Confidence interval on SEIR predictions

4/06/2024

Résumé

My abstract

1 Introduction

This is the introduction section. Here you can write the background of your study, the purpose of your research, and so on.

2 Methods

The data

As the goal of this study is to compare some forecasts on many different pandemics, we need to be able to generate synthetic pandemics, with a particular attention on the diversity of the pandemics generated.

Covasim

To generate the pandemics, we used Covasim [2], a python librairy that can simulate the evolution of a pandemic.

Covasim is an agent-based model that can model many different pandemics and has a high diversity of outputs.

This model takes as an input many parameters such as the population type, the population size, the age repartition ...and outputs a complete description of the pandemic, with real-time values of each relevant information, sucha as the number of severe, of asymptomatic... but also physical values such as the value of the reproduction number.

Covasim enables to generate a huge diversity of pandemics, thanks to the plurality of parameters that can be given as the input of the model, but also with interventions that can be planned by the users.

These interventions can simulate the impact of a vaccination campaign, with changes in the probability transmission, that can be different for all ages groups.

For the implementation and the first test of our models, we generated two pandemics. The first one focusing on the new deaths count and the second one focusing on the number of hospitalized count.

We will refer to these pandemics as the pandemic 1 and pandemic 2.

The parameters used to generate these pandemics are described in the table 1.

The parameters that are not specified are the default parameters of the Covasim librairy.

The different interventions were based on mobility reports from Västtrafik 1, the public transport company of the city of Gothenburg, which were reported during the Covid 19 pandemic and has been retrieved in [1].

These correspond to 53 relative weekly variations of the mobility, with a reference value of 1 for the first week of the report, which correspond to the 9-th week of 2020.

TABLE 1 – Table of the parameters used to simulate the firrst two pandemics.

Parameter	**Pandemic 1**	**Pandemic 2**
start day 1	2020-03-02	2020-03-02
end day	2020-07-01	2021-01-01
Population size	1000000	1000000
Interventions	interventions1	interventions2
population type	hybrid	hybrid
β initial	0.015	0.015
location	Sweden	Sweden
n infected initial	20	100

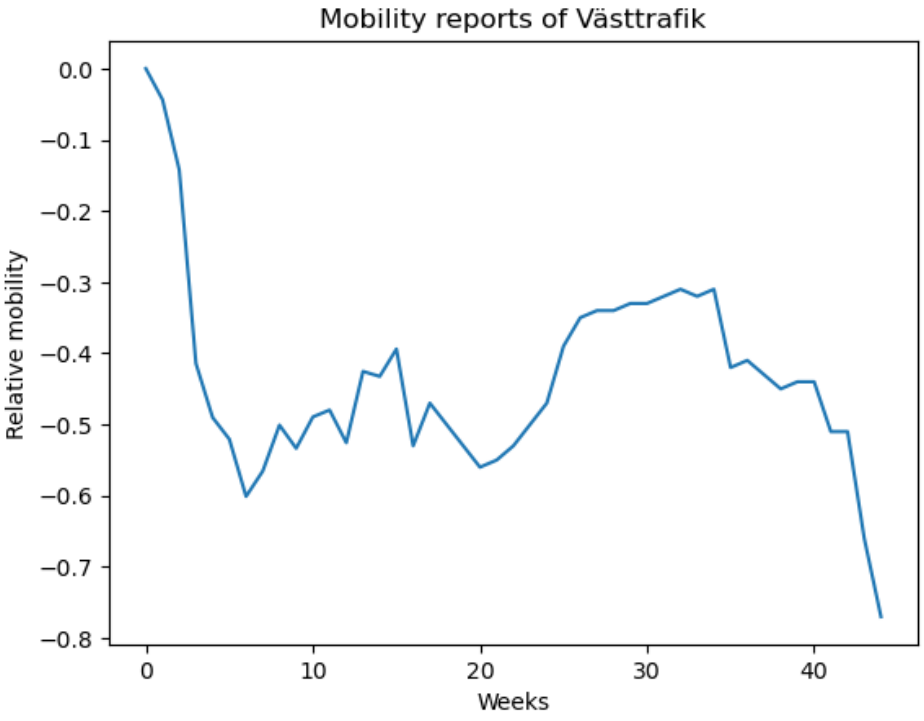

FIGURE 1 – Mobility reports from Västtrafik.

