre trois box-plots représentant la fenêtre observée : $Y = [0,1] \times [0,1]$ (gauche), $Y = [0, 0.8] \times [0,1]$ (milieu) ; $Y = [0, 0.6] \times [0,1]$ (droite).

	$(\ln(\rho_W), \ln(\gamma_{s_W}))$ pour $W_Y =$		
Paramètres réels $(\ln(\rho), \ln(\gamma_s))$	$[0,1] \times [0,1]$	$[0,0.8] \times [0,1]$	$[0,0.6] \times [0,1]$
$(5.7, -0.11)$	$(5.7, -0.06)$	$(5.65, -0.01)$	$(5.65, -0.01)$
$(5.7, -0.29)$	$(5.85, -0.44)$	$(5.75, -0.48)$	$(5.75, -0.47)$
$(5.7, -0.69)$	$(5.55, -0.68)$	$(5.5, -0.6)$	$(5.65, -0.6)$
$(5.7, -2.30)$	$(5.75, -2.3)$	$(5.65, -2.35)$	$(5.8, -2.5)$

TABLE 4.4 – Estimation de $(\ln(\rho_W), \ln(\gamma_{s_W}))$ sur W pour chaque couple de paramètres et chaque cas d'observation.

L'inférence menée avec cette méthode donne des résultats plutôt corrects, l'échantillonnage obtenu à l'aide des simulations semble presque toujours «corriger» le biais induit par la censure des données. En effet, la médiane des box-plots semble souvent plus proche de l'estimation menée sur les données complètes (en bleu) que ne l'est l'estimation menée sur l'observation seule. Avant de résumer tous ces résultats dans un seul et même tableau, nous détaillons les résultats pour les deux autres algorithmes proposés.

4.4.2.5 Résultats pour l'Algorithme 2

Cette fois-ci, nous avons extrait un $n_{\theta_Y} = 100-$échantillon de $\theta_Y \in p(\theta_Y|\mathbf{y})$ pour chaque cas. Chacun de ces θ_Y a été utilisé pour générer 1000 configurations sur W_X, puis la moyenne des statistiques suffisantes de ces configurations est utilisée comme entrée pour l'algorithme ABC Shadow avec le même réglage que précédemment. Le tableau 4.5 décrit les statistiques suffisantes obtenues sur W.

	$(\overline{n(\mathbf{y} \cup \mathbf{x})}, \overline{s_r(\mathbf{y} \cup \mathbf{x})})$ pour $W_Y =$		
Paramètres réels $(\ln(\rho), \ln(\gamma_s))$	$[0,1] \times [0,1]$	$[0,0.8] \times [0,1]$	$[0,0.6] \times [0,1]$
$(5.7, -0.11)$	$(313.29, 300.13)$	$(313.65, 307.97)$	$(312, 309.42)$
$(5.7, -0.29)$	$(239.59, 130.23)$	$(228.74, 117.18)$	$(221.66, 110.96)$
$(5.7, -0.69)$	$(179.43, 60.93)$	$(176.39, 63.37)$	$(192.49, 75.79)$
$(5.7, -2.30)$	$(144.96, 11.37)$	$(135.72, 9.08)$	$(144.28, 9.90)$

TABLE 4.5 – Valeurs de $(\overline{n(\mathbf{y} \cup \mathbf{x})}, \overline{s_r(\mathbf{y} \cup \mathbf{x})})$ pour chaque cas, complété avec des statistiques simulées.

À nouveau, la figure ci-dessous illustre les histogrammes obtenus pour $\ln(\rho_W)$ (gauche) et $\ln(\gamma_{s_W})$ (droite) pour $W_X = [1, 1.2] \times [0, 1]$ et $\ln(\gamma_s) = -2.30$. La ligne rouge représente le paramètre réel, la ligne jaune représente le mode de l'histogramme, la ligne verte représente l'estimation effectuée sur les données observées $\mathbf{y}$ et la ligne bleue représente l'estimation complète des données effectuée en 4.2. Nous remarquons une estimation très proche pour les paramètres ρ et γ_s. La figure 4.11 donne la dynamique de l'échantillonnage de la distribution jointe des paramètres et les séries temporelles des paramètres.

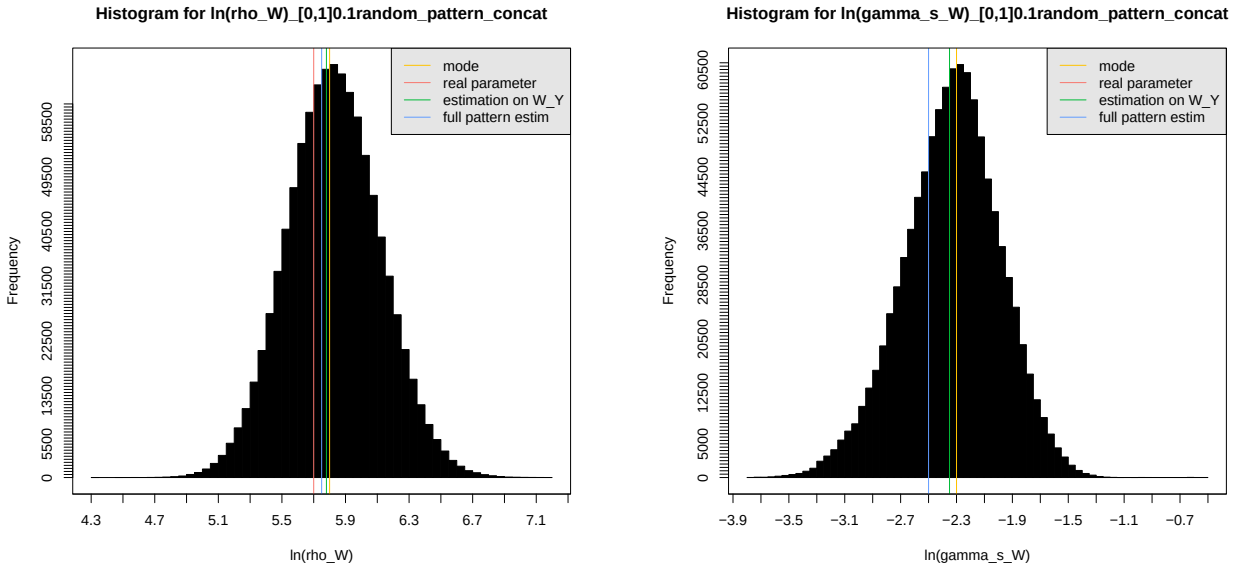


FIGURE 4.10 – Histogrammes pour les échantillons de $\log(\rho_W)$ (gauche) et $\log(\gamma_{s_W})$ (droite). La ligne rouge représente le paramètre réel, la ligne jaune représente le mode de l'histogramme, la ligne verte représente l'estimation effectuée sur les données observées $\mathbf{y}$ et la ligne bleue représente l'estimation complète des données résumée dans le tableau 4.2.

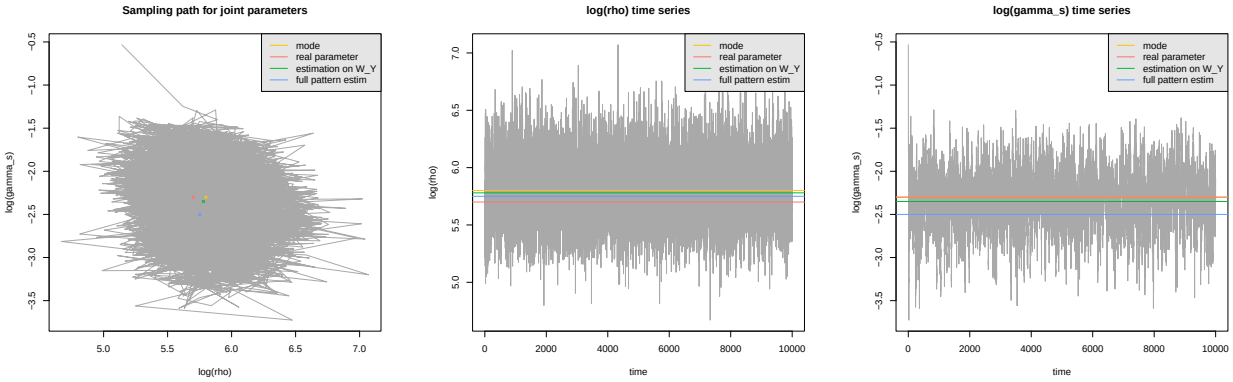


FIGURE 4.11 – Exemple de dynamique pour la loi jointe des deux paramètres (gauche) ; série chronologique de $\ln(\rho_W)$ (milieu) ; série chronologique de $\ln(\gamma_{s_W})$ (droite).

Nous constatons que l'inférence réalisée à l'aide des simulations a permis d'obtenir une nouvelle valeur pour l'estimation des paramètres. Cette fois-ci, l'écart entre l'estimation obtenue sur W_Y et celle obtenue avec la configuration complète est légèrement plus petit que celui entre l'estimation avec la configuration complète et l'estimation basée sur l'observation et les simulations. Les box-plots et le tableau ci-après résument les estimations obtenues.

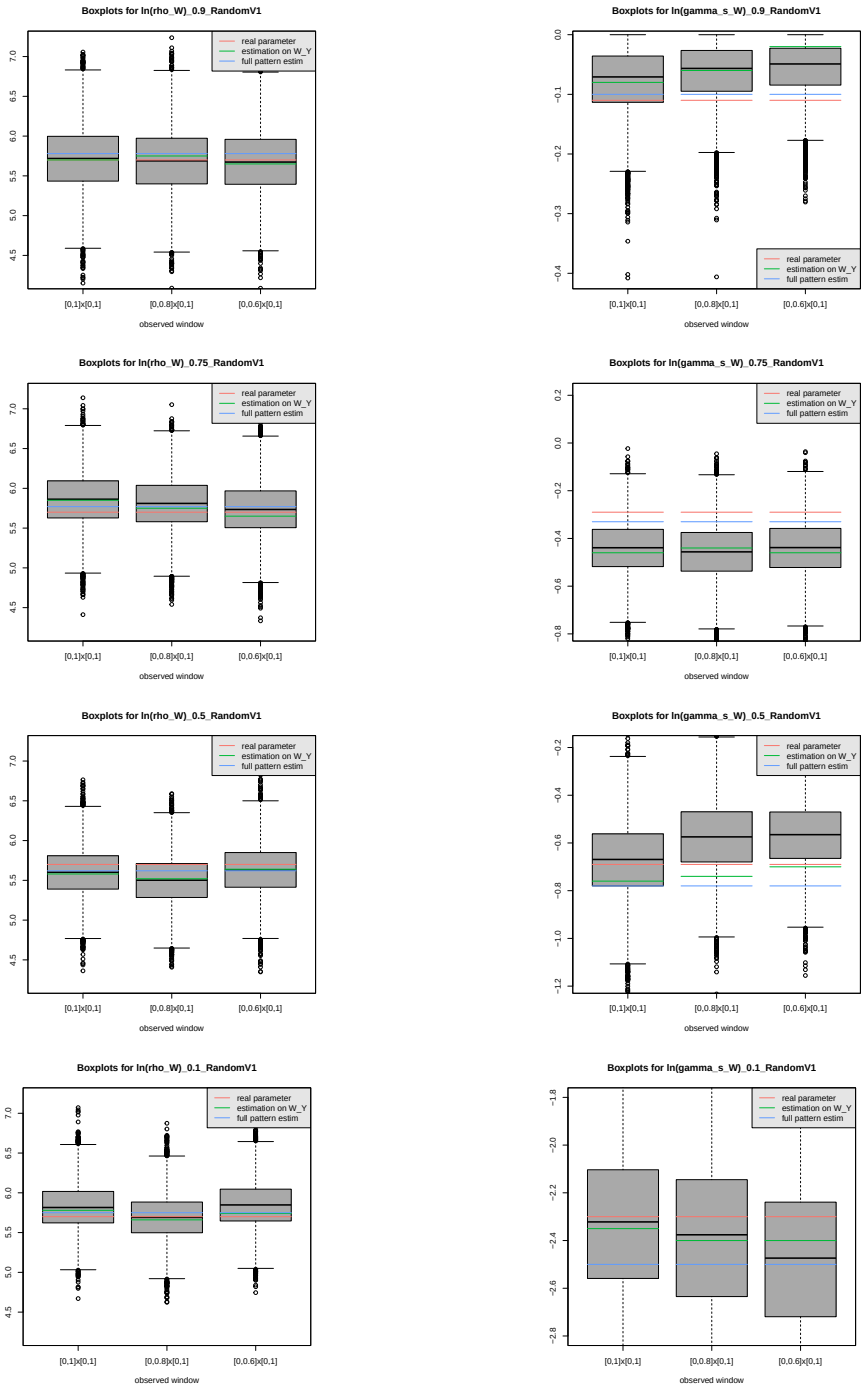


FIGURE 4.12 – Box-plots pour l'estimation de $\ln(\rho_W)$ (colonne gauche) et$\ln(\gamma_{s_W})$ (colonne droite) pour chaque couple de paramètres (du haut vers le bas : $(5.7, -0.11)$; $(5.7, -0.29)$; $(5.7, -0.69)$; $(5.7, -2.30)$). Chaque figure montre trois box-plots représentant la fenêtre observée : $Y = [0,1] \times [0,1]$ (gauche), $Y = [0,0.8] \times [0,1]$ (milieu) ; $Y = [0,0.6] \times [0,1]$ (droite).

