Le processus de Hawkes : inférence Bayésienne, sélection de variable et application en épidémiologie végétale

Proposition de stage niveau M2 à INRAE Jouy-en-Josas

Pour postuler, envoyez un CV et votre dernier relevé de notes à
katarzyna.adamczyk@inrae.fr et guillaume.konkamking@inrae.fr.

Contexte et motivation

Le processus de Hawkes est un modèle stochastique utilisé pour décrire des événements se produisant de manière séquentielle dans le temps, où chaque événement peut influencer la probabilité d'occurrence des événements futurs (Figure 1 gauche). Ce modèle non-markovien est particulièrement utile pour analyser des phénomènes auto-excitants, c'est-à-dire quand la survenue d'un événement augmente la probabilité de survenue d'événements similaires dans un avenir proche (par exemple, des répliques de séismes, la propagation d'infections, l'activité neuronale, etc.). En épidémiologie végétale, le processus de Hawkes permet de modéliser la propagation des maladies au sein des populations de plantes, en tenant compte des interactions complexes entre les agents pathogènes et les plantes hôtes. Dans le cadre de ce stage, il s'agit de construire un modèle prédictif de la propagation de la tavelure des pommiers, une maladie causée par le champignon Venturia Inaequalis qui libère des spores à intervalles irréguliers. Cette maladie peut altérer l'aspect des pommes et dégrader les récoltes. La sporulation et donc la propagation de la tavelure est influencée par les conditions climatiques et la variété des pommes, rendant sa gestion complexe. Un modèle prédictif utilisant le processus de Hawkes enrichi par des covariables serait particulièrement intéressant car il permettrait de prévoir les dates de libération des spores en fonction de différents facteurs d'influence. Un des enjeux est d'inclure, au-delà des données décrivant les événements de sporulation, un grand nombre de covariables hétérogènes de type environnemental, variétal, climatique, etc., pour améliorer le pouvoir prédictif du modèle.

Les études expérimentales sur la dynamique du champignon Venturia Inaequalis ont conduit à l'élaboration de modèles déterministes basés sur des règles de décision pour simuler le cycle biologique du champignon. Ces modèles, implémentés dans des logiciels comme RIMpro[1, 2], utilisent des données climatiques et de libération de spores pour évaluer le risque d'infection. Cependant, ces approches sont limitées par leur dépendance aux conditions expérimentales spécifiques et ne prennent pas en compte l'aléa inhérent à la dispersion des spores, c'est à dire la nature stochastique du processus.

Un modèle prédictif basé sur le processus de Hawkes, en revanche, permettrait d'intégrer de manière plus flexible les effets environnementaux et les interactions entre les agents pathogènes et les plantes hôtes. Ce modèle prédictif serait un outil précieux pour le traitement préventif des cultures et la réduction de la quantité de fongicides utilisée, car dans l'impossibilité de bien prédire les dates de sporulation on utilise actuellement de l'ordre de 20 traitements par saison, nombre que l'on pourrait réduire en améliorant la performance prédictive des modèles.

Objectifs du stage

Le processus de Hawkes est un processus ponctuel, au sens où une réalisation de ce processus correspond à une collection de points dans un ensemble S. Dans le cas univarié ($S = \mathbb{R}^+$), les points représentent par exemple les dates d'évènements de sporulation. La fonction d'intensité conditionnelle pour le processus de Hawkes est définie par l'équation :

$$\lambda(t) = \lambda_0 + \sum_{t_i<t} \mu(t - t_i) \tag{1}$$

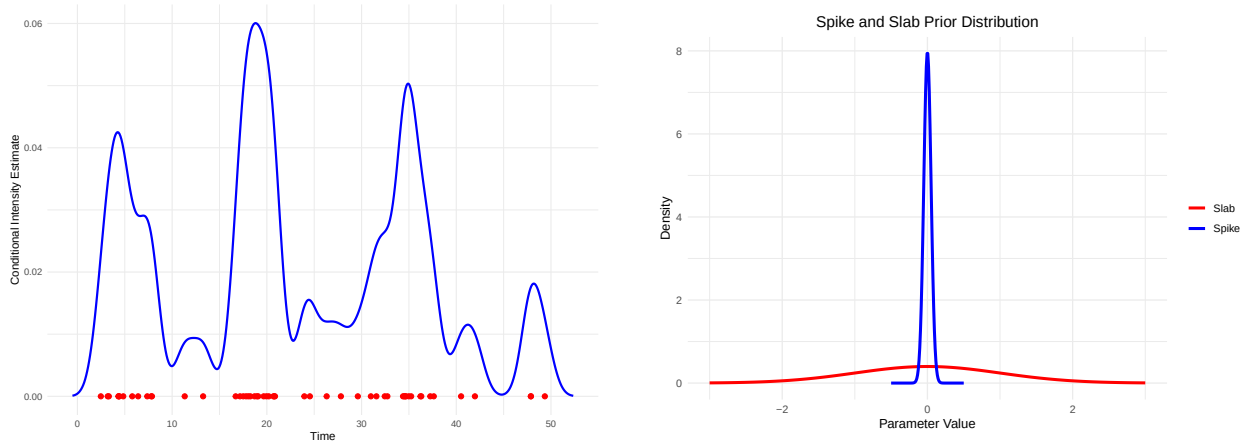

Figure 1: **Gauche** : simulation d'un processus de Hawkes univarié, avec estimation (fenêtre glissante) de la l'intensité conditionnelle. Les points représentent d'hypothétiques dates de sporulation. **Droite** : distribution a priori Spike and Slab (continue) pour des paramètres de régression.

