arXiv:2012.11542v1 [stat.ME] 21 Dec 2020

Uncertainty on the Reproduction Ratio in the SIR Model

S., ELLIOTT, [1] and C. GOURIEROUX [2]

December, 15, 2020
(Preliminary version)

The authors gratefully acknowledge financial support from the ACPR chair "Regulation and Systemic Risk", and the Agence Nationale de la Recherche (ANR-COVID) grant ANR-17-EUR-0010.

[1]University of Toronto.
[2]University of Toronto, Toulouse School of Economics, and CREST.

Uncertainty on the Reproduction Ratio in the SIR Model

Abstract

The aim of this paper is to understand the extreme variability on the estimated reproduction ratio R_0 observed in practice. For expository purpose we consider a discrete time stochastic version of the Susceptible-Infected-Recovered (SIR) model, and introduce different approximate maximum likelihood (AML) estimators of R_0. We carefully discuss the properties of these estimators and illustrate by a Monte-Carlo study the width of confidence intervals on R_0.

Keywords : SIR Model, Reproduction Ratio, COVID-19, Approximate Maximum Likelihood, EpiEstim, Final Size.

1 Introduction

In the standard epidemiological model, the reproduction ratio[3], that measures the expected number of persons that can be infected by a new infectious individual, plays a key role. Its value affects the explosive episode in the early phase of an epidemic, the size of the peak of infections as well as the epidemic final size [see e.g. Hethcote (2000), Ma, Earn (2006)]. It is followed daily, or weekly, as a simple indicator of either approaching, or receding from, the peak of an epidemic [see e.g. PHO (2020)], and is often used for sanitary policy. For instance, it may be used to fix the conditions of a partial lockdown, or to close the border to foreigners coming from other countries. "Alert levels are frequently based on this new totemic figure" [Adam (2020)].

The reproduction ratio is a forward looking notion whose definition involves a conditional expectation. This is a model based notion that depends on the information and dynamic model used to evaluate the expectation. This ex-ante notion has to be distinguished from the ex-post analogue counting retrospectively the number of persons infected by a given individual.[4] This model free ex-post notion cannot be computed without an accurate tracing process and is not immediately useful in a prediction perspective. [5]

In practice this ratio is approximated which generates a large uncertainty regarding its value [see Sanchez, Blauer (1997), the discussion and Table 2 in Obadia et al. (2012), webFigure 10 in Cori et al. (2013)]. For instance the first estimates for COVID-19 in Wuhan, China, were between 1.9 and 6.4 [see Li et al. (2020), Riou, Althaus (2020), Sanche et al. (2020), Wu et al. (2020)]. It is so important that "to calculate the official ratio of the United Kingdom, about ten groups present the results of their models to a dedicated government committee, which reaches consensus on a possible range. The individual models are not released" [Adam (2020)]. This uncertainty is due to the different interpretations and definitions of this ratio in models that underlie the estimation methods, to the estimation methods themselves [see Obadia et al. (2012), Cori et al. (2013) for standard estimation packages], and to the way they are adjusted to be applied in a rolling way [Wallinga,

[3]This terminology has been introduced in the epidemiological literature by McDonald (1952).

[4]The difference is similar to the difference between life expectancy and lifetime, or between volatility and realized volatility.

[5]See however White, Pagano (2008) for an application with well-documented influenza on two troop ships in late fall of 1918.

Teunis (2004), Cori et al. (2013)]. Moreover the estimates are generally provided without confidence bands, whereas these bands can be large, especially in the early phases of an epidemic, and, at the limit, these estimates can be non consistent of the reproductive ratio of interest even if applied to a large population.

The aim of this paper is to analyze precisely the uncertainty and lack of robustness of the estimated reproduction ratios. For expository purpose, we focus on the standard Susceptible-Infected-Recovered (SIR) initially introduced by Kermack, McKendrick (1927) and largely used in the literature. This model is used to define without ambiguity the reproductive ratio.

In Section 2, we introduce a discrete time stochastic version of the SIR model, discuss the possibility to aggregate the individual medical histories without loss of information. We also rigorously define the notions of reproduction ratios and how they evolve during the epidemic. Statistical inference of SIR model is the topic of Section 3. Since the binomial distributions that underlie the SIR model can be approximated by either Poisson, or Gaussian distributions depending on the structure of the population and on the transition probabilities, different approximate maximum likelihood estimators of the ratios are considered. They do not provide the same estimated values, nor do they have the same distributions when we perform the estimations in a Gaussian asymptotic framework. They can even be inconsistent in a Poisson asymptotic framework. This leads to Section 4 that contains a Monte-Carlo study to find confidence intervals valid for the different estimators and designs. The matrix variate definition of the reproductive number is introduced in Section 5 for a SIR model with heterogeneity. This leads to the introduction of within and between compartments reproductive ratios. Section 6 discusses an alternative definition of reproductive number, called instantaneous reproductive number introduced by Fraser (2007) which is based on a renewal equation for the evolution of infected individuals. This notion is the basis of a Bayesian estimation approach of the reproductive ratio, diffused by the EpiEstim R-package [Cori et al. (2013)]. The EpiEstim estimator is usually computed in a rolling way, but has to provide reasonable results in the standard SIR model. We discuss precisely why this approach considers a parameter of interest that does not correspond to the initial definition of the reproductive ratio and illustrate this feature by a Monte-Carlo study. We also discuss an alternative approach of the same type based on autoregressions of counts of new infected individual. Section 7 concludes. Appendix 1

provides a review of the main properties of the continuous time deterministic model and its Euler time discretization. Proofs of some estimation results and additional Monte-Carlo results are given in Appendices.

2 Model and Observations

We consider a discrete time stochastic version of the SIR model, with three states : S=1 susceptible, I=2, infected, infectious, R=3 recovered, immunized (or removed). We also discuss the aggregation of observations, and the notion of the reproductive ratio.

2.1 The model of individual histories

The model specifies the joint distribution of individual medical histories. For each individual $i, (i = 1, \ldots, n)$, and date $t, (t = 0, 1, \ldots, T)$, the variable Y_{it} provides the state $j = 1, 2, 3$ of individual i at date t.

Assumption A.1 : The individual histories $[Y_{i,t}, t = 0, 1, \ldots, T]$, $i = 1, \ldots, n$ are such that :

i) the variables $Y_{i,t}, i = 1, \ldots, n$ are independent conditional on past histories :

$$\underline{Y_{t-1}} = \left([Y_{i,t-1}, Y_{i,t-2}, \ldots, Y_{i,0}) , i = 1, \ldots, n \right].$$

ii) They have the same transition matrix :

$$P_t = (p_{jk}(t)),$$

where $p_{jk}(t)$ is the probability to migrate from state j at date $t-1$ to state k at date t, conditional on the past.

iii) The structure of the transition matrix is :

$$P_t = \begin{pmatrix} 1 - aN_2(t-1)/n & aN_2(t-1)/n & 0 \\ 0 & 1-c & c \\ 0 & 0 & 1 \end{pmatrix},$$

where $N_2(t-1)$ is the number of individuals in state $I = 2$ at date $t-1$ and a, c are parameters such that $a > 0$, $0 < c < 1$. The structure of the transition matrix characterizes the SIR model:

i) The last row of the matrix means that state $R = 3$ is an absorbing state implying an individual cannot be infected twice.

ii) The zero in the second row means that, after infection, the individual recovers, is immunized, and then cannot become at risk.

iii) The zero in the first row means that the individual cannot recover without being infected first.

iv) Parameter c is constant and represents the intensity of recovering.

v) Parameter a measures the contagion effect, and the intensity of being infected for an individual at risk is proportional to the proportion of infectious people

Under Assumption A.1, we deduce the joint distribution of $Y_{i,t}, i = 1, \ldots, n$, $t = 1, \ldots, T$ given the initial conditions $Y_{i,0}, i = 1, \ldots, n$. Nothing is said about the initial drawing of the $Y_{i,0}, i = 1, \ldots, n$. This conditional joint distribution is parameterized by two parameters a and c, that are assumed to be independent of both n, T.

2.2 Aggregated counts

Under Assumption A.1, it is possible to aggregate the individual data without loosing information on parameters a and c. We denote:

- $N_{jk}(t), j, k = 1, 2, 3$, the number of individuals transitioning from j to k between $t-1$ and t
- $N_j(t), j = 1, 2, 3$, the number of individuals in state j at date t
- $\hat{p}_{jk}(t) = N_{jk}(t)/N_j(t-1)$, the sample analogue of $p_{jk}(t)$
- $\hat{p}_j(t) = N_j(t)/n$, the proportion of individuals in state j at date t

It is known that the set of aggregates $\{N_{jk}(t), j, k = 1, 2, 3, t = 1, \ldots, T\}$ is a sufficient statistic for the analysis (see Appendix 2). Therefore the analysis can be based on these aggregates only. In the SIR framework, these aggregates are related as shown in Table 1.

Table 1 : The Aggregate Counts

	1	2	3	Total
1	$N_{11}(t)$	$N_{12}(t)$	0	$N_1(t-1)$
2	0	$N_{22}(t)$	$N_{23}(t)$	$N_2(t-1)$
3	0	0	$N_{33}(t)$	$N_3(t-1)$
Total	$N_1(t)$	$N_2(t)$	$N_3(t)$	n

In particular, the following relationships provide the cross-sectional counts in terms of the transition counts:

$$\begin{array}{rcl} N_1(t) & = & N_{11}(t), \\ N_2(t) & = & N_{12}(t) + N_{22}(t), \\ N_3(t) & = & N_{23}(t) + N_{33}(t), \\ N_1(t-1) & = & N_{11}(t) + N_{12}(t), \\ N_2(t-1) & = & N_{22}(t) + N_{23}(t), \\ N_3(t-1) & = & N_{33}(t). \end{array}$$

For the SIR model, these equations can be solved to get the transition counts in terms of marginal counts. We have :

$$\begin{array}{rcl} N_{11}(t) & = & N_1(t), \\ N_{12}(t) & = & N_1(t-1) - N_1(t) = -\Delta N_1(t), \\ N_{22}(t) & = & N_2(t) + \Delta N_1(t), \\ N_{23}(t) & = & N_2(t-1) - N_2(t) - \Delta N_1(t) = -\Delta N_1(t) - \Delta N_2(t) = \Delta N_3(t), \\ N_{33}(t) & = & N_3(t-1), \end{array}$$

where $\Delta = Id - L$ is the difference operator.

We deduce the following result :

Proposition 1 : For the SIR model of Assumption A.1, the sequence $N(t) = [N_1(t), N_2(t), N_3(t)]', t = 0, \ldots, T$, is also a sufficient statistic. Moreover the process $[N(t)]$ is an homogeneous Markov process.

Thus we have the same information in the transition counts and in the cross-sectional counts. This property is not satisfied in other epidemiological models.

2.3 Reproductive ratio

Other summaries of the development of a disease have been introduced in the epidemiological literature. An important concept is the reproductive (or reproduction) ratio (number). It is defined by computing the expected number of individuals at risk that a new infected individual will infect during his/her infectious period. In our framework with constant recovery intensity the length of the infection/infectious period is stochastic, with a geometric distribution with elementary probability : $P(X = x) = c(1-c)^{x-1}$, survivor function : $P[X \geq x] = (1-c)^{x-1}, x = 1, 2, \ldots$, and expectation : $EX = 1/c$.

We deduce the expected number of individuals infected by this individual newly infected at date t as [Farrington, Whitaker (2003)] :

$$\begin{aligned} R^*_{0,t} &= \frac{a}{n} \sum_{x=1}^{\infty} \left\{ E_t[N_1(t+x-1)](1-c)^{x-1} \right\} \\ &= \frac{a}{n} \sum_{x=0}^{\infty} \left\{ E_t[N_1(t+x)](1-c)^{x} \right] . \end{aligned} \tag{2.1}$$

This expectation depends on the transmission rate a, of the survival function of the infectious period, but also of the expected proportion of people at risk. For instance, if the population at risk disappears : $N_1(t) \simeq 0$, then $R_{0,t} = 0$ too. To adjust for the size of the population at risk and the medical notion of transmission, it is usually proposed to consider also :

$$R_{0,t} = \frac{a}{N_1(t)} \sum_{x=0}^{\infty} [E_t[N_1(t+x)](1-c)^{x}]. \tag{2.2}$$

These quantities are called basic reproductive and effective reproductive numbers for $R_{0,t}$ and $R^*_{0,t}$, respectively. Under Assumption A.1, the predictions $E_t N_1(t+x) = g[a, c, N_1(t), N_2(t), N_3(t)]$ by the homogeneous Markov property, where g is a nonlinear function independent of time. Therefore $R_{0,t}, R^*_{0,t}$ also depend on time through the marginal counts at time t.

In the literature, this time dependence is often disregarded by focusing at the very early phase (outbreak) of the epidemics. [see e.g. Hethcote (2000)].

At this date $t = 0$, it is assumed that :

i) $N_1(0) = n - \varepsilon, N_2(0) = \varepsilon, N_3(0) = 0$, where $\varepsilon, \varepsilon > 0$, is very small.

This ε corresponds to the first infected individuals, or the first cluster. Without this initial infection, the disease cannot appear in the population. In other words, the SIR model assumes a "closed economy", except at the initial date.

ii) During the following days $N_1(t) = n - \varepsilon(t)$, where $\varepsilon(t)$ is also small. An approximate formula for reproductive ratios is :

$$R_{0,0} = R^*_{0,0} \simeq a \sum_{x=0}^{\infty} (1-c)^x = \frac{a}{c}, \tag{2.3}$$

that is, the transmission rate times the expected length of the infection episode. This common value is called the initial reproductive ratio. However, during the epidemic, these measures can differ significantly.

2.4 Simulation

The conditional distributions of the count variables are easily deduced from Assumption A.1.