	$(\ln(\rho_W), \ln(\gamma_{s_W}))$ pour $W_Y =$		
Paramètres réels $(\ln(\rho), \ln(\gamma_s))$	$[0,1] \times [0,1]$	$[0,0.8] \times [0,1]$	$[0,0.6] \times [0,1]$
$(5.7, -0.11)$	$(5.75, -0.05)$	$(5.65, -0.01)$	$(5.65, -0.01)$
$(5.7, -0.29)$	$(5.85, -0.44)$	$(5.8, -0.46)$	$(5.75, -0.44)$
$(5.7, -0.69)$	$(5.6, -0.66)$	$(5.45, -0.58)$	$(5.6, -0.56)$
$(5.7, -2.30)$	$(5.8, -2.3)$	$(5.65, -2.35)$	$(5.8, -2.45)$

TABLE 4.6 – Estimation de $(\ln(\rho_W), \ln(\gamma_{s_W}))$ sur W pour chaque couple de paramètres et chaque cas d'observation.

À nouveau, l'inférence menée avec cette méthode donne des résultats plutôt corrects, l'échantillonnage ainsi obtenu à l'aide des simulations semble presque toujours «corriger» le biais induit par la censure des données. En effet, la médiane des box-plots semble souvent plus proche de l'estimation menée sur les données complètes (en bleu) que ne l'est l'estimation menée sur l'observation seule.

4.4.2.6 Résultats pour l'Algorithme 3

À partir des estimations résumées dans le tableau 4.3, nous avons lancé 10^6 itérations de l'algorithme MH. Les échantillons ont été conservés toutes les 1000 itérations, donnant 1000 configurations sur W_X pour chaque cas. Le tableau 4.7 décrit les statistiques complètes obtenues sur W que nous avons utilisées comme entrées de l'algorithme ABC Shadow.

	$(\overline{n(\mathbf{y} \cup \mathbf{x})}, \overline{s_r(\mathbf{y} \cup \mathbf{x})})$ pour $W_Y =$		
Paramètres réels $(\ln(\rho), \ln(\gamma_s))$	$[0,1] \times [0,1]$	$[0,0.8] \times [0,1]$	$[0,0.6] \times [0,1]$
$(5.7, -0.11)$	$(310.74, 290.30)$	$(320.41, 313.90)$	$(316.97, 314.14)$
$(5.7, -0.29)$	$(236.86, 125.82)$	$(226.20, 113.20)$	$(213.49, 101.10)$
$(5.7, -0.69)$	$(175.76, 57.09)$	$(173.39, 457.79)$	$(185.20, 64.57)$
$(5.7, -2.30)$	$(143.91, 10.74)$	$(135.11, 8.73)$	$(141.43, 9.53)$

TABLE 4.7 – Valeurs de $(\overline{n(\mathbf{y} \cup \mathbf{x})}, \overline{s_r(\mathbf{y} \cup \mathbf{x})})$ pour chaque cas, complété avec des statistiques simulées.

La comparaison de ces valeurs avec les statistiques suffisantes complètes (réelles) permet déjà de se faire une idée de l'impact de la perte d'informations sur W_X. Nous constatons que les simulations permettent d'obtenir des statistiques suffisantes assez proches des données complètes lorsque le nombre de points observés est élevé (e.g. pour $W_X = [0,1] \times [0,1]$), mais qu'elles s'éloignent dès que la taille de W_X augmente.

La figure ci-dessous illustre les histogrammes obtenus pour $\ln(\rho_W)$ (gauche) et $\ln(\gamma_{s_W})$ (droite) pour $W_X = [1, 1.2] \times [0,1]$ et $\ln(\gamma_s) = -2.30$. La ligne rouge représente le paramètre réel, la ligne jaune représente le mode de l'histogramme, la ligne verte représente l'estimation effectuée sur les données observées $\mathbf{y}$ et la ligne bleue représente l'estimation complète des données effectuée en 4.2. Nous remarquons une estimation très proche pour les paramètres ρ et γ_s. La figure 4.14 donne la dynamique de l'échantillonnage de la distribution jointe des paramètres et les séries temporelles des paramètres.

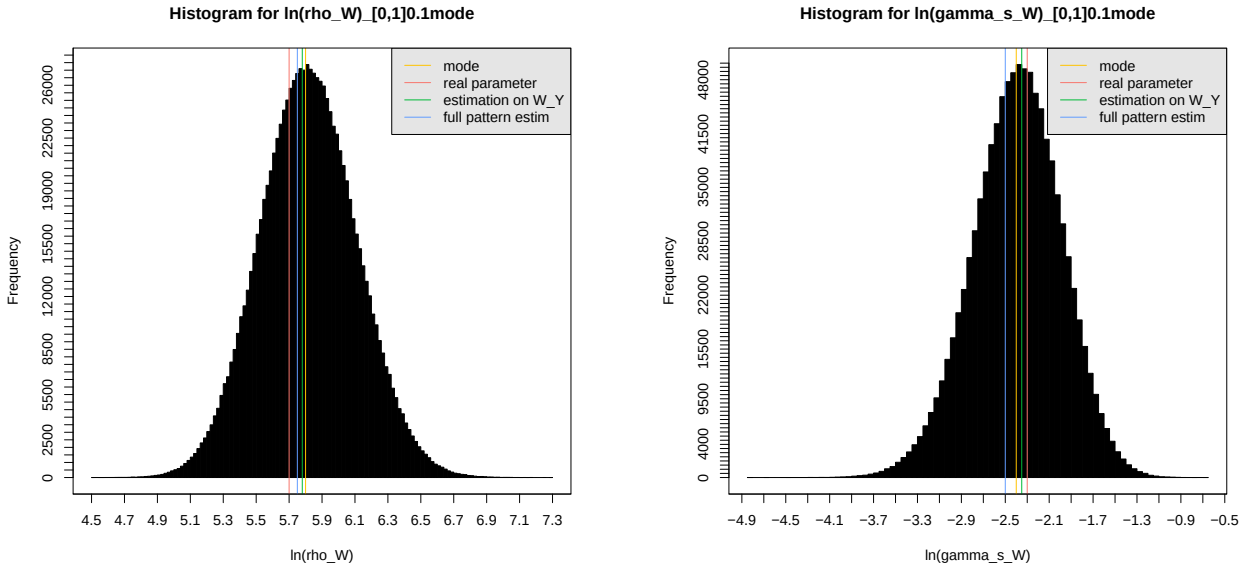


FIGURE 4.13 – Histogrammes pour les échantillons de $\log(\rho_W)$ (gauche) et $\log(\gamma_{s_W})$ (droite). La ligne rouge représente le paramètre réel, la ligne jaune représente le mode de l'histogramme, la ligne verte représente l'estimation effectuée sur les données observées $\mathbf{y}$ et la ligne bleue représente l'estimation complète des données résumée dans le tableau 4.2.

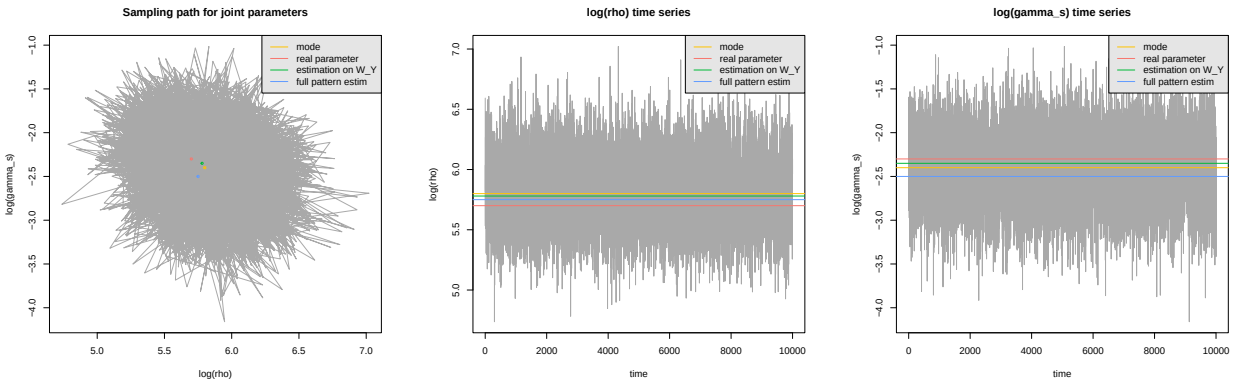


FIGURE 4.14 – Exemple de dynamique pour la loi jointe des deux paramètres (gauche) ; série chronologique de $\ln(\rho_W)$ (milieu) ; série chronologique de $\ln(\gamma_{s_W})$ (droite).

Nous constatons que l'inférence ainsi réalisée à l'aide des simulations a permis d'obtenir une nouvelle valeur pour l'estimation des paramètres. Dans le cas présent, l'écart entre l'estimation obtenue sur W_Y et celle obtenue avec la configuration complète semble plus grand que celui entre l'estimation avec la configuration complète et l'estimation basée sur l'observation et les simulations. Les box-plots et le tableau ci-après résument les estimations obtenues.

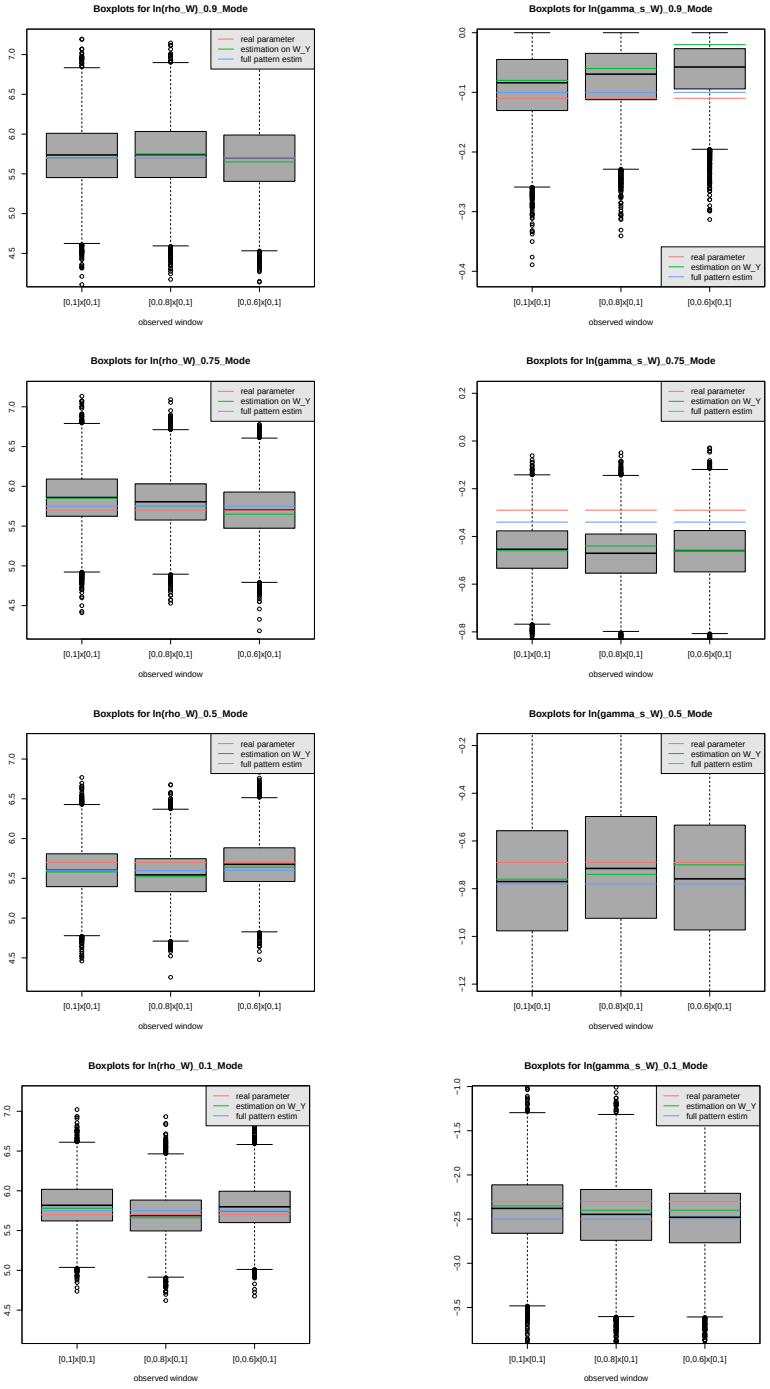


FIGURE 4.15 – Box-plots pour l'estimation de $\ln(\rho_W)$ (colonne gauche) et $\ln(\gamma_{s_W})$ (colonne droite) pour chaque couple de paramètres (du haut vers le bas : $(5.7, -0.11)$; $(5.7, -0.29)$; $(5.7, -0.69)$; $(5.7, -2.30)$). Chaque figure montre trois box-plots représentant la fenêtre observée : $Y = [0,1] \times [0,1]$ (gauche), $Y = [0,0.8] \times [0,1]$ (milieu) ; $Y = [0,0.6] \times [0,1]$ (droite)

