Fighting the pandemic: Rapid detection of SarsCov-2

Itziar Heriz Iturbe
Tommaso Mannelli Mazzoli
Rafael Gutiérrez García
Elena Jorge Alcover
Marta Maria Paz Rodriguez
Antoni Van Hul Miralles

June 2020

Abstract

COVID-19 is the issue that all countries are facing during 2020. The humanitarian loss consequences and the economical impact of this pandemic has the best researchers all over the world working on how to deal with every aspect of the virus.

The aims and goals of this project is to develop computationally efficient and accessible tools for rapid detection of SarsCov-2 using existing testing and computational infrastructure; the tools should be available for large-scale deployment in both developed and developing countries with different funding models of their health systems. The approach considered in this project is to consider SarsCov-2 detection using the so-called Full Blood Count test – a standard analysis that is routinely performed in hospitals and haematological laboratories world-wide. This test can be performed in under 5 minutes for a single patient, and the technology is accessible to countries with different economic backgrounds and populational profiles. The task, therefore, is to investigate this possibility for develop an algorithm for rapid assessment of SarsCov-2 infection using the Full Blood Count test.

Also, during this project, different mathematical equations will be studied in order to model the pandemic and therefore, study the effect of decisions made by governments such as the quarantine, testing populations...

Contents

I Development of the problem. 4

1 Modelling 5
- 1.1 The SIR model . . . 5
- 1.2 Maximum number of infected people . . . 6
- 1.3 SIR model with quarantine effect . . . 7
- 1.4 SIR model and relation with tests . . . 7
- 1.5 SIR model with delay time . . . 10
- 1.6 SEIR model . . . 11
- 1.7 SEIR model and relation with tests . . . 12

2 Data analysis 13
- 2.1 Descriptive analysis of blood test analysis . . . 13
- 2.2 Discriminant analysis . . . 17
  - 2.2.1 Introduction . . . 17
  - 2.2.2 Geometric criteria . . . 17
- 2.3 Linear discriminant analysis applied to the data . . . 18
  - 2.3.1 LDA . . . 18
  - 2.3.2 LDA with a previous PCA . . . 19
  - 2.3.3 Simulations . . . 20
- 2.4 Delay . . . 20
- 2.5 Rashomon effect . . . 21

II Conclusions 22

Part I

Development of the problem.

Chapter 1

Modelling

1.1 The SIR model

The SIR model was developed by Kermack and McKendrick in 1927. It is a system of three ordinary differential equations in which the entire population (which is assumed to be constant and closed) is divided into three groups: Susceptibles, Infectious and Removed.

SIR model

$$\begin{cases} \dot{S} = -\beta IS \\ \dot{I} = \beta IS - \gamma I \\ \dot{R} = \gamma I \end{cases} \tag{1.1}$$

Where

- β is the *infection rate*
- γ is the *recovery rate*.

and

- S is the number of *susceptible*.
- I is the number of *infected*.
- R is the number of *removed*.

If we sum the right hand sides of equations in (1.1), we obtain $\dot{(}S + I + R) = 0$, so

$$S(t) + I(t) + R(t) = \text{constant} \tag{1.2}$$

We assumed that $S(t) + I(t) + R(t) = 1$ for every t. This way, the variables could be understood as proportions of population instead of total number of people.
We define the R_0 parameter as

$$\mathcal{R}_0 := \frac{\beta}{\gamma} S_0 \tag{1.3}$$

$\mathcal{R}_0$ gives a good approach to the initial behaviour of the pandemic:

- If $\mathcal{R}_0 > 1$ the infected people function increases and therefore, the pandemic spreads.
- If $\mathcal{R}_0 < 1$ the pandemic does not spread.

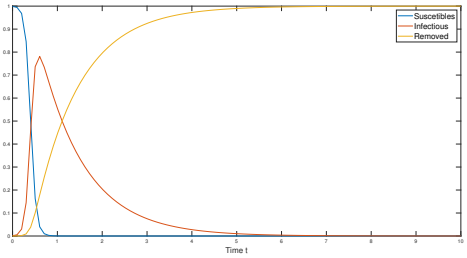
Furthermore, $\mathcal{R}_0$, known as "reproduction coefficient", measures the spread of the infection. In the one hand, if $\mathcal{R}_0$ takes a very large number, the infection will spread very fast. In the other hand, if $\mathcal{R}_0$ is greater than one but small enough, the infection rate will be such that the pandemic will spread slowly.
Calculating the infection rate can be easily done by making the derivative of I equal to 0):

$$\dot{I} = I(\beta S - \gamma) = 0 \implies \beta S - \gamma = 0 \tag{1.4}$$

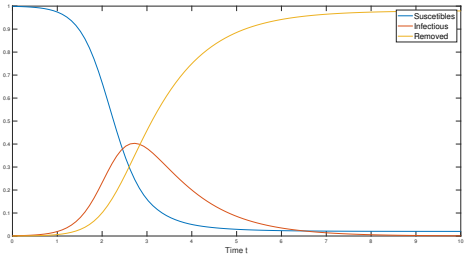
Since we are analysing the initial values, we can take S as $S_0 = S(0)$, getting

$$\frac{\beta S_0}{\gamma} - 1 = 0 \implies \mathcal{R}_0 = \frac{\beta S_0}{\gamma} \tag{1.5}$$

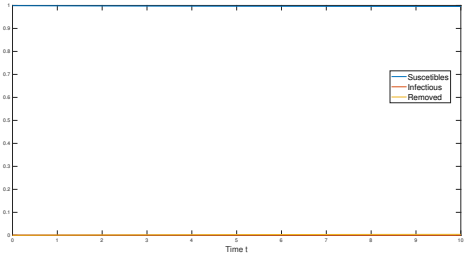
At GitHub MATLAB codes used for the project can be found.



(a) $\beta = 18$, $\gamma = 1$



(b) $\beta = 4$, $\gamma = 1$



(c) $\beta = 0.8$, $\gamma = 1$

Figure 1.1: Plot of I,R and S for $t \in [0, 20]$, $S(0) = 0.99$, $I(0) = 0.01$ and $R(0) = 0$

1.2 Maximum number of infected people

The maximum number of infected people as a function of $\mathcal{R}_0$ is

$$I_{\max}(\mathcal{R}_0) = I_0 + S_0 - \frac{S_0}{R_0}\Big(1 + \ln(\mathcal{R}_0)\Big) \tag{1.6}$$

We note that

$$\begin{aligned} \lim_{\mathcal{R}_0 \to 0^+} I_{\max}(\mathcal{R}_0) &= +\infty \\ \lim_{\mathcal{R}_0 \to +\infty} I_{\max}(\mathcal{R}_0) &= I_0 + S_0 \end{aligned} \tag{1.7}$$

If we differentiate, we obtain

$$I'_{\max}(\mathcal{R}_0) = \frac{S_0}{\mathcal{R}_0^2} \ln \mathcal{R}_0 > 0 \iff R_0 > 1 \tag{1.8}$$

In conclusion, the function grows if $\mathcal{R}_0 > 1$ and has a negative growth behaviour if $\mathcal{R}_0 \in (0, 1)$, so $I_{\max}(1) = I_0$ is the global minimum of the function.

