Research article

# The multilevel hierarchical data EM-algorithm. Applications to discrete-time Markov chain epidemic models

Divine Wanduku

*Department of Mathematical Sciences, Georgia Southern University, 65 Georgia Ave, Room 3309, Statesboro, GA, 30460, USA*

ARTICLE INFO

ABSTRACT

The theory of multilevel hierarchical data Expectation Maximization (EM)-algorithm is introduced via discrete time Markov chain (DTMC) epidemic models. A general model for a multilevel hierarchical discrete data is derived. The observed sample $Y$ in the system is a stochastic incomplete data, and the missing data $Z$ exhibits a multilevel hierarchical data structure. The EM-algorithm to find ML-estimates for parameters in the stochastic system is derived. Applications of the EM-algorithm are exhibited in the two DTMC models, to find ML-estimates of the system parameters. Numerical results are given for influenza epidemics in the state of Georgia (GA), USA.

## 1. Introduction

Deterministic mathematical models have a long history of applications in disease epidemiology, e.g. the Kermack and McKendrick SIR model [36], and more examples are given in Hethcote [32]. Since random events occur in epidemic dynamics, this implies that deterministic models are first approximations for reality. Indeed, they serve to estimate the *basic reproduction number* of diseases (cf. [21]); and to investigate factors affecting disease eradication and persistence in the population (cf. [9]). Also, deterministic models are more suitable for disease dynamics with large number of susceptible and infected states, where random epidemiological fluctuations can be ignored. However, in most epidemic dynamics, significant random variations occur, which are better represented by stochastic models. Compartmental epidemic models are used to investigate infectious disease dynamics. For instance, influenza, malaria and other types of compartmental disease models are studied in [24, 38, 40, 42, 43, 48, 49, 51, 55, 56, 57, 61]. In general, these compartmental models are classified as SIRS, SIR, SIS, SEIR and, SEIRS etc. models depending on the disease states involved in the disease dynamics [24, 48, 53, 54, 55, 61].

Probability modeling also has a long history, and a classic example is Bernoulli [12]. Also, stochastic epidemic models are investigated (cf. [10, 34]). Modeling with some counting processes such as continuous-time Markov chains (CTMC), where the state of the process is an integer valued random variable representing, for example, the disease states: susceptible, exposed, infectious or removed individuals, have wide applications in the literature [8, 11, 35]. Discrete-time Markov chain (DTMC) epidemic models on the other hand, have also received attention [6, 8, 29, 53, 61].

In the wide variety of DTMC epidemic models studied, there is (1) a class of DTMC epidemic models that are based on *birth-death processes*, where only one transition between disease states occurs over a small time step (cf. [8, 27, 35, 39, 47, 61]. These DTMC models approximate their corresponding CTMC epidemic models which have generator rate matrices that correspond with the rate functions of deterministic ordinary differential equation models that assume exponential holding lifetimes (cf. [8, 27, 35, 47, 61]. This class of DTMC epidemic models is the main case study for the hierarchical data EM-algorithm investigated in the paper, and other examples are [39, 61]. Another class of DTMC epidemic models is (2) DTMC epidemic models based on branching processes, such as, epidemic models based on Galton-Watson branching processes (cf. [8, 22]). The third class of DTMC epidemic models is (3) the class of *chain-binomial epidemic models*. Some classical examples of these models are the Greenwood [29] and Reed-Frost [6] models. Also see [28, 31, 52, 53, 58].

Employing mathematical models to obtain accurate predictions of epidemics, and deriving rational data-informed health policies necessitate the estimation of parameters of the models. For instance, the basic reproduction number (cf. [8]), $\Re_0$, is a complex parameter that depends on other simpler model parameters, which can be estimated. Significant progress has been made to derive and employ statistical and data-science techniques

to estimate and infer parameters of epidemic models, from epidemic data. A cross-section of these techniques is given in the following [7, 15, 18, 19, 26, 27, 35, 46, 47, 50, 60, 61, 62, 63]. Indeed, the least squares (LS) estimation method is applied to find parameter estimates for the epidemic models in [15, 16, 19]; maximum likelihood (ML) estimators are obtained in [26, 35, 47, 61, 63]; martingale estimators are obtained in [10, 26]; nonparametric estimation method is applied in [44]; Bayesian Markov chain Monte Carlo (MCMC) estimation technique is applied in [39, 45]; the expectation maximization (EM) algorithm technique is discussed in [10, 23]; the hierarchical linear modeling (HLM) method is applied in [50, 60].

Parameter estimation and inferences in epidemic models becomes more complex when only partial observation of the epidemic process is possible, and several important epidemiological variables are missing in the data collected. In such circumstances, the traditional estimation techniques such as the HLM, LS, ML and Bayesian estimation methods, based on complete data are limited; more sophisticated estimation techniques such as the martingale estimation (cf. [10]), Bayesian MCMC estimation (cf. [39]), and the EM algorithm (cf. [10, 23, 33]) are more useful.

The primary interest of this study is introducing the technique of the EM-algorithm in hierarchical data (HD) structures, to find the maximum likelihood (ML) estimators of parameters of epidemic models. The hierarchical data EM-algorithm derived in this study is suitably applied to all sorts of hierarchical data structures arising in both statistical, stochastic and deterministic systems. However, the examples to characterize the hierarchical data structure in this study will be drawn from DTMC SEIR epidemic models; the interest is finding the maximum likelihood (ML) estimators of the parameters of the SEIR epidemic models. Some examples of SEIR disease epidemics are Ebola and influenza (cf. [39]).

For DTMC epidemic models, the method of ML estimation (cf. [17]) is explored in [26, 61, 63], and for CTMC epidemic models in [35, 47]. The method of ML estimation employed in infectious disease dynamic systems, seeks to find estimates for a set of model parameters, from epidemic data, that maximize the chance of observing the data from the population distribution indexed by the parameters (cf. [17]). This estimation technique becomes challenging to apply, whenever minimizing the likelihood function leads to intractable results. The expectation-maximization algorithm (EM-algorithm) [13, 20, 30] is applied in circumstances, where incorporating missing or hidden information to the given data leads to a more tractable likelihood function. Some of the primary subjects of this paper are as follows. To derive a model for the hierarchical data (HD) structure; to derive the EM-algorithm for the HD structure; to derive a DTMC model for a SEIR disease epidemic, for instance, Ebola or influenza, as an example of a HD system; to explore the maximum likelihood estimation and EM-algorithm techniques to find ML-estimates for the parameters of the SEIR epidemic model.

The EM-algorithm introduced by Dempster et al. [20] is a well-established ML estimation tool in the literature, and it has been exploited extensively in many statistical studies (cf. [10]). Moreover, several advances have been made to improve the EM algorithm e.g. the MCEM algorithm [59]; the PX-EM algorithm [41]; the multiset EM algorithm [33] etc. Application of the EM algorithm in epidemic models to approximate model parameters is in the infancy state. Duncan and Gyongy [23] applied the EM-algorithm to obtain MLE's for parameters in a partial differential equation SEIR epidemic model with missing data. In [10], an illustration of the EM-algorithm is given for the classic Reed-Frost DTMC model [6], where hidden information is used to simplify the log-likelihood function.

Infectious disease transmission between susceptible and infectious individuals of a population is a very complex epidemiological process composed of several intricate microscopic hidden sub-processes. For instance, for a susceptible person to get infected, (1) there must be contacts between individuals of all states in the population; (2) some of the contacts are with infectious individuals, and likely to lead to infection; (3) given a set of possible infectious contacts, at least one leads to infection. In other words, the hidden microscopic sub-disease processes (1)-(3) define a hierarchical structure of nested distinct conditional contacting levels, where the first level involves all possible contacts between individuals of all states in the population; the second level that involves infectious contacts, conditioned on the first level of given a set of contacts with individuals of all states; the third level of contacting involves specific infectious contacts that successfully transmit the disease, conditioned on the second level of a given set of infectious contacts.

In a typical epidemic process, the hierarchical contacting data for the levels described in (1)-(3) are hidden. Furthermore, as remarked in [39], for SEIR models, the exposed state ($E$) in the population is unobserved, and hence, data on $E$ is missing. Thus, conducting ML estimation on DTMC SEIR models, where the disease transmission process satisfies (1)-(3), becomes computationally expensive because of an intractable likelihood function. And, the HD EM-algorithm is needed to find ML-estimates for the DTMC SEIR models.

Indeed, hierarchical data structures occur very often in medical research applications. For instance, in [50, 60], there are four HD levels consisting of measurements obtained from a "patient" category nested in a "hospital" category, which is further nested in a "city" category that is also nested in a "county" category. Given the high level of inter-dependency in a hierarchical data structure, the traditional statistical estimation methods are no longer feasible. For example, in [50, 60], a more complex extension of the ordinary least squares regression method, called *Hierarchical Linear Modeling (HLM)* technique, is proposed for deriving empirical linear models. Similarly, building upon the ideas of the studies [6, 25, 29, 39, 61], we shall propose a SEIR DTMC epidemic model as a case study for a new HD EM-algorithm derived in this study.

This paper is organized as follows. In Section 2, we derive the DTMC SEIR model used as case study for the HD EM-algorithm. In Section 3, we derive a mathematical model for the HD-structure, and give an example in the SEIR epidemic model. Section 4, we derive the HD EM-algorithm for an $M$-level HD structure. In Section 5, we apply the $M$-level HD EM-algorithm to a special example for a 4-level HD structure in the SEIR model, and find the $m^{th}$ step ML-estimate, in the EM-algorithm, for the parameter $\Theta^2$ for the SEIR model. In Section 6, we present numerical simulation results to test the algorithm.

## 2. Description and derivation of the general SEIR Markov chain

### 2.1. Derivation of the stochastic SEIR model

The SEIR model is based on the following assumptions: (A) We use a regular partition $t_0, t_1, t_2, \ldots, t_k = t_0 + (\Delta t)k, \forall k = 0, 1, 2, 3, \ldots$ to create discrete time intervals of length $\Delta t$ (*i.e.* $[t_k, t_{k+1}), \forall k \geq 0$), and count the number of people of each compartment in a SEIR epidemic namely: susceptible, exposed, infectious and recovered (acquired immune or death) in each time interval. That is, $x(t_k)$ is the number of people in state $x \in \{S, E, I, R\}$ present at the beginning of the epoch $k$ (i.e. in $[t_k, t_{k+1})$), or equivalently, at the end of the epoch $k-1$ (i.e. in $[t_{k-1}, t_k)$). Thus, $S(t_k), E(t_k), I(t_k), R(t_k) \in \mathbb{Z}_+$, and $S(t_0) > 0, I(t_0) > 0, \forall k \in \mathbb{Z}_+$.

(B) It is assumed the total human population of size $N(t_k) = S(t_k) + E(t_k) + I(t_k) + R(t_k) = N > 0$, ($N \in \mathbb{Z}_+$ is a constant) at time $t_k$ lives in a natural closed environment, where the outbreak of disease occurs. The susceptible people $S$ are vulnerable to infection and not yet infected. The exposed people $E$ are infected, but not infectious. The exposed incubate the disease for $T_1$ time units. The infected individuals spreading the disease

are denoted by $I$, and they spread the disease for $T_2$ time units, before becoming removed $R$ via recovery or death. In general, it is assumed that $T_1 \leq T_2$. Note that the time $t_k$ is synonymously used as time $k \geq 0$.

(C) Assume $C_{ij}(t_k)$, $i, j \in \{S, E, I, R\}$ is the number of epidemiological transition events that occur at time $k$, from state $i$ to state $j$, where $i, j \in \{S, E, I, R\}$. For example, $C_{SE}(t_k)$ is the number of newly infected people becoming exposed during the time $[t_k, t_{k+1})$, where the new infections are counted from $t_k$ until the onset of $t_{k+1}$. Similarly, $C_{EI}(t_k)$ and $C_{IR}(t_k)$ are the number of people converting from the exposed and infectious states, to the infectious and removed states, respectively, during time interval $[t_k, t_{k+1})$. It follows that $C_{ij}(t_k) \in \mathbb{Z}_+, \forall i, j \in \{S, E, I, R\}$. Also, it is easy to see that $0 \leq C_{SE}(t_k) \leq S(t_k)$, $0 \leq C_{EI}(t_k) \leq E(t_k)$, and $0 \leq C_{IR}(t_k) \leq I(t_k), \forall k \geq 0$. Observe that the $C_{ij}(t_k)$'s are mutually independent at any time $k \geq 0$.

From (A)-(C) above, the SEIR epidemic model is guided by the system (2.1)-(2.4). That is, given the states at time $k$, $(S(t_k), E(t_k), I(t_k), R(t_k))$, then the states at time $k + 1$ are given by

$$S(t_{k+1}) = S(t_k) - C_{SE}(t_k), \tag{2.1}$$

$$E(t_{k+1}) = E(t_k) + C_{SE}(t_k) - C_{EI}(t_k), \tag{2.2}$$

$$I(t_{k+1}) = I(t_k) + C_{EI}(t_k) - C_{IR}(t_k), \tag{2.3}$$

$$R(t_{k+1}) = R(t_k) + C_{IR}(t_k), \tag{2.4}$$

where similarly to [31, 39], it can be easily seen that

$$C_{SE}(t_k) \sim Binomial(S(t_k), P_{SE}(t_k)), \quad C_{EI}(t_k) \sim Binomial(E(t_k), P_{EI}(t_k)),$$

$$C_{IR}(t_k) \sim Binomial(I(t_k), P_{IR}(t_k)), \tag{2.5}$$

are random variables with binomial probability distributions; and in (2.5) denote by, $Binomial(n, \rho)$, where $n$ is the fixed number of Bernoulli trials, and $\rho$ is the success probability. That is, $P_{ij}(t_k)$, is the probability of transiting from state $i$ to state $j$, where $i \in \{S, E, I\}$, $j \in \{E, I, R\}$. To comprehend the intrinsic nature of hierarchical data (HD) structures in epidemic models such as (2.1)-(2.4), and also to elucidate all hidden or missing components of (2.1)-(2.4), detailed derivations of the probabilities $P_{ij}(t_k)$ are given in subsequent subsections.

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a complete probability space, where $\mathfrak{F}_{t_k}$ is a filtration (that is, sub $\sigma$- algebra $\mathfrak{F}_{t_k}$ that satisfies the following: given $t_1 \leq t_2 \Rightarrow \mathfrak{F}_{t_1} \subset \mathfrak{F}_{t_2}$; $E \in \mathfrak{F}_{t_k}, \exists k$, and $P(E) = 0 \Rightarrow E \in \mathfrak{F}_{t_0}$). Define a random vector measurable function

$$X : \mathbb{Z}_+ \times \Omega \to \mathbb{Z}_+^4, \quad X(t_k) = (S(t_k), E(t_k), I(t_k), R(t_k)) \in \mathbb{Z}_+^4, \forall k \in \mathbb{Z}_+. \tag{2.6}$$

Moreover, the states $S(t_k), E(t_k), I(t_k)$, and $R(t_k)$ at any time $k \geq 0$, are $\mathfrak{F}_{t_k}$-measurable and satisfy the equations (2.1)-(2.4). In addition, the random variables $C_{ij}(t_k) \in \mathbb{Z}_+, \forall i, j \in \{S, E, I, R\}$, are all discrete random variables on the probability space. Thus, $\mathfrak{F}_{t_k}$ is generated by the components of $X(t_0), X(t_1), \ldots, X(t_{k-1}), X(t_k)$. The collection

$$\{X(t_k) : k \geq 0\} = \{(S(t_k), E(t_k), I(t_k), R(t_k)) : k \geq 0\}, \tag{2.7}$$

defines a random process for the SEIR epidemic. In addition, if we introduce the notations[1]

$$C_{12}(t_k) \equiv C_{SE}(t_k), C_{23}(t_k) \equiv C_{EI}(t_k), C_{34}(t_k) \equiv C_{IR}(t_k), S(t_k) \equiv X_1(t_k),$$

$$S(t_{k+1}) \equiv X_1(t_{k+1}), E(t_k) \equiv X_2(t_k), E(t_{k+1}) \equiv X_2(t_{k+1}), I(t_k) \equiv X_3(t_k),$$

$$I(t_{k+1}) \equiv X_3(t_{k+1}), R(t_k) \equiv X_4(t_k), R(t_{k+1}) \equiv X_4(t_{k+1}),$$

then the transition events satisfy

$$C_{i,i+1}(t_k) = \sum_{j=1}^{i} (X_j(t_k) - X_j(t_{k+1})),$$

and $C_{34}(t_k) = X_4(t_{k+1}) - X_4(t_k)$, where, $i \in 1, 2, 3$.