	$(\ln(\rho_W), \ln(\gamma_{s_W}))$ pour $W_Y =$		
Paramètres réels $(\ln(\rho), \ln(\gamma_s))$	$[0,1] \times [0,1]$	$[0,0.8] \times [0,1]$	$[0,0.6] \times [0,1]$
$(5.7, -0.11)$	$(5.65, -0.06)$	$(5.7, -0.05)$	$(5.7, -0.01)$
$(5.7, -0.29)$	$(5.85, -0.45)$	$(5.8, -0.48)$	$(5.7, -0.46)$
$(5.7, -0.69)$	$(5.6, -0.8)$	$(5.5, -0.7)$	$(5.65, -0.78)$
$(5.7, -2.30)$	$(5.8, -2.4)$	$(5.66, -2.45)$	$(5.75, -2.5)$

TABLE 4.8 – Estimation de $(\ln(\rho_W), \ln(\gamma_{s_W}))$ sur W pour chaque couple de paramètres et chaque cas d'observation.

L'inférence menée avec cette méthode donne des résultats plutôt corrects, l'échantillonnage ainsi obtenu à l'aide des simulations semble presque toujours «corriger» le biais induit par la censure des données. En effet, la médiane des box-plots semble souvent plus proche de l'estimation menée sur les données complètes (en bleu) que ne l'est l'estimation menée sur l'observation seule. Nous allons maintenant résumer la totalité de l'inférence et comparer ces approches.

4.4.3 Comparaison des résultats

Le tableau ci-dessous résume l'inférence menée avec tous les algorithmes utilisés en 4.4.2.4, 4.4.2.5 et 4.4.2.6.

	$(\ln(\beta_W), \ln(\gamma_W))$ Estimation pour $W_Y =$		
Algorithme utilisé Paramètres réels	$[0,1] \times [0,1]$	$[0,0.8] \times [0,1]$	$[0,0.6] \times [0,1]$
Algorithme 1 $(5.7, -0.11)$	$(5.7, -0.06)$	$(5.65, -0.01)$	$(5.65, -0.01)$
Algorithme 1 $(5.7, -0.29)$	$(5.85, -0.44)$	$(5.75, -0.48)$	$(5.75, -0.47)$
Algorithme 1 $(5.7, -0.69)$	$(5.55, -0.68)$	$(5.5, -0.6)$	$(5.65, -0.6)$
Algorithme 1 $(5.7, -2.30)$	$(5.75, -2.3)$	$(5.65, -2.35)$	$(5.8, -2.5)$
Algorithme 2 $(5.7, -0.11)$	$(5.75, -0.05)$	$(5.65, -0.01)$	$(5.65, -0.01)$
Algorithme 2 $(5.7, -0.29)$	$(5.85, -0.44)$	$(5.8, -0.46)$	$(5.75, -0.44)$
Algorithme 2 $(5.7, -0.69)$	$(5.6, -0.66)$	$(5.45, -0.58)$	$(5.6, -0.56)$
Algorithme 2 $(5.7, -2.30)$	$(5.8, -2.3)$	$(5.65, -2.35)$	$(5.8, -2.45)$
Algorithme 3 $(5.7, -0.11)$	$(5.65, -0.06)$	$(5.7, -0.05)$	$(5.7, -0.01)$
Algorithme 3 $(5.7, -0.29)$	$(5.85, -0.45)$	$(5.8, -0.48)$	$(5.7, -0.46)$
Algorithme 3 $(5.7, -0.69)$	$(5.6, -0.8)$	$(5.5, -0.7)$	$(5.65, -0.78)$
Algorithme 3 $(5.7, -2.30)$	$(5.8, -2.4)$	$(5.66, -2.45)$	$(5.75, -2.5)$

TABLE 4.9 – Tableau récapitulatif des estimations.

À première vue, les valeurs ont l'air plutôt correctes et similaires selon les stratégies adoptées. Pour les comparer, nous allons nous référer à l'inférence obtenue en supposant que les données étaient complètes. Les tableaux 4.10 suivants donnent respectivement :

— la différence en valeur absolue entre les estimations basées sur les simulations et celles obtenues en prenant la configuration complète.
— la différence en valeur absolue entre les estimations basées seulement sur l'observation et celles obtenues en prenant la configuration complète.

	Différence entre les estimations via les stratégies et les estimations en données complètes		
Stratégie Paramètres réels	$[0,1]\times[0,1]$	$[0,0.8]\times[0,1]$	$[0,0.6]\times[0,1]$
Algorithme 1 $(5.7,-0.11)$	$(0.00,0.04)$	$(0.05,0.09)$	$(0.05,0.09)$
Algorithme 1 $(5.7,-0.29)$	$(0.10,0.10)$	$(0.00,0.14)$	$(0.00,0.13)$
Algorithme 1 $(5.7,-0.69)$	$(0.05,0.10)$	$(0.10,0.18)$	$(0.05,0.18)$
Algorithme 1$(5.7,-2.30)$	$(0.00,0.20)$	$(0.10,0.15)$	$(0.05,0.00)$
Algorithme 2 $(5.7,-0.11)$	$(0.05,0.05)$	$(0.05,0.09)$	$(0.05,0.09)$
Algorithme 2 $(5.7,-0.29)$	$(0.10,0.10)$	$(0.05,0.12)$	$(0.00,0.10)$
Algorithme 2 $(5.7,-0.69)$	$(0.00,0.12)$	$(0.15,0.20)$	$(0.00,0.22)$
Algorithme 2 $(5.7,-2.30)$	$(0.05,0.20)$	$(0.10,0.15)$	$(0.05,0.05)$
Algorithme 3 $(5.7,-0.11)$	$(0.05,0.04)$	$(0.00,0.05)$	$(0.00,0.09)$
Algorithme 3 $(5.7,-0.29)$	$(0.10,0.11)$	$(0.05,0.14)$	$(0.05,0.12)$
Algorithme 3 $(5.7,-0.69)$	$(0.00,0.02)$	$(0.10,0.08)$	$(0.05,0.00)$
Algorithme 3 $(5.7,-2.30)$	$(0.05,0.10)$	$(0.09,0.05)$	$(0.00,0.00)$

TABLE 4.10 – Différences en valeur absolue entre l'estimation en données complètes et les différentes stratégies.

	Différence entre les estimations sur W_Y et les estimations en données complètes		
Paramètres réels	$[0,1]\times[0,1]$	$[0,0.8]\times[0,1]$	$[0,0.6]\times[0,1]$
$(5.7,-0.11)$	$(0.00,0.02)$	$(0.05,0.04)$	$(0.05,0.08)$
$(5.7,-0.29)$	$(0.10,0.12)$	$(0.00,0.10)$	$(0.10,0.12)$
$(5.7,-0.69)$	$(0.02,0.02)$	$(0.08,0.04)$	$(0.04,0.08)$
$(5.7,-2.30)$	$(0.03,0.15)$	$(0.09,0.10)$	$(0.01,0.10)$

TABLE 4.11 – Différences en valeur absolue entre l'estimation en données complètes et l'estimation menée sur W_Y.

Nous constatons que les erreurs obtenues dans le Tableau 4.11 ne semblent pas exhiber de tendance ni avec la taille de la zone non observée ni avec la force de répulsion. Une première conclusion d'après ces résultats est que les erreurs trouvées ici semblent liées à la configuration en elle-même et d'à quel point les statistiques suffisantes restent proches de la moyenne des statistiques du modèle.
Un exemple pouvant illustrer ce phénomène est le suivant : pour une configuration du modèle de Poisson de paramètre $\rho = 200$ sur $[0,1]\times[0,1]$, nous attendons en moyenne 200 points. Si, pour cette configuration, 150 points sont dans $[0.5,1]\times[0,1]$, alors l'inférence basée sur la fenêtre $[0,0.5]\times[0,1]$ induira un biais par rapport à l'observation complète. À l'inverse, si les points sont bien répartis dans ces deux zones, l'inférence basée sur $[0,0.5]\times[0,1]$ sera très proche de celle basée sur la fenêtre complète.

En revanche, il paraît plus probant de comparer toutes ces données entre elles, entre stratégies et entre cas d'observations. Comparons d'abord les erreurs obtenues pour les différentes stratégies et celles obtenues avec l'estimation basée sur W_Y seulement.

— Dans le cas où nous avons utilisé l'Algorithme 3, 9 des 12 cellules du tableau 4.10 ont des

erreurs plus faibles que celles du tableau 4.11.
— Dans le cas des Algorithmes 1 et 2, seules 3 des 12 cellules du tableau ont des erreurs plus faibles ou égales à celles du tableau 4.11.

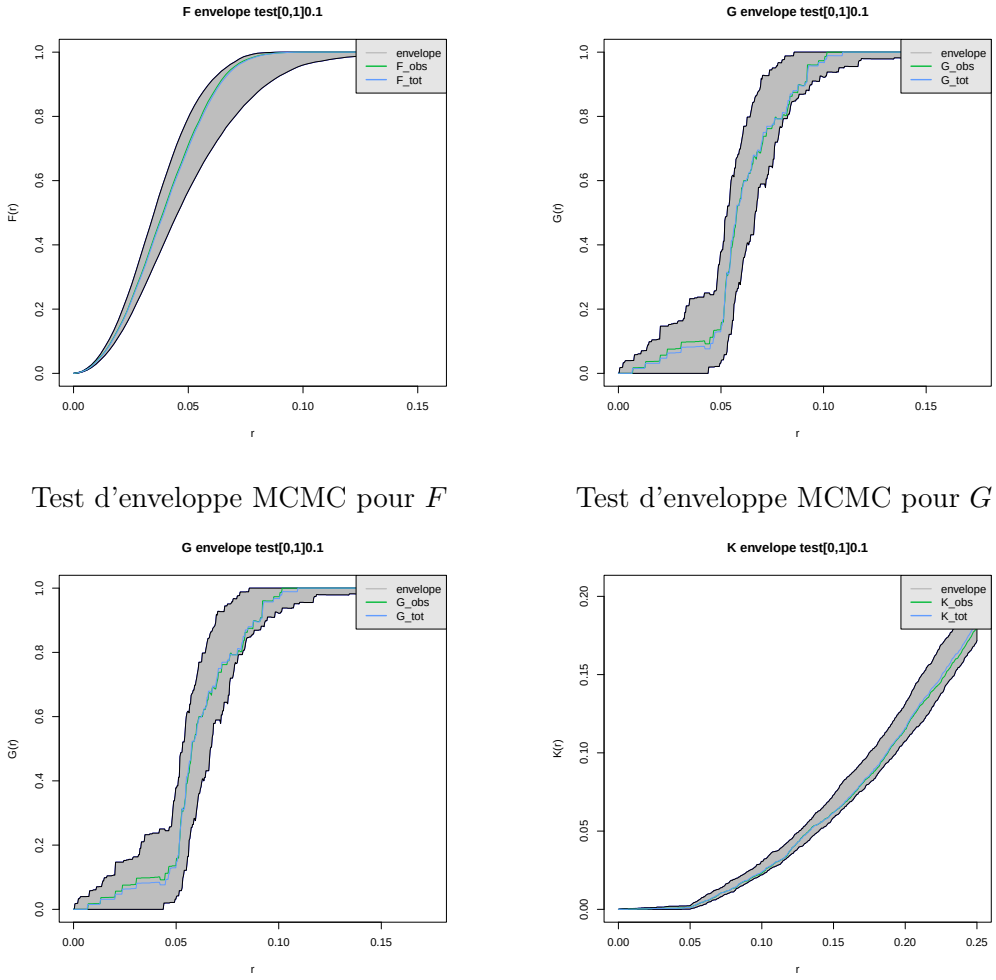
Ces résultats peuvent s'expliquer par le choix de la stratégie employée au sein des algorithmes. Pour les stratégies des algorithmes 1 et 2, nous avons extrait aléatoirement 100 valeurs de θ_Y dans l'échantillon de $(\theta_Y|\mathbf{y})$. L'inférence dépend alors grandement du choix de ces valeurs. Les résultats obtenus en utilisant l'algorithme 3 suggèrent que ces choix «aléatoires» pour θ_Y dans les algorithmes 1 et 2 devraient être fait proches du mode.

