Projet de MAP568

janvier-mars 2022

Josselin Garnier (Ecole polytechnique)

à remettre pour le 18 mars 2022

1 Introduction

Le but de ce projet est de conduire une analyse de quantification des incertitudes (y compris d'analyse de sensibilité) sur un modèle complexe, en utilisant les outils présentés dans le cours. L'étude d'un modèle épidémiologique compartimental (simplfié) est apparue comme un choix naturel pour le cours de cette année.

Le modèle SEIR (Susceptible-Exposed-Infected-Recovered) que nous présentons est un modèle compartimental déterministe connu depuis près d'un siècle et utilisé en épidémiologie depuis les années 80. C'est aussi celui utilisé par les épidémiologues étudiant la COVID-19, à l'Imperial College (équipe de Neil Ferguson [2]), à l'Institut Pasteur (équipe de Simon Cauchemez [4]) ou à l'INSERM (équipe de Vittoria Colizza [3]) par exemple. Ce modèle se présente sous la forme d'un système d'équations différentielles ordinaires couplées qui donnent l'évolution du nombre d'individus susceptibles (d'être infectés), du nombre d'individus exposés mais pas encore contagieux, du nombre d'individus infectés et contagieux, etc. Pour ceux que cela intéresse, vous pouvez par exemple regarder l'article [1] : ce modèle peut être considéré comme une approximation du comportement moyen d'un modèle compartimental stochastique si les périodes infectieuses sont supposées être i.i.d. avec une distribution exponentielle et si la population est supposée homogène à l'intérieur de chaque compartiment.

Afin d'être représentatif de cette épidémie, certaines caractéristiques spécifiques de la COVID-19 ont été prises en compte (présence d'individus asymptomatiques par exemple) et nous proposons donc d'étudier ici un modèle SEAIR (Susceptibles, Exposés, Asymptomatiques, Infectés, Rétablis). Nous ne considérerons pas un modèle complet comprenant stratification régionale et par âge, mais nous allons cependant exploiter des données d'hospitalisations recueillies par région.

2 Modèle SEIAR

2.1 Les compartiments

— S : les individus Susceptibles (d'être infectés). Durant l'évolution de l'épidémie, ils peuvent rester dans ce compartiment ou être infectés et passer dans le compartiment E.
— E : les individus Exposés. Ils sont en phase d'incubation et ne sont pas encore contagieux. Ils passent ensuite dans les compartiments A ou I.
— I : les individus Infectés (symptomatiques). Ils sont contagieux. Ils passent ensuite dans le compartiment R.
— A : les individus infectés Asymptomatiques. Ils sont contagieux (mais un peu moins que les individus du compartiment I). Ils passent ensuite dans le compartiment R.
— R : les individus rétablis, admis à l'hôpital, ou morts. Ils ne sont plus contagieux.

2.2 L'évolution de l'épidémie

L'évolution de l'épidémie est régie par le système d'équations différentielles ordinaires suivant (dans lequel l'unité de temps est le jour) qui donne le "nombre" d'individus présents dans chaque compartiment :

$$\frac{dS}{dt} = -\frac{\beta(t)}{N} S(I + \mu A), \tag{1}$$

$$\frac{dE}{dt} = \frac{\beta(t)}{N} S(I + \mu A) - \frac{E}{T_E}, \tag{2}$$

$$\frac{dI}{dt} = \frac{\alpha}{T_E} E - \frac{I}{T_I}, \tag{3}$$

$$\frac{dA}{dt} = \frac{1-\alpha}{T_E} E - \frac{A}{T_I}, \tag{4}$$

$$\frac{dR}{dt} = \frac{I + A}{T_I}. \tag{5}$$

Eqs. (1-2) : des individus Susceptibles deviennent Exposés par contamination par des individus Infectés symptomatiques ou Asymptomatiques.
Eqs. (2-3-4) : les individus Exposés passent en moyenne un temps T_E dans cette phase, et deviennent alors Infectés symptomatiques avec probabilité α ou Asymptomatiques avec probabilité $1 - \alpha$.
Eqs. (3-4-5) : les individus Infectés symptomatiques ou Asymptomatiques passent en moyenne un temps T_I dans cette phase, puis ils rentrent dans le compartiment final R.

Ces “nombres” sont en fait des réels dans le modèle. Ce système conserve la taille de la population :

$$N = S + E + I + A + R = \text{constante}.$$

Un dernier “compartiment” H est ajouté à des fins de calibration (voir plus loin). C'est le nombre d'hospitalisations journalières :

$$H(t) = \int_{t-1}^{t} \gamma \frac{I(s)}{T_I} ds, \tag{6}$$

qu'on peut prendre égal à $\gamma \frac{I(t)}{T_I}$ pour simplifier. Ce n'est pas un vrai compartiment, mais il est intéressant car *ce nombre est observé* et peut donc être utilisé à des fins de calibration.