2.1 Computing confidence intervals on the prediction

Assumption :

We suppose that the data of the pandemic observed follows the model h, of parameter $\theta^* \in \mathbb{R}^d$. Let Y_i, $i = 1, \ldots, n$ be the number of hospitalized at each day. We suppose that : $Y_i = h_{\theta^*}(i) + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, iid, and independent from all the other variables. The objective is to estimate θ^*. We use $\hat{\theta}$, the least square estimator of θ^* as an estimator of θ^* :

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - h_\theta(i))^2$$

Let :

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$h_\theta = \begin{pmatrix} h_\theta(1) \\ \vdots \\ h_\theta(n) \end{pmatrix}$$

We have :

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \|Y - h_\theta\|^2$$

Now, if θ is close enough to θ^*, we can write :

$$\forall i \in \{1, ..., n\} : h_\theta(i) = h_{\theta^*}(i) + (\theta - \theta^*)^T \nabla_\theta h_{\theta^*}(i)$$

which leads to :

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\|Y - h_{\theta^*} - (\theta - \theta^*)^T \nabla_\theta h_{\theta^*}\right\|^2$$

Let us define :

$$\tilde{Y} = Y - h_{\theta^*}$$
$$\beta = \theta - \theta^*$$
$$\hat{\beta} = \theta - \hat{\theta}$$

and let us define the matrix $A \in \mathbb{R}^{n \times d}$ such that $\forall i \in \{1, ..., n\}, \forall j \in \{1, ..., d\}, A_{i,j} = \frac{dh_{\theta^*}}{d\theta_j}(i)$.
The previous problem can be re-written as :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \left\|\tilde{Y} - A\beta\right\|^2$$

This is a regression linear problem.
Let us solve this problem in the general case.
Let $(A_i, \tilde{Y}_i)$ be the observations Let $\mathbb{P}$ be the law of A_i, and let us assume that $Y_i = A_i\beta^* + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
The solution of this problem is explicitly :

$$\hat{\beta} = (A^T A)^{-1} A^T \tilde{Y}$$

This least-square estimator is unbiased :

$$\mathbb{E}[\hat{\beta}] = \beta^*$$

$$\hat{\beta} = \left(\sum_{i=1}^{n} A_i^T A_i\right)^{-1} \times \left(\sum_{i=1}^{n} A_i^T \tilde{Y}_i\right)$$

$$\hat{\beta} = \frac{n}{n}\left(\sum_{i=1}^{n} A_i^T A_i\right)^{-1} \times \left(\sum_{i=1}^{n} A_i^T \tilde{Y}_i\right)$$

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} A_i^T A_i\right)^{-1} \times \left(\frac{1}{n}\sum_{i=1}^{n} A_i^T \tilde{Y}_i\right)$$

Let us denote :

$$\hat{D} = \frac{1}{n}\sum_{i=1}^{n} A_i^T A_i, \quad \text{and} \quad \hat{\delta} = \left(\frac{1}{n}\sum_{i=1}^{n} A_i^T \tilde{Y}_i\right)$$

We have :

$$\hat{\beta} = \hat{D}^{-1}\hat{\delta}$$

$$\hat{D} \underset{a.s}{\to} D = \mathbb{E}[A_i^T A_i]$$

$$\hat{\delta} \underset{a.s}{\to} \delta = \mathbb{E}[A_i^T \tilde{Y}_i]$$

$\hat{\beta} = \hat{D}^{-1}\hat{\delta} \underset{a.s}{\to} D^{-1}\delta$, as the following function ϕ is continuous :

$$\phi : \left\{ \begin{array}{ccc} \mathcal{GL}_n(\mathbb{R}) & \to & \mathcal{GL}_n(\mathbb{R}) \\ A & \mapsto & A^{-1} \end{array} \right.$$

