Statistics and Its Interface Volume 14 (2021) 73–81

Heterogeneity learning for SIRS model: an application to the COVID-19

Guanyu Hu*† and Junxian Geng

We propose a Bayesian Heterogeneity Learning approach for Susceptible-Infected-Removal-Susceptible (SIRS) model that allows underlying clustering patterns for transmission rate, recovery rate, and loss of immunity rate for the latest corona virus (COVID-19) among different regions. Our proposed method provides simultaneously inference on parameter estimation and clustering information which contains both number of clusters and cluster configurations. Specifically, our key idea is to formulates the SIRS model into a hierarchical form and assign the Mixture of Finite mixtures priors for heterogeneity learning. The properties of the proposed models are examined and a Markov chain Monte Carlo sampling algorithm is used to sample from the posterior distribution. Extensive simulation studies are carried out to examine empirical performance of the proposed methods. We further apply the proposed methodology to analyze the state level COVID-19 data in U.S.

Keywords and phrases: Bayesian nonparametric, Cluster learning, Infectious diseases, MCMC, Mixture of finite mixtures.

1. INTRODUCTION

The Coronavirus Disease 2019 (COVID-19) has created a profound public health emergency around world. It has become an epidemic with more than 5,000,000 confirmed infections worldwide as of May 21 2020. The spreading speed of COVID-19 which is caused by a new coronavirus is faster than severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS). Recently, the risk of COVID-19 has been a significant public-health concern and people pay more attention on precise and timely estimates and predictions of COVID-19. The Susceptible-Infectious-Recovered (SIR) model and its variation approaches, such as Susceptible-Infected-Removal-Susceptible (SIRS) [15, 16] and Susceptible-Exposed-Infected-Removal (SEIR) model [11], have been widely discussed to study the dynamical evolution of an infectious disease in a certain region. There are rich literatures producing early results on COVID-19 based on SIR model and its variations [26, 21, 24]. From statistician's perspectives, building a time-varying model under SIR and its variations is also fully discussed for COVID-19 [5, 23, 13]. In most existing literature, people focus more on dynamic regimes of the SIR models for COVID-19. They lack discussions on heterogeneity pattern of COVID-19 among different regions.

The aim of this paper is to propose a new hierarchical SIRS model for detecting heterogeneity pattern of COVID-19 among different regions under a Bayesian framework. Bayesian nonparametric methods such as Dirichlet process (DP) offer choices to do simultaneous inference on parameters' estimation and parameters' heterogeneity information which contains the number of clusters and clustering configurations. Compared with existing approaches such as finite mixtures models, Bayesian nonparametric approach does not need to pre-specify the number of clusters, which provides probabilistic framework for simultaneous inference of the number of clusters and the clustering labels. Miller and Harrison [17] points out that the estimation of the number of clusters under Dirichlet process mixture (DPM) model is inconsistent which will produce extremely small clusters. One remedy for over-clustering problem under DPM is mixture of finite mixtures model (MFM) [18]. The clustering properties of MFM are fully discussed in Miller and Harrison [18], Geng et al. [9] and it has been widely applied in different areas such as regional economics [12], environmental science [10], and social science [9]. Thus, the key idea of this paper is to assign MFM priors on different parameters of the SIRS model to capture the heterogeneity of parameters among different regions. The contribution of this paper are two-fold. First, we formulate a Bayesian heterogeneity learning model for SIRS under MFM. To our best knowledge, this is the first time when MFM is applied into epidemiology models such as SIRS. Our proposed Bayesian approach successfully captures the heterogeneity of three different parameters under the SIRS model among different regions while also considering uncertainty in estimation of the number of clusters. Several interesting findings based on our proposed method are discovered for COVID-19 data in US.

This paper is organized as follows. Section 2 presents the motivating data we analyze. We discuss our proposed Bayesian hierarchical model for heterogeneity learning under SIRS model framework in Section 3. The performance of our proposed method is illustrated via simulation studies

arXiv: 2007.08047

*Corresponding author.

†ORCID: 0000-0003-1410-1665.

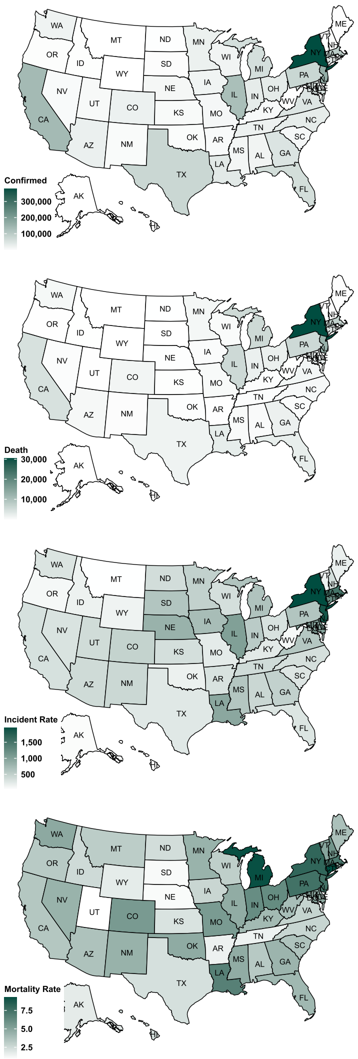

Figure 1. Exploratory Analysis of COVID-19 on June 10th.

in Section 4. Section 5 is devoted to the analysis of state level COVID-19 data in U.S. A brief discussion is presented in Section 6.

