Likelihood-free inference using Approximate Bayesian Computation(ABC)

Part I (Introduction to ABC)

Mechanistic models can be used to predict how systems will behave in a variety of circumstances. The expressiveness of programming languages facilitates the development of complex, high-fidelity simulations and the power of modern computing provides the ability to generate synthetic data from them. Unfortunately, these simulators are poorly suited for statistical inference. The source of the challenge is that the **probability density (or likelihood)** for a given observation—an essential ingredient for both frequentist and Bayesian inference methods—**is typically intractable**.Such models are often referred to as implicit models and contrasted against prescribed models where the likelihood for an observation can be explicitly calculated. The problem setting of statistical inference under intractable likelihoods has been dubbed likelihood-free inference—although it is a bit of a misnomer as typically one attempts to estimate the intractable likelihood, so we feel the term simulation-based inference is more apt.

The intractability of the likelihood is an obstruction for scientific progress as statistical inference is a key component of the scientific method. Two approaches were proposed to solve the intractability issue :

1) **Density estimation methods** are used to approximate the distribution of the summary statistics from samples generated by the simulator.
2) **Approximate Bayesian computation (ABC)** compares the observed and simulated data based on some distance measure involving the summary statistics.

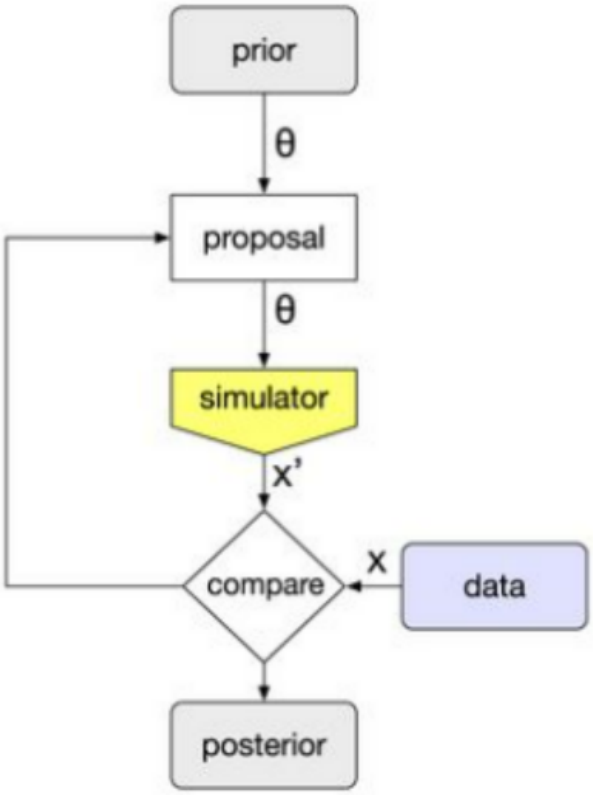
Simulator

For the purpose of our project, a **simulator is a computer program** that takes as **input a vector of parameters θ**, samples a series of internal states or latent variables $z_i \sim p_i(z_i|\theta, z_{<i})$, and finally **produces a data vector $x \sim p(x|\theta,z)$ as output**.

Methodology

The problem of inference without tractable likelihoods is not a new one, and two major approaches have been developed to address it. Arguably the most well known is ABC. Until recently, it was so established that the terms “likelihood-free inference” and “ABC” were often used interchangeably. In the simplest form of rejection ABC, the **parameters θ are drawn from the prior**, the simulator is run with those values to **sample $x_{sim} \sim p(\cdot|\theta)$**, and θ is retained as the posterior sample if the simulated data are sufficiently close to the observed data. In essence, the likelihood is approximated by the probability that the condition **$\rho(x_{sim}, x_{obs}) < \epsilon$** is satisfied,

where ρ is some distance measure and ϵ is a tolerance. The accepted samples then follow an approximate version of the posterior. We show a schematic workflow of this algorithm in fig shown below (for a more elaborate Markov chain Monte Carlo algorithm with a proposal function).