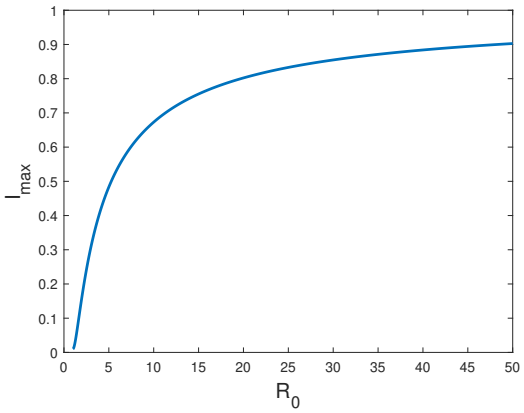


Figure 1.2: Plot of $I_{\max}$ for $R_0 \in (1, 50)$

1.3 SIR model with quarantine effect

So far, we have just gone through the basic model of any pandemic. However, many other factor could be added to better describe COVID-19. As a first approach, we consider the following variation of the SIR model, where the parameter α represents the fraction of the infected population that is put into quarantine. Notice that this model does not consider by which means this infected people are detected, it simply assumes that we are able to identify them.

SIR model with quarantine effect

$$\begin{cases} \dot{S} = -\beta IS \\ \dot{I} = \beta IS - \gamma I - \alpha I \\ \dot{R} = \alpha I + \gamma I \end{cases} \tag{1.9}$$

Where

- α is the proportion of infected people quarantined.

If we made α equal to 0, we would get our previous model, so this can be considered a generalization of SIR model.
To analyse this model, we begin with an analytical study of $\mathcal{R}_0$. $\mathcal{R}_0$ comes up to be the parameter that explains whether that the pandemic will increase from the beginning or not. As previously dine, we make $\dot{I} = 0$:

$$\dot{I} = 0 \iff \beta S - \gamma - \alpha = 0 \iff \frac{\beta S}{\gamma + \alpha} - 1 = 0 \tag{1.10}$$

which evaluated at the initial time gives

$$\frac{\beta}{\gamma + \alpha} S_0 - 1 = 0 \tag{1.11}$$

So, we have that

$$\mathcal{R}_0 = \frac{\beta}{\gamma + \alpha} S_0 \tag{1.12}$$

If $\mathcal{R}_0 = 1$, the pandemic reaches its maximum on the initial time and therefore the infected population decreases at every t.

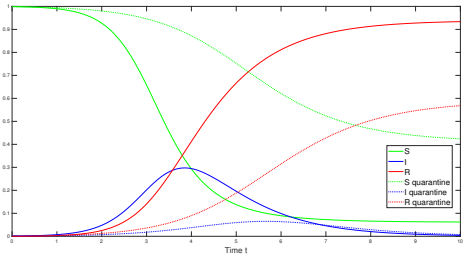


Figure 1.3: Solution with different values of α: $\alpha = 0.01$ (straight line), $\alpha = 0.99$ (dotted line)

1.4 SIR model and relation with tests

In the following variation of the SIR model we will take into account the effect of testing people. With this purpose, we are going to assume that the removed group is known to not be infected and for that reason, they won't be tested. Then, only a fraction of the infected ad susceptible population will be tested and put into quarantine in case they turn out to be positive in the test.

SIR model with false positive and false negative rate

$$\begin{cases} \dot{S} = -\beta IS - \alpha_S S \\ \dot{I} = \beta IS - \gamma I - \alpha_I I \\ \dot{R} = \gamma I + \alpha_S S + \alpha_I I \end{cases} \tag{1.13}$$

- FPR : *false positive rate*. For the COVID-19 test $FPR \approx 0.2553191$.
- FNR : *false negative rate*.For the COVID-19 test $FNR \approx 0.03846154$.
- α_S: proportion of susceptible people that are quarantined.
- α_I: proportion of infected people that are quarantined.

In this case the $\mathcal{R}_0$ parameter can be defined as

$$\mathcal{R}_0 = \frac{\beta}{\alpha_I + \gamma_I} S_0 \tag{1.14}$$

We define the rates of false positives and false negatives as:

$$FPR = \frac{FP}{FP + TN} \tag{1.15}$$

$$FNR = \frac{FN}{FN + TP} \tag{1.16}$$

where:

- FP is the number of *false positive* tests (people who are not infected but the test turns out positive).
- TN is the number of *true negative* tests (when someone is not infected, and the test say so).
- TP is the number of *true positive* tests (when someone is infected, and the test says so).
- FN is the number of *false negative* tests (when someone is infected, and test results says they are not).

We consider a "medical quarantine", which means that people who are in quarantine have been tested. So indeed, we are defining the value of α.
As explained before, we are assuming that the removed group is formed by people who cannot be infected, so the number of people who are tested is

$$\sigma(S + I), \tag{1.17}$$

where σ is the proportion of people tested.
We can now focus on studying the infected and susceptible groups.
The number of susceptible people who goes into quarantine are those whose test result are false positives, and this number can be easily calculated by using the σ and the FPR rates.
FP, and TN can only come from people who are not really infected (susceptible group). So, $TN + FP$ comes up to be the number of people in S who have been tested.
Now, using the (1.15) we can confirm that the S people that had a false positive result are $\sigma \cdot FPR \cdot S$ and therefore

$$\alpha_S = \sigma \cdot FPR \tag{1.18}$$

The same analysis can be applied to the infected people group. The number of infected people who are tested negative (FNR) is also the infected people who does not go into quarantine. Now, if α_I is the proportion of infected people that have accurate results (true positives), $\sigma - \alpha_I$ would be the proportion of infected people who had wrong results (false negatives). So, using (1.16):

$$\sigma - \alpha_I = \sigma \cdot FNR \implies \alpha_I = \sigma\,(1 - FNR) \tag{1.19}$$

In conclusion, the number of infected people that have false negative results is $I\sigma\,(1 - FNR)$.

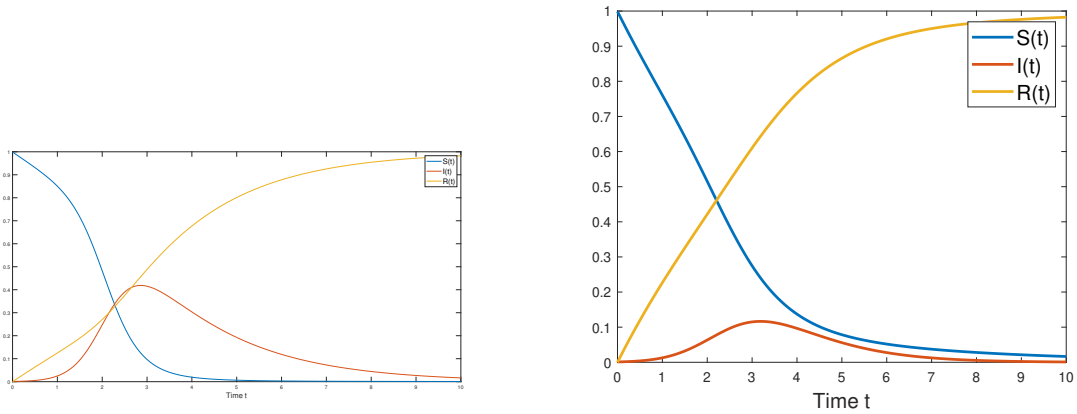
Finally, we offer the analytical solution of this problem, simplified by assuming that $\alpha_s = 0$.

