Journal of Data Science xx (xx), 1–6 DOI: xx.xxxx/xxxxxxxxx
July 2020

Supplementary Information to "Tracking Reproductivity of COVID-19 Epidemic with Varying Coefficient SIR Model"

Haoxuan Sun[1], Yumou Qiu[2], Han Yan[3], Yaxuan Huang[4], Yuru Zhu[5], Jia Gu[5], and Song Xi Chen[6,5]

[1]*Center for Data Science, Peking University*
[2]*Department of Statistics, Iowa State University*
[3]*School of Mathematical Sciences, Sichuan University*
[4]*Yuanpei College, Peking University*
[5]*Center for Statistical Science, Peking University*
[6]*Guanghua School of Management, Peking University*

Abstract

This supplementary material provides the detailed expressions for the estimation of the parameters in the varying coefficient SIR model (vSIR), and additional tables and figures for the main paper "Tracking Reproductivity of COVID-19 Epidemic with Varying Coefficient SIR Model".

1 Parameter estimation in vSIR

To estimate $\beta(t)-\gamma$ in equation (2), given a small window size $w>0$, we regress $\log\{I(t_1)\}$ on t_1 over the range $\{t_1 = t-w+1,\ldots,t\}$ for each $t \geq w$. Consider the linear regression

$$\log I(t_1) = \alpha(t) + (\beta(t)-\gamma)t_1 + \xi_{t_1},$$

for an error term ξ_{t_1} over the time range $\{t_1 = t-w+1,\ldots,t\}$. Let

$$\begin{aligned}\widehat{\beta(t)-\gamma} &= \frac{\sum_{t_1=t-w+1}^{t}(t_1-\bar{t}_1)\log I(t_1)}{\sum_{t_1=t-w+1}^{t}(t_1-\bar{t}_1)^2} \quad \text{and} \\ \hat{\alpha}(t) &= \frac{\sum_{t_1=t-w+1}^{t}\log I(t_1)}{w} - (\widehat{\beta(t)-\gamma})\bar{t}_1\end{aligned}$$

be the estimated slope and intercept of this local linear regression, respectively, where $\bar{t}_1 = \sum_{t_1=t-w+1}^{t} t_1/w$. Let

$$\widehat{\mathrm{Var}}(\widehat{\beta(t)-\gamma}) = \frac{\mathrm{MSE}_t}{\sum_{t_1=t-w+1}^{t}(t_1-\bar{t}_1)^2}$$

be the estimated variance of $\widehat{\beta(t)-\gamma}$, where $\mathrm{MSE}_t = \sum_{t_1=t-w+1}^{t}\{\log I(t_1)-\hat{\alpha}(t)-(\widehat{\beta(t)-\gamma})t_1\}^2/w$ is the mean square error.

The estimator of γ and its estimated variance are

$$\hat{\gamma} = \frac{\sum_{t=1}^{T-\delta}\Delta_\delta R_t I(t)}{\delta\sum_{t=1}^{T-\delta} I^2(t)} \quad \text{and} \quad \widehat{\mathrm{Var}}(\hat{\gamma}) = \frac{\mathrm{MSE}}{\delta^2\sum_{t=1}^{L-\delta} I^2(t)},$$

respectively, where $\mathrm{MSE} = \sum_{t=1}^{T-\delta}\{\Delta_\delta R_t - \hat{\gamma}I(t)\delta\}^2/(T-\delta)$ is the mean square error.

Table A1: The starting dates of the epidemic data for each province and city.