2. MOTIVATING DATA

Our motivating data comes from the COVID tracking project https://covidtracking.com. State Level COVID-19 Data are recorded for the 50 states plus Washington, DC. For simplicity, we refer to them as "51 states" in the rest of this paper. Up to June 10th, 2020, United States totally confirmed 2,043,031 cases. 114,533 people died because of COVID-19, and 607,279 people are recovered from COVID-19. The fatality rate of COVID-19 is 5.6% (i.e., the number of COVID-19 deaths divided by the number of confirmed COVID-19 cases).

Figure 1 shows state level confirmed numbers, death numbers, incident rate, and mortality rate. We can see that New York state has the highest confirmed number, death numbers, and incident rate; Connecticut has the highest mortality rate among 51 states; Montana has the least confirmed number; Alaska has the least death number; the incident rate of Hawaii is lowest among 51 states; and Texas has the lowest mortality rate.

3. METHOD

3.1 SIRS model

Compartment epidemic models are widely used to study infectious disease which spreads through human populations across a large region. SIR model [14] has been universally discussed for analyzing the dynamical evolution of an infectious disease in a large population. SIR model is extended to SIRS model for imperfect immunity situation [15, 16]. For a given time t, a fixed population can be split into three compartments: $S(t)$, $I(t)$, and $R(t)$, which denotes the number of susceptibles, the number of infectious, and the number of "recovereds" (which includes deaths), respectively. The dynamical process of SIRS model can be written as following nonlinear ordinary differential equations of three given compartments

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI/N + \phi R,\\ \frac{dI}{dt} &= \beta SI/N - \gamma I,\\ \frac{dR}{dt} &= \gamma I - \phi R,\end{aligned} \tag{1}$$

where β denotes the average rate of contact per unit time multiplied by the probability of disease transmission per contact between a susceptible and an infectious subject, γ denotes the rate of "recovery" per unit time, which is the rate at which infectious individuals are removed from being infectious due to recovery (or death), and ϕ denotes the rate of loss of immunity of recovered individuals per unit time, which is the rate at which recovered individuals become susceptible again [1, 28]. By adding the equations in (1), we notice that

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0.$$

Thus, the model postulates a fixed total population without entry and exits of demographic type. For example, there are no births or deaths caused by other than the disease we study in a certain time. The SIRS model assume the sum of all three compartments together is constant within a short period of time such that

$$S(t) + I(t) + R(t) = N, \tag{2}$$

where N is a fixed total population. In cases with discrete time $t = 1, \dots, T$ (in units of days), we have

$$\begin{aligned} S(t+1) &= S(t) - \beta S(t)I(t)/N + \phi R(t), \\ I(t+1) &= I(t) + \beta S(t)I(t)/N - \gamma I(t), \\ R(t+1) &= R(t) + \gamma I(t) - \phi R(t), \end{aligned} \tag{3}$$

with the same constraints as (2).

Based on the models in (3) and (2) and assumptions in [8], the data model of SIRS assumes conditional independent Poisson distributions evolving at discrete time points. For a given time $t = 1, \dots, T$, the data models are

$$Z_R(t)|P_R(t) \sim \text{Possion}(N \times P_R(t)), \tag{4}$$

and

$$Z_I(t)|P_I(t) \sim \text{Possion}(N \times P_I(t)), \tag{5}$$

where $Z_R(t)$ and $Z_I(t)$ are the observed number of "recovereds" (includes deaths) and infectious individuals at time t, respectively; N is known total number of population and $Z_S(t) = N - Z_I(t) - Z_R(t)$; $P_R(t)$ and $P_I(t)$ are underlying true rates of recovered and infectious individuals. Thus, our observed data are $\{(Z_R(t), Z_I(t))\} : t = 1, 2 \dots, T$. Based on the relationship between the number of "recovereds", infectious, and suspects, we have

$$P_R(t) + P_I(t) + P_S(t) = 1, \tag{6}$$

where $P_S(t)$ represents the underlying rate of susceptible individuals.

Similar to [28], we have following hidden processes:

$$\begin{aligned} P_R(t+1) &= P_R(t) + \gamma P_I(t) - \phi P_R(t), \\ P_I(t+1) &= P_I(t) + \beta P_S(t)P_I(t) - \gamma P_I(t), \\ P_S(t+1) &= P_S(t) - \beta P_S(t)P_I(t) + \phi P_R(t). \end{aligned} \tag{7}$$

In order to model the hidden uncertainties in SIRS model, we define following transformation of $P_R(t)$, $P_I(t)$ and $P_S(t)$ based on (6)

$$\begin{aligned} W_S(t) &\equiv \log(\frac{P_S(t)}{P_R(t)}), \\ W_I(t) &\equiv \log(\frac{P_I(t)}{P_R(t)}). \end{aligned} \tag{8}$$

