Stochastic SIR model with contact-tracing

Cornell 4^{th} Probability Summer School

Viet Chi TRAN
Université Lille 1 - France

Joint work with Stéphan Clémençon and Hector De Arazoza and Michael Blum

June 26, 2008

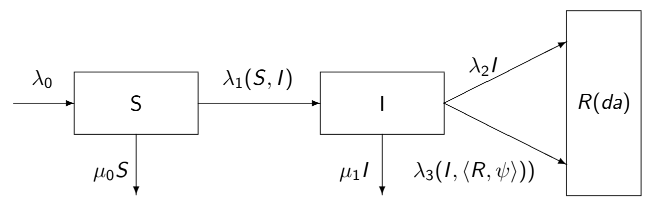
Cuban AIDS Epidemics

- AIDS has probably been introduced in Cuba since 1985
Today, we have a dataset with 8715 detected individuals, for which age, (possible) time of infection, times of detection, death, sex and sexual orientation etc. are available

- In order to detect seropositive individuals, two methods are used :
  - Random screening
  - Contact tracing

- Aim:
  - evaluate the efficiency of both methods.
  - propose estimates of PDEs of epidemiology that are linked to a population model at an individual level:
    - explains the noise in a natural way,
    - the formalism corresponds to the data of patients
    - the convergence and fluctuations of the individual-based model will give convergence and asymptotic normality for the estimates

SIR model



$$\lambda_1(S,I) = \lambda_1 SI, \quad \text{(mass action principle)}$$
$$= \lambda_1 SI/(I+S) \quad \text{or} \quad \lambda_1 I, \quad \text{(frequency dependence)}$$

$$\lambda_3(I,\langle R,\psi\rangle) = \lambda_3 I\langle R,\psi\rangle$$
$$= \lambda_3 I\langle R,\psi\rangle/(I+\langle R,\psi\rangle) \quad \text{or} \quad \lambda_3\langle R,\psi\rangle.$$

Age-structured component

- $R_t(da)$ is a point measure: $R_t(da) = \sum_{i=1}^{N_t} \delta_{a_i(t)}$

- Detected individuals age with speed 1 and contribute to the search of infectious people with rate $\psi(a)$.

Notation:

$$\langle R_t,\psi\rangle = \int_{\mathbb{R}_+} \psi(a) R_t(da) = \sum_{i=1}^{N_t} \psi(a_i(t)).$$

Examples of ψ:

$$\psi(a) = C_1 e^{-C_2 a}, \quad \text{or} \quad \psi(a) = \text{Gamma density function}$$

SDE and PDE approximation

- It is possible to describe $(S_t, I_t, R_t(da))$ as the solution of a SDE driven by a Poisson point measure.

- This allows easy simulations, which will be an advantage if dealing with estimation methods based on simulations.

- We can look at large population renormalization of the process and $(s_t^{(n)}, i_t^{(n)}, r_t^{(n)}(da))_{t\in\mathbb{R}_+} = (\frac{1}{n}S_t^{(n)}, \frac{1}{n}I_t^{(n)}, \frac{1}{n}R_t^{(n)}(da))_{t\in\mathbb{R}_+}$ converges to the weak solution of the following PDE:

$$\frac{ds_t}{dt} = \lambda_0 - \mu_0 s_t + \lambda_1(s_t, i_t)$$
$$\frac{di_t}{dt} = \lambda_1(s_t, i_t) - (\mu_1+\lambda_2) i_t - \lambda_3\left(i_t, \int_{\mathbb{R}_+} \psi(a)\rho_t(a)da\right)$$
$$\frac{\partial \rho_t}{\partial t}(a) = -\partial_a \rho_t(a)$$
$$\rho_t(0) = \lambda_2 i_t + \lambda_3\left(i_t, \int_{\mathbb{R}_+} \psi(a)\rho_t(a)da\right).$$

Central Limit Theorem

- The fluctuations is defined by:

$$\eta_t^{(n)} = \begin{pmatrix} \sqrt{n}(s_t^{(n)} - s_t) \\ \sqrt{n}(i_t^{(n)} - i_t) \\ \sqrt{n}(r_t^{(n)}(da) - r_t(da)) \end{pmatrix}$$

- It is valued in $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathcal{M}_S(\mathbb{R}_+)$ with $\mathcal{M}_S(\mathbb{R}_+)$ that can not be meterized with respect to the weak convergence topology.

