Statistical Inference Using GLEaM Model with Spatial Heterogeneity and Correlation between Regions

Yixuan Tan[1], Yuan Zhang[2], Xiuyuan Cheng[1], and Xiao-Hua Zhou[3,4,5]

[1]Department of Mathematics, Duke University
[2]School of Mathematical Sciences, Peking University
[3]Center for Statistical Sciences, Peking University
[4]Beijing International Center for Mathematical Research, Peking University
[5]Department of Biostatistics, School of Public Health, Peking University

Abstract

A better understanding of the various patterns in the coronavirus disease 2019 (COVID-19) spread in different parts of the world is crucial to its prevention and control. Motivated by the celebrated GLEaM model (Balcan et al., 2010[1]), this paper proposes a pioneering stochastic dynamic model to depict the evolution of COVID-19. The model allows spatial and temporal heterogeneity of transmission parameters and involves transportation between regions. Based on the proposed model, this paper also designs a two-step procedure for parameter inference, which utilizes the correlation between regions through a prior distribution that imposes graph Laplacian regularization on transmission parameters. Experiments on simulated data and real-world data in China and Europe indicate that the proposed model achieves higher accuracy in predicting the newly confirmed cases than baseline models.

1 Introduction

The outbreak of coronavirus disease 2019 (COVID-19) has impacted all aspects of the world significantly for a long. As of 26 Oct 2021, there have been over 243 million confirmed cases of COVID-19, including over 4 million deaths [2]. Therefore, it is essential to study the spread of COVID-19 for better prediction and prevention of the disease. This paper first proposes a stochastic dynamical model that describes the spread of COVID-19 in multiple regions and then develops an algorithm to estimate the transmission parameters and their posterior distributions.

The model proposed in this paper is inspired by The Global Epidemic and Mobility (GLEaM) model proposed in [1]. GLEaM is a stochastic dynamic model to depict the spread of epidemics, integrating multiple data layers that couple metapopulations at the global level. The model involves 3362 subpopulations in 220 countries obtained from Voronoi tessellation, centered around major airports. These subpopulations are connected by a multi-layer mobility network composed of processes from short-range communication in geographically close subpopulations to international flights. In each subpopulation, the transmission of epidemics is modeled by a Susceptible-Exposed-Infected-Removed (SEIR) compartmental model [3] that integrates short-range communication. After specifying the compartmental model and initial conditions, GLEaM can be run repeatedly to build a statistical ensemble for each point in the space of parameters, which can be further utilized for Monte Carlo analysis of epidemic parameters. A more detailed review of SEIR model and the GLEaM model is in Appendix A.

In the vast majority of GLEaM's applications [4, 5, 6, 7, 8, 9, 10, 11], the parameters are estimated based on [12]. [12] performed the maximal likelihood analysis of the reproduction number R_0 in the seed region Mexico. For each value of the reproduction number R_0, the method generated the distribution of arrival time of the influenza A(H1N1) in 12 countries produced by 2×10^3 GLEaM simulations. Then, the optimal the reproduction number R_0 was chosen by maximizing the likelihood function of arrival time. [12] and subsequent works following its settings [4, 5, 6, 7, 8, 9, 10, 11] assumed that the epidemic was seeded from one region and the transmission parameters or the other parameters (like the introduction date and location) were estimated through the maximal likelihood analysis of arrival time or other events. In particular, the method in [12] was adopted in [6] to estimate the posterior distribution of the reproduction number R_0 of COVID-19, which was assumed to be uniform for all subpopulations at all times. However, this setting is not suitable for the current scenario of the COVID-19 pandemic, since COVID-19 has lasted for a long time, and the community transmission has been widespread in most countries in the world [13]. To model the spread of COVID-19, both spatial and temporal heterogeneity of the transmission parameters are needed instead of directly modeling the reproduction number R_0 as a periodic function of time as in [12], since the social behaviors, containment measures, medical conditions, and other elements that might have effects on the spread of COVID-19 may vary among different countries and over time.

Recently, [14] improved the model described above by estimating the initially infected individuals in each subpopulation through microblogging data from Twitter and also estimating the reproduction number R_0 for USA, Italy, and Spain separately. Therefore, the method used in [14] involved spatial heterogeneity. However, in this study, international travel was not considered, and GLEaM was applied to each of the aforementioned countries (as an isolated system) independently. The transmission rates in all subpopulations of these countries were again presumed to be homogeneous. Furthermore, [14] also assumed that the initially infected individuals for each subpopulation were proportional to the total number of Twitter users in that subpopulation. Thus it still assumed that the severity of the pandemic at the initial outbreak of COVID-19 was uniform over the country, which is not the case for COVID-19.

In addition to the issues mentioned above, other potential concerns exist when using GLEaM to model the spread of COVID-19. As mentioned in the last section of [15], GLEaM can be used to simulate the spread of the epidemic under normal conditions since it uses the "steady-state" mobility data around the world. However, since the outbreak of COVID-19, the social order has been disrupted, and travel has been restricted in most countries. Thus GLEaM might not work well with its multi-layer mobility networks. Furthermore, the estimate of parameters using GLEaM is based on a large number of simulations to explore the space of parameters, which usually takes much computational time [1]. In addition, although the social behavior, medical conditions, and other factors that affect the spread of COVID-19 may vary among different regions, these factors for regions that are geographically close or have similarities in other aspects still bear some resemblance. Hence, the transmission rates for COVID-19 should not only have their own heterogeneity but also correlate with each other. However, neither of the features is reflected in GLEaM or most of its applications.

As the consequences of the possible constraints of GLEaM described above, most of the papers using GLEaM to model the epidemics mainly focus on estimating only the transmission parameter in the "seed region" at the very beginning of the outbreak. However, for the current long-lasting spread of COVID-19 pandemic all over the globe, the spatial and temporal heterogeneity of the transmission parameters is needed to be taken into full consideration.

In this paper, we propose a pioneering stochastic model based on the GLEaM framework that enables transmission parameters' spatial and temporal heterogeneity. In contrast to most applications of GLEaM, which mainly focus on the initial outbreak, the model we propose is able to model the long-lasting spread of COVID-19. For the inference of the parameters, this paper introduces an optimization algorithm that utilizes the correlation between districts. Furthermore, the posterior distribution of parameters

is estimated by Markov chain Monte Carlo (MCMC) sampling starting from an optimal choice of the parameters obtained from the optimization process. This takes a much shorter time than the methods based on GLEaM simulations.

The main contributions of our paper are:

- We propose a new stochastic model to describe the epidemic's lasting spread, allowing spatial and temporal heterogeneity of transmission parameters and transportation between districts.
- Based on the model, we also design an algorithm that first makes inference for the parameters through a two-step procedure combining the information of correlation between districts, which is equivalent to imposing graph Laplacian regularization on the transmission parameters. Then, the algorithm estimates the posterior distribution efficiently by MCMC sampling with the estimated parameters as the initial points.
- We compare the performance of the proposed model with the baseline models on both simulated data and real-world data.

  - For the simulated data, the results show that combining heterogeneity and transportation into the model helps improve the performance of trajectory prediction and parameter estimation. Moreover, our inference algorithm that integrates the correlation of districts leads to further improvement in predicting the future trajectories.

  - For the real-world data in China and Europe, the proposed model still outperforms the baselines in trajectory prediction.

Data sets used in this paper are publicly available. Our work focuses on proposing a new method for parameter inference and, consequently, answering questions regarding the long-lasting spread of COVID-19 in multiple regions.

A strength of the proposed model resides in introducing spatial and temporal heterogeneity of transmission parameters. We remark that there exist recent works which also involved different levels of heterogeneity in their models in various ways. [16, 17] utilized randomness in reproduction number to reflect heterogeneity of the population, using plate model with Bayesian method and heterogeneous well-mixed theory [18] with age-of-infection method [19], respectively. [20] and [21] used functional data analysis tools. Specifically, [20] captured two different epidemic patterns in different regions of Italy using probKMA algorithm [22], and [21] revealed different patterns of the epidemic across countries with functional principle component analysis. In addition, [23, 24, 25, 26, 27] adapted SEIR / Susceptible-Exposed-Infected (SEI) / Susceptible-Infected (SI) compartmental models similar to this paper. Among these works, [23, 24, 25] considered heterogeneity in the aspects of age groups, social links, and vaccination status separately. [26, 27], including spatial heterogeneity in one county and among states respectively, bore more similarity with this paper since they also allowed transmission parameters to be spatially heterogeneous and involved transportation between different regions. However, [26] only considered intracounty data, and the transportation was used to compute the effective size of compartments and did not affect the dynamic model. Furthermore, the transmission rates in [26] were determined by an SDE whose parameters were to be fitted. Therefore, [26] focused on a different scope from this paper. The settings of compartments in [27] are more realistic than the one considered in this paper by considering reporting rates. Nevertheless, compared with the model and inference algorithm described in Section 2, transmission rates in [27] were kept constant over time and were estimated using independent non-informative prior. Both [26] and [27] used Ensemble Kalman Filter, which samples particles in the state space according to the prior distribution and obtains the posterior distribution in the process of moving particles at each time step. This may be

computationally less efficient than directly applying MCMC according to the posterior distribution with the initial point maximizing the likelihood function, as implemented in this paper.

We list the default notations and parameters used throughout the paper in Table 1. The rest of the paper is structured in the following way: In Section 2, we introduce the stochastic dynamic model and corresponding inference algorithm. In Section 3, we compare the performance of trajectory prediction and parameter estimation of models with or without mobility, heterogeneity, or using correlation information in the inference part for the simulated data. Section 4 describes the real-world data used in this paper, presents the results and findings of applying our model to the COVID-19 data in China and Europe. We discuss our limitations and possible extensions in Section 5.

2 Method

2.1 Ethics Statement

The medical record data in China and Europe used in this paper are publicly available and from the official websites of the National Health Commission of the People's Republic of China [28], Chinese Center for Disease Control and Prevention [29], and European Centre for Disease Prevention and Control [30]. The collection of data is performed in compliance with local government regulations. More details of data sources can be found in Section 4.1.

2.2 Model Description

Compartmental model over multiple regions. In GLEaM ([1]) and other works ([26, 27]) involving transportation, the whole area is first divided into subdivisions. For example, the GLEaM model divides the total area of 220 countries into over 3300 subpopulations centered around major airports and [26] divided Milwaukee County and Dane County in the state of Wisconsin into several regions. In this paper, we consider abstract subdivisions in the whole area, which will be referred as "regions" hereinafter until further specifications in the later experiment sections. We denote by n the number of regions.

In our model, we use continuous time $t \in [0, T]$, where it is assumed that the evolution of the epidemic lasts within a period of T time units. The unit of time is fixed as one day throughout this paper. Note that when we introduce the transportation model in below the traveling matrix is assumed to be constant within each day, and also the observed data is collected on a daily basis. Thus we will use notation of discrete time (days) from $1, \ldots, T$ hereinafter, however, the evolution dynamic itself is modeled over continuous time.

For each region, we consider the following epidemiology compartments adapted from the SEIR model:

- $S_k(t)$: Susceptible.
- $E_k(t)$: Exposed and infectious.
- $H_k(t)$: Hospitalized.
- $R_k(t)$: Removed.

The subscript k of the states denotes that they belong to the k-th region, and dependence on the continuous time t is addressed through expressing the states as functions of $t \in [0, T]$. Removed individuals are those who have recovered or died.

At time t, we use $N_k(t) = S_k(t) + E_k(t) + H_k(t) + R_k(t)$ to denote the total population in the k-th region. $N_k(t)$ is allowed to be time-varying since the inter-region mobility is taken into consideration, especially

for the days before the implementation of travel restrictions. However, since the traveling volume is not comparable to the total population in a region, the fluctuation of the total population in a region is not obvious. In this paper, we assume that $N = \sum_{k=1}^{n} N_k(t)$ keeps constant over time, which means that we consider a closed system in this paper, where exported/imported cases are not considered. However, it is worth noting that we do allow the transportation of active virus carriers between regions within our system. We remark in advance that this assumption is reasonable for the real-world data sets considered in this paper. From January to February 2020, strict international travel restrictions were imposed in China. While for data in Europe, from May to August 2020, the local spread of the epidemic has reached a relatively high level, and the imported cases were not comparable to the indigenous cases. We also denote $(N_a)_k(t) = S_k(t) + E_k(t) + R_k(t)$ as the total population that are permitted to move in the k-th region, excluding the hospitalized ones.

Transportation between regions. Transportation plays an essential role in the spread of COVID-19. Actually, [31, 32] indicated that the travel restrictions were remarkably important in mitigating the transmission of COVID-19, especially in the early stage of the pandemic. Recently, as detailed in [33], the Omicron variant had spread to 110 countries and had become dominant in many of them by 22 December 2021, only one month after its first report from South Africa on 24 November 2021. This motivates us to also take transportation into consideration in this paper. Motivated by the GLEaM model [1], we introduce the transportation between regions. Specifically, we denote $(w_l)_{ij}$ as the traveling volume from region i to j on the l-th day ($l = 1, \ldots, T$). Then the traveling matrix W_l on the l-th day can be written as

$$W_l = \begin{pmatrix} (w_l)_{11} & (w_l)_{12} & \cdots & (w_l)_{1n} \\ (w_l)_{21} & (w_l)_{22} & \cdots & (w_l)_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ (w_l)_{n1} & (w_l)_{n2} & \cdots & (w_l)_{nn} \end{pmatrix}. \tag{2.1}$$

Then, inspired by [34], we consider the transition dynamic of $\xi(t) = \{S_1(t), E_1(t), H_1(t), R_1(t), \ldots, S_n(t), E_n(t), H_n(t), R_n(t)\} \in \{0, 1, 2, ..., N\}^{4n}$ over $t \in [0, T]$ described below:

- Transmission in the k-th region: A case from $E_k(t)$ chooses an individual from $N_k(t)$ randomly at Poisson rate $\lambda_k(t)$, and the individual chosen is infected if it is of state $S_k(t)$. Note that this is different from traditional SEIR models, since we assume that for the COVID-19 case, the pre-symptomatic patients from $E_k(t)$ can be contagious.
- Hospitalization in the k-th region: Each individual in $E_k(t)$ will be hospitalized with Poisson rate δ.
- Recovery or death in the k-th region: Each individual in $H_k(t)$ will transfer into $R_k(t)$ with Poisson rate γ_k. The rate γ_k owns spatial heterogeneity due to the uneven distribution of medical resources.
- Transportation between regions: At the end of the l-th day, all the individuals in region i except the ones in $H_i(l)$ have the same probability to travel from the i region to the j region, and the total traveling volume from the i region to the j region is $(w_l)_{ij}$. We assume that there are no transmissions happening during the transportation between regions. If we denote $\{\xi_{ij}^{[S,l]}, \xi_{ij}^{[E,l]}, \xi_{ij}^{[R,l]}\}$ as the number of people transported from $\{S_i(l), E_i(l), R_i(l)\}$ to $\{S_j(l), E_j(l), R_j(l)\}$ at the end of the l-th day, then $\{\xi_{ij}^{[S,l]}, \xi_{ij}^{[E,l]}, \xi_{ij}^{[R,l]}\}$ follows a multinomial distribution. Specifically,

$$P(\{\xi_{ij}^{[S,l]}, \xi_{ij}^{[E,l]}, \xi_{ij}^{[R,l]}\}) = \frac{((w_l)_{ij})!}{\left(\xi_{ij}^{[S,l]}\right)! \left(\xi_{ij}^{[E,l]}\right)! \left(\xi_{ij}^{[R,l]}\right)!} \left(\frac{S_i(l)}{(N_a)_i(l)}\right)^{\xi_{ij}^{[S,l]}} \left(\frac{E_i(l)}{(N_a)_i(l)}\right)^{\xi_{ij}^{[E,l]}} \left(\frac{R_i(l)}{(N_a)_i(l)}\right)^{\xi_{ij}^{[R,l]}}, \tag{2.2}$$

with $\xi_{ij}^{[S,l]} + \xi_{ij}^{[E,l]} + \xi_{ij}^{[R,l]} = (w_l)_{ij}$.

As a consequence, $N_k(t)$ is a piece-wise constant function of t which only changes at the end of each day. Specifically, for any time $t \geq 0$,

$$N_k(t) = N_k(0) + \sum_{l=1}^{\lfloor t \rfloor} \sum_{i=1}^{n} ((w_l)_{ik} - (w_l)_{ki}). \tag{2.3}$$

Differential equation with spatial heterogeneity. Following [35, 36], we consider the mean-field differential equation system defined by (2.4), which is a deterministic version of the stochastic model above:

$$\begin{cases} \dfrac{d\widetilde{S}_k(t)}{dt} = -\lambda_k \dfrac{\widetilde{S}_k(t)\widetilde{E}_k(t)}{\widetilde{N}_k(t)} + \displaystyle\sum_{i \neq k} \left[(w_{\lceil t \rceil})_{ik} \dfrac{\widetilde{S}_i(t)}{(\widetilde{N_a})_i(t)} - (w_{\lceil t \rceil})_{ki} \dfrac{\widetilde{S}_k(t)}{(\widetilde{N_a})_k(t)} \right] \\ \dfrac{d\widetilde{E}_k(t)}{dt} = \lambda_k \dfrac{\widetilde{S}_k(t)\widetilde{E}_k(t)}{\widetilde{N}_k(t)} - \delta \widetilde{E}_k(t) + \displaystyle\sum_{i \neq k} \left[(w_{\lceil t \rceil})_{ik} \dfrac{\widetilde{E}_i(t)}{(\widetilde{N_a})_i(t)} - (w_{\lceil t \rceil})_{ki} \dfrac{\widetilde{E}_k(t)}{(\widetilde{N_a})_k(t)} \right] \\ \dfrac{d\widetilde{H}_k(t)}{dt} = \delta \widetilde{E}_k(t) - \gamma_k \widetilde{H}_k(t) \\ \dfrac{d\widetilde{R}_k(t)}{dt} = \gamma_k \widetilde{H}_k(t) + \displaystyle\sum_{i \neq k} \left[(w_{\lceil t \rceil})_{ik} \dfrac{\widetilde{R}_i(t)}{(\widetilde{N_a})_i(t)} - (w_{\lceil t \rceil})_{ki} \dfrac{\widetilde{R}_k(t)}{(\widetilde{N_a})_k(t)} \right] \\ \dfrac{d(\widetilde{C_a})_k(t)}{dt} = \delta \widetilde{E}_k(t) \\ \dfrac{d(\widetilde{R_a})_k(t)}{dt} = \gamma_k \widetilde{H}_k(t) \\ \dfrac{d\widetilde{N}_k(t)}{dt} = \displaystyle\sum_i ((w_{\lceil t \rceil})_{ik} - (w_{\lceil t \rceil})_{ki}) \end{cases}, \tag{2.4}$$

where $\widetilde{S}_k(t)$ is the deterministic counterpart of $S_k(t)$ and so on. Note that since the traveling volume is a piece-wise constant function of t, $\widetilde{N}_k(t)$ evolves as a piece-wise linear function of t, and $\widetilde{N}_k(t)$ coincides with $N_k(t)$ when t takes integer values. $(\widetilde{C_a})_k(t)$ and $(\widetilde{R_a})_k(t)$ are the accumulated confirmed and removed cases determined by the ordinary differential equation (ODE) system (2.4) in the k-th region at time t respectively. $(\widetilde{C_a})_k(t)$ (or $(\widetilde{R_a})_k(t)$) only takes the increment of $\widetilde{E}_k(t)$ (or $\widetilde{R}_k(t)$) into account, meanwhile guaranteeing that each case is only accounted for once.

Furthermore, we assume that the accumulated confirmed and removed cases are available from data on a daily basis, which are denoted as $\{(C_a)_k(i)\}_{i=1}^T$ and $\{(R_a)_k(i)\}_{i=1}^T$, respectively. We also assume that $R_k(0)$ is 0, while $E_k(0)$ and $H_k(0)$ are left to be inferred for each $k = 1, \ldots, n$. For inference of parameters, we further denote $(\widetilde{\Delta C_a})_k(i) = (\widetilde{C_a})_k(i) - (\widetilde{C_a})_k(i-1)$ as the newly confirmed cases on the i-th day ($k = 1, \ldots, n, i = 2, \ldots, T$) determined by (2.4), and $(\Delta C_a)_k(i)$ as the newly confirmed cases computed from data, namely $(\Delta C_a)_k(i) = (C_a)_k(i) - (C_a)_k(i-1)$. The same convention holds for the definitions of $(\widetilde{\Delta R_a})_k(i)$ and $(\Delta R_a)_k(i)$. Note that the data $\{(C_a)_k(i), (R_a)_k(i), (\Delta C_a)_k(i), (\Delta R_a)_k(i)\}_{i=1}^T$ are random in nature.

