The EpiLPS project: a new Bayesian tool for estimating the time-varying reproduction number

Oswaldo Gressani [a], Jacco Wallinga [b,c], Christian Althaus [d], Niel Hens [a,e], Christel Faes [a]

28th Annual Meeting of the Royal Statistical Society of Belgium
October 22, 2021

[a] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Data Science Institute, Hasselt University, Belgium.
[b] Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands.
[c] Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, The Netherlands.
[d] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland.
[e] Centre for Health Economics Research and Modelling Infectious Diseases, Vaxinfectio, University of Antwerp, Belgium.

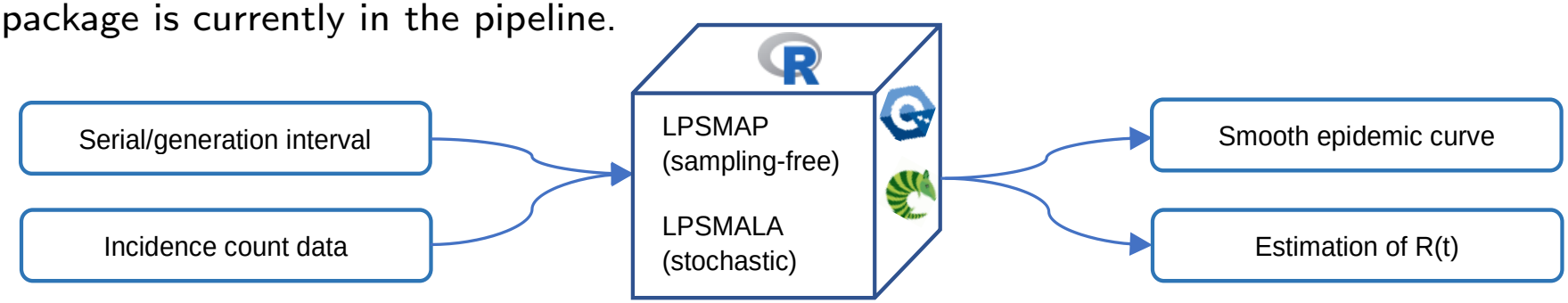
What is EpiLPS?

- **Epi**demiological modelling (tool) with **L**aplacian-**P**-**S**plines.
- Reproduction number $R(t)$: key metric to assess disease dynamics.
- A novel methodology for fast and flexible (approximate) Bayesian inference of $R(t)$.
  - Laplace approximations $\Rightarrow$ computationally very attractive.
  - P-splines $\Rightarrow$ flexible modeling framework with smooth estimates of the epidemic curve and $R(t)$.
- A set of (efficient) routines with intuitive functional calls for the end user.
- R package is currently in the pipeline.

What motivated EpiLPS?

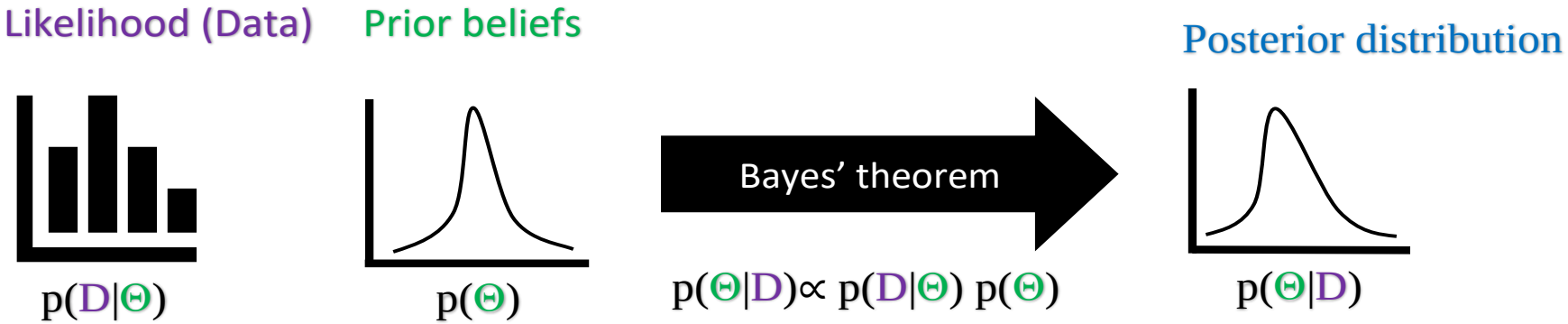

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES
HOME | ABOUT | SUBMIT | NEWS & NOTES | ALERTS / RSS
Search
Advanced Search
Comment on this paper
An approximate Bayesian approach for estimation of the reproduction number under misreported epidemic data
Oswaldo Gressani, Christel Faes, Niel Hens
doi: https://doi.org/10.1101/2021.05.19.21257438
This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.
Abstract | Full Text | Info/History | Metrics
Preview PDF

- Revisit model of Azmon et al. (2013)
- Will Laplace approximations do the job?



Bayes' theorem and the Laplace approximation in a nutshell

- Laplace approx. $\rightarrow$ key for making EpiLPS lightning fast, but what is it really?



