CRIStAL Centre de Recherche en Informatique, Signal et Automatique de Lille

Université de Lille

cnrs

Estimation of the reproduction number of the Covid19 pandemic

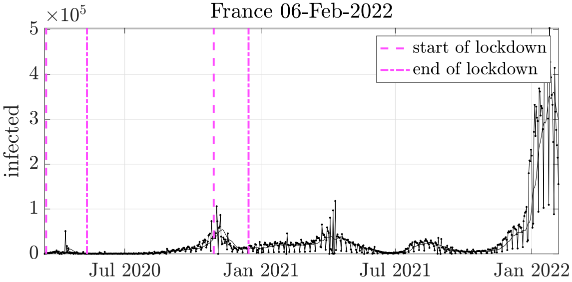
Maximum A Posteriori and credibility intervals

February 7th 2022

Barbara Pascal

Steniq, SigMA team meeting

Data: number of daily new infection counts



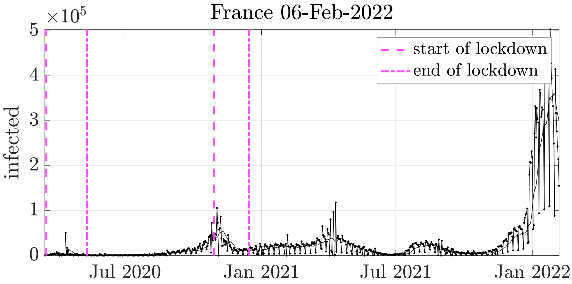
(partial) lockdown

Indicator: reproduction number $R0$

averaged number of people contaminated by one infected person

$R0 > 1$: the virus propagates,
$R0 < 1$: the epidemic slows down.

Taking, *e.g.*, lockdown, measures requires a **real-time**, daily, estimate R_t.



Reference: *Susceptible-Infected-Recovered* (SIR), among *compartmental models*

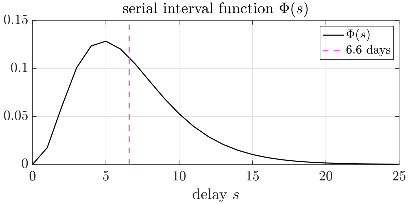
- refinement needed to get socially realistic model
- quadratic increase of the number of parameters
- Bayesian framework: heavy computational burden
- need consolidated and accurate datasets

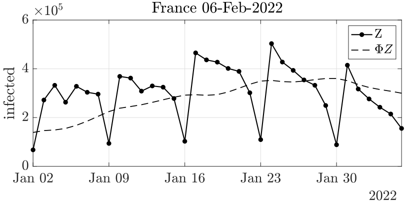
Poisson process accounting for random contamination

$$Z_t \sim \mathrm{Poiss}\left(p_t\right), \quad p_t = R_t \sum_{s=1}^{\tau_\Phi} \Phi(s) Z_{t-s}$$

$\Phi(s)$**: serial interval function** $\quad \tau_\Phi = 26$ days

random delay between onset of symptoms in primary and secondary cases





> modeled by a Gamma distribution with mean and variance of 6.6 and 3.5 days

Unknown parameters: $\boldsymbol{R} = (R_1, \ldots, R_T) \in \mathbb{R}_+^T$

Observed data: $\boldsymbol{Z} = (Z_1, \ldots, Z_T)$

Poisson distribution of parameter $p_t = R_t \sum_{s=1}^{\tau_\Phi} \Phi(s) Z_{t-s}$

$$\mathbb{P}(Z_t | \boldsymbol{Z}_{t-\tau_\Phi:t-1}, R_t) = \frac{p_t^{Z_t} \mathrm{e}^{-p_t}}{Z_t!}$$

