symmetry

Article

The Stochastic Approach for SIR Epidemic Models: Do They Help to Increase Information from Raw Data?

Alessandro Borri [1], Pasquale Palumbo [2,*] and Federico Papa [1]

[1] Institute for Systems Analysis and Computer Science "A. Ruberti" (IASI- CNR), 00185 Rome, Italy
[2] Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza, 20126 Milan, Italy
* Correspondence: pasquale.palumbo@unimib.it

Abstract: The recent outbreak of COVID-19 underlined the need for a fast and trustworthy methodology to identify the features of a pandemic, whose early identification is of help for designing non-pharmaceutical interventions (including lockdown and social distancing) to limit the progression of the disease. A common approach in this context is the parameter identification from deterministic epidemic models, which, unfortunately, cannot take into account the inherent randomness of the epidemic phenomenon, especially in the initial stage; on the other hand, the use of raw data within the framework of a stochastic model is not straightforward. This note investigates the stochastic approach applied to a basic SIR (Susceptible, Infected, Recovered) epidemic model to enhance information from raw data generated in silico. The stochastic model consists of a Continuous-Time Markov Model, describing the epidemic outbreak in terms of stochastic discrete infection and recovery events in a given region, and where independent random paths are associated to different provinces of the same region, which are assumed to share the same set of model parameters. The estimation procedure is based on the building of a loss function that symmetrically weighs first-order and second-order moments, differently from the standard approach that considers a highly asymmetrical choice, exploiting only first-order moments. Instead, we opt for an innovative symmetrical identification approach which exploits both moments. The new approach is specifically proposed to enhance the statistical information content of the raw epidemiological data.

Keywords: SIR models; parameter identification; stochastic approach

Citation: Borri, A.; Palumbo, P.; Papa, F. The Stochastic Approach for SIR Epidemic Models: Do They Help to Increase Information from Raw Data? *Symmetry* **2022**, *14*, 2330. https://doi.org/10.3390/sym14112330

Academic Editors: Jinyu Li and Mariano Torrisi

Received: 8 October 2022
Accepted: 1 November 2022
Published: 6 November 2022

1. Introduction

Since the very beginning of COVID-19 diffusion in early 2020, effective countermeasures to limit the spread of the disease have been non-pharmaceutical interventions such as lockdown, social distancing and limitation of economic activities, with an unavoidable worsening of social life and wealth [1–3]. Most of such unpopular decisions have been taken according to scientific committees striving to understand the real gravity of the diffusion and to forecast how these decisions would impact with the spreading of the pandemic. To this end, it has been soon clear that mathematical models could be of great help to identify the features of the COVID-19 spread as well as to make predictions and to support decisions [4–18]. Indeed, nowadays there can be found many models, ranging from the basic SIR epidemic model that accounts for Susceptible, Infected and Recovered individuals, to finer, multi-compartmental models accounting for the Exposed, Hospitalized, Vaccinated, etc. [19–21]. However, most of the existing models share the feature of being deterministic; therefore, they are not able to capture the inherent randomness of the phenomenon under investigation and to properly characterize the uncertainty of future possible scenarios.

In [22], we presented a first investigation on an alternative identification approach based on the Chemical Master Equation (CME) modeling framework, which was used to describe the COVID-19 epidemics in Italy. In that paper, the model parameters of the resulting stochastic SIR model, i.e., the relative infectivity rate and the per capita removal rate, were the target of the estimation approach. The proposed results exploited second-order moments and showed the potential effectiveness of the proposed approach by applying it to real data on the number of notified cases coming from the National Institute of Health of Italy (Istituto Superiore di Sanità, ISS) [23]. The CME modeling framework was originally introduced to describe the dynamics of chemical reaction networks (see for instance [24]), but the method is frequently used in biological applications, especially when the dynamics of the considered population is inherently stochastic, such as for epidemics [25], cellular biochemical systems [26] and growing tumor populations [27].

Differently from [22], in this work, we aim at investigating the estimation approach proposed in [22] on a theoretical ground, assuming to exploit in silico data provided by the model simulations. This allows us to stress the effectiveness of the proposed research with respect to different sets of pandemic parameters, not necessarily anchored to the ones related to the COVID-19 disease. According to the idea carried out in [22], we assume a given country affected by the pandemic divided in districts. Each district develops the pandemic according to the same model parameters, therefore providing independent random paths sampled from the same stochastic process: to this end, a Continuous-Time Markov Chain is exploited to deal with the discrete random events providing one more infected or one more recovered [28]. Random paths are simulated according to the Stochastic Sampling Algorithm (SSA) [29]. Different sets of model parameters provide different epidemic scenarios where to test the proposed methodology. First- and second-order moments are computed from in silico random paths, and they are compared to the theoretical first- and second-order moments computed from the CTMC theory [30]. As a matter of fact, an in silico model with known parameters allows one to evaluate more precisely the added value of second-order moments in the estimation. This is not possible in the presence of real data, because the real pandemic parameters would be unknown. For the moment computation, due to the nonlinear fashion of the model propensities (infected are supposed to occur at a rate proportional to both infected and susceptible individuals, as usual), we resort to the linear approximation, which is usually suggested by the conjecture that the pandemic shows very low infected individuals with respect to susceptibles. It is worth noticing that according to standard deterministic epidemic models, data coming from different sub-regions would be put together and therefore treated as averages.

Numerical results show that if only infectious data are available, the use of the second-order moment beyond the first-order one sensibly improves the parameter estimation, obtaining finite estimation errors when separate estimates of the relative infection and removal rate are attempted. Indeed, the first-order moment expression of the number of infected patients only depends on the difference between the model parameters, definitely making them not individually identifiable from the mean value of infected.

Conversely, when both infectious and removed data are available, the identification of the SIR parameters is always meaningful and benefits from a possible larger number of experimental samples, as happens when more provinces are considered. The removal rate is better estimated by the second-order procedure, differently from the infection rate, whose estimate worsens (on average). However, the model trajectories are scarcely sensitive to this estimation error, since both first and second-order experimental data are accurately reproduced by the second-order fitting. Furthermore, confidence intervals of both parameters are also reduced in this case, which suggest higher reliability of the second-order method in real scenarios.