Note that the model and the inference algorithm described below can be applied to estimate spatially heterogeneous transmission parameters as long as $\{(C_a)_k(i), k = 1, \ldots, n\}_{i=1}^T$, $\{(R_a)_k(i), k = 1, \ldots, n\}_{i=1}^T$, and $\{W_l\}_{l=1}^T$ are available. The availability of $\{(C_a)_k(i), k = 1, \ldots, n\}_{i=1}^T$ and $\{(R_a)_k(i), k = 1, \ldots, n\}_{i=1}^T$ is required in many works that use the SEIR model to estimate transmission rates of the epidemic [37, 38, 39],

Table 1: List of notations and parameters

Notation	Description	Notes
n	Total number of regions	
T	Total number of days	
$S_k(t)$	Susceptible individuals in the k-th region at time t, $1 \leq k \leq n, t \in [0, T]$	The same range of k and t also holds below.
$E_k(t)$	Exposed individuals in the k-th region at time t	$\{E_k(0)\}_{k=1}^n$ need to be inferred.
$H_k(t)$	Hospitalized individuals in the k-th region at time t	$\{H_k(0)\}_{k=1}^n$ need to be inferred.
$R_k(t)$	Removed individuals in the k-th region at time t	$R_k(0) = 0, \forall k$.
$N_k(t)$	Total individuals in the k-th region at time t	$\sum_{k=1}^n N_k(t)$ is assumed to be constant over time.
W_l	Traveling volume matrix on the l-th day	$(W_l)_{ij} = (w_l)_{i,j}, l \in \{1, \ldots, T\}$.
$\widetilde{S}_k(t)$	Deterministic counterpart of $S_k(t)$, determined by (2.4)	Other compartments follow the same convention of notations, $t \in [0, T]$.
$(\widetilde{C_a})_k(t)$	Accumulated confirmed cases determined by (2.4)	$t \in [0, T]$.
$(\widetilde{R_a})_k(t)$	Accumulated removed cases determined by (2.4)	$t \in [0, T]$.
$(C_a)_k(i)$	Accumulated confirmed cases on the i-th day	$i \in \{1, \ldots, T\}$.
$(R_a)_k(i)$	Accumulated removed cases on the i-th day	$i \in \{1, \ldots, T\}$.
$(\widetilde{\Delta C_a})_k(i)$	Newly confirmed cases determined by (2.4) on the i-th day	$i \in \{1, \ldots, T\}$.
$(\Delta C_a)_k(i)$	Newly confirmed cases on the i-th day	$i \in \{1, \ldots, T\}$.
λ_k	Infection rate in the k-th region	$\{\lambda_k\}_{k=1}^n$ need to be inferred and are allowed to be time-varying.
δ	Inverse of average incubation period	δ is prefixed as 0.14.
γ_k	Inverse of average removed time in the k-th region	$\{\gamma_k\}_{k=1}^n$ need to be inferred.
Θ	Set of parameters to be estimated	
W	The matrix characterizing the proximity between regions	
d	Total number of groups that n regions are divided into	
D_m	The m-th group of regions	$1 \leq m \leq d$.
μ	Penalty factor of Graph Laplacian regularization	
β	Parameter reducing correlation between inter-group regions	$\beta \in (0, 1)$
A	Affinity matrix constructed by W, μ, and β as in (2.10)	
σ	Parameter in the prior (2.8)	

and the transportation network is also used in GLEaM [1] and its applications. However, in contrast to the works based on GLEaM [12, 4, 5, 6, 15], we allow parameters to possess both spatial and temporal heterogeneity, and further utilize the correlation between regions in the inference of parameters. We remark that the spatial and temporal heterogeneity is reflected in the fact that the transmission parameters $\{\lambda_k\}$ are allowed to vary in both space and time in our model. The temporal heterogeneity is introduced in more detail for the real-world data in Section 4.2.

2.3 Estimation of Model Parameters

A two-step procedure of estimating parameters is described below.

First, note that δ, the inverse of the average time for a person from being exposed to hospitalized, is prefixed to be $\delta = 0.14$ for all regions and all time. According to [40], the mean duration of incubation period is 5.2 days. Furthermore, we assume that the average time for an individual from showing symptoms to being hospitalized is 2 days for simplicity. Thus, the mean duration for an individual from being exposed to being hospitalized is 7.2 days, whose inverse value is approximately 0.14.

Thus, the model parameters that need to be estimated are $\Theta := \{E_k(0), H_k(0), \lambda_k, \gamma_k\}_{k=1}^n$. We also denote $\overline{\Theta} := \{E_k(0), H_k(0), \lambda_k\}_{k=1}^n$, which excludes $\{\gamma_k\}_{k=1}^n$ from Θ.

A two-step analysis is adapted for the estimation. We first make inference for $\{\gamma_k\}_{k=1}^n$ utilizing the newly removed cases and the accumulated hospitalized cases. After the estimation of $\{\gamma_k\}_{k=1}^n$, $\overline{\Theta}$ are then estimated by maximizing the posterior distribution, where we introduce a prior distribution combining the information of correlation between regions. In addition, the marginal posterior distributions of $\overline{\Theta}$ can also be obtained by running MCMC sampling starting from the optimizer $\overline{\Theta}^*$ of the posterior distribution. Then we complete the inference of $\Theta = \{\{\gamma_k\}_{k=1}^n, \overline{\Theta}\}$.

Mathematically, given parameters Θ, we can run the ODE system (2.4) on $[0, T]$ and denote $x_i = (\{(\widetilde{\Delta C_a})_k(i)\}_{k=1}^n, \{(\widetilde{C_a})_k(i)\}_{k=1}^n, \{(\widetilde{\Delta R_a})_k(i)\}_{k=1}^n, \{(\widetilde{R_a})_k(i)\}_{k=1}^n)$ as the values determined by (2.4) on the i-th day, and $y_i = (\{(\Delta C_a)_k(i)\}_{k=1}^n, \{(C_a)_k(i)\}_{k=1}^n, (\Delta R_a)_k(i)\}_{k=1}^n, \{(R_a)_k(i)\}_{k=1}^n)$ as the observed counterpart of x_i.

Step 1: Estimate $\{\gamma_k\}_{k=1}^n$.

We first estimate $\{\gamma_k\}_{k=1}^n$ by maximizing $\prod_{i=1}^T P\left(\{(\Delta R_a)_k(i)\}_{k=1}^n \middle| \{\gamma_k\}_{k=1}^n, y_{i-1}\right)$ over γ_k for each k. By assuming that the newly removed cases in one day follow a Poisson distribution whose mean equals to the product of γ_k and the accumulated hospitalized cases (which is the difference between the accumulated confirmed cases and the accumulated removed cases, and thus is observable) the day before, we have that

$$\begin{aligned}
&\prod_{i=1}^T P\left(\{(\Delta R_a)_k(i)\}_{k=1}^n \middle| \{\gamma_k\}_{k=1}^n, y_{i-1}\right) \\
=&\prod_{i=1}^T \prod_{k=1}^n \text{Pois}\left((\Delta R_a)_k(i) \middle| \gamma_k\left((C_a)_k(i-1) - (R_a)_k(i-1)\right)\right) \\
=&\prod_{k=1}^n \prod_{i=1}^T \frac{\left(\gamma_k\left((C_a)_k(i-1) - (R_a)_k(i-1)\right)\right)^{(\Delta R_a)_k(i)}}{((\Delta R_a)_k(i))!} \exp\left(-\gamma_k\left((C_a)_k(i-1) - (R_a)_k(i-1)\right)\right).
\end{aligned}$$

Then, we estimate $\gamma_k^* = \arg\max_{\gamma_k} \prod_{i=1}^T \text{Pois}\left((\Delta R_a)_k(i) \middle| ((C_a)_k(i-1) - (R_a)_k(i-1))\gamma_k\right)$ for each k separately.

Step 2: Estimate $\overline{\Theta} = \{E_k(0), H_k(0), \lambda_k\}_{k=1}^n$ and the marginal posterior distributions of $\overline{\Theta}$.

Next, we estimate the remaining parameters $\overline{\Theta} = \{E_k(0), H_k(0), \lambda_k\}_{k=1}^n$ as follows.

First, note that λ_k is the transmission rate, which reflects the infectivity of the virus, the prevention and control measures, and the rates of close contacts in the k-th region.

Notice that the ODE system (2.4) is the mean-field version of our stochastic model, and $\{x_i\}_{i=1}^T$ (especially $\{(\widetilde{\Delta C_a})_k(i)\}_{k,i}$) are determined by the parameters, thus by the Markov property, $\{(\Delta C_a)_k(i)\}$ are all independent for $k = 1, \ldots, n$, $i = 1, \ldots, T$ conditioned on the parameters $\overline{\Theta}$. Furthermore we suppose that $(\Delta C_a)_k(i) \sim \text{Pois}((\widetilde{\Delta C_a})_k(i))$. Thus, the likelihood of $\overline{\Theta}$ can be written as

$$P\left(\{(\Delta C_a)_k(i)\}_{k,i}|\overline{\Theta}\right) = \prod_{i=1}^{T} P\left(\{(\Delta C_a)_k(i)\}_{k=1}^{n}\middle|\overline{\Theta}\right) = \prod_{i=1}^{T}\prod_{k=1}^{n} \frac{((\widetilde{\Delta C_a})_k(i))^{(\Delta C_a)_k(i)}}{((\Delta C_a)_k(i))!} e^{-(\widetilde{\Delta C_a})_k(i)}. \quad (2.5)$$

Now, we denote the posterior distribution of $\overline{\Theta}$ as $\pi(\overline{\Theta})$. Then by Bayesian formula,

$$\pi(\overline{\Theta}) = p\left(\overline{\Theta}\middle|\{(\Delta C_a)_k(i)\}_{k,i}\right) = \frac{1}{Z} p\left(\{(\Delta C_a)_k(i)\}_{k,i}\middle|\overline{\Theta}\right) p(\overline{\Theta}), \quad (2.6)$$

where $p(\overline{\Theta})$ is the prior distribution of $\overline{\Theta}$ to be determined and $Z = p\left(\{(\Delta C_a)_k(i)\}_{k,i}\right)$ is a constant irrelevant to $\overline{\Theta}$. We further denote $V(\overline{\Theta}) = -\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i}\middle|\overline{\Theta}\right)\right) - \log(p(\overline{\Theta}))$, then

$$\pi(\overline{\Theta}) = \frac{1}{Z}\exp(-V(\overline{\Theta})). \quad (2.7)$$

To fit the realistic evolution of the epidemic more precisely, $\overline{\Theta}$ is estimated as $\overline{\Theta}^* = \arg\max \pi(\overline{\Theta}) = \arg\min V(\overline{\Theta})$ with reasonable prior distribution $p(\overline{\Theta})$. Then, MCMC sampling scheme starting from $\overline{\Theta}^*$ is applied to get the posterior distribution for $\overline{\Theta}$. This process might possess higher computational efficiency than choosing the initial point for MCMC randomly or empirically.

Choosing prior distribution

The remaining problem is to choose the prior distribution $p(\overline{\Theta})$. Presuming that the transmission rates in the regions owning more similarities are closer, $p(\overline{\Theta})$ is designed to combine the information of correlations between regions. In particular, given a matrix A which characterizes the pairwise similarities between the regions, and if we denote $\lambda = (\lambda_1, \ldots, \lambda_n)^T \in \mathbb{R}^n$,

$$p(\overline{\Theta}) = \frac{1}{C_{\sigma,A}}\exp(-\lambda^T(D-A)\lambda - \sigma\|\lambda\|_2^2), \quad (2.8)$$

where $D = \text{diag}\{d_1, \ldots, d_n\}$ is the degree matrix of A with $d_i = \sum_{j=1}^{i} A_{ij}$, $C_{\sigma,A} = \int_{\mathbb{R}^n}\exp(-\lambda^T(D-A)\lambda - \sigma\|\lambda\|_2^2)d\lambda$ is a constant depending on σ and A. Here, a small σ is chosen for $p(\overline{\Theta})$ to be a probability measure without imposing much restriction on λ. Then,

$$\begin{aligned} V(\overline{\Theta}) &= -\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i}\middle|\overline{\Theta}\right)\right) + \lambda^T(D-A)\lambda + \sigma\|\lambda\|_2^2 + \log C_{\sigma,A} \\ &= -\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i}\middle|\overline{\Theta}\right)\right) + \frac{1}{2}\sum_{i,j} a_{ij}(\lambda_i - \lambda_j)^2 + \sigma\sum_i \lambda_i^2 + \log C_{\sigma,A}. \end{aligned} \quad (2.9)$$

The parameter estimation procedure could be extended to the case when $\{\lambda_k\}$ are time-varying by modifying (2.11), which we will introduce in more detail in Section 4.2 for the real-world data in China and Europe.

A is constructed from affinity matrix W by further addressing the correlation between regions with more similarities. To treat different data sets and W with a unified approach, we assume that $\max_{i,j} W_{ij} = 1$ (W can be re-scaled entry-wise if necessary).

For a given $W = (W_{ij})_{i,j\in\{1,\dots,n\}}$ whose choice is detailed later, the next step of attaining A is to divide the n regions into d groups ($D_1, ..., D_d$, $\cup_{m=1}^{d} D_m = \{1, \dots, n\}$, and $\forall i \neq j$, $D_i \cap D_j = \emptyset$) where the regions in the same groups have more similarities. Then, for given $\beta \in (0, 1)$ and a given penalty factor $\mu > 0$, a_{ij} is constructed as follows:

$$a_{ij} = \mu \begin{cases} W_{ij}, & \text{if regions } i \text{ and } j \text{ are in the same group,} \\ \beta W_{ij}, & \text{otherwise.} \end{cases} \tag{2.10}$$

We remark that β in (2.10) is taken to be 0.1 for all the experiments in this paper. By constructing A as in (2.10), correlations for regions in the same groups are further strengthened, whose transmission parameters are imposed with stronger restrictions.

Now we specify the choice of W for data sets that will be analyzed later in this paper. For simulated data and real-world data in China, in which cases the transportation data are available, we construct W from the traveling volume matrices $\{W_i\}_{i=1}^{T}$. Specifically, $W := \bar{W}/\max_{i,j} \bar{W}_{ij}$, where $\bar{W} := \frac{1}{2}((\frac{1}{T}\sum_{l=1}^{T} W_l) + (\frac{1}{T}\sum_{l=1}^{T} W_l)^T)$. Nevertheless, for real-world data in Europe, where we are not aware of traveling data publicly available that are sufficient for the proposed model, W is just the all 1 adjacency matrix. We remark that the affinity matrix W may also be obtained by ways other than using the transportation data, as long as it reflects the similarities between districts.

From the definition of a_{ij} in (2.10), the inference of $\overline{\Theta}$ can be equivalently written as follows

$$\begin{aligned} \overline{\Theta}^* &= \arg\min V(\overline{\Theta}) \\ &= \arg\min \left(-\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i} \middle| \overline{\Theta} \right) \right) + \frac{1}{2}\sum_{i,j} a_{ij}(\lambda_i - \lambda_j)^2 + \sigma \sum_i \lambda_i^2 \right) \\ &= \arg\min \Bigg(-\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i} \middle| \overline{\Theta} \right) \right) + \frac{\mu}{2}\Bigg(\sum_m \sum_{i,j\in D_m} W_{ij}(\lambda_i - \lambda_j)^2 \\ &\qquad + \beta \sum_{m_1<m_2} \sum_{i\in D_{m_1}, j\in D_{m_2}} W_{ij}(\lambda_i - \lambda_j)^2 \Bigg) + \sigma \sum_i \lambda_i^2 \Bigg), \end{aligned} \tag{2.11}$$

it can be seen that by choosing the prior distribution as in (2.8), a l_2 regularization term $\frac{1}{2}\sum_{i,j} a_{ij}(\lambda_i - \lambda_j)^2$ is imposed for better generalization.

Finally, after choosing $p(\overline{\Theta})$ determined by A and σ, the optimization process $\overline{\Theta}^* = \arg\min V(\overline{\Theta})$ is accomplished through Matlab function `fmincon`. In addition, the posterior distribution for $\overline{\Theta}$ could be obtained by classical MCMC sampling scheme starting from $\overline{\Theta}^*$ as mentioned above.

We summarize the procedure described in Section 2.3 in Algorithm 1 below.

Algorithm 1 GLEaM model with Graph Laplacian Prior Distribution

Input: Data of total confirmed cases and total removed cases $\{(C_a)_k(i)\}_{k,i}$, $\{(\widehat{R_a})_k(i)\}_{k,i}$, traveling volume matrices $\{W_t\}_{t=1}^T$, graph Laplacian penalty factor μ, parameter β, prior parameter σ and affinity matrix W, division of n provinces $D_1, \ldots, D_d$, number of MCMC iterations N_{ite}

Output: Inferred parameters Θ^*, estimated marginal posterior distributions of Θ, predicted deterministic trajectory $\hat{y}_{1:T}$

1: Compute newly confirmed cases and newly removed cases, $(\Delta C_a)_k(i) \leftarrow (C_a)_k(i) - (C_a)_k(i-1)$, $(\Delta R_a)_k(i) \leftarrow (R_a)_k(i) - (R_a)_k(i-1)$

2: Compute $\gamma_k^* \leftarrow \arg\max_{\gamma_k} \prod_{i=1}^T \text{Pois}\left((\Delta R_a)_k(i) \middle| ((C_a)_k(i-1) - (R_a)_k(i-1))\gamma_k\right)$

3: Construct affinity matrix A with W, μ, β, and divisions $D_1, \ldots, D_d$ as in (2.10). Compute the degree matrix D, $D_{ii} \leftarrow \sum_j a_{ij}$

4: Compute the inferred parameters $\overline{\Theta}^* \leftarrow \arg\min_{\overline{\Theta}} V(\overline{\Theta}) = \arg\min_{\overline{\Theta}} \left(-\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i} \middle| \overline{\Theta} \right) \right) + \lambda^T (D - A)\lambda + \sigma \|\lambda\|_2^2 \right)$. $V(\overline{\Theta})$ should be changed accordingly if λ is allowed to be time-varying

5: $\Theta^* = \{\{\gamma_k^*\}_{k=1}^n, \overline{\Theta}^*\}$

6: Compute the predicted deterministic trajectory $\hat{y}_{1:T}$ by applying explicit Euler's method to (2.4) with Θ^*

7: Run MCMC staring from $\overline{\Theta}^*$ for N_{ite} iterations

8: Compute the estimated marginal posterior distribution of $\theta \in \overline{\Theta}$ by $\hat{p}(\theta | \{(\Delta C_a)_k(i)\}_{k=1\ldots,n,i=1,\ldots,T}) \leftarrow \sum_{l=1}^{n_{bin}} \frac{c_l}{N} \mathbf{1}_{\theta \in H_l}$, where $\{H_l\}_{k=1}^{n_{bin}}$ are the bins and $\{c_l\}_{k=1}^{n_{bin}}$ are the numbers of elements in the bins in N_{ite} iterations

2.4 Prediction of the Epidemic Trajectories with the Estimated Parameters

Once the parameters $\Theta^* = \{\{\gamma_k^*\}_{k=1}^n, \overline{\Theta}^*\}$ are estimated from the optimizations $\gamma_k^* = \arg\max_{\gamma_k} \prod_{i=1}^T \text{Pois}\left((\Delta R_a)_k(i) \middle| ((C_a)_k(i-1) - (R_a)_k(i-1))\gamma_k \right)$ and $\overline{\Theta}^* = \arg\min V(\overline{\Theta})$ as described in Section 2.3, the trajectories of newly confirmed cases could be simulated according to the stochastic dynamic process with Θ^*. Furthermore, trajectories could also be sampled from the posterior distribution of Θ instead of using Θ^* alone, which also takes the randomness from $\overline{\Theta}$ into account. Particularly, this could be achieved by sampling $\overline{\Theta}$ from MCMC and then simulating trajectories with the sampled $\{\{\gamma_k^*\}_{k=1}^n, \overline{\Theta}\}$. Additionally, deterministic trajectories determined by (2.4) could also be computed by explicit Euler's method as in Algorithm 1.

3 Experimental Results for Simulated Data

Two specific cases are considered for simulated data. We first remark that the regions in Section 2 are called as provinces in this section. Section 3.2 considers four provinces separated into two groups (the provinces in the same group are assumed to have more similarities) and with traffic between each pair of the provinces. Section 3.3 considers thirty provinces randomly separated into three groups, with the other settings similar to the previous case. Section 3.1 includes more details of the experimental settings and sensitivity analysis.