Les conditions initiales du système sont données en termes de la date initiale t_0 et du nombre initial d'individus exposés E_0 à cette date. On fixe t_0 au 4 février (on agira sur le E_0 dans la calibration, pas sur le t_0). N est la taille de la population générale ($N = 67\,10^6$ pour la France). Les autres compartiments sont vides à t_0.

2.3 Les paramètres

— Le taux de transmission β est de la forme :

$$\beta(t) = \beta_0 \mathbf{1}_{t \le t_1} + \beta_0 \Big(1 - (1-f)\frac{t-t_1}{t_2}\Big) \mathbf{1}_{t_1 < t \le t_1+t_2} + \beta_0 f \mathbf{1}_{t_1+t_2 < t}.$$

L'hypothèse est qu'avant toute mesure de confinement, le taux de transmission est constant et décroît ensuite lorsque les mesures sont prises. On modélise ce comportement par une fonction continue et affine par morceaux, qui vaut β_0 avant t_1 et $f\beta_0$ après $t_1 + t_2$.

— β_0 : taux de transmission avant confinement.
— f : réduction effective du taux de transmission par confinement.
— μ : facteur de diminution de la contagiosité des asymptomatiques par rapport aux symptomatiques.
— α : proportion des individus symptomatiques parmi les infectés.
— T_E : temps moyen passé en phase d'incubation.
— T_I : temps moyen passé en phase infectieuse (contagieuse).
— γ : probabilité pour un infecté symptomatique de devoir aller à l'hôpital. Seuls les symptomatiques vont (éventuellement) à l'hôpital.

Remarques : 1) Le R_0 (le nombre moyen de contaminations par un individu infecté -symptomatique ou asymptomatique- lorsqu'on laisse l'épidémie se propager naturellement, en début d'épidémie) est

$$R_0 = (\alpha + (1-\alpha)\mu)\beta_0 T_I$$

Paramètre	Description	Loi
β_0	taux de transmission initial	$\beta_0 \sim \mathcal{U}(0.7, 1.1)$
f	réduction effective du taux de transmission par le confinement	$f \sim \mathcal{U}(0.1, 0.5)$
E_0	nombre d'exposés à t_0	$\ln E_0 \sim \mathcal{U}(-\ln 10, 2\ln 10)$
t_1	début des effets du confinement	$t_1 \sim \mathcal{U}(12 \text{ mars}, 22 \text{ mars})$
t_2	durée de mise place des effets du confinement	$t_2 \sim \mathcal{U}(4, 20)$ (jours)
α	proportion des symptomatiques parmi les infectés	$\alpha \sim \mathcal{U}(0.3, 0.7)$
μ	facteur de réduction de la contagiosité des asymptomatiques	$\mu \sim \mathcal{U}(0.3, 0.7)$
T_E	durée moyenne de la phase d'incubation	$T_E \sim \mathcal{U}(2.5, 3.5)$ (jours)
T_I	durée moyenne de la phase d'infection (contagieuse)	$T_I \sim \mathcal{U}(4, 5)$ (jours)
γ	probabilité pour un infecté symptomatique de devoir aller à l'hôpital	$\ln \gamma \sim \ln 0.03 + \mathcal{U}(-\ln 4, \ln 4)$

TABLE 1 – Lois des 10 paramètres d'entrée (les durées sont en jours).

2) On rappelle que les trois premiers cas (isolés) de covid-19 ont été détectés en France le 24 janvier. Après une série de différentes mesures, le confinement total a commencé le 17 mars et s'est achevé le 11 mai.

Les 10 paramètres $\beta_0, f, t_1, t_2, \mu, \alpha, T_E, T_I, \gamma, E_0$, constituent les paramètres d'entrée incertains du modèle. Lorsque ces 10 paramètres sont fixés, on obtient une trajectoire des variables de sortie S, E, I, A, R, H par résolution du système (1-6).

3 Propagation d'incertitudes

Les lois des 10 paramètres d'entrée (qu'on regroupe dans le vecteur d'entrée $\boldsymbol{x}$ dans la suite) sont données dans la table 1. On suppose les paramètres indépendants.

Question 1 : Programmez la résolution du système (1-6). Il faudra extraire les valeurs de S, E, ..., H, aux instants entiers. Il faudra faire un post-traitement pour extraire le nombre maximal d'hospitalisations journalières et la proportion finale (au 17 mai) de susceptibles.

Question 2 : Par échantillonnage Monte Carlo (en utilisant les lois des paramètres d'entrée de la table 1), donnez :
- la loi du nombre maximal d'hospitalisations journalières (histogramme, moyenne, quantiles à 10% et 90%),
- la loi de la proportion finale de susceptibles.

