Nonlinear Dyn (2021) 105:2775–2794
https://doi.org/10.1007/s11071-021-06707-6

ORIGINAL PAPER

Modeling and prediction of the transmission dynamics of COVID-19 based on the SINDy-LM method

Yu-Xin Jiang · Xiong Xiong · Shuo Zhang · Jia-Xiang Wang · Jia-Chun Li · Lin Du

Received: 25 January 2021 / Accepted: 4 July 2021 / Published online: 22 July 2021

Abstract The transmission dynamics of COVID-19 is investigated in this study. A SINDy-LM modeling method that can effectively balance model complexity and prediction accuracy is proposed based on data-driven technique. First, the Sparse Identification of Nonlinear Dynamical systems (SINDy) method is used to discover and describe the nonlinear functional relationship between the dynamic terms in the model in accordance with the observation data of the COVID-19 epidemic. Moreover, the Levenberg–Marquardt (LM) algorithm is utilized to optimize the obtained model for improving the accuracy of the SINDy algorithm. Second, the obtained model, which is consistent with the logistic model in mathematical form with small errors and high robustness, is leveraged to review the epidemic situation in China. Otherwise, the evolution of the epidemic in Australia and Egypt is predicted, which demonstrates that this method has universality for constructing the global COVID-19 model. The proposed model is also compared with the extreme learning machine (ELM), which shows that the prediction accuracy of the SINDy-LM method outperforms that of the ELM method and the generated model has higher sparsity.

Keywords COVID-19 · Transmission dynamics · SINDy · Data-driven · LM optimization algorithm

Y.-X. Jiang · X. Xiong · S. Zhang · J.-X. Wang · J.-C. Li · L. Du (✉)
School of Mathematics and Statistics, Northwestern Polythechnical University, Xi'an 710129, China
e-mail: lindu@nwpu.edu.cn

S. Zhang · X. Xiong · L. Du
MIIT Key Laboratory of Dynamics and Control of Complex Systems, Xi'an 710129, China

1 Introduction

At the beginning of 2020, Corona Virus Disease 2019 (COVID-19) occurred globally [2,3], and it is currently spreading worldwide on a large scale. Statistics from Johns Hopkins University in the United States showed that nearly 850,000 people have died of COVID-19 and more than 25,890,000 cases of infection have been confirmed in more than 180 countries and regions around the world as of September 31, 2020. The prediction of COVID-19 is an important task in the public health security. It can detect the development trend of the disease early and improve the predictability of the epidemic, which plays an important role in the prevention, treatment, and health decision making of the disease [4]. Therefore, the proposal of a universal prediction method for COVID-19 has important theoretical and practical significance for the guidance of epidemic prevention and control worldwide based on the development of the epidemic situation in China.

Model-driven and data-driven approaches are two main methods for the spread and prediction of COVID-19 at present. The traditional epidemiological models for studying infectious diseases consists of SIR and

SEIR models [1]. Teles used the SIR model to simulate the MERS epidemic in Korea for assessing the evolution of the curve of the number of COVID-19 cases in Portugal [5]. Some researchers used the SEIR model to analyze and identify the transmission dynamics of the COVID-19. Huang et al. completed the prediction of the COVID-19 epidemic in some Asian countries based on the transmission dynamics and universal SEIR model [4]. Tang et al. derived the basic reproduction number of COVID-19 through SEIR model analysis [6]. Gaurav et al. employed the SEIR model and regression method to analyze and predict the development of the epidemic in India [7]. Other researchers have also proposed some methods to improve and extend the SEIR model. He et al. used particle swarm algorithm to identify the parameters in the SEIR model and introduced seasonal and random infection parameters [8]. However, the abovementioned references ignore that COVID-19 has a long incubation period and strong infectivity; these characteristics give rise to great difficulty in predicting specific parameters in the traditional model. If a high-dimensional complex model is established to reduce the difficulty of estimation, then some problems such as the inconsistency of the development trend of the state variables in the model with the actual data will be induced.

In addition to traditional epidemiological models, academic circles at home and abroad have proposed many research methods that use machine learning algorithms to predict the spread of COVID-19 [10]. Javid used the extreme learning machine (ELM) to make predictions [11], but some issues such as low prediction accuracy and poor model interpretability still exist. Although machine learning methods have the abovementioned shortcomings, models based on data-driven methods in machine learning have been widely employed in nearly every branch of engineering and applied mathematics [12,13]. This framework serves as an alternative for the discovery of the dynamic equations for controlling the spread of infectious diseases. There have been some researchs that used data-driven methods to predict and analyze COVID-19 [14–16]. This data-driven method omits the complicated modeling process, which avoids the large error caused by parameter identification and has high practicability and universality [17–19]. Thus, it is often used in system identification [20]. However, data-driven modeling usually requires assumptions on the form of the model. Thus, the results are limited to linear dynamics, which can only produce valid results near the fixed point of dynamics [21]. In addition, if the data-driven method is used to establish a model to predict COVID-19, then it is also prone to problems such as over-fitting and low prediction accuracy in a long time. Therefore, designing a new data-driven method of the accurate prediction of the actual observation data of COVID-19, which could balance the complexity (interpretability) of the model of COVID-19 and the prediction accuracy, has become a key research topic and urgently needs a breakthrough at home and abroad.

The Sparse Identification of Nonlinear Dynamical systems (SINDy) method uses symbolic regression to discover and describe the nonlinear function of the relationship between variables and measured dynamic terms; this method also utilizes sparse representation to determine the correlation in an effective and scalable framework model item [17,18,22]. This method has been used for feature selection and parameter identification of differential equations in physical models [19]. It can also effectively solve the problem that the number of hidden neurons in the construction of deep neural networks cannot be determined [20]. Adopting this method to establish a model of COVID-19 based on data-driven strategy can effectively avoid errors caused by infectious disease coefficients in the model of COVID-19; thus, a more accurate prediction result can be obtained. Moreover, this method has better interpretability than the other data-driven modeling methods. However, this data-driven method may have problems of overfitting and low prediction accuracy. Accordingly, appropriate and reasonable improvements are needed in order to derive a more stable optimization method.

Levenberg-Marquardt (LM) algorithm has the advantages of reducing the probability of falling into a local minimum value, strong stability, and fast convergence [23]. It has become a standard technique for solving nonlinear least square problems [24] and has been employed to optimize the learning process of neural networks [25]. However, the accuracy of the LM algorithm depends heavily on the selection of the initial value [23]. If the SINDy data-driven modeling is combined with the LM algorithm to optimize the parameters in the model, then the accuracy of the SINDy algorithm can be improved first. Second, it can also effectively overcome shortcoming of the LM algorithm that relies too much on the initial value. The combination of the two methods can achieve a more complete review of

countries where the epidemic has ended and make more accurate predictions for countries where the epidemic has not ended. Thus, this combination can further solve the problems of determining the turning point of the epidemic.

This study presents a method that combines new data-driven methods with optimization algorithms the abovementioned discussion. The effect of balancing model complexity and interpretability simultaneously can be achieved by combining the SINDy method with the LM method. Focused on the transmission dynamics of infectious diseases, this study contributes to proposing a modeling method that can effectively balance interpretability and prediction accuracy based on a data-driven technique. The proposed modeling method could review the epidemic situation in Chinese mainland and predict the epidemic situation in other countries worldwide.

The paper is organized as follows. In Sect. 2, the basic principles of the nonlinear dynamic sparse recognition method and the LM algorithm are mainly introduced, and the "SINDy-LM" data-driven modeling method that combines the two is proposed. In Sect. 3, the method is first applied to establish a COVID-19 model suitable for Chinese mainland. This section also judges the location of the "epidemic turning point" and the SINDy-LM method is applied to other countries in the world (Australia, Egypt). Section 4 considers higher dimension cases and the robustness of the model. The comparison between SINDy-LM method and the ELM algorithm is also discussed in this section. The last section summarizes the conclusion and discusses the practical significance of the method, potential problems, and follow-up work.

2 Data-driven modeling based on SINDy-LM

In this section, the idea and algorithm procedure of SINDy method and LM algorithm are introduced in detail, and the two are combined to complete the establishment of the model.

2.1 Model setup based on SINDy method

The Sparse Identification of Nonlinear Dynamical systems (SINDy) method uses sparse regression and parameter identification to discover the reduced governing equations correctly from a large number of combined potential dynamic models. Steven L. Brunton and others introduced the background and partial applications of the SINDy method [17] and elaborated on using the SINDy method to study models and solve regression and selection problems [26].

This section introduce the SINDy method proposed by Steven L. Brunton [26]. The symbols used throughout the work are given as follows: lowercase letters (such as x) represent scalars, and bold lowercase letters (such as $\mathbf{x}$) represent vectors; bold uppercase letters (such as $\mathbf{X}$) define matrices; parentheses emphasize functions and vector functions (such as $f(\cdot)$); the relevance of variables with respect to time, such as $x(t)$, is emphasized when necessary.

SINDy method determines the governing equations in the infectious disease dynamics system by using the true observation data in many countries. The nonlinear dynamic system is expressed as Eq. (1):

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t)), \tag{1}$$

where vector $\mathbf{x}(t)$ represents the state of the system at time t, $\dot{\mathbf{x}}(t)$ represents the derivative of $\mathbf{x}(t)$ at time t, and $f(\mathbf{x}(t))$ is the evolution of $\mathbf{x}(t)$ with time to represent constraints of the governing equation in the dynamic system. In the established COVID-19 model, the main research goal is about the evolution of state variables such as the number of confirmed cases, deaths, cured cases and close contacts over time.

The specific form of the function f is determined by the observation data, and it requires collecting the time history of the state $\mathbf{x}(t)$. It needs to collect the COVID-19 time series data of $\mathbf{x}(t)$ at $t_1, t_2, \cdots, t_m$ and construct the relevant numerical matrix Eqs. (2) and (3):

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^T(t_1) \\ \mathbf{x}^T(t_2) \\ \vdots \\ \mathbf{x}^T(t_m) \end{bmatrix} = \begin{bmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_n(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_m) & x_2(t_m) & \cdots & x_n(t_m) \end{bmatrix} \begin{matrix} \downarrow time, \\ \\ \\ \\ \rightarrow state \end{matrix} \tag{2}$$

$$\dot{\mathbf{X}} = \begin{bmatrix} \dot{\mathbf{x}}^T(t_1) \\ \dot{\mathbf{x}}^T(t_2) \\ \vdots \\ \dot{\mathbf{x}}^T(t_m) \end{bmatrix}$$

$$= \begin{bmatrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) & \cdots & \dot{x}_n(t_1) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \cdots & \dot{x}_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{x}_1(t_m) & \dot{x}_2(t_m) & \cdots & \dot{x}_n(t_m) \end{bmatrix}, \tag{3}$$

where $\dot{\mathbf{X}}$ is approximated by the numerical differentiation of $\mathbf{X}$, and the data required in $\mathbf{X}$ are the observation data collected from the measurement in the real world. In this study, $x_1, x_2, \cdots, x_n$ will be substituted into state variables such as the number of confirmed cases and the number of deaths.