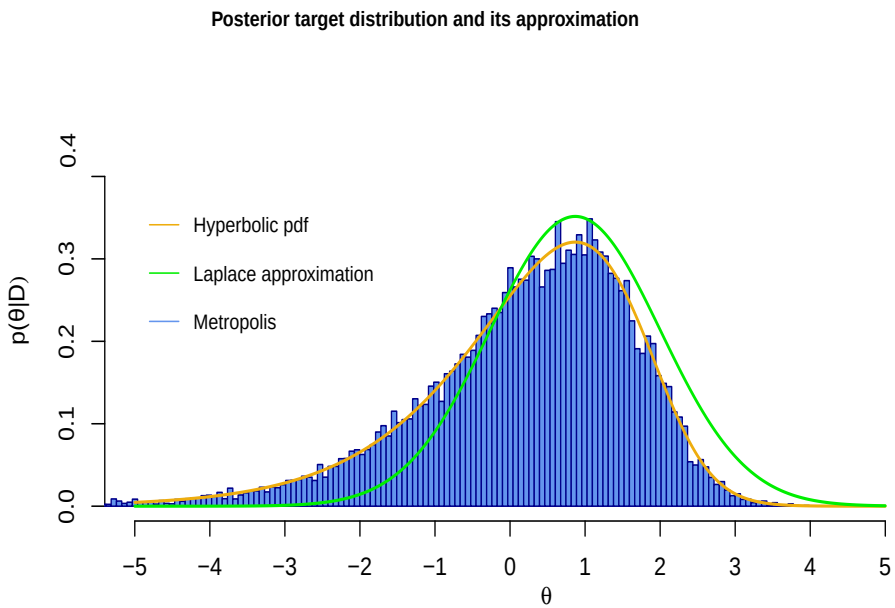
- Laplace approx.: simplify $p(\theta|\mathcal{D})$ but retain crucial features.
- Taylor expansion of $g(\theta) := \log p(\theta|\mathcal{D})$ around modal value $\hat{\theta}$:

$$g(\theta) \approx g(\hat{\theta}) + \frac{1}{2} g''(\hat{\theta})(\theta - \hat{\theta})^2.$$

- Apply exponential function on both sides:

$$\begin{aligned} \widetilde{p}_G(\theta|\mathcal{D}) &\propto \exp\left(-\frac{1}{2}\left(-g''(\hat{\theta})\right)(\theta - \hat{\theta})^2 \right) \\ &\Rightarrow \mathcal{N}\left(\hat{\theta}, \left(-g''(\hat{\theta})\right)^{-1}\right). \end{aligned}$$

- Ex: Hyperbolic distribution.



- Let $\{y_t, t = 1, \ldots, T\}$ be a time series of incidence counts.
- Negative Binomial model for y_t, i.e. $y_t \sim \text{NegBin}(\mu(t), \rho)$ following the parameterization of Piegorsch (1990):

$$p(y_t|\mu(t), \rho) = \frac{\Gamma(y_t + \rho)}{\Gamma(y_t + 1)\Gamma(\rho)} \left(\frac{\mu(t)}{\mu(t) + \rho}\right)^{y_t} \left(\frac{\rho}{\rho + \mu(t)}\right)^{\rho}, \ \mu(t) > 0, \ \rho > 0.$$

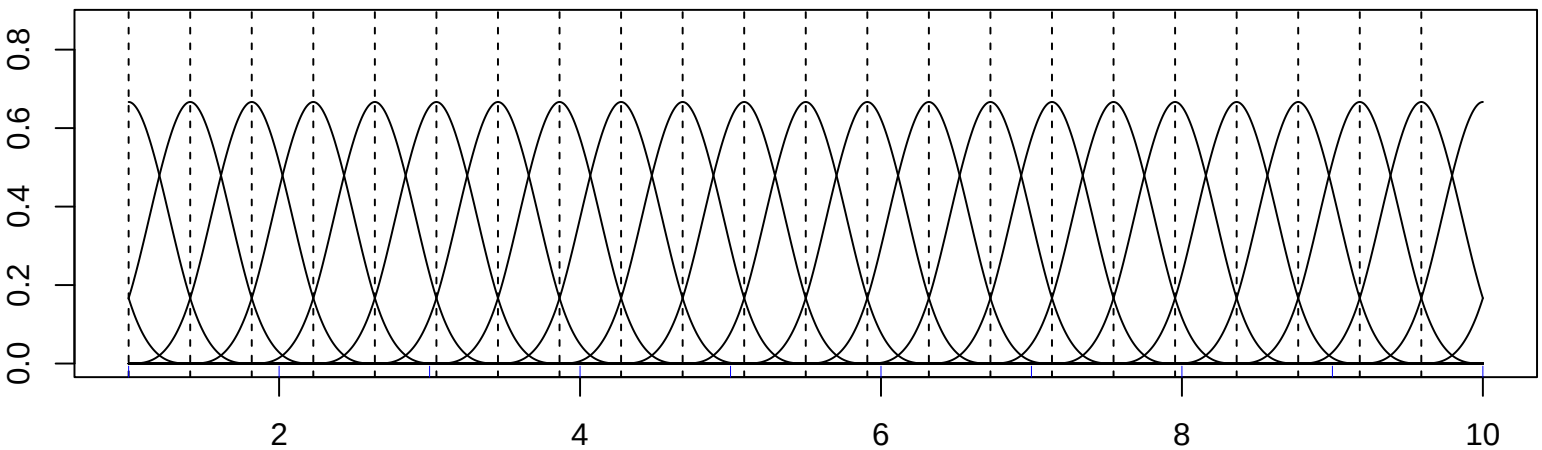
- Following Eilers and Marx (1996), we model $\mathbb{E}(y_t) = \mu(t)$ by means of cubic B-splines with amplitudes θ_k:

$$\log(\mu(t)) = \sum_{k=1}^{K} \theta_k b_k(t) = \boldsymbol{\theta}^\top b(t).$$

- Idea of P-splines: Fix K large enough and counterbalance the flexibility by introducing a roughness penalty on adjacent B-spline coefficients: $\lambda \boldsymbol{\theta}^\top P \boldsymbol{\theta}$.
- P is a penalty matrix and $\lambda > 0$ is the roughness penalty parameter to be maximized *a posteriori* $\rightarrow$ LPS**MAP**.

- B-spline basis defined on $\mathcal{T} = [r_l, T]$, where r_l is typically the first day of the epidemic (i.e. $r_l = 1$). In the **blapsr** package:

```
> lb <- 1   # Lower bound
> ub <- 10  # Upper bound
> xdom <- seq(lb,ub)
> Bsmat <- cubicbs(xdom, lb, ub, 25)
> plot(Bsmat) # Plot the basis
```



- In the Bayesian version Lang and Brezger (2004), the vector of B-spline coefficients is random $\Rightarrow$ Gaussian prior (based on random-walks).
- Also, priors on hyperparameters λ and ρ have to be specified.

- The Bayesian model formulation underlying EpiLPS is as follows:

$$
\begin{aligned}
y_t|\mu(t),\rho &\sim \text{NegBin}(\mu(t),\rho),\\
\log(\mu(t)) &= \boldsymbol{\theta}^\top b(t),\\
\boldsymbol{\theta}|\lambda &\sim \mathcal{N}_{\dim(\boldsymbol{\theta})}(0, Q_\lambda^{-1}),\\
\lambda|\delta &\sim \mathcal{G}\left(\phi/2, (\phi\delta)/2\right),\\
\delta &\sim \mathcal{G}(a_\delta, b_\delta),\\
\rho &\sim \mathcal{G}(a_\rho, b_\rho).
\end{aligned}
$$

- We fix $\phi = 2$, $a_\delta = b_\delta = 10$ and a proper (uninformative) prior on ρ with $a_\rho = b_\rho = 10^{-4}$.

- Denote by $\boldsymbol{\eta} := (\lambda, \rho)^\top$, the vector of model hyperparameters (to be optimized).

- Laplace approximations are used to approximate the posterior of $\boldsymbol{\theta}$ in three steps.

The "mechanics" of Laplacian-P-splines in 3 steps

- Objective: come up with approximated posteriors for $\boldsymbol{\theta}$ and $\boldsymbol{\eta} = (\lambda, \rho)^\top$.
- Gaussian prior $\boldsymbol{\theta}|\boldsymbol{\eta} \sim \mathcal{N}$ (LGM). Main goal is to approximate the joint posterior:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= \int p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathcal{D})\, d\boldsymbol{\eta} \\ &= \int p(\boldsymbol{\theta}|\boldsymbol{\eta}, \mathcal{D})\, p(\boldsymbol{\eta}|\mathcal{D})\, d\boldsymbol{\eta}. \end{aligned}$$

- In the philosophy of Tierney and Kadane (1986) and Rue et al. (2009):
  1. Laplace approximation to the conditional posterior:

$$p(\boldsymbol{\theta}|\boldsymbol{\eta}, \mathcal{D}) \to \widetilde{p}_G(\boldsymbol{\theta}|\boldsymbol{\eta}, \mathcal{D}).$$

  2. Approximation of the hyperparameter vector:

$$p(\boldsymbol{\eta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathcal{D})}{p(\boldsymbol{\theta}|\boldsymbol{\eta}, \mathcal{D})} \to \widetilde{p}(\boldsymbol{\eta}|\mathcal{D}) = \left.\frac{p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathcal{D})}{\widetilde{p}_G(\boldsymbol{\theta}|\boldsymbol{\eta}, \mathcal{D})}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(\boldsymbol{\eta})}.$$

  3. Approximation at the MAP for $\boldsymbol{\eta}$ (but other possibilities exist):

$$\widetilde{p}(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}_{\dim(\boldsymbol{\theta})}\left(\boldsymbol{\theta}^*(\boldsymbol{\eta}^*), \boldsymbol{\Sigma}^*(\boldsymbol{\eta}^*)\right).$$