4 Analyse de sensibilité

Question 3 : Faire une analyse de sensibilité pour les deux variables de sortie de la question 2 avec la méthode de Sobol (estimer les indices du premier ordre et les indices totaux).

Dans la suite, on regardera avec plus d'attention les sept paramètres suivants : β_0, f, E_0, t_1, t_2, γ et on fige les autres : $T_E = 3$, $T_I = 5$, $\mu = 0.5$, $\alpha = 0.5$.

5 Calibration

Pour calibrer le modèle, on va regarder les données collectées pendant la première vague. On va juste utiliser les données les plus fiables (les hospitalisations), car d'autres données (nombres de cas détectés, etc) sont plus difficiles à interpréter.

On collecte les données sur les hospitalisations depuis le site
`https://www.data.gouv.fr/fr/organizations/sante-publique-france/`
dans la base SIVIC qui fournit des données à partir de mars 2020. Téléchargez le fichier donnees-hospitalieres-classe-age-hebdo-covid19....csv qui contient des données sur les admissions hospitalières hebdomadaires (admissions journalières cumulées sur les 7 derniers jours) stratifiées par région et par âge, on fusionne les classes d'âge (on peut en fait prendre la classe d'âge "0", qui regroupe toutes les classes d'âge "9"=0-9 ans, "9"=10-19 ans, etc). Les codes des régions sont sur wikipedia ("11"=Ile-de-France, etc).

On trouve les populations des régions sur
`https://www.insee.fr/fr/statistiques/1893198/`

Question 4 : Dessiner les données collectées du 1er mars au 17 mai (vous devez trouver quelque chose qui ressemble à la figure 1).

5.1 Calibration déterministe

On se fixe une région, ou bien on prend la France tout entière. On note s_i, $i = 1, \ldots, n_{H7}$ (normalement, vous devez avoir $n_{H7} = 12$) les dates où les admissions hospitalières hebdomadaires sont recueillies. On cherche à ajuster au mieux le modèle

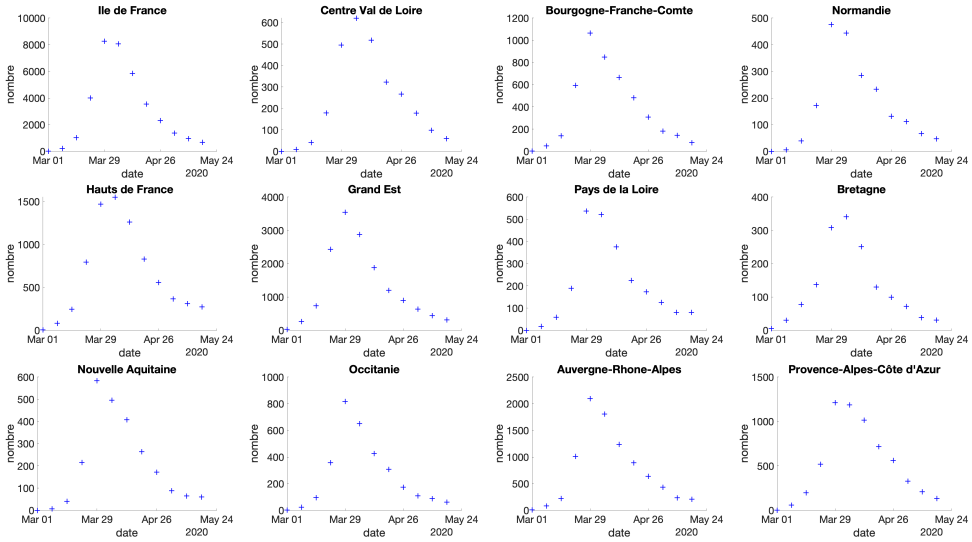

FIGURE 1 – Visualisation des données : admissions hospitalières hebdomadaires.

au sens des moindres carrés :

$$\boldsymbol{x}^* = \mathrm{argmin}_{\boldsymbol{x}} \mathcal{E}(\boldsymbol{x}),$$
$$\mathcal{E}(\boldsymbol{x}) = \sum_{i=1}^{n_{H7}} f_{H7}(\boldsymbol{x}, s_i)^2,$$

où les résidus sont définis par :

$$f_{H7}(\boldsymbol{x}, s_i) = \big(\mathcal{M}_{H7}(\boldsymbol{x}, s_i) - \mathrm{data}_{H7}(s_i)\big)/\sqrt{\mathcal{M}_{H7}(\boldsymbol{x}, s_i)},$$

$\mathcal{M}_{H7}(\boldsymbol{x}, t)$ est la prédiction des admissions hospitalières hebdomadaires à l'instant t du modèle SEIAR pour la région en question avec les paramètres $\boldsymbol{x}$ (on explique dans la prochaine section le choix de la normalisation en racine carrée), et $\mathrm{data}_{H7}(s)$ sont les données de la base SIVIC.