Proposition 2 : Under Assumption A.1,

i) $N_{12}(t)$ and $N_{23}(t)$ are independent given the past. $N_{12}(t)$ follows the binomial distribution $\mathcal{B}[N_1(t-1), a\dfrac{N_2(t-1)}{n}]$. $N_{23}(t)$ follows the binomial distribution $\mathcal{B}[N_2(t-1), c]$.

ii) The process $[N_1(t), N_2(t)]$ is a Markov process. Its conditional distribution is obtained from the distribution of $[N_{12}(t), N_{23}(t)]$ by the change of variable :

$$\begin{cases} N_1(t) &= N_1(t-1) - N_{12}(t), \\ \\ N_2(t) &= N_2(t-1) + N_{12}(t) - N_{23}(t). \end{cases}$$

These results can be used to simulate the aggregate counts for given parameter values a, c and given starting counts $N_1(0), N_2(0), N_3(0)$, along the following scheme, where $\overset{s}{\to}$ is a drawing in the binomial distributions

of Proposition 1 i), and $\stackrel{d}{\to}$ the application of the deterministic relation in Proposition 1 ii).

Table 2 : Simulation Scheme

$$\begin{array}{ccccc} [N_1(0), N_2(0), N_3(0)] & \stackrel{d}{\longrightarrow} & [N_1(1), N_2(1), N_3(1)] & \stackrel{d}{\to} & \\ \downarrow s & \nearrow d & \downarrow s & & \\ [N_{12}(1), N_{23}(1)] & & & & \end{array}$$

For simulations and by analogy with COVID-19, the parameter values can be fixed as :

$c = 0.07$, that corresponds to an expected infection[6] period of approximately 14 days, $R_{0,0} = a/c$ between 0.5 and 1.5, that means a between 0.095 and 0.105.

The initial structure for a population corresponding to the city of Toronto, say, can be $n = 3000000$ with a first cluster of $N_2(0) = 50$ [with $N_3(0) = 0$].

Thus for $a \simeq 0.1$, at date $t = 0, p_{12}(0) \simeq \dfrac{0.1 \ \ 50}{3000000} = \dfrac{1}{600000}$.

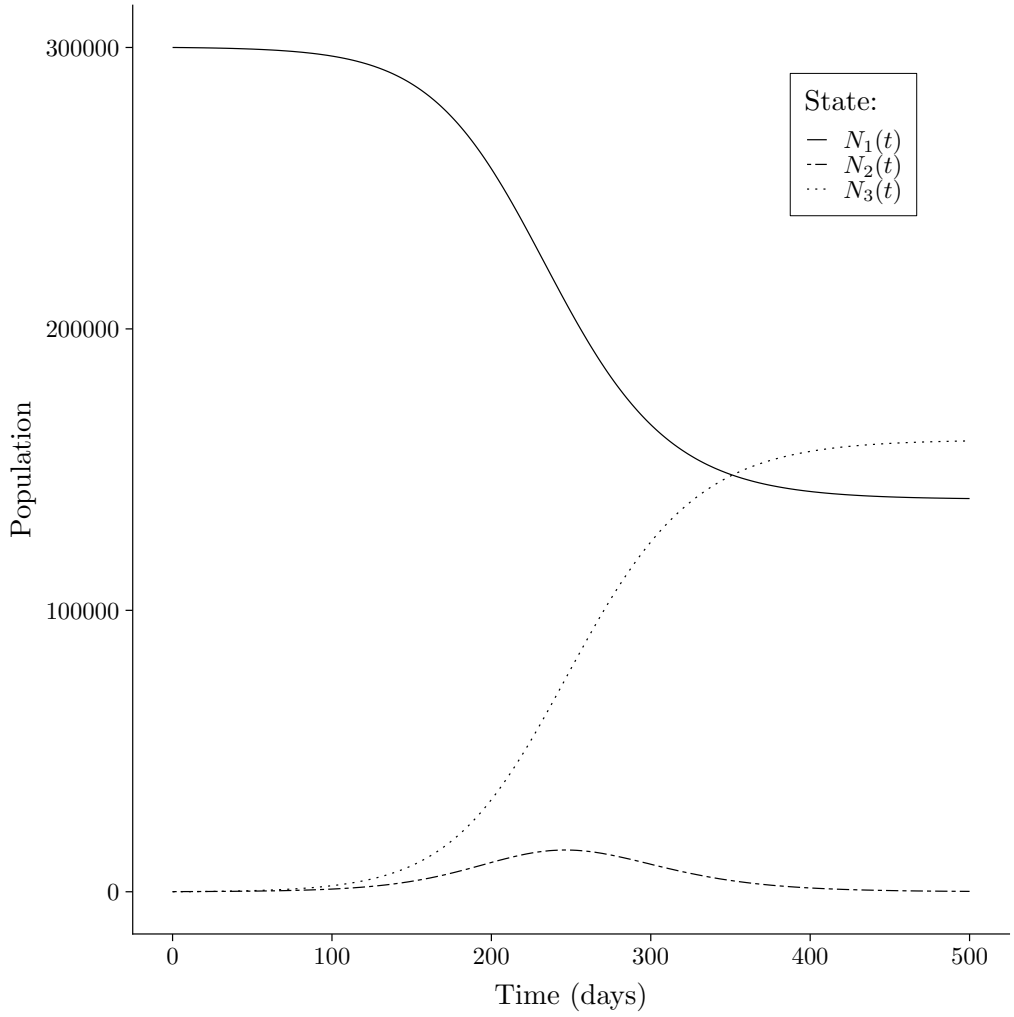
Thus $p_{23}(t)$ is small and $p_{12}(0)$ very small as well as $p_{12}(t)$ at the beginning of the epidemic.

A simulated path is given in Figure 1. We observe the standard patterns :

- A decreasing pattern for the size of the population at risk.
- An increasing pattern for the number of immunized people.
- The peak of the epidemic for the number of infected people, arising around one year in this simulation. The figure is given for a rather large number of days to highlight the asymptotic behaviour. For this SIR model, there is herd immunity [Allen (1994)], and the immunity ratio is around 55%.

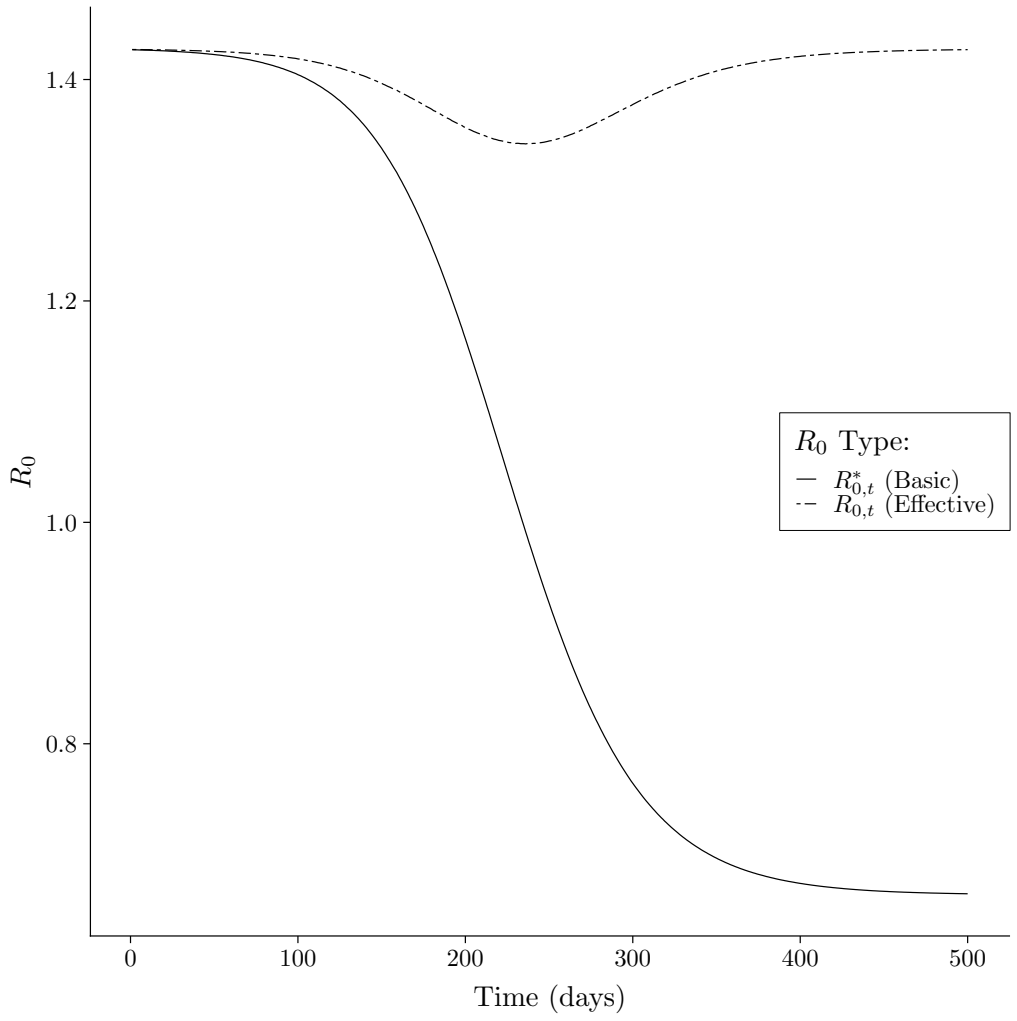
[6] In the SIR model the infection and infectious periods are assumed the same. This is not the case for COVID-19.

Figure 1: Simulated Path



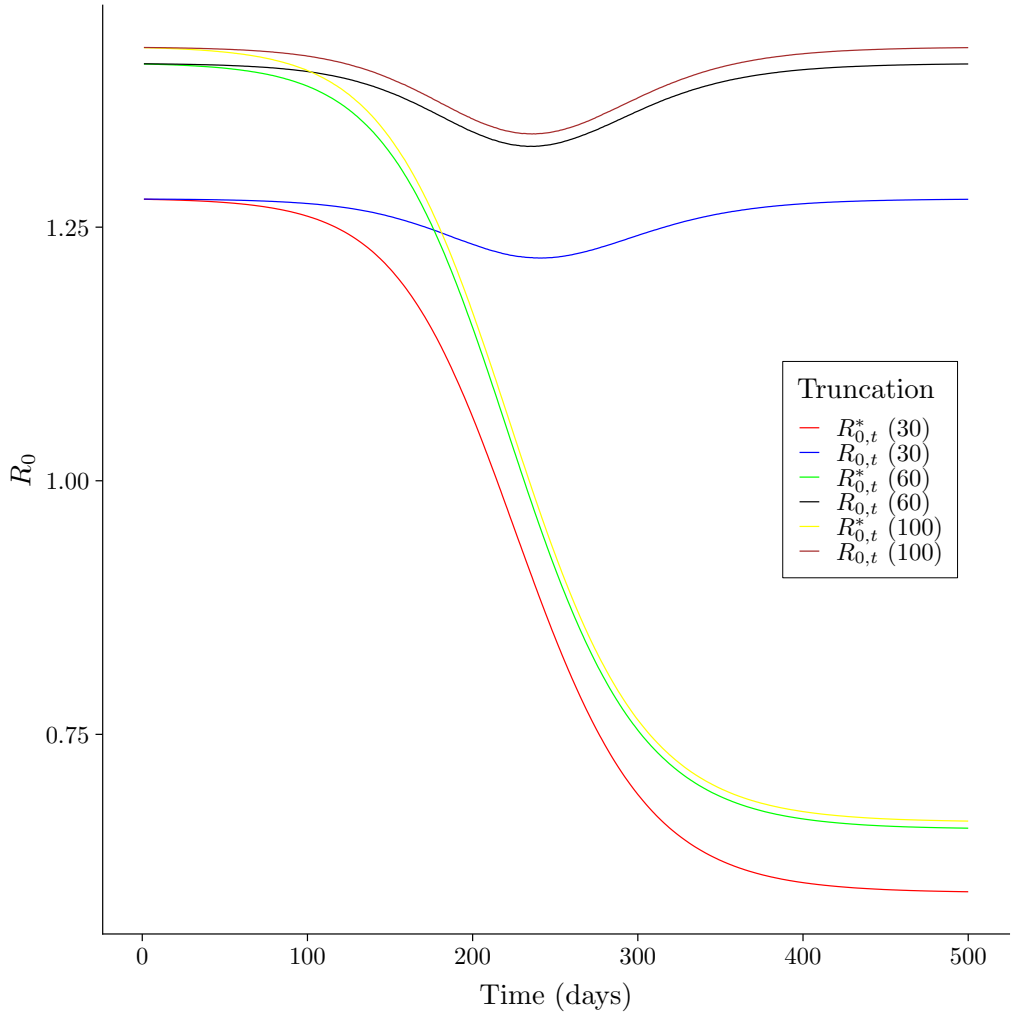
At each date t, we can simulate and average of several future paths $N_1(t+x), x = 1, \ldots, 30$, and compute the basic reproductive and effective reproductive numbers at t. These paths are reported in Figure 2 with a number of replications equal to $S = 100$. We observe that even the basic reproduction number, that is the number adjusted by the size of the population at risk, is not constant during the epidemics in the time discretized version of the SIR.

Figure 2 : Evolution of Basic and Effective Reproductive Ratios



We also observe that the final level of the effective reproduction number is equal to its starting value. Indeed for large t, the size of the population at risk coincides with the final size and then $R_0(\infty) = a/c$ too. The evolutions of Figure 2 are obtained with a length of 100 days for the future path of $N_1(t)$. In practice the sum can be truncated and such a truncation can have an impact on the evaluation of R_0. Figure 3 provides the evolutions of reproduction ratios computed with 30, 60, and 100 days, respectively.

Figure 3: Evolution of Reproductive Ratio Under Truncation



3 Estimation

3.1 Challenges

The estimation of a SIR model, and more generally of any epidemiological model, is a bit challenging for three main reasons.

i) The SIR model is a nonlinear dynamic model with chaotic properties [see e.g. Harko, Lobo, Mack (2014)]. This implies that small changes in parameter values a, c, in particular the estimation errors, can have a strong impact on the evolution of the process in the medium and long runs. It is known [see Allen (1994)] that the deterministic discrete time version of the SIR model satisfies herd immunity. However, in our stochastic framework the level of herd immunity, as well as the time at which it is reached, are very sensitive to the values of a, c, and to the initial conditions.

ii) The evolution of the disease is non-stationary, as seen in Figure 1. If $R_{0,0} > 1$, the proportion of infected individuals increases up to a peak, then decreases towards an asymptotic stationary state. This non-stationarity makes it difficult to analyze the properties of the estimators as functions of the number of observation dates T. Moreover, T is usually small, between 20-60 days, at the beginning of the epidemic.

iii) In contrast to the previous point, the cross-sectional dimension n is very large and we expect an asymptotic theory when n tends to infinity, T being fixed. However, Proposition 2 shows the key role of the binomial distributions $\mathcal{B}[N_1(t-1), p_{12}(t)]$ and $\mathcal{B}[N_2(t-1), c], t = 1, \ldots, T$. For an asymptotic analysis, what matters is not n, but rather the marginal counts $N_1(t-1), N_2(t-1)$. Whereas the susceptible population is often very large, at least at the beginning of the disease, the number of infected people is much smaller.