The renewal equation "plug-in" estimator

- Remember what we have so far: $\widetilde{p}(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}_{\dim(\boldsymbol{\theta})}\left(\boldsymbol{\theta}^*(\boldsymbol{\eta}^*), \boldsymbol{\Sigma}^*(\boldsymbol{\eta}^*)\right)$.
- Denote by $\boldsymbol{\varphi} = \{\varphi_1, \ldots, \varphi_k\}$ the serial interval distribution and let $\widehat{\boldsymbol{\theta}} := \boldsymbol{\theta}^*(\boldsymbol{\eta}^*)$.
- Start with the renewal equation: $y_t = \sum_s R(t)\varphi_s y_{t-s}$ and write it as:

$$R(t) = \begin{cases} y_t & ;\text{for } t = 1, \\ y_t \left(\sum_{s=1}^{t-1} \varphi_s y_{t-s} \right)^{-1} & ;\text{for } 2 \leq t \leq k, \\ y_t \left(\sum_{s=1}^{k} \varphi_s y_{t-s} \right)^{-1} & ;\text{for } k < t \leq T. \end{cases}$$

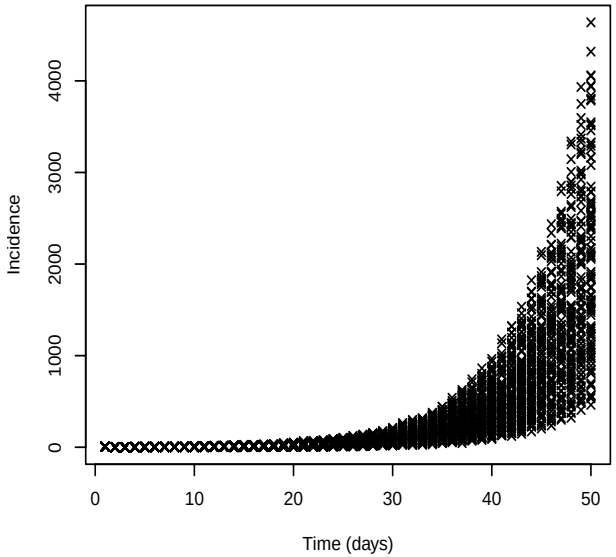
- Replace y_t by $\widehat{\mu}(t) = \exp\left(\widehat{\boldsymbol{\theta}}^\top b(t)\right)$ and write in compact notation:

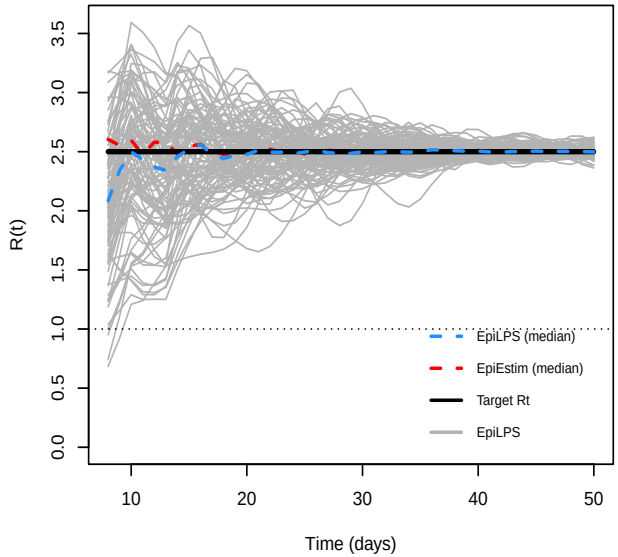
$$\begin{aligned} \widehat{R}(t) \quad = \quad & \exp\left(\widehat{\boldsymbol{\theta}}^\top b(t)\right) \Bigg\{ \mathbb{I}(t = 1) + \left(\sum_{s=1}^{t-1} \varphi_s \exp\left(\widehat{\boldsymbol{\theta}}^\top b(t-s)\right) \right)^{-1} \mathbb{I}(2 \leq t \leq k) \\ & + \left(\sum_{s=1}^{k} \varphi_s \exp\left(\widehat{\boldsymbol{\theta}}^\top b(t-s)\right) \right)^{-1} \mathbb{I}(k < t \leq T) \Bigg\}. \end{aligned}$$

