HIV with contact-tracing: a case study in Approximate Bayesian Computation

Michael Blum, Viet Chi Tran

HIV with contact-tracing: a case study in Approximate Bayesian Computation

Michael G.B. Blum[1], Viet Chi Tran[2]

[1] CNRS, Laboratoire TIMC-IMAG, Université Joseph Fourier, Grenoble, France

[2] Laboratoire P. Painlevé, Université Lille 1, France

July 13, 2009

Abstract

Statistical inference with missing data is a recurrent issue in epidemiology where the infection process is only partially observable. In this paper, Approximate Bayesian Computation, an alternative to data imputation methods such as Markov Chain Monte Carlo integration, is proposed for making inference in epidemiological models. This method of inference is not based on the likelihood function and relies exclusively on numerical simulations of the model. ABC consists in computing a distance between simulated and observed summary statistics and weighting the simulations according to this distance. We propose an original extension of ABC to path-valued summary statistics, corresponding to the cumulated number of detected individuals as a function of time. In a simple SIR model, we show that the posterior distributions obtained with ABC are similar to those obtained with MCMC. When detection times are binned or noisy, we introduce a vector of summary statistics for which several variants of the ABC can be applied. In a refined SIR model well-suited to the HIV contact-tracing program in Cuba, we perform a comparison between ABC with full and with binned data. The last section deals with the analysis of the Cuban HIV-AIDS data. We evaluate the efficiency of the detection system, and predict the evolution of the HIV-AIDS disease in the forthcoming years. We show in particular that the percentage of undetected infectious individuals among the contaminated population might be of the order of 40%.

Keywords: mathematical epidemiology, stochastic SIR model, contact-tracing, unobserved infectious population, Approximate Bayesian Computation, HIV-AIDS epidemics.
AMS Subject Classification: 92D30, 62F15, 62M05, 62N02, 60K99.

1 Introduction

Mathematical modelling in epidemiology plays an important role for understanding and predicting the spread of diseases, as well as for comparing and evaluating public health policies. It has been emphasized in the literature (*e.g.* [6, 33]) that although deterministic modelling can be a guide for describing epidemics, stochastic models have their importance in featuring realistic processes and in quantifying confidence in parameters estimates and

prediction uncertainty. Standard mathematical models in epidemiology consist in compartmental models, in which the population is structured in different classes composed of the susceptible, infectious and removed individuals (SIR models [31]). Parameter estimation for SIR models is usually a difficult task because of missing observations, which is a recurrent issue in epidemiology (*e.g.* [6, 24, 35, 36, 43]). Indeed, the infected population may be partially observed and the infection times may be missing. The computation of the likelihood in this context is numerically infeasible because it involves integration over all unobserved infection events.

Markov Chain Monte Carlo (MCMC) methods, that treat the missing data as extra parameters, have thus become increasingly popular for calibrating stochastic epidemiological models with missing data (*e.g.* [11, 21, 35, 36]). However, MCMC may be computationally prohibitive for high-dimensional missing observations (*e.g.* [12, 41]) and fine tuning of the proposal distribution is required for efficient MCMC algorithms [25]. In this paper, we show that SIR models with missing observations can be calibrated with the Approximate Bayesian Computation (ABC) approach, an alternative to MCMC, originally proposed for making inference in population genetics [37, 5]. This approach is not based on the likelihood function. It relies solely on numerical simulations of the model and comparisons between simulated and observed summary statistics. The idea of using simulations in models for which the distribution theory is intractable has been pioneered by Diggle and Gratton [16] in a frequentist setting. Interestingly, Silverman [39] in the discussion following their paper anticipated that compartmental models might constitute applications of the likelihood free approach.

The current work is motivated by the study of the Cuban HIV-AIDS database [14] that contains the dates of detection of the 8,662 individuals that have been found to be HIV positive in Cuba between 1986 and 2007. The database contains additional covariates including the manner by which an individual has been found to be HIV positive. The individuals can be detected either by the so-called *random screening* or *contact-tracing* methods. The latter is the mode of detection by which a person, that is found to be HIV positive, is invited to give the names of her/his sexual partners (*e.g.* [28]) so that they can in turn take a detection test. As usual for infectious disease data, the total number of infectious individuals as well as the infection times are unknown. Only data relative to the detected individuals are available.

The paper is organized as follows. In Section 2, we introduce the stochastic SIR model with contact-tracing that has been developed by Clémençon *et al.* [13]. Section 3 is devoted to ABC methods when detection times are assumed to be exactly observed. We propose an original extension of ABC to path-valued summary statistics consisting of the cumulated number of detections as a function of time. For this particular choice of summary statistics, ABC targets the true posterior distribution, and we compare numerically the posterior distributions obtained with ABC and MCMC. Section 4 deals with possibly noisy or binned detection times. When the previous path-valued statistics are unavailable, we introduce instead a finite-dimensional vector of summary statistics. We compare the statistical properties of various point estimates and credibility intervals obtained with the two different types of data. Finally, Section 5 concentrates on the analysis of the database for HIV-AIDS in Cuba. We address several questions concerning the dynamic of this epidemic: what is the percentage of the epidemic that is known [15, 28], or equivalently what is the efficiency of the detection system in Cuba [14]; how many new cases of HIV are expected in the forthcoming years; and what is the proportion of detections

that is expected in the contact-tracing program.

2 A stochastic SIR model for HIV-AIDS epidemics with contact-tracing

In this work, we restrict our study to the sexually-transmitted epidemic of HIV in Cuba. It has been inferred that 90% of the seropositive individuals have contracted the disease by sexual contacts [28]. For modelling the dynamics of the number of known and unknown HIV cases, we consider the SIR-type model developed in [13]. The population is divided into three main classes S, I and R corresponding to the *susceptible*, *infectious*, and *detected* individuals considered as *removed* because we assume that they do not transmit the disease anymore (see Figure 1). The population of the susceptible individuals, of size S_t, at time $t > 0$, consists of the sexually active seronegative (healthy) individuals. Individuals immigrate into the class S with a rate λ_0 and leave it by dying/emigrating, with rate $\mu_0 S_t$, or by becoming infected. The class of infectious individuals, of size I_t at time $t > 0$, corresponds to the seropositive individuals who have not taken a detection test yet and may thus contaminate new susceptible individuals. We assume that each individual may transmit the disease to a susceptible individual at rate λ_1 so that the total rate of infection is equal to $\lambda_1 S_t I_t$. Individuals leave the class I when they die/emigrate with a total rate of $\mu_1 I_t$, or when they are detected to be HIV positive.

The class R of the detected individuals, of size R_t at time t, is subdivided into two subclasses whether their seropositivity has been revealed by random screening or contact-tracing. As already mentioned, contact-tracing in Cuba consists in testing the sexual contacts of detected individuals [28]. In the following, we denote by R_t^1 (resp. R_t^2) the size, at time $t > 0$, of the population of removed individuals detected by random screening (resp. contact-tracing). We assume that the total rate of detection by random screening is $\lambda_2 I_t$. Concerning contact-tracing detection, the model shall capture the fact that the contribution of a removed individual to the rate of detection depends on the time elapsed since she/he has been found to be HIV positive. In the sequel, we will consider the two following expressions for the total rate of contact-tracing detection

$$\lambda_3 I_t \sum_{i \in R} \Psi(t - T_i) \quad \text{and} \quad \lambda_3 I_t \sum_{i \in R} \Psi(t - T_i)/(I_t + \sum_{i \in R} \Psi(t - T_i)), \tag{2.1}$$

where Ψ is a positive function and T_i denotes the time at which a removed individual i has been detected. The weight function Ψ determines the contribution of a removed individual i to the contact-tracing control according to the time $t - T_i$ she/he has been detected. In the following we will restrict our analysis to $\Psi(t) = e^{-ct}$, for $c > 0$, so that the contribution of a removed individual to the contact-tracing control decreases exponentially with the time elapsed since she/he has been detected. The first rate in (2.1) corresponds to a mass action principle, and is proportional to the sum of the contributions of detected individuals. The second rate in (2.1) corresponds to a model with frequency dependence. Further details and examples of more general infection and detection rates can be found in [13] as well as a more elaborate mathematical definition of the model based on random measures.

In the following we denote by $\theta = (\mu_1, \lambda_1, \lambda_2, \lambda_3, c)$ the multivariate parameter of the model. When there is no missing observation, Clémencon *et al.* [13] studied the maximum

likelihood estimators and established consistency and asymptotic normality.

2.1 Connection between the stochastic and the deterministic SIR model

In this section, we focus on the first contact-tracing detection rate that is proposed in equation (2.1) and which corresponds to a mass action principle. Similar results can be obtained for the second rate of equation (2.1). The evolution of the SIR process introduced in Section 2 can be represented by Stochastic Differential Equations (SDEs) driven by Poisson point measures and which describe the population at an individual level. These SDEs are given and studied in [13]. To link the stochastic SIR process with the classical equations of epidemiology [31], Clémençon *et al.* [13] show that in a large population renormalization, the individual-centered SIR process converges to the solution of the following system of PDEs

$$\left\{\begin{array}{lcl} \frac{ds_t}{dt} & = & \lambda_0 - \mu_0 s_t - \lambda_1 s_t i_t \\ \frac{di_t}{dt} & = & \lambda_1 s_t i_t - (\mu_1 + \lambda_2) i_t - \lambda_3 i_t \int_{\mathbb{R}_+} \Psi(a)\rho_t(a)\, da \\ \frac{\partial \rho_t}{\partial t}(a) & = & -\frac{\partial \rho_t}{\partial a}(a) \\ \rho_t(0) & = & \lambda_2 i_t + \lambda_3 i_t \int_{\mathbb{R}_+} \Psi(a)\rho_t(a)\, da \end{array}\right. \tag{2.2}$$

where s_t, and i_t denote the size of the susceptible and infectious populations at time $t \geq 0$, and $\rho_t(a)$ denotes the density of individuals having been detected since a time a at time t ($0 \leq a \leq t$). This PDE system with age provides an alternative to delay equations (*e.g.* [32, 46, 47]) and discrete stage structured models (*e.g.* [30]). With the exponential form for Ψ, the PDE system reduces to the following ODE

$$\left\{\begin{array}{lcl} \frac{ds_t}{dt} & = & \lambda_0 - \mu_0 s_t - \lambda_1 s_t i_t \\ \frac{di_t}{dt} & = & \lambda_1 s_t i_t - (\mu_1 + \lambda_2) i_t - \lambda_3 i_t r_t \\ \frac{dr_t}{dt} & = & \lambda_2 i_t + \lambda_3 i_t r_t - c r_t \end{array}\right. , \tag{2.3}$$

where $r_t = \int_0^t e^{-ca}\rho_t(a)\, da$, so that r_t measures, at time t, the contribution of the removed individuals to the rate of detection by contact-tracing.

