medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Estimation of local time-varying reproduction numbers in noisy surveillance data

Wenrui Li[1*], Katia Bulekova[2], Brian Gregor[2], Laura F. White[3], Eric D. Kolaczyk [1,4]

1 Department of Mathematics and Statistics, Boston University, Boston MA, USA

2 Research Computing Services, Information Services and Technology, Boston University, Boston MA, USA

3 Department of Biostatistics, Boston University, Boston MA, USA

4 Hariri Institute for Computing, Boston University, Boston MA, USA

* wenruili@bu.edu

Abstract

A valuable metric in understanding infectious disease local dynamics is the local time-varying reproduction number, i.e. the expected number of secondary local cases caused by each infected individual. Accurate estimation of this quantity requires distinguishing cases arising from local transmission from those imported from elsewhere. Realistically, we can expect identification of cases as local or imported to be imperfect. We study the propagation of such errors in estimation of the local time-varying reproduction number. In addition, we propose a Bayesian framework for estimation of the true local time-varying reproduction number when identification errors exist. And we illustrate the practical performance of our estimator through simulation studies and with outbreaks of COVID-19 in Hong Kong and Victoria, Australia.

Introduction

Epidemic modeling, while not at all new, has taken on renewed importance due to the COVID-19 pandemic. The local time-varying reproduction number, $R_*^{\text{local}}(t)$, is an

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

important quantity to monitor the infectiousness and transmissibility of diseases and, therefore, to design and adjust public health responses during an outbreak. Recent examples include monitoring transmission of the COVID-19 pandemic and demonstrating the efficacy of non-pharmaceutical interventions in more than 100 countries [1–4]. The value of $R_*^{\text{local}}(t)$ represents the expected number of secondary local cases arising from a primary case infected at time t. Different formal definitions of $R_*^{\text{local}}(t)$ have been proposed, and a number of methods are available to estimate this quantity. The most widely used is an estimator of the instantaneous reproduction number that is defined as the ratio of the expected number of incident locally infected cases at time t to the expected total infectiousness of infected individuals at time t [5, 6].

Distinguishing local cases from imported cases is essential to estimation of the local time-varying reproduction number. However, surveillance data generally is available only up to some level of error. For example, if we are unable to identify the correct source of infection from contact tracing or genetic information, imported cases might be misclassified as local cases, and vice versa. Such misclassification error is recognized as one limitation of estimating $R_*^{\text{local}}(t)$ in the COVID-19 outbreak [7, 8]. We investigate how identification error impacts on the estimation of the instantaneous reproduction number and, thus, on our understanding of diseases transmission dynamics.

Extensive work regarding improving inference of time-varying reproduction numbers has been done. For instance, there have been efforts to estimate the serial interval that is used to compute the total infectiousness for $R_*^{\text{local}}(t)$ estimation, including Bayesian parametric estimation using data augmentation Markov Chain Monte Carlo [9], and a cure model for limited follow-up data [10]. Many studies have explored the effects of imperfect detection and estimated the true infection prevalence [8, 11–13]. But, to our best knowledge, there has been little attention to date given towards accounting for identification errors of local and imported cases.

Our contribution in this paper is to quantify how such errors propagate to the local time-varying reproduction number, and to provide estimators for $R_*^{\text{local}}(t)$ when contact tracing survey information is available. Adopting the definition of $R_*^{\text{local}}(t)$ proposed by [5], we characterize the impact of identification errors on the bias of noisy local time-varying reproduction numbers. Our work shows that, in general, the bias can be expected to be nontrivial. Accordingly, we propose a Bayesian framework to estimate

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

the true local time-varying reproduction number. Numerical simulation suggests that high accuracy is possible for estimating local time-varying reproduction numbers in outbreaks of even modest size. We illustrate the practical use of our estimators in the context of COVID-19 pandemic in Hong Kong and Victoria, Australia.

The organization of this paper is as follows. In Methods Section we show the bias of the noisy local time-varying reproduction number, and propose a Bayesian hierarchical framework to estimate the true local time-varying reproduction number with imperfect knowledge. Results Section reports the practical performance of our estimators through simulation studies and with SARS-CoV-2 infections in Hong Kong and Australia. Finally, we conclude in Discussion Section with a discussion of future directions for this work.

Methods

In this section, we first quantify the bias of the noisy local time-varying reproduction number when misidentification occurs in the surveillance data. We then build a Bayesian hierarchical framework to estimate true local time-varying reproduction numbers. We also propose a method to estimate misidentification rates based on contact tracing survey data, which informs the prior distribution in the model.

Notation

We provide essential notation and background here. The number of newly infected cases at time t, $I_*(t)$, is the sum of the numbers of local ($I_*^{\text{local}}(t)$) and imported ($I_*^{\text{imported}}(t)$) cases. If one assumes independence between calendar time and the generation interval, $g(s)$, then the local time-varying reproduction number is defined as [5]

$$R_*^{\text{local}}(t) = \frac{\mu_*^{\text{local}}(t)}{\int_0^\infty g(s)\mu_*(t-s)ds}, \tag{1}$$

where $\mu_*^{\text{local}}(t) = \mathbb{E}[I_*^{\text{local}}(t)]$ and $\mu_*(t) = \mathbb{E}[I_*(t)]$.