Thus, the feasible region, denoted by, $\Omega^1_{X(t_{k+1})}$ for the state $X(t_{k+1})$ of the process at time $k + 1$, given the state of the process $X(t_k)$ at time $k$, is defined as follows:

$$\Omega^1_{X(t_{k+1})} = \{(s_{k+1}, e_{k+1}, i_{k+1}, r_{k+1}) \in \mathbb{Z}_+^4 | 0 \leq S(t_k) - s_{k+1} \leq S(t_k),$$

$$0 \leq E(t_k) - e_{k+1} + S(t_k) - s_{k+1} \leq E(t_k),$$

$$0 \leq I(t_k) - i_{k+1} + E(t_k) - e_{k+1} + S(t_k) - s_{k+1} \leq I(t_k),$$

$$\text{and} \quad 0 \leq r_{k+1} - R(t_k) \leq I(t_k)\}. \tag{2.8}$$

Clearly, the random process (2.7) is a Markov chain. Indeed, we introduce new notations in addition to (2.6) - (2.7).

**Notation 2.1.**

(i.) Let $x_k \in \mathbb{Z}_+^4$, where $x_k = (x_1^k, x_2^k, x_3^k, x_4^k) \equiv (s_k, e_k, i_k, r_k) \in \mathbb{Z}_+^4$. That is, $x_1^k \equiv s_k, x_2^k \equiv e_k, x_3^k \equiv i_k$ and $x_4^k \equiv r_k$. The vector $x_k = (s_k, e_k, i_k, r_k) \in \mathbb{Z}_+^4$ consists of non-negative integers for each $k \in \{0, 1, 2, 3, \ldots\}$, and $X(t_k) = x_k$ if and only if

---

[1] The notation "$\equiv$" signifies the operation of "mathematical equivalence" between two variables.

$$S(t_k) \equiv X_1(t_k) = x_1^k, S(t_{k+1}) \equiv X_1(t_{k+1}) = x_1^{k+1}, E(t_k) \equiv X_2(t_k) = x_2^k,$$

$$E(t_{k+1}) \equiv X_2(t_{k+1}) = x_2^{k+1}, I(t_k) \equiv X_3(t_k) = x_3^k, I(t_{k+1}) \equiv X_3(t_{k+1}) = x_3^{k+1},$$

$$R(t_k) \equiv X_4(t_k) = x_4^k, R(t_{k+1}) \equiv X_4(t_{k+1}) = x_4^{k+1},$$

where, $x_1^k \equiv s_k, x_2^k \equiv e_k, x_3^k \equiv i_k$ and $x_4^k \equiv r_k$.

(ii.) The notation $G(t_k)|H(t_k)$ denotes a conditional random variable $G(t_k)$, given the random variable $H(t_k)$ at each time $k \geq 0$. That is, for each $k \geq 0$, given a value for $H(t_k)$, then $G(t_k)$ is determined. Moreover, the collection $\{G(t_k)|H(t_k), k \geq 0\}$ is called a sub-stochastic process of the process $\{H(t_k), k \geq 0\}$.

Applying the Notation 2.1, the random process $\{X(t_k) : k \geq 0\}$ in (2.7) satisfies

$$P(X(t_{k+1}) = x^{k+1}|\mathfrak{F}_{t_k}) = P(X(t_{k+1}) = x^{k+1}|X(t_k) = x^k) = \prod_{i=1}^{3} P\left( C_{i,i+1}(t_k) = \sum_{j=1}^{i}(x_j^k - x_j^{k+1})|X(t_k) = x^k \right), \tag{2.9}$$

where $k \in \{0, 1, 2, \ldots\}$.

From (B) above, note that $N$ is constant, and $R(t_k) = N - S(t) - E(t_k) - I(t_k)$. Hence, the vector $(S(t_k), E(t_k), I(t_k))$ is sufficient for the model (2.1)-(2.4). Thus, we use the vectors $X(t_k) = (S(t_k), E(t_k), I(t_k), R(t_k))$ and $X(t_k) = (S(t_k), E(t_k), I(t_k))$ synonymously. Also, note that the distributions of the sub-stochastic processes $\{C_{i,i+1}(t_k)|X(t_k), k \geq 0\}, \forall i \in \{1, 2, 3\}$ are given in (2.5) without proof. However, to clearly understand the hierarchical data (HD) structure in the SEIR model (2.1)-(2.4), and also to understand the observed, and missing/hidden constituents of the data structure required for the EM algorithm, we explicitly derive the success probabilities $P_{ij}(t_k)$ of (2.5), where $i \in \{S, E, I\}$, $j \in \{E, I, R\}$; we completely characterize the distributions of $\{C_{i,i+1}(t_k)|X(t_k), k \geq 0\}, \forall i \in \{1, 2, 3\}$.

**Remark 2.1.** The EM-algorithm presented in this paper for the hierarchical data structure in the model (2.1)-(2.4), is for the case where the incubation and infectious periods, $T_1$ and $T_2$, respectively, are random with exponential lifetime distributions.

## 3. $M$-level hierarchical data structures

In this section, we characterize an $M$-level hierarchical data (HD) structure, where $M \geq 2$ is an integer. Some ideas from [14, 50] are used to construct the HD structure. Consider a data structure comprising of $M$-levels of nested hierarchical data, where at the $M^{th}$ level, there are $n_M \geq 1$ measurements, denoted by $Z_{k_M}$, obtained from $n_M$ distinct units present at this level, where $k_M = 1, 2, 3, \ldots n_M$. Let

$$C^{(M)} = \left\{ Z_{k_M} | k_M = 1, 2, 3, \ldots n_M \right\}, \tag{3.1}$$

denote the collection of the $n_M \geq 1$ measurements in the $M^{th}$ level.

Similarly, for each unit measurement $Z_{k_M}$, $k_M \in \{1, 2, 3, \ldots n_M\}$, there are $n_{(M-1)} \geq 1$ nested sub-measurements obtained in the $(M-1)^{th}$ level, denoted by $Z_{k_M, k_{(M-1)}}$, where $k_{(M-1)} = 1, 2, 3, \ldots n_{(M-1)}$. Let

$$C^{(M-1)} = \bigcup_{k_M=1}^{n_M} \left\{ Z_{k_M, k_{(M-1)}} | k_{(M-1)} = 1, 2, 3, \ldots n_{(M-1)} \right\},$$

be the collection of the $n_M \times n_{(M-1)} \geq 1$ measurements obtained from units in the $(M-1)^{th}$ level.

In a similar manner, for each unit measurement $Z_{k_M, k_{(M-1)}}$, where $k_{(M-1)} \in \{1, 2, 3, \ldots n_{(M-1)}\}$, there are $n_{(M-2)} \geq 1$ nested sub-measurements obtained in the $(M-2)^{th}$ level, denoted by $Z_{k_M, k_{(M-1)}, k_{(M-2)}}$, where $k_{(M-2)} = 1, 2, 3, \ldots n_{(M-2)}$. Let

$$C^{(M-2)} = \bigcup_{k_M=1}^{n_M} \bigcup_{k_{(M-1)}=1}^{n_{(M-1)}} \left\{ Z_{k_M, k_{(M-1)}, k_{(M-2)}} | k_{(M-2)} = 1, 2, 3, \ldots n_{(M-2)} \right\},$$

be the collection of the $n_M \times n_{(M-1)} \times n_{(M-2)} \geq 1$ measurements obtained in the $(M-2)^{th}$ level.

Inductively, it is easy to see that at the $(M-J)^{th}$ level, $J \in \{1, 2, 3, \ldots, M-1\}$, the set

$$C^{(M-J)} = \bigcup_{k_M=1}^{n_M} \bigcup_{k_{(M-1)}=1}^{n_{(M-1)}} \cdots \bigcup_{k_{[M-(J-1)]}=1}^{n_{[M-(J-1)]}} \left\{ Z_{k_M, k_{(M-1)}, \ldots, k_{(M-J)}, k_{(M-J)}} | k_{(M-J)} = 1, 2, 3, \ldots n_{(M-J)} \right\}, \tag{3.2}$$

is the collection of the $n_M \times n_{(M-1)} \times n_{(M-2)} \cdots \times n_{(M-J)} \geq 1$ measurements obtained in the $(M-J)^{th}$ level.

Thus, for each unit measurement $Z_{k_M, k_{(M-1)}, \ldots, k_{(M-J)}}$, where $k_{(M-J)} \in \{1, 2, 3, \ldots n_{(M-J)}\}$, there are $n_{[M-(J+1)]} \geq 1$ nested sub-measurements obtained in the $[M-(J+1)]^{th}$ level, denoted by $Z_{k_M, k_{(M-1)}, k_{(M-2)}, \ldots k_{(M-J)}, k_{[M-(J+1)]}}$, where $k_{[M-(J+1)]} = 1, 2, 3, \ldots n_{[M-(J+1)]}$. It follows that

$$C^{[M-(J+1)]} = \bigcup_{k_M=1}^{n_M} \bigcup_{k_{(M-1)}=1}^{n_{(M-1)}} \cdots \bigcup_{k_{(M-J)}=1}^{n_{(M-J)}} \left\{ Z_{k_M, k_{(M-1)}, k_{(M-2)}, \ldots k_{(M-J)}, k_{[M-(J+1)]}} | k_{[M-(J+1)]} = 1, 2, 3, \ldots n_{[M-(J+1)]} \right\},$$

is the collection of the $n_M \times n_{(M-1)} \times n_{(M-2)} \times \cdots n_{[M-(J+1)]} \geq 1$ measurements obtained in the $[M-(J+1)]^{th}$ level.

The Fig. 1 illustrates an example of a $M = 3$ hierarchical data (HD) structure. It is also easy to see that (3.2) presents a general expression to explicitly write all observations made at any level $J \in \{1, 2, 3, \ldots, M\}$. For example, in the second level of the hierarchical data structure, we obtain the following data collection
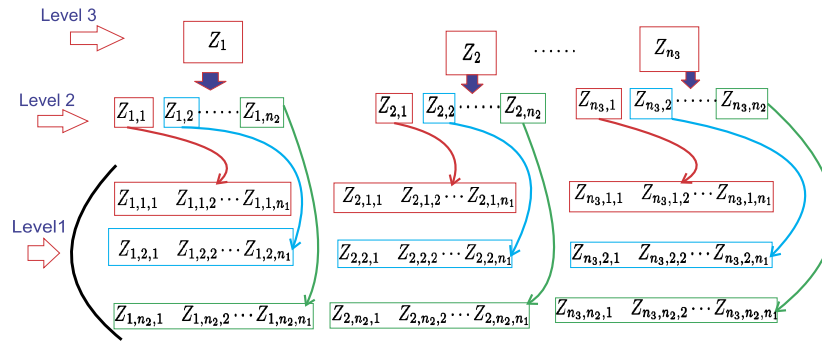
**Fig. 1.** Shows a hierarchical data structure with $M = 3$ levels of hierarchy. The arrows in this diagram point in the direction of nested lower hierarchical level data. For example, $Z_1 \to \{Z_{1,1}\} \to \{Z_{1,1,1} \quad Z_{1,1,2} \cdots Z_{1,1,n_1}\}$, signifies that $\{Z_{1,1,1} \quad Z_{1,1,2} \cdots Z_{1,1,n_1}\}$ is nested in $\{Z_{1,1}\}$, which is further nested in $\{Z_1\}$.

$$C^{(2)} = \bigcup_{k_M=1}^{n_M} \bigcup_{k_{(M-1)}=1}^{n_{(M-1)}} \cdots \bigcup_{k_3=1}^{n_3} \left\{ Z_{k_M,k_{(M-1)},k_{(M-2)},\ldots k_3,k_2} | k_2 = 1,2,3,\ldots n_2 \right\},$$

and in the first level of the HD structure, we obtain the following data collection

$$C^{(1)} = \bigcup_{k_M=1}^{n_M} \bigcup_{k_{(M-1)}=1}^{n_{(M-1)}} \cdots \bigcup_{k_3=1}^{n_3} \bigcup_{k_2=1}^{n_2} \left\{ Z_{k_M,k_{(M-1)},k_{(M-2)},\ldots k_3,k_2,k_1} | k_1 = 1,2,3,\ldots n_1 \right\}. \tag{3.3}$$

### 3.1. Example of a 4-level hierarchical data structure in the SEIR epidemic model

The distribution of $C_{SE}(t_k) \sim Binomial(S(t_k), P_{SE}(t_k))$ in (2.5) is based on the following assumptions for a mixture distribution given as definitions. These definitions will be recalled when applying the EM-algorithm to the model (2.1)- (2.4).

**Definition 3.1. Probability of passing infection from one infectious contact:**
1. At any time $t_k$, $k \geq 0$, assume that there is a constant probability $p$ of passing infection to a susceptible person after one interaction with an infectious person in the interval of time $[t_k, t_{k+1}), \forall k \geq 0$.

**Definition 3.2. The 4th level hierarchical data structure in the SEIR model:**
2. It is assumed that independent contacts occur between individuals in the population at the Poisson rate of $\lambda$ contacts per unit time. Suppose $n(t_k), k \geq 0$ is the number of contacts an individual makes with others in the population over the interval $[t_0, t_k]$, then the increment, denoted by $Z_{t_k}$, over the interval $[t_k, t_{k+1})$ has a Poisson distribution given by

$$Z_{t_k} = n(t_{k+1}) - n(t_k) \equiv n(t_{k+1} - t_k) = n(\Delta t) = n \sim Poisson(\lambda \Delta t), k \geq 0, \tag{3.4}$$

where $n \in \{0, 1, 2, \ldots\} \equiv \mathbb{Z}_+$.

It is easy to see that when the stochastic process is observed over times $t_k, k = 0, 1, 2, \ldots, T, T \in \mathbb{Z}_+$, then applying the design of the HD structure in (3.1)-(3.3), the collection of all possible contacts $Z_{t_k}$ an individual makes in the $(T + 1)$ subintervals $[t_k, t_{k+1}), k \in \{0, 1, 2, \ldots, T\}$ forms the fourth level (highest level $M = 4$) of the HD structure defined by

$$C^{(4)} = \left\{ Z_{t_k} | k = 0, 1, 2, \ldots, T \right\}. \tag{3.5}$$

**Definition 3.3. The 3rd level hierarchical data structure in the SEIR model:**
3. From (3.4), suppose in each interval $[t_k, t_{k+1})$, a susceptible individual makes $Z_{t_k} = n$ contacts with both infectious and noninfectious persons, where $n \in \{0, 1, 2, \ldots\}$. It is assumed that there is homogeneous mixing in the population, so that all contacts are equally likely to occur, regardless of the state (susceptible, exposed, infectious or removed) of an individual in the population.

Denote by $Z_{t_k,n}$, the number of infectious contacts that the susceptible person makes, given the $Z_{t_k} = n$ contacts with other individuals. It is easy to see that the distribution of $Z_{t_k,n}$ is binomial. That is,

$$Z_{t_k,n} \sim Binomial(n, \alpha_{t_k}), \tag{3.6}$$

where

$$\alpha_{t_k} = \frac{I(t_k)}{N},$$

is the probability that the susceptible individual contacts an infectious person in the population during the time interval $[t_k, t_{k+1})$. Moreover, the population has a total constant size $N$ at all time $k \geq 0$. Since the random variable $Z_{t_k,n}$ has binomial distribution with $n$ independent trials, it is easy to see that $Z_{t_k,n}$ takes values $Z_{t_k,n} = j$, where $j \in \{0, 1, 2, \ldots, n\}$.

Similarly, when the stochastic process is observed over times $t_k, k = 0, 1, 2, \ldots, T, T \in \mathbb{Z}_+$, then applying the design of the HD structure in (3.1)-(3.3), the collection of all possible infectious contacts an individual makes in the $(T + 1)$ subintervals $[t_k, t_{k+1}), k \in \{0, 1, 2, \ldots, T\}$, forms the third level of the HD structure defined by

$$C^{(3)} = \bigcup_{n=0}^{\infty} \left\{ Z_{t_k,n} | k = 0, 1, 2, \ldots, T \right\}. \tag{3.7}$$

**Definition 3.4. The 2nd level hierarchical data structure in the SEIR model:**

4. From (3.6) and (3.4), assume that during the time interval $[t_k, t_{k+1})$, a susceptible individual makes $Z_{t_k} = n$ contacts with other individuals in the population- both infectious and noninfectious individuals. Assume further that in the $Z_{t_k} = n$ total contacts made, there are $Z_{t_k,n} = j$ of the contacts that are with infectious individuals, where $j \in \{0, 1, 2, \ldots, n\}$, and $n \in \{0, 1, 2, \ldots\}$. It is clear that any one of the $Z_{t_k,n} = j$ infectious contacts with the susceptible individual can lead to infection.

