A grid search approach to find the death latency and case fatality rate of a disease during a pandemic

Stefan Strub

April 2025

1 Summary

During times of a pandemic it is important to not underestimate the severity of a disease, otherwise many governments ignore the danger at first which costs many lives. Unfortunately people often look at the case fatality rate (CFR) to estimate the danger of the disease, which is not as useful since during a pandemic deaths are usually under counted resulting in a lower CFR than the final CFR. During a pandemic it is only possible to estimate the final CFR, the real value is then unknown. Because knowing the final CFR is especially valuable during the pandemic we present in this work an approach using grid search to find a better estimate of the final CFR. We also estimate the death latency, which is the average time from a diagnosis until the death of a patient. These two numbers open also the possibility to make further predictions on the development of the pandemic. In the following we show an example of this method where we take the data from 5 years ago in April 2020 during the 2019-20 corona virus pandemic pretending to not know the data obtained afterwards.

2 Introduction

During times of a pandemic outbreak such as the 2019-20 corona virus pandemic it is essential to have good predictions of the pandemic development in order to make decisions on how to react to the outbreak. For example, people in charge of decisions, such as politicians, often take a look at the case fatality rate (CFR), the ratio of deaths to the total number of people diagnosed with the disease, to estimate the severity of the disease. This is often a bad idea since the during a pandemic of a high daily increase of infected people and a long resolution time could lead to under counting deaths resulting in a lower CFR than the final CFR, thus underestimating the severity of the disease, which could lead to taking not appropriate measures to slow the spread of the disease.

This underestimate of the pandemic unfortunately happened in 2019 and 2020 in most of the countries except for a few such as Taiwan and Hong Kong.

In this article we discuss a better estimate of the final CFR calculated during a pandemic where the final numbers of deaths and confirmed cases are unknown. This better estimate provides better information to world leaders, enabling them improved action against a disease, potentially saving lives.

For a better estimation of the final CFR we avoid the under counting of deaths due to the resolution time, by taking the ratio of deaths at a given day "x", to the total number of people diagnosed (confirmed cases) at day "x-resolution time" which is the resolution time before day x. To estimate the resolution time and and the final CFR we use a grid search to find the time shift (death latency) and scaling factor (final CFR) of the two noisy time series, namely the number of new cases at each day $C_{new}(t)$ and the number of new deaths at each day $D_{new}(t)$. Since the final CFR can be used as an estimate of the mortality, the probability that an infected person will die, we will call the the scaling factor mortality M and the time shift death latency δ for the rest of the article. The mortality is has a continuous value between 0 and 1, and since we have daily measurements of new cases and new deaths the death latency is discretized in days and has a positive value. To find the death latency and mortality where the overlap of the new cases and new deaths is optimal is an inverse problem, where the model parameter m is the death latency δ and mortality M. The goal is to obtain a probability density that describes the likelihood of $d = C_{new}(t - \delta) \cdot M$ being equal to $D_{new}(t)$. This is a non-linear relation of the model parameters to the data. Fortunately this equation can be solved quickly analytically, which motivates to use grid search for finding the likeliest model parameters.

As data prior, $p(d^{obs}|m) = p(d^{obs}|\delta, M)$, we use a Gaussian with standard deviation σ set equal to the noise level in the data, i.e.,

$$p(d^{obs}|m) = const. \exp\left[-\frac{1}{2N\sigma^2}\sum_{i=1}^{N}[D_{new}(t_i) - C_{new}(t_i - \delta) \cdot M]^2\right], \quad (1)$$

where N is the number of samples (days) in the time series. For the prior in model space, $p(m) = p(\delta, M)$, i.e., the prior information on δ and M, we choose constants. This corresponds to the assumption that we know very little about the death latency and mortality before looking at the actual time series.

3 Data

The used data is collected by the Johns Hopkins university (https://github.com/CSSEGISandData/COVID-19) where news reports of all countries are analysed and the total number of cases and total number of deaths at each day are stored in a data base. This

data base might has their errors since they are not directly the numbers provided by governments but by news reports. Some days later the data base gets synchronized with the numbers provided by the governments, therefore contain the most recent numbers an error additional to the error that not everybody is tested at every day. Furthermore it is also hard to trust the numbers provided by governments of countries other then democracies. Since dictators or one party systems have clear interests in keeping their number of cases and number of deaths low.

We consider a suitable country or region to make our estimates. This country or region has to full fill the following criteria: First, the country has to have an already declining number of new deaths in order to match the two time series. Second, the tests per population has to be high, in order consider the final CFR to be a good estimate of the mortality. Third, the number of deaths shall not be too low, otherwise statistical fluctuations could have a rather big influence in the ratio. The country which full fills all these criteria the best is Switzerland. Therefore only the data of Switzerland is considered for our estimate. The data space has a dimension of $2N$, where N is the number of days where data got collected. The factor of 2 is because we have two sets of data, once for the cases and once for the deaths.