In reality, we only know the serial interval and the number of diagnosed cases. Let $I(t)$, $I^{\text{local}}(t)$ and $I^{\text{imported}}(t)$ be the numbers of total diagnosed cases, local diagnosed cases, and imported diagnosed cases at time t, respectively. Then, we define a realistic

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

local time-varying reproduction number as

$$R^{\text{local}}(t) = \frac{\mu^{\text{local}}(t)}{\int_0^\infty w(s)\mu(t-s)ds}, \tag{2}$$

where $w(s)$ is the serial interval, $\mu^{\text{local}}(t) = \mathbb{E}[I^{\text{local}}(t)]$ and $\mu(t) = \mathbb{E}[I(t)]$. Note that the serial interval corresponds to date of symptom onset. One can estimate symptom onset dates by back calculation of report dates [14].

Realistically, we can expect identification of cases as local or imported to be imperfect. Let $\tilde{I}^{\text{local}}(t)$ and $\tilde{I}^{\text{imported}}(t)$ be the number of new local and imported cases reported at time t, with identification error. Thus, we define a noisy local time-varying reproduction number as

$$\tilde{R}^{\text{local}}(t) = \frac{\tilde{\mu}^{\text{local}}(t)}{\int_0^\infty w(s)\mu(t-s)ds}, \tag{3}$$

where $\tilde{\mu}^{\text{local}}(t) = \mathbb{E}[\tilde{I}^{\text{local}}(t)]$. The definition of $\tilde{R}^{\text{local}}(t)$ in (3) comes from an argument that mimics the original argument using Poisson arrivals in [15]. Specifically, we suppose that we observe a Poisson stream $\tilde{I}^{\text{local}}(t)$ that is a function of calendar time t in terms of the transmissibility, denoted $\tilde{\beta}^{\text{local}}(t,s)$, an arbitrary function of calendar time t and time since infection s. Then, $\tilde{\mu}^{\text{local}}(t)$ follows the so-called renewal equation

$$\tilde{\mu}^{\text{local}}(t) = \int_0^\infty \tilde{\beta}^{\text{local}}(t,s)\mu(t-s)ds. \tag{4}$$

Following [15], we have

$$\tilde{\beta}^{\text{local}}(t,s) = \tilde{R}^{\text{local}}(t)w(s). \tag{5}$$

Inserting (5) into (4) yields the definition of $\tilde{R}^{\text{local}}(t)$ in (3).

Our interest is in characterizing the manner in which the uncertainty in $\tilde{I}^{\text{local}}(t)$ and $\tilde{I}^{\text{imported}}(t)$ propagates to the local time-varying reproduction number, and providing estimators of $R^{\text{local}}(t)$ to account for identification errors.

Bias of the noisy local time-varying reproduction number

We quantify the bias of the noisy local time-varying reproduction number in (3) when misidentification occurs. We begin by defining a model for $\tilde{I}^{\text{local}}(t)$ and $\tilde{I}^{\text{imported}}(t)$. Let α_0 denote the probability that an imported case is misidentified as local, and α_1 the

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

probability that a local case is misidentified as imported. Then, a simple model is

$$\begin{aligned} \tilde{I}^{\text{local}}(t)|I^{\text{local}}(t), I^{\text{imported}}(t), \alpha_0, \alpha_1 &\sim \text{Bin}(I^{\text{local}}(t), 1-\alpha_1) + \text{Bin}(I^{\text{imported}}(t), \alpha_0), \\ \tilde{I}^{\text{imported}}(t) &= I^{\text{local}}(t) + I^{\text{imported}}(t) - \tilde{I}^{\text{local}}(t). \end{aligned} \tag{6}$$

Under independence, the first relationship in (6) is directly obtained by the definition of α_0 and α_1. And the second equation in (6) is due to the fact that the total number of cases reported at time t is not affected by the misidentification.

By (6), the relationship between $\tilde{\mu}^{\text{local}}(t)$ and $\mu^{\text{local}}(t)$ is

$$\tilde{\mu}^{\text{local}}(t) = (1-\alpha_1)\mu^{\text{local}}(t) + \alpha_0 \mu^{\text{imported}}(t), \tag{7}$$

where $\mu^{\text{imported}}(t) = \mathbb{E}(I^{\text{imported}}(t))$. Direct computation yields

$$\tilde{R}^{\text{local}}(t) = \Big(1 - \alpha_1 + \alpha_0 \frac{\mu^{\text{imported}}(t)}{\mu^{\text{local}}(t)}\Big) R^{\text{local}}(t) \tag{8}$$

when $\mu^{\text{local}}(t) \neq 0$. From (8), we can see that the bias of $\tilde{R}^{\text{local}}(t)$ depends on α_0, α_1 and the ratio of $\mu^{\text{imported}}(t)$ and $\mu^{\text{local}}(t)$. When $\mu^{\text{imported}}(t)/\mu^{\text{local}}(t) = 1$, we have $\tilde{R}^{\text{local}}(t) > R^{\text{local}}(t)$ if $\alpha_0 > \alpha_1$, and $\tilde{R}^{\text{local}}(t) < R^{\text{local}}(t)$ if $\alpha_0 < \alpha_1$.

Bayesian hierarchical modeling to account for misidentification

We propose a Bayesian framework to estimate $R^{\text{local}}(t)$ using noisy surveillance data. Following [5, 6, 15], we specify

$$I^{\text{local}}(t)|R^{\text{local}}(t), n(t-1), w(s) \sim \text{Pois}(R^{\text{local}}(t) \cdot \Lambda(t)), \text{ for } t > 0, \tag{9}$$

where $\Lambda(t) = \sum_{s=1}^{t} w(s) I(t-s)$ is the total infectiousness of infected individuals at time t, and $n(t-1)$ represent the historical data up to time $t-1$ (i.e., $I^{\text{local}}(0), I^{\text{imported}}(0), \cdots, I^{\text{local}}(t-1), I^{\text{imported}}(t-1)$). Note that $\Lambda(t)$ is undefined for $t = 0$. So, we assume that

$$I^{\text{local}}(0)|\mu^{\text{local}}(0) \sim \text{Pois}(\mu^{\text{local}}(0)). \tag{10}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