Thus, for $j \in \{0, 1, 2, \ldots, n\}$, and $n \in \{0, 1, 2, \ldots\}$, given that the susceptible person makes $Z_{t_k,n} = j$ infectious contacts, define a categorical random variable, $Z_{t_k,n,j}$, that indicates the $l^{th}$ infectious contact that potentially leads to infection, defined in Definition 3.1, and $l = 1, 2, \ldots, j$. That is,

$$Z_{t_k,n,j} = \begin{cases} 1, & \text{if the 1st person passes infection,} \\ 2, & \text{if the 2nd person passes infection,} \\ \vdots & \qquad\qquad \vdots \\ j, & \text{if the } j^{th} \text{ person passes infection.} \end{cases} \tag{3.8}$$

The distribution of $Z_{t_k,n,j}$ is characterized as follows. Define the sequence of probabilities (or masses) $(\gamma_1, \gamma_2, \ldots, \gamma_l)$, where $\sum_{l=1}^{j} \gamma_l = 1$. Under the assumption of homogenous mixing in the population, it is easy to see that all of the $j$ infectious contacts carry the same risk of infecting the susceptible person, that is, all $j$ infectious contacts are equally likely to lead to infection. Thus,

$$P\left( Z_{t_k,n,j} = l | Z_{t_k,n} = j, Z_{t_k} = n \right) = \gamma_l = \frac{1}{j}, \forall l \in \{1, 2, \ldots, j\},$$

is the probability that the $l^{th}$ infectious contact potentially leads to infection.

Similarly, when the stochastic process is observed over times $t_k, k = 0, 1, 2, \ldots, T$, $T \in \mathbb{Z}_+$, then applying the design of the hierarchical data structure in (3.1)-(3.3), the collection of all possible potential single effective infectious contacts a susceptible individual makes in the $(T + 1)$ subintervals $[t_k, t_{k+1}), k \in \{0, 1, 2, \ldots, T\}$ forms the second level of the hierarchical data structure defined by

$$C^{(2)} = \bigcup_{n=0}^{\infty} \bigcup_{j=0}^{n} \left\{ Z_{t_k,n,j} | k = 0, 1, 2, \ldots, T \right\}. \tag{3.9}$$

**Definition 3.5. The 1st level hierarchical data structure in the SEIR model:**

From (3.8), suppose it is given that a susceptible person makes $Z_{t_k} = n$ contacts, and $Z_{t_k,n} = j$, of the contacts are with infectious people, where $j \in \{0, 1, \ldots, n\}$; suppose it is also known that the $l^{th}$ infectious person potential passes the infection in the interval $[t_k, t_{k+1})$, where from (3.8), $Z_{t_k,n,j} = l$ is the indicator of the $l^{th}$ infectious person passing the infection in the interval $[t_k, t_{k+1})$, and $l \in \{1, 2, \ldots, j\}$. Define a new Bernoulli random variable, $Z_{t_k,n,j,l}$, which indicates the event that the $l^{th}$ infectious person passes the infection in $[t_k, t_{k+1})$. That is,

$$Z_{t_k,n,j,l} = \begin{cases} 1, & \text{if the } l^{th} \text{ person passes infection,} \\ 0, & \text{otherwise,} \end{cases} \tag{3.10}$$

where $l \in \{1, 2, \ldots, j\}$.

Recall Definitions 3.1-3.4, it is known that the probability of passing infection from one infectious contact is $p \in [0, 1]$. Thus, the distribution of $Z_{t_k,n,j,l}$ is given by

$$P\left( Z_{t_k,n,j,l} = 1 | Z_{t_k,n,j} = l, Z_{t_k,n} = j, Z_{t_k} = n \right) = p,$$

$$P\left( Z_{t_k,n,j,l} = 0 | Z_{t_k,n,j} = l, Z_{t_k,n} = j, Z_{t_k} = n \right) = 1 - p, \tag{3.11}$$

$\forall l \in \{1, 2, \ldots, j\}; \forall j \in \{0, 1, 2, \ldots, n\}; \forall n \in \{0, 1, 2, \ldots\}$.

Similarly, when the stochastic process is observed over times $t_k, k = 0, 1, 2, \ldots, T$, $T \in \mathbb{Z}_+$, then applying the design of the hierarchical data structure in (3.1)-(3.3), the collection of zeros and ones indicating success or failure of susceptible persons getting infected from all possible potential single effective infectious contacts in the $(T + 1)$ subintervals $[t_k, t_{k+1}), k \in \{0, 1, 2, \ldots, T\}$ forms the first level of the hierarchical data structure defined by

$$C^{(1)} = \bigcup_{n=0}^{\infty} \bigcup_{j=0}^{n} \bigcup_{l=1}^{j} \left\{ Z_{t_k,n,j,l} | k = 0, 1, 2, \ldots, T \right\}. \tag{3.12}$$

The following trivial result is given to simply show how the Definitions 3.1-3.4 lead to the distribution of $C_{SE}(t_k) \sim Binomial(S(t_k), P_{SE}(t_k))$.

**Theorem 3.1.** *Let the Definitions 3.1-3.4 be satisfied. The success probability $P_{SE}(t_k)$ of the binomial distribution $C_{SE}(t_k) \sim Binomial(S(t_k), P_{SE}(t_k))$ in (2.5) is given by*

$$P_{SE}(t_k) = 1 - e^{-p\alpha_{t_k}\lambda\Delta t}. \tag{3.13}$$

**Proof.** See Appendix A. □

**Remark 3.1.** Observe that the contact rate $\lambda$ and the probability of passing infection from one infectious contact $p$, give the effective contact rate $\beta(t_k) = \beta = p\lambda$ in [26, 39], where $\beta(t_k) = \beta = p\lambda$ is defined as the average number of effective contacts required to pass infection at time $k \geq 0$. Thus, the probability of getting infected at any time $k$, $P_{SE}(t_k), k \geq 0$ in (3.13) is exactly the same as in [26, 39].

### 3.2. Transition probabilities for the SEIR model with random incubation and infectious periods

It is assumed that the incubation and infectious periods, $T_1$ and $T_2$, are random variables. That is, a newly infected person remains in the exposed state ($E$) for an average time $\mathbb{E}(T_1) = \frac{1}{\delta_e}$, where $\delta_e$ is the average number of transitions from the exposed state ($E$) to the infectious stat ($I$) per unit time. Similarly, assume that a newly infectious person ($I$) is removed ($R$) from the infectious state after the average time $\mathbb{E}(T_1) = \frac{1}{\delta_r}$, where $\delta_r$ is the average number of transitions from the infectious state ($I$) to the removal state ($R$) per unit time. Clearly, if all transitions per unit time are independent, then $T_1$ and $T_2$ have exponential distributions. Moreover, $T_1$ and $T_2$ are independent random variables.

Trivially, using the survival distribution function $\bar{S}_x(t_k)$ of an individual in the state $x \in \{E, I\}$, the distribution of the transition events $C_{ij}(t_k)$ in (2.5) of the SEIR model (2.1)-(2.4) are explicitly given by

$$C_{SE}(t_k) \sim Binomial(S(t_k), P_{SE}(t_k)), \quad C_{EI}(t_k) \sim Binomial(E(t_k), P_{EI}(t_k)),$$

$$C_{IR}(t_k) \sim Binomial(I(t_k), P_{IR}(t_k)), \tag{3.14}$$

where $P_{SE}(t_k)$ is given in (3.13);

$$P_{EI}(t_k) = 1 - \frac{\bar{S}_E(t_{k+1})}{\bar{S}_E(t_k)} = 1 - e^{-\delta_e \Delta t}, \forall i = 1, 2, \ldots, e_k, \tag{3.15}$$

and

$$P_{IR}(t_k) = 1 - \frac{\bar{S}_I(t_{k+1})}{\bar{S}_I(t_k)} = 1 - e^{-\delta_r \Delta t}, \forall i = 1, 2, \ldots, i_k. \tag{3.16}$$

Further, from (2.9) and (3.14), an explicit formula for the transition probabilities are obtained in the following.

**Theorem 3.2.** *Let $T_1$ and $T_2$ be two independent exponential random variables with means $\mathbb{E}(T_1) = \frac{1}{\delta_e}$ and $\mathbb{E}(T_2) = \frac{1}{\delta_r}$, respectively, where rate parameters are $\delta_e$ and $\delta_r$, respectively. The SEIR Markov chain model $\{X(t_k), k = 0, 1, 2, \ldots\}$ has the following transition probabilities:*

$$P(X(t_{k+1}) = x^{k+1}|X(t_k) = x^k) = P((S(t_{k+1}), E(t_{k+1}), I(t_{k+1}) = (s_{k+1}, e_{k+1}, i_{k+1})|X(t_k) = x^k)$$

$$= \begin{cases} \binom{s_k}{s_{k+1}}(p^i(t_k))^{s_k - s_{k+1}}(1 - p^i(t_k))^{s_{k+1}} \times \\ \times \binom{e_k}{s_{k+1}+e_{k+1}}(P_{EI}^i(t_k))^{s_k+e_k-(s_{k+1}+e_{k+1})} \times \\ \times (1 - P_{EI}^i(t_k))^{s_{k+1}+e_{k+1}-s_k} \times \\ \times \binom{i_k}{s_{k+1}+e_{k+1}+i_{k+1}}(P_{IR}^i(t_k))^{s_k+e_k+i_k-(s_{k+1}+e_{k+1}+i_{k+1})} \times \\ \times (1 - P_{IR}^i(t_k))^{s_{k+1}+e_{k+1}+i_{k+1}-(s_k+e_k)}, & \text{for} \\ & (s_{k+1}, e_{k+1}, i_{k+1}) \in \Omega_X(t_{k+1}) \\ 0, & \text{otherwise.} \end{cases}$$

*Moreover, the feasible region for the process is given as follows:*

$$\Omega_X(t_{k+1}) = \{(s_{k+1}, e_{k+1}, i_{k+1}) \in \mathbb{Z}_+^3 | s_k \leq s_{k+1} + e_{k+1} \leq s_k + e_k,$$

$$s_k + e_k \leq s_{k+1} + e_{k+1} + i_{k+1} \leq s_k + e_k + i_k\}.$$

**Proof.** From (2.1)-(2.4) and (2.9), observe that

$$P(X(t_{k+1})|X(t_k)) = P((S(t_{k+1}), E(t_{k+1}), I(t_{k+1}) = (s_{k+1}, e_{k+1}, i_{k+1})|X(t_k) = x^k)$$

$$= P(C_{SE}(t_k) = (S(t_k) - S(t_{k+1}), C_{EI}(t_k) = (E(t_k) - E(t_{k+1}) + (S(t_k) - S(t_{k+1}),$$

$$C_{IR}(t_k) = (I(t_k) - I(t_{k+1}) + (E(t_k) - E(t_{k+1}) + (S(t_k) - S(t_{k+1})|X(t_k)). \tag{3.17}$$

The rest of the proof follows easily by applying the distribution of the transition events in (3.14). $\square$

### 3.3. Numerical results for the SEIR Markov chain models

Prior to deriving the EM-algorithm for hierarchical data in (3.1)-(3.3), and employing the algorithm to finding maximum likelihood estimates for the parameters of the two Markov chain model $\{X(t_k) : k \geq 0\}$ with transition probabilities in Theorem 3.2, from real epidemic data, we apply a theoretical approach to validate or assess the performance of the two Markov chain models. See Appendix B for the numerical results.

## 4. The hierarchical data EM-algorithm

### 4.1. The traditional EM algorithm and Jensen's inequality

The Expectation Maximization (EM) algorithm (cf. [13, 20, 30, 37]) is a computational method for finding the maximum likelihood estimate (denoted MLE or ML-estimate) of a parameter vector $\Theta$ of a given distribution. The EM-algorithm is most useful in: (1) when the given data for the ML technique has missing components at random; and (2) when the likelihood function is intractable to maximize, but adding hidden/missing data simplifies the process. In this study, the EM-algorithm is employed to remedy the two cases (1)-(2).

Kung et al. [37] present an explicit review of the traditional derivation and convergence of the EM-algorithm (see Section 3.2, [37]). Indeed, let $Y$ be observed data; $L(\Theta|Y) = P(Y|\Theta)$ is the likelihood function; let the vector $Z$ be the missing data or a missing component of the given data; and the complete data is $X = (Y, Z)$. The log-likelihood function $log(L(\Theta|X)) = log(P(Y, Z|\Theta))$ for the complete data is obtained and maximized to find the MLE of $\Theta$ in two basic algorithm steps, namely: the expectation (E)-step, and the maximization (M)-step.

The E-step of the algorithm consists of finding the conditional expected value of the complete log-likelihood function, denoted by

$$Q(\Theta|\Theta_m) = \mathbb{E}_{Z|Y;\Theta_m}[\log(L(\Theta|X))], \tag{4.1}$$

where

$$\mathbb{E}_{Z|Y;\Theta_m}[\log(L(\Theta|X))] = \mathbb{E}_{Z|Y;\Theta_m}[\log(P(Y, Z|\Theta))] = \sum_Z \log[P(Y, Z|\Theta)]P(Z|Y;\Theta_m),$$

and $\Theta_m$ is the estimate of $\Theta$ at the $m^{th}$ iteration or step, where $m = 0, 1, 2, \ldots$.

The M-step consists of maximizing the Q-function $Q(\Theta|\Theta_m)$ to find an estimate $\Theta_{m+1}$ for $\Theta$. This process is summarized in the following steps:

**Algorithm 4.1.**

1. Let $m = 0$ and $\Theta_m$ be an initial guess for $\Theta$.
2. For the observed data $Y$, and assuming that the guess $\Theta_m$ is correct, calculate the conditional expected log-likelihood $Q(\Theta|\Theta_m)$ in (4.1).
3. Find the $\Theta = \Theta_{m+1}$ that maximizes $Q(\Theta|\Theta_m)$, that is, find

$$\Theta_{m+1} = \underset{\Theta}{argmax}\left[Q(\Theta|\Theta_m)\right]. \tag{4.2}$$

4. Update $\Theta_m$ with $\Theta_{m+1}$, and repeat step 1 to step 3 until $\Theta$ stops noticeably changing.

Employing Jensen's inequality, the convergence of the EM-algorithm can be obtained. Indeed, recall Jensen's inequality [17] in the following:

**Lemma 4.1.** *Let $f$ be a convex function, and $X$ a random variable, then*

$$E[f(X)] \geq f(E[X]). \tag{4.3}$$

*Conversely, if you have a concave function (e.g. a logarithmic function), then*

$$E[f(X)] \leq f(E[X]). \tag{4.4}$$

By applying Lemma 4.1, it can be shown that the incomplete data log-likelihood function satisfies

$$\log(L(\Theta|Y)) = \log(P(Y|\Theta)) \geq Q(\Theta|\Theta_m) + R(\Theta_m|\Theta_m), \tag{4.5}$$

where $R(\Theta_m|\Theta_m)$ is the entropy term given by

$$R(\Theta_m|\Theta_m) = -E_{Z|Y;\Theta_m}\left[\log(P(Z|Y;\Theta_m))\right] = -\sum_Z \log[P(Z|Y;\Theta_m)]P(Z|Y;\Theta_m).$$

To verify (4.5), the reader is referred to [Subsection 3.2.2, Eq. (3.2.7), [37]].

**Remark 4.1.** There are several techniques to derive the Q-function (4.1) for the E-step of the EM-algorithm, for example, see [Subsection 3.2.1, Eq. (3.2.3), [37]]. However, the method employed in this paper to find an expression for the Q-function for the E-step utilizes the Jensen's inequality in Lemma 4.1, to obtain an expression similar to (4.5), which simultaneously gives the Q-function, the entropy, and also proves convergence of the EM-algorithm.

### 4.2. The EM algorithm in nested hierarchical data

In this paper, to obtain the Q-function $Q(\Theta|\Theta_m)$ in (4.1) of the EM-algorithm for the hierarchical data set presented subsequently, we employ the approach in [Subsection 3.2.2, Eq. (3.2.7), [37]], that consists of applying Jensen's inequality to obtain a similar relationship as (4.5) for the incomplete data log-likelihood function. The following definitions are made.

**Definition 4.1.**

(1.) Define $Y$, to be the observed data or observed random sample. In fact, $Y$ is a finite sequence of observed data written, for example, in the form

$$Y = \{Y_j | j = 1, 2, \ldots, n_M\},$$

where the values of the $Y_j$'s are real numbers, or vectors with real coordinates, generated from populations that exhibit either (1) independent and identical distributions, or (2) dependent and identical distributions. The latter case applies mostly to stochastic population data. Without loss of generality, in the derivation of the EM-algorithm, for the hierarchical data, $Y$ will be treated as a single random variable representing an observable characteristic in the population. Moreover, $Y$ is assumed to be a discrete random variable or vector, over $\mathbb{Z}_+$, the set of non-negative integers.