A candidate function library $\boldsymbol{\Theta}(\mathbf{x})$ consisting of candidate nonlinear functions Eq. (4), where each column represents the potential candidate of the element in $f(\cdot)$ to be discovered, is constructed. The function chosen to fill the library is arbitrary and can be composed of polynomial terms and trigonometric functions. Considering the interpretability of the model and the amount of calculation, the candidate function library only contains polynomial terms. Then, the nonlinear feature library is leveraged to find the least dynamic term that satisfies Eq. (4); thus, a COVID-19 model is generated as follows:

$$\dot{\mathbf{X}} = \boldsymbol{\Theta}(\mathbf{X})\boldsymbol{\Xi}, \tag{4}$$

$$\boldsymbol{\Theta}(\mathbf{X}) = \begin{bmatrix} | & | & | & | & \\ \mathbf{1} & \mathbf{X} & \mathbf{X}^{P_2} & \mathbf{X}^{P_3} & \cdots \\ | & | & | & | & \end{bmatrix}, \tag{5}$$

where element $\mathbf{1}$ represents a column vector composed of m ones, element $\mathbf{X}$ is defined in Eq. (2), and element $\mathbf{X}^{P_2}$ is the set of all quadratic polynomial functions of state vector $\mathbf{x}$. $\mathbf{X}^{P_3}$ is the set of cubic polynomial functions. The superscript P_2 is used to define the set of quadratic polynomial functions with a structure as in Eq. (6):

$$\mathbf{X}^{P_2} = \begin{bmatrix} x_1^2(t_1) & x_1x_2(t_1) & \cdots & x_2^2(t_1) & \cdots & x_n^2(t_1) \\ x_1^2(t_2) & x_1x_2(t_2) & \cdots & x_2^2(t_2) & \cdots & x_n^2(t_2) \\ \vdots & \vdots & & \vdots & & \vdots \\ x_1^2(t_m) & x_1x_2(t_m) & \cdots & x_2^2(t_m) & \cdots & x_n^2(t_m) \end{bmatrix}. \tag{6}$$

Each column of $\boldsymbol{\Theta}(\mathbf{X})$ represents the candidate function term on the right side of the equation. A greater degree of freedom is obtained when selecting coefficients in this alternative function library. However, other dynamic systems such as infectious disease transmission systems usually have only a few nonlinear terms in practical applications. Thus, the right side of the equation has high sparsity in the high-dimensional nonlinear function space. As emphasized in the literature [26], the assumption that the algorithm can converge to the true solution is that only a few elements make up the function $f(\cdot)$; this way makes it sparse in the space of possible functions [13]. This sparsity balances the complexity and accuracy of the model. In summary, a sparse regression problem can be established to determine the sparse vector of the coefficient matrix for obtaining the nonlinear effective function term. The objective function of SINDy is defined by ℓ_1 norm regression:

$$\boldsymbol{\xi}_\mathbf{i} = \arg\min_{\hat{\boldsymbol{\xi}}_\mathbf{i}} \left\| \dot{\mathbf{X}} - \hat{\boldsymbol{\xi}}_\mathbf{i}\boldsymbol{\Theta}(\mathbf{X}) \right\|_2 + \lambda \left\| \hat{\boldsymbol{\xi}}_\mathbf{i} \right\|_1, \tag{7}$$

where $\boldsymbol{\xi}_\mathbf{i}$ collects the coefficients of candidate function $\boldsymbol{\Theta}(\mathbf{X})$, which is the goal of minimization. The number of vectors is equal to the dimension n of the state vector with $i = 1, \cdots, n$. Only a few candidate functions are expected to affect system dynamics; thus, all vectors are expected to be sparse. Symbols $\|\cdot\|_1$ and $\|\cdot\|_2$ represent ℓ_1 norm and ℓ_2 norm, respectively. λ is a scalar multiplier and element $\lambda \left\| \hat{\boldsymbol{\xi}}_\mathbf{i} \right\|_1$ is a regularization term that penalizes coefficients different from 0 in a linear manner. It is the real promoter of sparsity in the minimizing problem.

For Eq. (7), convex optimization algorithms such as least absolute shrinkage and selection operator (LASSO) can be used to solve. The method of sequential threshold least squares can also be used instead. The sequential threshold least square method applies sparsity by "manually" setting all coefficients less than λ to 0 in an iterative manner. The result is very similar to that produced by LASSO. For the sake of brevity, this study takes the latter solution.

After all the sparse vectors are estimated, they can be collected in the sparse matrix $\boldsymbol{\Xi}$:

$$\boldsymbol{\Xi} = \begin{bmatrix} | & | & & | \\ \boldsymbol{\xi_1} & \boldsymbol{\xi_2} & \cdots & \boldsymbol{\xi_n} \\ | & | & & | \end{bmatrix}. \tag{8}$$

Given that $\boldsymbol{\Xi}$ is obtained, the model of each elements of the control equation Eq. (1) can be constructed as follows:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(\mathbf{x}) = \boldsymbol{\Theta}\left(\mathbf{x}^{\mathrm{T}}\right)\boldsymbol{\xi_i}, \tag{9}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ and $x_1, x_2, \ldots, x_n$ represent state variables such as the number of confirmed cases and the number of deaths. Unlike the data matrix

$\boldsymbol{\Theta}(\mathbf{X})$, $\boldsymbol{\Theta}(\mathbf{x}^{\mathrm{T}})$ is a vector of symbolic functions of elements of $\mathbf{x}$. In this way, a complete COVID-19 model is formed:

$$\dot{\mathbf{x}} = f(\mathbf{x}) = \boldsymbol{\Xi}^{\mathrm{T}} \left(\boldsymbol{\Theta} \left(\mathbf{x}^{\mathrm{T}} \right) \right)^{\mathrm{T}}. \tag{10}$$

In this study, the COVID-19 model is simplified to only one-dimension case and the reason for that will be illustrated in Sect. 4. That is, in Eq. (9), the governing equations followed by the components in the obtained COVID-19 model can be expressed as:

$$\begin{aligned} \frac{\mathrm{d}x_i}{\mathrm{d}t} &= \boldsymbol{\Theta}\left(x_i\right) \boldsymbol{\xi}_i \\ &= b_{0,i} + b_{1,i} x_i + b_{2,i} x_i^2 + b_{3,i} x_i^3 + \cdots b_{p,i} x_i^p, \end{aligned} \tag{11}$$

where $b_{0,i}, b_{1,i}, \cdots, b_{p,i}$ are the elements in the coefficient matrix $\boldsymbol{\Xi} = [\boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \cdots, \boldsymbol{\xi_n}]$ with each $\boldsymbol{\xi}_i = (b_{0,i}, b_{1,i}, \cdots, b_{p,i})^{\mathrm{T}}$.

2.2 Optimization of model coefficients with LM algorithm

The model generated by the SINDy method is more dependent on the selection of training data, and its advantage is that it has strong universality and flexibility. However, this data-driven method have problems with over-fitting and low prediction accuracy. Thus, appropriate and reasonable improvements are needed in order to derive a more stable optimization method.

The LM algorithm, which is a standard technique for solving nonlinear least square problems [23,24], has been used to optimize the learning process of neural networks [25]. This study combines SINDy method with LM algorithm to optimize the parameters in the model.

According to the abovementioned SINDy method, Eq. (4) can be solved as Eq. (10), which is the specific coefficient matrix $\boldsymbol{\Xi} = [\boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \cdots, \boldsymbol{\xi_n}]$ of the model. This study combines the SINDy method with the LM algorithm to iteratively optimize the items in $\boldsymbol{\Xi}$ for improving the accuracy of the coefficients $\boldsymbol{\xi_i}$ of each function in the model. As a result, better prediction results can be obtained.

The coefficients of each term in the model Eq. (11) directly obtained by the SINDy method are set as the initial value of the iteration, which is set as $\boldsymbol{\xi}_i^{(0)}$. Then, the principle of the iterative procedure is

$$\boldsymbol{\xi}_i^{(k+1)} = \boldsymbol{\xi}_i^{(k)} + \Delta \boldsymbol{\xi}_i, \tag{12}$$

where $\boldsymbol{\xi}_i^{(k)}$ is the vector at the kth iteration($k = 1, 2, ..., M$); $\boldsymbol{\xi}_i^{(k+1)}$ is the vector at the $(k+1)$th iteration; $\Delta \boldsymbol{\xi}_i$ is the variation during the two iterations. Next, for $\boldsymbol{\xi}_i$ at the kth iteration (the superscript (k) is temporarily omitted for brevity), assume that

$$\Delta \boldsymbol{\xi}_i = -\left[\nabla^2 E\left(\boldsymbol{\xi}_i\right) \right]^{-1} \nabla E\left(\boldsymbol{\xi}_i\right), \tag{13}$$

$$\begin{aligned} E\left(\boldsymbol{\xi}_i\right) &= \left\| \boldsymbol{e}\left(\boldsymbol{\xi}_i\right) \right\|_2 \\ &= \frac{1}{2} \sum_{j=1}^{m} \left[f\left(x_i\left(t_j\right); \boldsymbol{\xi}_i\right) - \dot{x}_i\left(t_j\right) \right]^2, \end{aligned} \tag{14}$$

where $\boldsymbol{e}(\boldsymbol{\xi}_i) = (e_1, e_2, \ldots, e_m)$ is the error and $E(\boldsymbol{\xi}_i)$ is the $\ell_2$norm of error. $\nabla E(\boldsymbol{\xi}_i)$ is the gradient, and $\nabla^2 E(\boldsymbol{\xi}_i)$ is the Hessian matrix of $E(\boldsymbol{\xi}_i)$, as shown in the following equations:

$$\nabla E(\boldsymbol{\xi}_i) = \left[\begin{array}{cccc} \frac{\partial E}{\partial b_{0,i}} & \frac{\partial E}{\partial b_{1,i}} & \cdots & \frac{\partial E}{\partial b_{p,i}} \end{array} \right]^{\mathrm{T}}, \tag{15}$$

$$\nabla^2 E(\boldsymbol{\xi}_i) = \left[\begin{array}{ccccc} \frac{\partial^2 E}{\partial^2 b_{0,i}} & \frac{\partial^2 E}{\partial b_{0,i} b_{1,i}} & \cdots & \frac{\partial^2 E}{\partial b_{0,i} b_{p,i}} \\ \frac{\partial^2 E}{\partial b_{1,i} b_{0,i}} & \frac{\partial^2 E}{\partial^2 b_{1,i}} & \cdots & \frac{\partial^2 E}{\partial b_{1,i} b_{p,i}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 E}{\partial b_{p,i} b_{0,i}} & \frac{\partial^2 E}{\partial b_{p,i} b_{1,i}} & \cdots & \frac{\partial^2 E}{\partial^2 b_{p,i}} \end{array} \right]. \tag{16}$$