And we assume the imported case counts follow a Poisson distribution:

$$I^{\text{imported}}(t)|\mu^{\text{imported}}(t) \quad \sim \quad \text{Pois}(\mu^{\text{imported}}(t)). \tag{11}$$

Next, we define relevant prior distributions. We assume a distribution for $R^{\text{local}}(t)$ of the form

$$R^{\text{local}}(t)|n(t-1), w(s) \quad \sim \quad \text{Gamma}(a^{\text{local}}_{t|t-1}, b^{\text{local}}_{t|t-1}), \text{ for } t > 0. \tag{12}$$

This choice is similar to that in [5], but differs in that we specify gamma conditioned on the history, rather than marginally. The conditioning reflects the expectation that the evolution of $R^{\text{local}}(t)$ is likely to depend on the course of infection in the population and intervention measures that may result. Analogously, we also assume gamma distributed priors for $\mu^{\text{imported}}(t)$ and $\mu^{\text{local}}(0)$, that is,

$$\begin{aligned} \mu^{\text{imported}}(t) \quad &\sim \quad \text{Gamma}(a^{\text{imported}}_t, b^{\text{imported}}_t), \\ \mu^{\text{local}}(0) \quad &\sim \quad \text{Gamma}(a^{\text{local}}_0, b^{\text{local}}_0). \end{aligned} \tag{13}$$

In addition, we assume the convention that the misidentification rates are beta distributed, and hence given by

$$\begin{aligned} \alpha_0 \quad &\sim \quad \text{Beta}(\zeta_{\alpha_0}, \xi_{\alpha_0}), \\ \alpha_1 \quad &\sim \quad \text{Beta}(\zeta_{\alpha_1}, \xi_{\alpha_1}). \end{aligned} \tag{14}$$

By using Markov chain Monte Carlo (MCMC) simulation, we can get both estimates of $R^{\text{local}}(t)$ and its uncertainty. We implement MCMC using the R package, NIMBLE [16–18] with the default assignment of sampler algorithms. The samplers assigned to the variables are as follows: Gibbs samplers are assigned to $\mu^{\text{local}}(0)$ and $\mu^{\text{imported}}(t)$, $t \geq 0$, which have conjugate relationships between their prior distribution and the distributions of their stochastic dependents; slice samplers [19] are used for $I^{\text{local}}(t)$ and $I^{\text{imported}}(t)$, $t \geq 0$; Metropolis-Hastings adaptive random-walk samplers are set to α_0, α_1 and $R^{\text{local}}(t)$, $t > 0$.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Estimating misidentification rates

Without any information on the misidentification rates, it is difficult to get an accurate estimator of $R^{\text{local}}(t)$. However, contact tracing data could provide adequate information to estimate the misidentification rates.

Let p_i be the probability that we think individual i is a local case based on the survey. Then, p_i is a mixture of α_0 and $1-\alpha_1$. Note that $\alpha_1 \sim \text{Beta}(\zeta_{\alpha_1}, \xi_{\alpha_1})$ implies $1-\alpha_1 \sim \text{Beta}(\xi_{\alpha_1}, \zeta_{\alpha_1})$. We thus model the distribution of p_i as a mixture of two beta distributions:

$$p_i \quad \sim \quad \pi_0 \text{Beta}(\zeta_{\alpha_0}, \xi_{\alpha_0}) + (1-\pi_0)\text{Beta}(\xi_{\alpha_1}, \zeta_{\alpha_1}), \tag{15}$$

where π_0 can be interpreted as the fraction of the diagnosed cases that are imported. By using the expectation–maximization (EM) algorithm, we can obtain estimators $\hat{\zeta}_{\alpha_0}, \hat{\xi}_{\alpha_0}, \hat{\zeta}_{\alpha_1}$ and $\hat{\xi}_{\alpha_1}$.

Note that, if $1-\zeta_{\alpha_0}/(\zeta_{\alpha_0}+\xi_{\alpha_0}) - \zeta_{\alpha_1}/(\zeta_{\alpha_1}+\xi_{\alpha_1}) \neq 0$, we obtain unbiased estimators of $I^{\text{local}}(t)$ and $I^{\text{imported}}(t)$

$$\begin{aligned} \hat{I}^{\text{local}}(t) &= \frac{[1-\zeta_{\alpha_0}/(\zeta_{\alpha_0}+\xi_{\alpha_0})] \cdot \tilde{I}^{\text{local}}(t) - \zeta_{\alpha_0}/(\zeta_{\alpha_0}+\xi_{\alpha_0}) \cdot \tilde{I}^{\text{imported}}(t)}{1-\zeta_{\alpha_0}/(\zeta_{\alpha_0}+\xi_{\alpha_0}) - \zeta_{\alpha_1}/(\zeta_{\alpha_1}+\xi_{\alpha_1})}, \\ \hat{I}^{\text{imported}}(t) &= \frac{[1-\zeta_{\alpha_1}/(\zeta_{\alpha_1}+\xi_{\alpha_1})]\tilde{I}^{\text{imported}}(t) - \zeta_{\alpha_1}/(\zeta_{\alpha_1}+\xi_{\alpha_1})\tilde{I}^{\text{local}}(t)}{1-\zeta_{\alpha_0}/(\zeta_{\alpha_0}+\xi_{\alpha_0}) - \zeta_{\alpha_1}/(\zeta_{\alpha_1}+\xi_{\alpha_1})}. \end{aligned} \tag{16}$$

Thus, good initial values of $I^{\text{local}}(t)$ and $I^{\text{imported}}(t)$ in MCMC are estimators of $\hat{I}^{\text{local}}(t)$ and $\hat{I}^{\text{imported}}(t)$ based on the estimated misidentification rates, i.e., replacing $\zeta_{\alpha_0}, \xi_{\alpha_0}, \zeta_{\alpha_1}, \xi_{\alpha_1}$ in (16) by $\hat{\zeta}_{\alpha_0}, \hat{\xi}_{\alpha_0}, \hat{\zeta}_{\alpha_1}, \hat{\xi}_{\alpha_1}$.