(2.) Define $Z$ to be the missing data or missing components with a hierarchical data structure consisting of $M$-levels ($M > 1$), similar to (3.1)-(3.3), given by

$$Z = \left\{ C^{(M)}, C^{(M-1)}, \ldots, C^{(2)}, C^{(1)} \right\}. \tag{4.6}$$

That is, it is assumed that the missing data $Z$ has $M$ nested levels of hierarchy, which is similarly structured as (3.1)-(3.3). Without loss of generality, $Z$ will be treated as hierarchical data collection of discrete random variables over $\mathbb{Z}_+$, the set of non-negative integers.

(3.) Define $X$ to be the complete data containing both $Y$ and $Z$. That is, $X = (Y, Z)$.

(4.) Define $\Theta = (\theta_1, \theta_2, \ldots, \theta_\tau)$ to be the vector of unknown parameters that defines the true distribution of $Y$, where $\tau$ is positive integer.

(5.) Define $l(\Theta|Y) = \log[L(\Theta|Y)] = \log[P(Y|\Theta)]$ be the log-likelihood function of the incomplete data.

Note that the Q-function in (4.1) is a conditional expectation over all possible values of the missing random variable $Z$. Thus, since we can only average one random variable at a time, we apply a step-by-step approach to obtain the Q-function by averaging over the different levels of the hierarchical missing data $Z$ in (4.6). That is, we obtain the corresponding Q-function with the effects of $C^{(M)}$ in step 1, then consider the effects of $C^{(M-1)}$ in the step 2; $C^{(M-2)}$ in the step 3, and so on, until the first level $C^{(1)}$ of the hierarchical data structure.

In addition, we apply an inductive approach for all positive integers $M > 1$ as follows. To organize the solution process, we present in the following the results for $M = 3$, and generalize in the next result.

**Theorem 4.1.** *From (3.1)-(3.3), let $M = 3$, that is, we assume that the missing data $Z$ in (4.6) consists of*

$$Z = \left\{ C^{(3)}, C^{(2)}, C^{(1)} \right\},$$

*where from (3.1)-(3.3),*

$$C^{(3)} = \left\{ Z_{k_3} | k_3 = 1, 2, 3, \ldots n_3 \right\},$$

$$C^{(2)} = \bigcup_{k_3=1}^{n_3} \left\{ Z_{k_3, k_2} | k_2 = 1, 2, 3, \ldots n_2 \right\},$$

$$C^{(1)} = \bigcup_{k_3=1}^{n_3} \bigcup_{k_2=1}^{n_2} \left\{ Z_{k_3, k_2, k_1} | k_1 = 1, 2, 3, \ldots n_1 \right\}. \tag{4.7}$$

*Also, let $Y$, $\Theta$ and $l(\Theta|Y)$ be as defined in Definition 4.1. It follows that the log-likelihood function of the incomplete data $l(\Theta|Y)$ satisfies*

$$l(\Theta|Y) = \log[L(\Theta|Y)] \geq Q_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta|\Theta_m) + R_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta_m|\Theta_m) + R_{Z_{k_3}, Z_{k_2}}(\Theta_m|\Theta_m) + R_{Z_{k_3}}(\Theta_m|\Theta_m),$$

*where*

$$Q_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta|\Theta_m) = \mathbb{E}_{Z_{k_3}|Y,\Theta_m} \left[ \mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m} \left[ \mathbb{E}_{Z_{k_1}|Y,Z_{k_3},Z_{k_2},\Theta_m} \left[ \log \left[ P(Y, Z_{k_3}, Z_{k_2}, Z_{k_1}|\Theta) \right] \right] \right] \right], \tag{4.8}$$

$$R_{Z_{k_3}}(\Theta_m|\Theta_m) = -\mathbb{E}_{Z_{k_3}|Y,\Theta_m} \left[ \log P(Z_{k_3}|Y,\Theta_m) \right], \tag{4.9}$$

$$R_{Z_{k_3}, Z_{k_2}}(\Theta_m|\Theta_m) = -\mathbb{E}_{Z_{k_3}|Y,\Theta_m} \left[ \mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m} \left[ \log \left( P(Z_{k_2}|Y,Z_{k_3},\Theta_m) \right) \right] \right], \tag{4.9}$$

*and*

$$R_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta_m|\Theta_m) = -\mathbb{E}_{Z_{k_3}|Y,\Theta_m} \left[ \mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m} \left[ \mathbb{E}_{Z_{k_1}|Y,Z_{k_3},Z_{k_2},\Theta_m} \left[ \log \left[ P(Z_{k_1}|Y,Z_{k_3},Z_{k_2},\Theta_m) \right] \right] \right] \right]. \tag{4.10}$$

*That is, the log-likelihood function of the incomplete data $l(\Theta|Y)$ is bounded from below by the Q-function $Q_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta|\Theta_m)$ which includes the average effects of all the missing hierarchical information $C^{(3)}$, $C^{(2)}$ and $C^{(1)}$ in $Z$, and the entropy terms $R_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta_m|\Theta_m)$, $R_{Z_{k_3}, Z_{k_2}}(\Theta_m|\Theta_m)$, and $R_{Z_{k_3}}(\Theta_m|\Theta_m)$.*

**Proof.** The complete proof of Theorem 4.1 is given in Appendix C. $\square$

**Remark 4.2. E- and M-steps of the EM-algorithm, and convergence:**

(1.) Note that it can be shown, applying a similar direct approach of deriving the Q-function exhibited in [Subsection 3.2.2, Eq. (3.2.3), [37]], and the step-by-step method of adding and averaging over the missing hierarchical information $C^{(3)}$, $C^{(2)}$ and $C^{(1)}$ that the log-likelihood function of the incomplete data, $l(\Theta_m|Y)$, for $\Theta = \Theta_m$ satisfies

$$l(\Theta_m|Y) = \log\left[L(\Theta_m|Y)\right] = Q_{Z_{k_3},Z_{k_2},Z_{k_1}}(\Theta_m|\Theta_m) + R_{Z_{k_3},Z_{k_2},Z_{k_1}}(\Theta_m|\Theta_m) + R_{Z_{k_3},Z_{k_2}}(\Theta_m|\Theta_m) + R_{Z_{k_3}}(\Theta_m|\Theta_m), \qquad (4.11)$$

where $Q_{Z_{k_3},Z_{k_2},Z_{k_1}}(\Theta_m|\Theta_m)$, $R_{Z_{k_3},Z_{k_2},Z_{k_1}}(\Theta_m|\Theta_m)$, $R_{Z_{k_3},Z_{k_2}}(\Theta_m|\Theta_m)$, and $R_{Z_{k_3}}(\Theta_m|\Theta_m)$ are given in (4.8)-(4.10). The proof of (4.11) is omitted to reduce the complexity of the derivation process of the EM-algorithm in this paper, especially because no additional knowledge is gained. Furthermore, for any $\Theta_{max}$ that satisfies

$$\Theta_{max} = \underset{\Theta}{argmax}\left[Q_{Z_{k_3},Z_{k_2},Z_{k_1}}(\Theta|\Theta_m)\right], \qquad (4.12)$$

it implies that the following holds

$$Q_{Z_{k_3},Z_{k_2},Z_{k_1}}(\Theta_{max}|\Theta_m) \geq Q_{Z_{k_3},Z_{k_2},Z_{k_1}}(\Theta_m|\Theta_m), \qquad (4.13)$$

and as a result the following relationship is satisfied

$$l(\Theta_{max}|Y) \geq l(\Theta_m|Y). \qquad (4.14)$$

The relationships in (4.13)-(4.14) imply that, at the $(n+1)^{th}$ iteration, as we continuously replace the value of $\Theta_m$ in the Q-function $Q_{Z_{k_3},Z_{k_2},Z_{k_1}}(\Theta|\Theta_m)$ in (4.8), by the new value $\Theta_{max}$ obtained from (4.12), we move closer to the true value of $\Theta$ that maximizes the log-likelihood $l(\Theta|Y)$.

(2.) Therefore, from above convergence property in (4.14), we conclude that

**E-step:** the E-step of the EM-algorithm consists of finding the Q-function $Q_{Z_{k_3},Z_{k_2},Z_{k_1}}(\Theta|\Theta_m)$ in (4.8), and

**M-step:** the M-step consists of finding $\Theta_{max}$ in (4.12).

The mathematical patterns observed in the Q-functions (4.8) and entropy terms (4.9)-(4.10), given in Theorem 4.1 are easy to generalize for the $M$-level hierarchical missing data $Z$ in (4.6). The major issue with the generalization is the complex nature of the conditional expectation notation, when the level of hierarchy $M > 1$ is arbitrary. We introduce new notations for the conditional expectation in the level $M > 1$.

**Notation 4.1.**

(1.) The sequence of random variables $(Z_{k_M}, Z_{k_{(M-1)}}, \ldots, Z_{k_3}, Z_{k_2}, Z_{k_1})$, is denoted by $(Z_{k_M} - Z_{k_1})$; the sequence $(Z_{k_M}, Z_{k_{(M-1)}}, \ldots, Z_{k_3}, Z_{k_2})$, is denoted by $(Z_{k_M} - Z_{k_2})$. And in general, for each $j = 1, 2, \ldots, M-1$, the sequence $(Z_{k_M}, Z_{k_{(M-1)}}, \ldots, Z_{k_{(j+1)}})$, is denoted by $(Z_{k_M} - Z_{k_{(j+1)}})$.

(2) Denote by $\mathbb{E}^*_{Z_{k_1}|Y,Z_{k_M}-Z_{k_2},\Theta_m}[\cdots]$, the conditional expectation with respect to the mass of $Z_{k_1}$, given the sequence $(Y, Z_{k_M}, Z_{k_{(M-1)}}, \ldots, Z_{k_3}, Z_{k_2}, \Theta_m)$, that is,

$$\mathbb{E}^*_{Z_{k_1}|Y,Z_{k_M}-Z_{k_2},\Theta_m}[\cdots] \equiv \mathbb{E}_{Z_{k_1}|Y,Z_{k_M},Z_{k_{(M-1)}},\ldots,Z_{k_3},Z_{k_2},\Theta_m}[\cdots],$$

where the notation "$\equiv$" signifies "equivalence" between two identities. And in general, for each $j = 1, 2, \ldots, M-1$, denote by $\mathbb{E}^*_{Z_{k_j}|Y,Z_{k_M}-Z_{k_{j+1}},\Theta_m}[\cdots]$, the conditional expectation with respect to the mass of $Z_{k_j}$, given the sequence $(Y, Z_{k_M}, Z_{k_{(M-1)}}, \ldots, Z_{k_3}, Z_{k_{j+1}}, \Theta_m)$, that is,

$$\mathbb{E}^*_{Z_{k_j}|Y,Z_{k_M}-Z_{k_{(j+1)}},\Theta_m}[\cdots] \equiv \mathbb{E}_{Z_{k_1}|Y,Z_{k_M},Z_{k_{(M-1)}},\ldots,Z_{k_3},Z_{k_2},\Theta_m}[\cdots], j \in \{1, 2, \ldots, M-1\},$$

$$\mathbb{E}^*_{Z_{k_M}|Y,\Theta_m}[\cdots] \equiv \mathbb{E}_{Z_{k_M}|Y,\Theta_m}[\cdots].$$

A generalization of Theorem 4.1 is presented subsequently, and the proof follows immediately by applying the method of mathematical induction.

**Proposition 4.1.** *Let $M > 1$, $Y$, $Z$, $\Theta$ and $l(\Theta|Y)$ be as defined in Definition 4.1. It follows that the log-likelihood function of the incomplete data $l(\Theta|Y)$, satisfies*

$$l(\Theta|Y) = \log\left[L(\Theta|Y)\right]$$

$$\geq Q_{Z_{k_M}-Z_{k_1}}(\Theta|\Theta_m) + R_{Z_{k_M}-Z_{k_1}}(\Theta_m|\Theta_m)$$

$$+ R_{Z_{k_M}-Z_{k_2}}(\Theta_m|\Theta_m) + R_{Z_{k_M}-Z_{k_3}}(\Theta_m|\Theta_m)$$

$$+ R_{Z_{k_M}-Z_{k_4}}(\Theta_m|\Theta_m) + R_{Z_{k_M}-Z_{k_5}}(\Theta_m|\Theta_m)$$

$$+ \ldots + R_{Z_{k_M}-Z_{k_{(M-1)}}}(\Theta_m|\Theta_m) + R_{Z_{k_M}}(\Theta_m|\Theta_m), \qquad (4.15)$$

*where*

$$Q_{Z_{k_M}-Z_{k_1}}(\Theta|\Theta_m) = \mathbb{E}^*_{Z_{k_M}|Y,\Theta_m}\Big[\, \mathbb{E}^*_{Z_{k_{(M-1)}}|Y,Z_{k_M},\Theta_m}\Big[\, \mathbb{E}^*_{Z_{k_{(M-2)}}|Y,Z_{k_M}-Z_{k_{(M-1)}},\Theta_m}\Big[\, \ldots \mathbb{E}^*_{Z_{k_4}|Y,Z_{k_M}-Z_{k_5},\Theta_m}\Big[$$

$$\mathbb{E}^*_{Z_{k_3}|Y,Z_{k_M}-Z_{k_4},\Theta_m}\Big[\, \mathbb{E}^*_{Z_{k_2}|Y,Z_{k_M}-Z_{k_3},\Theta_m}\Big[$$

$$\mathbb{E}^*_{Z_{k_1}|Y,Z_{k_M}-Z_{k_2},\Theta_m}\Big[\log\Big[P(Y, Z_{k_M}, Z_{k_{(M-1)}}, \ldots, Z_{k_3}, Z_{k_2}, Z_{k_1}|\Theta)\Big]\Big]\,\Big]\,\Big], \qquad (4.16)$$

*is the generalized Q-function which considers the effects of all levels of missing information in $Z$ given by (4.6). In addition, the entropy terms are given as follows.*

$$R_{Z_{k_M}}(\Theta_m|\Theta_m) = -\mathbb{E}_{Z_{k_M}|Y,\Theta_m}\left[\log P(Z_{k_M}|Y,\Theta_m)\right],\tag{4.17}$$

$$R_{Z_{k_M}-Z_{k_{(M-1)}}}(\Theta_m|\Theta_m) = -\mathbb{E}_{Z_{k_M}|Y,\Theta_m}\left[\mathbb{E}_{Z_{k_{(M-1)}}|Y,Z_{k_M},\Theta_m}\left[\log\left(P(Z_{k_{(M-1)}}|Y,Z_{k_M},\Theta_m)\right)\right]\right],$$

$$R_{Z_{k_M}-Z_{k_{(M-2)}}}(\Theta_m|\Theta_m) = -\mathbb{E}^*_{Z_{k_M}|Y,\Theta_m}\left[\mathbb{E}^*_{Z_{k_{(M-1)}}|Y,Z_{k_M},\Theta_m}\left[\mathbb{E}^*_{Z_{k_{(M-2)}}|Y,Z_{k_M}-Z_{k_{(M-1)}},\Theta_m}\left[\log\left[P(Z_{k_{(M-2)}}|Y,Z_{k_M}-Z_{k_{(M-1)}},\Theta_m)\right]\right]\right]\right],$$

*and in general, for $j = 0, 1, 2, \ldots, M-2$, the other entropy terms are obtained using the formula*

$$R_{Z_{k_M}-Z_{k_{(j+1)}}}(\Theta_m|\Theta_m) = -\mathbb{E}^*_{Z_{k_M}|Y,\Theta_m}\left[\,\mathbb{E}^*_{Z_{k_{(M-1)}}|Y,Z_{k_M},\Theta_m}\left[\,\mathbb{E}^*_{Z_{k_{(M-2)}}|Y,Z_{k_M}-Z_{k_{(M-1)}},\Theta_m}\right.\right.$$

$$\mathbb{E}^*_{Z_{k_{(M-3)}}|Y,Z_{k_M}-Z_{k_{(M-2)}},\Theta_m}\left[\,\ldots\mathbb{E}^*_{Z_{k_{(j+2)}}|Y,Z_{k_M}-Z_{k_{(j+3)}},\Theta_m}\left[\,\mathbb{E}^*_{Z_{k_{(j+1)}}|Y,Z_{k_M}-Z_{k_{(j+2)}},\Theta_m}\right.\right.$$

$$\log\left[P(Z_{k_{(j+1)}}|Y,Z_{k_M}-Z_{k_{(j+2)}},\Theta_m)\right]\,]\,]\,].\tag{4.18}$$

*For example, using (4.18), when $j = 0$, we obtain $R_{Z_{k_M}-Z_{k_1}}(\Theta_m|\Theta_m)$ given by*

$$R_{Z_{k_M}-Z_{k_1}}(\Theta_m|\Theta_m) = -\mathbb{E}^*_{Z_{k_M}|Y,\Theta_m}\left[\,\mathbb{E}^*_{Z_{k_{(M-1)}}|Y,Z_{k_M},\Theta_m}\left[\,\mathbb{E}^*_{Z_{k_{(M-2)}}|Y,Z_{k_M}-Z_{k_{(M-1)}},\Theta_m}\right.\right.$$

$$\mathbb{E}^*_{Z_{k_{(M-3)}}|Y,Z_{k_M}-Z_{k_{(M-2)}},\Theta_m}\left[\,\ldots\mathbb{E}^*_{Z_{k_4}|Y,Z_{k_M}-Z_{k_5},\Theta_m}\left[\,\mathbb{E}^*_{Z_{k_3}|Y,Z_{k_M}-Z_{k_4},\Theta_m}\right.\right.$$

$$\mathbb{E}^*_{Z_{k_2}|Y,Z_{k_M}-Z_{k_3},\Theta_m}\left[\,\mathbb{E}^*_{Z_{k_1}|Y,Z_{k_M}-Z_{k_2},\Theta_m}\left[\,\log\left[P(Z_{k_1}|Y,Z_{k_M}-Z_{k_2},\Theta_m)\right]\,]\,]\,].\tag{4.19}$$

*In other words, (4.15) signifies that, the log-likelihood function of the incomplete data $l(\Theta|Y)$ is bounded from below by the Q-function $Q_{Z_{k_M}-Z_{k_1}}(\Theta|\Theta_m)$ which includes the average effects of all the missing hierarchical information $C^{(M)}\cdots, C^{(2)}$ and $C^{(1)}$ in $Z$, and the entropy terms in (4.17)-(4.19), which represent the uncertainty in the different levels of the missing data in $Z$.*

**Proof.** The proof of this result follows immediately by the method of mathematical induction from Theorem 4.1. Thus, the proof is omitted. $\square$

**Remark 4.3.** It can be inferred from the convergence property exhibited in Remark 4.3 that when there are $M$ hierarchical levels in the missing data $Z$ in (4.6), the E- & M- steps of the EM-algorithm consist of the following.