However, for large $N_1(t-1), N_2(t-1)$, we may apply the standard asymptotic results for a binomial distribution. That is, the possibility to approximate it by either a Poisson, or a Gaussian distribution. Thus, the approximation of $\mathcal{B}(N, p)$, say, is either $\mathcal{P}(Np)$, if $N \to \infty, p \to 0$, such that $Np \to \lambda > 0$, or $N[Np, Np(1-p)]$, if $N \to \infty$, p being fixed. In our framework both $p_{23}(t) = c$ and $p_{12}(t)$ are small. The choice between the approximations depend on the magnitudes of $N_1(t-1)p_{12}(t), N_2(t-1)p_{23}(t)$,

$t = 1, \ldots, T,$. That is, the numbers of new infected and new recovered, respectively.

Loosely speaking, if they are smaller than 45-50, say, the Poisson approximation can be used, the Gaussian approximation, otherwise. But at the beginning of the epidemic and also at the end of the epidemic $N_{12}(t)$ and $N_{23}(t)$ are rather small, while being larger in the peak of the epidemic. Therefore, the approximation will depend on the observation date. They also depend on the size n of the population of interest. For instance, this size is smaller if we want to consider a subpopulation of Toronto, say, males older than 75.[7]

3.2 Mechanistic model

A major part of the literature is based on a deterministic dynamic model, that assumes implicitly the possibility to closely approximate the theoretical transition probabilities by their frequency counterparts. That is, to use the Gaussian approximation.

More precisely, under Assumption A.1, we have :

$$E_{t-1}\hat{p}(t) = P[\hat{p}_2(t-1)]'\hat{p}(t-1). \tag{3.1}$$

Therefore if $\hat{p}(t) \sim p(t)$, we get the following deterministic dynamic model for the $p(t)'s$:

$$p(t) = P[p_2(t-1)]'p(t-1). \tag{3.2}$$

This is often called the mechanistic model [see Breto et al. (2009) and Appendix 1 for its link with the continuous time SIR model].

3.3 (Approximate) Maximum Likelihood Estimator

In our framework, the log-likelihood function $L(a, c)$ can be decomposed as a sum $L(a, c) = L_1(a) + L_2(c)$. This allows us to estimate separately a and c by focusing on the first and second rows of the (observed) transition matrix, respectively (see Appendix 2). Different log-likelihood functions can

[7] See also Zhang et al. (2020) for an analysis restricted to the analysis of the epidemic on the Diamond Princess cruise ship.

be considered such as, the true one based on the binomial distributions, or approximate ones based on either Poisson, or Gaussian approximations.

3.3.1 Binomial log-likelihood

We have :

$$L_1(a) = \sum_{t=1}^{T} \{N_{11}(t) \log[1 - a\hat{p}_2(t-1)] + N_{12}(t) \log[a\hat{p}_2(t-1)]\}, \tag{3.3}$$

$$L_2(c) = \sum_{t=1}^{T} \{N_{22}(t) \log(1-c) + N_{23}(t) \log c\}. \tag{3.4}$$

The ML estimator of a is the solution of the first-order conditions :

$$-\sum_{t=1}^{T} \left[\frac{N_{11}(t)\hat{p}_2(t-1)}{1 - \hat{a}\hat{p}_2(t-1)} \right] + \frac{1}{\hat{a}} \sum_{t=1}^{T} N_{12}(t) = 0, \tag{3.5}$$

and has no closed form expression.

The ML estimator of c is :

$$\begin{aligned} \hat{c} &= \sum_{t=1}^{T} N_{23}(t) / \sum_{t=1}^{T} N_2(t-1) \\ &= \sum_{t=1}^{T} \left\{ \frac{N_2(t-1)}{\sum_{t=1}^{T} N_2(t-1)} \hat{p}_{23}(t) \right\}. \end{aligned} \tag{3.6}$$

This is a weighted combination of the dated transition frequencies.

3.3.2 Poisson approximate log-likelihood

We have :

$$L_1^P(a) \propto \sum_{t=1}^{T}\{N_{12}(t)\log[aN_1(t-1)\hat{p}_2(t-1)] - aN_1(t-1)\hat{p}_2(t-1)]\}, \quad (3.7)$$

$$L_2^P(c) \propto \sum_{t=1}^{T}\{N_{23}(t)\log[N_2(t-1)c] - N_2(t-1)c]\}. \quad (3.8)$$

We get Poisson approximate maximum likelihood (AML) estimators with closed form expressions :

$$\hat{a}_P = n\sum_{t=1}^{T} N_{12}(t)/\sum_{t=1}^{T}[N_1(t-1)N_2(t-1)], \quad (3.9)$$

$$\hat{c}_P = \sum_{t=1}^{T} N_{23}(t)/\sum_{t=1}^{T} N_2(t-1) = \hat{c}. \quad (3.10)$$

The first formula shows that $\hat{a}_P$ is a weighted average of the dated estimated transition coefficients : $\hat{a}_t = N_{12}(t)/[N_1(t-1)\hat{p}_2(t-1)]$, with weights proportional to $N_1(t-1)\hat{p}_2(t-1)$.

We deduce an analytical formula for the corresponding estimator of the initial reproductive number :

$$\hat{R}_{0,P} = \frac{n\sum_{t=1}^{T} N_{12}(t)\sum_{t=1}^{T} N_2(t-1)}{\sum_{t=1}^{T}[N_1(t-1)N_2(t-1)]\sum_{t=1}^{T} N_{12}(t)}. \quad (3.11)$$

This formula can be used if $\sum_{t=1}^{T} N_{23}(t)$ is non zero, that is if recovery has been observed.

3.3.3 Gaussian approximate log-likelihood

We have :

$$L_1^G(a) \propto -\frac{1}{2}\sum_{t=1}^{T}\log(a\hat{p}_2(t-1)[1-a\hat{p}_2(t-1)]) - \frac{1}{2}\sum_{t=1}^{T}N_1(t-1)\frac{[\hat{p}_{12}(t)-a\hat{p}_2(t-1)]^2}{a\hat{p}_2(t-1)[1-a\hat{p}_2(t-1)]}, \tag{3.12}$$

$$L_2^G(c) \propto -\frac{T}{2}\log[c(1-c)] - \frac{1}{2}\sum_{t=1}^{T}N_2(t-1)\frac{[\hat{p}_{23}(t)-c]^2}{c(1-c)}. \tag{3.13}$$

3.3.4 Unfeasible Gaussian approximate log-likelihood

The approximate log-likelihood is obtained by replacing the variance $a\hat{p}_2(t-1)[1-a\hat{p}_2(t-1)]$ by the estimate $\hat{p}_{12}(t)[1-\hat{p}_{12}(t)]$.[8] We get :

$$L_1^{UG}(a) = -\frac{1}{2}\sum_{t=1}^{T}\left\{N_1(t-1)\frac{(\hat{p}_{12}(t)-a\hat{p}_2(t-1)]^2}{\hat{p}_{12}(t)(1-\hat{p}_{12}(t)]}\right\}. \tag{3.14}$$

We get a closed form expression for $\hat{a}_{UG}$ that corresponds to an unfeasible Generalized Least Squares (GLS) estimator of a :

$$\hat{a}_{UG} = \sum_{t=1}^{T}(N_1(t-1)\hat{p}_2(t-1)/[1-\hat{p}_{12}(t))/\sum_{t=1}^{T}\left[\frac{N_1(t-1)\hat{p}_2(t-1)^2}{\hat{p}_{12}(t)[1-\hat{p}_{12}(t)]}\right]. \tag{3.15}$$

3.3.5 Poisson/Gaussian approximate log-likelihood

When n is large, p small and np large, the Poisson distribution $\mathcal{P}(np)$ can be approximated by a Gaussian distributions $N(np, np)$. Thus compared to the approximation in 3.3.3, the term in p^2 in the variance is disregarded. We have :

$$L_1^{PG}(a) \propto -\frac{1}{2}\sum_{t=1}^{T}\log[a\hat{p}_2(t-1)] - \frac{1}{2}\sum_{t=1}^{T}\left\{N_1(t-1)\frac{[\hat{p}_{12}(t)-a\hat{p}_2(t-1)]^2}{ap_2(\hat{t-1})}\right\} \tag{3.16}$$

$$L_2^{PG}(c) \propto -\frac{1}{2}T\log c - \frac{1}{2}\sum_{t=1}^{T}\left\{N_2(t-1)\frac{[\hat{p}_{23}(t)-c]^2}{c}\right\} \tag{3.17}$$

[8] This can be inconsistent when n tends to infinity.

Then the AML estimates are positive solutions of polynomial equations of degree 2, that are :

$$\frac{1}{T}\sum_{t=1}^{T}\{N_1(t-1)\hat{p}_2(t-1)\}\,a^2 + a - \frac{1}{T}\sum_{t=1}^{T}\{N_1(t-1)\hat{p}_{12}(t)\} = 0,$$

and

$$\frac{1}{T}\sum_{t=1}^{T}N_2(t-1)nc^2 + c - \frac{1}{T}\sum_{t=1}^{T}\{N_2(t-1)\hat{p}_{23}(t)\} = 0, \text{ respectively.}$$

To summarize, we get as many AML estimators of a, c and of the initial reproduction number $R_{0,0} = a/c$ as (approximated) log-likelihoods. This can also explain the different approximations of $R_{0,0}$ published even when applied to the same series of aggregate counts.

3.4 Properties of the AML estimators

The properties of the AML estimators can be derived by Monte-Carlo as shown in Section 4. Their asymptotic properties depend on either the Poisson, or Gaussian asymptotics, depending on which is the most appropriate, and on the selected estimators. For instance, we may have chosen a Poisson AML estimator when the Gaussian asymptotic conditions were satisfied. In this case, whereas $B(N,p)$ is well approximated by $N[Np, Np(1-p)]$, it has been replaced by $\mathcal{P}(Np)$, which is close to $N(Np, Np)$. Therefore we have not used the right Gaussian approximation and have neglected the term in p^2.

For illustration we consider below two cases :

i) The behaviour of the Poisson AML estimator $\hat{a}_P$, when Poisson asymptotics are valid.

ii) The behaviour of the binomial ML estimator $\hat{a}$, when Gaussian asymptotics are valid.

3.4.1 Poisson AML and Poisson asymptotics

Let us consider the case $T = 1$, that is two observations of the aggregates. The main results below will be valid for any finite T. Then we have :

$$\begin{aligned} \hat{a}_P &= nN_{12}(1)/N_1(0)N_2(0), \\ \hat{c}_P &= N_{23}(1)/N_2(0), \\ \hat{R}_{0,P} &= \hat{a}_P/\hat{c}_P = \frac{N_{12}(1)}{N_{23}(1)}\frac{n}{N_1(0)}. \end{aligned} \tag{3.18}$$

Conditional on $[N_1(0), N_2(0)]$, the estimators $\hat{a}_P$ and $\hat{c}_P$ are independent such that $\frac{N_1(0)N_2(0)}{n}\hat{a}_P \sim \mathcal{P}[a\frac{N_1(0)N_2(0)}{n}]$, $N_2(0)\hat{c}_P \sim \mathcal{P}[cN_2(0)]$.

We deduce that : $E_0\hat{a}_P = a, E_0\hat{c}_P = c$, that shows that the Poisson AML estimators are unbiased for $T = 1$. Their variances are :

$$V_0\hat{a}_P = \frac{an}{N_1(0)N_2(0)}, V_0\hat{c}_P = \frac{c}{N_2(0)}. \tag{3.19}$$

In practice $N_1(0) \simeq n$, and $N_2(0)$ is rather small (< 30 or 40, say) for Poisson asymptotics to be valid. Therefore both $V_0(\hat{a}_P)$ and $V_0(\hat{c}_P)$ are not small, even for large n and we cannot expect the consistency of $\hat{a}_P, \hat{c}_P$ for n large under Poisson asymptotics.

Moreover, at the very beginning of an epidemic, the infected individuals have not yet recovered meaning $N_{23}(1) = 0$. We deduce that : $\hat{R}_{0P} = \hat{a}_P/\hat{c}_P = \hat{a}_P/0 = \infty$. This illustrates the lack of accuracy on the basic reproductive ratio during the initial phase of the outbreak.

Remark 1 : The unbiasedness property is specific to the case $T = 1$. If $T = 2$, we have :

$$\hat{a}_P = \frac{n[N_{12}(1) + N_{12}(2)]}{N_1(0)N_2(0) + N_1(1)N_2(1)}$$

We deduce its expectation at date 1 :

$$E_1(\hat{a}_P) = n\frac{N_{12}(1) + aN_1(1)N_2(1)}{N_1(0)N_2(0) + N_2(1)N_2(1)},$$

and by iterated expectation,

$$E_0(\hat{a}_P) = nE_0\left[\frac{N_{12}(1) + aN_1(1)N_2(1)}{N_1(0)N_2(0) + N_1(1)N_2(1)}\right],$$

which is the expectation of a complicated nonlinear function of counts $N_1(1), N_2(1), N_{12}(1)$.