Now, let us show that $\hat{\beta}$ is asymptotically normal :

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta^*) &= \sqrt{n}(\hat{D}^{-1}\hat{\delta} - \beta^*) \\
&= \sqrt{n}(\hat{D}^{-1}\hat{\delta} - \hat{D}^{-1}\hat{D}\beta^*) \\
&= \sqrt{n}\hat{D}^{-1}(\hat{\delta} - \hat{D}\beta^*) \\
&= \sqrt{n}\hat{D}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} A_i^T \tilde{Y}_i - \frac{1}{n}\sum_{i=1}^{n} A_i^T A_i \beta^*\right) \\
&= \frac{\sqrt{n}}{n}\hat{D}^{-1}\left(\sum_{i=1}^{n} A_i^T (\tilde{Y}_i - A_i\beta^*)\right) \\
&= \frac{1}{\sqrt{n}}\hat{D}^{-1}\left(\sum_{i=1}^{n} A_i^T \epsilon_i\right)
\end{aligned}$$

This line is made of two terms. Let's show that each one of them converges in law.

$$\begin{aligned}
\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} A_i^T \epsilon_i'\right) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} A_i^T \epsilon_i'\right) \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} A_i^T \epsilon_i' - 0\right) \\
&\xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}(A_i^T \epsilon_i))
\end{aligned}$$

Yet, as ϵ_i and A_i are independant, and $\mathbb{E}[A_i^T \epsilon_i'] = 0$, $\text{Var}(A_i^T \epsilon_i) = \mathbb{E}[A_i A_i^T \epsilon_i^2] = \mathbb{E}[A_i A_i^T]\sigma'^2$.
Finally, $\frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} A_i^T \epsilon_i' \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, D\sigma'^2)$.
On the other hand, $\hat{D}^{-1} \xrightarrow{\mathcal{L}} D^{-1}$, which is constant.
Finally, with Slutsky, we obtain that :

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta^*) &\xrightarrow{\mathcal{L}} D^{-1}\mathcal{N}(0, D\sigma'^2) \\
&\xrightarrow{\mathcal{L}} \mathcal{N}(0, D^{-1}(D\sigma'^2)(D^{-1})^T) \\
&\xrightarrow{\mathcal{L}} \mathcal{N}(0, D^{-1}D\sigma'^2(D^{-1})^T) \\
&\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma'^2 D^{-1}) \\
&\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma'^2 (A^T A)^{-1})
\end{aligned}$$

Let's get back to the first problem :
As $\beta^* = 0$ and $\hat{\beta} = \hat{\theta} - \theta^*$, we have :

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 (A^T A)^{-1})$$

and,

$$\hat{\theta} \sim \mathcal{N}(\theta^*, \frac{\sigma^2}{n} (A^T A)^{-1})$$

As a first conclusion, we have that $\hat{\theta}$ is asymptotically normal.
Let Σ be the covariance matrix estimated from the computation of $\hat{\theta}$. In our case, $\Sigma = \frac{\sigma^2}{n}(A^T A)^{-1}$.
As $\hat{\theta}$ is asymptotically normal, we can apply the delta-method :

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\ \Sigma)$$

$$\sqrt{n}(h_{\hat{\theta}} - h_{\theta^*}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \nabla_\theta h_\theta^T \Sigma \nabla_\theta h_\theta)$$

And finally :

$$h_{\hat{\theta}} \to \mathcal{N}(h_{\theta^*}, \frac{1}{n} \nabla_\theta h_\theta^T \Sigma \nabla_\theta h_\theta)$$

By estimating $\frac{1}{n}\Sigma$ from `curve_fit`, we can compute the confidence interval of the prediction with the quantiles of the normal distribution.

Results

Discussion

Conclusion

Références

[1] Philip Gerlee et al. "Predicting regional COVID-19 hospital admissions in Sweden using mobility data". In : *Scientific reports* 11.1 (2021), p. 24171.

[2] Cliff C Kerr et al. "Covasim : an agent-based model of COVID-19 dynamics and interventions". In : *PLOS Computational Biology* 17.7 (2021), e1009149.