Cantó et al. *Advances in Difference Equations* (2017) 2017:33
DOI 10.1186/s13662-017-1078-5

Advances in Difference Equations
a SpringerOpen Journal

RESEARCH Open Access

Estimation of parameters in a structured SIR model

Begoña Cantó*, Carmen Coll and Elena Sánchez

*Correspondence:
bcanto@mat.upv.es
Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera, 14, Valencia, 46022, Spain

Abstract

In this paper, an age-structured epidemiological process is considered. The disease model is based on a SIR model with unknown parameters. We addressed two important issues to analyzing the model and its parameters. One issue is concerned with the theoretical existence of unique solution, the identifiability problem. The second issue is how to estimate the parameters in the model. We propose an iterative algorithm to study the identifiability of the system and a method to estimate the parameters which are identifiable. A least squares approach based on a finite set of observations helps us to estimate the initial values of the parameters. Finally, we test the proposed algorithms.

Keywords: epidemic process; mathematical modeling; parameter estimation; identifiability; discrete-time system

1 Introduction and problem statement

In the past decades, dynamical systems have been used to develop models in physics, biology, chemistry, engineering and epidemiology; see for instance [1, 2] and [3]. Usually equations modeling these phenomena depend on several parameters. Some of them have a scientific meaning and others might come from approximations. Unfortunately, most of the parameters are unknown. Then the parameter estimation is essential for modeling biological systems. Moreover, to execute a parameter estimation task, first one needs to ensure the identifiability of the system, since if the number of unknown parameters is very large, it is often impossible to find a unique solution to this problem.

Identification of systems deals with the problem of modeling of systems in which previous information available is not sufficient to determine all the parameters involved in the model. The identification property has been studied on topics related with dynamic systems [4–6]. Parameter estimation is the process of trying to calculate the model parameters based on a dataset. Often, some of the parameters can be measured, while the rest can only be fitted. A crucial tool in the fitting process is assigning of the parameter values so that the errors between the measured variables and the corresponding model predictions are minimized. The process consists in assuming that the values of the parameters of a given system are unknown, but that we have recorded inputs and outputs over a time interval. The usual estimation methods include the projection algorithm, gradient algorithm, and least squares algorithm [7, 8].

Therefore, real data are needed to construct and validate models. The identification and estimation problems will be used to find the model that best fits the data from a set of candidates. Finally, it is necessary to evaluate and to validate if the model satisfies the process properties.

In this paper, parameter estimation for an epidemic model has been tackled in the framework of control theory. The algorithm is developed by exploiting the special structure of our model.

Consider a nonlinear system representing an epidemiological model given by

$$x(k+1) = f\big(x(k), \mathbf{p}\big), \quad k \in \mathbb{Z},$$

where $x(k) \in \mathbb{R}^n$ is the state vector, $\mathbf{p} \in \mathbb{R}^l$ is the parameter vector and $f : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}^n$ is a continuously differentiable function. This system can be linearized around the disease free equilibrium point which is defined as the equilibrium point where no disease is present in the population. In our problem this linear system is of the type

$$x(k+1) = A(\mathbf{p})x(k) + b, \quad k \in \mathbb{Z}, \tag{1}$$

where $A(\mathbf{p}) \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. The matrix coefficients have a fixed structure and it is now interesting to obtain the parameters required by the model structure. This property is known as the identifiability problem. Given a parameterized model it is important to uniquely identify the parameters since it is necessary to do experimental design and to estimate the unknown parameters of the model using experimental data.

The identifiability problem is based on the determination of all parameter sets which give the same input-output response. The identifiability of the system (1) depends on if the parameters can be determined uniquely from a known output of the system [9]. That is, given two parameter vectors $\mathbf{p}$, $\bar{\mathbf{p}}$, if we denote by $x_{\mathbf{p}}(k)$ and $x_{\bar{\mathbf{p}}}(k)$ the output of the system (1), respectively, then the equation $x_{\mathbf{p}}(k) = x_{\bar{\mathbf{p}}}(k)$, for all $k \geq 0$, implies $\mathbf{p} = \bar{\mathbf{p}}$.

For linear models there are many well-established techniques to analyze structural identifiability; see, for example, [10–12] and the references therein.

After analyzing the identifiability property we consider the estimation problem. Parameter estimation is an important issue in biological systems because it is useful for obtaining predictions of computer models of biological systems step. This problem is usually addressed by fitting model simulations to the observed experimental dataset $ob(i) \in \mathbb{R}^n$, $i = 1, \ldots, K$. The filter is well known in control and estimation theory and has application in a wide range of fields such as epidemiology, weather forecasting and economy.

To solve the estimation problem we rewrite the system (1) obtaining

$$x(k+1) = M(k)\mathbf{p} + N(k), \quad k \in \mathbb{Z}.$$

Define $e(i) = (ob(i) - x(i))$, $i = 1, \ldots, K$ $e_K = \operatorname{col}(e(i))_{i=0}^{K-1}$, and

$$d_K = \operatorname{col}\big(d(i)\big)_{i=1}^{K} = \operatorname{col}\big(ob(i) - N(i-1)\big)_{i=1}^{K}, \qquad H_K = \operatorname{col}\big(M(i)\big)_{i=0}^{K-1}.$$

Then, for K observations, we want to find the parameter vector which minimizes the quadratic cost function

$$\begin{aligned} J_K(\mathbf{p}) &= \frac{1}{2}\sum_{i=1}^{K} e(i)^T e(i) \\ &= \frac{1}{2} e_K^T e_K \\ &= \frac{1}{2}(d_K - H_K\mathbf{p})^T(d_K - H_K\mathbf{p}). \end{aligned}$$

Efficient parameter estimation methods can be found in the literature. A common principle for most of them is to minimize the error between the observed and predicted quantities, often reaching a local optimum, and getting the solution requires intensive computation. One of the most usual methods for estimating parameters is a gradient-based regression algorithm. A good overview of different methods developed to estimate the parameters can be found in [13].