3.1 More Details of Experimental Settings and Sensitivity Analysis

Experimental settings

The results in Section 3 are for 100 replicas. In each replica, three random trajectories are sampled independently according to the stochastic model with prefixed parameters, part of which are treated as the ground truth training, validation, and testing trajectory, respectively (see more details in Appendix B.1.1). For each model, we first fit the parameters using the training trajectory and then predict the testing trajectory using the estimated parameters. Note that we detail the choice of hyper-parameters for the model proposed in Appendix B.2. In particular, the penalty factor $\mu > 0$ is chosen by cross-validation and chosen as the value minimizing the validation error, since for the simulated data, the validation error are identically distributed as the testing error. The comparison of trajectory prediction is from one typical realization, for which we compare the ground truth training and testing trajectories with the fitted training and predicted testing trajectories for all the models. Additionally, parameter estimation and quantitative evaluations are compared with mean and standard deviation over all 100 replicas. The detailed computations of training, validation, and testing errors can be found in Appendix C.

Sensitivity analysis of σ

Note that parameter inference with the proposed model involves the parameter σ, as shown in (2.11). Therefore, the sensitivity analysis for the parameter σ in (2.11) is performed for the four provinces case. Specifically, the results for σ varying from 10^{-6} to 10^{0} are presented and compared. Details can be found in Sections 3.2.4 and 3.2.5. Similar results are obtained for other data sets, and details are omitted.

Mismatched partition of regions

We also note that the graph Laplacian penalty of the proposed model depends on the partitioning the regions into several groups, as described in Section 2.3. Since the graph knowledge is usually not fully known, it is a question whether our methods can still perform well without accurate prior knowledge. For the thirty provinces case, we report the results of the proposed model with a mismatch between the partition of the regions and the ground truth division, the details of which can be found in Section 3.3.

3.2 Four Provinces Case

3.2.1 Data Description

In this simulated study, we let $n = 4$, $T = 20$, and set the threshold T_{th} separating training and testing data to be 10 (more detailed can be seen in Appendix B.1.1). The other prefixed parameters are listed below:

- For $k \in \{1, 2, 3, 4\}$, $N_k(0) = 10^6$, $E_k(0) = 30$, $H_k(0) = 10$.
- For $l \in \{1, \ldots, T\}$, $i, j \in \{1, 2, 3, 4\}$ and $i \neq j$, $(W_l)_{ij} = 5 \times 10^3$.
- $\lambda_1 = 0.5$, $\lambda_2 = 0.47$, $\lambda_3 = 0.4$, $\lambda_4 = 0.37$, $\delta = \gamma_1 = \cdots = \gamma_4 = 0.14$.

The four provinces are divided into two groups, with the first group consisting of province 1 and 2 and the second group consisting of province 3 and 4. The similarities within groups are reflected in the settings that the values of $\{\lambda_k\}$ are closer for provinces in the same group.

3.2.2 Models to Compare

We first specify the models to be compared below. The last one is the proposed model, and the first four models serve as baselines with different settings.

1. The model with uniform prior distribution, without heterogeneity or migration.
2. The model with uniform prior distribution, without heterogeneity but with migration.
3. The model with uniform prior distribution, with heterogeneity but without migration.
4. The model with uniform prior distribution, with both heterogeneity and migration.
5. The model with prior distribution based on graph Laplacian, with both heterogeneity and migration.

For better illustration and comparison between the models in the experiment results, the Models 1-5 are summarized in Table 2 below.

Model	Migration	Heterogeneity	Prior of Θ
1	×	×	uniform prior
2	✓	×	uniform prior
3	×	✓	uniform prior
4	✓	✓	uniform prior
5	✓	✓	graph Laplacian prior

Table 2: Models to be compared when the transportation data are available. Model 5 is the proposed model in this paper, and Models 1-4 are baseline models with different settings.

First, the models with uniform prior distributions themselves (Models 1-4) are compared according to whether two key assumptions exist in the model:

(1) Whether the transmission rates $\{\lambda_k\}$ are allowed to vary over regions.

(2) Whether there exists transportation between regions.

Then, the model with prior distribution based on graph Laplacian (Model 5) is compared with those using uniform distributions as prior distributions (Models 1-4). The former one utilizes the correlation between subpopulations by adding a l_2 regularization term for the model. In contrast, only lower and upper bounds are imposed on parameters without other prior information being used in the latter ones.

For Model 5, we conduct the sensitivity analysis for parameter σ in (2.8) and present results for $\sigma = 10^{-6}, 10^{-3}$ and 10^0 respectively. Then for Model 5, by (2.11), $\overline{\Theta}$ are estimated by the optimization problem in (3.1):

$$\overline{\Theta}^* = \arg\min(-\log p(y_{1:T_{th}}|\overline{\Theta}) + \mu\left((\lambda_1-\lambda_2)^2 + (\lambda_3-\lambda_4)^2 + \beta \sum_{(i,j)\neq(1,2)\text{or}(3,4)} (\lambda_i-\lambda_j)^2\right) + \sigma\|\lambda\|_2^2), \tag{3.1}$$

where β is taken to be 0.1.

We remark that in Models 1-4, the estimation of $\Theta = \{\{\gamma_k\}_{k=1}^n, \overline{\Theta}\}$ still follows a similar two-step procedure as in Model 5, and the first step of obtaining $\gamma_k^* = \arg\max_{\gamma_k} \prod_{i=1}^T \text{Pois}\big((\Delta R_a)_k(i)\big|((C_a)_k(i-1)-(R_a)_k(i-1))\gamma_k\big)$ remains formally the same. The difference lies in the optimization object of $\overline{\Theta}$. First, the l_2 regularization term becomes prior knowledge of the parameters' upper and lower bounds. Second, for models without heterogeneity of parameters, $\{\lambda_k\}_{k=1}^n$ are forced to be the same in (2.4). For models without transportation between regions, terms involving W_t disappear in (2.4). Additionally, for the other

data sets considered in the following sections, the parameter estimation methods for the baseline models are similar and thus will not be repeated.

Finally, note that the model in [12] is similar to Models 1 and 2, since they all assume a spatially homogeneous transmission parameter. However, [12] assumed that the epidemic was seeded from one seed region while Models 1 and 2 do not make such assumption. Moreover, [12] focused more on the spread of the epidemic from the seed region at the early stage of the pandemic, and only the introduction dates in the other regions were utilized for the estimation of transmission parameters. In comparison, the estimation of transmission rate in Models 1 and 2 exploits the data in all regions in the whole process.

3.2.3 Results of Trajectory Prediction

First, we remark that in Model 5, μ is chosen to minimize the mean of validation errors over 100 replicas. The weighted (simply averaged) validation errors, $\mathrm{MAE}_{(w)}^{[\mathrm{Val}]}$ and $\mathrm{MSE}_{(w)}^{[\mathrm{Val}]}$ ($\mathrm{MAE}_{(s)}^{[\mathrm{Val}]}$ and $\mathrm{MSE}_{(s)}^{[\mathrm{Val}]}$), are defined as in Appendix C. We remark that the superscript $^{[\mathrm{Val}]}$ refers to when the error is computed on validation data, and the subscripts $_{(w)}$ and $_{(s)}$ denote that the errors are the weighted and simple average of relative errors over time respectively.

The mean of weighted validation errors $\mathrm{MAE}_{(w)}^{[\mathrm{Val}]}$ and $\mathrm{MSE}_{(w)}^{[\mathrm{Val}]}$ are shown in Figure 1 below, and the mean of simply averaged validation errors $\mathrm{MAE}_{(s)}^{[\mathrm{Val}]}$ and $\mathrm{MSE}_{(s)}^{[\mathrm{Val}]}$ are shown in Figure A1. μ is chosen to be $10^{2.7}$, which minimizes the mean of validation errors, as can be seen from Figure 1 and Figure A1.

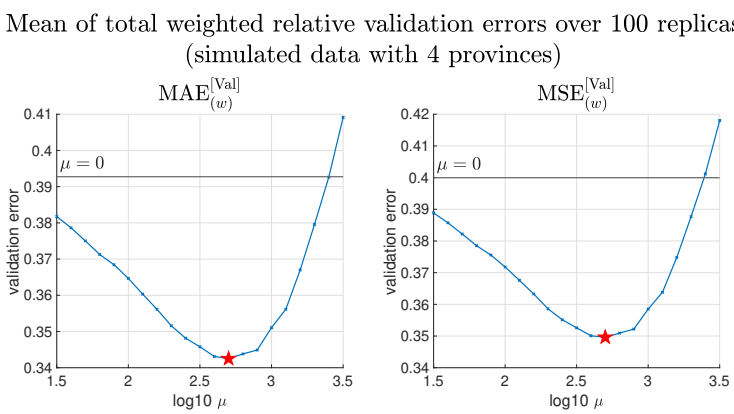

Figure 1: Mean of weighted validation errors $\mathrm{MAE}_{(w)}^{[\mathrm{Val}]}$ and $\mathrm{MSE}_{(w)}^{[\mathrm{Val}]}$ against μ (computed by (A5)) over 100 replicas for simulated data (four provinces). The horizontal lines show the values of the mean of the validation errors when $\mu = 0$.

Then, The trajectories of a typical realization are plotted in Figure 2 and the differences between the true and fitted trajectories are shown in Figure 3. As shown in these two figures, heterogeneity helps improve the prediction of testing data more than transportation, while introducing migration without heterogeneity of parameters worsens the estimate as can also be noticed from Table 4. More explanations can be found in Section 3.2.5.

Additionally, Model 5 with prior distrbution based on graph Laplacian lowers the absolute errors of predicted trajectories compared with Model 4.

As shown in Figures 2 and 3, Models 1 and 2 have slightly better generalization accuracy than Model 5 for Province 3. On the one hand, for data in this replica, the estimated λ_3 is 0.3860 using Model 5 and 0.4640 using Models 1 or 2 (recalling that the ground truth λ_3 is 0.4). On the other hand, due to

the randomness of the generated testing data, the sampled newly confirmed cases are much more than the deterministic ones in Province 3 obtained by running (2.4) with the ground truth parameters. Hence, although all these estimates of λ_3 are biased from the ground truth 0.4, estimates using Models 1 and 2, which are biased up, lead to less absolute errors.

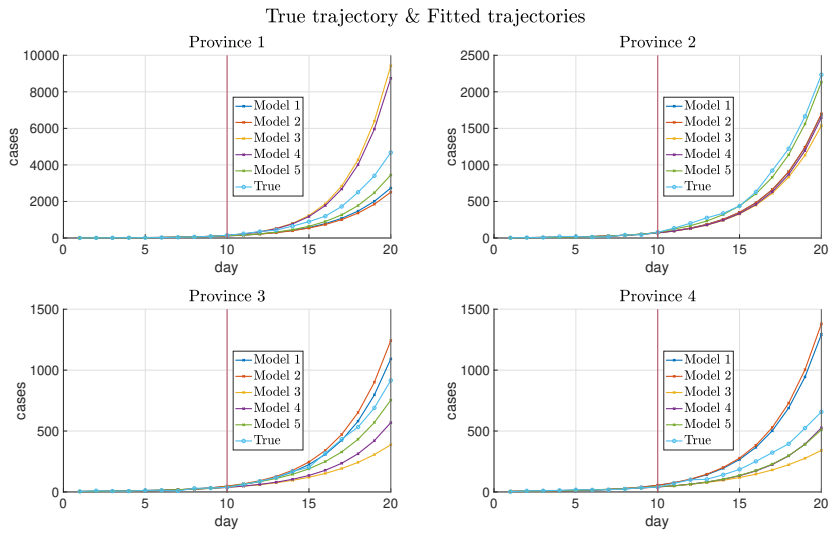

Figure 2: True and fitted trajectories of Models 1-5 in the four provinces. The vertical lines show the threshold $T_{th} = 10$.

3.2.4 Results of Parameter Estimation

The mean and standard deviation of $\{\lambda_i\}$ estimated by the five models for four provinces case are reported in Table 3. It can be observed that models allowing heterogeneity estimate parameters more accurately, and Model 5 that integrates the correlation leads to slightly better estimate for λ_2. We can see that compared to Model 4, the estimates of smaller λ_k's (such as $\lambda_2, \lambda_3, \lambda_4$) become larger, and the estimate of λ_1 which has the largest value becomes smaller, since the graph Laplacian penalty tends to make $\{\lambda_k\}_{k=1}^n$ closer to each other.

Moreover, we performed a sensitivity analysis for the hyper-parameter σ to check that the results are robust to σ. The last two rows of Table 3 show the parameter estimation results for Model 5 with $\sigma = 10^{-3}$ and 10^0 respectively (more values of $\sigma \in [10^{-8}, 10^0]$ are tested and the results are similar as well). We observe that the variation of parameters estimated by Model 5 with σ varying from 10^{-6} to 10^0 does not exceed 1%. Therefore, the parameter estimation results are not sensitive to the choice of σ as long as σ is not too large.

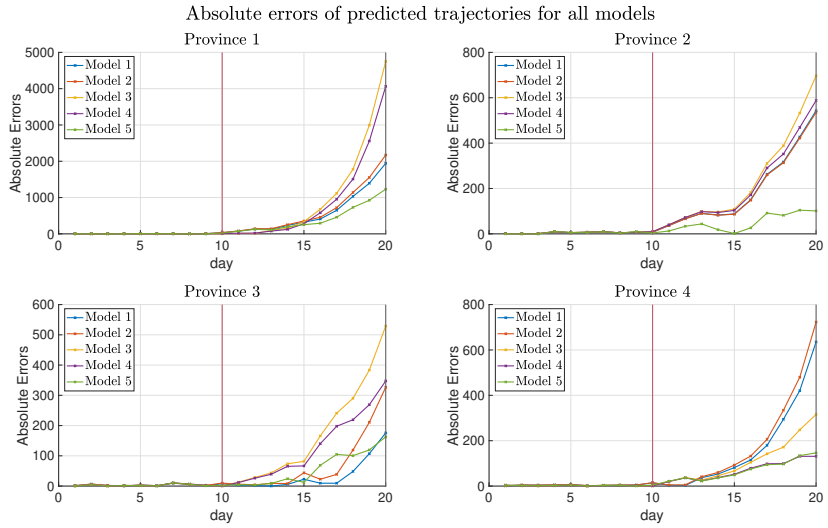

Figure 3: Absolute errors of fitted trajectories of Models 1-5 in the four provinces. The vertical lines show the threshold $T_{th} = 10$.

Model	M	H	GL prior	λ_1	λ_2	λ_3	λ_4
1	×	×	×	0.4458 ± 0.0192			
2	✓	×	×	0.4459 ± 0.0192			
3	×	✓	×	0.4927 ± 0.0331	0.4620 ± 0.0370	0.3988 ± 0.0389	0.3705 ± 0.0349
4	✓	✓	×	0.4980 ± 0.0337	0.4643 ± 0.0376	0.3922 ± 0.0407	0.3605 ± 0.0366
	✓	✓	✓($\sigma = 10^{-6}$)	0.4805 ± 0.0254	0.4654 ± 0.0260	0.3979 ± 0.0256	0.3879 ± 0.0231
5	✓	✓	✓($\sigma = 10^{-3}$)	0.4805 ± 0.0254	0.4654 ± 0.0260	0.3979 ± 0.0256	0.3879 ± 0.0231
	✓	✓	✓($\sigma = 10^{0}$)	0.4800 ± 0.0254	0.4648 ± 0.0260	0.3972 ± 0.0256	0.3871 ± 0.0230

Table 3: Estimated λ_i of five models. Recall that the ground truth is $\lambda_1 = 0.5$, $\lambda_2 = 0.47$, $\lambda_3 = 0.4$, $\lambda_4 = 0.37$. Models in Column 1 are detailed in Table 2. Columns 2-4 indicate whether the model permits **migration** between provinces (for "**M**"), the **heterogeneity** of parameters (for "**H**") and the **Graph Laplacian prior** (for "**GL prior**"), respectively. λ_i are inferred for each of the 100 replicas, and then the mean and standard deviation are presented in the table above. In addition, $\mu = 10^{2.7}$ ($\sigma = 10^{-6}, 10^{-3}, 10^{0}$ respectively) in Model 5 for inference of parameters as defined in (3.1), which minimizes mean of relative validation errors over 100 replicas as shown in Figure 1 and Figure A1.

3.2.5 Further Model Evaluation

The training and testing errors, $\text{MAE}^{[\text{Tr}]}_{(w)}$, $\text{MAE}^{[\text{Te}]}_{(w)}$, $\text{MSE}^{[\text{Tr}]}_{(w)}$, $\text{MSE}^{[\text{Te}]}_{(w)}$, as defined in Appendix C (in the equation (A5)) , are listed below in Table 4 with mean and standard deviation. We remind the readers that the superscripts $^{[\text{Tr}]}$ and $^{[\text{Te}]}$ represent the errors are computed on training and testing data respectively, and the subscript $_{(w)}$ denotes that the error is weighted average of the daily relative errors over time. It can be seen from Table 4 that the presence of both heterogeneity and transportation helps reduce the training and testing errors by comparing the first four models. By comparing Model 4 and Model 5, it can be seen that using the graph Laplacian prior could further reduce the testing errors.

Model	M	H	GL prior	$\text{MAE}^{[\text{Tr}]}_{(w)}$	$\text{MAE}^{[\text{Te}]}_{(w)}$	$\text{MSE}^{[\text{Tr}]}_{(w)}$	$\text{MSE}^{[\text{Te}]}_{(w)}$
1	×	×	×	0.1728 ± 0.0273	0.4700 ± 0.1879	0.2379 ± 0.0375	0.4801 ± 0.1887
2	✓	×	×	0.1828 ± 0.0296	0.5432 ± 0.2048	0.2487 ± 0.0390	0.5551 ± 0.2067
3	×	✓	×	0.1417 ± 0.0232	0.4095 ± 0.1692	0.1997 ± 0.0348	0.4173 ± 0.1683
4	✓	✓	×	0.1418 ± 0.0229	0.3834 ± 0.1672	0.1997 ± 0.0343	0.3900 ± 0.1665
	✓	✓	✓($\sigma = 10^{-6}$)	0.1481 ± 0.0230	0.3300 ± 0.1519	0.2069 ± 0.0355	0.3365 ± 0.1513
5	✓	✓	✓($\sigma = 10^{-3}$)	0.1481 ± 0.0230	0.3300 ± 0.1519	0.2069 ± 0.0355	0.3365 ± 0.1513
	✓	✓	✓($\sigma = 10^{0}$)	0.1482 ± 0.0230	0.3283 ± 0.1498	0.2071 ± 0.0356	0.3347 ± 0.1490

Table 4: Training and testing error with standard deviation for simulated data (four provinces) of five models. The formulas of errors are detailed in Appendix C (equation (A5) or its modifications to compute testing errors). Models in Column 1 are detailed in Table 2. Columns 2-4 indicate whether the model permits **migration** between provinces (for "**M**"), the **heterogeneity** of parameters (for "**H**") and the **Graph Laplacian prior** (for "**GL prior**"), respectively. $\text{MAE}^{[\text{Tr}]}_{(w)}$, $\text{MAE}^{[\text{Te}]}_{(w)}$, $\text{MSE}^{[\text{Tr}]}_{(w)}$ and $\text{MSE}^{[\text{Te}]}_{(w)}$ are computed for each of the 100 replicas, then the mean and standard deviation are presented in the table above. In addition, $\mu = 10^{2.7}$ ($\sigma = 10^{-6}, 10^{-3}, 10^{0}$ respectively) in Model 5 for inference of parameters as defined in (3.1), which minimizes relative validation errors as shown in Figure 1 and Figure A1.

In addition, it can be seen from Table 4 that the errors increase greatly after transportation is included while heterogeneity remains absent. A possible explanation for this might be that without heterogeneity of parameters and transportation between provinces, the estimated λ_k's are lower than the true λ_k's for group 1, which leads to that the estimated newly confirmed cases are fewer than the true ones in group 1. For the same reason, the estimated newly confirmed cases are higher than the true ones in group 2. When the transportation is considered, more confirmed cases in group 1 are transferred to group 2 than the cases transported in the opposite direction. As a result, when the transmission parameters do not have heterogeneity, migration between provinces will worsen the prediction performance compared to the case without migration.

Furthermore, the last two rows of Table 4 report the training and testing errors for Model 5 with the same $\mu = 10^{2.7}$ while $\sigma = 10^{-3}$ and $\sigma = 10^{0}$ respectively. As a consequence of the robustness of the parameter estimation regarding σ, the errors of Model 5 are also robust to σ. The similar analysis is also performed for the other data sets and the similar results can be obtained which we do not report repetitively. Hereinafter, the results are presented with $\sigma = 10^{-6}$.

