Supplementing local R_t estimation during periods of geographical dropout

Draft outline

Last revised: December 2, 2024

1 Introduction

- General intro to Rt estimation
- Discussion of real-time
- Emphasis on "local" methodologies
- Extension to estimation across many locations simultaneously
- Not all locations have the same reporting frequency
- High-level description of the methods and a "killer figure" that shows what we do.

2 Methods

Bayesian local-global methodology

We assume $p \geq 2$ locations (the local level) compose a region (the global level) with random mixing within each location. We do not model interactions among locations.

- Let $I_{t,\ell}$ be incidence in location $1 \leq \ell \leq p$ at time t (usually in days).
- Let I_t be the incidence over the entire region with $I_t = \sum_\ell y_{t,\ell}$.

Consequently, we can employ a renewal model to estimate R_t from time $t = 1, \ldots, T + H$ as:

$$\mathbb{P}\left(R_t \mid I_1^t\right) \text{ given } I_t \sim \text{Noise}\left(R_t \Lambda_t\right) \text{ and } \Lambda_t = \sum_{i=1}^{t-1} \omega_i I_{t-i}$$

Λ_t is the total infectiousness in the region, the weights ω_i compose the generation time distribution of the disease and $\sum_{i=1}^{\infty} \omega_i = 1$. 'Noise' is some noise distribution of choice.
We consider that there is dropout i.e., a loss of data, in location ℓ from time $t = T$.

- All locations have the same generation time distribution and $\Lambda_{t,\ell} = \sum_{i=1}^{t} \omega_i I_{t-i,\ell}$.
- The location with dropout has unknown $I_{\ell,t}$ for $T \leq t \leq T + H$.
- Local-level renewal models provide the posterior $\mathbb{P}\left(R_{t,\ell} \mid I_{1,\ell}^{t}\right)$ only up to $t = T$.

There are two options for providing estimates during the dropout time period. For any time in this period $1 \leq h \leq H$ we can either:

1. Baseline: project from our last local posterior $\mathbb{P}\left(R_{T,\ell} \mid I_{1,\ell}^{T}\right)$ to time $T + h$.
2. Correction: borrow information from the global posteriors $\mathbb{P}\left(R_{T+h} \mid I_{1}^{T+h}\right)$.

Our proposed correction uses the last local posterior as a prior and then draws information from the global posteriors (which exclude ℓ) to update this into the posterior:

$$\mathbb{P}\left(R_{T+1,\ell} \,\middle|\, I_{1,\ell}^{T}, \sum_{m \neq \ell} I_{T+1,m}\right) \propto \mathbb{P}\left(R_{T+1,\ell} \,\middle|\, \sum_{m \neq \ell} I_{T+1,m}\right) \mathbb{P}\left(R_{T+1,\ell} \mid I_{1,\ell}^{T}\right)$$

This follows due to the conditional independence of the infections in locations. Later time steps are obtained by iterating this process sequentially (with the prior reset to the dropout posterior from the last time step). We have not yet implemented this full version and so test a 'deterministic dropout correction' via the following sequence:
Input $\{R_t\}_{t=1}^{T+H}$, $\{R_{t,\ell}\}_{t=1}^{T}$, $\Lambda_{\ell,t}$. For $h = 1, \ldots, H$, do

1. Predict $\hat{I}_{T+h,\ell} = R_{T+h}\Lambda_{T+h,\ell}$ (expectations from renewal models).
2. Convolve $\Lambda_{T+h,\ell} = \sum_{i=1}^{T+h-1} \omega_i \tilde{I}_{T+h-i}$, where $\tilde{I}_j = I_j \mathbb{1}[j \leq T] + \hat{I}_j \mathbb{1}[j > T]$.
3. Estimate local $\hat{R}_{T+h,\ell} = \epsilon_h R_{T+h} \frac{\Lambda_{T+h,\ell}}{\Lambda_{T+h+1,\ell}}$, with ϵ_h as a factor we currently set to 1.

Current and suggested implementations

The above deterministic algorithm was modified during actual implementation as follows:

- It is not necessary to sequentially compute $\hat{R}_{T+h,\ell}$ and this procedure ignores it.
- Because location ℓ is unavailable for $t = T + 1, \ldots, T + H$, we also cannot compute Λ_t for $t = T + 2, \ldots, T + H$. For $h \geq 1$ we used the correction $1 - \Lambda_{\ell,t}/\Lambda_t$, to rescale regional incidence (and convolved incidence).
- Once we have the sequence $\{\hat{I}_{\ell,t}\}$ for $t = 1, \ldots, T + H$, we simply input this into the local R_t estimation routine as pseudodata.

We are predicting $\hat{I}_{T+h,\ell}$ using R_{T+h} and $\Lambda_{T+h,\ell}$, both of which are smooth. Then we reconvolve and iterate. It may be more productive to view this as a process with 2 modules, a `forecaster` and an `Rtestimator`. Then the meta procedure is

1. Use `forecaster` to produce $\{\Lambda_{T+h,\ell}\}$ for $2 \leq h \leq H$. We do this conditional on all available information (incidence at the location, the region, or some related quantity).

2. We can supplement these forecasts with auxiliary signals (wastewater, deaths etc.). Currently, we have a minimal `forecaster` (uses global R_t and local convolved incidence).

3. Calculate $\hat{I}_{T+h,\ell}$ by taking the first (backward) differences of $\{\Lambda_{T+h,\ell}\}$.

4. Use `Rtestimator` once across the pseudo time series to produce $\{R_{t,\ell}\}_{t=1}^{T+H}$

Potential paper outline

This study could combine the algorithm maybe as a small R function that users can apply to data, or equivalently we can setup the `forecaster` module such that it offers outputs that can simply be integrated within user-preferred `Rtestimator`modules. There are several questions to investigate that could complete the paper.

- What is the value of combining local and global information relative to projecting forward locally? When should this work?

- How should we define the global scale? Perhaps pick the areas with a history of having similarly synched epidemics? Nearest neightbours?

- Given p locations in a region, how many can we tolerate simulatanous dropout in? What about other patterns? How long (relative to the generation time) does dropout need to be before there is no chance of correction?

- Can we inform the additional parameter ϵ_h in some way to improve the correction? Does this require other data e.g., mobility?

Practically, we could combine some more simulated examples using a combination of the `rtestim` and `EpiFilter` code that we already have developed that can provide insights into when and how global information can supplement local data scarcity as well as prvide guidelines for when this global information will not work. This could be about 2 figures. Then about 3 figues with the more impactful content that uses real examples, perhaps from the CFA pipeline pr the other data sources we discussed. As baselines we can always compare to local forecasting and hold out real data to use for validation.