Results

In this section, we conduct some simulations to illustrate the performance of the proposed estimation methods. And we apply our method to two real data sets. One is surveillance data of COVID-19 in Hong Kong that includes contact tracing information, including travel history data [20]. They collected information on 1,038 SARS-CoV-2 cases confirmed between 23 January and 28 April 2020. And they identified 355 local cases and 683 imported cases. The other data set is from the COVID-19 pandemic in Victoria, Australia, studied in [21]. There they had 1,333 laboratory-confirmed cases of COVID-19 between 6

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

January and 14 April 2020. After excluding duplicate patients from cases, they identified 345 local cases and 558 imported cases.

We consider two settings, a simulation setting and an application setting. In the simulation setting, we first use surveillance data from Hong Kong and Victoria to create realistic simulated data, and then we add identification errors to the 'true' local and imported cases derived from the simulated epidemics, finally we estimate the local time-varying reproduction number using the noisy local and imported cases counts. In the application setting, we assume that identified local and imported cases in the real data sets are with some error. The former results allow us to understand what properties can be expected of our estimators, while the latter are reflective of what would be observed in practice with such data.

Simulation study

In this simulation study, we used Covasim [22], a stochastic individual-based model for transmission of SARS-CoV-2, calibrated to the epidemics in Hong Kong and Victoria. Fig 1 shows the average daily local and imported diagnosed counts over 1,000 trials. The noisy $\tilde{I}^{\text{local}}(t)$ and $\tilde{I}^{\text{imported}}(t)$ are generated according to (6). We set $\alpha_0 \sim \text{Beta}(2, 18)$ (mean of 0.1), and $\alpha_1 \sim \text{Beta}(2, 8)$ (mean of 0.2), $\text{Beta}(4, 8)$ (mean of 0.33), or $\text{Beta}(8, 8)$ (mean of 0.5) to see the effect of small α_0 and large α_1. This might happen if the definition of imported cases relies on travel history collected in the case investigation and some people are infected locally, even though they have a travel history within 14 days prior to symptom onset. We also consider $\alpha_1 \sim \text{Beta}(2, 18)$, and $\alpha_0 \sim \text{Beta}(2, 8)$, $\text{Beta}(4, 8)$, or $\text{Beta}(8, 8)$ (corresponding to small α_1 and large α_0, which might occur if cases are defined as local when we are not sure about their source of infection.) We assume that both α_0 and α_1 are unknown.

We evaluate the estimate for $R^{\text{local}}(t)$ in terms of a corresponding posterior, and 95% credible intervals. Fig 2 and 3 show the simulation results, in which we run MCMC chains of 10,000 samples for each of 1,000 simulated epidemic trials. Fig 2 assumes that we are more likely to misclassify local cases as imported cases and Fig 3 assumes that we are more likely to misclassify imported cases as local cases. For comparison purposes, we compute $R_*^{\text{local}}(t)$ and $R^{\text{local}}(t)$ defined in (1) and (2) by approximating

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

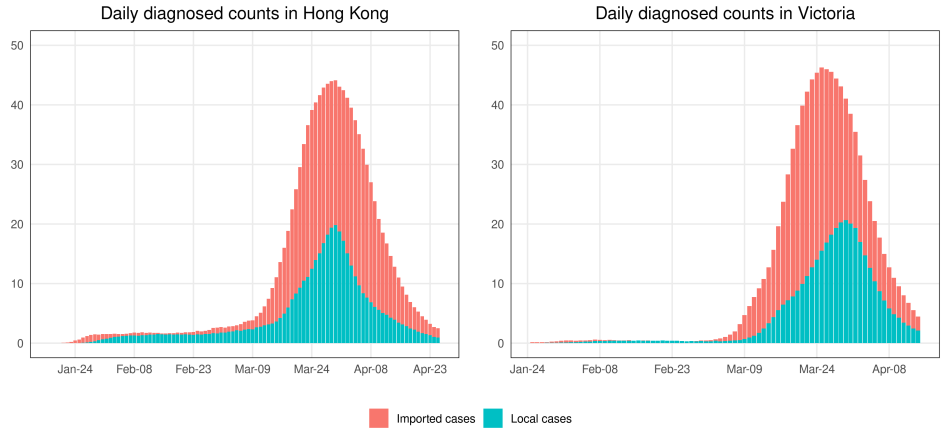

Fig 1. The means of daily local and imported diagnosed counts in 1,000 simulation trials for epidemics in Hong Kong and Victoria.

$\mu_*^{\text{local}}(t)$, $\mu_*(t)$, $g(s)$, $\mu^{\text{local}}(t)$, $\mu(t)$, $w(s)$ using 1,000 simulation trials. And we calculate the most widely used estimator of $\tilde{R}^{\text{local}}(t)$ defined in (3), which is implemented in the R package, EpiEstim [23]. We view it as a representative estimator that does not account for misidentification, i.e., it treats the noisy local and imported cases as true.