	Provinces	Starting dates		Provinces	Starting dates
1	Anhui	Jan-21-2020	24	Liaoning	Jan-22-2020
2	Beijing	Jan-20-2020	25	Neimenggu	Jan-24-2020
3	Chongqing	Jan-21-2020	26	Ningxia	Jan-22-2020
4	Enshizhou	Jan-24-2020	27	Qinghai	Jan-25-2020
5	Ezhou	Jan-24-2020	28	Shaanxi	Jan-23-2020
6	Fujian	Jan-22-2020	29	Shandong	Jan-21-2020
7	Gansu	Jan-24-2020	30	Shanghai	Jan-20-2020
8	Guangdong	Jan-20-2020	31	Shanxi	Jan-22-2020
9	Guangxi	Jan-22-2020	32	Shiyan	Jan-23-2020
10	Guizhou	Jan-22-2020	33	Sichuan	Jan-21-2020
11	Hainan	Jan-22-2020	34	Suizhou	Jan-24-2020
12	Hebei	Jan-22-2020	35	Tianjin	Jan-21-2020
13	Heilongjiang	Jan-22-2020	36	Tianmen	Jan-24-2020
14	Henan	Jan-21-2020	37	Wuhan	Jan-16-2020
15	Huanggang	Jan-20-2020	38	Xiangyang	Jan-25-2020
16	Huangshi	Jan-25-2020	39	Xianning	Jan-25-2020
17	Hubei	Jan-16-2020	40	Xiantao	Jan-23-2020
18	Hunan	Jan-21-2020	41	Xiaogan	Jan-23-2020
19	Jiangsu	Jan-22-2020	42	Xinjiang	Jan-23-2020
20	Jiangxi	Jan-21-2020	43	Yichang	Jan-24-2020
21	Jilin	Jan-23-2020	44	Yunnan	Jan-21-2020
22	Jingmen	Jan-22-2020	45	Zhejiang	Jan-21-2020
23	Jingzhou	Jan-22-2020			

Table A2: The prediction error in percentage, $\{\hat{I}(t) - I(t)\}/I(t)$, for February 17, 18 and 19 by the proposed prediction method in Section 3.4 with the empirically estimated γ, using data to February 16 2020.

Province	Feb 17	Feb 18	Feb 19
Hubei	2.22%	4.37%	5.98%
Guangdong	4.72%	4.46%	5.82%
Zhejiang	5.16%	4.76%	5.38%
Beijing	3.29%	3.01%	3.32%
Chongqing	7.06%	8.32%	11.49%
Hunan	7.06%	8.45%	10.95%
Guangxi	3.92%	2.45%	2.71%
Shanghai	9.70%	14.46%	20.70%
Jiangxi	4.07%	-0.30%	-4.14%
Sichuan	3.58%	4.73%	6.28%
Shandong	5.94%	7.00%	9.58%
Anhui	4.96%	3.48%	3.36%
Fujian	3.75%	3.52%	3.45%
Henan	5.26%	5.54%	6.21%
Jiangsu	7.15%	7.77%	9.83%
Hainan	7.15%	4.22%	7.11%
Tianjin	7.16%	12.35%	16.56%
Yunnan	5.49%	7.27%	11.10%
Shaanxi	4.59%	4.06%	4.35%
Heilongjiang	3.03%	4.47%	6.69%
Liaoning	5.99%	3.70%	3.89%
Guizhou	8.84%	11.20%	16.86%
Jilin	4.16%	4.08%	6.49%
Ningxia	6.75%	4.60%	5.55%
Hebei	6.23%	5.86%	7.34%
Gansu	10.22%	9.37%	10.79%
Xinjiang	2.30%	0.90%	-1.89%
Shanxi	4.84%	2.34%	0.09%
Neimenggu	-1.45%	-2.72%	-4.92%
Qinghai	14.21%	1.88%	1.19%

