arXiv:2206.05161v2 [stat.ME] 6 Mar 2023

Approximating optimal SMC proposal distributions in individual-based epidemic models

Lorenzo Rimella, Christopher Jewell and Paul Fearnhead

March 7, 2023

Abstract

Many epidemic models are naturally defined as individual-based models: where we track the state of each individual within a susceptible population. Inference for individual-based models is challenging due to the high-dimensional state-space of such models, which increases exponentially with population size. We consider sequential Monte Carlo algorithms for inference for individual-based epidemic models where we make direct observations of the state of a sample of individuals. Standard implementations, such as the bootstrap filter or the auxiliary particle filter are inefficient due to mismatch between the proposal distribution of the state and future observations. We develop new efficient proposal distributions that take account of future observations, leveraging the properties that (i) we can analytically calculate the optimal proposal distribution for a single individual given future observations and the future infection rate of that individual; and (ii) the dynamics of individuals are independent if we condition on their infection rates. Thus we construct estimates of the future infection rate for each individual and then use an independent proposal for the state of each individual given this estimate. Empirical results show order of magnitude improvement in efficiency of the sequential Monte Carlo sampler for both SIS and SEIR models.

1 Introduction

The use of dynamical disease transmission models to inform disease control policy has increased throughout the 21st century for both human and livestock outbreaks, for example, SARS and H1N1 pandemic influenza in humans (Zhou et al., 2004), avian influenza in poultry (Van der Goot et al., 2005), and foot-and-mouth disease in cloven-hoofed livestock (Zhou et al., 2004; Jewell et al., 2009). Most recently, they have been central in informing national-level decisions on social distancing and vaccination strategies for the SARS-CoV-2 pandemic (Brooks-Pollock et al., 2021; Funk et al., 2020). Besides outbreaks, such models are also useful for studying the dynamics of endemic diseases, with the ability to explain random fluctuations around an otherwise stable case incidence, particularly in highly heterogeneous populations (Britton, 2010).

In essence, disease transmission models belong to the class of state transition models, described by a directed (though not necessarily acyclic) graph. For example, the SIS model proposes individuals as existing as "susceptible" or "infected", and individuals are allowed to transition from either state to the other. In a stochastic setting, it is natural to assume that an individual in the population experiences a hazard rate of progressing from some source state to a destination state. This setup has particular relevance for the case when transition hazard rates depend on the individuals' characteristics as well as the characteristics of their relationship with each other. Many applications demand individual-level granularity, particularly when observations are of specific individuals or where disease interventions are targeted to particular individuals (Chapman et al., 2020; Jewell et al., 2009; Cocker et al., 2022).

Inference for such models is, however, challenging due to the presence of partial- or total-censoring of transition events, for which the state-space increases exponentially with population size. For example, in an SIS model, we may have noisy observations of which individuals exist in either the S or I states at particular times, but no direct observation of when state transitions occur.

Following Rimella et al. (2022), we consider sequential Monte Carlo (SMC) methods for inference for such models. We show that standard implementations of SMC (Gordon et al., 1993; Pitt and Shephard, 1999) are inefficient for these individual-based epidemic models. In particular, they struggle to propose states for all individuals that will be consistent with future obser-

vations. Ju et al. (2021) consider how to improve the efficiency of SMC for individual-based epidemic models, but they consider observations of e.g. the number of infected individuals, and their approach does not obviously apply to the observation models we consider.

To improve SMC, we develop a novel proposal distribution that takes account of future observations. The key idea is based on two properties of the dynamics of individual-based epidemic models. First, calculating the conditional distribution of the state of a single individual, given future observations and the future infection rate of the individual is tractable. This can be calculated using standard recursions for finite-state hidden Markov models Rabiner and Juang (1986) together with the fact that the state-space for a single individual is small (e.g. 2 for an SIS model or 4 for an SEIR model). Second, there is a form of conditional independence across individuals: if we condition on the future infection rates for each individual then the dynamics of the state for individuals are independent of each other. In the models, we consider the infection rate for each individual just depends on the total number of infectious individuals. Thus we can use ideas from Whiteley and Rimella (2021) to estimate the future number of infectious individuals. Then conditioning on this estimate, and the corresponding infection rates for each individual, we have a proposal distribution that is independent across individuals, and for each individual is equal to the true conditional distribution of the state given the estimated future infection rates and observations for that individual.

The computational cost of using this proposal is proportional to the number of time-steps at which we have future observations. In practice, we can implement such a proposal distribution just conditioning on future observations over a suitable time window. We show empirically that using this proposal distribution can lead to an order of magnitude improvement in Monte Carlo efficiency, even after accounting for the increased computational cost.

2 Preliminaries

We use bold lowercase letters for vectors, e.g. $\mathbf{a}$, and bold uppercase letters for matrices, e.g. $\mathbf{A}$. We use $\mathbf{A}^{(i,j)}$ for the (i,j)-th element of $\mathbf{A}$ and we use $\mathbf{A}^{(i,\bullet)}$ (or $\mathbf{A}^{(\bullet,j)}$) to represent the column vector given by the (i)-th column (or the (j)-th row) of matrix $\mathbf{A}$. With $\circ$ and $/$ we denote the elementwise product and ratio between vectors or matrices. $\mathbf{1}_M$ denotes the M-dimensional vector

of ones. Given $t, s \in \mathbb{N}$ with $t > s$ we use $[s : t]$ for the set $\{s, \dots, t\}$, which also applies on indexing as a shorthand, e.g. for $t \in \mathbb{N}$ we use $y_{[1:t]}$ for $\{y_1, \dots, y_t\}$. The notation for the main probability distributions is reported in Table 4.

Distribution	Categorical	Bernoulli	Binomial	Gaussian	Uniform	Multinomial
Notation	$\mathcal{C}at_M(i\|\mathbf{p})$	$\mathcal{B}e(i\|q)$	$\mathcal{B}in(i\|N, q)$	$\mathcal{N}(a\|\mu, \sigma^2)$	$\mathcal{U}nif(q\|a, b)$	$\mathcal{M}ult(\mathbf{c}\|N, \mathbf{p})$

Table 1: Notation table for probability mass and density functions.

3 Model

3.1 Individual-based epidemic models

In this article we consider individual-based models defined by: the number of compartments M, the population size N, the initial probability of an individual being assigned to a compartment $(\mathbf{p}_{n,0})_{n\in[1:N]}$ and the probability of an individual to transition from one compartment to the other $(\mathbf{K}_{n,\bullet})_{n\in[1:N]}$, where the stochastic transition matrix $\mathbf{K}_{n,\bullet}$ is defined as a function of an M-dimensional vector $\mathbf{c}$, i.e. $\mathbf{c} \to \mathbf{K}_{n,\mathbf{c}}$. In practice, $\mathbf{c}^{(i)}$ is the number of individuals in compartment i, and so the transition matrix $\mathbf{K}_{n,\mathbf{c}}$ depends on the compartments' state only, however, more general versions are possible and briefly discussed in Section 6 (e.g. spatial models). We use $(\mathbf{x}_t)_{t\geq 0}$ for the population state and $(\mathbf{c}_t)_{t\geq 0}$ for the compartments' state, following:

$$\text{Time 0: } \mathbf{x}_0^{(n)} \sim \mathcal{C}at_M(\bullet|\mathbf{p}_{n,0}) \text{ for } n \in [1:N] \text{ and } \mathbf{c}_0^{(i)} = \sum_{n=1}^N \mathbb{I}_{\mathbf{x}_0^{(n)}}(i) \text{ for } i \in [1:M];$$

$$\text{Time } t\text{: } \mathbf{x}_t^{(n)}|\mathbf{x}_{t-1} \sim \mathcal{C}at_M\left(\bullet\middle|\mathbf{K}_{n,\mathbf{c}_{t-1}}^{(\mathbf{x}_{t-1}^{(n)},\bullet)}\right) \text{ for } n \in [1:N] \text{ and } \mathbf{c}_t^{(i)} = \sum_{n=1}^N \mathbb{I}_{\mathbf{x}_t^{(n)}}(i) \text{ for } i \in [1:M].$$

SIS example We can make the SIS model heterogeneous by following the construction in Ju et al. (2021). Suppose that we have $d \in \mathbb{N}$ covariates for each individual, we can then define $(\mathbf{w}_n)_{n\in[1:N]}$ as the collection of d-dimensional vectors gathering the individual-specific covariates, from which

we can compute for $n \in [1:N]$:

$$\mathbf{p}_{n,0} = \begin{bmatrix} 1 - \frac{1}{1+\exp(-\beta_0^{\mathrm{T}} w_n)} \\ \frac{1}{1+\exp(-\beta_0^{\mathrm{T}} w_n)} \end{bmatrix}, \quad \mathbf{K}_{n,c} = \begin{bmatrix} 1 - \frac{1}{1+\exp(-\beta_\lambda^{\mathrm{T}} w_n)} \frac{\mathbf{c}^{(2)}}{N} & \frac{1}{1+\exp(-\beta_\lambda^{\mathrm{T}} w_n)} \frac{\mathbf{c}^{(2)}}{N} \\ \frac{1}{1+\exp(-\beta_\gamma^{\mathrm{T}} w_n)} & 1 - \frac{1}{1+\exp(-\beta_\gamma^{\mathrm{T}} w_n)} \end{bmatrix}$$

with $\beta_0 \in \mathbb{R}^d$ and $\beta_\lambda, \beta_\gamma \in \mathbb{R}^d$. In this model we have individual-specific probabilities of infection and recovery.

3.2 Observation model

The observation process is denoted by $(\mathbf{y}_t)_{t\geq 1}$ and given $(\mathbf{q}_{n,t})_{n\in[1:N],t\geq 1}$ with $\mathbf{q}_{n,t} \in [0,1]^M$ we generate observations per each time step t as follows:

$$\mathbf{y}_t^{(n)} = \mathbf{x}_t^{(n)} \mathbf{r}_t^{(n)} \text{ with } \mathbf{r}_t^{(n)} \sim \mathcal{B}e\left(\bullet \middle| \mathbf{q}_{n,t}^{(\mathbf{x}_t^{(n)})}\right) \text{ for } n \in [1:N], \tag{1}$$

which we refer to as the "granular observations model". Note that $\mathbf{y}_t^{(n)} \in [0:M]$ meaning that we either report the state of individual n as it is ($\mathbf{y}_t^{(n)} = \mathbf{x}_t^{(n)}$) or we do not report it at all ($\mathbf{y}_t^{(n)} = 0$). This model includes observations from random samples of the population, where each component of $\mathbf{q}_{n,t}$ is the same and equal to the probability that individual n is included in the sample at time t, as well as situations where observations are preferentially made for certain states (such as observations being of infected farms for foot-and-mouth disease). To simplify the notation and derivations, in this paper we focus on individual homogeneous reporting rates, i.e. $\mathbf{q}_{n,t} = \mathbf{q}_t$, and under this assumption $\sum_{n\in[1:N]} \mathbb{I}_{\mathbf{y}_t^{(n)}}(i) \sim \mathcal{B}in(\bullet|\mathbf{c}_t^{(i)}, \mathbf{q}_t^{(i)})$ for any $i \in [1:M]$, which recover the binomial observation model (Whiteley and Rimella, 2021; Ju et al., 2021).

SIS example Given $\mathbf{q}_t \in [0,1]^2$, $\mathbf{q}_t^{(1)}$ is the probability of reporting a susceptible, while $\mathbf{q}_t^{(2)}$ is the probability of reporting an infected.

3.3 Inference in individual-based models with granular observations

In epidemiology, we are interested in inferring both the unknown state of the population $\mathbf{x}_t$ and the parameters of the epidemic θ. Given the time

horizon t, the individual-based model with granular observation $(\mathbf{x}_s, \mathbf{y}_s)_{s\in[1:t]}$ is by construction a hidden Markov model (HMM). We can hence compute filtering distribution $p(\mathbf{x}_s|\mathbf{y}_{[1:s]}, \theta)$ and marginal likelihood $p(\mathbf{y}_{[1:s]}|\theta)$ with the forward algorithm (Rabiner and Juang, 1986). The parameters can be then inferred through, for example, the EM algorithm (Yang et al., 2017).

The forward algorithm requires marginalizing over the whole state-space, making it unfeasible for our individual-based model, where marginalizations are $\mathcal{O}(M^N)$. Alternatively, Sequential Monte Carlo (SMC) algorithms can be employed to obtain particle approximations of $p(\mathbf{x}_s|\mathbf{y}_{[1:s]}, \theta)$ and $p(\mathbf{y}_{[1:s]}|\theta)$ (Ionides et al., 2006; Kucharski et al., 2020) at a cost that is linear in the number of particles and time horizon. Given a number of particles $P \in \mathbb{N}$, at each time step s an SMC algorithm proposes instances $(\mathbf{x}_s^p)_{p\in[1:P]}$ of the latent process $(\mathbf{x}_s)_{s\in[1:t]}$ through the proposal distribution $q(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[1:t]})$, with $q(\mathbf{x}_0|\mathbf{x}_{-1}, \mathbf{y}_{[1:t]}) \coloneqq q(\mathbf{x}_0|\mathbf{y}_{[1:t]})$ and $q(\mathbf{x}_0|\mathbf{y}_{[1:t]})$ proposal distribution at time $s = 0$, and it assigns weights $(w_s^p)_{p\in[1:P]}$ to the particles to produce an importance sample that approximates the filtering distribution. Before moving to the next step, the algorithm uses a resampling scheme $r_s(i)$, a distribution over the particles' indexes $[1 : P]$, to discard low-weight particles. At the end of the procedure, particle estimates of the filtering distribution, $p(\mathbf{x}_s|\mathbf{y}_{[1:s]}, \theta) \approx (\sum_{\tilde{p}\in[1:P]} w_s^p)^{-1} \sum_{p\in[1:P]} w_s^p \delta_{\mathbf{x}_s^p}(\mathbf{x}_s)$, and the marginal likelihood, $p(\mathbf{y}_{[1:t]}|\theta) \approx \prod_{s\in[1:t]} \sum_{p\in[1:P]} w_s^p$, are generated.