SIR model with false positive and false negative rate

$$\begin{cases} \dot{S} = -\beta IS \\ \dot{I} = \beta IS - I(\gamma + \alpha_I) \\ \dot{R} = I(\gamma + \alpha_I) \end{cases} \tag{1.20}$$

If we define a primitive $\xi(t)$ of $\beta I(t)$, we can write

$$\xi(t) := \beta \int_0^t I(\hat{t})\,\mathrm{d}\hat{t} \tag{1.21}$$



(a) $\sigma = 0.01$ (b) $\sigma = 0.99$

Figure 1.4: Solutions with $\gamma = 1$, $\beta = 4$ and different values of σ

Firstly, we will assume that $R(0) = 0$. So that, $S_0 = 1 - I_0$ (being N the total population). If we take the fist equation of the system we get

$$\dot{S} + \beta IS = 0 \tag{1.22}$$

Now, using the previous ξ expression, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(S(t)\,\mathrm{e}^{\beta\int_0^t I(\hat{t})\,\mathrm{d}\hat{t}}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(S(t)\,\mathrm{e}^{\xi(t)}\right) = 0 \tag{1.23}$$

So $S(t)\,\mathrm{e}^{\xi(t)}$is constant and therefore is equal to $S(0)\,\mathrm{e}^{\xi(0)}$

$$S(t)\,\mathrm{e}^{\xi(t)} = S(0) \implies S(t) = S(0)\,\mathrm{e}^{-\xi t} = (N - I_0)\,\mathrm{e}^{-\xi t} \tag{1.24}$$

Taking the definition of ξ we know that $\dot{\xi} = \beta I$.
Furthermore, we have that $\xi(0) = 0 = R(0)$, and:

$$\dot{R} = I(\gamma + \alpha_I) \tag{1.25}$$

So, that means that

$$R = \frac{\dot{\xi}}{\beta}(\gamma + \alpha_I) \tag{1.26}$$

Finally, since $I + R + S = 1$, we get

$$I = 1 - R - S = 1 - \frac{\dot{\xi}}{\beta}(\gamma + \alpha_I) - S_0\,\mathrm{e}^{-\xi(t)} \tag{1.27}$$

So, this results in the following differential equation and its initial condition.

$$\begin{cases} \dot{\xi} = \beta I = \beta\left(1 - \frac{\dot{\xi}}{\beta}(\gamma + \alpha_I) - S_0\,\mathrm{e}^{-\xi(t)}\right) \\ \xi(0) = 0 \end{cases} \tag{1.28}$$

Now, a final suggestion could be applied to all the models shown above. A country might not afford having too many people into quarantine (due to the economical impact that this may cause). Any country in this situation, could try to get the optimal rate of citizens that should be quarantined. With this purpose we could implement a new model that introduces a new group in the system.

SIR model with false positive and false negative rate

$$\begin{cases} \dot{S} = -\beta IS - \alpha_S S \\ \dot{I} = \beta IS - \gamma_I I - \alpha_I I \\ \dot{Q} = \alpha_I I + \alpha_S S - \gamma_Q Q \\ \dot{R} = \gamma_I I + \gamma_Q Q \end{cases} \tag{1.29}$$

Where

- Q is the number of people quarantined.

Furthermore,we must consider that the recovery rate is different for infected people who are in quarantine and those who are not in quarantine. So,

- γ_I is the rate of infected people that were not in quarantine and go to the removed group.
- γ_Q is the rate of infected people that were in quarantine and go to the removed group.

1.5 SIR model with delay time

Delay-differential equations (DDE's) are a really important tool in dynamical systems. They usually appear in engineering problems such as technological control problems, where there is a delay between the observations and the adjustments. They also appear frequently in pandemic models.
There are different types of DDE's, but we are going to focus on the delay-differential equation with only one constant of delay. Its general form is the following.

$$\dot{\boldsymbol{x}} = \boldsymbol{f}\left(\boldsymbol{x}(t),\, \boldsymbol{x}(t-\tau)\right), \tag{1.30}$$

where $\tau \in \mathbb{R}$ is the delay time.

SIR model with delay time

$$\begin{cases} \dot{S}(t) = -\beta\, I(t)\, S(t) \\ \dot{I}(t) = \beta\, I(t)\, S(t) - \gamma_I\, I(t) - \alpha\, I(t-\tau)\, \operatorname{sgn} I(t) \\ \dot{R}(t) = \gamma_I\, I(t) + \alpha I(t-\tau)\, \operatorname{sgn} I(t) \end{cases} \tag{1.31}$$

Where

- τ is the delay time, the time that happens between the beginning of infectiousness and the final diagnosis.

and $\operatorname{sgn} a$ represents the function defined by

$$\operatorname{sgn}(a) = \begin{cases} 1 & \text{if } a > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{1.32}$$

As we have done with the previous models, we are going to study $\mathcal{R}_0$. In order to do that, we will use the Taylor series of $I(t-\tau)$ around t, so we get

$$I(t-\tau) = I(t) - \tau \dot{I}(t) + \frac{\tau^2}{2}\ddot{I}(t) + \mathcal{O}(\tau^3). \tag{1.33}$$

Taking the first order term we can easily obtain an approximation of $\mathcal{R}_0$, i.e.,

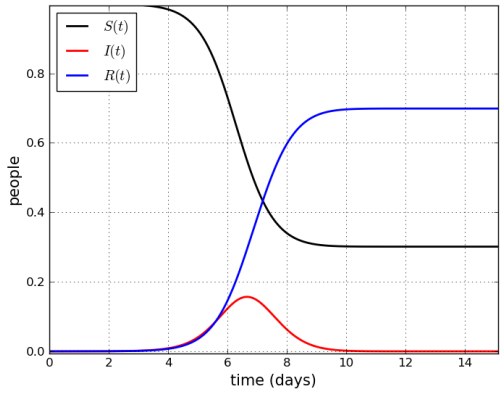
$$\dot{I}(t) = \beta I(t) S(t) - \gamma_I I(t) - \alpha(I(t) - \tau \dot{I}(t)) \tag{1.34}$$

$$\dot{I} = I\left(\frac{\beta S - \gamma_I \alpha}{1 + \alpha\tau}\right). \tag{1.35}$$

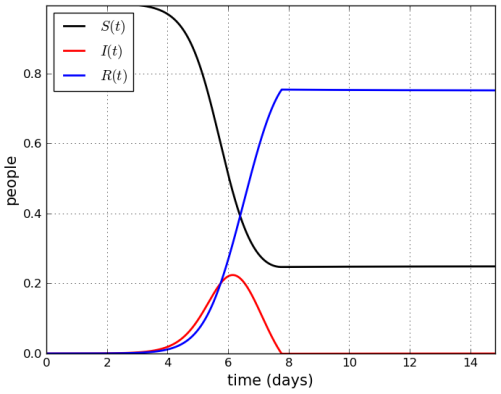
Using the previous analysis we obtain the same result as in the previous model.

$$\mathcal{R}_0 = \frac{\beta}{\gamma_I + \alpha} S_0, \tag{1.36}$$

Some simulations are shown now, depending on the value of τ.



