Assessment of COVID-19 hospitalization forecasts from a simplified SIR model[*]

P.-A. Absil[§] Ousmane Diao[§] Mouhamadou Diallo[¶]

July 14, 2020

Abstract

We propose the SH model, a simplified version of the well-known SIR compartmental model of infectious diseases. With optimized parameters and initial conditions, this time-invariant two-parameter two-dimensional model is able to fit COVID-19 hospitalization data over several months with a remarkably good accuracy. Some COVID-19 hospitalization forecasts are also observed to be excellent over a period as long as two months. However, the model is less successful at predicting the time and height of the hospitalization peak.

Key words: COVID-19 prediction; COVID-19 forecast; SIR model; SH model; hidden variable; hospitalization prediction

1 Introduction

The SIR model [KMW27] is a simple compartmental model that is widely used to model infectious diseases [Het00]. Letting $S(t)$, $I(t)$, and $R(t)$ denote the number of susceptible, infectious and removed (or recovered) individuals at time t, and letting $\dot{S}(t)$, $\dot{I}(t)$, and $\dot{R}(t)$ denote their time derivatives, the SIR model consists in the following three-dimensional continuous-time autonomous dynamical system

$$\dot{S}(t) = -\frac{\beta}{N} S(t)I(t) \tag{1a}$$

$$\dot{I}(t) = \frac{\beta}{N} S(t)I(t) - \gamma I(t) \tag{1b}$$

$$\dot{R}(t) = \gamma I(t), \tag{1c}$$

where $N = S(t) + I(t) + R(t)$ is the constant total population and β and γ are parameters. The SIR model, and several (sometimes deep) variations thereof, have been applied in several works to model the COVID-19 dynamics (see, e.g., [Atk20, Koz20, Nes20, CNP20]) with known limitations (see [RVHL20, BFG$^+$20, WF20]). Sometimes, an SIR-like model is used to make long-term predictions (see [BD20]). However, we are not aware of studies where the SIR model parameters and initial conditions are learned on a "train" part of the available data in order to predict a "test" part of the data, making it possible to assess the prediction accuracy of the model.

In this paper, we adapt the SIR model to the situation where (i) S, I and R are hidden variables but $I(t)$ is observed through a "proxy" $H(t) = \alpha I(t)$, where α is unknown but constant, and (ii) not only β and γ but also the total population N are unknown and have thus to be estimated. In the context of the COVID-19 application, H will stand for the total number of lab-confirmed hospitalized patients. The proposed adapted SIR model, which we term *SH model*, is given in (8).

[*] The second author is supported by a fellowship awarded by UCLouvain's Conseil de l'action internationale.

[§] ICTEAM Institute, UCLouvain, B-1348 Louvain-la-Neuve, Belgium (http://sites.uclouvain.be/absil/).

[¶] Molecular Biology Unit/Bacteriology-Virology Lab, CNHU A. Le Dantec / Université Cheikh Anta Diop, Dakar, Sénégal

It has two state variables ($\bar{S}$—a scaled version S—and H) and two parameters ($\bar{\beta}$—which lumps together the parameters β, N, and α—and γ).

We leverage the proposed SH model as follows in order to make hospitalization predictions. Given observed values $(H_o(t))_{t=t_i,\dots,t_c}$, we estimate the parameters $\bar{\beta}$, γ, and the initial conditions $\bar{S}(t_i)$ and $H(t_i)$ of the SH model. Then we simulate the SH model in order to predict $(H(t))_{t=t_c+1,\dots,t_f}$ for a specified final prediction time t_f.

This work connects with the areas of parameter estimation (for obvious reasons), data assimilation (for the generation of the initial conditions) and machine learning (for the train-test approach). Combining these concepts in order to make COVID-19 hospitalization predictions using an SIR-like model appears to be new.

2 Data

In Section 4, we will use use COVID-19 datasets for Belgium[1] and France[2] that provide us with the following data for $t = t_s, \dots, t_e$:

- $H_o(t)$: number of COVID-19 hospitalized patients on day t;
- $E_o(t)$: number of COVID-19 patients entering the hospital (number of lab-confirmed hospital intakes) on day t;
- $L_o(t)$: number of COVID-19 patients discharged from the hospital on day t.