The plots of the mean of weighted and simply averaged testing errors $\text{MAE}^{[\text{Te}]}_{(w)}$, $\text{MSE}^{[\text{Te}]}_{(w)}$, $\text{MAE}^{[\text{Te}]}_{(s)}$ and $\text{MSE}^{[\text{Te}]}_{(s)}$ against varying μ are shown in Figure 4 and Figure A2 respectively. Recall that the subscripts (w) and (s) denote the weighted and simple average respectively. Note that Model 5 with $\mu = 10^{2.7}$, which minimizes validation errors as shown in Figure 1 and Figure A3, achieves the minimal values of testing errors $\text{MAE}^{[\text{Te}]}_{(w)}$ and $\text{MSE}^{[\text{Te}]}_{(w)}$ (also $\text{MAE}^{[\text{Te}]}_{(s)}$ and $\text{MSE}^{[\text{Te}]}_{(s)}$) . This is because validation and testing errors have the same distribution in this case. We also note that the testing error of Model 4 is nearly the same as that of Model 5 with small μ, since Model 4 is basically the same as Model 5 with $\mu = 0$.

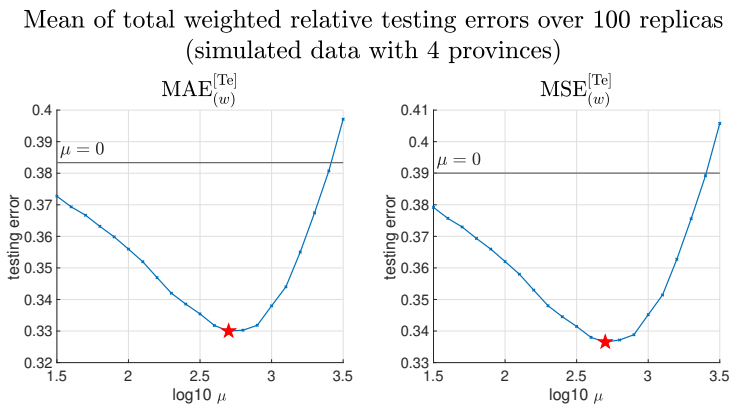

Figure 4: Mean of weighted testing errors $\mathrm{MAE}^{[\mathrm{Te}]}_{(w)}$ and $\mathrm{MSE}^{[\mathrm{Te}]}_{(w)}$ against μ (computed by (A5)) over 100 replicas for simulated data (four provinces). The horizontal lines show the values of the mean of the testing errors when $\mu = 0$.

3.3 Thirty Provinces Case

In this simulation study, we present the results for the simulated data involving a larger number of regions.

3.3.1 Data Description

We set the total number of regions $n = 30$, the total days considered $T = 30$, and set the threshold that separates the training and testing set as $T_{th} = 10$. The other prefixed parameters are listed as below:

- For $k \in \{1, 2, \ldots, n\}$, $N_k(0) = 10^6$, $E_k(0) = 30$, $H_k(0) = 10$.
- For $l \in \{1, \ldots, T\}$, $i, j \in \{1, 2, \ldots, n\}$ and $i \neq j$, $(W_l)_{ij} = 5 \times 10^3$.
- For provinces in Group 1:

  $\lambda_4 = 0.307, \lambda_5 = 0.295, \lambda_6 = 0.303, \lambda_{13} = 0.299, \lambda_{16} = 0.296, \lambda_{21} = 0.296, \lambda_{25} = 0.294, \lambda_{28} = 0.293, \lambda_{29} = 0.305.$

  For provinces in Group 2:

  $\lambda_1 = 0.416, \lambda_2 = 0.410, \lambda_7 = 0.405, \lambda_8 = 0.393, \lambda_9 = 0.400, \lambda_{11} = 0.402, \lambda_{12} = 0.406, \lambda_{15} = 0.407, \lambda_{18} = 0.395, \lambda_{19} = 0.410, \lambda_{22} = 0.396, \lambda_{23} = 0.419.$

  For provinces in Group 3:

  $\lambda_3 = 0.493, \lambda_{10} = 0.512, \lambda_{14} = 0.514, \lambda_{17} = 0.503, \lambda_{20} = 0.507, \lambda_{24} = 0.498, \lambda_{26} = 0.488, \lambda_{27} = 0.495, \lambda_{30} = 0.482.$

Note that the thirty provinces are randomly assigned to three groups, and we denote this ground truth partition as P:

- D_1: Provinces 4, 5, 6, 13, 16, 21, 25, 28, 29.
- D_2: Provinces 1, 2, 7, 8, 9, 11, 12, 15, 18, 19, 22, 23.

- D_3: Provinces 3, 10, 14, 17, 20, 24, 26, 27, 30.

The transmission rates of the provinces in the same groups are closer as listed above. Specifically, the transmission rates of provinces in D_1, D_2, and D_3 are randomly sampled from normal distribution with mean 0.3, 0.4, and 0.5 respectively and all with standard deviation 0.01.

3.3.2 Models to Compare

We also compare the Models 1-5 listed in Table 2 as detailed in Section 3.2.2.

Note that for Model 5, the parameter σ in (2.8) is still fixed to be 10^{-6}. When we assume that the ground truth group division P as listed above is known, in model performance evaluation as well as the optimization of μ for Model 5, by (2.11), $\overline{\Theta}$ are estimated by the optimization problem in (3.2):

$$\overline{\Theta}^* = \arg\min \Bigg(-\log\left(p\left(y_{1:T_{th}}|\overline{\Theta}\right)\right) + \frac{\mu}{2}\Bigg(\sum_{m=1}^{3} \sum_{i,j\in D_m} (\lambda_i - \lambda_j)^2 + \beta \sum_{1\le m_1<m_2\le 3} \sum_{i\in D_{m_1}, j\in D_{m_2}} (\lambda_i - \lambda_j)^2 \Bigg) + \sigma \sum_i \lambda_i^2 \Bigg), \tag{3.2}$$

where $\beta = 0.1$.

Moreover, for Model 5, we also report results when there are mismatches in partition of the provinces, since in real-world applications the graph information may not be fully known. Specifically, since most of the results reported in the following sections are for Provinces 4 (in D_1), 1 (in D_2), and 3 (in D_3), we consider the other two partitions $P' = \{D_1', D_2', D_3'\}$ and $P'' = \{D_1'', D_2'', D_3''\}$, which are different from P in the assignments of Provinces 4, 1, and 3, and another three provinces respectively. Comparison between the results using the ground truth partition and the results of these two different kinds of mismatches may reflect the potentially different impact of these mismatches on the results. For Model 5 with partitions P' or P'', $\overline{\Theta}$ is estimated by (3.2) with $P = \{D_1, D_2, D_3\}$ replaced by $P' = \{D_1', D_2', D_3'\}$ or $P'' = \{D_1'', D_2'', D_3''\}$.

The partition $P' = \{D_1', D_2', D_3'\}$ deviates from P in the assignment of Provinces 4 (in D_1), 1 (in D_2), and 3 (in D_3):

- D_1': Provinces **3**, 5, 6, 13, 16, 21, 25, 28, 29.
- D_2': Provinces **4**, 2, 7, 8, 9, 11, 12, 15, 18, 19, 22, 23.
- D_3': Provinces **1**, 10, 14, 17, 20, 24, 26, 27, 30.

The partition $P'' = \{D_1'', D_2'', D_3''\}$ deviates from P in the assignment of Provinces 5 (in D_1), 2 (in D_2), and 10 (in D_3):

- D_1'': Provinces 4, **10**, 6, 13, 16, 21, 25, 28, 29.
- D_2'': Provinces 1, **5**, 7, 8, 9, 11, 12, 15, 18, 19, 22, 23.
- D_3'': Provinces 3, **2**, 14, 17, 20, 24, 26, 27, 30.

For mismatched partitions P' and P'', the provinces assigned to the wrong groups are marked in bold.

We remark that the parameter inference for Models 1-4 is the same as detailed in Section 3.2.2.

3.3.3 Results of Trajectory Prediction

Note that the validation and testing errors have the same distribution for the simulated data. Therefore, for Model 5 (with the ground truth partition P), we choose $\mu = 10^{1.8}$, which minimizes the mean of the validation errors as marked with red pentagrams in Figure 5 and Figure A3. The following results for Model 5 are all obtained with $\mu = 10^{1.8}$.

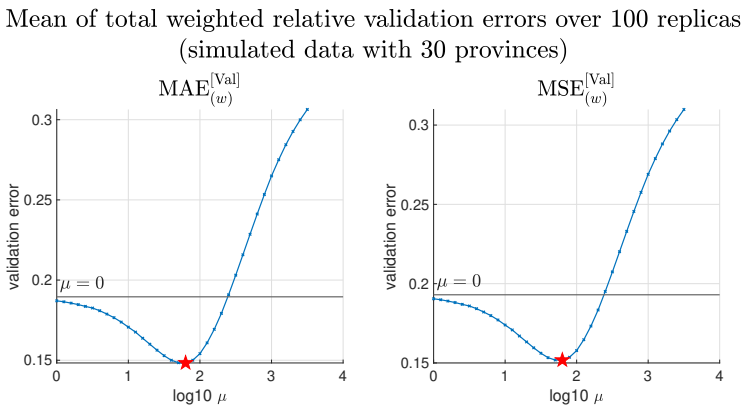

Figure 5: Mean of weighted validation errors $\mathrm{MAE}_{(w)}^{[\mathrm{Val}]}$ and $\mathrm{MSE}_{(w)}^{[\mathrm{Val}]}$ against μ (computed by (A5)) over 100 replicas for simulated data (thirty provinces). The horizontal lines show the values of the mean of the validation errors when $\mu = 0$.

Then, we choose three provinces from the total thirty provinces (one province from each group), and plot the prediction of the trajectories and also the absolute errors from one specific replica as in Figures 6 and 7. It can be observed from Figure 7 that Model 5 (the green lines) achieve the most accurate prediction results due to the graph Laplacian regularization. Model 4 (the purple lines) also have good generalization performance, thanks to the heterogeneity and transportation involved in this model.

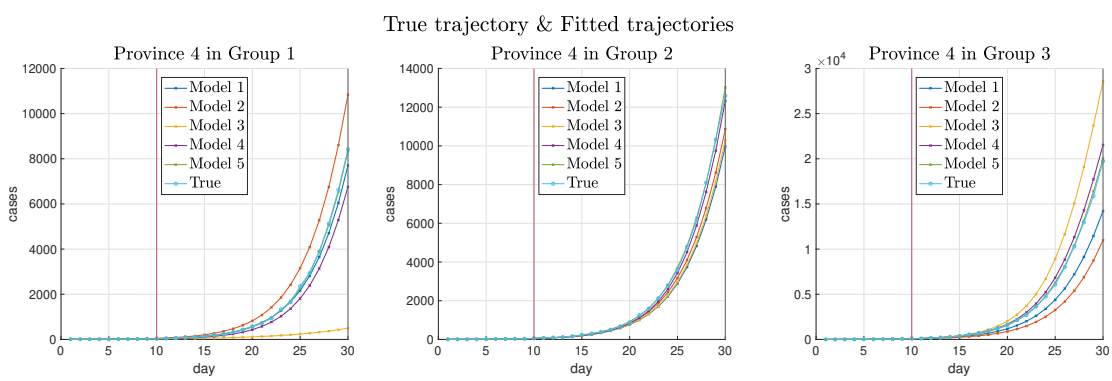

Figure 6: True and fitted trajectories of Models 1-5 for three provinces chosen from the total thirty provinces. The vertical lines show the threshold $T_{th} = 10$.

We note that Model 3 also behaves much worse than the best models, especially in Provinces 4 (in D_1) and 3 (in D_3). In fact, the estimates for the λ_k's in these three provinces using Model 3 are $0.2908, 0.4198$, and 0.4639, respectively, which are not much different from the ground truth values. Nevertheless, Model

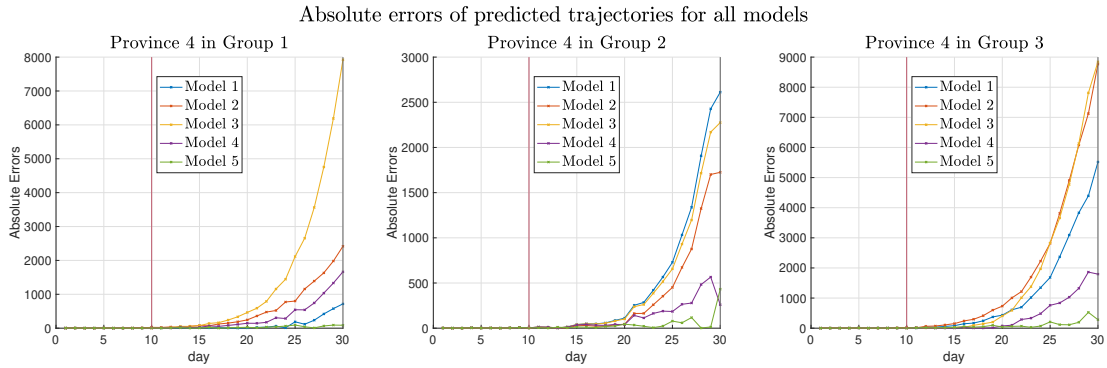

Figure 7: Absolute errors fitted trajectories of Models 1-5 for three provinces chosen from the total thirty ones. The vertical lines show the threshold $T_{th} = 10$.

3 does not involve transportation between provinces. This leads to the confirmed cases of Province 3 in D_3 (which are much more than the cases of provinces in the other two groups) not output to other provinces; hence the predicted newly confirmed cases are much more than the truth. A similar situation happens in the case of Province 4 (in D_1). Namely, Model 3 does not take the imported cases from the provinces in the other two groups into account, causing the predicted newly confirmed cases to be much less than the truth.

3.3.4 Results of Parameter Estimation

The mean and standard deviation of λ_k's in Provinces 4 (in D_1), 1 (in D_2), and 3 (in D_3) are reported in Table 5. We can still see that Models 3-5 estimate the transmission parameters more accurately than Models 1 and 2 especially for Provinces 4 and 3. In addition, the estimates by Model 5 of λ_k are closer than those by Model 4, due to the existence of the regularization term.

Model	M	H	GL prior	λ_4	λ_1	λ_3
1	×	×	×		0.4223 ± 0.0067	
2	✓	×	×		0.4227 ± 0.0067	
3	×	✓	×	0.3743 ± 0.0349	0.4101 ± 0.0284	0.4537 ± 0.0324
4	✓	✓	×	0.3051 ± 0.0383	0.4093 ± 0.0350	0.4902 ± 0.0326
(P)	✓	✓	✓	0.3332 ± 0.0172	0.4071 ± 0.0175	0.4784 ± 0.0209
5 (P')	✓	✓	✓	0.3669 ± 0.0148	0.4335 ± 0.0180	0.4375 ± 0.0218
(P'')	✓	✓	✓	0.3430 ± 0.0174	0.4051 ± 0.0174	0.4761 ± 0.0210

Table 5: Estimated λ_i of five models. Recall that the ground truth is $\lambda_4 = 0.307$ (in D_1), $\lambda_1 = 0.416$ (in D_2), $\lambda_3 = 0.493$ (in D_3). Remarks for Columns 1-4 and estimates of λ_k's are the same as those in Table 3. In addition, $\mu = 10^{1.8}$ in Model 5 for inference of parameters as defined in (3.2), which minimizes mean of relative validation errors over 100 replicas (computed with partition P) as shown in Figure 5 and Figure A3. The partitions of Model 5 are P, P', and P'' respectively. P is the ground truth underlying graph structure, and P' and P'' are mismatched partitions introduced in Section 3.3.2.

The last three rows show the parameter estimates using Model 5 with three partitions P, P', and P''

respectively. Recall that P is the ground truth underlying graph structure, and P' and P'' are mismatched partitions introduced in Section 3.3.2. It can be seen that compared with the estimated parameters using Model 5 with P' that differs from P in the grouping of Provinces 4, 1, and 3, the estimates using Model 5 with P'' that differs from P in the grouping of Provinces 5, 2, and 10 are closer to the estimates using Model 5 with the ground truth partition P. Hence, the results imply that for the parameter inference of simulated data generated from a certain partition, the incorrect division causes more discrepancy in the mismatched regions when comparing the estimates using the ground truth partition.

3.3.5 Further Model Evaluation

The mean and standard deviation of the weighted training and testing errors (formulas can be found in Appendix C, specifically in (A5)) are presented in Table 6. We see that Model 5 with $\mu = 10^{1.8}$ achieves the minimum testing errors among all the models, and hence have the best generalization performance.

Model	M	H	GL prior	$\text{MAE}^{[\text{Tr}]}_{(w)}$	$\text{MAE}^{[\text{Te}]}_{(w)}$	$\text{MSE}^{[\text{Tr}]}_{(w)}$	$\text{MSE}^{[\text{Te}]}_{(w)}$
1	×	×	×	0.1739 ± 0.0092	0.1704 ± 0.0543	0.2445 ± 0.0135	0.1744 ± 0.0534
2	✓	×	×	0.2427 ± 0.0136	0.3387 ± 0.0573	0.3128 ± 0.0171	0.3420 ± 0.0566
3	×	✓	×	0.1541 ± 0.0082	0.4920 ± 0.0594	0.2212 ± 0.0137	0.4973 ± 0.0599
4	✓	✓	×	0.1549 ± 0.0083	0.1960 ± 0.0647	0.2200 ± 0.0139	0.1994 ± 0.0649
(P)	✓	✓	✓	0.1648 ± 0.0083	0.1495 ± 0.0579	0.2312 ± 0.0135	0.1531 ± 0.0570
5 (P')	✓	✓	✓	0.1702 ± 0.0084	0.1612 ± 0.0522	0.2364 ± 0.0135	0.1650 ± 0.0514
(P'')	✓	✓	✓	0.1707 ± 0.0085	0.1648 ± 0.0539	0.2372 ± 0.0134	0.1685 ± 0.0532

Table 6: Training and testing error with standard deviation for simulated data (thirty provinces) of five models. The formulas of errors are detailed in Appendix C (equation (A5) or its modifications to compute testing errors). Remarks for Columns 1-4 and presented results of errors are the same as those in Table 4. In addition, $\mu = 10^{1.8}$ in Model 5 for inference of parameters as defined in (3.1) which minimizes mean of relative validation errors over 100 replicas (computed with partition P) as shown in Figure 5 and Figure A3. The partitions of Model 5 are P, P', and P'' respectively. P is the ground truth underlying graph structure, and P' and P'' are mismatched partitions introduced in Section 3.3.2.

The last three rows show the comparison of errors for Model 5 with three partitions P, P', and P'' respectively. We may see that the testing errors of Model 5 with the incorrect partitions are basically the same and larger than those with the ground truth partitions. Nevertheless, these errors are still smaller than the other baseline models. Therefore, the incorrect division might worsen the generalization performance of Model 5; however, when the division does not deviate from the ground truth much, it still behaves better than Models 1-4 with uniform prior.

Moreover, as shown in Table 6, both Model 2, which allows transportation but not heterogeneity, and Model 3, that introduces heterogeneity but no transportation, behave worse than the baseline Model 1. Model 2 having more significant testing errors is due to the same reason as the case of four provinces, as has been analyzed in Section 3.2.5. The reason for Model 3 not being able to predict well is explained in detail in Section 3.3.3.

However, we also note that Model 4, which involves both transportation and heterogeneity of parameters, has greater testing errors than Model 1 on average. This phenomenon may be partly explained by the fact that, on the one hand, there may be more infected cases imported from regions with higher λ_k's to those with lower transmission rates. On the other hand, although the transmission parameters are the same and all-around 0.42 in Model 1, E_0 and H_0 are still allowed to be spatially heterogeneous, which

makes the trajectories not that far away from the ground truth. Meanwhile, though the λ_k's are allowed to be spatially heterogeneous in Model 4, it often underestimates λ_k's for provinces in D_1 and overestimates λ_k's for provinces in D_3, making the trajectories deviate from the ground truth.

Figures 8 and A4 plot the change of weighted and simply averaged testing errors $\mathrm{MAE}^{[\mathrm{Te}]}_{(w)}$, $\mathrm{MSE}^{[\mathrm{Te}]}_{(w)}$, $\mathrm{MAE}^{[\mathrm{Te}]}_{(s)}$ and $\mathrm{MSE}^{[\mathrm{Te}]}_{(s)}$ regarding μ (recall that the subscripts (w) and (s) denote weighted and simple average respectively). We can still observe that Model 5 with the minimizers of validation errors also has the minimal testing errors, since the errors have the same distribution for the simulated data.