In the simulated epidemics for both Hong Kong and Victoria, if we ignore the misidentification, we will underestimate $R^{\text{local}}(t)$ when the mean of α_0 is small and the mean of α_1 is relatively large (Fig 2), and overestimate $R^{\text{local}}(t)$ when the mean of α_1 is small and the mean of α_0 is relatively large (Fig 3), with the biases increasing when the means of α_0 and α_1 increase. The results are consistent with (8) implying that the biases will lead to inappropriate public health response, i.e., inadequate interventions or overreaction. We correct the bias by our Bayesian hierarchical framework. The biases of our estimators are close to zero in all cases. The 95% credible intervals of our estimators are wide in the first two months because the number of incident cases are very low. For the last month or so when the diagnosed counts are relatively high, the 95% credible intervals are narrow.

Application

We apply our proposed methods to surveillance data of COVID-19 in Hong Kong and Victoria. Fig 4 (a) and (b) show the daily local and imported cases counts in Hong Kong and Victoria. For Hong Kong data, [20] calculated the serial intervals using a gamma distribution and estimated shape and rate parameters of 2.23 and 0.37, respectively

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

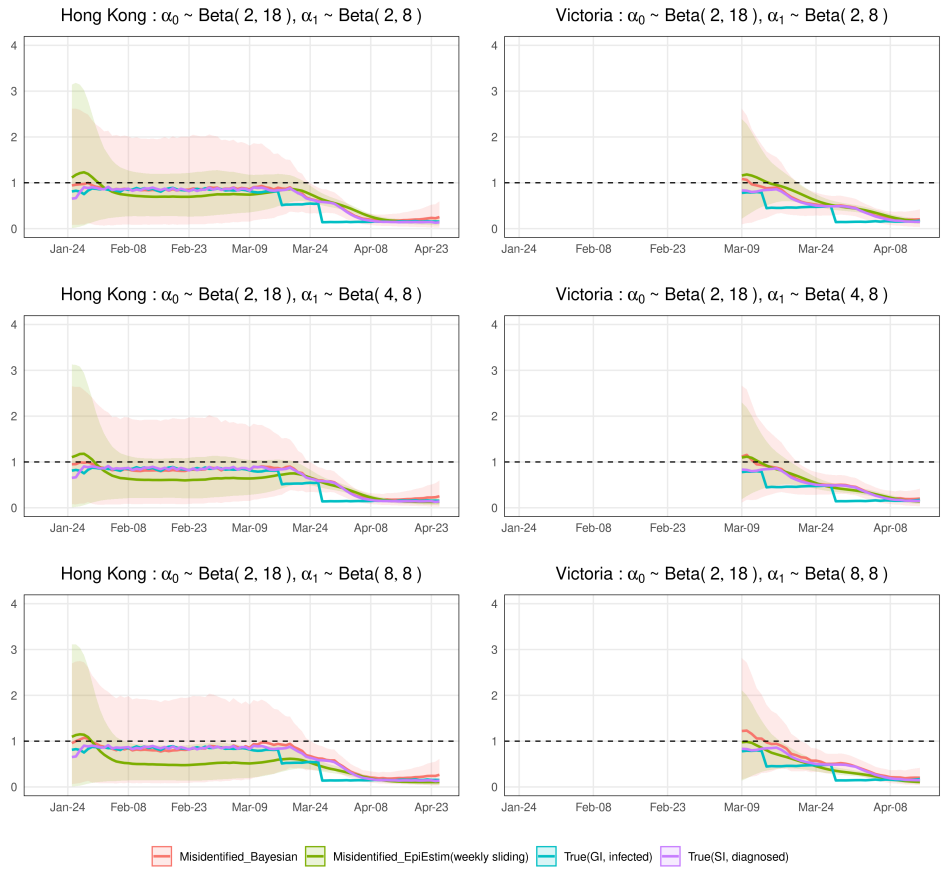

Fig 2. Estimations of local time-varying reproduction numbers in simulated epidemics for Hong Kong and Victoria under three sets of error misidentification rates: $\alpha_0 \sim \text{Beta}(2, 18)$, and $\alpha_1 \sim \text{Beta}(2, 8)$, $\text{Beta}(4, 8)$, or $\text{Beta}(8, 8)$. The error bands are the averages of 95% credible intervals over 1,000 trials. Note that the differences between the blue curve ($R_*^{\text{local}}(t)$) and the purple curve ($R^{\text{local}}(t)$) are due to the differences among infected dates, symptom onset dates, diagnosed dates.

(corresponding to a mean of around 6 days and standard deviation of around 4 days). There is no specific serial interval that has been calculated for Victoria. Considering the epidemic curve in Victoria is relatively similar to that in Hong Kong, we use the same serial interval distribution when we estimate $R^{\text{local}}(t)$ in Victoria.

Fig 4 (c) and (d) show estimates for $R^{\text{local}}(t)$ under three scenarios: 1) no identification error, 2) small α_0 and large α_1, 3) small α_1 and large α_0. We run MCMC chains of 10,000 samples and the error bands are the 95% credible intervals. We can see that the estimated local time-varying reproduction numbers are quite different when the two identification error rates are about 10% and 30%. If we think we are more likely to misclassify local cases as imported, then we should trust the curve corresponding to scenario 2). If imported cases are more likely to be misidentified as local, then the curve

