The Bayesian framework of back-calculation and nowcasting based on line-list data

Tenglong Li

July 9, 2020

1 Introduction

Surveillance plays a pivotal role in controlling the Covid-19 pandemic and is being used to provide guidance for government responses to the pandemic. A prerequisite for effective surveillance is to have reliable epidemic curve. This is challenging for many reasons, including the use of case report dates and challenges with the timeliness of case reporting.

Case report dates are typically used to track the disease and indicate when a case was reported. This can occur at many points in the progression of disease, depending on when an individual is tested (e.g. there can be a substantial difference between individuals detected from surveillance testing and those detected upon admission to the hospital with advanced disease). It is more informative to use a date that is linked to the progression of the disease, and in that case, the observable symptom onset date is preferred. This date is more useful in estimating reproductive numbers and is more proximate in time to the infection date, the most biologically relevant date, but typically not observable. Unfortunately, the symptom onset dates are often missing for many cases in case notification data, rendering the transformation of epidemic curve based on the original reported dates impossible without statistical adjustment.

Case report data also suffers from lags in reporting (REFs). Obtaining test results and then having those reported to appropriate public health authorities takes a variable amount of time. This means that the reported cases on any given day are necessarily missing the most recent cases. Again, to have a more timely and accurate understanding of current case burden requires statistical adjustment.

There are at least two steps to recover epidemic curve based on the reporting curve to address these two issues. The first step is back-calculation, which requires one to group cases based on their onset dates and recount the group sizes. The second step is nowcasting, which is only necessary for the most recent portion of the back-calculated curve. The reason for doing nowcasting for the last part is that the case counts whose onset dates close to the last reporting day are incomplete and those numbers will be increased as case reports are filled in in the future.

Most previous work focused on obtaining the epidemic curve directly from reporting curve based on the distribution of reporting delay, which is defined as the lag between the onset date and the reporting date for a case (REFs). This delay distribution is typically estimated using...(can you fill in what type of data is used for these approaches?). The distribution of reporting delay is built on the observed reporting delays and typically is time invariant. When line-list data, which has information at individual level, is available the approach can be modified to impute individual missing onset dates using the reporting delay distribution. In this case we can assume that onset dates are most likely missing at random conditional on reporting dates and inference on reporting delay distribution can account for this kind of missing mechanism for imputation (you can REF our 2009 H1N1 paper where we do this crudely).

In this project, we develop an analytical framework for estimating the reporting delay distribution and recovering the epidemic curve using individual line list data. This framework has three components: (1) Bayesian inference of the reporting delay distribution; (2) Back-calculation based on the estimated reporting delay distribution; (3) Nowcasting. The Bayesian inference is built on a model where parameters of the reporting delay distribution depend on the reporting date and spatial region. Missing onset dates are imputed based on the estimated reporting delay distribution as well. Figure 1 illustrates the workflow of this framework.

2 Method

2.1 Model setup

For line-list data, we denote individual reported date and onset dates as r_i and o_i, respectively, for $i = 1, \dots, n$. Therefore, the individual reporting delay is defined as $d_i = r_i - o_i$ and we assume $d_i \geq 0$ throughout this paper. Furthermore, we define X_1 as the nxp matrix contains weekly indicators and X_2 as the nxp matrix of weekend indicators, where p indicates the number of weeks of data. Note that the indicators in X_1 can be modified to any other appropriate time scale.

In general, the reporting delay distribution is defined as:

$$d \sim \mathrm{NB}(\mu, r), \mu = e^{\alpha + X_1\beta + X_2\gamma} \tag{1}$$

where r and μ are the size (dispersion) and mean parameters for negative binomial distribution. α is the region-specific intercept.

2.2 Bayesian inference of reporting delay distribution

Built on the negative binomial regression (1), the posterior distribution is defined as:

$$f(\alpha_g, \beta_t, r, d^{\text{miss}} | d^{\text{obs}}, X_1, X_2) \propto f(\alpha) f(\beta) f(\gamma) f(r) f(d^{\text{miss}}) f(d^{\text{obs}} | \alpha, \beta, \gamma, r, d^{\text{miss}}, X_1, X_2) \tag{2}$$

where d^{miss} denotes missing reporting delays and d^{obs} denotes observed reporting delays.

