Introduction aux Processus Stochastiques

Projet Chaîne de Markov en temps discret

Analyse de la propagation d'un virus par chaîne de Markov

Profs.: Céline Esser et Pierre Geurts

Année académique 2020-2021

Ce travail est à réaliser par groupe de 2 étudiants. Le rapport et le code source sont à remettre via la plateforme de soumission de Montefiore pour **le vendredi 7 mai 2021 à 23h59** au plus tard.

Contexte général et objectifs

Dans ce projet, et en lien avec l'actualité, on se propose de modéliser la propagation d'un virus au sein d'une population à l'aide de processus de Markov en temps discret.

Le projet est découpé en deux parties. La **première partie** vous guidera pas à pas dans l'utilisation de chaînes de Markov pour modéliser la propagation d'un virus sur un graphe connectant les individus de la population. On considèrera d'abord "sur papier" le cas très simplifié d'une bulle de deux individus. On considéra ensuite deux techniques de simulation (au niveau des individus ou plus macroscopiques) de la chaîne pour traiter un nombre d'individus plus important.

Dans la **seconde partie** du projet, on utilisera le même type d'outils de modélisation pour étudier un cas plus réaliste où il s'agira de prédire l'évolution du virus au sein de la population belge et d'utiliser ces prédictions pour faire des recommendations.

Remarque : la démarche qu'on vous fait explorer dans le cadre du projet est réaliste mais les modèles utilisés sont fortement simplifiés par rapport aux modèles utilisés par les experts et pas nécessairement les plus appropriés. La principale raison de cette simplification est de vous permettre de réaliser le projet dans un temps raisonnable et également de minimiser les temps de calcul de vos simulations.

1 Modèles de propagation d'un virus sur un graphe

On suppose une population de N individus pouvant être dans trois catégories par rapport à un virus [1] :

1. Il s'agit d'un modèle compartimental de type SIR. Voir par exemple `https://fr.wikipedia.org/wiki/Modèles_compartimentaux_en_épidémiologie`

— Susceptibles (S) : des individus sains susceptibles d'attraper le virus

— Infectieux (I) : des individus infectés par le virus et donc malades

— Immunisés (R) : des invidus ayant été infectés précédemment mais guéris et (momentanément) immunisés contre une réinfection.

Ce modèle est très simplifié. Il ignore par exemple les individus décédés et il ne tient pas compte non plus du fait qu'un individu est mis en quarantaine dès qu'il montre les symptômes de la maladie. Ces deux simplifications n'auront cependant pas un grand impact sur la propagation de la maladie, du moins dans le cas où les individus ne peuvent pas perdre leur immunité. En effet, du point de vue de la propagation du virus, les personnes décédées et les personnes en quarantaine sont équivalentes à des personnes immunisées qui ne peuvent plus transmettre la maladie.

Contrairement à des modèles plus macroscopiques basés sur des équations différentielles [2], on souhaite dans ce projet étudier la propagation de la maladie au niveau des individus eux-mêmes (ou sous-groupes d'individus dans un second temps) en prenant en compte les interactions entre eux. Pour modéliser ces interactions, on supposera donné un graphe (non dirigé) reliant ces individus, sous la forme d'une matrice d'adjacence $W \in \{0,1\}^{N\times N}$. Un élément $W_{i,j}$ vaudra 1 si les individus i et j sont susceptibles de se rencontrer (et donc de se transmettre le virus), 0 sinon.

Etant donné ces hypothèses, le modèle (stochastique) de propagation en temps discret proposé est le suivant :

— Un individu infectieux au temps t a une probabilité β d'infecter au temps $t+1$ chacun des individus susceptibles auquel il est connecté dans le graphe

— Un individu infectieux au temps t a une probabilité μ de guérir, et de devenir immunisé, au temps $t+1$. La guérison d'un individu au temps $t+1$ n'aura pas d'impact sur sa faculté à infecter d'autres individus au temps t selon la première règle.