2 Age-structured SIR model. Identifiability

In this work we study the parameter estimation of an age-structured SIR model where the individuals are organized in compartments from an age range. The age structure is critical in modeling epidemics caused by certain common diseases such as measles and influenza or sexually transmitted diseases (STD); one that many people are worried about getting is HIV. The choice of this type of structure is because the outcome of the epidemic may depend sensitively on the contact structure, the recovery rate, and the death rate. For example, to analyze an infectious disease such as measles, the population can be divided into five age groups or compartments, the age grouping 0-2, 3-6, 7-12, 13-16, 16+, corresponding to the main school grades in Spain. So, we propose a discrete age-structured SIR model on the basis of age has great influence on the spread of infectious disease. It is formulated using the usual parameters in mathematical epidemiology. For different values of this parameter we can get different types of infections. That is, we divide thee population into m compartments according to their ages but these compartments need not have the same range of ages. Thus, we have susceptible individuals S_i, infected individuals I_i and recovered individuals R_i at the ith age compartment, for $i = 1, \dots, m$.

We consider that only the individuals of the same age range are in contact, so the susceptible individuals only are infected from the infected individuals of its compartment.

We take account the transference of individuals from the ith compartment to the $(i+1)$th compartment when they change the age range in the dynamic process and, moreover, we consider the entry of new individuals in S_1 proportional to the size of the population $\beta(k)N$, in order for the size the population N to remain constant.

Therefore, the dynamic process is described in Figure 1, where there are 10 different parameters, five of them are associated to each age compartment and the rest are associated to the transference of individuals between consecutive compartments, which are defined in Table 1.

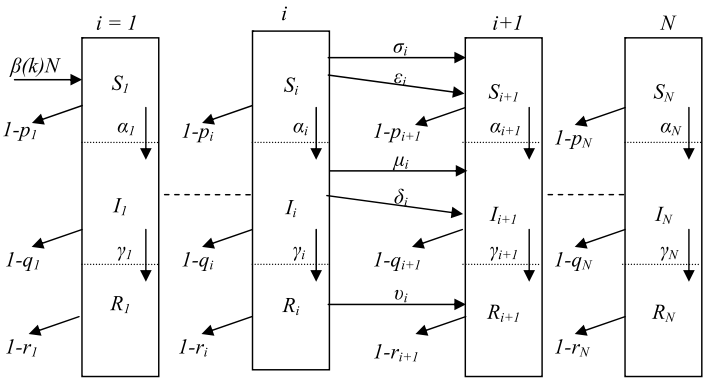


Figure 1 Dynamic process between *m* age compartments at time *k*.

Table 1 Parameters in an age compartment-structured SIR model.

Parameters at *i*th compartment	
p_i, q_i, r_i	Survival rates of S_i, I_i, R_i.
α_i	Exposition rate of susceptible individuals S_i by contact with infected individuals I_i.
γ_i	Rate of infected individuals becoming recovered individuals.
σ_i, μ_i, ν_i	Rate individuals changing of age-compartment without changing the state.
ϵ_i, δ_i	Rate individuals changing of age-compartment with changing the state.

Without loss of generality we consider the model with three compartments, $m = 3$, and its mathematical representation is given by the nonlinear discrete-time system $x(k+1) = f(x(k))$:

$$
\begin{aligned}
i = 1 \quad & \begin{cases} S_1(k+1) = (p_1 - \sigma_1 - \alpha_1 \frac{I_1(k)}{N}) S_1(k) + \beta(k) N, \\ I_1(k+1) = (q_1 - \gamma_1 - \mu_1) I_1(k) + (1-\epsilon_1)\alpha_1 \frac{I_1(k)}{N} S_1(k), \\ R_1(k+1) = (r_1 - \nu_1) R_1(k) + (1-\delta_1)\gamma_1 I_1(k), \end{cases} \\
i = 2 \quad & \begin{cases} S_2(k+1) = \sigma_1 S_1(k) + (p_2 - \sigma_2 - \alpha_2 \frac{I_2(k)}{N}) S_2(k), \\ I_2(k+1) = \mu_1 I_1(k) + (q_2 - \gamma_2 - \mu_2) I_2(k) + \epsilon_1 \alpha_1 S_1(k) \frac{I_1(k)}{N} + (1-\epsilon_2)\alpha_2 S_2(k) \frac{I_2(k)}{N}, \\ R_2(k+1) = \nu_1 R_1(k) + (r_2 - \nu_2) R_2(k) + \delta_1 \gamma_1 I_1(k) + (1-\delta_2)\gamma_2 I_2(k), \end{cases} \\
i = 3 \quad & \begin{cases} S_3(k+1) = \sigma_2 S_2(k) + (p_3 - \alpha_3 \frac{I_3(k)}{N}) S_3(k), \\ I_3(k+1) = \mu_2 I_2(k) + (q_3 - \gamma_3) I_3(k) + \epsilon_2 \alpha_2 S_2(k) \frac{I_2(k)}{N} + \alpha_3 S_3(k) \frac{I_3(k)}{N}, \\ R_3(k+1) = \nu_2 R_2(k) + r_3 R_3(k) + \delta_2 \gamma_2 I_2(k) + \gamma_3 I_3(k), \end{cases}
\end{aligned}
$$

with

$$
\beta(k) N = N - \sum_{i=1}^{3} p_i S_i(k) - \sum_{i=1}^{3} q_i I_i(k) - \sum_{i=1}^{3} r_i R_i(k).
$$

