Analyse de survie appliquée à la modélisation de la transmission des maladies infectieuses : mesurer l'impact des interventions

Génia Babykina [1] & Simon Cauchemez [2]

[1] *Université de Lille, Faculté Ingénierie et Management de la Santé (UFR ILIS), CERIM, 42 rue Ambroise Paré, 59120 Loos, France, evgeniya.babykina@univ-lille2.fr*

[2] *Institut Pasteur, Unité de Modélisation Mathématique des Maladies Infectieuses, 28 rue du Dr Roux, Paris 75015, France, simon.cauchemez@pasteur.fr*

Résumé. Dans l'article l'approche d'analyse de survie est proposée pour la modélisation de la transmission des maladies infectieuses dans les ménages. Le risque d'infection tient compte des caractéristiques des individus du ménage, de la communauté, de l'infectivité de l'individu contagieux, mais également des interventions mises en place pour réduire la transmission. L'approche Bayésienne est adaptée pour l'estimation des paramètres. Une technique d'augmentation de données est appliquée afin de tenir compte des instants d'infection non observés. Les résultats des simulations de Monte Carlo sont présentés et discutés.

Mots-clés. analyse de survie, MCMC, inférence Bayésienne, augmentation de données, modélisation des maladies infectieuses

Abstract. In this article a survival analysis approach is proposed to model the transmission of infectious diseases within households. The risk of infection accounts for the characteristics of household members, of the community and of individual infectivity. It also accounts for the impact of interventions aiming to reduce transmission. The Bayesian approach is used for parameter estimation. Data augmentation technique is used to account for unobservable instants of infections. The Monte Carlo simulations results are presented and discussed.

Keywords. survival analysis, MCMC, Bayesian inference, data augmentation, modeling of infectious diseases

1 Introduction

Dans l'article la modélisation de la transmission des maladies infectieuses au sein de ménages est approchée par l'analyse de données de survie. Le risque de la transmission dans le modèle proposé dépend des caractéristiques des individus du ménage, de la communauté et de l'infectivité de l'individu contagieux. L'inférence se fait dans un cadre Bayésien, où

une stratégie d'augmentation de données est mise en place pour gérer les problèmes de données manquantes, les instants d'infection n'étant pas observés. L'algorithme MCMC est utilisé pour explorer la distribution jointe a posteriori des paramètres et des données augmentées. Dans le passé, ce type d'approches a été utilisé pour estimer les paramètres décrivant la transmission de la grippe dans les ménages [1, 2]. Ici, nous explorons la performance de ces méthodes pour évaluer l'impact d'interventions (usage de masques ou traitements antiviraux) visant à réduire la transmission dans le ménage.

2 Notations et modèle

Nous considérons un temps discret (une échelle journalière), un ménage m de taille N_m, un membre j du ménage m, $j \in (1, \cdots, N_m)$ et le jour d'infection du sujet j, d_j. Notons D le temps de censure (fin d'observation) et posons $d_j = D$ pour les individus n'ayant pas présenté de symptômes pendant la période d'observation.

Il est alors possible de modéliser le risque d'infection de l'individu j pour le jour d par un individu i, infecté le jour d_i avec $d > d_i$, i et j étant les membres d'un même ménage m comme (voir [1]):

$$P^m_{i\to j}(d) = f_1(\boldsymbol{Y}^m; \boldsymbol{\beta})\, f_2\left(\boldsymbol{X}^i; \boldsymbol{\gamma}_1\right) f_3\left(\boldsymbol{X}^j; \boldsymbol{\gamma}_2\right) f_4(d - d_i; \boldsymbol{\delta}), \tag{1}$$

avec f_1 une fonction des caractéristiques du ménage $\boldsymbol{Y}^m$, f_2 et f_3 les fonctions des caractéristiques de l'individu infectieux ($\boldsymbol{X}^i$) et de l'individu susceptible ($\boldsymbol{X}^j$), f_4 une fonction décrivant l'infectivité au cours du temps qui est une approximation de la charge virale et peut être spécifiée par une fonction simple décroissante en fonction du temps, $(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\delta}) \in \mathbb{R}^p$ le vecteur de paramètres à estimer.