**E-step:** the E-step of the EM-algorithm consists of finding the Q-function $Q_{Z_{k_M}-Z_{k_1}}(\Theta|\Theta_m)$ in (4.16), and
**M-step:** the M-step consists of finding $\Theta_{max}$ below.

$$\Theta_{max} = \underset{\Theta}{argmax}\left[Q_{Z_{k_M}-Z_{k_1}}(\Theta|\Theta_m)\right].$$

Recall Definitions 3.2-3.5, there are four levels of hidden hierarchical data in the SEIR model. Thus, the application of the EM-algorithm in Theorem 4.1 and Proposition 4.1 to the SEIR model is exhibited in the next section for the case where the missing data has $M = 4$ - levels of hierarchy.

## 5. Application of the hierarchical data EM-algorithm

### 5.1. The traditional maximum likelihood estimation technique in the SEIR model

Note that the hierarchical data EM-algorithm proposed in Theorem 4.1 and Proposition 4.1 has wider applications than in the SEIR epidemic model presented in this study as a case study. In fact, the 4-level hierarchical missing data given in Section 3.1 serves mainly as an example to understand the application of the EM algorithm in hierarchical data sets. Also, note that applying the EM-algorithm to the SEIR epidemic model is intended more to be informative about the process of applying the algorithm, and not the process that generates the data for the SEIR epidemic model. The process that generates the data for the SEIR epidemic model described below in Remark 5.1(2.) assumes that contact tracing (cf. [4]) is employed during the epidemic, and this is not applicable in all settings of SEIR epidemics.

In this section, we derive estimators for the true parameters of the model in Theorems 3.2, given any observed data for the state of the process over time. The maximum likelihood estimation technique (cf. [17]) has been applied in similar stochastic systems (cf. [26, 35, 47, 61]). We employ ideas in these studies to find maximum likelihood estimators (MLE's) for the parameter $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ for the SEIR model in Theorem 3.2. Observe that the SEIR model in Theorem 3.2 has four parameters $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$. To facilitate understanding of the application of the 4-level EM-algorithm in Theorem 4.1 and Remark 4.2, we find MLE's for the four parameter model in Theorem 3.2.

Note that in a stochastic system, the parameter $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ represents fixed measures in the population at each time step $t_k$, where the population at any time $t_k, k \geq 0$ is defined by the random vector

$$X(t_k) = (S(t_k), E(t_k), I(t_k)),\tag{5.1}$$

which satisfies the model equations (2.1)-(2.4). Specific values of the random vector $X(t_k)$ are denoted

$$\hat{x}(t_k) = (\hat{s}_k, \hat{e}_k, \hat{i}_k), \forall k = 0, 1, 2, \ldots.\tag{5.2}$$

**Remark 5.1.**

(1.) Recall Remark 3.1, the contact rate $\lambda$ and the probability of passing infection from one infectious contact $p$, give the effective contact rate $\beta(t_k) = \beta = p\lambda$ in [26, 39]. Note that in practice, the parameter $\beta = p\lambda$ is identifiable (cf. [26, 39]), but the parameters $p$ and $\lambda$ represent measures of microscopic events that are not easily identified.

(2.) While it is obvious that the number of random transitions between states denoted by $C_{ij}(t_k)$, $i,j \in \{S, E, I, R\}$, are sufficient to analyze and conduct inferences in the stochastic process governed by (2.1)-(2.4), the approach to find a ML-estimate for $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ in this paper centers on observation of some components of the state $X(t_k)$ of the system in (5.1). Indeed, for populations with fixed and finite sizes $N$ defined in Subsection 2.1, the following are true: (a) the epidemic dies eventually (cf. [8]); (b) it may be possible, for reasonable sizes of the population $N$, to track and observe some components of the state $X(t_k)$, for example, the number of individuals in the susceptible and infectious states $S(t_k)$ and $I(t_k)$, respectively, can be measured at any time $t_k$, $k \geq 0$. For example, new contact tracing programs are considered effective control strategies for epidemics such as Ebola, HIV, tuberculosis, Covid-19 (cf. [4]), whenever the population size is relatively small. Thus, epidemics in small populations, where contact tracing programs are applied as control strategy, can often lead to concrete observations of the states $S(t_k)$ and $I(t_k)$.

The ML-estimation method presented in this section is based on the premises in Remark 5.1[(2.)]. That is, it is assumed that the population size $N$ is relatively small, and the states $S(t_k)$ and $I(t_k)$ are traced and observed for a finite amount of time. This strong assumption of contact tracing ultimately amounts to several of the quantities $C_{ij}(t_k)$, $i,j \in \{S, E, I, R\}$ in (2.1)-(2.4) being observed over time, which further differentiates significantly the procedure in this paper from the earlier studies [31, 39], which are based on observing the $C_{ij}(t_k)$'s. Furthermore, as in the Markov chain models in [31, 39], some components in the vector $X(t_k)$ are not observable. For example, the component $E(t_k) = \hat{e}_k$, representing those who are in the latent phase, with no visible symptoms, and during the period over which the disease is still incubating, is not an observable state. Thus, only the state $S(t_k) = \hat{s}_k \in \mathbb{Z}_+$ representing those who are healthy with no visible clinical symptoms, and the state $I(t_k) = \hat{i}_k \in \mathbb{Z}_+$ representing those with clear clinical symptoms can be considered as observable components of $X(t_k)$, at any time $t_k$, $k \geq 0$. Thus, as in [31, 39], $E(t_k) = \hat{e}_k$ will be treated as part of the missing information. However, unlike [31, 39], the EM-algorithm framework is applied here to treat the missing data.

For the traditional ML-estimation technique, the population $X(t_k)$ is partially observed over $t_k$, $k = 0, 1, 2, \ldots, T$, where the initial state $X(t_0) = \hat{x}(t_0)$ is assumed to be known. That is, a path of the collection

$$H_T = \{X(t_0), X(t_1), X(t_2), \ldots, X(t_T)\} = \{\mathfrak{S}_T, \mathfrak{E}_T, \mathfrak{I}_T\}, \tag{5.3}$$

given by

$$\hat{H}_T = \{\hat{x}(t_0), \hat{x}(t_1), \hat{x}(t_2), \ldots, \hat{x}(t_T)\} = \{\hat{\mathfrak{S}}_T, \hat{\mathfrak{E}}_T, \hat{\mathfrak{I}}_T\}, \tag{5.4}$$

is obtained, where only the values of $S(t_k)$ and $I(t_k)$ are observed, and the sub-collections $\mathfrak{S}_T, \mathfrak{E}_T, \mathfrak{I}_T$ and $\hat{\mathfrak{S}}_T, \hat{\mathfrak{E}}_T, \hat{\mathfrak{I}}_T$ contain the corresponding components $S(t_k), E(t_k), I(t_k)$ of the vector $X(t_k)$ over $t_k$, $k = 0, 1, 2, \ldots, T$.

We use the sample path $\hat{H}_T$ of the process $\{X(t_k); k = 0, 1, 2, \ldots\}$ to find the maximum likelihood estimates for the parameter $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$, applying the EM-algorithm. From this point onwards in the description of the traditional ML-estimation approach, only the SEIR model in Theorem 3.2 with parameter $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ is discussed.

If all components in $\hat{H}_T$ were observed, then it is easy to see from (5.2), (5.3), and (5.4), that the complete likelihood function of $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ would be given as follows:

$$L_1(\Theta^2|\hat{H}_T) = L_1(p, \lambda, \delta_e, \delta_r|\hat{H}_T) = P(H_T = \hat{H}_T|p, \lambda, \delta_e, \delta_r) = P(X(t_T) = \hat{x}(t_T), X(t_{T-1}) = \hat{x}(t_{T-1}), \ldots, X(t_0) = \hat{x}(t_0)|p, \lambda, \delta_e, \delta_r). \tag{5.5}$$

But, $\{X(t_k), k = 0, 1, 2, \ldots\}$ a Markov chain, and since it is assumed $X(t_0)$ is known, it is easy to see that (5.5) reduces to

$$L_1(p, \lambda, \delta_e, \delta_r|\hat{H}_T) = \prod_{k=1}^{T} P(X(t_k) = \hat{x}(t_k)|X(t_{k-1}) = \hat{x}(t_{k-1}); p, \lambda, \delta_e, \delta_r). \tag{5.6}$$

It follows from (5.6), that

$$L_1(p, \lambda, \delta_e, \delta_r|\hat{H}_T) = \prod_{k=1}^{T} P(S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k|X(t_{k-1}) = \hat{x}(t_{k-1}); p, \lambda, \delta_e, \delta_r). \tag{5.7}$$

From (5.7), the log-likelihood function is given as follows

$$l_1(\Theta^2|\hat{H}_T) = \log L_1(p, \lambda, \delta_e, \delta_r|\hat{H}_T) = \sum_{k=1}^{T} \log[P(S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k|X(t_{k-1}) = \hat{x}(t_{k-1}); p, \lambda, \delta_e, \delta_r)]. \tag{5.8}$$

Note that in equation (5.7), the state $E(t_k) = e_k$, $k = 0, 1, 2, \ldots, T$ is missing data. As a result the traditional ML-estimation becomes limited.

Also note that even with the assumption that the component $E(t_k) = e_k$, $k = 0, 1, 2, \ldots, T$ is observed in the path $\hat{H}_T$, attempting to maximize the likelihood function $L$ in (5.7) by computing the derivatives of $L$ with respect to $p, \lambda, \delta_e,$ and $\delta_r$, and setting the results to zero, leads to intractable system of equations. Thus, we apply the expectation maximization (EM) algorithm in Theorem 4.1 to resolve the limitations of the traditional ML-estimation technique to find an appropriate MLE for the parameter vector $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$.

### 5.2. Application of the hierarchical data EM-algorithm in the SEIR model

To apply the EM-algorithm in Theorem 4.1, Proposition 4.1 and Remark 4.2, we define the observable and missing data in the following.

**Definition 5.1. Observed data in the SEIR model:**
(1) From Definition 4.1[1.], the observed data $Y$ for the SEIR model is the time series data for the observable components $S(t_k)$ and $I(t_k)$ of the vector $X(t_k)$, given by the collection

$$Y = \{\hat{\mathfrak{E}}_T, \hat{\mathfrak{I}}_T\},$$ 
(5.9)

where $\hat{\mathfrak{E}}_T$, and $\hat{\mathfrak{I}}_T$ are defined in (5.4).

**Definition 5.2. Missing non-hierarchical data in the SEIR model:**

(1.) From Definition 4.1[2.], the time series for the exposed state $E(t_k)$, over the times $k = 1, 2, \ldots, T$, which is denoted by $\hat{\mathfrak{E}}_T$ and given in (5.4) is also missing data. However, the missing data $\hat{\mathfrak{E}}_T$ is not of the hierarchical data structure in Definition 4.1[2.].
(2.) Also, it is assumed that $E(t_0) = e_0$ is known.
(3.) For each $k \in \{0, 1, 2, \ldots, T\}$, it is easy to see that the feasible region $\Omega^1_{E(t_{k+1})}$ for $E(t_{k+1})$, given $E(t_k)$ from (2.8), is given by

$$\Omega^1_{E(t_{k+1})} = \{e_{k+1} \in \mathbb{Z}_+ | s_k - s_{k+1} \le e_{k+1} \le e_0 + \sum_{j=1}^{k+1}(s_{j-1} - s_j)\}.$$
(5.10)

Thus, the combined feasible region for $E(t_k)$ over $k \in \{0, 1, 2, \ldots, T\}$, denoted by $\Omega^1_E$, is given by

$$\Omega^1_E = \bigcup_{k=0}^{T} \Omega^1_{E(t_k)}.$$

To impute the missing data $E(t_k)$, over the times $k = 1, 2, \ldots, T$, to the log-likelihood function, we introduce a new notation for the super-sum over $\Omega^1_E$. Denote by

$$\sum_{\Omega^1_E}[\cdots] = \sum_{k=0}^{T} \sum_{\Omega^1_{E(t_k)}}[\cdots].$$
(5.11)

**Definition 5.3. Missing hierarchical data in the SEIR model:**
(1.) The missing hierarchical data $Z$ defined in (4.6), in the SEIR epidemic model, is given by Definitions 3.2-3.4. Indeed, for the sequence of times $t_k$, $k = 1, 2, \ldots T$ that $Y$ in (5.9) is observed, the number of contacts that a susceptible person makes in the interval $[t_k, t_{k+1})$, denoted by $Z_{t_k}$, which is defined in (3.4) is hidden information.

Also, for the given number of contacts $Z_{t_k} = n$, $n \in \mathbb{Z}_+$ that susceptible person makes in the interval $[t_k, t_{k+1})$, there are $Z_{t_k,n} = j$ contacts with infectious individuals in the population, where $j \in \{0, 1, \ldots, n\}$, and $Z_{t_k,n}$ is defined in (3.6). Moreover, the data on $Z_{t_k,n}$, $k = 1, 2, \ldots T$ is also hidden information.

In addition, given that the susceptible person makes $Z_{t_k} = n$ contacts, and $Z_{t_k,n} = j$, of the contacts are with infectious people, where $j \in \{0, 1, \ldots, n\}$, the information of the $l^{th}$ person who likely passes the infection in the interval $[t_k, t_{k+1})$ is also hidden, where information $l \in \{1, 2, \ldots, j\}$. That is, the categorical random variable $Z_{t_k,n,j} = l$, $l \in \{1, 2, \ldots, j\}$ defined in (3.8) is hidden data.

Finally, given that the susceptible person makes $Z_{t_k} = n$ contacts, and $Z_{t_k,n} = j$, of the contacts are with infectious people, where $j \in \{0, 1, \ldots, n\}$, it is easy to see that taking note of the $l^{th}$ infectious person likely passing the infection in the interval $[t_k, t_{k+1})$, where $l \in \{1, 2, \ldots, j\}$, the information of the event that the $l^{th}$ person passes the infection is also hidden. That is, the collection of Bernoulli random variables $Z_{t_k,n,j,l} = \iota$, $\iota = 0, 1$, and $l \in \{1, 2, \ldots, j\}$, defined in (3.12) is hidden data.

Hence, from above, the hierarchical missing data $Z$ in (4.6), in the SEIR epidemic model, is given by Definitions 3.2-3.5. That is,

$$Z = \{C^{(4)}, C^{(3)}, C^{(2)}, C^{(1)}\},$$
(5.12)

given in (3.5), (3.7), (3.9) and (3.12).

**Definition 5.4. Unknown parameters in the SEIR model:**
From Definition 4.1[4.], the unknown parameter is $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ for the SEIR model in Theorem 3.2.