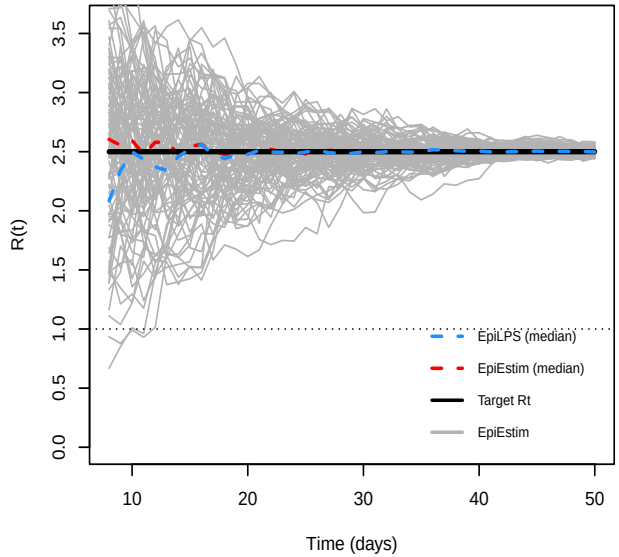
Validation of LPSMAP via simulations

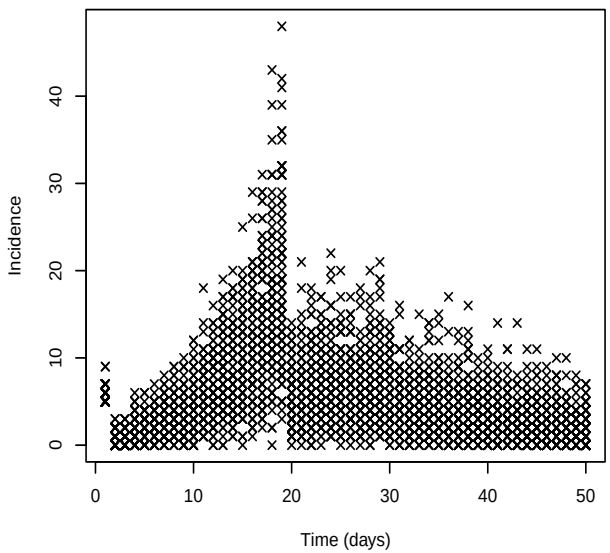
- Consider an epidemic of $T = 50$ days.
- Data generating process based on Poisson counts and the renewal equation $\rightarrow$ repeat $S = 100$ times.
- Compare with *estimate_R()* routine of **EpiEstim** package Cori et al. (2013).
- For **EpiEstim**, we need to specify a sliding window, (here 7 days).
- Four different scenarios:
  - Scenario 1: constant $R(t)$.
  - Scenario 2: assess impact of intervention, step function for $R(t)$.
  - Scenario 3: curved $R(t)$.
  - Scenario 4: decaying $R(t)$.

