Stochastic SIR model with contact-tracing

Cornell 4th 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

Age-structured component

• $R_i(da)$ is a point measure: $R_i(da) = \sum_{i=1}^{N_i} \delta_{a_i(t)}$

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

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

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

SIR model

$$\lambda_0 S \quad \lambda_1 S (I) \quad \lambda_2 I \quad R(da)$$

$$\mu_0 S \quad \mu_1 I \quad \lambda_3 (I, (R, \psi))$$

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

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

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 $(S_t(n), I_t(n), R_t(da))_{t \in \mathbb{R}_+} = (\frac{1}{n} S_t(n), \frac{1}{n} I_t(n), \frac{1}{n} R_t(da))_{t \in \mathbb{R}_+}$, converges to the weak solution of the following PDE:

$$\frac{dS_t}{dt} = \lambda_0 - \mu_0 S_t + \lambda_1 (S_t, I_t)$$
$$\frac{dI_t}{dt} = \lambda_2 S_t I_t - (\mu_1 + \lambda_2) I_t - \lambda_3 \left( \int_{\mathbb{R}_+} \psi(a) \rho_t(a) da \right)$$
$$\frac{\partial \rho_t}{\partial t}(a) = -\partial_a \rho_t(a)$$
$$\rho_t(0) = \lambda_2 I_t + \lambda_3 \left( \int_{\mathbb{R}_+} \psi(a) \rho_t(a) da \right).$$
Central Limit Theorem

- The fluctuations are defined by:
  \[ \eta_t^{(n)} = \begin{pmatrix} \sqrt{n}(s_t^{(n)} - s_t) \\ \sqrt{n}(i_t^{(n)} - i_t) \\ \sqrt{n}(c_t^{(n)}(da) - n_t(da)) \end{pmatrix} \]

- It is valued in \( \mathbb{R} \times \mathbb{R} \times \mathcal{M}_2(\mathbb{R}_+) \) with \( \mathcal{M}_2(\mathbb{R}_+) \) that can not be meterized with respect to the weak convergence topology.

- Following work of Mélard (98), we consider an embedding of this space in well-chosen distribution spaces:
  \[ W_0^{2.0} \hookrightarrow W_0^{1,0} \hookrightarrow W_0^{2,-1} \hookrightarrow C^{0,1} \]
  \[ C^{-0.1} \hookrightarrow W_0^{2,-1} \hookrightarrow C^{-1.0} \hookrightarrow C^{-2.0} \hookrightarrow W_0^{-3.0} \]

The sequence of fluctuation process (\( \eta_t^{(n)} \)) converges to a Gaussian semi-martingale

Simulations

Cuban AIDS Epidemics and Contact-tracing

Estimation with complete data

Estimation with unobserved data

Limit theorems for the MLE

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

\[ \frac{1}{n} \left( I_t^{(n)}(\theta^*) - I_t(\theta) \right) \to K(\theta, \theta^*), \]

where:

\[ K(\theta, \theta^*) = \int_0^T \lambda_2(\vartheta) \Phi \left( \lambda_2(\theta^*) \right) dt \]

\[ + \int_0^T \lambda_2(\vartheta) \Phi \left( \lambda_2(\theta^*) \right) dt, \]

where \( \Phi(x) = \log(x) + 1/x - 1 \) and \( (s_t^*, i_t^*, r_t^*(da)) \) is the solution of the PDE system with rate functions associated with \( \theta^* \).

Law of large numbers: Under identifiability and regularity assumptions, if the parameter space \( \Theta \) is compact, the MLE is consistent:

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

Maximum likelihood estimation

Aim: estimation of \( \lambda_2 \) and \( \lambda_3(S, I) \)

Complete likelihood

\[ L_T^{(n)}(\theta) = \exp \left[ -\int_0^T \left( \lambda_1(\vartheta) + \mu + a_n \right) \right. \]

where:

\[ \int E_t \left( \int_0^T \left( \lambda_1(\vartheta) + \mu + a_n \right) \right. \]

\[ \times \frac{1}{n} \left( I_t^{(n)}(\theta^*) - I_t(\theta) \right) \left. \right] \]

Maximum likelihood estimator

\[ \hat{\theta}_n = \arg \max_{\theta \in \Theta} L_T^{(n)}(\theta) = \arg \max_{\theta \in \Theta} I_T^{(n)}(\theta). \]

Asymptotic normality: Under the proper regularity assumptions:

\[ \sqrt{n} \left( \hat{\theta}_n - \vartheta^* \right) \Rightarrow \mathcal{N}(0, I_0^{-1}), \text{ as } n \to \infty. \]

where the Fisher information matrix is given by:

\[ I_\theta = - \int_0^T \left( \lambda_2(\vartheta) \right. \]

where \( \text{log} \) denotes the hessian matrix of any twice differentiable function \( \theta \in \Theta \to g(\theta) \).
Examples

Model (A): \[ \lambda_3(i, (r, \psi)) = \lambda_2(r, \psi), \]
Model (B): \[ \lambda_3(i, (r, \psi)) = \lambda_1(r, \psi) + \gamma, \]
Model (C): \[ \lambda_3(i, (r, \psi)) = \lambda_1(r, \psi). \]

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

\[ \lambda^{(n)}_2 = \frac{\text{number of detections by random screening}}{\int_0^T \kappa^0(\lambda^{(n)}(\psi), \psi) d\psi}, \]

Model (A):

\[ \lambda^{(n, A)}_2 = \frac{\text{number of detections by contact tracing}}{\int_0^T \kappa^0(\lambda^{(n, A)}(\psi), \psi) d\psi}, \]

Model (B):

\[ \lambda^{(n, B)}_2 = \frac{\text{number of detections by contact tracing}}{\int_0^T \kappa^0(\lambda^{(n, B)}(\psi), \psi) d\psi}, \]

Model (C):

\[ \lambda^{(n, C)}_2 = \frac{\text{number of detections by contact tracing}}{\int_0^T \kappa^0(\lambda^{(n, C)}(\psi), \psi) d\psi} = \lambda^{(n)}_2. \]

Results

\[ \psi(a) = e^{-a} \]

<table>
<thead>
<tr>
<th>model</th>
<th>parameter</th>
<th>estimated value</th>
<th>asymptotic std</th>
<th>log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>( \lambda_2 )</td>
<td>9.57 \times 10^{-3}</td>
<td>4.34 \times 10^{-3}</td>
<td>-2115</td>
</tr>
<tr>
<td>(B)</td>
<td>( \lambda_2 )</td>
<td>4.50 \times 10^{-3}</td>
<td>2.67 \times 10^{-4}</td>
<td>-2119</td>
</tr>
<tr>
<td>(C)</td>
<td>( \lambda_2 )</td>
<td>1.85 \times 10^{-5}</td>
<td>1.09 \times 10^{-6}</td>
<td>-2117</td>
</tr>
</tbody>
</table>

Table: Estimated parameters and asymptotic standard deviations.

Cuban AIDS Epidemics and Contact-tracing

Estimation with complete data

Estimation with unobserved data

Bayesian approach

- Idea:
  - There is a prior law with density \( \pi(\theta) \) for the parameter \( \theta \)
  - 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 \( \theta \) in the prior
- simulate trajectories of the process with parameter \( \theta \), 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)

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.