






Time fused coefficient SIR model with application to COVID-19 epidemic in the United States

Hou-Cheng Yang ^{a*}, Yishu Xue ^{b*}, Yuqing Pan^c, Qingyang Liu^b and Guanyu Hu ^d

^aDepartment of Statistics, Florida State University, Tallahassee, FL, USA; ^bDepartment of Statistics, University of Connecticut, Storrs, CT, USA; ^cMicrosoft Corporation, Redmond, WA, USA; ^dDepartment of Statistics, University of Missouri Columbia, Columbia, MO, USA

ABSTRACT

In this paper, we propose a Susceptible–Infected–Removal (SIR) model with time fused coefficients. In particular, our proposed model discovers the underlying time homogeneity pattern for the SIR model's transmission rate and removal rate via Bayesian shrinkage priors. MCMC sampling for the proposed method is facilitated by the **nimble** package in R. Extensive simulation studies are carried out to examine the empirical performance of the proposed methods. We further apply the proposed methodology to analyze different levels of COVID-19 data in the United States.

ARTICLE HISTORY

Received 8 August 2020
Accepted 23 May 2021

KEYWORDS

Time fusion; homogeneity pursuit; infectious diseases; MCMC; shrinkage prior

2010 MATHEMATICS

SUBJECT CLASSIFICATION
62P10


1. Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first identified in December 2019, and then rapidly spread across the world, causing the current global pandemic of coronavirus disease 2019 (COVID-19). As of July 23, the novel coronavirus has spread to 216 countries and territories, with a total of more than 14 million confirmed infections and 600,000 fatal cases worldwide [26]. Eight months after the initial outbreak, large numbers of new cases are still reported from many major countries, resulting in not only public health crises, but also severe economic and political ramifications. As the pandemic rages on with no end in sight, it is of urgent necessity for epidemiologists to quantify and interpret the trajectories of the COVID-19 pandemic, so as to help formulate more effective public policies.

The Susceptible–Infectious–Recovered [SIR; 12] model and its variants, such as Susceptible–Infected–Removed–Susceptible [SIRS; 13,14] and Susceptible–Exposed–Infected–Removal [SEIR; 9] models are commonly used to describe the dynamics of an infectious disease in a certain region. In the basic SIR model, a population is segregated into three time-dependent compartments including Susceptible ($S(t)$), Infectious ($I(t)$),

CONTACT Guanyu Hu  gh7mr@missouri.edu

*These authors contributed equally to this work.

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02664763.2021.1936467>

and Recovered/removed ($R(t)$). One who does not have the disease at time t , but may be infected due to contact with an infected person belongs to the susceptible compartment. The infected compartment is made up of those who have a disease at time t , and can potentially get a susceptible individual infected by contact. The recovered compartment include those who are either recovered or dead from the disease, and are no longer contagious, i.e. removed from the infectious compartment, at time t . Removal can be due to several possible reasons, including death, recovery with immunity against reinfection, and quarantine and isolation from the rest of the population. A recovered/removed individual will not be back into the susceptible compartment anymore. Such model assumption match well with the COVID-19 outbreak, and therefore we adopt the SIR model as our basic model in this paper.

From the statistical perspective, the key study objective is the inference of transmission and recovery rates from the model. Regarding time-invariant SIR and SEIR models, there have been timely applications to early epidemic data right after the breakout of COVID-19 [18,22,27]. In order to differentiate evolutionary patterns of COVID-19 among different regions, Hu and Geng [10] developed a Bayesian heterogeneity learning methodology for SIRS. As the epidemic continued to spread rampantly, statisticians proposed time-dependent models based on SIR to elucidate the temporal dynamics of this disease [4,11,21]. Estimated by various assumptions on temporal smoothness, the transmission and recovery rates of these models constantly alter over time, which limits their ability to effectively detect abrupt changes. In contrast, we consider the scenario in which the transmission and recovery rates are constant within locally stationary periods segmented by a collection of change points, which is aligned with the fact that different stages of epidemic progression are naturally partitioned. This motivates us to estimate a piecewise constant model.

The fused lasso [23], with L_1 sparsity-inducing penalty imposed on all successive differences, is one of the most popular methods for time fusion and change point detection. Motivated by the frequentist L_1 fusion penalty, Kyung *et al.* [1] proposed its Bayesian counterpart, namely Bayesian fused lasso, which imposed independent Laplace priors [16] on the differences. To solve the posterior inconsistency problem of Bayesian fused lasso, Song and Cheng [19] used heavier tailed student- t priors for Bayesian fusion estimation. In addition to the Laplace and student- t priors, other Bayesian shrinkage priors with different statistical properties, such as spike-and-slab [7] and horseshoe priors [3], can also be adopted to induce time fusion.

The contributions of this paper are in three-fold. First, we apply three different types of shrinkage priors to capture the time homogeneity patterns of infectious and removal rates under the SIR framework. Second, it is noticed that our proposed method can be easily implemented by the **nimble** package [6] in R. A straightforward tutorial on using **nimble** to obtain shrinkage priors under the SIR framework is provided in the supplemental material. Finally, several interesting findings are discovered through analysis of COVID-19 data, including including national level, state level, and county level.

The remainder of this paper is organized as follows. In Section 2, the COVID-19 data of selected state and county are introduced. We briefly review the SIR model, and then present our model framework in Section 3. Simulation studies are conducted in Section 4. Applications of the proposed methods to COVID-19 data are presented in Section 5. Section 6 concludes the paper with a discussion.

2. Motivating data

The COVID-19 data is obtained from the R package **COVID19** [8]. We consider the observations recorded from 2020-05-14 to 2020-07-23, a 71-day long period. US nationwide aggregated data, as well as data for five states: New York (NY), California (CA), Florida (FL), South Dakota (SD), and Wyoming (WY) are our focus in this study. We also consider the county-level data including: Los Angeles, Miami-Dade and New York City. The data is reported daily, with variables including the population size, the number of confirmed cases, the number of recoveries, and the number of deaths, etc.