Taking $x(k) = \operatorname{col}(x_i(k))_{i=1}^{3}$ with $x_i(k) = (S_i(k)\ I_i(k)\ R_i(k))^T$, $i = 1, 2, 3$, and linearizing around the disease-free equilibrium point $x^\star = f(x^\star, \mathbf{p})$, which is given by $P_f = (S_1^f, 0, 0, S_2^f, 0, 0, S_3^f, 0, 0)$ with

$$
S_1^f = \frac{N(1 - p_2 + \sigma_2)(1 - p_3)}{K}, \qquad S_2^f = \frac{\sigma_1 N (1 - p_3)}{K}, \qquad S_3^f = \frac{\sigma_1 \sigma_2 N}{K},
$$

where $K = (1 - p_2 + \sigma_2 + \sigma_1)(1 - p_3) + \sigma_2\sigma_1$, we obtain the following linear discrete-time system:

$$x(k+1) = A(\mathbf{p})x(k) + b, \tag{2}$$

where $b = (N\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)^T$ and

$$A(\mathbf{p}) = \begin{pmatrix} -\sigma_1 & -(f_1+q_1) & -r_1 & -p_2 & -q_2 & -r_2 & -p_3 & -q_3 & -r_3 \\ 0 & h_1-\gamma_1 + (1-\epsilon_1)f_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma_1(1-\delta_1) & r_1-\nu_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma_1 & 0 & 0 & p_2-\sigma_2 & -f_2 & 0 & 0 & 0 & 0 \\ 0 & \mu_1+\epsilon_1 f_1 & 0 & 0 & h_2-\gamma_2 + (1-\epsilon_2)f_2 & 0 & 0 & 0 & 0 \\ 0 & \gamma_1\delta_1 & \nu_1 & 0 & \gamma_2(1-\delta_2) & r_2-\nu_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_2 & 0 & 0 & p_3 & -f_3 & 0 \\ 0 & 0 & 0 & 0 & \mu_2+\epsilon_2 f_2 & 0 & 0 & q_3-\gamma_3 + f_3 & 0 \\ 0 & 0 & 0 & 0 & \gamma_2\delta_2 & \nu_2 & 0 & \gamma_3 & r_3 \end{pmatrix}$$

with

$$h_1 = q_1 - \mu_1, \qquad h_2 = q_2 - \mu_2,$$
$$f_1 = \alpha_1\frac{S_1^*}{N}, \qquad f_2 = \alpha_2\frac{S_2^*}{N}, \qquad f_3 = \alpha_3\frac{S_3^*}{N}.$$

From the above equations we obtain some sufficient conditions to ensure that the system has positive solutions, that is, $S_i(k)$, $I_i(k)$, and $R_i(k)$, $i = 1, 2, 3$, are nonnegative for all initial conditions. For this purpose, we give the next result.

Proposition 1 *Consider the parameters p_i, q_i, r_i, α_i, γ_i, ν_i, $i = 1, 2, 3$, and σ_j, μ_j, $j = 1, 2$ of the model. If $p_i \geq \sigma_i + \alpha_i$, $q_i \geq \gamma_i + \mu_i$, $i = 1, 2$, $p_3 \geq \alpha_3$, $q_3 \geq \gamma_3$ and $I(k) \leq \theta N$ where $\theta = \min_{i=1,2,3}\{\frac{p_i-\sigma_i}{\alpha_i}\}$, with $\sigma_3 = 0$, then $S_i(k)$, $I_i(k)$, $R_i(k)$, $i = 1, 2, 3$, are nonnegative for all initial conditions.*

Note that these conditions on the parameters are consistent. It is logical that the rate of individuals who move from one compartment to another plus the rate of individuals who recover from the disease do not exceed the rate of survival in each compartment.

2.1 Algorithm to identifiability of the parameters

We assume that the rates σ_i, μ_i and ν_i, $i = 1, 2, 3$, and the rates ϵ_i and δ_i, $i = 1, 2$, are known. Then the parameters to identify are the following:

$$\mathbf{p} = (p_i, q_i, r_i, \gamma_i, \alpha_i, i = 1, 2, 3).$$

Model predictions depend on the parameters, some of which must be estimated from experimental data. The most important characteristic is that the parameters have physical significance and that it is possible to determine their values from observed data. The

identifiability helps us test the unique relationship between parameter sets and model response and guarantees that the parameters can be estimated under ideal conditions. It is important to check the identifiability property since if a model is not correctly formulated, problems can appear in the parameter estimation.

Denoting the solution of the system (2) by $x_{\mathbf{p}}(k)$ we have $x_{\mathbf{p}}(k) = A(\mathbf{p})^k x(0) + \sum_{i=0}^{k-1} A(\mathbf{p})^i b$. To identify the parameters we consider different initial states $x(0)$ and suppose $x_{\mathbf{p}}(k) = x_{\bar{\mathbf{p}}}(k)$ for all $k \geq 0$ from two parameters $\mathbf{p}$ and $\bar{\mathbf{p}}$. Then we want to prove that $\mathbf{p} = \bar{\mathbf{p}}$. Specifically, we identify all the parameters, except p_1.

In order to solve this identifiability problem in general, that is, when we have $m \geq 3$ age compartments, we give the following algorithm.

Algorithm

Step 1 Input data: m (number of compartments), N (population); σ_i, μ_i and ν_i, $i = 1, 2, 3$; ϵ_i and δ_i, $i = 1, 2$.

Step 2 Input variable parameters: p_i, q_i, r_i, γ_i, f_i, $i = 1, 2, 3$, and $\bar{p}_i$, $\bar{q}_i$, $\bar{r}_i$, $\bar{\gamma}_i, \bar{f}_i$, $i = 1, 2, 3$.

Step 3 Obtain $n = 3m$ and introduce canonical vectors $\{e_i\}_{i=1}^n$.

Step 4 Construct b, $A(\mathbf{p})$, and $A(\bar{\mathbf{p}})$ as in (2).