For high-dimension cases, computing the second derivative will be very complicated. In the Gauss-Newton method, the Hessian matrix is not calculated directly, but is fitted via the Jacobian matrix. They are rewritten by

$$\nabla^2 E(\boldsymbol{\xi}_i) \approx \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{\xi}_i) \boldsymbol{A}(\boldsymbol{\xi}_i), \tag{17}$$

However, if the Jacobian matrix is used to fit the Hessian matrix, the calculated result is not necessarily reversible. So on this basis, an identity matrix is introduced:

$$\nabla^2 E(\boldsymbol{\xi}_i) \approx \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{\xi}_i) \boldsymbol{A}(\boldsymbol{\xi}_i) + \mu \boldsymbol{I}. \tag{18}$$

Among them, $\boldsymbol{I}$ is the identity matrix, $\mu > 0$ is the damping coefficient and a constant. $\boldsymbol{A}(\boldsymbol{\xi}_i^{(k)})$ denotes the Jacobian matrix of $\boldsymbol{e}(\boldsymbol{\xi}_i)$ at the kth iteration, as shown in the following equation:

$$\begin{aligned} \boldsymbol{A}_k &\triangleq \boldsymbol{A}(\boldsymbol{\xi}_i^{(k)}) \\ &= \left[\begin{array}{ccccc} \frac{\partial e_1(\boldsymbol{\xi}_i^{(k)})}{\partial b_{0,i}} & \frac{\partial e_1(\boldsymbol{\xi}_i^{(k)})}{\partial b_{1,i}} & \cdots & \frac{\partial e_1(\boldsymbol{\xi}_i^{(k)})}{\partial b_{p,i}} \\ \frac{\partial e_2(\boldsymbol{\xi}_i^{(k)})}{\partial b_{0,i}} & \frac{\partial e_2(\boldsymbol{\xi}_i^{(k)})}{\partial b_{1,i}} & \cdots & \frac{\partial e_2(\boldsymbol{\xi}_i^{(k)})}{\partial b_{p,i}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial e_m(\boldsymbol{\xi}_i^{(k)})}{\partial b_{0,i}} & \frac{\partial e_m(\boldsymbol{\xi}_i^{(k)})}{\partial b_{1,i}} & \cdots & \frac{\partial e_m(\boldsymbol{\xi}_i^{(k)})}{\partial b_{p,i}} \end{array} \right]. \end{aligned} \tag{19}$$

In summary, the expression of the LM algorithm is as Eq.(20) and (21):

$$\Delta\boldsymbol{\xi}_i = -\left(\boldsymbol{A}_k^{\mathrm{T}}\boldsymbol{A}_k + \mu\boldsymbol{I}\right)^{-1}\boldsymbol{A}_k^{\mathrm{T}}\boldsymbol{e}_k, \tag{20}$$

$$\boldsymbol{\xi}_i^{(k+1)} = \boldsymbol{\xi}_i^{(k)} - \left(\mathbf{A}_k^{\mathrm{T}}\boldsymbol{A}_k + \mu\boldsymbol{I}\right)^{-1}\boldsymbol{A}_k^{\mathrm{T}}\boldsymbol{e}_k. \tag{21}$$

The error function value $E(\boldsymbol{\xi}_i^{(k)})$ needs to be recalculated and μ_k needs to be set as the value of μ in the k th iteration. Suppose $\nu > 1$, if $E(\boldsymbol{\xi}_i^{(k+1)}) < E(\boldsymbol{\xi}_i^{(k)})$, make $\mu_{k+1} = \mu_k/\nu$ and stop decreasing μ when $E(\boldsymbol{\xi}_i^{(k+1)}) > E(\boldsymbol{\xi}_i^{(k)})$ is satisfied; if $E(\boldsymbol{\xi}_i^{(k+1)}) > E(\boldsymbol{\xi}_i^{(k)})$, make $\mu_{k+1} = \mu_k \cdot \nu$ and stop updating μ when $E(\boldsymbol{\xi}_i^{(k+1)}) < E(\boldsymbol{\xi}_i^{(k)})$ is satisfied. Finally, the specific coefficient matrix $\boldsymbol{\Xi}^{(M)} = \left[\boldsymbol{\xi}_1^{(M)}, \boldsymbol{\xi}_2^{(M)}, \ldots, \boldsymbol{\xi}_n^{(M)}\right]$ after M iterations is obtained, as well as the optimized COVID-19 model.

The complete modeling procedures are shown in the Fig. 1.

3 Main Results

3.1 Review work in China

According to the infection cases in Chinese mainland reported by the National Health Commission of China, the cumulative number of confirmed cases, cumulative cures, cumulative deaths, and other epidemic data from 10 January to 30 June, 2020 can be obtained.

On the basis of the proposed modeling method that combines the SINDy data-driven method and the LM optimization algorithm introduced above, 47 days of actual observation data from 10 January, 2020 to 25 February, 2020 are set as the training set. The data from 26 February, 2020 to 1 April, 2020 are used as the testing set to solve the COVID-19 model in Chinese mainland, which can realize the review of the epidemic.

By using the epidemic data from Chinese mainland, the coefficients of each function term at the right side of Eq. (10) are shown in Table 1.

Table 1 shows the coefficients of each function item at the right side of the governing equation followed by the cumulative number of confirmed cases, deaths, cured, and close contacts. The function items represented by the numbers on the horizontal axis are x, x^2, x^3, and x^{-1}. The coefficients of some function terms have been omitted in the table because they are all zero.

The results in Table 1 show that the coefficients of the dynamic terms in the COVID-19 model obtained using the SINDy-LM method have high sparsity, which is consistent with the underlying laws followed by other infectious disease systems. It also demonstrates that the model obtained by this method has a certain interpretability compared with the model obtained by the general neural network algorithm.

Let N_c denote the cumulative number of confirmed cases changing over time and N_d denote the number of deaths. By substituting the data of the cumulative number of confirmed cases into Eq. (10), the specific expression of Eq. (11) is obtained by solving

$$\frac{\mathrm{d}N_c}{\mathrm{d}t} = b_{1,1}N_c + b_{2,1}N_c^2, \tag{22}$$

where $b_{1,1} = 0.2391$, $b_{2,1} = -3.0234e^{-6}$.

Utilize the LM algorithm to iteratively optimize the coefficients, and obtain the optimized governing equation:

$$\frac{\mathrm{d}N_c}{\mathrm{d}t} = \hat{b}_{1,1}N_c + \hat{b}_{2,1}N_c^2, \tag{23}$$

where $\hat{b}_{1,1} = 0.2204$, $\hat{b}_{2,1} = -2.6790e^{-6}$.

By solving Eq. (23), the analytical expression of the cumulative number of confirmed cases in the model can be obtained as Eq. (24):

$$N_c(t) = \frac{\alpha}{1 + e^{-\beta(t-\tau)}}, \tag{24}$$

where $\alpha = 82269$, $\beta = 0.2204$, $\tau = 30.4959$.

By comparing the Eq.(25) and the classic Logistic model [27],

$$N_c(t) = \frac{K}{1 + \left(\frac{K}{N_0} - 1\right)e^{-r(t-t_0)}}, \tag{25}$$

where K is the population capacity, N_0 is the number of the population at time t_0, and r is the growth rate.

The mathematical form of Eq. (24) is exactly the same as that of Eq. (25). Thus, the mathematical model followed by the cumulative number of confirmed cases is the logistic model with $K = 82269, r = 0.2204$. The Chinese government employed effective anti-epidemic policies after the outbreak. Thus, it strictly controlled the number of people entering and leaving the country. Therefore, Chinese mainland can be regarded as a unit, in which the vast majority of cases occurred without any major "import" or "export" events. Therefore, the logistic model is indeed suitable for the prediction and analysis of the development of the epidemic in Chinese mainland.

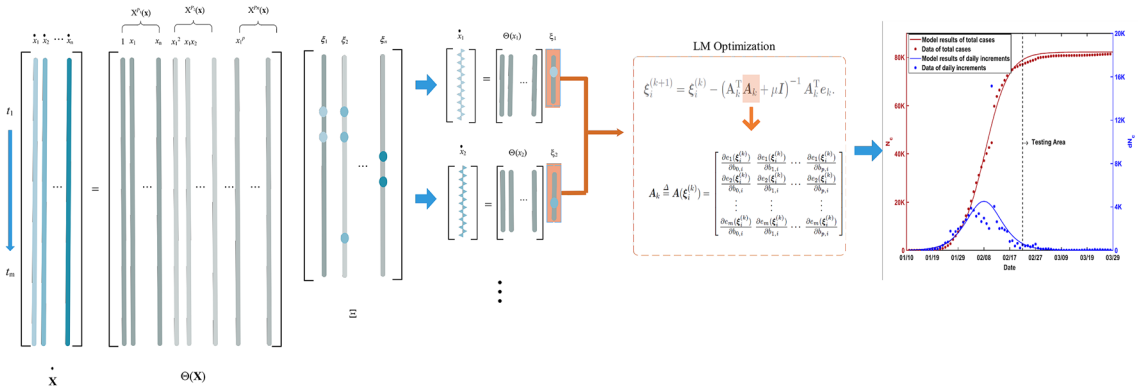


Fig. 1 Complete modeling procedures of SINDy-LM method

Table 1 Coefficients of each function item in the COVID-19 model

Function item	x_i	x_i^2
Confirmed cases	0.2204	$-2.6790e^{-6}$
Deaths	0.1715	$-5.6026e^{-5}$
Cured cases	0.2123	$-5.4154e^{-6}$
Close contacts	0.2122	$-3.3610e^{-7}$

The governing equation for the cumulative number of deaths in the COVID-19 model constructed with the SINDy-LM method is as shown in Eq. (26),

$$\frac{\mathrm{d}N_d}{\mathrm{d}t} = \hat{b}_{1,2}N_d + \hat{b}_{2,2}N_d^2, \tag{26}$$

where $\hat{b}_{1,2} = 0.1715$, $\hat{b}_{2,2} = -5.6026e^{-5}$.