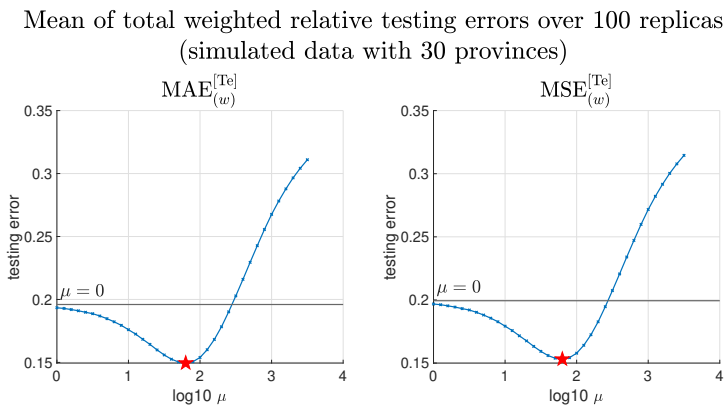

Figure 8: Mean of weighted testing errors $\mathrm{MAE}^{[\mathrm{Te}]}_{(w)}$ and $\mathrm{MSE}^{[\mathrm{Te}]}_{(w)}$ against μ (computed by (A5)) over 100 replicas for simulated data (thirty provinces). The horizontal lines show the values of the mean of the testing errors when $\mu = 0$.

4 Experimental Results on COVID-19 Data

In this section, we apply our model to two real-world data sets, the COVID-19 data in China (from January to early February 2020) and Europe (from May to August 2020). We remark that the regions in Section 2 refer to provinces or municipalities for the data in China and refer to countries for the data in Europe. First, same as in Section 3, we show the predicted trajectories of the proposed model (both deterministic and stochastic ones) and the baseline models (only deterministic ones) for regions in China and Europe. We remark that regions refer to provinces or municipalities for COVID-19 data in China and countries for COVID-19 data in Europe. This shows that the proposed model can capture and further predict the trend of the true trajectories and behaves better than the baseline models, mainly due to the introduction of heterogeneity of transmission parameters and the employment of correlation between provinces or countries. After that, we show the estimate of the posterior distribution of transmission parameters in a selected province or country using the proposed model. Finally, we evaluate the models quantitatively by computing the training and testing errors as described in Appendix C, which further demonstrates the benefits of the proposed model compared to the baselines.

For the real-world data in Europe, since there might be no unique proper division, we presented the results from two different partitions, both of which are based on geographical locations of the countries. The results imply that in comparison to the simulation study where the data are generated in the artificially divided regions, the generalization performances for the real-world data of both the partitions are similar and both much better than the baseline models.

4.1 Data Sources

Real-world data studied in this section involve the data in China and Europe.

Real-world Data in China

Data in China include the number of confirmed cases and removed cases (consisting of recovered cases and fatalities) in China's major provinces and municipalities from January 21st to March 28th (2020). These publicly available data are downloaded from the websites of the National Health Commission of the People's Republic of China [28] and Chinese Center for Disease Control and Prevention [29]. The corresponding total population in each province or municipality is from [41].

We also utilize two transportation data sets in China extracted from Baidu Qianxi [42], lasting from January 10th to February 10th in 2020. The first data set, called as migration indexes, reflects the number of people moving out from the provinces. The second data set reflects the percentage of people moving to the corresponding destinations from their starting points. The traveling volume is estimated by combining these two data sets.

Real-world Data in Europe

Data in Europe include the number of confirmed and removed cases (consisting of recovered cases and fatalities) in the following 11 countries from May 1st to August 31st in 2020: Denmark, Finland, Norway, Austria, Germany, Switzerland, Italy, Spain, Belgium, France, and Ireland. The data are downloaded from [43, 13], where the data are collected from European Center for Disease Prevention and Control [30]. The data of total population in each country is obtained from [44]. During the study, we are not able to obtain suitable transportation data between these European countries needed by our model, and thus we take simplification by incorporating only spatial and temporal heterogeneity and no transportation in this experiment, details in Section 4.4.

We remark that the data in China are from the initial outbreak of the epidemic. On the contrary, the data in Europe are from when the epidemic has been ongoing about six months. The choice of the aforementioned time period of interest is to illustrate that the proposed model can predict both the case when the epidemic breaks out from some seed region and the case when the epidemic has already established sizable local spread within different regions.

4.2 Model Extension by Allowing Time-Varying Parameters

To capture the trend that the transmission parameters might change in real-world data due to travel restrictions or other factors, we first extend the model and parameter estimation procedure described in Section 2 so that $\{\lambda_k\}$ are allowed to be time-varying.

Specifically, suppose that λ_k is a piece-wise constant function of t, and $\lambda_k = \lambda_k^{(1)}$ for days before some threshold T_k, and $\lambda_k = \lambda_k^{(2)}$ afterwards.

We denote $\lambda^{(\mathbf{1})} = (\lambda_1^{(1)}, \ldots, \lambda_n^{(1)})^T$ and $\lambda^{(\mathbf{2})} = (\lambda_1^{(2)}, \ldots, \lambda_n^{(2)})^T$ as the vectors of transmission rates in the two periods respectively. The stochastic dynamic model introduced in Section 2.2 remains unchanged except that the transmission rates will vary as time increases.

As for the parameter estimation, the first step of obtaining $\gamma_k^* = \arg\max_{\gamma_k} \prod_{i=1}^T \mathrm{Pois}\big((\Delta R_a)_k(i) \big| ((C_a)_k(i-1) - (R_a)_k(i-1))\gamma_k\big)$ also stays the same as described in Section 2.3.

Then, when estimating $\overline{\Theta} = \{E_k(0), H_k(0), \lambda_k^{(1)}, \lambda_k^{(2)}\}_{k=1}^n$ and the marginal posterior distributions of $\overline{\Theta}$, we modify the prior distribution $p(\overline{\Theta})$ as below,

$$p(\overline{\Theta}) = \frac{1}{C_{\sigma,A}} \exp(-\lambda^{(\mathbf{1})^T}(D-A)\lambda^{(\mathbf{1})} - \sigma\|\lambda^{(\mathbf{1})}\|_2^2 - \lambda^{(\mathbf{2})^T}(D-A)\lambda^{(\mathbf{2})} - \sigma\|\lambda^{(\mathbf{2})}\|_2^2), \tag{4.1}$$

where σ, D and A have the same meaning as in (2.8), and $C_{\sigma,A} = \int_{\mathbb{R}^{2n}} \exp(-\lambda^{(\mathbf{1})^T}(D-A)\lambda^{(\mathbf{1})} - \sigma\|\lambda^{(\mathbf{1})}\|_2^2 - \lambda^{(\mathbf{2})^T}(D-A)\lambda^{(\mathbf{2})} - \sigma\|\lambda^{(\mathbf{2})}\|_2^2) d\lambda^{(\mathbf{1})} d\lambda^{(\mathbf{2})}$ is still the normalizing constant. Consequently,

$$\begin{aligned}
V(\overline{\Theta}) &= -\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i}\middle|\overline{\Theta}\right)\right) - \log(p(\overline{\Theta})) \\
&= -\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i}\middle|\overline{\Theta}\right)\right) + \lambda^{(\mathbf{1})^T}(D-A)\lambda^{(\mathbf{1})} + \sigma\|\lambda^{(\mathbf{1})}\|_2^2 + \lambda^{(\mathbf{2})^T}(D-A)\lambda^{(\mathbf{2})} \\
&\quad + \sigma\|\lambda^{(\mathbf{2})}\|_2^2 + \log C_{\sigma,A} \\
&= -\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i}\middle|\overline{\Theta}\right)\right) + \frac{1}{2}\sum_{i,j} a_{ij}(\lambda_i^{(1)} - \lambda_j^{(1)})^2 + \sigma\|\lambda^{(\mathbf{1})}\|_2^2 + \frac{1}{2}\sum_{i,j} a_{ij}(\lambda_i^{(2)} - \lambda_j^{(2)})^2 \\
&\quad + \sigma\|\lambda^{(\mathbf{2})}\|_2^2 + \log C_{\sigma,A}. \qquad (4.2)
\end{aligned}$$

Therefore, by the definition of a_{ij} in (2.10), the estimation of $\overline{\Theta}$ by minimizing $V(\overline{\Theta})$ can be equivalently written as

$$\begin{aligned}
\overline{\Theta}^* &= \arg\min V(\overline{\Theta}) \\
&= \arg\min\Bigg(-\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i}\middle|\overline{\Theta}\right)\right) + \frac{1}{2}\sum_{i,j} a_{ij}(\lambda_i^{(1)} - \lambda_j^{(1)})^2 + \sigma\|\lambda^{(\mathbf{1})}\|_2^2 \\
&\quad + \frac{1}{2}\sum_{i,j} a_{ij}(\lambda_i^{(2)} - \lambda_j^{(2)})^2 + \sigma\|\lambda^{(\mathbf{2})}\|_2^2\Bigg) \\
&= \arg\min\Bigg(-\log\left(p\left(\{(\Delta C_a)_k(i)\}_{k,i}\middle|\overline{\Theta}\right)\right) + \frac{\mu}{2}\Bigg(\sum_m \sum_{i,j\in D_m} W_{ij}\left((\lambda_i^{(1)} - \lambda_j^{(1)})^2 + (\lambda_i^{(2)} - \lambda_j^{(2)})^2\right) \\
&\quad + \beta \sum_{m_1<m_2} \sum_{i\in D_{m_1}, j\in D_{m_2}} W_{ij}\left((\lambda_i^{(1)} - \lambda_j^{(1)})^2 + (\lambda_i^{(2)} - \lambda_j^{(2)})^2\right)\Bigg) + \sigma\sum_i((\lambda_i^{(1)})^2 + (\lambda_i^{(2)})^2)\Bigg). \qquad (4.3)
\end{aligned}$$

After choosing W and σ, $\overline{\Theta}$ can be inferred through (4.3). Then, we could estimate the marginal posterior distributions of $\overline{\Theta}$ by MCMC sampling with the initial point $\overline{\Theta}^*$.

We would like to remark that same as in the simulation study, since the proposed model depends on the penalty factor μ as shown in (4.3), μ is chosen so that the corresponding validation error achieves or is slighted higher than the minimum. Specifically, results from the multiple choices of μ are present, since for the real-world data, the validation data may not have the same distribution as the testing data (the choice of the training, validation, and testing sets is detailed in Appendix B and Sections 4.3.1 and 4.4.1).

4.3 Results for COVID-19 Data in China

4.3.1 Data Description

After the selection process described in Appendix D, $n = 21$ provinces or municipalities are taken into consideration, which are Anhui, Beijing, Fujian, Gansu, Guangdong, Guangxi, Hebei, Henan, Hubei, Hunan, Jiangsu, Jiangxi, Liaoning, Ningxia, Shandong, Shan-Xi, Shanxi, Shanghai, Sichuan, Zhejiang, and Chongqing.

We set $T = 31$, from January 10th to February 10th 2020. However, we remark that only the data after January 21st are available. For measuring the fitting and generalization performance, we set the threshold $T_{th} = 27$ (February 5st). The data from $T_{th,1} = 11$ (January 21st) to the T_{th}-th day (February

5st) are treated as the training set, and the data from $T_{th}+1$ (February 6th) to the t-th day (February 10th) are treated as the testing set. Furthermore, for cross-validation of parameter μ of the proposed model, we choose another threshold $T_{th,2}=17$ (January 27th) and another ending data $T_2=20$ (January 30th). The parameters are first fitted on the data from $T_{th,1}$ (January 21st) to $T_{th,2}$ (January 27th), and then the validation errors are computed on the data from $T_{th,2}$ (January 27th) to T_2 (January 30th). μ is chosen according to validation errors.

In addition, from observation of data from January 10th to February 10th, the transmission rates in the selected provinces or municipalities change after some specific time point. Thus, we allow $\{\lambda_k\}$ to be time-varying for this data set with appropriate changes to the model described in Section 2, which are detailed in Section 4.2. Recall that T_k denotes the day when λ_k changes in the k-th region. For real-world data in China, noting that multiple containment measures, including domestic travel restrictions, contact tracing and epidemiological investigation, and even city lock down were implemented since late January/ early February, we choose $T_k=24$ (February 2nd) for Anhui, Beijing, Henan and Hubei, and $T_k=20$ (January 29th) for the other provinces from the initial observation of the data.

Furthermore, the data from Baidu Qianxi might not be the exact traveling volumes between municipalities and provinces. We assume that the actual traveling volume from one starting point to one destination is proportional to Baidu migration index (which reflects numbers of people departed from the stating point) and the percentage of population traveling from this origin to the destination. The corresponding scaling parameter α also needs to be inferred from the data for all the models.

4.3.2 Models to Compare

As in Section 3.2.2, the same Models 1-5 are compared for COVID-19 data in China. Recall that Model 5 is the model proposed in this paper and the other four are the baseline models for comparison.

Note that for Model 5 with prior distribution based on graph Laplacian, the provinces are chosen to be divided into two groups, which are Hubei and other provinces except Hubei, due to the overwhelmed medical resources in Wuhan at the early stage of the pandemic [45]. For Model 5, by (4.3), the time-varying extension of (2.11) in Section 4.2, $\overline{\Theta}=\{E_k(0),H_k(0),\lambda_k^{(1)},\lambda_k^{(2)}\}_{k=1}^n$ and scaling parameter α are estimated by the optimization problem in (4.4):

$$\begin{aligned}\overline{\Theta}^*,\alpha^* = \arg\min\Bigg(&-\log p(y_{1:T_{th}}|\overline{\Theta},\alpha)+\frac{\mu}{2}\Bigg(\sum_{m=1}^{2}W_{ij}\left((\lambda_i^{(1)}-\lambda_j^{(1)})^2+(\lambda_i^{(2)}-\lambda_j^{(2)})^2\right)\\ &+\beta\sum_{i\in D_1,j\in D_2}W_{ij}\left((\lambda_i^{(1)}-\lambda_j^{(1)})^2+(\lambda_i^{(2)}-\lambda_j^{(2)})^2\right)\Bigg)+\sigma\sum_i((\lambda_i^{(1)})^2+(\lambda_i^{(2)})^2)\Bigg),\end{aligned} \tag{4.4}$$

where $\beta=0.1$, and $\lambda_k^{(1)}$ and $\lambda_k^{(2)}$ are the transmission rates in the k-th province before and after the T_k-th day respectively, and σ is chosen to be 10^{-6}. $W_{ij}=\bar{W}_{ij}/\max_{i,j}\bar{W}_{ij}$, where $\bar{W}_{ij}=\frac{1}{2}(\frac{1}{T}\sum_{l=1}^{T}(w_l)_{ij}+(\frac{1}{T}\sum_{l=1}^{T}(w_l)_{ji})$ is the averaged traveling volume between the i-th and j-th province with rescaling, treated as the proximity between the two provinces.

For Models 1-4, if the model does not have heterogeneity of transmission parameters, then $\{\lambda_1^{(1)},\ldots,\lambda_n^{(1)}\}$ are forced to be the same and so do $\{\lambda_1^{(2)},\ldots,\lambda_n^{(2)}\}$, while the transmission rates in the two periods are allowed to be different; if the model does not allow transportation between provinces, then terms involving W_t disappear in (2.4) as described in Section 3.2.2.

4.3.3 Results of Trajectory Prediction

We first remark that for Model 5, three values of μ are chosen, and part of the results below are present from these three choices to perform the sensitivity analysis of μ. The first value is $\mu = 10^{2.4}$, which minimizes the relative validation errors ($\text{MAE}_{(w)}^{[\text{Val}]}$, $\text{MSE}_{(w)}^{[\text{Val}]}$, $\text{MAE}_{(s)}^{[\text{Val}]}$, and $\text{MSE}_{(s)}^{[\text{Val}]}$), marked with black pentagrams in Figure 9 below and Figure A5. The other two values are $\mu = 10^2$ and $\mu = 10^{2.7}$, whose validation errors are slightly larger, marked with red and green pentagrams in Figure 9 and Figure A5 respectively. Since for the real-world data in China, the validation error does not necessarily have the same distribution as the testing error, some of the results from various choices of μ are presented for better comparison.

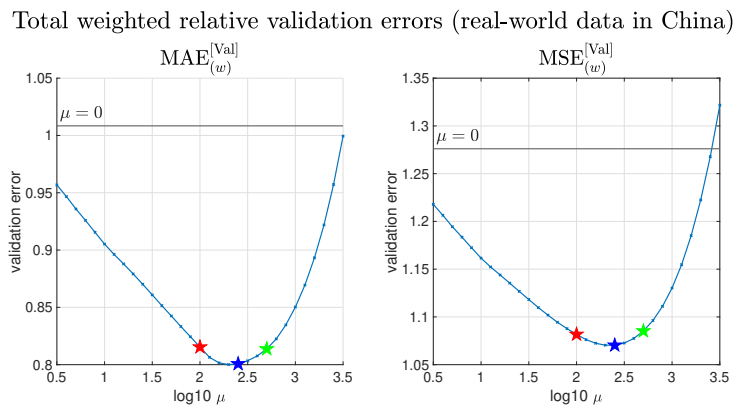

Figure 9: Weighted validation errors $\text{MAE}_{(w)}^{[\text{Val}]}$ and $\text{MSE}_{(w)}^{[\text{Val}]}$ against μ (computed by (A5)) for real-world data in China. The horizontal lines show the values of the validation errors when $\mu = 0$.

Figure 10 shows the true and predicted trajectories for newly confirmed cases in Hubei. In Figure 10,

- The red line shows the true trajectory.
- The black lines show the predicted deterministic trajectories obtained by running (2.4) with Θ^* inferred using (4.4) with $10^2, 10^{2.4}, 10^{2.7}$ respectively.
- The black scatter plots show 100 stochastic trajectories with Θ^* inferred using (4.4) with $10^{2.4}$
- The red scatter plots show 100 stochastic trajectories sampled from the posterior distribution of $\overline{\Theta}$, which is estimated by 5×10^5 MCMC iterations staring from $\overline{\Theta}^*$ inferred using (4.4) with $10^{2.4}$

The deterministic trajectories obtained by the other four models and the corresponding absolute errors are also plotted for better comparison. It can be seen that all the predicted deterministic trajectories generally capture the trend of the true trajectory. Moreover, sampling trajectories from the posterior distribution of $\overline{\Theta}$ generates more randomness than sampling trajectories with Θ^*. Besides, it can be seen from Figure 10 that the performances of all the five models (including Model 5 with the three choices of μ) are similar for Hubei. This may be explained by the fact that the cases in Hubei outnumber those in other provinces or municipalities, which makes all the models tend to fit the trajectory of Hubei best.

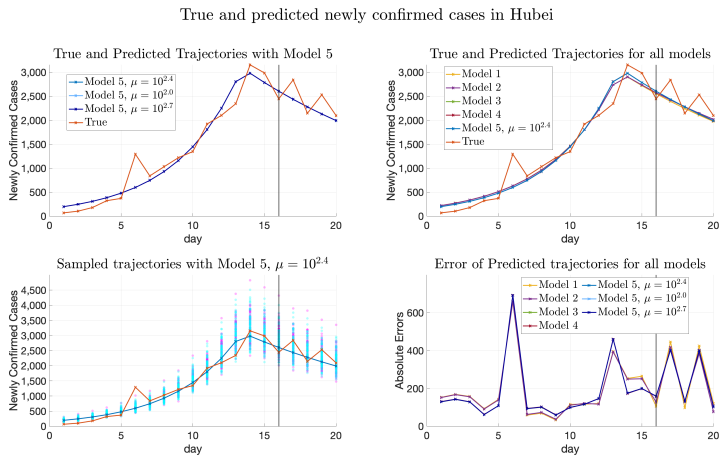

Figure 10: True and fitted trajectories in Hubei. The red line shows the true trajectory, the black lines show the predicted deterministic trajectories with Θ^* ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{2.4}, 10^2, 10^{2.7}$ respectively), black and red scatter plots show 100 stochastic trajectories with Θ^* and sampled from the posterior distribution of $\overline{\Theta}$ respectively ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{2.4}$). The vertical lines show the threshold $T_{th} = 27$ for the training data and testing data.

Similarly, Figure 11 shows the predicted trajectories in Henan. As can be seen in Figure 11, unlike Hubei, heterogeneity of parameters helps the model fit and predict the trajectory of Henan in a more significant manner. When the transmission parameters in the two periods are forced to be the same in all provinces, the increment of the newly confirmed cases of the predicted trajectory is faster than the trend shown in the true trajectory in the first period. Meanwhile, the decrement of the newly confirmed cases of the fitted deterministic trajectory in the second period is also slower than the trend reflected in the true trajectory. Thus, heterogeneity helps improve the performance of fitting and predicting. Additionally, although $\mu = 10^{2.4}$ minimizes the validation errors, it can be seen from Figure 11 that the deterministic trajectory obtained by Model 5 with $\mu = 10^2$ achieves better performance in prediction than the one obtained by Model 5 with $\mu = 10^{2.4}$. Therefore, the result suggests that the choice of μ is not necessarily limited to the minimizer of the validation error. Instead, we may also compare trajectories with μ that have slightly larger validation error for possibly better generalization performance.