(a) $\tau = 0.5$



(b) $\tau = 1$

The longer the delay time is, the higher the infected rate, $I_{\max}$, gets. That could be an expected result, because during that delay period, during which the infected person does not know that they are contagious, they carry the virus around and they are bound to infect more people.

1.6 SEIR model

It is very important in order to make a realistic model to consider the proportion of people who are infected, but still don't know it because they have not developed any symptom. In order to do this, we introduce a new variable E. This group is known as the "exposed" group and the resulting system is

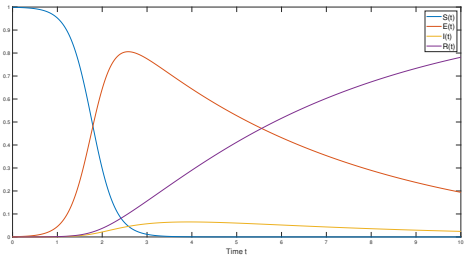
SEIR model

$$\begin{cases} \dot{S} = -\beta\, I\, S - \beta E\, S \\ \dot{E} = \beta(E + I)S - aE - \gamma_E E \\ \dot{I} = aE - \gamma I \\ \dot{R} = \gamma I + \gamma_E E \end{cases} \tag{1.37}$$

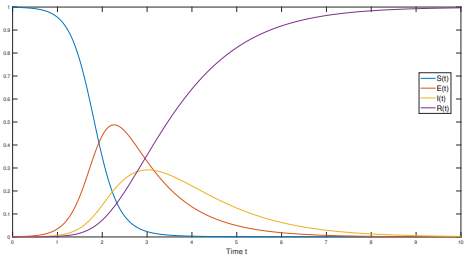
Where

- a is the rate of people who go from exposed to infected.
- γ_E is the rate of exposed people that go to removed group because they did not develop any symptom.

γ and γ_E are different rates since a person who has the infection without symptoms is bound to recover earlier than a person who has symptoms.



(c) $a = 0.01$



(d) $a = 0.99$

Figure 1.5: SEIR model for different values of a

Once again we focus on the $\mathcal{R}_0$ value. Now the only variables related to $\mathcal{R}_0$ are E and I. The matrix of the subsystem formed by I and E is:

$$\begin{pmatrix} \beta S - (a + \gamma_E) & \beta S \\ a & -\gamma \end{pmatrix} \tag{1.38}$$

Note that for people in group E that go to group I before they finally recover and go to R, $\gamma_E = \gamma$. So,

$$\dot{E} + \dot{I} = \beta(E + I)S - \gamma(E + I) \tag{1.39}$$

In this special case, we could consider another new variable called Z that would be representing $E + I$, getting the following:

$$\dot{Z} = \beta ZS - \gamma Z \tag{1.40}$$

Repeating the process done in previous cases, and we obtain that

$$\dot{Z} = 0 \iff \beta ZS_0 - \gamma Z = 0 \iff Z(\beta S_0 - \gamma) = 0 \iff \mathcal{R}_0 = \frac{\beta}{\gamma} S_0 \tag{1.41}$$

That is actually an easier case than the one we are interested in.
Going back to the Jacobian matrix of the linearized dynamic subsystem formed by E and I: (1.38).
Exponential growth occurs when one eigenvalue of (1.38) is positive. So, since the independent term of the characteristic polynomial the determinant of (1.38); then, we get the following sufficient condition.

$$(a + \gamma_E - \beta S)\gamma - a\beta S < 0 \iff a\beta S - (a + \gamma_E - \beta S)\gamma > 0 \tag{1.42}$$

$\mathcal{R}_0$ is consequently defined as

$$\mathcal{R}_0 := \frac{a\beta S_0 + \beta S_0 \gamma}{a\gamma + \gamma\gamma_E}. \tag{1.43}$$

1.7 SEIR model and relation with tests

SEIR model with quarantine (after test)

$$\begin{cases} \dot{S} = -\beta\, I\, S - \beta E\, S - \alpha_S S \\ \dot{E} = \beta(E + I)S - aE - \gamma_E E - \alpha_E E \\ \dot{I} = aE - \gamma I - \alpha_I I \\ \dot{R} = \gamma I + \gamma_E E + \alpha_S S + \alpha_E E + \alpha_I I \end{cases} \tag{1.44}$$

Where:

- α_S is the rate of susceptible people that are put into quarantine (so, they got FP results).
- α_E is the rate of exposed people that are put into quarantine (so, they got TP results).
- α_I is the rate of infected people that are put into quarantine (so, they got TP results).

In this case, the σ parameter is split into two different parameters.

- σ_1 is the proportion of people without any symptom that are tested. So, $\sigma_1(S + E)$ is all this group that is tested.
- σ_2 is the proportion of infected people (people with symptoms) that are tested, $\sigma_2 I$.

This seems more realistic because the health system surely does not act the same way with people who seem to be infected and with those who seem healthy.
Then $\sigma_1(E + S)$ are asymptomatic people who are tested. And $\sigma_2 I$ are symptomatic people who are tested.
Now, let's look at the α_I, α_S and α_E new expressions. α_S does not change at all. Using (1.15) we have that the proportion of people in group S who get false positive results are:

$$\sigma_1 \cdot E \cdot FNR \tag{1.45}$$

$$\alpha_S = \sigma_1 \cdot FPR \tag{1.46}$$

Now, if α_E is the fraction of infected asymptomatic people who are quarantined; then it is also the fraction of this group who got TP results in their tests. Using (1.16):

$$\sigma_1 - \alpha_E = \sigma_1 FNR \implies \alpha_E = \sigma_1(1 - FNR) \tag{1.47}$$

Finally, to get α_I, the procedure is similar to the one previously done.

$$\sigma_2 - \alpha_I = \sigma_2 FNR \implies \alpha_I = \sigma_2(1 - FNR) \tag{1.48}$$

So, in conclusion, we have that:

$$\begin{cases} \alpha_S = \sigma_1 FPR \\ \alpha_E = \sigma_1(1 - FNR) \\ \alpha_I = \sigma_2(1 - FNR) \end{cases} \tag{1.49}$$

Chapter 2

Data analysis

Once we have studied different models that could be applied to the pandemic, and the importance of tests has been proved, the objective of the project is to develop a cheap test based on the results of a simple blood exam. First of all, we begin by analyzing the data we are going to work with. This data shows the results of blood tests conducted over a total of 2048 patients in a hospital of Sao Paolo, Brazil. This means we start with 2048 observations, and 39 variables for all of them. This variables are divided in two groups.

- **Blood test variables.** This variables include 14 parameters including the amount of leukcytes, eosinphils and other blood cells.
- **Virus test variables.** A total of 24 binary variables that show if patients turned out positive or negative in virus tests such as Influenza, Endovirus, etc...