The time-varying process of $W_R(t)$ and $W_I(t)$ is defined as

$$\begin{aligned} W_S(t+1) &= \mu_S(t) + \epsilon_S(t+1), \\ W_I(t+1) &= \mu_I(t) + \epsilon_I(t+1), \end{aligned} \tag{9}$$

where $\epsilon_S(t) \sim N(0, \sigma_S^2)$ and $\epsilon_I(t) \sim N(0, \sigma_I^2)$ for $t = 1, 2, \dots, T$. Based on (6) and (7), we have

$$\begin{aligned} \mu_S(t) =& W_S(t) \\ &+ \log(1 + \frac{\phi}{\exp(W_S(t))} \\ &- \frac{\beta \exp(W_I(t))}{1 + \exp(W_S(t)) + \exp(W_I(t))}) \\ &+ \log(\frac{1}{1 + \gamma \exp(W_I(t)) - \phi}) \end{aligned} \tag{10}$$

and

$$\begin{aligned} \mu_I(t) =& W_I(t) \\ &+ \log(1 - \gamma + \frac{\beta \exp(W_S(t))}{1 + \exp(W_S(t)) + \exp(W_I(t))}) \\ &+ \log(\frac{1}{1 + \gamma \exp(W_I(t)) - \phi}) \end{aligned} \tag{11}$$

Based on the transformation in (8), we have our data in (4) and (5) as

$$\begin{aligned} &Z_R(t)|W_S(t), W_I(t) \\ &\sim \text{Poisson}\left(N \times \frac{1}{1 + \exp(W_S(t)) + \exp(W_I(t))}\right), \\ &Z_I(t)|W_S(t), W_I(t) \\ &\sim \text{Poisson}\left(N \times \frac{\exp(W_I(t))}{1 + \exp(W_S(t)) + \exp(W_I(t))}\right). \end{aligned} \tag{12}$$

For the simplicity, we refer the model from (9) to (12) as $\{(Z_R(t), Z_I(t), N), t = 1, 2 \dots, T\} \sim \text{SIRS}(\beta, \gamma, \phi, \sigma_S^2, \sigma_I^2)$. Based on the transmission rate and recover rate, the basic reproduction number, R_0, can be calculated by

$$R_0 = \frac{\beta}{\gamma}. \tag{13}$$

3.2 Heterogeneity learning

In Section 2, our motivating data is at state level in US and we are interested in whether there are heterogeneity patterns on the parameters of interest among different states. As an assumption, we believe that different states might have different parameters, however, some states will share similar pattern in transmission rate, recovery rate, or loss of immunity rate. Next, we introduce nonparametric Bayesian methods for heterogeneity learning of SIRS parameters over n different regions. In this section, we focus on the the transmission rate β for different regions. Recovery rate and loss of immunity rate can be parameterized in the same way.

Let $z_1, \dots, z_n \in \{1, \dots, k\}$ denote clustering labels of n regions and $\beta_1, \dots, \beta_n$ denote the corresponding parameters in SIRS model for n regions. Our goal is to cluster $\beta_1, \dots, \beta_n$ into k clusters with distinct values $\beta_1^*, \dots, \beta_k^*$, which is usually unknown in practice. A popular solution for unknown

k is to introduce the Dirichlet process mixture prior models [2] as following:

$$\beta_i \sim G, \quad G \sim DP(\alpha G_0), \tag{14}$$

where G_0 is a base measure and α is a concentration parameter. If a set of values of $\beta_1, \dots, \beta_n$ are drawn from G, a conditional prior can be obtained by integration [3]:

$$\begin{aligned} & p(\beta_{n+1} \mid \beta_1, \dots, \beta_n) \\ & = \frac{1}{n+\alpha} \sum_{j=1}^{n} \delta_{\beta_j}(\beta_{n+1}) + \frac{\alpha}{n+\alpha} G_0(\beta_{n+1}). \end{aligned} \tag{15}$$

Here, $\delta_{\beta_j}(\beta_\ell) = I(\beta_\ell = \beta_j)$ is a point mass at β_j. We can obtain the following equivalent models by introducing cluster membership z_j's and letting the unknown number of clusters k go to infinity [19].

$$\begin{aligned} z_i \mid \boldsymbol{\pi} &\sim \text{Discrete}(\pi_1, \dots, \pi_k), \\ \beta_c^* &\sim G_0 \\ \boldsymbol{\pi} &\sim \text{Dirichlet}(\alpha/k, \dots, \alpha/k) \end{aligned} \tag{16}$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$. In addition, the distribution of z_i can be marginally given by a stick-breaking representation [22] of Dirichlet process (DP) as

$$\begin{aligned} z_i &\sim \sum_{h=1}^{\infty} \pi_h \delta_h, \\ \pi_h &= \nu_h \prod_{\ell \le h} (1 - \nu_\ell), \\ \nu_h &\sim \text{Beta}(1, \alpha), \end{aligned} \tag{17}$$

where δ_h is the Dirac function with mass at h.

However, [18] proved that the DP mixture model produces extraneous clusters in the posterior leading to inconsistent estimation of the *number of clusters* even when the sample size grows to infinity. A modification of DP mixture model called Mixture of finite mixtures (MFM) model is proposed to circumvent this issue [18]:

$$\begin{aligned} k &\sim p(\cdot) \\ (\pi_1, \dots, \pi_k) \mid k &\sim \text{Dirichlet}(\eta, \dots, \eta) \\ z_i \mid k, \boldsymbol{\pi} &\sim \sum_{h=1}^{k} \pi_h \delta_h, \quad i = 1, \dots, n, \end{aligned} \tag{18}$$

where $p(\cdot)$ is a proper probability mass function (p.m.f.) on $\{1, 2, \dots, \}$.

The conditional prior of β's under MFM can be stated as below:

$$\begin{aligned} & p(\beta_{n+1} \mid \beta_1, \dots, \beta_n) \\ & \propto \sum_{k=1}^{w} (n_k + \eta) \delta_{\beta_k^*} + \frac{V_n(w+1)}{V_n(w)} \eta G_0(\beta_{n+1}). \end{aligned} \tag{19}$$

where $\beta_1^*, \dots, \beta_w^*$ are the distinct values taken by $\beta_1, \dots, \beta_n$, w is the number of existing clusters and n_k is the size of cluster k.