The performance of SMC algorithms heavily depends on the proposal distributions $(q(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[1:t]}))_{s\in[0:t]}$ and the resampling scheme $(r_s(i))_{s\in[0:t]}$, incautious choices of these quantities might lead to high variance of the marginal likelihood estimator, particles/weights degeneracy and even observation mismatch, which might cause the failure of the algorithm. The Bootstrap Particle Filter (BPF) (Gordon et al., 1993; Candy, 2007) proposes new particles through the transition kernel and it resamples according to the current weights, i.e. $q(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[1:t]}) = p(\mathbf{x}_s|\mathbf{x}_{s-1}, \theta)$, with $q(\mathbf{x}_0|\mathbf{y}_{[1:t]}) = p(\mathbf{x}_0|\theta)$, and $r_s(i) = \mathcal{C}at_P\left(i|\left[w_s^1, \ldots, w_s^P\right]\right)$. BPF is known to perform poorly in high-dimensional scenarios (Bickel et al., 2008) and with informative observation, especially when simulated particles have to match certain paths. An easy fix is to include the information from the current observations in the proposal distributions, to avoid mismatch at the current time step when proposing new particles. The resulting algorithm is called the auxiliary particle filter (APF) (Pitt and Shephard, 1999; Carpenter et al., 1999; Johansen and Doucet, 2008) and it arises by picking $q(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[1:t]}) = p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_s, \theta)$,

with $q(\mathbf{x}_0|\mathbf{y}_{[1:t]}) = p(\mathbf{x}_0|\theta)$, and $r_s(i) = \mathcal{C}at_P\left(i|\left[w_s^1, \ldots, w_s^P\right]\right)$. We illustrate graphically in Figure 1 the comparison between BPF and APF in an individual-based model. The BPF fails after three iterations because the proposed particle mismatch the observed state for individuals 1 and 3, indeed we observe $\mathbf{y}_3^{(1)} = 2$ and $\mathbf{y}_3^{(3)} = 2$, but the BPF proposes $(\mathbf{x}_3^p)^{(1)} = 1$ and $(\mathbf{x}_3^p)^{(3)} = 1$ (green lines). On the contrary, the APF is able to propose particles that are constrained to match the observation, because it includes the current data in the proposal. However, the APF's proposal is still inefficient as it does not take into account future observations. In Figure 1 this is seen by it tending to propose a switch to the infected state immediately before the observation of an infected individual, whereas often an individual becomes infected one or more time-steps earlier. For more complicated models, such as the SEIR model we consider in Section 5.2, the APF can also suffer from mismatch, as the transition to an observed state may not be possible for the current state of a particle.

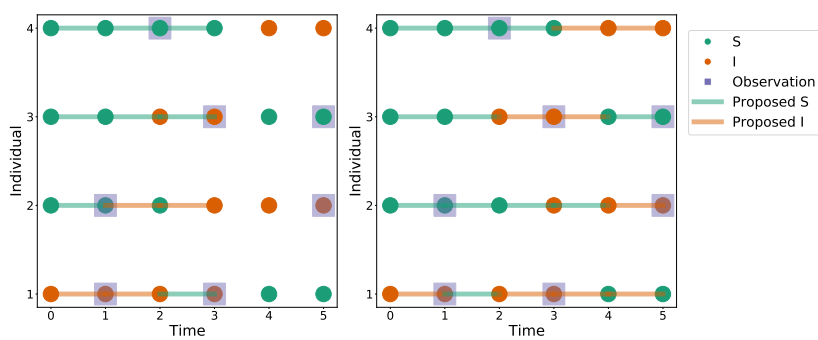

Figure 1: Illustration of BPF (left) and APF (right), in an SIS scenario. Colored dots show the state of each individual, with green for susceptible and red for infected. Dots in grey squares are observations. Horizontal lines from $s-1$ to s are used for the proposed states in s.

In the next section, we show how to build for any $s \in [0:t]$ an approximation of $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta)$ for the individual-based model with granular observations. Given that computing $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta)$ requires $(\mathbf{c}_{\tilde{s}})_{\tilde{s}\in[s:t-1]}$, the main idea consists of approximating $(\mathbf{c}_{\tilde{s}})_{\tilde{s}\in[s:t-1]}$ with the expectation of a precomputed multinomial distribution (Whiteley and Rimella, 2021) and to propagate backward the observation $\mathbf{y}_{[s:t]}$ to inform the proposal in s.

4 Optimal proposal distributions for individual-based models

The optimal proposal for an SMC is $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta)$ and it can be compute recursively:

Time t: $p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta) = \sum_{\mathbf{x}_t \in [1:M]^N} p(\mathbf{y}_t|\mathbf{x}_t, \theta) p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta)$ and
$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t, \theta) = \frac{p(\mathbf{y}_t|\mathbf{x}_t,\theta)p(\mathbf{x}_t|\mathbf{x}_{t-1},\theta)}{p(\mathbf{y}_t|\mathbf{x}_{t-1},\theta)}$;

Time s: $p(\mathbf{y}_{[s:t]}|\mathbf{x}_{s-1}, \theta) = \sum_{\mathbf{x}_s \in [1:M]^N} p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s, \theta) p(\mathbf{y}_s|\mathbf{x}_s, \theta) p(\mathbf{x}_s|\mathbf{x}_{s-1}, \theta)$ and
$p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta) = \frac{p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s,\theta)p(\mathbf{y}_s|\mathbf{x}_s,\theta)p(\mathbf{x}_s|\mathbf{x}_{s-1},\theta)}{p(\mathbf{y}_{[s:t]}|\mathbf{x}_{s-1},\theta)}$;

Time 0: $p(\mathbf{y}_{[1:t]}, \theta) = \sum_{\mathbf{x}_0 \in [1:M]^N} p(\mathbf{y}_{[1:t]}|\mathbf{x}_0, \theta) p(\mathbf{x}_0, \theta)$ and
$p(\mathbf{x}_0|\mathbf{y}_{[1:t]}, \theta) = \frac{p(\mathbf{y}_{[1:t]}|\mathbf{x}_0,\theta)p(\mathbf{x}_0,\theta)}{p(\mathbf{y}_{[1:t]}|\theta)}$.

See Fearnhead (2008) for a review of the optimal proposal for importance sampling, Chopin et al. (2020) for a discussion on optimal proposal for Particle Filters and Whiteley and Lee (2014) for a more technical discussion.

A marginalization over the whole state-space is required resulting in a computational cost of $\mathcal{O}(M^N)$ per each step. Observe that at the beginning of the recursion we can exploit the factorization over the individuals at time t, components of $\mathbf{x}_t$, of the transition kernel and emission distribution to reduce the computational cost of the marginalization to $\mathcal{O}(NM)$:

$$p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta) = \prod_{n \in [1:N]} \sum_{\mathbf{x}_t^{(n)} \in [1:M]} \mathbf{K}_{n,\mathbf{c}_{t-1}}^{(\mathbf{x}_{t-1}^{(n)}, \mathbf{x}_t^{(n)})} \left(\mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(\mathbf{x}_t^{(n)})} \left(1 - \mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(0)}. \tag{2}$$

Notice that $\mathbf{y}_t^{(n)}$ is not conditionally independent given $\mathbf{x}_{t-1}^{(n)}$ because of the dependence of the transition kernel on the compartments' state $\mathbf{c}_{t-1}$. This breaks the computational trick because we cannot express $p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta)$ as a product over the individuals at time $t-1$, and so the cheap marginalization has to be repeated for each state of $\mathbf{x}_{t-1}$, leading to $\mathcal{O}(NM^{N+1})$. However, $\mathbf{y}_t^{(n)}$ is conditionally independent given $\mathbf{x}_{t-1}^{(n)}$ and $\mathbf{c}_{t-1}$, meaning that if an estimate of $\mathbf{c}_{t-1}$ is available a priori the factorization is preserved and the same trick can be iterated in the next time steps.

4.1 A priori estimates of the compartments' states

Whiteley and Rimella (2021) proposes an efficient way to approximate the smoothing distribution $p(\mathbf{c}_s|\mathbf{y}_{[1:t]}, \theta)$ with a multinomial distribution $\mathcal{M}ult(\mathbf{c}_s|N, \mathbf{m}_{s|t})$ whose parameters are computed recursively with a forward and a backward step through the data at a computational cost $\mathcal{O}(tM^3)$. In the multinomial approximation, there are two key assumptions: the homogeneity of the individuals and a binomial observation model of the form $\mathcal{B}in(\bullet|\mathbf{c}_s^{(i)}, \mathbf{q}_s^{(i)})$. We can recover homogeneity in the individual-based model with granular observation by defining the mean initial distribution $\bar{\mathbf{p}}_{n,0}$ and the mean transition kernel $\bar{\mathbf{K}}_{\mathbf{c}_s}$:

$$\bar{\mathbf{p}}_{n,0}^{(i)} := \frac{1}{N} \sum_{n\in[1:N]} \mathbf{p}_{n,0}^{(i)}, \text{ for } i \in [1:M],$$

$$\bar{\mathbf{K}}_{\mathbf{c}_s}^{(i,j)} := \frac{1}{N} \sum_{n\in[1:N]} \mathbf{K}_{n,\mathbf{c}_s}^{(i,j)}, \text{ for } i,j \in [1:M].$$

We remark that recovering homogeneity by approximating the individuals' transition kernel with an average is also a key step in Ju et al. (2021), where the transition probabilities are approximated by averaging over the individuals to avoid an exponential computational cost in the population size.

We already have $\sum_{n\in[1:N]} \mathbb{I}_{\mathbf{y}_s^{(n)}}(i) \sim \mathcal{B}in(\bullet|\mathbf{c}_s^{(i)}, \mathbf{q}_s^{(i)})$ for $i \in [1:M]$, from which we can define the cumulative observations per each compartment as the vector $\mathbf{o}_s$ with components $\mathbf{o}_s^{(i)} := \sum_{n\in[1:N]} \mathbb{I}_{\mathbf{y}_s^{(n)}}(i)$.

Using the aforementioned approximate dynamic and observation model, the multinomial approximation in Whiteley and Rimella (2021) scans the data forward and backward and computes multinomial approximations of the filtering and smoothing distribution (the full algorithm is reported in the appendix). The forward pass consists of a prediction step and an update step preserving the multinomial form, precisely, starting from $\mathbf{m}_{0|0} := \bar{\mathbf{p}}_{n,0}$, we have:

$$\mathbf{m}_{s-1|s} := \left(\mathbf{m}_{s-1|s-1}^{\mathrm{T}} \bar{\mathbf{K}}_{\mathbf{m}_{s-1}}\right)^{\mathrm{T}}, \quad \mathbf{m}_{s|s} := \frac{\mathbf{o}_s}{N} + \left(1 - \frac{\mathbf{1}_M^{\mathrm{T}} \mathbf{o}_s}{N}\right) \frac{\mathbf{m}_{s-1|s} \circ (\mathbf{1}_M - \mathbf{q}_s)}{1 - \mathbf{m}_{s-1|s}^{\mathrm{T}} \mathbf{q}_s},$$

which gives an approximation for the filtering distribution $p(\mathbf{c}_s|\mathbf{y}_{[1:s]}) \approx \mathcal{M}ulti(\mathbf{c}_s|N, \mathbf{m}_{s|s})$. The backward pass implements a reverse kernel and applies it backward:

$$\mathbf{L}_s := \left\{ \left[(\mathbf{m}_{s|t}\mathbf{1}_M^{\mathrm{T}}) \circ \bar{\mathbf{K}}_{\mathbf{m}_s}\right] / \left[\mathbf{1}_M(\mathbf{m}_{s|t}^{\mathrm{T}} \bar{\mathbf{K}}_{\mathbf{m}_s})\right] \right\}^{\mathrm{T}}, \quad \mathbf{m}_{s|t} := \left(\mathbf{m}_{s+1|T}^{\mathrm{T}} \mathbf{L}_s\right)^{\mathrm{T}},$$

outputting the M-dimensional probability vector $\mathbf{m}_{s|t}$ and so approximating the smoothing distribution with $p(\mathbf{c}_s|\mathbf{y}_{[1:t]},\theta) \approx \mathcal{M}ulti(\mathbf{c}_s|N,\mathbf{m}_{s|t})$. Given the multinomial approximations we can approximate the compartments' state with:

$$\mathbf{c}_s \approx \mathbb{E}_{\mathcal{M}ult(\mathbf{c}_s|N,\mathbf{m}_{s|t})}(\mathbf{c}_s) = N\mathbf{m}_{s|t}.$$

We have imposed a restriction on the emission distribution by assuming a uniform reporting probability for all individuals. However, our approach can be extended to accommodate a more general scenario where $\mathbf{q}_{t,n}$ varies with n. To do so, we can compute the mean reporting rate $\bar{\mathbf{q}}_t := (N)^{-1}\sum_{n\in[1:N]}\mathbf{q}_{t,n}$ when running Whiteley and Rimella (2021) and then substitute back $\mathbf{q}_{t,n}$ when computing the approximation to the optimal proposal.

