

UNIVERSITÉ DES ANTILLES ET DE LA GUYANE

THÈSE

pour obtenir le grade de
Docteur de l'Université des Antilles et de la Guyane

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES

présentée par

Florencia CHIMARD
(florencia.chimard@univ-ag.fr)

Mélanges de Processus Ponctuels spatio-temporels et approche bayésienne semi-paramétrique

soutenue publiquement le jeudi 16 Décembre 2010

Composition du jury :

Président :	Professeur Jeffrey ROSENTHAL	University of Toronto, Canada
Rapporteur :	Professeur Lance WALLER	Emory University, Atlanta USA
Rapporteur :	Professeur Patrick BROWN	University of Toronto, Canada
Directeur :	Professeur Jean VAILLANT	Université des Antilles et de la Guyane
Co-directeur :	Professeur Richard EMILION	Université d'Orléans, France
Examineur :	Professeur Alex MERIL	Université des Antilles et de la Guyane

*A mes parents,
A mon frère Thony qui nous a quitté prématurément.*

Résumé

Mots-clés : Processus ponctuel, processus de Dirichlet, loi a priori stick-breaking, modèle bayésien hiérarchique, méthodes MCMC

Les processus ponctuels sont souvent utilisés comme modèles de répartitions spatiales ou spatio-temporelles d'occurrences. Pour notre part, dans cette thèse, nous nous intéressons à l'analyse statistique bayésienne de ces processus dans le cas où nous disposons de k cartes exhaustives à des dates d'observations échelonnées dans le temps. Nous proposons deux contextes d'étude.

Tout d'abord, nous considérons des occurrences constituant la réalisation d'un processus de Cox spatio-temporel dont l'intensité est associée à un processus shot noise généralisé. Le modèle correspond à une mesure d'intensité liée à des contributions générées par un processus caché de Poisson et qui suivent un processus de Dirichlet centré sur la loi Gamma. A partir des positions spatiales des occurrences observées entre plusieurs paires de dates d'observations consécutives, nous proposons d'inférer sur les paramètres d'intérêt à l'aide de méthodes MCMC dans le cadre d'un modèle bayésien hiérarchique. Un algorithme avec augmentation des données est proposé et testé sur des jeux de données artificielles.

D'autre part, nous analysons la situation où l'ensemble d'étude est un ensemble discret de positions possibles pour chaque occurrence du phénomène. La présence/absence d'une unique occurrence en une position donnée implique que nous aurons des données binaires. Par conséquent, nous développons un modèle de mélange de lois de Bernoulli avec un paramètre d'intensité d'arrière plan suivant un processus autorégressif d'ordre 1 log-gaussien. Nous utilisons une approche bayésienne hiérarchique pour mener à

bien l'inférence statistique de notre modèle. Nous développons un algorithme Metropolis-within-Gibbs pour calculer la loi a posteriori des paramètres d'intérêt. Des tests sont effectués sur des données artificielles et sur des données concernant le virus de la feuille jaune de la canne à sucre.

Abstract

Keywords : Point process, Dirichlet process, stick-breaking prior, hierarchical bayesian model, MCMC methods

Point processes are often used as tools for describing spatial or spatio-temporal point patterns. In this Phd dissertation, we give an overview of bayesian statistical analysis for point processes and recent tools like the Dirichlet process and its diverse extensions. We focus on situations where the available data are maps of the studied point process at different observations dates. Two contexts are considered.

Firstly, we consider occurrences of events in a studied area forming the realization of a spatio-temporal Cox process directed by a generalized shot noise intensity measure. A hidden Poisson process generates contributions to the intensity measure which are distributed according to a Dirichlet process centered on the Gamma distribution. For data consisting of spatial locations of occurrences between several pairs of consecutive observation dates, we develop statistical inference about the parameters of interest by means of MCMC methods within the framework of hierarchical bayesian modeling. A data augmentation algorithm is introduced and tested on artificial data.