Note that the removal group for county-level data only contain deaths and there is no information available for recoveries. Similar to in Sun *et al.* [21], a three-point moving average filter is applied to the infectious group $I(t)$ and removal group $R(t)$ to reduce noise. Due to the large size of the susceptible group, the group sizes are visualized on a natural log scale in Figure 1 and Figure 2. The infectious and removal numbers are much smaller

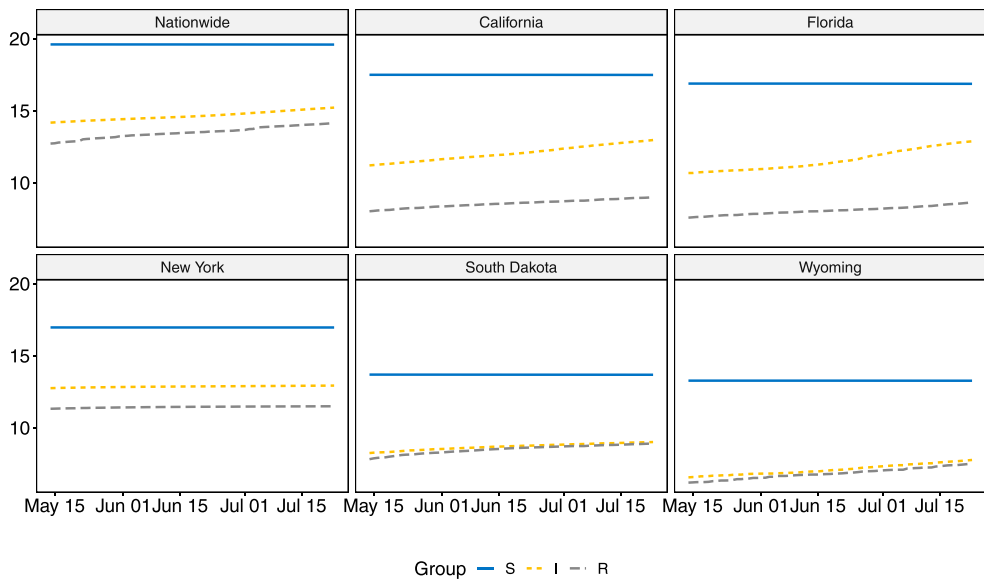


Figure 1. Visualizations for $S(t)$, $I(t)$ and $R(t)$ in US nationwide and five individual states on natural log scale.

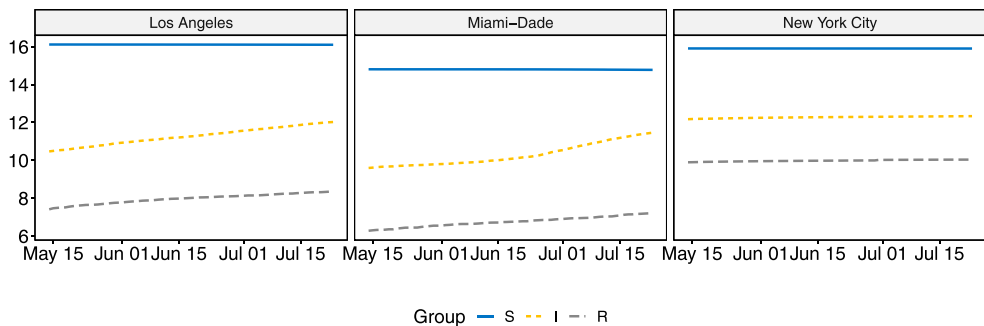


Figure 2. Visualizations for $S(t)$, $I(t)$ and $R(t)$ in the three selected counties on natural log scale.

in SD and WY when compared to other states. As during the studied period, NY is still under lock-down, both the infectious and removal groups experienced slow increases. For CA and FL, however, potentially due to re-open in early May, their infectious and removal groups saw rapid increases. The three counties selected are the metropolitan areas in CA, FL and NY, and the trends observed are similar to those in their respective states.

3. Method

3.1. The SIR and vSIR models

In the SIR model, we consider a fixed total population of size N . By ‘fixed’, we assume that the population size does not vary over time. The effect of natural death or birth are not considered here, as the outstanding period of an infectious disease is much shorter than human average lifetime. Denote, at time t ($t \geq 1$), the counts of susceptible, infectious, and recovered/removed persons within a given region as $S(t)$, $I(t)$ and $R(t)$, respectively, and the relationship $N = S(t) + I(t) + R(t)$ always holds.

Two parameters in the SIR model are time-invariant: the transmission rate β , and the recovering rate γ . The transmission rate β controls how much the disease can be transmitted through exposure. It is jointly determined by the chance of contact and the probability of disease transmission. The recovering rate γ stands for the rate at which infected individuals recover or die. Time-varying property of these two parameters is ignored in traditional SIR modeling, which is a rather strong simplifying assumption that hurdles the model’s prediction power for disease trend. Therefore, we adopt the time-varying SIR [vSIR; 3] framework, where both β and γ are functions of time t .

The vSIR model can be viewed as both a deterministic model and a stochastic model. The deterministic vSIR model allows us to describe the number of people in each compartment with the ordinary differential equations (ODEs). A generalized version of the deterministic vSIR model having infectious rate $\beta(t)$ and removal rate $\gamma(t)$ with respect to time can be described as follows:

$$\begin{aligned}\frac{dS(t)}{dt} &= \frac{-\beta(t)I(t)S(t)}{N}, \\ \frac{dI(t)}{dt} &= \frac{\beta(t)I(t)S(t)}{N} - \gamma(t)I(t), \\ \frac{dR(t)}{dt} &= \gamma(t)I(t).\end{aligned}\tag{1}$$

While the deterministic vSIR model seems appealing due to its simplicity, the spread of a disease, however, is naturally stochastic. Disease transmission between two individuals is random rather than deterministic. The stochastic formulation of the vSIR model is, therefore, preferred for epidemic modeling purposes, as it allows for randomness in the disease spreading process.

3.2. Time fusion SIR model

Consider the the vSIR-Poisson process framework of Sun *et al.* [21] with two time-varying parameters, $\beta(t)$ and $\gamma(t)$. Let N be the total population, $M(t) = I(t) + R(t)$ denote the

cumulative number of diagnosed cases and $\Delta M(t) = M(t) - M(t - 1)$, $\Delta R(t) = R(t) - R(t - 1)$ represent the daily changes of $M(t)$ and $R(t)$. The initial values $\Delta M(1)$ and $\Delta R(1)$ are defaulted, respectively, to $M(1)$ and $R(1)$. Let $t = 1, \dots, T$ denote the time domain. Hence we have

$$\begin{aligned} \Delta M(t) &\sim \text{Poisson} \left(\frac{\beta(t)S(t)I(t)}{N} \right), \\ \Delta R(t) &\sim \text{Poisson} (\gamma(t)I(t)), \quad t = 2, \dots, T. \end{aligned} \tag{2}$$

For most infectious diseases, the infectious and removal rates $\beta(t)$ and $\gamma(t)$ do not always change smoothly over time, as they can be influenced by certain government policies in a notable manner. In other words, $\beta(t)$ and $\gamma(t)$ can fluctuate around a fixed value within a specific time period, and then with the inception of a policy, fluctuate around another value in a period that follows. Identifying the subpopulation structure of these two parameters with time fusion patterns in the SIR model will enhance our understanding of infectious diseases such as COVID-19. In this paper, we assume that successive differences of the infectious rate $\Delta\beta(t) = \beta(t) - \beta(t - 1)$ and removal rate $\Delta\gamma(t) = \gamma(t) - \gamma(t - 1)$ both have an unknown clustered pattern with respect to time. Both $\Delta\beta(1)$ and $\Delta\gamma(1)$ are defaulted to 0. For example, with a cluster of 0's in the successive differences, $\beta(t)$ would remain constant over the corresponding time period. Toward this end, we use three different shrinkage priors on both $\Delta\beta(t)$ and $\Delta\gamma(t)$ to detect such clusters, including the student- t prior, horseshoe prior and spike-and-slab prior [see, 17,19 for more discussion]. As our proposed model focuses on time fusion, we name it hierarchical time fusion SIR (tf-SIR).