The previous analysis shows that the cumulative number of deaths in China is also consistent with the logistic model, with model parameters $K = 3061.3104$, $r = 0.1715$. Therefore, the COVID-19 model obtained is compatible with the traditional empidemiological model and conforms to the natural law followed by the infectious disease transmission dynamic system.

Compared with the prediction results that directly uses the logistic model and identifies the parameters, the SINDy-LM method simplifies the process of analyzing many assumptions and is obviously more concise and powerful. Compared with other machine learning algorithms such as deep neural networks, SINDy-LM also avoids problems such as the uncertainty of neurons in the hidden layer. The resulting model is more interpretable.

According to the differential equations (such as Eqs. (23) and (26)) followed to obtain the state variables, the results of the review and prediction of the cumulative number of confirmed cases and deaths in Chinese mainland are shown in Fig. 2. The goodness of fit R^2 is greater than 0.95, which confirms that the model generated by the SINDy-LM method has high accuracy. Moreover, the curve trend of the cumulative number of confirmed cases and deaths obtained is consistent with the actual trend. The relative error is also within the controllable range.

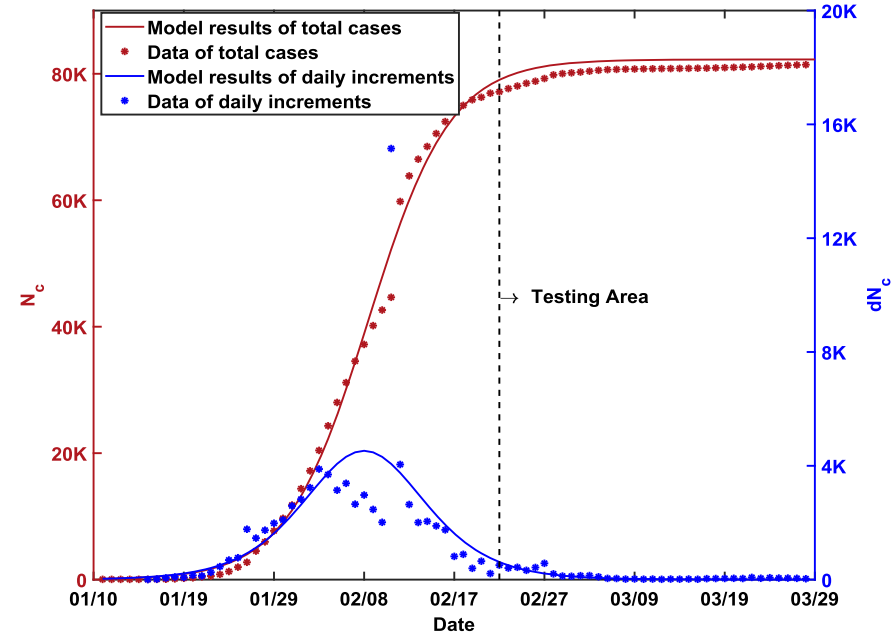
As shown in Fig. 2, the data of cumulative cases is demonstrated in red and the data of daily increments is demonstrated in blue. The left side represents the training set, and the right side represents the testing set. The dots are the actual data, and the lines are the predicted data.

3.2 Forecasting work of the other countries

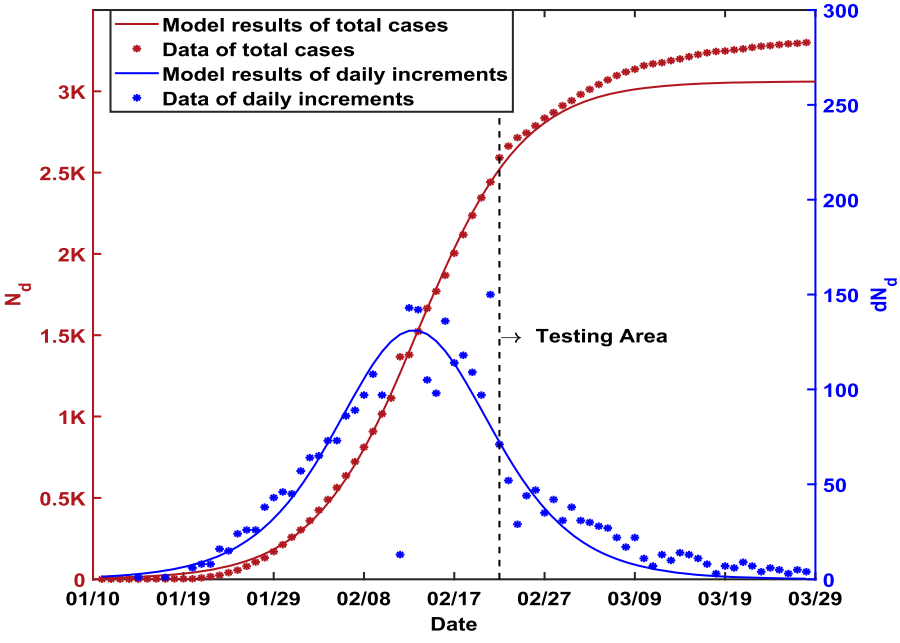
In addition to China, this section also forecasts and analyzes the evolution of the epidemic situation in other countries to verify whether this method is universal in other areas of the world. From the "Baidu Pandemic Real-time Big Data Report" website, relevant epidemic data in Australia and Egypt can be obtained during the 180 days after the outbreak. Based on the SINDy-LM method introduced above, the COVID-19 model is established and solved with the observation data.

The coefficients of the dynamic terms in Eq. (23) of the COVID-19 model followed in Australia are

$$\hat{b}_{1,1} = 0.1637, \hat{b}_{2,1} = -5.0489e^{-5}.$$



(a)



(b)

Fig. 2 Review results of Chinese mainland in the COVID-19 model followed by (a) the cumulative number of confirmed cases and (b) the cumulative number of deaths

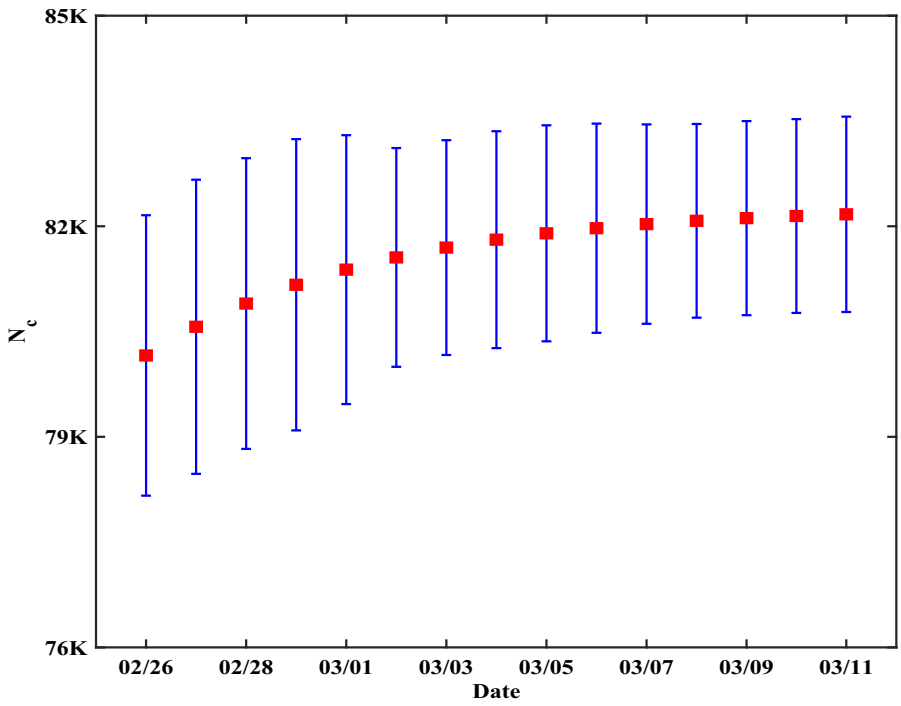


Fig. 3 Error bar graph obtained by predicting the cumulative number of confirmed cases in the next 15 days

The figure of the cumulative number of confirmed cases changing over time is shown in Fig. 4.

The prediction of the number of confirmed cases in Australia can be obtained using the SINDy-LM method. Fig. 4 shows that the results (total cases and daily increments) using this method are consistent with the actual results on the testing set. The trend of the curve is also roughly the same as the actual curve trend, which almost coincides in September, and the goodness of fit R^2 is 0.99926. The relative error of the prediction is small, which demonstrates the accuracy of the model is high. Similarly, the dynamic coefficients of the cumulative number of confirmed cases in Egypt in Eq. (23) are

$$\hat{b}_{1,1} = 0.0627, \hat{b}_{2,1} = -6.3115e^{-7}.$$

Fig. 5 shows the review and prediction of the epidemic in Egypt. The results are in line with the true situation. The goodness of fit is $R^2 = 0.99959$, which indicates that the relative error of the prediction results is also within the controllable range.

The prediction of the COVID-19 epidemic in Australia and Egypt shows that the SINDy-LM method can make accurate review and prediction on the development of epidemic with high-precision in other countries. Thus, the prediction of the epidemic via this method has some universality to a certain extend.

3.3 Exploring the "epidemic turning point"

The epidemic turning point is one of the signs that many researchers pay attention to ([16], [28], [29]). There are various definitions of the turning point. A common definition is the date when the number of newly confirmed cases each day peaks and then drops, that is, the point where the cumulative number of confirmed cases increases the fastest, which is the one used by some research groups [30]. However, Norden E. Huang et.al [16] proposed the fact that the number of new infections reached a peak and then declined does not necessarily mean that the epidemic has "turned from danger to safety", because the number of infected people is still increasing and there is still an urgent need for additional medical resources. In addition, locating this peak is highly susceptible to data failures and changes in diagnostic definitions. For example, on 12 February, when Hubei changed the definition of confirmed infection from the standard of nucleic acid gene sequencing test to clinical observation and radiology chest scan, more than 14,000 newly infection cases were added

