A stochastic data-adapted model for epidemics.

Enrique M. Cabaña

Seminario de Probabilidad y Estadística –Seminario COVID
Universidad de la República

2020/07/03

1- A Markov chain version of the basic SIR model

The classes of the well known Susceptible-Infected-Removed (SIR) model are splitted into subclasses in order to adapt the model to the observation of the dates in which each individual transits the Covid when he or she is identified as been infected with Sars-CoV-2. Non-identified infected individuals are also included in hidden classes of the model.

The kind of data expected to be available for each identified person are the dates of

- infection,
- first symptoms,
- hospital confinement,
- transfer to intensive care unit,
- discharge from hospital,
- complete recovery or
- death,

whatever be appliable.

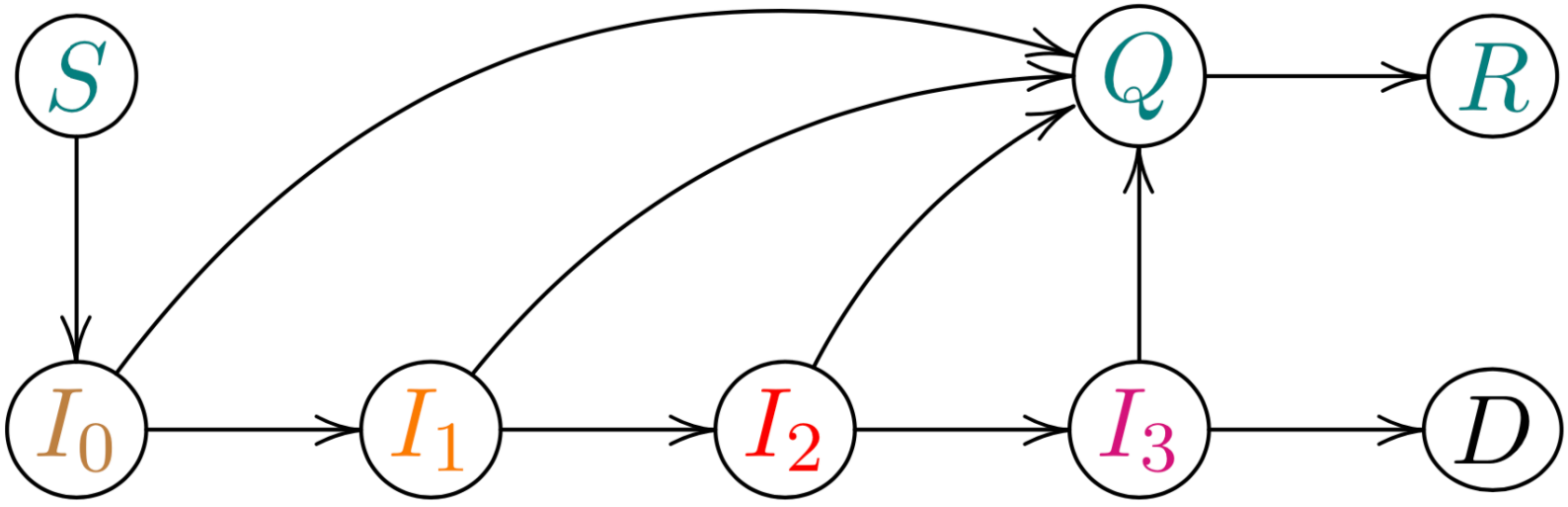
Consequently we split the population assumed of constant size N in the following classes integrated by persons with the indicated characteristics:

- S: Susceptible.
- I_0: Infected not (yet) detected.
- I_1: Infected, detected by the Health System, put into quarantine.
- I_2: In hospital confinement.
- I_3: Transferred to intensive care unit.
- Q: Discharged from hospital, still infectious, in quarantine.
- R: Recovered (non infectious, at least for a reasonable period of time).
- D: Deceased.

We assume that each individual in the population of size N is initially susceptible except for one or a few individuals infected that start the process of contagion. The former ones constitute the class S, and the latter the class I_0.

- A person in S that is infected moves to class I_0.
- A person in I_0 can be detected and moves to I_1 or be recovered after remaining some time infected and infectious.
- A person in I_1 can recover and move to Q, R or get worse and be confined to I_2 and eventually recover to Q or get even worse and move to I_3 and so forth.
- ...

The possible paths followed by the members of the population are indicated in next figure.


S
Q
R
I_0
I_1
I_2
I_3
D

Our main assumptions are:

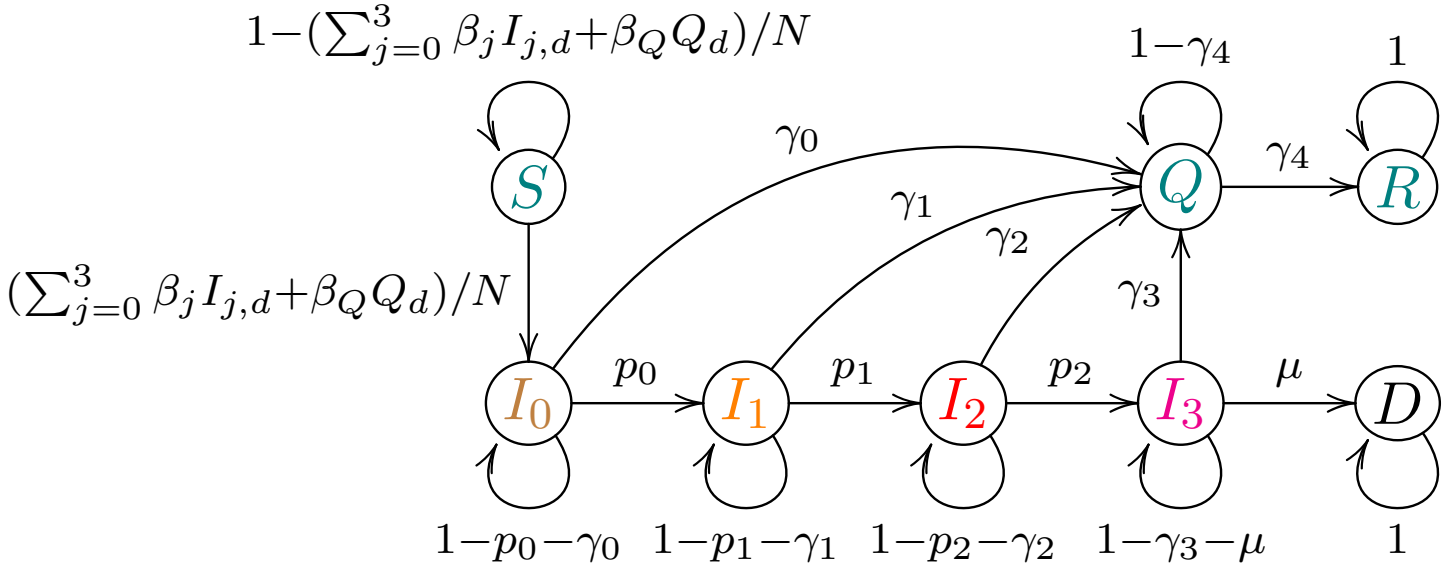
1. Each individual follows a Markov Chain with daily transitions from one class to another with the probabilities indicated in next diagram.

   The amount of persons in a class at day d is denoted by the name of the class with subscript d.

   The transit through the chain of persons not entering I_1 is not observed. If they arrive to I_1 their path is registered up to the end of the period of observation $d_I, d_{I+1}, \ldots, d_F$.

2. The individual paths are independent copies of each other.

REMARK: As a consequence, the time in each class has a geometric probability distribution with expected time equal to the inverse of the probability of leaving the class.



Parameters: $\gamma_1, \gamma_2, \gamma_3, \gamma_4, p_0, p_1, p_2, \mu, \beta_0, \beta_1.\beta_2, \beta_3, \beta_Q, N$

Let us denote $C_\nu := (C_{\nu,d})_{d=d_I,d_{I+1},\ldots,d_F}$ the chain indicating the class $C_{\nu,d}$ where the ν-th person is at day d, and $\boldsymbol{C} := (C_1, C_2, \ldots, C_N)$ the chain that joins the information of the entire population. If any day two individuals arbitrarily chosen are interchanged, the distribution of the chain $\boldsymbol{C}$ remains the same. Then the family of random vectors

$$X_d = (S_d, I_{0,d}, I_{1,d}, I_{2,d}, I_{3,d}, Q_d, R_d, D_d) = (X_{d,1}, X_{d,2}, \ldots, X_{d,8}),$$

one for each day, with components

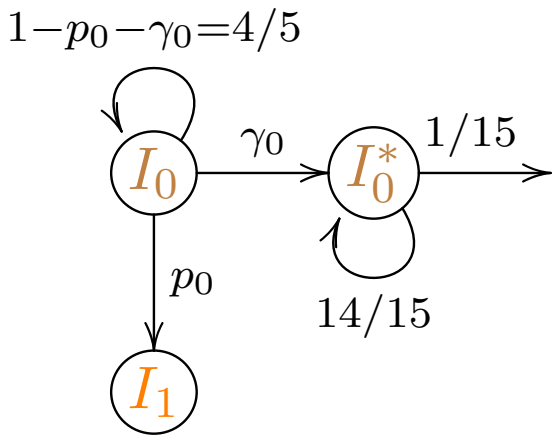
$$X_{d,i} = \sum_{\nu=1}^{N} \mathbf{1}_{\{C_{\nu,d}=i\}}, \quad i = 1, 2, \ldots, 8$$

equal to the amount of individuals at each class at day d is a Markov chain with state space equal to the subset of $\boldsymbol{N}^8$, $\boldsymbol{N} = \{0, 1, 2, \ldots, N\}$ composed by the vectors with sum of components equal to N.

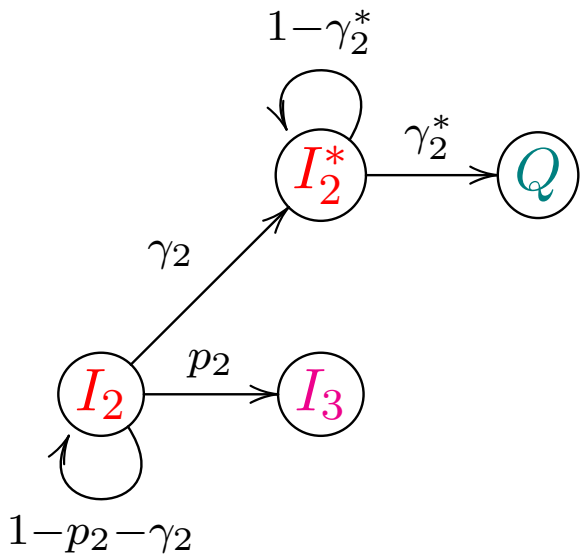
We introduce now two modifications to the basic chain due to the fact that the expected time that each person remains in one class may be different according to the class to which shall be directed in the next transition.