The remaining variable is a 'target' variable since it shows whether the patient is positive or negative in SarsCov-2. The objective of the data analytics in the problem is to provide the model with some reliable parameters based on the data of this 2048 observations. Secondly, in order to make proper predictions, it is necessary to previously process the data. This means that any 'NaN' register should be somehow treated, either by deleting the observation, or imputing data if the information could be obtained. When deleting all the observations that contained any missing value, the total amount of data was reduced to less than 100 observations. In order to maintain a large enough set of data in order to make predictions, the procedure was the following.

1. Remove all the binary variables, most of them, virus test variables.
2. With the remaining data, proceed to exclude all the observations containing any 'Nan'.
3. This resulted into 598 observations, which we considered a much better number of total observations.
4. Conducted a LDA, with a previous PCA, and also without this previous PCA.

2.1 Descriptive analysis of blood test analysis

The following tables show the different characteristics of the variables we are studying. It might be of special interest to point out that the proportion of ones in the target variable is much smaller than the proportion of zeros. This must be taken into account when conducting any data analysis since some methods could leads us to predictions of few 'positives', which for sure is going to give a small error. However, the false negatives are something we should try to avoid due to the importance of not considering any infected person as healthy.

The FREQ Procedure

SARS_Cov_2_exam_result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	517	86.45	517	86.45
1	81	13.55	598	100.00

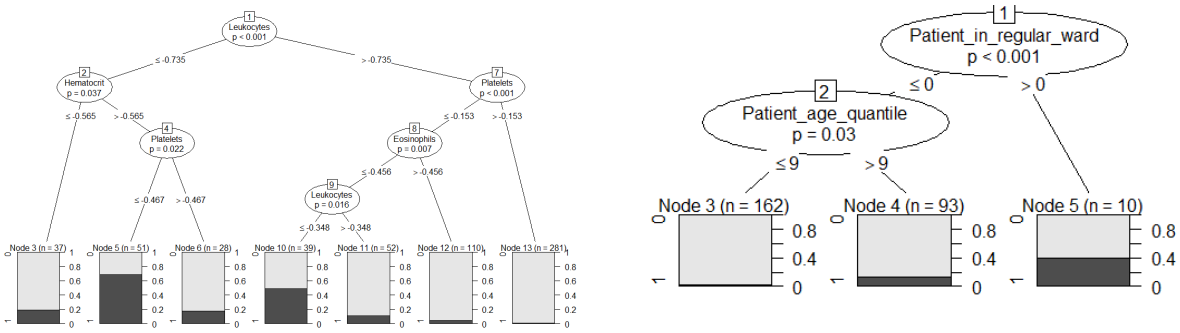
Figure 2.1: Target variable

In Figure 2.3a we can see a Decision Tree about Blood Test.
As you can see in Figure 2.4, the most important variables are Leukocytes, Eosinophils and Platelets in that order. This explains that the same variables are important nodes in the decision tree.
Even if what we used for predictions were only the blood test variables, it is also interesting to consider the effect of the virus tests. Therefore, we also proceed to analyse this part of the data.

The MEANS Procedure

Variable	N	N Miss	Mean	Median	Std Dev	Minimum	Maximum
F1	598	0	1147.58	1234.00	661.5776535	0	2047.00
SARS_Cov_2_exam_result	598	0	0.1354515	0	0.3424917	0	1.0000000
Hematocrit	598	0	0.0030748	0.0534070	1.0024288	-4.5014195	2.6627038
Hemoglobin	598	0	0.0056351	0.0403160	1.0012546	-4.3456030	2.6718678
Platelets	598	0	0.0065699	-0.1091542	0.9947519	-2.5524261	9.5320339
Mean_platelet_volume	598	0	-0.000017846	-0.1015171	1.0016735	-2.4575746	3.7130520
Red_blood_Cells	598	0	-0.0094385	0.0138521	0.9878378	-3.9706082	3.6457062
Lymphocytes	598	0	-0.000817384	-0.0142670	1.0025351	-1.8650696	3.7640996
Mean_corpuscular_hemoglobin_conc	598	0	0.0100232	-0.0545852	0.9919569	-5.4318085	3.3310707
Leukocytes	598	0	0.0066731	-0.2087048	1.0003084	-2.0203025	4.5220418
Basophils	598	0	0.000985958	-0.2237665	1.0029573	-1.1401438	11.0782194
Mean_corpuscular_hemoglobin	598	0	0.0229335	0.1259032	0.9505376	-5.9376040	4.0985460
Eosinophils	598	0	0.0039679	-0.3298351	1.0027817	-0.8355077	8.3508759
Mean_corpuscular_volume	598	0	0.0217649	0.0760592	0.9549945	-4.5808120	3.4109800
Monocytes	598	0	-0.0036822	-0.1151911	0.9980545	-2.1637213	4.5333972
Red_blood_cell_distribution_widt	598	0	-0.0193383	-0.1827903	0.9696220	-1.5980943	6.9821839

Figure 2.2: Blood test variables



(a) Decision Tree Blood Test

(b) Decision Tree Virus Test

Figure 2.3: Decision Trees

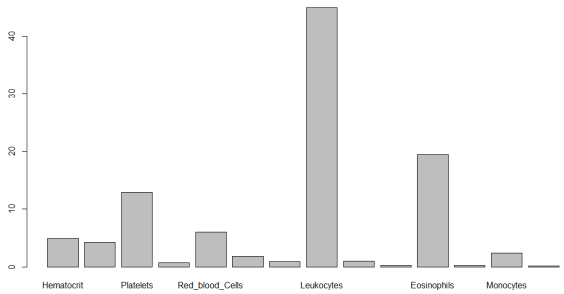


Figure 2.4: Importance of the variables

In Figure 2.5 we can compare the number of records with and without COVID. The Table 2.1 is a contingency table with frequency of zeros and ones, that is, if the patients turned out positive or negative in different virus test. As you can see, all the variables are binary.

SARS_Cov_2_exam_result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	246	92.83	246	92.83
1	19	7.17	265	100.00

Figure 2.5: Target variable

Variable	Frequency of 0s	Frequency of 1s
Patient_in_regular_ward	255	10
Patient_in_semi_intensive_unit	247	18
Respiratory_Syncytial_Virus	247	18
Influenza_A	259	6
Influenza_B	237	28
Parainfluenza_1	264	1
CoronavirusNL63	261	4
Rhinovirus_Enterovirus	188	77
Coronavirus_HKU1	263	2
Parainfluenza_3	265	0
Chlamydophila_pneumoniae	264	1
Adenovirus	260	5
Parainfluenza_4	260	5
Coronavirus229E	264	1
CoronavirusOC43	262	3
Inf_A_H1N1_2009	228	37
Bordetella_pertussis	265	0
Metapneumovirus	262	3
Influenza_B_rapid_test	250	15
Influenza_A_rapid_test	246	19

Table 2.1: Virus test variables

The figure 2.3 is a Decision Tree. The most important variables seem to be the age and having been in a regular ward, and consequently, this allows to classify the data.
To end this chapter, we wanted to comment some interesting graphics we got. In Figure 2.7 we can see that the COVID-19 is high related with the age of the patient.
In addition, Figure 2.8 shows that the percent of patients with COVID who have remained in critical unit is greater than the percent of the patients who have stayed in critical unit but not because of COVID.