Secondly, we analyse the case where the point process support is discrete with at most one occurrence for a given element of the support. For such binary data, we present and discuss models based on Bernoulli distribution mixture with a background intensity following a log-gaussian. The statistical inference for these models is developed by using a hierarchical bayesian approach. Tests are carried out on artificial data and data from Yellow Leaf Sugarcane Virus observations.

Remerciements

Je voudrais en premier lieu exprimer toute ma gratitude au professeur Jean VAILLANT, pour sa disponibilité, son soutien, ses encouragements, au cours de ces trois années, mais aussi pour avoir fait preuve, à plusieurs reprises, de sa confiance en mon travail de recherche.

Je remercie également le professeur Richard EMILION, co-directeur de thèse, pour ses précieuses remarques et les conseils qu'il m'a prodigués.

Je voudrais exprimer toute ma reconnaissance au professeur Jeffrey ROSENTHAL qui m'a fait l'honneur de participer et présider ce jury de thèse. Je le remercie de m'avoir accueilli au sein de son département à l'Université de Toronto pour un séjour de recherche auquel cette thèse doit beaucoup.

Je remercie sincèrement les professeurs Patrick BROWN et Lance WALLER, ainsi que le professeur Alex MERIL, pour les remarques et suggestions qui m'ont permis d'améliorer cette thèse.

Je tiens à remercier toute ma famille et tous mes amis pour avoir contribué à un environnement favorable à mes activités de recherche.

Pour finir, je remercie toutes les personnes qui ont contribué de près ou de loin à la concrétisation de cette thèse.

Table des matières

Résumé	3
Abstract	5
Remerciements	7
Notations	13
Abréviations et sigles	15
Liste des figures	19
Liste des algorithmes	21
Introduction	23
1 Processus Ponctuels	39
1.1 Le contexte mathématique	39
1.2 Formalisme des processus ponctuels	40
1.2.1 Définitions et généralités	40
1.2.2 Caractérisations d'un processus ponctuel	43
1.3 Processus Ponctuels fondamentaux	46
1.3.1 Processus de Bernoulli, Processus Binomial	46
1.3.2 Processus de renouvellement	46
1.3.3 Processus de Poisson homogène	47
1.3.4 Processus de Poisson non homogène	48
1.3.5 Processus de Cox	48
1.4 Processus Vraisemblance d'un PP	49
1.5 Statistiques élémentaires des PP	50

1.5.1	Echantillonnage des PPST	50
1.5.2	Estimation standard	51
1.5.3	Statistique de balayage	51
1.5.4	Tests d'homogénéité spatiale	51
1.6	Quelques exemples illustratifs	53
1.6.1	Exemple 1 : Loh et Zhu (2007, [40])	53
1.6.2	Exemple 2 : Zhuang (2006, [74])	54
1.6.3	Exemple 3 : Diggle et al. (2005, [17])	54
1.6.4	Exemple 4 : Vaillant (1992, [66])	55
2	Éléments d'analyse bayésienne hiérarchique	57
2.1	Approche bayésienne paramétrique	57
2.1.1	Introduction au choix bayésien	57
2.1.2	Des informations a priori aux lois a priori	59
2.1.3	Modèles hiérarchiques	59
2.1.4	Mélange de lois	61
2.1.5	Modèles de mélange dans le cadre bayésien	62
2.1.6	Atouts et justifications	64
2.1.7	Lois a priori conjuguées	64
2.1.8	Décompositions conditionnelles	66
2.2	Méthodes de calculs en analyse bayésienne	67
2.2.1	Contraintes d'élaboration	67
2.2.2	Intégration numérique	68
2.2.3	Méthodes de Monte-Carlo standard	69
2.2.4	Méthode de l'échantillonnage préférentiel	71
2.2.5	Simulation	72
2.2.6	Chaînes de Markov	73
2.2.7	Propriétés des chaînes de Markov	75
2.2.8	Méthodes MCMC = Markov Chain Monte Carlo	77
2.2.9	Impact des méthodes MCMC sur la statistique bayésienne	80
3	Classes de processus Stick-Breaking	81
3.1	Préliminaires	81
3.2	Loi de Dirichlet et sa généralisée	84
3.3	Modèle d'urne de Polya	86
3.4	Distribution a posteriori et estimation bayésienne	87
3.4.1	Échantillonnage multinomial	87
3.4.2	Naturelle conjuguée	88
3.5	Processus de Dirichlet	89
3.5.1	Définitions	89
3.5.2	Modèle d'urne de Blackwell-MacQueen et règle prédictive	92