- A person in I_0 can be detected as bearing Sars CoV-2 and be put in quarantine in I_1, or remain undetected and carry the disease until completely recovered. This last case implies a stay of about twenty days at I_0, while been detected can occur in about five days. This is taken into account in our model by adding a new class I_0^* between I_0 and the continuation towards recuperation. An expected time of five days in I_0 and fifteen days in I_0^* fixes the chain and leads to a local scheme like this:

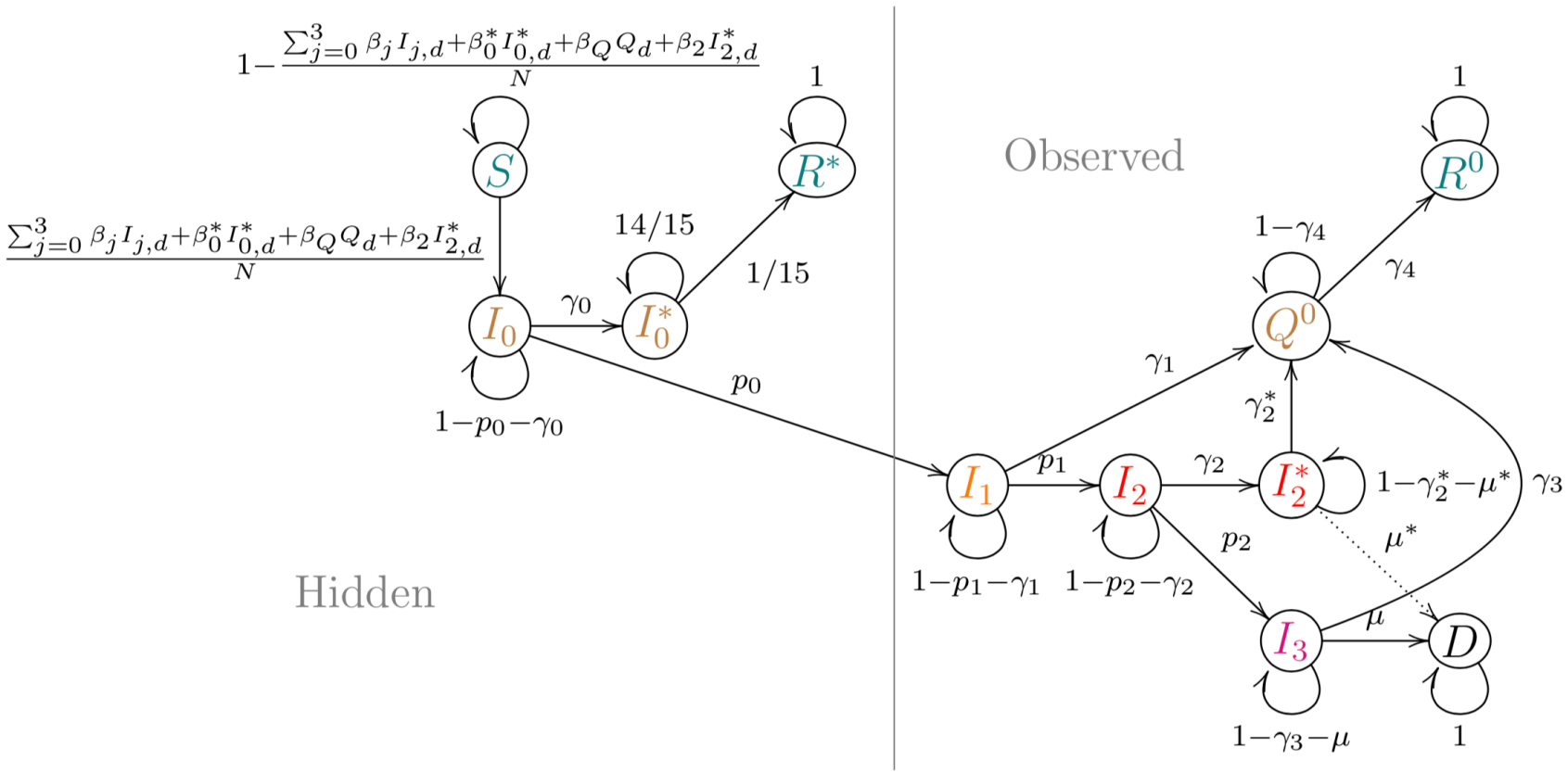
$$\gamma_0 + p_0 = 1/5$$



- Similar situations may appear at other classes. We shall assume that the transit when a patient leaves I_2 can occur either almost immediately after been confined in the hospital if the disease gets worse and leads him or her to I_3, or the stay in I_2 can last longer until been transferred to Q. For this case the local scheme to apply is as follows:



We make one last modification to separate the hidden states from the observed ones. In particular, this implies to split R into R^0 that contains the recovered patients after been identified as infected, and R^* constituted by the recovered persons after going through the disease undetected:



2-Parameter estimation

2-1- Estimation of the probabilities of observed transitions.

Let K denote the number of observable classes. Each individual in class x_i at day $d < d_F$ provides a realization of a multinomial random variable with parameters $(1, \boldsymbol{p}_i = (p_{i,j})_{j=1,2,\ldots,K})$, namely, the indicator of the class of the same individual at day $d+1$. These variables are independent and contain all the information about the parameter $\boldsymbol{p}_i$ provided by the observation of the epidemic.

The joint probability of all variables associated to class i is

$$\prod_{\nu=1}^{N} \prod_{d=d_I}^{d_F-1} p_{C_{\nu,d},C_{\nu,d+1}} = \prod_{i,j=1}^{K} p_{i,j}^{n_{i,j}},$$

with

$$n_{i,j} = \sum_{\nu=1}^{N} \sum_{d=d_I}^{d_F-1} \mathbf{1}_{\{C_{\nu,d}=i, C_{\nu,d+1}=j\}},$$