2. A Stochastic SIR Model

In the present work, our investigation focuses on the early spread of the epidemics, which is a relatively short time horizon where the system is open-loop (no restriction measures have not been taken yet); in particular, we focus on an identification problem in a situation of scarce data; for this reason, we decided to keep our modeling setting as simple as possible, and we selected the classical SIR model [31] as the more appropriate framework for our study, by neglecting for example the exposition dynamics (see e.g., [32]) and the presence of delayed transmission and lags/dead times in the effects of control actions (see e.g., [33] for a comprehensive review on this topic, not limited to biological systems).

According to the standard SIR modeling choice, the whole population is divided into Susceptibles (the ones prone to get infected), Infected (the ones actually infected and responsible for further infections) and Recovered (the ones healed and no more infectious, or death). The whole population (Susceptibles + Infected + Recovered) is not supposed to vary for birth/death processes different from the ones provoked by the pandemics: its constant amount is denoted by N. Due to the constraint on the whole population, the state of the system evolves, thus, according to a pair of components. In this work, we consider the *currently infectious* patients I and the *removed* individuals R, and their dynamics is described by exploiting the formalism of Continuous-Time Markov Models (CTMC) [28]. After deriving the evolution of the pair I, R, the dynamical behavior of the number of individuals susceptible to the infection S is straightforwardly obtained by the relation

$$S = N - I - R. \tag{1}$$

Note that according to the chosen modeling setting, the state variables are actually countable variables, which is a more appropriate representation than the unrealistic continuity assumption made by the classical deterministic SIR formulation.

Going into the mathematical details, the state variation of I and R, as well as of the other derived variables, is due to a pair of discrete stochastic events, namely the infection, E_1, and recovery (by healing or death), E_2, events:

$$\begin{aligned} E_1 &: \begin{cases} I \to I+1, \\ S \to S-1, \end{cases} \qquad w_1 = \frac{\beta}{N}(N-I-R)I, \\ E_2 &: \begin{cases} I \to I-1, \\ R \to R+1, \end{cases} \qquad w_2 = \gamma I, \end{aligned} \tag{2}$$

where w_1 and w_2 are the propensities of the modeled events (i.e., the time derivatives of the transition probabilities, time^{-1}), β is the relative infectivity (depending on the infection probability of an infected-susceptible contact and on the contact frequency), and γ is the per capita rate of removal (healing plus death) from the infection.

The dynamical equation of the first- and second-order moments cannot be directly written in a closed form from the propensity expressions due to the nonlinear form of w_1 [30]. In order to obtain closed forms of higher-order moments, we consider the simplifying assumption

$$S \approx N,$$

which is actually a realistic assumption when the epidemic is at the beginning of its outbreak or when it is controlled, so that the total cases are sensibly lower than the whole population number. This assumption simplifies the expression of w_1 as

$$w_1 \simeq \beta I,$$

that becomes actually linear with respect to the state variables. The simplified expression of w_1 allows one to write the first-order moment equations as

$$\begin{aligned} \frac{d\langle I(t)\rangle}{dt} &= (\beta-\gamma)\langle I(t)\rangle, \\ \frac{d\langle R(t)\rangle}{dt} &= \gamma\langle I(t)\rangle, \end{aligned} \tag{3}$$

that provide the explicit solutions

$$
\begin{aligned}
\langle I(t)\rangle &= \langle I(0)\rangle e^{(\beta-\gamma)t}, \\
\langle R(t)\rangle &= \langle R(0)\rangle + \frac{\gamma}{\beta-\gamma}\left(e^{(\beta-\gamma)t}-1\right)\langle I(0)\rangle.
\end{aligned} \tag{4}
$$

The ODE system (3), representing the first-order moments dynamics when $S/N \approx 1$, gives back the standard structure of the deterministic SIR formulation [31].

Indeed, these first-order Equations (3) and (4) are usually exploited to identify the relative infectivity rate β and the per capita removal rate γ according to the measured infected or cumulative cases available from local or national Institutes of Health. In [22], we proposed a way to exploit the inherent randomness of the events and the straightforward stochastic SIR model in order to gain information from the available data. The key point is to gather more random paths from a given scenario. To this end, we assume to divide a given country (affected by a pandemic) into m districts. Each district has its own population (N_i, $i = 1, \dots, m$) and faces the same pandemic whose spreading depends on a unique set of model parameters: in other words, each district shares the same pair (β, γ). By keeping the approximation $S \approx N_i$, $i = 1, \dots, m$, each district shares as well also the approximated propensities. A further hypothesis is that any district is isolated with any other. This way, the stream of data acquired from the districts is independent random paths, according to which higher-order moments can be computed and exploited to increase the reliability of the estimates.

Therefore, according to [30], due to the approximation $S_i \approx N_i$, we write the second-order moments in closed forms as

$$
\begin{aligned}
\frac{d\left\langle I^2(t)\right\rangle}{dt} &= \left\langle \left((I(t)+1)^2 - I^2(t)\right)w_1\right\rangle + \left\langle \left((I(t)-1)^2 - I^2(t)\right)w_2\right\rangle, \\
\frac{d\langle I(t)R(t)\rangle}{dt} &= \left\langle \left((I(t)+1)R(t) - I(t)R(t)\right)w_1\right\rangle \\
&\quad + \left\langle \left((I(t)-1)(R(t)+1) - I(t)R(t)\right)w_2\right\rangle, \\
\frac{d\left\langle R^2(t)\right\rangle}{dt} &= \left\langle \left((R(t)+1)^2 - R^2(t)\right)w_2\right\rangle,
\end{aligned} \tag{5}
$$

that is, after computations:

$$
\begin{aligned}
\frac{d\left\langle I^2(t)\right\rangle}{dt} &= 2(\beta-\gamma)\left\langle I^2(t)\right\rangle + (\beta+\gamma)\langle I(t)\rangle, \\
\frac{d\langle I(t)R(t)\rangle}{dt} &= \gamma\left\langle I^2(t)\right\rangle + (\beta-\gamma)\langle I(t)R(t)\rangle - \gamma\langle I(t)\rangle, \\
\frac{d\left\langle R^2(t)\right\rangle}{dt} &= 2\gamma\langle I(t)R(t)\rangle + \gamma\langle I(t)\rangle,
\end{aligned} \tag{6}
$$

that give the explicit solutions

$$
\begin{aligned}
\left\langle I^2(t)\right\rangle &= e^{(\beta-\gamma)t}\left(e^{(\beta-\gamma)t}\left\langle I^2(0)\right\rangle + \tfrac{\beta+\gamma}{\beta-\gamma}\left(e^{(\beta-\gamma)t}-1\right)\langle I(0)\rangle\right), \\
\langle I(t)R(t)\rangle &= e^{(\beta-\gamma)t}\left(\langle I(0)R(0)\rangle + \tfrac{\gamma}{\beta-\gamma}\left(e^{(\beta-\gamma)t}-1\right)\left\langle I^2(0)\right\rangle\right. \\
&\qquad \left. + \tfrac{\gamma}{\beta-\gamma}\left(\tfrac{\beta+\gamma}{\beta-\gamma}\left(e^{(\beta-\gamma)t}-1\right) - 2\beta t\right)\langle I(0)\rangle\right), \\
\left\langle R^2(t)\right\rangle &= \left\langle R^2(0)\right\rangle + \tfrac{2\gamma}{\beta-\gamma}\left(e^{(\beta-\gamma)t}-1\right)\langle I(0)R(0)\rangle \\
&\quad + \tfrac{\gamma^2}{(\beta-\gamma)^2}\left(e^{2(\beta-\gamma)t} - 2e^{(\beta-\gamma)t} + 1\right)\left\langle I^2(0)\right\rangle \\
&\quad + \tfrac{\gamma}{(\beta-\gamma)^2}\left(\tfrac{\gamma(\beta+\gamma)}{\beta-\gamma}e^{2(\beta-\gamma)t} + (\beta+\gamma)e^{(\beta-\gamma)t} - 4\beta\gamma t e^{(\beta-\gamma)t} - \tfrac{\beta(\beta+\gamma)}{\beta-\gamma}\right)\langle I(0)\rangle.
\end{aligned} \tag{7}
$$

Remark 1. *The initial conditions $\langle I(0)\rangle$, $\langle R(0)\rangle$, $\langle I^2(0)\rangle$, $\langle I(0)R(0)\rangle$, and $\langle R^2(0)\rangle$ have to be considered as unknown to be identified as well as the pair (β,γ). However, in case of the very beginning of the epidemic spread, the initial recovered could be thought of as a deterministic value equal to 0, so that first- and second-order moment equations involving $R(0)$ simplify as follows:*

$$\langle R(t)\rangle = \frac{\gamma}{\beta-\gamma}\left(e^{(\beta-\gamma)t}-1\right)\langle I(0)\rangle \tag{8}$$

$$\begin{aligned}\langle I(t)R(t)\rangle = e^{(\beta-\gamma)t}\Bigg(&\tfrac{\gamma}{\beta-\gamma}\left(e^{(\beta-\gamma)t}-1\right)\langle I^2(0)\rangle \\ &+\tfrac{\gamma}{\beta-\gamma}\left(\tfrac{\beta+\gamma}{\beta-\gamma}\left(e^{(\beta-\gamma)t}-1\right)-2\beta t\right)\langle I(0)\rangle\Bigg),\end{aligned} \tag{9}$$

$$\begin{aligned}\langle R^2(t)\rangle = &\tfrac{\gamma^2}{(\beta-\gamma)^2}\left(e^{2(\beta-\gamma)t}-2e^{(\beta-\gamma)t}+1\right)\langle I^2(0)\rangle \\ &+\tfrac{\gamma}{(\beta-\gamma)^2}\left(\tfrac{\gamma(\beta+\gamma)}{\beta-\gamma}e^{2(\beta-\gamma)t}+(\beta+\gamma)e^{(\beta-\gamma)t}-4\beta\gamma t e^{(\beta-\gamma)t}-\tfrac{\beta(\beta+\gamma)}{\beta-\gamma}\right)\langle I(0)\rangle.\end{aligned} \tag{10}$$

3. The Identification Procedure

Let us assume that an epidemic is spreading throughout a given country and that the stochastic modeling framework proposed above represents the main features of the disease. Let us also consider the different regions belonging to the country, assuming that their different geographical positions and local cultures may affect the value of the parameter pair (β,γ). However, the epidemiological data coming from different sub-regions, the aforementioned districts that we will call provinces or counties hereafter, for simplicity, can be seen as different realizations of the regional CTMC. In other words, each province initially shares the same pair (β,γ) of the region, and its data can be regarded as a random path of the region where it belongs. Obviously, fluxes of people within provinces are supposed to be neglected.

Differently from [22], where we assumed that the daily number of cumulative cases was the only available data coming from the provinces, here, we assume two different scenarios depending on the information level:

(a) We assume knowledge of periodic measurements of currently *infectious* individuals;
(b) We assume knowledge of periodic measurements of currently *infectious and removed* individuals.

More formally, in this paper, we focus the attention on a single region with $m \in \mathbb{N}$ provinces, which are tracked for consecutive K samples, and for which the data set of currently infectious individuals $\{i_{l,k}\}_{k=0}^{K-1}$, $l = 1,\ldots,m$, is always available and the data set of removed individuals $\{r_{l,k}\}_{k=0}^{K-1}$, $l = 1,\ldots,m$, is available only in the scenario (b) defined above; those data sets denote the measurements of infectious and removed for province l at discrete time $t = k\Delta$, $k = 0,...,K-1$, where Δ is usually a 1-day interval. We assume $i_{l,0} \geq 1$ for any province l, i.e., there is already one infectious in each province at the initial time. This is necessary to obtain meaningful random paths, since the condition $I = 0$ characterizes the absorbing states of the CTMC, preventing the occurrence of both infections and removal events (since $I = 0$ implies $w_1 = w_2 = 0$). We also assume $r_{l,0} = 0$ for any province l, which is compatible with the early spreading of the epidemic.

