Kalman filtering for noisy observations of partially observed epidemics and inference in mixed effects models

Romain NARCI[1]

Maud Delattre[2], Catherine Larédo[1] et Elisabeta Vergu[1]

(1) INRAE, Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE), Jouy-en-Josas
(2) UMR AgroParisTech-INRA MIA, Paris

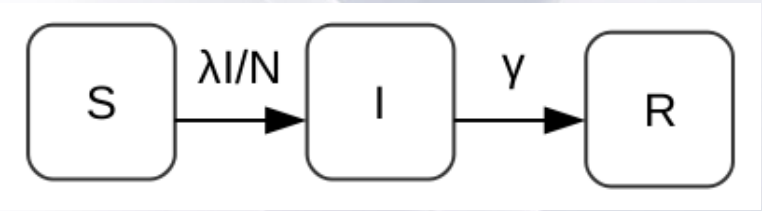
May 27th, 2020

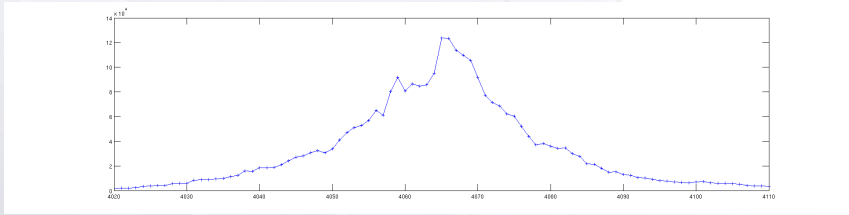
Plan

General context

- Epidemics $\rightarrow$ fast increase (number of new cases/unit of time) of a disease incidence in a given place at given time
- Understanding and predicting the epidemic dynamics $\rightarrow$ major role of mechanistic dynamical models to describe epidemic processes
- Here: epidemic dynamics modeled by stochastic processes
- Main issue: how to deal with partially observed epidemics (incomplete and noisy) for estimating the key parameters of the models ?

A simple mechanistic model for one outbreak: *SIR*





- Compartmental model where S, I, R = numbers of **susceptible**, **infectious**, **recovered** individuals
- N: population size **known** and **fixed**
- Key parameters:
  - λ : **transmission** rate
  - γ : **recovery** rate ; $d = \frac{1}{\gamma}$: **infectious period**
  - $R_0 = \frac{\lambda}{\gamma}$: **basic reproduction number**

Deterministic and stochastic models

- Deterministic model $\begin{cases} ds(t) & = -\lambda s(t) i(t) dt \\ di(t) & = (\lambda s(t) i(t) - \gamma i(t)) dt \\ (s(0), i(0)) & = (s_0, i_0) \end{cases}$
- Bidimensional Markov jump process on $\{0, \ldots, N\}^2$

  $(S, I) \to (S-1, I+1)$ at the rate $\lambda \frac{SI}{N}$

  $(S, I) \to (S, I-1)$ at the rate γI

Observations

- Observations in practice:
  - → number of infectious (or newly infected individuals)
  - → daily or weekly observations
  - → with measurement errors (reporting and diagnostic errors, etc.)
- SIR model:
  - → only one compartment is observed: I
  - → discretized
  - → noisy

$\Longrightarrow$ the inference of parameters is not direct

State of the art and objective

- Sophisticated existing inference methods (Maximum Iterated Filtering (Ionides et al. 2006); Approximate Bayesian Computation based on sequential Monte Carlo (Sisson et al. 2007); Particle Markov Chain Monte Carlo (Andrieu et al. 2010)) perform well but have some limitations in practice:
  - rely on data completion via computer simulations
  - substantial computation times
  - numerous tuning parameters
- Objective: propose a generic inference method easily practicable and able to deal with discrete, incomplete and noisy outbreak data
- Originality: approach based on a diffusion approximation with small variance coefficient + observation errors

Our approach

- Two-stage Gaussian approx.