The first prior we consider is student- t prior. Despite the popularity of the Laplace prior, it has a light tail, suffers from posterior inconsistency issues [19,20], and often leads to smoothly varying estimation results, i.e. it cannot identify the clustered structure. The student- t prior, with its heavier tail, induces stronger shrinkage effect, and enjoys a nice posterior consistency property. The student- t shrinkage prior on the successive differences can be written as:

$$\begin{aligned} \Delta\beta(t) \mid \sigma_\beta^2 &\sim t_{df_\beta}(l_\beta\sigma_\beta), \quad \sigma_\beta^2 \sim \text{IG}(a_{\sigma_\beta}, b_{\sigma_\beta}), \quad \beta(1) \mid \sigma_\beta^2 \sim \text{N}(0, \sigma_\beta^2\lambda_1), \\ \Delta\gamma(t) \mid \sigma_\gamma^2 &\sim t_{df_\gamma}(l_\gamma\sigma_\gamma), \quad \sigma_\gamma^2 \sim \text{IG}(a_{\sigma_\gamma}, b_{\sigma_\gamma}), \quad \gamma(1) \mid \sigma_\gamma^2 \sim \text{N}(0, \sigma_\gamma^2\eta_1), \quad t = 2, \dots, T, \end{aligned}$$

where T denotes the termination time of observation, $t_{\omega_1}(\omega_2)$ denotes the student- t distribution with degree of freedom ω_1 and scale parameter ω_2 , $\text{N}()$ stands for the normal distribution, and $\text{IG}()$ stands for the inverse gamma distribution. Note that the above student- t distribution can be rewritten as an inverse gamma scaled Gaussian mixture, and hence the tf-SIR model with student- t prior for the sequential differences in $\beta(t)$ and $\gamma(t)$ can be alternatively formulated as:

$$\begin{aligned} \Delta\beta(t) \mid \sigma_\beta^2, \lambda_t &\sim \text{N}(0, \lambda_t\sigma_\beta^2), \quad \lambda_t \sim \text{IG}(a, b), \\ \Delta\gamma(t) \mid \sigma_\gamma^2, \eta_t &\sim \text{N}(0, \eta_t\sigma_\gamma^2), \quad \eta_t \sim \text{IG}(c, d), \\ \sigma_\beta^2 &\sim \text{IG}(a_{\sigma_\beta}, b_{\sigma_\beta}), \quad \sigma_\gamma^2 \sim \text{IG}(a_{\sigma_\gamma}, b_{\sigma_\gamma}), \quad t = 2, \dots, T, \end{aligned} \tag{3}$$

where a, b satisfy conditions $df_\beta = 2a$ and $l_\beta = \sqrt{\frac{b}{a}}$. Similarly, c, d satisfy conditions $df_\gamma = 2c$ and $l_\gamma = \sqrt{\frac{d}{c}}$. Both σ_β^2 and σ_γ^2 are global parameters that shrink the successive differences toward 0. Following common practices, we assume inverse gamma priors for both to impart heavy tails, and keep the probability distribution further from 0 than the Gamma distribution. Different levels of sparsity can be achieved by varying the values of σ_β^2 and σ_γ^2 , with smaller values inducing stronger shrinkage toward 0.

The second prior is the horseshoe prior [2,3], which is a continuous shrinkage prior, and is one of the so called global-local shrinkage prior. It has exhibited ideal theoretical characteristics, and demonstrated good empirical performance [5,24]. Our tf-SIR model with the horseshoe prior can be expressed as:

$$\begin{aligned} \Delta\beta(t) \mid \sigma_\beta^2, \lambda_t &\sim N(0, \lambda_t^2 \sigma_\beta^2), \quad \lambda_t \sim C^+(0, 1), \quad \beta(1) \mid \sigma_\beta^2 \sim N(0, \sigma_\beta^2 \lambda_1), \\ \Delta\gamma(t) \mid \sigma_\gamma^2, \eta_t &\sim N(0, \eta_t^2 \sigma_\gamma^2), \quad \eta_t \sim C^+(0, 1), \quad \gamma(1) \mid \sigma_\gamma^2 \sim N(0, \sigma_\gamma^2 \eta_1), \\ \sigma_\beta^2 &\sim \text{IG}(a_{\sigma_\beta}, b_{\sigma_\beta}), \quad \sigma_\gamma^2 \sim \text{IG}(a_{\sigma_\gamma}, b_{\sigma_\gamma}), \quad t = 2, \dots, T, \end{aligned} \tag{4}$$

where σ_β^2 and σ_γ^2 are same as defined above, and λ_t and η_t are both local parameters following the half-Cauchy distribution $C^+(0, 1)$ that allows, respectively, some $\Delta\beta(t)$ and $\Delta\gamma(t)$ to escape from the shrinkage.

Finally, our third prior of choice is the spike-and-slab prior [7,15]. It is a two component discrete mixture prior. In this paper, we write it as a two-component mixture of Gaussian distributions, and the model is expressed as:

$$\begin{aligned} \Delta\beta(t) \mid \sigma_\beta^2, \lambda_t &\sim \lambda_t N(0, \sigma_\beta^2) + (1 - \lambda_t) N(0, \epsilon^2), \\ \lambda_t &\sim \text{Ber}(p), \quad \beta(1) \mid \sigma_\beta^2 \sim N(0, \sigma_\beta^2 \lambda_1), \\ \Delta\gamma(t) \mid \sigma_\gamma^2, \eta_t &\sim \eta_t N(0, \sigma_\gamma^2) + (1 - \eta_t) N(0, \epsilon^2), \\ \eta_t &\sim \text{Ber}(\pi), \quad \gamma(1) \mid \sigma_\gamma^2 \sim N(0, \sigma_\gamma^2 \eta_1), \\ \sigma_\beta^2 &\sim \text{IG}(a_{\sigma_\beta}, b_{\sigma_\beta}), \quad \sigma_\gamma^2 \sim \text{IG}(a_{\sigma_\gamma}, b_{\sigma_\gamma}), \quad t = 2, \dots, T, \end{aligned} \tag{5}$$

where $\epsilon \ll \sigma_\beta^2$ and $\epsilon \ll \sigma_\gamma^2$, λ_t and η_t are indicators that take values in $\{0, 1\}$, $\text{Ber}()$ denotes the Bernoulli distribution, and again σ_β^2 and σ_γ^2 are same as defined above. In this paper, we fix the inclusion probabilities p and π . In some cases, ϵ is set to 0 so that the spike is taken to a point mass at the origin δ_0 . This distribution can be sensitive to prior choices of the slab width or prior inclusion probability, and therefore we choose the normal distribution with a small variance centered at 0 as the spike.