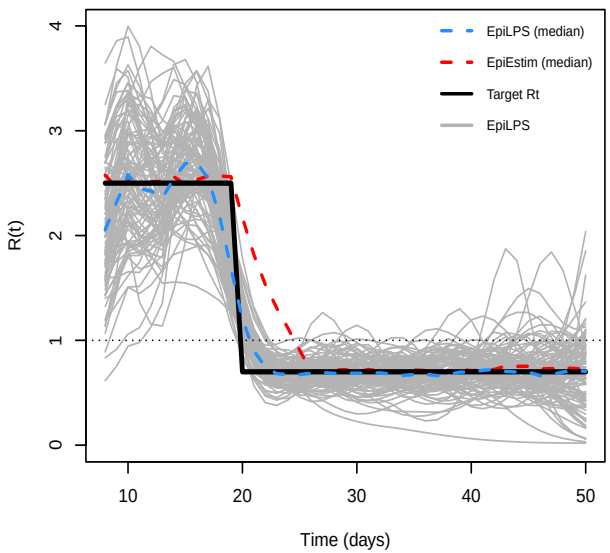
Scenarios 1 and 2

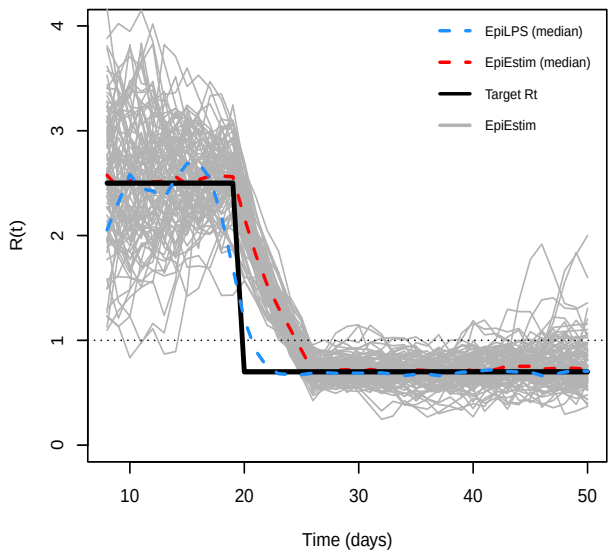




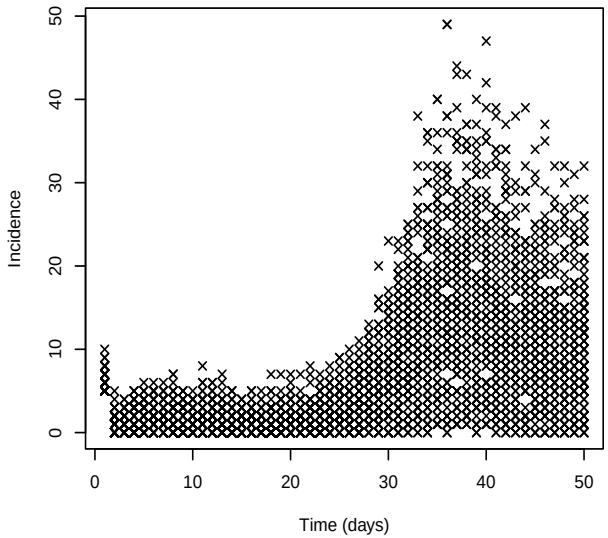


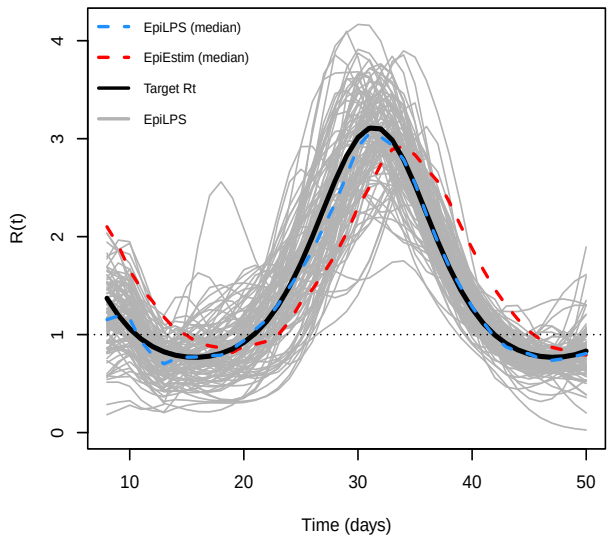


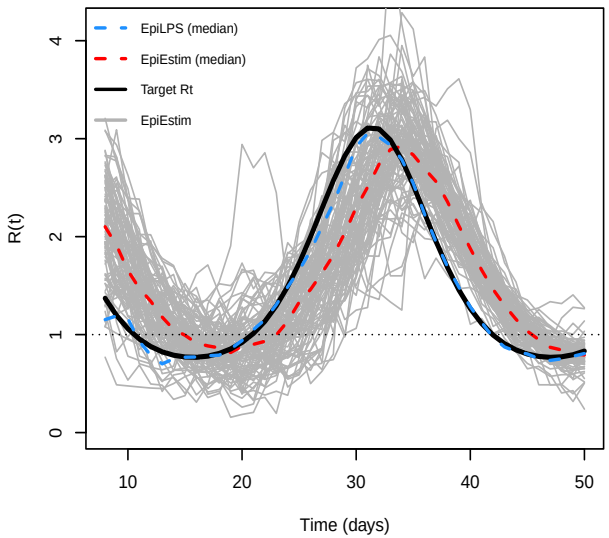


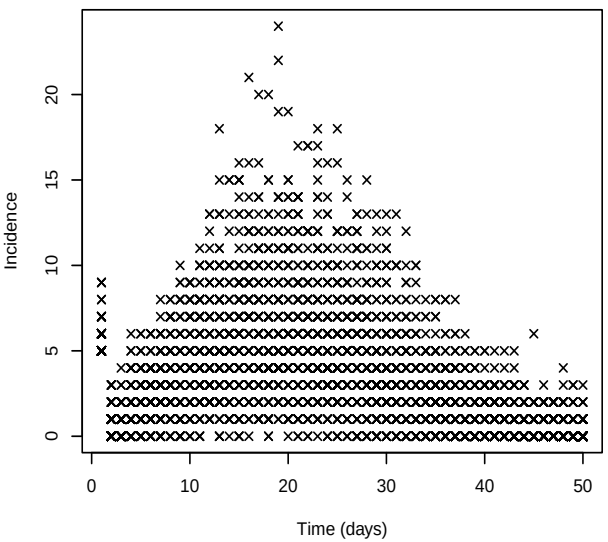


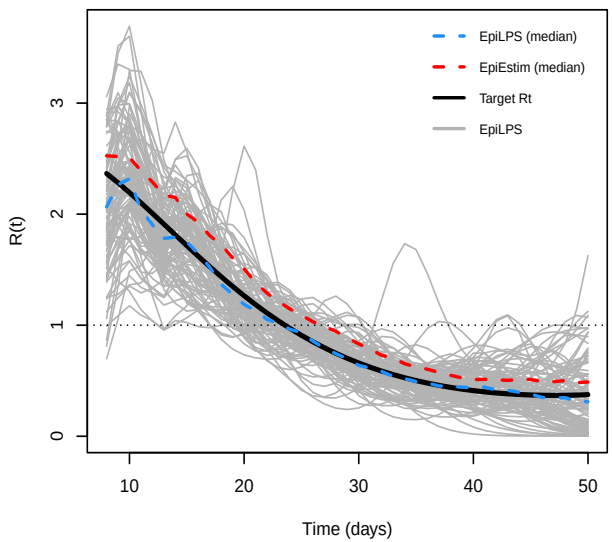
Scenarios 3 and 4