Question 5 : Choisissez une région ou la France tout entière et évaluez numériquement $\boldsymbol{x}^*$ dans le domaine de $\mathbb{R}^6$ déterminé par le support des lois a priori (si vous prenez la France tout entière, mulitpliez les bornes du domaine de E_0 par 10). Comparez sur une figure les données et les prédictions $\mathcal{M}_{H7}(\boldsymbol{x}^*, \cdot)$. Donnez les valeurs R_0^* et γ^* ainsi obtenues.
Remarques : Il y a des routines d'optimisation dans python ! Attention, la fonction $\mathcal{E}$ peut posséder des minima locaux !

5.2 Calibration bayésienne

Le modèle statistique est le suivant :

- On connaît la loi a priori des paramètres $\boldsymbol{x} \in \mathbb{R}^6$ (voir table 1).

- Pour évaluer la vraisemblance, on suppose un modèle statistique de la forme

$$\text{data}_{H7}(s_i) = \mathcal{M}_{H7}(\boldsymbol{x}, s_i) + \sqrt{\mathcal{M}_{H7}(\boldsymbol{x}, s_i)}\epsilon_{H7,i},$$

où $\epsilon_{H7,i} \sim \mathcal{N}(0, \sigma^2)$ sont indépendants en i. Ce modèle statistique est obtenu par approximation d'un modèle dans lequel on suppose que les $\text{data}_{H7}(s_i)$ sont des réalisations de loi de Poisson de paramètre $\mathcal{M}_{H7}(\boldsymbol{x}, s_i)$ et on approche une loi de Poisson de paramètre M (grand) par une loi $\mathcal{N}(M, M)$. On aurait alors $\sigma = 1$, mais on prend en compte en plus une erreur de modèle qui a la même structure que l'erreur d'observation, d'où le σ. On obtient ainsi une expression de la vraisemblance $p(\mathbf{data}|\boldsymbol{x}, \sigma)$, $\mathbf{data} = (\text{data}_{H7}(s_i))_{i=1}^{n_{H7}}$:

$$p(\mathbf{data}|\boldsymbol{x}, \sigma) = \frac{1}{(2\pi)^{n_{H7}/2}\sigma^{n_{H7}}} \exp\Big[-\frac{1}{2}\sum_{i=1}^{n_{H7}} \frac{f_{H7}(\boldsymbol{x}, s_i)^2}{\sigma^2}\Big].$$

Le point $\boldsymbol{x}^*$ obtenu dans la calibration déterministe est le maximum de vraisemblance lorsque σ est fixé.
On détermine l'hyper-paramètre σ par une méthode du maximum de vraisemblance : on fixe la valeur de σ à σ^* telle que la vraisemblance $p(\mathbf{data}|\boldsymbol{x}^*, \sigma)$ est maximale.

Question 6 : Vérifiez qu'on a $(\sigma^*)^2 = \frac{1}{n_{H7}} \sum_{i=1}^{n_{H7}} f_{H7}(\boldsymbol{x}^*, s_i)^2$.

La loi a posteriori de $\boldsymbol{x}$ n'a pas d'expression explicite puisqu'elle implique des appels au modèle compartimental dans la vraisemblance. Par conséquent, nous devons recourir à des algorithmes d'échantillonnage. Ici, nous suggérons d'utiliser un algorithme de Metropolis-Hastings.

Question 7 : Générez un échantillon de la loi a posteriori des paramètres $\boldsymbol{x}$. Tracez des histogrammes des lois a posteriori de R_0 et de γ.

Question 8 : Calibrez avec toutes les données des 12 régions de France métropolitaine. On négligera les communications entre les régions. Dans ces conditions, le E_0 dépend de la région et on peut supposer que les autres paramètres sont communs, si bien que le $\boldsymbol{x}$ devient de dimension $5 + 12 = 17$.
Les deux hypothèses de communications nulles et de paramètres communs sont discutables...

Références

[1] F. Brauer, Mathematical epidemiology : Past, present, and future, Infectious Disease Modelling, Vol. 2, pp. 113–127 (2017). [Modèles compartimentaux] 1

[2] Imperial College COVID-19 Response Team (16 March 2020), "Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand" ; Imperial College COVID-19 Response Team (30 March 2020), "Estimating the number of infections and the impact of nonpharmaceutical interventions on COVID-19 in 11 European countries". Cf http ://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/ [Modélisations de la covid-19, mars 2020] 1

[3] L. Di Domenico et al., Impact of lockdown on COVID-19 epidemic in Ile-de-France and possible exit strategies, BMC Medicine, Vol. 18, 240 (2020) [Modélisations de la covid-19, avril 2020] 1

[4] H. Salje et al., Estimating the burden of SARS-CoV-2 in France, Science, Vol. 369, Issue 6500, pp. 208-211 [Modélisations de la covid-19, mai 2020] 1