4. Simulation

4.1. Simulation designs

We use the R package **SimInf** [25] to generate data. Four designs over a time span of 80 days are considered. The time domain is divided into four equally sized pieces each spanning for 20 days, where both the infectious rate $\beta(t)$ and the removal rate $\gamma(t)$ are piecewise constant within each of them. The four designs have different $\beta(t)$, $\gamma(t)$, and population

Table 1. Parameters used in data generation under the four simulation settings.

Design	$\beta(t)$ on pieces 1, 2, 3, 4	$\gamma(t)$ on pieces 1, 2, 3, 4	Population size
Design 1	(0.15, 0.20, 0.10, 0.05)	(0.05, 0.09, 0.10, 0.08)	10^6
Design 2	(0.10, 0.15, 0.10, 0.05)	(0.05, 0.09, 0.10, 0.08)	10^6
Design 3	(0.07, 0.09, 0.08, 0.05)	(0.02, 0.04, 0.06, 0.07)	10^7
Design 4	(0.05, 0.08, 0.05, 0.07)	(0.02, 0.05, 0.04, 0.03)	10^7

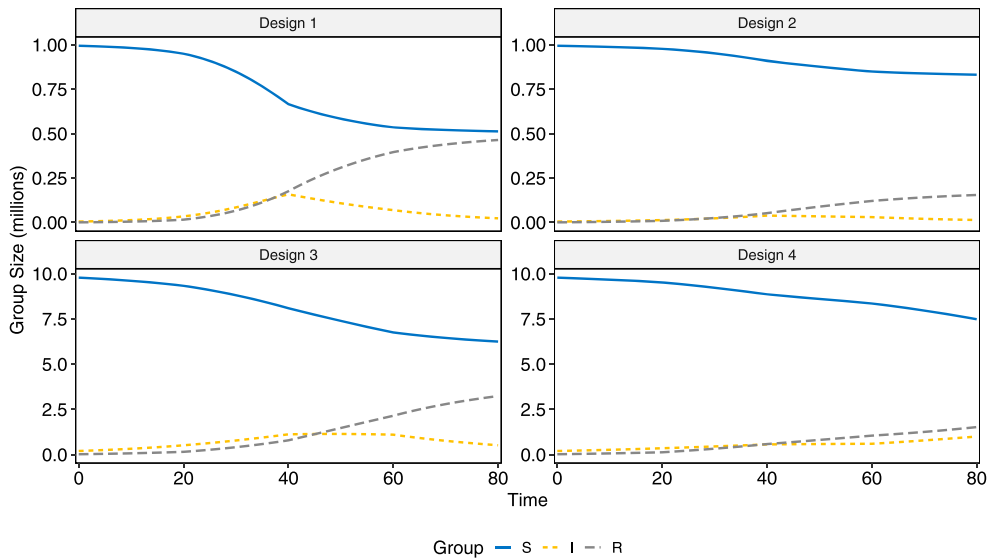


Figure 3. Visualization of example datasets generated under each of the two simulation designs.

size N , and the numerical values are listed in Table 1. Four example datasets, one for each design, are visualized in Figure 3. Design 1 corresponds to a fairly high infectious disease with high removal/recovery rate and design 2 corresponds to a mildly infectious disease with similar removal/recovery rate. Designs 3 and 4 have larger population sizes, and a disease with smaller numerical values for $\beta(t)$ can infect a large portion of the population, such as demonstrated for Design 3. Design 4, with small $\beta(t)$ and $\gamma(t)$, exhibits slow overall development. A total of 100 replicates are performed for each design. In each replicate, the length of the MCMC chain is set to 50,000. As the numerical values for $\beta(t)$ and $\gamma(t)$ in all four designs are not large, it is essential that we get independent posterior samples. To ensure minimal correlation between draws, we set the thinning interval to 10 and set the burn-in to 3000, which leaves us 2000 samples to perform inference.

4.2. Performance measures

The parameter estimates for $\beta(t)$ and $\gamma(t)$ for the 100 replicates are visualized respectively in Figures 4 and 5 as grey lines, and the true underlying values are also plotted in dashed lines. The first observation is that, under all four designs, all three models yield quite accurate parameter estimation performance, as the grey band formed by 100 parameters lie close to or around the dashed line in both plots. Secondly, as the population size in Design 3 and Design 4 is 10 times that in Design 1 and 2, their corresponding grey bands are, overall,

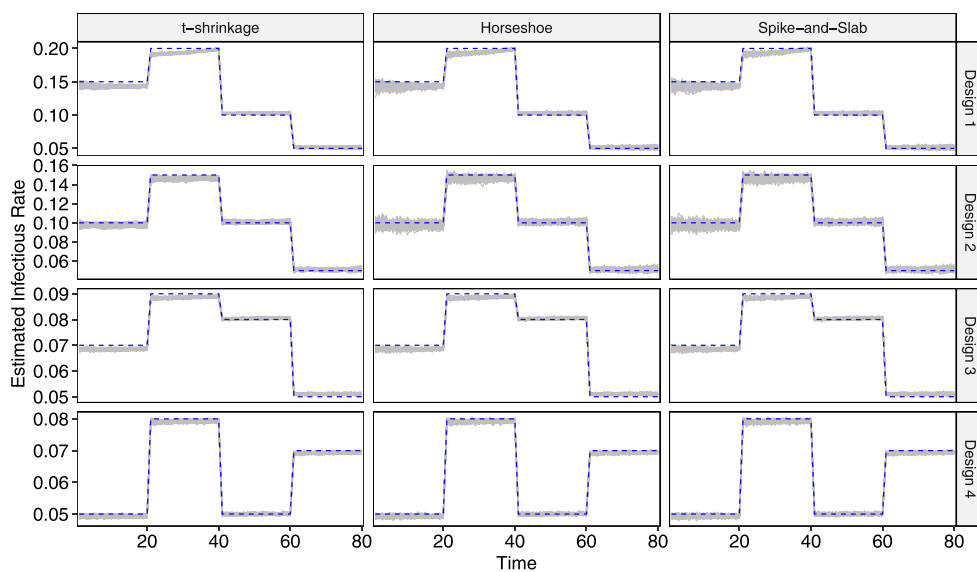


Figure 4. Plot of estimated $\beta(t)$ in 100 replicates for each combination of design and prior. True values are overlaid in dashed lines.

