Inference for Epidemic Data using Diffusion processes with small diffusion coefficient

Romain GUY with C. Laredo and E. Vergu

Laboratoire de Probabilités et modèles aléatoires (Paris Diderot)
Unité Mathématiques et Informatique Appliquées, INRA Jouy-en-Josas

1st April 2011

Outline

Plan

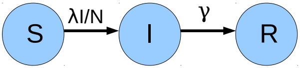
Notations, and model assumptions

Notations

N : population size
m : initial invectives
λ : transmission rate
γ : recovery rate
R_0 : basic reproduction number
$S(t), I(t)$: numbers of susceptibles, infecteds, $s(t) = \frac{S(t)}{N}, i(t) = \frac{I(t)}{N}$: proportion of susceptibles, infecteds

Assumptions

- Homogenous mixing in closed population
- Discrete observations of S and I on a fixed interval $[0, T]$, with sampling interval Δ $(T = n\Delta)$



Markov Pure Jump Model

Let $X_0 = (N - m, m)$ and $X_t = (S_t, I_t)$.

Transitions and holding time

$(S, I) \xrightarrow{\frac{\lambda}{N} SI} (S - 1, I + 1)$
$(S, I) \xrightarrow{\gamma I} (S, I - 1)$
Exponentials holding times

Maximum Likelihood Estimators from complete observations (all jumps)

$$\hat{\lambda}_{MLE} = N \frac{N - m - S(T)}{\int_0^T S(t) I(t) dt}, \; \hat{\gamma}_{MLE} = \frac{N - S(T) - I(T)}{\int_0^T I(t) dt}$$

Asymptotic Normality

$$\sqrt{N} \left(\begin{pmatrix} \hat{\lambda}_{MLE} - \lambda_0 \\ \hat{\gamma}_{MLE} - \gamma_0 \end{pmatrix} \right) \underset{N \to \infty}{\longrightarrow} \mathcal{N} \left(0, \begin{pmatrix} var(\lambda_0) & 0 \\ 0 & var(\gamma_0) \end{pmatrix} \right)$$

with $var(\lambda_0) = \frac{\lambda_0^2}{(1 - \frac{m}{N})(1 - s(T))}, var(\gamma_0) = \frac{\gamma_0^2}{(1 - \frac{m}{N})(1 - \frac{m}{N} - s(T) - i(t))}$

ODE Model

Let $x_{\lambda,\gamma}(t) = (s(t), i(t))$, $(s(0), i(0)) = (1 - \frac{m}{N}, \frac{m}{N})$

Classical ODE System

$$\frac{ds}{dt} = -\lambda si$$

$$\frac{di}{dt} = \lambda si - \gamma i$$

Do not depend on the population size !

Observations

Discrete observations at times $t_k = k\Delta$, $k = 0, ..., n$ $X_{t_k} = x_{\lambda,\gamma}(t_k) + \epsilon_k$ with $\epsilon_k \underset{iid}{\sim} \mathcal{N}_2\left(0, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\right)$

Statistical Inference for ODE

Least Square Estimator

$$LSE(\lambda,\gamma) = \sum_{k=0}^{n} (X_{t_k} - x_{\lambda,\gamma}(t_k))^2, \; (\hat{\lambda}_{LSE}, \hat{\gamma}_{LSE}) = \underset{(\lambda,\gamma)\in\Theta}{argmin} LSE(\lambda,\gamma)$$

Asymptotic Normality

$$\sqrt{n}\left(\begin{pmatrix} \hat{\lambda}_{LSE} - \lambda_0 \\ \hat{\gamma}_{LSE} - \gamma_0 \end{pmatrix}\right) \underset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\right)$$

Diffusion approximation model

Let $X_t = (s_t, i_t)$, B_1, B_2 two independent Brownians motions, $(s(0), i(0)) = (1 - \frac{m}{N}, \frac{m}{N})$

Stochastic Differential Equation

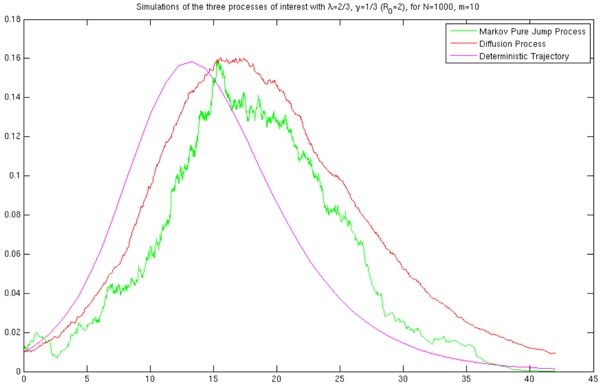
$$ds_t = -\lambda s_t i_t dt + \frac{1}{\sqrt{N}} \sqrt{\lambda s_t i_t} dB_1(t)$$

$$di_t = (\lambda s_t i_t - \gamma i_t) dt - \frac{1}{\sqrt{N}} \sqrt{\lambda s_t i_t} dB_1(t) + \frac{1}{\sqrt{N}} \sqrt{\gamma i_t} dB_2(t)$$