- Following work of Métivier (84) Méléard (98), we consider an embedding of this space in well-chosen distribution spaces.

$$W_0^{3,0} \hookrightarrow C^{2,0} \hookrightarrow C^{1,0} \hookrightarrow W_0^{1,1} \hookrightarrow C^{0,1}$$
$$C^{-0,1} \hookrightarrow W_0^{-1,1} \hookrightarrow C^{-1,0} \hookrightarrow C^{-2,0} \hookrightarrow W_0^{-3,0}$$

The sequence of fluctuation process $(\eta^{(n)})_{n\in\mathbb{N}^*}$ converges to a Gaussian semi-martingale

Simulations

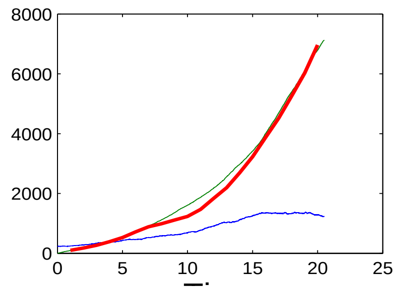


Figure: *Simulations using the microscopic model: Green: simulated cumulated number of detected individuals. Blue: simulated cumulated number of infectious individuals. Red: Observed cumulated number of detected individuals.*

Cuban AIDS Epidemics and Contact-tracing

Estimation with complete data

Estimation with unobserved data

Maximum likelihood estimation

Aim: estimation of λ_2 and $\lambda_3(S, I)$

Complete likelihood

$$\mathcal{L}_T^{(n)}(\theta) = e^{nT - \int_{u=0}^T (n\lambda_0 + \mu_0 n s_u^{(n)} + n\lambda_1(s_u^{(n)}, i_u^{(n)}) + (\mu_1 n + \lambda_2(\theta) n) i_u^{(n)} + n\lambda_3(i_u^{(n)}, \langle r_u^{(n)}, \psi\rangle, \theta))du} \times \prod_{k=1}^{K_T^{(n)}} L_\theta(E_k, (s_{T_k}^{(n)}, i_{T_k}^{(n)}, r_{T_k}^{(n)}(da))),$$

where: $L_\theta(E, (s, i, r(da))) = \lambda_0^{\mathbf{1}_{\{E=0\}}} (\mu_0 s)^{\mathbf{1}_{\{E=1\}}} \lambda_1(s, i)^{\mathbf{1}_{\{E=2\}}} (\lambda_2(\theta) i)^{\mathbf{1}_{\{E=3\}}} \lambda_3(i, \langle r, \psi\rangle, \theta)^{\mathbf{1}_{\{E=4\}}} (\mu_1 i)^{\mathbf{1}_{\{E=5\}}}.$

Maximum likelihood estimator
$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \log \mathcal{L}_T^{(n)}(\theta) = \arg\max_{\theta\in\Theta} l_T^{(n)}(\theta).$

Limit theorems for the MLE

For all $T > 0$ and any $(\theta^*, \theta) \in \Theta^2$, as $n \to \infty$, we have the following convergence in $\mathbb{P}_{\theta^*}$-probability,

$$\frac{1}{n}\{l_T^{(n)}(\theta^*) - l_T^{(n)}(\theta)\} \to K(\theta, \theta^*),$$

$$\text{where: } K(\theta, \theta^*) = \int_{t=0}^T \lambda_2(\theta^*) i_t^* \Phi\left(\frac{\lambda_2(\theta^*)}{\lambda_2(\theta)}\right) dt + \int_{t=0}^T \lambda_3(i_t^*, \langle r_t^*, \phi\rangle, \theta^*) \Phi\left(\frac{\lambda_3(i_t^*, \langle r_t^*, \phi\rangle, \theta^*)}{\lambda_3(i_t^*, \langle r_t^*, \phi\rangle, \theta)}\right) dt,$$

where $\Phi(x) = \log(x) + 1/x - 1$ and where $(s^*, i^*, r^*(da))$ is the solution of the PDE system with rate functions associated with θ^*.

Law of large numbers: Under identifiability and regularity assumptions, if the parameter space Θ is compact, the MLE is consistent:

$$\lim_{n\to\infty} \hat{\theta}_n = \theta^*, \text{ in } \mathbb{P}_{\theta^*} - \text{probability}.$$

Asymptotic normality: Under the proper regularity assumptions:

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \Rightarrow \mathcal{N}(0, \mathcal{I}_{\theta^*}^{-1}), \text{ as } n \to \infty.$$

where the *Fisher information matrix* is given by:

$$\mathcal{I}_\theta = -\int_{u=0}^T \Big\{ \mathcal{H}_\theta \lambda_2(\theta) i_u^* \left(\frac{\lambda_2(\theta^*)}{\lambda_2(\theta)} - 1\right) - \nabla_\theta \lambda_2(\theta) \cdot^t \nabla_\theta \lambda_2(\theta) \frac{\lambda_2(\theta^*) i_u^*}{\lambda_2(\theta)^2} + \mathcal{H}_\theta \lambda_3(i_u^*, \langle r_u^*, \psi\rangle, \theta) \left(\frac{\lambda_3(i_u^*, \langle r_u^*, \psi\rangle, \theta^*)}{\lambda_3(i_u^*, \langle r_u^*, \psi\rangle, \theta)} - 1\right) - \nabla_\theta \lambda_3(i_u^*, \langle r_u^*, \psi\rangle, \theta) \cdot^t \nabla_\theta \lambda_3(i_u^*, \langle r_u^*, \psi\rangle, \theta) \frac{\lambda_3(i_u^*, \langle r_u^*, \psi\rangle, \theta^*)}{\lambda_3(i_u^*, \langle r_u^*, \psi\rangle, \theta)^2} \Big\} du.$$

where$\mathcal{H}_\theta g$ denotes the hessian matrix of any twice differentiable function $\theta \in \Theta \mapsto g(\theta)$.