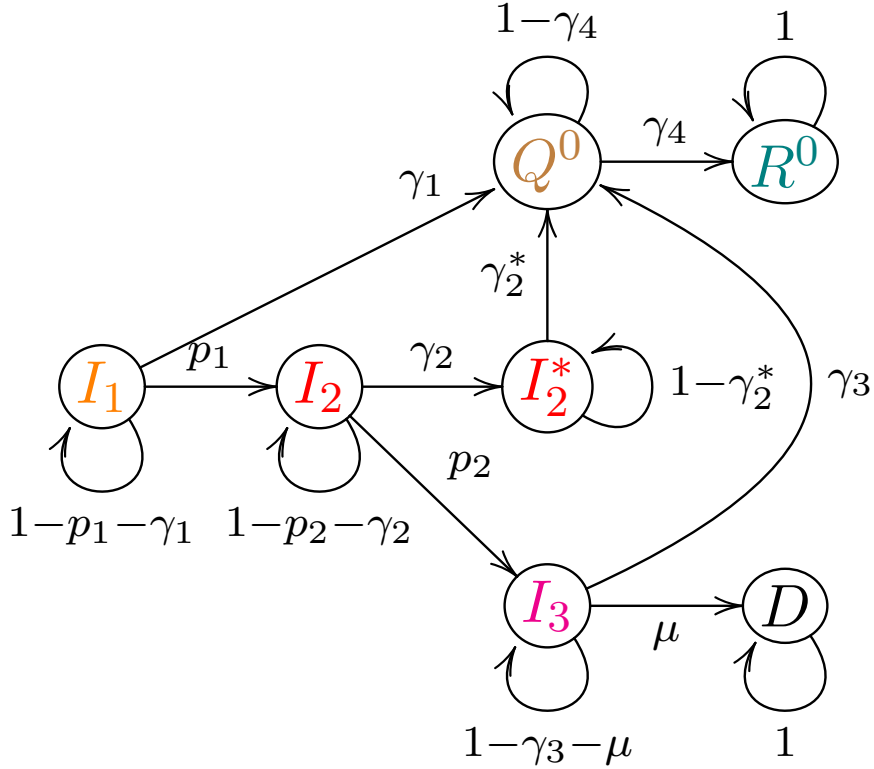
hence the maximum likelihood estimators of $p_{i,j}$ are

$$\hat{p}_{i,j} = \frac{n_{i,j}}{\sum_{j=1}^{K} n_{i,j}}$$

This solves the estimation problem for the probabilities

$$\gamma_1, \gamma_2, \gamma_2^*, \gamma_3, p_1, p_2, \mu.$$

The transit from Q^0 to R^0 may not be observed, because a common practice is to discharge patients from hospitals when they are recovered from symptoms but still infectious, and sent to home quarantine. In that case γ_4 is estimated according to the usual practice of the health system, as the inverse of the duration of quarantine.



2-2- Probabilities estimation for the hidden part of the chain

The hidden classes S, I_0, I_0^*, R_0^* are linked to the observed part of the chain through the inputs

$$H_d = \sum_{\nu=1}^{N} \mathbf{1}_{\{C_{\nu,d-1}=I_0, C_{\nu,d}=I_1\}}$$

assumed to be observed.
The parameters to be estimated are the probabilities γ_0, p_0 and the coefficients $\beta_0, \beta_0^*, \beta_1, \beta_2, \beta_3, \beta_Q$ involved in the calculation of the probability

$$p_{S,I_0} = \frac{\sum_{j=0}^{3} \beta_j I_{j,d} + \beta_0^* I_{0,d}^* + \beta_Q Q_d + \beta_2 I_{2,d}^*}{N}.$$

The rate of contagion increases with the amount of infectious individuals, and the expression adopted for p_{S,I_0} is a linear approach for that dependence.

All coefficients and consequently β depend on the social isolation measures adopted, and the sanitary precautions applied to reduce contagions, so that they may change over time. These changes can be reflected by adopting simple time varying expressions for those parameters, for instance, sectionally constant functions, or with sectionally constant increments.

We shall introduce a partition $d_0 = d_I < d_1 < \cdots < d_M = d_F$, and assume that each coefficient is constant on each of the intervals $d_{m-1} \leq d \leq d_m$.

The knowledge of the operating provisions applicable to enhance social isolation or the re-establishment of activities and services suggest where to place the end-points of the intervals of constancy of the coefficients or their increments, to be established a priori as initial approximations for optimization algorithms.

Some examples of end-points of such intervals are in Uruguay

- 03/13 to 03/19 the government adopts strong measures of social isolation,
- 04/4-12 Easter motivates a relaxation of the voluntary quarantine,
- 04/13 the construction sector resumes its activities,
- 04/23 rural schools restart face-to-face classes,
- 06/09 shopping centres are reopened,
- 06/1-15-29 gradual return of pre-school, primary, secondary and technical education activities,
- ...

We propose two different approaches in order to estimate the parameters of the hidden part of the model:

- a least squares approximation between the expected and observed values of the inputs H_d to I_1, and
- a maximum likelihood procedure based on the application of a Viterbi algorithm.

We start by considering the former one, which is much simpler than the latter.

2-2-a Least squares probabilities estimation

Let us introduce the notation $p_{S,I_0} = \beta_d I_d/N$ where
$I_d = I_{0,d} + I^*_{0,d} + \check{I}_d \qquad (\text{with } \check{I}_d := I_{1,d} + I_{2,d} + I^*_{2,d} + I_{3,d} + I_{Q,d})$
is the total amount of infected individuals.
The paths of the hidden part of the chain satisfy:

$$\begin{aligned}
S_d &= S_{d-1} - B_{d-1}, B_{d-1} \sim \text{Bin}(S_{d-1}, \beta I_{d-1}/N) \\
I_{0,d} &= M_{d-1,1} + B_{d-1}, M_{d-1} \sim \text{Mult}(I_{0,d-1}, (1-p_0-\gamma_0, p_0, \gamma_0)) \\
I^*_{0,d} &= M_{d-1,3} + B^*_{d-1}, B^*_{d-1} \sim \text{Bin}(I^*_{d-1}, 14/15) \\
H_d &= M_{d-1,2}
\end{aligned}$$

and the conditional expectations given the past $\mathcal{A}_{d-1}$ up to the day $d-1$ are

$$\begin{aligned}
\mathbf{E}(S_d|\mathcal{A}_{d-1}) &= S_{d-1} - \beta S_{d-1} I_{d-1}/N \\
\mathbf{E}(I_{0,d}|\mathcal{A}_{d-1}) &= I_{0,d-1}(1-p_0-\gamma_0) + \beta S_{d-1} I_{d-1}/N \\
\mathbf{E}(I^*_{0,d}|\mathcal{A}_{d-1}) &= I_{0,d-1}\gamma_0 + I^*_{d-1} \times 14/15 \\
\mathbf{E}(H_d|\mathcal{A}_{d-1}) &= I_{0,d-1} p_0
\end{aligned}$$