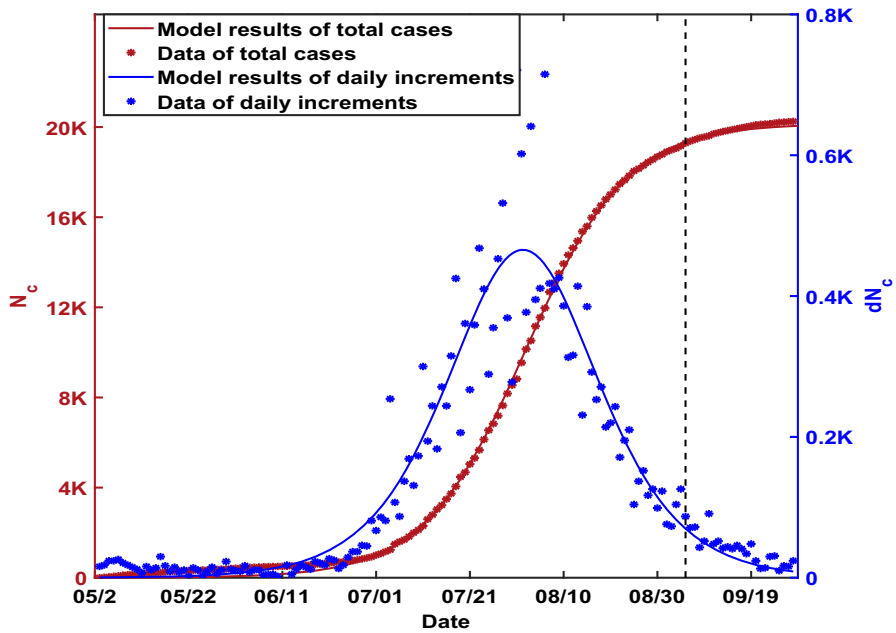


Fig. 4 Prediction results of Australia in the COVID-19 model (the number of confirmed cases)

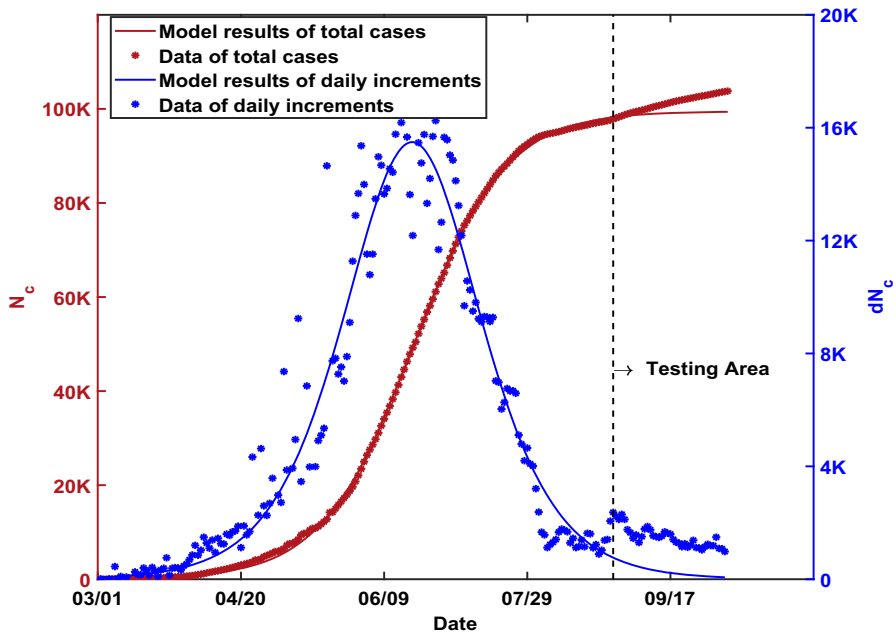


Fig. 5 Prediction results of Egypt in the COVID-19 model (the number of confirmed cases)

that day, thus forming a peak that has not yet appeared. But in other parts of China, the definition of "infected person" has not changed.

A more meaningful turning point should be the date when the number of confirmed cases reached a peak and then began to decline. Norden E. Huang et.al [16] proved that for the ongoing COVID-19 epidemic, the determination of this turning point is not sensitive to past data issues, including the sharp increase in $N_c(t)$ when Hubei changed its definition of "confirmed infected person" on 12 February .

Therefore, this definition is used to judge the epidemic turning point in this study. Subtracting the daily recovered cases each day $dN_r(t)$ from the daily confirmed cases $dN_c(t)$, we can obtain the number of newly infected cases each day. This physical quantity

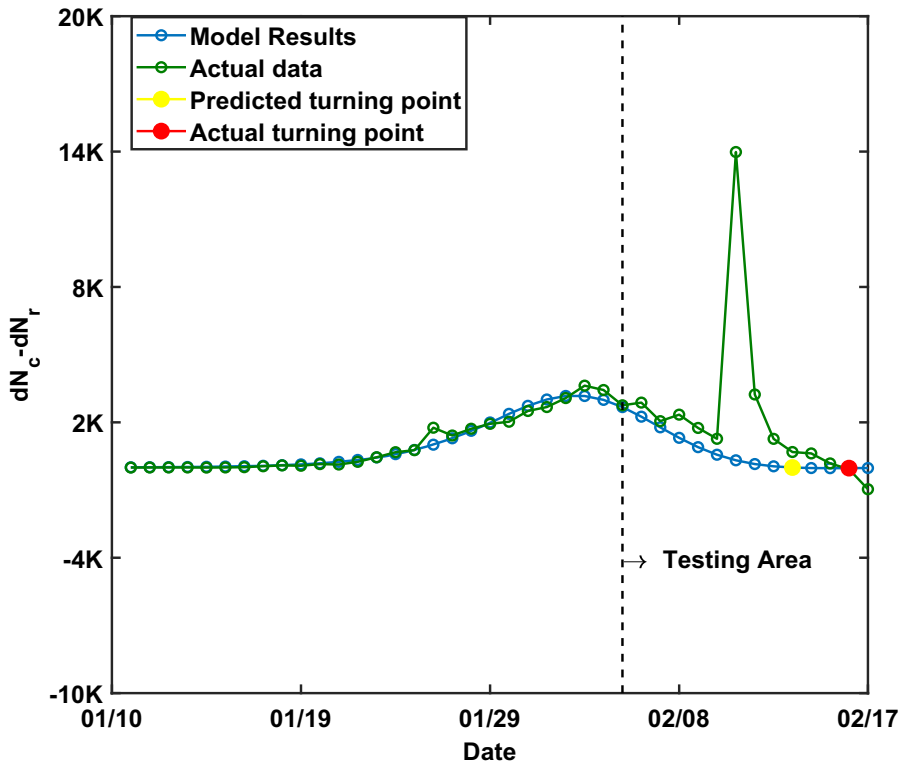


Fig. 6 Forecast of the epidemic turning point

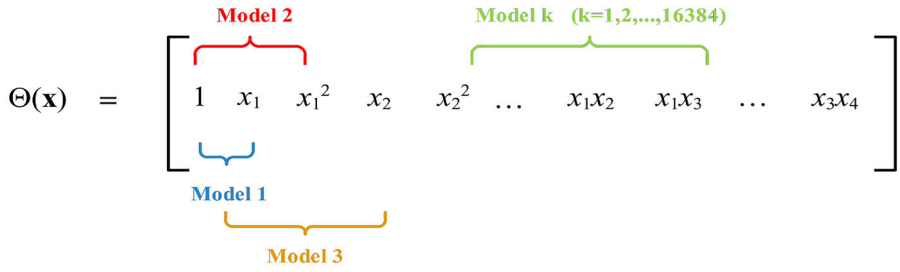


Fig. 7 Select different subsets of function libraries and generate models

Table 2 R^2 of 10 models generated by different candidate function libraries

Model	1	**2**	3	4	5
R^2	0.45844	**0.97758**	0.87692	0.86179	0.45961
Model	6	7	8	9	10
R^2	0.17508	0.04079	0.45850	0.45851	0.45850

is called Existing Infected Cases (EIC) [16], and the epidemic turning point t_p is the date when EIC $= 0$. In other words, t_p should satisfy:

$$\mathrm{d}N_c\left(t_p\right) = \mathrm{d}N_r\left(t_p\right). \tag{27}$$

Using the data of 27 days as the training set, the COVID-19 model derived by the SINDy-LM method can judge the epidemic turning point t_p 7 days in advance. According to Eq.(11), the equations followed by the number of confirmed cases and the number of recovered cases can be obtained. Then on the basis of Eq.(27), the epidemic turning point can be determined, as shown in Fig. 6.

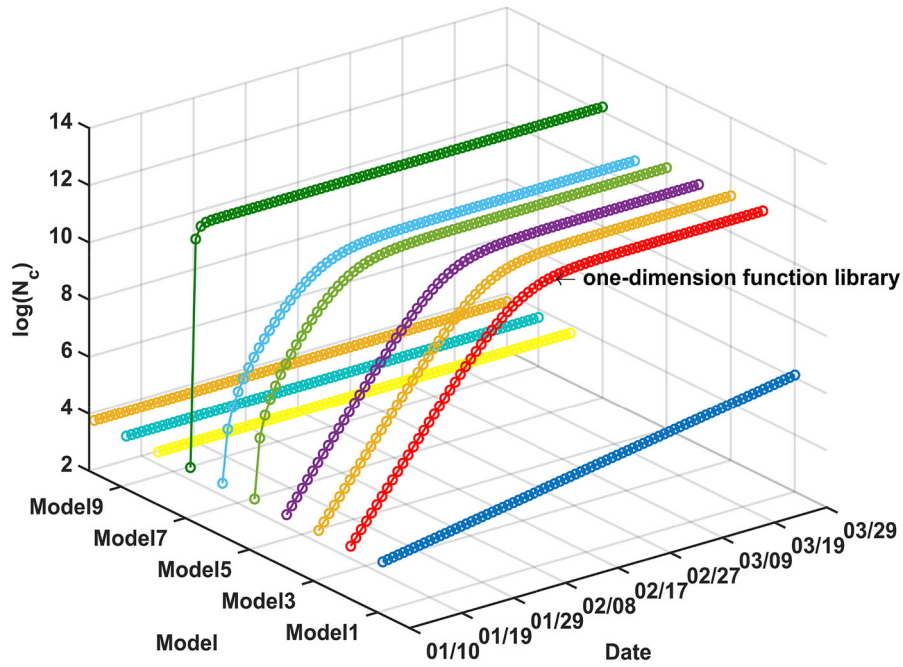


Fig. 8 Prediction of the cumulative number confirmed cases in Chinese mainland obtained by the models generated by 10 different kinds of candidate function libraries (after taking the logarithm)

The red dot in Fig. 6 denotes the true epidemic turning point (16 February) defined by Eq.(27), and the yellow dot denotes the predicted epidemic turning point using the SINDy-LM method (13 February). The difference of 3 days between the two shows that SINDy-LM method can make a reliable prediction of the epidemic turning point. In addition, it should be noted that the prediction results of the COVID-19 model using SINDy-LM method for the next 10 days after February 6 are relatively reliable, except for February 11, when Hubei changed its definition of "confirmed infected" since it used the newest diagnostics.