Step 5 For each $i = 1, \ldots, n$, construct the output $x_{\mathbf{p}}(1)$ and $x_{\bar{\mathbf{p}}}(1)$ of the system (2) obtained from $x(0) = Ne_i$, which we denote by $\{x(i, 1, \mathbf{p})\}_{i=1}^n$ and $\{x(i, 1, \bar{\mathbf{p}})\}_{i=1}^n$, respectively.

Step 6 For $i = 1, \ldots, n$:

- Step 6.1 For each $l = 1, \ldots, n$, the lth row of $x(i, 1, \mathbf{p})$ is denoted as $x(i, 1, \mathbf{p})^l$.
- Step 6.2 If $x(i, 1, \mathbf{p})^l \geq 0$ for all l, then solve $x(i, 1, \mathbf{p}) = x(i, 1, \bar{\mathbf{p}})$ and save the identified variable parameters.
- Step 6.3 Else check if all parameters are identified and go to Step 8. Otherwise, go to Step 7.

Step 7 $t = i$

- Step 7.1 Input initial data $\{S_h^0, I_h^0\}$ such that $S_h^0 + I_h^0 = N$ being $t = 3(h-1) + 2$ and construct initial conditions $x(0) = S_h^0 e_{t-1} + I_h^0 e_t$, denoted by $\hat{x}(t, 0)$.
- Step 7.2 Construct the output of the system (2) at time $k = 1$ obtained from $\hat{x}(t, 0)$ for each parameter $\mathbf{p}$, and $\bar{\mathbf{p}}$ which we denote by $\{\hat{x}(t, 1, \mathbf{p})\}_{t=1}^n$ and $\{\hat{x}(t, 1, \bar{\mathbf{p}})\}_{i=1}^n$, respectively.
- Step 7.3 Solve $\hat{x}(t, 1, \mathbf{p}) = \hat{x}(t, 1, \bar{\mathbf{p}})$ and save the identified parameters.
- Step 7.4 $i = t + 1$. If $i = n + 1$, then go to Step 8. Otherwise, return to Step 6.1.

Step 8 Using the definition of S_i^f and f_i, the parameters α_i are identified.

Step 9 All parameters are identified, except p_1, END.

In the process specified in the algorithm we have considered the nonnegativity of the solution. For this purpose, we have had to use the solution of the system to the initial conditions constructed at Step 7 to obtain nonnegativity outputs.

It is clear that we not only want to know if the model is or is not identifiable. Even if it is not, we want to know the parameters which can be identified, because in some cases we need only estimate some of the parameters to verify that it fulfills a hypothesis. Further, if a parameter is not identifiable, it is not estimable. Observe that the parameter p_1 is not identified due to the condition that the population is kept constant N at each time. In addition, the identifiability of the parameters does not guarantee that can be estimated as an even in the case that a parameter is identifiable, it may be difficult to estimate.

3 Parameter estimation

Now, we consider that the survival rates $p_i = p$, $q_i = q$ and $r_i = r$ are known. It assumes what ecologists refer to as Type II mortality, which is a constant mortality rate over the entire life span. This pattern is approached by most birds and some mammals [14]. Basically, Type II mortality is a good approximation for the survival rate of human populations in the developed world. Furthermore, we suppose known the transference rates between consecutive compartments σ_i, μ_i, and ν_i, $i = 1, 2, 3$, and ϵ_i and δ_i, $i = 1, 2$. Thus, from an observed dataset our goal is to find an approximate value of the parameters f_i (from them we have α_i) and the parameters γ_i, using the mathematical model given by (2). Note that α_i and γ_i are the most important rates since they give us information as regards the disease under consideration. That is, for each age range, we want to have an estimated value of the rate of infection of a susceptible individual and the rate of recovery of an infected individual, which allow us to draw conclusions on the incidence of the disease according to the age of the individual.

From an initial observation $ob(0)$, we consider an observed dataset

$$\left\{ob(k)\right\}_{k=1}^{K} = \left\{\begin{pmatrix} S_1(k) & S_2(k) & S_3(k) & I_1(k) & I_2(k) & I_3(k) & R_1(k) & R_2(k) & R_3(k) \end{pmatrix}^T\right\}_{k=1}^{K},$$

in K steps, $K \geq 1$, and, on the other hand, we have the fit mathematical model

$$x(k+1) = A(\mathbf{p})x(k) + B, \qquad x(0) = ob(0), \quad k \geq 1,$$

where, from now on, the parameter vector to estimate is

$$\mathbf{p} = \begin{pmatrix} f_1 & f_2 & f_3 & \gamma_1 & \gamma_2 & \gamma_3 \end{pmatrix}^T.$$

Rewriting the system (2) we have

$$x(k+1) = M(k)\mathbf{p} + N(k) = \begin{pmatrix} M_1(k) & M_2(k) \\ M_3(k) & M_4(k) \\ M_5(k) & M_6(k) \end{pmatrix} \mathbf{p} + N(k), \tag{3}$$

where

$$M_1(k) = \begin{pmatrix} -I_1(k) & 0 & 0 \\ (1-\epsilon_1)I_1(k) & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad M_2(k) = \begin{pmatrix} 0 & 0 & 0 \\ -I_1(k) & 0 & 0 \\ (1-\delta_1)I_1(k) & 0 & 0 \end{pmatrix},$$

$$M_3(k) = \begin{pmatrix} 0 & -I_2(k) & 0 \\ \epsilon_1 I_1(k) & (1-\epsilon_2)I_2(k) & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad M_4(k) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -I_2(k) & 0 \\ \delta_1 I_1(k) & (1-\delta_2)I_2(k) & 0 \end{pmatrix},$$