4.2 Approximate optimal proposals for individual-based models

Conditioning on $\mathbf{c}_{\tilde{s}} = N\mathbf{m}_{\tilde{s}|t}$ for $\tilde{s} \in [s:t]$ makes the individuals evolve independently from each other and so it allows an analytical computation of $p(\mathbf{y}_{[s:t]}|\mathbf{x}_{s-1},\theta)$ at a cost $\mathcal{O}(NM)$. Starting again from (C):

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{x}_{t-1},\theta) &\approx \prod_{n\in[1:N]} \sum_{\mathbf{x}_t^{(n)}\in[1:M]} \mathbf{K}_{n,N\mathbf{m}_{t-1|t}}^{(\mathbf{x}_{t-1}^{(n)},\mathbf{x}_t^{(n)})} \left(\mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(\mathbf{x}_t^{(n)})} \left(1-\mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(0)} \\ &=: \prod_{n\in[1:N]} \boldsymbol{\xi}_{n,t-1}^{(\mathbf{x}_{t-1}^{(n)})}, \end{aligned}$$

where we define the quantities $\boldsymbol{\xi}_{n,t-1}$ for each individual n as the approximate probability of observing the future observation $\mathbf{y}_t^{(n)}$ given the state at time $t-1$. We can then follow a similar argument and approximate $p(\mathbf{y}_{[s:t]}|\mathbf{x}_{s-1},\theta)$ as follows:

$$\begin{aligned} p(\mathbf{y}_{[s:t]}|\mathbf{x}_{s-1},\theta) &\approx \prod_{n\in[1:N]} \sum_{\mathbf{x}_s^{(n)}\in[1:M]} \boldsymbol{\xi}_{n,s}^{(\mathbf{x}_s^{(n)})} \mathbf{K}_{n,N\mathbf{m}_{s-1|t}}^{(\mathbf{x}_{s-1}^{(n)},\mathbf{x}_s^{(n)})} \left(\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(\mathbf{x}_s^{(n)})} \left(1-\mathbf{q}_s^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(0)} \\ &=: \prod_{n\in[1:N]} \boldsymbol{\xi}_{n,s-1}^{(\mathbf{x}_{s-1}^{(n)})}, \end{aligned}$$

where $\boldsymbol{\xi}_{n,s-1}$ is the approximate probability for each individual n of observing the future observation $\mathbf{y}_{[s:t]}$ given the state at time $s-1$. Note that the marginalization is repeated for all the states of $\mathbf{x}_{s-1}^{(n)}$ and not $\mathbf{x}_{s-1}$, which reduces the cost from $\mathcal{O}(NM^{N+1})$ to $\mathcal{O}(NM^2)$. We can now build our proposal distribution for SMC and approximate $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta)$ as:

$$p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta) \approx \prod_{n\in[1:N]} \frac{\boldsymbol{\xi}_{n,s}^{(\mathbf{x}_s^{(n)})} \mathbf{K}_{n,\mathbf{c}_{s-1}}^{(\mathbf{x}_{s-1}^{(n)},\mathbf{x}_s^{(n)})} \left(\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(\mathbf{x}_s^{(n)})} \left(1-\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(0)}}{\tilde{\boldsymbol{\xi}}_{n,s}^{(\mathbf{x}_{s-1}^{(n)})}},$$

$$p(\mathbf{x}_0|\mathbf{y}_{[1:t]}, \theta) \approx \prod_{n\in[1:N]} \frac{\boldsymbol{\xi}_{n,0}^{(\mathbf{x}_0^{(n)})} \mathbf{p}_{n,0}^{(\mathbf{x}_0^{(n)})}}{\tilde{\xi}_{n,0}},$$

for $s \in [1:t]$ and with:

$$\tilde{\boldsymbol{\xi}}_{n,s}^{(\mathbf{x}_{s-1}^{(n)})} := \sum_{\mathbf{x}_s^{(n)}\in[1:M]} \boldsymbol{\xi}_{n,s}^{(\mathbf{x}_s^{(n)})} \mathbf{K}_{n,\mathbf{c}_{s-1}}^{(\mathbf{x}_{s-1}^{(n)},\mathbf{x}_s^{(n)})} \left(\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(\mathbf{x}_s^{(n)})} \left(1-\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(0)},$$

$$\tilde{\xi}_{n,0} := \sum_{\mathbf{x}_0^{(n)}\in[1:M]} \boldsymbol{\xi}_{n,0}^{(\mathbf{x}_0^{(n)})} \mathbf{p}_{n,0}^{(\mathbf{x}_0^{(n)})},$$

for $n \in [1:N]$.

It is crucial to understand the difference between $\boldsymbol{\xi}_{n,s}$ and $\tilde{\boldsymbol{\xi}}_{n,s}$. $\boldsymbol{\xi}_{n,s}$ is used to approximate $p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s, \theta)$ without knowing $\mathbf{x}_s$ and so it is computed by substituting $\mathbf{c}_s$ with $N\mathbf{m}_{s|t}$. $\tilde{\boldsymbol{\xi}}_{n,s}$ is used to approximate $p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s, \theta)$ when knowing $\mathbf{x}_s$ and so having access to the actual $\mathbf{c}_s$. The latter is important because when considering the proposal distribution of an SMC we know the latest particles and we want to propose the next time step given the last. It is worth mentioning the special case $s = 0$, here we have no latest particles hence the recursion looks different, in particular, $\tilde{\xi}_{n,0}$ is a scalar and it can be used to approximate the marginal likelihood $p(\mathbf{y}_{[1:t]}|\theta)$. Note that the marginal likelihood approximation could be a useful tool, for example, it can be employed in pseudo-likelihood methods (Andrieu and Roberts, 2009) or implemented in a delayed acceptance particle MCMC (Golightly et al., 2015).

$\boldsymbol{\xi}_{n,s}$ and $\tilde{\boldsymbol{\xi}}_{n,s}$ are the only quantities needed to compute our approximate proposal distribution and they can be precomputed before running the SMC

Algorithm 1 Computation of $(\boldsymbol{\xi}_{n,h,s}, \tilde{\boldsymbol{\xi}}_{n,h,s})_{n\in[1:N]}$

Require: $(\mathbf{K}_{n,\bullet})_{n\in[1:N]}$, $(\mathbf{q}_{\tilde{s}})_{\tilde{s}\in[s+1:s+h]}$, $(\mathbf{m}_{\tilde{s}|t})_{\tilde{s}\in[s:s+h-1]}$, $\mathbf{y}_{[s+1:s+h]}$, **if** $s \neq 0$ **add** $\mathbf{y}_s, \mathbf{c}_{s-1}$

for $n = 1, \dots, N$ **do**
  $\boldsymbol{\xi}_{n,h,s+h} \leftarrow \mathbf{1}_M$
  for $\tilde{s} = s+h-1, \dots, s$ **do**
    $\boldsymbol{\xi}^{\mathrm{T}}_{n,h,\tilde{s}} \leftarrow \mathbf{K}^{(\bullet,i)}_{n,N\mathbf{m}_{\tilde{s}|t}} \mathbf{q}^{(i)}_{\tilde{s}+1} \boldsymbol{\xi}^{(i)}_{n,h,\tilde{s}+1} \mathbb{I}_{\mathbf{y}^{(n)}_{\tilde{s}+1}}(i) + \mathbf{K}_{n,N\mathbf{m}_{\tilde{s}|t}}(\mathbf{1}_M - \mathbf{q}_{\tilde{s}+1} \circ \boldsymbol{\xi}_{n,h,\tilde{s}+1}) \mathbb{I}_{\mathbf{y}^{(n)}_{\tilde{s}+1}}(0)$
  if $s \neq 0$ **then**
    $\tilde{\boldsymbol{\xi}}^{\mathrm{T}}_{n,h,s} \leftarrow \mathbf{K}^{(\bullet,i)}_{n,\mathbf{c}_{s-1}} \mathbf{q}^{(i)}_{s} \boldsymbol{\xi}^{(i)}_{n,h,s} \mathbb{I}_{\mathbf{y}^{(n)}_{s}}(i) + \mathbf{K}_{n,\mathbf{c}_{s-1}} \left(\mathbf{1}_M - \mathbf{q}_s \circ \boldsymbol{\xi}_{n,h,s}\right) \mathbb{I}_{\mathbf{y}^{(n)}_{s}}(0)$
  else
    $\tilde{\xi}_{n,h,0} \leftarrow \mathbf{p}^{\mathrm{T}}_{0,n} \boldsymbol{\xi}_{n,h,1}$

at a computational cost $\mathcal{O}(tNM^2)$. However, this also requires a memory cost of $\mathcal{O}(tNM^2)$, because they have to be accessible when running the SMC and it is a considerable issue when t is large. As an alternative, we can compute $\boldsymbol{\xi}_{n,s}$ and $\tilde{\boldsymbol{\xi}}_{n,s}$ at each step of the SMC, which requires a computational cost of $\mathcal{O}(Pt^2NM^2)$. A quadratic in t computational cost is still undesirable, hence we can reduce it by using the observations from the closest future instead of the whole sequence. We can indeed focus on approximating $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:s+h]}, \theta)$ for $h \in \mathbb{N}$ and $h \ll t$. Given that we have presented our approximation for an arbitrary t, approximating $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:s+h]}, \theta)$ is like approximating $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:t]}, \theta)$ for $t = s+h$, but we make the dependence on h explicit by defining $\boldsymbol{\xi}_{n,h,s}$ and $\tilde{\boldsymbol{\xi}}_{n,h,s}$ as the $\boldsymbol{\xi}_{n,s}, \tilde{\boldsymbol{\xi}}_{n,s}$ obtained from the algorithm when looking h steps ahead. The whole procedure is summarized in Algorithm 1 and it requires a computational cost of $\mathcal{O}(hNM^2)$. Embedding this algorithm in an SMC demands a computational cost of $\mathcal{O}(PthNM^2)$, which can be controlled by the users depending on the computational resources and application. We can then conclude the section by stating our

optimal proposal distribution:

$$q(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[1:t]}, \theta) = \prod_{n\in[1:N]} \frac{\boldsymbol{\xi}_{n,h,s}^{(\mathbf{x}_s^{(n)})} \mathbf{K}_{n,\mathbf{c}_{s-1}}^{(\mathbf{x}_{s-1}^{(n)},\mathbf{x}_s^{(n)})} \left(\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(\mathbf{x}_s^{(n)})} \left(1-\mathbf{q}_s^{(\mathbf{x}_s^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_s^{(n)}}(0)}}{\tilde{\boldsymbol{\xi}}_{n,h,s}^{(\mathbf{x}_{s-1}^{(n)})}},$$

$$q(\mathbf{x}_0|\mathbf{y}_{[1:t]}, \theta) = \prod_{n\in[1:N]} \frac{\boldsymbol{\xi}_{n,h,0}^{(\mathbf{x}_0^{(n)})} \mathbf{p}_{n,0}^{(\mathbf{x}_0^{(n)})}}{\tilde{\xi}_{n,h,0}}.$$

4.3 Resampling

The resampling scheme $(r_s(i))_{s\in[0:t]}$ is not trivial, indeed choosing resampling schemes that are not looking into future observation vanishes all the effort in building optimal proposals (Fearnhead, 2008). Ideally, resampling should be done according to the smoothing distribution $p(\mathbf{x}_s|\mathbf{y}_{[1:t]}, \theta)$ (Scharth and Kohn, 2016):

$$p(\mathbf{x}_s|\mathbf{y}_{[1:t]}, \theta) = \frac{p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s, \theta)p(\mathbf{x}_s|\mathbf{y}_{[1:s]}, \theta)}{p(\mathbf{y}_{[s+1:t]}|\mathbf{y}_{[1:s]}, \theta)} \propto p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s, \theta)p(\mathbf{x}_s|\mathbf{y}_{[1:s]}, \theta),$$

which is a combination of the probability of observing the future observations given the current sample $\mathbf{x}_s$ and the filtering distribution. The equivalent of the low-cost case where the proposal distribution approximate $p(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[s:s+h]}, \theta)$ follows trivially for $t = s + h$.

The quantities involved in the optimal resampling cannot be computed in closed form and they need to be approximated. SMC outputs a particle approximation $(N)^{-1}\sum_{p\in[1:P]} w_s^p \delta_{\mathbf{x}_s^p}(\mathbf{x}_s)$ of the filtering distribution $p(\mathbf{x}_s|\mathbf{y}_{[1:s]}, \theta)$ and at the same time algorithm 1 gives an approximation $\prod_{n\in[1:N]} \tilde{\boldsymbol{\xi}}_{n,h,s+1}^{(\mathbf{x}_s^{(n)})}$ for $p(\mathbf{y}_{[s+1:t]}|\mathbf{x}_s, \theta)$. It then follows that the approximate optimal resampling is:

$$r_s(i) \propto w_s^i \prod_{n\in[1:N]} \tilde{\boldsymbol{\xi}}_{n,h,s+1}^{((\mathbf{x}_s^i)^{(n)})} \quad \text{for } i \in [1:P],$$

with $(\mathbf{x}_s^p)_{p\in[1:P]}$ being the sampled particles at time s.

5 Experiments

In this section, we analyse the performance of SMC algorithms when using our approximation of the optimal proposal and resampling scheme. We consider simulated data from two compartmental models: Susceptible-Infected-Susceptible (SIS), Susceptible-Exposed-Infected-Removed (SEIR), which are analysed in Section 5.1 and Section 5.2 respectively. For each model we follow an experimental routine inspired by Ju et al. (2021), comparing our method with the BPF and APF: (i) compare methods based on the effective sample size (ESS) $1/\sum_{i\in[1:P]}(r_s(i))$; (ii) compare methods based on the standard deviation of the estimate of marginal likelihood; (iii) study of the marginal likelihood surface on a grid of parameter values for different t when using our method.