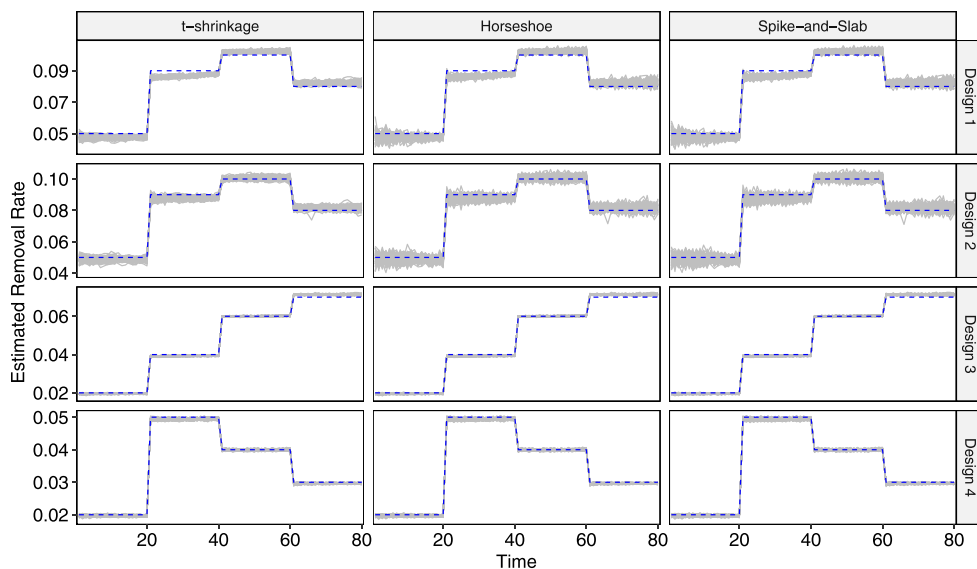


Figure 5. Plot of estimated $\gamma(t)$ in 100 replicates for each combination of design and prior. True values are overlaid in dashed lines.

tighter. Thirdly, in both figures, the grey bands corresponding to the t -shrinkage prior is narrower than those for horseshoe and spike-and-slab, indicating overall relatively stable estimation performance.

The estimation performance, in addition to visually, is also measured numerically. For $\beta(t)$, we apply the following three metrics:

$$\text{MAB}_{\beta}(t) = \frac{1}{100} \sum_{\ell=1}^{100} \left| \hat{\beta}_{\ell}(t) - \beta(t) \right|, \tag{6}$$

$$\text{MSE}_{\beta}(t) = \frac{1}{100} \sum_{\ell=1}^{100} \left(\hat{\beta}_{\ell}(t) - \beta(t) \right)^2, \tag{7}$$

$$\text{SD}_{\beta}(t) = \frac{1}{99} \sum_{\ell=1}^{100} \left(\hat{\beta}_{\ell}(t) - \bar{\hat{\beta}}(t) \right)^2, \tag{8}$$

where $\hat{\beta}_{\ell}(t)$ is the posterior estimate of β at time t in the ℓ th replicate for $\ell = 1, \dots, 100$ and $t = 1, \dots, 80$, and $\bar{\hat{\beta}}(t) = \frac{1}{100} \sum_{\ell=1}^{100} \hat{\beta}_{\ell}(t)$. The metrics for γ are defined in a similar manner, and therefore we omit the details.

The three models are compared in terms of the three performance metrics in Figures 6 and 7. One interesting observation is that the MAB and MSE tend to be large near when $t \in \{20, 40, 60\}$, which corresponds to when changes in parameters occur. They then stabilize as the disease continues to develop. For relatively larger values of the true parameter, the MAB and MSE are larger than for small values of true parameters. As can be observed from the third column in both Figures 6 and 7, the horseshoe and spike-and-slab priors perform similarly in terms of MAB, MSE and SD. When the sample size is 10^6 , the t -shrinkage prior

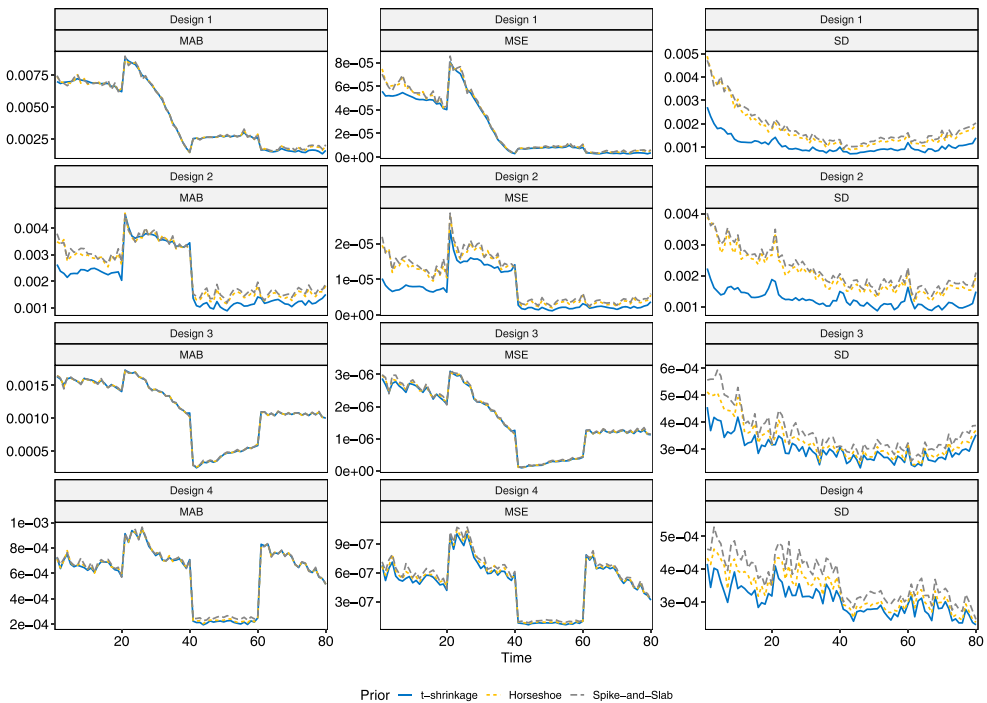


Figure 6. Plot of MAB, MSE and SD of parameter estimate for $\beta(t)$ under different designs.