$$M_5(k) = \begin{pmatrix} 0 & 0 & -I_3(k) \\ 0 & \epsilon_2 I_2(k) & I_3(k) \\ 0 & 0 & 0 \end{pmatrix}, \qquad M_6(k) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -I_3(k) \\ 0 & \delta_2 I_2(k) & I_3(k) \end{pmatrix},$$

and $N(k) = \operatorname{col}(N_i(k))_{i=1}^{9}$ where

$$N_1(k) = N - p\sum_{i=2}^{3} S_i(k) - q\sum_{i=1}^{3} I_i(k) - r\sum_{i=1}^{3} R_i(k) - \sigma_1 S_1(k),$$

$$N_2(k) = h_1 I_1(k), \qquad N_3(k) = (r-\nu_1)R_1(k),$$

$$N_4(k) = \sigma_1 S_1(k) + (p-\sigma_2)S_2(k), \qquad N_5(k) = \mu_1 I_1(k) + h_2 I_2(k),$$

$$N_6(k) = \nu_1 R_1(k) + (r-\nu_2)R_2(k), \qquad N_7(k) = \sigma_2 S_2(k) + pS_3(k),$$

$$N_8(k) = \mu_2 I_2(k) + qI_3(k), \qquad N_9(k) = \nu_2 R_2(k) + rR_3(k).$$

From the K data of the observed dataset we want to estimate the value of $\mathbf{p}$, that is, we want to find the parameter vector which minimizes the quadratic function

$$J_K(\mathbf{p}) = \frac{1}{2}(d_K - H_K\mathbf{p})^T(d_K - H_K\mathbf{p}).$$

Thus, $\mathbf{p}$ satisfies

$$\frac{\partial J_K(\mathbf{p})}{\partial \mathbf{p}} = H_K^T H_K\mathbf{p} - H_K^T d_K = 0.$$

Note that if $S_K = H_k^T H_K$ is nonsingular, then the solution is $\mathbf{p} = S_K^{-1}H_K^T d_K$, and if it is singular, then $\mathbf{p} = S_K^{\dagger}H_K^T d_K$ where † denotes the M-Penrose generalized inverse matrix. In this last case, $\mathbf{p}$ is not identifiable since we have infinite values for the parameter and a unique output of the mathematical model.

From the structure of the matrices we can establish the following result.

Proposition 2 *Let the system be* (3). *The estimation problem has a unique solution if and only if for each* i, $i = 1,2,3$, *there exists* k_i, $0 \le k_i \le K$ *such that* $I_i(k_i) \neq 0$.

Proof If for each i, $i = 1,2,3$ there exists k_i, $0 \le k_i \le K$ such that $I_i(k_i) \neq 0$ then $\operatorname{rank}(H_{k_0}) = 6$ for $k_0 = \max\{k_i\}$. Hence, $\operatorname{rank}(S_K) = 6$. Conversely, if $\operatorname{rank}(S_K) = 6$, from $S_K = H_K^T H_K$ and using the structure of the matrices H_K and $M(k)$ the condition is proved. Therefore, we can ensure that S_K is definite positive, that is, all eigenvalues are positive and there exists S_K^{-1} for all $K > k_0$. □

Under the above assumption, we could obtain an approximated value of $\mathbf{p}$, for instance, using the descendent gradient method. That is, from an initial $\mathbf{p}_0$ the $(i+1)$th step provides us

$$\begin{aligned}\mathbf{p}_{i+1} &= \mathbf{p}_i - a_i\left(S_K\mathbf{p}_i - H_K^T d_K\right)\\ &= \mathbf{p}_i + a_i H_K^T(d_K - H_K\mathbf{p}_i),\end{aligned}$$

where a_i is the minimum of the curve $h(a) = J_K(\mathbf{p}_i - a(S_K\mathbf{p}_i - H_K^T d_K))$. Thus, we have given a numerical procedure to achieve the best fit between observed data and the parameters of the model. This algorithm is based on iterative local search in a down-hill direction from the initial point.

The parameter $\mathbf{p} \geq 0$ chosen in this epidemiological model satisfies $\|\mathbf{p}\|_2 < 1$, where $\|\cdot\|_2$ denote the spectral norm. In the next result we establish a condition on the observed dataset in order to keep this property in the process of parameter estimation.

Proposition 3 *Let the system be* (3). *For each* i, $i = 1, 2, 3$, *we suppose that there exists* k_i, $0 \leq k_i \leq K$ *such that* $I_i(k_i) \neq 0$.

$$\text{If } \|d_K\|_2 < \frac{1}{\rho(S_K^{-1})\sqrt{\rho(S_K)}} \text{ then } \|\mathbf{p}_K\|_2 < 1,$$

where $\mathbf{p}_K = S_K^{-1} H_K^T d_K$ *is the parameter which minimizes the problem associated with* K *observation data and* $\rho(\cdot)$ *denoting the spectral radius*.

Proof From $\mathbf{p}_K = S_k^{-1} H_K^T d_K$ and taking into account that S_K is a symmetric matrix

$$\|\mathbf{p}_K\|_2 = \left\|S_K^{-1} H_K^T d_K\right\|_2 \leq \rho\left(S_K^{-1}\right)\|H_K\|_2\|d_k\|_2 < \rho\left(S_K^{-1}\right)\sqrt{\rho(S_K)}\|d_K\|_2 < 1. \qquad \square$$

3.1 Adding more observations. Algorithm

Consider K data of the observed dataset, such that $I_i(k_i) \neq 0$ for some k_i, $0 \leq k_i \leq K$ and for each i, $i = 1, 2, 3$. This fact implies that there exists k_0 such that $H(k_0)$ is full rank. Now, we want to improve the approximated value of parameter $\mathbf{p}$ adding one observation $ob(K+1)$ and fitting the mathematical model to the $K+1$ data of the observed dataset. Using that

$$H_{K+1}^T H_{K+1} = H_K^T H_K + M^T(K) M(K),$$
$$H_{K+1}^T d_{K+1} = H_K^T d_K + M^T(K) d(K+1),$$

we obtain the following discrete-time variable system to represent the dynamic of the parameter vector:

$$\mathbf{p}_{K+1} = A_K \mathbf{p}_K + B_K, \quad K \geq 1, \tag{4}$$

where $A_K = S_{K+1}^{-1} S_K$ and $B_K = S_{K+1}^{-1} M^T(K) d(K+1)$, where $S_K = H_K^T H_K$.