  1) Gaussian approximation of the **epidemic density-dependent Markovian jump process** (Ethier & Kurtz (2005), Guy & al. 2015), using a diffusion based approach, with small coefficient (population size $N \to +\infty$) on a fixed interval $[0, T]$

  $\to$ Convergence of the normalized Markov jump process (**LLN**) to an explicit ODE solution and then, to a Gaussian process (**CLT**)

  2) Gaussian approximation of the **observation model** accounting for systematic noise

- Interest:

  – use of Kalman filter approaches to compute the log-likelihood of the observations

  – estimate model parameters

Gaussian approximation of the state model

- Diffusion approximation of the multidimensional normalized Markov jump process $X_N(t) := X(t)/N$:

$$dX_N(t) = b(X_N(t))dt + N^{-1/2}\sigma(X_N(t))dB(t); \qquad X_N(0) = \xi$$

- **Proposition. (Gaussian approximation)** Taylor expansion (Freidlin & Wentzell (1978)) of $X_N(t)$:

$$X_N(t) = x(t) + N^{-1/2}g(t) + N^{-1/2}R_N(t),$$

where $\sup_t \|R_N(t)\| \to 0$ in probability as $N \to +\infty$

$x(.) =$ ODE solution

$g(.) =$ Gaussian process depending on $b(.)$ and $\sigma(.)$

Gaussian approximation of the observation model

- Observations O:
  - discrete observations of I at times $t_k = k\Delta$, $k = 0, \ldots, n$ where $n =$ number of observations and $\Delta =$ sampling interval
  - noisy: use of a Binomial distribution with parameter p (reporting rate) for modeling the measurement errors
    $O(t_k) \sim \mathcal{B}(I(t_k), p)$
- Derivation of a conditional Gaussian approximation of the observation model taking into account the measurement errors

Kalman framework

- Vector of parameters $\theta = (\lambda, \gamma, p, s_0, i_0)$
- Linear Gaussian state space model:

$$X_k = F_k(\theta, \Delta) + A_{k-1}(\theta, \Delta)X_{k-1} + N^{-1/2} C_k(\theta, \Delta) U_k$$

$\rightarrow$ $F_k(\theta, \Delta) = x(\theta, t_k) - \Phi(\theta, t_k, t_{k-1})x(\theta, t_{k-1})$

$\rightarrow$ $A_{k-1}(\theta, \Delta) = \Phi(\theta, t_k, t_{k-1})$; Φ = resolvent matrix

$\rightarrow$ $C_k(\theta, \Delta) \approx \sqrt{\Delta}\sigma(\theta, x(\theta, t_k))$ for Δ small enough

- Approximate observation model:

$$Y_k = pI_N(t_k) + \sqrt{N^{-1}p(1-p)i(\theta, t_k)}V_k,$$

- $\{U_k\}_{k\geq 0}$, $\{V_k\}_{k\geq 0}$: independent standard Gaussian random variables

log-likelihood computation

- Approximate log-likelihood of the observations $y_0, \ldots, y_n$ given by:

$$\mathcal{L}(y_0, \ldots, y_n; \theta) = \log f(y_0; \theta) + \sum_{i=1}^{n} \log f(y_i | y_{0:i-1}; \theta)$$