Le taux (ou intensité) d'infection pour un individu j du ménage m le jour d est ainsi défini comme

$$\lambda_j(d) = \sum_{i=1, i\neq j}^{N_m} \left(P^m_{i\to j}(d)\right) + P^c_j(d) \tag{2}$$

Le premier terme de l'Eq.(2), sous forme d'une somme, est le risque d'infection de l'individu j à l'intérieur du ménage, par tous les membres infectés avant le jour d, le deuxième terme de l'Eq.(2) est le risque d'infection communautaire le jour d qui peut dépendre des caractéristiques de la communauté $\boldsymbol{Z}^c$ et du jour d. Le risque communautaire est paramétré par le vecteur $\boldsymbol{\alpha} \in \mathbb{R}^k$: $P^c_j(d) = f_5(\boldsymbol{Z}^c, d; \boldsymbol{\alpha})$.

L'instant exact de l'infection d'un individu n'est en général pas observé. Les données disponibles contiennent l'instant (date, jour) d'apparition des symptômes chez l'individu. Les techniques d'augmentation des données [2] sont alors utilisées pour l'inférence.

Dans la suite nous notons $\tilde{d}_j$ le jour d'apparition des symptômes chez l'individu j avec $\tilde{d}_j \geq d_j$, et $\mu\left(\tilde{d}_j - d_j, \boldsymbol{\eta}\right)$ une loi de distribution de la durée entre l'infection et l'apparition des symptômes (période d'incubation), paramétrée par le vecteur $\boldsymbol{\eta} \in \mathbb{R}^l$.

3 Inférence statistique

La probabilité conjointe des observations W, des variables non-observées ψ et des paramètres ζ s'écrit comme [3]

$$\mathbb{P}\left[W, \psi, \zeta\right] = \mathbb{P}\left[W|\psi\right]\mathbb{P}\left[\psi|\zeta\right]\mathbb{P}\left[\zeta\right]. \tag{3}$$

Le premier terme multiplicatif de l'Eq.(3) correspond au *niveau d'observation* assurant la cohérence entre les données observées et augmentées, le deuxième terme multiplicatif est le *niveau de transmission* modélisant la transmission de maladie au sein des ménages en supposant que les instants d'infections sont connus, le dernier terme est la distribution *a priori* des paramètres.

Conditionnellement aux observations le jour 0 (première personne malade au sein d'un ménage) et au vecteur de paramètres $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\delta}, \boldsymbol{\eta})$, $\boldsymbol{\theta} \in \mathbb{R}^{p+k+l}$ (p étant le nombre de paramètres caractérisant la transmission à l'intérieur d'un ménage, k le nombre de paramètres caractérisant la transmission inter-ménages et l le nombre de paramètres de la fonction décrivant la période d'incubation), la fonction de vraisemblance augmentée pour un individu j dont les symptômes apparaissent le jours $\tilde{d}_j$ s'écrit alors comme :

$$L_j^m\left(\tilde{d}_j, d_j|\boldsymbol{\theta}\right) = \left[1 - \exp\left(-\lambda_j\left(d_j\right)\right)\right]^{\mathbb{1}_{d_j<D}} \exp\left(-\sum_{0<u<d_j}\lambda_j\left(u\right)\right)\left[\mu\left(\tilde{d}_j - d_j\right)\right]^{\mathbb{1}_{d_j<D}}, \tag{4}$$

λ_j et μ étant fonctions de $\boldsymbol{\theta}$ et $\mathbb{1}_A$ une fonction indicatrice d'un événement A.

La méthode de Monte Carlo Markov Chain (MCMC) avec l'algorithme de Metropolis-Hasting est utilisée pour explorer la distribution jointe a posteriori des paramètres du modèle $\boldsymbol{\theta}$ et des données augmentées.

4 Résultats numériques

Les propriétés d'estimateurs sont évaluées empiriquement par les simulations de Monte Carlo. Les données sont simulées à partir de la structure de données sur la transmission de la grippe recueillies au Bangladesh : les tailles des ménages ainsi que les caractéristiques personnelles des membres ont été conservées. Il y a 1185 ménages de taille 4.95 personnes en moyenne. Nous supposons qu'une intervention (traitement, vaccination) a un effet sur l'infectivité d'un individu. Plus précisément la fonction de l'infectivité aura une forme

différente pour les individus traités. La distribution de Weibull à deux paramètres est adoptée pour décrire ce phénomène.

