Bayesian inference of reporting delays based on line-list data

Tenglong Li

6/22/2020

Approach 1: Adjusting the epidemic curve based on the distribution of report delays

Introduction

Delayed reporting is common during the Covid-19 pandemic we are experiencing now. In this project, we are trying to develop analytical framework for adjusting report delays and correcting the epidemic curve. The central task is to derive the distribution of report delays based on known report delays, and this further require imputation of missing report delays as well. We offer a Bayesian framework which allows us to impute the missing report delays and distribution of report delays simultaneously. Based on the MCMC estimates of distributional parameters of report delays, we can then do back-calculation based on the reporting case numbers and nowcasting.

List of the tasks

- Task 1: Estimate the distribution of report delays using Bayesian approach
- Task 2: Back-calculation: Adjust the epidemic curve given distribution of report delays
- Task 3: Estimate time-varying reproductive number based on adjusted epidemic curve
- Task 4: Gather recent reproductive numbers for nowcasting
- Task 5: Nowcasting based on reproductive numbers and known serial interval
- Optional Task: Correct daily case counts for inconsistent testing ratio

Notation

Line-list Data

- $\mathbf{X}$: The regional, week and weekend indicators for each case
- d_i: The report delay for the i^{th} case
- d^{miss}: missing report delays
- d^{obs}: observed report delays
- G: The total number of regions
- W: The number of weeks
- T: The last day of reporting
- D_r: Date of Reporting
- D_o: Date of onset symptom

Epidemic Curve

- N_t: The raw reported number on day t, also is used to represent all daily case counts

Adjusted Epidemic Curve

- $\tilde{N}_t$: The adjusted daily counts based on dates of onset symptom

Parameters

- α_g: the parameters for geographical region, for $g = 1, \ldots, G$
- β_t: the parameters for the week of report, for $t = 1, \ldots, W$
- γ: the parameter for reporting on weekend
- r: the dispersion parameter

Reproductive numbers

- R_t: the time-varying reproductive number with the moving window ends at day t.
- $\hat{R}_t$: the time-varying reproductive number used for nowcasting only.

User-choice hyperparameters

- k: the maximum serial interval
- p_j for $j = 1, \ldots, k$: serial interval density
- l: the maximum report delay
- z: the moving window size for **EpiEstim**

Note that we are only interested in estimating the $\tilde{N}_t$ and R_t.

Data Description

The line list data should have at least five columns (variables): date of report, date of sympotom onset, region, week of report, indicator of whether the case was reported on weekend. Note that the last two variable can be easily derived from the date of report. Region should be number from 1 through G and week of report should be number from 1 through W.

Task 1: Estimate the distribution of report delays

The posterior distribution is:

$$f(\alpha_g, \beta_t, \gamma, r, d^{\text{miss}} | \mathbf{X}, d^{\text{obs}}) \propto f(\alpha_g) f(\beta_t) f(\gamma) f(r) f(d^{\text{miss}}) f(d^{\text{obs}} | \alpha_g, \beta_t, \gamma, r, d^{\text{miss}}, \mathbf{X})$$

Gibbs sampler:

1. sample from $f(d^{\text{miss}} | \alpha_g, \beta_t, \gamma, r, d^{\text{obs}}, \mathbf{X})$
2. sample from $f(\alpha_g | \beta_t, \gamma, r, d^{\text{miss}}, d^{\text{obs}}, \mathbf{X})$
3. sample from $f(\beta_t | \alpha_g, \gamma, r, d^{\text{miss}}, d^{\text{obs}}, \mathbf{X})$
4. sample from $f(\gamma | \alpha_g, \beta_t, r, d^{\text{miss}}, d^{\text{obs}}, \mathbf{X})$
5. sample from $f(r | \alpha_g, \beta_t, \gamma, d^{\text{miss}}, d^{\text{obs}}, \mathbf{X})$

Prior:

- $\alpha_g, \beta_t, \gamma \sim \text{Unif}(a, b)$

- $r \sim \text{Unif}(0, c)$
- $P(d^{\text{miss}} = j) = \frac{1}{l}$, for j = $1, \ldots, l$

Likelihood:

$$d^{\text{obs}}, d^{\text{miss}} \sim \text{NB}(\mathbf{X}\boldsymbol{\beta}, r)$$

where $\boldsymbol{\beta}$ is the collection of $\alpha_g, \beta_t, \gamma$.

The above Bayesian approach has been coded in `mcmc.cpp` and will output estimates of $\alpha_g, \beta_t, \gamma, r$. Note that missing report delays will not be outputted as we only needs to know the distribution of report delays.

Task 2: Back-calculation

$P_{t+j}(d = j)$ is the probability that report delay is j days for cases reported on the t+j day. By definition: $P_{t+j}(d = j) = P(D_o = t | D_r = t + j)$.