Apart from the inherent stochastic nature of epidemic propagation that has already been pointed out, and that may be particularly important for small populations (see *e.g.* [19]), considering a stochastic SIR model rather than its deterministic counterpart can present at least two important advantages for parameter calibration. First, it is quite straightforward to perform exact simulations from the stochastic model (see Section 2.2) and this is one motivation for considering ABC methods. Second, the individual-centered stochastic process suits the formalism of statistical methods, which are based on samples of individual data. Within this formalism, problems such as missing or noisy data can be tackled with the arsenal of statistical methods. Since the estimates of the stochastic process converge to the parameters of the PDEs (see [13]), this provides new alternative approaches for calibrating the parameters of PDEs and ODEs (*e.g.* [2, 3, 10, 33]).

2.2 Exact path simulation of the SIR model with contact-tracing

The main difficulty for simulating the SIR model with contact-tracing lies in the fact that the rate of detection by contact-tracing evolves with time. Here, we consider an

acceptance-rejection technique for simulating the SIR-type process between time 0 and the end of the observation period at time $t = T$ (see [18, 20, 45] for similar algorithms).

In order to simplify the algorithm, we assume that the values of the parameter λ_0 and μ_0 are such that the size of the population of susceptible individuals remains constant during the observation period.

The algorithm can be described iteratively as follows:

1. Start with a population of $S = S_0$ susceptible individuals, $I = I_0$ infectious individuals, no detected individuals, and a vector of ages since detection set to the null vector.

2. Assume that we have already simulate k events, and that the k^{th} event occurs at time t_k. We describe, in the following, how to simulate the $k+1^{\text{th}}$ event. Let $\tau = t_k$ be the current time.

   (a) Simulate an independent exponential random variable $\mathcal{E}$ with parameter

   $$C_k = \lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k} + \lambda_3 I_{t_k} R_{t_k}$$

   which is an upper bound for the sum of the rates of occurrence of all possible events. The time of the next putative event is defined as $\tau' = \tau + \mathcal{E}$.

   (b) Increment the ages by $\mathcal{E}$

   (c) If $\tau' > T$, then stop the simulation process.
   Else, simulate an independent uniform random variable U in $[0, C_k]$.
   If $0 \leq U < \lambda_1 S_{t_k} I_{t_k}$ then a susceptible individual is removed, and an infectious individual is added.
   If $\lambda_1 S_{t_k} I_{t_k} \leq U < \lambda_1 S_{t_k} I_{t_k} + \mu_1 I_{t_k}$ then an infectious individual is removed.
   If $\lambda_1 S_{t_k} I_{t_k} + \mu_1 I_{t_k} \leq U < \lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k}$ then an infectious individual is removed, an individual detected by random screening is added, and a zero is added to the vector of ages since detection.
   If $\lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k} \leq U < \lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k} + \lambda_3 I_{t_k} \sum_{i \in R} \Psi(\tau' - T_i)$ then an infectious individual is removed, an individual detected by contact-tracing is added, and a zero is added to the vector of ages since detection.
   Else, nothing happens. Return to step 2a with the current time set equal to τ'.

The complexity of the algorithm scales with the total number of events, so that it depends on the parameter vector θ. When simulating the Cuban HIV epidemic, $t = 0$ corresponds to the beginning of the epidemic in 1986 and the simulations are performed until the end of the observation period at time $T = 21.5$, in July 2007.

3 ABC with sufficient summary statistics for epidemic models

We start this section by giving arguments to motivate Bayesian analysis in epidemiology. Then, we introduce the main principles of ABC and we give the standard rejection ABC algorithm. The census HIV data in Cuba are presented as well as the summary statistics

that will be considered in the ABC algorithm. In this section, we assume that the curves of the cumulated numbers of removed individuals, as functions of time, are available and we use these curves as path-valued summary statistics. The section ends with a comparison of the posterior distributions obtained with MCMC and ABC methods.

Bayesian approaches have been widely used in epidemiology (*e.g.* [12, 11, 23, 24, 36, 42, 43]) and the reasons to adopt a Bayesian approach are manyfold. First, Bayesian methods offer a convenient way to handle missing observations, such as unobserved infection times and unknown number of infectious individuals. A second reason is that epidemiological models may contain nuisance parameters that should be integrated over when making inference. In the context of the SIR model with contact-tracing, the parameters c and μ_1 are considered as the nuisance parameters whereas the parameters of interest are λ_1, λ_2, and λ_3. Thirdly, high dimensional integration, typically involved in Bayesian methods, might be more convenient than high dimensional optimization when the likelihood function is flat in certain directions. A last reason is that Bayesian algorithms usually provide samples from the posterior distribution from which it is straightforward to obtain credibility intervals, whereas frequentist confidence intervals may be based on asymptotic derivations that might be poor approximations when the sample is too small or highly correlated. However, we lay absolutely no claim that Bayesian methods are the most relevant methods for SIR-type models and applications of the frequentist likelihood-free method of Diggle and Gratton [16], for instance, might be of great interest in this setting.

3.1 Sampling from the posterior

Let us sum up the main principle of ABC. For simplicity, we deal here with densities, but the following description also holds when dealing with measures that are not absolutely continuous w.r.t. the Lebesgue measure. Let $\mathbf{x}$ be the available data and $\pi(\theta)$ be the prior density. Two approximation steps are at the core of ABC.

Replacing observations with summary statistics First, instead of focusing on the posterior density $p(\theta \,|\, \mathbf{x})$, ABC aims at a possibly less informative *target* density $p(\theta \,|\, S(\mathbf{x}) = s_{obs}) \propto \Pr(s_{obs}|\theta)\pi(\theta)$ where S is a summary statistic that takes its values in a normed space, and s_{obs} denotes the observed summary statistic. The summary statistic S can be a d-dimensional vector or an infinite-dimensional variable such as a L^1 function. Of course, if S is sufficient, then the two conditional densities are the same and no approximation is involved in this first step. In the following, the target distribution will also be coined as the *partial posterior distribution*.

Simulation-based approximations of the posterior Once the summary statistics have been chosen, the second approximation arises when estimating the partial posterior density $p(\theta \,|\, S(\mathbf{x}) = s_{obs})$ and sampling from this distribution. This step involves non-parametric kernel estimation and possibly correction refinements due to Beaumont *et al.* [5] and Blum and François [8]. These corrections are given in Section 4.2.

Let us now describe how the ABC method with smooth rejection [5] generates random draws from the target distribution. We assume here that the summary statistics s_{obs} have been calculated for the data. The algorithm can be described as follows

1. Generate N random draws (θ_i, s_i), $i = 1 \dots N$, where θ_i is generated from the prior distribution and where s_i is the vector of summary statistics calculated for the i^{th} synthetic data set, simulated from the generative model with parameter θ_i (see Section 2.2).

2. Associate with the i^{th} simulation the weight $W_i = K_\delta(s_i - s_{obs})$, where δ is a tolerance threshold and K_δ a (possibly multivariate) smoothing kernel.

3. Then the distribution $(\sum_{i=1}^N W_i \delta_{\theta_i})/(\sum_{i=1}^N W_i)$, in which δ_θ denotes the Dirac mass at θ, approximates the target distribution. In other words, the resulting weighted sample (θ_i, W_i) forms a sample from a distribution close to the partial posterior distribution $p(\theta \,|\, S(\mathbf{x}) = s_{obs})$.

When K_δ is a function that takes 0 or 1 values, the ABC (non-smooth) rejection algorithm simply consists in keeping the simulations for which the simulated summary statistics are close enough to the observed ones (see [37]).

3.2 Point estimation

Once a sample from the target distribution has been obtained, several estimators may be considered for point estimation of each one-dimensional parameter λ_j, $j = 1, 2, 3$. We will consider here the means, medians and modes of the marginal posterior distributions.

- Using the weighted sample $(\lambda_{j,i}, W_i)$, $i = 1 \dots N$, the *mean* of the target distribution $p(\lambda_j | s_{obs})$ is estimated by

$$\widehat{\lambda_j} = \frac{\sum_{i=1}^N \lambda_{j,i} W_i}{\sum_{i=1}^N W_i} = \frac{\sum_{i=1}^N \lambda_{j,i} K_\delta(s_i - s_{obs})}{\sum_{i=1}^N K_\delta(s_i - s_{obs})}, \quad j = 1, 2, 3 \tag{3.1}$$

  which is the well-known Nadaraya-Watson regression estimator of the conditional expectation $\mathbb{E}(\lambda_j \,|\, s_{obs})$ [34, 48].

- The *medians* of the marginal target distributions are estimated by computing the medians of a non-weighted sample that has been obtained by sampling with replacement in the weighted sample (θ_i, W_i), $i = 1 \dots N$.

- The *modes* are estimated by maximizing the estimates $\widehat{p}(\lambda_j \,|\, s_{obs})$ of the marginal distributions of λ_j, $j = 1, 2, 3$, obtained as

$$\widehat{p}(\lambda_j \,|\, s_{obs}) = \frac{\sum_{i=1}^N K_\Delta(\lambda_{j,i} - \lambda_j) K_\delta(s_i - s_{obs})}{\sum_{i=1}^N K_\delta(s_j - s_{obs})}, \quad \lambda_j > 0, \quad j = 1, 2, 3 \tag{3.2}$$

  where Δ is a bandwith parameter for the density estimation. It can be seen that formula (3.2) corresponds to a standard kernel smoothing for conditional density estimation (*e.g.* [27]).

For computing the 95% credibility intervals, we estimate the 97.5% and the 2.5% quantiles of the marginal target distributions in the same manner as the medians.

3.3 Data and choice of summary statistics

Data Starting at the time of the first detection in 1986, the Cuban HIV-AIDS data consists principally of the detection times at which the individuals have been found to be HIV positive. At the time of the last detection event that is considered, in July 2007, there is a total of 8,662 individuals in the database. For each detection event, there is a label indicating if the individual has been detected by random screening or contact-tracing.

Summary statistics We consider in this section the two (infinite-dimensional) statistics $(R^1_t, t \in [0,T])$ and $(R^2_t, t \in [0,T])$. Because the data consists of the detection times for the two different types of detection, these two statistics can simply be viewed as a particular coding of the whole dataset. This implies that these two statistics are sufficient with respect to the parameter θ and that the partial posterior distribution is the same as the posterior distribution $p(\theta \,|\, \mathbf{x})$.