Des résultats similaires s'obtiennent en faisant la somme totale des erreurs obtenues : seule l'algorithme 3 possède des erreurs plus faibles pour les couples de paramètres (e.g. 0.29 pour la ligne «Algorithme 3 $(5.7, -2.30)$» du premier tableau et 0.48 pour cette même ligne du deuxième).

4.4.4 Contrôle de l'inférence

Nous souhaitons désormais savoir si les estimations trouvées permettent de resimuler des configurations préservant la structure spatiale observée sur W_Y. Pour cela, nous utilisons à nouveau des tests d'enveloppes MCMC. Plus précisément, nous simulons des réalisations du processus de Strauss avec les paramètres estimés indiqués dans le tableau 4.9 sur $W = [0, 1.2] \times [0, 1]$. Pour chacune de ces réalisations, nous la restreignons à la fenêtre observée W_Y pour chaque couple de paramètre et chaque cas de censure. L'hypothèse H_0 supposant que le modèle observé est une réalisation du processus de Strauss avec les paramètres estimés sera donc rejetée si la statistique observée n'appartient pas à l'enveloppe créée par les statistiques simulées.

La figure 4.16 illustre les tests d'enveloppe des fonctions F, G, g et K pour $\ln(\gamma_s) = -2.30$ et $W_Y = [0, 1] \times [0, 1]$. Les lignes vertes et bleues correspondent respectivement aux estimations pour la configuration observée et la configuration complète, pour l'estimation effectuée via l'algorithme 3. À chaque fois, les modèles observés et réels pour les estimateurs F, G, g et K se trouvent à l'intérieur des enveloppes, ce qui indique une estimation plutôt correcte. Les résultats obtenus sont similaires pour les deux autres méthodes utilisées et pour chaque couple de paramètres considéré. Les simulations ont été réalisées à l'aide de la bibliothèque `C++ DRLib` (https ://gitlab.univ-lorraine.fr/labos/iecl/drlib) et les estimations pour ces fonctions ont été obtenues à l'aide de la bibliothèque `R spatstat`[Baddeley et al., 2015].



Test d'enveloppe MCMC pour F | Test d'enveloppe MCMC pour G

Test d'enveloppe MCMC pour g | Test d'enveloppe MCMC pour K

FIGURE 4.16 – Test d'enveloppe pour les fonctions F, G, g et K (500 simulations) pour la configuration observée (ligne verte) et la configuration entièrement observée (ligne bleue) - les enveloppes Monte Carlo (zone grise).

La figure 4.17 présente les box-plots pour les statistiques suffisantes du modèle obtenues à partir de 500 simulations Monte Carlo, en effectuant deux «tests d'enveloppe» basés respectivement sur le nombre de points et le nombre de voisins r. À nouveau, les simulations ont été restreintes à W_Y afin d'être comparées à l'observation $\mathbf{y}$. La ligne verte représente les statistiques suffisantes observées. Pour chaque scénario d'estimation, les statistiques observées se trouvent à l'intérieur des enveloppes Monte Carlo représentées par les boîtes à moustaches, ce qui indique que les statistiques des configurations simulées englobent les statistiques observées du modèle. Les figures à gauche représentent les box-plots pour le nombre de points et les figures à droite représentent les box-plots pour le nombre de $r-$voisins.

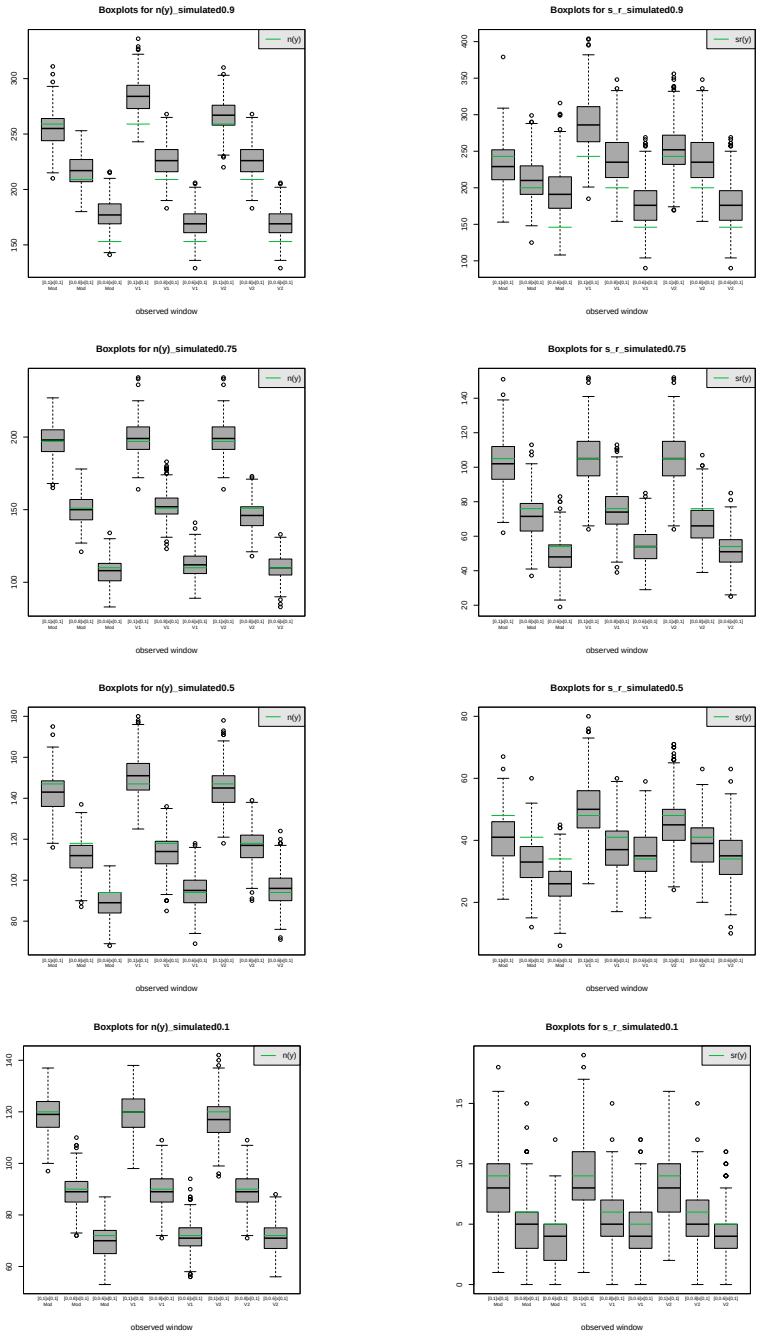


FIGURE 4.17 – «Tests d'enveloppe» MCMC pour les statistiques suffisantes du modèle (500 simulations) pour le nombre de points (gauche) et le nombre de $r-$voisins (droite). La ligne verte représente les statistiques observées sur W_Y pour chaque cas. Pour chaque figure, les trois premiers box-plots en partant de la gauche sont obtenus avec l'algorithme 3 lorsque $W_Y = [0,1] \times [0,1]$, $W_Y = [0,0.8] \times [0,1]$ et $W_Y = [0,0.6] \times [0,1]$ respectivement. Les trois suivants sont obtenus avec l'algorithme 2 lorsque $W_Y = [0,1] \times [0,1]$, $W_Y = [0,0.8] \times [0,1]$ et $W_Y = [0,0.6] \times [0,1]$ respectivement. Enfin, les trois derniers représentent les mêmes fenêtres observées et l'algorithme 1.

Chapitre 5

Conclusions et perspectives

Les travaux de cette thèse ont abordé trois problèmes principaux :

- La modélisation multi-interaction pour traiter des jeux de données exhibant plusieurs structures galactiques (filaments cosmiques).
- La modélisation multi-échelle pour prendre en compte le fait que les galaxies sont distribuées comme les perles d'un collier le long des filaments et des jeux de données volumineux.
- Le problème d'inférence en données incomplètes, abordée en utilisant le contexte Bayésien.

L'objectif de cette thèse est de proposer des travaux méthodologiques pour l'utilisation des processus ponctuels de Gibbs appliqués à des données cosmologiques. Pour cela, nous avons présenté trois modélisations et proposé une extension de l'inférence en données incomplètes au cadre Bayésien.

Le premier modèle introduit en 3.1 repose sur une inhomogénéité induite par les structures galactiques (ici, les filaments) et une superposition de modèles à interactions sur la même plage de voisinage. Le modèle semble s'ajuster de manière convenable aux données, les tests d'enveloppe MCMC indiquent des comportements similaires en termes de distances inter-points. La méthodologie présentée à travers la création du modèle et de la brève étude descriptive des données semble prometteuse et pose un premier jalon pour se diriger vers un approfondissement de l'étude de ce jeu de données. La méthode d'échantillonnage *a posteriori* semble également robuste en présence de modèles mettant en jeu de nombreuses interactions entre les points et en présence d'inhomogénéité.

Néanmoins, les caractéristiques du deuxième ordre (la fonction de corrélation par paire et la fonction K) ne sont pas parfaitement respectées dans le test d'enveloppe. Les estimations indiquent également que le paramètre pour le modèle de Strauss est, à l'exception du cas $(r_s, r_a) = (0.01, 0.01)$, très proche de 0, indiquant que le modèle n'est pas parfaitement adapté à ces données. Il semble que l'inconvénient de considérer une superposition de modèles est que deux interactions potentiellement opposées (lorsque le paramètre d'Area-Interaction correspond à de l'agglomération) ont une influence sur le même «territoire».

Ces premiers travaux invitent à utiliser une méthodologie similaire pour la partie non traitée du jeu de données en Figure 3.1. Une première continuation possible de ces travaux serait de raffiner l'analyse exploratoire pour identifier des meilleurs rayons d'interactions pour les modèles

considérés. Une telle analyse a été le point de départ du deuxième modèle proposé dans cette thèse.

Pour finir, l'étude menée dans cette section se base sur la détection des filaments galactiques dans des données. Cette étude pourrait être étendue à toutes les structures que les astrophysiciens parviennent à détecter (e.g des murs, des clusters, des super-clusters [Einasto, 2025]). Ainsi, ce type modélisation pourrait être enrichi par des mécanismes de conditionnement adaptés à la structure considérée.

L'enrichissement des travaux autour de la modélisation par les processus de Gibbs s'est aussi fait à travers les deux modélisations proposées pour le jeu de données présenté dans la section 3.2. Cette partie présente l'analyse exploratoire, la modélisation et l'inférence pour une étude locale de ce grand jeu de données. Ce travail est une continuation des études effectuées par [Hurtado-Gil et al., 2021] et de la modélisation présentée en 3.1. La superposition des modèles de Strauss et d'Area-Interaction introduite dans ces travaux n'existe que dans une seule zone et peut constituer, comme mentionné plus haut, une forme de compétition entre les modèles. Les nouveaux modèles proposent quant à eux des interactions multiples avec attraction autour des points et répulsion dans un voisinage en couronne autour de ces derniers, ce qui permet de cibler les interactions dans des zones plus précises. Pour cela, nous avons introduit deux nouveaux modèles, les modèles StraussCrown et GeyerCrown, appliqués pour la première fois à des données cosmologiques.

Les modèles s'ajustent plutôt bien aux données en se basant sur un simple test d'enveloppe et les résultats indiquent qu'il est important de permettre différents types d'interaction à différentes échelles. Les tests d'enveloppe des fonctions g indiquent que le comportement à petite échelle est meilleur lorsque nous utilisons le modèle Area-Interaction. Les «erreurs» observées à petite échelle peuvent être expliquées par le fait que la densité de galaxies au sein de clusters est très élevée, ce qui avait mené à des difficultés lors de la modélisation pour l'étude menée dans [Hurtado-Gil et al., 2021]. Comme perspective, il serait alors de considérer un modèle Area-Interaction GeyerCrown, où, comme pour la première modélisation, le modèle Area-Interaction est utilisé entre 0 et r_1. Les composantes de StraussCrown seraient alors remplacées par des composantes GeyerCrown. De plus, l'avantage d'utiliser le modèle GeyerCrown est que les phénomènes de bords liés à la définition des modèles (e.g. pour le modèle de Strauss) sont supprimés puisque le modèle de saturation de Geyer est défini sur $\mathbb{R}$ tout entier.