3.4.2 Binomial ML and Gaussian asymptotics

This is the standard asymptotic theory when the Law of Large Numbers and the Central Limit Theorem are applicable. The sample frequencies tend to their theoretical counterparts : $\hat{p}_{jk}(t) \to p_{jk}(t)$, $\hat{p}_j(t) \to p_j(t), j, k = 1, 2, 3$, when n tends to infinity. The ML estimators tend to the true parameter values : $\hat{a} \to a, \hat{c} \to c, \hat{R}_0 = \hat{a}/\hat{c} \to a/c$, at speed $1/\sqrt{n}$. $\hat{a}, \hat{c}$ are asymptotically independent, asymptotically normal and their variances are consistently estimated by :

$$\hat{V}(\hat{a}) = \left\{\sum_{t=1}^{T}\left(\frac{N_{11}(t)\hat{p}_2(t-1)}{[1-\hat{a}\hat{p}_2(t-1)]^2}\right) + \frac{1}{\hat{a}^2}\sum_{t=1}^{T}N_{12}(t)\right\}^{-1}, \quad (3.20)$$

$$\hat{V}(\hat{c}) = \frac{\hat{c}(1-\hat{c})}{\sum_{t=1}^{T}N_2(t-1)}. \quad (3.21)$$

Remark 2 : The Gaussian asymptotics can also be applied to other AML estimators as the Poisson AML. In this case the Poisson AML estimator of a is still consistent, asymptotically normal. However, since the approximate log-likelihood is misspecified, its asymptotic variance is obtained by a sandwich formula, that involves the two expressions of the information matrix [see Huber (1967)].

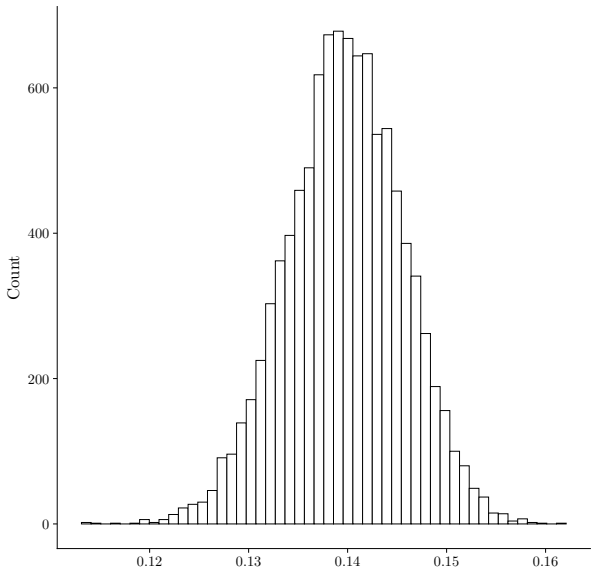
4 Monte-Carlo Study

Even when the Gaussian asymptotics can be used, we do not know if they are accurate for determining the confidence intervals on the different parameters a, c, R_0. In this section, we perform a Monte-Carlo analysis for some of the estimators introduced in Section 3. We fix the design as follows :

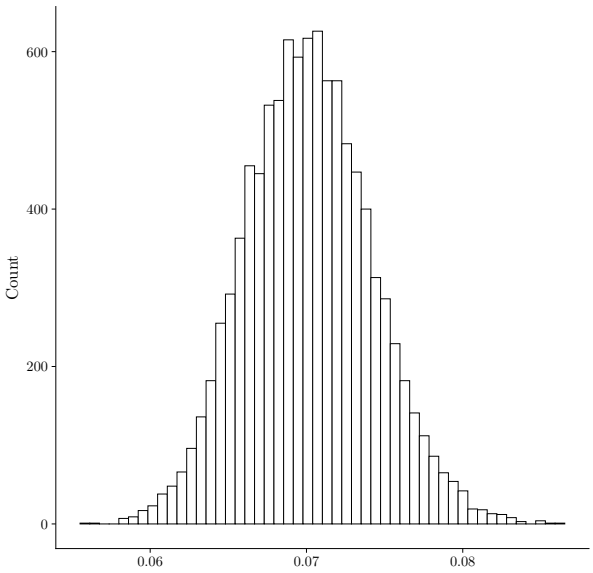
$$N_1(0) = 3000000, N_2(0) = 100, 1000, T = 20, c = 0.07, R_0 = 2$$

This corresponds to estimators computed on the period $[0, T]$. Note that the process of marginal counts is Markov. Therefore it also applies to a rolling estimator computed on $(t, t+T)$, say, where the marginal counts at t are the counts fixed for $N_1(0), N_2(0)$. This explains why we allow for a large value of $N_2(0)$ in the design. Figures 4 and 5 correspond to the parameters estimated by Approximated Poisson likelihood with $N_2(0) = 100, 1000$, respectively. They provide the finite sample distributions of parameters $a, c, R_0 = a/c$.

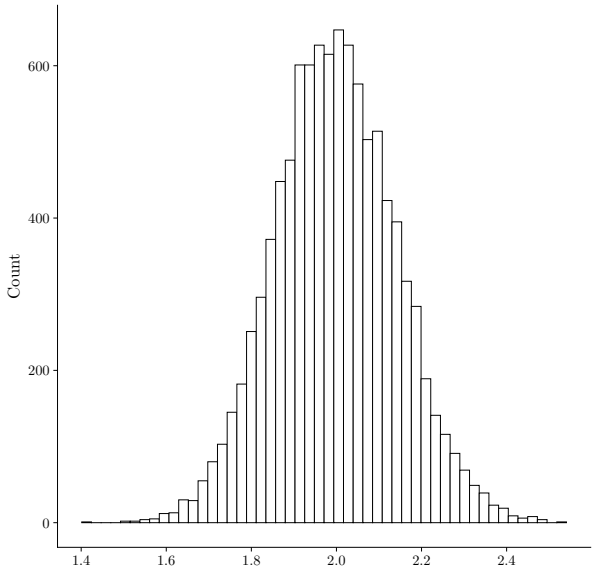
Figure 4: Distributions of Approximate Poisson Estimators $N_2(0) = 100$.



$\hat{a}$

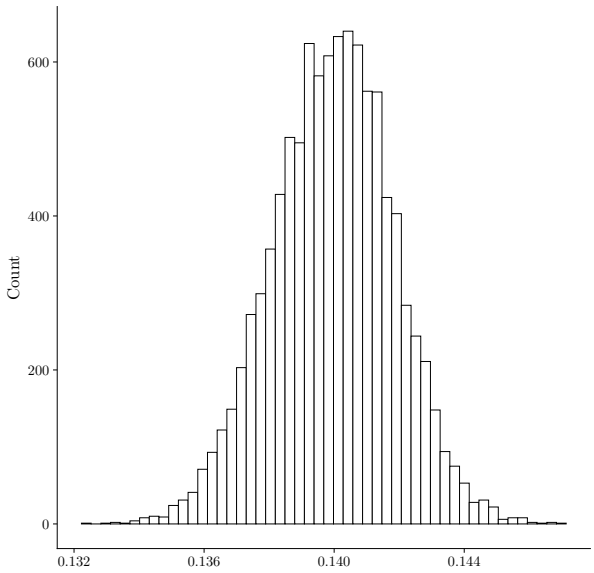


$\hat{c}$

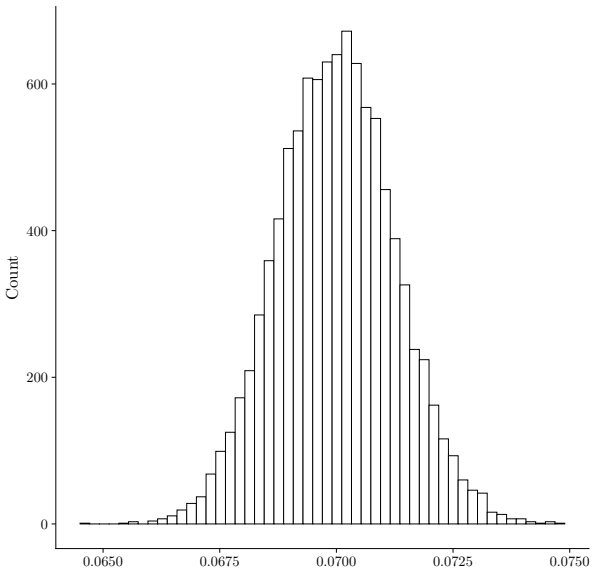


$\hat{R}_0$

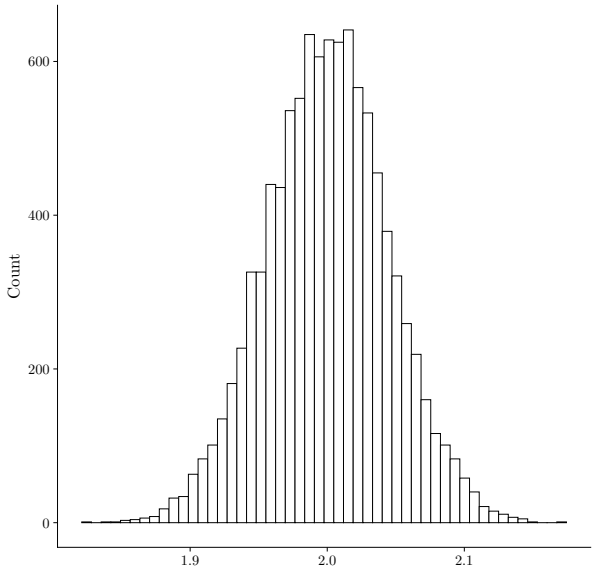
Figure 5: Distributions of Approximate Poisson Estimators $N_2(0) = 1000$.



$\hat{a}$



$\hat{c}$



$\hat{R}_0$

$N_2(0)$	T	a	c	R_0	mean($\hat{a}$)	var($\hat{a}$)	median($\hat{a}$)	$\rho(\hat{a}, \hat{c})$
5	20	0.035	0.07	0.5	0.03115	0.00045922	0.03044	-0.112
5	20	0.140	0.07	2.0	0.13119	0.00099481	0.13539	-0.246
5	40	0.105	0.07	1.5	0.09677	0.00051789	0.10100	-0.380
5	40	0.140	0.07	2.0	0.13326	0.00044708	0.13732	-0.489
100	20	0.140	0.07	2.0	0.13969	0.00003447	0.13973	-0.005
100	40	0.070	0.07	1.0	0.06963	0.00001785	0.06977	0.006
200	20	0.070	0.07	1.0	0.06982	0.00001722	0.06986	-0.027
200	40	0.070	0.07	1.0	0.06982	0.00000899	0.06990	-0.009
300	20	0.070	0.07	1.0	0.06994	0.00001169	0.07000	-0.008
300	40	0.035	0.07	0.5	0.03492	0.00000545	0.03496	0.000

Table 3 : Random Selection of $\hat{a}$ Summary Statistics

$N_2(0)$	T	a	c	R_0	mean($\hat{c}$)	var($\hat{c}$)	median($\hat{c}$)	$\rho(\hat{a}, \hat{c})$
50	40	0.035	0.07	0.5	0.07091	0.00006506	0.07034	-0.004
100	40	0.070	0.07	1.0	0.07034	0.00001723	0.07015	0.006
100	40	0.105	0.07	1.5	0.07019	0.00000806	0.07008	-0.007
200	20	0.105	0.07	1.5	0.07010	0.00001175	0.07000	-0.007
200	20	0.140	0.07	2.0	0.07007	0.00000809	0.07004	-0.007
300	20	0.035	0.07	0.5	0.07012	0.00001469	0.07008	-0.004
500	20	0.035	0.07	0.5	0.07005	0.00000902	0.07002	-0.003
500	20	0.105	0.07	1.5	0.07005	0.00000461	0.07000	0.008
500	40	0.035	0.07	0.5	0.07010	0.00000609	0.07002	0.012
1000	20	0.035	0.07	0.5	0.07006	0.00000433	0.07004	0.006

Table 4 : Random Selection of $\hat{c}$ Summary Statistics

$N_2(0)$	T	a	c	R_0	mean($\hat{R}_0$)	var($\hat{R}_0$)	median($\hat{R}_0$)
5	40	0.035	0.07	0.5	0.43255	0.08842763	0.42888
50	20	0.035	0.07	0.5	0.49951	0.01497913	0.49202
50	20	0.140	0.07	2.0	1.99277	0.04051883	1.98941
100	20	0.070	0.07	1.0	0.99856	0.01403924	0.99422
100	40	0.070	0.07	1.0	0.99327	0.00689571	0.99369
200	20	0.105	0.07	1.5	1.49857	0.00917040	1.49868
300	40	0.035	0.07	0.5	0.49858	0.00160204	0.49925
500	40	0.035	0.07	0.5	0.49956	0.00096347	0.49970
500	40	0.070	0.07	1.0	0.99871	0.00137583	0.99869
1000	20	0.070	0.07	1.0	0.99950	0.00137986	0.99882

Table 5 : Random Selection of $\hat{R}_0$ Summary Statistics

Whereas a significant skewness is observed for the estimation of contagion parameter, this feature largely disappears for the reproduction number. This is due to the nonlinear transformation to compute R_0, but also to the dependence between $\hat{a}$ and $\hat{c}$. R_0 is known at $\pm 20\%$ for $N_2(0) = 100$, at $\pm 10\%$ for $N_2(0) = 1000$.

To have more insight on the finite sample properties of these estimators, we provide summary statistics including the correlation (ρ) between $\hat{a}$ and $\hat{c}$, for different designs (a,c), initial $N_2(0)$, and number of observations T in Tables 3-5. Finite sample distributions for the estimators computed by unfeasible Gaussian approximate likelihood are given in Appendix 3.

5 The Reproductive Number Under Heterogeneity

5.1 Model with heterogeneity

Another source of variability for the estimated R_0 is due to latent heterogeneity and concerns the definition of R_0 itself. For illustration, we consider a situation with two homogeneous populations, population 1 and population 2, say. Then the SIR model to a $(\text{SIR})^2$ model with six states : $S_1I_1R_1S_2I_2R_2$ in the terminology of Gourieroux, Jasiak (2020)b, Appendix 1. The (6,6) transition matrix is block diagonal with diagonal blocks given by :

$$P_{j,t} = \begin{pmatrix} 1 - a_{j1}\dfrac{N_2^1(t-1)}{N^1} - a_{j2}\dfrac{N_2^2(t-1)}{N^2} & a_{j1}\dfrac{N_2^1(t-1)}{N^1} + a_{j2}\dfrac{N_2^2(t-1)}{N^2} & 0 \\ 0 & 1 - c_j & c_j \\ 0 & 0 & 1 \end{pmatrix},$$

for $j = 1, 2$, where $N_2^j(t)$ (resp. n^j) is the number of infected people in population j (resp. the size of population j). Typically, the two populations can correspond to two age categories, young and old, say. Now the contagion parameter has a matrix form : $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$. Indeed there is contagion within each population : a_{11}, a_{22}, and between the populations a_{12}, a_{21}.