The L^1 norm between the simulated paths R^l_i $(l = 1, 2,\ i = 1 \dots N)$ and the observed ones R^l_{obs} $(l = 1, 2)$ is

$$\|R^l_{obs} - R^l_i\|_1 = \int_0^T |R^l_{obs,s} - R^l_{i,s}|\, ds \quad ,\ l = 1, 2,\ i = 1 \dots N. \tag{3.3}$$

For computing the weights W_i, we choose a product kernel so that $W_i = K_{\delta_1}(\|R^1_{obs} - R^1_i\|_1) K_{\delta_2}(\|R^2_{obs} - R^2_i\|_1)$ where δ_1, δ_2 are 2 possibly different tolerance thresholds. An Epanechnikov kernel is considered for both K_{δ_1} and K_{δ_2}. In practice, rather than dealing with a tolerance threshold δ, we set a tolerance rate P_δ that corresponds to the percentage of accepted simulations. Here, we consider the same tolerance rates $P_{\delta_1} = P_{\delta_2}$ for the two summary statistics. This amounts to choosing the P_{δ_1} (resp. P_{δ_2}) quantile of the distances $\|R^1_{obs} - R^1_i\|_1$ (resp. $\|R^2_{obs} - R^2_i\|_1$), $i = 1 \dots N$, for the tolerance threshold δ_1 (resp. δ_2).

There is a correspondence between ABC and deterministic approaches in which parameter calibration relies on the minimization of a *cost* function (*e.g.* [3, 10]). When considering Epanechnikov kernels for K_{δ_1} and K_{δ_2}, the value of the θ_i's that maximizes the weights W_i's maximizes $(1 - \|R^1_{obs} - R^1_i\|_1/\delta_1)^2 (1 - \|R^2_{obs} - R^2_i\|_1/\delta_2)^2$. For small differences between the simulated and observed paths, this amounts to minimizing the cost function $\|R^1_{obs} - R^1_i\|_1/\delta_1 + \|R^2_{obs} - R^2_i\|_1/\delta_2$.

3.4 Comparison between ABC and MCMC methods for a standard SIR model

Even for standard SIR model, the posterior distributions have no explicit expressions. Following Beaumont *et al.* [5] a performance indicator for ABC techniques consists in their ability to replicate likelihood-based results given by MCMC when the latter can be obtained. Here the situation is particularly favorable for comparing the two methods since the partial and the full posterior distributions are the same so that both methods aim at sampling from the same distribution. In the following examples, we choose samples of small sizes ($n = 3$ and $n = 29$) so that the dimension of the missing data is reasonable and MCMC achieves fast convergence. For large sample sizes with high-dimensional missing data, MCMC convergence might indeed be a serious issue ([12, 41]).

We consider the standard SIR model which corresponds to setting $\lambda_3 = 0$ in the model of Section 2. We choose gamma distributions for the priors of λ_1 and λ_2 with shape pa-

rameters of 0.1 and rate parameters of 1 and 0.1 respectively. The data consists of the detection times and we assume that the infection times are unobserved. We implement the MCMC algorithm of O'Neill and Roberts [36] that is based on Gibbs and Metropolis-Hastings updating rules. A total of $10,000$ steps are considered for MCMC containing an initial burn-in of $5,000$ steps. For ABC, the summary statistic consists of the cumulative number of detections as a function of time. We use an L^1 distance (equation (3.3)) for comparing the simulated trajectories and the observed one. A total of $100,000$ simulations are performed for ABC.

The first example was previously considered by O'Neill and Roberts [36]. They simulated detection times by considering one initial infectious individual and by setting $S_0 = 9$, $\lambda_1 = 0.12$, and $\lambda_2 = 1$. The resulting detection times are then shifted so that the first detection is found at time 0. They obtained the three following times $t_1 = 0$, $t_2 = 1.52292$, $t_3 = 1.15504$, and the end of the observation period was $T = 1.7$. As displayed by Figure 2, the posterior distributions obtained with ABC are extremely closed to the ones obtained with MCMC provided that the tolerance rate is sufficiently small. Figure 2 clearly shows that the tolerance rate may change importantly the posterior distribution obtained with ABC (see the posterior distributions for λ_2).

In a second example, we simulate a standard SIR trajectory with $\lambda_1 = 0.12$, $\lambda_2 = 1$, $S_0 = 30$ and $I_0 = 1$. The data now consists of 29 detection times during the time interval $[0, T]$ with $T = 5$ (the data are given in the Supplementary Material). Once again, the results (see Figure 2b) show that the ABC posteriors are close to the MCMC posteriors provided that the tolerance rate is small enough. We also note that ABC tends to produce posterior distributions with larger tails compared to MCMC, especially for the largest tolerance rate. This can be explained by considering the extreme scenario in which the tolerance threshold δ goes to infinity. This amounts to giving a weight of 1 to all the simulations so that ABC targets the prior distribution instead of the posterior. Because the prior has typically larger tails than the posterior, ABC may inflate the tails for tolerance rates that are too large (see also [5]).

4 Comparison between ABC with full and binned data

When there is noise in the detection times or when the detection times have been binned, the full observations $(R^1_t, t \in [0,T])$ and $(R^2_t, t \in [0,T])$ may be unavailable. In this section, we replace these summary statistics by a vector of summary statistics such as the numbers of detections per year during the observation period. Since these new summary statistics are not sufficient anymore, the new partial posterior distribution may be different from the standard posterior $p(\theta \,|\, \mathbf{x})$. In the following, we compare point estimates and credibility intervals obtained from ABC procedures with full and binned data. Additionally, we investigate the effect of the threshold δ on the estimated partial posterior distribution.

4.1 A new set of summary statistics

We capture the data with a d-dimensional vector consisting of three different types of statistics. First, we compute the numbers R^1_T and R^2_T of individuals detected by random

screening and contact-tracing by the end of the observation period, after $T = 21.5$ years. Second, for each year j, we compute the numbers of individuals that have been found to be HIV positive $R^l_{j+1} - R^l_j$, $l = 1, 2$. Last, we take advantage of the reconstruction of the infectious times that has been performed by Clemençon *et al.* [13] for the six first years of the epidemic. This reconstruction is based on a preliminary estimation of the incubation time between infection and AIDS, and relies on the assumption that almost every infectious individual before 1992 has been detected and has developed AIDS by 2007. This allows to reconstruct 150 missing infection times out of 848 observations between 1986 and 1992. Here we assume that this reconstruction is exact. Based on this reconstruction, it is possible to compute the third type of summary statistics consisting of the numbers of new infectious for each of the the sixth first years $I_{j+1} - I_j$ for $j = 0, \ldots, 5$, as well as the mean time during which an individual is infected but has not been detected yet. This mean time corresponds to the mean sojourn time in the class I for the sixth first years of the epidemic.

In order to compute the weights $W_i, i = 1 \ldots N$, required for performing the algorithm with smooth rejection of Section 3, we consider the following spherical kernel $K_\delta(x) \propto K(\|\mathbf{H}^{-1}x\|/\delta)$. Here K denotes the one-dimensional Epanechnikov kernel, $\|.\|$ is the Euclidian norm of $\mathbb{R}^d$ and $\mathbf{H}^{-1}$ a matrix. Because the summary statistics may span different scales, $\mathbf{H}^{-1}$ is taken equal to the diagonal matrix with the inverse of the variance of each one-dimensional summary statistic on the diagonal [5]. Concerning the tolerance threshold δ, we again rather set a tolerance rate P_δ that corresponds to the percentage of accepted simulations.

4.2 Curse of dimensionality and regression adjustments

In the case of a d-dimensional vector of summary statistics, it is known (*e.g.* [9, 17]) that when $N \to +\infty$, the estimator of the conditional mean (equation (3.1)) is convergent if the tolerance rate satisfies $\lim_{N\to+\infty} \delta_N = 0$, so that its bias converges to 0, and $\lim_{N\to+\infty} N\delta_N^d = +\infty$, so that its variance converges to 0. As d increases, a larger tolerance threshold shall be chosen to keep the variance constant. This implies that the bias of the estimators may increase as the number of summary statistics is increased. This phenomenon known as the *curse of dimensionality* [27] may be a serious issue for the ABC-rejection approach. The following paragraph presents the correction originally introduced by Beaumont *et al.* [5] and refined by Blum and François [8] to cope with the curse of dimensionality.

For large thresholds δ, the rejection method may retain couples (θ_i, s_i) with summary statistics far from s_{obs} meaning heuristically that the associated θ_i's may not be considered as random draws from the distribution $p(\theta \,|\, s_{obs})$ anymore. To overcome this fact, Beaumont *et al.* [5] adjusted the θ_i's in (3.1) and (3.2) so that the corrected values denoted by θ_i^* are random draws close to the partial posterior $p(\theta|s_{obs})$. Doing this, they found that the resulting target distributions were numerically insensitive to the tolerance threshold δ for a large range of small enough values of δ.

We present here the adjustment principle in a general setting within which the corrections of [5] and [8] can be derived. Correction adjustments aim at obtaining from a random couple (θ_i, s_i) a random variable distributed according to $p(\theta \,|\, s_{obs})$. The idea is to construct a coupling between the distributions $p(\theta|s_i)$ and $p(\theta|s_{obs})$, through which we can shrink the simulations θ_i, $i = 1 \ldots N$ to a sample of i.i.d. draws from $p(\theta|s_{obs})$. In

the following, we describe how to perform the corrections for each of the one-dimensional components separately. In the remaining part of subsection 4.2, θ_i will thus denote an arbitrary one-dimensional component of the vector of parameters. Correction adjustments are obtained by assuming a relation $\theta = G(s, \varepsilon) =: G_s(\varepsilon)$ between parameters and summary statistics, where G is a (possibly complicated) function and ε a r.v. with a distribution that conditionally to s does not depend on s. A possibility is to choose $G_s = F_s^{-1}$ the (generalized) inverse of the cumulative distribution function of $p(\theta|s)$. In this case, $\varepsilon = F_s(\theta)$ is a uniform r.v. on $[0, 1]$. Then, the corrected sample is:

$$\theta_i^* = F_{s_{obs}}^{-1}(F_{s_i}(\theta_i)) \quad i = 1 \dots N. \tag{4.1}$$

The fact that the θ_i^*'s are i.i.d. with density $p(\theta|s_{obs})$ arises from the standard inversion algorithm. For other choices of G, provided G_s^{-1} is properly defined, $\theta_i^* = G_{s_{obs}}(G_{s_i}^{-1}(\theta_i))$.

Of course, the function G is unknown and shall be approximated in practice. As a consequence, the adjusted i.i.d. simulations θ_i^*, $i = 1 \dots N$, constitute an approximate sample of $p(\theta \mid s_{obs})$. The ABC algorithm with regression adjustment is as follows

1. Simulate, as Step 1 of the rejection algorithm, a sample $(\theta_i, s_i)_{i \in [\![1,N]\!]}$ of i.i.d. r.v.

2. By making use of the sample of the (θ_i, s_i)'s weighted by the W_i's, approximate the function G such that $\theta_i = G(s_i, \varepsilon_i)$ in the vicinity of s_{obs}.

3. Replace the θ_i's by the corrected θ_i^*'s. The resulting weighted sample (θ_i^*, W_i), $i = 1 \dots N$, form a sample from the target distribution. The weights $W_i = K_\delta(s_i - s_{obs})$ heuristically give less importance to values for which the adjustment has been more important.