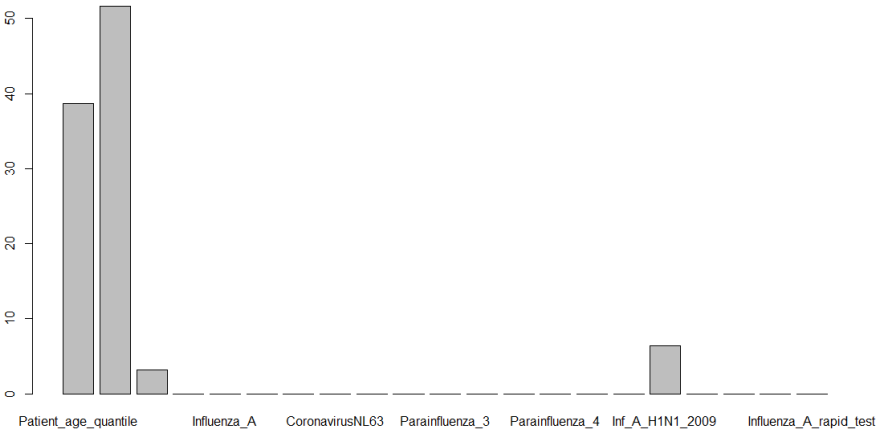


Figure 2.6: Importance of the variables

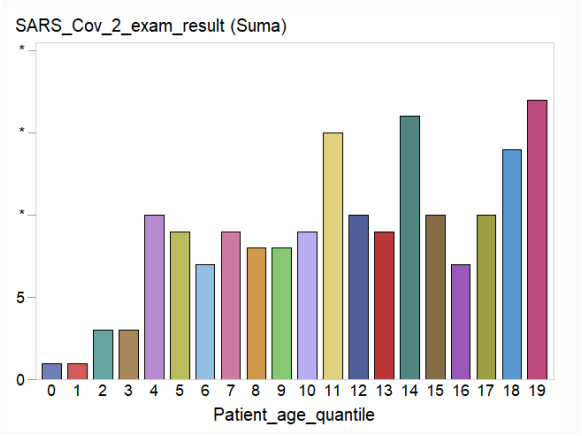


Figure 2.7: Patient age

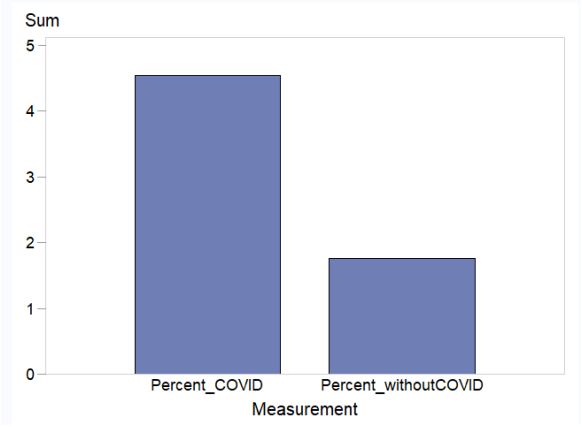


Figure 2.8: Patient in critical unit

2.2 Discriminant analysis

2.2.1 Introduction

Discriminant analysis starts with quantitative data of individuals that are going to be classified into $k \geq 2$ groups. It is about partitioning the space of individuals in k parts, in such a way that we will assign a new individual to group j, if the vector corresponding to its evaluation belongs to part j. The membership in one or the other group is introduced into analysis by using a categorical variable which takes as many values as existing groups. In the discriminant analysis this variable plays the role of dependent variable.

To begin, we will need to decide on the most appropriate type of separation function. Linear functions are the most frequently used, so we will use this type of separation to analyse our problem.

Suppose that an individual w belongs to one of k possible groups $G_1, \ldots, G_k$ of population P, and also, we have the following information:

1. The values of the n individuals in the population take about p variables $x_1, \ldots, x_p$ in addition to the group to which each one belongs (n_j being the size of each group).
2. Information of an individual w in respect to the same variables.

We are going to obtain a decision rule that allows assigning w to one of the groups. In general, the classification is solved by constructing a function of the variables $F = g(x_1, \ldots, x_p)$, called the discriminant function. The group to which each individual belongs will be decided from the value that the function takes.

There are different criteria to obtain the discriminant functions that separate a k sets. In this chapter we will see the approach and development of some of them.

2.2.2 Geometric criteria

This classification criterion is very intuitive since it consists of assigning w to the closest group. To find the distances between the individual and a set of data we must take into account that:

1. Each group is represented by its centroid or mean vector.
2. The distance to be used must be that of Mahalanobis, since we have not assumed that we have incorrect and standardized variables.

The Mahalanobis distance, represented by D^2, was proposed by this author in 1936 and is a generalization of the Euclidean distance, which takes into account the matrix of intragroup covariances.

Let $G_1, \ldots, G_k$ k groups of $n_1, \ldots, n_k$ individuals each, respectively. For each individual p variables are known, $x_1, \ldots, x_p$. A matrix of observations corresponds to each group, so there are k matrices, each one of dimension $n_j \times p$, $j = 1, \ldots, k$. From this data, the Mahalanobis distance between the individual w of coordinates $x = (x_1, \ldots, x_p)$ and the group G_j is given by the formula:

$$D^2(w, G_j) = (x - \mu_j)' V^{-1} (x - \mu_j).$$

Being μ_j the column vector that contains the means of the considered variables of group j and V the intragroup variance-covariance matrix, that is

$$V = \frac{\sum_{j=1}^{k} n_j V_j}{\sum_{j=1}^{k} n_j}$$

where V_j covariance matrix of j group.

The Euclidean distance is the particular case of the Mahalanobis distance in which $V = I$. Thus, the Euclidean distance does not take into account the dispersion of variables or the relationships between them, while the distance of Mahalanobis does consider it when introducing the inverse of the matrix into his calculation of intragroup covariances.

If we consider that we only have two groups, the decision rule will be:

$$\begin{cases} w \in G_1 & \text{if } D^2(w, G_1) - D^2(w, G_2) < 0 \\ w \in G_2 & \text{in other case} \end{cases}$$

Doing some maths,

$$D^2(w, G_1) - D^2(w, G_2) = (\mu_1 - \mu_2)' V^{-1}(\mu_1 + \mu_2) - 2(\mu_1 - \mu_2)' V^{-1} x$$

Taking $\overline{\mu} = \dfrac{\mu_1 + \mu_2}{2}$, the discriminant hyperplane will be:

$$H \equiv (\mu_1 - \mu_2)' V^{-1} \overline{\mu} - (\mu_1 - \mu_2)' V^{-1} x = 0.$$

Usually, mean vectors and covariance matrix are unknown, being replaced by unbiased estimates; the sample mean vectors $\overline{x}_1$ y $\overline{x}_2$ and by

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

respectively, being S_1^2 and S_2^2 the sample variance-covariance of each group.

2.3 Linear discriminant analysis applied to the data

For this analysis two different tests were made. One with the data variables as input and another one with the principal components as input.