4 Methods

The new cases and new deaths calculated from the data are pictured in Fig. 1 and Fig. 2 as the blue data points. Since finding the optimal overlap of these two very noisy time series we average the data points over 7 days, where we take additionally the 3 days before and after a given day. Therefore the most recent 3 days are cut off. These averages are shown in orange. Our goal is to find the death latency and mortality such that these two functions overlap the best.

The two model parameters are a mix of discretized and continuous parameters. The dicretisation of the death latency is given by the measurements which are taken once a day. For the mortality we choose a discretisation from 0.001 to 0.5 with a step width of 0.001. For measurements of 85 days this gives a model space dimension of $n = 42500$. The mortality would be between 0 and 1 but for the 2019-20 corona virus outbreak we can assume to be lower than 0.5 for the death latency we can also take some prior knowledge form patients which died from the disease where we collect the data on the date of infection and day of death. For this estimate we do not take the effort to obtain additional information from other sources. This inverse problem is mixed-determined because the data points of the data set contains the information about both parameters. To solve the problem we use grid search since the forward calculations are inexpensive and can be solved quickly. To determine the optimal parameters we take the posterior maximum-likelihood model m_{max}. There is also no need for a regularisation since we can easily compute the posterior $p(m|d^{obs})$. For computing the probability that the model parameters m are falling in a specific

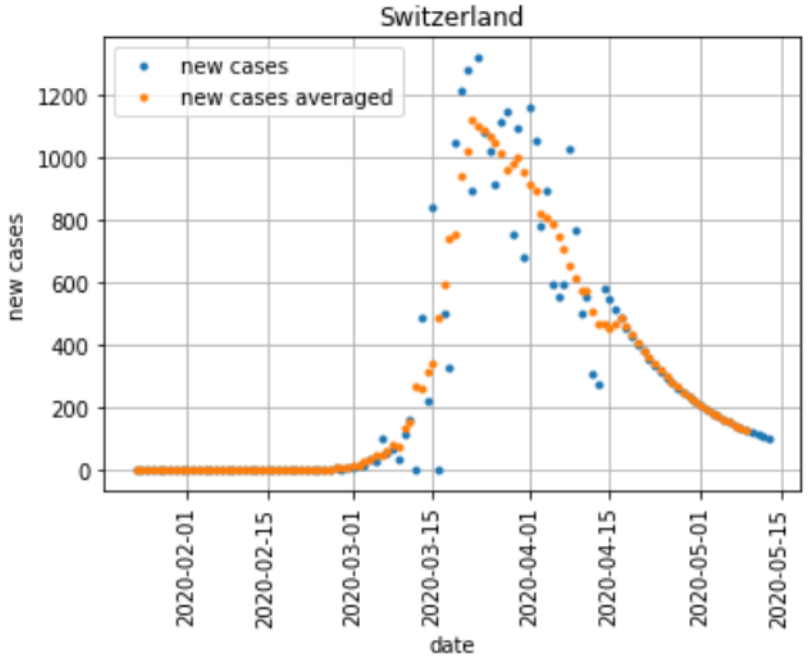

Figure 1: New cases of in confirmed to have the corona virus at each day in Switzerland. The data beyond 2020-04-15 is a prediction of new cases. For this problem we do not need numbers of future dates, so this can be ignored.

sub volume M' of the model space $\mathbb{M}$ we can calculate:

$$P(M \in \mathbb{M}) = \int_{\mathbb{M}} p(m|d^{obs})\, dm. \tag{2}$$

5 Results

Plotting the posterior $p(m|d^{obs})$ shown in Fig. 3 shows clear a clear maximum-likelihood model $m_{max} = (\delta = 9\,\text{days}, M = 0.054)$. Indeed when we manipulate the new cases average according to our result we obtain a good overlap with the new deaths average function, as we can see in Fig. 4.

6 Discussion and conclusions

The data source could be improved by taking the data from the Statistisches Amt Zurich, because their data is directly collected from the cantons of Switzerland. This result can be used further to determine on how good or baldy other countries are testing and therefore give an estimate on how large the dark number of infected people (untested positive) is. Tho the mortality for the corona