In the limit ϵ→0, inference with ABC becomes exact, but for continuous data the acceptance probability vanishes. In practice, small values of ϵ require unfeasibly many simulations. For large ϵ, sample efficiency is increased at the expense of inference quality. Similarly, the sample efficiency of ABC scales poorly to high-dimensional data x. Since the data immediately affect the rejection process (and in more advanced ABC algorithms the proposal distribution), inference for new observations requires repeating the entire inference algorithm. ABC is thus best suited for the case of a single observation or at most a few i.i.d. data points.

Part II (Implementation of ABC)

Now, we will implement a paper which is based on the concept of Approximate Bayesian Computation (ABC) in order to gain understanding of how to generate priors and do inference.

Paper which we will try to implement :
https://michaelgutmann.github.io/assets/papers/Lintusaari2019a.pdf

Title of the paper : Resolving outbreak dynamics using approximate bayesian computation for stochastic birth-death models

Abstract : Earlier research has suggested that approximate Bayesian computation (ABC) makes it possible to fit simulator-based intractable birth–death models to investigate communicable disease outbreak dynamics with accuracy comparable to that of exact Bayesian methods. However, recent findings have indicated that key parameters, such as the **reproductive number R, may remain poorly identifiable with these models**. Here we show that this identifiability **issue can be resolved by taking into account disease-specific characteristics of the transmission process** in closer detail. Using tuberculosis (TB) in the San Francisco Bay area as a case study, we consider a model that generates genotype data from a mixture of three stochastic processes, each with its own distinct dynamics and clear epidemiological interpretation. We show that **our model allows for accurate posterior inferences about outbreak dynamics** from aggregated annual case data with genotype information. **As a byproduct of the inference, the model provides an estimate of the infectious population size at the time the data were collected**. The acquired estimate is approximately two orders of magnitude smaller than assumed in earlier related studies, and it is much better aligned with epidemiological knowledge about active TB prevalence. Similarly, the reproductive number R related to the primary underlying transmission process is estimated to be nearly three times larger than previous estimates, which has a substantial impact on the interpretation of the fitted outbreak model.

Introduction to paper

Birth–death processes are flexible models used for numerous purposes, in particular for characterizing the spread of infections under the so-called Susceptible–Infectious–Removed (SIR) formulation of an epidemic process. Under circumstances where a disease outbreak occurs but where daily, weekly or even monthly incidence counts are not directly applicable or available, the estimation of key epidemiological parameters, such as the reproductive number R, has to be based on alternative sources of information. **Likelihood-based inference could provide an alternative to standard outbreak investigations relying solely on incident count data**, but it is often considerably more challenging.

In an earlier research done, a large infectious population size of n = 10,000 was required for the BD simulator to produce similar levels of genetic diversity to those observed in the San

Francisco Bay data. Because it has not been observed, this assumption is difficult to justify when the acquired estimates depend on it. Here we introduce an alternative formulation of the BD model that resolves the identifiability issue of R.

The proposed model does not require any assumptions about the underlying infectious population size, instead providing an estimate for that value as a byproduct of the inference. The model incorporates epidemiological knowledge about the TB infection and disease activation processes by assuming that the observed genotype data represent a mixture of **three birth–death processes**, each with clearly distinct characteristics. In the new model, we consider latent and active TB infections separately, as only the latter lead to new transmission events. Transmission clusters are formed by recent infections that rapidly progress to active TB and spread further through the host population. Due to the rapid onset of symptoms in a new active case, the fingerprint of the pathogen remains the same throughout the transmission process, and its patients consequently form an epidemiological cluster. If, on the other hand, an infection remains latent, the pathogen undergoes mutations and thus acquires a new genetic fingerprint over the years.

The model

Our model is based on a birth–death (BD) process.
Birth Event : An event that corresponds to the appearance of a new case of active TB.
Death Event : An event that corresponds to any event that makes an existing host non-infectious. Such events include death, sufficient treatment, quarantine, and relocation away from the community under investigation.