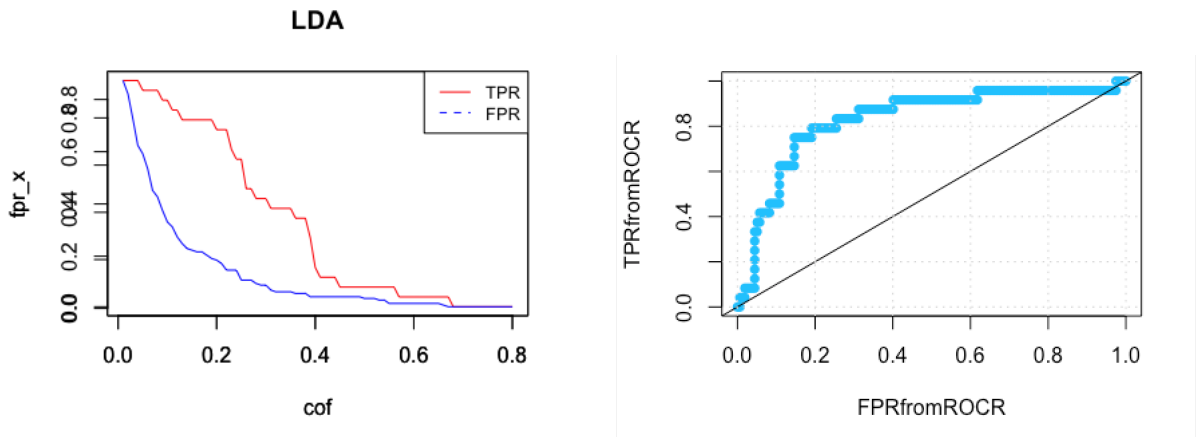
2.3.1 LDA

The first step to be able to make predictions is to divide our data set into *training* and *test*. The training set will be used to train the data we have and create the LDA algorithm that will later on be used to test the rest of the data. The proportion chosen for this data division in 70% for training and 30% for testing. Next, the

Data	
df_bt	598 obs. of 15 variables
testing	176 obs. of 15 variables
training	422 obs. of 15 variables

Figure 2.9: Proportion of training training set and test set

discriminant analysis is developed with the training set. For this purpose, we have used *Rstudio*. It is possible to adjust different parameters in order to obtain the maximum number of true positives and the minimum of false positives. This means, our objective is to maximize the distance between these rates. There is an important parameter when it comes to LDA, the *cutoff*, which indeed is related to this maximization. The optimal cutoff provides us with the maximum distance between the false positive rate and the false negative rate. Having said that, our first task is to obtain the *confusion matrix* for different cutoffs, and after that, the optimal cutoff will be chosen.



(a) True positive rate and false positive rate

(b) ROC curve

Figure 2.10

In addition, it is relevant to show the ROC curve, and in this case the area that remains under the curve is 0.8288.

As a result, we obtain the optimal cutoff, which we could have easily guessed by the graph of the rates as 0.22. This cutoff provides us with a final confusion matrix. In this matrix, the false negative rate is of high interest since we do not want our test to classify as not infected any infected individual.

		Predicted	
		0	1
Real Data	0	134	23
	1	6	18

Table 2.2: Confusion matrix

And from this results we can calculate the false positive rates and the true positive rates. However, for more accurate results, we are going to randomly select the training data more times, up to 100 different sets, and get the rates from this simulations. We will go deeper into this point in a following chapter.

2.3.2 LDA with a previous PCA

The difference between this analysis and the previous one is that, in this case, we are firstly going to group all the variables into the *principal components*. The principal components analysis, from now on PCA, tries to put variables that could be somehow related into the same group. This way, the variables are cut down into components that describe different aspects of the data. A good example that could be applied to our blood test data, is the separation of red cells and white cells. The white cells, related to the immune response of the body, could probably have an important weight in some components. Having done this previous PCA, the components come up to be the new input to our LDA, and the following procedure is just the same as before.

PCA

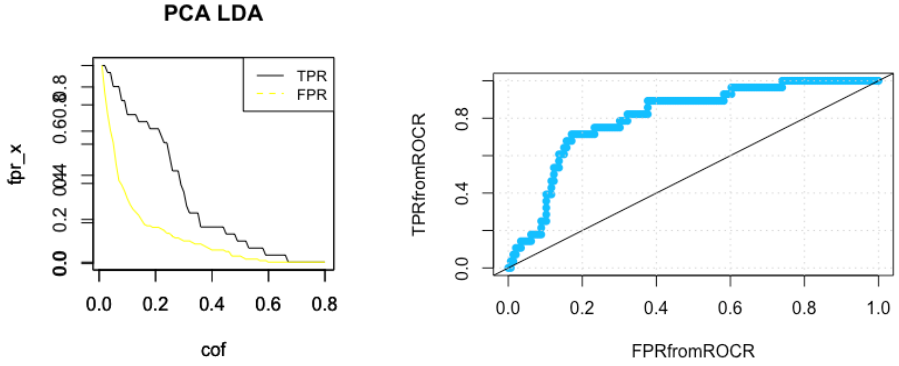
Once again, RStudio offer the results of the PCA, and here it is important to make the decision of which components should be taken. We are trying to take the components that explain, at least, 85% of the variance, and that means choosing the 8 first principal components.

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6    PC7
Standard deviation     1.7739 1.5591 1.3252 1.2265 1.06597 0.97615 0.9150
Proportion of Variance 0.2248 0.1736 0.1254 0.1075 0.08116 0.06806 0.0598
Cumulative Proportion  0.2248 0.3984 0.5238 0.6313 0.71245 0.78051 0.8403