Le risque de transmission de l'Eq.(1) défini dans les simulations est

$$P^m_{i\to j}(d) = \begin{cases} c_1 \times \mathcal{W}(d - d_i;\, k_1, m_1) & \text{si l'individu } i \text{ est traité} \\ c_0 \times \mathcal{W}(d - d_i;\, k_0, m_0) & \text{si l'individu } i \text{ n'est pas traité,} \end{cases}$$

$\mathcal{W}(\cdot;\, k_h, m_h)$ étant la distribution de Weibull discrète avec k_h le paramètre de forme et m_h la médiane.

Le vrai vecteur de paramètres $\boldsymbol{\theta} = (c_0, k_0, m_0, c_1, k_1, m_1)$ est fixé à

$$\boldsymbol{\theta} = (0.1, 1.2, 3, 0.05, 1.6, 1.9).$$

Les courbes d'infectivité correspondantes sont présentées sur la Figure 1. L'individu traité est en général moins contagieux ($c_1 < c_0$), de plus sa force d'infection est concentrée au début de la maladie. L'individu non traité a une force d'infection en général plus importante et est contagieux pendant une période plus longue.

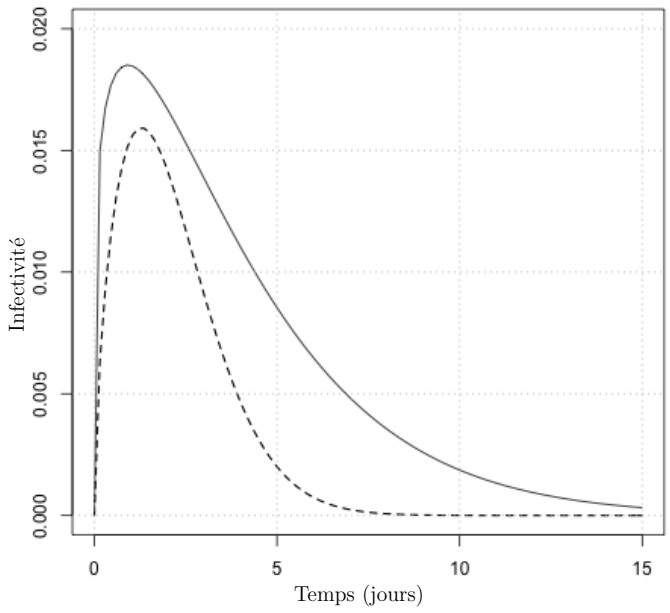

Figure 1: Courbes simulées : l'infectivité des individus non traités (trait plein) et l'infectivité des individus traités (trait en pointillé)

Les données ainsi simulées résultent en taux d'attaque secondaire moyen (le nombre moyen d'individus infectés par un cas) de 0.41, soit en moyenne 482.9 contacts infectés

dans les ménages. Les données simulées contiennent les dates d'apparition de symptômes de 5871 membres de 1185 ménages observés pendant 17 jours après l'apparition de symptômes du premier individu du ménage. 5000 itérations de MCMC avec la période de *burn-out* de 1000 sont réalisées sur chacun des 10 jeux de données simulées. Les distributions *a priori* non informatives des paramètres sont utilisées.

Les résultats d'estimations des paramètres sont présentés sur la Figure 2.