Nous avons ensuite orienté les travaux vers l'inférence avec des données partiellement observées, ce qui est très fréquent pour des données spatialisées et dans notre cas, en cosmologie. Pour aborder l'inférence avec des données incomplètes, nous avons proposé deux distributions dont la marginale est la distribution *a posteriori* d'intérêt. Les solutions pour échantillonner ces distributions ont été motivées par des résultats théoriques. À partir de ces résultats, nous avons dérivé des algorithmes plus simples à utiliser en termes d'efficacité numérique. Ces stratégies ont ensuite été testées dans le cadre d'une étude par simulation, en utilisant les réalisations du modèle de Strauss avec 4 paramètres différents et 3 cas pour la fenêtre observée W_Y. Les résultats indiquent une estimation des paramètres satisfaisante, tout en permettant de fournir une moyenne pour les statistiques suffisantes sur la région manquante. Des tests d'enveloppe MCMC utilisant les fonction F, G, g et K ainsi que les statistiques du modèle ont été utilisés pour contrôler les résultats de l'inférence, indiquant toujours une inférence cohérente.

Les algorithmes permettent donc d'obtenir de plutôt bons résultats. Les algorithmes 1 et 2, basés sur des valeurs aléatoires de l'échantillon $f(\theta_Y|\mathbf{y})$, n'ont été testées que pour 100 paramètres différents, ce qui, dans un échantillon de taille 10^6 semble trop faible. Les résultats obtenus pour l'Algorithme 3 suggèrent une étude plus approfondie en prenant ces valeurs autour du mode de l'échantillon de $f(\theta_Y|\mathbf{y})$. Cette approche semble être une piste intéressante pour la suite de ces travaux : le choix de multiples valeurs de l'échantillon de $f(\theta_Y|\mathbf{y})$ semble adéquat pour approcher la vraisemblance des données incomplètes. En effet, l'intégrale présente au numérateur de cette dernière indique que cette vraisemblance doit prendre en compte les différentes configurations possibles sur W_X, ce que nous tentons de traduire à travers l'approximation de $\mathbb{E}_{\widehat{\theta_Y}}[t(X)|Y = \mathbf{y}]$. Cette même intégrale est indirectement approchée à l'aide des simulations conditionnelles dans la solution proposée par [Gelfand and Carlin, 1993] à travers le ratio de vraisemblance, ce qui a également motivé notre proposition de loi *a posteriori*.

Ces travaux constituent une première contribution à l'élargissement de l'inférence en données incomplètes au cadre Bayésien. Nous nous appuyons fortement sur l'algorithme ABC Shadow, qui permet d'obtenir un échantillon de paramètres candidats pour la simulation conditionnelle sur W_X sachant la configuration $\mathbf{y}$. De plus, nous obtenons un échantillon pour cette loi, ce qui constitue une perspective intéressante quant à l'étude de son comportement asymptotique et des éventuelles applications qui découleraient de ce comportement (variance asymptotique, variance asymptotique MCMC, ...)

Des pistes intéressantes d'approfondissement découlent de ces travaux. Nous mentionnons d'abord l'enrichissement du contrôle de l'estimation (analyse résiduelle, tests d'enveloppes globaux). De plus, la validation et le choix de modèles sont une piste d'enrichissement prometteuse concernant les modélisations proposées. Les modélisations pourraient également être étendues aux processus ponctuels marqués pour prendre en compte les caractéristiques des galaxies disponibles déjà dans les catalogues. Bien évidement, il faudrait étendre ce cadre de travail aux données 3D, pour pouvoir s'attaquer réellement à la modélisation des grands catalogues des données massives disponibles et à venir. Enfin, aboutir à une modélisation prenant en compte les structures galactiques et des modèles multi-échelles sur des données réelles et utilisée dans le cadre des données incomplètes serait l'autre aboutissement remarquable de ces travaux. A ce propos, l'article [Gabriel et al., 2023] peut être intéressant pour l'estimation de l'influence de ces structures galactiques, latentes, dans les données cosmologiques dans les régions comportant des données manquantes. Ceci dit, les hypothèses d'indépendance utilisées par les auteurs apparaissent comme une limitation quant à l'application immédiate telle quelle de leur approche.

Comme perspectives plus immédiates, nous souhaitons poursuivre les travaux sur les données incomplètes en les appliquant aux données simulées par les astrophysiciens. Ces données issues d'une projection sur la sphère, sont proches du cadre que nous avons posé où deux régions disjointes peuvent être considérées pour les régions observées et non observées. Ces travaux sont en cours et feront l'objet d'une publication dans un journal. Plus généralement, des applications aux données spatio-temporelles pour pouvoir effectuer de la prédiction peuvent être envisagées. Enfin, transférer les méthodes proposées pour les données cosmologiques à d'autres domaines peut constituer un sujet de recherche à part entière.

Bibliographie

[Alquier et al., 2014] Alquier, P., Friel, N., Everitt, R., and Boland, A. (2014). Noisy monte carlo : Convergence of markov chains with approximate transition kernels.

[Ambler and Silverman, 2009] Ambler, G. K. and Silverman, B. W. (2009). Perfect simulation of spatial point processes using dominated coupling from the past with application to a multiscale area-interaction point process. *arXiv e-prints*, page arXiv :0903.2651.

[Atchade et al., 2014] Atchade, Y., Fort, G., and Moulines, E. (2014). On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*.

[Atchadé et al., 2017] Atchadé, Y., Fort, G., and Moulines, É. (2017). On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10) :1–33.

[Atchadé et al., 2013] Atchadé, Y. F., Lartillot, N., and Robert, C. (2013). Bayesian computation for statistical models with intractable normalizing constants. *Brazilian Journal of Probability and Statistics*, 27(4).

[Ba et al., 2023] Ba, I., Coeurjolly, J.-F., and Cuevas-Pacheco, F. (2023). Pairwise interaction function estimation of stationary Gibbs point processes using basis expansion. *The Annals of Statistics*, 51(3) :1134 – 1158.

[Baddeley et al., 2013a] Baddeley, A., Coeurjolly, J.-F., Rubak, E., and Waagepetersen, R. (2013a). Variational estimators for the parameters of gibbs point process models. *Bernoulli*, 19(4) :1761–1798.

[Baddeley et al., 2015] Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns : Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press.

[Baddeley et al., 2013b] Baddeley, A., Turner, R., Mateu, J., and Bevan, A. (2013b). Hybrids of gibbs point process models and their implementation. *Journal of Statistical Software*, 55(11) :1–43.

[Baddeley and Van Lieshout, 1995] Baddeley, A. and Van Lieshout, M. (1995). Area-interaction point processes. *Annals of the Institute of Statistical Mathematics*, 47 :601–619.

[Beaumont et al., 2009] Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate bayesian computation. *Biometrika*, 96(4) :983–990.

[Bedford and Berg, 1997] Bedford, T. and Berg, J. V. D. (1997). A remark on the van lieshout and baddeley j-function for point processes. *Advances in Applied Probability*, 29(1) :19–25.

[Berman and Turner, 1992] Berman, M. and Turner, T. (1992). Approximating point process likelihoods with glim. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1) :31–38.

[Besag, 1974] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society : Series B (Methodological)*, 36(2) :192–225.

[Besag, 1978] Besag, J. E. (1978). Some methods of statistical analysis for spatial data (with discussion). *Bulletin of the International Statistical Institute*, 47 :77–92.

[Billiot et al., 2008] Billiot, J.-M., Coeurjolly, J.-F., and Drouilhet, R. (2008). Maximum pseudo-likelihood estimator for exponential family models of marked gibbs point processes. *Electronic Journal of Statistics*, 2 :234–264.

[Blei et al., 2017] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference : A review for statisticians. *Journal of the American Statistical Association*, 112(518) :859–877.

[Caimo and Friel, 2011] Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1) :41–55.

[Celeux et al., 1995] Celeux, G., Chauveau, D., and Diebolt, J. (1995). On stochastic versions of the em algorithm. Research Report 2514, INRIA.

[Chafai and Malrieu, 2018] Chafai, D. and Malrieu, F. (2018). Recueil de modèles aléatoires.

[Coeurjolly et al., 2012] Coeurjolly, J.-F., Dereudre, D., Drouilhet, R., and Lavancier, F. (2012). Takacs–fiksel method for stationary marked gibbs point processes. *Scandinavian Journal of Statistics*, 39(3) :416–443.

[Coeurjolly and Drouilhet, 2010] Coeurjolly, J.-F. and Drouilhet, R. (2010). Asymptotic properties of the maximum pseudo-likelihood estimator for stationary gibbs point processes including the lennard-jones model. *arXiv preprint arXiv :1007.1894*.

[Coeurjolly et al., 2016] Coeurjolly, J.-F., Guan, Y., Khanmohammadi, S., and Waagepetersen, R. P. (2016). Towards optimal takacs–fiksel estimation. *Spatial Statistics*, 18 :396–411.

[Coeurjolly and Lavancier, 2017] Coeurjolly, J.-F. and Lavancier, F. (2017). Parametric estimation of pairwise gibbs point processes with infinite range interaction. *Bernoulli*, 23(2) :953–986.

[Coles and Jones, 1991] Coles, P. and Jones, B. (1991). A lognormal model for the cosmological mass distribution. *Monthly Notices of the Royal Astronomical Society*, 248 :1–13.

[Cox, 1955] Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2) :129–164.

[Cronie and Särkkä, 2011] Cronie, O. and Särkkä, A. (2011). Some edge correction methods for marked spatio-temporal point process models. *Computational Statistics & Data Analysis*, 55(7) :2209–2220.

[Daley and Vere-Jones, 2003] Daley, D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Volume 1 : Elementary Theory and Methods, Second Edition*. Springer.

[Daley and Vere-Jones, 2008] Daley, D. and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes, Volume 2 : General Theory and Structure, Second Edition*. Springer.

[Daudel, 2021] Daudel, K. (2021). *Méthodes de Monte-Carlo adaptatives pour les modèles complexes*. Thèse de doctorat, Institut Polytechnique de Paris (Télécom Paris), Palaiseau, France. Dirigée par François Roueff et Randal Douc.

[Delyon et al., 1999] Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1) :94 – 128.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) :1–22.

[Dereudre and Lavancier, 2016] Dereudre, D. and Lavancier, F. (2016). Consistency of likelihood estimation for Gibbs point processes. *Annals of Statistics*.

[Descombes et al., 1999] Descombes, X., Morris, R. D., Zerubia, J., and Berthod, M. (1999). Estimation of markov random field prior parameters using markov chain monte carlo maximum likelihood. *IEEE Transactions on Image Processing*, 8 :954–963.

[Diggle, 1985] Diggle, P. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(2) :138–147.

[Diggle, 1983] Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. Academic Press, London.

[Diggle et al., 1994] Diggle, P. J., Fiksel, T., Grabarnik, P., Ogata, Y., Stoyan, D., and Tanemura, M. (1994). On parameter estimation for pairwise interaction point processes. *International Statistical Review*, 62(1) :99–117.

[Diggle and Gratton, 1984] Diggle, P. J. and Gratton, R. J. (1984). Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2) :193–227.

[Doroshkevich and Shandarin, 1978] Doroshkevich, A. G. and Shandarin, S. F. (1978). A statistical approach to the theory of galaxy formation. *Soviet Astronomy*, 22 :653–660.

[D'Angelo et al., 2024] D'Angelo, G., Adelfio, G., Mateu, J., and Cronie, O. (2024). Semi-parametric profile pseudolikelihood via local summary statistics for spatial point pattern intensity estimation. *arXiv preprint arXiv :2404.10344*.

[Einasto, 2025] Einasto, M. (2025). Galaxy superclusters and their complexes in the cosmic web. *Universe*, 11(6) :167.

[Fearnhead and Prangle, 2012] Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate bayesian computation : semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 74(3) :419–474.

[Fiksel, 1984] Fiksel, T. (1984). Estimation of parameterized pair potentials of marked and non-marked gibbsian point processes. *Elektronische Informationsverarbeitung und Kybernetik*, 20(6) :270–278.

[Fiksel, 1988] Fiksel, T. (1988). Estimation of interaction potentials of gibbsian point processes. *Statistics*, 19(1) :77–86.

[Fort and Moulines, 2003] Fort, G. and Moulines, E. (2003). Convergence of the monte carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4) :1220–1259.

[Gabriel et al., 2023] Gabriel, E., Rodríguez-Cortés, F. J., Coville, J., Mateu, J., and Chadœuf, J. (2023). Mapping the intensity function of a non-stationary point process in unobserved areas. *Stochastic Environmental Research and Risk Assessment*, 37 :327–343. Also arXiv :2111.14403 [stat.ME].