The $(\text{SIR})^2$ model can be constrained by introducing degrees of infectiveness and of infection vulnerability, denoted α_j and β_j, respectively. Then, the contagion matrix A is equal to : $A = \beta\alpha'$. This matrix has reduced rank equal to 1.

The existence of between and within contagions will modify the notion of the reproductive number which now must account for the different types of contagions from a new infected individual of type 1 (resp. 2) to individuals at risk of either type 1, or 2. The initial reproductive number now has a matrix form :

$$\begin{aligned} R_{0,0} &= \beta\tilde{\alpha}', \\ \text{with } \tilde{\alpha}_j &= \alpha_j/c_j, j = 1, 2. \end{aligned}$$

The diagonal elements of matrix $R_{0,0}$ can be very different. For instance, if one segment includes the super-spreaders, the reproductive number can pass from a value around 2 [WHO (2020)] to a value between 4.5 and 11.5 [Kochanczik et al. (2020)].

5.2 Omitted heterogeneity

Let us now assume such an underlying $(\text{SIR})^2$ model and aggregate the two subpopulations in $S = S_1 U S_2$, $I = I_1 U I_2, R = R_1 U R_2$. There is an aggregation bias that implies that the cross-sectional counts :

$$N_1(t) = N_1^1(t) + N_1^2(t), N_2(t) = N_2^1(t) + N_2^2(t), N_3(t) = N_3^1(t) + N_3^2(t),$$

no longer define a Markov process. However, it is still possible to compute the transition matrix at horizon 1. Let us for instance consider the probability for an individual at risk at date $t-1$ (i.e. in state S at $t-1$) to be infected at date t by a new infectious individual. By the Bayes formula, we get :

P[infected at t | at-risk at $t-1$]
$= P$[infected at t | at-risk $t-1$, in Pop 1] P [at-risk $t-1$, Pop 1 | at-risk at $t-1$]
$+ P$ [infected at t | at risk at $t-1$, in Pop 1] P[at risk at $t-1$, in Pop 2 | at risk at $t-1$]

$$
\begin{aligned}
&= \frac{N_1^1(t-1)}{N_1(t-1)}\left[a_{11}\frac{N_2^1(t-1)}{N^1}+a_{12}\frac{N_2^2(t-1)}{N^2}\right]\\
&+ \frac{N_1^2(t-1)}{N_1(t-1)}\left[a_{21}\frac{N_2^1(t-1)}{N^1}+a_{22}\frac{N_2^2(t-1)}{N^2}\right]\\
&= \left[\beta_1\frac{N_1^1(t-1)}{N_1(t-1)}+\beta_2\frac{N_1^2(t-1)}{N_1(t-1)}\right]\left[\alpha_1\frac{N_1^1(t-1)}{N^1}+\alpha_2\frac{N_2^2(t-1)}{N^2}\right]\\
&= a_t\frac{N_2(t-1)}{N},
\end{aligned}
$$

where a_t is the dated transmission parameter in the SIR model with omitted heterogeneity. Therefore, using the standard SIR model when there is heterogeneity implies a time varying contagion parameter. A similar effect, known as the mover-stayer phenomenon, exists for the intensity to recover from the infection state and leads to a time varying c_t, and therefore on the reproductive number : $R_{0,0,t} = a_t/c_t$.

This type of decomposition can easily be extended to more than two homogeneous subpopulations [see e.g. Alipoor, Boldea (2020)].

6 Instantaneous Reproductive Number

There exists on the market different packages to estimate a reproductive number, usually in a rolling way. We discuss below one set of estimation methods to approximate the instantaneous reproductive number, a notion that differs from the basic reproduction number.[9] This type of computation and the associated software can be found in [Cori et al. (2013), with the EpiEstim package, the time dependent reproduction number in the RO

[9] A proposed alternative is to define R as an exponential rate of diffusion of the disease usually estimated by either log-regression, or Poisson regression. [see e.g. Lipsitch et al (2003), Wallinga, Lipsitch (2007), Boelle et al. (2009)]. Other approaches are based on some assumption of a network of contagion, as in Wallinga, Teunis (2004). However their proposed methodology assumes a static equilibrium network, tries to reconstitute a tracing ex-post, without really taking into account the dynamic of the disease. This implies a right censoring bias [Cauchemez et al. (2006)]

package Obadia et al. (2012)] and are used, for instance, in the official reproductive number provided by Public Health Ontario [PHO (2020)]. Even if this approach is presented to estimate time varying reproduction number, the methodology is expected to work also in a framework of a weakly time dependent reproduction number. This is why the discussion is done under the SIR model.

6.1 A generic estimator

Alternative estimation approaches of the reproductive number have been introduced in the literature and in the software. They are often presented as almost model free and are popular among practitioners since they are simple to use. An example of such a generic approach has been introduced in Fraser (2007), Cori et al. (2013) following a similar idea appeared in Wallinga, Teunis (2004), p511. The method requires the knowledge of the sequence of new infections only, $N_{12}(t)$, with t varying. The count of time t is written on the lagged counts as :

$$N_{12}(t) \simeq \sum_{s=1}^{S} \gamma_s N_{12}(t-s),$$

and the regression coefficients can be normalized as : $\gamma_s = w_s \gamma$, where $\sum_{s=1}^{S} w_s = 1$.

The estimated "instantaneous reproduction number" is defined in EpiEstim as [see Cori et al. (2020), p2] :

$$\hat{R}_t^i = \frac{N_{12}(t)}{\sum_{s=1}^{t} N_{12}(t-1)\hat{w}_s}, \tag{6.1}$$

where the sum in the denominator starts at the first time of infections and $\hat{w}_s$ is a Bayesian estimate of the infectiousness profile. [10]

[10] Sometimes the infectiveness profile of w_s is even not really estimated, but fixed ex-ante, possibly through a prior [see e.g. Cory et al. (2013), webappendix 4 and the discussion below]. The results will significantly depend on this selected sequence.

Such a simple procedure is not necessarily robust: it depends on the length of the estimation period, the number of lags in the sum appearing in the denominator, on the choice of the infectiveness profile w_s and on its estimate. But more importantly, any generic approach will work well under some implicit assumptions and if the notion on interest is correctly defined under these assumptions.

Let us illustrate the properties of the EpiEstim approach. This estimator is usually computed in a rolling way. It is based on a Bayesian assumption with a prior on the distribution of the serial interval, that is, the time from symptom onset in a primary case (infector) to symptom onset in a secondary case (infectee). This prior depends on two parameters, that are a mean and a standard deviation. In EpiEstim1 we have retained the same log-normal prior : mean = 4.5 days, standard deviation = 2.5 days chosen by PHO (2020). It is close to the prior in Nishura et al (2020) [mean = 4.7 days, standard deviation = 2.5, based on 18 pairs of infector - infectee], but different from the prior in Du et al. (2020) [mean = 3.96, standard deviation = 4.15, based on 468 pairs].

In Figure 6 we display different estimates computed from a simulated series satisfying the SIR model. The EpiEstim1 estimate is calculated on a window of seven days. The approximate ML estimates [Binomial, Poisson, Unfeasible Gaussian] are computed at each date t using all the data from the outbreak. The Poisson and Binomial estimates cannot be distinguished. All estimates have poor properties at the beginning, when the number of new infections is rather small and there are almost no recoveries. The ML estimators show a variability which becomes rather small after 30 days, and they converge to the true value of the basic reproductive number.

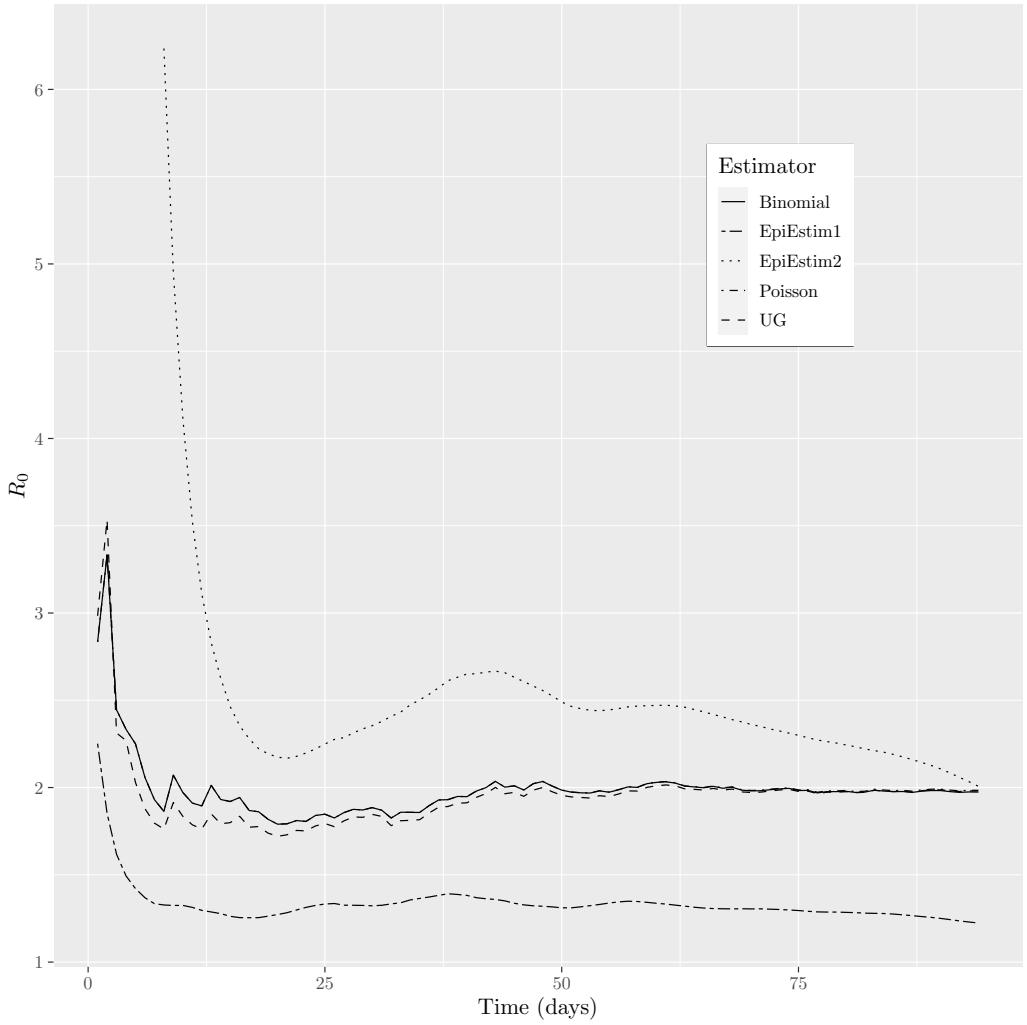
Let us now discuss the evolution of the EpiEstim1 estimator. This evolution is strongly dependent on the Bayesian approach that is used. Even if the estimate is computed in a rolling way, only seven observations are taken into account at each date t, that gives a significant weight to the prior. This explains the weak variability of this estimate over time. Moreover the level of the estimate is strongly dependent on the selected prior and clearly it is not varying around the true value of R_0, even if it accounts for the information in the counts of new infected. In EpiEstim1, we have followed the current practice in which the prior relies on preexisting estimates of the serial interval distribution. These estimates can correspond to another disease, to the same disease in another country, or in the case of COVID-19, a small number of observations: 18 pairs in Nishura et al. (2020), endogenously se-

lected (12 among these pairs correspond to transmission within family, then to short transmissions). These are estimated using the definition of the serial interval as the time between symptomatic cases [Thompson et al. (2019)] which will underestimate the mean and uncertainty due to the presence of asymptomatic infection periods and/or individuals.

Finally the choice of a log-normal prior instead of a gamma prior, that is, of a thin tail prior instead of a fat tail prior can also lead to an underestimation of the level. A further implication of the Bayesian approach can be observed when one provides the software with a zero vector instead of a vector containing new infections. In this scenario, the process will simply return a reproduction number which is constant over time.

In order to check the role of the prior, we also display in Figure 6 the plot corresponding to the EpiEstim estimator with a log-normal prior with the same mean and standard deviation as the geometric distribution with mean 14 days. This is an unfeasible estimator assuming that the infectivity profile is fixed at its true value [see the discussion in 6.4.2, and formula (6.19)]. A convergence to the true value R_0 is now observed. These drawbacks of the EpiEstim approach have been recently mentioned by some authors of the R software package [Thompson et al. (2019)], who propose an improved version of their package. We will discuss it later on, however this recent version has not yet been implemented.

Figure 6 : Comparison Using EpiEstim on Simulated SIR Model Data



The objective of the following sections is to discuss the origin of the EpiEstim approach in order to explain the differences in the estimates observed in Figure 6.

6.2 The underlying model

To understand formula (6.1), we have to extend the basic SIR model. We retain a constant contagion parameter a but introduce a stochastic duration

of infectiousness D, which is not necessarily geometrically distributed. Its distribution is characterized by its survival function denoted:

$$\gamma(s) = P[D \geq s], s = 1, 2, ... \tag{6.2}$$

Then the expression of the basic reproductive number is easily derived (see Section 2.3). It becomes :

$$R_{0,t} = \frac{a}{N_1(t)} \sum_{s=0}^{\infty} \{E_t(N_1(t+s)\gamma(s)\}. \tag{6.3}$$

Let us now write this expression in terms of new infections. We have :

$$N_1(t) - N_1(t-1) = -N_{12}(t), \tag{6.4}$$

and then :

$$N_1(t+s) = N_1(t) - \sum_{k=1}^{s} N_{12}(t+k). \tag{6.5}$$

By replacing $N_1(t+s)$ by this expression in equation (6.3), we get :

$$\begin{aligned} R_{0,t} &= \frac{a}{N_1(t)} \sum_{s=0}^{\infty} \{\gamma(s)[N_1(t) - E_t[\sum_{k=1}^{s} N_{12}(t+k)]\} \text{ (with the convention } \sum_{k=1}^{0} = 0) \\ &= a \sum_{s=0}^{\infty} \gamma(s) - \frac{a}{N_t(t)} \sum_{s=1}^{\infty} \sum_{k=1}^{s} [\gamma(s) E_t(N_{12}(t+k))] \\ &= a \sum_{s=0}^{\infty} \gamma(s) - \frac{a}{N_1(t)} \sum_{k=1}^{\infty} [E_t N_{12}(t+k) \sum_{s=k}^{\infty} \gamma(s)]. \end{aligned}$$

The partial sums of the survival function $\gamma(s)$ can be rewritten in terms of moments of the stochastic duration of infectiousness. We get :

$$R_{0,t} = aE(D) - \frac{a}{N_1(t)} \sum_{k=1}^{\infty} \{E[(D-k)^+] E_t(N_{12}(t+k)]\}, \tag{6.6}$$

where $x^+ = Max(x, 0)$.