The subscript $_o$ stands for "observed".

2.1 Discussion

In the data, there is usually a mismatch between $H_o(t)$ and $H_o(t-1)+E_o(t)-L_o(t)$, even if $E_o(t)$ and $L_o(t)$ are replaced by $E_o(t-1)$ and $L_o(t-1)$. For Belgium, $H_o(t_s)+\sum_{t=t_s+1}^{t_e} E_o(t)-L_o(t)$ is significantly larger than $H_o(t_e)$. This can be due to the patients who get infected at the hospital (they would be counted in H_o without appearing in E_o) and to the patients who die at the hostpital (they would be removed from H_o withoug appearing in L_o).

In order to remedy this mismatch for the Belgian data, we redefine $L_o(t)$ by $L_o(t) := -H_o(t)+H_o(t-1)+E_o(t)$. For the French data, we sum the "rad" and "dc" columns to get $L_o(t)$, and we define $E_o(t) = H_o(t) - H_o(t-1) + L_o(t)$.

Several other COVID-19 data are available. In particular, the daily number of infected people, $I_o(t)$, is also reported by health authorities. However, the graph of the I_o time series is visually less smooth than H_o. We thus suspect that it lends itself less nicely to being explained by an SIR model. A possible cause is that I_o is affected by two major sources of noise: not all infected persons are tested, and the tests are not perfectly accurate. In contrast, the reported number of COVID-19 hospitalized people, H_o, is expected to be much more accurate. Moreover, for the authorities, predicting H is more crucial than predicting I, as H has a more direct impact on the required number of beds or ventilators. Therefore, as in [Koz20], we focus on H.

3 Models and methods

3.1 Case hospitalization ratio

We assume that, for all t,

$$H(t) = \alpha I(t) \tag{2}$$

[1] `https://epistat.sciensano.be/Data/COVID19BE_HOSP.csv` obtained from `https://epistat.wiv-isp.be/covid/`

[2] donnees-hospitalieres-covid19-2020-07-10-19h00.csv obtained from `https://www.data.gouv.fr/en/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/`

where α is unknown but constant over time. In other words, (2) posits that a constant fraction of the infected people is hospitalized.

3.2 Observation models

We assume the following observation models with additive noise:

$$H_o(t) = H(t) + \epsilon_H(t) \tag{3a}$$

$$E_o(t) = E(t) + \epsilon_E(t) \tag{3b}$$

$$L_o(t) = L(t) + \epsilon_L(t). \tag{3c}$$

Assuming that the ϵ noises are independent Gaussian centered random variables confers a maximum likelihood interpretation to some subsequent estimators, but this assumption is obviously very simplistic.

3.3 Proposed SH model

Multiplying (1a) and (1b) by α, and multiplying the numerator and denominator of (1a) by α, we obtain

$$\alpha\dot{S}(t) = -\frac{\beta}{N\alpha}\,\alpha S(t)\,\alpha I(t) \tag{4}$$

$$\alpha\dot{I}(t) = \frac{\beta}{N\alpha}\,\alpha S(t)\,\alpha I(t) - \gamma\alpha I(t). \tag{5}$$

Letting

$$\bar{S} := \alpha S \tag{6}$$

$$\bar{\beta} := \frac{\beta}{N\alpha} \tag{7}$$

and using (2), we obtain the simplified SIR model

$$\dot{\bar{S}}(t) = -\bar{\beta}\bar{S}(t)H(t) \tag{8a}$$

$$\dot{H}(t) = \bar{\beta}\bar{S}(t)H(t) - \gamma H(t) \tag{8b}$$

which we term the *SH model*. (The "S" in this SH model can be interpreted as the number of individuals susceptible of being hospitalized.) The SH model has only two parameters ($\bar{\beta}$ and γ), one hidden state variable ($\bar{S}$) and one observed state variable (H) with observation model (3a).