Essentially, we are modeling $P_{t+j}(d = j)$ as $d \sim \text{NB}(\mathbf{X_{t+j}}\boldsymbol{\beta}, r)$, where $\mathbf{X_{t+j}}$ are the reporting week and weekend indicator for date $t + j$. The above distribution should be truncated as we have specified the maximum delay as l. Note that $P_{t+j}(d = j)$ can be written as $P(D_o = t | D_r = t + j)$ given D_r is the "parameter" defining the distribution of D_o.

The back-calculated case counts for day t is denoted as y_t, which is computed as $y_t = \sum_{j=0}^{l} N_{t+j} P_{t+j}(d = j)$ for $t = -l, \ldots, T$, assuming the first day of reporting is $t = 0$. Note that the adjustment is complete for $t \leq T - l$ and the adjustment is incomplete for $t = T - l + 1, \ldots, T$.

Formally, the adjusted daily counts $\tilde{N}_t = y_t$ for $t \leq T - l$; $\tilde{N}_t = y_t + s_t$ for $t = T - l + 1, \ldots, T$, which are addressed by nowcasting.

Theory: Connection between back-calculation and nowcasting

Using Bayes formula and treating the reporting day D_r as distribution parameter:

$$P(D_r = t + j | D_o = t) = \frac{P(D_o = t | D_r = t + j) \cdot P(D_r = t + j)}{\sum_{k=0}^{l} P(D_o = t | D_r = t + k) \cdot P(D_r = t + k)}$$

Given flat prior for D_r, the above equation is simplied as:

$$P(D_r = t + j | D_o = t) = \frac{P(D_o = t | D_r = t + j)}{\sum_{k=0}^{l} P(D_o = t | D_r = t + k)} = \frac{P_{t+j}(d = j)}{\sum_{k=0}^{l} P_{t+k}(d = k)}$$

Task 3: Estimate time-varying reproductive number based on adjusted epidemic curve

The time-varying reproductive number R_t is computed by the package **EpiEstim** (for smoothing window size z). Note that we can only trust R_t estimate for $t \leq T - l$. And we need to estimate R_t based on nowcasted case counts for $t > T - l$.

Task 4: Gather recent reproductive numbers for nowcasting

There are three different methods for projection:

- Fixed $\hat{R}_t$: Either takes a fixed value of R_t for all nowcasting days or assumes $\hat{R}_t = \hat{R}_{T-l}$ for $t > T - l$.
- Iterative/Dynamic method: get nowcasted case counts $\tilde{N}_t$ based on $\hat{R}_{t-1}$ and obtain $\hat{R}_t$ based on $\tilde{N}_t$. For this method, input of R_t is not required.
- Specific method: User gives a vector of R_t used for nowcasting, each for one single day. User could use any model they want to produce such vector of reproductive numbers.

Task 5: Nowcasting based on reproductive numbers and known serial interval

The nowcasted case counts are computed as follows: $\tilde{N}_t = y_t + s_t$, $s_t \sim \text{Pois}(\mu_t), \mu_t = [R_{t-1} \sum_{j=1}^{\min(t,k)} \tilde{N}_{t-j} p_j] \cdot w_t$, for $t = T - l + 1, \ldots, T$.

w_t is computed as:

$$w_t = P(D_r > T | D_o = t) = \frac{\sum_{j>T-t}^{l} P(D_o = t | D_r = t + j)}{\sum_{k=0}^{l} P(D_o = t | D_r = t + k)} = \frac{\sum_{j>T-t}^{l} P_{t+j}(d = j)}{\sum_{k=0}^{l} P_{t+k}(d = k)}$$

The standard error of nowcasted case count on date t is approximatedly computed as: $\sigma_t = \sqrt{\sum_{i=T-l+1}^{t} \mu_i}$ for $t = T - l + 1, \ldots, T$.

We will estimate the time varying reproductive number R_t based on $\tilde{N}_t$ for $t = T - l + 1, \ldots, T$ as well.

Note that task 2 through task 5 need to be done for each draw from the joint posterior distribution and we take the 2.5, 50 and 97.5 percentile of $\tilde{N}_t$ and R_t as the point and confidence interval estimate. For nowcasted case counts and time-varying reproductive number, we take a conservative approach:

1. For each posterior draw, we compute the lower and upper bound of $\tilde{N}_t$ as $\tilde{N}_t \pm 2\sigma_t$.
2. For each posterior draw, we find the lower and upper bound of R_t based on $\tilde{N}_t \pm 2\sigma_t$.
3. Across all posterior draws, we take the 2.5 percentile of the lower bound and the 97.5 percentile of the upper bound of $\tilde{N}_t$ (and R_t) as the final confidence interval. We widen the confidence interval intentionally to incorporate uncertainty about back-calculation.

Simulation

1. Generate true epidemic curve based on reproductive number and serial interval
2. Create complete line-list data
3. Make some of the onset dates missing at random
4. Use Bayesian MCMC to recover the distribution of reporting delays
5. Back-calculation
6. Nowcasting
7. Compare the estimate of case counts and reproductive numbers to the known ones.