Beaumont *et al.* local linear regressions (LOCL) The case where G is approximated by the linear model $G(s, \varepsilon) = \alpha + s^t\beta + \varepsilon$, was considered in [5]. The parameters α and β are inferred by minimizing the weighted least-square $\sum_{i=1}^N K_\delta(s_i - s_{obs})(\theta_i - (\alpha + (s_i - s_{obs})^T\beta))^2$. The estimator $\hat{\alpha}$ corresponds to the estimation of $\mathbb{E}(\theta|s_{obs})$ obtained with standard local polynomial regression (see *e.g.* [17]). In this case, the correction amounts to adjusting the expectation of $p(\theta \mid s_i)$ to $\hat{\alpha}$. Using equation 4.1, the correction of [5] is derived as

$$\theta_i^* = \theta_i - (s_i - s_{obs})^T\hat{\beta}, \; i = 1 \dots N. \tag{4.2}$$

Blum [7] proved the asymptotic consistency of the estimator given by equation (3.2) when the θ_i's are adjusted with the formula (4.2). Interestingly, similar adjustments have been proposed by Hansen [26] and Hyndman et al. [29] in the context of conditional density estimation. More generally, we note that ABC can be viewed from the angle of conditional density estimation, the major difference being that ABC aims at simulating replicates from the conditional density and not at estimating the conditional density.

Blum and François' nonlinear conditional heteroscedastic regressions (NCH) To relax the assumptions of homoscedasticity and linearity inherent to the local linear regression model, Blum and François [8] approximated G by $G(s, \varepsilon) = m(s) + \sigma(s) \times \varepsilon$ where $m(s)$ denotes the conditional expectation $\mathbb{E}(\theta|s)$, and $\sigma^2(s)$ the conditional variance. The estimators $\hat{m}$ and $\log \hat{\sigma}^2$ of the conditional expectation and of the logarithm of the conditional variance are found by adjusting two feed-forward neural networks [38] using

weighted least squares. To motivate the choice of neural networks, Blum and François [8] emphasized that the regression layer is not performed on the (possibly high dimensional) subspace of the summary statistics but on a subspace of lower dimension found via internal projections. For the NCH model, parameter adjustment is performed as follows

$$\theta_i^* = \hat{m}(s_{obs}) + (\theta_i - \hat{m}(s_i)) \times \frac{\hat{\sigma}(s_{obs})}{\hat{\sigma}(s_i)},\ i = 1 \dots N.$$

In practical applications of the NCH model, we train $L = 10$ neural networks for each conditional regression (expectation and variance) and we average the results of the L neural networks to provide the estimates $\hat{m}$ and $\log \hat{\sigma}^2$.

Reparameterization In both regression adjustment approaches, the regressions can be performed on transformations of the responses θ_i rather that on the responses themselves. Parameters whose prior distributions have finite supports are transformed via the logit function and non-negative parameters are transformed via the logarithm function. These transformations guarantee that the θ_i^*'s lie in the support of the prior distribution. The transformations have the additional advantage of stabilizing the variance [1].

4.3 Comparison for synthetic datasets

We now simulate synthetic datasets for which the true values of the parameters are known and we compare the point estimates obtained with the different ABC algorithms. Our purpose is to see if the replacement of a sufficient summary statistic by a vector of non-sufficient summary statistics may still lead to close approximations of the posterior distribution. The quality of the calibration with respect to the choice of the threshold parameter δ is also considered.

We simulate $M = 200$ synthetic data sets for given values of the parameters. In order to work on data similar to the Cuban database for the HIV-AIDS epidemic (see [13]), we choose $\mu_1 = 2 \times 10^{-6}$, $\lambda_1 = 1.14 \times 10^{-7}$, $\lambda_2 = 3.75 \times 10^{-1}$, $\lambda_3 = 6.55 \times 10^{-5}$, and $c = 1$. The initial conditions are set to $S_0 = 6 \times 10^6$, the size of the Cuban population in the age-group 15-49 [14], $I_0 = 232$ and $R_0 = 0$. When analyzing the synthetic data sets, we simulate only 6 years of the epidemics.

We study four variants of ABC for estimating λ_1, λ_2, and λ_3. When using the finite dimensional vector of summary statistics, we perform the smooth rejection approach as well as the LOCL and NCH corrections with a total of 21 summary statistics: the 18 summary statistics corresponding to the yearly increases of R^1, R^2, and I; the final numbers of detected individuals R_6^1 and R_6^2; and the mean sojourn time in the class I. When considering the two trajectories $(R_t^1, t \in [0, T])$ and $(R_t^2, t \in [0, T])$ as the summary statistics, we perform the smooth rejection approach using the product kernel for the computation of the weights W_i. Each of the $M = 200$ estimations of the partial posterior distributions are performed using a total of $N = 5000$ simulations of the SIR model with the mass action principle (first rate in equation (2.1)).

Prior distributions The prior distributions of the parameters $\mu_1, \lambda_1, \lambda_2$ and λ_3 are chosen to be uniform on a log scale. The prior distributions are defined on a log scale to reflect our uncertainty about the order of magnitude of the parameters. More specifically,

the prior distribution for $\log_{10}(\mu_1)$ is $\mathcal{U}(-6,-4)$ where $\mathcal{U}(a,b)$ denotes the uniform distribution on the interval (a,b). The prior distribution for $\log_{10}(\lambda_1)$ is $\mathcal{U}(-9,-6)$, the prior distribution for $\log_{10}(\lambda_2)$ is $\mathcal{U}(-4,3)$ and the prior distribution for $\log_{10}(\lambda_3)$ is $\mathcal{U}(-8,2)$. The bounds of the uniform distributions are set to keep the simulations from being degenerate. The prior for the parameter c is $\log_{10}(2)/\mathcal{U}(1/12,5)$. This prior is chosen so that the half time of Ψ, which measures the contribution to the rate of contact-tracing of a detected individual, is uniformly distributed between 1/12 and 5 years. In ABC, the choice of the prior is of considerable importance since the approximation of the partial posterior, given by equation (3.2), might be loose for very diffuse prior. Blum [7] showed that the variance of the estimator (3.2) is inversely proportional to the probability of targeting *a priori* the observed summary statistics $p(s_{obs}) = \int_\theta p(s_{obs} \mid \theta)\pi(\theta)\, d\theta$.

Point estimates of θ and credibility intervals Figure 3 displays the boxplots of the 200 estimated modes, medians, 2.5% and 97.5% quantiles of the posterior distribution for λ_1 as a function of the tolerance rate P_δ. The corresponding figures for λ_2 and λ_3 can be found in the Supplementary Material.

First, we find that the medians and the modes are equivalent except for the rejection method with 21 summary statistics for which the mode is less biased. For the lowest tolerance rates, the point estimates obtained with the four possible methods are close to the value $\lambda_1 = 1.14 \times 10^{-7}$ used in the simulations, with smaller credibility intervals for the LOCL and NCH variants. When increasing the tolerance rate, the bias of the point estimates obtained with the rejection method with 21 summary statistics slightly increases. By contrast, up to tolerance rates smaller than 50%, the biases of the point estimates obtained with the three other methods remain small. As could be expected, the widths of the credibility intervals obtained with the rejection methods increase with the tolerance rate while they remain considerably less variable for the LOCL and NCH variants.

Mean square error To further investigate the differences between the statistical properties of the different methods, we compute the rescaled mean square errors (RMSEs). The RMSEs are computed on a log scale and rescaled by the range of the prior distribution so that

$$\mathrm{RMSE}(\lambda_j) = \frac{1}{M}\sum_{k=1}^{M} \frac{(\log(\widehat{\lambda}_j^k) - \log(\lambda_j))^2}{\mathrm{Range}(\mathrm{prior}(\lambda_j))^2}, \quad j = 1,2,3, \tag{4.3}$$

where $\widehat{\lambda}_j^k$ is a point estimate obtained with the k^{th} synthetic data set. Figure 4 displays the values of the RMSEs as functions of the tolerance rate. We find that the smallest values of the RMSEs are usually reached for the lowest value of the tolerance rate (but see the RMSEs for λ_2 in Figure 4). For λ_1 and λ_2, the RMSEs of the point estimates obtained with the four different methods are comparable for the lowest tolerance rate. However, the smallest values of the RMSEs are always found when performing the rejection method with the two sufficient summary statistics R^1 and R^2. This finding is even more pronounced for the parameter λ_3. This is due to the fact that the trajectories R^1 and R^2 are the most informative summary statistics for estimating the λ_j, $j = 1,2,3$.

Rescaled mean credibility intervals To compare the whole posterior distributions obtained with the four different methods, we additionally compare the different credibil-

ity intervals. As displayed by Figure 3 and the Figures 1 and 2 of the Supplementary Material, the credibility intervals obtained with the smooth rejection schemes increase importantly with the tolerance rate whereas such an important increase is not observed for the regression approaches. To further compare the 95% credibility intervals obtained with the different methods, we computed the rescaled mean credibility intervals ($RMCI$) defined as follows

$$\text{RMCI} = \frac{1}{M}\sum_{k=1}^{M}\frac{|IC_j^k|}{\text{Range}(\text{prior}(\lambda_\text{j}))}, \quad j = 1, 2, 3, \tag{4.4}$$

where $|IC_j^k|$ is the length of the k^{th} estimated credibility interval for the parameter λ_j.

As displayed by Figure 5, credibility intervals obtained with the NCH method are clearly the thinnest, those obtained with the rejection methods are the widest and those obtained with the LOCL method have intermediate width. We additionally find that the RMCIs obtained with the regression methods also increase with the tolerance rate. This phenomenon is partly due to the increase of the variance of the extreme quantiles that is observed when the tolerance rate increases.

In the following, we perform the NCH correction when considering the finite-dimensional vector of summary statistics. This choice is motivated by the small RMSEs and RMCIs obtained with the NCH method (Figures 4 and 5).

5 Application to the Cuban HIV-AIDS epidemic

In this section, we calibrate the SIR model with contact-tracing of [13] by using the Cuban HIV-AIDS database presented in Section 3.3. We apply two methods: the smooth rejection ABC with the two path-valued summary statistics and the NCH ABC with the vector of summary statistics described in Section 4.2. When considering the Cuban data, with 21 years of observations, the latter consists of a vector of 51 summary statistics.

5.1 Parameter calibration and goodness of fit

To fit the SIR model to the Cuban HIV-AIDS data, we use a total of 100,000 simulations. We consider the two different rates of contact-tracing detection (2.1) and we use the same initial conditions and priors as in Section 4.3.

To set the value of the tolerance rate P_δ for each of the two procedures, we consider the 15^{st} years of the epidemic as the training data set and we choose the value of the tolerance rate P_δ that minimizes the prediction error at the end of the epidemic, which corresponds to $T = 21.5$. The prediction error is defined as

$$\text{Pred Error} = \mathbb{E}_{P_\delta}\left[\frac{|R^1_{21.5}(P_\delta) - R^1_{obs,21.5}|}{R^1_{obs,21.5}} + \frac{|R^2_{21.5}(P_\delta) - R^2_{obs,21.5}|}{R^2_{obs,21.5}}\right] \tag{5.1}$$

where $\mathbb{E}_{P_\delta}$ denotes the expectation with respect to the partial posterior distribution found with a tolerance rate set to P_δ.