[Gelfand and Carlin, 1993] Gelfand, A. E. and Carlin, B. P. (1993). Maximum-likelihood estimation for constrained-or missing-data models. *Canadian Journal of Statistics*, 21(3) :303–311.

[Geyer and Møller, 1994] Geyer, C. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, 21(4) :359–373.

[Geyer, 1994] Geyer, C. J. (1994). On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1) :261–274.

[Geyer, 1999] Geyer, C. J. (1999). Likelihood inference for spatial point processes. In Barndorff-Nielsen, O., Kendall, W., and van Lieshout, M., editors, *Stochastic Geometry, Likelihood and Computation*. CRC Press/Chapman and Hall, Boca Raton.

[Geyer and Møller, 1993] Geyer, C. J. and Møller, J. (1993). Simulation procedures and likelihood inference for spatial point processes. Technical Report 260, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, Aarhus, Denmark.

[Geyer and Thompson, 1992] Geyer, C. J. and Thompson, E. A. (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3) :657–699.

[Gillot et al., 2023] Gillot, N., Stoica, R. S., and Gemmerlé, D. (2023). Study the galaxy distribution characterisation via bayesian statistical learning of spatial marked point processes. In *Proceedings RING*. Preprint HAL, hal-04163649.

[Gillot et al., 2024] Gillot, N., Stoica, R. S., Särkkä, A., and Gemmerlé, D. (2024). Spatial point process modelling and Bayesian inference for large data sets. In *RING Meeting*, Nancy, France. École nationale supérieure de géologie (ENSG) Nancy.

[Goulard et al., 1996] Goulard, M., Grabarnik, P., and Särkkä, A. (1996). Parameter estimation for gibbs point processes. *Scandinavian Journal of Statistics*, 23(3) :365–379.

[Gourieroux and Monfort, 1989] Gourieroux, C. and Monfort, A. (1989). *Statistique et modèles économétriques / Christian Gourieroux,... Alain Monfort,...* Collection Économie et statistiques avancées Série École nationale de la statistique et de l'administration économique et Centre d'études des programmes économiques. Économica, Paris.

[Green, 1995] Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4) :711–732.

[Gregori et al., 2004] Gregori, P., van Lieshout, M., and Mateu, J. (2004). Mixture formulae for shot noise weighted point processes. *Statistics & Probability Letters*, 67(4) :311–320.

[Grelaud et al., 2009] Grelaud, A., Marin, J.-M., Robert, C. P., Rodolphe, F., and Taly, J.-F. (2009). Abc likelihood-free methods for model choice in gibbs random fields. *Bayesian Analysis*, 4(2).

[Gu and Li, 1998] Gu, M. G. and Li, S. (1998). A stochastic approximation algorithm for maximum-likelihood estimation with incomplete data. *The Canadian Journal of Statistics*, 26(4) :567–582.

[Gu and Zhu, 2002] Gu, M. G. and Zhu, H.-T. (2002). Maximum Likelihood Estimation for Spatial Models by Markov Chain Monte Carlo Stochastic Approximation. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 63(2) :339–355.

[Guttorp et al., 2023] Guttorp, P., Illian, J., Kostensalo, J., Kuronen, M., Myllymäki, M., Särkkä, A., and Thorarinsdottir, T. L. (2023). What you see is not what is there : Mechanisms, models, and methods for point pattern deviations.

[Hastings, 1970] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1) :97–109.

[Huang and Ogata, 1999] Huang, F. and Ogata, Y. (1999). Improvements of the maximum pseudo-likelihood estimators in various spatial statistical models. *Journal of Computational and Graphical Statistics*, 8(3) :510–530.

[Hurtado-Gil et al., 2021] Hurtado-Gil, L., Stoica, R. S., Martínez, V., and Arnalte-Mur, P. (2021). Morphostatistical characterization of the spatial galaxy distribution through Gibbs point processes. *Monthly Notices of the Royal Astronomical Society*, 507(2) :1710–1722.

[Icke and van de Weygaert, 1987] Icke, V. and van de Weygaert, R. (1987). Fragmenting the universe. *Astronomy and Astrophysics*, 184 :16–32.

[Icke and van de Weygaert, 1991] Icke, V. and van de Weygaert, R. (1991). The galaxy distribution as a Voronoi foam. *Royal Astronomical Society*, 32 :85–112.

[Illian et al., 2008] Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Spatial Analysis and Modelling of Spatial Point Patterns*. Wiley-Interscience.

[Jahnel, 2024] Jahnel, B. (2024). The variational principle for a marked gibbs point process. *arXiv preprint arXiv :2406.10914*.

[Jansson and Cronie, 2024] Jansson, A. and Cronie, O. (2024). Comparison of point process learning and its special case takacs-fiksel estimation. *arXiv preprint arXiv :2405.19523*.

[Jensen and Kunsch, 1994] Jensen, J. L. and Kunsch, H. R. (1994). On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Annals of the Institute of Statistical Mathematics*, 46 :475–486.

[Keerin and Boongoen, 2022] Keerin, P. and Boongoen, T. (2022). Estimation of missing values in astronomical survey data : An improved local approach using cluster directed neighbor selection. *Information Processing & Management*, 59(2) :102881.

[Kelly and Ripley, 1976] Kelly, F. P. and Ripley, B. D. (1976). A note on strauss's model for clustering. *Biometrika*, 63(2) :357–360.

[Kendall and Møller, 2000] Kendall, W. S. and Møller, J. (2000). Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability*, 32(3) :844–865.

[Kuhn and Lavielle, 2004] Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM : Probability and Statistics*, 8 :115–131.

[Liang, 2010] Liang, F. (2010). A double metropolis–hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80 :1007–1022.

[Liang and Jin, 2013] Liang, F. and Jin, I. H. (2013). A monte carlo metropolis-hastings algorithm for sampling from distributions with intractable normalizing constants. *Neural computation*, 25.

[Liang et al., 2016] Liang, F., Jin, I. H., Song, Q., and Liu, J. S. (2016). Adaptive exchange algorithm for sampling from distributions with intractable normalizing constants. *Journal of the American Statistical Association*, 111 :377–393.

[van Lieshout and Stoica, 2006] van Lieshout, M. N. M. and Stoica, R. (2006). Perfect simulation for marked point processes. *Computational Statistics & Data Analysis*, 51 :679–698.

[van Lieshout and Stoica, 2003] van Lieshout, M. N. M. and Stoica, R. S. (2003). The Candy model : properties and inference. *Statistica Neerlandica*, 57(2) :177–206.

[Lu and Friel, 2024] Lu, C. and Friel, N. (2024). Bayesian strategies for repulsive spatial point processes.

[Lund and Rudemo, 2000] Lund, J. and Rudemo, M. (2000). Models for point processes observed with noise. *Biometrika*, 87(2) :235–249.

[Marin et al., 2011] Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2011). Approximate bayesian computational methods. *Statistics and Computing*, 22(6) :1167–1180.

[Martinez and Saar, 2002] Martinez, V. and Saar, E. (2002). Clustering statistics in cosmology. In *Astronomical Data Analysis II*, volume 4847, page 86. SPIE.

[Martinez and Saar, 2001] Martinez, V. J. and Saar, E. (2001). *Statistics of the galaxy distribution*. Chapman & Hall/CRC.

[Mase, 2000] Mase, S. (2000). Marked gibbs processes and asymptotic normality of maximum pseudo-likelihood estimators. *Mathematische Nachrichten*, 209(1) :151–169.

[Matérn, 1966] Matérn, B. (1966). *Spatial Variation ; Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations*. Stockholm. Statens Skogsforskningsinstitut. Meddelanden. University of Sweden.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6) :1087–1092.

[Meyn and Tweedie, 1993] Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*, volume 92. Springer-Verlag.

[Miao et al., 2023] Miao, Z., Chen, Y.-C., and Dobra, A. (2023). Bayesian finite mixtures of ising models.

[Moka et al., 2021] Moka, S. B., Juneja, S., and Mandjes, M. R. H. (2021). Rejection and importance sampling based perfect simulation for gibbs hard-sphere models.

[Monfort, 1997] Monfort, A. (1997). *Cours de statistique mathématique*. Economie et statistiques avancées. Economica.

[Moyeed and Baddeley, 1991] Moyeed, R. A. and Baddeley, A. J. (1991). Stochastic approximation of the mle for a spatial point pattern. *Scandinavian Journal of Statistics*, 18(1) :35–46.

[Murray et al., 2006] Murray, I., Ghahramani, Z., and MacKay, D. (2006). Mcmc for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press.

[Myllymäki et al., 2016] Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2016). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 79(2) :381–404.

[Møller and Cuevas-Pacheco, 2018] Møller, J. and Cuevas-Pacheco, F. (2018). Log gaussian cox processes on the sphere.

[Møller and Helisová, 2010] Møller, J. and Helisová, K. (2010). Likelihood inference for unions of interacting discs. *Scandinavian Journal of Statistics*, 37(3) :365–381.

[Møller and Jensen, 1991] Møller, J. and Jensen, J. L. (1991). Pseudo-likelihood for exponential family models of spatial point processes. *Scandinavian Journal of Statistics*, 18(1) :67–84.

[Møller et al., 2006] Møller, J., Pettitt, A., Reeves, R., and Berthelsen, K. (2006). An efficient mcmc method for distributions with intractable normalising constants. *Biometrika*, 93.

[Møller et al., 1998] Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3) :451–482.

[Møller and Waagepetersen, 2004] Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC.

[N. Friel and Wit, 2009] N. Friel, A. N. Pettitt, R. R. and Wit, E. (2009). Bayesian inference in hidden markov random fields for binary data defined on large lattices. *Journal of Computational and Graphical Statistics*, 18(2) :243–261.

[Neath, 2013] Neath, R. C. (2013). On convergence properties of the monte carlo em algorithm. In Jones, G. and Shen, X., editors, *Advances in Modern Statistical Theory and Applications : A Festschrift in Honor of Morris L. Eaton*, volume 10, pages 43–62. Institute of Mathematical Statistics. IMS Collections.

[Neyman and Scott, 1958] Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society : Series B (Methodological)*, 20(1) :1–29.

[Ogata and Tanemura, 1981] Ogata, Y. and Tanemura, M. (1981). Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Annals of the Institute of Statistical Mathematics*, 33 :315–338.

[Ogata and Tanemura, 1984] Ogata, Y. and Tanemura, M. (1984). Likelihood analysis of spatial point patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3) :496–518.

[Ogata and Tanemura, 1985] Ogata, Y. and Tanemura, M. (1985). Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method. *Biometrics*, 41 :421–433.

[Ogata and Tanemura, 1989] Ogata, Y. and Tanemura, M. (1989). Likelihood estimation of soft-core interaction potentials for gibbsian point patterns. *Annals of the Institute of Statistical Mathematics*, 41(3) :583–600.

[Pandey, 2010] Pandey, B. (2010). Statistically significant length-scale of filaments as a robust measure of galaxy distribution. *Monthly Notices of the Royal Astronomical Society*, 401(4) :2687–2696.

[Paranjape and Alam, 2020] Paranjape, A. and Alam, S. (2020). Voronoi volume function : a new probe of cosmology and galaxy evolution. *Monthly Notices of the Royal Astronomical Society*, 495(3) :3233–3251.

[Peebles, 1973] Peebles, P. J. E. (1973). Statistical Analysis of Catalogs of Extragalactic Objects. I. Theory. *Astrophysical Journal*, 185 :413–440.

[Penttinen, 1984] Penttinen, A. (1984). Modelling interaction in spatial point patterns : parameter estimation by the maximum likelihood method. Technical report, Jyväskylä Studies in Computer Science, Economics and Statistics (Thesis / Report). Jyväskylä Univ.

[Perkins et al., 2017] Perkins, N. J., Cole, S. R., Harel, O., Tchetgen Tchetgen, E. J., Sun, B., Mitchell, E. M., and Schisterman, E. F. (2017). Principled Approaches to Missing Data in Epidemiologic Studies. *American Journal of Epidemiology*, 187(3) :568–575.

[Pons-Bordería et al., 1999] Pons-Bordería, M.-J., Martínez, V. J., Stoyan, D., Stoyan, H., and Saar, E. (1999). Comparing estimators of the galaxy correlation function. *The Astrophysical Journal*, 523(2) :480.

[Preston, 1975] Preston, C. (1975). Spatial birth and death processes. *Advances in Applied Probability*, 7 :465 – 466.

[Propp and Wilson, 1996] Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Struct. Algorithms*, 9(1–2) :223–252.

[Raynal et al., 2018] Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2018). ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10) :1720–1728.

[Reype, 2022] Reype, C. (2022). *Modélisation probabiliste et inférence bayésienne pour l'analyse de la dynamique des mélanges de fluides géologiques : détection des structures et estimation des paramètres*. Theses, Université de Lorraine.

[Ripley, 1976] Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2) :255–266.

