SEIR Modeling, Simulation, Parameter Estimation, and Their Application for COVID-19 Epidemic Prediction †

Elham Taghizadeh 1,* and Ali Mohammad-Djafari 2,3

1 Department of Mathematics, University of Mazandaran, Babolsar 47415-95447, Iran
2 International Science Consulting and Training (ISCT), 91440 Bures sur Yvette, France
3 Scientific Leader of Shanfeng Blower Company, Shangyu District, Shaoxing 312300, China
* Correspondence: e.taghizadeh@stu.umz.ac.ir

Abstract: In this paper, we consider the SEIR (Susceptible-Exposed-Infectious-Removed) model for studying COVID-19. The main contributions of this paper are: (i) a detailed explanation of the SEIR model, with the significance of its parameters. (ii) calibration and estimation of the parameters of the model using the observed data. To do this, we used a nonlinear least squares (NLS) optimization and a Bayesian estimation method. (iii) When the parameters are estimated, we use the models for the prediction of the spread of the virus and compute the probable number of infections and deaths of individuals. (iii) We show the performances of the proposed method on simulated and real data. (iv) Remarking that the fixed parameter model could not give satisfactory results on real data, we proposed the use of a time-dependent parameter model. Then, this model is implemented and used on real data.

Keywords: SEIR model; ordinary differential equations; nonlinear least squares; Bayesian inference; COVID-19

1. Introduction

Mathematical models and computer simulations are useful experimental tools for building and testing theories, assessing quantitative conjectures, determining sensitivities to changes in parameter values, and estimating key parameters from data. Understanding the transmission characteristics of infectious diseases in communities, regions, and countries can lead to better approaches to controlling the transmission of those diseases. Mathematical models are utilized in comparing, planning, implementing, evaluating, and optimizing numerous detection, prevention, therapy, and management programs. Epidemiology modeling can contribute to the design and analysis of epidemiological surveys, recommend crucial data that should be collected, determine trends, make general forecasts, and estimate the uncertainty in forecasts [1–4].

There exist a number of models for infectious diseases; as for compartmental models, starting from the very classical SIR model to additional complicated proposals [5]. Many research works have been reported. They show that those SIS, SIR and SEIR models can replicate the dynamics of various epidemics. These models have also been used to model COVID-19 [6–8]. To mention a few of them, Tang et al. [9] investigated a general SEIR epidemiological model where quarantine, isolation, and treatment are considered. Wang et al. [10] applied the phase-adjusted estimation for the amount of COVID-2019 cases in Wuhan.

Our main contribution is to consider the SEIR compartmented model and to give the significance of its parameters; methods for the estimation of its parameters and its calibration, and finally, use it for prediction. For the estimation of its parameter, we used two methods: a deterministic nonlinear least squares (NLS) optimization and a Bayesian
estimation method. Once the parameters are estimated, we use the model for the prediction of the spread of the virus and calculate the probable number of infections and deaths of individuals. We show the performances of the proposed method on simulated data. Remarking that the fixed-parameter model could not provide satisfactory results on real data, we proposed the use of a time-dependent parameter model.

The rest of this paper is organized as follows. Section 2 describes the construction of the mathematical SEIR model and the comprehension of its parameters. Section 3 focuses on the estimation of the parameters of SEIR using simulated data. In Section 4, the focus is on the application of this model for the available data. We will see how to adapt the parameter estimation to the available data. Remarking that the evolution of the COVID-19 epidemic can not be modeled by a fixed-parameter model, we adopt a time-varying parameter model and adapt it to the available data. In Section 5, we describe the main conclusions.

2. A Continuous-Time SEIR Models

SEIR is a classical model for the problem of modeling epidemics. This model is composed of four components, shown in Figure 1.

Figure 1. The four components of the SEIR model and their relations.

The dynamic equations relating these compartments are given by:

\[
\frac{dS}{dt} = -\beta S \frac{I}{N}, \quad \frac{dE}{dt} = \beta S \frac{I}{N} - \sigma E(t), \quad \frac{dI}{dt} = \sigma E - \gamma I, \quad \frac{dR}{dt} = \gamma I
\]  

(1)

In this model:

- Susceptible (S) is the variety of people susceptible to contracting the infection;
- Exposed (E) the variety of individuals exposed who are alive but not infected;
- Infectious (I) is the variety of individuals that are infected;
- Recovered (R) is the cumulative variety of individuals that recovered from the disease.

The total population at time t is represented by \( N = S + E + I + R \). The significance of the parameters are: \( \beta \) is the infection rate or the speed of spread, \( \sigma \) is the incubation rate or the rate of latent individuals becoming infectious (average period of incubation is \( 1/\sigma \)), and \( \gamma \) is the recovery rate or mortality rate. If the duration of infection is \( D \) then \( \gamma = 1/D \). This model has the initial conditions \( S(0) \geq 0, I(0) \geq 0, E(0) \geq 0, \) and \( R(0) \geq 0 \). Note that \( S + E + I + R = N \).