Statistical moments (up to the second order) at each time are readily computed from data as

$$\overline{I_k} = \frac{1}{m}\sum_{l=1}^{m} i_{l,k}, \qquad \overline{R_k} = \frac{1}{m}\sum_{l=1}^{m} r_{l,k}, \tag{11}$$

$$\overline{I_k^2} = \frac{1}{m}\sum_{l=1}^{m} i_{l,k}^2, \qquad \overline{IR_k} = \frac{1}{m}\sum_{l=1}^{m} i_{l,k}r_{l,k}, \qquad \overline{R_k^2} = \frac{1}{m}\sum_{l=1}^{m} r_{l,k}^2. \tag{12}$$

For the following developments, we also define the quantities $I_k := \langle I(k) \rangle$, $R_k := \langle R(k) \rangle$, $I_k^2 := \langle I^2(k) \rangle$, $IR_k := \langle I(k)R(k) \rangle$, $R_k^2 := \langle R^2(k) \rangle$, which are the explicit moment Expressions (4) and (7) computed at general time $t = k\Delta$ and starting from the initial conditions I_0, R_0, I_0^2, IR_0, R_0^2, where we set $R_0 = IR_0 = R_0^2 = 0$, to make the theoretical moment Expressions (4) and (7) consistent with the data assumption $r_{l,0} = 0, \forall l$.

3.1. Scenario (a): Parameter Estimation from Infectious Data

In this scenario, we assume that only the expressions of $\overline{I_k}$ and $\overline{I_k^2}$ from (11) and (12) are known. We can define the following running costs, accounting for the Euclidean distance between statistical and theoretical first- and second-order moments of the infectious at each time $k = 0, \ldots, K-1$:

$$J_k^1(\beta, \gamma, I_0) = \frac{\left(I_k - \overline{I_k}\right)^2}{\overline{I_k}^2}, \qquad J_k^2(\beta, \gamma, I_0, I_0^2) = \frac{\left(I_k^2 - \overline{I_k^2}\right)^2}{\overline{I_k^2}^2}, \tag{13}$$

where we highlighted the dependence of J_k^1 and J_k^2 on the parameters and on the initial conditions to estimate, i.e., I_0 and I_0^2.

3.2. Scenario (b): Parameter Estimation from Infectious and Removed Data

In this scenario, all the statistical moment expressions in (11) and (12) are known, so the running costs at all times k will also account for the first- and second-order moments of the removed and for the joint second-order moment of infectious and removed:

$$J_k^1(\beta, \gamma, I_0) = \frac{\left(I_k - \overline{I_k}\right)^2}{\overline{I_k}^2} + \frac{\left(R_k - \overline{R_k}\right)^2}{\overline{R_k}^2}, \tag{14}$$

$$J_k^2(\beta, \gamma, I_0, I_0^2) = \frac{\left(I_k^2 - \overline{I_k^2}\right)^2}{\overline{I_k^2}^2} + \frac{\left(IR_k - \overline{IR_k}\right)^2}{\overline{IR_k}^2} + \frac{\left(R_k^2 - \overline{R_k^2}\right)^2}{\overline{R_k^2}^2}. \tag{15}$$

Note that in spite of the different information set with respect to the scenario (a), the Expressions (14) and (15) of the running costs in the scenario (b) depend on the same parameters, since only the non-zero moment initial conditions I_0 and I_0^2 need to be estimated.

3.3. The Optimization Problem

The goal of the identification step is to minimize the following cost, which is a weighted normalized version of the root mean square error:

$$J_\alpha(\beta, \gamma, I_0, I_0^2) = \sqrt{\frac{1}{K} \sum_{k=0}^{K-1} \alpha J_k^1(\beta, \gamma, I_0) + (1 - \alpha) J_k^2(\beta, \gamma, I_0, I_0^2)}, \tag{16}$$

where $\alpha \in [0, 1]$ weighs the relative importance of tracking effectively first-order with respect to second-order moments, and the running costs are defined according to the chosen scenario, i.e., Equation (13) for scenario (a) and Equations (14) and (15) for scenario (b). Looking at the family of functions defined by (16), the standard approach consists of considering a highly asymmetrical choice, exploiting only first-order moments ($\alpha = 1$). Instead, we opt for an innovative symmetrical identification approach which exploits both moments (first and second order), choosing $\alpha = 0.5$ that symmetrically poses in between with respect to the use of them. In other words, the proposed estimation procedure is based on a loss function that symmetrically balances the information content provided by first- and second-order moments.

For a fixed α, the optimization problem returns the fitted parameters $\hat{\beta}$ and $\hat{\gamma}$ and the initial conditions $\hat{I}_0$ and $\hat{I}_0^2$ achieving the minimum cost $\hat{J}_\alpha = J_\alpha(\hat{\beta}, \hat{\gamma}, \hat{I}_0, \hat{I}_0^2)$, i.e., satisfying

$$(\hat{\beta}, \hat{\gamma}, \hat{I}_0, \hat{I}_0^2) = \arg \min_{\substack{(\beta,\gamma,I_0,I_0^2)\in\mathbb{R}_+^4 \\ \text{subject to Equations (4),(7)}}} J_\alpha(\beta, \gamma, I_0, I_0^2), \tag{17}$$

where J_α is defined in (16) and where we included obvious positivity constraints on the parameters and on the initial conditions, which are required to belong to the positive orthant $\mathbb{R}_+^4$ of the $four$-dimensional Euclidean space.

4. Results and Discussion

The identification method described in the previous section has been applied to an ideal setting, where data have been generated according to the stochastic CTMC model (2). This is a substantial difference with the work [22], where methods have been applied to the real data of the COVID-19 pandemic in Italy, because the goal of the present work is a more rigorous validation of the identification method in a general epidemic scenario.