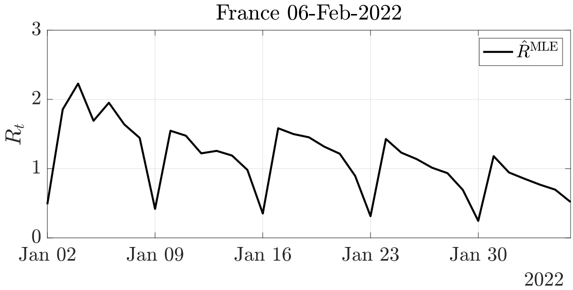
> negative log-likelihood

$$\begin{aligned} -\ln\left(\mathbb{P}(Z_t | \boldsymbol{Z}_{t-\tau_\Phi:t-1}, R_t)\right) &= p_t - Z_t \ln(p_t) + \ln(Z_t!) \\ &\underset{Z_t \gg 1}{\simeq} p_t - Z_t \ln(p_t) + Z_t \ln(Z_t) - Z_t \\ &\underset{\text{(def.)}}{=} \mathrm{d}_{\mathrm{KL}}(Z_t | p_t) \quad \text{Kullback-Leibler divergence} \end{aligned}$$

> maximizing the likelihood is equivalent to minimizing $-\ln \mathbb{P}$

$$\widehat{\boldsymbol{R}}^{\mathrm{MLE}} = \underset{\boldsymbol{R}\in\mathbb{R}_+^T}{\operatorname{argmin}} \sum_{t=1}^{T} \mathsf{d}_{\mathsf{KL}}\left(Z_t \,|\, R_t(\Phi Z)_t\right), \quad (\Phi Z)_t \triangleq \sum_{s=1}^{\tau_\Phi} \Phi(s) Z_{t-s}$$

Explicit solution: $\widehat{\boldsymbol{R}}_t^{\mathrm{MLE}} = Z_t/(\Phi Z)_t$

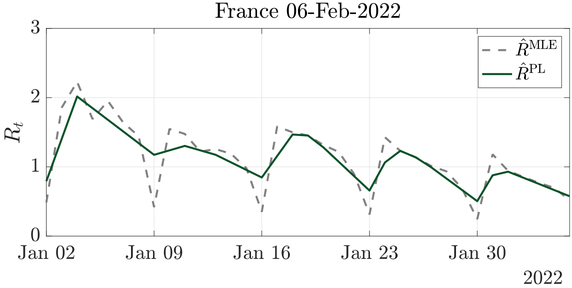


not realistic! pseudo-periodicity, irregularity, no local trend

> favor piecewise linear behavior

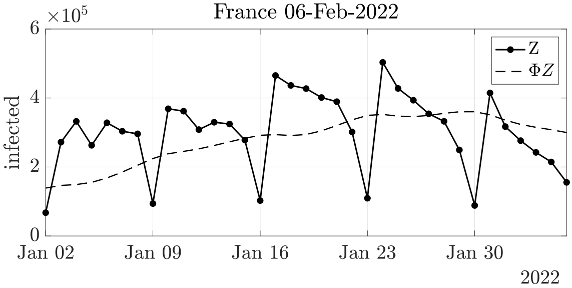
$$\widehat{\boldsymbol{R}}^{\mathsf{PL}} = \underset{\boldsymbol{R} \in \mathbb{R}_+^T}{\operatorname{argmin}} \sum_{t=1}^{T} \mathsf{d}_{\mathsf{KL}}\left(Z_t \,|R_t(\Phi Z)_t\right) + \lambda_{\mathsf{time}} \|\mathbf{D}_2 \boldsymbol{R}\|_1$$

$(\mathbf{D}_2 \boldsymbol{R}) = R_{t+1} - 2R_t + R_{t-1}$ second order discrete derivative



better, but still pseudo-oscillations

Managing low quality of the data



New infection counts $\boldsymbol{Z} = (Z_1, \ldots, Z_T)$ are corrupted by

- missing samples,
- non meaningful negative counts,
- retrospected cumulated counts spread over few days,
- pseudo-seasonality effects, with less counts on non working days, ...