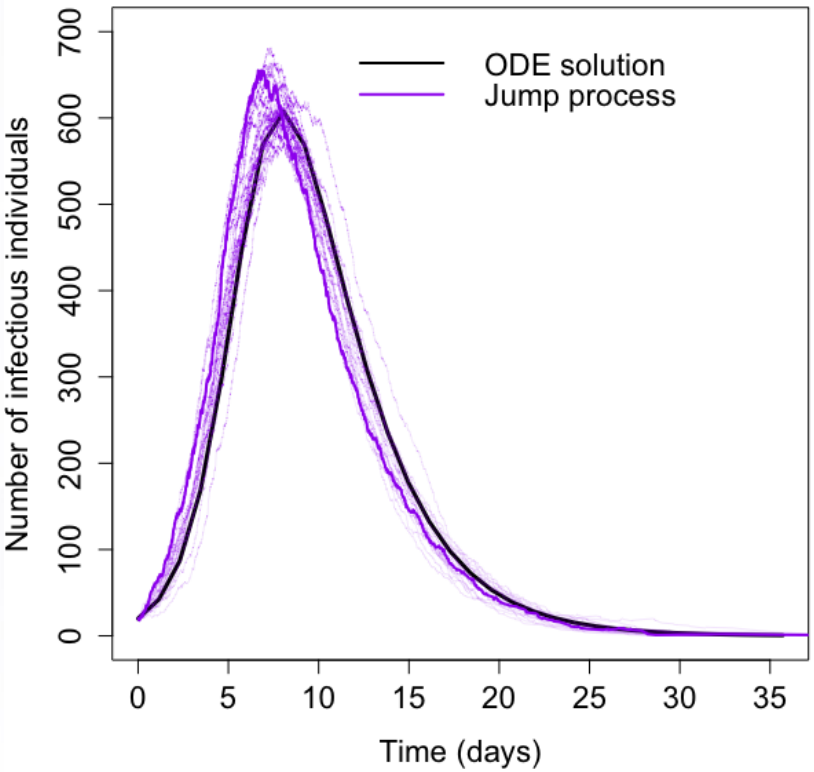
- Computing $\mathcal{L}(y_0, \ldots, y_n; \theta)$ requires an expression for each term $\log f(\ldots, \theta)$
- The distributions $Y_i | Y_{0:i-1}; \theta$ are Gaussian distributions with **mean** and **variance** computable using conditional moments $E(X_i | Y_{0:i-1}; \theta)$ and $V(X_i | Y_{0:i-1}; \theta)$ (Cappé, Moulines & Rydén, 2005)
- The Kalman filter is diverted to compute recursively the conditional densities $\log f(\ldots, \theta)$

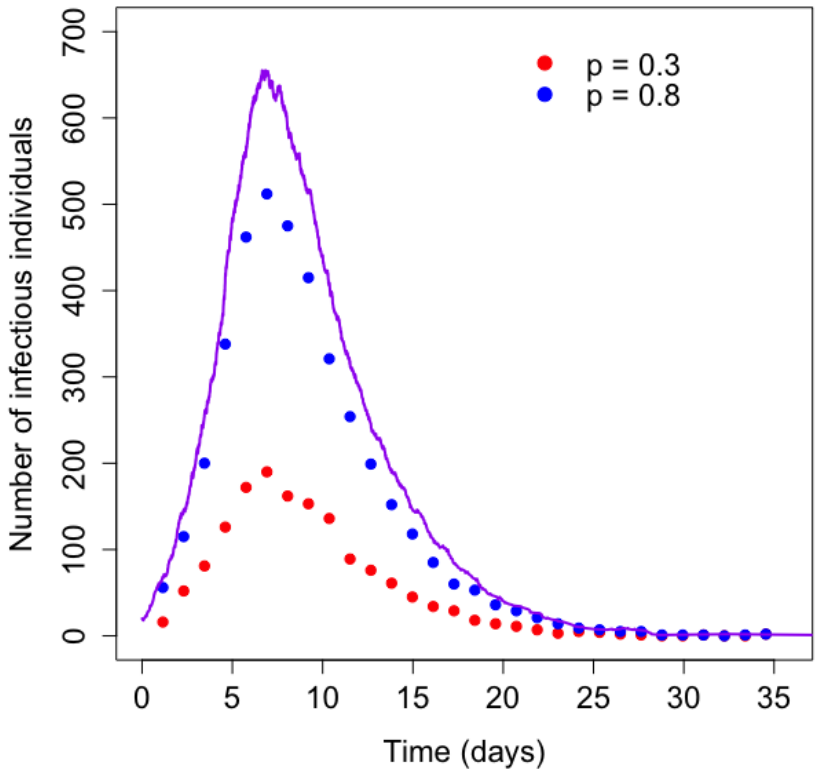
Simulation setting

- Exact simulation (Gillespie algorithm) of the Markov jump process associated to the SIR model: $X(t) = (S(t), I(t))$
- Simulation of the observations: $O_k \sim \mathcal{B}(I(t_k), p)/N$
- Parameters values: $\lambda = 1$, $\gamma = 1/3$, $s_0 = 0.99$, $i_0 = 0.01$
- 500 runs on $[0, T]$, T depending on the scenario
  - N (population size): 1000, 2000, 10000
  - n (number of observations): 10, 30, 100
  - p (reporting rate): 0.3, 0.8
- For each scenario (N, n, p): point estimators