All the experiments are run on 32gb Tesla V100 GPU available on the HEC (High-End Computing) facility from Lancaster University. The code can be found in the GitHub repository "Optimal_IBM_proposal" (https://github.com/LorenzoRimella/Optimal_IBM_proposal).

5.1 Susceptible-infected-susceptible

The SIS model is used in epidemiology to model the spread of a disease in a population when herd immunity is not possible. As already mentioned in Section 3.1, we can formulate an individual-based model by having individuals-specific covariates $(\mathbf{w}_n)_{n\in[1:N]}$ and use these covariates to define a unique dynamic per each individual. We have covariates of the form $\mathbf{w}_n = [\mathbf{w}_n^{(1)}, \mathbf{w}_n^{(2)}]^{\mathrm{T}}$ where $\mathbf{w}_n^{(1)} = 1$ and $\mathbf{w}_n^{(2)} \sim \mathcal{N}(\bullet|0,1)$ independently for all $n \in [1:N]$. If not specified otherwise we consider $N = 100$, time horizon $t = 100$ and data generating parameters (DGP) given by: $\boldsymbol{\beta_0} = [-\log(N-1), 0]^{\mathrm{T}}$, $\boldsymbol{\beta_\lambda} = [-1, 2]^{\mathrm{T}}$, $\boldsymbol{\beta_\gamma} = [-1, -1]^{\mathrm{T}}$ and $\mathbf{q}_t = \mathbf{q}$ with $\mathbf{q} = [0.8, 0.8]^{\mathrm{T}}$.

The first experiment consists of measuring the ESS for BPF, APF, and $h = 1, 5, 10, 20, 50$ when $P = 512$. Figure 2 displays our findings for a number of particles $P = 512$. The BPF fails in sampling any epidemics trajectory, this is due to the mismatch problem mentioned in Section 3.3, indeed it is enough to mismatch a single individual out of N to assign 0 probability to the associated particle. APF corrects the proposal by looking at the current observation and so it avoids mismatch. Even though this is a significant improvement compared to BPF the ESS is still very low. Our approximate

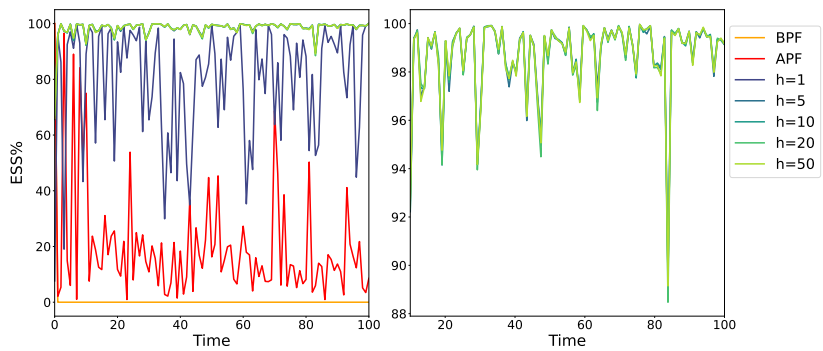

Figure 2: ESS percentage over time for BPF, APF and our method when $h = 1, 5, 10, 20, 50$. Different colors correspond to different methods. The left plot shows all the listed methods, while the right one considers only $h = 10, 20, 50$ and zoom-in.

optimal proposal reaches a significantly better ESS than APF by just looking at the next step in the future ($h = 1$). We also observe that choosing $h > 5$ does not improve much the performance, this is due to the forgetting property of our HMM (Douc et al., 2009).

In the next experiment, we look at the standard deviation of the marginal likelihood estimates. We consider two frameworks: one using the data generating parameters and the other substituting β_λ with $[-3, 0]^{\mathrm{T}}$. Standard deviations are computed over 100 runs. The APF is $3 - 4$ times faster than our method when $h = 5$, but the standard deviation is, in both frameworks, $10 - 20$ times higher than $h = 5$ for small P and even $20 - 30$ times higher than $h = 5$ for big P. Again, we do not notice a substantial improvement when using $h > 5$. The computational cost highly depends on the implementation, our scripts run on GPUs and parallelize each step of the SMC across individuals and particles, hence we do not report significant changes in the running time when increasing P.

Now suppose we want to infer β_λ or β_γ. We start by setting β_λ in a 2-dimensional grid on $[-4, 4]^2$ and the other parameters to the data generating parameters (including β_γ). We then compute estimates of the marginal likelihood with an SMC employing our proposal and resampling scheme when $P = 512$. The procedure is then replicated for β_γ. Both experiments are run for $h = 5, 10$ and $t = 50, 100$, with new data generated per each value of t. Marginal likelihood contour plots are reported in Figure 3 in log scale and

Table 2: Table reporting standard deviation for the APF and our method when $h = 5, 10, 20$ under the data generating process (DGP) and non data generating process (NDGP) with $P = 128, 512, 2048$. The mean computational cost of a single step of the SMC is reported in the first row with the name of the algorithm.

	APF	0.7s	h=5	2.5s	h=10	3.94s	h=20	6.61s
	DGP	NDGP	DGP	NDGP	DGP	NDGP	DGP	NDGP
P	std	std	std	std	std	std	std	std
128	4.99	9.89	0.3	0.92	0.31	1.0	0.37	0.89
512	4.01	6.66	0.17	0.48	0.18	0.49	0.18	0.48
2048	2.83	6.23	0.11	0.25	0.11	0.22	0.11	0.22

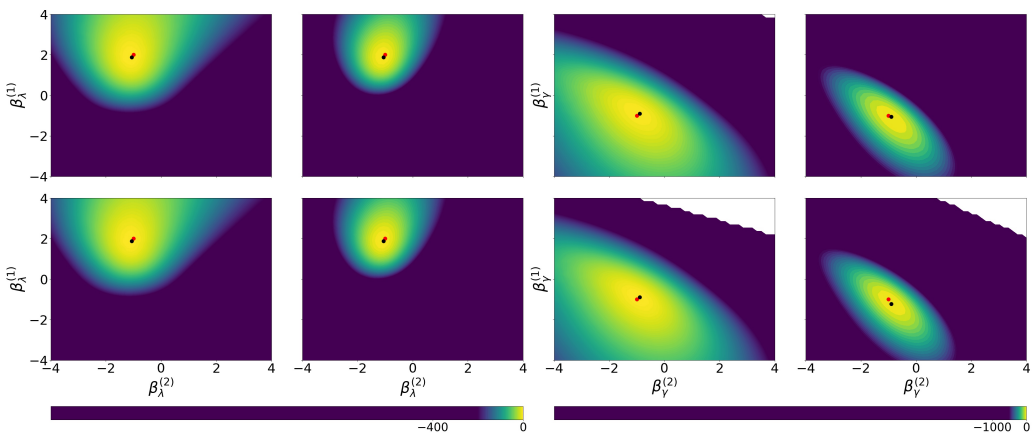

Figure 3: Marginal likelihood contour plots on a β_λ grid and a β_γ grid in log-scale. The first and second columns refer to $t = 50, 100$ from left to right for β_λ. The third and fourth columns refer to $t = 50, 100$ from left to right for β_γ. Rows refer to $h = 5, 10$ from top to bottom. The colorbars are common across parameters and their maximum is set to 0. In red are the data-generating parameters and in black are the MLE on the grid.

normalized to have their max in zero.

In both figures, we can observe that increasing the time concentrates the likelihood around the data-generating parameters. Choosing $h = 10$ does not improve much inference over β_λ, but it helps to infer β_γ by removing some combination of the parameters from the inference (white spaces).

5.2 Susceptible-exposed-infected-removed

The SEIR model is another popular model in epidemiology (He et al., 2020; Deguen et al., 2000; Porter and Oleson, 2013), it is used when the disease is expected to have a latent period (exposed compartment) and herd immunity (removed compartment). The SEIR case is significantly more challenging than the SIS because the transition kernel constrains the dynamic on $S \to E \to I \to R$ and so if in our SMC at time $t-1$ we have a particle with individual n in compartment S and we then observe the same individual at time t in compartment I or R the SMC assigns 0 probability to that particle.

As for the SIS case, a heterogeneous SEIR model is obtained by including a collection of covariates defining $(\mathbf{w}_n)_{n\in[1:N]}$. The initial distribution $\mathbf{p}_{n,0}$ is defined on compartments 1 (S) and 3 (I) as for the SIS case, with zeros for compartments 2 (E) and 4 (I). Similarly, $(\mathbf{K}_{n,\bullet})_{n\in[1:N]}$ is defined as the SIS for transitions $1,2$ (S, E) and $3,4$ (I, R), with the additional transition $2,3$ (E, I) given by $1-\exp(-\rho)$. Full definitions of $\mathbf{p}_{n,0}$ and $(\mathbf{K}_{n,\bullet})_{n\in[1:N]}$ are available in the appendix. The emission distribution follows (3.2).

We have covariates of the form $\mathbf{w}_n = [\mathbf{w}_n^{(1)}, \mathbf{w}_n^{(2)}]^{\mathrm{T}}$ where $\mathbf{w}_n^{(1)} = 1$ and $\mathbf{w}_n^{(2)} \sim \mathcal{N}(\bullet|0,1)$ independently for all $n \in [1:N]$. If not specified otherwise we consider $N = 1000$, time horizon $t = 100$ and data generating parameters given by: $\boldsymbol{\beta_0} = [-\log(N/10-1), 0]^{\mathrm{T}}$, $\boldsymbol{\beta_\lambda} = [1,2]^{\mathrm{T}}$, $\rho = 0.2$, $\boldsymbol{\beta_\gamma} = [-1,-1]^{\mathrm{T}}$ and $\mathbf{q}_s = \mathbf{q}$ with $\mathbf{q} = [0,0,0.4,0.6]^{\mathrm{T}}$.

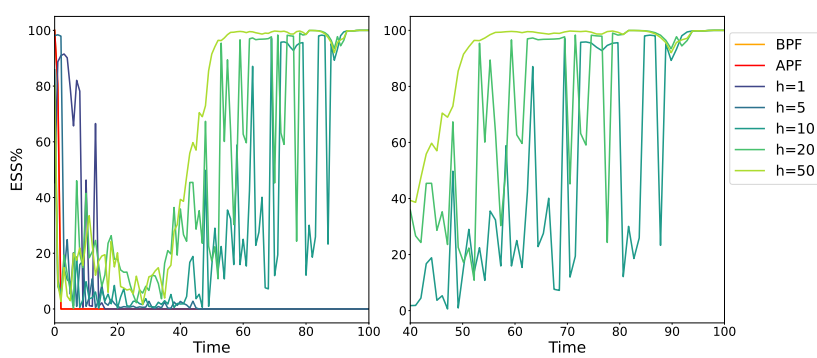

Figure 4: ESS percentage over time for BPF, APF, and our method when $h = 1, 5, 10, 20, 50$. Different colors correspond to different methods. The left plot shows the listed methods, while the right one considers only $h = 5, 10, 20, 50$ and zoom-in.

As for the SIS case, we start by analysing the ESS for BPF, APF, and $h = 1, 5, 10, 20, 50$ with $P = 512$. In Figure 4 both the BPF and the APF fail due to a mismatch between the proposed particles and the observations. Even our method fails for $h \leq 5$, but when choosing $h \geq 10$ we are able to avoid mismatch and get an increasing in time ESS.

Table 3: Table reporting standard deviation for our method when $h = 5, 10, 20, 50$ under the data generating parameters (DGP) and non data generating parameters (NDGP) with $P = 128, 512, 2048$. The mean computational cost of a single step of the SMC is reported in the first row with the name of the algorithm.

	h=5	0.9s	h=10	3.5s	h=20	5.45s	h=50	9.03s
	DGP	NDGP	DGP	NDGP	DGP	NDGP	DGP	NDGP
P	std	std	std	std	std	std	std	std
128	58.18	68.2	20.47	32.6	9.59	18.32	6.93	11.71
512	48.23	74.78	18.37	28.64	6.39	15.76	6.23	10.72
2048	42.7	58.37	15.03	24.68	5.69	13.25	4.57	10.45

We then investigate the standard deviation and computational cost of our method when h changes, and we report our results in Table 3. Observe there is a significant improvement in the standard deviation when increasing h up to 50, with the jump from $h = 20$ to $h = 50$ being less substantial. Clearly, there is a trade-off, a decrease in standard deviation has to be paid for by an increase in computational cost, but it seems worth it for $h < 50$, given that halving the standard deviation is associated with less than doubling the computational cost.

We conclude by reproducing the marginal likelihood surfaces of β_λ and β_γ on grids for the SEIR scenario. The experiments are run for $h = 10, 20$ and $t = 50, 100$, with new data generated per each value of t. Figure 5 shows the marginal likelihood contour plots on a log scale and normalized to have their max in zero. As for the SIS case, an increase in t concentrates the likelihood around the DGP as can be seen in both figures. Notice that the log-likelihood surface of β_λ is multi-modal, this is due to observing neither susceptible nor exposed individuals, which makes inference on this parameter significantly harder. Choosing $h = 20$ seems to smooth the likelihood surface and it also avoids failure close to the data generating parameters, as seen by white holes in the surface for $h = 10$ and $t = 100$. β_γ has a smoother surface compare to β_λ and again increasing h seems to improve the shape.