Like the stick-breaking representation in (17) of Dirichlet process, the MFM also has a similar construction. If we choose $k - 1 \sim \text{Poisson}(\lambda)$ and $\eta = 1$ in (18), the mixture weights $\pi_1, \cdots, \pi_k$ is constructed as follows:

1. Generate $\eta_1, \eta_2, \cdots \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$,
2. $k = \min\{j : \sum_{i=1}^{j} \eta_i \ge 1\}$,
3. $\pi_i = \eta_i$, for $i = 1, \cdots, k-1$,
4. $\pi_k = 1 - \sum_{i}^{k-1} \pi_i$.

For ease of exposition, we refer the stick-breaking representation of MFM above as MFM(λ) with default choice of $p(\cdot)$ being Poisson(λ) and $\eta = 1$.

3.3 Hierarchical model

In order to allow for simultaneously heterogeneity learning of three parameters in SIRS model, the MFM prior is introduced for parameters β, γ and ϕ in the SIRS model. Our observed data are $\{(Z_R(t, \boldsymbol{s}_i), Z_I(t, \boldsymbol{s}_i), N_i) : t = 1, 2, \dots, T, i = 1, 2, \dots, n\}$, where t denotes each discrete time point, i denotes each state, and $\boldsymbol{s}_i$ denotes the location for state i. The hierarchical SIRS model with MFM is given as

$$\begin{aligned} & \{(Z_R(t, \boldsymbol{s}_i), Z_I(t, \boldsymbol{s}_i), N_i), t = 1, 2 \dots, T\} \sim \\ & \text{SIRS}(\beta_{z_i^\beta}, \gamma_{z_i^\gamma}, \phi_{z_i^\phi}, \sigma_{S,i}^2, \sigma_{I,i}^2), i = 1, 2, \dots, n \\ & z_i^\beta \sim \text{MFM}(\lambda_\beta), i = 1, 2, \dots, n, \\ & z_i^\gamma \sim \text{MFM}(\lambda_\gamma), i = 1, 2, \dots, n, \\ & z_i^\phi \sim \text{MFM}(\lambda_\phi), i = 1, 2, \dots, n, \\ & \beta_{z_i^\beta} \sim G_\beta, \\ & \gamma_{z_i^\gamma} \sim G_\gamma, \\ & \phi_{z_i^\phi} \sim G_\phi, \\ & \sigma_{S,i}^2, \sigma_{I,i}^2 \sim \text{IG}(0.01, 0.01), i = 1, 2, \dots, n, \end{aligned} \tag{20}$$

where z^β, z^γ, and z^ϕ denote the cluster assignments of parameter β, γ, and ϕ, respectively. G_β, G_γ, and G_ϕ is the base distribution for parameter β, γ, and ϕ, respectively. The choices of G_β, G_γ, and G_ϕ will be discussed in Section 3.4.

3.4 Prior and posterior

For the hierarchical SIRS model with MFM introduced in Section 3.3, the set of parameters is denoted as $\Theta = \{\beta_{z_i^\beta}, \gamma_{z_i^\gamma}, \phi_{z_i^\phi}, \sigma_{S,i}^2, \sigma_{I,i}^2, \lambda_\beta, \lambda_\gamma, \lambda_\phi : i = 1, 2 \dots, n\}$. To complete the model, we now specify the joint prior distribution for the parameters. Based on the natural constraints generated by (3), we have following distribution for bases distribution G_β, G_γ and G_ϕ, respectively:

$$\begin{aligned} \beta_{z_i^\beta} &\sim \text{Uniform}(0,1), \\ \gamma_{z_i^\gamma} &\sim \text{Uniform}(0,1), \\ \phi_{z_i^\phi} &\sim \text{Uniform}(0,1). \end{aligned} \tag{21}$$

For the hyperparameters for three MFM processes, we assign gamma prior Gamma$(1,1)$ on $\lambda_\beta, \lambda_\gamma, \lambda_\phi$. With the joint prior distributions $\pi(\Theta)$, the posterior distribution of these parameters based on the data $D = \{(Z_R(t,\boldsymbol{s}_i), Z_I(t,\boldsymbol{s}_i), N_i) : t = 1,2,\ldots,T, i = 1,2,\ldots,n\}$ is given as

(22)

$$\begin{aligned} &\pi(\Theta|(Z_R(t,\boldsymbol{s}_i), Z_I(t,\boldsymbol{s}_i), N_i) : t = 1,2,\ldots,T, i = 1,2,\ldots,n) \\ &\propto L(D|\Theta) \times \pi(\Theta), \end{aligned}$$

where $L(D|\Theta)$ is the full data likelihood given from the model (9) to (12). The analytical form of the posterior distribution of $\pi(\Theta|(Z_R(t,\boldsymbol{s}_i), Z_I(t,\boldsymbol{s}_i), N_i) : t = 1,2,\ldots,T, i = 1,2,\ldots,n)$ is unavailable. Therefore, we carry out the posterior inference using the MCMC sampling algorithm to sample from the posterior distribution and then obtain the posterior estimates of the unknown parameters. Computation is facilitated by the **nimble** package [7] in R which generates **C++** code for faster computation.