The following assumptions simplify the model:

- The contagion due to infected persons in quarantine or in hospital confinement is negligible compared with the one due to infected individuals non detected and consequently with unrestricted contacts with susceptible persons. Then p_{S,I_0} reduces to $\frac{\beta_{0,d} I_{0,d} + \beta^*_{0,d} I^*_{0,d}}{N}$.

- Let us identify the coefficient $\beta^*_{0,d}$ with $\beta_{0,d}$, since the classes I_0 and I_0^* are the result of artificially splitting the class of unobserved infected individuals to allow for different times of transition towards I_1 or R^*. This implies that β becomes equal to β_0.

These simplifications reduce the equations of the conditional expectations to

$$\begin{aligned}
\mathbf{E}(S_d|\mathcal{A}_{d-1}) &= S_{d-1} - \beta S_{d-1}(I_{0,d-1} + I^*_{0,d-1})/N \\
\mathbf{E}(I_{0,d}|\mathcal{A}_{d-1}) &= I_{0,d-1}(1 - p_0 - \gamma_0) + \beta S_{d-1}(I_{0,d-1} + I^*_{0,d-1})/N \\
\mathbf{E}(I^*_{0,d}|\mathcal{A}_{d-1}) &= I_{0,d-1}\gamma_0 + I^*_{d-1} \times 14/15 \\
\mathbf{E}(H_d|\mathcal{A}_{d-1}) &= I_{0,d-1}p_0.
\end{aligned}$$

The independence of the binomial and multinomial variables imply that the expectation of products of variables in the right-hand members are equal to the products of their expectations, except for the product $S_d I_{0,d}$.

The estimate

$$\begin{aligned} & \mathbf{E}S_d\mathbf{E}I_{0,d} - \mathbf{E}(S_dI_{0,d}) = \mathbf{E}(B_{d-1}^2) - (\mathbf{E}B_{d-1})^2 \\ = \; & \mathbf{Var}\mathbf{E}(B_{d-1}|\mathcal{A}_{d-1}) + \mathbf{E}\mathbf{Var}(B_{d-1}|\mathcal{A}_{d-1}) \\ < \; & \mathbf{Var}\beta S_{d-1}(I_{0,d-1} + I_{0,d-1}^*)/N + \mathbf{E}\beta S_{d-1}(I_{0,d-1} + I_{0,d-1}^*)/N \\ < \; & \mathbf{E}\beta^2(I_{0,d-1} + I_{0,d-1}^*)^2 + \mathbf{E}\beta(I_{0,d-1} + I_{0,d-1}^*) \end{aligned}$$

shows that while the proportion of infected persons is very small, the substitution of $\mathbf{E}S_d\mathbf{E}I_{0,d}$ for $\mathbf{E}(S_dI_{0,d})$ produces a negligible relative error of the order of $1/N$.

By replacing the random transitions by their approximate expectations, the expected paths of the hidden part of the chain, are computed by means of the recurrence

$$\begin{aligned}
\mathbf{E}S_d &= \mathbf{E}S_{d-1} - \beta \mathbf{E}S_{d-1}(\mathbf{E}I_{0,d-1} + \mathbf{E}I^*_{0,d-1})/N \\
\mathbf{E}I_{0,d} &= \mathbf{E}I_{0,d-1}(1 - p_0 - \gamma_0) + \beta \mathbf{E}S_{d-1}(\mathbf{E}I_{0,d-1} + \mathbf{E}I^*_{0,d-1})/N \\
\mathbf{E}I^*_{0,d} &= \mathbf{E}I_{0,d-1}\gamma_0 + \mathbf{E}I^*_{d-1} \times 14/15 \\
\mathbf{E}H_d &= \mathbf{E}I_{0,d-1}p_0.
\end{aligned}$$

Let us recall that given the partition $d_0 = d_I < d_1 < \cdots < d_M = d_F$, we assume that $\beta_{0,d} = b_m$ for $d_{m-1} \leq d \leq d_m$ is sectionally constant. Therefore the simplifications that we have introduced reduce the parameters to be estimated to

$$\boldsymbol{b} := (b_1, \ldots, b_M), \quad \boldsymbol{d} := (d_1, \ldots, d_{M-1}), \gamma_0, p_0.$$

Let us compute the least squares estimates of the parameters

$$(\tilde{\boldsymbol{b}}, \tilde{\boldsymbol{d}}, \tilde{\gamma}_0, \tilde{p}_0) = \arg\min SS(\boldsymbol{b}, \boldsymbol{d}, \gamma_0, p_0)$$

for $SS = \sum_{d=d_I}^{d_F} (\mathbf{E}H_d - h_d)^2$ where

- $h_d = \sum_{\nu=1}^{N} \mathbf{1}_{\{C_{\nu,d-1}=I_0, C_{\nu,d}=I_1\}}$ is the observed value of the input to I_1 at day d, $(d_I \leq d \leq d_F)$, and
- $\mathbf{E}H_d$ are the expectations obtained from the equations of the expected path with parameters $\boldsymbol{b}, \boldsymbol{d}, \gamma_0, p_0$.