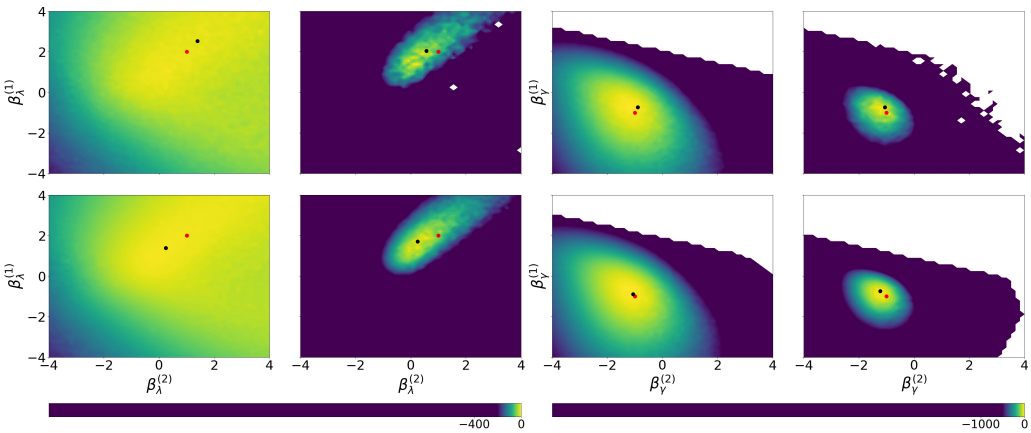

Figure 5: Marginal likelihood contour plots on a β_λ grid and a β_γ grid in log-scale. The first and second columns refer to $t = 50, 100$ from left to right for β_λ. The third and fourth columns refer to $t = 50, 100$ from left to right for β_γ. Rows refer to $h = 5, 10$ from top to bottom. The colorbars are common across parameters and their maximum is set to 0. In red are the data-generating parameters and in black are the MLE on the grid.

6 Discussion

Our findings demonstrate the difficulties in fitting individual-based epidemic models in the presence of censored data and highlight the significance of incorporating future observations to guide the choice of proposal distributions in SMC algorithms. The underlying framework in which proposal distributions are built is general and the algorithm requires only obtaining, for each individual, estimates of the transition rates at the times $[t+1, t+h]$, which are then propagated backwards to build a proposal distribution that includes future observations.

While the overall procedure has been presented as an algorithm to compute good proposal distributions, it seems like several aspects of the work could be used to improve existing methods. For example, as already mentioned, our backward recursion method could be used to compute the marginal likelihood approximation in pseudo-likelihood methods (Andrieu and Roberts, 2009), or as the first approximate model stage in the delayed acceptance scheme of Golightly et al. (2015).

Our implementation focuses on the case of homogeneous reporting rates in a fully connected population, however, the work can be extended beyond this setting. Indeed, it is straightforward to use these techniques for heterogeneous reporting rates, as discussed at the end of Section D, and for spatial epidemic models we would simply need to obtain an estimate of the spatial risk of infection to be able to run the recursion. Epidemic models with an open population (e.g. migration or births-deaths) and misreporting can also be included in the class of models we can deal with, by substituting the multinomial approximation (Whiteley and Rimella, 2021) with alternative approximations (Whitehouse et al., 2022).

Supplementary Materials

The supplementary materials are divided in five sections reporting:

1. the main notation and conventions;
2. an introduction to compartmental model;
3. the main computation in HMM and SMC;
4. the algorithm by Whiteley and Rimella (2021);
5. some additional experiments and extra details on some experiments from the main paper.

Acknowledgements

This work is supported by EPSRC grants EP/R018561/1 (Bayes4Health) and EP/R034710/1 (CoSInES). The authors thank Simon Spencer for a useful discussion, which initiated the idea.

References

Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(3), 269–342.

Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics 37*(2), 697–725.

Bickel, P., B. Li, and T. Bengtsson (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, pp. 318–329. Institute of Mathematical Statistics.

Brauer, F. (2008). Compartmental models in epidemiology. In *Mathematical Epidemiology*, pp. 19–79. Springer.

Britton, T. (2010). Stochastic epidemic models: A survey. *Mathematical Biosciences 225*(1), 24–35.

Brooks-Pollock, E., L. Danon, T. Jombart, and L. Pellis (2021). Modelling that shaped the early COVID-19 pandemic response in the UK. *Phil. Trans. R. Soc. B 376*, 20210001.

Candy, J. V. (2007). Bootstrap particle filtering. *IEEE Signal Processing Magazine 24*(4), 73–85.

Carpenter, J., P. Clifford, and P. Fearnhead (1999). Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation 146*(1), 2–7.

Chapman, L. A. C., S. E. F. Spencer, T. M. Pollington, C. P. Jewell, D. Mondal, J. Alvar, T. D. Hollingsworth, M. M. Cameron, C. Bern, and G. F. Medley (2020). Inferring transmission trees to guide targeting of interventions against visceral leishmaniasis and post–kala-azar dermal leishmaniasis. *Proceedings of the National Academy of Sciences 117*(41), 25742–25750.

Chopin, N., O. Papaspiliopoulos, et al. (2020). *An introduction to sequential Monte Carlo*, Volume 4. Springer.

Cocker, D., M. Sammarro, K. Chidziwisano, N. Elviss, S. Jacob, H. Kajumbula, L. Mugisha, D. Musoke, P. Musicha, A. Roberts, B. Rowlingson, A. Singer, R. Byrne, T. Edwards, R. Lester, C. Wilson, B. Hollihead, N. Thomson, C. Jewell, T. Morse, and N. Feasey (2022, February). Drivers of Resistance in Uganda and Malawi (DRUM): a protocol for the evaluation of One-Health drivers of Extended Spectrum Beta Lactamase (ESBL) resistance in Low-Middle Income Countries (LMICs). *Wellcome Open Res.*.

Deguen, S., G. Thomas, and N. P. Chau (2000). Estimation of the contact rate in a seasonal SEIR model: application to chickenpox incidence in France. *Statistics in Medicine 19*(9), 1207–1216.

Douc, R., E. Moulines, and Y. Ritov (2009). Forgetting of the initial condition for the filter in general state-space hidden Markov chain: a coupling approach. *Electronic Journal of Probability 14*, 27–49.

Fearnhead, P. (2008). Computational methods for complex stochastic systems: a review of some alternatives to MCMC. *Statistics and Computing 18*(2), 151–171.

Fintzi, J., J. Wakefield, and V. N. Minin (2021). A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. *Biometrics*.

Funk, S., S. Abbott, B. Atkins, M. Baguelin, J. Baillie, P. Birrell, J. Blake, N. Bosse, J. Burton, J. Carruthers, N. Davies, D. De Angelis, L. Dyson, W. Edmunds, R. Eggo, N. Ferguson,

K. Gaythorpe, E. Gorsich, G. Guyver-Fletcher, J. Hellewell, E. Hill, A. Holmes, T. House, C. Jewell, M. Jit, T. Jombart, I. Joshi, M. Keeling, E. Kendall, E. Knock, A. Kucharski, K. Lythgoe, S. Meakin, J. Munday, P. Openshaw, C. Overton, F. Pagani, J. Pearson, P. Perez-Guzman, L. Pellis, F. Scarabel, M. Semple, K. Sherratt, M. Tang, M. Tildesley, E. Van Leeuwen, L. Whittles, C. C.-. W. Group, I. C. C.-. R. Team, and I. Investigators (2020). Short-term forecasts to inform the response to the COVID-19 epidemic in the UK. *medRxiv*.

Golightly, A., D. A. Henderson, and C. Sherlock (2015). Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statistics and Computing 25*(5), 1039–1055.

Gordon, N. J., D. J. Salmond, and A. F. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F-Radar and Signal Processing*, Volume 140, pp. 107–113. IET.

He, S., Y. Peng, and K. Sun (2020). SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dynamics 101*(3), 1667–1680.

Ionides, E. L., C. Bretó, and A. A. King (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences 103*(49), 18438–18443.

Jewell, C. P., M. J. Keeling, and G. O. Roberts (2009, Dec). Predicting undetected infections during the 2007 foot-and-mouth disease outbreak. *J R Soc Interface 6*(41), 1145–1151.

Jewell, C. P., T. Kypraios, P. Neal, and G. O. Roberts (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis 4*(3), 465–496.

Johansen, A. M. and A. Doucet (2008). A note on auxiliary particle filters. *Statistics & Probability Letters 78*(12), 1498–1504.

Ju, N., J. Heng, and P. E. Jacob (2021). Sequential Monte Carlo algorithms for agent-based models of disease transmission. arXiv:2101.12156.

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

Kucharski, A. J., T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, et al. (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases 20*(5), 553–558.

Lekone, P. E. and B. F. Finkenstädt (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics 62*(4), 1170–1177.

Pitt, M. K. and N. Shephard (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association 94*(446), 590–599.

Porter, A. T. and J. J. Oleson (2013). A path-specific SEIR model for use with general latent and infectious time distributions. *Biometrics 69*(1), 101–108.

Rabiner, L. and B. Juang (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine 3*(1), 4–16.

Rimella, L., M. Sammarro, S. Alderton, N. Feasey, P. Fearnhead, and C. Jewell (2022). Inference on extended-spectrum beta-lactamase *Escherichia coli* and *Klebsiella pneumoniae* data through SMC^2. *In Preparation*.

Scharth, M. and R. Kohn (2016). Particle efficient importance sampling. *Journal of Econometrics 190*(1), 133–147.

Van der Goot, J., G. Koch, M. De Jong, and M. Van Boven (2005). Quantification of the effect of vaccination on transmission of avian influenza (H7N7) in chickens. *Proceedings of the National Academy of Sciences 102*(50), 18141–18146.

Whitehouse, M., N. Whiteley, and L. Rimella (2022). Consistent and fast inference in compartmental models of epidemics using poisson approximate likelihoods. *arXiv preprint arXiv:2205.13602*.

Whiteley, N. and A. Lee (2014). Twisted particle filters.

Whiteley, N. and L. Rimella (2021). Inference in stochastic epidemic models via multinomial approximations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1297–1305. PMLR.

Yang, F., S. Balakrishnan, and M. J. Wainwright (2017). Statistical and computational guarantees for the Baum-Welch algorithm. *The Journal of Machine Learning Research 18*(1), 4528–4580.

Zhou, Y., Z. Ma, and F. Brauer (2004). A discrete epidemic model for SARS transmission and control in China. *Mathematical and Computer Modelling 40*(13), 1491–1506.

A Preliminaries

Given the vector $\mathbf{a}$ we use $\mathbf{a}^{(i)}$ for the (i)-th element of $\mathbf{a}$ (vectors are always assumed to be column vectors), given the matrix $\mathbf{A}$ we use $\mathbf{A}^{(i,j)}$ for the (i,j)-th element of $\mathbf{A}$ and we use $\mathbf{A}^{(i,\bullet)}$ (or $\mathbf{A}^{(\bullet,j)}$) to represent the column vector given by the (i)-th column (or the (j)-th row) of matrix $\mathbf{A}$. We use $\mathbf{A}^{\mathrm{T}}$ for the transpose of matrix $\mathbf{A}$ (similarly for vectors). The dot product between two matrices $\mathbf{A}, \mathbf{B}$ is denoted by $\mathbf{AB}$ (similarly for vectors). Sums or differences between vectors or matrices have to be thought of as elementwise, similarly when using vectors or matrices and a scalar. We employ the convention $\mathbb{R}_+$ for the positive real numbers. We use the notation $\mathbb{I}_a(b)$ for the indicator function, with a, b being scalars, vectors, or matrices.

Given the probability mass (or density) function of a random variable $\mathcal{D}(x|\theta)$ evaluated in x with parameters θ as in table 4, we use $X \sim \mathcal{D}(\bullet|\theta)$ to say the random variable X is drawn from the corresponding probability random variable and we denote with $\mathbb{E}_{\mathcal{D}(X|\theta)}[X]$ the corresponding expectation.

Distribution	Categorical	Bernoulli	Binomial	Gaussian	Uniform	Multinomial
Notation	$\mathcal{C}at_M(i\|\mathbf{p})$	$\mathcal{B}e(i\|q)$	$\mathcal{B}in(i\|N,q)$	$\mathcal{N}(a\|\mu,\sigma^2)$	$\mathcal{U}nif(q\|a,b)$	$\mathcal{M}ult(\mathbf{c}\|N,\mathbf{p})$
Domain	$[1:M]$	$\{0,1\}$	$[0:N]$	$\mathbb{R}$	$[a,b]$	$\{\mathbf{c}:\sum_{i\in[1:M]}\mathbf{c}^{(i)}=N\}$

Table 4: Notation table for probability mass and density functions along with domains.

B Introduction to compartmental models

In epidemiology, compartmental models are used to describe the dynamics of an epidemic in a population, where the compartments represent different stages of the disease (Brauer, 2008). A closed population stochastic compartmental model is fully defined by: the number of compartments M, the population size N, the initial probability of being assigned to a compartment $\mathbf{p}_0$, and the probability of transition from one compartment to the other $\mathbf{K}_\bullet$, which is a stochastic transition matrix $\mathbf{c} \to \mathbf{K}_\mathbf{c}$ defined as a function of the compartments' state $\mathbf{c}$. The dynamics of a compartmental model is generally described by a discrete-time Markov chain $(\mathbf{c}_t)_{t\geq 0}$ counting the number of individuals in each compartment, i.e. $\mathbf{c}_t$ is an M-dimensional vector with $\mathbf{c}_t^{(i)}$ being the number of individuals in compartment i (Keeling and Rohani, 2011). The evolution of $\mathbf{c}_t$ can be specified from an individual perspective by defining the discrete-time Markov chain $(\mathbf{x}_t)_{t\geq 0}$, where $\mathbf{x}_t$ is an N-dimensional vector representing the state of the population at time t (Whiteley and Rimella, 2021), i.e. $\mathbf{x}_t^{(n)}$ is the compartment of individual n at time t. The evolution of $(\mathbf{x}_t)_{t\geq 0}$ and $(\mathbf{c}_t)_{t\geq 0}$ can be then represented as the following flow:

Time 0: $\mathbf{x}_0^{(n)} \sim \mathcal{C}at_M(\bullet|\mathbf{p}_0)$ for $n \in [1:N]$ and
$\mathbf{c}_0^{(i)} = \sum_{n=1}^N \mathbb{I}_{\mathbf{x}_0^{(n)}}(i)$ for $i \in [1:M]$;

Time t: $\mathbf{x}_t^{(n)}|\mathbf{x}_{t-1} \sim \mathcal{C}at_M\left(\bullet\middle|\mathbf{K}_{\mathbf{c}_{t-1}}^{(\mathbf{x}_{t-1}^{(n)},\bullet)}\right)$ for $n \in [1:N]$ and
$\mathbf{c}_t^{(i)} = \sum_{n=1}^N \mathbb{I}_{\mathbf{x}_t^{(n)}}(i)$ for $i \in [1:M]$;

where we first make all the individuals move (simulate $\mathbf{x}_t$) and then count the individuals in each compartment (compute $\mathbf{c}_t$).

