HIV with contact tracing: a case study in approximate Bayesian computation

Michael G. B. Blum, Viet Chi Train

Yali Wan STAT 518

May 2, 2013

The SIR Model

Our study is restricted to the sexually transmitted epidemic of HIV in cuba.

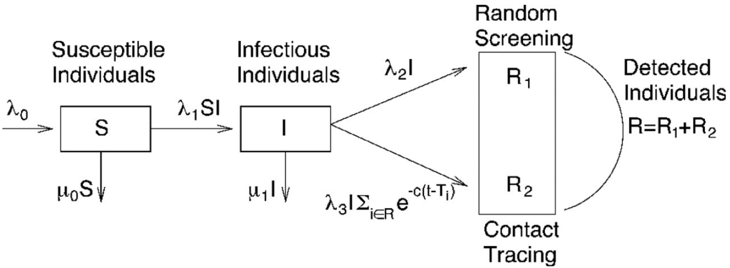


Figure : Schematic description of the SIR model with contact tracing

Parameter of interest: λ_1,λ_2 and λ_3.

Why Not MCMC

Marcov Chain Monte Carlo Method is not always good with SIR models

- Computationally prohibitive for high-dimensional missing observations (Cauchemez and Ferguson,2008; Chis Ster and others, 2009)
- Fine-tuning of the proposal distribution is required for efficient algorithms (Gilks and Roberts, 1996)

What If We Use MCMC

Observed data: removal time: $\Gamma_1 = 0, \Gamma_2, \cdots, \Gamma_n$, $\Gamma_i \in [0, T]$
Missing data: infectious time: $I_1, I_2, \cdots, I_m$,$I_1 < 0$
Prior:
$\lambda_1 \sim Gamma(a_1, v_1)$
$\lambda_2 \sim Gamma(a_2, v_2)$,
$y \sim \theta exp(\theta y) I(y < 0)$, y is the density of I_1

Metropolis Hasting Algorithm with Gibbs sampling

Sampling posteriors

- $f(\Gamma, I|\lambda_1, \lambda_2, I_1) = \prod_{i=1}^{n} \lambda_2 Y_{\Gamma_j^-} \prod_{j=2}^{m} \lambda_1 X_{I_j^-} Y_{I_j^-} exp\{-\int_{I_1}^{T} (\lambda_1 X_t Y_t + \lambda_2 Y_t) dt\}$
- $\pi(\lambda_1|\Gamma, I, I_1, \lambda_2) \sim \Gamma(a_1 + \int_{I_1}^{T} X_t Y_t dt, m - 1 + v_1)$
- $\pi(\lambda_2|\Gamma, I, I_1, \lambda_1) \sim \Gamma(a_2 + \int_{I_1}^{T} Y_t dt, n + v_2)$

M-H step inside

abbreviate $f(\Gamma, I|\lambda_1, \lambda_2, I_1)$ by $f(I)$

- Moving an infection time: $\frac{f(I-\{s\}+\{t\})}{f(I)} \wedge 1$, t is sampled uniformly on(I_1, T)
- Removing an infection time: $\frac{f(I-\{s\})m}{f(I)(T-I_1)} \wedge 1$
- Adding a new infection time: $\frac{f(I+\{t\})(T-I_1)}{f(I)(m+1)} \wedge 1$

Approximate Bayesian Computation

Two Approximation are at the core of ABC

- Replacing observations with summary statistics: Use posterior $p(\theta|S(x) = S_{obs})$ instead of $p(\theta|x)$

  In a fully observed SIR model, Summary statistics are R_t^1 and R_t^2, where $R_t^1 + R_t^2 = R_t$, $t \in [0, T]$. R_t^1 and R_t^2 are sufficient statistics.
- Simulation-based approximations of the posterior. (More will be described on the partially observed model.)

The Algorithm

1. Generating N random draws$(\theta_i, s_i), i = 1, \cdots, N$. The parameter θ_i is generated from the prior distribution π, and the vector of summary statistics s_i is calculated for the ith data set that is simulated from the generative model with parameter θ_i.

2. Associate to the ith simulation, the weight $W_i = K_\delta(s_i - s_{0bs})$, where δ is a tolerance threshold and K_δ a (possibly multivariate)smoothing kernel.

3. 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.

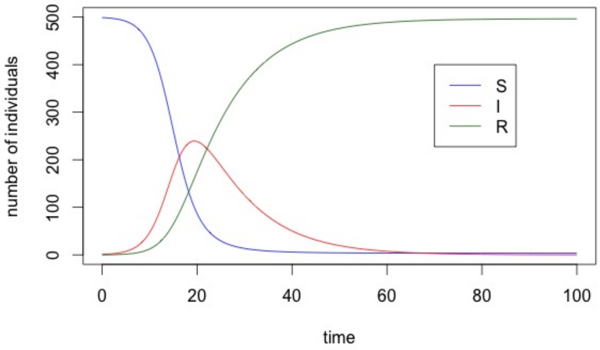
$\lambda_j, j = 1, 2, 3$, is estimated by $\hat{\lambda}_j = \sum_{i=1}^N \lambda_{j,i} W_i / \sum_{i=1}^N W_i$.

How to simulate summary statistics using stochastic SIR models

Main difficulty: the rate of detection changes with time.
Assumption: λ_0, μ_0 are such that the size of the population of S remains constant. Q: How come?

1. $S = S_0, I = I_0, R = 0$ Current time: Γ
2. Assume k events has already be simulated, now come to (k+1)th event simulate $\epsilon \sim exp(C_k)$, where $C_k = \lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k} + \lambda 3 I_{t_k} R_{t_k}$, $\Gamma' = \Gamma + \epsilon$
3. Stop if $\Gamma' > T$
otherwise, simulate $U \sim unif(0, C_k)$

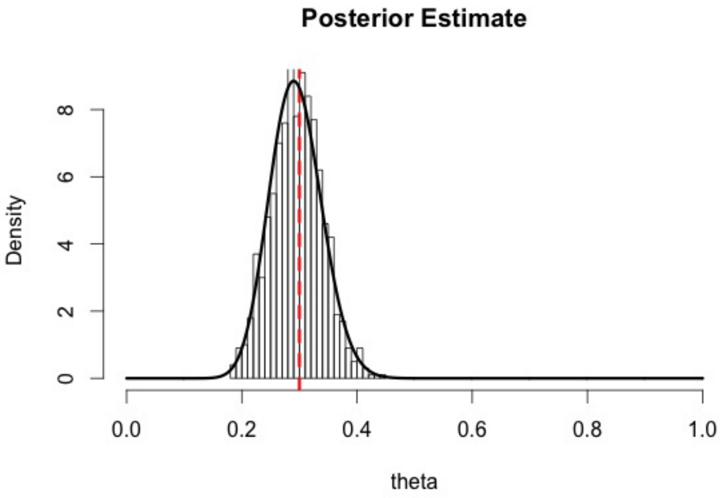
Example: Simulation from deterministic SIR model



What to do if we obtain the simulated summary statistics

ABC - rejection-sampling algorithm.
ABC - smoothing kernel

Example



Challenges

Things I will do next:

- applying abc method to with the second approximation.
- applying abc method to the real data: Cuban database.

The End

Thank you all for the attention!