Examples

$$\begin{aligned}
\text{Model (A):}\quad \lambda_3(i,\langle r,\psi\rangle) &= \lambda_3\langle r,\psi\rangle,\\
\text{Model (B):}\quad \lambda_3(i,\langle r,\psi\rangle) &= \lambda_3\frac{\langle r,\psi\rangle i}{\langle r,\psi\rangle + i},\\
\text{Model (C):}\quad \lambda_3(i,\langle r,\psi\rangle) &= \lambda_3\langle r,\psi\rangle i.
\end{aligned}$$

Here $\theta = (\lambda_2,\lambda_3) \in \Theta \subset \mathbb{R}_+^{*2}$, and the true parameter is denoted $\theta^* = (\lambda_2^*,\lambda_3^*)$.

$$\widehat{\lambda}_2^{(n)} = \frac{\text{number of detections by random screening}}{n\int_{v=0}^T i_v^{(n)}dv},$$

Model (A):

$$\widehat{\lambda}_3^{(n,A)} = \frac{\text{number of detections by contact tracing}}{n\int_0^T \langle r_v^{(n)},\psi\rangle dv}.$$

Model (B):

$$\widehat{\lambda}_3^{(n,B)} = \frac{\text{number of detections by contact tracing}}{n\int_0^T \frac{i_v^{(n)}\langle r_v^{(n)},\psi\rangle}{(i_v^{(n)}+\langle r_v^{(n)},\psi\rangle)}dv}$$

Model (C):

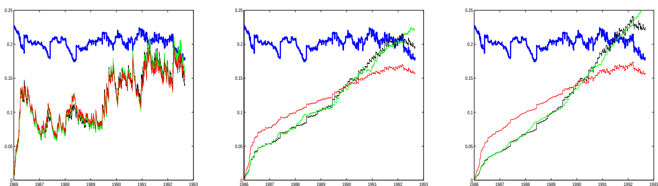
$$\widehat{\lambda}_3^{(n,C)} = \frac{\text{number of detections by contact tracing}}{n\int_0^T i_v^{(n)}\langle r_v^{(n)},\psi\rangle dv}.$$

Results

$\psi(a) = e^{-ca}$

model	parameter	estimated value	asymptotic std	log-likelihood
	λ_2	$9.57\,10^{-4}$	$4.34\,10^{-5}$	
		$c = 10^{-2}$		
(A)	λ_3	$3.90\,10^{-3}$	$2.30\,10^{-4}$	-2115
(B)	λ_3	$4.50\,10^{-3}$	$2.67\,10^{-4}$	-2119
(C)	λ_3	$1.85\,10^{-5}$	$1.09\,10^{-6}$	-2117
		$c = 10^{-3}$		
(A)	λ_3	$6.56\,10^{-4}$	$3.87\,10^{-5}$	-2138
(B)	λ_3	$1.30\,10^{-3}$	$7.72\,10^{-5}$	-2143
(C)	λ_3	$3.11\,10^{-6}$	$1.83\,10^{-7}$	-2140
		$c = 3\,10^{-4}$		
(A)	λ_3	$4.37\,10^{-4}$	$2.58\,10^{-5}$	-2144
(B)	λ_3	$1.10\,10^{-3}$	$6.54\,10^{-5}$	-2146
(C)	λ_3	$2.07\,10^{-6}$	$1.22\,10^{-7}$	-2147

Table: *Estimated parameters and asymptotic standard deviations.*



Evolution of the instantaneous detection rates $t \mapsto \widehat{\lambda}_2^{(n)} i_t^{(n)}$ *(in thick blue line) and* $t \mapsto \widehat{\lambda}_3^{(n)}(i_t^{(n)},\langle r_t^{(n)},\psi\rangle)$ *(in green (resp. red and black) line for Model (A) (resp. (B) and (C))). Left:* $c = 10^{-2}$*, middle:* $c = 10^{-3}$*, right:* $c = 3\,10^{-4}$

Cuban AIDS Epidemics and Contact-tracing

Estimation with complete data

Estimation with unobserved data

Unobserved data

- In practice, only the detected population is observed and present in our database.
- This gives us an information on the past size of the infectious population.
- Methods of completion of our database linked with simulations of the unobserved population conditionally to the observed one are complicated: no simple proposal law because of interactions.

Bayesian approach

- Idea:
  - There is a prior law with density $\pi(\theta)$ for the parameter θ
  - We want to approximate the conditional a posterior law with density $\pi(\theta\,|\,\text{observations})$
- Approximate Bayesian computation (Beaumont *et al.*, 2002): approximate $\pi(\theta\,|\,\text{summary statistics})$ instead of the posterior. If the summary statistics are sufficient, then it is the same. For this:
  - compute the summary statistics on the real sample
  - simulate parameters θ in the prior
  - simulate trajectories of the process with parameter θ, and compute the summary statistics for each simulated trajectory
  - Compute the conditional distribution (weights *à la* Nadaraya-Watson to compute the conditional expectation for instance)