— Un individu immunisé au temps t a une probabilité α de redevenir susceptible au temps $t+1$.

La probabilité β modélise le fait qu'un individu ne rencontre pas nécessairement tous ses contacts à chaque pas de temps et qu'interagir avec une autre individu ne va pas nécessairement mener à une infection. μ^{-1} et α^{-1} représentent le nombre moyen de pas de temps nécessaires respectivement à la guérison d'un individu I et à la perte d'immunité d'un individu R.

Tel que décrit, le processus est un processus de Markov en temps discret. Les états de la chaîne correspondante sont représentés par les catégories, parmi 3, de l'ensemble des N individus de la population. Le nombre d'états de la chaîne est donc 3^N.

Un modèle très similaire a été proposé par exemple dans cet article [1].

1.1 Modèle(s) exact(s) à deux individus

Dans un premier temps, on se propose d'étudier "sur papier" le modèle exact dans le cas où il n'y a que deux individus, connectés entre eux. On supposera qu'initialement un des deux individus est infectieux et l'autre est susceptible, les deux individus ayant la même probabilité d'être infectieux.

2. Par exemple, ici : `http://gabgoh.github.io/COVID/index.html`.

Questions. Répondez aux questions suivantes :

1. Justifiez que le modèle proposé est bien un processus de Markov en temps discret. Déterminez les états de la chaîne correspondante et représentez le graphe de transition associé à cette chaîne.
2. Caractérisez de la manière la plus précise possible cette chaîne. Est-elle (a)périodique, irréductible, régulière, absorbante ?
3. En fixant β, μ et α respectivement à 0.5, 0.1, 0.05, calculez et tracez sur un graphe l'évolution en fonction du temps du nombre moyen d'individus dans chacune des trois catégories (S, I, et R).
4. Calculez sur base de la matrice de transition avec les mêmes valeurs de paramètres qu'au point précédent le temps moyen (en pas de temps) nécessaire à la disparition totale du virus (plus aucune personne infectieuse). Discutez (en l'illustrant) l'impact des paramètres β, μ, et α sur ce temps.
5. Considérons maintenant le cas où la première personne a en fait des contacts avec l'extérieur (par exemple au travail) et a donc une probabilité δ (qu'on supposera invariante dans le temps) d'être infectée à chaque pas de temps (si elle est susceptible).
   (a) Modifiez le graphe de transition pour prendre en compte δ et caractérisez la nouvelle chaîne de Markov ainsi obtenue.
   (b) En supposant $\delta = 0.05$, quelle pourcentage de temps chacune des deux personnes passeront-elles dans l'état infectieux en régime stationnaire ?
6. Dans le cas où $\delta = 0$, les deux individus sont indistinguables. Si on ne s'intéresse qu'à l'évolution du nombre d'individus dans chacune des catégories, S, I ou R, il est possible de modéliser le système par une chaîne de Markov alternative dont les états sont identifiés par trois variables S_t, I_t et R_t donnant le nombres d'individus dans chaque catégorie à chaque pas de temps t.
   (a) Déterminez les états de cette nouvelle chaîne de Markov et représentez son graphe de transition.
   (b) Reproduisez les expériences des sous-questions 3 et 4 avec cette chaîne et vérifiez que vous obtenez bien les mêmes résultats qu'avec la chaîne précédente.

1.2 Simulations au niveau des individus

Calculer explicitement la matrice de transition comme dans la section précédente n'est possible que pour des valeurs de N faibles, le nombre d'états devenant rapidement trop élevé. Il est néanmoins toujours possible d'estimer les mêmes courbes et statistiques que dans la section précédente en se basant sur des simulations de la chaîne de Markov. Pour faire cela, on vous demande d'écrire dans cette section un programme permettant de générer une réalisation aléatoire de la chaîne, représentée par l'évolution au cours du temps de l'état des N individus, étant donné une matrice d'adjacence W modélisant leurs contacts et des valeurs de β, μ et α fixées a priori. Sur base de réalisations, vous devrez être capables de mesurer à chaque pas de temps le nombre moyen d'individus dans les trois classes et de calculer le temps nécessaire à la disparition du virus. Pour répondre aux questions ci-dessous, une matrice d'adjacence W^{sf} (fichier `Wscalefree.txt`) vous est fournie, représentant un graphe "scale-free" défini sur 2000 individus comportant 1999 arêtes.