4 Discussion

4.1 Discussion of high-dimension case

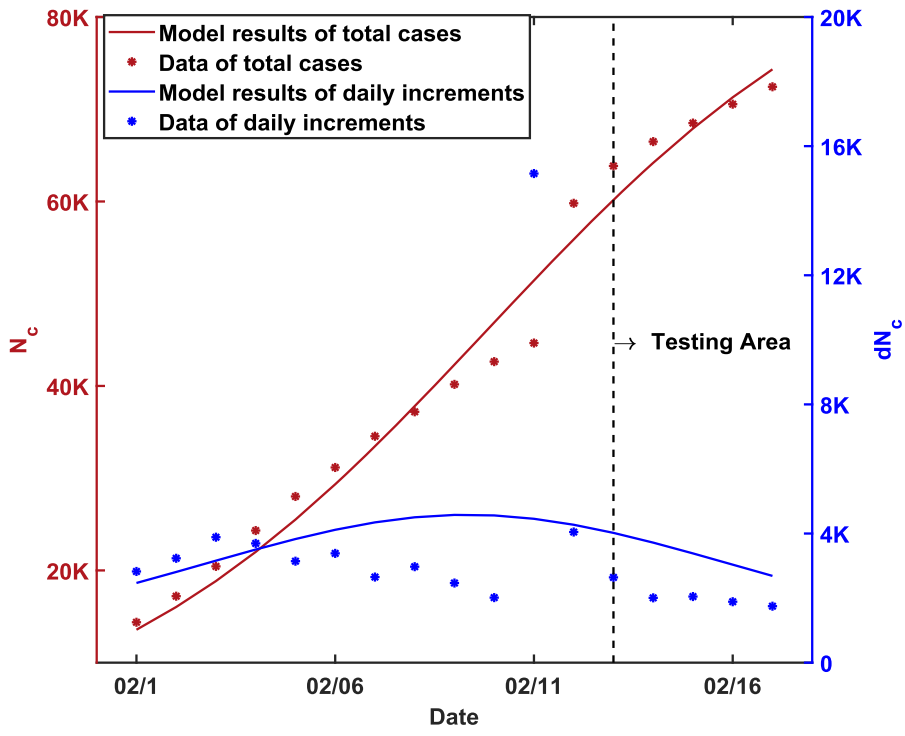
In Sect. 2.1, the candidate function library $\boldsymbol{\Theta}(\mathbf{x})$ (Eq. (5)) is simplified to one-dimension case, namely $\boldsymbol{\Theta}(x_i)$. This section will discuss how to establish a COVID-19 model in a higher dimensions case to describe the epidemic and prove that the simplified method in Sect. 2.1 is reasonable.

In the high-dimension case, the candidate library $\boldsymbol{\Theta}(\mathbf{x})$ will be expressed as Eq. (28). (It should be noted that $\boldsymbol{\Theta}(\mathbf{x})$, which is a row vector, is different from the numerical matrix $\boldsymbol{\Theta}(\mathbf{X})$, which contains time series data.)

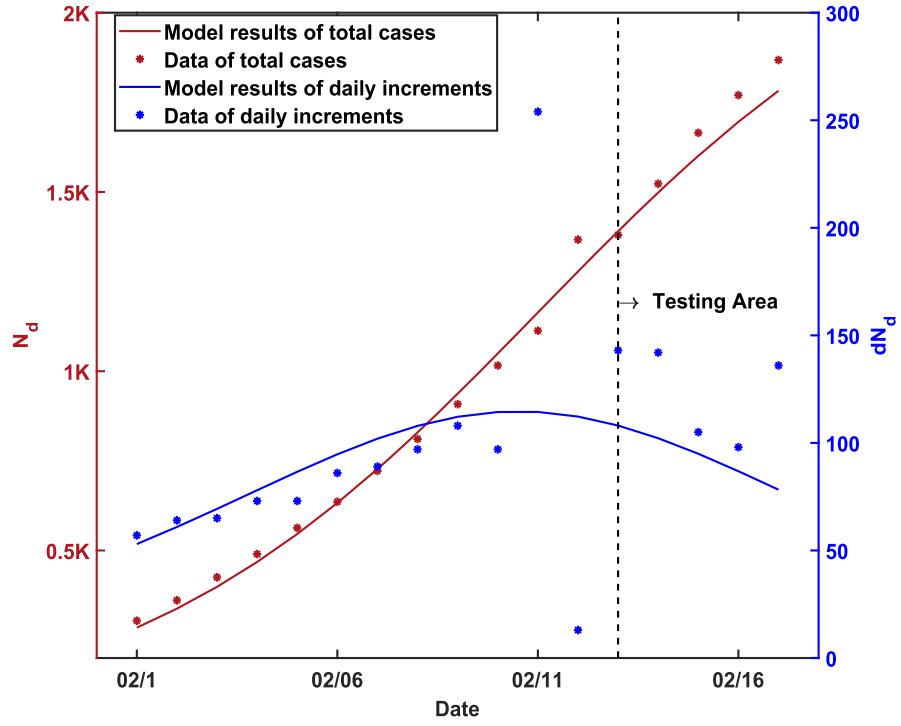
$$\boldsymbol{\Theta}(\mathbf{x}) = \left[1 \; x_1 \; \cdots \; x_n \; x_1^2 \; x_1x_2 \; x_1x_3 \; \cdots \; x_1^p \; \cdots \right] \quad (28)$$

Considering the interpretability of the model and the relatively accurate results in Sect. 3.1, p is set as 2. At the same time, a permutation and combination method is adopted to select the functions in the candidate function library, forming different subsets of function libraries ("sub-libraries") and generate models to verify the influence of each dynamic term in high-dimension situation. In the case of $n = 4$, there are $\sum_{i=0}^{p} C_i^n = 14$ terms in the entire high-dimension function library, which can form $2^{14} = 16384$ sub-libraries according to whether each term is in the subset of function library. That is, 16384 sub-models can be generated, as shown in Fig. 7.

Based on the interpretability and simplification of the model, 10 sub-libraries that are most likely to generate the Logistic model are explored further, including one-dimension function libraries, and the SINDy-LM method is used to generate 10 different models. Solving their corresponding Eqs. (4) and (9), the prediction results of the cumulative number confirmed cases in Chinese mainland are shown in Fig. 8. The model corresponding to the red curve pointed by the arrow in the figure (Model 2) uses the one-dimension function



(a)



(b)

Fig. 9 Forecast results of short term using 35-day of data followed by (a) the cumulative number of confirmed cases (b) the cumulative number of deaths

Table 3 Predicted results and relative error of short term(the cumulative number of confirmed cases and deaths)

Confirmed cases	Forecast	Reality	Relative error
15 February	64189.36	66492	0.03463
16 February	67907.59	68500	0.008648
17 February	71291.87	70548	0.010544
18 February	74327.53	72436	0.026113
Deaths	Forecast	Reality	Relative error
15 February	1498.372	1523	0.016171
16 February	1600.558	1665	0.038704
17 February	1695.521	1770	0.042079
18 February	1782.367	1868	0.045842

library, which is consistent with the model obtained after simplification in Sect. 2.1.

Through the goodness of fit of these 10 models (Table 2), it can be found that the R^2 of Model 2, which uses the one-dimension function library, is significantly higher than other models with high-dimension function libraries. On the basis of the interpretability and sparsity of model , nearly other 200 models generated by the function libraries which may obtain logistic models are also calculated for their goodness of fit. Moreover, the results indicate that the R^2 of these models are all inferior to that of the model generated by one-dimension function library, so it can be demonstrated that the one-dimension candidate function library $\boldsymbol{\Theta}(x_i)$ in Sec. 2.1 (Eq. (5)) is reasonable.

4.2 Robustness analysis

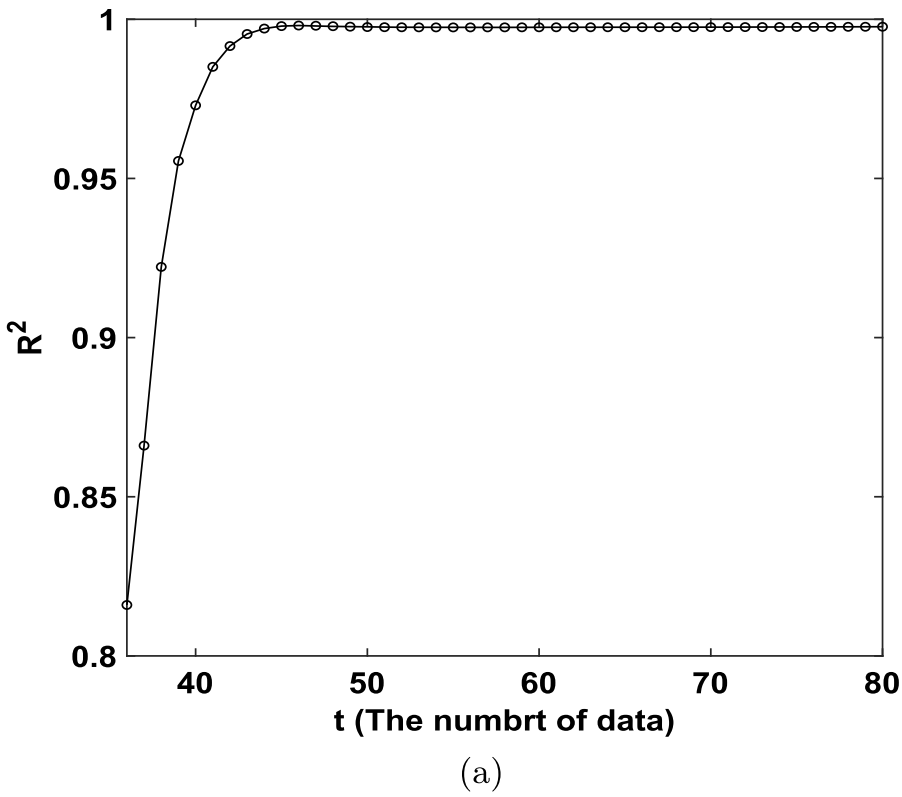
The effect of the model obtained via the data-driven method usually depends on the choice of the training set, we choose actual observation data in different time periods as the training set to verify the robustness of the model.

First, using 35-day data as the training set, it is found that the obtained model has relatively good results in the forecast of short term. Fig. 9 shows the prediction obtained by adopting this model to forecast the cumulative number of confirmed cases and the cumulative number of deaths in the future 4 days. The relative error is shown in Table 3. According to these results, we can draw the conclusion: when the training set is required to have less data, the model generated by the SINDy-LM method can be used for short-term forecast and obtain relatively reliable results.