In the following simulations, we consider a region with a possibly variable number of provinces of random sizes (N_i, $i = 1, \dots, m$), which are independently sampled according to a discrete uniform distribution in the interval $[25 \dots 10^4; 75 \dots 10^4]$, where we denoted by $[a; b] := [a, b] \cap \mathbb{N}$ an interval on the line of natural numbers. This results in random sizes which are comparable to many Italian provinces. Similarly, the initial infectious are independently sampled according to a discrete uniform distribution in the interval $[1; I_{0,\max}]$, where we set $I_{0,\max} = 10$. Each random path associated to a province of the simulated region is built according to the Stochastic Sampling Algorithm (SSA) [29].

We choose an observation interval of $K = 14$ days, which is a period compatible with the assumption of early-stage epidemics (see also [8]), inducing an approximate exponential increase of the number of infectious, by virtue of the simplifying assumption $S \approx N$.

For each of the two scenarios described in the previous section, we will consider two cases:

(1) Best fitting limited to the first-order moments, corresponding to the case $\alpha = 1$ in the index $J_\alpha(\beta, \gamma, I_0, I_0^2)$ in (16); this case (optimization of $J_1(\beta, \gamma, I_0, I_0^2)$) will be shortly referred to as *best first-order fit*.

(2) Best fitting of a balanced combination of first- and second-order moments corresponding to the case $\alpha = 0.5$ in the index $J_\alpha(\beta, \gamma, I_0, I_0^2)$ in (16); this case (optimization of $J_{0.5}(\beta, \gamma, I_0, I_0^2)$) will be shortly referred to as *best second-order fit*, in the sense that it exploits (in a balanced way) the moments up to the second order.

All the simulations have been performed in the MATLAB® environment, exploiting the function lsqcurvefit" (Optimization Toolbox).

4.1. Scenario (a): Parameter Estimation from Infectious Data

In scenario (a) of parameter estimation from infectious data, we consider $M = 100$ simulations of a region with $m = 5$ provinces, with each simulation associated to a point of the grid $(\beta, \gamma) \in \{(0.05k, 0.005j) : k, j \in [1; 10]\}$.

The simulation results show that in this scenario, the estimation of (β, γ) obtained by the best second-order fitting procedure ($J_{0.5}$ minimization) outperforms the analogous estimation achieved by the best first-order fitting procedure (J_1 minimization). In particular, if reasonable upper bounds of the parameters are not provided to the numerical optimizer, the best first-order procedure returns estimated values that can be arbitrarily large in general. Instead, the estimation of the difference $(\beta - \gamma)$ is quite accurate in both cases.

The described behavior can be readily justified by looking at the expression of $\langle I(t) \rangle$ and $\langle I^2(t) \rangle$ from (4) and (7), respectively:

$$\langle I(t) \rangle = \langle I(0) \rangle e^{(\beta-\gamma)t}, \tag{18}$$

$$\left\langle I^2(t) \right\rangle = e^{(\beta-\gamma)t} \left(e^{(\beta-\gamma)t} \left\langle I^2(0) \right\rangle + \frac{\beta+\gamma}{\beta-\gamma} \left(e^{(\beta-\gamma)t} - 1 \right) \langle I(0) \rangle \right). \tag{19}$$

As a matter of fact, the first-order moment expression (18) only depends on the difference $(\beta - \gamma)$, which makes the two parameters not individually identifiable by using only this equation. On the other hand, although the second-order moment expression in (19) formally depends on both $(\beta - \gamma)$ and $(\beta + \gamma)$, the actual value of $\langle I^2(t) \rangle$ is scarcely affected by the sum $(\beta + \gamma)$, multiplying the value $\langle I(0) \rangle$, while it is much more sensitive to the difference $(\beta - \gamma)$, which affects the growth rate of the exponential terms. This leads to a rather large relative error (more than 30%, on average, in our simulations), even in the best second-order fit of parameter γ.

As an example, the fitting results for the realization related to the "true values" $(\beta, \gamma) = (0.3, 0.03)$ are shown in Figure 1. The figure shows the comparison between the sample moments (blue circles), which are obtained by applying Equations (11) and (12) to the raw data, and the model predictions (lines), obtained from Equations (4) and (7), using the parameter values estimated with $\alpha = 1$ (red dashed line) or $\alpha = 0.5$ (yellow solid line). As expected, only the best second-order fit (yellow solid line) is able to track both the sequence of first- and second-order data, I_k and I_k^2, respectively. In this simulation, the estimates of the parameter pair are $(\hat{\beta}_1, \hat{\gamma}_1) = (1.695, 1.402)$ for the best first-order fit and $(\hat{\beta}_2, \hat{\gamma}_2) = (0.262, 0.022)$ for the best second-order fit.