**Remark 5.2.** Since the missing hierarchic data in the SEIR model in Definition 5.3 has four levels of hierarchy, we employ the hierarchical data EM-algorithm deduced from Theorem 4.1, Proposition 4.1 and Remark 4.2. Furthermore, explicit knowledge of the distribution of the missing hierarchical data in (5.12) is given in Definitions 3.2-3.5.

**Theorem 5.1.** *Suppose the observed data $Y$ in the SEIR epidemic model in Theorem 3.2, is as given in Definition 5.1. Let the missing non-hierarchical data $E(t_k)$ over $k = 1, 2, \ldots, T$ be as defined in Definition 5.2; let the hierarchical hidden data in the SEIR epidemic model over $k = 1, 2, \ldots, T$ be as defined in Definition 5.3. The log-likelihood function $l(\Theta^2|Y)$ of the unknown parameter $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ for the SEIR model in Theorem 3.2 satisfies the following inequality.*

$$l(\Theta^2|Y) \ge \sum_{k=1}^{T} Q_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}(\Theta^2|\Theta^2_m)$$

$$+ \sum_{k=1}^{T} R_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}(\Theta^2_m|\Theta^2_m)$$

$$+ \sum_{k=1}^{T} R_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}}(\Theta_m^2 | \Theta_m^2) + \sum_{k=1}^{T} R_{E(t_k), Z_{t_k}, Z_{t_k,n}}(\Theta_m^2 | \Theta_m^2)$$

$$+ \sum_{k=1}^{T} R_{E(t_k), Z_{t_k}}(\Theta_m^2 | \Theta_m^2) + \sum_{k=1}^{T} R_{E(t_k)}(\Theta_m^2 | \Theta_m^2), \tag{5.13}$$

*where for each $k = 1, 2, \ldots, T$, the Q-function of the E-step of the EM-algorithm is given as follows.*

$$Q_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}(\Theta^2 | \Theta_m^2) = \mathbb{E}_{E(t_k) | S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\mathbb{E}_{Z_{t_k} | S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [ \mathbb{E}_{Z_{t_k,n} | S(t_k), E(t_k), Z_{t_k}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\mathbb{E}_{Z_{t_k,n,j} | S(t_k), E(t_k), Z_{t_k}, Z_{t_k,n}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [ \mathbb{E}_{Z_{t_k,n,j,l} | S(t_k), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$l_{1,k}(\Theta^2 | Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}, S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}))]]]]] \tag{5.14}$$

*and*

$$l_{1,k}(\Theta^2 | Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}, S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}))$$
$$= \log \left\{ P\left( Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}, S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k | S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \right) \right\}. \tag{5.15}$$

*Also, the entropy terms are given as follows.*

$$R_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}(\Theta_m^2 | \Theta_m^2) = -\mathbb{E}_{E(t_k) | S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\mathbb{E}_{Z_{t_k} | S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [ \mathbb{E}_{Z_{t_k,n} | S(t_k), E(t_k), Z_{t_k}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\mathbb{E}_{Z_{t_k,n,j} | S(t_k), E(t_k), Z_{t_k}, Z_{t_k,n}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [ \mathbb{E}_{Z_{t_k,n,j,l} | S(t_k), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\log \mathbb{P}(Z_{t_k,n,j,l} | Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2)]]]]]$$

$$R_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}}(\Theta_m^2 | \Theta_m^2) = -\mathbb{E}_{E(t_k) | S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\mathbb{E}_{Z_{t_k} | S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [ \mathbb{E}_{Z_{t_k,n} | S(t_k), E(t_k), Z_{t_k}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\mathbb{E}_{Z_{t_k,n,j} | S(t_k), E(t_k), Z_{t_k}, Z_{t_k,n}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\log \mathbb{P}(Z_{t_k,n,j} | Z_{t_k}, Z_{t_k,n}, S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2)]]]]$$

$$R_{E(t_k), Z_{t_k}, Z_{t_k,n}}(\Theta_m^2 | \Theta_m^2) = -\mathbb{E}_{E(t_k) | S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\mathbb{E}_{Z_{t_k} | S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [ \mathbb{E}_{Z_{t_k,n} | S(t_k), E(t_k), Z_{t_k}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\log \mathbb{P}(Z_{t_k,n} | Z_{t_k}, S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2)]]]$$

$$R_{E(t_k), Z_{t_k}}(\Theta_m^2 | \Theta_m^2) = -\mathbb{E}_{E(t_k) | S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\mathbb{E}_{Z_{t_k} | S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} [$$

$$\log \mathbb{P}(Z_{t_k} | S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2)]]$$

*and*

$$R_{E(t_k)}(\Theta_m^2 | \Theta_m^2) = -\mathbb{E}_{E(t_k) | S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} \left[ \log \mathbb{P}(E(t_k) | S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2) \right].$$

**Proof.** The idea of the proof of Theorem 5.1 is similar to the proof of Theorem 4.1, where we apply a step-by-step approach to impute hidden data in the incomplete log-likelihood function, and average over the distributions of the hidden non-hierarchical data $E(t_k)$, $k = 1, 2, \ldots, T$, and the hidden hierarchical data $Z = \{C^{(4)}, C^{(3)}, C^{(2)}, C^{(1)}\}$, defined in Definitions 5.2-5.3, respectively. This method applies the Jensen's inequality, similarly as in the proof Theorem 4.1, to simultaneously obtain the Q-function and entropy terms in (5.13); as well as show convergence of the EM-algorithm. The complete proof of Theorem 5.1 is given in Appendix D. □

**Remark 5.3. The hierarchical data EM-algorithm in the SEIR model in Theorem 3.2:**
Similarly to Remark 4.3, it is obvious from Theorem 5.1 that applying the EM-algorithm to find MLE for the unknown parameter $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ for the SEIR model in Theorem 3.2, consists of the following E- & M- steps.

E-step: the E-step of the EM-algorithm consists of finding the Q-function

$$Q(\Theta^2 | \Theta_m^2) = \sum_{k=1}^{T} Q_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}(\Theta^2 | \Theta_m^2), \tag{5.16}$$

where $Q_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}(\Theta^2 | \Theta_m^2)$ is defined in (5.14), and

M-step: the M-step consists of finding $\Theta^2_{max}$ below.

$$\Theta^2_{max} = \underset{\Theta^2}{argmax} \left[ \sum_{k=1}^{T} Q_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta^2_m) \right] = \underset{\Theta^2}{argmax} \left[ Q(\Theta^2|\Theta^2_m) \right]. \tag{5.17}$$

In the next set of results we apply the steps for the EM-algorithm in Remark 5.3 to find MLE for the parameter $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ for the SEIR model in Theorem 3.2. First, we present a separate result for the E-step of the EM-algorithm.

**Theorem 5.2.** *Suppose the assumptions of Theorem 5.1 are satisfied, that is, the log-likelihood function for $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ satisfies (5.13), where $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ is the parameter vector for the SEIR model in Theorem 3.2 and $\Theta^2_m = (p^m, \lambda^m, \delta_e^m, \delta_r^m)$ is the $m^{th}$ step estimate for $\Theta^2$ in the EM-algorithm characterized in Remark 5.3, where $m \in \mathbb{Z}_+$. For each $k \in \{1, 2, 3, \ldots, T\}$, let the missing non-hierarchical data $E(t_k)$ over $k = 1, 2, \ldots, T$ be as defined in Definition 5.2; let the hierarchical hidden data in the SEIR epidemic model over $k = 1, 2, \ldots, T$ be as defined in Definition 5.3. Define the following random variable that is independent of the parameter $\Theta^2$, but depends on the other random variables $E(t_{k-1}), E(t_k)Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1$*

$$\mathbb{K}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1)$$
$$= \log\left[ \binom{\hat{s}_{k-1}}{\hat{s}_{k-1} - \hat{s}_k} \right] + \log\left[ \binom{E(t_{k-1})}{E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k} \right]$$
$$+ \log\left[ \binom{\hat{i}_{k-1}}{\hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k} \right] + \log\left( \frac{1}{(Z_{t_k})!} \right)$$
$$+ \log\left[ \binom{Z_{t_k} = n}{Z_{t_k,n} = j} \left( \alpha_{t_k} \right)^{Z_{t_k,n}} \left( 1 - \alpha_{t_k} \right)^{Z_{t_k} - Z_{t_k,n}} \right] + \log\left( \frac{1}{Z_{t_k,n}} \right).$$

*Also define the following random variable that dependents on both the parameter $\Theta^2$, and the other random variables $E(t_{k-1}), E(t_k)Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1$*

$$\mathbb{J}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; \Theta^2)$$
$$= \left( \hat{s}_{k-1} - \hat{s}_k \right) \log p + \left( \hat{s}_k \right) \log(1-p)$$
$$+ \left( E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k \right) \log\left( P_{EI}(t_{k-1}) \right) + \left( E(t_k) - [\hat{s}_{k-1} - \hat{s}_k] \right) \log\left( 1 - P_{EI}(t_{k-1}) \right)$$
$$+ \left( \hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k \right) \log\left( P_{IR}(t_{k-1}) \right)$$
$$+ \left( \hat{i}_k - [E(t_{k-1}) - E(t_k)] - [\hat{s}_{k-1} - \hat{s}_k] \right) \log\left( 1 - P_{IR}(t_{k-1}) \right)$$
$$+ Z_{t_k} \log(\lambda \Delta t) - \lambda \Delta t + \log(p). \tag{5.18}$$

*The Q-function $Q(\Theta^2|\Theta^2_m)$ in (5.16) of the EM-algorithm in Remark 5.3, is given as follows. Let $Q_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta^2_m)$ be as defined in (5.14). For each $k \in \{1, 2, 3, \ldots, T\}$, the following conditional expected value term is defined.*

$$\mathbb{E}_{Z_{t_k,n,j,l}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \left[ l_{1,k}(\Theta^2|Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}, S(t_k), E(t_k), I(t_k); \right.$$
$$\left. S(t_{k-1}), I(t_{k-1})) \right]$$
$$= \left[ \mathbb{K}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1) + \mathbb{J}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; \Theta^2) \right] \times$$
$$\times p^m.$$

*Thus, from (5.14), it follows that*

$$Q_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta^2_m) = Q^1_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta^2_m) + Q^2_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta^2_m),$$

*where*

$$Q^1_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta^2_m) = \mathbb{E}_{E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \Big[$$
$$\mathbb{E}_{Z_{t_k}|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \Big[ \mathbb{E}_{Z_{t_k,n}|S(t_k),E(t_k),Z_{t_k},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \Big[$$
$$\mathbb{E}_{Z_{t_k,n,j}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \Big[ \mathbb{E}_{Z_{t_k,n,j,l}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \Big[$$
$$\mathbb{K}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1) \Big] \Big] \Big] \Big] \Big] \tag{5.19}$$

*and*

$$Q^2_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta^2_m) = \mathbb{E}_{E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \Big[$$
$$\mathbb{E}_{Z_{t_k}|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \Big[ \mathbb{E}_{Z_{t_k,n}|S(t_k),E(t_k),Z_{t_k},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \Big[$$
$$\mathbb{E}_{Z_{t_k,n,j}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \Big[ \mathbb{E}_{Z_{t_k,n,j,l}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta^2_m} \Big[$$
$$\mathbb{J}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; \Theta^2) \Big] \Big] \Big] \Big] \Big], \tag{5.20}$$

*and from (5.16), it follows that the Q-function is given by*

$$Q(\Theta^2|\Theta_m^2) = \sum_{k=1}^{T} Q^1_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta_m^2) + \sum_{k=1}^{T} Q^2_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta_m^2). \tag{5.21}$$

**Proof.** The idea of the proof of Theorem 5.2 is to explicitly compute the complex conditional expectation expression of the Q-function $Q(\Theta^2|\Theta_m^2)$ in (5.16) in a step-by-step manner, by averaging over the nested conditional distributions of the hierarchical data $Z_{t_k} = n, Z_{t_k,n} = j, Z_{t_k,n,j} = l$, $\forall k = 1, 2, \dots T$, given in Definitions 3.2-3.5&5.3, and indexed by the $m^{th}$ step parameter estimate $\Theta_m^2$; and also over the distributions of the non-hierarchical data $E(t_k) = \hat{e}_k$, $\forall k = 1, 2, \dots T$. Some basic probability rules, algebraic manipulations and the discrete probability distributions in (3.14) are used. The complete proof of Theorem 5.2 is given in Appendix E $\quad\square$

**Remark 5.4.** From Theorem 5.2, it is easy to see that the M-step in (5.17) reduces to finding $\Theta_{max}^2$ below.

$$\Theta_{max}^2 = \underset{\Theta}{argmax}\left[ \sum_{k=1}^{T} Q^2_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta_m^2) \right]$$

where $Q^2_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta_m^2)$ is defined in (5.20). This is because $Q^1_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta_m^2)$ in (5.19) is independent of the parameter $\Theta^2$. The M-step of the EM-algorithm in Remark 5.3 is given in the next result.

**Theorem 5.3.** *Suppose the E-step of the EM-algorithm which consists of finding the Q-function is as stated in Theorem 5.2, where the Q-function is given (5.21). Let $\tilde{\mathbb{E}}_{k;\Theta_m^2}[\cdots]$ denote the following complex conditional expectation operator.*

$$\tilde{\mathbb{E}}_{k;\Theta_m^2}[\cdots] = \mathbb{E}_{E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\Big[$$
$$\mathbb{E}_{Z_{t_k}|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\Big[\mathbb{E}_{Z_{t_k,n}|S(t_k),E(t_k),Z_{t_k},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\Big[$$
$$\mathbb{E}_{Z_{t_k,n,j}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\Big[$$
$$\cdots]]]]. \tag{5.22}$$

*Assume further that for each $m \geq 0$, $\Theta_m^2$ is the $m^{th}$ step estimate of the parameter $\Theta^2 = (p, \lambda, \delta_e, \delta_r)^T$ of the SEIR model in Theorem 3.2. The M-step of the EM-algorithm consists of finding*

$$\Theta_{m+1}^2 = \Theta_{max}^2 = \underset{\Theta}{argmax}\left[Q(\Theta^2|\Theta_m^2)\right],$$

*where $Q(\Theta^2|\Theta_m^2)$ is given in (5.21). Moreover,*

$$\Theta_{m+1}^2 = \Theta_{max}^2 = \begin{pmatrix} \dfrac{\sum_{k=1}^{T}[\hat{s}_{k-1}-\hat{s}_k+1]}{\sum_{k=1}^{T}[\hat{s}_{k-1}-\hat{s}_k+1]+\sum_{k=1}^{T}[\hat{s}_k]} \\ \dfrac{\sum_{k=1}^{T}\tilde{\mathbb{E}}_{k;\Theta_m^2}[Z_{t_k}]}{\left(\sum_{k=1}^{T}[\Delta t]\right)} \\ \dfrac{1}{\Delta t}\log[A_e+1] \\ \dfrac{1}{\Delta t}\log[B_r+1] \end{pmatrix}$$

*where*

$$A_e = \frac{\sum_{k=1}^{T}\tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1})-E(t_k)\right)\right] + \sum_{k=1}^{T}\left[\left(\hat{s}_{k-1}-\hat{s}_k\right)\right]}{\sum_{k=1}^{T}\tilde{\mathbb{E}}_{k;\Theta_m^2}\left[E(t_k)\right] - \sum_{k=1}^{T}\left[\left(\hat{s}_{k-1}-\hat{s}_k\right)\right]},$$

*and*

$$B_r = \frac{\sum_{k=1}^{T}\left[\left(\hat{i}_{k-1}-\hat{i}_k\right)\right] + \sum_{k=1}^{T}\tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1})-E(t_k)\right)\right] + \sum_{k=1}^{T}\left[\left(\hat{s}_{k-1}-\hat{s}_k\right)\right]}{\sum_{k=1}^{T}\left[\hat{i}_k\right] - \sum_{k=1}^{T}\tilde{\mathbb{E}}_{k;\Theta_m^2}\left[E(t_{k-1})-E(t_k)\right] - \sum_{k=1}^{T}\left[\hat{s}_{k-1}-\hat{s}_k\right]}.$$

**Proof.** The basic idea of the proof of Theorem 5.3 is simply to apply the tools of differential calculus, and matrix algebra to find a value for the parameter $\Theta^2$ that maximizes the Q-function in (5.21). The complete proof of Theorem 5.3 is given in Appendix F. $\quad\square$

From Theorem 5.3 and Algorithm 4.1, we state in the following, the complete steps of the EM-algorithm for finding the MLE of the parameter $\Theta^2$ for the SEIR epidemic model in Theorem 3.2.