Questions. Répondez aux questions suivantes dans le rapport :

1. En vous mettant dans les mêmes conditions qu'aux sous-questions 3 et 4 de la section 1.1 (modèle à deux individus et mêmes valeurs de paramètres), générez un nombre suffisant de réalisations de la chaîne de Markov et reportez sur un graphe l'évolution du nombre moyen d'individus dans les trois catégories. Calculez également la moyenne du temps de disparition des individus infectieux sur ces réalisations. Vérifiez que ces résultats confirment les résultats de la section précédente.
2. Calculez les mêmes courbes pour le graphe W^{sf} fourni en utilisant des valeurs $\beta = 0.5$, $\mu = 0.2$ et $\alpha = 0.0$ et en supposant qu'initialement 0.5% des individus (au hasard) sont infectés. Faites des moyennes sur un nombre suffisant de réalisations pour que vos estimations soient stables.
3. En utilisant les autres paramètres fixés comme au point précédent, étudiez l'impact du paramètre α du modèle. Vous pouvez pour cela utiliser les deux graphes fournis. Testez des valeurs relativement faibles croissantes de α et observez l'impact sur la convergence du nombre d'infectieux. Discutez ces résultats en fonction de ce que prédit la théorie.

1.3 Simulations macroscopiques

L'utilisation de simulations numériques permet de traiter des plus grands graphes mais reste très lourde dans le cas où le nombre d'individus est très grand. Dans le cas d'un graphe complètement connecté, les individus deviennent cependant indistinguables et il est alors possible d'utiliser la même idée qu'au point 6, c'est-à-dire simuler directement l'évolution du nombre d'individus dans chaque catégorie, plutôt que de maintenir l'état de tous les individus.

Questions. Dans votre rapport, répondez aux questions suivantes :

1. Si on note S_t, I_t, et R_t le nombre d'individus dans chaque catégorie au temps t, expliquez comme générer (efficacement) de nouvelles valeurs S_{t+1}, I_{t+1} et R_{t+1} selon le modèle de propagation du virus.
2. Implémentez un simulateur de la chaîne basé sur ce principe et comparez les nombres moyens par catégorie obtenus avec ce simulateur avec ceux obtenus avec la chaîne de Markov à la section 1.1 (questions 3) et avec le simulateur de la section 1.2 (question 1).
3. Effectuez une simulation avec un graphe de taille $N = 2000$ et comparez le résultat obtenu exactement dans les mêmes conditions avec le simulateur de la section précédente. Choisissez pour cette expérience des valeurs de paramètres représentatives.

2 Propagation d'un virus dans la population belge

Dans cette seconde partie, on propose d'utiliser les mêmes outils que dans la partie 1 pour aborder un scénario un peu plus réaliste, où un épidémie se déclare et on fait appel à vous pour prédire son évolution et recommander des mesures.

2.1 Scénario et modèle

Un nouveau virus est apparu en Belgique, dont on ne sait pas a priori la dangerosité. Il a été observé il y a 25 jours pour la première fois. On a pu déterminer qu'une dizaine d'individus

Bruxellois revenant d'un voyage aux Etats-Unis l'avaient amené sur le territoire belge. Depuis lors, chaque commune belge (il y en a 589) permet aux personnes qui se sentent malades de se faire tester. Pour le moment, la ville de Liège n'est pas touchée (aucun cas positif n'a été détecté) mais le bourgmestre est inquiet et il fait appel à vous, en tant qu'expert en processus stochastiques, pour essayer d'analyser la situation au mieux.