For the chosen tolerance rate P_δ, we now investigate the goodness of fit of the SIR-type model. By simulating paths of the SIR model associated with parameters θ sampled from the partial posterior distribution, we check if the model reproduces *a posteriori* the observed summary statistics. In Figure 6, we display the Posterior Predictive Distributions (PPD) (*e.g.* [22]) of $R^1_{21.5}$, $R^2_{21.5}$, I_6, as well as the mean sojourn time in the class I. Figure 6 has been obtained when using the two sufficient summary statistics R^1 and R^2, using the optimal tolerance rates of $P_{\delta_1} = P_{\delta_2} = 1\%$, and considering the model of frequency dependence (second rate in equation (2.1)) As displayed by Figure 6, $R^1_{obs,21.5}$ is close to the mode of the PPD and $R^2_{obs,21.5}$ is smaller than the mode but still contained in the PPD. By contrast, the mean sojourn time in the class I is not contained in the PPD and the observed number of infectious individuals is in the lower tail of the PPD. An explanation might be that an age-structure has to be taken into account for the infection rate in order to capture the non-Markovian effects (*e.g.* [43]). A model with an increasing infectious rate could diminish the mean sojourn time in the class I and increase by compensation the number of infections to maintain the infection pressure constant. When considering the model with a mass action principle (first rate in equation (2.1)), we observed (see Supplementary Material) that the statistic $R^1_{obs,21.5}$ is not contained in the PPD. With a mass action principle, the rate of contact-tracing detection increases linearly with the total contribution of the detected individuals, which may be too rapid in comparison with the observed data.

Concerning the PPDs obtained with the NCH method and the 51 summary statistics, we find that they have extremely wide supports for both the model with a mass action principle and the model with frequency dependence (see Supplementary Material). These large PPDs suggest that the summary statistics measuring the detections and the infections may contain conflicting signals, which results in a large variance of the partial posterior distribution.

In the following, to provide point estimates and credibility intervals, we consider the model with frequency dependence that has been fitted with the two trajectories R^1 and R^2. The point estimates of Table 1 have been obtained by maximizing the estimated marginal posterior distributions.

Parameter	Point estimates	Lower bound of the 95% credibility interval	Upper bound of the 95% credibility interval
λ_1	$5.4\,10^{-8}$	$3.9\,10^{-8}$	$2.3\,10^{-7}$
λ_2	0.13	0.007	1.17
λ_3	0.19	0.03	0.82

Table 1: *Point estimates and 95% credibility intervals for the infection rate* λ_1 *and the detection rates* λ_2 *and* λ_3.

The point estimate of the rate of infection λ_1 implies that the net rate of infection per infectious individuals $\lambda_1 S$ is equal to 0.32 (95%CI $= 0.23 - 1.37$). This means that the waiting time before an infectious individual, that has not been detected yet, infects an other individual is 3.1 years (95%CI $= 0.72 - 4.34$).

5.2 The dynamic of the Cuban HIV-AIDS epidemic

Reconstruction of the cumulative numbers of detections Figure 7 displays the dynamics predicted by the SIR model for the numbers of individuals detected by random screening $(R_t^1, t \in [0,T])$, by contact-tracing $(R_t^2, t \in [0,T])$ and for the number of unknown infectious individuals $(I_t, t \in [0,T])$. Interestingly, there is a really good fit between the real and predicted numbers of individuals detected by random screening except between 1992 and 1995. This period corresponds to the period of crisis that followed the collapse of the Soviet Union and during which the HIV detection system received less attention [13]. We also find that there is a slight discrepancy in the recent years (2000-2007) between the real and predicted numbers of individuals detected by contact-tracing. The SIR model predicts a larger number of contact-tracing detections which may reveal a weakening in the contact-tracing system. An explanation might be linked with the fact that a new way of detection, related to contact-tracing detection has appeared in the same time and is still treated as random detection. This new type of detection is promoted by the family doctors who ask to their patients the names of the individuals at risk (H. De Arazoza, personnal communication).

Performance of the contact-tracing system When testing for the performance of the contact-tracing system, Hsieh *et al.* [28] computed the coverage of the epidemic defined as the percentage of infectious individuals that have been detected $(R^1+R^2)/(I+R^1+R^2)$. As displayed by Figure 8, the SIR model predicts a coverage of 62% (95%CI = 36%−66%) in 2000 that is much lower than a coverage of 83% (75% − 87%) as inferred in [28] (the confidence interval is given in [15]). However, since the PPDs of Figure 6 show that the SIR model predicts less infectious individuals than observed, the coverage might still be overestimated and would consequently be even smaller than 62%.

Using this estimation of the coverage, we can compare the rates of detection by random screening and contact-tracing per infectious individual. The estimated per capita rate of random screening is $\lambda_2 = 0.13$. The per capita rate of contact-tracing is slightly more complicated and is equal to $\lambda_3 \sum_{i\in R} \Psi(t-T_i)/(I_t + \sum_{i\in R} \Psi(t-T_i))$. Using a zero-order expansion for Ψ, we find that this rate can be approximated by the product of λ_3 with the coverage of the epidemic. Hence, the per capita rate of contact-tracing can be estimated as $0.19 \times 0.62 \approx 0.12$ that is almost equal to λ_2.

Predictions Additionally, simulations of the SIR model provide predictions for the evolution of the HIV dynamic in the forthcoming years. Obviously, predictions for the year 2015, for instance, should be interpreted with caution since they rely on the assumption that the different rates remain constant. The SIR model predicts that in 2015, 42,000 (95%CI = 29,000 − 107,000) individuals will be infected since the beginning of the epidemic in Cuba. Among these infected individuals, a proportion of 45% (95%CI = 29% − 46%) will be detected by random screening an a proportion of 21% (95%CI = 10% − 22%) will be detected by contact-tracing. As displayed by Figure 7, the SIR-type model with contact-tracing predicts that the total proportion, among the detected individuals, of individuals detected by contact-tracing would reach an asymptote of 32% (95%CI = 25% − 33%) in 2015 whereas the counting data reveals a drop in the proportion of individuals detected by contact-tracing. The total number of infected individuals in 2015 corresponds to 27,000 (95%CI = 19,000 − 80,000) new cases of HIV

between July 2007 ($T = 21.5$) and January 2015 ($T = 29$). In the same period of time, the SIR model predicts that $12,000$ individuals ($95\%\text{CI} = 9,000 - 24,000$) will be detected by random screening and $6,000$ individuals ($95\%\text{CI} = 4,000 - 8,000$) will be detected by contact-tracing.

6 Conclusions

In the context of temporal epidemiological data, we show that Approximate Bayesian Computation (ABC) techniques can provide reasonable estimates of the parameters of interest such as the infection and detection rates. ABC inference relies solely on the simulations of the model and can therefore be applied to various epidemiological models defined in terms of an implicit stochastic mechanism [16]. In its broad lines, ABC consists in rejecting the simulations that produce summary statistics too different from the observations. Practically, ABC offers a convenient way for making inference since it can be applied to different variants of SIR-type model without modifications. Additionally, in the context of partially observed population and missing infectious times, MCMC methods require to reconstruct the unknown data which can be highly computationally intensive for large populations. For instance, [36] and [43] considered MCMC algorithms for populations consisting of about 100 individuals whereas the Cuban HIV-AIDS database contains almost 10,000 known HIV positive individuals, which makes the total (known and unknown) number of infectious individuals even larger. When the dimension of the missing data, the infection times here, is both large and unknown, data imputation with MCMC can be computationally very demanding (*e.g.* Cauchemez and Ferguson [12], Chis Ster *et al.* [41]). In [12], an analytically tractable Cox-Ingersoll-Ross diffusion was introduced to approximate the SIR model and avoid data augmentation. In [41] (Appendix A), where the order of magnitude of the missing data is similar to ours, a Reversible Jump MCMC (RJMCMC) was implemented. This RJMCMC required millions of iterations and the production of a single chain took 4-5 days on a parallelized system. ABC is in this perspective much more easy to implement and the model is fit within an hour or two.

In this paper, we consider both finite and infinite dimensional summary statistics. The infinite-dimensional summary statistics consist of the cumulative number of detected individuals as a function of time. When considering the standard SIR model, *i.e.* removing the contact-tracing detections, comparisons between MCMC and ABC show that ABC produces posterior distributions similar to MCMC provided that the tolerance rate is small enough. For tolerance rates that are too large, ABC posteriors may be wider than MCMC posteriors and the modes obtained with ABC may also be shifted from the modes obtained with MCMC. When comparing different ABC methods, we find that the point estimates of the parameters λ_1, λ_2, λ_3, with the smallest quadratic errors are obtained with the rejection method based on the infinite-dimensional statistics. However, the 95% credibility intervals obtained with this method are large and critically depend on the tolerance rate. By contrast, regression-based adjustment methods, and the NCH method more particularly, considerably shorten the credibility intervals and are less sensitive to the tolerance rate. Applications of regression-based ABC methods [5, 8] constitute therefore a solution for "stabilizing" the credibility intervals. However, no ABC-regression based methods have been developed so far for infinite-dimensional summary statistics.

In the last section of the paper, we calibrate the SIR model to the Cuban HIV-AIDS

data that contains the times at which the Cuban individuals have been found to be HIV positive. By comparing the obtained posterior predictive distributions, we find that the model with a frequency-dependent rate of contact-tracing provides the best fit to the data. We suggest here one possible improvement of the SIR model that could ameliorate the fit to the Cuban HIV-AIDS data. Because the mean time during which an infected individual has not been detected yet, is not well predicted by the SIR model, detection rates that depend on the time elapsed since infection could be considered. Such a model would contain additional parameters and the simple rejection scheme might require a much higher number of simulations for targeting in the partial posterior distribution. For models with a substantial numbers of parameters, adaptive ABC algorithms [40, 4, 44], that use the simulations to modify the sampling distribution of the parameter θ, might constitute interesting ways to explore for the future of ABC in epidemiology.

Concerning the epidemic of HIV-AIDS in Cuba, it was previously oberved by De Arazoza *et al.* [15] that one detection out of three is due to the contact-tracing program (see Figure 7). However, when comparing the present-day per capita rates of contact-tracing and random screening, we find that they are almost the same (see also [13]) and here almost equal to 0.13. This means that the predicted waiting time before an individual infected today will be detected is equal to $1/(2 \times 0.13) \approx 3.8$ years and both types of detection are equally probable at the time of detection. Although contact-tracing detection contributes importantly to HIV screening in Cuba, our results suggest that the screening might have been largely incomplete. We find that the percentage of undetected individuals among the infectious individuals might have been underestimated ([28]) and would be of the order of 40%.