SIS example The susceptible-infected-susceptible model (SIS) is a well-known compartmental model used to model the spread of a disease in a population where herd immunity is not possible, i.e. the individuals can be re-infected. A stochastic SIS with closed population N can be represented in the previous framework by: $M = 2$, $\mathbf{p}_0$ in the 2-dimensional simplex (with $\mathbf{p}_0^{(1)}$ probability of being susceptible at time 0) and $\mathbf{K}_\bullet$ a 2by2-dimensional stochastic transition matrix. A popular choice of $\mathbf{K}_\bullet$ is:

$$\mathbf{K}_c = \begin{bmatrix} e^{-\beta\frac{\mathbf{c}^{(2)}}{N}} & 1-e^{-\beta\frac{\mathbf{c}^{(2)}}{N}} \\ 1-e^{-\gamma} & e^{-\gamma} \end{bmatrix}$$

with $\beta, \gamma \in \mathbb{R}_+$ transmission and recovery parameters.

Focusing on the compartments' state automatically assumes homogeneous individuals, which is a significant simplification of the real world, where each individual often has their own covariates, e.g. age. individual-based compartmental models or simply individual-based models relax the homogeneity assumption and look at the disease from an individual perspective. The main difference resides in the individual-specific $(\mathbf{p}_{n,0})_{n\in[1:N]}$ and $(\mathbf{K}_{n,\bullet})_{n\in[1:N]}$ representing the heterogeneous dynamic of the individuals.

B.1 Considerations on non-granular observation models

Compartmental models in epidemiology are generally treated as latent and accompanied by an observation model representing the conditional distribution of the observations given the current compartments' state often refer as the emission distribution. Popular choices are the Negative binomial distribution (Fintzi et al., 2021) and the binomial distribution (Lekone and Finkenstädt, 2006), with the latter being used by Ju et al. (2021) in an individual-based model framework. The main limitation of these emission distributions is that they are formulated over the aggregated population, while for an individual-based model, it is fair to assume individual-based observations.

C Hidden Markov models

In an HMM $(\mathbf{x}_t, \mathbf{y}_t)_{t\geq 1}$ the recursive computation of $p(\mathbf{x}_t|\mathbf{y}_{[1:t]}, \theta)$ and $p(\mathbf{y}_{[1:t]}|\theta)$ is known as forward algorithm and it consists of the following steps:

Time 0: $p(\mathbf{x}_0|\theta)$ from the initial distribution;

Time t: $p(\mathbf{x}_t|\mathbf{y}_{[1:t-1]}, \theta) = \sum_{\mathbf{x}_{t-1}\in[1:M]^N} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta)p(\mathbf{x}_{t-1}|\mathbf{y}_{[1:t-1]}, \theta)$ and
$p(\mathbf{y}_t|\mathbf{y}_{[1:t-1]}, \theta) = \sum_{\mathbf{x}_t\in[1:M]^N} p(\mathbf{y}_t|\mathbf{x}_t, \theta)p(\mathbf{x}_t|\mathbf{y}_{[1:t-1]}, \theta)$ and
$p(\mathbf{x}_t|\mathbf{y}_{[1:t]}, \theta) = \frac{p(\mathbf{y}_t|\mathbf{x}_t,\theta)p(\mathbf{x}_t|\mathbf{y}_{[1:t-1]},\theta)}{p(\mathbf{y}_t|\mathbf{y}_{[1:t-1]},\theta)}$ and
$p(\mathbf{y}_{[1:t]}|\theta) = p(\mathbf{y}_{[1:t-1]}|\theta)p(\mathbf{y}_t|\mathbf{y}_{[1:t-1]}, \theta)$.

A close-form solution for the filtering is often not available and an SMC algorithm can be employed to compute particle estimates of both $p(\mathbf{x}_t|\mathbf{y}_{[1:t]}, \theta)$ and $p(\mathbf{y}_{[1:t]}|\theta)$. A general SMC algorithm is presented in 2 and it consists of: resampling the previous particles according to the resampling scheme and correcting the weights accordingly, proposing new particles with the proposal distribution, update the weights. Note that it is required $\sum_{p\in[1:P]} \tilde{w}_s^p = 1$ with $\tilde{w}_s^p \propto w_s^{i_s^p}/r_s(i_s^p)$, with i_s resampled indices, which is then updated according to the transition kernel, emission distribution and proposal distribution. The likelihood estimate is then given by $\hat{p}(y_{[1:s]}|\theta) \leftarrow \hat{p}(y_{[1:s-1]}|\theta) \sum_{p\in[1:P]} w_s^p$, where w_s^p are the weights after update.

Algorithm 2 A general sequential Monte Carlo algorithm

Require: P, θ, $\mathbf{y}_{[1:t]}$, $(q(\mathbf{x}_s|\mathbf{x}_{s-1}, \mathbf{y}_{[1:t]}))_{s\in[0:t]}$, $(r_s(i))_{s\in[0:t]}$

Sample $\mathbf{x}_0^p \sim q(\bullet|\mathbf{y}_{[1:t]})$

Compute $w_0^p \leftarrow \frac{p(\mathbf{x}_0^p|\theta)}{q(\mathbf{x}_0^p|\mathbf{y}_{[1:t]})}$ for $p \in [1:P]$

for $s = 1, \dots, t$ **do**

  Resample $i_{s-1}^p \sim r_{s-1}(\bullet)$ and set $\tilde{\mathbf{x}}_{s-1}^p \leftarrow \mathbf{x}_{s-1}^{i_{s-1}^p}$ for $p \in [1:P]$

  Correct the weights $\tilde{w}_{s-1}^p \propto \frac{w_{s-1}^{i_{s-1}^p}}{r_{s-1}(i_{s-1}^p)}$ and normalize for $p \in [1:P]$

  Propose $\mathbf{x}_s^p \sim q(\bullet|\tilde{\mathbf{x}}_{s-1}, \mathbf{y}_{[1:t]})$

  Compute $w_s^p \leftarrow \tilde{w}_{s-1}^p \frac{p(\mathbf{x}_s^p|\tilde{\mathbf{x}}_{s-1}^p,\theta)p(\mathbf{y}_s|\mathbf{x}_s^p,\theta)}{q(\mathbf{x}_s^p|\tilde{\mathbf{x}}_{s-1}^p,\mathbf{y}_{[1:t]})}$ for $p \in [1:P]$

$\hat{p}(y_{[1:t]}|\theta) \leftarrow \prod_{s\in[1:t]} \sum_{p\in[1:P]} w_s^p$

In the HMM terminology the initial distribution $p(\mathbf{x}_0|\theta)$, the transition kernel $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta)$ and the emission distribution $p(\mathbf{y}_t|\mathbf{x}_t, \theta)$ for our individual-based model with granular observation are given by:

$$p(\mathbf{x}_0|\theta) = \prod_{n\in[1:N]} \mathbf{p}_{n,0}^{(\mathbf{x}_0^{(n)})}, \quad p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta) = \prod_{n\in[1:N]} \mathbf{K}_{n,\mathbf{c}_{t-1}}^{(\mathbf{x}_{t-1}^{(n)},\mathbf{x}_t^{(n)})},$$

$$p(\mathbf{y}_t|\mathbf{x}_t, \theta) = \prod_{n\in[1:N]} \left(\mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(\mathbf{x}_t^{(n)})} \left(1 - \mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(0)}.$$

These definitions are useful to see how the transition kernel and the emission distribution can be used to simplify the formulation of $p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta)$:

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{x}_{t-1}, \theta) &= \sum_{\mathbf{x}_t\in[1:M]^N} \prod_{n\in[1:N]} \mathbf{K}_{n,\mathbf{c}_{t-1}}^{(\mathbf{x}_{t-1}^{(n)},\mathbf{x}_t^{(n)})} \left(\mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(\mathbf{x}_t^{(n)})} \left(1 - \mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(0)} \\ &= \prod_{n\in[1:N]} \sum_{\mathbf{x}_t^{(n)}\in[1:M]} \mathbf{K}_{n,\mathbf{c}_{t-1}}^{(\mathbf{x}_{t-1}^{(n)},\mathbf{x}_t^{(n)})} \left(\mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(\mathbf{x}_t^{(n)})} \left(1 - \mathbf{q}_t^{(\mathbf{x}_t^{(n)})}\right)^{\mathbb{I}_{\mathbf{y}_t^{(n)}}(0)}. \end{aligned} \tag{3}$$

D A priori estimate of the compartments' state

The approximate homogeneous dynamic considered in the main paper is:

Time 0: $\mathbf{x}_0^{(n)} \sim \mathcal{C}at_M(\bullet|\bar{\mathbf{p}}_{n,0})$ for $n \in [1:N]$ and
$\mathbf{c}_0^{(i)} = \sum_{n=1}^N \mathbb{I}_{\mathbf{x}_0^{(n)}}(i)$ for $i \in [1:M]$;

Algorithm 3 Multinomial approximation by Whiteley and Rimella (2021)

Require: $(\bar{\mathbf{p}}_{n,0})_{n\in[1:N]}$, $(\bar{\mathbf{K}}_{\bullet})_{n\in[1:N]}$, $(\mathbf{q}_s)_{s\in[t]}$, $(\mathbf{o}_s)_{s\in[t]}$

$\mathbf{m}_{0|0} \leftarrow \bar{\mathbf{p}}_{n,0}$ ▷ Forward step

for $s = 1, \dots, t$ **do**

  $\mathbf{m}_{s-1|s} \leftarrow \left(\mathbf{m}_{s-1|s-1}^{\mathrm{T}} \bar{\mathbf{K}}_{\mathbf{m}_{s-1}}\right)^{\mathrm{T}}$

  $\mathbf{m}_{s|s} \leftarrow \frac{\mathbf{o}_s}{N} + \left(1 - \frac{\mathbf{1}_M^{\mathrm{T}} \mathbf{o}_s}{N}\right) \frac{\mathbf{m}_{s-1|s} \circ (\mathbf{1}_M - \mathbf{q}_s)}{1 - \mathbf{m}_{s-1|s}^{\mathrm{T}} \mathbf{q}_s}$

for $s = t-1, \dots, 0$ **do** ▷ Backward step

  $\mathbf{L}_s \leftarrow \left\{ \left[(\mathbf{m}_{s|t} \mathbf{1}_M^{\mathrm{T}}) \circ \bar{\mathbf{K}}_{\mathbf{m}_s} \right] / \left[\mathbf{1}_M (\mathbf{m}_{s|t}^{\mathrm{T}} \bar{\mathbf{K}}_{\mathbf{m}_s}) \right] \right\}^{\mathrm{T}}$

  $\mathbf{m}_{s|t} \leftarrow \left(\mathbf{m}_{s+1|T}^{\mathrm{T}} \mathbf{L}_s\right)^{\mathrm{T}}$

Time t: $\mathbf{x}_t^{(n)} | \mathbf{x}_{t-1} \sim \mathcal{C}at_M \left(\bullet \middle| \bar{\mathbf{K}}_{\mathbf{c}_{t-1}}^{(\mathbf{x}_{t-1}^{(n)}, \bullet)} \right)$ for $n \in [1:N]$ and
$\mathbf{c}_t^{(i)} = \sum_{n=1}^{N} \mathbb{I}_{\mathbf{x}_t^{(n)}}(i)$ for $i \in [1:M]$.

The full algorithm by (Whiteley and Rimella, 2021) is reported in algorithm 3.

E Additional experiments

E.1 Susceptible-infected-susceptible

	APF	0.7s	h=5	2.5s	h=10	3.94s	h=20	6.61s
	DGP	NDGP	DGP	NDGP	DGP	NDGP	DGP	NDGP
P	std	std	std	std	std	std	std	std
64	7.02	10.97	0.4	1.26	0.48	1.35	0.46	1.19
128	4.99	9.89	0.3	0.92	0.31	1.0	0.37	0.89
256	5.24	8.3	0.27	0.72	0.24	0.67	0.35	0.63
512	4.01	6.66	0.17	0.48	0.18	0.49	0.18	0.48
1024	3.42	7.17	0.15	0.34	0.21	0.33	0.15	0.35
2048	2.83	6.23	0.11	0.25	0.11	0.22	0.11	0.22

Table 5: Table reporting standard deviation for APF and our method when $h = 5, 10, 20$ under DGP and NDGP with $P = 64, 128, 256, 512, 1024, 2048$. The mean computational cost is reported in the first row with the name of the algorithm.