Remark 3 : In the standard SIR model, formula (6.6) becomes :

$$R_{0,t} = (a/c)\{1 - \sum_{k=1}^{\infty}[(1-c)^k E_t(N_{12}(t+k))]\}.$$

Let us now discuss the conditional expectation E_t. In the SIR framework, the conditioning set includes the current and lagged values of the $N_{jk}(t), j, k = 1, 2, 3$, or equivalently of the cross-sectional counts $N_k(t), k = 1, 2, 3$. Therefore the sufficient summary of the past information requires two sequences of counts.

By considering a single sequence of counts, i.e. the counts of new infected people, the generic approach is changing the information set and modifies the definition of the dated reproductive number (see the discussion in Section 6.3).

With this restricted information set, the new reproductive number is :

$$R_{0,t}^N = aED - \frac{a}{N_1(t)} \sum_{k=1}^{\infty}\{E[(D-k)^+]E[N_{12}(t+k)|\underline{N_{12}(t)}]\}, \tag{6.7}$$

where index N indicates the restriction to new infections.

Can we expect a linear prediction formula for the prediction of the counts of new infected people, such as :

$$E[N_{12}(t+k)|\underline{N_{12}(t)}] = \sum_{h=0}^{\infty} \beta_{kh} N_{12}(t-h), \tag{6.8}$$

with time independent β_{kh} coefficients? Likely not, due to the nonlinear dynamics of a contagion model and its analysis during a non-stationary episode.

6.3 Which Definition of Reproduction Number

To understand the significant difference between the formula (6.1) for $\hat{R}_t^i$ and the formula (6.7) for R_{0t}^N, it is useful to come back on the paper in which the notion of instantaneous reproductive number was introduced [see Fraser (2007)]. Fraser's approach is based on a renewal equation :

$$I(t) = \sum_{s=1}^{\infty} \beta(t,s) I(t-s), \tag{6.9}$$

where $I(t)$ is the incidence proportion [11] at t, or attack rate (approximated by $N_{12}(t)/N_1(t-1)$) and $\beta(t,s)$ is the effective contact rate between infectious and susceptible individuals taking into account the generation of new infected people. Both the SIR model and the renewal equation appear in the same paper of Kermack-McKendrick (1927) and are compatible. Under the SIR model, the contact rate $\beta(t,s)$ is a complicated nonlinear function of the sufficient summary counts, that is, the new infected and new recovered counts between dates $t-s$ and t. Therefore, in the SIR framework, the renewal equation (6.9) involves a "lagged endogenous" contact rate, in fact an equilibrium contact rate.

Let us now give the definitions of reproduction ratios in Fraser (2007). Two notions called "case reproductive ratio" and "instantaneous reproductive ratio", respectively, are introduced with the main objective to get a ready-to-use measure based on simple analytical formulas. It is important to note that they have new names, since they significantly differ from the standard basic and effective reproductive numbers. It is particularly important to note that they do not have the same interpretation. For instance, the instantaneous reproductive number is defined from (6.9) by considering what reproduction can be expected if "the conditions remain unchanged" in the past, i.e. $I(t-s) = I, s = 1, 2$. The ratio is then defined as [see eq.(3) in Fraser (2007)] :

$$R_t^i = \frac{I_t}{I} = \sum_{s=1}^{\infty} \beta(t,s). \tag{6.10}$$

This practice disregards the endogeneity of the contact rates. Indeed the contact rates also depend on the evolution of the number of new infected individuals which has assumed to be unchanged in the "linear" component of the renewal equation but not in the (nonlinear) contact rate. Moreover, the assumption of unchanged condition is not necessarily compatible with the evolution with peak corresponding to a SIR model and to the observations of $I(t)$, or $N_{12}(t)$. In fact, one objective of this definition was to reveal in the measure the expected sudden decrease of R resulting from a new effective control undertaken at time t.

Finally to derive the expression (6.1), it is also assumed a decomposition of the contact rate as :

[11] See CDC (2012) for the different definitions of incidence depending on the selected denominator.

$$\beta(t,s) = R_t^i w(s), \tag{6.11}$$

where the $w(s), s = 1, \ldots, S$ sum up to 1.

By taking into account this reduced rank condition, the renewal equation (6.9) is equivalent to :

$$R_t^i = I_t / \sum_{s=1}^{S} (I_{t-s} w(s)], \tag{6.12}$$

which explains the generic estimate (6.1) (if $N_1(t)$ is not changing a lot, see the discussion in Section 6.4) and its interpretation as the ratio of new infections by the total infectiousness of infected individuals up to time $t-1$.

6.4 Sources of Bias

Let us now make explicit the three main sources of bias when a formula such as (6.12) is used to approximate the basic reproductive number. The discussion is done under the assumption [12] of a SIR model with constant parameters $a, c, R_{00} = a/c$. As usual in epidemiology it is important to distinguish between the stochastic models of the observations and its associated deterministic (or mechanistic) model corresponding to a virtual population of infinite size [Breto et al. (2009), Smieszek (2009), Funk et al. (2018)].

6.4.1 The mechanistic model

Let us derive a mechanistic model of infection derived from the SIR model. As in Section 3.2, we denote $p_1(t), p_{12}(t)$ the theoretical probabilities corresponding to the frequencies $\hat{p}_1(t) = N_1(t)/n, \hat{p}_{12}(t) = N_{12}(t)/n$. We assume that the frequencies $\hat{p}$ tends to the $p's$ when n tends to infinity. In this case, $p(t)$ is also equal to the (unconditional) expectation of $\hat{p}(t)$. Let us focus on the mechanistic component of the model for infection, that is without considering recovery.

[12] Some Monte-Carlo studies have been performed in the literature under some specific renewal model, comparing R_t^i with its estimate [see e.g. Cori et al. (2013)]. Such an analysis is misleading since the right comparison is between the estimate of R_t^i and the basic R_{00} to measure the bias that can result from the approximations described in Section 6.3.

When n varies, we need to appropriately adjust the contagion parameter to derive the mechanistic model, i.e. to replace a by $a_n = a/n$, say. Then we have :

$$E_{t-1}\left(\frac{N_{12}(t)}{n}\right) = a\frac{N_1(t-1)}{n}\frac{N_2(t-1)}{n}. \tag{6.13}$$

Let us now decompose the count $N_2(t-1)$ as :

$$N_2(t-1) = \sum_{s=1}^{t} N_2(t-1;s), \tag{6.14}$$

where $N_2(t-1;s)$ is the number of individuals infected at $t-s$ for the first time and still infectious at $t-1$. In the SIR model with geometric duration of infection, we have :

$$E_{t-1}\left[\frac{N_2(t-1;s)}{n}\right] = \frac{N_{12}(t-s)}{n}(1-c)^{s-1}, \tag{6.15}$$

$$\text{then: } E\left[\frac{N_2(t-1,s)}{n}\right] = (1-c)^{s-1}E\left(\frac{N_{12}(t-s)}{n}\right). \tag{6.16}$$

Making n tend to infinity in these relations and using the fact that the limit of the $p's$ are deterministic, we get the deterministic recursive equation :

$$p^*_{12}(t) = ap_1(t-1)\sum_{s=1}^{t}[(1-c)^{s-1}p^*_{12}(t-s)], \tag{6.17}$$

or equivalently,

$$p^*_{12}(t) = a[1-\sum_{s=1}^{t} p^*_{12}(t-s)]\sum_{s=1}^{t}[(1-c)^{s-1}p^*_{12}(t-s)], \tag{6.18}$$

where $p^*_{12}(t) = \lim_{n\to\infty}[N_{12}(t)/n]$. $p^*_{12}(t)$ differs from $p_{12}(t)$, by the denominator n instead of $N_1(t-1)$, except at the beginning of the disease. From (6.18), we see that the series $p^*_{12}(t) = E(N_{12}(t)/n)$ satisfies a quadratic recursive equation with an order that tends to infinity with t.

6.4.2 The linearization bias

A first approximation assumes that $p_1(t-1)$ is close to 1. This approximation is reasonable and standard at the beginning of the disease, but can induce biases in the medium run (when looking for the peak) and in the long run (when looking for final size and herd immunity). Under this approximation, we get :

$$\begin{aligned} p^*_{12}(t) &\simeq a\sum_{s=1}^{t}[(1-c)^{s-1}p^*_{12}(t-s)] \\ &= \frac{a}{c}\sum_{s=1}^{t}[w(s)p^*_{12}(t-s)], \text{ or} \\ p^*_{12}(t) &= R_{0,0}\sum_{s=1}^{t}[w(s)p^*_{12}(t-s)], \end{aligned} \tag{6.19}$$

with $w(s) = c(1-c)^{s-1}$.

The relation (6.19) on the expected new infection rates is the basis of the methodology introduced in Fraser (2007).

6.4.3 The causality bias

In a deterministic equation as (6.19), the fact that a variable is in the right hand side, or in the left-hand side of the equation is not important. However this becomes important, when the population size is large, but finite, and the probabilities replaced by their frequency analogues. We can always deduce from (6.19) a relation on observations as :

$$N_{12}(t) = -R_{0,0}\sum_{s=1}^{t}[w(s)N_{12}(t-s)] + u(t), \tag{6.20}$$

where $u(t)$ are errors, but there is no reason for the $u(t)'s$ to be independent of one another and more importantly, to be independent of the lagged $N_{12}(t-s), (s = 1, \ldots)$. In fact, under the SIR model, they are dependent and correlated with the lagged counts. This can induce a bias when a (Poisson) least squares is applied to estimate $R_{0,0}$ (assuming either fixed $w(s)$, i.e. given c, or $w(s)$ estimated by OLS).

6.4.4 The bias and lack of efficiency of the two-step approach

Presenting an estimation approach of a simple OLS type and based only on the counts of new detected individuals has made some success for the EpiEstim approach. However, this presentation is a bit misleading. Indeed the weights $w(s)$ (i.e. parameter c) are unknown. Either they are fixed in some arbitrary way (see the description in the EpiEstim package) and this bias on the weights will imply a bias in the computation of R_0, as seen in the example of Figure 6, or they have to be estimated. This estimation will require more complicated approaches and the efficient use of other count series such as the total count of infected individuals, or the counts of new recovered individuals (see Section 2.2), or alternatively, tracing data on infector/infectee pairs (see the discussion in the conclusion).

In fact, the mechanistic equation (6.19) shows that there is an identification issue for large population sizes and automatically a poor accuracy for an estimate based on the observation of counts $N_{12}(t)$ only. Indeed, by introducing the lag-operator L, equation (6.19) can be written as :

$$\begin{aligned} p_{12}^*(t) &= R_{0,0} \sum_{s=1}^{t} [c(1-c)^{s-1} L^s p_{12}^*(t)] \\ \Leftrightarrow \quad p_{12}^*(t) &= R_{0,0} cL \Big(\sum_{s=0}^{t-1} (1-c)^s L^s\Big) p_{12}^*(t) \\ &\sim R_{0,0} cL \frac{1}{1-(1-c)L} p_{12}^*(t) \quad \text{(for large } t\text{)} \end{aligned}$$

This is equivalent to :

$$\begin{aligned} & [1-(1-c)L] p_{12}^*(t) = cR_{0,0} L p_{12}^*(t) \\ \Leftrightarrow \quad & p_{12}^*(t) = [1 + c[R_{0,0} - 1] p_{12}^*(t). \end{aligned}$$

Due to the linearization highlighted in Section (6.4.2), the mechanistic model tends to an exponential pattern for large t. Moreover, for a large number of observation T, we can essentially identify the "rate of explosion" equal to $1 + c[R_{0,0} - 1]$, not separately $R_{0,0}$ and c.

6.5 The Autoregression Estimate

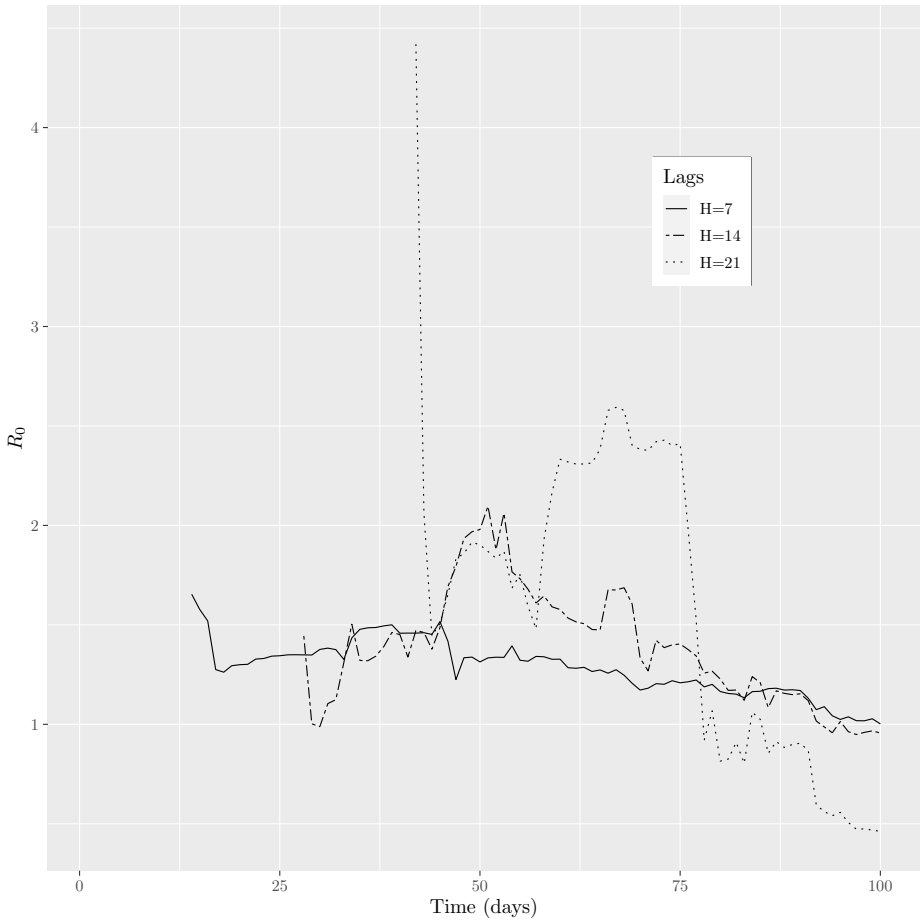
An alternative estimate of the reproductive number can also be introduced based on the approximate asymptotic relation (6.19). This estimator depends only on the counts of new infected individuals and is easy to compute as follows :

First, select an autoregressive order H and then regress $N_{12}(t)$ on $N_{12}(t-1), \ldots, N_{12}(t-H)$ [without intercept] by OLS, for $t = H+1, \ldots, T$. If $\hat{\gamma}(s), s = 1, \ldots, H$, are the estimated regression coefficients, define the estimator of the reproductive number as :

$$\hat{R}_{00}^{AR} = \sum_{s=1}^{H} \hat{\gamma}(s). \tag{6.21}$$

This estimator has a variance that will increase with H, since more underlying parameters have to be estimated. It has also the drawback of being computable only after at least $2H+1$ days, due to the lag and the minimal number of observations necessary to identify the autoregressive parameter. These estimates have been computed for $H = 7, 14, 21$ days in a non-rolling way on the same set of simulated data used to generate Figure 6.