$>$ parametric modeling out of reach

Nonstationary Poisson process with *outliers*

$$Z_t \sim \mathrm{Poiss}\left(R_t(\Phi Z)_t + O_t\right)$$

O_t: significant values, concentrated on specific days (Sundays, day-offs, ...)

Unknown parameters: $(\boldsymbol{R}, \boldsymbol{O}) = (R_1, \dots, R_T, O_1, \dots, O_T) \in \mathbb{R}_+^T \times \mathbb{R}^T$

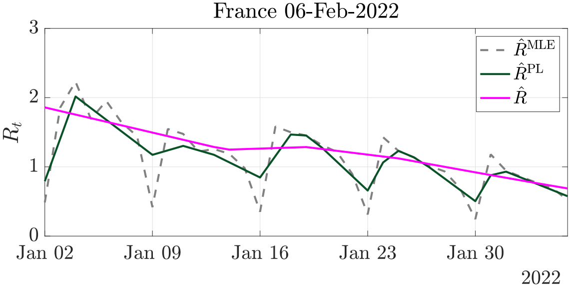
Extended penalized log-likelihood

$$\left(\widehat{\boldsymbol{R}}, \widehat{\boldsymbol{O}}\right) = \underset{(\boldsymbol{R},\boldsymbol{O})\in\mathbb{R}_+^T\times\mathbb{R}^T}{\mathrm{argmin}} \sum_{t=1}^{T} \mathrm{d_{KL}}\left(Z_t \,|R_t(\Phi Z)_t + O_t\right) + \lambda_{\mathrm{time}}\|\mathbf{D}_2\boldsymbol{R}\|_1 + \lambda_{\mathrm{O}}\|\boldsymbol{O}\|_1$$

$>$ favors piecewise linear reproduction number and sparse outliers

Nonsmooth convex optimization for the estimation of $\boldsymbol{R}$

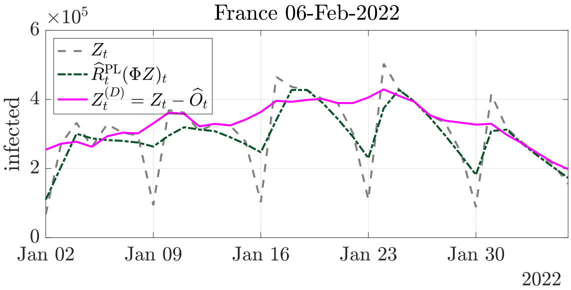
$$\left(\widehat{\boldsymbol{R}}, \widehat{\boldsymbol{O}}\right) = \underset{(\boldsymbol{R},\boldsymbol{O})\in\mathbb{R}_+^T\times\mathbb{R}^T}{\operatorname{argmin}} \ \mathrm{D}_{\mathrm{KL}}\left(\boldsymbol{Z}\,|\,\boldsymbol{R}\cdot\Phi\boldsymbol{Z}+\boldsymbol{O}\right) + \lambda_{\mathrm{time}}\|\mathbf{D}_2\boldsymbol{R}\|_1 + \lambda_{\mathrm{O}}\|\boldsymbol{O}\|_1$$



> no more pseudo-seasonality, local trends well captured, smooth behavior

As a byproduct

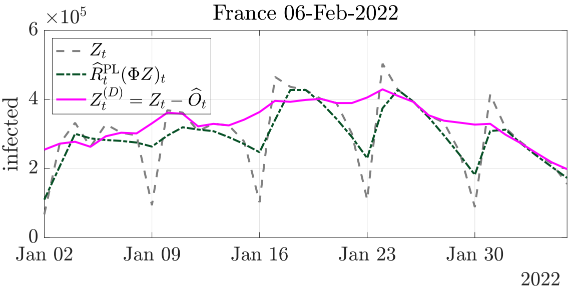
$$\widehat{\mathbf{Z}}^{(\mathrm{D})} = \mathbf{Z} - \widehat{\mathbf{O}}$$



Denoised infection counts

As a byproduct

$$\widehat{\boldsymbol{Z}}^{(\mathrm{D})} = \boldsymbol{Z} - \widehat{\boldsymbol{O}}$$



> level of confidence in the reproduction number $\widehat{R}_t$ and corrected count $\widehat{Z}_t^{(D)}$?