How difficult is the optimization procedure obviously depends on the number and kind of parameters included in the model. With the selected parameters, $M = 4$ and data resembling the evolution of the Covid-19 epidemic in Uruguay, the sum of squares presents more than one relative minima.

2-2-b Maximum likelihood probabilities estimation

The hidden Markov chain with states $X_d = (S_d, I_{0,d}, I^*_{0,d}, H_d)$ in the state space $\boldsymbol{N}^4$ has transition probabilities depending on the parameters β, γ_0, p_0 and emits $H_d \in \boldsymbol{N}$ with probability one.

We keep the notations and simplifications introduced in §2-2-a.

Then the well known Viterbi algorithm may be used to obtain the paths $\boldsymbol{X}_d := (X_{d_I}, X_{d_I+1}, \ldots, X_d)$ with $X_d = x$ that maximize the conditional probabilities given that $\boldsymbol{H}_d := (H_{d_I}, H_{d_I+1}, \ldots, H_d)$ is equal to $\boldsymbol{h}_d := (h_{d_I}, h_{d_I+1}, \ldots, h_d)$:

$$m(x, d, \beta, \gamma_0, p_0, \boldsymbol{h}) = \max_{\boldsymbol{x}_d, x_d = x} \mathbf{P}\{\boldsymbol{X}_d = \boldsymbol{x}_d | H_d = \boldsymbol{h}_d\}$$

The standard algorithm applies a forward recurrence in d to determine that maxima, and the maximum probability attained by a complete path is $\ell(\beta, \gamma_0, p_0) = \max_x m(x, d_F, \beta, \gamma_0, p_0, \boldsymbol{h})$.

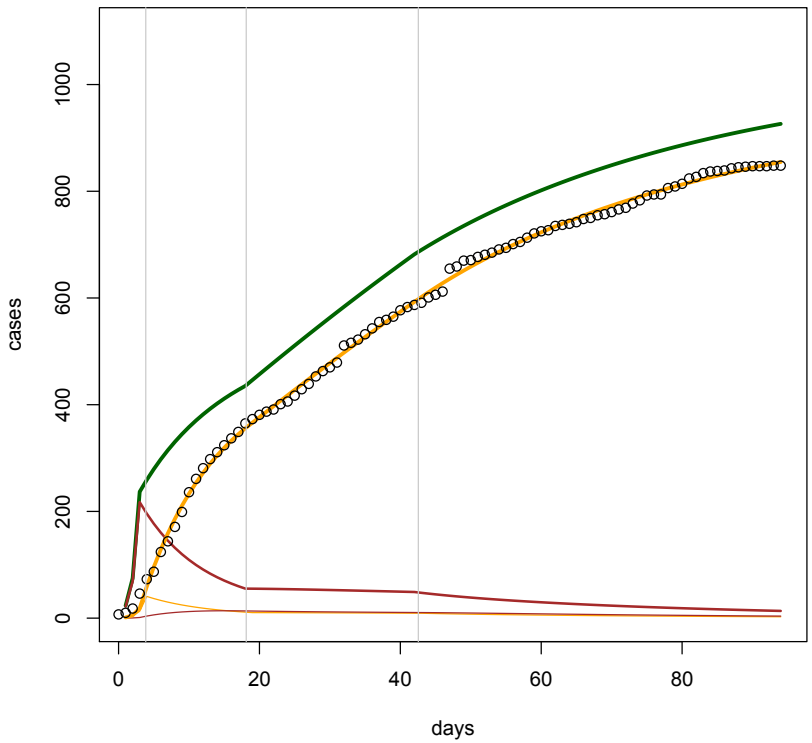
The maximum likelihood estimators are then

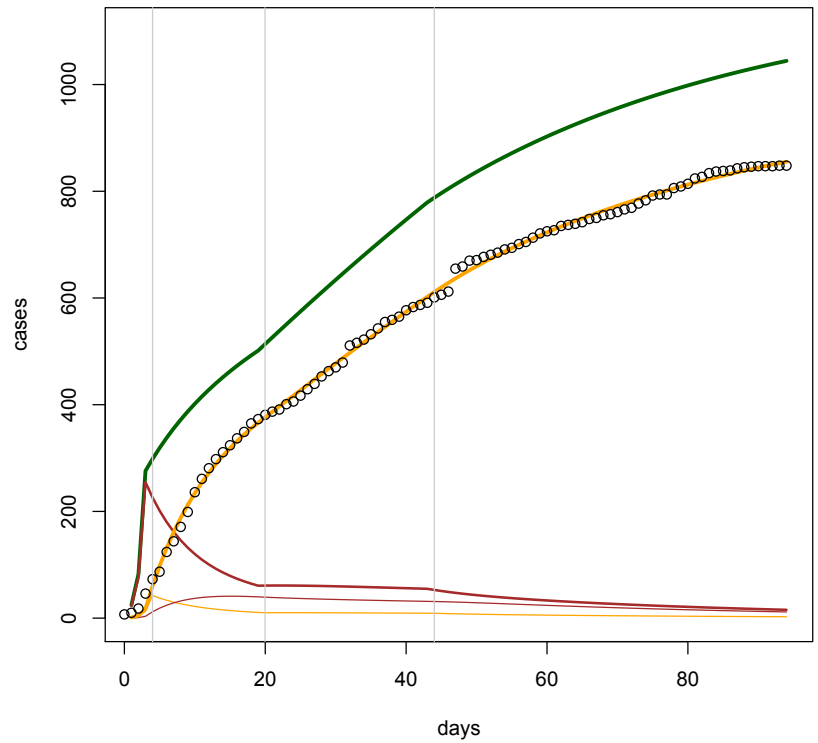
$$(\hat{\beta}, \hat{\gamma}_0, \hat{p}_0) = \arg\max \ell(\beta, \gamma_0, p_0).$$

The algorithm adds a backward recurrence in order to find the maximizing path, thus obtaining a maximum likelihood estimation of the sizes of the hidden classes along the path, but this can be avoided if our purpose is just to obtain the probabilities estimations.