Figure 2 shows an example simulation of this model.

Figure 2. A typical SEIR Model simulated solution. This result is obtained with the following parameters: \( \beta = 0.5, \sigma = 0.02, \gamma = 0.01, days = 420 \), and the initial values: \( N = 65,256,000, E(0) = 5000, I(0) = 0, R(0) = 0 \).
3. Calibration of the Epidemic Model

With such a set of differential equations, the mathematical problems to solve are mainly in two categories: forward and inverse problems. The forward problem consists of computing the outputs when the parameters and the initial conditions are fixed. The inverse problems consist of: (i) estimating the parameters when a set of observable data related to the outputs are given; (ii) predict unobserved parts of outputs (future values) from the observed (passed data) when the parameters are estimated. Let us first explain these in a more formal and general mathematical formulation.

3.1. Forward Problem

Consider now the following general SEIR

\[ x(t) = \chi(x(t), \theta), \]
\[ x(t_0) = x_0 \]  

(2)

in which

- \( x(t) = [S(t), E(t), I(t), R(t)]^T \) is the 4-dimensional state vector;
- the initial state \( x_0 = x(0) \) is a constant 4-dimensional vector;
- \( \theta = [\beta, \sigma, \gamma] \) is the 3-dimensional vector of the model parameters;
- \( \chi : \Omega \subset \mathbb{R}^4 \times \mathbb{R}^3 \rightarrow \mathbb{R}^4 \) is a continuous map, which is described on the open set

\[ \Omega = \{(x(t), \theta) \in \mathbb{R}^4 \times \mathbb{R}^3|x_n(t) > 0 \text{ and } \theta_m > 0, \text{ for } n = 1, \cdots 4, m = 1, \cdots, 3 \} \]  

The forward problem consists in computing the model response \( x(t) \) given the initial conditions \( x_0 \) and a set of parameters \( \theta \).

3.2. Inverse Problem

The inverse problem is the opposite of the forward problem: Given a set of observed data \( z(t), t = 1, \cdots, M \), we try to estimate parameter \( \theta \), and then, the state vector \( x(t) \) for all \( t \), and thus, the unobserved values of \( z(t), t > M \) (prediction). There are mainly two approaches to inverse problems: deterministic and probabilistic. In the following, we describe these two approaches through two methods: i) the nonlinear least squares (NLS) and the Bayesian maximum a posteriori (MAP).

3.3. Nonlinear Least Squares (NLS) Solution

Figure 3 schematically shows the forward and inverse problems in a deterministic approach.

\[ \theta \rightarrow \begin{array}{c}
\begin{array}{c}
\text{Forward model} \\
x(t) = \chi(x(t), \theta) \\
x(t_0) = x_0
\end{array}
\end{array} \rightarrow x(t) \rightarrow \begin{array}{c}
\begin{array}{c}
\text{Observables} \\
z(t) = \Phi(\theta, x(t))
\end{array}
\end{array} \rightarrow z(t), t = 1, \cdots, M
\]

\[ \theta \leftarrow \begin{array}{c}
\begin{array}{c}
\text{Inversion criteria} \\
J(\theta) = \sum_{t=1}^{M} \|z(t) - \Phi(\theta, x(t))\|^2
\end{array}
\end{array} \leftarrow z(t), t = 1, \cdots, M
\]

Figure 3. Forward and inverse problems. Here, we assume that the initial conditions \( x_0 \) are known.

The nonlinear least squares method (NLS) is used numerically to approximate a solution for the inverse problem (3) within the least-squares fitting, where we look for the \( \theta^* \) value of the model parameter \( \theta \), which minimizes the least squares criterion

\[ J(\theta) = \sum_{t=1}^{M} \|z(t) - \Phi(\theta, x(t))\|^2 \]  

(4)
where $x(t)[\theta]$ is the output of the model for a given set of parameters $\theta$ and $z = \phi(\theta, x(t))$ is the observable outputs.

Such a problem is clearly a nonlinear least-squares problem, since the dependence of a solution system on the parameter $\theta$ is through a highly nonlinear system of differential equations.

3.4. Bayesian Estimation Framework

In a Bayesian framework, first, we may consider Equation (4) as the log-likelihood of parameters $\theta$ in the data:

$$p(z|\theta) \propto \exp\left\{-\frac{1}{2} v_{\epsilon} J(\theta)\right\}$$

(5)

This means that we assume that there is some uncertainty on the observed data $z(t)$ and that the only thing we know about those errors is that they are centered (zero mean), identically and independently distributed with fixed variance $v_{\epsilon}$. In fact, this formal prior knowledge is translated by the maximum entropy principle (MEP) to a Gaussian distribution [11].