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

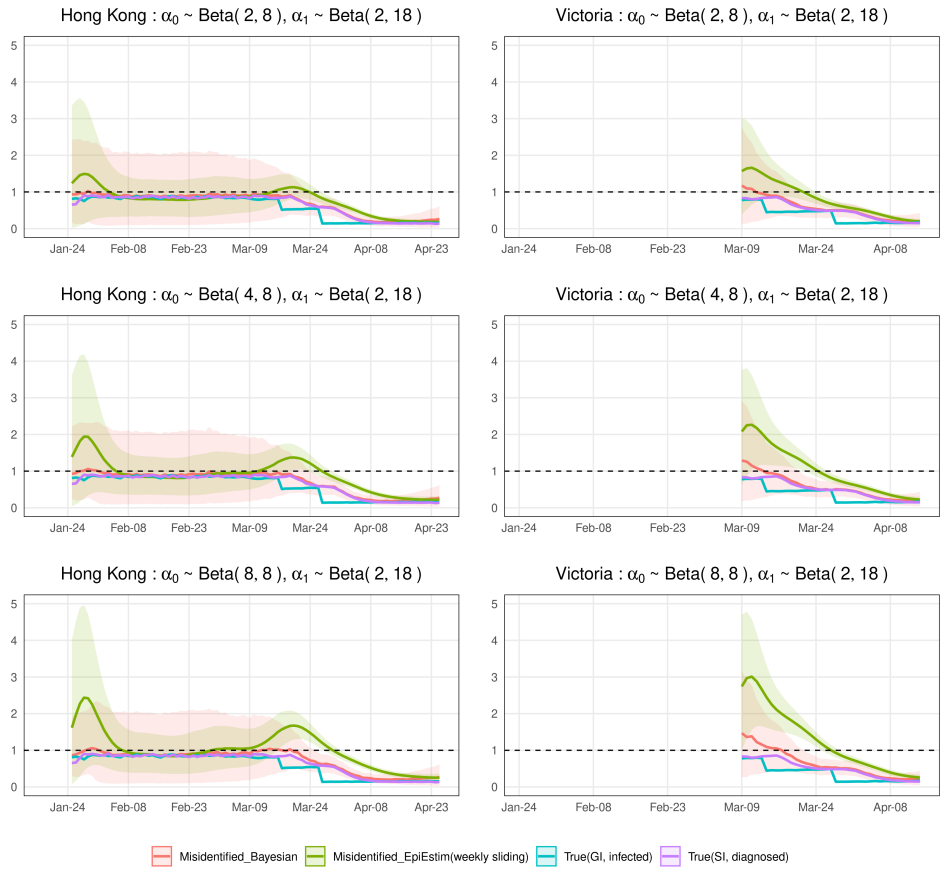

Fig 3. Estimations of local time-varying reproduction numbers in simulated epidemics for Hong Kong and Victoria under three sets of error misidentification rates: $\alpha_1 \sim \text{Beta}(2, 18)$, and $\alpha_0 \sim \text{Beta}(2, 8)$, $\text{Beta}(4, 8)$, or $\text{Beta}(8, 8)$. The error bands are the averages of 95% credible intervals over 1,000 trials.

corresponding to scenario 3) is reliable. And if we believe the identification error is close to zero, we should trust the estimate under scenario 1).

Ultimately, we see that the ability to account for identification error appropriately in reporting the local time-varying reproduction number can lead to substantially different conclusions than use of the original, noisy local time-varying reproduction number. These differences can then in turn be translated to decision making for public health response.

Discussion

We have developed a general framework for estimation of the true local time-varying reproduction numbers in contexts wherein one has identified local and imported case counts with some error. Simulations demonstrate that substantial inferential accuracy