3.5 Group inference via MCMC samples

After obtaining posterior samples, an important task is to do inference on posterior samples. Using posterior mean or posterior median for grouping label $\boldsymbol{z}$ is not suitable. Instead, inference on the clustering configurations is obtained employing the modal clustering method of [6]. The inference is based on the membership matrices of posterior samples, $B^{(1)}, \ldots, B^{(M)}$, where $B^{(t)}$ for the t-th post-burn-in MCMC iteration is defined as:

(23)

$$B^{(t)} = [B^{(t)}(i,j)]_{i,j\in\{1:n\}} = 1(z_i^{(t)} = z_j^{(t)})_{n\times n}, \quad t = 1,\ldots,M.$$

Here $1(\cdot)$ denotes the indicator function, which means $B^{(t)}(i,j) = 1$ indicates observations i and j are in the same cluster in the t-th posterior sample post burn-in. After obtaining the membership matrices of the posterior samples, a Euclidean mean for membership matrices is calculated by:

$$\bar{B} = \frac{1}{M}\sum_{t=1}^{M} B^{(t)}.$$

Based on $\bar{B}$ and $B^{(1)}, \ldots, B^{(M)}$, we find the iteration with the least squares distance to $\bar{B}$ as

$$C_L = \text{argmin}_{t\in(1:M)} \sum_{i=1}^{n}\sum_{j=1}^{n}\{B(i,j)^{(t)} - \bar{B}(i,j)\}^2. \tag{24}$$

The estimated parameters, together with the cluster assignments $\boldsymbol{z}$, are then extracted from the C_L-th post burn-in iteration.

4. SIMULATION

In this section, we investigate the performance of the hierarchical SIRS model with MFM from a variety of measures.

4.1 Simulation settings and evaluation metrics

In order to mimic the real dataset we analyze, we choose $n = 51$ and the population for each location is assigned as the real data population for 51 states. The time length T equals 30 for all the simulation replicates. The total number of replicates in our simulation study is 100. For each parameter, we have two different groups and we set the true values of the parameters $\beta_1 = 0.06, \beta_2 = 0.6$, $\phi_1 = 0.06, \phi_2 = 0.6$, and $\gamma_1 = 0.06, \gamma_2 = 0.6$. We randomly assign the labels to 51 locations and fix them over 100 replicates. The grouping labels for three parameters is given in Figure 2.

For each replicate, we have 15,000 iterations MCMC samples and have first 5,000 iterations burn-in in order to obtain samples from every 5th iteration of the last 10,000 iterations.

The performance of the posterior estimates of parameters were evaluated by the mean bias (MB) and the mean standard deviation (MSD) in the following ways, take β as an example:

$$\begin{aligned} \text{MB} &= \frac{1}{100}\sum_{r=1}^{100}\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}^r(\boldsymbol{s}_i) - \beta(\boldsymbol{s}_i)\right\}, \\ \text{MSD} &= \sqrt{\frac{1}{100}\sum_{r=1}^{100}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}^r(\boldsymbol{s}_i) - \bar{\hat{\beta}}(\boldsymbol{s}_i)\right)^2\right\}}, \end{aligned}$$

where $\bar{\hat{\beta}}(\boldsymbol{s}_i)$ is the mean of the posterior estimate over 100 replicates.

For clustering estimation evaluation, the estimated number of clusters $\hat{K}$ for each replicate is summarized from the MCMC iteration picked by Dahl's method. Rand Index [RI; 20] is applied to evaluate cluster configuration. The RI is calculated by R-package **fossil** [25]. A higher value of the RI represents higher accuracy of clustering. The average RI (MRI) was calculated as the mean of RIs over the 100 replicates.

4.2 Simulation results

The parameter estimation performance and clustering performance results are shown in Table 1 and Table 2.

From the results shown in Table 1, we see that the MBs and MSDs of the parameters are both within a reasonable range. In general, performance of posterior estimates achieve a good target.

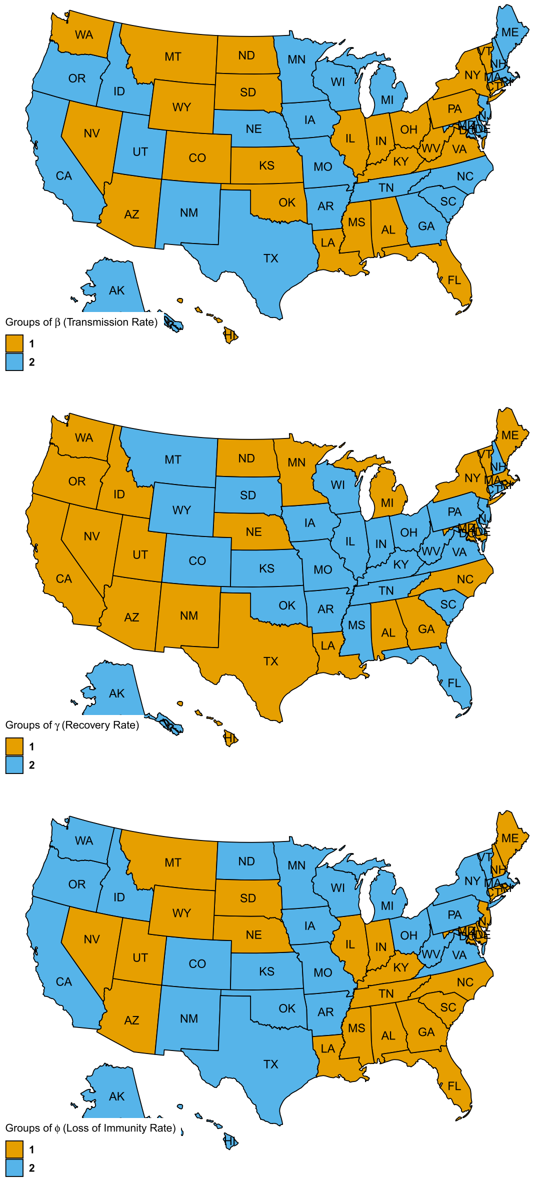
Groups of β (Transmission Rate)
1
2
Groups of γ (Recovery Rate)
1
2
Groups of ϕ (Loss of Immunity Rate)
1
2

Figure 2. Grouping Labels for β, γ, and ϕ.