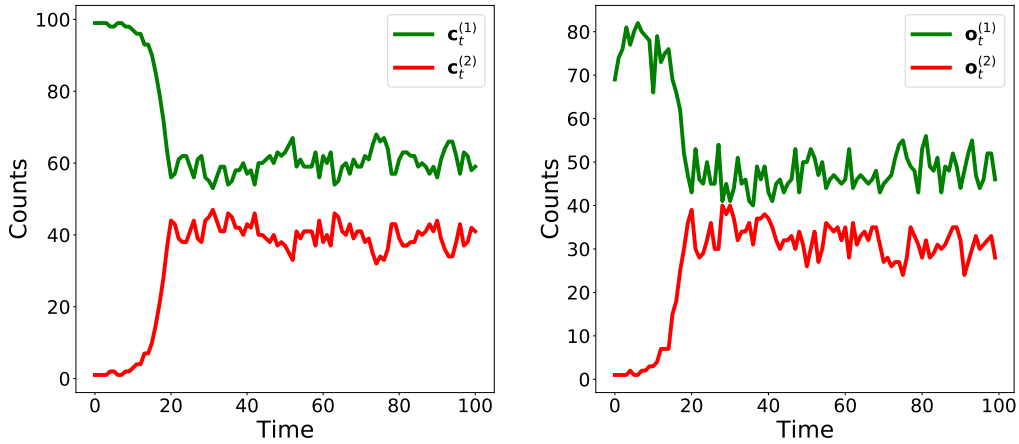

Figure 6: SIS epidemic, on the left the unobserved process, on the right the observed data.

Outcomes from the SIS model are shown in figure 6. The full table on the marginal likelihood standard deviation is reported in table 5. Complete marginal likelihood contour plot are reported in figure 7 and figure 8 in log scale and normalized to have their max in zero.

We measure the sensitivity to $\mathbf{q}$ of APF and our method when $h = 1, 5, 10, 20$ when $P = 512$. We study the performance of the considered methods when $\mathbf{q}$ varies, precisely we choose a scenario where data are generated according to $\mathbf{q} = [0.8, 0.8]^{\mathrm{T}}$, but the algorithms use $\mathbf{q} = [i, i]^{\mathrm{T}}$ with $i \in [0.1, \ldots, 0.9]$. As for the previous experiment we compute the standard deviation and ESS bands by running 100 times algorithm 2 for each framework. Results are reported in figure 9. APF is always associated with a higher standard deviation of the marginal likelihood estimate and to a lower ESS. As for the previous experiments, we have a significant improvement in both standard deviation and ESS when choosing $h = 1$, this gets even better when $h \geq 5$, especially from an ESS perspective.

We now infer $(\beta_0, \beta_\lambda, \beta_\gamma, \mathbf{q})$ from data generated from DGP. For this experiment, we employ a Particle marginal Metropolis-Hastings (PMMH) (Andrieu et al., 2010) using an SMC with our proposal distribution and resampling scheme along with the following set of priors: $\mathcal{N}(\beta_0^{(1)}|0, 3)$, $\mathcal{N}(\beta_0^{(2)}|0, 3)$, $\mathcal{N}(\beta_\lambda^{(1)}|0, 3)$, $\mathcal{N}(\beta_\lambda^{(2)}|0, 3)$, $\mathcal{N}(\beta_\gamma^{(1)}|0, 3)$, $\mathcal{N}(\beta_\gamma^{(1)}|0, 3)$, $\mathcal{U}nif(\mathbf{q}^{(1)}|0, 1)$, $\mathcal{U}nif(\mathbf{q}^{(2)}|0, 1)$. As proposal distribution for the parameters, we use a Gaussian random walk on the log-parameters with variance chosen to match the optimal acceptance rate of 23%.

We run the PMMH for 100000 iterations and we used a burn-in period of 10000 and thinning. Marginal posterior distributions are reported in figure 10, we can notice that posterior are "peaky" and close to the DGP, with the exception of β_0 which has a flatter posterior, which is more difficult

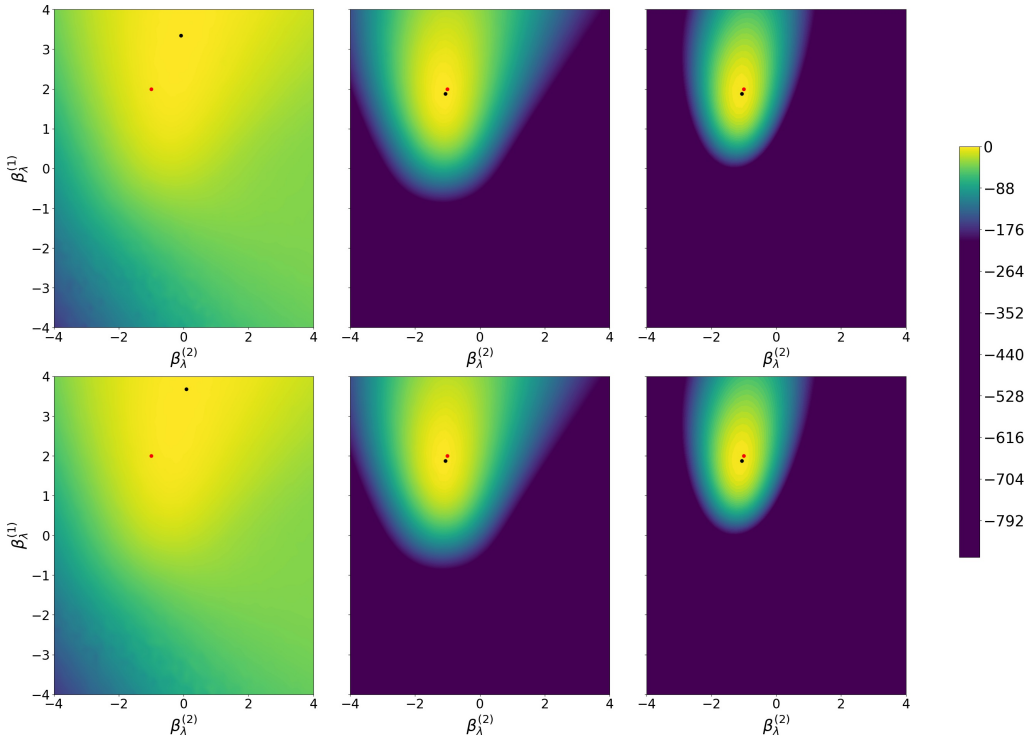

Figure 7: Marginal likelihood contour plots on a β_λ grid in log-scale. Columns refer to $t = 10, 50, 100$ from left to right. Rows refer to $h = 5, 10$ from top to bottom. The colorbar is common across the plots and in each plot, the maximum is set to 0. In red is the DGP and in black is the MLE on the grid.

to learn given that the only information is derived from $\mathbf{y}_1$. To push our study one step further we also plot $R_0^{(n)}$ distribution and posterior predictive in figure 11. We can observe that the PMMH learned the $R_0^{(n)}$ distribution on a global level (first column) and on an individual level (second column), with $R_0^{(n)}$ from the DGP almost indistinguishable. For the posterior predictive we sample 200 times from the posterior and we then simulate 200 epidemics with those parameters. Figure 11 shows good coverage of the aggregated data (fourth column) and of the aggregated latent data (third column). The latter is not available during the inference process, but it is stored at simulation time to add an additional level of comparison.

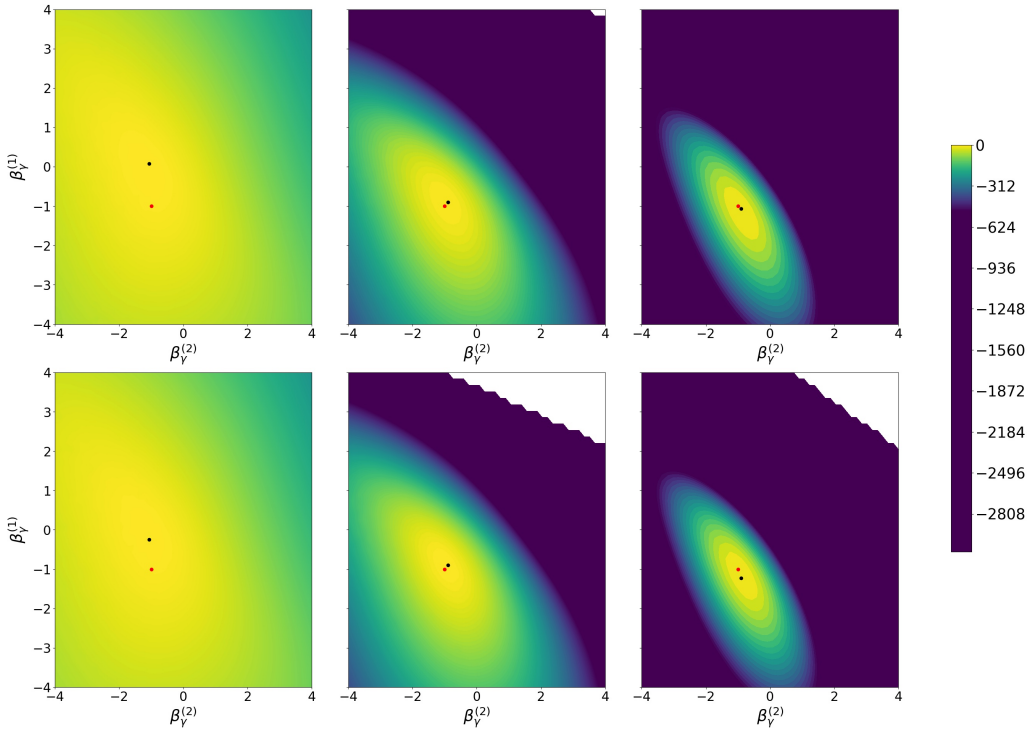

Figure 8: Marginal likelihood contour plots on a β_γ grid in log-scale. Columns refer to $t = 10, 50, 100$ from left to right. Rows refer to $h = 5, 10$ from top to bottom. The colorbar is common across the plots and in each plot, the maximum is set to 0. In red is the DGP and in black is the MLE on the grid.

E.2 Susceptible-exposed-infected-removed

We can make the SEIR model heterogeneous by considering $d \in \mathbb{N}$ and by defining $(\mathbf{w}_n)_{n\in[1:N]}$ as the collection of d-dimensional vectors collecting the individual-specific covariates. From $\mathbf{w}_n$ we can define the initial distribution:

$$\mathbf{p}_{n,0} = \begin{bmatrix} 1 - \frac{1}{1+\exp(-\beta_0^{\mathrm{T}} w_n)} \\ 0 \\ \frac{1}{1+\exp(-\beta_0^{\mathrm{T}} w_n)} \\ 0 \end{bmatrix} \text{ for } n \in [1:N] \text{ and } \beta_0 \in \mathbb{R}^d$$

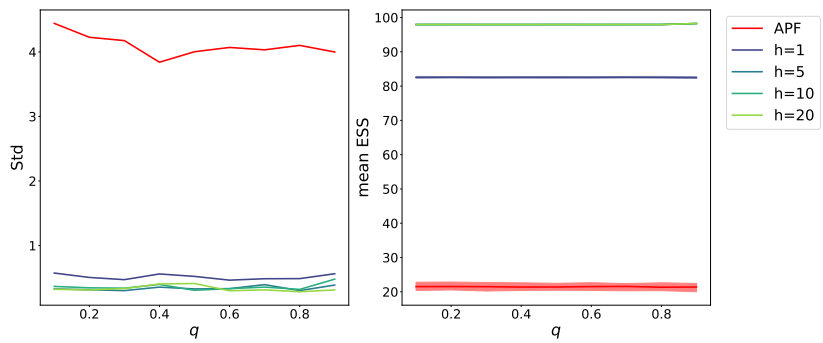

Figure 9: Effective sample size and marginal likelihood standard deviation sensitivity to $\mathbf{q}$. The first column shows the marginal likelihood standard deviation, while the second column reports ESS where mode, 5%-quantile, and 95%-quantile are reported.

