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

Michael G. B. Blum, Viet Chi Train

Yali Wan STAT 518

April 16, 2013

The SIR Model

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

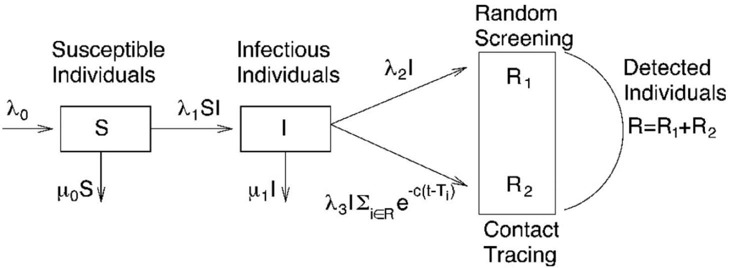


Figure : Schematic description of the SIR model with contact tracing

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)

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.

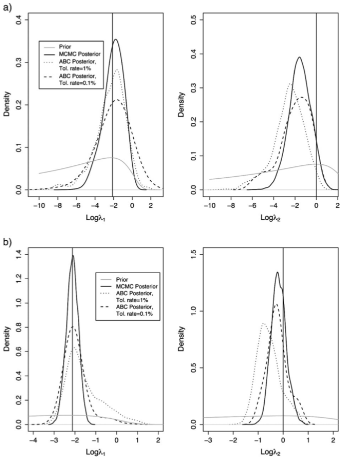
3. The distribution $\sum_{i=1}^N W_i \delta_{\theta_i} / \sum_{i=1}^N W_i$, in which δ_θ denotes the Dirac mass at θ, approximates the target distribution.

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

The Result From A Fully Observed Model

Part a: The data consists 3 detection time

Part b: The data consists 29 detection time



Approximate Bayesian Computation

When full observations are unavailable, summary statistics is composed of:

- R^1_T and R^2_T,
- $R^l_{j+1} - R^l_j$, $l = 1, 2$, for each year j.
- $I_{j+1} - I_j$, for $j = 0, \cdots, 5$.
- Mean time during which an individual is infected but has not been detected yet.

"Curse of dimensionality"

The Second Core Approximation

Correction Adjustment: $\theta_i^* = G_{s_{obs}}^{-1}(G_{s_i}(\theta_i))$, $i = 1, \cdots, N$.

Available methods for choosing G:

- Local linear regressions of Beaumont and others (LOCL)
- Nonlinear conditional heteroscedastic regressions of Blum and Francois (NCH)

Discussion

- ABC applications have been restricted to models with moderate number of parameters.
- Statisticians are more experienced with MCMC.
- No ABC with regression adjustment have been developed so far for infinite- dimensional summary statistics.

The End

Thank you all for listening!