Least squares estimation for the hidden chain

parameters								sum of squares
b_0	b_1	b_2	b_3	d_1	d_2	d_3	p_0	
2.335593	0.101232	0.159153	0.138412	4	19	43	0.187314	77.365
2.494468	0.086325	0.122817	0.102434	4	20	44	0.166285	78.226

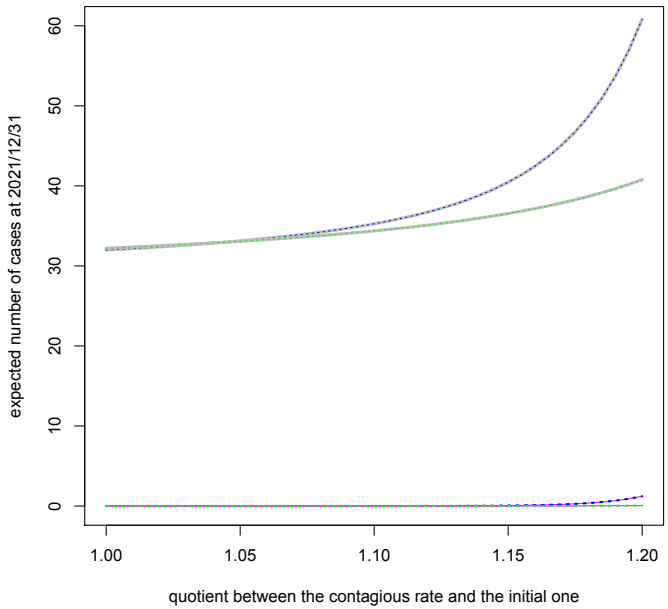




number of cases on d_F		
S	I_0	I_0^*
2999074	13.53	3.85
2998956	15.61	11.61

proportion of undetected infections
6.343%
16.9575%

The difference between the scenarios of the previous page is quantitatively small, but the following graph shows that the second one (green dotted lines) resists better than the first (blue dotted lines) the effect of increasing the contagious rate after the end of the period of observation.



The lines on gray background correspond to the expected accumulated number of deaths by 2021/12/31, and the lines on magenta background indicate the expected number of patients in ICU at the same time.

The simulation of the estimated paths of the system is aided by an interactive application.

`https://emcabana.shinyapps.io/pred9june/`

COVID-19 Stochastic Model for Uruguay

Days since 9 June 2020

0 206 571

0 58 116 174 232 290 348 406 464 522 571

Seed

Variables to include

- [x] N-S (black)
- [x] I0 (brown)
- [x] I1 (orange)
- [x] I2 (red)
- [x] I3 (magenta)
- [x] R (green)
- [x] Active (yellow)
- [x] D (gray)

Add random paths for one variable

- [x]

Which one?

- [ ] N-S
- [ ] I0
- [ ] I1
- [ ] I2
- [x] I3
- [x] D

Number of simulated random paths

0 19 50

0 5 10 15 20 25 30 35 40 45 50

State of the system on 2020/6/9

Scale maximum

Deceased number

Select the upper limit of cases axis

1 1,000

1 101 201 301 401 501 601 701 801 901 1,000

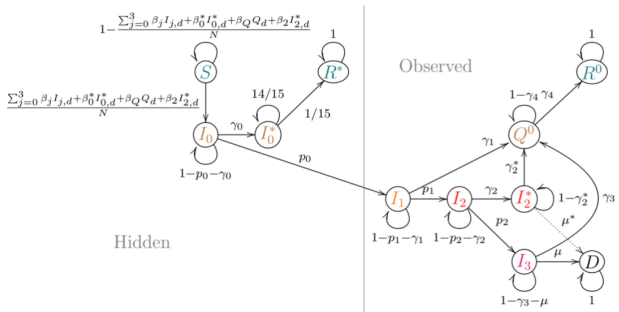
Magnification factor of the initial contagion rate

1 2

1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2

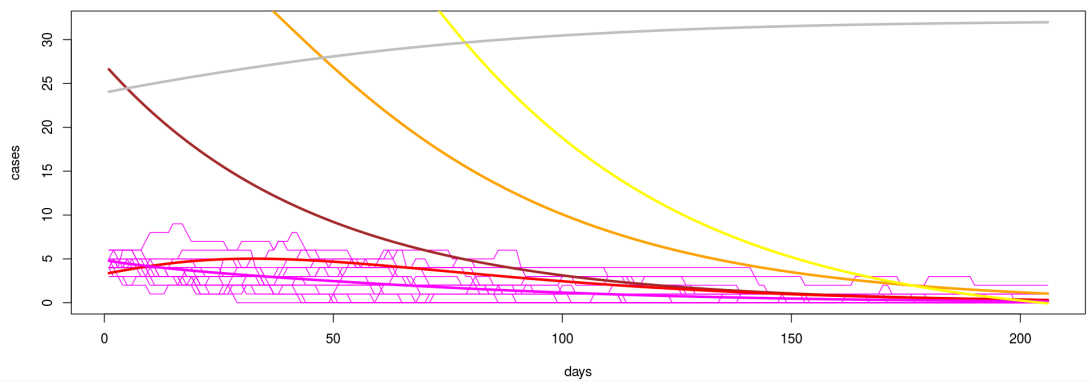
Evolution of the epidemics with constant contagious rate starting 9 June 2020, with simulated initial data resembling the situation in Uruguay

by Enrique M. Cabaña



S: Susceptible.
I_0, I_0^*: Infected not (yet) detected.
I_1: Infected, detected by the Health System, put into quarantine.
I_2, I_2^*: In hospital confinement.
I_3: Transferred to intensive care unit.
Q^0: Discharged from hospital, still infectious, in quarantine.
R^0, R^*: Recovered (non infectious, at least for a reasonable period of time).
D: Deceased.