The solution of this system is

$$\mathbf{p}_K = \Phi_A(K, k_0)\mathbf{p}_{k_0} + \sum_{j=k_0+1}^{K-1} \Phi_A(K, j+1) B_j, \quad K > k_0,$$

with the monodromy matrix $\Phi_A(K, k_0)$ defined as $\Phi_A(K, k_0) = A_{K-1} \cdots A_{k_0}$ if $K > k_0$ and $\Phi_A(K, k_0) = I$ if $K = k_0$.

Note that if A_K is asymptotically stable, that is, $\rho(A_K) < 1$ for all $K > k_0$, then the monodromy matrix is also asymptotically stable, since $\rho(\Phi_A(K, k_0)) = \|\Phi_A(K, k_0)\|_2 \leq \prod_{k=k_0}^{K-1} \rho(A_k) < 1$ (this is followed from the symmetry of the matrix S_K for $K > k_0$). Hence, we can ensure that the recurrence sequence of the parameter vector $\{\mathbf{p}_K\}_{K \geq 1}$ obtained as solution of (4) is bounded if B_K is also bounded.

Finally, we establish a condition on the new data $K+1$ in order that the consecutive approximations of the parameter are sufficiently close.

Proposition 4 *Let the system be* (4). *Suppose that there exists* k_0 *such that* $\operatorname{rank}(H(k_0))$ *is full rank, and* $\rho(A_K) < 1$ *for all* $K > k_0$. *Consider the observation data such that* $\|ob(K+1) - x(K+1)\|_2 < \frac{\epsilon}{\rho(S_{K+1}^{-1})\rho(M^T(K)M(K))}$, *for some* $\epsilon > 0$. *Then*

$$\|\mathbf{p}_{K+1} - \mathbf{p}_K\|_2 < \epsilon.$$

Proof Given K observation data and $A_K = S_{K+1}^{-1}S_K$, $B_K = S_{K+1}^{-1}M^T(K)d(K+1)$ we have

$$\begin{aligned}
\|\mathbf{p}_{K+1} - \mathbf{p}_K\|_2 &= \|A_K\mathbf{p}_K + B_K - \mathbf{p}_K\|_2 \\
&\leq \left\|S_{K+1}^{-1}\right\|_2 \left\|(S_K - S_{K+1})\mathbf{p}_K + M^T(K)d(K+1)\right\|_2 \\
&\leq \rho\left(S_{K+1}^{-1}\right)\left\|M^T(K)\right\|_2 \left\|d(K+1) - M(K)\mathbf{p}_K\right\|_2 \\
&= \rho\left(S_{K+1}^{-1}\right)\rho\left(M^T(K)M(K)\right)\left\|ob(K+1) - x(K+1)\right\|_2 < \epsilon. \qquad \square
\end{aligned}$$

Remark 1 The parameters involved in an epidemiological process are not always known. To obtain a value of these sufficiently reliable, it is necessary to know if it can be identified from a set of observations of the process, and then estimate its value. In the literature, there exist several approaches to the identifiability problem and to the estimation problem. From a set of observations, for instance in engineering, it is usual to consider the transfer matrix, in chemicals, if we consider the input-output response and the parameters of Markov [4, 9], or directly from the solution of the system. Generally the estimation approach is based in the gradient algorithm and the least squares algorithm, [7, 8]. In our case, we identify the case using the structure of the matrices and asking whether the performance of the estimation process is constructive, using a least squares algorithm. Then the algorithm proposed can be used to identify and estimate the parameters of other time-discrete age-structured models taking into account only the structure one has when performing the steps of our algorithm.

4 Numerical example

Consider an age-structured population in three compartments which may suffer a contagious disease. We consider that the survival rate of susceptible, infected, and recovered individuals are independent from the age. Concretely, we have $p = 0.99$, $q = 0.95$, $r = 0.98$. Let us consider that the rates of individuals changing compartment due to increasing age are known. Specifically, we consider $\sigma_i = \mu_i = 0.01$ and $\epsilon = \nu_i = \delta_i = 0.01$, for all i.

In this process we want to obtain an estimation of the exposition and the recovered rates which we suppose different according to the age of the individual.

We make an experiment on a sample of size $N = 90{,}000$ from the initial condition $ob(0) = (67{,}000, 1{,}850, 0, 14{,}700, 1{,}100, 0, 4{,}300, 1{,}050, 0)$ and an observed dataset $\{ob(k), k = 1, \ldots, 15\}$, given in Table 2.

We see that the matrix H_1 is full rank and the coefficient matrix A_k of the discrete-time linear system given by (4) is stable, $\rho(A_k) < 1$, for all $k = 1, \ldots, 15$. Applying this recurrence equation we obtain the approximations of the parameter vector $\mathbf{p}$ given in Table 3.