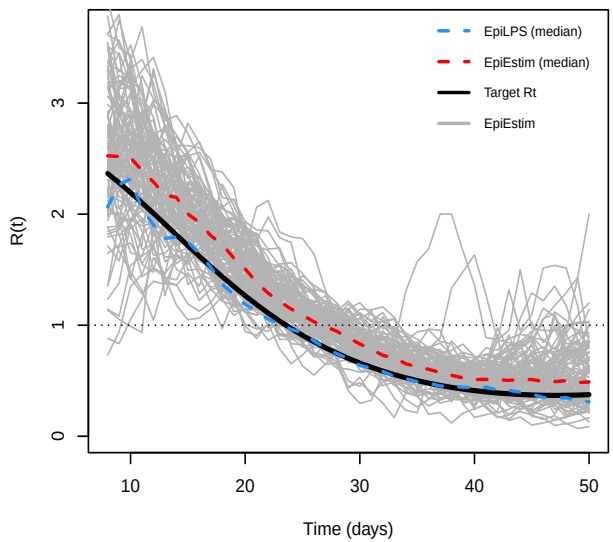








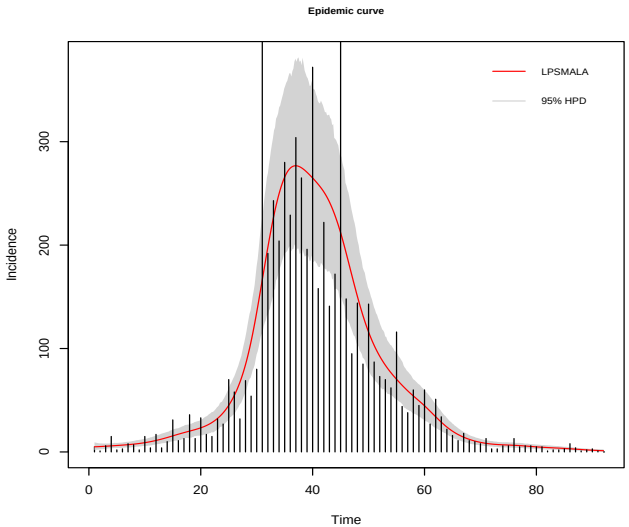


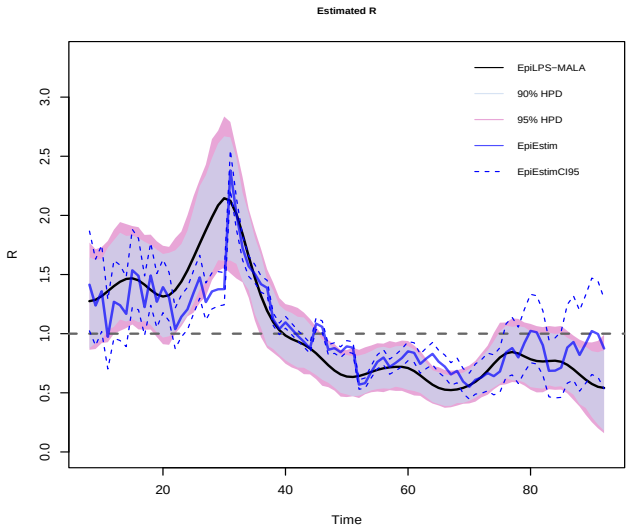


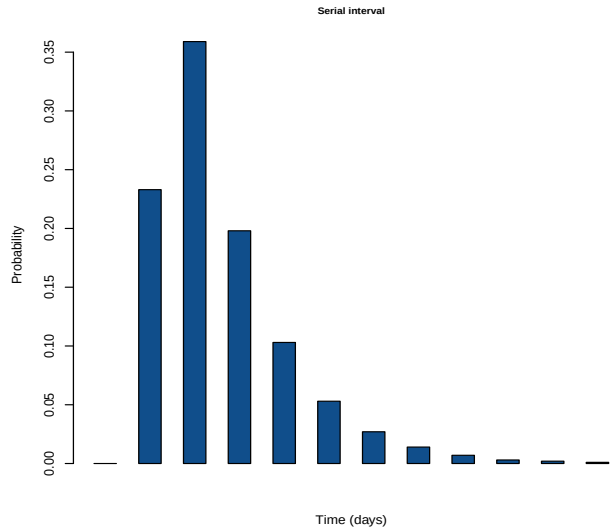
Real data applications with LPSMALA

We revisited historical outbreaks analyzed in Cori et al. (2013) with LPSMALA.

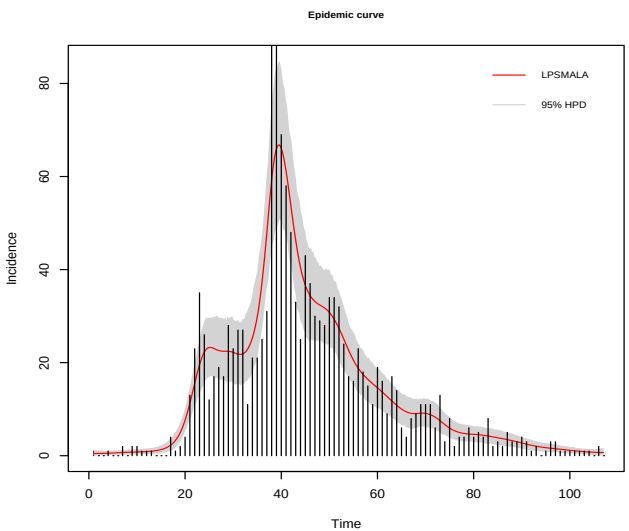
- Influenza pandemic in Baltimore (1918)