Remarks

- Classic Approximation : studies asymptotic properties of Pure Jump process (Ethier and Kurz) or Van Kampen approximation of Master Equation
- MLE untractable when discretely observed
- Multidimensionnal diffusion processes
- Small noise $\sim \frac{1}{\sqrt{N}}$ in large population
- Parameters (λ, γ) both in drift and diffusion coefficient

Exemple of trajectory : proportion of infecteds over time



Plan

Theoretical model and existing results

Let X_t^ϵ be the unique strong solution of the SDE

- $dX_t^\epsilon = b(\alpha, X_t^\epsilon)dt + \epsilon\sigma(\beta, X_t^\epsilon)dB_t, \; X_0 = x_0 \in \mathbb{R}^p$
- We observe X_t^ϵ at times $t_k = k\Delta$ on a fixed interval $[0, T]$ $(T = n\Delta)$
- $\sigma(\beta, x) \in M_p(\mathbb{R}), b(\alpha, x) \in \mathbb{R}^p, \Sigma(\beta, x) = {}^t\sigma(\beta, x)\sigma(\beta, x) \in GL_p(\mathbb{R})$

Existing estimation result for high-frequency data (Gloter and Sorensen (2009))

Under the condition $\exists \rho > 0, \frac{1}{\epsilon n^\rho}$ bounded
For a class of contrast processes, associated Minimum Contrast Estimators (MCEs) are consistent and :

$$\begin{pmatrix} \epsilon^{-1}(\hat{\alpha}_{\epsilon,n} - \alpha_0) \\ \sqrt{n}(\hat{\beta}_{\epsilon,n} - \beta_0) \end{pmatrix} \underset{n\to\infty,\epsilon\to 0}{\longrightarrow} N\left(0, \begin{pmatrix} I_b^{-1}(\alpha_0, \beta_0) & 0 \\ 0 & I_\sigma^{-1}(\alpha_0, \beta_0) \end{pmatrix}\right)$$

$I_b^{-1}(\alpha_0, \beta_0)$ being optimal

Main Idea of our inference approach (Generalization of Genon-Catalot(90))

Use of Taylor's Stochastic Expansion formula (Azencott (82))

$$X_t^\epsilon = x_\alpha(t) + \epsilon g_{\alpha,\beta}(t) + \epsilon^2 R_{\alpha,\beta}^\epsilon(t)$$

where $x_\alpha(t)$ is the deterministic solution $\frac{dx_\alpha(t)}{dt} = b(\alpha, x_\alpha(t)),\ x(0) = x_0 \in \mathbb{R}^p$

$$dg_{\alpha,\beta}(t) = \frac{\partial b}{\partial x}(\alpha, x_\alpha(t))g_{\alpha,\beta}(t)dt + \sigma(\beta, x_\alpha(t))dB_t,\ \ g_{\alpha,\beta}(0) = 0_{\mathbb{R}^p}$$

where $R_{\alpha,\beta}^\epsilon$ satisfies :

$$\sup_{t\in[0,T]} \{\|\epsilon R_{\alpha,\beta}^\epsilon(t)\|\} \underset{\mathbb{P},\epsilon\to 0}{\longrightarrow} 0$$

Let Φ_α be the invertible matrix solution of
$\frac{d\Phi_\alpha}{dt}(t, t_0) = \frac{\partial b}{\partial x}(\alpha, x_\alpha(t))\Phi_\alpha(t, t_0),\ \ \Phi_\alpha(t_0, t_0) = I_p$

Properties of $g_{\alpha,\beta}$

- $g_{\alpha,\beta}$ is a gaussian process (and we can obtain is analytic expression)
- $g_{\alpha,\beta}(t_k) = \Phi_\alpha(t_k, t_{k-1})g_{\alpha,\beta}(t_{k-1}) + Z_k^{\alpha,\beta}$
- $Z_k^{\alpha,\beta}$ independent gaussian variables

Contrast process derived from $Z_k^{\alpha,\beta}$

$$U_{\Delta,\epsilon}(\alpha,\beta)) = \sum_{k=1}^{n} log\left[det\left(\Sigma(\beta, X_{t_{k-1}})\right)\right] + \frac{1}{\epsilon^2\Delta}\sum_{k=1}^{n} {}^tN_k(\alpha)\Sigma^{-1}(\beta, X_{t_{k-1}})N_k(\alpha)$$

$$\text{with } N_k(\alpha) = X_{t_k} - x_\alpha(t_k) - \Phi_\alpha(t_k, t_{k-1})\left[X_{t_{k-1}} - x_\alpha(t_{k-1})\right].$$

$$(\hat{\alpha}_{\epsilon,\Delta}, \hat{\beta}_{\epsilon,\Delta}) = \underset{(\alpha,\beta)\in\Theta}{argmin}\ U_{\Delta,\epsilon}(\alpha,\beta)$$