The model incorporates **two Birth-Death processes** and **one pure birth process** that have epidemiologically based interpretations. As in a standard BD process, these events are assumed to be independent of one another and to occur at specific rates. **The time between two events** is assumed to **follow the exponential distribution** specified by the **rate of occurrence**, causing the number of events to **follow the Poisson distribution**. The timescale considered here is one calendar year. The evolution of the infectious population is simulated by drawing events according to their rates.

The observations are collected for a **time interval of 2 years** that matches that of the observed data of San Francisco. Also, observations are collected from the simulated process after a sufficient warmup period so that the process can be expected to have reached stable properties.

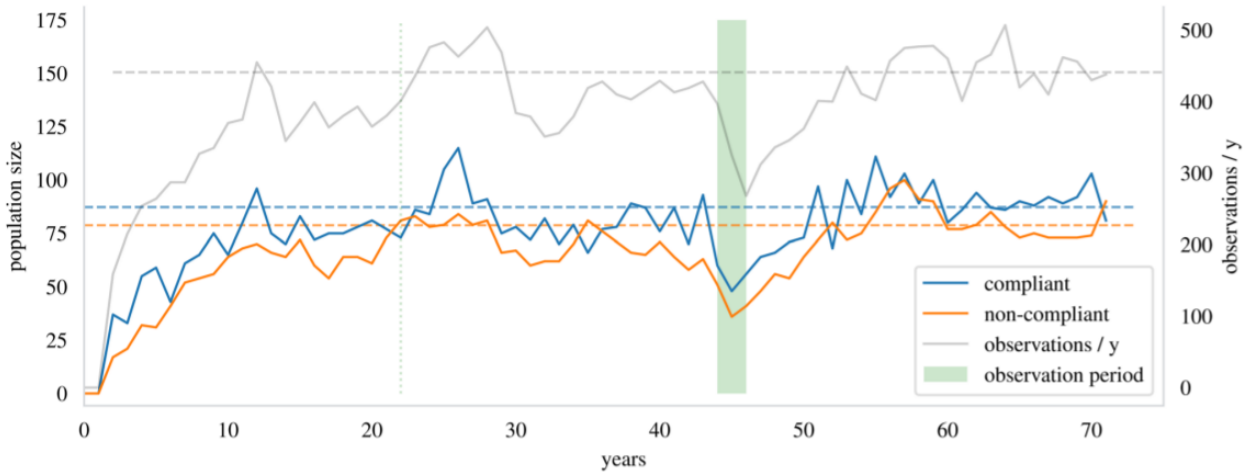


Fig. An illustration of simulated compliant and non-compliant populations as observed at the end of each year

In the figure, the dashed lines are the balance values. The population sizes fluctuate around them after the process has matured. Both populations surpass their balance values at least once by the 22-year mark. The observation period is the green patch.

We introduce a **burden parameter β** that reflects the rate at which new active TB cases with a previously unseen pathogen fingerprint appear in the community. This is the pure birth process of the model, and it represents reactivation of TB from latent cases as well as new pathogen fingerprints introduced by immigration. In the simulation, each such case receives a new cluster index that has not been assigned to any earlier case. We also introduce two distinct birth–death processes for cases that are either compliant or noncompliant with treatment. These birth–death processes are parametrized with birth rates τ_i and death rates δ_i , where i = 1 denotes the non-compliant population and i = 2 the compliant population.

We assume that a new **TB case is non-compliant with therapy with probability p1** . At transmission (birth event in the simulation), this probability is used to determine the patient type of the new case. We also assume that the epidemic is at a steady state by requiring that compliant cases have a reproductive number $R_2 = \tau_2 / \delta_2 < 1$ and that **the reproductive number R_1 of the non-compliant cases is constrained such that the population does not grow without limit**. The steady state assumption is motivated by the tuberculosis incidence counts in the United States during the data collection period.

Statistical Analysis of the model

Let subscript i = 1 and 2 denote the non-compliant and compliant subpopulation respectively. Size of a subpopulation follows a compound birth–death process whose birth rate is a linear function of the burden rate and of the birth rates of the two subpopulations at their respective present sizes.

The birth rate of the non-compliant subpopulation = $p_1 (\beta + \tau_1 n_1 + \tau_2 n_2)$,

where n_1 and n_2 are the current subpopulation sizes
p1 is the probability of a case being non-compliant.