Note that these approximations satisfy $\|\mathbf{p}_k\|_2 \approx 0.08$ and $\|\mathbf{p}_{k+1} - \mathbf{p}_k\|_2 < 10^{-2}$, for all $k = 1, \ldots, 15$. In fact, at time k, the condition established in Proposition 4 holds when we compare the observed data $ob(k)$ and the output $x(k)$ of the system (3) obtained from $\mathbf{p}_k$:

$$\left\|ob(k) - x(k)\right\|_2 < \frac{10^{-2}}{\rho(S_k^{-1})\rho(M^T(k-1)M(k-1))}, \quad k = 2, \ldots, 15.$$

Table 2 Observed dataset of SIR individuals in a population.

k	Observed SIR data at time k, $ob(k)$: $(S_1(k), I_1(k), R_1(k), S_2(k), I_2(k), R_2(k), S_3(k), I_3(k), R_3(k))$
$k = 1$	(66,701, 1,680, 74, 15,075, 986, 67, 4,400, 964, 53)
$k = 2$	(66,398, 1,522, 140, 15,442, 886, 126, 4,505, 881, 100)
$k = 3$	(66,090, 1,385, 198, 15,798, 796, 176, 4,615, 797, 145)
$k = 4$	(65,778, 1,256, 248, 16,141, 717, 220, 4,723, 738, 179)
$k = 5$	(65,466, 1,144, 295, 16,476, 642, 260, 4,835, 670, 212)
$k = 6$	(65,152, 1,040, 336, 16,791, 584, 294, 4,947, 615, 241)
$k = 7$	(64,836, 960, 370, 17,105, 516, 326, 5,066, 544, 277)
$k = 8$	(64,521, 870, 400, 17,408, 474, 347, 5,185, 505, 290)
$k = 9$	(64,210, 772, 429, 17,705, 447, 368, 5,310, 450, 309)
$k = 10$	(63,892, 721, 449, 17,981, 395, 386, 5,439, 410, 327)
$k = 11$	(63,580, 665, 468, 18,264, 353, 401, 5,543, 384, 342)
$k = 12$	(63,269, 620, 484, 18,529, 292, 415, 5,679, 358, 354)
$k = 13$	(62,960, 545, 498, 18,800, 268, 425, 5,808, 331, 365)
$k = 14$	(62,645, 510, 509, 19,043, 260, 430, 5,930, 301, 372)
$k = 15$	(62,351, 459, 519, 19,298, 218, 440, 6,067, 266, 382)

Table 3 Estimated values of the parameter p.

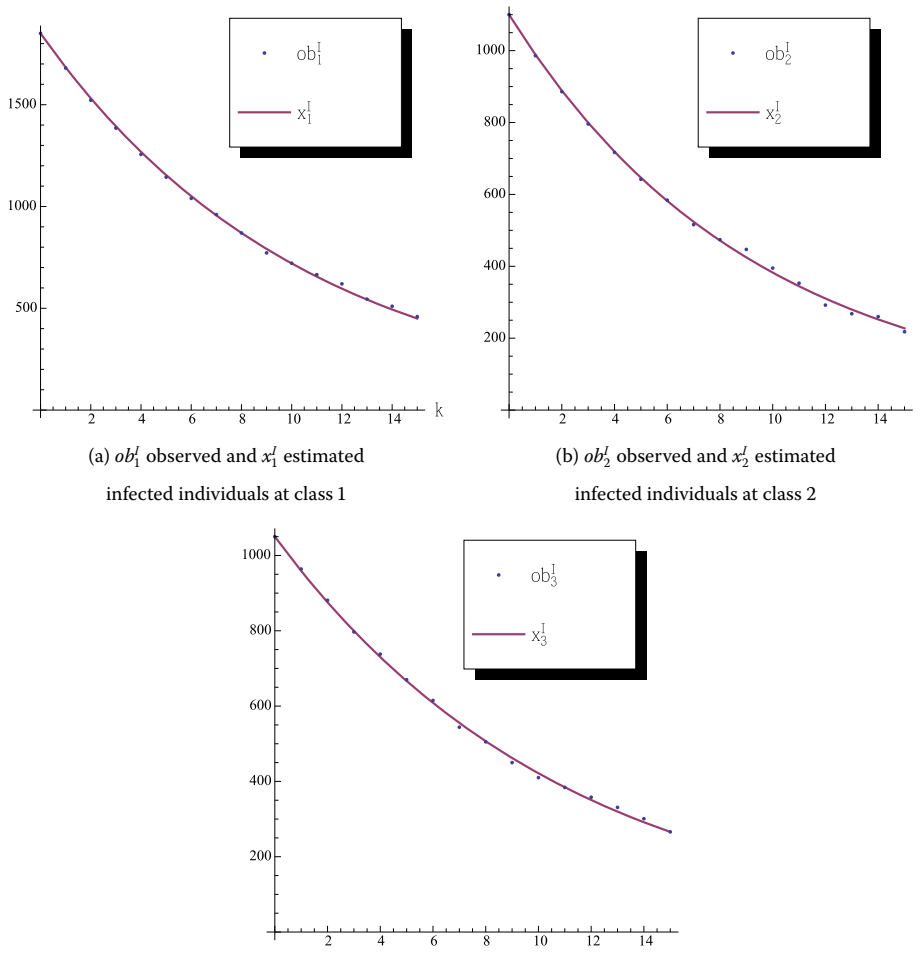
Estimated value $\mathbf{p}_i$ of p from dataset $\{ob(k)\}_{k=1}^{i}$
$\mathbf{p}_1 = (0.00956693, 0.000751478, 0.00521651, 0.0407453, 0.0610622, 0.0490052)$
$\mathbf{p}_2 = (0.00907106, 0.000369321, 0.00390195, 0.0410328, 0.0604876, 0.049093)$
$\mathbf{p}_3 = (0.00933411, 0.000142124, 0.00248374, 0.0407704, 0.0600328, 0.0503554)$
$\mathbf{p}_4 = (0.00927112, 0.000579406, 0.00335454, 0.0407454, 0.0599345, 0.0491128)$
$\mathbf{p}_5 = (0.00938896, 0.000491274, 0.00304479, 0.040709, 0.060185, 0.0493809)$
$\mathbf{p}_6 = (0.00941313, 0.00179466, 0.00343295, 0.0407295, 0.0603492, 0.0493287)$
$\mathbf{p}_7 = (0.00981096, 0.0015837, 0.00287703, 0.0403433, 0.0609501, 0.0508397)$
$\mathbf{p}_8 = (0.00970246, 0.00194832, 0.00285248, 0.0403566, 0.0604301, 0.0498377)$
$\mathbf{p}_9 = (0.00927398, 0.00247651, 0.00222736, 0.0407216, 0.0597542, 0.0499973)$
$\mathbf{p}_{10} = (0.00967561, 0.00282768, 0.00192291, 0.0404139, 0.0601675, 0.0499162)$
$\mathbf{p}_{11} = (0.0100975, 0.00258414, 0.00326115, 0.0415127, 0.0617572, 0.0517111)$
$\mathbf{p}_{12} = (0.0105733, 0.00229971, 0.00304286, 0.0423655, 0.0637888, 0.0525927)$
$\mathbf{p}_{13} = (0.0105774, 0.00208498, 0.00314224, 0.0435171, 0.0644926, 0.0535962)$
$\mathbf{p}_{14} = (0.0110997, 0.00268321, 0.00341103, 0.0441655, 0.0650516, 0.0546224)$
$\mathbf{p}_{15} = (0.0110427, 0.00218739, 0.00309953, 0.0447639, 0.0659736, 0.0555352)$