Idea: interpret the minimization problem

$$\underset{\boldsymbol{R},\boldsymbol{O}}{\text{minimize}} \quad \mathsf{D}_{\mathsf{KL}}\left(\boldsymbol{Z}\,|\,\boldsymbol{R}\cdot\Phi\boldsymbol{Z}+\boldsymbol{O}\right)+\lambda_{\mathsf{time}}\|\mathbf{D}_2\boldsymbol{R}\|_1+\lambda_{\mathrm{O}}\|\boldsymbol{O}\|_1$$

as a Maximum A Posteriori estimate of the parameter $\boldsymbol{\theta}=(\boldsymbol{R},\boldsymbol{O})$.

$>$ $\boldsymbol{\theta}$, $\boldsymbol{Z}$ realizations of random vectors whose distributions are to be specified

Purpose: reformulation of the estimation in a Bayesian framework

$$\widehat{\boldsymbol{\theta}}=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\,\pi(\boldsymbol{\theta}|\boldsymbol{Z})$$

π the density of the a posteriori distribution.

Quantiles of the distribution $\pi \Longrightarrow$ **credibility intervals**

A posteriori distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{Z}) \sim \prod_{t=1}^{T} \underbrace{\mathbb{P}(Z_t|\boldsymbol{Z}_{1:t-1}, R_t, O_t)}_{\text{likelihood}} \; \underbrace{\mathbb{P}(R_t, O_t|\boldsymbol{R}_{1:t-1}, \boldsymbol{O}_{1:t-1})}_{\text{prior}}$$

Likelihood: standard Kullback-Leibler based for Poisson model

$$\mathbb{P}(Z_t|\boldsymbol{Z}_{1:t-1}, R_t, O_t) \propto \exp(-\mathsf{d}_{\mathsf{KL}}\,(Z_t\,|R_t(\Phi Z)_t + O_t\,))$$

Prior: ***R*** and ***O*** supposed *mutually independent*

- Laplace auto-regressive AR(2) for R_t, $\forall t > 2$

$$\mathbb{P}(R_t|\boldsymbol{R}_{t-2:t-1}) = \frac{\lambda_{\text{time}}}{2}\exp\left(-\lambda_{\text{time}}|R_t - 2R_{t-1} + R_{t-2}|\right)$$

- independent Laplace for O_t, $\forall t > 0$

$$\mathbb{P}(O_t) = \frac{\lambda_{\text{O}}}{2}\exp\left(-\lambda_{\text{O}}|O_t|\right)$$

Markov chain Monte Carlo $\boldsymbol{\theta} = (\boldsymbol{R}, \boldsymbol{O})$

Purpose: sample from the distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{Z}) \sim \exp\left(-\sum_{t=1}^{T} \mathrm{d}_{\mathrm{KL}}\left(Z_t \,|R_t(\Phi Z)_t + O_t\,\right) - \lambda_{\mathrm{time}} \sum_{t=3}^{T} |R_t - 2R_{t-1} + R_{t-2}| - \lambda_{\mathrm{O}} \sum_{t=1}^{T} |O_t| \right)$$

To be compared with

$$\underset{\boldsymbol{R},\boldsymbol{O}}{\text{minimize}} \quad \mathrm{D}_{\mathrm{KL}}\left(\boldsymbol{Z}\,|\boldsymbol{R}\cdot\boldsymbol{\Phi}\boldsymbol{Z} + \boldsymbol{O}\,\right) + \lambda_{\mathrm{time}}\|\mathbf{D}_2\boldsymbol{R}\|_1 + \lambda_{\mathrm{O}}\|\boldsymbol{O}\|_1$$