Results for high frequency data ($\Delta \to 0$)

Under the condition $\epsilon^2 n \underset{\epsilon,\Delta\to 0}{\longrightarrow} 0$

$$\begin{pmatrix} \epsilon^{-1}(\alpha\hat{_{\epsilon,\Delta}} - \alpha_0) \\ \sqrt{n}(\beta\hat{_{\epsilon,\Delta}} - \beta_0) \end{pmatrix} \underset{n\to\infty,\epsilon\to 0}{\longrightarrow} N\left(0, \begin{pmatrix} I_b^{-1}(\alpha_0,\beta_0) & 0 \\ 0 & I_\sigma^{-1}(\alpha_0,\beta_0) \end{pmatrix}\right)$$

.

Results for low frequency data (Δ and n being fixed)

n fixed : no asymptotic results for $\hat{\beta}_{\epsilon,\Delta}$

β known

We only consider $\hat{\alpha}_{\epsilon,\Delta}(\beta_0) = \underset{\alpha\in\Theta_a}{argmin}\ U_{\Delta,\epsilon}(\alpha,\beta_0)$

and then $\epsilon^{-1}\left(\hat{\alpha}_{\epsilon,\Delta}(\beta_0) - \alpha_0\right) \underset{\epsilon\to 0}{\longrightarrow} \mathcal{N}(0, I_\Delta^{-1}(\alpha_0,\beta_0))$

with $I_\Delta(\alpha_0,\beta_0) \underset{\Delta\to 0}{\longrightarrow} I_b(\alpha_0,\beta_0)$

β unknown

We modify the contrast process in a "conditional least square" contrast :

$$U_\epsilon\left(\alpha,(X_{t_k})_{k\in\{1,..,n\}}\right) = \frac{1}{\epsilon^2}\sum_{k=1}^{n} {}^tN_k(X,\alpha)N_k(X,\alpha) \quad (1)$$

then $\hat{\alpha}_\epsilon = \underset{\alpha\in\Theta_a}{argmin}\ U_\epsilon(\alpha)$

satisfies :$\epsilon^{-1}\left(\hat{\alpha}_\epsilon - \alpha_0\right) \underset{\epsilon\to 0}{\longrightarrow} \mathcal{N}(0, \tilde{I}_\Delta^{-1}(\alpha_0,\beta_0))$

Plan

Return on the diffusion model

Stochastic Differential Equation

$$\begin{aligned} ds_t &= -\lambda s_t i_t dt + \frac{1}{\sqrt{N}} \sqrt{\lambda s_t i_t} dB_1(t), \\ di_t &= (\lambda s_t i_t - \gamma i_t) dt - \frac{1}{\sqrt{N}} \sqrt{\lambda s_t i_t} dB_1(t) + \frac{1}{\sqrt{N}} \sqrt{\gamma i_t} dB_2(t) \end{aligned}$$

Simulations

$N \in [1000; 10000]$, $\Delta = 1$ (1 observation/ day)

- $\epsilon = \frac{1}{\sqrt{N}} << 1$
- Δ is fixed
- $\alpha = (\lambda, \gamma) = \beta \Rightarrow$ Special case : Results for known β hold if we replace each β occurence with α.

Simulation study (using Matlab)

Algorithm

1. Exact simulation of an epidemic with Markov Pure Jump process (Gillespie algorithm with choice of N, m, λ, γ)
2. Calculation of $\hat{\lambda}_{MLE}, \hat{\gamma}_{MLE}$ (observation of the whole path of the process)
3. Observations of discrete data on a fixed interval (1 observation/day) up to extinction time
4. Estimation phase for *LSE*, Gloter and Sorensen contrast, our method for unknown β (Conditionnaly least square contrast), and for $\alpha = \beta$, using minimization function of Matlab (fminsearch)

Presented results

We repeat 100 times this algorithm to build empiric confidence intervals and avoid early extinction events

Remark

Step 4 : (Analytic power) Short time of estimation

Simulation results ($R_0 = 2$)

For $N = 1000$, $m = 10$, $\lambda = 2/3$, $\gamma = 1/3$, 1 *data/day* $\Rightarrow$ 40 *observations*