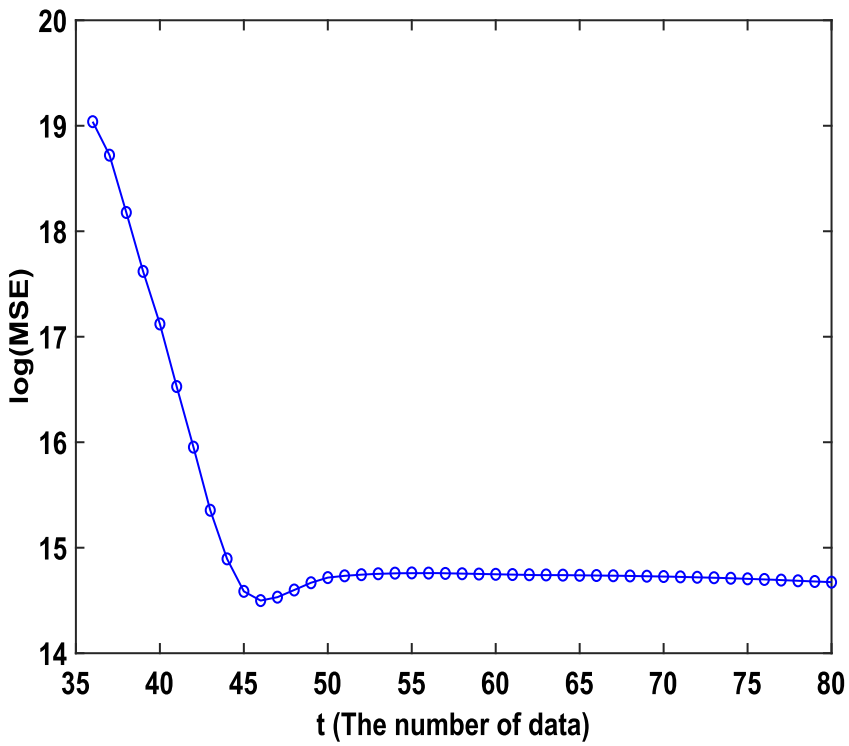
Then, we observe the effect of fitting again using some statistics. In this study, the goodness of fit is used to measure the degree of agreement between the model and the actual observation data. The statistic for measuring the goodness of fit is the coefficient of determination R^2, which is defined as Eq. (29):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} [y_i - \bar{y}]^2}{\sum_{i=1}^{n} [f(t_i) - y_i]^2}. \tag{29}$$

In Eqs.(29) and (30), $f(t_i)$ represents the fitted value, and y_i represents the observed value. According to the definition of Eq. (29), the model fits the observations better when R^2 is closer to 1. On the contrary, the fit is worse and the model is unreliable when the value of R^2 is smaller. The true observation data from 10 January, 2020 to 15 February, 2020 are selected as the original training set and The data from 16 February, 2020 to 1 April, 2020 are used as the initial testing set. The SINDy-LM method is utilized to establish a COVID-19 model, and a function graph with the coefficient of determination varying with the amount of data used in the training set is obtained as shown in Fig. 10(a). According to Fig. 10(a), the coefficient of determination R^2 shows an increasing trend with the amount of data used in the first 40 days. This result is due to the lack of data in the previous period and the unstable model. However, R^2 has exceeded 0.95 after using 40 days of data. The fluctuation range has shrunk, and it has remained stable in an infinitely close range. The model has been very stable and robust after the testing set data volume exceeds 40 days. This result

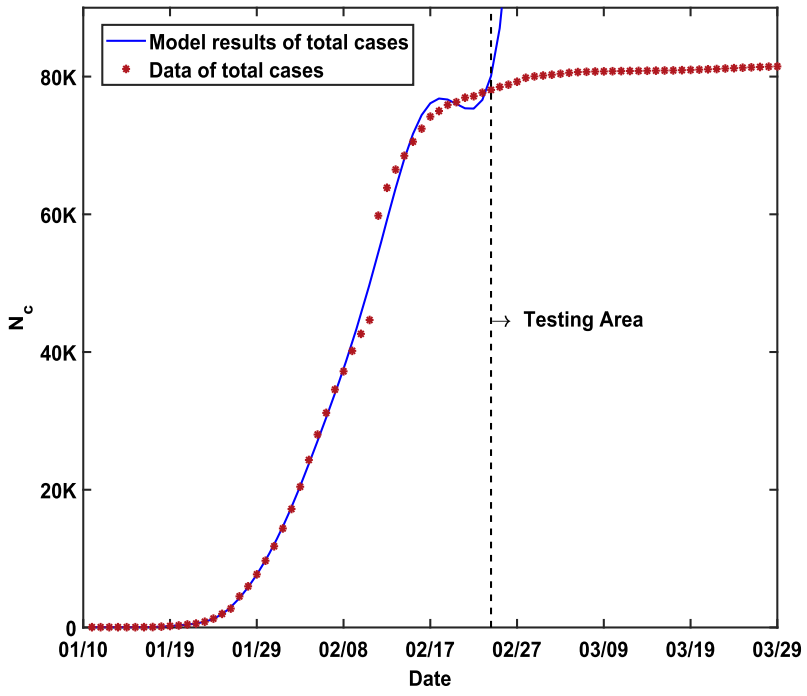


(a)

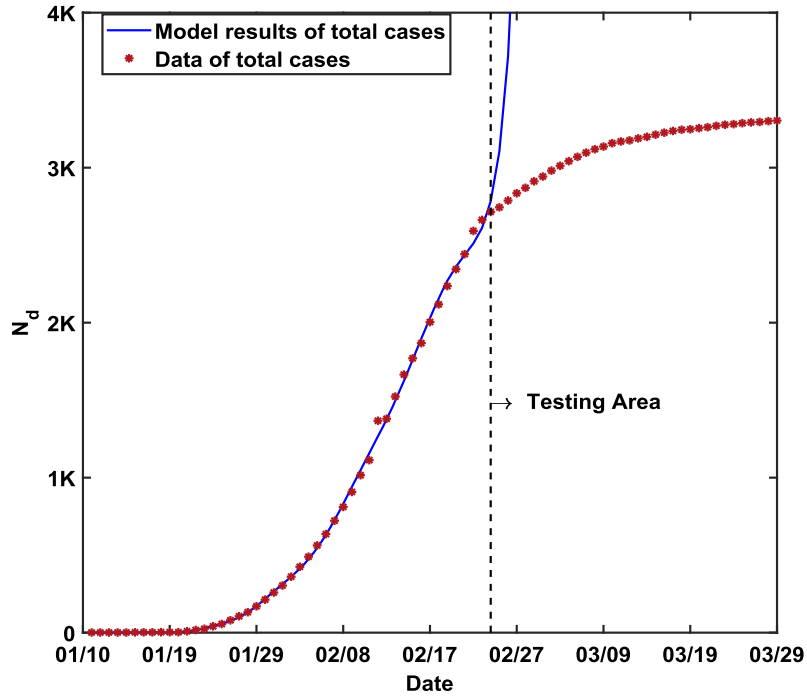


(b)

Fig. 10 Robustness analysis of the model followed by (a) variation in goodness of fit (R^2) with the amount of data used and (b) the variation of mean square error (MSE) with the amount of data used (after taking the logarithm)



(a)



(b)

Fig. 11 Review results of Chinese mainland using ELM method followed by (a) the cumulative number of confirmed cases and (b) the cumulative number of deaths

Table 4 Relative error of the SINDy-LM method and ELM method

Forecast days	1 day	3 days	5 days	7 days
SINDy-LM	0.0156	0.041	0.079	0.123
ELM	0.551	0.721	0.891	0.919

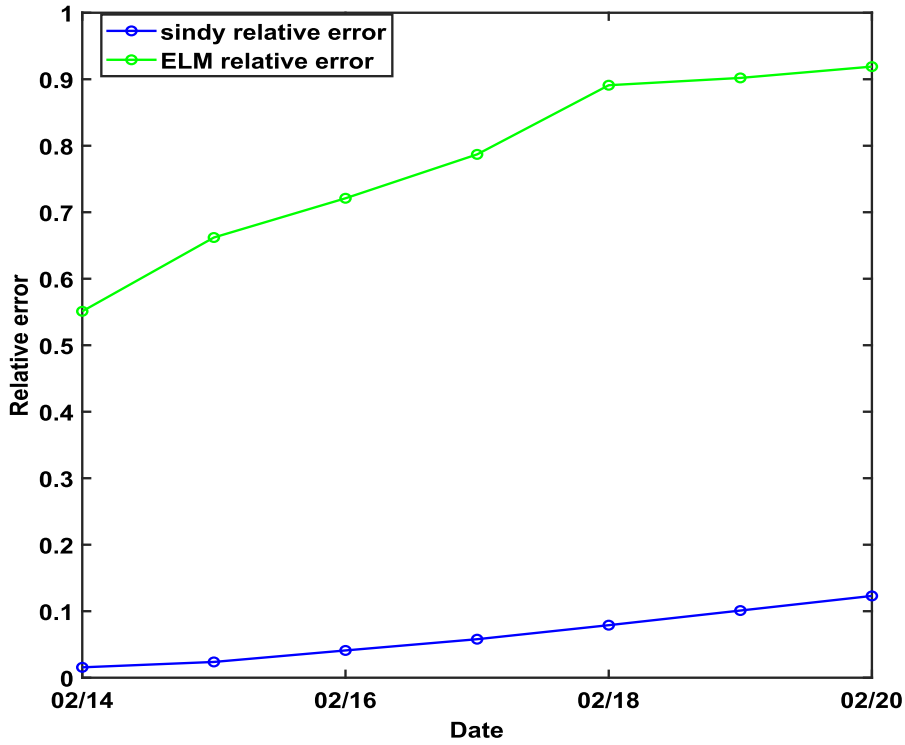


Fig. 12 Relative error of the ELM and SINDy-LM methods varies with time

implies that the SINDy-LM method only needs to utilize limited data to build a COVID-19 model with high prediction accuracy. This property has a very positive effect on the prediction and control of the middle and late stages of the epidemic.

In addition to the goodness of fit, the mean squared error (MSE) is often regarded as an important indicator to measure the prediction results in practical applications. MSE represents the average of the sum of squares of the difference between the predicted and true values, and its mathematical expression is

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}[f\,(t_i) - y_i]^2\,. \tag{30}$$

Given Eq. (30), MSE is not fixed. It will have a certain change depending on the selected data changes. We make a chart of MSE with the variation in amount of data used and also study the changes of accuracy in the model.

Figure 10(b) is a figure of the mean square error with the amount of data used. The mean square error shows a decreasing trend as the amount of data increases in the first 35 days. This result is due to the lack of data in the previous period and the low degree of model fit and accuracy. As the amount of data increases, the model tends to stabilize, and the degree of fit and accuracy are significantly improved. After using 46 days of data, the mean square error increases slightly. This increase is due to that the base number of confirmed cases has reached a large value after 45 days, and the mean square error is slowly increasing because of its influence. After using 53 days of data, the mean square error has reached a relatively small value, and it begins to decline slowly. The accuracy of the model also continues to improve.