Note that, in (8), the number of patients entering the hospital by unit of time is

$$E(t) := \bar{\beta}\bar{S}(t)H(t) \tag{9}$$

and the number of patients leaving the hospital by unit of time is

$$L(t) := \gamma H(t). \tag{10}$$

3.4 Estimation of the SH model parameters and initial conditions

The goal is now to leverage the SH model (8) in order to predict future values of H based on its past and current observations $(H_o(t))_{t=t_{\underline{s}},\dots,t_c}$. To this end, we have to estimate (or "learn") four estimand variables: the two parameters $\bar{\beta}$ and γ and the two initial values $\bar{S}(t_i)$ and $H(t_i)$, where t_i is the chosen initial time for the SH model (8). One possible approach is to minimize some error measure between the simulated values $(H(t))_{t=t_i,\dots,t_c}$ and the observed values $(H_o(t))_{t=t_i,\dots,t_c}$ as a function of the four estimand variables. However, the error measure is not available as a closed-form expression of the four estimands, and this makes this four-variable optimization problem challenging. We now show that it is possible to estimate $H(t_i)$ and γ separately. This leaves us with an optimization problem in the two variables $\bar{\beta}$ and $\bar{S}(t_i)$, making it possible to visualize the objective function by means of a contour plot.

3.4.1 Train and test sets

To recap, we have $t_s \leq t_i < t_c < t_e$. The provided datasets go from t_s to t_e. The *test set* is $(H_o(t), E_o(t), L_o(t))_{t\in[t_c+1,t_e]}$, and this data cannot be used to estimate the variables and simulate the SH model. The SH model is initialized at t_i, and we refer to the data $(H_o(t), E_o(t), L_o(t))_{t\in[t_i,t_c]}$ as the *train set*, though it is legitimate to widen it to $t \in [t_s, t_c]$.

3.4.2 Estimation of $H(t_i)$

It is reasonable to believe that ϵ_H in (3a) is small in practice. Hence we simply take

$$H(t_i) := H_o(t_i).$$

3.4.3 Estimation of γ

We have $L(t) = \gamma H(t)$, see (10). In view of the observation model (3), we can estimate γ by a ratio of means:

$$\hat{\gamma}^{\mathrm{RM}} = \frac{\sum_{t=t_i}^{t_c} L_o(t)}{\sum_{t=t_i}^{t_f} H_o(t)}.$$

Several other estimators are possible, such as the least square estimator, or the total least squares estimator which is the maximum likelihood estimator of γ for the iid Gaussian noise model (3).

Note that t_i in the expressions of $\hat{\gamma}$ can legitimately be replaced by any time between t_s and t_c. Only data in the test set, i.e., occurring after t_c, are unavailable in the variable estimation phase.

3.4.4 Estimation of $\bar{\beta}$ and $\bar{S}(t_i)$

We now have to estimate the two remaining variables, namely $\bar{\beta}$ and $\bar{S}(t_i)$. We choose the following sum-of-squared-errors objective function

$$\phi(\bar{\beta}, \bar{S}(t_i)) = c_H \sum_{t=t_i}^{t_c} (H(t) - H_o(t))^2 + c_E \sum_{t=t_i}^{t_c} (E(t) - E_o(t))^2 + c_L \sum_{t=t_i}^{t_c} (L(t) - L_o(t))^2, \tag{11}$$

where the c coefficients are parameters, all set to 1 in our experiments, unless otherwise stated. In (11), $H(t)$, $E(t)$ as in (9), and $L(t)$ as in (10), are given by the (approximate) solution of the SH model (8) in which (i) $H(t_i)$ and γ take the values estimated as above, and (ii) $\bar{\beta}$ and $\bar{S}(t_i)$ take the values specified in the argument of ϕ. In order to compute the required approximate solution of the SH model (8), we use the explicit Euler integration with a time step of one day, yielding, for $t = t_i, \ldots, t_c - 1$,

$$\bar{S}(t+1) = \bar{S}(t) - \bar{\beta}\bar{S}(t)H(t) \tag{12a}$$

$$H(t+1) = \bar{H}(t) + \bar{\beta}\bar{S}(t)H(t) - \gamma H(t). \tag{12b}$$

Now that the objective function ϕ is defined (also termed "cost function" or "loss function"), we estimate $\bar{\beta}$ and $\bar{S}(t_i)$ by the (approximate) minimizer of ϕ returned by some derivative-free optimization solver.