Table 1. Estimation Performance for Simulation Study (β (transmission rate), γ (recovery rate), ϕ (loss of immunity rate)

Parameter	MB	MSD
β_1	0.008	0.021
β_2	-0.072	0.152
γ_1	0.007	0.017
γ_2	-0.068	0.151
ϕ_1	0.012	0.023
ϕ_2	-0.069	0.149

Table 2. Grouping Performance for Simulation Study (β (transmission rate), γ (recovery rata), ϕ (loss of immunity rate)

Parameter	MRI	S.D of RI	$\hat{K}$	S.D. of $\hat{K}$
β	0.854	0.058	2.12	0.33
γ	0.857	0.057	2.33	0.55
ϕ	0.847	0.059	2.31	0.54

And we see that our proposed methods successfully recover the number of groups and grouping labels within a reasonable range for all three parameters from Table 2. Average rand index for all parameters around 0.85 indicate our proposed method truly recovers the group labels for all three parameters. The mean of the estimated number of groups is close to true number of groups over 100 replicates.

5. REAL DATA ANALYSIS

5.1 30-day analysis from April 1st

We analyze 30-Day data from April 1st, 2020. The reason why we analyze this time period data is that U.S. Government announced the suspension of entry as immigrants and nonimmigrants of certain additional persons who pose a risk of transmitting corona-virus https://www.whitehouse.gov/presidential-actions/ on March 11th, 2020. From the April 1st, we can assume that there are very limited imported cases from outside U.S. We analyze 30-day data based on the model in (20) and use the priors discussed in Section 3.4. We run 50,000 MCMC iterations and burnin the first 20,000 iterations in order to obtain samples from every 10th iteration of the last 30,000 iterations. The group labels are obtained by Dalh's method in Section 3.5.

For β, one group is identified. $\beta = 0.079$ with 95% Highest Probability Density (HPD) interval [4] $(0.058, 0.098)$. For γ, three groups are identified with $\gamma_1 = 0.0054$ with HPD interval $(0.0021, 0.0207)$, $\gamma_2 = 0.0419$ with HPD interval $(0.0022, 0.0609)$ and $\gamma_3 = 0.0164$ with HPD interval $(0.0035, 0.0241)$. The basic reproduce numbers of three groups are given: 31 states in Group 1 with $R_0 = 14.62$ and HPD interval $(1.048, 21.200)$, 11 states in Group 2 with $R_0 = 1.88$ and HPD interval $(0.039, 15.773)$, and 7 states in Group 3 with $R_0 = 4.82$ and HPD interval $(0.892, 11.116)$, respectively. The grouping labels are shown in Figure 3.

For ϕ, one group is identified. $\phi = 0.0015$ with HPD interval $(1.181 \times 10^{-7}, 0.0047)$.

5.2 30-day analysis from May 1st

The second time period we analyze is from May 1st, 2020. Other settings are same with previous analysis.

For β, one group is identified. $\beta = 0.0042$ with 95% HPD interval $(3.056 \times 10^{-8}, 0.1083)$. Compared with previous 30-day data, in this time period, the transmission rate decreases a lot. For γ, two groups are identified with $\gamma_1 = 0.0381$ with HPD interval $(0.0048, 0.3713)$ and $\gamma_2 = 0.0007$ with

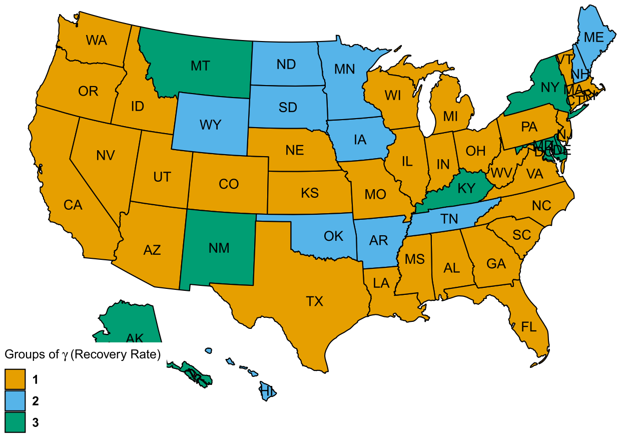

Figure 3. Group Labels for γs of April 1st data.

HPD interval $(0.0003, 0.0013)$. Two states, Oregon and Vermont, are identified in group 1. Other states are identified in group 2. For ϕ, one groups is identified. $\phi = 0.0006$ with HPD interval $(2.747 \times 10^{-7}, 0.0026)$.