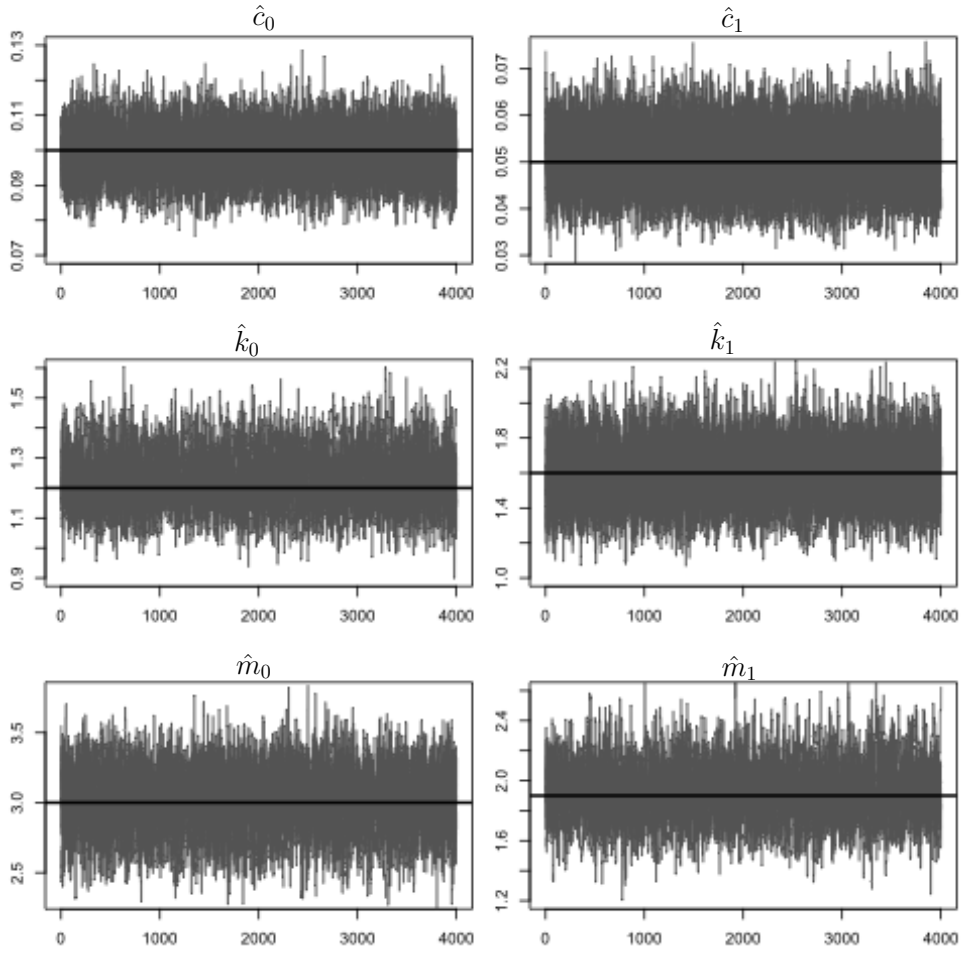

Figure 2: Estimations des paramètres à partir de données simulées : résultats de 10 simulations, le vrai paramètre est présenté par le trait noir

La convergence d'estimateurs et l'absence du biais sont visuellement confirmés. Les résultats numériques d'estimations basées sur un des 10 jeux de données sont donnés dans le Tableau 1. L'effet d'intervention (traitement) sur les paramètres du modèle est distingué en termes d'intervalles de crédibilité.

5 Conclusion et perspectives

Le modèle de survie est proposé pour l'étude de la transmission des maladies infectieuses. Les techniques Bayésiennes de MCMC sont utilisées pour l'estimation des paramètres, les instants d'infections non observés sont traités par l'augmentation de données. L'étude de Monte Carlo montre la convergence et l'absence de biais dans les estimations. Cette analyse de simulation démontre que ces méthodes peuvent être utilisées pour évaluer l'impact d'interventions visant à réduire la transmission dans les ménages. Nous allons utiliser cette approche pour étudier un jeu de données décrivant la transmission de la grippe dans des ménages au Bangladesh où les personnes malades ont reçu un traitement antiviral. L'impact du traitement sur la transmission de la grippe sera évalué.

Table 1: Résultats d'estimations sur un jeu de données : moyennes empiriques et intervalles de crédibilité de 95%

Paramètre	θ_j	$\hat{\theta}_j$ ($\hat{\sigma}_{\hat{\theta}_j}$)	95% IC
c_0	0.10	0.10 (0.005)	$[0.09, 0.11]$
c_1	0.05	0.05 (0.005)	$[0.04, 0.06]$
k_0	1.20	1.18 (0.62)	$[1.06, 1.30]$
k_1	1.60	1.63 (0.14)	$[1.36, 1.90]$
m_0	3.00	2.80 (0.17)	$[2.46, 3.13]$
m_1	1.90	2.17 (0.16)	$[1.86, 2.47]$

References

[1] S. Cauchemez, Ch.A. Donnelly, C. Reed, A.C. Ghani, Ch. Fraser, Ch. Kent, L. Finelli et N. Ferguson, Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States, *New England Journal of Medicine*, 361(27), pp. 2619–2627, 2009

[2] S. Cauchemez, F. Carrat, C. Viboud, A.-J. Valleron et P.-Y. Boelle, A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data, *Statistics in medicine*, 23(22), pp. 3469–3487, 2004

[3] K. Auranen, E. Arjas, T. Leino et A.K. Takala, Transmission of pneumococcal carriage in families: a latent Markov process model for binary longitudinal data, *Journal of the American Statistical Association*, 95(452), pp. 1044–1053, 2000