The corresponding death rate is δ1n1 . Using this approach, we can determine the balance sizes b1 and b2 of the subpopulations-that is, the values of n1 and n2 that make the birth rate equal to the death rate in each subpopulation. In this steady state, the subpopulation sizes neither shrink nor grow. We obtain expressions for b1 and b2 by solving the following set of linear equations:

$$\delta_1 b_1 = p_1 (\beta + \tau_1 b_1 + \tau_2 b_2)$$
$$\delta_2 b_2 = p_2 (\beta + \tau_1 b_1 + \tau_2 b_2)$$

where p2 = 1 – p1 is the probability of a new case being compliant. The linear equations yield the following solution:

$$b_1 = \frac{p_1 \beta \delta_2}{\delta_2 \delta_1 - p_2 \tau_2 \delta_1 - p_1 \tau_1 \delta_2}$$
$$b_2 = \frac{b_1(\delta_1 - p_1 \tau_1) - p_1 \beta}{p_1 \tau_2}.$$

Given this solution, the balance values b1 and b2 exist when

$$R_1 < 1 / p_1$$

and

$$R_2 < (1 - p_1 R_1) / p_2 .$$

Assuming, for instance, that **p2 = 0.95** , we would have **R1 < 20**.

Also,mean number of observed cases per year can be approximated as follows:

$$\hat{n}_{obs} = p_{obs} (\delta_2 b_2 + \delta_1 b_1).$$

We used approximate Bayesian computation to carry out parameter inference due to the unavailability of the likelihood function. The **result is a sample from the approximate posterior distribution $p(R_1, t_1, R_2, \beta \mid y_0)$**.

We used the **Engine for Likelihood-Free Inference (ELFI) software** to perform our inference. Using rejection sampling, we selected 1000 parameter values from a total of 6M simulations. This large number of simulations was possible due to the fact that we implemented a computationally efficient, vectorized version of the simulator in Python. We utilized rejection sampling in order to incorporate priors appropriate to our model structure.

Priors

We set priors over the **burden rate β**, **reproductive numbers R_1 and R_2** , and **the net transmission rate $t_1 = \tau_1 - \delta_1$** of the non-compliant subpopulation.

We fix compliant population **death rate $\delta_2 = 5.95$** and use it to calculate the net transmission rate $t_2 = \delta_2 (R_2 - 1)$. We also fixed the probability of being observed to **$p_{obs} = 0.8$** and the probability of a new case being non-compliant to **$p_1 = 0.05$**.

Priors are set as follows :

$$\beta \sim N(200,30)$$
$$R_1 \sim Unif(1.01, 20)$$
$$R_2 \mid R_1 \sim Unif(0.01, (1 - 0.05)/0.95)$$
$$t_1 \sim Unif(0.01,30)$$

Given the observed data, we set the following additional constraints to optimize computation:

$$n_{obs} < 350 \text{ and } \tau_1 < 40$$

Summary Statistics

In descriptive statistics, summary statistics are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible. These summary statistics aim to capture meaningful properties of the observed data given the new model.

We identified **8 summary statistics** that will be helpful for our experiment. The first summary is the **number of observations**, which is here allowed to vary. Five of the summaries are related to the clustering structure, where a cluster is defined as a group of TB cases with the same genetic fingerprint: **the total number of clusters**, **the relative number of singleton clusters**, **the relative number of clusters of size two**, **the size of the largest cluster**, and **the mean of the successive differences in size among the four largest clusters**. These were chosen

specifically to emphasize the most stable properties of the clustering structure. The remaining two summaries are related to the observation times of the largest cluster. Observation times were not included in earlier studies, and here they prove useful for identifying the net transmission rate t_1 . The summary statistics in question are the **number of months from the first observation to the last** and **the number of months in which at least one observation was made from the largest cluster**.