With the estimated values of β and γ, the basic reproduction number, R_0, is calculated among different states. There are two different groups for the basic reproduction number. One group include Oregon and Vermont with $R_0 = 0.1102$ and HPD interval $(6.532 \times 10^{-7}, 2.656)$. The other group includes other 49 states with $R_0 = 5.4619$ with HPD interval $(3.26 \times 10^{-5}, 118.36)$. Comparing to the 30 days period from April 1st, we can see a decrease for R_0 in general. The partial reason for the decreasing of R_0 is that most states have higher daily new confirmed and resource of hospitals is of shortage in April. The estimated transmission rate is higher and the recovery rate is lower in April.

The 14-day average growth rates from June 1st of Oregon and Vermont are 2.25×10^{-5} and 1.65×10^{-5}, respectively. The 14-day average growth rates of some representative states such as New York and South Dakota are 3.69×10^{-5} and 7.21×10^{-5}, respectively. These growth rate results are consistent with our grouping detection.

6. DISCUSSION

In this paper, we develop a nonparametric Bayesian heterogeneity learning method for SIRS model based on Mixture of Finite Mixtures model. This statistical framework was motivated by the heterogeneity of COVID-19 pattern among different regions. Our simulation results indicate that the proposed method can recover the heterogeneity pattern of parameters among different regions. Illustrated by the analysis of COVID-19 data in U.S., our proposed methods reveal the heterogeneity pattern among different states.

In addition, three topics beyond the scope of this paper are worth further investigation. First, we can add spatially dependent structure [12, 27] on the heterogeneity of different states. Second, there is one limitation of our proposed method. Our model assumes parameters are constant over a certain time period. For a future time window, the heterogeneity over different regions is not predictable. Building heterogeneity learning model with time varying coefficients is an interesting future work. Finally, proposing a measurement error correction for SIRS devotes another interesting future work.

APPENDIX A. ADDITIONAL SIMULATION RESULTS

In this section, we provide the simulation results with $K = 3$ for three parameters. The time length T equals 30 for all the simulation replicates. The total number of replicates in our simulation study is 100. For each parameter, we have three different groups and we set the true values of the parameters $\beta_1 = 0.1, \beta_2 = 0.3, \beta_3 = 0.7$, $\gamma_1 = 0.1, \gamma_2 = 0.3, \gamma_3 = 0.7$, and $\phi_1 = 0.1, \phi_2 = 0.3, \phi_3 = 0.7$. The simulation results for grouping performance is presented in Table 3.

Table 3. Grouping Performance for Simulation Study ($K = 3$) (β (transmission rate), γ (recovery rata), ϕ (loss of immunity rate)

Parameter	MRI	S.D of RI	$\hat{K}$	S.D. of $\hat{K}$
β	0.766	0.052	3.09	0.28
γ	0.755	0.043	3.95	0.78
ϕ	0.743	0.051	3.15	0.41

From the results shown in Table 3, we see that the clustering performance becomes a little bit worse than $K = 2$. The partial reason is that the number of observations in each cluster decreases which makes the clustering harder than previous cases.

APPENDIX B. NIMBLE CODE FOR MCMC

In the appendix, the nimble code is listed to demonstrate the MCMC algorithm. With nimble, we can write our own code in R but in BUGS syntax and then nimble can compile our code into C++.

```
piFun2 <- nimbleFunction(
  run = function(r = double(1)) {
    rlength <- length(r)
    rsum <- rep(0, rlength)
    pi <- rep(0, rlength)
    rsum[1] <- pi[1] <- r[1]
    for (i in 2:rlength) {
      rsum[i] <- rsum[i - 1] + r[i]
      if (rsum[i] >= 1) {
        pi[i] <- 1 - rsum[i - 1]
      }
      else {pi[i] <- r[i]}
    }
    for (i in 1:rlength) {
      if (pi[i] < 0) {pi[i] <- 0}
    }
    returnType(double(1))
    return(pi)
  }
)
SIRSCode <- nimbleCode({


  for (i in 1:n){
    ZR[1,i]~dpois(lambda = probR[1,i]*population[i])
    ZI[1,i]~dpois(lambda = probI[1,i]*population[i])
    probR[1,i]<-1/(1+exp(WS[1,i])+exp(WI[1,i]))
    probI[1,i]<-exp(WI[1,i])/(1+exp(WS[1,i])+exp(WI[1,i]))
    WS[1,i]~dnorm(muws[1,i],var=sigma2s[i])
    WI[1,i]~dnorm(muwi[1,i],var=sigma2i[i])
    muws[1,i]<-init_s[i]
    muwi[1,i]<-init_i[i]
  for (t in 2:T){
      ZR[t,i]~dpois(lambda = probR[t,i]*population[i])
      ZI[t,i]~dpois(lambda = probI[t,i]*population[i])
      probR[t,i]<-1/(1+exp(WS[t,i])+exp(WI[t,i]))
      probI[t,i]<-exp(WI[t,i])/(1+exp(WS[t,i])+exp(WI[t,i]))
      WS[t,i]~dnorm(muws[t,i],var=sigma2s[i])
      WI[t,i]~dnorm(muwi[t,i],var=sigma2i[i])
      muws[t,i]<-WS[t-1,i]+
    log(1+phi[i]/exp(WS[t-1,i])
     -beta[i]*exp(WI[t-1,i])/
     (1+exp(WS[t-1,i])+exp(WI[t-1,i])))
    +log(1/(1+gamma[i]*exp(WI[t-1,i])-phi[i]))

    muwi[t,i]<-WI[t-1,i]+
   log(1-gamma[i]+beta[i]*exp(WS[t-1,i])/
   (1+exp(WS[t-1,i])+exp(WI[t-1,i])))
   +log(1/(1+gamma[i]*exp(WI[t-1,i])-phi[i]))

  }

  }

  for (i in 1:n){
    sigma2s[i]~ dinvgamma(0.25,0.4)
    sigma2i[i]~ dinvgamma(0.25,0.4)
    beta[i]<-hatbeta[latent_beta[i]]
    phi[i]<-hatphi[latent_phi[i]]
    gamma[i]<-hatgamma[latent_gamma[i]]
    latent_beta[i]~ dcat(pi_beta[1:M])
    latent_phi[i]~ dcat(pi_phi[1:M])
    latent_gamma[i]~ dcat(pi_gamma[1:M])
  }


  for (i in 1:M) {
    r_beta[i] ~ dexp(rate = lambda_beta)
    r_phi[i] ~ dexp(rate = lambda_phi)
    r_gamma[i] ~ dexp(rate = lambda_gamma)
  }
  pi_beta[1:M] <- piFun2(r = r_beta[1:M])
  pi_phi[1:M] <- piFun2(r = r_phi[1:M])
  pi_gamma[1:M] <- piFun2(r = r_gamma[1:M])

  lambda_beta ~ dgamma(1,1)
  lambda_phi ~ dgamma(1,1)
  lambda_gamma ~ dgamma(1,1)
  for (k in 1:M) {
    hatbeta[k]~dunif(0,1)
    hatphi[k]~ dunif(0,1)
    hatgamma[k]~dunif(0,1)
  }


})
```