3.5.3	Processus de Dirichlet et mélanges	95
3.5.4	Classification et métaphore du "restaurant chinois"	98
3.6	Méthodes d'approximation de la loi a posteriori	99
3.6.1	Echantillonneur de Gibbs avec lois a priori conjuguées	99
3.6.2	Processus de Dirichlet hiérarchisé	101
3.7	Classe des lois a priori Stick-Breaking	102
3.7.1	Construction du stick-breaking	102
3.7.2	Troncature des mesures $G_\infty(a, b, .)$	104
3.7.3	Application du stick-breaking dans le Bayésien	105
3.8	Processus Stick-breaking à noyaux	106
3.8.1	Motivation	106
3.8.2	Formulation du modèle	107
3.8.3	Propriétés conditionnelles et marginales	108
3.8.4	Représentation alternative et troncature	109
3.8.5	Loi prédictive	109
3.9	Processus de Dirichlet oblique	112
3.9.1	Définitions	112
3.9.2	Skewed Polya sequences	113
4	Mélanges de Processus Ponctuels et estimations	117
4.1	Données observables	117
4.2	Introduction	120
4.3	Notations et contexte	121
4.4	Processus de Dirichlet et mélanges	124
4.5	Modélisation bayésienne hiérarchique semi paramétrique	125
4.6	Quelques propriétés distributionnelles	130
4.7	Génération de la loi a posteriori	132
4.7.1	Fonctions de mise à jour	132
4.7.2	Algorithme MCMC	133
4.8	Étude des propriétés statistiques à partir de données artificielles	135
4.8.1	Données artificielles	135
4.8.2	Simulations et résultats	136
5	Modèle bayésien pour répartition sur une grille	147
5.1	Modèles en épidémiologie végétale	147
5.1.1	Modèle en temps continu	147
5.1.2	Modèle en temps discret	148
5.1.3	Inférence a posteriori	152
5.2	Simulation et algorithmes MCMC	156
5.2.1	Génération des données artificielles	156
5.2.2	Algorithmes MCMC	157

5.2.3	Traitement statistique des données artificielles	160
5.3	Une illustration : La maladie de la feuille jaune de la canne à sucre	167
5.3.1	La canne à sucre	167
5.3.2	Le virus SCYLV	168
5.3.3	Données observées	168
5.3.4	Modélisation hiérarchique	169
6	Conclusions et perspectives	175
A	Simulation d'un Processus Ponctuel	179
A.1	Processus Ponctuel Temporel	179
A.2	Processus Ponctuel Spatial	180
A.3	Processus Ponctuel Spatio-Temporel	180
B	Processus gaussien et log-gaussien	181
B.1	Densité normale univariée	181
B.2	Densité log-normale	181
B.3	Densité normale multivariée	182
B.4	Processus gaussien et log-gaussien	182
C	Rappels mathématiques	183
C.1	Processus auto-régressif	183
C.2	Mouvement brownien	183
C.3	Introduction aux processus stochastiques	184
C.3.1	Processus stochastiques	184
C.3.2	Processus stationnaires	186
C.3.3	Martingales et Semi-martingales	187
D	Exemples d'intensité cumulée	189
E	Programmes informatiques complémentaires	197
E.1	Simulation d'un PP	197
E.1.1	Simulation d'un PP temporel	197
E.1.2	Simulation d'un PP spatial	198
E.1.3	Simulation d'un PP spatio-temporel	199
E.2	Script R du chapitre 4	199
E.2.1	Script pour le tracé de l'intensité cumulée	199
E.2.2	Script pour la génération des données artificielles	200
E.2.3	Script pour la génération de la la loi a posteriori	201
	Bibliographie	211