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

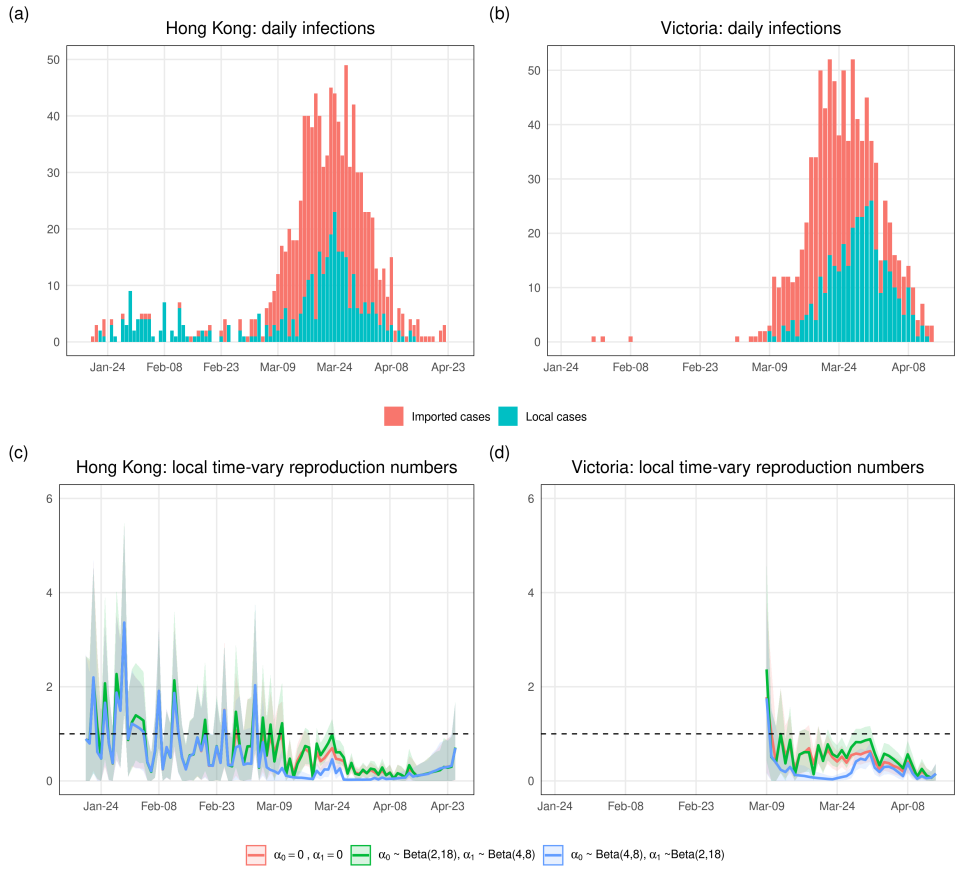

Fig 4. Epidemic curves of COVID-19 cases and estimations of local time-varying reproduction numbers in Hong Kong and Victoria. (a) The epidemic curve of daily cases of laboratory-confirmed SARS-CoV-2 infection in Hong Kong by symptom onset date and colored by case category. Asymptomatic cases are included here by date of confirmation. (b) The epidemic curve of the coronavirus disease cases in Victoria by sample collection date and colored by case category. (c) and (d) Estimations of local time-varying reproduction numbers under three scenarios: 1) no identification error, 2) $\alpha_0 \sim \text{Beta}(2, 18)$ and $\alpha_1 \sim \text{Beta}(4, 8)$ (around 10% imported cases are misclassified as local and around 33.3% local cases are misclassified as imported), 3) $\alpha_0 \sim \text{Beta}(4, 8)$ and $\alpha_1 \sim \text{Beta}(2, 18)$ (around 33.3% imported cases are misclassified as local and around 10% local cases are misclassified as imported). The bands are the 95% credible intervals.

by our estimators is possible when nontrivial error is present. And our application to epidemics in Hong Kong and Victoria shows that the gains offered by our approach over presenting the noisy local instantaneous reproduction number can be pronounced.

We have shown examples on a state/province level, but our method could be useful for cities, or more local settings, such as a university trying to determine if there is substantial local transmission occurring. Our approach requires daily numbers of local and imported cases, serial interval, and contact tracing data or other data to provide

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

adequate information to estimate the misidentification rates.

We have pursued a Bayesian approach to the problem of estimating the local instantaneous reproduction number. The credible intervals are relatively wide when the number of cases is low. To improve the performance at low case incidence, Kalman filtering is a natural approach. Estimating the time-vary reproduction number by Kalman filtering is an emerging topic. For instance, [24] constructed a recursive Bayesian smoother for estimating the effective reproduction number from the incidence of an infectious disease in real time and retrospectively. However, one typically does not distinguish between local and imported cases in this setting.

The identification errors are informed by contact tracing survey data in our approach. If the data from the survey is categorical (e.g., we ask people where they were infected and attach some qualitative measure of our confidences that we think they are local cases), we can transform them into numerical values. For example, [25] proposed a method that converts categorical variables to numerical data for Gaussian distribution. We could modify the method to convert categorical variables to Beta distributed data. If the survey data is unavailable, using genomic data is a natural alternative. Genomic surveillance has been used to detect transmission clusters and to provide information on the possible source of individual cases [26–31].

We have showed the results of retrospective estimation. And it is computationally feasible to run MCMC on each day to obtain real time estimators; it takes about 5 minutes for the MCMC chain of 10,000 samples. To reduce the computational cost, one approach is adaptive MCMC methods [32, 33], which use the covariance structure of the posterior distribution to design proposal distributions. Other methods include stochastic Newton [34] and Riemannian manifold MCMC [35], which construct efficient proposals by local derivative information.

Data Accessibility

No primary data are used in this paper. Secondary data sources are taken from [20, 21]. These data and the code necessary to reproduce the results in this paper are available at `https://github.com/KolaczykResearch/EstimLocalRt`.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Funding

This work was supported in part by ARO award W911NF1810237. This work was also supported by National Institutes of Health, R01 GM122878.

References

1. You C, Deng Y, Hu W, Sun J, Lin Q, Zhou F, et al. Estimation of the time-varying reproduction number of COVID-19 outbreak in China. International Journal of Hygiene and Environmental Health. 2020; p. 113555. doi:10.1101/2020.02.08.20021253.

2. Li Y, Campbell H, Kulkarni D, Harpur A, Nundy M, Wang X, et al. The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: a modelling study across 131 countries. The Lancet Infectious Diseases. 2020;doi:10.1016/s1473-3099(20)30785-4.

3. Rubin D, Huang J, Fisher BT, Gasparrini A, Tam V, Song L, et al. Association of social distancing, population density, and temperature with the instantaneous reproduction number of SARS-CoV-2 in counties across the United States. JAMA network open. 2020;3(7):e2016099–e2016099. doi:10.1001/jamanetworkopen.2020.16099.

4. Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Research. 2020;5(112):112. doi:10.12688/wellcomeopenres.16006.1.

5. Thompson RN, Stockwin JE, van Gaalen RD, Polonsky JA, Kamvar ZN, Demarsh PA, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics. 2019;doi:10.1016/j.epidem.2019.100356.

6. Cori A, Ferguson NM, Fraser C, Cauchemez S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. Am J Epi. 2013;178(9). doi:10.1093/aje/kwt133.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

7. Chong KC, Cheng W, Zhao S, Ling F, Mohammad KN, Wang M, et al. Transmissibility of coronavirus disease 2019 in Chinese cities with different dynamics of imported cases. PeerJ. 2020;8:e10350. doi:10.7717/peerj.10350.

8. Arroyo Marioli F, Bullano F, Kučinskas S, Rondón-Moreno C. Tracking R of COVID-19: A New Real-Time Estimation Using the Kalman Filter. Available at SSRN 3581633. 2020;doi:10.1101/2020.04.19.20071886.

9. Reich N, Lessler J, Cummings D, Brookmeyer R. Estimating incubation period distributions with coarse data. Stat Med. 2009;28(22).

