西北疫情研究报告

摘要

2021 年 10 月 13 日，内蒙古锡林郭勒盟二连浩特市报告第一例本土确诊病例，随后疫情蔓延至西安、兰州、甘肃、北京、重庆和成都等多地，持续时长长达一个月之久，我们收集了本次疫情日新增确诊数据，采用了贝叶斯方法估计有效再生数 R_t

关键字：有效再生数、贝叶斯估计

一、 研究背景

二、 研究方法

2.1 数据来源

中华人民共和国国家卫生与健康委员会 [1]

2.2 模型介绍

2.2.1 符号说明

符号	含义
D_t	第 t 天新增确诊数
I_t	第 t 天新增发病数
T	从发病到确诊的时间
$p_{ij} = P(i-j \leqslant T < i-j+1)$	第 i 天确诊，第 j 天发病的概率
R_t	第 t 天有效再生数

2.2.2 假设

- 从发病到确诊的时间 T 服从 Weibull 分布
- 每一天的新增发病数互相独立

2.2.3 迭代算法计算每日新增发病病例

我们从中华人民共和国国家安全卫生与健康委员会得到了从十月十三日到十一月十五日三十四天来的每日新增确诊病例。十月十三日，内蒙古锡林郭勒盟二连浩特市报告第一例确诊病例，十一月十五日报告了来源于内蒙古的最后一例新增确诊病例。假设每日新增发病数 I_t(第 t 天) 服从 Poisson 分布 $P(\lambda_t)$(以 λ_t 为均值), 通过我们得到的每日新增确诊病例，我们希望得到每日新增发病数，即估计参数 λ_t。我们使用了 Richardson-Lucy 的迭代算法 [2] 解决这一问题，给出迭代公式如下

$$D_i^{(n)} = \sum_{j=t_1}^{i} p_{ij}\lambda_j^{(n)}, \quad \lambda_j^{(n+1)} = \frac{\lambda_j^{(n)}}{q_j} \sum_{i=\max|j,r_1\}}^{r_m} \frac{p_{ij}D_i}{D_i^{(n)}} \tag{1}$$

其中，假设我们要根据自 r_1 至 r_m 共 m 天的每日新增确诊数，即 $\{D_i\}_{i=r_1,\ldots,r_m}$ 估计自 t_1 至 t_l 共 l 天的每日新增发病数，即 $\{\lambda_j\}_{j=t_1,\ldots,t_l}$ 那么公式中的 q_j 满足

$$q_j = \begin{cases} 0, j < t_1 or j > t_l \\ \sum\limits_{i=\max\{j,s_1\}}^{s_m} p_{ij} \end{cases}$$

而 $D_i^{(n)}$ 与 $\lambda_j^{(n)}$ 表示 C_i 与 λ_j 迭代第 n 次的结果。

令

$$\chi^2 = \frac{1}{m}\sum_{i=r_1}^{r_m}\frac{\left(D_i^{(n)} - D_i\right)^2}{D_i^{(n)}}$$

则当 χ^2 足够小时停止迭代。

2.2.4 贝叶斯方法估计有效再生数

为了估计有效再生数 R_t，我们参考了文献 [3] 中的模型：

$$E[I_t] = R_t\sum_{j=1}^{t} w_j I_{t-j} \tag{2}$$

其中

- E 代表随机变量的数学期望；
- w_j 表示自感染后第 j 天的传染性；

我们不难得知，w_j 与被感染的时刻无关，且当 w_j 最大时，表示一个被感染的患者在第 j 天传染性最强, 考虑到在实际问题中 w_j 难以观测，所以用症状出现时间的分布 p_j 来估计从首发病例出现症状到其继发病例出现症状的时间间隔的分布，换句话说，p_j 最大表示从首发病例出现症状经过 j 天继发病例出现症状的概率最大，也就表示病例在第 j 天的传染性最强，所以我们得到了公式：

$$E[I_t] = R_t\sum_{j=1}^{t} p_j I_{t-j} \tag{3}$$

进一步，我们通过文献 [3] 中提供的贝叶斯方法来估计 R_t 的值，假设 R_t 在时间间隔 $[t-\tau, t]$ 内保持不变，且 R_t 以伽马分布为先验分布。即

$$P(R_t = x) = f(x, \beta, \alpha) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, x > 0 \tag{4}$$

则 R_t 的后验分布为

$$\pi(R_t|I) = R_{t,\tau}^{\alpha+\sum_{s=t-\tau+1}^{t} I_s - 1}e^{-R_{t,\tau}\left(\sum_{s=t-\tau+1}^{t}\Lambda_s + \frac{1}{\beta}\right)}\prod_{s=t-\tau+1}^{t}\frac{\Lambda_s^{I_s}}{I_s!}\frac{1}{\Gamma(\alpha)\beta^\alpha} \tag{5}$$

其中，$\Lambda_t = \sum_{j=1}^{t} I_{t-j} p_j$

上述 $\pi(R_t|I)$ 正比于

$$R_{t,\tau}{}^{\alpha+\sum_{s=t-\tau+1}^{t} I_s - 1} e^{-R_{t,\tau}\left(\sum_{s=t-\tau+1}^{t} \Lambda_s + \frac{1}{\beta}\right)} \prod_{s=t-\tau+1}^{t} \frac{\Lambda_s^{I_s}}{I_s!}$$

即 $R_{t,\tau}$ 的后验分布服从以 $(\alpha + \sum_{s=t-\tau+1}^{t} I_s)$ 为形状参数，以 $(\frac{1}{\frac{1}{\beta}+\sum_{s=t-\tau+1}^{t} \Lambda_s})$ 为尺度参数的伽马分布。