Modèle. Pour étudier de manière "réaliste" la propagation du virus au niveau des différentes régions de la Belgique et pouvoir faire une prédiction la plus précise pour Liège, il vous semble important de prendre en compte les déplacements des individus entre villes. En cherchant sur internet, vous avez trouvé un fichier reprenant, pour chaque paire de communes belges, le nombre d'individus vivants dans une commune et travaillant dans une autre [3]. Soit C le nombre de communes belges, l'information de ce fichier peut être résumée dans une matrice $W \in \mathbb{N}^{C\times C}$, dont l'élément $W_{i,j}$ est le nombre d'individus de la commune i qui travaillent dans la commune j. On notera également N_i le nombre d'individus qui vivent dans la commune i. Pour simplifier le modèle, on ne prendra en compte que les individus qui travaillent et donc $N_i = \sum_{j=1}^{C} W_{i,j}$ et $N = \sum_{i=1}^{C} N_i$.

Sur base des idées explorées dans la première partie, le modèle simplifié de propagation du virus qu'on se propose d'implémenter est le suivant (on supposera qu'un pas de temps correspond à un jour) :

— Chaque individu infectieux au temps t a une probabilité β_1 d'infecter au temps $t+1$ chacun des individus susceptibles qui vivent dans la même commune que lui et une probabilité β_2 d'infecter au temps $t+1$ chacun des individus susceptibles qui travaillent dans la même commune que lui.

— Un individu infectieux au temps t a une probabilité μ d'être guéri et de devenir immunisé au temps $t+1$.

Comme on s'intéresse à l'évolution du virus à court terme, on supposera qu'un patient immunisé le reste indéfiniment ($\alpha = 0$). L'utilisation de deux paramètres β_1 et β_2 permet de prendre en compte des différences potentielles de taux d'infection lors du travail et dans le cercle familial ou d'amis. Le modèle fait l'hypothèse très simplificatrice que chaque individu est connecté avec l'entiereté des individus de sa commune et l'entiereté des personnes qui travaillent dans la même commune que lui. Cette hypothèse est peu réaliste mais elle vous permettra d'utiliser le modèle de la section 1.3 au niveau des sous-groupes d'individus vivant et travaillant dans les mêmes communes, ce qui devrait rendre les simulations faisables, malgré le grand nombre d'individus (plusieurs millions). L'utilisation de valeurs de β_1 et β_2 faibles devrait permettre également de prendre en compte le nombre de contacts réduit de chaque individu.

Au niveau des tests, vous faites l'hypothèse que parmi les personnes infectieuses, seuls les individus nouvellement infectés vont potentiellement se faire tester à chaque pas de temps et que le test a un taux de faux positifs nul mais un taux de faux negatifs non négligeable. Chaque individu nouvellement infecté a donc une probabilité γ d'être détecté positif, alors qu'aucun autre individu ne peut être détecté positif. Notez qu'à ce jour, ne connaissant pas la dangerosité du virus, les individus infectieux ne sont pas mis en quarantaine.

3. Le fichier en question est disponible à cette adresse : `https://statbel.fgov.be/fr/open-data/census-2011-matrice-des-deplacements-domicile-travail-par-sexe`. Il a été collecté lors d'un recensement en 2011.

Paramètres. Dans la suite, on pourra supposer que $\mu = 0.1$, $\gamma = 0.4$ et que $\beta_2 = 2\beta 1$. On sait que deux variants du virus existent mais on ne sait pas lequel est arrivé en Belgique. Le premier variant correspond à $\beta_1 = 10^{-6}$ et le second, plus virulent, à $\beta_1 = 2 \cdot 10^{-6}$.