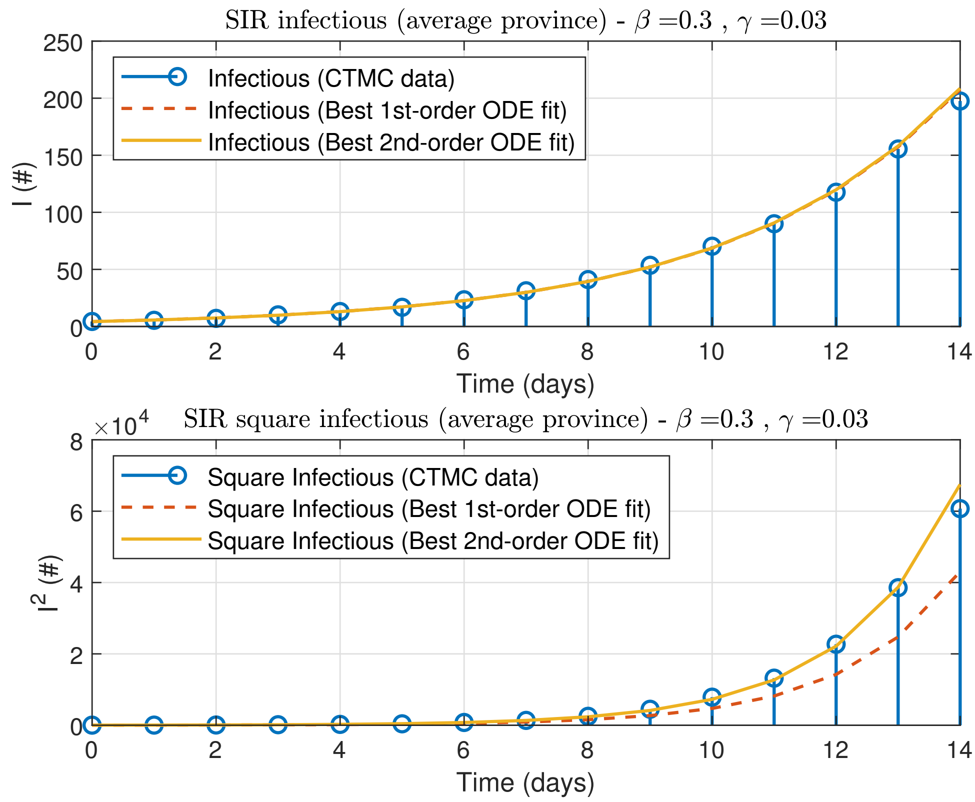

Figure 1. Scenario (a): estimation from infectious data. Fitting of the first-order moment (**top panel**) and of the second-order moment (**bottom panel**) of the infectious individuals from a region including $m = 5$ provinces, in the case $(\beta, \gamma) = (0.3, 0.03)$: CTMC data (blue circles), best first-order fit (red dashed line), best second-order fit (yellow solid line).

4.2. Scenario (b): Parameter Estimation from Infectious and Removed Data

In scenario (b) of parameter estimation from infectious and removed data, we consider a more extensive simulation setup, namely $M = 100$ simulations in two distinct cases of a region with $m = 5$ and $m = 10$ provinces, and the same grid of scenario (a). The results are summarized in Table 1, evaluated by means of the mean square error (MSE) of the best first- and second-order fit and by the length of the corresponding estimated 95% confidence intervals (ci). Confidence intervals have been estimated by means of the Matlab function nlparci", exploiting the Jacobian matrix computed numerically by function lsqcurvefit" and evaluated at the estimation point.

Table 1. Scenario (b). Estimation results evaluated in terms of mean square error (MSE) of the best first-order fit and second-order fit and by the length of the corresponding estimated 95% Confidence Intervals (CI), with $m = 5, 10$ provinces.

m	β MSE (1st order)	β MSE (2nd order)	β CI (1st order)	β CI (2nd order)
5	9.99	18.08	0.17	0.13
10	4.91	8.37	0.08	0.03
m	γ MSE (1st order)	γ MSE (2nd order)	γ CI (1st order)	γ CI (2nd order)
5	23.01	17.62	0.014	0.013
10	11.14	5.30	0.012	0.008

Overall, the first remark is that the two parameters β and γ are now individually identifiable, since the complete sets of theoretical moments (4), (7) and statistical moments (11) and (12) are now exploited. The two sub-cases do not exhibit particular qualitative differences in the estimation when varying the number of provinces m, while from the quantitative viewpoint, there is a net reduction of the MSE in the case $m = 10$ with respect to the case $m = 5$, which is probably due to the double number of samples in the case $m = 10$.

By keeping fixed the number of provinces, the parameter γ is fitted better (lower MSE) by the second-order procedure ($J_{0.5}$ minimization) with respect to the first-order one (J_1 minimization). A possible explanation for such behavior is that the second-order moment equation in (6) introduces another dynamics $\langle R^2(t) \rangle$ (in addition to $\langle R(t) \rangle$) depending on γ only, which is probably useful for an improved estimation of this parameter.

Initially, the parameter β is fitted better, on average, by the first-order procedure. This drawback of the second-order procedure is compensated by a higher robustness of the estimation of both β and γ, expressed in terms of the 95% confidence intervals, whose length is substantially reduced, with a larger reduction in the case of larger m. Notice that a higher confidence level of the estimated parameters is crucial when applying the identification procedure to real settings, where the true parameter values are unknown.

As an example, the fitting results for the realization $(\beta, \gamma) = (0.5, 0.03)$ are shown in Figures 2 and 3. As expected and symmetrically to the analogous simulations related to the scenario (a), only the best second-order fit (yellow solid line) is able to track accurately both the sequence of first- and second-order data.

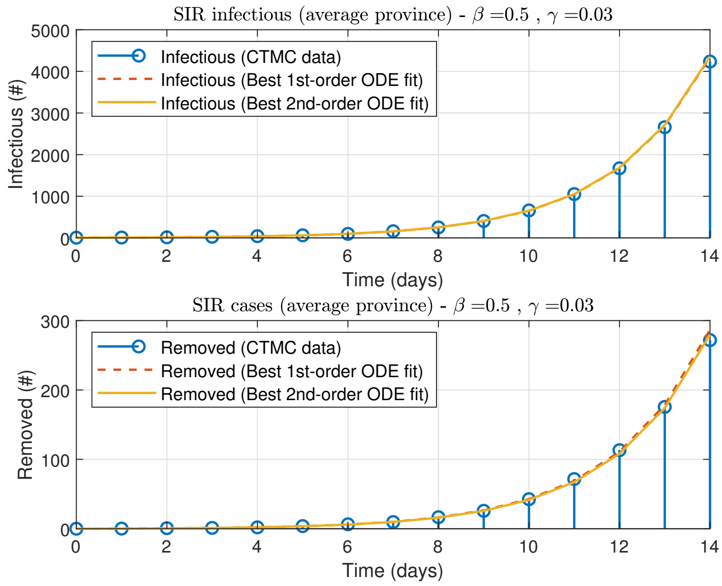

Figure 2. Scenario (b): estimation from infectious and removed data. Fitting of the first-order moments of the infectious individuals (**top panel**) and of the removed individuals (**bottom panel**) from a region including $m = 5$ provinces, in the case $(\beta, \gamma) = (0.5, 0.03)$: CTMC data (blue circles), best first-order fit (red dashed line), best second-order fit (yellow solid line).