Moreover, from Figure 12 which shows the fitted trajectories in Anhui, it can still be seen that heterogeneity helps the models predict the testing trajectory better. Furthermore, introducing correlation helps Model 5 (with $\mu = 10^{2.4}$) predict the trajectory marginally better than the other models.

4.3.4 Results of Parameter Estimation

To illustrate the estimate of the posterior of parameters, Figure 13 shows the estimated posterior distributions of $\lambda_k^{(1)}$ and $\lambda_k^{(2)}$ in Hubei using Model 5, in which the red lines show the inferred $\lambda_k^{(1)}$ and $\lambda_k^{(2)}$ with $\mu_1 = 10^{-3.5}$, and Table 7 shows the inferred $\lambda_k^{(1)}$ and $\lambda_k^{(2)}$ with the mean and standard deviation of their estimated posterior distributions from MCMC.

By comparing the transmission rates in Hubei in the two periods, it can be seen that the transmission rate decreases greatly after February 2nd, which indicates that the containment measures adopted in Hubei against the COVID-19 outbreak are effective.

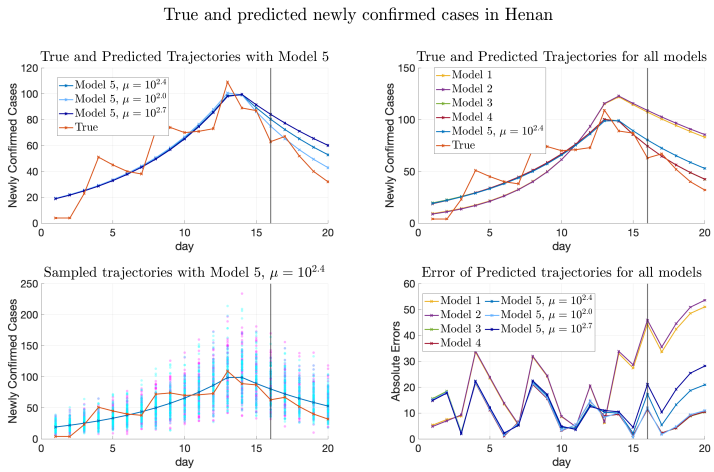

Figure 11: True and fitted trajectories in Henan. The red line shows the true trajectory, the black lines show the predicted deterministic trajectories with Θ^* ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{2.4}, 10^2, 10^{2.7}$ respectively), black and red scatter plots show 100 stochastic trajectories with Θ^* and sampled from the posterior distribution of $\overline{\Theta}$ respectively ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{2.4}$). The vertical lines show the threshold $T_{th} = 27$ for the training data and testing data.

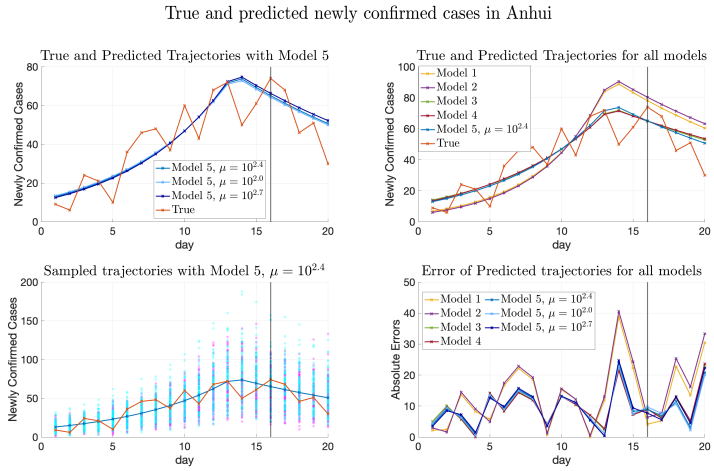

Figure 12: True and fitted trajectories in Anhui. The red line shows the true trajectory, the black lines show the predicted deterministic trajectories with Θ^* ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{2.4}, 10^2, 10^{2.7}$ respectively), black and red scatter plots show 100 stochastic trajectories with Θ^* and sampled from the posterior distribution of $\overline{\Theta}$ respectively ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{2.4}$). The vertical lines show the threshold $T_{th} = 27$ for the training data and testing data.

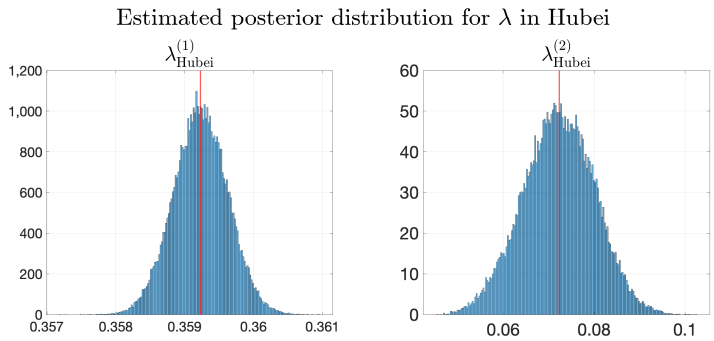

Figure 13: Estimated posterior distribution of λ in Hubei with $\mu_1 = 10^{2.4}$. The vertical red lines represent the values of corresponding θ^*.

	θ^*	mean of estimated posterior	standard deviation of estimated posterior
$\lambda^{(1)}_{\text{Hubei}}$	0.3592	0.3592	3.932×10^{-4}
$\lambda^{(2)}_{\text{Hubei}}$	0.0724	0.0722	0.0078

Table 7: Estimated paramters in Hubei. θ^* is inferred using (4.4) with $\mu = 10^{2.4}$, which minimizes validation errors as shown in Figures 9 and A5. The posterior distribution is estimated by 5×10^5 MCMC iterations.

4.3.5 Further Model Evaluation

Table 8 shows the training errors and testing errors of the five models computed as detailed in Appendix C. As can be seen from Table 8, both heterogeneity of parameters and transportation between provinces help reduce the training errors and the testing errors as the simulated data case. The utilization of graph Laplacian prior helps further reduce the testing errors. Note that the testing errors of Model 4 are close to those of Model 3, which might be explained as after the traveling restrictions are imposed on January 23rd [46], transportation no longer has much influence on the spread of the epidemic.

Model	M	H	GL prior	$\text{MAE}^{[\text{Tr}]}_{(w)}$	$\text{MAE}^{[\text{Te}]}_{(w)}$	$\text{MSE}^{[\text{Tr}]}_{(w)}$	$\text{MSE}^{[\text{Te}]}_{(w)}$
1	×	×	×	0.313	0.473	0.580	0.600
2	✓	×	×	0.311	0.465	0.580	0.593
3	×	✓	×	0.267	0.442	0.526	0.565
4	✓	✓	×	0.266	0.440	0.523	0.562
	✓	✓	✓ ($\mu = 10^{2.4}$)	0.270	0.424	0.538	0.536
5	✓	✓	✓ ($\mu = 10^{2}$)	0.267	0.423	0.531	0.535
	✓	✓	✓ ($\mu = 10^{2.7}$)	0.274	0.427	0.546	0.540

Table 8: Training and testing error for real-world data in China of five models. The formulas of errors are detailed in Appendix C (equation (A5) or its modifications to compute testing errors). The remarks for Columns 1-4 are the same as in Table those in Table 4. In addition, for parameter inference with Model 5 as defined in (4.4), as shown in Figures 9 and A5, $\mu = 10^{2.4}$ achieves the minimal validation errors, and 10^2 and $10^{2.7}$ correspond to slightly greater validation errors.

Figures 14 and A6 plot weighted and simply averaged testing errors against varying μ respectively. We can see from both Table 8 and Figure 14 that though $\mu = 10^{2.4}$ achieves the minimal validation error, it has slightly larger testing error than $\mu = 10^2$.

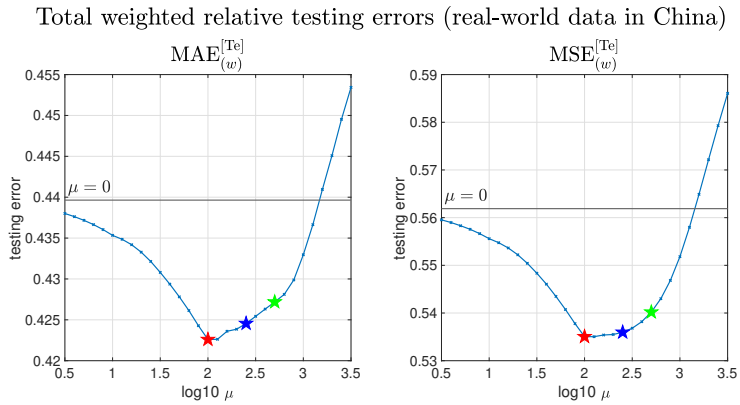

Figure 14: Weighted testing errors $\text{MAE}^{[\text{Te}]}_{(w)}$ and $\text{MSE}^{[\text{Te}]}_{(w)}$ against μ (computed by (A5)) for real-world data in China. The horizontal lines show the values of the testing errors when $\mu = 0$.

4.4 Results for COVID-19 Data in Europe

4.4.1 Data Description

11 countries in Europe are considered in the experiments, which are Denmark, Finland, Norway, Austria, Germany, Switzerland, Italy, Spain, Belgium, France, and Ireland. Some countries such as Iceland and Luxembourg are excluded since their total population is too small, and some other countries like Sweden, Holland, and England are ignored due to that the removed cases in these countries are zero (from the data sources [43]) and thus the two-step procedure for parameter estimation in this paper may not be applied.

Since the original data oscillate much, especially over the weekends, the data are first smoothed by

taking a simple moving average (which treats the unweighted average of data in 7 days around the i-th day as the new data point for the i-th day). After the smoothing, $T = 116$.

Additionally, from the observation of newly confirmed cases from May 1st to August 1st, the transmission rates change after some threshold in the selected countries in Europe. Thus, we still allow $\{\lambda_k\}$ to be time-varying for this data set as described in Section 4.2. $T_k = 50$ is chosen to be the day when λ_k changes for all countries except Norway, Italy and France. In addition, $T_{Nowway} = 75$, $T_{Italy} = 85$ and $T_{France} = 60$. Furthermore, for evaluating fitting and generalization performance, we set $T_{th} = 95$. For choosing μ (the details can be found in Appendix B.2), we set $T_{th,2} = 18$ and $T_2 = 36$. Namely, the validation errors are computed by comparing the data from $T_{th,2}$-th day to T_2-th day with the predicted trajectories using parameters fitted on the data from the 1st day to $T_{th,2}$-th day. Then, the choice of μ is based on validation errors.

4.4.2 Models to Compare

For COVID-19 data in Europe, only three models detailed below are compared, which allow heterogeneity of parameters but do not include transportation. The last model is the proposed one in this paper, and the first two are baseline models. The transportation data needed in this model are not available to the best of our knowledge. However, the impact of transportation is expected to be less significant due to travel restrictions [46].

1'. Model with uniform prior distribution, without heterogeneity or migration.

2'. Model with uniform prior distribution, with heterogeneity but without migration.

3'. Model with prior distribution based on graph Laplacian, with heterogeneity but no migration.

For Model 3', 11 countries are partitioned into $d = 4$ groups according to geographical locations:

- D_1 (Northern Europe): Denmark, Finland, Norway;
- D_2 (Central Europe): Austria, Germany, Switzerland;
- D_3 (Southern Europe): Italy, Spain;
- D_4 (Western Europe): Belgium, France, Ireland.

We denote this partition as P. Since this might not be the unique appropriate partition for the $n = 11$ countries, we also present results from another partition denoted as P' in the following sections. P' also groups the countries that are geographically close together:

- D'_1: Finland, Norway;
- D'_2: Denmark, Austria, Germany;
- D'_3: Spain, Belgium, France, Ireland;
- D'_4: Switzerland, Italy.

The results for Model 3' in Section 4.4.5 are presented with both the partitions P and P'.

Besides, by the same time-varying extension (4.3) detailed in Section 4.2, for Model 3', $\overline{\Theta} = \{E_k(0), H_k(0), \lambda_k^{(1)}, \lambda_k^{(2)}\}_{k=1}^n$ are estimated by the optimization problem in (4.5):

$$\overline{\Theta}^* = \arg\min \Bigg(-\log p(y_{1:T_{th}}|\overline{\Theta}) + \frac{\mu}{2}\Bigg(\sum_m \sum_{i,j\in D_m} \Big((\lambda_i^{(1)} - \lambda_j^{(1)})^2 + (\lambda_i^{(2)} - \lambda_j^{(2)})^2 \Big) + \beta \sum_{m_1<m_2} \sum_{i\in D_{m_1}, j\in D_{m_2}} \Big((\lambda_i^{(1)} - \lambda_j^{(1)})^2 + (\lambda_i^{(2)} - \lambda_j^{(2)})^2 \Big) \Bigg) + \sigma \sum_i \Big((\lambda_i^{(1)})^2 + (\lambda_i^{(2)})^2 \Big) \Bigg), \quad (4.5)$$

or by (4.5) with the partition $P = \{D_1, D_2, D_3, D_4\}$ replaced by another partition $P' = \{D_1', D_2', D_3', D_4'\}$. Here, σ is still taken to be 10^{-6}, and β is still 0.1.

4.4.3 Results of Trajectory Prediction

We still remark that for Model 3' (with P), as in the case of real-world data in China, part of the results reported are from three choices of μ for careful consideration, since the validation error may not have the same distribution as the testing error. Note that Figures 15 and A7 plot the weighted and simply averaged relative validation errors, $\text{MAE}_{(w)}^{[\text{Val}]}$, $\text{MSE}_{(w)}^{[\text{Val}]}$, $\text{MAE}_{(s)}^{[\text{Val}]}$ and $\text{MSE}_{(s)}^{[\text{Val}]}$ respectively. One choice is $\mu = 10^{3.9}$, which minimizes the validation errors (computed with P), marked with black pentagrams in Figures 15 and A7. The other two choices are $\mu = 10^{3.2}$ and $\mu = 10^5$, which correspond to slightly larger validation errors and are marked in red and green pentagrams respectively in Figures 15 and A7.

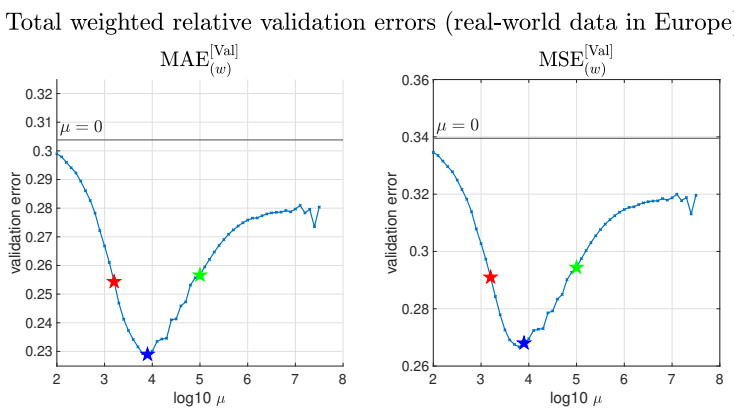

Figure 15: Weighted validation errors $\text{MAE}_{(w)}^{[\text{Val}]}$ and $\text{MSE}_{(w)}^{[\text{Val}]}$ against μ (computed by (A5)) for real-world data in Europe. The horizontal lines show the values of the validation errors when $\mu = 0$.

Figure 16, 17 and 18 present the true and predicted trajectories in Austria, Germany, and Italy, respectively. As can be seen from Figure 16, 17 and 18, heterogeneity of transmission parameters in Model 2' (green lines) and Model 3' (black lines) helps improve the performance of fitting and generalization. Furthermore, utilization of correlation between countries in Model 3' further improves the prediction of the trajectories. We also note that for Model 3', compared to the results for $\mu = 10^{3.9}$, which minimizes the validation errors, $\mu = 10^5$ predicts the trajectories of Austria and Italy more accurately. Hence, the results for Model 3' are sensitive to μ, and it might be better to compare the trajectories from multiple choices of μ.

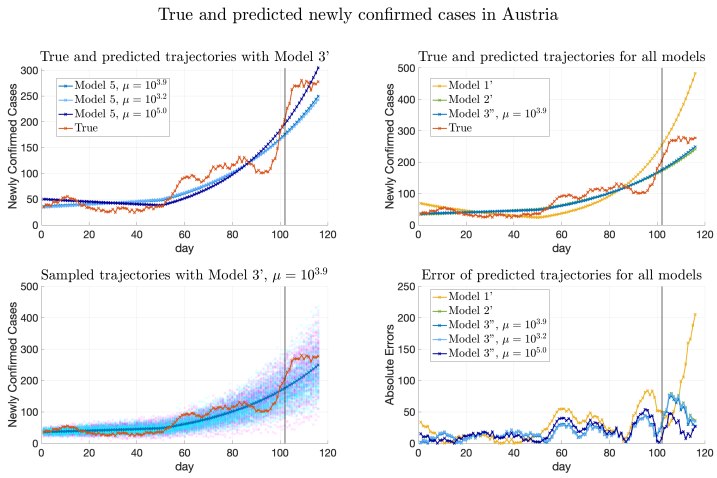

Figure 16: True and fitted trajectories in Austria. The red line shows the true trajectory, the black lines show the predicted deterministic trajectories with Θ^* ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{3.9}, 10^{3.2}, 10^5$ respectively), black and red scatter plots show 100 stochastic trajectories with Θ^* and sampled from the posterior distribution of $\overline{\Theta}$ respectively ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{3.9}$). The vertical lines show the threshold $T_{th} = 102$ for the training data and testing data.

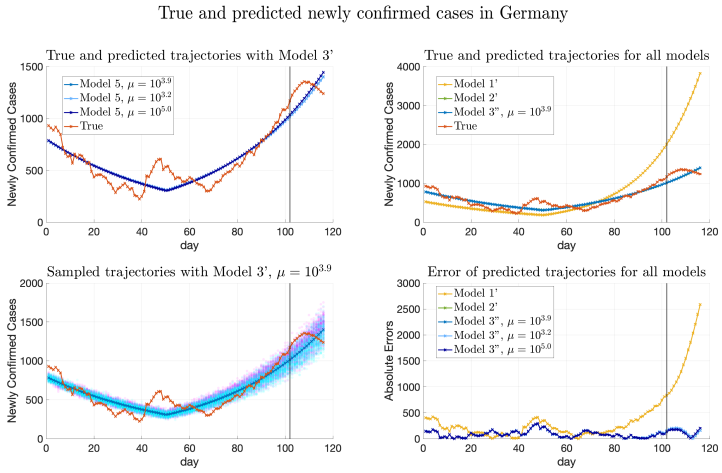

Figure 17: True and fitted trajectories in Germany. The red line shows the true trajectory, the black lines show the predicted deterministic trajectories with Θ^* ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{3.9}, 10^{3.2}, 10^5$ respectively), black and red scatter plots show 100 stochastic trajectories with Θ^* and sampled from the posterior distribution of $\overline{\Theta}$ respectively ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{3.9}$). The vertical lines show the threshold $T_{th} = 102$ for the training data and testing data.

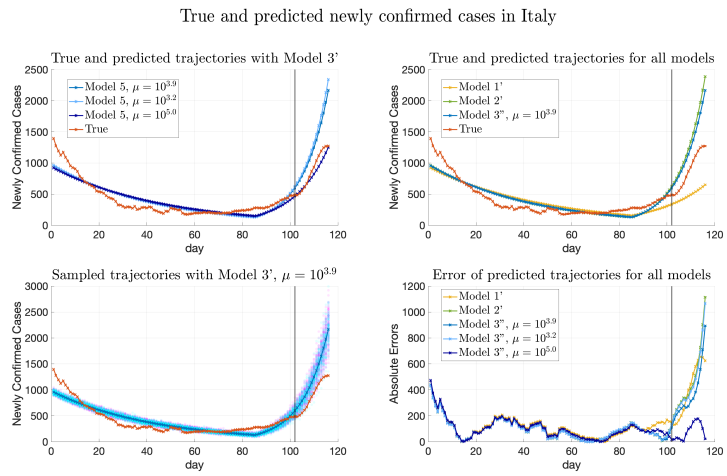

Figure 18: True and fitted trajectories in Italy. The red line shows the true trajectory, the black lines show the predicted deterministic trajectories with Θ^* ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{3.9}, 10^{3.2}, 10^5$ respectively), black and red scatter plots show 100 stochastic trajectories with Θ^* and sampled from the posterior distribution of $\overline{\Theta}$ respectively ($\overline{\Theta}^*$ inferred using (4.4) with $\mu = 10^{3.9}$). The vertical lines show the threshold $T_{th} = 102$ for the training data and testing data.