MCMC principle: $\pi(\boldsymbol{\theta}|\boldsymbol{Z})$ is **intractable**, thus

generate a sequence $\{\boldsymbol{\theta}^n,\, n \geq 0\}$ with π as invariant measure

$\Longrightarrow$ for n sufficiently large, $\boldsymbol{\theta}^n \sim \pi(\boldsymbol{\theta}|\boldsymbol{Z})$ are representative samples

Metropolis Hastings with Gaussian proposals

Drift term: $\boldsymbol{\xi}^{n+1} \sim \mathcal{N}(\mathbf{0}_{2T}, \mathbf{C})$, and $\mathbf{C} \in \mathbb{R}^{2T\times 2T}$ covariance matrix

$$\boldsymbol{\theta}^{n+1/2} = \boldsymbol{\mu}(\boldsymbol{\theta}^n) + \boldsymbol{\xi}^{n+1}$$

Langevin: $\boldsymbol{\mu} = \mathbf{I} - \gamma\nabla(-\ln\pi)$ drives the chain toward *high probability* regions

for the Covid19 application $\ln\pi$ **not differentiable**

> we proposed a *proximal* Langevin strategy, with $\nabla \to \partial f$ *sub-differential*

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathrm{prox}_{\gamma\|[\mathbf{D}_2,\mathbf{I}]\cdot\|_1}\left(\boldsymbol{\theta} - \gamma\nabla \mathsf{D}_{\mathsf{KL}}(\boldsymbol{\theta})\right)$$

Acceptance-rejection Metropolis mechanism: q Gaussian kernel

> set $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^{n+1/2}$ with probability

$$\max\left(1, \frac{\pi(\boldsymbol{\theta}^{n+1/2})}{\pi(\boldsymbol{\theta}^n)}\frac{q(\boldsymbol{\theta}^{n+1/2},\boldsymbol{\theta}^n)}{q(\boldsymbol{\theta}^n,\boldsymbol{\theta}^{n+1/2})}\right)$$

5% credibility intervals for Covid19 reproduction number

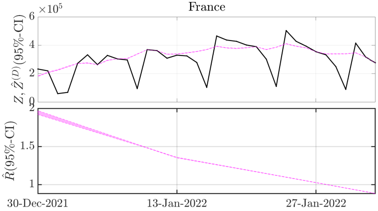
Data from *Johns Hopkins University* https://coronavirus.jhu.edu/

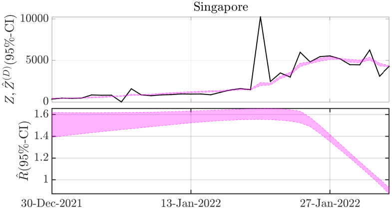
- from National Health Authorities
- 200+ countries
- since the outbreak of the pandemic
- updated on a daily basis

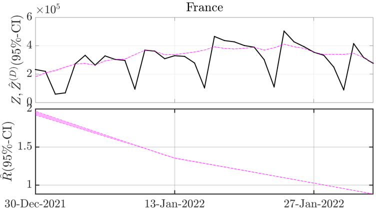
Credibility intervals for $\boldsymbol{R}$ and $\boldsymbol{Z}^{(D)}$

$>$ $\lambda_{\text{time}} = 3.5 \times \text{std}(\boldsymbol{Z})$, λ_{O} for **all** countries/time periods

$>$ 10 million points in the Markov chain $\{\boldsymbol{\theta}^n, n \geq 0\}$







- real-time credibility interval estimate of R_t
  > really informative for Health Authorities
- leverage connection between *variational* and *Bayesian* formulations
- using nonsmooth optimization tools smarter than random walk

> Estimate of R_t from convex optimization:
https://perso.ens-lyon.fr/patrice.abry/

> Credibility interval estimates of R_t:
https://perso.math.univ-toulouse.fr/gfort/