It is assumed that the individuals in a population of size N (=3000000) follow independent paths of the Markov chain indicated in the diagram. Next plots allow to compare two scenarios slightly different, corresponding to two alternative estimations based on the available data for the initial period of the disease (3/7 to 6/9).



Initial contagion rate = 0.102381902478001 . Proportion of infected persons detected = 0.831425

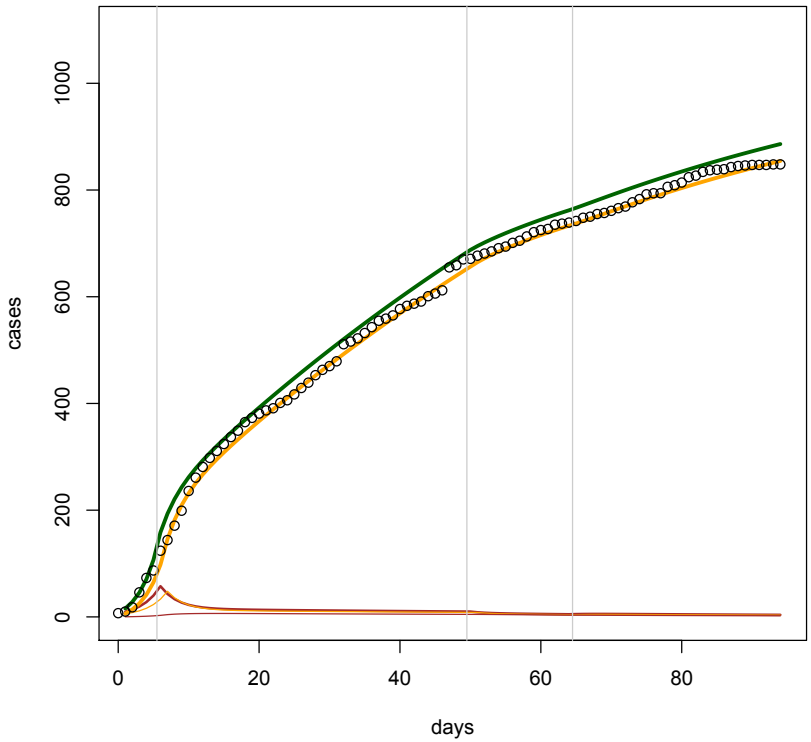
Complement to the least squares estimation

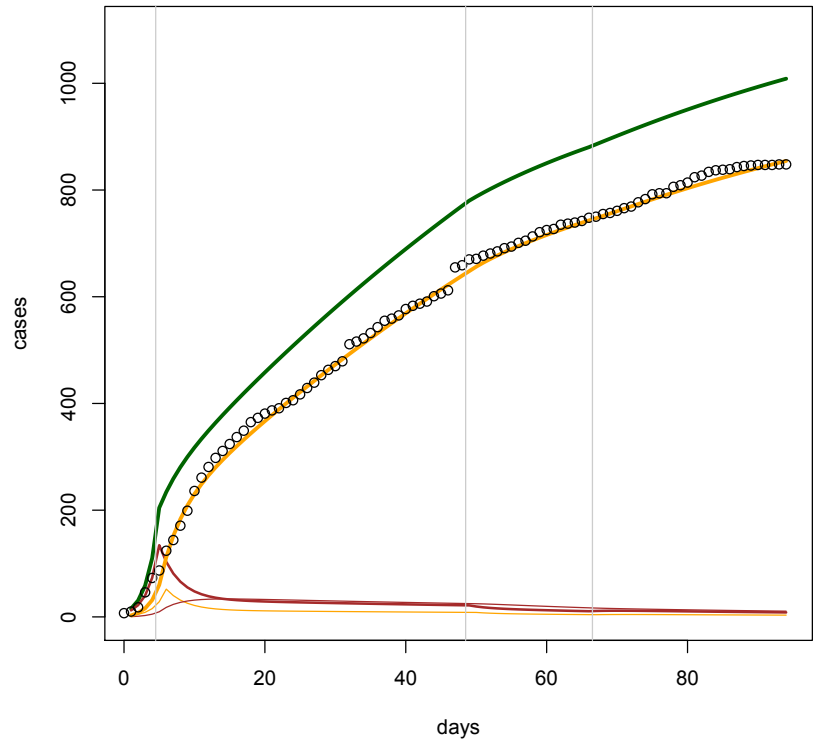
(added after the 3 July presentation in the Seminar)

We have relaxed the imposition $p_0 + \gamma_0 = 0.2$ due to assume that the mean time in I_0 is five days, and have included p_0 and γ_0 as free parameters in the least squares estimation. This leads to several new minima, from which we select the following two that differ mainly in the estimation of the mean number of days spent before infected people are identified:

parameters									sum of
b_0	b_1	b_2	b_3	d_1	d_2	d_3	p_0	γ_0	squares
1.243393	0.578258	0.495824	0.552874	5.82	49.69	65.00	0.833906	0.027178	72.576026
1.204182	0.206181	0.165161	0.192428	4.63	49.00	66.77	0.387359	0.066053	84.884535

These new estimations are added as Scenarios 3 and Scenario 4 in the Shiny application.





number of cases on d_F			proportion of undetected infections
S	I_0	I_0^*	
2999114	3.86	2.05	3.156%
2.998991	8.39	10.74	14.568%

Again the scenario with larger proportion of undetected infections resist better an increase of the contagious rate, as can be verified by using the interactive appliation.