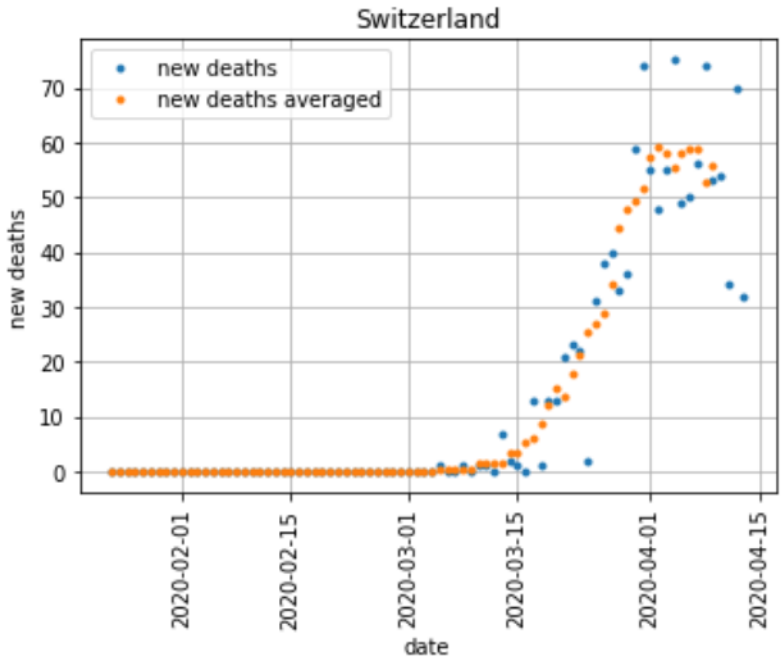

Figure 2: New deaths of people confirmed to have the corona virus at each day in Switzerland.

virus specifically is larger for older people, such estimates have to be made carefully, for example Italy has a rather old population. Furthermore, the mortality and death latency can be used to predict the new deaths at a point later in time, since the number of new cases is shifted by the death latency. This prediction also gets the advantage to predict the developments of new deaths because only looking at the past data points to make the prediction might not take into account most recent changes in behavior. For example if a country goes into a look down, the cases show an effect earlier than the deaths. Therefore can the effect on the deaths be better predicted by shifting and shrinking the confirmed cases.

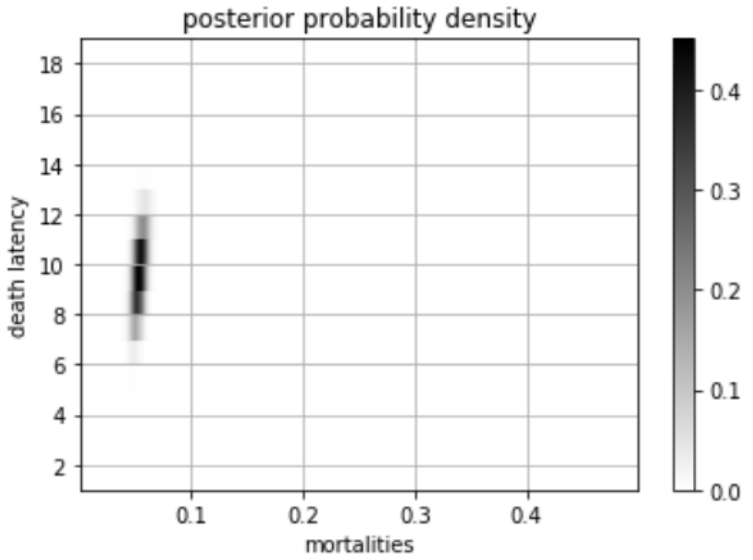

Figure 3: The posterior $p(m|d^{obs})$ plotted for the two model parameters death latency and mortality.

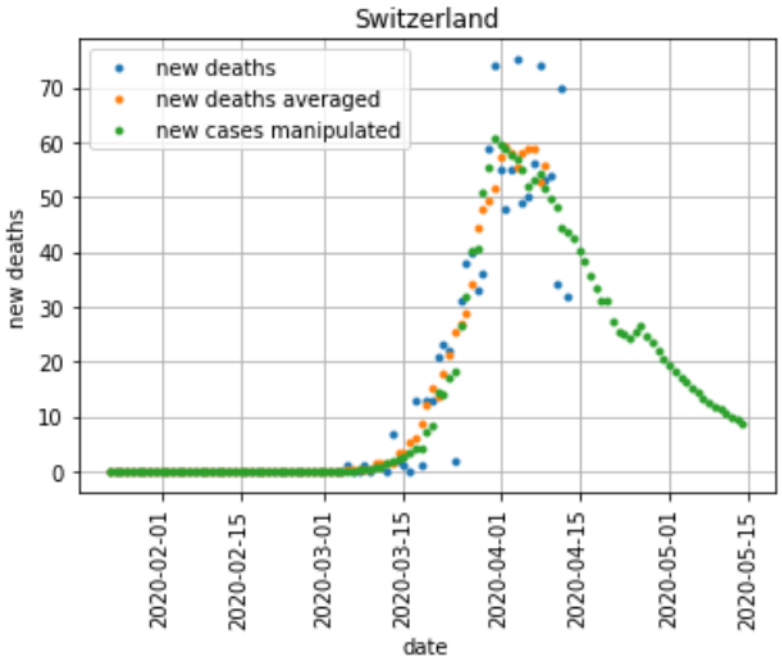

Figure 4: New deaths of people confirmed to have the corona virus at each day in Switzerland.