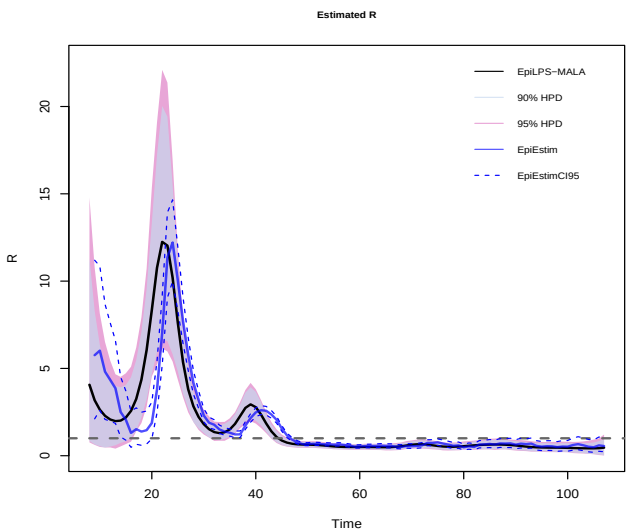


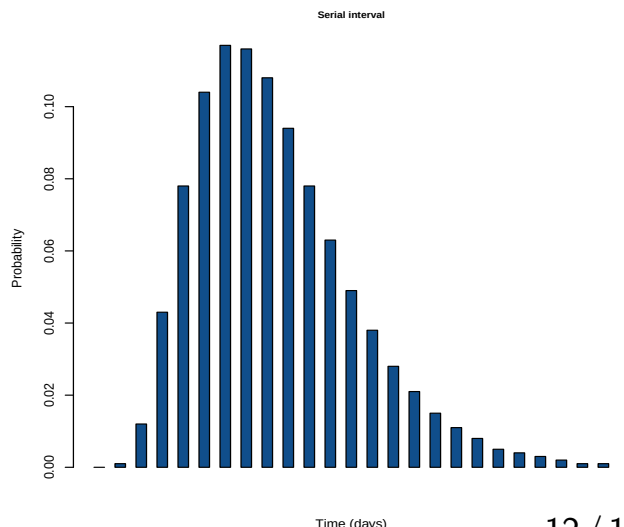




- SARS epidemic in Hong Kong (2003)







Main methodological differences between EpiLPS and EpiEstim:

- EpiLPS extracts a signal from the epidemic curve (smoothed version of mean incidence counts) and injects it into the renewal equation. Priors imposed on overdispersion parameter and smoothing parameter.
- EpiEstim imposes a (conjugate) Gamma prior on $R(t)$, so that *a posteriori* $R(t)$ is also Gamma distributed. In Cori et al. (2013), they assume a Gamma prior with mean 5 and standard deviation 5.
- In EpiLPS, we show that *a posteriori* $R(t)$ has a log-normal distribution.
- EpiEstim uses a Poisson distribution to model transmissions, while EpiLPS uses a Negative Binomial model (accounts for overdispersion).
- In EpiLPS no need to choose a time window, simply estimate $R(t)$ over the whole time interval of interest. EpiEstim requires a specification of a time window as an input (what is the optimal sliding window?).

Concluding remarks and take home message

- EpiLPS is a novel methodology for fast and flexible approximate Bayesian inference of the epidemic curve and the time-varying reproduction number $R(t)$.
- The computational efficiency is mainly due to Laplace approximations and its associated routine coded in C++ and integrated via Rcpp.
- The Bayesian P-splines framework allows to get smooth estimates of the epidemic curve and the latter information is used in conjuction with the epidemic renewal equation to provide a smooth estimate of $R(t)$.
- Credible intervals of $R(t)$ can be efficiently computed via the delta method.
- EpiLPS provides both a sampling-free approach (LPSMAP) and a fully stochastic approach based on Langevin diffusions (LPSMALA) to estimate $R(t)$.

Acknowledgments: This project is funded by the European Union's Research and Innovation Action under the H2020 work programme, EpiPose (grant number 101003688).


Cori, A., Ferguson, N.M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. American Journal of Epidemiology, 178(9):1505–1512.


Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. Statistical Science, 11(2):89-102.


Gressani, O., Faes, C. and Hens, N. (2021). Laplacian P-splines for Bayesian inference in the mixture cure model. ArXiv preprint. arxiv.org/abs/2103.01526


Gressani, O., Faes, C. and Hens, N. (2021). An approximate Bayesian approach for estimation of the reproduction number under misreported epidemic data. MedRxiv preprint. doi.org/10.1101/2021.05.19.21257438


Gressani, O. and Lambert, P. (2018). Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. Computational Statistics & Data Analysis, 124:151-167.


Gressani, O. and Lambert, P. (2021). Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines. Computational Statistics & Data Analysis, 154.


Piegorsch, Walter W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. Biometrics, 863-867.