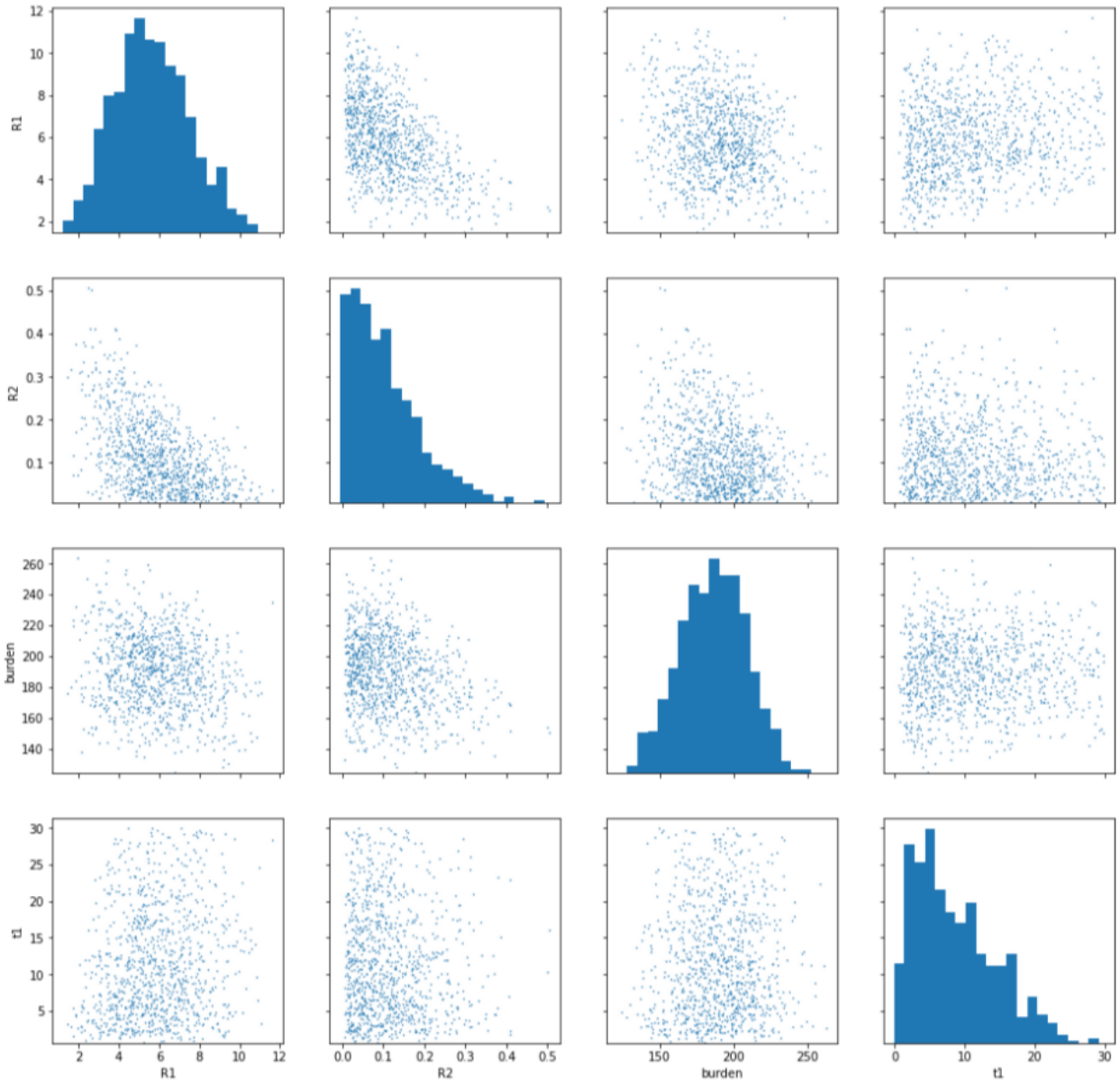
We weighted our summary statistics to adjust for and even out differences in their magnitudes. The final summary statistics and weights perform well in the evaluation of the model. The **resulting acceptance/rejection threshold is ε = 31.7**, while the smallest distance observed in our simulations is 12.5. Like our summary statistics, this threshold was selected from our trial runs of inference on synthetic data. We chose a value that struck a good balance between run time, acceptance rate, and the resulting Monte Carlo error rate.