                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.83085 0.79443 0.76190 0.57381 0.05469 0.03235 0.01973
Proportion of Variance 0.04931 0.04508 0.04146 0.02352 0.00021 0.00007 0.00003
Cumulative Proportion  0.88962 0.93470 0.97617 0.99968 0.99990 0.99997 1.00000
```

Figure 2.11: Results of the PCA

Now, our data set, consists of the same number of observations, but with the principle components as variables. We are now ready to conduct the LDA.



(a) True positive rate and false positive rate with PCA

(b) ROC curve with the previous PCA

Now, the area under the ROC curve is 0.803816.
In this case, the optimal cutoff turns out to be 0.17, and the resulting confusion matrix is the following one.

	Predicted		
Real Data		0	1
	0	121	25
	1	8	20

Table 2.3: Confusion matrix

2.3.3 Simulations

To get the previous results, we have fixed a seed which is used to split the dataset into a training set and a validation set. Using a any other seed will lead us to slightly different results. That is why we decided to make 100 simulations.
When we got all the simulations, our goal was to minimize the False Negative Rate(FNR) because in this context, we aim to avoid any misclassification of an infected person.
The results are the following:

$$TPR(\text{True Positive Rate}) = TP/(TP + FN) = 0.9615385$$

$$FPR(\text{False Positive Rate}) = FP/(FP + TN) = 0.2553191$$

$$FNR(\text{False Negative Rate}) = FN/(FN + TP) = 0.03846154,$$

where TP is the number of True Positives, FN is the number of False Negatives, TN is the number of True Negatives and FP is the number of False Positives.
However, we also though that other crieria, like taking the mean of the rates could be better. The results were the following.

$$TPR(\text{True Positive Rate}) = TP/(TP + FN) = 0.8288872$$

$$FPR(\text{False Positive Rate}) = FP/(FP + TN) = 0.2103967$$

$$FNR(\text{False Negative Rate}) = FN/(FN + TP) = 0.1711128,$$

2.4 Delay

Our purpose in this chapter is to study the *delay*. The *delay* is the time window between test and results. To get a good estimation, we calculate how many times we have to do the test in order to get a false negative rate close to zero. Assuming independence, we can calculate the chance of not detecting an infected person as follows: FNR^{times}, where *times* are the times the test is made.

We got the following graphic in which we can see that our *delay* is three units of time, because from this point on, the curve stabilizes.

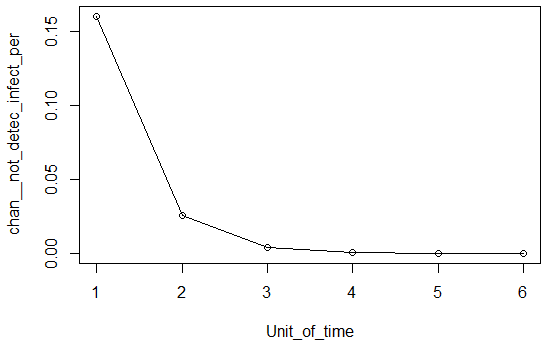


Figure 2.12: *Delay*

2.5 Rashomon effect

In this chapter we have taken into account the *Rashomon effect*, which means, we have done an exhaustive analysis of the principal components in order to choose the ones that give better results when it comes to minimizing the false negative rates. With this purpose we consider all the possible combinations of principal components, such that, they explain at least 85% of the variance. Once we have all this lists of principal components, we proceed to do the LDA with these PCA and choose the best of all results. In this case, when talking about the best result, we are referring to the one that offers the lowest FNR. Particularly, there are 4 combinations of principal components that give a $FNR = 0$. Therefore, to compare them, we analyse the FPR that this combinations of principal components offer, and choose the lowest one. The final principal components obtained in this search of optimality is $1, 2, 4, 5, 6, 7, 8, 9, 10, 11$.
Once we have the principal components we are going to use, we can actually run a 100 simulations with different seeds and calculate the mean of the rates. These are the final results.

$$TPR(\text{True Positive Rate}) = TP/(TP + FN) = 0.8427493$$

$$FPR(\text{False Positive Rate}) = FP/(FP + TN) = 0.2236289$$

$$FNR(\text{False Negative Rate}) = FN/(FN + TP) = 0.1572507$$

However, the variance is not always the best indicator of which components must be chosen. Therefore, we also tried all the possible combinations of principal components, even those that did not get to explain 85% of the variance. In this case, the optimal combination of principal components was $1, 3, 4, 11$. To sum up, we attach this summary table.

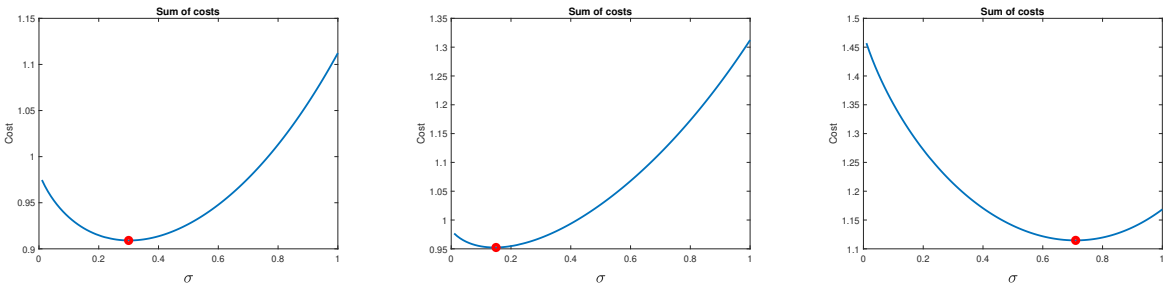
Models	**TPR**	**FPR**	**FNR**
LDA	0.8288872	0.2103967	0.1711128
PCA-LDA	0.8400366	0.2025678	0.1599634
PCARash(>= 85%)-LDA	0.8427493	0.2236289	0.1572507
PCARash-LDA	0.8157532	0.2763271	0.1842468

Table 2.4: Final models

Part II

Conclusions

The aim of this project was to create a cheap test that could lead to reliable results in order to correctly model the pandemic in any country and therefore, make recommendations for some of the most important sanitary and economical decisions.
With this objective, we focus on the optimal proportion of tests we should make. The aim is to consider a human cost for the people infected and the economic cost of having people in quarantine.



(a) Costs equally valued (b) Priority on saving money (c) Priority saving lifes

Figure 2.13a shows the optimal number of tests that should be done in a country in which the government tries to give the same importance to human loss and to economical impact. We should make tests less than 40% of the population.
The second case could be representing a country whose priority is to save money. This government would try to test a small amount of the population and consider that enough in order to stop the pandemic without a much bigger economical impact. In Figure 2.13b we can see that the recommendation would be to test less than the 20% of the population.
In the last case, in Figure 2.13c the priority is to save lives. The government would accept to do as many tests as recommended no matter what the costs would be. In this case, we should make our test to 70% of the population.
These results are quite useful because they allow us to make a good decision depending on the priorities of a country.
It is interesting to point out that the accuracy of the test we are explaining in this project is high enough, 84% of TPR. Not only the results are good, but also, the test is cheap (between 30 and 60 pounds) and could be done anywhere since it is based in a simple blood exam. It is not a test just for developed countries, it is a test that makes us all equal.

To end this project, we would like to present same open questions and future projects that could be applied to get better results.

First of all, our database was to small and based on a single hospital. It is well known that the virus has had a very different behaviour in every country, so reliable results would require a wider range of data.

For the modelling part delay time could be added to SEIR model. In addition, it could be interesting to study the effect of putting every symptomatic person into quarantine and testing only some of the asymptomatic population.

As we can see in Figure 2.7 the age quantile of patients with COVID is high, which means that the number of people who need to be hospitalized having COVID is related to the patients age. One of the biggest problems in this pandemic has been hospitals space and lack of material, so taking into account the clear correlation between ages and hospitalization, it could be recommendable to repeat this project considering different ages.

Bibliography

[1] Brugnano Luigi, Iavernaro Felice (2020), *A multi-region variant of the SIR model and its extensions*, ArXiV;

[2] Miller C. Joel (2016), *Mathematical models of SIR disease spread with combined non-sexual and sexual transmission routes*;

[3] Jones H.James (2007), *Notes on* $\mathcal{R}_0$, Standford University;

[4] Roussel R. Marc(2005), *Delay differential equations*, University of Lethbridge;

[5] Benyamin Ghojogh, Mark Crowley (2019), *Linear and Quadratic Discriminant Analysis: Tutorial*;

[6] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo2 (May 2020), *An interpretable mortality prediction model for COVID-19 patients*, Nature;

[7] https://en.wikipedia.org/.