3.5 Alternative: optimizing over the four estimand variables

An alternative is to consider the following objective function in the four estimand variables:

$$\tilde{\phi}(\bar{\beta}, \bar{S}(t_i), \gamma, H(t_i)) = c_H \sum_{t=t_i}^{t_c} (H(t) - H_o(t))^2 + c_E \sum_{t=t_i}^{t_c} (E(t) - E_o(t))^2 + c_L \sum_{t=t_i}^{t_c} (L(t) - L_o(t))^2, \tag{13}$$

In (13), $H(t)$, $E(t)$ as in (9), and $L(t)$ as in (10), are given by the solution of the discrete-time SH model (12) where the parameters $\bar{\beta}$ and γ and the initial conditions $\bar{S}(t_i)$ and $H(t_i)$ take the

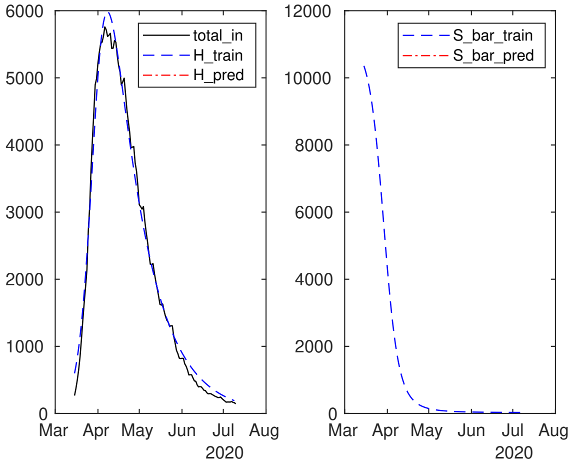

Figure 1: Belgium, fitting the SH model to the H_o (total hospitalized) curve. In this experiment, the train set is the whole dataset, hence there there is no test (prediction) curve. Reproduce with SHR__12PA__BEL__traintstart1__traintstop117__c100.zip.

values specified in the argument of $\tilde{\phi}$. Minimizing $\tilde{\phi}$ is a more challenging problem. It may be essential to give a good initial point to the optimization solver, and a natural candidate for this is the values obtained by the procedure described in the previous subsections.

In our preliminary experiments, we have found that this alternative does not present a clear advantage in terms of the prediction mean absolute percentage error (MAPE). The results reported in Section 4 are obtained with the sequential prediction approach of Section 3.4, unless otherwise specified.

3.6 Prediction of H

Recall that the time range between t_i and t_c is the train period and the time range between $t_c + 1$ and t_e is termed the test period.

In order to predict the values of H over the test period, we apply the above procedure to estimate the four estimand variables $\bar{\beta}$, γ, $\bar{S}(t_i)$, and $H(t_i)$, and we compute the solution $H(t)$ of (12) for t from t_i to t_e. The prediction is then $(H(t))_{t=t_c+1,\ldots,t_e}$.

Note that the method only uses the values of $H_o(t)$ and $L_o(t)$ over the train period. The discrepancy between $(H(t))_{t=t_c+1,\ldots,t_e}$ and $(H_o(t))_{t=t_c+1,\ldots,t_e}$ thus reveals the quality of the prediction.

4 Results

We now apply the method of Section 3 to the data of Section 2 available for Belgium and France. For Belgium, the dataset start date t_s is 2020-03-15 and the end date t_e is 2020-07-10. For France, t_s is 2020-03-18 and t_e is also 2020-07-10.

The method is implemented in MATLAB version R2019a. The code to reproduce the results is available from `https://sites.uclouvain.be/absil/2020.05`. PATODO: make code available.

4.1 Fitting experiment

We first check how well the SH model (8) can fit the available data for Belgium. For this experiment, we use the method of Section 3.5 with $c_E = c_L = 0$ in order to get the best possible fit (in the least squares sense) to the H_o curve. The result is shown in Figure 1.