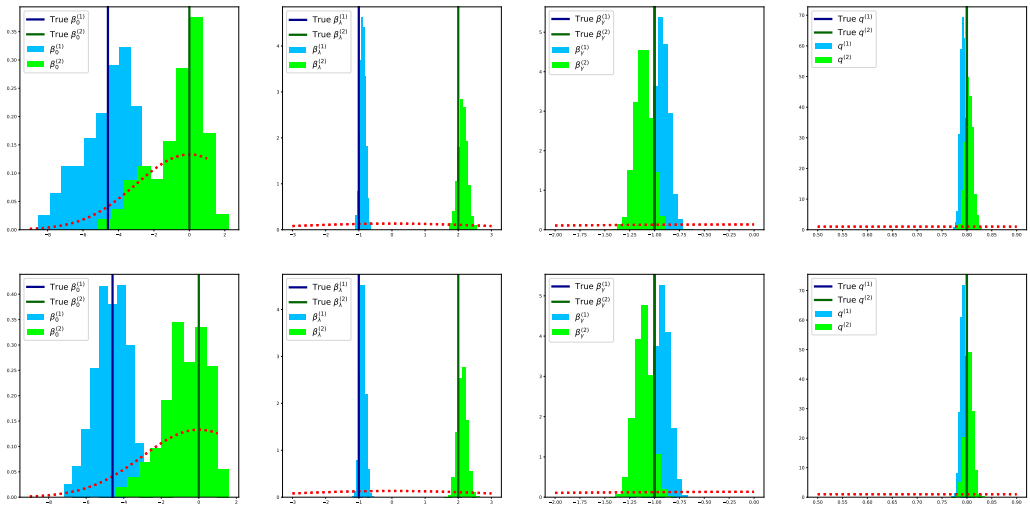

Figure 10: Posterior distribution over the parameters of interest. On the columns $(\beta_0, \beta_\lambda, \beta_\gamma, \mathbf{q})$ from left to right. On the rows $h = 5, 10$ from top to bottom. Priors are reported in red dotted lines. Vertical lines are the DGP.

and calculate $(\mathbf{K}_{n,\bullet})_{n\in[1:N]}$ as:

$$\mathbf{K}_{n,c} = \begin{bmatrix} 1 - \frac{1}{1+\exp(-\beta_\lambda^{\mathrm{T}} w_n)} \frac{\mathbf{c}^{(3)}}{N} & \frac{1}{1+\exp(-\beta_\lambda^{\mathrm{T}} w_n)} \frac{\mathbf{c}^{(3)}}{N} & 0 & 0 \\ 0 & \exp(-\rho) & 1-\exp(-\rho) & 0 \\ 0 & 0 & 1 - \frac{1}{1+\exp(-\beta_\gamma^{\mathrm{T}} w_n)} & \frac{1}{1+\exp(-\beta_\gamma^{\mathrm{T}} w_n)} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

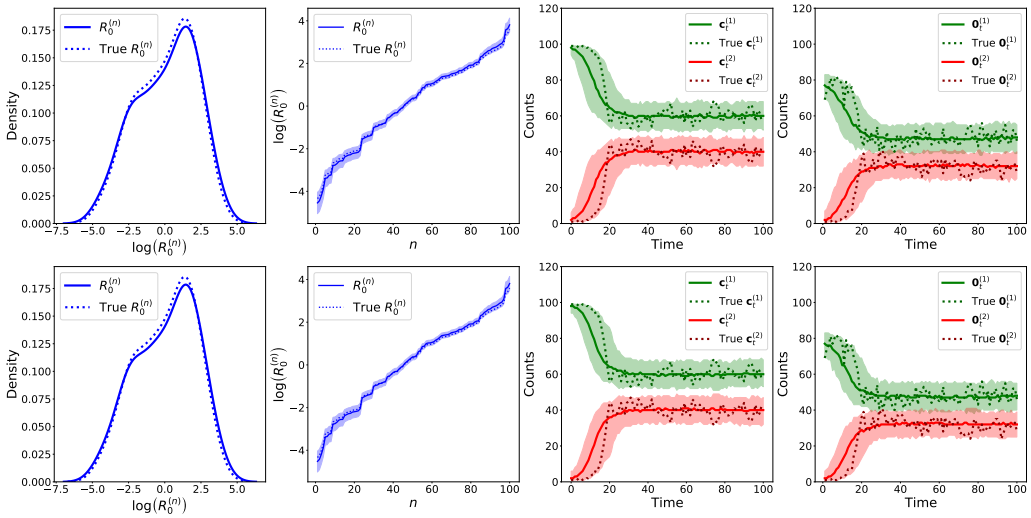

Figure 11: R_0 distribution and posterior predictive. First column distribution of the mean $R_0^{(n)}$. Second column $R_0^{(n)}$ and credible bands for each individual (sorted for increasing $R_0^{(n)}$). The third and fourth columns are posterior predictive over $\mathbf{x}_t$ and $\mathbf{y}_t$. Rows from top to bottom $h = 5, 10$.

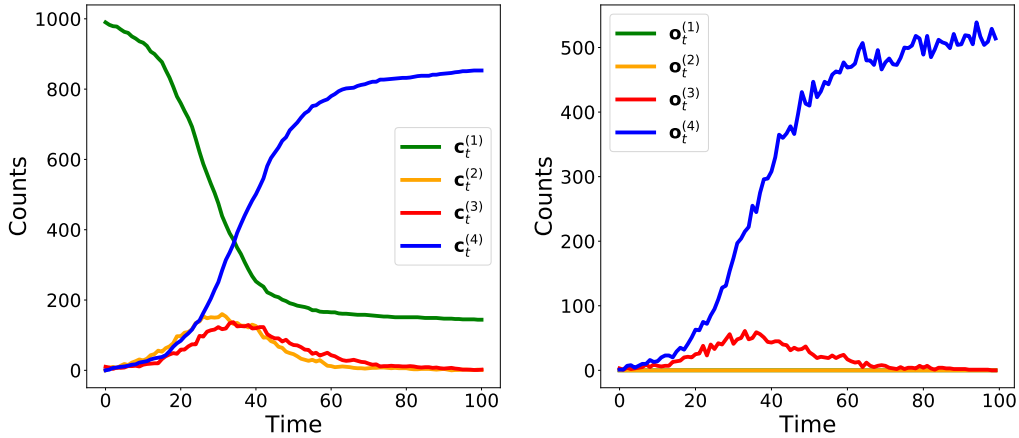

Figure 12: SEIR epidemic, on the left the unobserved process, on the right the observed data.

	h=5	0.9s	h=10	3.5s	h=20	5.45s	h=50	9.03s
	DGP	NDGP	DGP	NDGP	DGP	NDGP	DGP	NDGP
P	std	std	std	std	std	std	std	std
64	49.1	29.59	23.36	34.47	8.44	18.13	8.48	13.93
128	58.18	68.2	20.47	32.6	9.59	18.32	6.93	11.71
256	43.99	72.48	21.3	31.35	7.56	16.07	6.77	11.16
512	48.23	74.78	18.37	28.64	6.39	15.76	6.23	10.72
1024	45.64	66.12	19.36	29.75	5.71	14.1	5.38	10.5
2048	42.7	58.37	15.03	24.68	5.69	13.25	4.57	10.45

Table 6: Table reporting standard deviation for our method when $h = 5, 10, 20$ under DGP and NDGP with $P = 64, 128, 256, 512, 1024, 2048$. The mean computational cost is reported in the first row with the name of the algorithm.

for $n \in [1 : N]$ and with $\beta_\lambda, \beta_\gamma \in \mathbb{R}^d$ and $\rho \in \mathbb{R}_+$. In this model, we have individual-specific probabilities of infection and recovery and a homogeneous latent period of $1/\rho$. Outcomes from the model are shown in figure 12.

The full table on the marginal likelihood standard deviation is reported in table 6. Complete marginal likelihood contour plots are reported in figure 13 and figure 14 in log scale and normalized to have their max in zero.

Sensitivity to $\mathbf{q}$ is measured as for the SIS case, with data generated with $\mathbf{q} = [0, 0, 0.4, 0.6]^{\mathrm{T}}$ and our algorithm is run using $P = 512$ and $\mathbf{q} = [0, 0, i, i]^{\mathrm{T}}$ with $i \in [0.1, \ldots, 0.9]$. Figure 15 reports the results for different choices of h. As expected an increase in h is associated with a smaller standard deviation and a bigger mean ESS. It is also important to mention small values of h are more likely to fail even if they are close to the DGP, see $h = 5$, while higher values of h report $-\infty$ when they are far from the DGP, see $h = 50$ on the values $0.1, 0.9$.

Department of Mathematics and Statistics, Lancaster University, UK

E-mail: l.rimella@lancaster.ac.uk

Department of Mathematics and Statistics, Lancaster University, UK

E-mail: c.jewell@lancaster.ac.uk

Department of Mathematics and Statistics, Lancaster University, UK

E-mail: p.fearnhead@lancaster.ac.uk

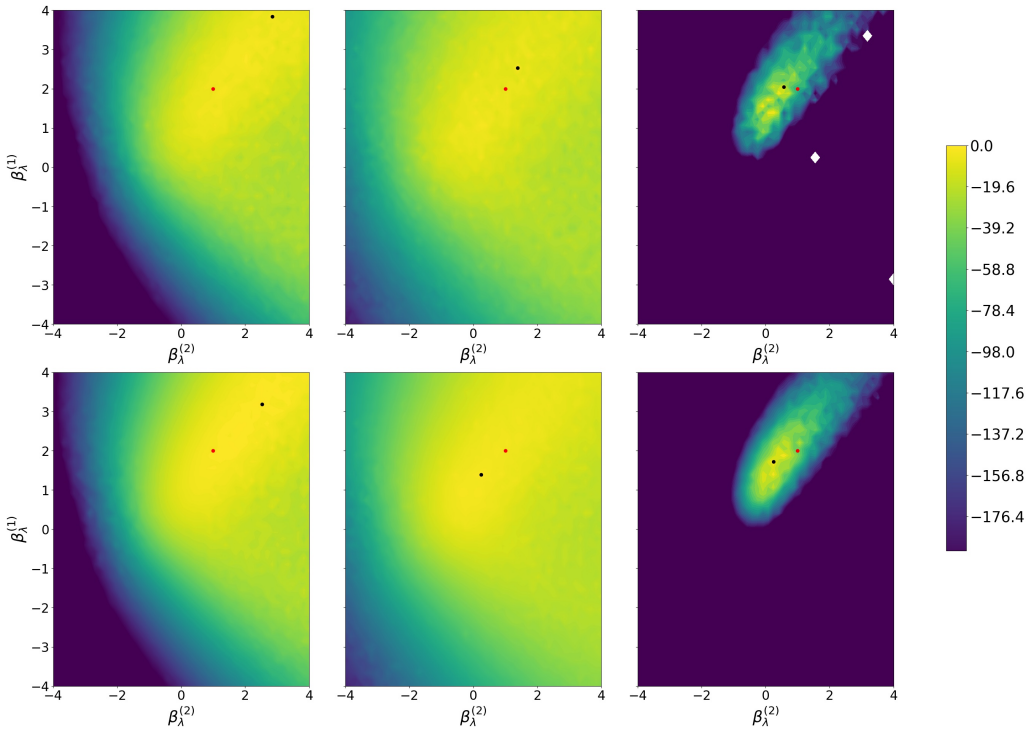

Figure 13: Marginal likelihood contour plot on a β_λ grid in log-scale. Columns refer to $t = 25, 50, 100$ from left to right. Rows refer to $h = 10, 20$ from top to bottom. The colorbar is common across the plots and in each plot, the maximum is set to 0. In red is the DGP and in black is the MLE on the grid.

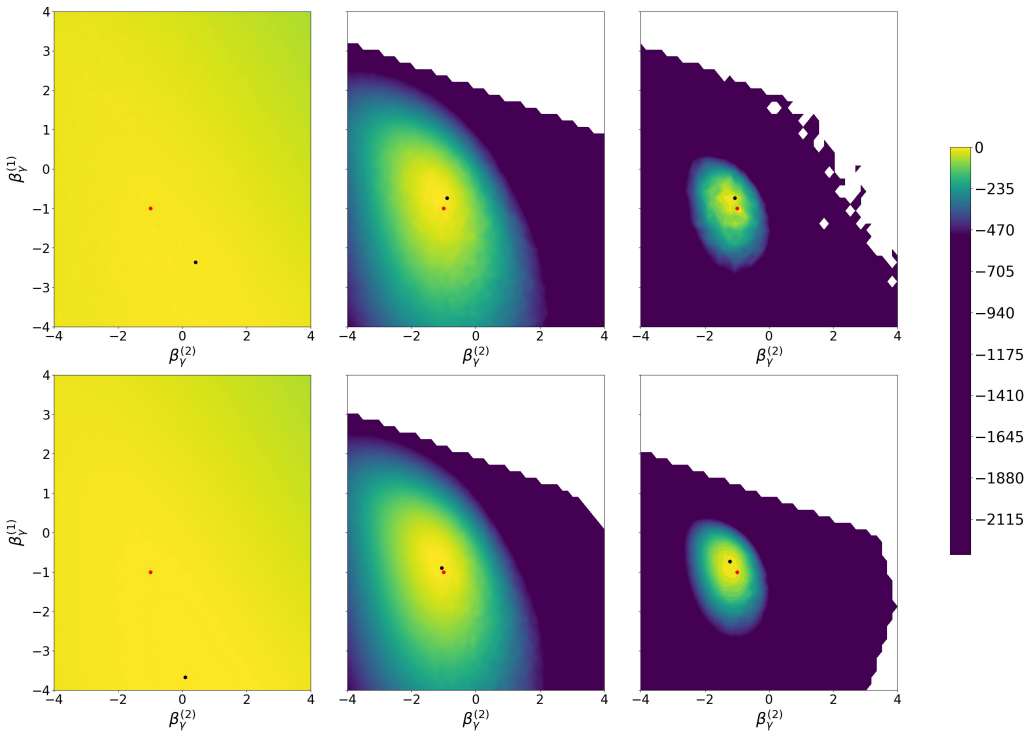

Figure 14: Marginal likelihood contour plot on a β_γ grid in log-scale. Columns refer to $t = 25, 50, 100$ from left to right. Rows refer to $h = 10, 20$ from top to bottom. The colorbar is common across the plots and in each plot, the maximum is set to 0. In red is the DGP and in black is the MLE on the grid.

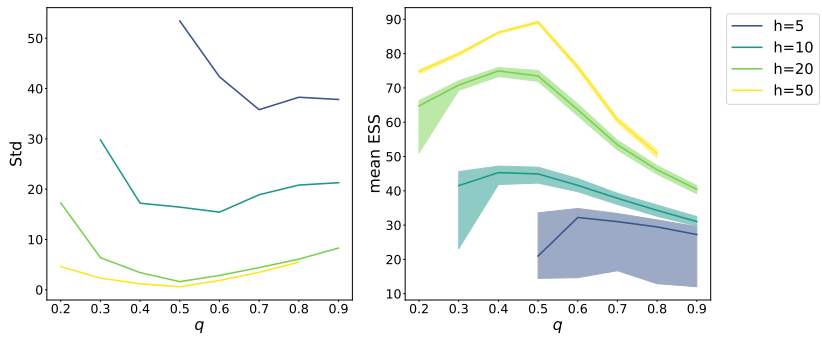

Figure 15: Effective sample size and marginal likelihood standard deviation sensitivity to **q**. The first column shows the marginal likelihood standard deviation, while the second column reports ESS where mode, 5%-quantile, and 95%-quantile are reported.