Acknowledgment: The authors are grateful to Pr. H. De Arazoza of the University of La Havana and to Dr. J. Perez of the National Institute of Tropical Diseases in Cuba for granting them access to the HIV-AIDS database. This work has been supported by the French Agency for Research through the ANR MAEV and ANR Viroscopy, and by the Rhône-Alpes Institute of Complex Systems (IXXI). The authors thank the anonymous referees for their valuable suggestions.

References

[1] G.E.P Box an D.R. Cox. An analysis of transformations. Journal of the Royal Statistical Society: Series B, 26:211–246, 1964.

[2] R.M. Anderson and R.M.C. May. Infectious Diseases of Humans: dynamics and Control. Oxford University Press, Oxford, 1991.

[3] H.T. Banks, J.A. Burns, and E.M. Cliff. Parameter estimation and identification for systems with delays. SIAM Journal of Control and Optimization, 19(6):791–828, 1981.

[4] M.A. Beaumont, J.-M. Marin, J.-M. Cornuet, and C.P. Robert. Adaptivity for ABC algorithms: the ABC-PMC scheme. 2008. `arXiv:0805.2256v6`.

[5] M.A. Beaumont, W. Zhang, and D.J. Balding. Approximate Bayesian computation in population genetics. Genetics, 162:2025–2035, 2002.

[6] N.G. Becker and T. Britton. Statistical studies of infectious disease incidence. Journal of the Royal Statistical Society: Series B, 61(2):287–307, 1999.

[7] M.G. Blum. Approximate Bayesian Computation: a non-parametric perspective. 2009. Arxiv:0904.0635.

[8] M.G.B. Blum and O. François. Non-linear regression models for Approximate Bayesian Computation. Statistics and Computing, 2009. in press.

[9] D. Bosq and J.-P. Lecoutre. Théorie de l'estimation fonctionnelle. Ecole Nationale de la Statistique et de l'Administration Economique et Centre d'Etudes des Programmes Economiques. Economica, 1987.

[10] J.A. Burns, E.M. Cliff, and S.E. Doughty. Sensitivity analysis and parameter estimation for a model of Chlamydia Trachomatis infection. Journal of Inverse Ill-Posed Problems, 15:19–32, 2007.

[11] S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boelle. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statistics in Medicine, 23:3469–3487, 2004.

[12] S. Cauchemez and N.M. Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in london. Journal of the Royal Society Interface, 5:885–897, 2004.

[13] S. Clemençon, V.C. Tran, and H. De Arazoza. A stochastic SIR model with contact-tracing: large population limits and statistical inference. Journal of Biological Dynamics, 2(4):391–414, 2008.

[14] H. de Arazoza, J. Joanes, R. Lounes, C. Legeai, S. Clémençon, J. Perez, and B. Auvert. The HIV/AIDS epidemic in Cuba: description and tentative explanation of its low prevalence. BMC Infectious Disease, 7:130, 2007.

[15] H. de Arazoza, R. Lounes, J. Perez, and T. Hong. What percentage of the cuban HIV-AIDS epidemic is known. Rev Cubana Med Trop, 55:30–37, 2003.

[16] P.J. Diggle and R.J. Gratton. Monte Carlo methods of inference for implicit statistical models. Journal of the Royal Society: Series B, 46:193–227, 1984.

[17] J. Fan. Design-adaptive nonparametric regression. Journal of the American Statistical Association, 87(420):998–1004, 1992.

[18] R. Ferrière and V.C. Tran. Stochastic and deterministic models for age-structured populations with genetically variable traits. ESAIM: Proceedings, 27:289–310, 2009.

[19] B.F Finkenstädt, O.N. Bjørnstad, and B.T. Grenfell. A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. Biostatistics, 3:493–510, 2002.

[20] N. Fournier and S. Méléard. A microscopic probabilistic description of a locally regulated population and macroscopic approximations. Ann. Appl. Probab., 14(4):1880–1919, 2004.

[21] C. A. Gilligan G. J. Gibson, A. Kleczkowski. Bayesian analysis of botanical epidemics using stochastic compartmental models. Proceedings of the national academy of sciences, 101:12120–12124, 2004.

[22] A. Gelman and X-L Meng. Model checking and model improvment. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, Markov chain Monte Carlo in practice. Chapman & Hall, 1996.

[23] G.J. Gibson. Markov chain Monte-Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. Applied Statistics, 46:215–233, 1997.

[24] G.J. Gibson and E. Renshaw. Likelihood estimation for stochastic compartmental models using Markov chain methods. Statistics and Computing, 11:347–358, 2001.

[25] W.R. Gilks and G.O. Roberts. Strategies for improving MCMC. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, Markov chain Monte Carlo in practice. Chapman & Hall, 1996.

[26] B.E. Hansen. Nonparametric conditional density estimation. 2004. Working paper available at `http://www.ssc.wisc.edu/~bhansen/papers/ncde.pdf`.

[27] W. Härdle, M. Muller, S. Sperlich, and A. Werwatz. Nonparametric and Semiparametric Models. Springer, 2004.

[28] Y.H. Hsieh, H. de Arazoza, S.M. Lee, and C.W. Chen. Estimating the number of Cubans infected sexually by human immunodeficiency virus using contact tracing data. Int. J. Epidemiol., 31(3):679–83, 2002.

[29] R.J. Hyndman, D.M. Bashtannyk, and G.K. Grunwald. Estimating and visualizing conditional densities. Journal of Computing and Graphical Statistics, 5(4):315–336, 1996.

[30] V. Isham. Stochastic models for epidemics with special reference to AIDS. The Annals of Applied Probability, 3(1):1–27, 1993.

[31] W.O. Kermack and A.G. McKendrick. Contributions to the mathematical theory of epidemics. Bulletin of Mathematical Biology, Springer New York, 53(1-2):33–55, 1991.

[32] Y. Kuang. Delay Differential Equations with Applications to Population Dynamics, volume 151 of Lecture Notes in Statistics. Springer, New York, 2000.

[33] D. Mollison. Epidemic models: their structure and relation to data. Chap. 2: The structure of Epidemic Models. Cambridge University Press, 3 edition, 1995.

[34] E.A. Nadaraya. On estimating regression. Theory of Probability and its Applications, 10:186–190, 1964.

[35] P.D. O'Neill. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Mathematical Biosciences, 180:103–114, 2002.

[36] P.D. O'Neill and G.O. Roberts. Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society: Series A, 162:121–129, 1999.

[37] J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun, and M.W. Feldman. Population growth of human Y chromosomes: a study of y chromosome microsatellites. Molecular Biology and Evolution, 16:1791–1798, 1999.

[38] B.D. Ripley. Pattern recognition and neural networks. Cambridge University Press, Cambridge, 1996.

[39] B.W. Silverman. Discussion of the paper by Diggle and Gratton. Journal of the Royal Statistical Society: Series B, 46:212–214, 1984.

[40] S.A. Sisson, Y. Fan, and M. Tanaka. Sequential Monte Carlo without likelihoods. Proc. Nat. Acad. Sci. USA, 104:1760–1765, 2007.

[41] I. Chis Ster, B.K. Singh, and N.M. Ferguson. Epidemiological inference for partially observed epidemics: the example of the 2001 foot and mouth epidemic in Great Britain. Epidemics, 1:21–34, 2009.

[42] G. Streftaris and G.J. Gibson. Bayesian analysis of experimental epidemics of foot-and-mouth disease. Journal of the Royal Society: Series B, 271:1111–1117, 2004.

[43] G. Streftaris and G.J. Gibson. Bayesian inference for stochastic epidemics in closed population. Statistical Modelling, 4(1):63–75, 2004.

[44] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M.P. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of The Royal Society Interface, 6:187–202, 2009.

[45] V.C. Tran. Modèles particulaires stochastiques pour des problèmes d'évolution adaptative et pour l'approximation de solutions statistiques. PhD thesis, Université Paris X - Nanterre, 2006. `http://tel.archives-ouvertes.fr/tel-00125100`.

[46] P. van den Driessche. Some epidemiological models with delays. In Differential Equations and Applications to Industry (Claremont, CA, 1994), pages 507–520, River Edge, NJ, 1996. World Sci. Publishing.

[47] P. van den Driessche. Time delay in epidemic models. In C. Castillo-Chavez, S. Blower, P. van den Driessche, D. Kirschner, and A.A. Yakubu, editors, Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An Introduction, volume 125, pages 119–128, New York, 2002. IMA Volumes in Mathematics and its Applications, Springer.

[48] G.S. Watson. Smooth regression analysis. Shankya Series A, 26:359–372, 1964.

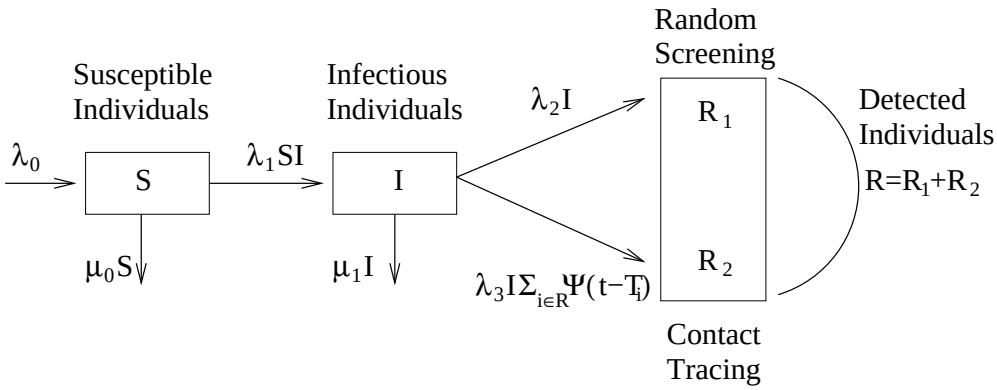
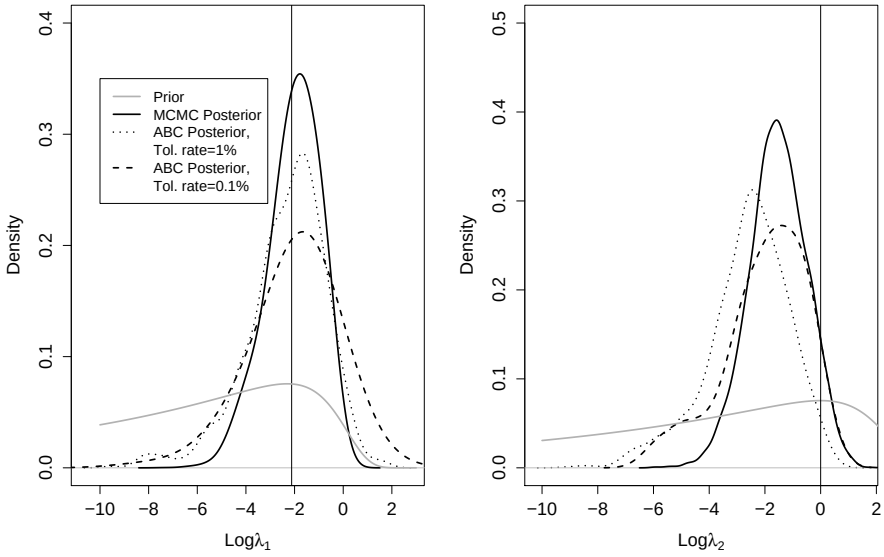