Simulated data: $\lambda = 1$, $\gamma = 1/3$, $s_0 = 0.99$, $i_0 = 0.01$

Population size $N = 2000$ and number of observations $n = 30$





Comparison with Maximum Iterated Filtering (MIF)

1) Inference method (Ionides et al. (2006, 2011, 2015)) in the general framework of the partially observed Markov processes, implemented in the R package **POMP** (King et al. (2017)):
   - maximizes the likelihood obtained by Sequential Monte Carlo, also known as the particle filter
   - provides a Monte Carlo estimation of the maximum likelihood
   - requires the setting of several tuning parameters (number of particles, number of iterations, etc.)
2) Comparison:
   - Point estimators for each parameter (not shown here)
   - Boxplot of the relative bias for each parameter

Numerical results (relative bias) on 500 epidemics: $\lambda = 1$

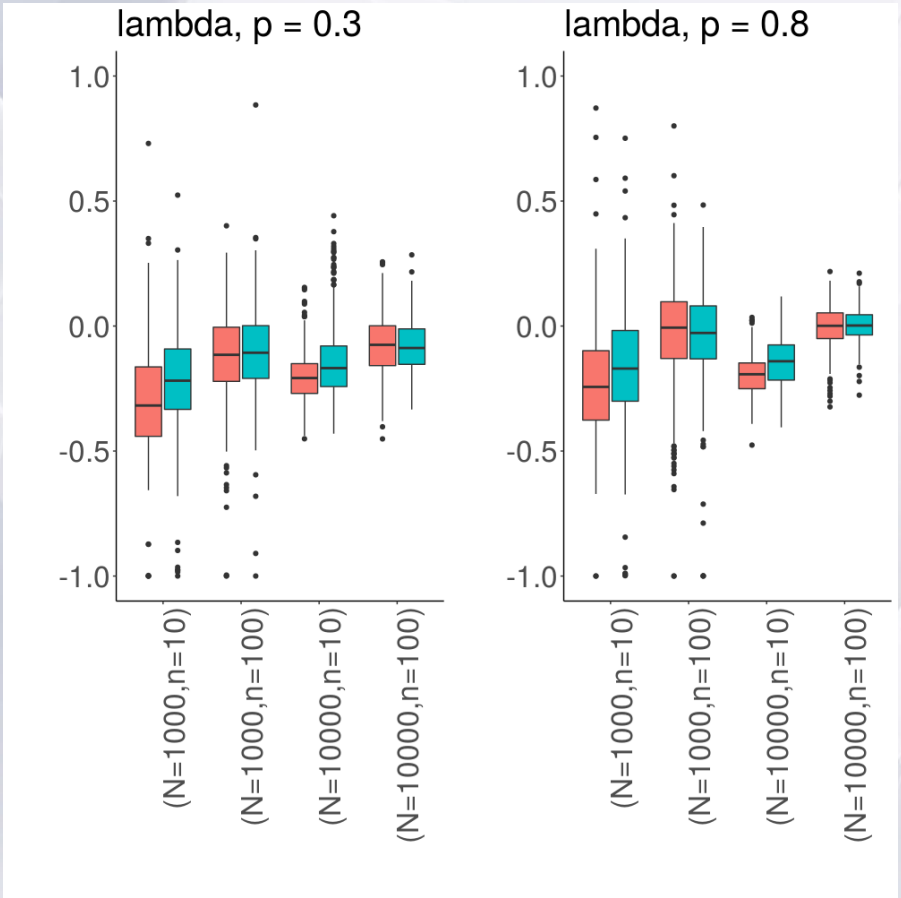


Figure: Relative bias of $\hat{\lambda}$ for the Kalman (•) and MIF (•) inference methods as a function of (N, n).