**Algorithm 5.1. The EM-algorithm for the SEIR epidemic model in Theorem 3.2:**

1. Let $m = 0$ and $\Theta_m^2 = \Theta_0^2$ be an initial guess for $\Theta^2$.
2. Find the $m^{th}$ step estimate $\Theta_{m+1}^2$, $m \geq 1$ iteratively, where

$$\Theta_{m+1}^2 = \begin{pmatrix} \dfrac{\sum_{k=1}^{T}[\hat{s}_{k-1}-\hat{s}_k+1]}{\sum_{k=1}^{T}[\hat{s}_{k-1}-\hat{s}_k+1]+\sum_{k=1}^{T}[\hat{s}_k]} \\ \dfrac{\sum_{k=1}^{T}\tilde{\mathbb{E}}_{k;\Theta_m^2}[Z_{t_k}]}{\left(\sum_{k=1}^{T}[\Delta t]\right)} \\ \dfrac{1}{\Delta t}\log[A_e^m+1] \\ \dfrac{1}{\Delta t}\log[B_r^m+1] \end{pmatrix} \tag{5.23}$$

where

$$A_e^m = \frac{\sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1}) - E(t_k)\right)\right] + \sum_{k=1}^T \left[\left(\hat{s}_{k-1} - \hat{s}_k\right)\right]}{\sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[E(t_k)\right] - \sum_{k=1}^T \left[\left(\hat{s}_{k-1} - \hat{s}_k\right)\right]},$$

and

$$B_r^m = \frac{\sum_{k=1}^T \left[\left(\hat{i}_{k-1} - \hat{i}_k\right)\right] + \sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1}) - E(t_k)\right)\right] + \sum_{k=1}^T \left[\left(\hat{s}_{k-1} - \hat{s}_k\right)\right]}{\sum_{k=1}^T \left[\hat{i}_k\right] - \sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[E(t_{k-1}) - E(t_k)\right] - \sum_{k=1}^T \left[\hat{s}_{k-1} - \hat{s}_k\right]}. \qquad (5.24)$$

3. Update $\Theta_m$ in (5.23) with $\Theta_{m+1}$, and repeat step 2 until $\Theta_{m+1}^2$ stops noticeably changing.

**Remark 5.5.** We are left to find expressions for the expected values of the missing data at random $\sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1}) - E(t_k)\right)\right]$, $\sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[Z_{t_k}\right]$, and $\sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[E(t_k)\right]$ in (5.23)-(5.24).

1. It is easy to see from the model equations (2.1)-(2.4), that we iteratively obtain the following

$$[E(t_{k-1}) - E(t_k)] = [S(t_{k-1}) - S(t_k)] + [I(t_{k-1}) - I(t_k)] - [C_{IR}(t_{k-1})],$$

$$E(t_k) = E(t_0) + \sum_{j=1}^k \left([S(t_{j-1}) - S(t_j)] + [I(t_{j-1}) - I(t_j)]\right) - \sum_{j=1}^k [C_{IR}(t_{j-1})]. \qquad (5.25)$$

Recall Definition 5.2[2.], that $E(t_0) = e_0$ is assumed known. Furthermore, from (3.14)-(3.16), $C_{IR}(t_k) \sim Binomial(I(t_k), P_{IR}(t_k))$. Thus, applying the conditional expected value operator $\tilde{\mathbb{E}}_{k;\Theta_m^2}$ defined in (5.22), to (5.25), we obtain the following.

$$\begin{aligned}
\sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1}) - E(t_k)\right)\right] &= \sum_{k=1}^T [s_{k-1} - s_k] + \sum_{k=1}^T [i_{k-1} - i_k] - \sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}[C_{IR}(t_{k-1})] \\
&= \sum_{k=1}^T [s_{k-1} - s_k] + \sum_{k=1}^T [i_{k-1} - i_k] - \sum_{k=1}^T [I(t_{k-1})P_{IR}(t_{k-1})] \\
&= \sum_{k=1}^T [s_{k-1} - s_k] + \sum_{k=1}^T [i_{k-1} - i_k] - \sum_{k=1}^T \left[i_{k-1}\left(1 - e^{\delta_r^m \Delta t}\right)\right].
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[E(t_k)\right] &= \sum_{k=1}^T e_0 + \sum_{k=1}^T \sum_{j=1}^k \left([s_{j-1} - s_j] + [i_{j-1} - i_j]\right) - \sum_{k=1}^T \sum_{j=1}^k \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[C_{IR}(t_{j-1})\right] \\
&= \sum_{k=1}^T e_0 + \sum_{k=1}^T \sum_{j=1}^k \left([s_{j-1} - s_j] + [i_{j-1} - i_j]\right) - \sum_{k=1}^T \sum_{j=1}^k \left[i_{j-1}\left(1 - e^{\delta_r^m \Delta t}\right)\right].
\end{aligned}$$

2. We use the following argument to find a plausible expression for $\sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}[Z_{t_k}]$. From Definition 3.2, it can be deduced that when $\Theta_m^2$ and all other terms in the conditional expected value operator $\tilde{\mathbb{E}}_{k;\Theta_m^2}$ in (5.22) are given, $Z_{t_k} \sim Poisson(\lambda_{1,k}^m)$, where $\lambda_{1,k}^m$ is selected such that it is proportional to $\lambda^m \Delta t$, i.e. $\lambda_{1,k}^m \propto \lambda^m \Delta t$. Hence, from (5.22), the conditional expected value sum

$$\sum_{k=1}^T \tilde{\mathbb{E}}_{k;\Theta_m^2}[Z_{t_k}] \propto \sum_{k=1}^T [\lambda_{1,k}^m]. \qquad (5.26)$$

## 6. A theoretical example for the EM-algorithm in the SEIR epidemic model

In this section, we present a theoretical numerical example on influenza epidemics, to explore the application of the EM-algorithm given in Algorithm 5.1. Since this study is focused on introducing the multi-level hierarchical data EM-algorithm, only the parameter estimation for the SEIR model in Theorem 3.2 is presented. That is, we find a ML-estimate (MLE) for $\Theta^2 = (p, \lambda, \delta_e, \delta_r)^T$. Note that this example is not intended to be informative about influenza epidemics, but simply to exhibit the applicability of the Algorithm 5.1. A more proper study of the statistical inferences in influenza SEIR models, with real data, and applying the EM-algorithm will appear elsewhere.

### 6.1. Testing the EM-algorithm on simulated influenza epidemic data

The influenza data [1, 5] for the state of Georgia (GA), U.S.A. for the $2017 - 2018$ flu season is used to illustrate the EM-algorithm presented in Algorithm 5.1. The flu data consists of 52 weeks of measurements collected by the WHO and the *NREVSS (National Respiratory And Enteric Virus Surveillance System)* in collaboration with laboratories networking with CDC. Using the data, the ML-estimate of the parameter $\Theta^2 = (p, \lambda, \delta_e, \delta_r)^T$ for the SEIR Markov chain $\{X(t_k; k \geq 0)\}$, with governing equations (2.1)-(2.4), and transition probabilities in Theorem 3.2. Note that this example is hypothetical, and the following assumptions are used.

(1) The population consists of $n = 10519475$ individuals, similar to the size of the population of GA, USA in 2018 [5]. Furthermore, the population size is fixed over the 52 weeks. (2) The variant of influenza virus considered is *type A virus*; the weekly new infections are exhibited in Fig. 2 obtained from [1]. All reported cases are assumed to be infectious, i.e. the exposed class is not observed. It is also assumed that there is no readily supply of

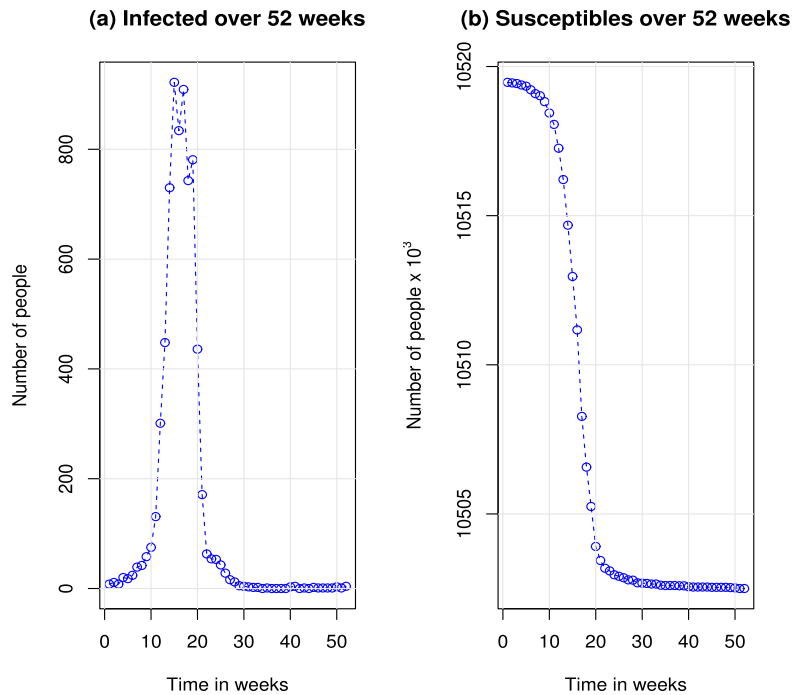**(a) Infected over 52 weeks** **(b) Susceptibles over 52 weeks**



Fig. 2. Shows a time-series over a period of 52 weeks for the number of people susceptible (b) and infected (a) with influenza type A virus.

the vaccine for the virus during this period. Moreover, the estimates for the susceptible population comparable to GA in 2018 are summarized in Fig. 2.

(3) It is assumed that the incubation and infectious periods, $T_1$ and $T_2$, respectively, are random, and beyond this period, the exposed person becomes infectious, while the infectious person fully recovers and removed, respectively. New births are not considered. Hence, from (2)-(3) it is easy to see that the initial susceptible population decreases continuously over the 52 weeks considered because of the continuous infection and removal processes. Hence, the number of infectious individuals present in any week are those infected at the beginning of the week.

(4) It is also assumed for simplicity that, initially only one person is exposed, that is, $e_0 = 1$ is exposed (cf. [39]); 8 people are initially infectious, and nobody is assumed to be removed in the population initially.

From the above assumptions (1)-(4), and the data in [1, 5] exhibited in Fig. 2, the following are obtained by employing the Algorithm 5.1 to the influenza data.

Note that employing Algorithm 5.1 to the influenza data to find the ML- estimate $\Theta^2_{m+1}$ in the $(m+1)^{th}$ step is basically substituting values for $I(t_k) = i_k$, $S(t_k) = s_k$ over $k = 0, 1, 2, \ldots, T$, to the formulas in (5.23)-(5.24) and Remark 5.5, where in the influenza data in Fig. 2, the weekly counts of infectious cases are values for $I(t_k) = i_k$, and the values for $S(t_k) = s_k$ are the weekly count for the susceptible state. Furthermore, the time $k = 0, 1, 2, \ldots, T$ represents the 52 weeks.

Denote the final step ML-estimate of $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ by $\hat{\Theta}^2 = (\hat{p}, \hat{\lambda}, \hat{\delta}_e, \hat{\delta}_r)$. It is easy to see that the values of $\hat{\Theta}^2 = (\hat{p}, \hat{\lambda}, \hat{\delta}_e, \hat{\delta}_r)$ are obtained when the algorithm converges. Indeed, employing the Algorithm 5.1 to the influenza data in Fig. 2, leads to the estimates in Table 1.

Note from Algorithm 5.1 that the ML-estimate for $\lambda$ is obtained iteratively by applying the steps in the algorithm, until convergence, utilizing the formula for $\lambda^{m+1}$ in (5.23) and (5.26), that is,

$$\lambda^{m+1} = \frac{\sum_{k=1}^{T} \bar{\mathbb{E}}_{k;\Theta^2_m}[Z_{t_k}]}{\left(\sum_{k=1}^{T}[\Delta t]\right)}.$$

In addition, the algorithm converges after 41 steps, to the ML-estimate for $\lambda$, denoted by $\hat{\lambda} = \lambda^{41} = 2.0/day$. The convergence to $\hat{\lambda} = \lambda^{41} = 2.0/day$ is shown in Fig. 3.

Also, the ML-estimate for $p$ is obtained iteratively by applying the steps in the algorithm, until convergence, utilizing the formula for $p^{m+1}$ in (5.23) and (5.26), that is,

$$p^{m+1} = \frac{\sum_{k=1}^{T} \left[\hat{s}_{k-1} - \hat{s}_k + 1\right]}{\sum_{k=1}^{T} \left[\hat{s}_{k-1} - \hat{s}_k + 1\right] + \sum_{k=1}^{T} \left[\hat{s}_k\right]}.$$

The algorithm converges to the ML-estimate for $p$, denoted by $\hat{p} = 0.3176$ per week = 0.0454 per day.

The ML-estimates for $\delta_e$ and $\delta_r$ are also iteratively obtained by applying the steps in the algorithm, until convergence, utilizing the formula for $\delta_e^{m+1}$ and $\delta_r^{m+1}$ in (5.23) and (5.26), that is,

$$\delta_e^{m+1} = \frac{1}{\Delta t} \log\left[A_e^m + 1\right],$$

$$\delta_r^{m+1} = \frac{1}{\Delta t} \log\left[B_r^m + 1\right],$$
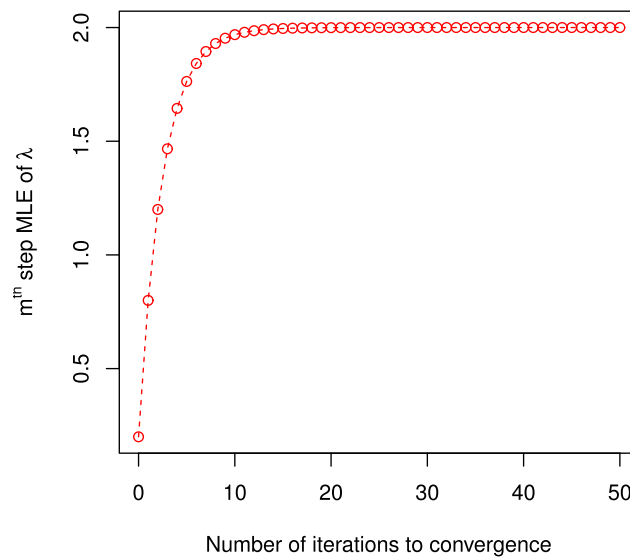
where

**Fig. 3.** Shows the $m^{th}$ step estimate for the MLE over the number of steps $m = 0, 1, 2, 2, \dots$. The initial estimate for $\lambda^0 = 0.2$. Beyond the 41st step, the EM-algorithm converges to the MLE of $\hat{\lambda} = 2.0/day$.
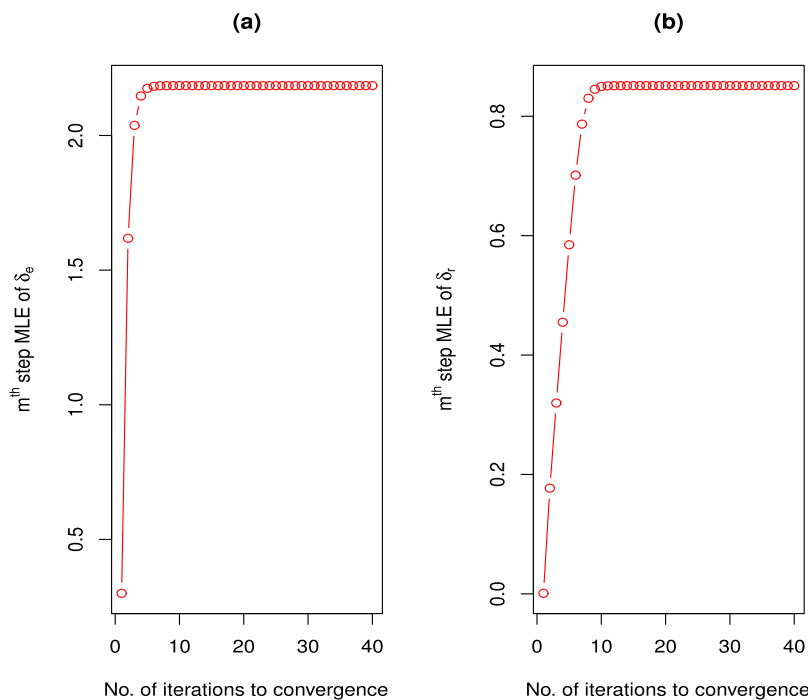
**(a)**                                                          **(b)**



**Fig. 4.** Shows the $m^{th}$ step estimate of the MLE of the parameters $\delta_e$ in (a), and $\delta_r$ in (b), over the number of steps $m = 0, 1, 2, 3, \dots$. The initial estimate for $\delta_e^0 = \delta_r^0 = 0.01$. The convergence to $\hat{\delta}_e = 2.185118$ occurs after 13 iterations; the convergence to $\hat{\delta}_r = 0.8513226$ occurs after 18 iterations.

$$A_e^m = \frac{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1}) - E(t_k)\right)\right] + \sum_{k=1}^{T}\left[\left(\hat{s}_{k-1} - \hat{s}_k\right)\right]}{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[E(t_k)\right] - \sum_{k=1}^{T}\left[\left(\hat{s}_{k-1} - \hat{s}_k\right)\right]},$$

and

$$B_r^m = \frac{\sum_{k=1}^{T}\left[\left(\hat{i}_{k-1} - \hat{i}_k\right)\right] + \sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1}) - E(t_k)\right)\right] + \sum_{k=1}^{T}\left[\left(\hat{s}_{k-1} - \hat{s}_k\right)\right]}{\sum_{k=1}^{T}\left[\hat{i}_k\right] - \sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[E(t_{k-1}) - E(t_k)\right] - \sum_{k=1}^{T}\left[\hat{s}_{k-1} - \hat{s}_k\right]}.$$

It is easy to see from Fig. 4 that the algorithm converges to the ML-estimates for $\delta_e$ and $\delta_r$, respectively, given by $\hat{\delta}_e = 2.185$ per week and $\hat{\delta}_r = 0.851$ per week. Moreover, the convergence to $\hat{\delta}_e = 2.185118$ occurs after 13 iterations; the convergence to $\hat{\delta}_r = 0.8513226$ occurs after 18 iterations.