Figure 7 : The Autoregression Estimate



The variability effect and the impossibility to use it at the beginning of the disease are clearly observed. We also note that these estimators do not converge to the true value. Indeed this approach is also subject to the causality and linearization biases. The bias is observed in Figure 7 with an underestimation of R_0.

7 Concluding Remarks

The estimated reproductive numbers are used as basic tools to follow the progression of an epidemic such as COVID-19 and monitor the (changes in)

health policy. For instance, specific partial lockdown policies may be introduced if the estimated R_0 is larger than 1.5. Such policies neglect the variability of both this notion and its approximations (estimates). We have considered this question in the framework of a time discretized SIR model and shown that this variability can be due to the definition itself which is time dependent, sometime author dependent, or to an omitted underlying heterogeneity. It is also a consequence of the different estimation methods that are used, with bias and uncertainty that depend on the available information.

As a by-product we have shown that the estimate of R_0 based on the Poisson approximate likelihood of the SIR model, used in a rolling way, with possibly a prior on parameters (see Appendix 4 for the Bayesian estimation) is as simple as the approach suggested in the standard EpiEstim, with two advantages: it is using the information of both new infected and currently infected people, and does not fix arbitrarily the infectiousness profile.

As mentioned in the text, Thompson et al. (2019) highlight issues related to the use of the standard EpiEstim and propose an improved version of the package to correct some of the drawbacks of the standard one. They propose to use in real time two series of data: the counts of new infected people and "up-to-date observations of serial intervals". In this approach there will be (as in the rolling approach based on SIR) a larger information set, and this updating will also introduce path varying mean and standard deviation of the distribution of the serial interval. With this improved version, the two approaches will not differ greatly by the underlying model on which they are based, (see the discussion in 6.4.1, 6.4.2) but instead by the observations they are using to calibrate the parameters. That is, the counts of new infected and currently infected people in the SIR model based estimator and the counts of new infected and data on pairs of infector/infectee obtained by tracing.

The choice of approach should largely depend on the availability (and cost) of such data, especially at the beginning of the epidemic, and on their reliability. In particular, the available data are currently incomplete, since they do not account for the undetected asymptomatic people [see Gourieroux, Jasiak (2020 a], and they are left and right censored for tracing of the pairs infector/infectee.

The SIR model has been chosen since the different estimation approaches were implicitly based on this model and it has facilitated the discussions and comparisons. Clearly, to obtain a more complete picture, similar exercises would have to be done on models with more features. The aim of this extended analysis would be to account for the difference between the infec-

tion and infectious period and also incorporate a stochastically time varying contagion parameter component [see e.g. Gourieroux, Lu (2020)].

References

[1] Adam, D. (2020) : A Guide to R : The Pandemic's Misunderstood Metric", Nature, 583, 346-348.

[2] Alipoor, A., and O., Boldea (2020) : "The Role of Elementary Schools in SARS-CoV-2 Transmission", DP Tilburg Univ.

[3] Allen, L. (1994) : "Some Discrete Time SI, SIR and SIS Epidemic Models", Mathematical Biosciences, 124, 83-105.

[4] Aronson, J., Brasset, J., and K., Mahtani (2020) : "When Will it Be Over ? An Introduction to Viral Reproduction Numbers", Oxford COVID-19 Evidence Service.

[5] Bailye, N. (1975) "The Mathematical Theory of Infectious Diseases and its Applications", 2nd edition, London Griffin.

[6] Becker, N., Glass, K., Barnes, B., Caley, P., Philip, D., McCaw, J., McVernon, J., and J., Woods (2006) : "The Reproduction Number Using Mathematical Models to Assess Responses to an Outbreak of an Emerged Viral Respiratory Disease", National Center for Epidemiology and Population Health, April.

[7] Bettencourt, L., and R., Ribeiro (2008) : "Real Time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases", Plos One, 3:e2185.

[8] Boelle, P., Bernillon, P, and J., Desencios (2009) : "A Preliminary Estimation of the Reproduction Ratio for New Influenza A(H1N1) from the Outbreak in Mexico, March-April 2009", Euro Surveill., 14, pli = 19205.

[9] Breto, C., He, D., Ionides, E., and A., King (2009) : "Time Series Analysis via Mechanistic Models", Annals of Applied Statistics, 3, 319-348.

[10] Cauchemez, S., Boelle, P., and C., Donnelly (2006) : "Real Time Estimates in Early Detection of SARS", Emerg. Infect. Dis., 12, 110-113.

[11] Center for Disease Control and Prevention (CDC) (2012) : "Principles of Epidemiology in Public Health Practice : An Introduction to Applied Methodology and Biostatistics", Third Edition, Section 3.

[12] Cori, A., Ferguson, N., Fraser, C., and S., Cauchemez (2013) : A New Framework and Software to Estimate Time Varying Reproduction Numbers During Epidemics", American Journal of Epidemiology, 178, 1505-1512.

[13] Dietz, K. (1993) : "The Estimation of the Basic Reproduction Number for Infectious Diseases", Statistical Methods in Medical Research, 2, 23-45.

[14] Diekmann, O., Heesterbeck, N., and J., Metz (1990) : "On the Definition and the Computation of the Basic Reproduction Ratio R_0 in Models for Infectious Diseases in Heterogeneous Populations ", Journal of Mathematical Biology, 28, 365-382.

[15] Du, Z., Xu, X., Wu, Y., Wang, L., Cowling, B., and L., Ancel Meyers (2020) : "Serial Interval of COVID-19 Among Publicly Reported Confirmed Cases", Emerg. Infect Dis., 26, 1341-1343.

[16] Farrington, P., and H., Whitaker (2003) : "Estimation of Effective Reproduction Numbers for Infectious Diseases Using Serological Survey Data", Biostatistics, 4, 621-632.

[17] Fraser, C. (2007) : "Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic", Plos One, 2, e758.

[18] Funk, S., Camacho, A., Kucharsky, A., Eggo, R., and J., Edmunds (2018) : "Real-Time Forecasting of Infectious Disease Dynamics with a Stochastic Semi-Mechanistic Model", Epidemics, 22, 56-61.

[19] Galvani, A., Lei, X., and N., Jewell (2003): "Severe Acute Respiratory Syndrome : Temporal Stability and Geographic Variation in Case-Fatality Rates and Doubling Times", Emerg. Infect. Dis, 9, 991-994.

[20] Gourieroux, C., and J., Jasiak (2020)a : "Time Varying Markov Process with Partially Observed Aggregate Data : An Application to Coronavirus", forthcoming Journal of Econometrics.

[21] Gourieroux, C., and J., Jasiak (2020)b : "Analysis of Virus Transmission : A Transition Model Representation of Stochastic Epidemiological Models", forthcoming Annals of Economics and Statistics.

[22] Gourieroux, C., and Y., Lu (2020) : "SIR Model with Stochastic Transmission", CREST DP.

[23] Harko, T., Lobo, s., and M., Mak (2014) : "Exact Analytical Solution of the Susceptible-Infected-Recovered (SIR) Epidemic Model and the SIR Model with Equal Death and Birth Rates", Applied Mathematics and Computations, 236, 184-194.

[24] Hethcote, H. (2000) : "The Mathematics of Infectious Diseases", SIAM Review, 42, 599-653.

[25] Huber, P. (1967) : "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions", Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 221-233.

[26] Jones, J. (2007) : "Notes on R_0", Stanford Univ.

[27] Kermack, W., and A., Mc Kendrick (1927) : "A Contribution to the Mathematical Theory of Epidemics", Proceedings of the Royal Statistical Society, A, 115-700-721.

[28] Kochanczik, M., Grabowski, F., and T., Lipniacki (2020) : "Accounting for Super-Spreading Gives the Basic Reproduction Number R_0 of Covid-19 that is Higher than Initially Estimated", University of Warsaw.

[29] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y. et al. (2020) : "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia", The New Englant Journal of Medicine, 382, 1199-1207.

[30] Lipsitch, M., Cohen, T., Cooper, B., Robins, J., Ma, S., James, L. Gopalakrisna, G., Chew, S, Tan, C., Samore, M., Fisman, D., and M., Murray (2003) : "Transmission Dynamics and Control of Severe Acute Respiratory Syndrome", Science, 300, 1966-1970.

[31] Ma, J. (2020) : "Estimating Epidemic Exponential Growth Rate and Basic Reproduction Number", Infectious Disease Modelling, 5, 129-141.

[32] Ma, J., and D., Earn (2006) : "Generality of the Final Size Formula for an Epidemic of a Newly Invading Infectious Disease", Bulletin of Mathematical Biology, 68, 679-702.

[33] McDonald, G. (1952) : "The Analysis of Equilibrium in Malaria", Tropical Diseases Bulletin, 49, 813-829.

[34] Nishiura, H., Linton, N., and A., Akhmetzhanov (2020) :"Serial Interval of Novel Coronavirus (COVID-19) Infections", Int. J. Infect. Dis, 93, 284-286.

[35] Obedia, T., Haneef, R., and P.Y., Boelle (2012) : "The RO Package : A Toolbox to Estimate Reproduction Numbers for Epidemic Outbreaks", BMC Medical Informatics and Decision Making, 12, 147-1-9.

[36] Public Health Ontario (PHO) (2020) : "Epidemiological Summary : Evolution of COVID-19 Case Growth in Ontario", June 12.

[37] Riou, J., and C., Althaus (2020) : "Pattern of Early Human to Human Transmission of Wuhan (2010) Novel Coronavirus, December 2019 to January 2020", Eurosurveillance, 25, 4.

[38] Sanche, S., Liu, Y., Xu, C., Romero-Severson, E., Hengartner, N., and K., Rulan" (2020) : "Hight Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronoavirus 2", Emerging Infection Diseases, 26.

[39] Sanchez, M., and S., Blauer (1997) : "Uncertainty and Sensitivity Analysis of the Basic Reproduction Rate : Tuberculosis as an Example", Amer. J. of Epid., 145, 1127-1137.

[40] Smieszek, T. (2009) : "A Mechanistic Model of Infection : Why Duration and Intensity of Contacts Should be Included in Models of Disease Spread", Theoretical Biology and Medical Modelling, 6, 25.

[41] Thompson, R., Stockwin, J., Van Gaalen, R., Polonsky, J, Kamvar, Z., Demarsh, P., Dahlqwist, J., Li, S., Miguel, E., Jombart, T., Lessler, J., Cauchemez, S., and A., Cori (2019) : "Improved Inference of Time Varying Reproduction Numbers During Infectious Disease Outbreaks", Epidemics, 29, 10036.

[42] Toda, A. (2020) : "Susceptible-Infected-Recovered (SIR) : Dynamics of COVID-19 and Economic Impact", ArXiv : 2003. 11221v2.

[43] Wallinga, J., and M., Lipsitch (2007) : "How Generation Intervals Shape the Relationship Between Growth Rates and Reproductive Numbers', Proceedings of the Royal Society B: Biological Sciences, 274, 599.

[44] Wallinga, J., and P., Teunis (2004) : "Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures", American Journal of Epidemiology, 160, 509-516.

[45] White, L., and M., Pajano (2008) : "A Likelihood-Based Method for Real Time Estimation of the Scrial Interval and Reproduction Number of an Epidemic", Statist. Med., 27, 2999-3016.

[46] World Health Organisation (WHO) (2020) : "Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)".

[47] Wu, J., Leung, K., Bushman, M., Kishore, N., Niehus, R. et al. (2020) " Estimating Clinical Security of COVID-19 from the Transmission Dynamics in Wuhan, China", Nature Medicine, 506-510.

[48] Zhang, S., Diao, M., Yu, W., Pei, L., Lin, Z., and D., Chen (2020) : "Estimation of the Reproductive Number of Novel Coronavivus (COVID-19) and the Probable Outbreak Size on the Diamond Princess Cruise Ship: A Data Driven Analysis", International Journal of Infectious Disease, 93, 201-204.

Appendix 1

The Continuous Time SIR Model

The SIR model is usually written as a continuous time deterministic model. The standard notations are :

$$x(t) = p_1(t), y(t) = p_2(t), z(t) = p_3(t).$$

This model defines the dynamics of the cross-sectional structure by the system of differential equations :

$$\begin{cases} \dfrac{dx(t)}{dt} &= -\alpha x(t)y(t), \\ \dfrac{dy(t)}{dt} &= \alpha x(t)y(t) - \gamma y(t), \\ \dfrac{dz(t)}{dt} &= \gamma y(t), \end{cases} \tag{6.1}$$

where α, γ are positive parameters.

This differential system admits a closed form solution derived rather recently [Harko, Lobo, Mak (2014), Section 2, eq (17)-(20)]. This solution depends on parameters α, γ and on starting values $x(0), y(0), z(0)$.