Numerical results (relative bias) on 500 epidemics: $\gamma = 1/3$

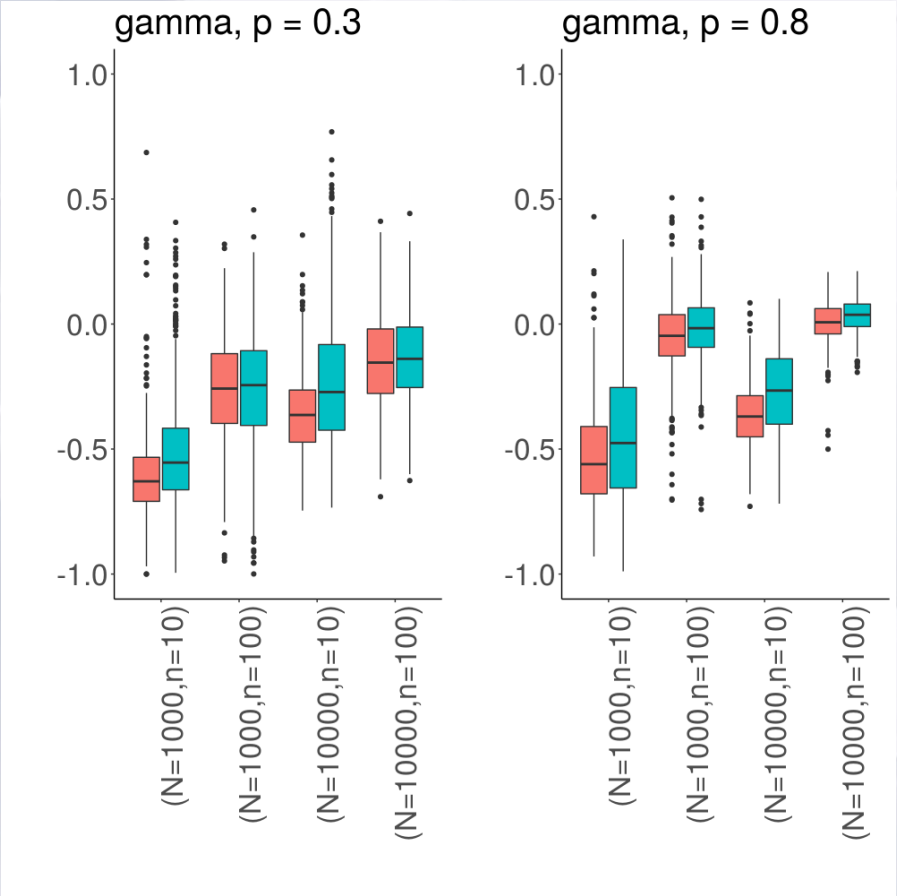


Figure: Relative bias of $\hat{\gamma}$ for the Kalman (•) and MIF (•) inference methods as a function of (N, n).

Partial conclusion

- Advantages of our approach:
  - accounting for specificities of the available data
  - can be extended to other mechanistic models
  - yields promising results: easy to implement and satisfying performances
- Pre-print available on HAL: https://hal.archives-ouvertes.fr/hal-02475936
- Application of our inference method on real data in progress

Context

- Recurrent outbreaks:

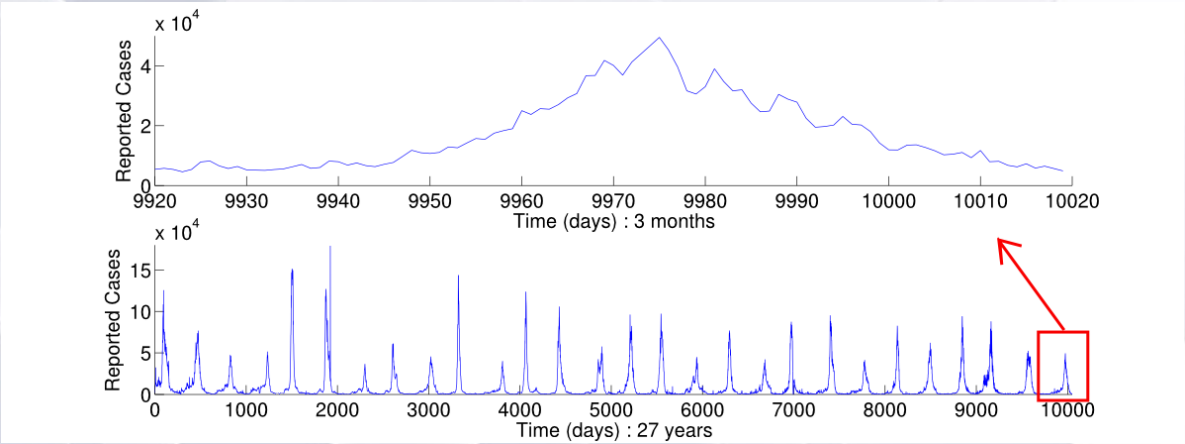


Figure: Numbers of new cases developing flu symptoms in Ile-de-France (Réseau Sentinelles, http://www.sentiweb.fr/)

- Multisite outbreaks: **Covid-19** in various regions in France
- Main issue: take directly into account the variability between epidemic events by mixed effects models in order to estimate key parameters

Framework

- Notations: **observations** $y = (y_u, 1 \leq u \leq U)$, unobserved **individual parameters** $\Phi = (\Phi_u, 1 \leq u \leq U)$ and a **vector of parameters** θ (variability intra- and inter-population)
- Example (noisy ODE):

$$Y_{u,k}|\Phi_u \sim \mathcal{N}(i_u(t_k), \sigma^2)$$
$$\lambda_u \sim \mathcal{N}(\lambda_{pop}, \omega_\lambda^2)$$
$$\gamma_u \sim \mathcal{N}(\gamma_{pop}, \omega_\gamma^2)$$

where u is the epidemic index, $i_u(t_k)$ is the ODE solution (infectious) at time t_k for an epidemic u, $\Phi_u = (\lambda_u, \gamma_u)$ and $\theta = (\lambda_{pop}, \gamma_{pop}, \omega_\lambda, \omega_\gamma, \sigma)$

Issue

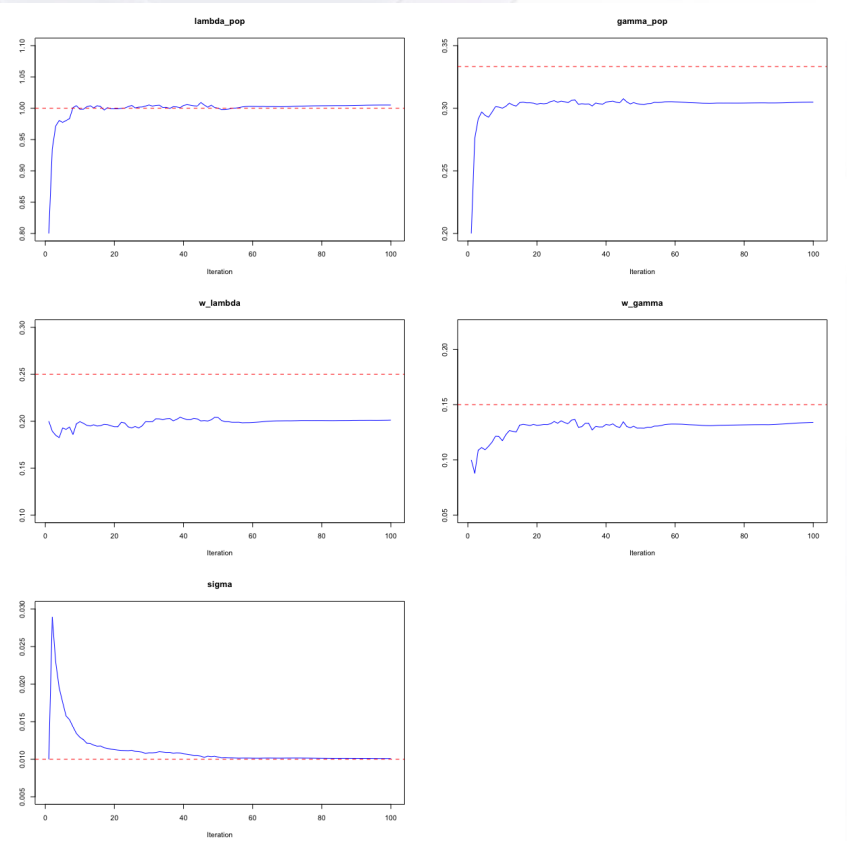
- Objective: estimate θ by maximizing the log-likelihood of the observations

