Inference for epidemic data using diffusion processes with small diffusion coefficient

Romain Guy, Catherine Laredo, Elisabeta Vergu

Inference for epidemic data using diffusion processes with small diffusion coefficient

Romain GUY[1,2] with C. Laredo[1,2] and E. Vergu[2]

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

4èmes Rencontres Des Jeunes Statisticiens (Aussois)
7 septembre 2011

Introduction

A work guided by data

- Estimation of key parameters of the epidemic dynamics
- Often low frequency time series
- Often aggregated and incomplete datasets.

Objective

Provide estimators with good properties for the diffusion model approximating the epidemic dynamics

Outline

1. Classical SIR epidemic models
2. Parametric inference for discretely observed diffusion processes
3. Back to the epidemics and simulations results

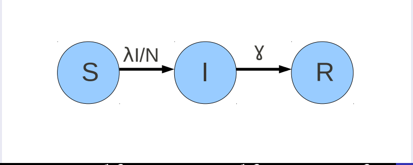
Notations and model assumptions

Notations

N : population size ; m : initial infectives
λ : transmission rate ; γ : recovery rate (=1/mean holding time in infective state)
$R_0 = \lambda/\gamma$: basic reproduction number (=mean number of secondary infections generated by an infective in an entirely susceptible population)
$S(t), I(t)$: numbers of susceptibles, infectives ; $s(t) = \frac{S(t)}{N}, i(t) = \frac{I(t)}{N}$: proportion of susceptibles, infectives

Assumptions

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



Markov pure jump model and Inference

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

Transitions

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

Observations

All jumps are observed

Maximum Likelihood Estimators (MLE) and asymptotic normality (Andersson and Britton 2000)

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

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

with $var(\lambda_0), var(\gamma_0)$ being explicit

ODE model and Inference (as a first approximation)

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

Classical ODE system (for N large)

$$\frac{ds}{dt} = -\lambda si \text{ and } \frac{di}{dt} = \lambda si - \gamma i,$$

Observations

Discrete observations $X_{t_k} = x_{\lambda,\gamma}(t_k) + \epsilon_k$ at times $t_k = k\Delta$ $(k = 0, ..., n)$, with $\epsilon_k \underset{iid}{\sim} \mathcal{N}_2(0, \Sigma)$

Least Square Estimator (LSE) and asymptotic normality

$$LSE(\lambda,\gamma) = \frac{1}{n}\sum_{k=0}^{n}(X_{t_k} - x_{\lambda,\gamma}(t_k))^2, \; (\hat{\lambda}_{LSE}, \hat{\gamma}_{LSE}) = \underset{(\lambda,\gamma)\in\Theta}{argminLSE}(\lambda,\gamma)$$

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

Diffusion approximation model

Let $X_t = (s_t, i_t)$, $(s_0, i_0) = (1 - \frac{m}{N}, \frac{m}{N})$ and B_1, B_2, two independent Brownians motions.

Stochastic Differential Equation

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

Remarks

- Classical approximation : before passing to the limit ($N \to \infty$) in the normalized system of the Markov jump process
- MLE untractable when the path is discretely observed

Framework

- Multidimensionnal diffusion processes
- Small noise $\sim \frac{1}{\sqrt{N}}$ in large population
- Parameters (λ, γ) both in drift and diffusion coefficients

General diffusion 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 $[0, T]$ $(T = n\Delta)$
- $\sigma(\beta, x) \in M_p(\mathbb{R}), b(\alpha, x) \in \mathbb{R}^p, \Sigma(\beta, x) = \sigma(\beta, x)\,{}^t\sigma(\beta, x)$ invertible.

Existing estimation results 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

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

with $I_b^{-1}(\alpha_0, \beta_0)$ explicit and optimal (= asymptotic variance for continuous observations on $[0, T]$).

For epidemics : $\epsilon^{-1} = \sqrt{N}$

Main idea of our inference approach (extension of Genon-Catalot(90))

Use of Taylor's stochastic expansion formula (Azencott (82))

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

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

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

and $R^\epsilon_{\alpha,\beta}$ satisfies

$$\underset{t\in[0,T]}{sup}\ \{\|\epsilon R^\epsilon_{\alpha,\beta}(t)\|\} \underset{\mathbb{P},\epsilon\to 0}{\longrightarrow} 0.$$

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

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

- $g_{\alpha,\beta}$ is a Gaussian process for which we can obtain the analytic expression.
- $g_{\alpha,\beta}(t_k) = \Phi_\alpha(t_k, t_{k-1})g_{\alpha,\beta}(t_{k-1}) + \sqrt{\Delta}Z_k^{\alpha,\beta}$
- $Z_k^{\alpha,\beta}$ are independent $\mathcal{N}\left(0, S_k^{\alpha,\beta}\right)$ variables.