For the priors, we choose:

1. $\alpha, \beta, \gamma \sim \text{Unif}(-b, b)$
2. $r \sim \text{Unif}(0, c)$
3. $P(d^{\text{miss}} = j) = \frac{1}{M}$, for j $= 1, \ldots, M$

where b, c, M are large positive numbers and indicate the priors are uninformative.

The following Gibbs sampler will iterates:

1. sample from $f(d^{\text{miss}} | \alpha, \beta, \gamma, r, X_1, X_2)$
2. sample from $f(\alpha | \beta, \gamma, r, d^{\text{miss}}, d^{\text{obs}}, X_1, X_2)$
3. sample from $f(\beta | \gamma, r, d^{\text{miss}}, \beta, d^{\text{obs}}, X_1, X_2)$
4. sample from $f(\gamma | r, d^{\text{miss}}, \alpha, \beta, d^{\text{obs}}, X_1, X_2)$
5. sample from $f(r | d^{\text{miss}}, \alpha, \beta, \gamma, d^{\text{obs}}, X_1, X_2)$

2.3 Back-calculation

Back-calculation needs one to set a maximum reporting delay l, then the probability of onset date $D_o = t - j$ (i.e., reporting delay is j) given reporting date $D_r = t$ is

$$P(D_o = t - j | D_r = t) = \frac{P_t(d = j)}{\sum_{k=0}^{l} P_t(d = k)} \tag{3}$$

where $P_t(d = k)$ is the probability mass function of the negative binomial distribution corresponding to reporting date t and reporting delay k.

Given reported curve $\{N_t | t = 0, 1, \ldots, T\}$, the back-calculated curve is:

$$y_t = \sum_{j=0}^{l} N_{t+j} P(D_o = t | D_r = t + j). \tag{4}$$

We note that the epidemic curve, $\{\tilde{N}_t | t = -l, -l+1, \ldots, T\}$ will be very incomplete as t approaches T, i.e.

$$\begin{aligned} \tilde{N}_t &= y_t, \text{ for } t = -l, \ldots, T - l \\ \tilde{N}_t &\geq y_t, \text{ for } t = T - l + 1, \ldots, T. \end{aligned} \tag{5}$$

We address this issue using nowcasting for the back-calculated curve $\{y_t | t = -l, -l + 1, \ldots, T\}$.

2.4 Nowcasting

2.4.1 The parametric approach

The parametric approach typically needs knowledge on reproductive numbers, and then it simulates $\tilde{N}_t, t = T - l + 1, \ldots, T$ based on the following distribution:

$$\tilde{N}_t \sim \text{Pois}(\mu_t), \mu_t = R_t \cdot \sum_{j=1}^{\min(t,s)} \tilde{N}_{t-j} p_j \tag{6}$$

where s is the maximum serial interval and p_j is the probability that serial interval equal to j. The serial interval distribution and s must be given by users. We provide three possible options for R_t specifications:

1. Dynamic method: We compute $\hat{R}_{t-1}$ (the estimate of R_{t-1}) using a method such as that proposed by Cori et al (REF) and use it as the input for R_t, and then generate $\tilde{N}_t$ based on (6). This algorithm will iterate until $\tilde{N}_T$ is obtained.
2. Fixed method: We use a fixed value as R_t for $t = T - l + 1, \ldots, T$. This fixed value can be provided by users or be $\hat{R}_{T-l}$.
3. Specific method: Users input $R_t, t = T - l + 1, \ldots, T$, presumably based on a hypothetical scenario or obtained from other data.

The 95% confidence interval can be approximately computed as $\mu_t \pm 2 \cdot \sqrt{\sum_{i=T-l+1}^{t} \mu_t}$. The dynamic method does not require knowledge about reproductive numbers while the other two methods do.