Données. Vous disposez de la matrice mentionnée ci-dessus (fichier `WBelgium.txt` sur Ecampus). Pour réduire les temps de calcul, nous avons mis à zéro dans cette matrice toutes les valeurs $W_{i,j} < 100$. Elle concerne finalement 587 communes, 3,343,106 individus ($= \sum_{i=1}^{C} \sum_{j=1}^{C} W_{i,j}$) et contient 5757 valeurs non nulles (sur les 344569 paires de communes possibles).

Le fichier `daily_positive_tests.txt` contient pour les 25 derniers jours le nombre d'individus testés positifs dans chaque commune.

2.2 Questions

Dans votre rapport, répondez aux questions suivantes :

1. Implémentez le modèle de simulation proposé. Expliquez son principe général dans le rapport.
2. Sur base de votre simulateur, tracez les courbes d'évolution du nombre total moyen (sur 25 simulations au moins) d'individus dans chaque catégorie en fonction du temps, sur la Belgique entière et puis dans les communes suivantes : Bruxelles, Anvers et Liège. Utilisez les valeurs de paramètres par défaut mentionnées ci-dessus (pour l'un des variants au choix). Ajoutez sur vos courbes un intervalle de confiance à 95% autour de cette moyenne. Discutez ces courbes.
3. A partir du nombre d'individus testés positifs dans chaque commune, déterminez quel variant du virus est le plus probablement présent en Belgique. Justifiez votre réponse.
4. Tracez les mêmes courbes qu'au point 2 pour le bon variant si ce n'est pas celui que vous aviez utilisé précédemment. Informez le bourgmestre de Liège du nombre maximum de patients infectés auquel il peut s'attendre et dites-lui quand le pic devrait avoir lieu.
5. On se rend compte que la maladie est finalement plus grave que prévu et qu'un pourcentage non négligeable de personnes infectieuses doivent être hospitalisées en soins intensifs. Votre analyse précédente ayant convaincu le bourgmestre de Liège, il vous recommande pour intégrer une task force fédérale chargée de déterminer les mesures les moins contraignantes possibles permettant d'éviter la saturation des soins intensifs. On estime qu'approximativement 10% des personnes infectées doivent être hospitalisées et que chaque commune dispose de 400 lits en soins intensifs par 100.000 habitants. Sur base de ces contraintes, proposez, modélisez et étudiez l'impact d'au moins deux mesures telles que par exemple :
   - La mise en quarantaine des personnes testées positives.
   - Le port du masque.
   - Un confinement des travailleurs.
   - La vaccination.

   Dans chaque cas, identifiez un paramètre pertinent modélisant le degré d'application ou l'impact de la mesure (par exemple, le pourcentage de personnes qui n'iront pas travailler ou le nombre de personnes à vacciner par jour) et déterminez la valeur minimale de ce

paramètre pour que l'application de la mesure permette d'atteindre l'objectif (si c'est possible). Vous pouvez concentrer votre analyse sur le niveau national et/ou sur les communes qui seront les plus touchées d'après votre modèle.

3 Resources

L'énoncé du projet et les fichiers mentionnés ci-dessus sont disponibles sur Ecampus (rubrique "Projet"). Les fichiers `Communecoord.txt` et `Communename.txt` donnent respectivement les noms et les coordonnées des communes belges (à des fins de visualisation par exemple).

4 Soumission

Vous devez nous fournir un rapport *au format pdf* contenant vos réponses, concises mais précises, aux questions posées ainsi que le code que vous avez utilisé pour y répondre, le tout sous forme d'archive zip. L'archive doit être soumise sur la plateforme de soumission de Montefiore pour **le vendredi 7 mai 2021 à 23h59** au plus tard.

Pour vos expériences, vous pouvez utiliser le langage de programmation que vous souhaitez (avec notre accord, si ce n'est pas matlab, Mathematica, R, Python, Java ou C). Quel que soit le langage, on vous demande d'implémenter les fonctions de simulations par vous-mêmes. Vous ne pouvez pas utiliser une boîte à outils existante qui ferait exactement ce qui est demandé. En cas de doute, contactez le professeur.

Références

[1] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.*, 10(4), January 2008.