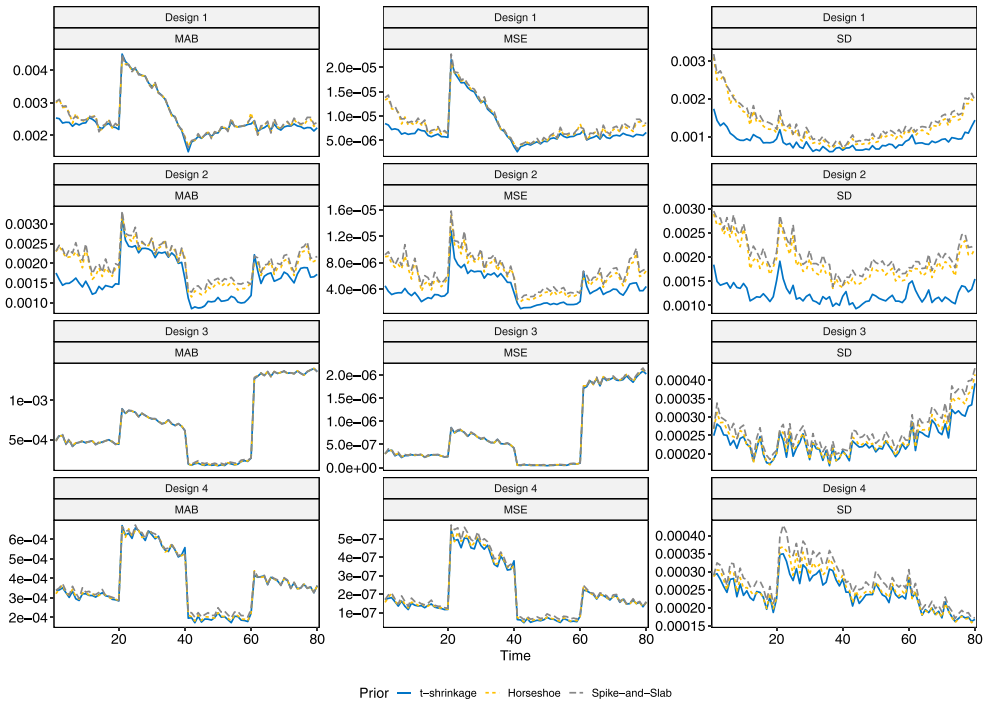


Figure 7. Plot of MAB, MSE and SD of parameter estimate for $\gamma(t)$ under different designs.

yields parameter estimates that are overall more stable and have smaller SD than the other two, which is consistent with the third observation for the grey bands. This difference, however, decreases with increase in sample size.

5. Real data analysis

The proposed methodology is applied on COVID-19 data for both state-level and county level introduced in Section 2. Analysis for other states and counties can be conducted in the same way, which is omitted in this paper. Similar as the simulation studies, the chain length is set to 50,000 with thinning 10, and the first 3000 samples after thinning are treated as burn-in. The estimated infectious rates and removal rates, together with their 95% highest posterior density (HPD) intervals are shown in Figures 8 – 11.

In both state-level and county-level, the three different priors yield similar results. From Figures 8 and 9, we find the infectious rate for NY is smaller than other states during this period. Also, the relatively stable $\hat{\beta}(t)$ for NY after June 6th indicates a potential cluster. FL witnesses a pump peak after mid June, which results from the aggressive reopen in FL. The clustered pattern in other states, however, is not as clear as that in NY, which is partially due to the fact that testing and reporting are conducted timely in NY as it is one of the initial hotspots in March and April that experienced high growth of COVID-19 cases, while in the other states that we considered, there is more delay in testing and reporting. As the number of cumulative cases and daily new cases are large in CA, FL and NY, the HPD band for $\hat{\beta}(t)$ is tight, while in SD and WY, where daily

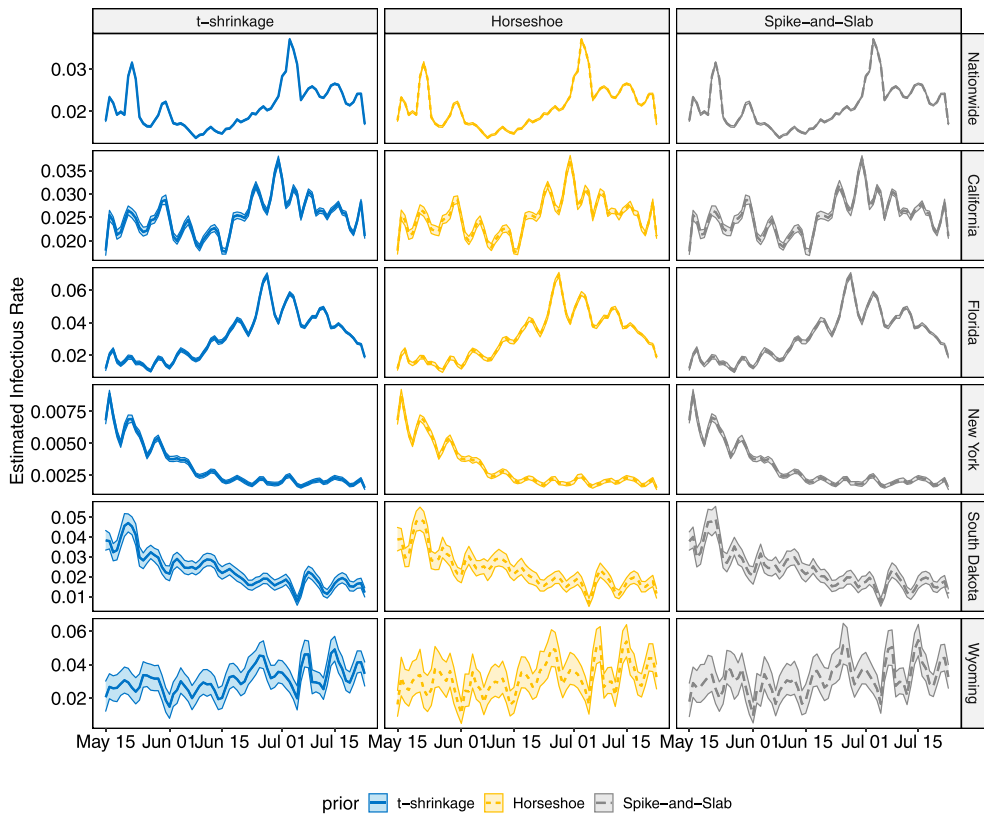


Figure 8. Plot for the estimated infectious rate $\beta(t)$ with 95% HPD intervals for US nationwide and five individual states over the studied time frame.

new cases do not exceed, respectively, 100 and 50, the estimated HPD band is much wider. As for the removal rate, NY, FL and CA have similar result. Despite the differences in numerical values, the trends of $\hat{\gamma}(t)$ for NY and FL are similar, and display a weekly seasonality, with the estimated removal rates being smaller than average on weekends. This is due to the fact that reporting is less active during weekends than during the week. Note that for SD, the estimated removal rate clearly shows a relatively stable pattern between June 20th and July 1st, indicating the existence of a potential cluster. The nation-wide removal rate estimate is constant across the time and has few peaks during this time period, as a few states release the recovered cases in a cumulative manner on a certain day. On the county level, $\hat{\beta}(t)$ for New York City becomes stable after June 15th. The infectious rate estimate for Miami-Dade remained relatively low before June 22nd but started to increase, and remained relatively high. For Los Angeles, $\hat{\beta}(t)$ experiences fluctuations with weekly seasonality, but the overall trend remains stable. The estimated removal rates are very small in all three counties, and all have wide HPD bands, since the removal group only contains deaths in county-level data. The sudden jump in Figure 11 for New York City corresponds to the release of 633 death cases, which was due to data anomaly.