Let us consider the integral equations :

$$\begin{aligned} t &= \int_{\exp[-\frac{\alpha}{\gamma}z(0)]}^{u(t)} \frac{dv}{v[-\alpha - \gamma \log v + \alpha x(0) \exp[\frac{\alpha}{\gamma}z(0]v]}, \\ &\equiv G[u(t); \alpha, \gamma, p(0)], \end{aligned} \tag{6.2}$$

where $p(0) = [x(0), y(0), z(0)]'$ is the initial structure. Then the solution is :

$$\begin{cases} x(t) &= x(0)\exp[\frac{\alpha}{\gamma}z(0)]G^{-1}[t;\alpha,\gamma,p(0)], \\ \\ y(t) &= 1-x(t)-z(t), \\ \\ z(t) &= -\frac{\alpha}{\gamma}G^{-1}[t;\beta,\gamma,p(0)]. \end{cases}$$

The knowledge of the solution allows to derive the following results.

i) $x(t)$ decreases to a limiting value $x(\infty)$.
ii) $z(t)$ increases to a limiting value $z(\infty)$.
iii) $y(t)$ usually increases to a peak, then decreases to $y(\infty)=0$.
iv) There is herd immunity that is $x(\infty)>0$, and this final size is equal to the solution of :

$$x(\infty)-x(0)-y(0)-\frac{c}{a}\log[x(\infty)/x(0)]=0.$$

v) The herd immunity can be reached in a finite time.

[see e.g. Kermack, McKendrick (1927), Hethcote (2000), or Ma, Earn (2006) for the expression of the final size, and Gourieroux, Lu (2020), for property v)].

An analogue discrete time deterministic model is :

$$\begin{cases} x(t) &= x(t-1)-ax(t-1)y(t-1), \\ \\ y(t) &= y(t-1)+ax(t-1)y(t-1)-cy(t-1), \\ \\ z(t) &= z(t-1)+cy(t-1). \end{cases}$$

This analogue is not the exact time discretized continuous time SIR. In particular the parameters a,c have interpretations that slightly differ from α,γ, and may depend on the time-step of the discretization. Moreover, in nonlinear dynamic systems, such a Euler discretization might change the dynamic properties of the trajectories. However, it is known that properties i), ii), iii) of the trajectories are still satisfied and that there is always herd immunity [Allen (1994)]. However, the herd immunity cannot be reached in

a finite time and the expression of the final size is not known under closed form.

This discrete time analogue is exactly the mechanistic model derived in Section 3.2.

Appendix 2

Statistical Inference

1. Likelihood function.

The individual histories are equivalently characterized by the sequence of dummy variables :

$$\begin{aligned} z_{jit} &= 1, \text{ if individual } i \text{ is in state } j \text{ at date } t, \\ & \\ & 0, \text{ otherwise.} \end{aligned}$$

Then, by applying the Bayes formula, the likelihood is equal to :[13]

$$\begin{aligned} l(a,c) &= \Pi_{j=1}^{2}\Pi_{k=1}^{3}\Pi_{i=1}^{n}\Pi_{t=1}^{T}\left[p_{jk}(t;a,c)^{z_{ij,(t-1)}z_{ikt}}\right] \\ & \\ &= \Pi_{j=1}^{2}\Pi_{k=1}^{3}\left[p_{kj}(t;a,c)^{\sum_{i=1}^{n}\sum_{t=1}^{T} z_{ij(t-1)}z_{ikt}}\right] \\ & \\ &= \Pi_{j=1}^{2}\Pi_{k=1}^{3} p_{jk}(t;a,c)^{\sum_{t=1}^{T} N_{jk}(t)}, \end{aligned}$$

where the transition probabilities may depend on $N_2(t-1)$. This explains why the $N_{jk}(t), j,k = 1,2,3$ define a sufficient statistics.

2. Decomposition of the log-likelihood function

We deduce :

[13] with the appropriate convention for treating the absorbing state.

$$
\begin{aligned}
L(a,c) &= \log l(a,c) \\
&= \sum_{k=1}^{3}\left[\sum_{t=1}^{T} N_{1k}(t)\log p_{1k}(t,a)\right] \\
&+ \sum_{k=1}^{3}\left[\sum_{t=1}^{T} N_{2k}(t)\log p_{2k}(t,c)\right] \\
&\equiv L_1(a) + L_2(c),
\end{aligned}
$$

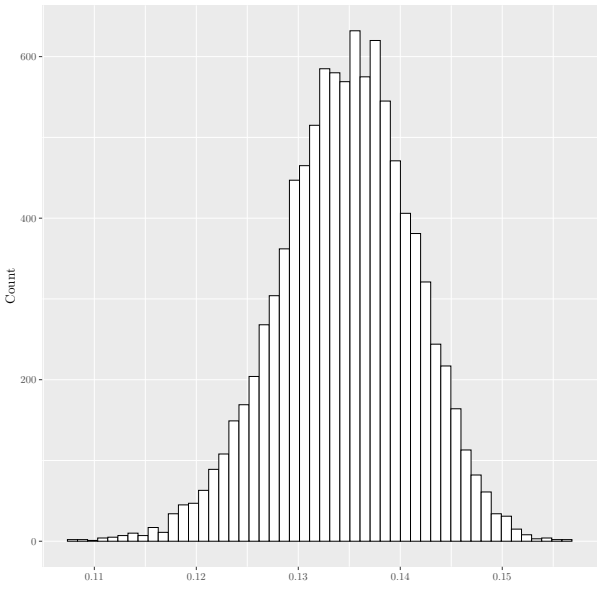
noting that the transition probabilities of the first row (resp. the second row) depend on a [resp. c] only.

Appendix 3

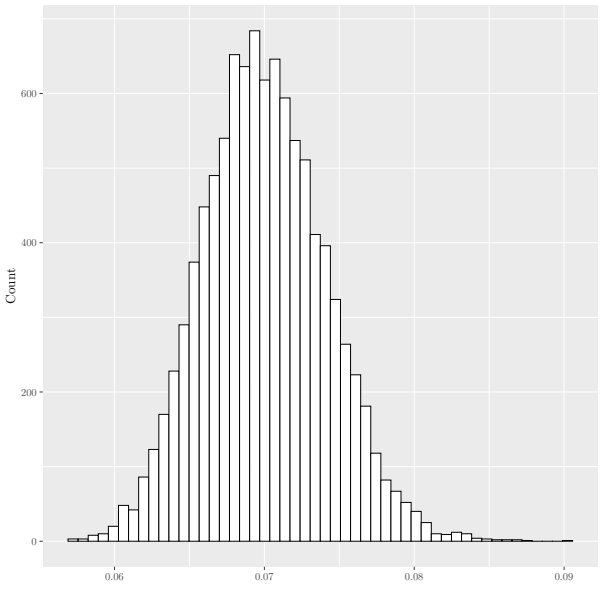
Finite Sample Properties of the Unfeasible Gaussian ML Estimators

We provide for the unfeasible Gamma ML estimator the Figures a.1, a.2, that are the analogues of Figures 4, 5 given in the text for the approximate Poisson ML estimator. These distributions are similar to the distributions for the Poisson . Nevertheless Figure 6 shows that their evolutions with the number of observations are highly different.

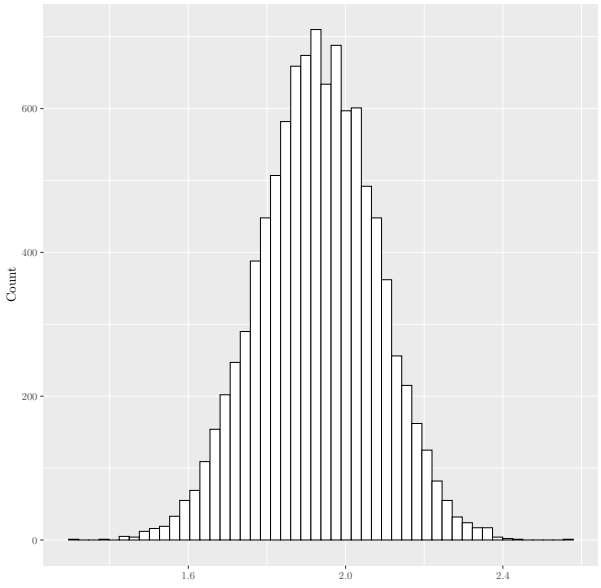
Figure a.1 : Distribution of Approximate Unfeasible Gaussian Estimators, $N_2(0) = 100$



$\hat{a}$

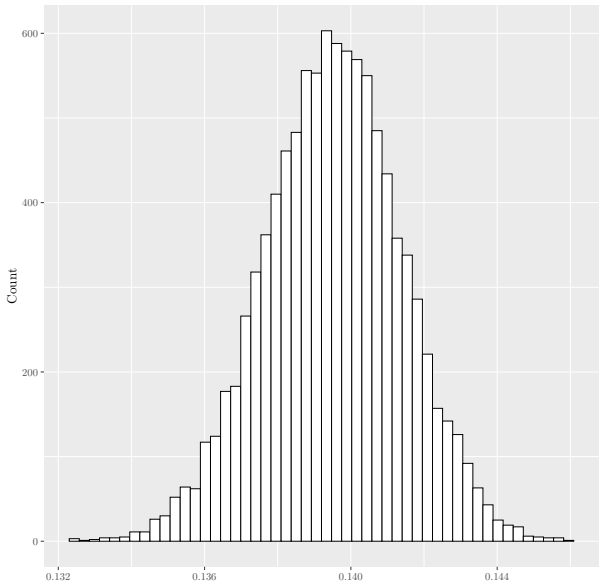


$\hat{c}$

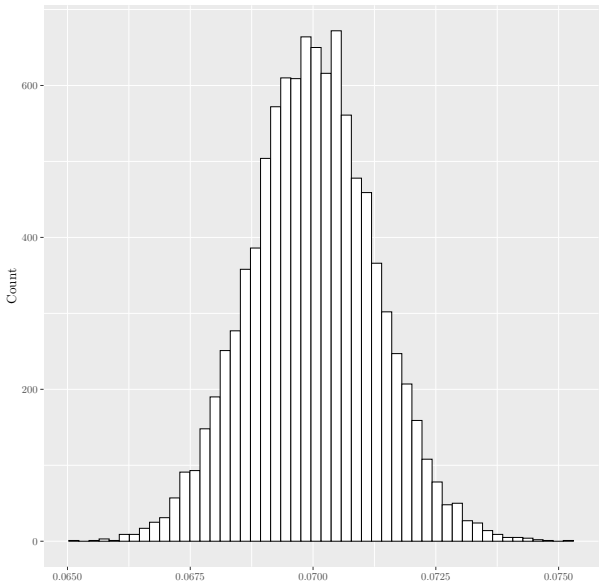


$\hat{R}_0$

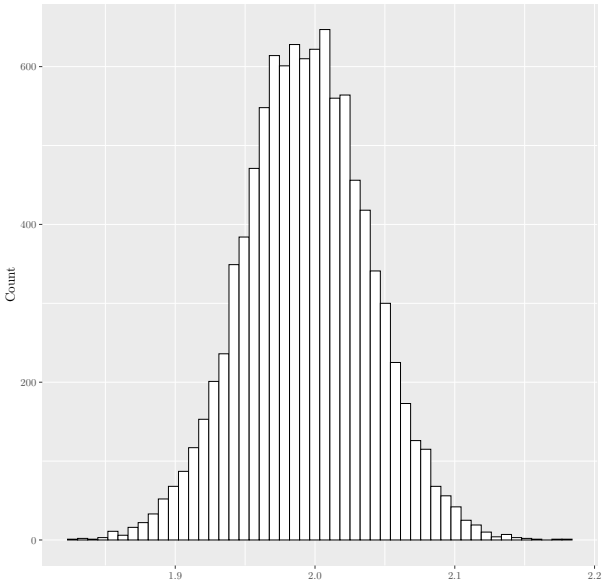
Figure a.2 : Distribution of Approximate Unfeasible Gaussian Estimators, $N_2(0) = 1000$



$\hat{a}$



$\hat{c}$



$\hat{R}_0$

Appendix 4

Bayesian Estimators

Several authors have considered Bayesian estimation approaches [see e.g. Cori et al. (2019), Webappendix 1]. To facilitate the comparisons, we consider below Bayesian estimation approaches for the Poisson approximate likelihood. As noted in the literature the Poisson likelihood has an expression in parameters a, c, that allows for a conjugate prior for these parameters. More precisely, the following result is easily derived.

Proposition : For the Poisson approximate likelihood,

i) A conjugate prior for a, c is such that a and c are independent with gamma distributions $\gamma(\nu_a, \lambda_a)$ and $\gamma(\nu_c, \lambda_c)$, respectively.

ii) For this prior the posterior is such that : a and c are independent with gamma distributions :

$$\gamma[\sum_{t=1}^{T} N_{12}(t) + \nu_a, \sum_{t=1}^{T}[N_1(t-1)\hat{p}_2(t-1)] + \lambda_a],$$

$$\gamma[\sum_{t=1}^{T} N_{23}(t) + \nu_c, \sum_{t=1}^{T}[N_2(t-1) + \lambda_c],$$

respectively.

iii) Let us denote $\nu_a(t), \nu_c(t), \lambda_a(t), \lambda_c(t)$ the degrees of freedom and scales of the posterior distributions of a and c. Then the posterior distribution of $R_0 = a/c$ is such that: $\dfrac{\lambda_a(t)\nu_c(t)}{\lambda_c(t)\nu_a(t)} R_0$ follows a Fisher distribution $F(2\nu_a(t), 2\nu_c(t))$.

This Bayesian analysis differs from the derivation in Cori et al. (2013). Indeed the conjugate prior is naturally introduced on parameters a and c by gamma distributions, whereas they introduce a nonconjugate prior on R_0^i and the infection profile $w(s)$, characterized by c in the SIR framework. This modifies significantly the posterior of R_0. The posterior mean of R_0 is : $\dfrac{\lambda_c(t)\nu_a(t)}{\lambda_a(t)} \dfrac{\nu_c(t)}{\nu_c(t)-1}$, if $\nu_c(t) > 1$, and does not exist, otherwise.

The reason for the non existence of the posterior mean is similar to the reason for the nonexistence of the Poisson approximate maximum likelihood estimator. If we observe no recovery, or if there is a prior for a long infectious period, the ML or Bayesian approaches can provide posterior distributions with fat tail.