4.3 Comparison with other methods (ELM algorithm)

In the prediction of the COVID-19 epidemic, many institutions use machine learning methods except analytical modeling methods to build prediction models for COVID-19. Among them, Johns Hopkins University used the method of ELM [11]. The ELM is a learning algorithm based on single hidden layer feedforward neural networks (SLFNs), which can omit complex analysis and modeling procedure and simplify calculations. However, when the original data are mixed with a large number of noise variables, the classification and regression accuracy of the ELM algorithm are greatly reduced, and the robustness of the obtained model is poor. In addition, this neural network-based machine learning algorithm cannot produce a model with interpretability.

To further explore the advantages of the SINDy-LM method compared with machine learning algorithms, this study compares the two methods by reviewing and predicting the epidemic in China.

First, the data of the first 45 days of the epidemic in China are taken as the training set, and the data of the last 35 days as the testing set. The activation function in SLFNs is composed of elementary function libraries, which must be infinitely differentiable. The input weights and hidden layer deviations of SLFNs can be randomly assigned according to the relevant theory of ELMs. Therefore, SLFNs can be simply considered a linear system. We can analytically calculate the output weights connecting the hidden and output layers through simple generalized inverse operations of the hidden layer output matrix to determine the structure of SLFNs and give the prediction results. The review and prediction results of the epidemic situation in Chinese mainland by using the ELM method are shown in Fig. 11.

In Fig. 11, the results in (a) are obtained using the ELM method to obtain the review and prediction results of the cumulative number of confirmed cases in China; those in (b) are obtained using the ELM method to obtain the review and prediction results of the cumulative number of deaths in China. The blue line represents the model result, the red dots represent the true data, the left side of the black dotted line is the training set, and the right side is the testing set.

According to the results in Fig. 11, the gap between the predicted results obtained using the ELM method and the actual results becomes quite large when predicting the cumulative number of confirmed cases and cumulative number of deaths in China in the next 35 days. Moreover, the development of the cumulative number of confirmed cases and deaths does not match the facts. Similar results are obtained after changing the amount of data in the training set, which shows that the robustness of the model is quite bad. The relative error produced by the short-term prediction is quantified and compared.

Table 4 and Fig. 12 show that the prediction accuracy obtained by SINDy-LM method in epidemic prediction outperforms that of the ELM method. However, the relative error of the ELM method will also accumulate quickly, and large error and instabilities will occur when making long-term predictions. Thus, the accuracy and robustness of the model are poor. To be brief, the relative error obtained by the SINDy-LM method is smaller, and it is more stable over time. Therefore, the obtained model by SINDy-LM has better accuracy and robustness than that by ELM.

5 Conclusion

In this study, a SINDy-LM method is proposed to model and study the COVID-19 transmission system, which balances complexity and prediction accuracy simultaneously. First, the prediction results of this method in Chinese mainland, Australia, and Egypt are given, which indicate high accuracy and a certain universality of the derived method. Especially, the studied model accurately determines that the " the epidemic turning point" in China will appear on 13 February, 2020. Second, the COVID-19 model produced by this method has strong sparsity, and the result has less fluctuation with the amount of data used. It also demonstrated that, in terms of describing COVID-19 epidemic, to simplify the candidate function library to a one-dimension case is reasonable. Finally, comparing the "SINDy-LM" and ELM methods, it can be demonstrated that the former is better than the latter in terms of model interpretability and prediction accuracy, which plays an important role in providing action guidelines to do a good job in epidemic prevention and control under nomal conditions.

Although our method is effective for COVID-19 epidemic modeling and can be extended to data-driven modeling of other complex systems, there are still some aspects that need to be improved in the future.

First, the derivatives in Eq. (4) rely on numerical differentiation. In this study, derivatives are taken using finite differences for clean data. However, numerical approximations of derivatives are inherently unstable due to the introduction of truncation and round-off errors [31]. Thus, some methods with high accuracy and stability, such as automatic differentiation, can be utilized in the modeling process in SINDy-LM method.

In addition, it is noteworthy that the performance of the SINDy-LM method is not very satisfactory when modeling multi-phase epidemics on a global scale, and the results obtained are not as good as the high-dimensi- onal differential system with composite traditional models [32]. The features of data at different stages are quite diverse, which is a disadvantage for most data-driven methods. However, for countries where multi-phased evolution appeared, the data set can be divided and SINDy can be utilized in different stages, which will be studied for further work and expected to obtain relatively good results for each stage.

Moreover, we will also attempt to apply the SINDy-LM algorithm to the parameter identification in the traditional analytical modeling, in order to solve the problem of difficult parameter identification in classical analytical modeling and make up for the poor interpretability of data-driven modeling simultaneously.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Nos. 11972292, 11902252, 11672233), the foundation of National Key Laboratory of Science and Technology on Aero-dynamic Design and Research (No. 614220119040101) and the National Training Program of Innovation and Entrepreneurship for Undergraduates (No. S202010699139). The authors also thanks Jiumei He and Hao Yuan for their discussion.

Data Availability Statement The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Kermack, W.O., McKendrick, A.G.: 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)
2. Anderson, R.M., May, R.M.: Population biology of infectious diseases: Part I. Nature **280**(5721), 361–367 (1979)
3. Coronavirus: Common symptoms,preventive measures, and how to diagnose it. Caringly Yours [Online]. Available: https://www.caringlyyours.com/content /2020/01/28 (2020)
4. Huang, S.Z., Peng, Z.X., Jin, Z.: Research on novel coronavirus pneumonia epidemic control strategy: efficiency evaluation and recommendations. Sci. China: Math. **50**(06), 885–898 (2020)
5. Teles, P.: Predicting the evolution of Covid-19 in Portugal using an adapted SIR model previously used in South Korea for the MERS outbreak. medRxiv [Online]. Available: https://www.medrxiv.org/content/early/2020/03/25/2020.03.18.20038612 (2020)
6. Tang, B., Wang, X., Li, Q., et al.: Estimation of the transmission risk of 2019-nCov and its implication for public health interventions. J. Clin. Med. **9**(2), 462 (2020)
7. Gaurav, P., Poonam, C., Rajan, G., et al.: SEIR and Regression Model based COVID-19 outbreak predictions in India. arXiv preprint arXiv: 2004.00958 (2020)
8. He, S., Peng, Y., Sun, K.: SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dyn. **101**(3), 1667–1680 (2020)
9. Hou, C., Chen, J., Zhou, Y., et al.: The effectiveness of quarantine of Wuhan city against the Corona Virus Disease 2019 (COVID-19): A well-mixed SEIR model analysis. J. Med. Virol. **92**(7), 841–848 (2020)
10. Yang, Z., Zeng, Z., Wang, K., et al.: Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. Journal of Thoracic Disease. **12**(3), 165 (2020)
11. Javid, A.M., Liang, X., Venkitaraman, A., et al.: Predictive Analysis of COVID-19 Time-series Data from Johns Hopkins University. arXiv preprint arXiv:2005.05060 (2020)
12. Loiseau, J.C., Brunton, S.L.: Constrained sparse Galerkin regression. arXiv preprint arXiv:1611.03271 (2016)
13. Schaeffer, H.: Learning partial differential equations via data discovery and sparse optimization. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. **473**(2197), 1364–5021 (2017)
14. Huang, N.E., Qiao, F., Tung, K.K.: A data-driven tool for tracking and predicting the course of COVID-19 epidemic as it evolves. medRxiv [Online]. Available:https://www.medrxiv.org/content/early/2020/05/23/2020.03.28.20046177 (2020)
15. Hermanowicz, S.W.: Simple model for Covid-19 epidemics-back-casting in China and forecasting in the US. medRxiv [Online]. Available:https://www.medrxiv.org/content/2020/04/03/2020.03.31.20049486 (2020)
16. Huang, N.E., Qiao, F.: A data driven time-dependent transmission rate for tracking an epidemic: a case study of 2019-nCoV. ence Bulletin. 65(6), 425-427 (2020)
17. Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Nat. Acad. Sci. **113**(15), 3932–3937 (2016)
18. Mangan, N.M., Kutz, J.N., Brunton, S.L., et al.: Model selection for dynamical systems via sparse regression and information criteria. Proceedings Mathematical Physical and Engineering Sciences. **473**(2204), 1364–5021 (2017)

19. Corbetta, M.: Application of sparse identification of nonlinear dynamics for physics-informed learning. 2020 IEEE Aerospace Conference. IEEE, 1-8 (2020)
20. Ljung, L.: System identification: theory for the user. Tsinghua University Press, Tsinghua (2002)
21. Suhubi, E.S.: Nonlinear oscillations, dynamical systems, and bifurcations of vector fields: Applied Mathematical Science, Vol. 42, J. Guckenheimer and P. Holmes, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo (1983). XVI + 453 pp. 206 figs, DM 104. International Journal of Engineering Science. 26(2), 221-222 (1988)
22. Schmidt, M., Lipson, H.: Distilling free-form natural laws from experimental data. science. 324(5923), 81-85 (2009)
23. Wang, B.X.: Research on LM optimization algorithm and neural network predictive control in nonlinear systems. Taiyuan University of Technology Press, Taiyuan (2016)
24. Lourakis, M.I.A.: A brief description of the Levenberg-Marquardt algorithm implemented by levmar. Found. Res. Technol. **4**(1), 1–6 (2005)
25. Wilamowski, B.M., Yu, H.: Improved computation for levenberg-marquardt training. IEEE Trans. Neural Net. **21**(6), 930–7 (2010)
26. Brunton, S.L., Kutz, J.N.: Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. (2019)
27. Wu, K., Darcet, D., Wang, Q., et al.: Generalized logistic growth modeling of the COVID-19 outbreak in 29 provinces in China and in the rest of the world. arXiv preprint arXiv:2003.05681 (2020)
28. Kermack, W., McKendrick, A.: A contribution to the mathematical theory of epidemics. Proc. Roy. Soc. London A **115**, 700–721 (1927)
29. Heymann, D.L., Shindo, N.: COVID-19: what is next for public health? Lancet **395**(10224), 542–545 (2020)
30. Huang, N.E., Qiao, F.: A data driven time-dependent transmission rate for trackingan epidemic: a case study of 2019-nCoV. Sci. Bull. **65**, 425–427 (2020)
31. Atilim G. B., Barak A P., Alexey A. R., et al: Automatic differentiation in machine learning. arXiv preprint arXiv:1502.05767 (2015)
32. Liu, X., Zheng, X., Balachandran, B.: COVID-19: data-driven dynamics, statistical and distributed delay models, and observations. Nonlinear Dyn. **101**(3), 1527–1543 (2020)