Summary statistic	Explanation	Weight	y_0
n_{obs}	Number of observations.	1	473
$n_{clusters}$	Number of clusters.	1	326
r_{c1}	Relative number of singleton clusters. Computed as $r_{c1} = n_{c1} / n_{obs}$, where n_{c1} is the number of clusters of size 1. The value of r_{c2} is computed likewise.	100/0.60	0.60
r_{c2}	Relative number of clusters of size 2.	100/0.04	0.04
largest	Size of the largest cluster.	2	30
mean_largest_diff	Mean of the successive differences in size among the four largest clusters.	10	6.67
month_period	Number of months from the first observation to the last in the largest cluster.	10	24
obs_months	The number of months in which at least one observation was made from the largest cluster.	10	17

Table : The summary statistics, their weights, and their values for the observed data y_0

Results

Below figure shows a sample of 1000 values from the joint approximate posterior distribution $p(R_1, t_1, R_2, \beta \mid y_0)$.



The pairwise sample clouds seem reasonably concentrated, and do not extend to the edges of the axes. The histograms and scatter plots are fairly normally shaped, with the only minor exception being that the net transmission rate of the non-compliant population t_1 has a slight tail

towards high values. A visual comparison of the posterior against the prior, together with the above observations, suggests that the model is identifiable for the San Francisco dataset.

The posterior means, medians and 95% credible intervals are given in below table

Parameter	Mean	Median	95% CI
R1	5.84	5.75	(3.64, 8.12)
t1	11.71	11.32	(6.54, 17.8)
R2	0.11	0.11	(0.05, 0.17)
β	190	190	(168, 214)

Table : Posterior summaries

The means and medians are similar to one another, which indicates that the posterior distributions are symmetrical. t_1 has the largest discrepancy due to the presence of the small tail mentioned above.

As a byproduct of the above model, we also got Balance sizes (compliant, non-compliant) = (87.298982605842141, 78.765999343616869).

Discussion

We have proposed a stochastic birth–death model to expand on several previous studies examining the use of simulator based inference to investigate the spread of active TB within a community. Outbreaks of TB are characterized by epidemiologically linked clusters of patients with active TB that emerge within a relatively short time interval.

Earlier approaches suffered from the inability to reproduce these large clusters with an appropriate level of heterogeneity in cluster sizes without the prior assumption of a very large infectious population (to the order of 10,000 individuals). This assumption has a considerable effect on the estimate of the reproductive number R. However, epidemiological knowledge of TB does not support the existence of such a large infectious population in the study region during the observation period. **Under our model, a prior estimate of the infectious population size is not needed**. This model has a different parametrization for which estimates can be found from the literature. **As a byproduct of the inference, the model also yields estimates for the infectious population size at the end of the data collection period**.

For each subpopulation, the basic reproductive number (R_1 or R_2) represents the average number of infections caused by a single infectious case that rapidly progresses to active TB. This value excludes latent infections, which are indirectly captured via the burden rate parameter β. We estimate that the basic reproductive number of non-compliant patients is **R_1 = 5.88** with a **95% credible interval (CI) of (3.64, 8.12)** and for compliant patients to be **R_2 = 0.09 with a 95% CI of (0.05, 0.17)**.

References

1. Resolving outbreak dynamics using approximate Bayesian computation for stochastic birth–death models
   https://michaelgutmann.github.io/assets/papers/Lintusaari2019a.pdf

2. The frontier of simulation-based inference
   https://www.pnas.org/content/117/48/30055

3. Fundamentals and Recent Developments in Approximate Bayesian Computation
   https://academic.oup.com/sysbio/article/66/1/e66/2420817

4. The ABCs of Approximate Bayesian Computation
   https://towardsdatascience.com/the-abcs-of-approximate-bayesian-computation-bfe11b8ca341

5. ELFI: Engine for Likelihood Free Inference
   https://github.com/elfi-dev/elfi
   https://elfi.readthedocs.io/en/latest/