2.4.2 The nonparametric approach

The nonparametric approach is based on the idea that we can estimate the probability of $D_r = t + j$ given $D_o = t$. Note that in order to generate this estimate, we need to consider two factors: (1) whether the reporting delay distribution will change for future days; (2) what the reported daily counts will be for future l days. The probability of $D_r = t + j$ given $D_o = t$ can be easily obtained:

$$P(D_r = t + j) = \frac{N_{t+j}}{\sum_{k=0}^{l} N_{t+k}}, \ j = 0, 1, \ldots, l.$$

$$P(D_r = t + j | D_o = t) = \frac{P(D_o = t | D_r = t + j) \cdot P(D_r = t + j)}{\sum_{k=0}^{l} P(D_o = t | D_r = t + k) \cdot P(D_r = t + k)}, \ t = T - l + 1, \ldots, T. \tag{7}$$

Drawing on (7), we will able to estimate $\tilde{N}_t$ as:

$$\begin{aligned} w_t = P(D_r \leq T | D_o = t) &= \frac{\sum_{j=0}^{T-l} P(D_o = t | D_r = t+j) \cdot P(D_r = t+j)}{\sum_{k=0}^{l} P(D_o = t | D_r = t+k) \cdot P(D_r = t+k)} \\ \tilde{N}_t &= \frac{y_t}{w_t},\ t = T-l+1, \ldots, T. \end{aligned} \tag{8}$$

The knowledge on future reported case counts $\{N_t | t = T+1, \ldots, T+l\}$ is needed for the nonparametric approach. One can either make assumption or projections about $\{N_t | t = T+1, \ldots, T+l\}$. Time-series forecasting for future reported case counts is also appropriate in this case. On the other hand, the future reporting delay distribution could be assumed constant as time goes by.

In general, nowcasting is more challenging than back-calculation since it depends on future reported counts which are unavailable in this context. The final confidence interval of $\tilde{N}_t$ (and possibly $\hat{R}_t$)should take the uncertainty in estimating the Bayesian inference of reporting delay distribution as well.

3 Simulation study

For this simulation study, the time-varying reproductive number R_t is estimated via `EpiEstim` for a sliding window size τ:

$$\begin{aligned} \hat{R}_t &= \frac{(\sum_{k=t-\tau}^{t} \tilde{N}_k) + 1}{(\sum_{k=t-\tau}^{t} \Lambda_k(p_j)) + 0.2} \\ \Lambda_k(p_j) &= \sum_{j=1}^{\min(k,s)} \tilde{N}_{k-j} p_j \end{aligned} \tag{9}$$

In general, the simulation study is set up based on the reported curve $\{N_t | t = 0, 1, \ldots, T+l\}$ and the reporting delay distribution. We have two simulation scenarios:

1. Scenario 1: The distribution of future reporting delay is close to the distribution of reporting delay in the most recent week (time).

2. Scenario 2: The distribution of future reporting delay is different from the distribution of reporting delay in the most recent week (time).

In each simulation scenario, we compare seven different approaches for nowcasting: 1-dynamic method; 2-specific method with biased input; 3-specific method with unbiased input; 4-fixed method with biased input; 5-fixed method with unbiased input; 6-nonparametric method with biased input; 7-nonparametric method with unbiased input.

The simulation procedure is described below:

1. Choose scenario.
2. Set up the reported curve, parameters of reporting delay distribution, serial interval, the maximum delay, and τ.
3. Obtain the true epidemic curve from $\{N_t | t = 0, 1, \ldots, T + l\}$.
4. Compute the true $\hat{R}_t$ based on (9) and τ.
5. Create line-list data and get onset dates based on reporting delay distribution.
6. Randomly make some onset dates missing for each reporting date.
7. Bayesian inference of reporting delay distribution given the simulated line-list data.
8. Back-calculation.
9. Do nowcasting using the seven approaches mentioned above.
10. Generate estimated epidemic curve and $\hat{R}_t$ base on it.
11. Repeat step 3 through step 10 many times.
12. Compare all estimates with the true epidemic curve and $\hat{R}_t$

4 Example: CDC data

5 Discussion