Figure 1: Schematic description of the SIR model with contact-tracing.

a)



b)

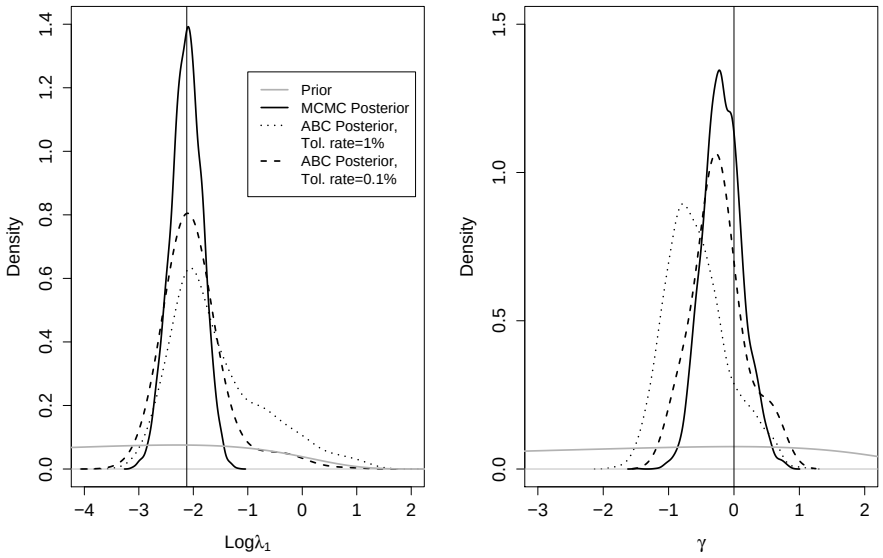


Figure 2: Comparison of the posterior densities obtained with MCMC and ABC. The vertical lines correspond to the values of the parameters used for generating the synthetic data. a) The data consists of 3 detection times that have been simulated by O'Neill and Roberts [36]. b) The data consists of 29 detection times that we have simulated by setting $\lambda_1 = 0.12$, $\lambda_2 = 1$, $S_0 = 30$, $I_0 = 1$, and $T = 5$ (see Supplementary Material for the 29 detection times).

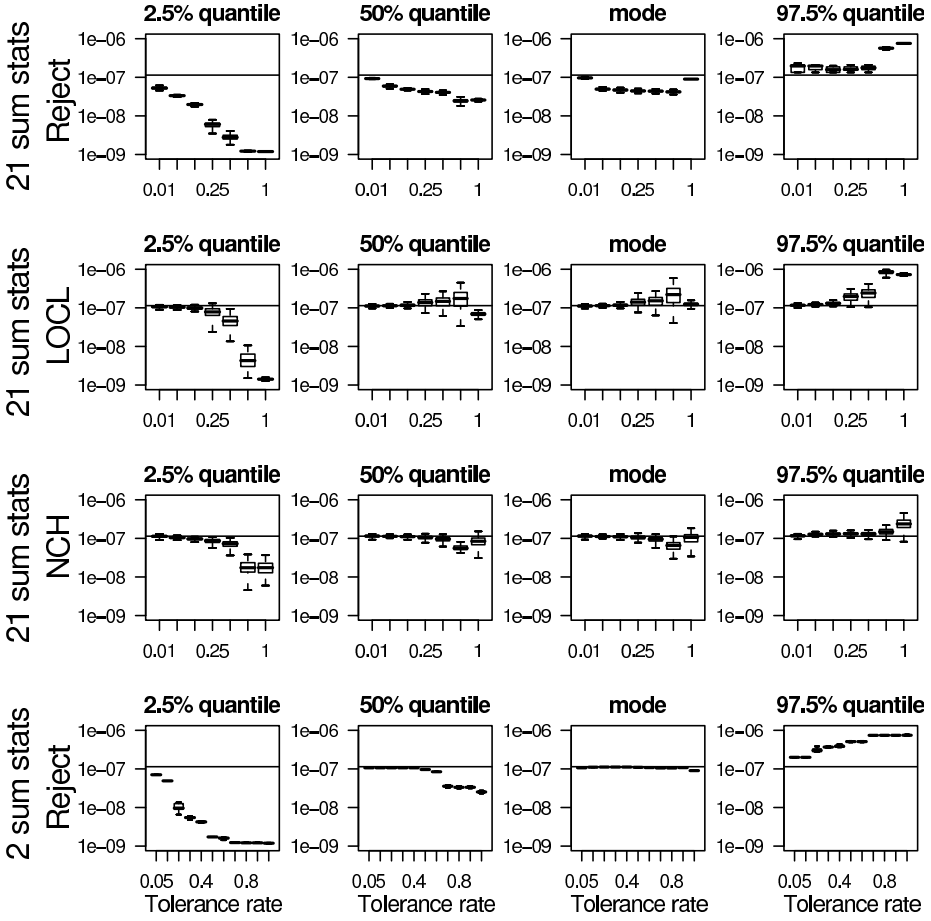


Figure 3: Boxplots of the $M = 200$ estimated modes and quantiles (2.5%, 50%, and 97.5%) of the partial posterior distributions of λ_1. For each ABC method and each value of the tolerance rate, 200 posterior distributions are computed for each of the 200 synthetic data sets. The horizontal lines correspond to the true value $\lambda_1 = 1.14 \times 10^{-7}$ used when simulating the 200 synthetic data sets. The different tolerance rates are 0.01, 0.05, 0.10, 0.25, 0.50, 0.50, 0.75, and 1 for all the ABC methods except the rejection scheme with the two summary statistics. For the latter method, the tolerance rates are 0.007, 0.02, 0.06, 0.13, 0.27, 0.37, 0.45, 0.53, 0.66, 0.80, 1.

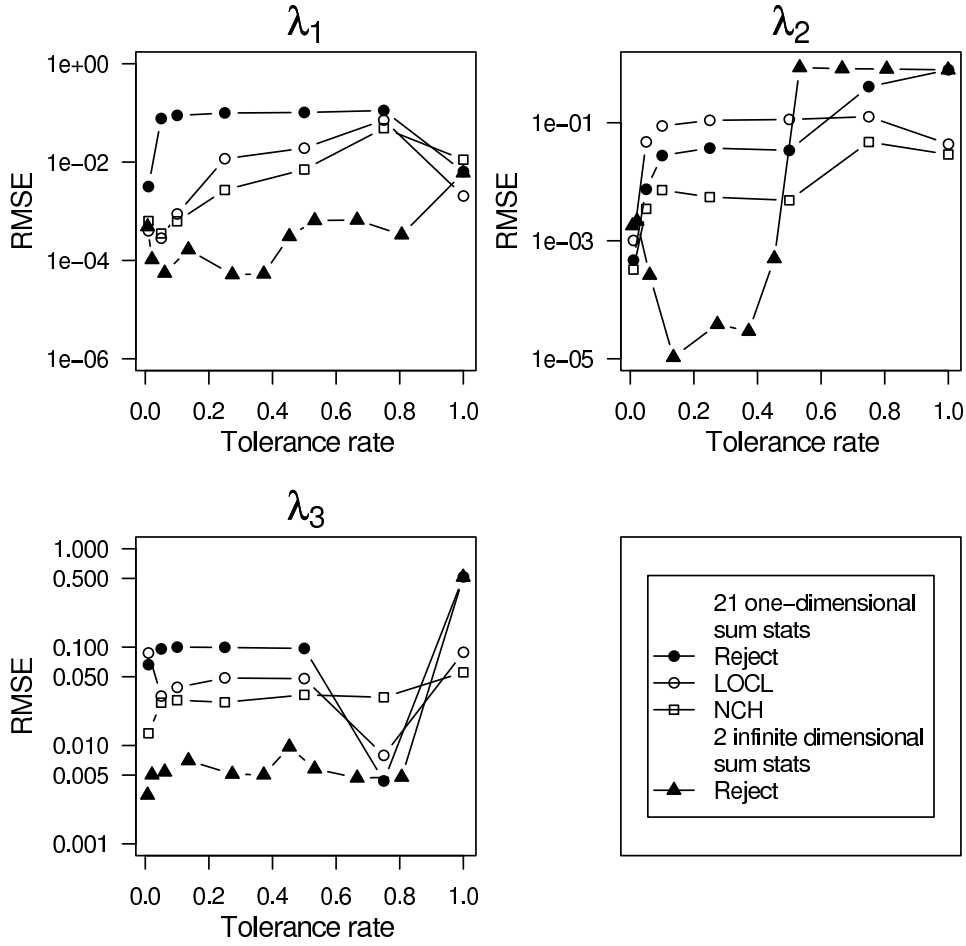


Figure 4: Rescaled mean squared error (RMSE) of the mode of the partial posterior distributions when estimating the three parameters λ_1, λ_2, and λ_3. The RMSE's are plotted as a function of the tolerance rate.

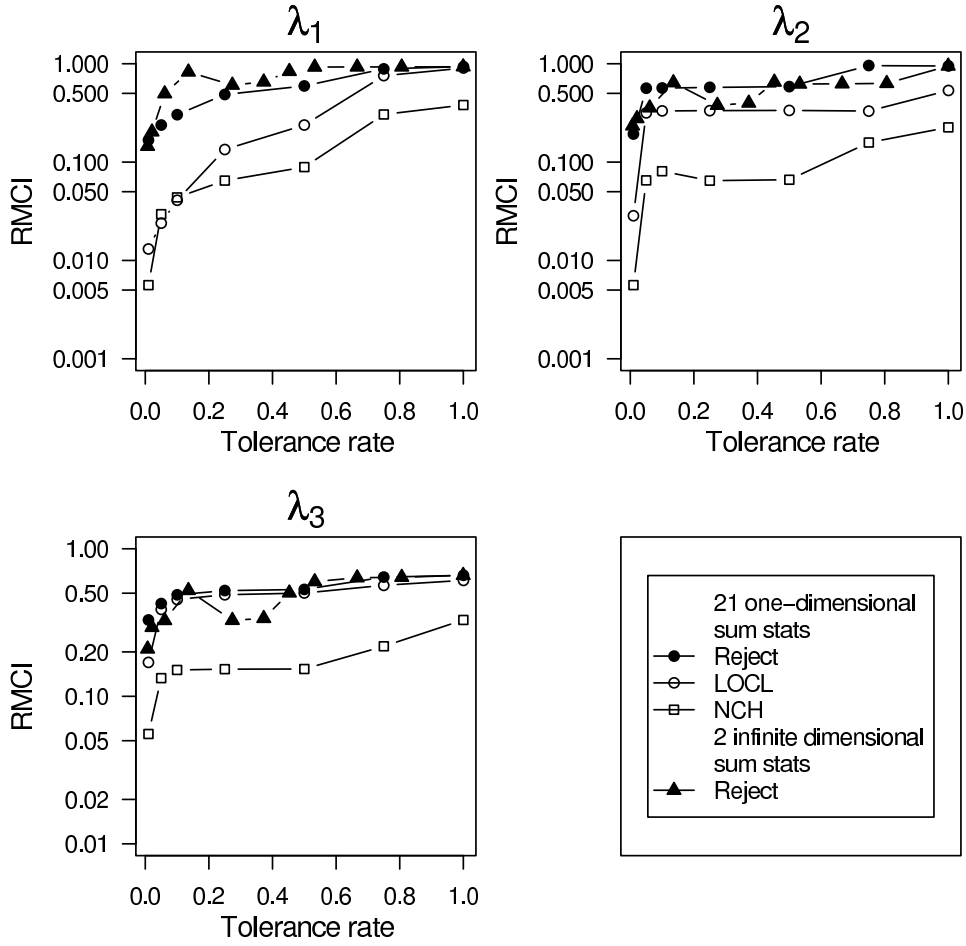


Figure 5: Rescaled mean credibility interval (RMCI) of the partial posterior distributions when estimating the three parameters λ_1, λ_2, and λ_3. The RMCI's are plotted as a function of the tolerance rate.