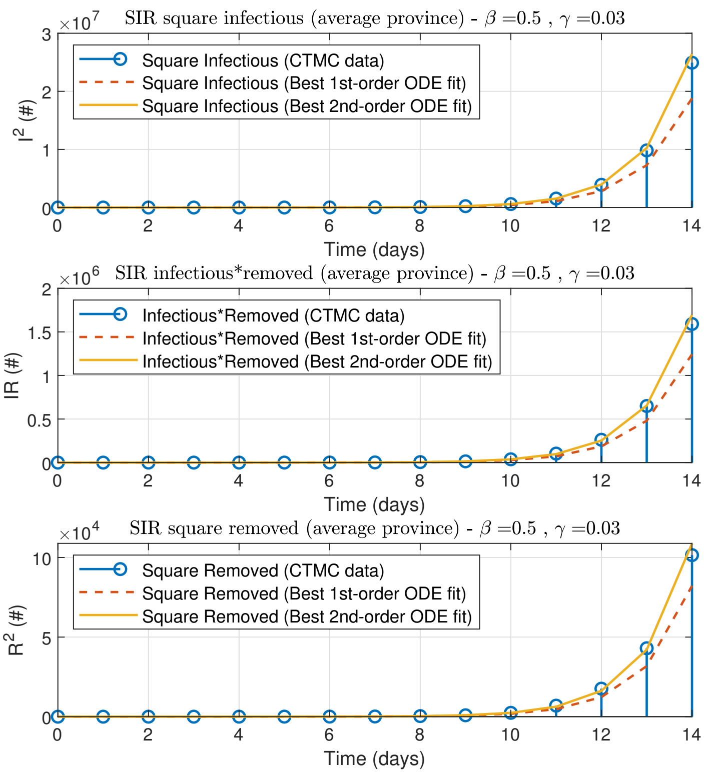

Figure 3. Fitting of the second-order moments of the infectious-removed population from a region including $m = 5$ provinces, in the case $(\beta, \gamma) = (0.5, 0.03)$: CTMC data (blue circles), best first-order fit (red dashed line), best second-order fit (yellow solid line).

5. Conclusions

This paper proposes a stochastic approach for the SIR modeling and identification, which is aimed at increasing the information content carried out by raw epidemiological data. The identification method is based on a modeling framework describing new infections and removals as discrete stochastic events and it assumes the epidemiological data related to the provinces of a given region as independent random paths drawn from the same Continuous-Time Markov Chain describing the epidemics spread in a given region. The estimation procedure is based on the building of a loss function that symmetrically weighs first-order and second-order moments, differently from the highly asymmetrical choice of standard approaches that exploit only first-order moments.

The numerical results show that when the infectious data are the only data available, adding the information content of the second-order moment to the first-order one is crucial to sensibly improve the parameter estimation, allowing one to obtain finite estimation errors attempting to separately estimate the relative infection and removal rates. Indeed, the first-order moment expression of the number of infected patients only depends on the difference between the model parameters, definitely making them not individually identifiable from the mean value of infected.

In the scenario where more information is available, namely both data of infectious and removed, the identification of the single parameters is always meaningful and benefits from a possible larger number of experimental samples, as happens when more provinces are considered. The removal rate is better estimated by the second-order procedure, differently from the infection rate, whose estimate becomes worse (on average). However, the model trajectories seem to be scarcely sensitive to this estimation error, since both first and second-order experimental data are accurately reproduced by the second-order fitting. Furthermore, confidence intervals of both parameters are also reduced in this case, which suggest higher reliability of the second-order method in real scenarios.

In conclusion, in view of the preliminary results shown in this paper and highlighting the potential of the approach, further investigation will be devoted to the topic of parameter estimation from real data exploiting information deriving from higher-order moment equations. This will be the object of future work.

Author Contributions: Conceptualization, A.B., P.P. and F.P.; Investigation, A.B., P.P. and F.P.; Methodology, A.B., P.P. and F.P.; Software, A.B.; Writing—original draft, A.B., P.P. and F.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

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

References

1. Mandel, A.; Veetil, V. The Economic Cost of COVID Lockdowns: An Out-of-Equilibrium Analysis. *Econ. Disasters Clim. Chang.* **2020**, *4*, 431–451. [CrossRef] [PubMed]
2. Spelta, A.; Flori, A.; Pierri, F.; Bonaccorsi, G.; Pammoli, F. After the lockdown: Simulating mobility, public health and economic recovery scenarios. *Nature* **2020**, *10*, 1–13. [CrossRef] [PubMed]
3. Saladino, V.; Algeri, D.; Auriemma, V. The Psychological and Social Impact of COVID-19: New Perspectives of Well-Being. *Front. Psychol.* **2020**, *11*, 2550. [CrossRef] [PubMed]
4. Gatto, M.; Bertuzzo, E.; Mari, L.; Miccoli, S.; Carraro, L.; Casagrandi, R.; Rinaldo, A. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 10484–10491. [CrossRef] [PubMed]
5. Calafiore, G.C.; Novara, C.; Possieri, C. A time-varying SIRD model for the COVID-19 contagion in Italy. *Annu. Rev. Control.* **2020**, *50*, 361–372. [CrossRef]
6. Molnar, T.G.; Singletary, A.W.; Orosz, G.; Ames, A.D. Safety-Critical Control of Compartmental Epidemiological Models With Measurement Delays. *IEEE Control. Syst. Lett.* **2021**, *5*, 1537–1542. [CrossRef]
7. Morato, M.M.; Bastos, S.B.; Cajueiro, D.O.; Normey-Rico, J.E. An optimal predictive control strategy for COVID-19 (SARSCoV- 2) social distancing policies in Brazil. *Annu. Rev. Control* **2020**, *50*, 417–431. [CrossRef]