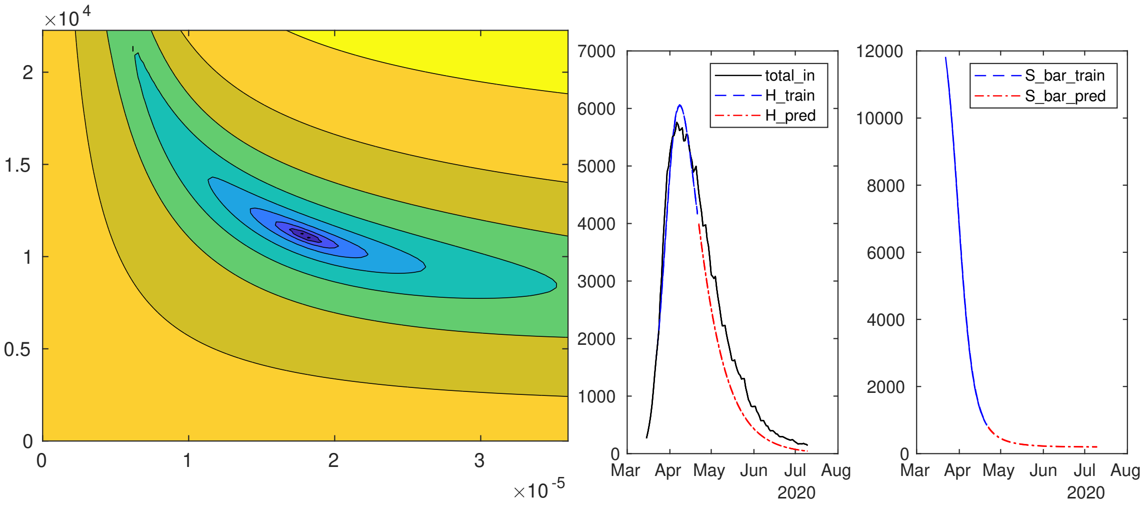

Figure 2: Belgium, 80-day test period. Left: contour plot of ϕ (11). Right: fitting and predictions with the SH model. Reproduce with SHR_12PA_BEL_traintstart8_traintstop38_c111.zip.

The fitting error is remarkably small. For the French data, we have obtained a similar conclusion.

Note that the parameters of the SH model are constant with respect to time. This contrasts with [Koz20] where there are two phases, and with [Nes20] where the infection rate is piecewise constant with several pieces.

We stress that Figure 1 tells us little about the prediction capability of the model. If the fit over some period is bad, then predictions (i.e., forecasts) over that period can only be bad. But if the fit is good (as it is the case here), the predictions can still be bad due to the sensitivity of the estimands with respect to the data preceding the to-be-predicted period. For example, a better fit (in the RMSE sense) than in Figure 1 can be obtained with a polynomial of degree 8; however, its prediction capability is abysmal.

In order to assess the prediction capability of the model, we have to learn the estimand variables over a train period, use the learned model in order to predict H over a forthcoming test period, and finally compare the prediction with the held-back (test) data. This is what we proceed to do in the rest of this Section 4.

4.2 Belgium, 80-day test period

We first consider the Belgian data, for which t_e is 2020-07-10. We choose t_c to be 2020-04-21, which gives a test period of 80 days. In order to give a sense of the sensitivity of the results, we superpose the three curves obtained for a train period duration equal to 29, 30, and 31 days.

A contour plot of the objective function ϕ (11) is given in Figure 2 for Belgium. Based on a visual inspection, we choose (1e-5,1e4) as the initial point of the optimization solver. For simplicity, in this preliminary report, the solver is MATLAB's `fminsearch`, which uses the Nelder–Meade direct search method.

The middle plot of Figure 2 shows $(H_o(t))_{t=t_s,\dots,t_e}$ (observed hospitalizations, in black), $(H(t))_{t=t_i,\dots,t_c}$ (hospitalizations given by the model over the train period, in blue), and $(H(t))_{t=t_c+1,\dots,t_e}$ (hospitalizations predicted over the test period, in red). The right-hand plot of Figure 2 shows the evolution of $\bar{S}(t)$.

4.3 Belgium, 60-day test period

We proceed as in Section 4.2 but now we choose t_c to be 2020-05-11, which gives a test period of 60 days. The results are shown on Figure 3.