Le scalaire λ_0 représente le taux d'occurrence de base, indépendant du temps. Le deuxième terme traduit l'influence du passé du processus via la fonction μ qui dépend du délai entre le temps t et les temps d'occurrence t_i des évènements antérieurs. Cette intensité peut ensuite être intégrée, généralement de façon analytique, et donner une vraisemblance explicite. Une manière naturelle d'inclure des covariables dans ce modèle consiste à autoriser le taux d'occurrence de base λ_0 ou la fonction μ à dépendre de ces covariables, de manière linéaire ou non. En présence d'un grand nombre de covariables, il se pose naturellement la question de savoir s'il faut toutes les inclure et de comment gérer les défis classiques la régression en grande dimension. La sélection d'un nombre restreint de covariables et la construction de modèles parcimonieux sont généralement essentiels pour obtenir une bonne performance prédictive.

L'inférence Bayésienne a récemment démontré son efficacité pour la sélection de variables en grande dimension dans les modèles non-linéaires[3]. Une stratégie classique consiste à utiliser des priors induisant la parcimonie, tels que les priors Spike-and-Slab et Horseshoe. Le prior Spike-and-Slab (Figure 1 droite) est un mélange de distributions, l'une très concentrée autour de zéro ("spike") et l'autre plus large ("slab"), favorisant ainsi l'exclusion des variables non pertinentes tout en permettant aux variables pertinentes de prendre des valeurs non nulles, sans le biais fort typiquement associé à l'approche LASSO. La flexibilité de l'inférence Bayésienne permet d'utiliser ce type de prior au sein d'un modèle hiérarchique pour induire de la parcimonie sur la structure choisie, en l'occurrence sur le nombre de covariables impactant la fonction d'intensité conditionnelle. De plus, l'inférence Bayésienne est particulièrement adaptée à l'objectif de prédiction de ce projet, car elle offre une méthode naturelle de quantification de l'incertitude de prédiction, tenant compte de l'incertitude d'estimation des divers paramètres et notamment sur les variables sélectionnées.

La popularité des méthodes Bayésiennes de sélection de variables a conduit à de nombreuses avancées en statistique computationnelle pour échantillonner dans la distribution a posteriori[4, 5]. Dans ce stage, nous proposons d'explorer d'abord des méthodes de type Monte Carlo Markov Chain, dont l'efficacité a été observée pour des problèmes allant jusqu'à 1000 covariables. Nous pourrions ensuite passer à des méthodes applicables à une plus grande dimension.

Profil recherché

Le candidat doit être en formation de M2 (ou une formation équivalente) en probabilités et statistique. Un intérêt pour la modélisation statistique, des notions d'apprentissage statistique et de programmation en `R` ou Python sont nécessaires.

Compétences acquises à l'issue du stage

Le stagiaire développera une variété de savoirs recherchés dans le monde académique et industriel : familiarité avec l'inférence bayésienne, maîtrise d'un langage de programmation probabiliste et des notions de statistique computationnelle.

Conditions du stage

Laboratoire d'accueil

UR 1404 Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE), INRAE, 78352 Jouy-en-Josas

Le centre INRAE de Jouy-en-Josas propose un éventail d'activités sportives et culturelles (`https://adas-jouy.hub.inrae.fr/`.)

Encadrants

Katarzyna Adamczyk, : `katarzyna.adamczyk@inrae.fr`
Guillaume Kon Kam King : `guillaume.konkamking@inrae.fr`

Durée : à partir de 5-6 mois

Gratification environ 610 euros nets par mois. Il existe des possibilités de logement à tarif préférentiel à l'entrée du centre, mais leur disponibilité n'est pas garantie.

References

[1] L. Jamar and M. Lateur, 'Strategies to reduce copper use in organic apple production', in I international symposium on organic apple and pear 737, 2006, pp. 113–120.

[2] A. Duval-Chaboussou and A. Leblois, 'Synthesis of 4 years of evaluation of natural substances in the control of apple scab with a view to reduce copper doses', in XXXI international horticultural congress (IHC2022): International symposium on sustainable control of pests and diseases 1378, 2022, pp. 97–104.

[3] M. Naveau, G. Kon Kam King, R. Rincent, L. Sansonnet, and M. Delattre, 'Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the SAEM algorithm', Stat Comput, vol. 34, no. 1, p. 53, Dec. 2023, doi: 10.1007/s11222-023-10367-4.

[4] K. Ray, B. Szabo, and G. Clara, 'Spike and slab variational Bayes for high dimensional logistic regression', in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 14423–14434.

[5] G. Zanella and G. Roberts, 'Scalable Importance Tempering and Bayesian Variable Selection', Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 81, no. 3, pp. 489–517, Jul. 2019, doi: 10.1111/rssb.12316.

[6] P. Jankowski and S. Masny, 'Influence of moisture on maturation rate of the Venturia inaequalis (Cooke) Wint. ascospores in central Poland', Journal of Plant Diseases and Protection, vol. 127, no. 2, pp. 155–163, 2020.