We may then assign or choose a prior $p(\theta)$ to translate any prior knowledge we may have about $\theta$. This can be just flat priors on the parameters or, rather, some reference, noninformative, or conjugate priors. In this work, as all the parameters are the rates (infection rate, incubation rate, and recovery rate), we used flat priors for them.

When the likelihood and the prior are fixed, we can use the Bayesian rule to obtain the posterior law:

$$p(\theta|z) = \frac{p(z|\theta) p(\theta)}{p(z)}$$

(6)

from which we can infer $\theta$. If we choose the maximum a posteriori (MAP) solution

$$\hat{\theta} = \arg \max_{\theta} \{p(z|\theta) p(\theta)\} = \arg \min_{\theta} \{-\ln(p(z|\theta)) + \ln(p(\theta))\}$$

(7)

we arrive at

$$\hat{\theta} = \arg \min_{\theta} \{J_{MAP}(\theta)\}$$

(8)

where $J_{MAP}(\theta) = J(\theta) + \lambda R(\theta)$ with $J(\theta)$ is given by (4), $R(\theta) \propto -\ln(p(\theta))$ and $\lambda > 0$ is a hyperparameter depending on the variance of the errors $v_{\epsilon}$ and some parameters of the prior law.

Comparing the expressions in (4) and (8), we see that the MAP solution can be considered as a regularized solution to the standard NLS solution with $\lambda$ as the regularization parameter. In our case, as we have chosen flat priors, $\ln p(\theta) = 0$, and so the criterion to optimize (8) becomes the same as criterion (4).

There are many other Bayesian estimators, for example, the posterior mean, but its inference needs MCMC sampling methods. The computational cost of such methods is huge for this kind of application. We stand here with the more classical techniques of Chi-squared, AIC (Akaike Information Criterion), and BIC (Bayesian Information Criterion) for which standard techniques of computations are, in general, included in many optimization tools available, for example, in Python language packages [12]. A Jupyter notebook of all the simulations of this paper will be available after its publication on github.

3.5. Adaptation of the Inversion Methods to Available Data

Now we confront the modeling of the previous section with accessible data that can not be a whole $[S, E, I, R](t)$ time series but a subset or alternative combination of them. As we work with the COVID-19 data from the Johns Hopkins University at GitHub [13], the available data are:

- Confirmed (C)
- Dead (D)
Recovered (R) as cumulative integer-valued time series for days from 22 January 2020. Note that these values are absolute numbers, not relative to a total population. Note that the unconfirmed cases and also the Susceptibles are not accessible at all, whereas the Confirmed contain the Dead and the Recovered from earlier days. For example, using the Johns Hopkins data, they supply C, D, and R individually while not stating the most necessary variables \( E = C - D - R \), \( S = N - (C + R + D) \). We currently use a model that works with the Johns Hopkins data while not having the ability to use \( S \) in the SEIR model. Since the SEIR model does not distinguish between recoveries and deaths, we set the following relations:

\[
\begin{align*}
\text{cR}_{\text{Model}} & \iff R_{\text{Cov}}, \quad I_{\text{Model}} \iff C_{\text{Cov}}, \\
E_{\text{Model}} & \iff C_{\text{Cov}} - D_{\text{Cov}} - R_{\text{Cov}}, \quad S_{\text{Model}} \iff N - (C_{\text{Cov}} + D_{\text{Cov}} + R_{\text{Cov}}).
\end{align*}
\]

Now, returning to the SEIR model, we see the subsequent relations:

\[
\begin{align*}
\gamma &= \frac{R}{I} \approx \frac{R_{n+1} - R_n}{I_n} = \gamma_n, \\
\sigma &= \frac{I \gamma + I}{E} \approx \frac{(R + C)_{n+1} - (R + C)_n}{E_n} = \sigma_n, \\
q &= \frac{\beta S}{N} = \frac{\dot{E} + \sigma E}{I} \approx \frac{(2C - D)_{n+1} - (2C - D)_n}{I_n} = q_n
\end{align*}
\]

defining time series \( \gamma_n \) and \( q_n \) that model \( \gamma \) and \( q = \beta S_{\text{Model}} / N \) without knowing SEIR. This is equivalent to the model

\[
\begin{align*}
\gamma_n I_n &= R_{n+1} - R_n, \\
\sigma_n E_n &= (R + C)_{n+1} - (R + C)_n, \\
q_n I_n &= (2C - D)_{n+1} - (2C - D)_n
\end{align*}
\]

Then, for the estimation of the parameters with its adaptation Figure 4 indicates the conversion from S-E-I-R to C-D-R.