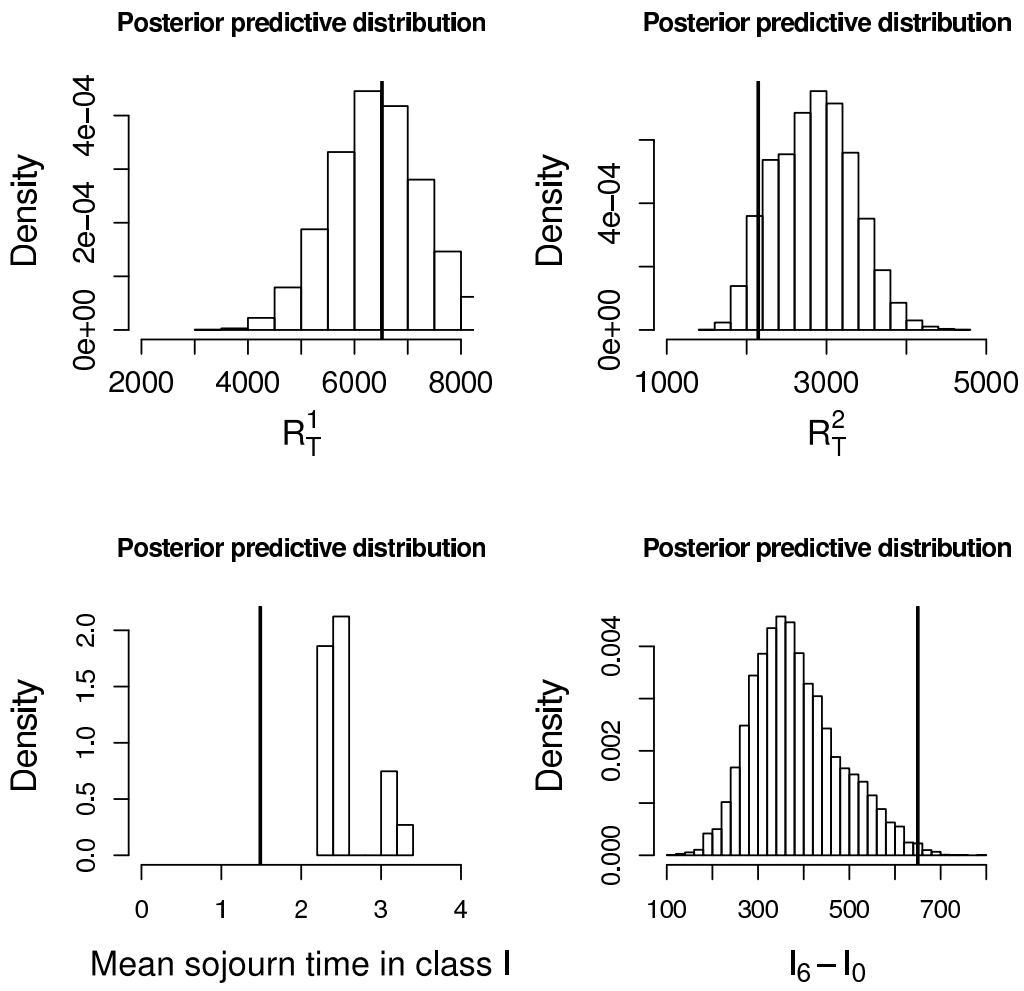


Figure 6: Bayesian posterior predictive distributions of $R^1_{21.5}$, $R^2_{21.5}$, I_6, and the mean sojourn time in the class I. The SIR model corresponds to the model with frequency dependence for contact-tracing detection. The partial posterior samples are obtained with the smooth rejection ABC algorithm by making use of the 2 infinite-dimensional summary statistics R^1 and R^2. Tolerance rate of $P_{\delta_1} = P_{\delta_2} = 1\%$ are considered for each summary statistic.

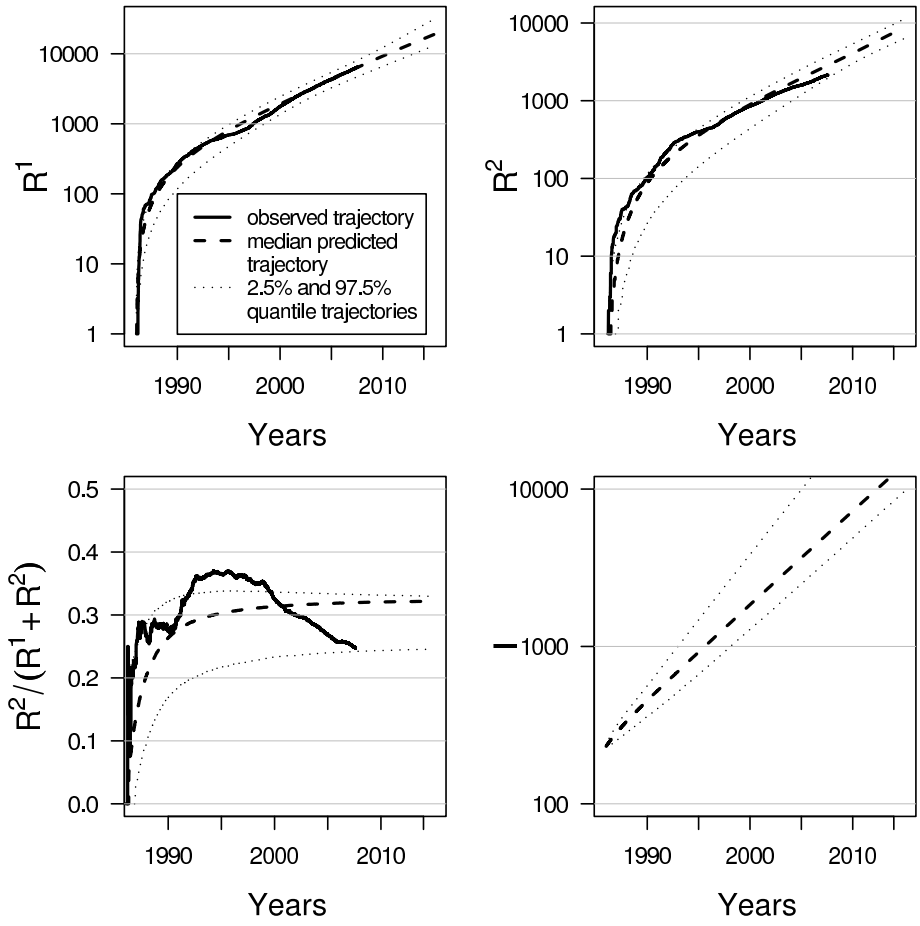


Figure 7: Median and 95% credibility intervals of the posterior predictive distributions of R_t^1, R_t^2, $R_t^1/(R_t^1 + R_t^2)$, and I_t from $t = 0$ (1986) to $T = 29$ (2015). The posterior samples are generated by the rejection scheme with the two summary statistics. A tolerance rate of $P_\delta = 1\%$ is considered for each summary statistic.

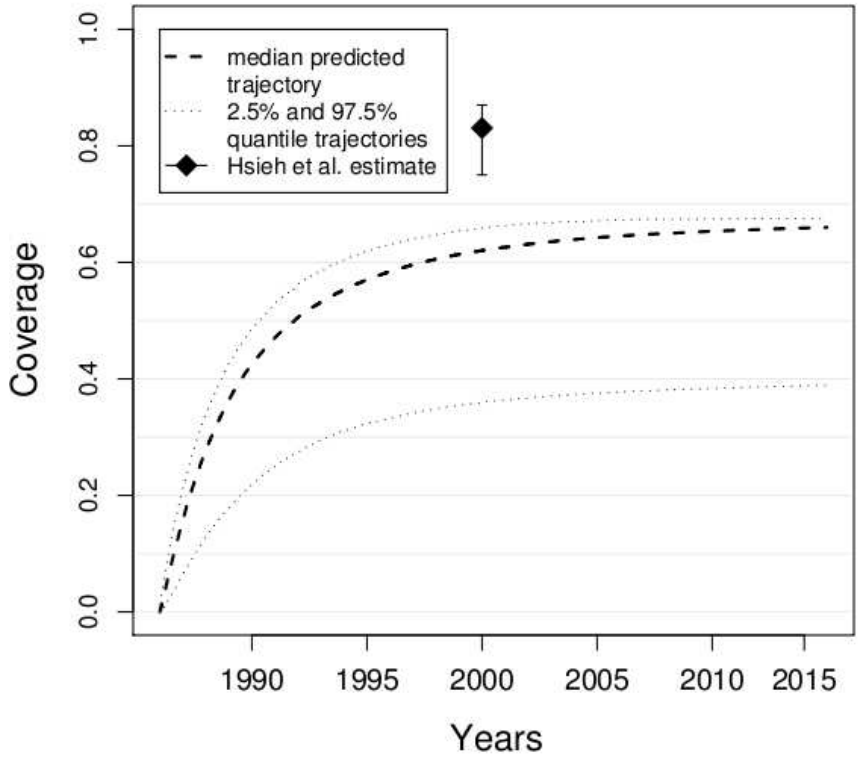


Figure 8: Median and 95% credibility intervals of the posterior predictive distribution of the coverage of the epidemic from $t = 0$ (1986) to $T = 29$ (2015). The coverage is defined as the proportion of known HIV positive individuals. The posterior samples are generated by the rejection scheme with the two summary statistics R^1 and R^2. A tolerance rate of $P_\delta = 1\%$ is considered for each summary statistic.