Then a good estimation of the vector can be $\mathbf{p} = (0.01, 0.002, 0.003, 0.04, 0.06, 0.05)$, such as we can see in the Figure 2, where the infected individuals of each age class are compared with the signal obtained when the above value of $\mathbf{p}$ is considered.

Note that from the definition of $\mathbf{p} = (f_1\, f_2\, f_3\ \gamma_1\ \gamma_2\ \gamma_3)^T$ and f_1, f_2, f_3 we see that the estimated value to the exposition rate are $\alpha_1 = 0.02$, $\alpha_2 = 0.008$, $\alpha_3 = 0.012$, and the recovery rates are $\gamma_1 = 0.04$, $\gamma_1 = 0.06$, $\gamma_1 = 0.05$.

5 Conclusions

An infectious disease acting on a population has been considered. This population is structured at age compartments with susceptible, infected, and recovered individuals. The epidemiological process is modeled by a dynamic system with unknown parameters. First, we have established a condition to ensure the nonnegativity of the solution. Next, the identifiability problem has been analyzed and an algorithm to identify the parameters has been constructed. The following issue considered has been how to estimate the parameters in the model. Using a least square method we have showed the descendent gradient method and a condition to ensure that the estimated parameter has norm less than 1. Moreover, we have constructed a recurrence equation when more observed data are considered and



(a) ob_1^I observed and x_1^I estimated infected individuals at class 1

(b) ob_2^I observed and x_2^I estimated infected individuals at class 2

(c) ob_3^I observed and x_3^I estimated infected individuals at class 3

Figure 2 Evolution of infected individuals at each age compartment.

we have established a condition to ensure that the consecutive approximations of the parameter are sufficiently close. Finally, an illustrative example has been showed.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The work is a product of the intellectual environment of the whole team and each author has participated equally in the work.

Acknowledgements
The authors would like to thank the referees and the editor for their comments and useful suggestions for improvement of the manuscript. This work has been partially supported by Spanish Grant MTM2013-43678-P.

Received: 21 December 2015 Accepted: 6 January 2017 Published online: 27 January 2017

References

1. Strogatz, S, Friedman, M, Mallinck-Rodt, AJ, McKay, S: Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. Perseus Books, Washington (1994)
2. De La Sen, M, Quesada, A: Some equilibrium, stability, instability and oscillatory results for an extended discrete epidemic model with evolution memory. Adv. Differ. Equ. **2013**, 234 (2013)

3. Han, Q, Wang, Z: On extinction of infectious diseases for multi-group SIRS models with satured incidence rate. Adv. Differ. Equ. **2015**, 333 (2015)
4. Cantó, B, Coll, C, Sánchez, E: Structural identifiability of a model of dialysis. Math. Comput. Model. **50**, 733-737 (2009)
5. Cantó, B, Coll, C, Sánchez, E: Identifiability of a class of discretized linear partial differential algebraic equations. Math. Probl. Eng., 1-12 (2011)
6. Craciun, G, Pantea, C: Identifiability of chemical reaction networks. J. Math. Chem. **44**, 244-259 (2008)
7. Malik, MB, Salman, M: State-space least mean square. Digit. Signal Process. **18**, 334-345 (2008)
8. Ding, F, Liu, PX, Liu, G: Multiinnovatiovation least-squares identification for system modeling. IEEE Trans. Syst. Man Cybern., Part B, Cybern. **18**(3), 767-778 (2010)
9. Ben-Zvi, A, McLellan, PJ, McAuley, KB: Identifiability of linear time-invariant differential-algebraic systems, I. The generalized Markov parameter approach. Ind. Eng. Chem. Res. **42**, 6607-6618 (2003)
10. Boyadjiev, C, Dimitrova, E: An iterative method for model parameter identification. Comput. Chem. Eng. **29**, 941-948 (2005)
11. Ben-Zvi, A, McLellan, PJ, McAuley, KB: Identifiability of linear time-invariant differential-algebraic systems, 2. The differential-algebraic approach. Ind. Eng. Chem. Res. **43**, 1251-1259 (2004)
12. Dion, JM, Commault, C, van der Woude, J: Generic properties and control of linear structured systems: a survey. Automatica **39**, 1125-1144 (2003)
13. Chou, IC, Voit, EO: Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Math. Biosci. **219**, 57-83 (2009)
14. Schmitz, OJ: Ecology and Ecosystems Conservation. Island Press, Washington (2013)