![Figure 4. Conversion from S-E-I-R to C-D-R.](image)

This straightforward model has at least two drawbacks:

- It assumes that the number of the population \( N \) does not change during the epidemic;
- It does not account for possible context modification of the sickness (fixed parameters during the whole analysis).

Numerous extensions are supplied to enhance it. The primary extension of the model with a further block is utilized to account for the specific fact that the whole number of the population is not fixed [7,14–22]. The second categories propose time-dependent parameters [23–26].
4. Constructing Predictions

To use the proposed model to perform prediction, the process we propose is, first estimate the parameters of the model from the real available data using the method presented in Figure 5 using either the nonlinear least squares or Bayesian methods. Then, once the model has been identified from the data, we can use it for prediction. To be more precise, we first consider that we have real data over the horizon \( 1, \ldots, T \), where 1 denotes the first day of the estimation period and \( T \) the last one. We can initialize the state of the system with \( \mathbf{x}(t_0, t_0) = 1 \) from the real data. Then, with an initial value for the parameters, we can compute \( \mathbf{x}(t > t_0) = [S(t), E(t), I(t), R(t)] \), then \( \mathbf{z}(t) = [C(t), D(t), R(t)] \) and finally the value of criterion \( J(\theta) \), which can then be iteratively optimized to obtain the final estimation of the parameters. When the parameters are obtained, we can use them to perform any prediction of state \( \mathbf{x}(t) \) and then \( \mathbf{z}(t) \) for any time interval, particularly for any \( t > T \).

Figure 5. Prediction results with the SEIR model (time-varying parameter).

Analysis of COVID-19 Data for France

Daily overall incidences of COVID-19 cases were collected from information sources on github. The records showed the simplest date of the first positive Corona test, date of release from hospital, and typically, date of first negative Corona test. Cases that did not arise via transmission started at \( t = 0 \); and cases that arose via transmission were marked on the day of the first positive Corona test. The optimal parameter values obtained using the procedure given in Section 3.3 are \( \beta = 6.52731360, \sigma = 0.00304030, \gamma = 0.57801349 \). In Figure 5, we present the results of simulations of the variety of S, I, and R listed through time for France with a fixed-parameter model.

However, the fixed-parameter model could not give satisfactory results on real data, so we used a time-dependent \( \beta \) parameter model. Specifically, a straightforward periodic function \( \beta(t) \):

\[
\beta(t) = a(1 + b \sin(\omega t)),
\]

(14)

Parameter \( a \) denotes the baseline or average transmission rate, \( \omega \) is the period of the forcing, and \( b \) represents the amplitude of the disease.

In Figures 5 and 6, we present the results of prediction of the variety of S, I, and R listed through time for France with a constant and a varying parameter model.

Figure 6. Prediction results with the SEIR model (constant parameters).

Then, we used Equation (9) to perform the conversion of S-E-I-R to C-D-R. Figure 7 indicates this conversion.
Figure 7. Left: Conversion from estimated S-E-I-R to C-D-R (time-varying parameter). Right: Conversion from S-E-I-R to C-D-R on real data.

5. Conclusions and Discussion

The SEIR model we used here appears to be useful for understanding the propagation of the virus. Two main difficulties appeared:

- First, the available data (C,D,R) could not determine in a unique way the S,E,I,R. In fact, going from (S,E,I,R) to (C,D,R) is unique, but the inverse is not. Therefore, we had to be careful when directly using the original SEIR model. We explained this in the previous section when we wanted to estimate the parameters of the model from the real (C,D,R) data.

- The second difficulty appeared when we saw that the fixed parameters model could not predict the real data well. This is explained due to the fact that this simple model does not account for many real situations. In particular, a fixed-parameter model does not correspond to reality. Many works have been performed to overcome these difficulties. Between them, we chose a simple method in which parameter $\beta$ varies over time. We used a periodic time dependent for this parameter, as it was also used by other researchers [23–26]. Finally, using appropriate values for the parameters $a$ and $b$ of this time-dependent model, we could obtain good predictions with real data.

Author Contributions: Conceptualization, E.T. and A.M.-D.; methodology, E.T. and A.M.-D.; software, E.T. and A.M.-D.; validation, A.M.-D.; formal analysis, E.T. and A.M.-D.; investigation, E.T. and A.M.-D.; resources, E.T. and A.M.-D.; data curation, E.T.; writing—original draft preparation, E.T.; writing—review and editing, A.M.-D.; visualization, E.T. and A.M.-D.; supervision, A.M.-D.; project administration, A.M.-D.; funding acquisition, A.M.-D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The presented data is available at https://github.com/CSSEGISandData/COVID-19.

Conflicts of Interest: The authors declare no conflict of interest.

References


25. Araujo, M.B.; Naimi, B. Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate. *medRxiv* **2020**. [CrossRef]