[Ripley, 1977] Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(2) :172–212.

[Ripley, 1981] Ripley, B. D. (1981). *Spatial Statistics*. John Wiley & Sons, Chichester.

[Ripley, 1988] Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press.

[Robert and Casella, 2000] Robert, C. and Casella, G. (2000). Monte carlo statistical method. *Technometrics*, 42.

[Ruth, 2024] Ruth, W. (2024). A review of monte carlo-based versions of the em algorithm. *arXiv preprint*, abs/2401.00945.

[Särkkä, 1995] Särkkä, A. (1995). Pseudo-likelihood approach for gibbs point processes in connection with field observations. *Statistics*, 26(1) :89–97.

[Shirota and Gelfand, 2017] Shirota, S. and Gelfand, A. E. (2017). Approximate bayesian computation and model assessment for repulsive spatial point processes. *Journal of Computational and Graphical Statistics*, 26(3) :646–657.

[Simkus et al., 2023] Simkus, V., Rhodes, B., and Gutmann, M. U. (2023). Variational gibbs inference for statistical model estimation from incomplete data.

[Snethlage et al., 2002] Snethlage, M., Martínez, V., Stoyan, D., and Saar, E. (2002). Point field models for the galaxy point pattern - modelling the singularity of the two-point correlation function. *Astronomy and Astrophysics*, 388(3) :758–765.

[Soares-Santos et al., 2010] Soares-Santos, M., de Carvalho, R. R., Annis, J., Gal, R. R., La Barbera, F., Lopes, P. A. A., Wechsler, R. H., Busha, M. T., and Gerke, B. F. (2010). The voronoi tessellation cluster finder in 2+1 dimensions. *The Astrophysical Journal*, 727(1) :45.

[Sorce et al., 2023] Sorce, J. G., Stoica, R. S., and Tempel, E. (2023). Processus ponctuel de Gibbs et inférence Bayésienne pour réduire les biais observationnels : cas des catalogues de vitesse de galaxies. In *GRETSI 2023 XXIXème Colloque Francophone de Traitement du Signal et des Images*, Grenoble, France.

[Stoica et al., 2021] Stoica, R., Deaconu, M., Philippe, A., and Hurtado-Gil, L. (2021). Shadow Simulated Annealing : A new algorithm for approximate Bayesian inference of Gibbs point processes. *Spatial Statistics*.

[Stoica et al., 2005a] Stoica, R., Gregori, P., and Mateu, J. (2005a). Simulated annealing and object point processes : Tools for analysis of spatial patterns. *Stochastic Processes and their Applications*, 115(11) :1860–1882.

[Stoica, 2001] Stoica, R. S. (2001). *Processus ponctuels pour l'Extraction de Réseaux Linéaires*. PhD thesis, Université de Nice Sophia-Antipolis, Nice, France.

[Stoica, 2010] Stoica, R. S. (2010). Marked point processes for statistical and morphological analysis of astronomical data. *European Physical Journal Special Topics*, 186(1) :123–165.

[Stoica, 2014] Stoica, R. S. (2014). *Modélisation probabiliste et inférence statistique pour l'analyse des données spatialisées*. Habilitation à Diriger des Recherches thesis - Université de Lille.

[Stoica, 2025] Stoica, R. S. (2025). *Random Patterns and Structures in Spatial Data*. Chapman and Hall.

[Stoica et al., 2002] Stoica, R. S., Descombes, X., van Lieshout, M. N. M., and Zerubia, J. (2002). An application of marked point processes to the extraction of linear networks from images. In Mateu, J. and Montes, F., editors, *Spatial Statistics through Applications*, pages 1–12. WIT Press, Southampton, UK.

[Stoica et al., 2004] Stoica, R. S., Descombes, X., and Zerubia, J. (2004). A gibbs point process for road extraction in remotely sensed images. *International Journal of Computer Vision*, 57 :121–136.

[Stoica et al., 2007] Stoica, R. S., Gay, E. E., and Kretzschmar, A. (2007). Cluster pattern detection in spatial data based on Monte Carlo inference. *Biometrical Journal*, 49(4) :505–519.

[Stoica et al., 2005b] Stoica, R. S., Martinez, V. J., Mateu, J., and Saar, E. (2005b). Detection of cosmic filaments using the Candy model. *Astronomy and Astrophysics - Astronomy and Astrophysics*, 434(2) :423–432.

[Stoica et al., 2017] Stoica, R. S., Philippe, A., Gregori, P., and Mateu, J. (2017). ABC Shadow algorithm : a tool for statistical analysis of spatial patterns. *Statistics and Computing*, 27(5) :1225–1238.

[Stoica et al., 2015] Stoica, R. S., Tempel, E., Liivamägi, L. J., Castellan, G., and Saar, E. (2015). Spatial patterns analysis in cosmology based on marked point processes. In Fraix-Burnet, D. and Valls-Gabaud, D., editors, *Statistics for astrophysics. Methods and applications of the regression*. European Astronomical Society Publication Series EDP Sciences.

[Stoica, R. S. et al., 2005] Stoica, R. S., Martínez, V. J., Mateu, J., and Saar, E. (2005). Detection of cosmic filaments using the candy model. *Astronomy and Astrophysics*, 434(2) :423–432.

[Stoyan and Stoyan, 1994] Stoyan, D. and Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields : Methods of Geometrical Statistics*. Wiley.

[Strauss, 1975] Strauss, D. J. (1975). A model for clustering. *Biometrika*, 62 :467–475.

[Särkkä, 1993] Särkkä, A. (1993). *Pseudo-likelihood approach for pair potential estimation of Gibbs processes*. PhD thesis, University of Jyväskylä, Jyväskylä, Finland.

[Takács, 1986] Takács, L. (1986). Estimator for the pair potential of a gibbsian point process. *Statistics*, 17(3) :367–370.

[Takács and Fiksel, 1986] Takács, R. and Fiksel, T. (1986). Interaction pair-potentials for a system of ant's nests. *Biometrical Journal*, 28(8) :1007–1013.

[Taty Moukati et al., 2024] Taty Moukati, F., Stoica, R., Bonneau, F., Wu, X., and Caumon, G. (2024). From fault likelihood to fault networks : Stochastic seismic interpretation through a marked point process with interactions. *Mathematical Geosciences*, 57.

[Tempel et al., 2014] Tempel, E., Kipper, R., Saar, E., Bussov, M., Hektor, A., and Pelt, J. (2014). Galaxy filaments as pearl necklaces ? *Astronomy and Astrophysics*, 572(A8) :1–8.

[Tempel et al., 2018] Tempel, E., Kruuse, M., Kipper, R., Tuvikene, T., Sorce, J. G., and Stoica, R. S. (2018). Bayesian group finder based on marked point processes. method and application to the 2mrs data set. *Astronomy and Astrophysics*, 618(A61) :1–18.

[Tempel et al., 2016] Tempel, E., Stoica, R. S., Kipper, R., and Saar, E. (2016). Bisous model - detecting filamentary patterns in point processes.

[Tempel et al., 2013] Tempel, E., Stoica, R. S., and Saar, E. (2013). Evidence for spin alignment of spiral and elliptical/S0 galaxies in filaments. *Monthly Notices of the Royal Astronomical Society*, 428 :1827–1836.

[Tierney, 1994] Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4) :1701–1728.

[Totsuji and Kihara, 1969] Totsuji, H. and Kihara, T. (1969). The Correlation Function for the Distribution of Galaxies. *Astronomical Society of Japan*, 21 :221.

[van Lieshout, 2000] van Lieshout, M. N. M. (2000). *Markov Point Processes and Their Applications*. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO.

[van Lieshout, 2019] van Lieshout, M. N. M. (2019). *Theory of Spatial Statistics : A concise Introduction*. Chapman & Hall.

[van Lieshout and Stoica, 2003] van Lieshout, M. N. M. and Stoica, R. S. (2003). The candy model revisited : properties and inference. *Statistica Neerlandica*, 57 :1–30.

[van Lieshout and van Zwet, 2000] van Lieshout, M. N. M. and van Zwet, E. W. (2000). Maximum likelihood estimation for the bombing model. Technical Report 0008, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, The Netherlands.

[van Lieshout and van Zwet, 2001] van Lieshout, M. N. M. and van Zwet, E. W. (2001). Exact sampling from conditional boolean models with applications to maximum likelihood inference. *Advances in Applied Probability*, 33(2) :339–353.

[Vihrs et al., 2020] Vihrs, N., Møller, J., and Gelfand, A. (2020). Approximate bayesian inference for a spatial point process model exhibiting regularity and random aggregation. *Scandinavian Journal of Statistics*, 49.

[Voronoi, 1908] Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les parallélloèdres primitifs. *Journal für die reine und angewandte Mathematik*, 134 :198–287.

[Wang et al., 2019] Wang, T., Schofield, M., Bebbington, M., and Kiyosugi, K. (2019). Bayesian Modelling of Marked Point Processes with Incomplete Records : Volcanic Eruptions. *Journal of the Royal Statistical Society Series C : Applied Statistics*, 69(1) :109–130.

[Wei and Tanner, 1990] Wei, G. C. G. and Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411) :699–704.

[Wijayawardhana et al., 2024] Wijayawardhana, A., Gunawan, D., and Suesse, T. (2024). Variational bayes inference for spatial error models with missing data.

[Winkler, 2003] Winkler, G. (2003). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer.

[Younes, 1988] Younes, L. (1988). Estimation and annealing for Gibbsian fields. *Annales de l'I.H.P. Probabilités et statistiques*, 24(2) :269–294.

[Younes, 1989] Younes, L. (1989). Parametric inference for imperfectly observed gibbsian fields. *Probability Theory and Related Fields*, 82 :625–645.

[Yuan and Wang, 2024] Yuan, W. and Wang, G. (2024). Markov chain monte carlo without evaluating the target : an auxiliary variable approach.

[Zaninetti, 2006] Zaninetti, L. (2006). On the large-scale structure of the universe as given by the voronoi diagrams. *Chinese Journal of Astronomy and Astrophysics*, 6(4) :387–395.

[Zaninetti, 2018] Zaninetti, L. (2018). Filaments of galaxies and voronoi diagrams. *Open Astronomy*, 27(1) :335–340.

[Zel'dovich, 1970] Zel'dovich, Y. B. (1970). Gravitational instability : An approximate theory for large density perturbations. *Astronomy and Astrophysics*, 5 :84–89.

Annexe A

Preuves

A.1 Preuve du théorème 2.3.1

Démonstration. **(i)** Les deux densités s'annulant en dehors de $b(\theta, \Delta/2)$, la propriété est vraie dans ce cas. Soit $\psi \in b(\theta, \Delta/2)$, le théorème de la moyenne appliqué au dénominateur de q_Δ donne l'existence d'un $\theta^* \in b(\theta, \Delta/2)$ tel que $I(\theta, \Delta, \mathbf{x}) = V_\Delta \frac{f(\mathbf{x}|\theta^*)}{c(\theta^*)}$. Comme nous avons supposé Θ compact de $\mathbb{R}^r$, la continuité de f donne l'existence d'un minimum $m(\mathbf{x}) := \inf_{\phi \in \Theta} f(\mathbf{x}|\phi)$ avec $m(\mathbf{x}) > 0$. Pour $A \in \mathcal{T}_\Theta$ il vient :

$$\begin{aligned}
\int_A |q_\Delta(\theta \to \psi - U_\Delta(\theta \to \psi)| d\psi &= \int_{A \cap b(\theta, \Delta/2)} \left| \frac{\frac{f(\mathbf{x}|\psi)}{c(\psi)}}{V_\Delta \frac{f(\mathbf{x}|\theta^*)}{c(\theta*)}} - \frac{1}{V\Delta} \right| d\psi \\
&\leq \frac{1}{V_\Delta} \sup_{\phi \in \Theta} \frac{c(\phi)}{f(\mathbf{x}|\phi)} \int_{A \cap b(\theta, \Delta/2)} \left| \frac{\frac{f(\mathbf{x}|\psi)}{c(\psi)}}{V_\Delta \frac{f(\mathbf{x}|\theta^*)}{c(\theta*)} - \frac{1}{V\Delta}} \right| d\psi \\
&\leq \frac{\mu(A \cap b(\theta, \Delta/2))}{V_\Delta} m(\mathbf{x})^{-1} \times \sup_{d(\psi, \theta^*) < \Delta} \left| \frac{f(\mathbf{x}|\psi)}{c(\psi)} - \frac{f(\mathbf{x}|\theta^*)}{c(\theta^*} \right| \\
&\leq m(\mathbf{x})^{-1} \sup_{d(\psi, \theta^*) < \Delta} \left| \frac{f(\mathbf{x}|\psi)}{c(\psi)} - \frac{f(\mathbf{x}|\theta^*)}{c(\theta^*)} \right|
\end{aligned}$$

où le supremum tend vers 0 lorsque Δ tend vers 0.
Si $f(\mathbf{x}|\cdot) \in \mathcal{C}^1(\Theta)$, nous obtenons, en utilisant l'inégalité des accroissements finis :

$$\begin{aligned}
\int_A |q_\Delta(\theta \to \psi - U_\Delta(\theta \to \psi)| d\psi &\leq m(\mathbf{x})^{-1} \Delta \sup_{\psi^* \in \Theta} ||D_\Theta f(\mathbf{x}|\psi*)|| \\
&:= C_1(\mathbf{x}, f, \Theta)\Delta
\end{aligned}$$

où $C_1(\mathbf{x}, f, \Theta)$ est une constante dépendant uniquement de $\mathbf{x}$, f et Θ.