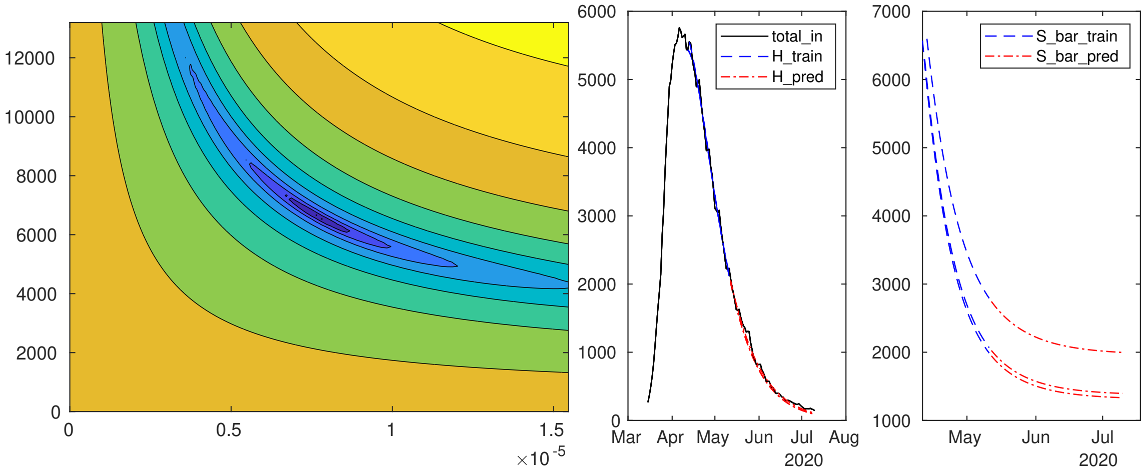

Figure 3: Belgium, 60-day test period. Left: contour plot of ϕ (11). Right: fitting and predictions with the SH model. Reproduce with SHR_12PA_BEL_traintstart28_traintstop58_c111.zip.

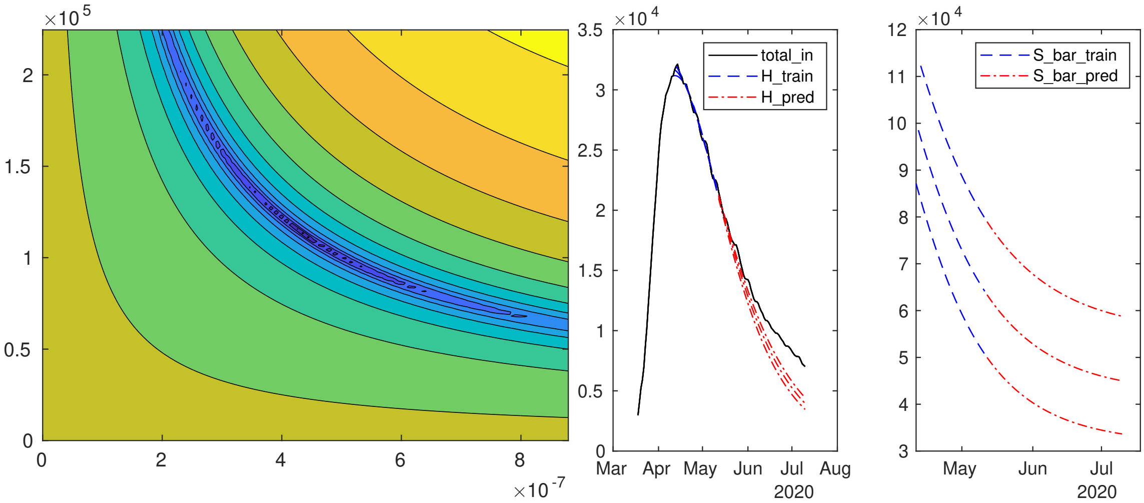

Figure 4: France, 60-day test period. Left: contour plot of ϕ (11). Right: fitting and predictions with the SH model. Reproduce with SHR_12PA_FRA_traintstart25_traintstop55_c111.zip.

4.4 France, 60-day test period

We proceed as in Section 4.3 but now we use the data for France. The results are shown on Figure 4.

4.5 Discussion

Some long-term predictions for Belgium are remarkably accurate. Figure 3 is a representative case. The best of the three shown curves has an MAPE_test of 7.8221%. However, we see in the right-hand plot of Figure 3 that $\bar{S}$ is highly sensitive to the data. This is corroborated by the oblong shape of the level curves of ϕ shown in the left-hand plot. We have observed that the sensitivity of $\bar{S}$ gets even worse when the fitting is performed further away from the peak of H_o.