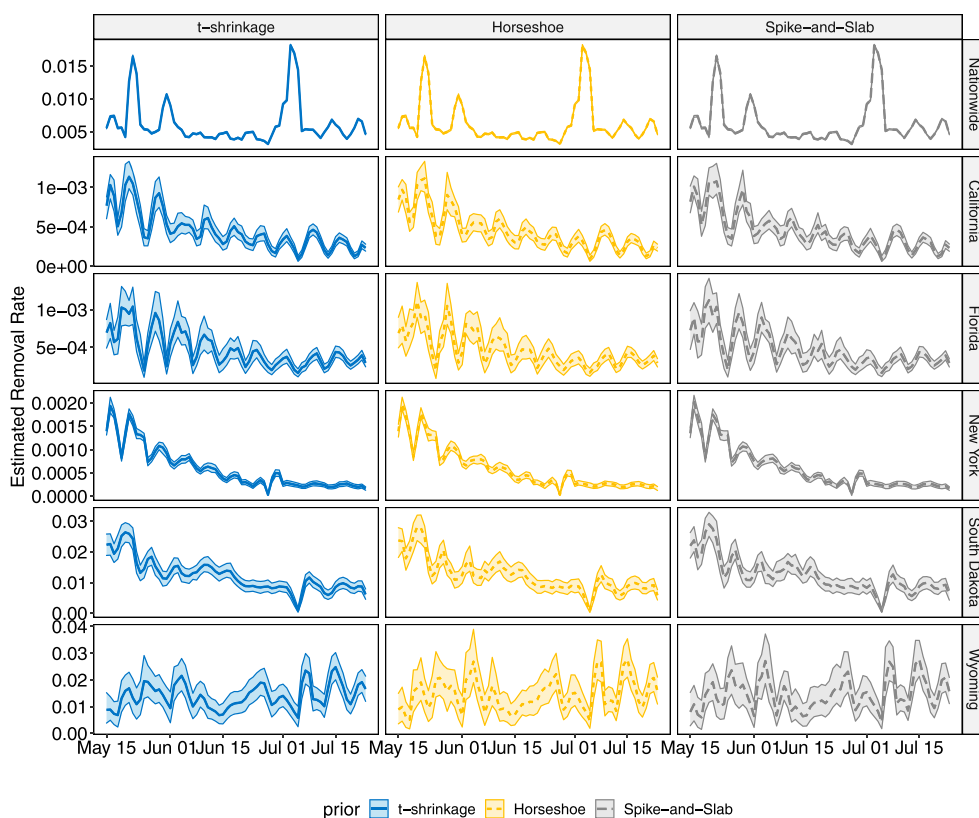


Figure 9. Plot for the estimated removal rate $\gamma(t)$ with 95% HPD intervals for US nationwide and five individual states over the studied time frame.

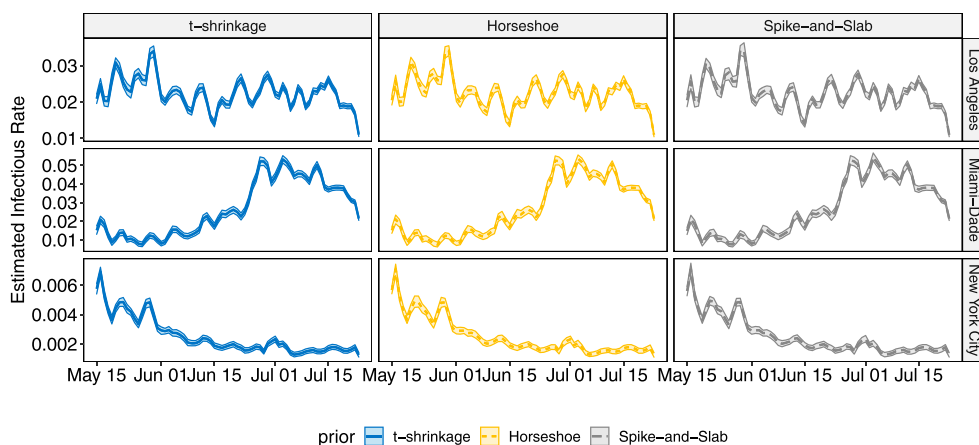


Figure 10. Plot for the estimated infectious rate $\beta(t)$ with 95% HPD intervals for the three selected counties over the studied time frame.

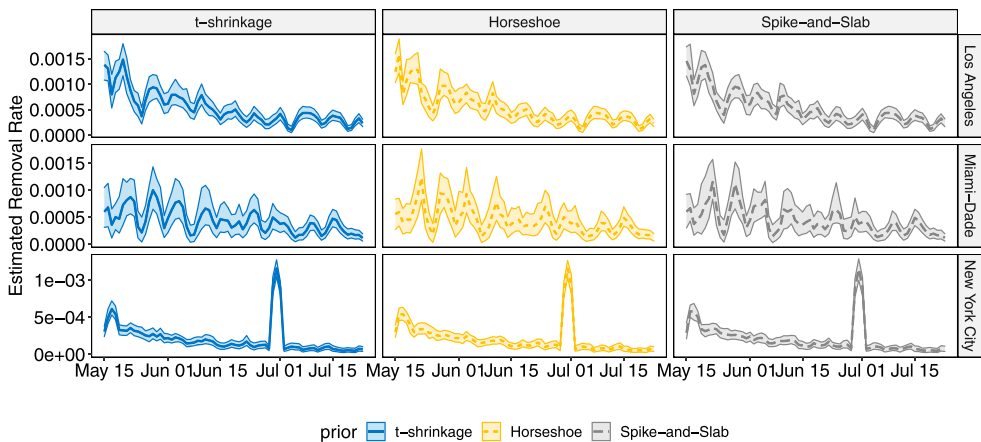


Figure 11. Plot for the estimated removal rate $\gamma(t)$ with 95% HPD intervals for the three selected counties over the studied time frame.

6. Conclusion

In this paper, we proposed the tf-SIR model to capture group structure for infectious rate and removal rate of different time period by using Bayesian shrinkage priors. To our best knowledge, this is the first attempt in literature to use Bayesian shrinkage to recover unknown grouping structure for SIR model. Our simulation results indicate that the proposed method has reasonable performance and ability to capture the group pattern of infectious rate and removal rate. The analysis of COVID-19 data also brings in new understanding of the infectious disease such as COVID-19. Also, our tf-SIR model can not only be used to model and assess COVID-19 pandemic but also other epidemic.