我们将后验分布的均值作为 R_t 的估计值，容易知道，R_t 的后验分布的均值为 $\frac{\alpha+\sum_{s=t-\tau+1}^{t} I_s}{\frac{1}{\beta}+\sum_{s=t-\tau+1}^{t} \Lambda_s}$ 。

三、 结果呈现

3.1 每日新增发病

根据公式 (1)，我们实现了迭代算法，根据自 r_1 至 r_m 共 m 天的每日新增确诊数，即 $\{D_i\}_{i=r_1,\ldots,r_m}$ 我们估计了自 t_1 至 t_l 共 l 天的每日新增发病数，即 $\{\lambda_j\}_{j=t_1,\ldots,t_l}$

将结果呈现在下图中：

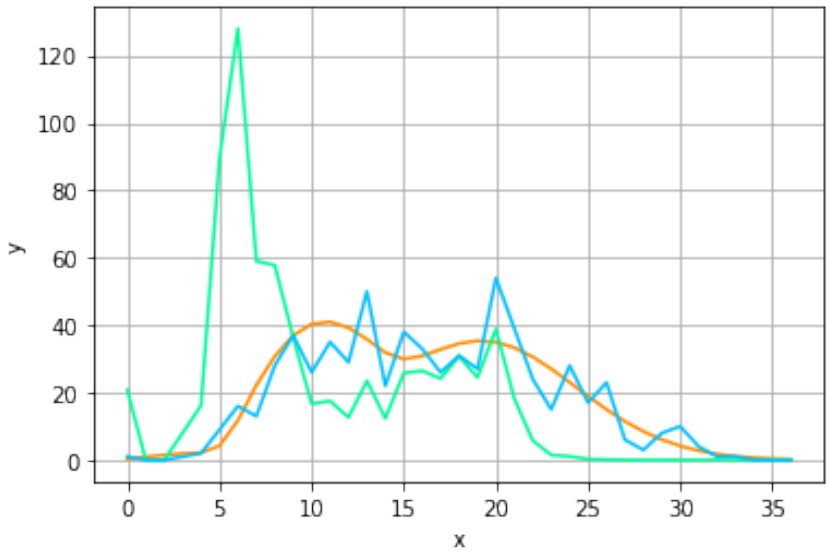

图 **1** 每日新增发病：图中蓝色折线代表每日新增确诊数，橙色曲线代表迭代 **200** 次之后的每日新增确诊，绿色曲线代表我们估计的每日新增发病数

3.2 有效再生数 R_t 的估计

我们选择序列区间（从首发病例症状出现到继发病例症状出现）服从均值为 5 天，标准差为 3 天的伽马分布 [4]，并设置 $\alpha = 1, \beta = 5$。为了更好地检测 R_t 的变化，我们选择了一个相对较小的时间窗口 $\tau = 3$。

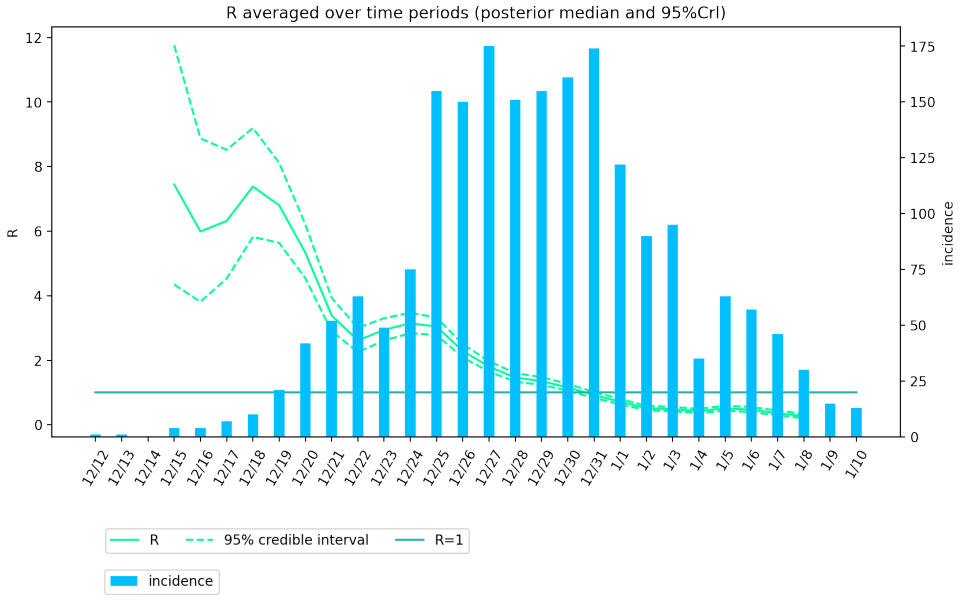

图 **2**　R_t 的估计结果：绿色实线代表 R_t 的均值，绿色虚线代表置信水平为 **95%** 的区间估计

参考文献

[1] https://www.dydata.io/datastore/detail/1888217184434524160/

[2] Goldstein E, Dushoff J, Ma J, Plotkin JB, Earn DJ, Lipsitch M. Reconstructing influenza incidence by deconvolution of daily mortality time series. Proc Natl Acad Sci U S A. 2009 Dec 22;106(51):21825-9. doi: 10.1073/pnas.0902958106. Epub 2009 Dec 18. PMID: 20080801; PMCID: PMC2796142.

[3] Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013 Nov 1;178(9):1505-12. doi: 10.1093/aje/kwt133. Epub 2013 Sep 15. PMID: 24043437; PMCID: PMC3816335.

[4] Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. Int J Infect Dis. 2020 Apr;93:284-286. doi: 10.1016/j.ijid.2020.02.060. Epub 2020 Mar 4. PMID: 32145466; PMCID: PMC7128842.