4.4.4 Results of Parameter Estimation

Same as before, Figure 19 shows the estimated posterior distributions of $\lambda^{(1)}_{\text{Italy}}$ and $\lambda^{(2)}_{\text{Italy}}$ using Model 3', in which the red lines show the inferred $\lambda^{(1)}_{\text{Italy}}$ and $\lambda^{(2)}_{\text{Italy}}$ with $\mu = 10^{3.9}$. Table 9 shows the inferred $\lambda^{(1)}_{\text{Italy}}$ and $\lambda^{(2)}_{\text{Italy}}$ and the mean and standard deviation of their estimated posterior distribution from MCMC. The estimated $\lambda^{(2)}_{\text{Italy}}$ is larger than the estimated $\lambda^{(1)}_{\text{Italy}}$, which is consistent with the trend that the newly confirmed cases in Italy first decrease and then increase.

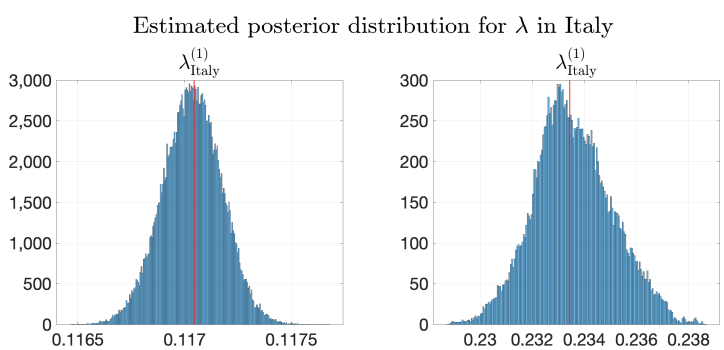

Figure 19: Estimated posterior distributions of λ in Italy with $\mu = 10^{3.9}$. The vertical red lines represent the values of corresponding θ^*.

	θ^*	mean of estimated posterior	standard deviation of estimated posterior
$\lambda^{(1)}_{\text{Italy}}$	0.1170	0.1170	1.3794×10^{-4}
$\lambda^{(2)}_{\text{Italy}}$	0.2336	0.2334	0.0015

Table 9: Estimated paramters in Italy. θ^* is inferred using (4.5) with $\mu = 10^{3.9}$, which minimizes validation errors as shown in Figures 15 and A7. The posterior distribution is estimated by 5×10^5 MCMC iterations.

4.4.5 Further Model Evaluation

Table 10 shows training and testing errors of the three models. It can be seen from Table 10 that Models 2' and 3' have smaller both training and testing errors than Model 1', due to the introduction of heterogeneity of transmission parameters. We can also see that compared with Model 2', Model 3' that adds Graph Laplacian prior has better generalization performance. Especially, Model 3' with $\mu = 10^5$ greatly reduces the testing errors, although it does not have the minimal validation errors.

Model	M	H	GL prior	$\text{MAE}^{[\text{Tr}]}_{(w)}$	$\text{MAE}^{[\text{Te}]}_{(w)}$	$\text{MSE}^{[\text{Tr}]}_{(w)}$	$\text{MSE}^{[\text{Te}]}_{(w)}$
1'	×	×	×	0.362	0.716	0.460	0.775
2'	×	✓	×	0.180	0.531	0.240	0.601
	×	✓	✓ ($\mu = 10^{3.9}$)	0.181	0.462	0.243	0.523
3' (P)	×	✓	✓ ($\mu = 10^{3.2}$)	0.180	0.508	0.240	0.574
	×	✓	✓ ($\mu = 10^{5}$)	0.212	0.327	0.282	0.380
	×	✓	✓ ($\mu = 10^{3.9}$)	0.181	0.462	0.243	0.523
3' (P')	×	✓	✓ ($\mu = 10^{3.2}$)	0.180	0.510	0.240	0.577
	×	✓	✓ ($\mu = 10^{5}$)	0.214	0.322	0.283	0.373

Table 10: Training and testing error for real-world data in Europe of three models. The formulas of errors are detailed in Appendix C (equation (A5) or its modifications to compute testing errors). The remarks for Columns 2-4 are the same as in Table those in Table 4. In addition, for parameter inference with Model 3' as defined in (4.5), as shown in Figures 15 and A7, $\mu = 10^{3.9}$ achieves the minimal validation errors, and $10^{3.2}$ and 10^5 correspond to slightly greater validation errors. The partitions in Model 3' are P and P' respectively. P and P' are two different partitions of countries in Europe introduced in Section 4.4.2.

Recall that P and P' are two different partitions of countries in Europe introduced in Section 4.4.2, both based on geographical locations. By comparing the corresponding results in Table 10 for Model 3' with partitions P and P' respectively, it can be seen that the testing errors are almost the same. This is different from the implications of the simulated data. A possible explanation for this might be that the regions in the real world have more complicated correlations than the simulated regions where clustering information is uniquely and artificially prefixed. Although reasonable groupings may not always be unique for real-world cases, the proposed model could still predict the trajectories more accurately than the baseline models.

Figure 20 below and Figure A8 plot the weighted and simply averaged testing errors (whose definitions are in Appendix C) against varying μ respectively, from which one may see that by imposing regularization properly through choosing a moderate μ helps improve the generalization performance. By comparing

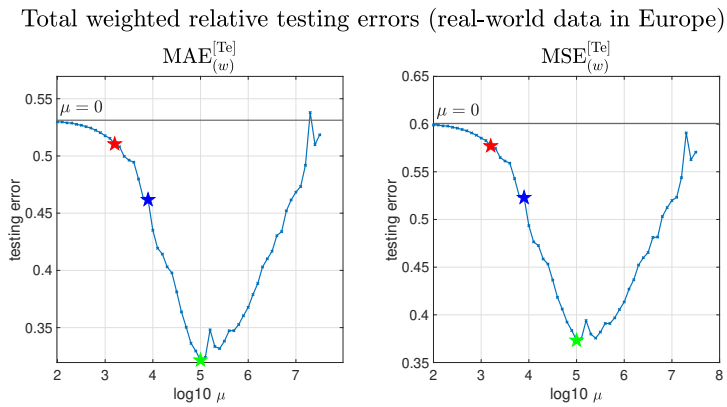

Figure 20: Weighted testing errors MAE$^{[Te]}_{(w)}$ and MSE$^{[Te]}_{(w)}$ against μ (computed by (A5)) for real-world data in Europe. The horizontal lines show the values of the testing errors when $\mu = 0$.

Figure 20 (and Figure A8) and Figure 15 (and Figure A7), we see that the change in validation and testing errors concerning increasing μ are not the same. This further verifies the indications from the previous findings that it would be better to examine and compare the results from multiple choices of μ with relatively low validation errors for Model 3' to choose a better μ for the specific data sets.

5 Discussion

In this paper, we propose a stochastic dynamic model inspired by [1], considering the inter-district transportation and both spatially and temporally heterogeneous transmission parameters, which can model the ongoing and lasting spread of the epidemic in multiple districts. Based on this model, we also introduce a two-step procedure for estimating the parameters, which maximizes over posterior distribution with a prior distribution that utilizes graph Laplacian regularization to address the correlation between districts. Experiments on both simulated and real-world data show that compared with the baseline models, this new model with transportation, heterogeneity of parameters, and usage of the correlation in parameter inference improves the performance of fitting and generalization.

We acknowledge that there are limitations to this work. First, the stochastic dynamic model proposed in this paper might not be comprehensive enough to characterize the spread of COVID-19 in reality. For example, the model does not account for the ascertainment rate of the positive cases, the asymptomatic virus carriers, and the infectious latent period. Second, the proposed model does not consider the heterogeneity with respect to transmission risk in different populations, for example, populations with different ages, jobs, or health conditions, and needs further modification to incorporate the large-scale COVID-19 vaccination as well as waning of immunity over time. Third, the parameter inference depends on the correlation between regions determined by the clustering patterns among regions. However, such division is usually not fully known in the real-world cases, and the inference of graph structure is yet to be explored in this paper. Finally, if more complete and accurate transportation data are available, then the proposed model might predict the trend of the epidemic more accurately.

The current methods have possible extensions in the following several directions in our future study. First, the dynamic model is to be refined to be closer to reality, for example, taking the asymptomatic virus carriers and contact tracking into account and modeling the change of the transmission parameters over time in a more sophisticated way. Second, the model can be modified to accommodate the hetero-

geneity of various populations, including different age groups and vaccination statuses. For example, the compartments are to be further divided into subgroups, and the parameters such as transmission rates and mortality rates are allowed to be different; and the dynamic model is also to be adjusted accordingly. Third, the method could further include detecting the underlying graph structure of the regions so that the construction of the graph Laplacian matrix could be more self-contained and systematic. At last, the method could be further used to assess the influence of containing measures taken by different countries, for example, by adjusting the traveling volume to make it different from the actual transportation data and then analyzing the corresponding influence on the size and trend of the pandemic.

References

[1] Balcan, D. *et al.* Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of computational science* **1**, 132–145 (2010).

[2] World Health Organization (WHO). `https://covid19.who.int`.

[3] Keeling, M. J. & Rohani, P. *Modeling infectious diseases in humans and animals* (Princeton university press, 2011).

[4] Gomes, M. F. *et al.* Assessing the international spreading risk associated with the 2014 west african ebola outbreak. *PLoS currents* **6** (2014).

[5] Pastore-Piontti, A. *et al.* Real-time assessment of the international spreading risk associated with the 2014 west african ebola outbreak. In *Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases*, 39–56 (Springer, 2016).

[6] Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science* **368**, 395–400 (2020).

[7] Colizza, V. *et al.* Estimate of novel influenza a/h1n1 cases in mexico at the early stage of the pandemic with a spatially structured epidemic model. *PLoS currents* **1**, RRN1129–RRN1129 (2009).

[8] Balcan, D. *et al.* Modeling the critical care demand and antibiotics resources needed during the fall 2009 wave of influenza a(h1n1) pandemic. *PLoS currents* **1**, RRN1133–RRN1133 (2009).

[9] Bajardi, P. *et al.* Modeling vaccination campaigns and the fall/winter 2009 activity of the new a(h1n1) influenza in the northern hemisphere. *Emerging health threats journal* **2**, e11–e11 (2009;2008;2010;).

[10] Tizzoni, M. *et al.* Real-time numerical forecast of global epidemic spreading: Case study of 2009 a/h1n1pdm. *BMC medicine* **10**, 165–165 (2012).

[11] Poletto, C. *et al.* Assessment of the middle east respiratory syndrome coronavirus (mers-cov) epidemic in the middle east and risk of international spread using a novel maximum likelihood analysis approach. *Eurosurveillance* **19**, 20824 (2014).

[12] Balcan, D. *et al.* Seasonal transmission potential and activity peaks of the new influenza a (h1n1): a monte carlo likelihood analysis based on human mobility. *BMC medicine* **7**, 1–12 (2009).

[13] Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases* **20**, 533–534 (2020).

[14] Zhang, Q. *et al.* Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th international conference on world wide web*, 311–319 (2017).

[15] Vespignani, A. Predicting the behavior of techno-social systems. *Science* **325**, 425–428 (2009).

[16] Donnat, C. & Holmes, S. Modeling the heterogeneity in covid-19's reproductive number and its impact on predictive scenarios. *Journal of Applied Statistics* 1–29 (2021).

[17] Tkachenko, A. V. *et al.* Time-dependent heterogeneity leads to transient suppression of the covid-19 epidemic, not herd immunity. *Proceedings of the National Academy of Sciences* **118** (2021).

[18] Moreno, Y., Pastor-Satorras, R. & Vespignani, A. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems* **26**, 521–529 (2002).

[19] Kermack, W. O. & McKendrick, A. G. 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**, 700–721 (1927).

[20] Boschi, T., Di Iorio, J., Testa, L., Cremona, M. A. & Chiaromonte, F. Functional data analysis characterizes the shapes of the first covid-19 epidemic wave in italy. *Scientific reports* **11**, 1–15 (2021).

[21] Carroll, C. *et al.* Time dynamics of covid-19. *Scientific reports* **10**, 1–14 (2020).

[22] Cremona, M. A. & Chiaromonte, F. Probabilistic k-mean with local alignment for clustering and motif discovery in functional data. *arXiv preprint arXiv:1808.04773* (2018).

[23] Hilton, J. & Keeling, M. J. Estimation of country-level basic reproductive ratios for novel coronavirus (sars-cov-2/covid-19) using synthetic contact matrices. *PLoS computational biology* **16**, e1008031 (2020).

[24] Szapudi, I. Heterogeneity in sir epidemics modeling: superspreaders and herd immunity. *Applied Network Science* **5**, 1–12 (2020).

[25] Volpert, V., Banerjee, M. & Sharma, S. Epidemic progression and vaccination in a heterogeneous population. application to the covid-19 epidemic. *Ecological Complexity* 100940 (2021).

[26] Hou, X. *et al.* Intracounty modeling of covid-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race. *Proceedings of the National Academy of Sciences* **118** (2021).

[27] Chen, S., Li, Q., Gao, S., Kang, Y. & Shi, X. State-specific projection of covid-19 infection in the united states and evaluation of three major control measures. *Scientific reports* **10**, 1–9 (2020).

[28] National Health Commission of the People's Republic of China. `http://en.nhc.gov.cn/antivirusfight.html`.

[29] Chinese Center for Disease Control and Prevention. `https://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm`.

[30] European Centre for Disease Prevention and Control. `https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases`.

[31] Tian, H. *et al.* An investigation of transmission control measures during the first 50 days of the covid-19 epidemic in china. *Science* **368**, 638–642 (2020).

[32] Kraemer, M. U. *et al.* The effect of human mobility and control measures on the covid-19 epidemic in china. *Science* **368**, 493–497 (2020).

[33] World Health Organization (WHO). Enhancing response to omicron sars-cov-2 variant. https://www.who.int/publications/m/item/enhancing-readiness-for-omicron-(b.1.1.529)-technical-brief-and-priority-actions-for-member-states.

[34] Zhang, Y. *et al.* Prediction of the covid-19 outbreak based on a realistic stochastic model. *medRxiv* (2020).

[35] Kurtz, T. G. Solutions of ordinary differential equations as limits of pure jump markov processes. *Journal of applied Probability* **7**, 49–58 (1970).

[36] Kurtz, T. G. Limit theorems for sequences of jump markov processes approximating ordinary differential processes. *Journal of Applied Probability* **8**, 344–356 (1971).

[37] He, S., Peng, Y. & Sun, K. Seir modeling of the covid-19 and its dynamics. *Nonlinear dynamics* **101**, 1667–1680 (2020).

[38] López, L. & Rodo, X. A modified seir model to predict the covid-19 outbreak in spain and italy: simulating control scenarios and multi-scale epidemics. *Results in Physics* **21**, 103746 (2021).

[39] Yang, Z. *et al.* Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *Journal of thoracic disease* **12**, 165 (2020).

[40] He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature medicine* **26**, 672–675 (2020).

[41] National Bureau of Statistics of China. Annual data by province (2019). http://www.stats.gov.cn/english/Statisticaldata/AnnualData/.

[42] Baidu Qianxi. https://qianxi.baidu.com/#/2020chunyun.

[43] Johns Hopkins University. Covid-19 data repository by the center for systems science and engineering (csse) at johns hopkins university (2020). https://github.com/CSSEGISandData/COVID-19.

[44] Wikipedia. List of european countries by population. https://en.wikipedia.org/wiki/List_of_European_countries_by_population.

[45] Information Office of Hubei Provincial People's Government. Prevention and control of pneumonia outbreak of new coronary virus infection (2020). https://www.hubei.gov.cn/hbfb/xwfbh/202002/t20200210_2023490.shtml.

[46] Blavatnik School of Government. Covid-19 government response tracker. https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker.

[47] Kendall, D. G. Deterministic and stochastic epidemics in closed populations. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Health*, 149–165 (University of California Press, 1956).

[48] Bailey, N. T. A simple stochastic epidemic. *Biometrika* 193–202 (1950).

[49] Bartlett, M. Some evolutionary stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)* **11**, 211–229 (1949).

[50] Britton, T. *et al. Stochastic epidemic models with inference* (Springer, 2019).

Appendix A Review of SEIR Model and the GLEaM Model

In this section of Appendix, we provide a more detailed introduction to the SEIR Model and the GLEaM Model so as to provide a background of our study and augment the main text.

To depict the evolution of the epidemics, [19] proposed the celebrated Susceptible-Infected-Removed (SIR) model and characterized the development of the pandemic with a deterministic ordinary differential equation (ODE). There are many extensions of the SIR model, including the Susceptible-Exposed-Infected-Removed (SEIR) model for diseases with a latent period, the Susceptible-Infected-Susceptible(SIS) model for diseases that do not gain immunity after recovery, etc.

These deterministic transmission models are constructed certain assumptions, including that the population is large, closed, and homogeneous. Due to the random nature of the transmission process, many stochastic dynamic models are developed [47, 48, 49]. Under certain rather generalized conditions, the deterministic models can be seen as the mean-field equations of the corresponding stochastic processes. However, this approximation may not hold when the size of the outbreak has not grown up to the same order of the total population, which is the case in many applications [50]. More details can be found in [34] and the references therein.

The Global Epidemic and Mobility (GLEaM) model proposed in [1] used a meta-population scheme which balanced between the agent-based stochastic models and the deterministic compartmental models. Specifically, [1] adapted a high-resolution population database that divided the surface of the earth with cells of 15min $\times$ 15min of arc, and then used Voronoi tessellation to assign each cell to one of the major airports around the world. The obtained subdivisions were then called subpopulations.

The stochastic dynamic in the subpopulations was then coupled with two layers of mobility flows apart from the infection dynamic within each subpopulation. The first layer was the worldwide airport network between the airports in the subpopulations, which could be seen as a weighted graph whose edges represented the number of passengers between each pair of airports. This layer was integrated into the model through stochastic transportation between subpopulations. The second layer was the commuting network that connected subpopulations graphically close. This layer was integrated through being used to compute the effective population and infection in each subpopulation. More details can be found in [1].

Appendix B Data Splitting and Choice of Hyper-parameters

In this section, we first specify the choice of training, validation, and testing data, and explain the choice of the hyper-parameter μ through cross-validation in more detail for simulated and real-world data in Section B.1. Then, we detail the choice of the algorithmic hyper-parameters in Section B.2, including σ in (2.8), the penalty factor μ, the parameter $\beta \in (0, 1)$ in (2.10), and partition of regions.

B.1 Construction of Training/Testing/Validation Data

In this section, we detail how the training, testing, and validation data are chosen for the simulated and real-world study in Sections B.1.1 and B.1.2 respectively. Then, we also explain how the fitted/predicted trajectories are computed for the training/testing/validation data. The corresponding errors are calculated

as the (weighted) average of daily relative errors for the fitted/predicted trajectories compared with the ground truth data.

We remark that the choice of training and testing data are for evaluating the model performance through training and testing errors. In addition, choosing validation data aims to help select hyper-parameters in the proposed model through validation errors, as will be mentioned in Section B.2. The details of computation of errors can be found in Section C.

B.1.1 Simulated Data

For simulated data, three random trajectories are sampled independently according to the stochastic model with prefixed parameters in one replica. For some threshold T_{th}, data from the first T_{th} days of one trajectory are treated as the training data $Y^{[\mathrm{Tr}]} = \{(\Delta C_a)_k^{[\mathrm{Tr}]}(i), k = 1, \ldots, n\}_{i=1}^{T_{th}}$, data from the last $T - T_{th}$ days of the second trajectory are treated as the validation data $Y^{[\mathrm{Val}]} = \{(\Delta C_a)_k^{[\mathrm{Val}]}(i), k = 1, \ldots, n\}_{i=T_{th}+1}^{T}$, and data from the last $T - T_{th}$ days of the last trajectory are treated as the testing data $Y^{[\mathrm{Te}]} = \{(\Delta C_a)_k^{[\mathrm{Te}]}(i), k = 1, \ldots, n\}_{i=T_{th}+1}^{T}$. This guarantees that the training and validation data are independent from the testing data and that validation data have the same distribution as testing data.

The fitting performance is assessed by comparing the training data $Y^{[\mathrm{Tr}]}$ with the fitted deterministic trajectory $Y^{[\mathrm{Tr},\Theta]} = \{(\widetilde{\Delta C_a})_k^{[\mathrm{Tr},\Theta]}(i), k = 1, \ldots, n\}_{i=1}^{T_{th}}$. $Y^{[\mathrm{Tr},\Theta]}$ is obtained by integrating ODE system (2.4) for T_{th} days with parameters $\Theta = \{E_k(0), H_k(0), \lambda_k, \gamma_k\}_{k=1}^{n}$ inferred by the corresponding model. The model can be either any baseline model or the newly proposed model.