8. Borri, A.; Palumbo, P.; Papa, F.; Possieri, C. Optimal design of lock-down and reopening policies for early-stage epidemics through SIR-D models. *Annu. Rev. Control* **2020**, *51*, 511–524. [CrossRef]
9. Castanos, F.; Mondié, S. Observer-based predictor for a susceptible-infectious-recovered model with delays: An optimal control case study. *Int. J. Robust Nonlinear Control* **2021**, *31*, 5118–5133. [CrossRef]
10. Roda, W.C.; Varughese, M.B.; Han, D.; Li, M.Y. Why is it difficult to accurately predict the COVID-19 epidemic? *Infect. Model.* **2020**, *5*, 271–281. [CrossRef]
11. Bertozzi, A.L.; Franco, E.; Mohler, G.; Short, M.B.; Sledge, D. The challenges of modeling and forecasting the spread of COVID- 19. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 16732–16738. [CrossRef] [PubMed]
12. Chinazzi, M.; Davis, J.T.; Ajelli, M.; Gioannini, C.; Litvinova, M.; Merler, S.; Y Piontti, A.P.; Mu, K.; Rossi, L.; Sun, K.; et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **2020**, *368*, 395–400. [CrossRef] [PubMed]
13. Kucharski, A.J.; Russell, T.W.; Diamond, C.; Liu, Y.; Edmunds, J.; Funk, S. Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *Lancet Infect. Dis.* **2020**, *20*, 553–558. [CrossRef]
14. Flaxman, S.; Mishra, S.; Gandy, A.; Unwin, H.J.; Mellan, T.A.; Coupland, H.; Whittaker, C.; Zhu, H.; Berah, T.; Eaton, J.W.; et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **2020**, *584*, 257–261. [CrossRef]
15. Giamberardino, P.D.; Iacoviello, D.; Papa, F.; Sinisgalli, C. Dynamical evolution of COVID-19 in Italy with an evaluation of the size of the asymptomatic infective population. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 1326–1332. [CrossRef]
16. Montefusco, F.; Procopio, A.; Bulai, I.M.; Amato, F.; Pedersen, M.G.; Cosentino, C. Interacting with COVID-19: How population behavior, feedback and memory shaped recurrent waves of the epidemic. *IEEE Control Syst. Lett.* **2022**, *7*, 583–588. [CrossRef]
17. Hadi, M.A.; Ali, H.I. Control of COVID-19 system using a novel nonlinear robust control algorithm. *Biomed. Signal Process. Control.* **2021**, *64*, 102317. [CrossRef]
18. Hadi, M.A.; Amean, Z.M. New strategy to control COVID-19 pandemic using lead/lag compensator. *Biomed. Signal Process. Control.* **2021**, *68*, 102669. [CrossRef]
19. Giordano, G.; Blanchini, F.; Bruno, R.; Colaneri, P.; Filippo, A.D.; Matteo, A.D.; Colaneri, M. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat. Med.* **2020**, *26*, 855–860. [CrossRef]
20. Giordano, G.; Colaneri, F.M.; Filippo, A.D.; Blanchini, F.; Bolzern, P.; Nicolao, G.D.; Sacchi, P.; Colaneri, P.; Bruno, R. Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy. *Nat. Med.* **2021**, *27*, 993–998. [CrossRef]
21. Verrelli, C.M.; Rossa, F.D. Two-Age-Structured COVID-19 Epidemic Model: Estimation of Virulence Parameters to Interpret Effects of National and Regional Feedback Interventions and Vaccination. *Mathematics* **2021**, *9*, 2414. [CrossRef]
22. Borri, A.; Palumbo, P.; Papa, F. Spread/removal parameter identification in a SIR epidemic model. In Proceedings of the 60th IEEE Conference on Decision and Control (CDC), Austin, TX, USA, 14–17 December 2021; pp. 2079–2084.
23. Italian Civil Protection Department; Morettini, M.; Sbrollini, A.; Marcantoni, I.; Burattini, L. COVID-19 in Italy: Dataset of the Italian Civil Protection Department. *Data Brief* **2020**, *30*, 105526.
24. Cardelli, L.; Kwiatkowska, M.; Laurenti, L. Stochastic analysis of Chemical Reaction Networks using Linear Noise Approximation. *Biosystems* **2016**, *149*, 26–33. [CrossRef] [PubMed]
25. Jenkinson, G.; Goutsias, J. Numerical Integration of the Master Equation in Some Models of Stochastic Epidemiology. *PLoS ONE* **2012**, *7*, e36160. [CrossRef]
26. Liang, J.; Qian, H. Computational Cellular Dynamics Based on the Chemical Master Equation: A Challenge for Understanding Complexity. *J. Comput. Sci. Technol.* **2010**, *25*, 154–168. [CrossRef]
27. Borri, A.; Palumbo, P.; Papa, F. Deterministic vs stochastic formulations and qualitative analysis of a recent tumour growth model. *IFAC-PapersOnLine* **2020**, *53*, 16418–16423. [CrossRef]
28. van Kampen, N.G. *Stochastic Processes in Physics and Chemistry*, 3rd ed.; North Holland Personal Library; Elsevier: Amsterdam, The Netherlands, 2007.
29. Gillespie, D.T. Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* **1977**, *81*, 2340–2361. [CrossRef]
30. Hespanha, J.P.; Singh, A. Stochastic models for chemically reacting systems using polynomial stochastic hybrid systems. *Int. J. Robust Nonlinear Control* **2005**, *15*, 669–689. [CrossRef]
31. Kermack, W.O.; McKendrick, A.G. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. London. Ser.* **1927**, *115*, 700–721.
32. Iannelli, M.; Pugliese, A. Mathematical modeling of epidemics. In *An Introduction to Mathematical Population Dynamics*; Springer: Cham, Switzerland, 2014; pp. 209–264.
33. Normey-Rico, J.E.; Camacho, E.F. *Control of Dead-Time Processes*; Springer: London, UK, 2007.