Main idea of our inference approach (extension of Genon-Catalot(90))

Contrast process derived from $(Z_k^{\alpha,\beta})_k$-likelihood

$$\begin{aligned}
U_{\Delta,\epsilon}(\alpha,\beta) &= \epsilon^2 \sum_{k=1}^{n} log \left[det \left(S_k^{\alpha,\beta} \right) \right] \\
&+ \frac{1}{\Delta} \sum_{k=1}^{n} {}^t N_k(\alpha) (S_k^{\alpha,\beta})^{-1} N_k(\alpha) \\
\text{with } N_k(\alpha) &= X_{t_k} - x_\alpha(t_k) - \Phi_\alpha(t_k, t_{k-1}) \left[X_{t_{k-1}} - x_\alpha(t_{k-1}) \right].
\end{aligned}$$

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

Results for low frequency data (Δ and n fixed)

No asymptotic results for $\hat{\beta}_{\epsilon,\Delta}$

β known

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

$$\tilde{U}_{\epsilon,\Delta}\left(\alpha,(X_{t_k})_{k\in\{1,..,n\}}\right) = \frac{1}{\Delta}\sum_{k=1}^{n} {}^tN_k(\alpha)N_k(\alpha).$$

Then $\hat{\alpha}_\epsilon = \underset{\alpha\in\Theta_a}{argmin}\ \tilde{U}_{\epsilon,\Delta}\left(\alpha\right)$

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

For epidemics : $\alpha = (\lambda,\gamma) = \beta \Rightarrow$ Special case, results for known β hold if we replace each β occurence with α.

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

Contrast process

Using $\|S_k^{\alpha_0,\beta_0} - \Sigma(\beta_0, X_{t_{k-1}})\| \xrightarrow[\epsilon,\Delta\to 0]{} 0$, we consider :

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

Asymptotic Normality

Under the condition $\epsilon^2 n \xrightarrow[\epsilon,\Delta\to 0]{} 0$

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

.

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. Discrete observations on a fixed interval
4. Estimation phase for *LSE*, Gloter and Sorensen (2009) contrast, our MCE for $\alpha = \beta$, and for unknown β (conditionnaly least square contrast). Numerical optimisation using fminsearch.

Results

Mean of the point estimation on 1000 runs, theoretical confidence intervals for MCEs, empirical confidence interval for LSE, for each scenario and for each of the four estimators (+ MLE as a reference)

Simulated values

$N \in [1000; 10000]$, $\Delta \in [T/10; 1; T/100]$, $\gamma = 1/3$, $R_0 \in [1.2; 2]$

Example of trajectories : proportion of infectives over time ($N = 1000$, $R_0 = 2$)

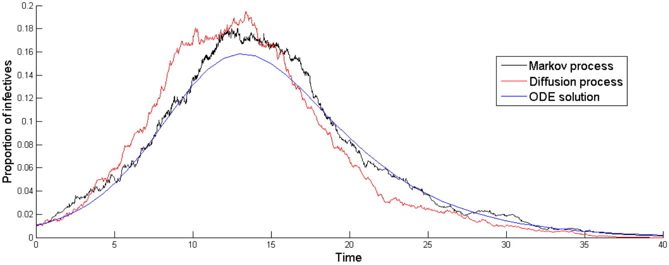
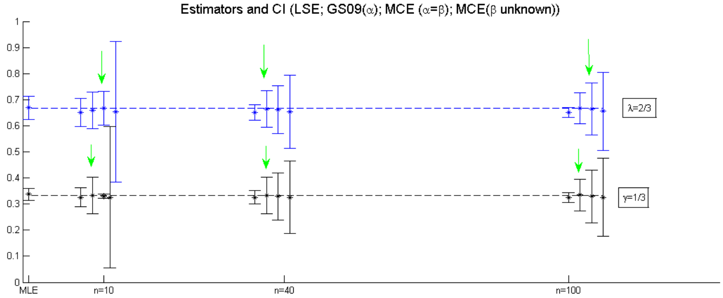