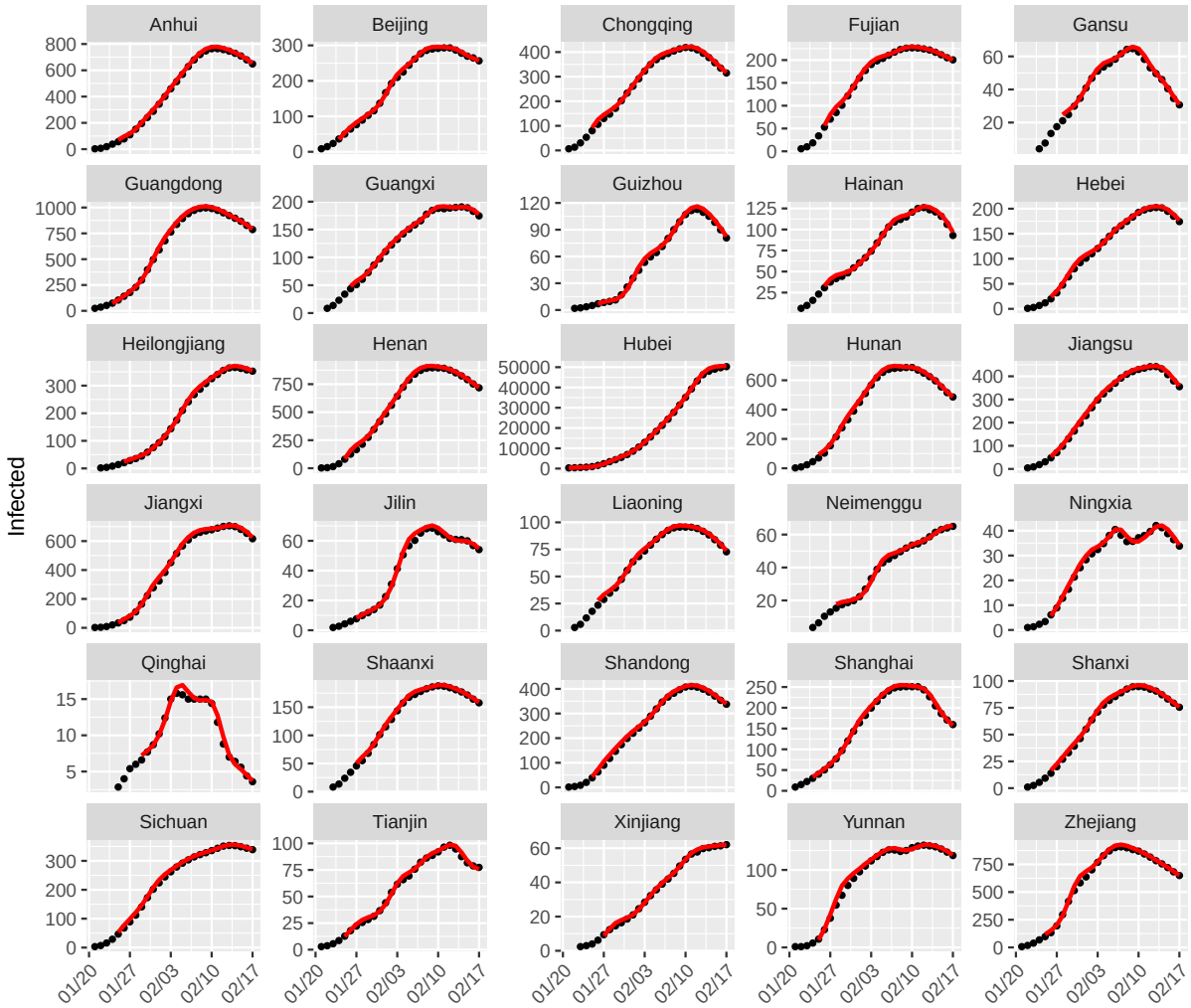

Figure S1: The Fitted curve (red) for the number of the infected patients ($I(t)$, black dot) in 30 provinces by the proposed varying coefficient SIR model.