Received 4 July 2020

REFERENCES

[1] Anderson, R. M., Anderson, B., and May, R. M. (1992). *Infectious diseases of humans: dynamics and control.* Oxford University Press.

[2] Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 2*(6), 1152–1174. MR0365969

[3] Blackwell, D., MacQueen, J. B., et al. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics 1*(2), 353–355. MR0362614

[4] Chen, M.-H. and Shao, Q.-M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics 8*(1), 69–92. MR1705909

[5] Chen, Y.-C., Lu, P.-E., Chang, C.-S., and Liu, T.-H. (2020). A time-dependent SIR model for COVID-19 with undetectable infected persons. *arXiv preprint arXiv:2003.00122*.

[6] Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference for Gene Expression and Proteomics 4*, 201–218. MR2706330

[7] de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics 26*(2), 403–413. MR3640196

[8] Dukic, V., Lopes, H. F., and Polson, N. G. (2012). Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *Journal of the American Statistical Association 107*(500), 1410–1426. MR3036404

[9] Geng, J., Bhattacharya, A., and Pati, D. (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association 114*(526), 893–905. MR3963189

[10] Geng, J., Shi, W., and Hu, G. (2019). Bayesian nonparametric nonhomogeneous Poisson process with applications to USGS earthquake data. *arXiv preprint arXiv:1907.03186*.

[11] Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review 42*(4), 599–653. MR1814049

[12] Hu, G., Geng, J., Xue, Y., and Sang, H. (2020). Bayesian spatial homogeneity pursuit of functional data: an application to the US income distribution. *arXiv preprint arXiv:2002.06663*. MR3873724

[13] Jo, H., Son, H., Jung, S. Y., and Hwang, H. J. (2020). Analysis of COVID-19 spread in South Korea using the SIR model with time-dependent parameters and deep learning. *medRxiv*.

[14] Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 115*(772), 700–721.

[15] Kermack, W. O. and McKendrick, A. G. (1932). Contributions to the mathematical theory of epidemics. II. – The problem of endemicity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 138*(834), 55–83.

[16] Kermack, W. O. and McKendrick, A. G. (1933). Contributions to the mathematical theory of epidemics. III. – Further studies of the problem of endemicity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 141*(843), 94–122.

[17] Miller, J. W. and Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems*, pp. 199–206.

[18] Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*(521), 340–356. MR3803469

[19] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265. MR1823804

[20] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

[21] Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A., and Jewell, C. P. (2020). Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. *MedRxiv*.

[22] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 639–650. MR1309433

[23] Sun, H., Qiu, Y., Yan, H., Huang, Y., Zhu, Y., Gu, J., and Chen, S. X. (2020). Tracking reproductivity of COVID-19 epidemic in China with varying coefficient SIR model [discussion paper]. *Journal of Data Science*, 2.

[24] Tang, B., Wang, X., Li, Q., Bragazzi, N. L., Tang, S., Xiao, Y., and Wu, J. (2020). Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *Journal of Clinical Medicine 9*(2), 462.

[25] Vavrek, M. J. (2011). Fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica 14*(1), 16.

[26] Wu, J. T., Leung, K., and Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet 395*(10225), 689–697.

[27] Zhao, P., Yang, H.-C., Dey, D. K., and Hu, G. (2020). Bayesian spatial homogeneity pursuit regression for count value data. *arXiv*

preprint arXiv:2002.06678. MR4011757

[28] Zhuang, L. and Cressie, N. (2014). Bayesian hierarchical statistical SIRS models. *Statistical Methods & Applications 23*(4), 601–646. MR3278930

Guanyu Hu
146 Middlebush Hall
Columbia, MO 65211-6100
US
E-mail address: gh7mr@missouri.edu

Junxian Geng
900 Ridgebury Rd
Ridgefield, CT
US
E-mail address: junxian.geng@boehringer-ingelheim.com