Similarly, the generalization performance is assessed by comparing the testing data $Y^{[\mathrm{Te}]}$ with the predicted deterministic trajectory $Y^{[\mathrm{Te},\Theta]} = \{(\widetilde{\Delta C_a})_k^{[\mathrm{Te},\Theta]}(i), k = 1, \ldots, n\}_{i=T_{th}+1}^{T}$. $Y^{[\mathrm{Te},\Theta]}$ is computed by running (2.4) for T days with parameters $\Theta = \{E_k(0), H_k(0), \lambda_k, \gamma_k\}_{k=1}^{n}$ and then taking out the data of the last $T - T_{th}$ days.

Moreover, to help determine the hyper-parameter in the proposed model, we evaluate the model's performance with parameters fitted using training data on the validation data set. Specifically, for a given hyper-parameter, we first fit the parameters $\Theta = \{E_k(0), H_k(0), \lambda_k, \gamma_k\}_{k=1}^{n}$ by the training data $Y^{[\mathrm{Tr}]}$ and (2.11). Then, we compute the predicted deterministic trajectory for the validation data, $Y^{[\mathrm{Val},\Theta]} = \{(\Delta C_a)_k^{[\mathrm{Val},\Theta]}(i), k = 1, \ldots, n\}_{i=T_{th}+1}^{T}$, by integrating ODE system (2.4) for T days with the fitted Θ_μ and taking out the data of the last $T - T_{th}$ days.

B.1.2 Real-world Data

For real-world data, only one observable trajectory is available, thus the training and testing data ($Y^{[\mathrm{Tr}]}$ and $Y^{[\mathrm{Te}]}$) are the two parts of the true trajectory split by the T_{th}-th day. Moreover, the validation data $Y^{[\mathrm{Val}]}$ are part of the training data $Y^{[\mathrm{Tr}]}$, lasting from the $T_{th,2}$-th day to the T_2-th day, for some other threshold $T_{th,2}$ and end point T_2.

In contrast to the case for simulated data, for real-world data, the training data and validation data are not independent from the testing data. Instead, to measure the fitting and generalization performance for some model with the fitted parameters Θ, a deterministic trajectory for all the T days is computed with Θ, and then the data for the first T_{th} days and for the last $T - T_{th}$ days are treated as $Y^{[\mathrm{Tr},\Theta]}$ and $Y^{[\mathrm{Te},\Theta]}$. Moreover, to choose hyper-parameters for the proposed model in this paper, we first fit the parameters Θ with (4.3) using the the first $T_{th,2}$ days of training data and a given hyper-parameter. Then, we compute the predicted validation trajectory for this hyper-parameter, $Y^{[\mathrm{Val},\Theta]}$, as the last $T_2 - T_{th,2}$ days of the deterministic trajectory of total T_2 days determined by (2.4) with the parameters Θ.

We remark that for real-world data, T_{th} should not be earlier than the maximum of T_k , which is the day that λ_k changes in the k-th region, $\forall k = 1, \ldots, n$. Moreover, we choose $T_{th,2}$ such that it is earlier

than the minimum of T_k due to that the time span of the real-world data is not long enough especially the data in China.

B.2 Choice of Hyper-parameters

It can be seen from Section 2.3 and Algorithm 1 that there are several hyper-parameters that determine the model proposed, including σ in the prior distribution (2.8) along with the factor μ for graph Laplacian penalty, β, and partition of regions that all decide the affinity A through (2.10). We remark on the choice of these hyper-parameters as follows.

- σ: the factor for regularization on l_2 norm of the vector(s) of transmission parameters, appearing in both (2.11) and (4.3). In this paper, σ is chosen to be 10^{-6} for all the data sets. We present the results of sensitivity analysis performed for σ on the simulated data with four provinces in Section 3.2. The results are not sensitive to the choice of σ so long as σ is not too large. Similar patterns show for other data sets, thus details are not reported repetitively and the choice of σ is fixed as 10^{-6} throughout the experiments.

- Partition of regions: the partition determining construction of the affinity matrix A, such that correlation between regions from the same group in the partition is more addressed. For simulated data, experimental results are mainly reported with underlying graph information known. For real-world data, the partition of provinces in China is based on the medical overrun in Hubei at the initial outbreak of the epidemics, and the partition of countries in Europe is grounded on the geographical locations. Moreover, to test whether the model works when the graph information is not fully known, we also present results from mismatched/alternative partitions for simulated/real-world data in Section 3.3 and 4.4. The results show that the proposed model still predicts more accurately than the baseline models if the mismatch is not too serious (for the simulated data) or the reasonable partition may not be unique (for the real-world data).

- μ: the penalty factor of the graph Laplacian regularization, appearing in both (2.11) and (4.3). μ is chosen through cross-validation. Specifically, we separate out some validation data, then fit parameters Θ_μ for each μ and compute validation errors by comparing the actual validation trajectories with the predicted ones with fitted Θ_μ. Then, the choice of μ depends on the validation errors. In the Sections B.1.1 and B.1.2 above, we have specified how to choose validation set for simulated and real-world data in detail respectively. In addition, the formulas of validation errors can be found in Section C. For the case of simulated data, validation errors are identically distributed as testing errors, which are used to evaluate the prediction performance of models, μ is chosen to minimize the validation errors. On the contrary, for real-world data when validation errors may have different distribution from testing errors, results from multiple choices of μ with relative small validation errors are presented. Details of results are in corresponding sections of experimental results.

- β: the parameter taking values in $(0, 1)$ that reduces the correlation between inter-group regions through (2.10). In this paper, β is fixed to be 0.1 for all the experiments. We remark that other reasonable choices of $\beta \in (0, 1)$ would also lead to similar results, which might not be exactly the same as those with $\beta = 0.1$ though. β can also be chosen through cross-validation, similarly to the choice of μ, and we omit the details in this paper.

Appendix C Model Evaluation Metric

In Section B, we have specified the choice of ground truth training/testing/validation data and explained how they are estimated using models for both simulated and real-world data.

In the current subsection, we define the training, validation, and testing errors for model evaluation and help choose hyper-parameters of the model. Specifically, we introduce the definitions of mean relative error (which will be abbreviated as MAE hereinafter) and root mean squared relative error (which will be abbreviated as MSE for short), simply averaged or weighted.

In the paper, the superscripts $^{[\mathrm{Tr}]}$, $^{[\mathrm{Val}]}$, and $^{[\mathrm{Te}]}$ refer to when the error (MAE and MSE) is computed on the training, validation and testing data respectively. We thus introduce the definition of error here omitting the superscript for simplicity.

We compute the error over the days from $i = 1$ to T_e. Given the data of true daily confirmed cases ΔC_a and those estimated from the model $\widetilde{\Delta C_a}$, we define

$$\mathrm{Err}_k(i) = \frac{|(\Delta C_a)_k(i) - (\widetilde{\Delta C_a})_k(i)|}{(\Delta C_a)_k(i)} \tag{A1}$$

as the relative error in the k-th region on the i-th day. Then, the total error for the k-th region are the weighted sum of $\{\mathrm{Err}_k(i)\}_{i=1}^{T_e}$. We use weighted average in computing the error, and adopt the following two types of weights $\alpha_k(i)$ cross region k and time i. The first type of weight, denoted with subscript $_{(w)}$, is defined as

$$\alpha_{(w),k}(i) = \frac{(\Delta C_a)_k(i)}{\sum_{i'=1}^{T_e}(\Delta C_a)_k(i')}, \tag{A2}$$

which share the same trend of increasing or decreasing with the newly confirmed cases in the k-th region and give larger weights to errors with more cases. The second type of weight conducts simple average over time, namely,

$$\alpha_{(s),k}(i) = \frac{1}{T_e}, \quad \forall k \in \{1, \ldots, n\}, \quad i \in \{1, \ldots, T_e\}. \tag{A3}$$

With the two types of weights (A2) and (A3), we define the weighted MAE and MSE for the k-th region as

$$\mathrm{MAE}_{(w),k} = \sum_{i=1}^{T_e} \alpha_{(w),k}(i)\mathrm{Err}_k(i), \quad \mathrm{MSE}_{(w),k} = \left(\sum_{i=1}^{T_e} \alpha_{(w),k}(i)\left(\mathrm{Err}_k(i)\right)^2\right)^{1/2}. \tag{A4}$$

The errors weighted by $\alpha_{(s),k}(i)$ are defined similarly and denoted with subscript $_{(s)}$ instead. Note that, strictly speaking, the error for each day and region before taking average is a relative error, c.f. (A1). We use the name 'MAE' and 'MSE' following the convention.

The definition (A4) is for each region k. The total error over all regions is defined by taking an average over regions, that is,

$$\mathrm{MAE}_{(w)} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{MAE}_{(w),k}, \quad \mathrm{MSE}_{(w)} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{MSE}_{(w),k}. \tag{A5}$$

The simple-average weight counterparts $\mathrm{MAE}_{(s)}$ and $\mathrm{MSE}_{(s)}$ are defined in the same way from $\mathrm{MAE}_{(s),k}$ and $\mathrm{MSE}_{(s),k}$ respectively.

Appendix D Preprocessing of COVID-19 data in China

Before the data is analyzed, an initial screening over provinces is applied to the COVID-19 data in China. In our study, we find this preprocessing step necessary for stability of inferring the model parameters.

Specifically, only the provinces or municipalities in China that satisfy the following conditions are considered:

1. The accumulated confirmed cases by February 10th 2020 are above 50. If the accumulated confirmed cases is too small, the newly confirmed cases touch zero frequently and the compartment model considered in this paper may not be appropriate to characterize the spread of COVID-19 in such region.

2. The estimated removal rate δ_k is not less than 10^{-10}. If the newly removed cases are too small, then the two-step procedure proposed in this model may not estimate the parameters appropriately.

3. The change of transmission parameter occurs no later than February 2nd 2020. Note that the data in China last to February 10th 2020, and the training data end on February 5th 2020. If the changing point is too late, the training data could not capture the transmission parameter in the second period.

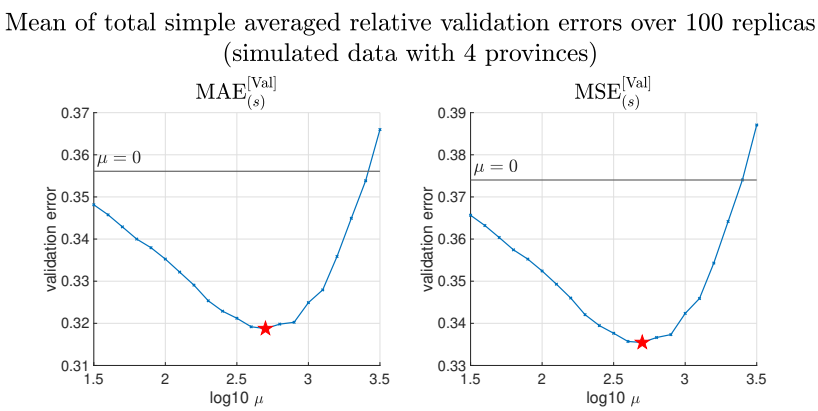

Figure A1: Mean of simply averaged validation errors $\mathrm{MAE}_{(s)}^{[\mathrm{Val}]}$ and $\mathrm{MSE}_{(s)}^{[\mathrm{Val}]}$ against μ (computed by (A5)) over 100 replicas for simulated data (four provinces). The horizontal lines show the values of the mean of the validation errors when $\mu = 0$.

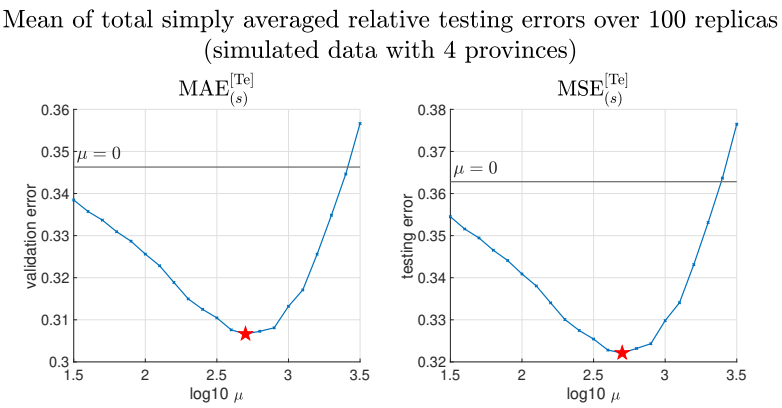

Figure A2: Mean of simply averaged testing errors $\mathrm{MAE}_{(s)}^{[\mathrm{Te}]}$ and $\mathrm{MSE}_{(s)}^{[\mathrm{Te}]}$ against μ (computed by (A5)) over 100 replicas for simulated data (four provinces). The horizontal lines show the values of the mean of the testing errors when $\mu = 0$.

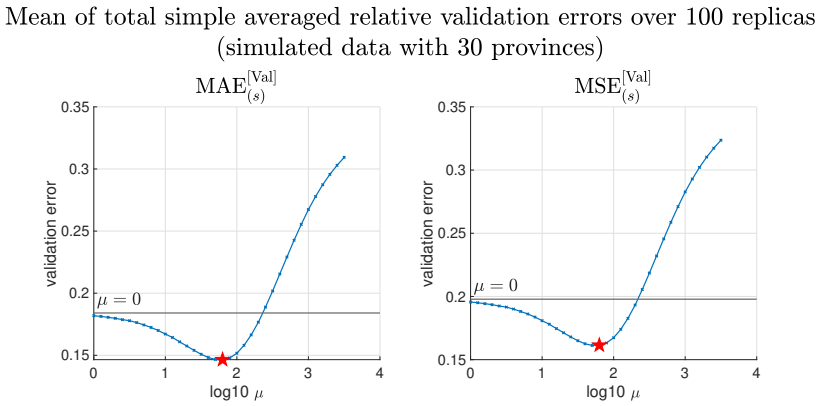

Figure A3: Mean of simply averaged validation errors $\mathrm{MAE}_{(s)}^{[\mathrm{Val}]}$ and $\mathrm{MSE}_{(s)}^{[\mathrm{Val}]}$ against μ (computed by (A5)) over 100 replicas for simulated data (thirty provinces). The horizontal lines show the values of the mean of the validation errors when $\mu = 0$.

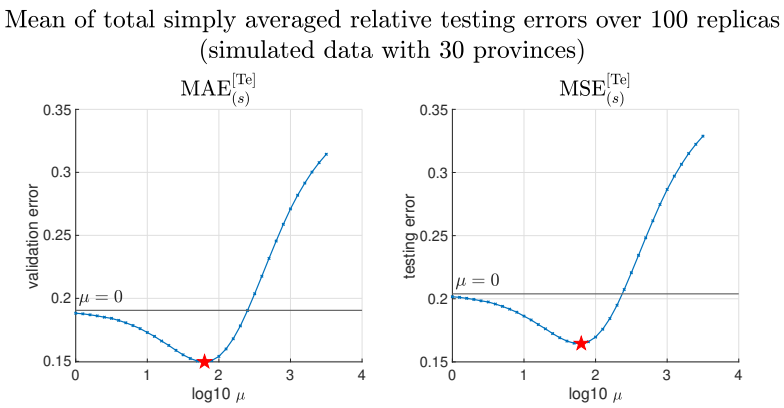

Figure A4: Mean of simply averaged testing errors $\mathrm{MAE}_{(s)}^{[\mathrm{Te}]}$ and $\mathrm{MSE}_{(s)}^{[\mathrm{Te}]}$ against μ (computed by (A5)) over 100 replicas for simulated data (thirty provinces). The horizontal lines show the values of the mean of the testing errors when $\mu = 0$.

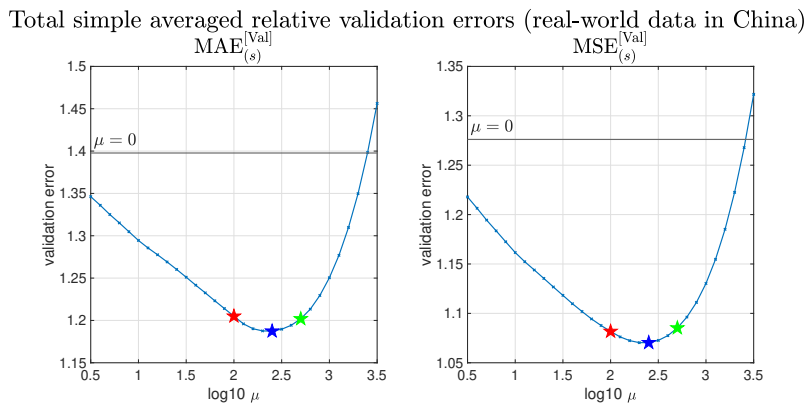

Figure A5: Simply averaged validation errors $\mathrm{MAE}_{(s)}^{[\mathrm{Val}]}$ and $\mathrm{MSE}_{(s)}^{[\mathrm{Val}]}$ against μ (computed by (A5)) for real-world data in China. The horizontal lines show the values of the validation errors when $\mu = 0$.

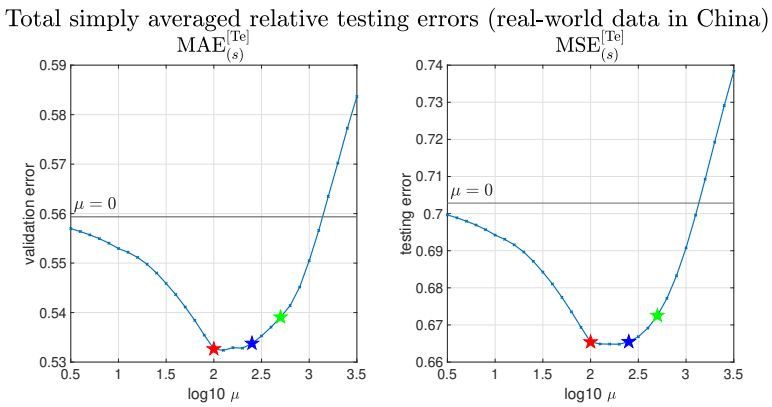

Figure A6: Simply averaged testing errors MAE$^{[\mathrm{Te}]}_{(s)}$ and MSE$^{[\mathrm{Te}]}_{(s)}$ against μ (computed by (A5)) for real-world data in China. The horizontal lines show the values of the testing errors when $\mu = 0$.

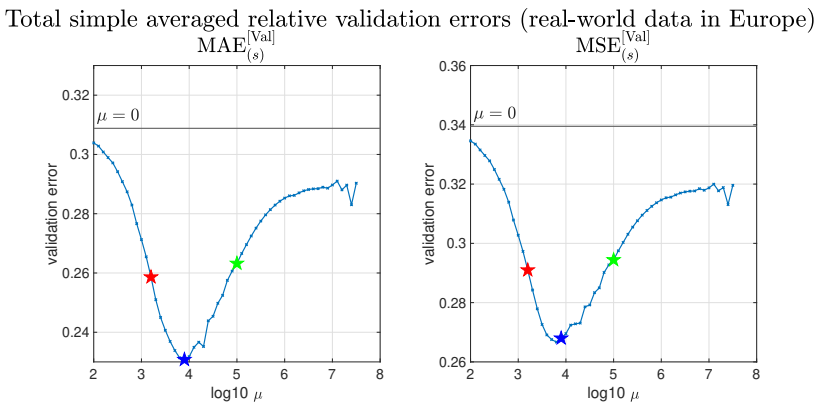

Figure A7: Simply averaged validation errors MAE$^{[\mathrm{Val}]}_{(s)}$ and MSE$^{[\mathrm{Val}]}_{(s)}$ against μ (computed by (A5)) for real-world data in Europe. The horizontal lines show the values of the validation errors when $\mu = 0$.

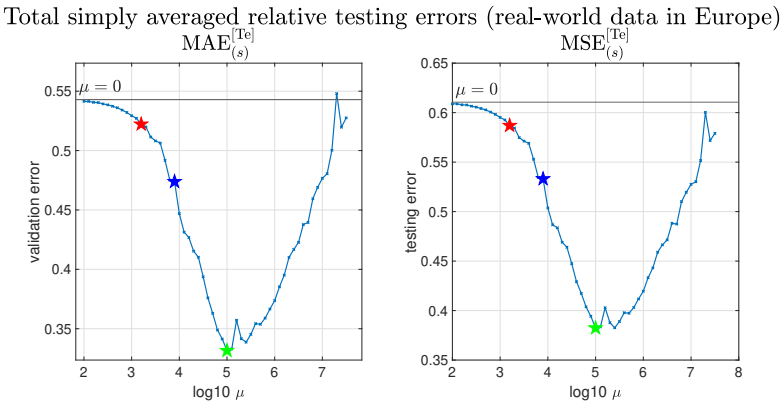

Figure A8: Simply averaged testing errors $\mathrm{MAE}_{(s)}^{[\mathrm{Te}]}$ and $\mathrm{MSE}_{(s)}^{[\mathrm{Te}]}$ against μ (computed by (A5)) for real-world data in Europe. The horizontal lines show the values of the testing errors when $\mu = 0$.