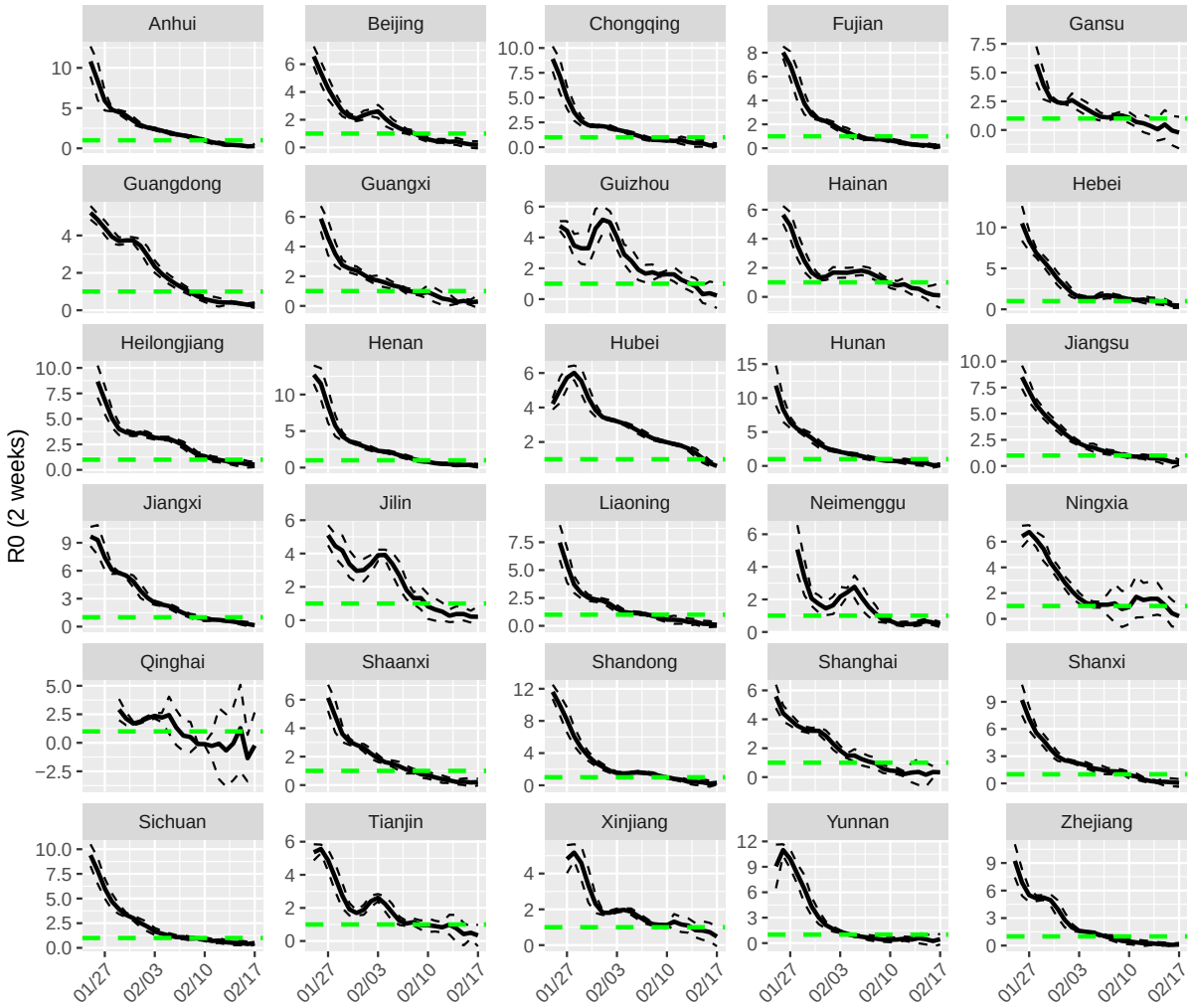

Figure S2: The estimated effective reproductive number $R_t^{14} = \hat{\beta}(t) \times 14$ with its 95% confidence interval for 30 provinces, where 14 days is taken as the average infectious period. The green horizontal line is 1.