One interesting consideration, as suggested by one anonymous reviewer, is to use a supermartingale structure, i.e. non-increasing function with respect to time, to model the infectious rate $\beta(t)$ because of the implementation of policy intervention. While such assumption might not be met by the development of COVID-19 in the United States, it is a quite reasonable assumption when studying the course of development for COVID-19 in countries where quarantine and stay-at-home policies are strictly enforced such as China, Singapore, and Vietnam.

In addition, three topics beyond the scope of this paper are worth further investigation. First, in our real data application, a moving average approach is applied to deal with measurement errors for observed data. Proposing a measurement error model with SIR is an interesting future work. Furthermore, different states may have similar infection and removal pattern. Subgroup detection for different states will help the government design its policies. Finally, discovering theoretical guarantees such as posterior concentration rates of proposed methods is also devoted to future research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Hou-Cheng Yang  <http://orcid.org/0000-0002-8679-4280>

Yishu Xue  <http://orcid.org/0000-0002-9660-6087>

Guanyu Hu  <http://orcid.org/0000-0003-1410-1665>

References

- [1] G. Casella, M. Ghosh, J. Gill, and M. Kyung, *Penalized regression, standard errors, and Bayesian lassos*, *Bayesian Anal.* 5 (2010), pp. 369–411.
- [2] C.M. Carvalho, N.G. Polson, and J.G. Scott, *Handling Sparsity via the Horseshoe*, in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, Vol. 5, D. van Dyk and M. Welling, eds., *Proceedings of Machine Learning Research*, pp. 16–18 Apr, Hilton Clearwater Beach Resort, Clearwater Beach, Florida, USA, PMLR, 2009, pp. 73–80.
- [3] C.M. Carvalho, N.G. Polson, and J.G. Scott, *The horseshoe estimator for sparse signals*, *Biometrika* 97 (2010), pp. 465–480.
- [4] Y.C. Chen, P.E. Lu, C.S. Chang and T.H. Liu, *A time-dependent SIR model for COVID-19 with undetectable infected persons*, *IEEE Trans. Netw. Sci. Eng.* 7 (2020), pp. 3279–3294.
- [5] J. Datta and J.K. Ghosh, *Asymptotic properties of Bayes risk for the horseshoe prior*, *Bayesian Anal.* 8 (2013), pp. 111–132.
- [6] P. de Valpine, D. Turek, C.J. Paciorek, C. Anderson-Bergman, D.T. Lang, and R. Bodik, *Programming with models: Writing statistical algorithms for general model structures with NIMBLE*, *J. Comput. Graph. Stat.* 26 (2017), pp. 403–413.
- [7] E.I. George and R.E. McCulloch, *Variable selection via Gibbs sampling*, *J. Amer. Statist. Assoc.* 88 (1993), pp. 881–889.
- [8] E. Guidotti and D. Ardia, *COVID-19 data hub*, 2020, working paper. Available at <https://covid19datahub.io>.
- [9] H.W. Hethcote, *The mathematics of infectious diseases*, *SIAM Rev.* 42 (2000), pp. 599–653.
- [10] G. Hu and J. Geng, *Heterogeneity learning for SIRS model: An application to the COVID-19*, *Stat. Interface* 14 (2021), pp. 73–81.
- [11] S.Y. Jung, H. Jo, H. Son, and H.J. Hwang, *Real-world implications of a rapidly responsive COVID-19 spread model with time-dependent parameters via deep learning: Model development and validation*, *J. Med. Internet. Res.* 22 (2020), p. e19907.
- [12] W.O. Kermack and A.G. McKendrick, *A contribution to the mathematical theory of epidemics*, *Proc. R. Soc. Lond. Ser. A.* 115 (1927), pp. 700–721.
- [13] W.O. Kermack and A.G. McKendrick, *Contributions to the mathematical theory of epidemics. II. The problem of endemicity*, *Proc. R. Soc. Lond. Ser. A* 138 (1932), pp. 55–83.
- [14] W.O. Kermack and A.G. McKendrick, *Contributions to the mathematical theory of epidemics. III. Further studies of the problem of endemicity*, *Proc. R. Soc. Lond. Ser. A* 141 (1933), pp. 94–122.
- [15] T.J. Mitchell and J.J. Beauchamp, *Bayesian variable selection in linear regression*, *J. Amer. Statist. Assoc.* 83 (1988), pp. 1023–1032.
- [16] T. Park and G. Casella, *The Bayesian lasso*, *J. Amer. Statist. Assoc.* 103 (2008), pp. 681–686.
- [17] J. Piironen and A. Vehtari, *Sparsity information and regularization in the horseshoe and other shrinkage priors*, *Electron. J. Stat.* 11 (2017), pp. 5018–5051.
- [18] J.M. Read, J.R. Bridgen, D.A. Cummings, A. Ho, and C.P. Jewell, *Novel coronavirus 2019-nCoV: Early estimation of epidemiological parameters and epidemic predictions*, *MedRxiv* (2020). doi:10.1101/2020.01.23.20018549.
- [19] Q. Song and G. Cheng, *Bayesian fusion estimation via t shrinkage*, *Sankhya A* (2019), pp. 1–33.
- [20] Q. Song and F. Liang, *Nearly optimal Bayesian shrinkage for high dimensional regression*, *arXiv preprint arXiv:1712.08964* (2017).
- [21] H. Sun, Y. Qiu, H. Yan, Y. Huang, Y. Zhu, and S.X. Chen, *Tracking reproducibility of COVID-19 epidemic in China with varying coefficient SIR model*, *J. Data Sci.* 18 (2020), pp. 455–472.

- [22] B. Tang, X. Wang, Q. Li, N.L. Bragazzi, S. Tang, Y. Xiao, and J. Wu, *Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions*, *J. Clin. Med.* 9 (2020), pp. 462.
- [23] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso*, *J. R. Stat. Soc. Ser. B. (Stat. Methodol.)* 67 (2005), pp. 91–108.
- [24] S.L. Van Der Pas, B.J. Kleijn, and A.W. Van Der Vaart, *The horseshoe estimator: Posterior concentration around nearly black vectors*, *Electron. J. Stat.* 8 (2014), pp. 2585–2618.
- [25] S. Widgren, P. Bauer, R. Eriksson, and S. Engblom, *SimInf: An R package for data-driven stochastic disease spread simulations*, *J. Stat. Softw.* 91 (2019), pp. 1–42.
- [26] World Health Organization, *WHO coronavirus disease (COVID-19) dashboard*, 2020. Available at <https://covid19.who.int/> (accessed 20 July 2020).
- [27] J.T. Wu, K. Leung, and G.M. Leung, *Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study*, *Lancet* 395 (2020), pp. 689–697.