$$\mathcal{L}(y;\theta) = \log p(y;\theta) = \int p(y,\Phi;\theta)\ d\Phi$$

- Problem: as Φ is not observed, the expression of the log-likelihood of the observations is not explicit
- When the relationship between observations y and individual parameters Φ is linear: **EM** algorithm
- If not: **SAEM** algorithm $\rightarrow$ needs to compute the joint distribution and to simulate from the conditional distribution $p(.|y;\theta)$

Numerical results

- $U = 50$ epidemics
- Initial values of the proportion of susceptibles/infectious $s_0 = 0.95,\ i_0 = 0.01$
- Parameters values: $\lambda_{pop} = 1$, $\gamma_{pop} = 1/3$, $\omega_\lambda = 0.25$, $\omega_\gamma = 0.15$, $\sigma = 0.01$
- Vector of the observation times: $(1, 3, 5, 7, 9, 12, 16, 20, 24, 28, 32, 36)$ days



Perspectives

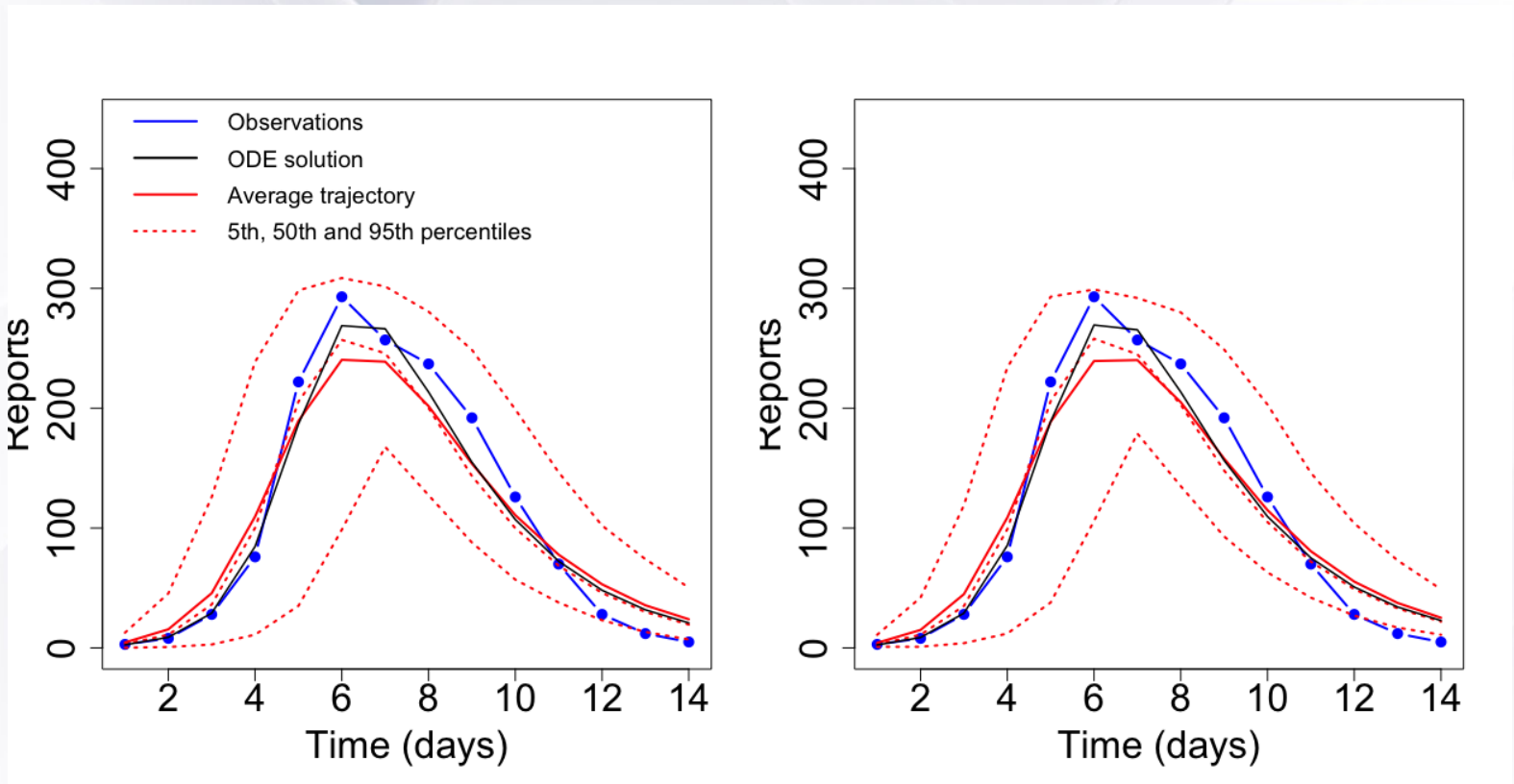
- Applications on real data in progress
- Perform inference when the observations are partially observed and noisy (first part of my PhD) by mixing the inference based on the **Kalman filter** with the **SAEM** algorithm:
  - → take into account the small variance coefficient in the SAEM algorithm
  - → investigate the theoretical properties in such a framework (convergence speed, etc.)