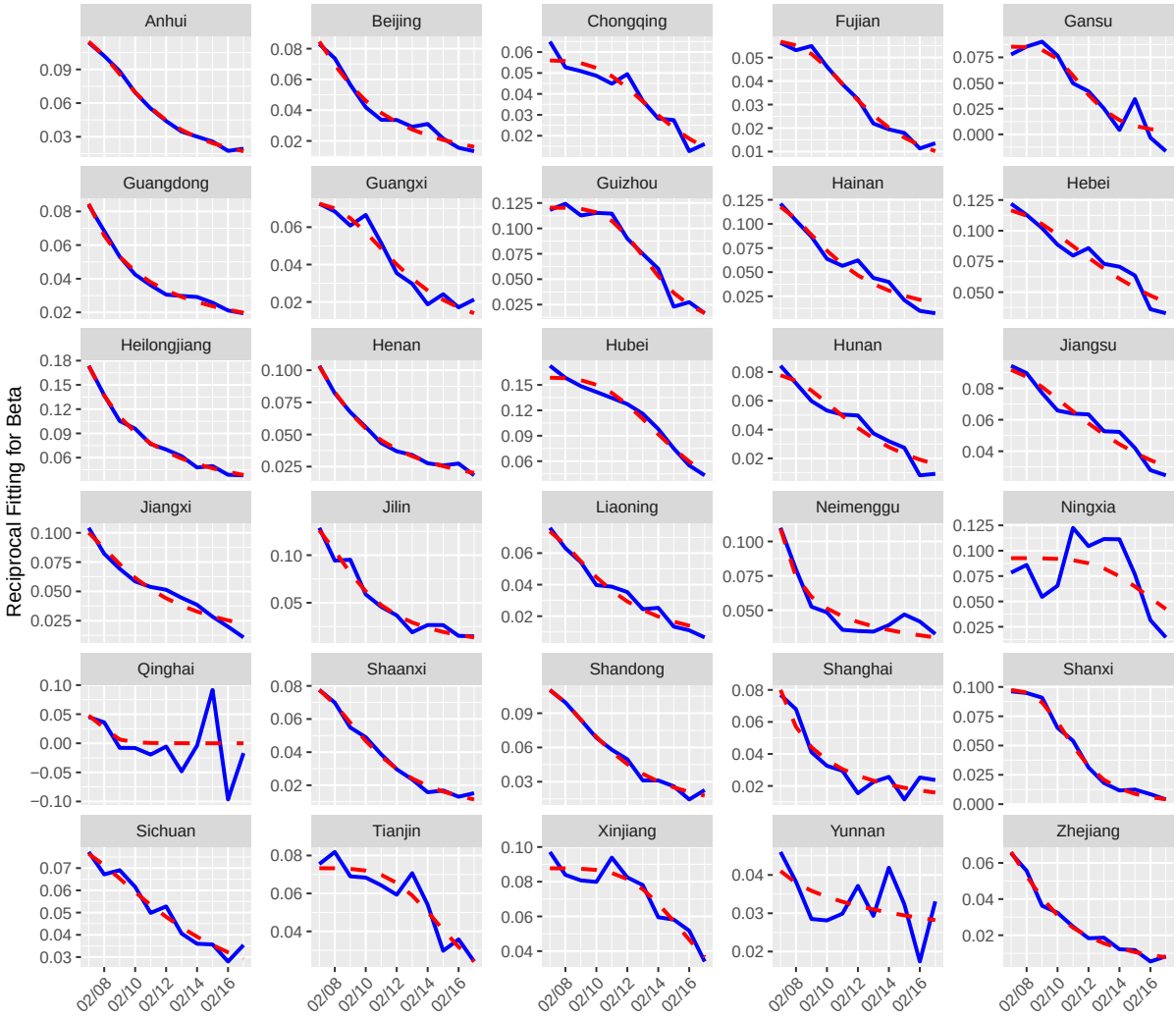

Figure S3: The estimated $\hat{\beta}(t)$ (in blue) from the varying coefficient SIR model (1) versus the fitted line $\tilde{\beta}(t)$ (in red) of $\hat{\beta}(t)$ by the reciprocal regression model (4) for the data from Feb 6th to Feb 17th 2020 for 30 provinces.