(ii) Comme précédemment, le théorème de la moyenne est utilisé pour obtenir le résultat :

$$\sup_{\psi\in\Theta}\left|\frac{q_\Delta(\theta\to\psi|x)}{q_\Delta(\psi\to\theta|x)}-\frac{\frac{f(\mathbf{x}|\psi)}{c(\psi}\mathbf{1}_{b(\theta,\Delta/2)}(\psi)}{\frac{f(\mathbf{x}|\psi)}{c(\psi}\mathbf{1}_{b(\theta,\Delta/2)}(\theta)}\right| \leq \sup_{\psi\in b(\theta,\Delta/2)}\left|\frac{\frac{f(\mathbf{x}|\psi)}{c(\psi)}/(V_\Delta\frac{f(\mathbf{x}|\theta^*)}{c(\theta^*)})}{\frac{f(\mathbf{x}|\theta)}{c(\theta)}/(V_\Delta\frac{f(\mathbf{x}|\psi^*)}{c(\psi^*)})}-\frac{\frac{f(\mathbf{x}|\psi)}{c(\psi)}}{\frac{f(\mathbf{x}|\theta)}{c(\theta)}}\right|$$

$$\leq \sup_{\psi\in b(\theta,\Delta/2)}\left|\frac{\frac{f(\mathbf{x}|\psi)}{c(\psi)}}{\frac{f(\mathbf{x}|\theta)}{c(\theta)}}\right|\left|\frac{\frac{f(\mathbf{x}|\psi^*)}{c(\psi^*)}}{\frac{f(\mathbf{x}|\theta^*)}{c(\theta^*)}}-1\right|$$

$$\leq M(\mathbf{x})m(\mathbf{x})^{-2}\sup_{d(\theta^*,\psi^*)\leq\Delta}\left|\frac{f(\mathbf{x}|\psi^*)}{c(\psi^*)}-\frac{f(\mathbf{x}|\theta^*)}{c(\theta^*)}\right|$$

avec $M(\mathbf{x}) := \sup_{\Phi\in\Theta} f(\mathbf{x}|\Phi) < \infty$, $\theta^* \in b(\theta, \Delta/2)$ et $\psi^* \in b(\psi, \Delta/2)$ les valeurs obtenues par théorème de la moyenne. Comme précédemment, si $f(\mathbf{x}|\cdot) \in \mathcal{C}^1(\Theta)$, nous avons alors :

$$\sup_{\psi\in\Theta}\left|\frac{q_\Delta(\theta\to\psi|x)}{q_\Delta(\psi\to\theta|x)}-\frac{\frac{f(\mathbf{x}|\psi)}{c(\psi}\mathbf{1}_{b(\theta,\Delta/2)}(\psi)}{\frac{f(\mathbf{x}|\psi)}{c(\psi}\mathbf{1}_{b(\theta,\Delta/2)}(\theta)}\right| \leq M(\mathbf{x})m(\mathbf{x})^{-2}\Delta\sup_{\psi^*\in\Theta}||D_\Theta f(\mathbf{x}|\psi^*)||$$

$$:= C_2(\mathbf{x}, f, \Theta)\Delta$$

où $C_2(\mathbf{x}, f, \Theta)$ est une constante dépendant uniquement x, p et Θ □

A.2 Preuve de la proposition 2.3.3

Démonstration. Si $n = 1$, la définition de noyaux de transitions, l'introduction de $U_\Delta(\theta \to \psi)\alpha_{ideal}(\theta \to \psi) - U_\Delta(\theta \to \psi)\alpha_{ideal}(\theta \to \psi)$ dans la première valeur absolue, l'inégalité triangulaire et le caractère borné des fonctions $\mathbf{1}_A(\cdot)$, $\alpha_{ideal}(\cdot)$ et $\alpha_{shadow}(\cdot)$ permettent d'écrire :

$$\begin{aligned}
&|P_{ideal}(\theta, A) - P_{shadow}(\theta, A)| \\
&\leq \int_{\psi\in b(\theta,\Delta/2)} |q_\Delta(\theta\to\psi)\alpha_{ideal}(\theta\to\psi) - U_\Delta(\theta\to\psi)\alpha_{shadow}(\theta\to\psi)|d\psi \\
&+ \mathbf{1}_A(\theta)\int_{\psi\in b(\theta,\Delta/2)} |q_\Delta(\theta\to\psi)[1-\alpha_{ideal}(\theta\to\psi)] - U_\Delta(\theta\to\psi)[1-\alpha_{shadow}(\theta\to\psi)]|d\psi \\
&\leq 3\int_{\psi\in b(\theta,\Delta/2)} |q_\Delta(\theta\to\psi) - U_\Delta(\theta\to\psi)|d\psi + 2\int_{\psi\in b(\theta,\Delta/2)} U_\Delta(\theta\to\psi)|\alpha_{ideal}(\theta\to\psi) \\
&- \alpha_{shadow}(\theta\to\psi)|d\psi
\end{aligned}$$

Étant sous les hypothèses du théorème 2.3.1, appliquons le point **(i)** et le corollaire 2.3.2. Soit $\epsilon > 0$, il existe alors Δ_ϵ tel que $\forall\Delta > 0$,

$$[\Delta < \Delta_\epsilon \Rightarrow \forall\theta \in \Theta, \leq \int_{\psi\in b(\theta,\Delta/2)} |q_\Delta(\theta\to\psi) - U_\Delta(\theta\to\psi)|d\psi < \epsilon/6$$

et

$$\sup_{\psi\in\Theta} |\alpha_{ideal}(\theta\to\psi) - \alpha_{shadow}(\theta\to\psi)| < \epsilon/2]$$

D'où l'existence de $\Delta_\epsilon > 0$ tel que $|P_{ideal}(\theta, A) - P_{shadow}(\theta, A)| < \epsilon$ si $\Delta \leq \Delta_\epsilon$ indépendamment de θ et de A.

Si nous supposons désormais que $f(\mathbf{x}|\cdot) \in \mathcal{C}^1(\Theta)$, les inégalités précédentes deviennent donc par inégalité des accroissements finis :

$$|P_{ideal}(\theta, A) - P_{shadow}(\theta, A)| \leq 3C_1(\mathbf{x}, f, \Theta)\Delta + 2C_2(\mathbf{x}, f, \Theta)\Delta := C_3(\mathbf{x}, f, \Theta)\Delta.$$

Étant donné cette majoration, un candidat pour $\Delta_0(\epsilon, 1) := C_4(x, p, \Theta)\epsilon$. Les constantes C_3 et C_4 dépendant uniquement de $\mathbf{x}$, f et Θ. Pour $n > 1$, un raisonnement par induction amène :

$$\begin{aligned}
&|P_{ideal}^{(n)} - P_{shadow}^{(n)}| \\
&\leq |P_{ideal}^{(n)} - P_{ideal}^{(n-1)}P_{shadow}^{(1)} + P_{ideal}^{(n-1)}P_{shadow}^{(1)} - P_{shadow}^{(n)}| \\
&\leq |P_{ideal}^{(n-1)}||P_{ideal}^{(1)} - P_{shadow}^{(1)}| + |P_{ideal}^{(n-1)} - P_{shadow}^{(n-1)}||P_{shadow}^{(n-1)}| \\
&\leq |P_{ideal}^{(1)} - P_{shadow}^{(1)}| + |P_{ideal}^{(n-1)} - P_{shadow}^{(n-1)}| \\
&\leq \cdots \leq n|P_{ideal}^{(1)} - P_{shadow}^{(1)}| < \epsilon
\end{aligned}$$

uniformément en θ et A pour tout Δ tel que $\Delta \leq \Delta_0(\epsilon, n) := \Delta_0(\epsilon/n, 1) = C_4(x, p, \Theta)\frac{\epsilon}{n}$.

□

Annexe B

Présentation de la librairie C++ DRLib

DRLib est une libraire C++ destinée à la modélisation, la simulation et l'inférence de processus ponctuels marqués. Elle vise à compléter les outils existants, tels que la bibliothèque spatstat dans R [Baddeley et al., 2015], avec un code C++ fiable et efficace. Cette librairie trouve son origine dans la bibliothèque MPPLIB développée principalement par [van Lieshout and Stoica, 2006], où des algorithmes de simulation exacts pour les processus ponctuels marqués ont été programmés en C++ afin de réaliser des études de simulation à grande échelle. La bibliothèque DRLib est un projet construit par Radu Stoica et Didier Gemmerlé. Ce projet est en plein déroulement. Ce dernier permet d'effectuer la modélisation, la simulation par l'algorithme de Metropolis-Hastings ajout/retrait de points et l'inférence par l'algorithme ABC Shadow pour les processus ponctuels cités dans ce manuscrit. Les modèles implémentés dans la librairie disponible en ligne sont :

— Processus de Poisson.
— Processus de Strauss.
— Processus Area-Interaction.
— Processus de Strauss superposé avec Area-Interaction.

J'ai intégré dans cette librairie les travaux présentés dans cette thèse : les modèles de saturation de Geyer, StraussCrown et GeyerCrown et l'élargissement de la simulation de ces modèles à la simulation conditionnelle.

La version actuelle de la bibliothèque est disponible à l'adresse suivante :
https ://gitlab.univ-lorraine.fr/labos/iecl/drlib.

Annexe C

Points clés du déroulement de la thèse

C.1 Séjour scientifique

— Séjour de 2 mois en Suède Chalmers, Université de Göteborg pour une collaboration avec Pr. Aila Särkkä, mars - mai 2024.

C.2 Liste des publications

Ces travaux ont donné lieu à trois publications dans des proceedings de conférence :

— Nathan Gillot, Radu S. Stoica, Aila Särkkä, Didier Gemmerlé. Application of Approximate Bayesian Computation algorithms for parameter estimation for Gibbs point processes based on partly missing data. RING Meeting 2025, École nationale supérieure de géologie (ENSG) Nancy, Sep 2025, Nancy, France. ⟨hal-05184911⟩
— Nathan Gillot, Radu S. Stoica, Aila Särkkä, Didier Gemmerlé. Spatial point process modelling and Bayesian inference for large data sets. RING Meeting, École nationale supérieure de géologie (ENSG) Nancy, Sep 2024, Nancy, France. ⟨hal-04645186⟩
— Nathan Gillot, Radu S. Stoica, Didier Gemmerlé. Study the galaxy distribution characterisation via Bayesian statistical learning of spatial marked point processes. RING Meeting, École nationale supérieure de géologie (ENSG) Nancy, Sep 2023, Nancy, France. ⟨hal-04163649v2⟩

C.3 Liste des présentations et séminaires

— SSIAB 16 (Workshop on Spatial Statistics and Image Analysis in Biology) : «Bayesian algorithms for parameter estimation for Gibbs point processes based on partly missing data», Smögen, juin 2025
— Exposés au RING Meeting, Nancy, septembre 2023-2024-2025.
  — «Study the galaxy distribution characterisation via Bayesian statistical learning of spatial marked point processes» 2023 (Poster)
  — «Spatial point process modelling and Bayesian infernce for large data sets» 2024 (Exposé)
  — «Application of Approximate Bayesian Computation (ABC) algorithms for parameter estimation for Gibbs point processes based on partly missing data» 2025 (Exposé)

- Guest lecture, autour des méthodes MCMC (Échantillonneur de Gibbs) et méthodes Bayésiennes (ABC Shadow) à des étudiants de Master à Chalmers, mai 2024 et mai 2025. Pr Ottmar Cronie.
- Exposés au sein de l'équipe Inria PASTA :
  - «L'algorithme EM pour variables latentes.»
  - «Autour des constantes de normalisations.»
  - «Cartographie et Mathématiques : existe-t-il une carte «parfaite» ?»
  - «Introduction à la modélisation, simulation et inférence pour les processus ponctuels.»
- TP d'introduction à la `DRLib`, École d'Été à Haapsalu (Estonie), en collaboration avec R.S. Stoica et D.Gemmerlé. juillet 2025