Method	$\hat{\lambda}$	CI_{95} *empiric*	CI_{95} *theoretical*
MLE(all data)	0.657	[0.645; 0.669]	[0.643; 0.671]
LSE	0.643	[0.618; 0.668]	[0.633; 0.653]
Gloter Sorensen$(\hat{\alpha}_{\epsilon,n})$	0.622	[0.611; 0.634]	[0.622; 0.622]
$\hat{\alpha}_{\epsilon,\Delta}(\alpha = \beta)$	0.656	[0.651; 0.660]	[0.656; 0.657]
$\hat{\alpha}_{\epsilon}(\beta$ *unknown*$)$	0.645	[0.642; 0.649]	[0.644; 0.646]
Method	$\hat{\gamma}$	CI_{95} *empiric*	CI_{95} *theoretical*
MLE(all data)	0.336	[0.330; 0.342]	[0.330; 0.342]
LSE	0.329	[0.314; 0.343]	[0.321; 0.337]
Gloter Sorensen$(\hat{\alpha}_{\epsilon,n})$	0.386	[0.367; 0.404]	[0.386; 0.386]
$\hat{\alpha}_{\epsilon,\Delta}(\alpha = \beta)$	0.336	[0.333; 0.338]	[0.335; 0.336]
$\hat{\alpha}_{\epsilon}(\beta$ *unknown*$)$	0.331	[0.330; 0.333]	[0.330; 0.331]

Global remarks

- $\hat{\beta}_{\epsilon,n}$ and $\hat{\beta}_{\epsilon,\Delta}$ do not provide satisfying results (not shown)
- Red : True value of parameters not in the CI
- Green : best point estimation

Simulations results ($R_0 = 1.2$)

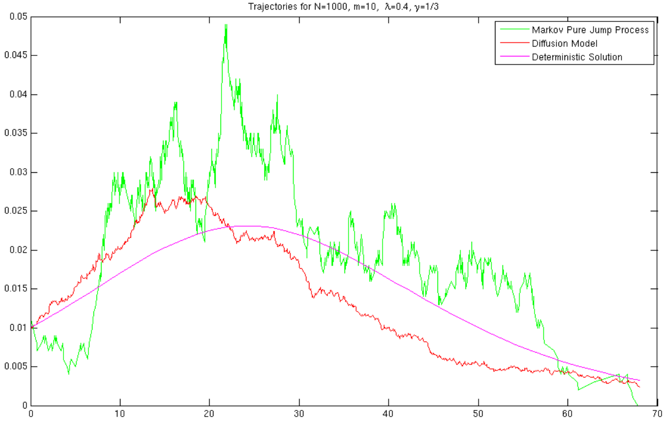
For $N = 10000$, $m = 100$, $\lambda = 0.4$, $\gamma = 1/3$,
1 *observation/day* $\Rightarrow$ 115 *observations*

Method	$\hat{\lambda}$	*empiric* CI_{95}	$\hat{\gamma}$	*empiric* CI_{95}
MLE(all data)	0.397	[0.395; 0.399]	0.337	[0.336; 0.338]
LSE	0.387	[0.377; 0.398]	0.328	[0.319; 0.337]
Gloter Sorensen($\hat{\alpha}_{\epsilon,n}$)	0.410	[0.409; 0.411]	0.330	[0.330; 0.331]
$\hat{\alpha}_{\epsilon,\Delta}(\alpha = \beta)$	0.396	[0.396; 0.397]	0.329	[0.329; 0.330]
$\hat{\alpha}_{\epsilon}(\beta$ *unknown*$)$	0.396	[0.396; 0.397]	0.336	[0.336; 0.337]

For $N = 1000$, $m = 10$, $\lambda = 0.4$, $\gamma = 1/3$,
1 *observation/day* $\Rightarrow$ 65 *observations*

Method	$\hat{\lambda}$	*empiric* CI_{95}	$\hat{\gamma}$	*empiric* CI_{95}
MLE(all data)	0.387	[0.381; 0.393]	0.362	[0.346; 0.377]
LSE	0.402	[0.364; 0.440]	0.353	[0.322; 0.384]
Gloter Sorensen($\hat{\alpha}_{\epsilon,n}$)	0.382	[0.381; 0.383]	0.336	[0.334; 0.338]
$\hat{\alpha}_{\epsilon,\Delta}(\alpha = \beta)$	0.396	[0.394; 0.397]	0.357	[0.355; 0.359]
$\hat{\alpha}_{\epsilon}(\beta$ *unknown*$)$	0.392	[0.385; 0.399]	0.363	[0.359; 0.367]

Exemple of trajectory for R_0=1.2 and N=1000



Limits and perspectives

Limits

1. Limits of the SIR model
2. The total number of infecteds is assumed observed (instead of incidences, more realistic assumption)
3. The two coordinates (s_t, i_t) are assumed observed (which is not often the case)

Next Directions

1. Results hold for any autonomous system (SEIR,...)
2. Modifying the diffusion model (observe (u_t, v_t) with $u_t = s_t i_t, v_t = i_t$) and observe integrated diffusion
3. Work to do

Thank you !