However, when the train period is located in the increasing part of the H curve, the long-term predictions become much less accurate. It appears that the high sensitivity of $\bar{S}$ has a strong impact on the height and position of the peak in the H curve.

The predictions that we obtained for France are in general less accurate. We also considered some departments separately, with a similar conclusion.

The ratio by which $\bar{\beta}$—and thus β in view of (7)—has to be multiplied to reach the threshold between and increase and a decrease of $H(t)$ (the state of the model and γ being equal) is

$$\frac{L(t)}{E(t)} = \frac{\gamma}{\bar{\beta}\bar{S}(t)}$$

where the equality comes from (9) and (10). In view of the high variability of $\bar{S}(t)$ observed in most experiments, this ratio is very difficult to estimate accurately. Furthermore, the impact of the various prevention measures on β is largely unknown. Hence no recommendation regarding relaxing or strengthening prevention measures can be deduced from our study. In any case, even if the ratio was known, trying to steer the system close to the threshold would seem very unwise.

5 Conclusion

The experiments in Section 4 have shown that the proposed method has a remarkably good fitting capability over the whole available data, and also a remarkably good predictive value over certain time ranges for the Belgian data. However, there are also time ranges where the prediction is very inaccurate, and the accuracy is also found to be lower for the French data. The predictions returned by the model should thus be taken with much caution. In keeping with this warning, we refrained from displaying predictions beyond the end time of the datasets. However, the code is freely available to make such predictions, but there is no warranty on the accuracy of the predictions.

Another source of caution is that there is no guarantee that the considered objective functions are unimodal. The optimization solver might thus get stuck in a local nonglobal minimum, yielding a suboptimal fit of the train data and possibly a poorer prediction than what an omniscient solver would achieve. Moreover, even if the objective function is unimodal, the stopping criterion of the solver may trigger before an accurate approximation of the minimum is reached.

References

[Atk20] Andrew Atkeson. What will be the economic impact of COVID-19 in the US? Rough estimates of disease scenarios. Working Paper 26867, National Bureau of Economic Research, March 2020. doi:10.3386/w26867.

[BD20] Gyan Bhanot and Charles DeLisi. Predictions for europe for the covid-19 pandemic from a SIR model. *medRxiv*, 2020. doi:10.1101/2020.05.26.20114058.

[BFG$^+$20] Jackie Baek, Vivek F. Farias, Andreea Georgescu, Retsef Levi, Tianyi Peng, Deeksha Sinha, Joshua Wilde, and Andrew Zheng. The limits to learning an SIR process: Granular forecasting for Covid-19, 2020. arXiv:arXiv:2006.06373.

[CNP20] Giuseppe C. Calafiore, Carlo Novara, and Corrado Possieri. A modified SIR model for the COVID-19 contagion in Italy, 2020. arXiv:arXiv:2003.14391.

[Het00] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000. URL: https://doi.org/10.1137/S0036144500371907, arXiv:https://doi.org/10.1137/S0036144500371907, doi:10.1137/S0036144500371907.

[KMW27] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927. doi:10.1098/rspa.1927.0118.

[Koz20] Gregory Kozyreff. Hospitalization dynamics during the first COVID-19 pandemic wave: SIR modelling compared to Belgium, France, Italy, Switzerland and New York City data, 2020. `arXiv:arXiv:2007.01411`.

[Nes20] Yurii Nesterov. Online prediction of COVID19 dynamics. Belgian case study. CORE Discussion Paper 2020/22, UCLouvain, 2020. URL: `https://uclouvain.be/en/research-institutes/lidam/core/core-discussion-papers.html`.

[RVHL20] Weston C. Roda, Marie B. Varughese, Donglin Han, and Michael Y. Li. Why is it difficult to accurately predict the COVID-19 epidemic? *Infectious Disease Modelling*, 5:271 – 281, 2020. `doi:https://doi.org/10.1016/j.idm.2020.03.001`.

[WF20] Meimei Wang and Steffen Flessa. Modelling Covid-19 under uncertainty: what can we expect? *The European Journal of Health Economics*, 2020. `doi:10.1007/s10198-020-01202-y`.