Figure: Trajectories of the three processes for N=1000 $R_0 = 2$, $\gamma = 1/3$, $T = 40$ and 1% of initial infectives

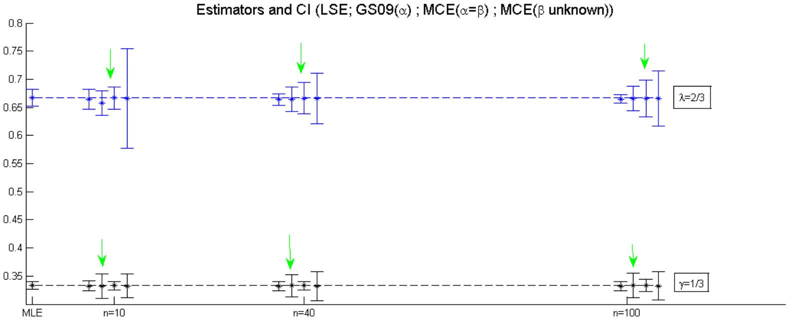
Simulation results for $N = 1000$, $R_0 = 2$, $\gamma = 1/3$, $T = 40$ and $n = 10, 40, 100$

Notations

green arrows = best ponctual estimator for one scenario
For each scenario, and each parameter in order : LSE, $GS09(\alpha)$, MCE($\alpha = \beta$), MCE(β unknown)



Simulation results for $N = 10000$, $R_0 = 2$, $\gamma = 1/3$, $T = 40$ and $n = 10, 40, 100$



Another Case : $R_0 = 1.2$

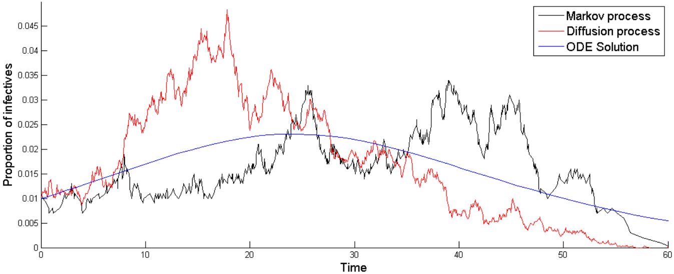
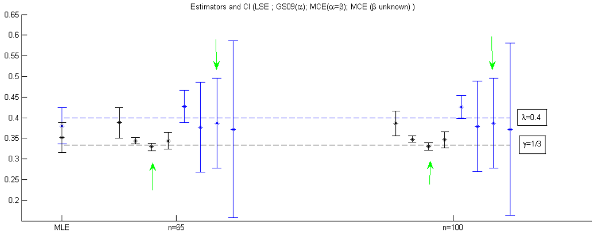
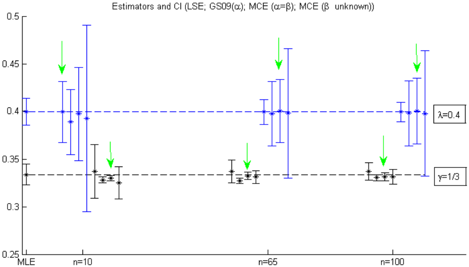


Figure: Trajectories of the three processes for $N = 1000$, $R_0 = 1.2$, $\gamma = 1/3$, $T = 65$ and 1% of initial infectives

Simulation results for $N = 1000, 10000$, $T = 65$ and $n = 10, 65, 100$

N=1000





N=10000

Limits and perspectives

To be refined

1. Limits of the SIR model : acceptable as a first attempt, possible to be refined to integrate more realistic assumptions (e.g. increase the state space dimension)
2. Idealized statistical framework : here the two system coordinates (s_t, i_t) are assumed observed (instead of a function of i_t, a more realistic assumption)

Next directions (work in progress)

1. Complexify the model : no difficulty if the new system is still an autonomous system, regardless to its size
2. Modify the statistical framework : observe integrated diffusion

References

References

H. Andersson and T. Britton.
Stochastic epidemic models and their statistical analysis.
Springer, 2000.

S.N. Ethier and T.G. Kurtz.
Markov processes : characterization and convergence.
Wiley, 1986.

V. Genon-Catalot.
Maximum contrast estimation for diffusion processes from discrete observations.
Statistics, 1990.

A. Gloter and M. Sorensen.
Estimation for stochastic differential equations with a small diffusion coefficient.
Stochastic Processes and their Applications, 2009.

Thank you for your attention !