10. Ma Y, Jenkins HE, Sebastiani P, Ellner JJ, Jones-Lòpez EC, Dietze R, et al. Using cure models to estimate the serial interval of tuberculosis with limited follow-up. Am J Epidemiol. 2020;189(11):1421–1426. doi:https://doi.org/10.1093/aje/kwaa090.

11. Miller DA, Talley BL, Lips KR, Campbell Grant EH. Estimating patterns and drivers of infection prevalence and intensity when detection is imperfect and sampling error occurs. Methods in Ecology and Evolution. 2012;3(5):850–859. doi:10.1111/j.2041-210x.2012.00216.x.

12. McClintock BT, Nichols JD, Bailey LL, MacKenzie DI, Kendall WL, Franklin AB. Seeking a second opinion: uncertainty in disease ecology. Ecology letters. 2010;13(6):659–674. doi:10.1111/j.1461-0248.2010.01472.x.

13. Cui N, Chen Y, Small DS. Modeling parasite infection dynamics when there is heterogeneity and imperfect detectability. Biometrics. 2013;69(3):683–692. doi:10.1111/biom.12050.

14. Li T, White LF. Bayesian back-calculation and nowcasting for line list data during the COVID-19 pandemic. medRxiv. 2020;doi:10.1101/2020.12.08.20238154.

15. Fraser C. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. PlosOne. 2007;2(8). doi:10.1371/.

16. de Valpine P, Turek D, Paciorek C, Anderson-Bergman C, Temple Lang D, Bodik R. Programming with models: writing statistical algorithms for general model

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

structures with NIMBLE. Journal of Computational and Graphical Statistics. 2017;26:403–413. doi:10.1080/10618600.2016.1172487.

17. de Valpine P, Paciorek C, Turek D, Michaud N, Anderson-Bergman C, Obermeyer F, et al.. NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling; 2020. Available from: `https://cran.r-project.org/package=nimble`.

18. de Valpine P, Paciorek C, Turek D, Michaud N, Anderson-Bergman C, Obermeyer F, et al.. NIMBLE User Manual; 2020. Available from: `https://r-nimble.org`.

19. Neal RM. Slice sampling. Annals of statistics. 2003; p. 705–741. doi:10.1214/aos/1056562461.

20. Adam DC, Wu P, Wong JY, Lau EHY, Tsang TK, Cauchemez S, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nature Medicine. 2020;26(11):1714–1719. doi:10.1038/s41591-020-1092-0.

21. Seemann T, Lane C, Sherry N, Duchene S, da Silva AG, Caly L, et al. Tracking the COVID-19 pandemic in Australia using genomics. medRxiv. 2020;doi:10.1101/2020.05.12.20099929.

22. Kerr CC, Stuart RM, Mistry D, Abeysuriya RG, Hart G, Rosenfeld K, et al. Covasim: an agent-based model of COVID-19 dynamics and interventions. medRxiv. 2020;doi:10.1101/2020.05.10.20097469.

23. Cori A, Kamvar ZN, Stockwin JE, Jombart T, Thompson RN, Dahlqwist E. EpiEstim; 2020.

24. Parag KV. Improved estimation of time-varying reproduction numbers at low case incidence and between epidemic waves. medRxiv. 2020;doi:10.1101/2020.09.14.20194589.

25. Patki N, Wedge R, Veeramachaneni K. The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE; 2016. p. 399–410. Available from: `https://doi.org/10.1109/dsaa.2016.49`.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

26. Leavitt SV, Lee RS, Sebastiani P, Horsburgh CR, Jenkins HE, White LF. Estimating the relative probability of direct transmission between infectious disease patients. International journal of epidemiology. 2020;doi:https://doi.org/10.1093/ije/dyaa031.

27. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of healthcare associated COVID-19: a prospective genomic surveillance study. The Lancet infectious diseases. 2020;20(11):1263–1272. doi:10.1016/s1473-3099(20)30562-4.

28. Deng X, Gu W, Federman S, du Plessis L, Pybus OG, Faria N, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science. 2020;.

29. Poon AF, Gustafson R, Daly P, Zerr L, Demlow SE, Wong J, et al. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. The lancet HIV. 2016;3(5):e231–e238. doi:10.1016/s2352-3018(16)00046-1.

30. Sansone M, Andersson M, Gustavsson L, Andersson LM, Nordén R, Westin J. Extensive hospital in-ward clustering revealed by molecular characterization of influenza A virus infection. Clinical Infectious Diseases. 2020;doi:10.1093/cid/ciaa108.

31. Peters PJ, Pontones P, Hoover KW, Patel MR, Galang RR, Shields J, et al. HIV infection linked to injection use of oxymorphone in Indiana, 2014–2015. New England Journal of Medicine. 2016;375(3):229–239. doi:10.1056/nejmoa1515195.

32. Haario H, Saksman E, Tamminen J, et al. An adaptive Metropolis algorithm. Bernoulli. 2001;7(2):223–242. doi:10.2307/3318737.

33. Roberts GO, Rosenthal JS. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. Journal of applied probability. 2007;44(2):458–475. doi:10.1017/s0021900200117954.

34. Martin J, Wilcox LC, Burstedde C, Ghattas O. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic

medRxiv preprint doi: https://doi.org/10.1101/2021.04.23.21255958; this version posted April 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

inversion. SIAM Journal on Scientific Computing. 2012;34(3):A1460–A1487. doi:10.1137/110845598.

35. Girolami M, Calderhead B. Riemann manifold langevin and hamiltonian monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011;73(2):123–214. doi:10.1111/j.1467-9868.2010.00765.x.