Real data set

- Influenza outbreak in a Britain boarding school in the north of England in January, 1978
- $N = 763$ boys were at risk and one boy from Hong-Kong became infectious from 15 to 18 January $\Longrightarrow$ $S(0) = 762$ and $I(0) = 1$
- Observations: number of infectious boys over 14 days with one observation per day $\Longrightarrow$ $I(t_k)$, $k = 0, \ldots, 14$ and $n = 14$ observations (SIR model)
- Estimate parameters : λ, γ and p

Results: post predictive check

$\hat{\lambda} = 1.72$; $\hat{\gamma} = 0.48$; $\hat{p} = 1.00$ $\hat{\lambda} = 1.71$; $\hat{\gamma} = 0.45$; $\hat{p} = 0.95$



Kalman **MIF**

Solution

When the relationship between observations y and individual parameters Φ is:

- linear $\rightarrow$ EM algorithm: given some initial values θ_0, iteration k updates θ_{k-1}^{EM} to θ_k^{EM} with the following steps:
  - **E-step**: Evaluate the quantity

$$Q_k^{EM}(\theta) = \mathbb{E}\left[\log p(y, \Phi; \theta) | y; \theta_{k-1}^{EM}\right]$$

  - **M-step**: Update the estimation of θ:

$$\theta_k^{EM} = \arg\max_\theta Q_k^{EM}(\theta)$$

- nonlinear $\rightarrow$ Monte Carlo EM (MCEM) ; **Stochastic Approximation EM (SAEM)**

A first mixed effects model (SIR)

- Model description:

$$Y_{u,v}|\Phi_u \sim \mathcal{N}(i_u(t_v), \sigma^2)$$
$$\lambda_u \sim \mathcal{N}(\lambda_{pop}, \omega_\lambda^2)$$
$$\gamma_u \sim \mathcal{N}(\gamma_{pop}, \omega_\gamma^2)$$

  where u is the pop. index, $i_u(t_v)$ is the ODE sol. (infectious) at time t_v for an epidemic u, $\Phi_u = (\lambda_u, \gamma_u)$ and $\theta = (\lambda_{pop}, \gamma_{pop}, \omega_\lambda, \omega_\gamma, \sigma)$

- Nonlinear relationship between obs. y and individual parameters Φ
  $\Longrightarrow$ use of the SAEM algo. which replaces the E-step by:
  - **Simulation step**: For $u = 1, 2, ..., U$, draw $\Phi_u^{(k)}$ from the conditional distribution $p(\Phi_u|y_u; \theta_{k-1})$
  - **Stochastic approx.**: Update $Q_{k-1}(\theta)$ according to $Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\log p(y, \Phi^{(k)}; \theta) - Q_{k-1}(\theta)\right)$ where (γ_k) is a decreasing seq. of positive numbers such that $\gamma_1 = 1$