The ML-estimates $\hat{\delta}_e = 2.185$ per week and $\hat{\delta}_r = 0.851$ per week, respectively, for $\delta_e$ and $\delta_r$ are plausible. Indeed, these estimates imply that the estimate for the mean incubation period $\mu_{T_1} = \mathbb{E}[T_1] = \frac{1}{\delta_e}$ for the influenza virus is $\hat{\mu}_{T_1} = \frac{1}{\hat{\delta}_e} = 0.458\text{weeks} = 3.2\text{days}$. Also, the estimate for the mean infectious period or mean influenza contagious period, $\mu_{T_2} = \mathbb{E}[T_2] = \frac{1}{\delta_r}$, for an individual in this population, without treatment, is

**Table 1.** Shows the MLE for the parameters $(p, \lambda, \delta_e, \delta_r)$ obtained for the influenza exhibited in Fig. 2. A measure of variation using the standard deviation (sd) is also given for each estimate.

| Parameter | Definition | MLE |
|---|---|---|
| $p$ | transmission rate/ contact/day | 0.0454 |
| $\lambda$ | Mean number of contacts/ day | 2.0 (see Fig. 3) |
| $\beta$ | transmission rate/day | $\hat{p}\hat{\lambda} = 0.0907 (sd = 0.081)$ |
| $\frac{1}{\delta_e}$ | Mean incubation period (days) | $3.2 (sd = 0.307)$ |
| $\frac{1}{\delta_r}$ | Mean infectious period (days) | $8.2 (sd = 0.446)$ |

$\hat{\mu}_{T_2} = \frac{1}{\delta_r} = 1.175$ weeks $= 8.2$ days. Observe from the sources [2, 3] that these estimates for $\mu_{T_1} = \mathbb{E}[T_1]$ and $\mu_{T_2} = \mathbb{E}[T_2]$ fall in the expected range of values for these parameters, for typical influenza epidemics. That is, the estimate $\hat{\mu}_{T_1} = 3.2$ days lies in the expected range of 1- 4 days (cf. [2, 3]) for symptoms of influenza to appear; the estimate $\hat{\mu}_{T_2} = 8.2$ days lies in the expected range of values of 7-10 days (cf. [2, 3]), for an infected person to remain contagious. The ML-estimates for $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ are summarized in the Table 1. Note that finding an analytical measure of variation for the estimators $\Theta^2_m, m \geq 1$ in Algorithm 5.1, from a given sample path of the stochastic process, $Y$, given in Definition 5.1, is not trivial because of the high interdependence of the state of the system over time. However, the following approach was used to estimate the variances/standard deviation (sd) of the ML-estimates $\hat{\Theta}^2 = (\hat{p}, \hat{\lambda}, \hat{\delta}_e, \hat{\delta}_r)$, given in Table 1.

Using the values of the MLEs in Table 1 in place of the parameters $(p, \lambda, \delta_e, \delta_r)$ in the stochastic model $\{X(t_k; k \geq 0)\}$ in Theorem 3.2, a thousand sample paths for the stochastic process were generated in a similar manner as the example in [Section 3.3, Fig. B.6]. For every sample path generated over time $t_k, k = 0, 1, 2, \ldots, T = 500$, the estimates $\hat{\Theta}^2 = (\hat{p}, \hat{\lambda}, \hat{\delta}_e, \hat{\delta}_r)$ were obtained by applying Algorithm 5.1. The sample variances for the MLEs $\hat{\Theta} = (\hat{p}, \hat{\lambda}, \hat{\delta}_e, \hat{\delta}_r)$ were estimated from the sample variances of the 1000 observations for $\hat{\Theta}^2 = (\hat{p}, \hat{\lambda}, \hat{\delta}_e, \hat{\delta}_r)$ from the 1000 sample paths. The standard deviation (sd) of the estimates are given in Table 1. The choice for $T = 500$ is guided by the existence of a quasi-stationary ergodic distribution for the process over sufficiently long time.

Indeed, it is easy to see from Theorem 3.2 that the stochastic model $\{X(t_k; k \geq 0)\}$ has a finite state space $\Omega_X(t_k), k \geq 0$. Moreover, every state is accessible from each other. In addition, self-loops are possible in the Markov chain, since the chance that the state remains the same over consecutive discrete times is positive. Therefore, the Markov chain is positive recurrent and aperiodic. Also, since the total population size remains constant, i.e. $N(t_k) = S(t_k) + E(t_k) + I(t_k) + R(t_k) = N > 0$, the epidemic ultimately ends (cf. [8]), and the process is absorbed in a state, where $\lim_{k \to \infty} I(t_k) = 0$. See [8] for more details. Thus, for the process $\{X(t_k; k \geq 0)\}$, there exists an ergodic quasi-stationary distribution.

## 7. Conclusion

In this study, a general model for multi-level hierarchical data (HD) is presented for discrete data. The EM-algorithm scheme to find ML-estimates for the parameters of the system is derived, where the observed sample $Y$ is either a finite sequence of independent and identically distributed random variables, or a sequence of dependent and identically distributed random variables (such as a time series in a stochastic process); the missing data in the system follows the multi-level HD structure derived. DTMC SEIR epidemic models are used to formulate examples for the multi-level HD. Moreover, the multi-level HD EM-algorithm derived is employed to find ML-estimates of the parameters of the DTMC SEIR model, where missing data are of two different categories, namely, (1) part of the missing data is non-hierarchical in structure, and the other part of the missing data is (2) HD composed of hidden components in the disease transmission rate of the DTMC SEIR model.

Simulated data for influenza from the state of Georgia (GA), USA, is used to create a theoretical study for an influenza epidemic, where the DTMC SEIR model is applied, and the HD EM-algorithm is employed to find ML-estimates for the model parameters. Note, the choice for the DTMC SEIR model for exhibiting the steps of the HD EM-algorithm is simply for convenience. Indeed, the scenarios for employing the HD EM-algorithm to find ML-estimates are enormous. However, where other types of estimators apart from ML-estimates are desired, it is easy to see that the HD EM-algorithm is a very complex choice for inferring parameters in (2.1)-(2.4), because of the complex nature of the hidden dynamic components in the disease transmission rate of the DTMC SEIR model. Other less complex alternatives for estimating the parameters include Bayesian methods, such as, MCMC algorithm and Gibbs sampling (cf. [39]). Other limitations are given in Remark 5.1.

## Declarations

### Author contribution statement

### Funding statement

### Data availability statement

No data was used for the research described in the article.

### Declaration of interests statement

The authors declare no competing interests.

*Additional information*

No additional information is available for this paper.

## Appendix A. Proof of Theorem 3.1

**Proof.** Define an indicator random variable $Y(t_k)$ to represent the event that a susceptible person gets infected in the interval $[t_k, t_{k+1})$. That is, $Y(t_k) = 1$, when the susceptible person is infected and $Y(t_k) = 0$, otherwise. It is easy to see from Definitions 3.1-3.4 that

$$
\begin{aligned}
P_{SE}(t_k) &= P(Y(t_k) = 1) \\
&= \sum_{n=0}^{\infty} \sum_{j=0}^{n} \sum_{l=l}^{j} P(Y(t_k) = 1, Z_{t_k,n,j} = l, Z_{t_k,n} = j, Z_{t_k} = n), \\
&= \sum_{n=0}^{\infty} \sum_{j=0}^{n} \sum_{l=l}^{j} P(Y(t_k) = 1 | Z_{t_k,n,j} = l, Z_{t_k,n} = j, Z_{t_k} = n) \times \\
&\quad \times P(Z_{t_k,n,j} = l | Z_{t_k,n} = j, Z_{t_k} = n) \times P(Z_{t_k,n} = j | Z_{t_k} = n) \times P(Z_{t_k} = n).
\end{aligned}
\tag{A.1}
$$

From Definitions 3.1&3.4, observe that $P(Y(t_k) = 1 | Z_{t_k,n,j} = l, Z_{t_k,n} = j, Z_{t_k} = n) = p$. The distribution of $P(Z_{t_k,n,j} = l | Z_{t_k,n} = j, Z_{t_k} = n)$ is given by Definition 3.4; the distribution of $P(Z_{t_k,n} = j | Z_{t_k} = n)$ is given by Definition 3.3; the distribution of $P(Z_{t_k} = n)$ is given by Definition 3.2. Thus, (A.1) simplifies to (3.13). □

## Appendix B. Numerical results for the SEIR Markov chain models

In this section, we apply a theoretical approach to validate the performance of the Markov chain model $\{X(t_k) : k \geq 0\}$ with transition probabilities in Theorem 3.2. For selected values of $p$, $\lambda$, $\delta_e$ and $\delta_e$, multiple sample paths or time-series for the process $\{X(t_k) : k \geq 0\}$ are generated; and different sample statistics are analyzed to determine whether the process leads to plausible predictions for the SEIR disease epidemic. Note, in Section 6, some ideas in this section will be used to estimate a measure of variation for the estimators derived in this paper.

Three sample paths are depicted in Fig. B.5 (contrasted with red, blue and purple lines) for each state $S, E, I, R$ of the SEIR Markov chain epidemic model $\{X(t_k) : k \geq 0\}$ in Theorem 3.2. The conditions utilized are: $p = 0.0055$, and $\lambda = 10$; the average incubation $T_1$ and infectious $T_2$ periods are respectively, $E(T_1) = 10$ and $E(T_2) = 20$. Furthermore, the initial conditions used are: $S(t_0) = 200,000$, $E(t_0) = 500$, $I(t_0) = 1000$, and $R(t_0) = 0$. Based on 1000 sample realizations for the states $S, E, I, R$ of the process $\{X(t_k) : k \geq 0\}$ at the time $t_{40}$, histograms are given in Fig. B.6. Also, the intervals for the populations means of the states $S, E, I, R$ at time $t_{40}$, at the 95% confidence level are respectively, $197783.9 < \mathbb{E}[S(t_{40})] < 197794.2$, $607.8549 < \mathbb{E}[E(t_{40})] < 611.7331$, $1112.502 < \mathbb{E}[I(t_{40})] < 1118.106$ and $1983.614 < \mathbb{E}[R(t_{40})] < 1988.112$. Note $\mathbb{E}[.]$ is the expectation operator.

Fig. B.5 shows that infection rises in the population over time for all disease states $E$ and $I$. The rise in infection in the population is matched by a decline in the susceptible state $S$ overtime, and a rise in recovery $R$, overtime. Fig. B.6 shows summaries for the distributions of the states $S, E, I, R$ at the $40^{th}$ time-step $t_{40}$ for the stochastic process. Moreover, inferring the populations for $S, E, I, R$ at the $40^{th}$ time-step $t_{40}$ using the 95% confidence intervals, it is easy to see that there are significantly more people in the $S$ and $R$ states than in the disease states $E$ and $I$. Although, there are relatively more individuals in the $E$ and $I$ at the $40^{th}$ time-step $t_{40}$ than at $t_0$, the disease is not yet very aggressive to the $S$ state by the time $t_{40}$.

## Appendix C. Proof of Theorem 4.1
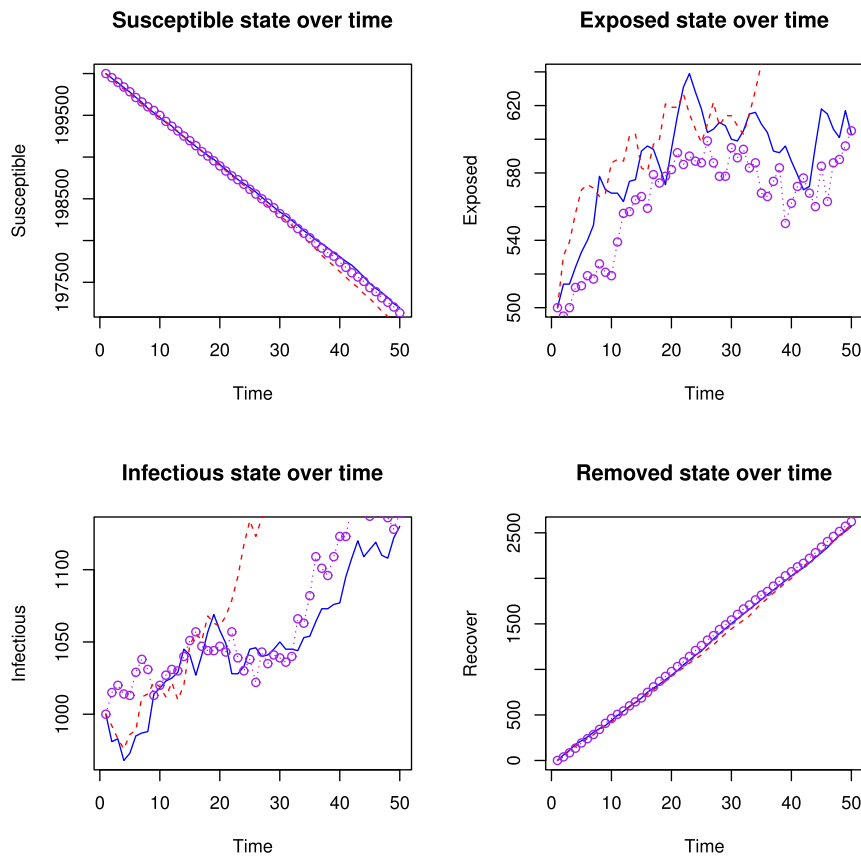
**Proof.** From Definition 4.1, to average over the missing component $C^{(3)}$, the log-likelihood function of the incomplete data is expressed as follows.

$$
\begin{aligned}
l(\Theta | Y) &= \log[L(\Theta | Y)] = \log[P(Y | \Theta)] \\
&= \log[P(Y | \Theta)] = \log \sum_{Z_{k_3}} \left[ P(Y, Z_{k_3} | \Theta) \right] \\
&= \log \sum_{Z_{k_3}} \left[ P(Y, Z_{k_3} | \Theta) \right] \times \frac{P(Z_{k_3} | Y, \Theta_m)}{P(Z_{k_3} | Y, \Theta_m)} \\
&= \log \mathbb{E}_{Z_{k_3} | Y, \Theta_m} \left[ \frac{P(Y, Z_{k_3} | \Theta)}{P(Z_{k_3} | Y, \Theta_m)} \right] \\
&\geq \mathbb{E}_{Z_{k_3} | Y, \Theta_m} \log \left[ \frac{P(Y, Z_{k_3} | \Theta)}{P(Z_{k_3} | Y, \Theta_m)} \right] \\
&= \mathbb{E}_{Z_{k_3} | Y, \Theta_m} \left[ \log \left[ P(Y, Z_{k_3} | \Theta) \right] \right] - \mathbb{E}_{Z_{k_3} | Y, \Theta_m} \left[ \log P(Z_{k_3} | Y, \Theta_m) \right] \\
&= Q_{Z_{k_3}}(\Theta | \Theta_m) + R_{Z_{k_3}}(\Theta_m | \Theta_m),
\end{aligned}
\tag{C.1}
$$

where the inequality in (C.1) follows from Jensen's inequality, and the Q-function with the effects of $C^{(3)}$ is given by

$$
Q_{Z_{k_3}}(\Theta | \Theta_m) = \mathbb{E}_{Z_{k_3} | Y, \Theta_m} \left[ \log \left[ P(Y, Z_{k_3} | \Theta) \right] \right],
\tag{C.2}
$$

and

**Susceptible state over time**

**Exposed state over time**

**Infectious state over time**

**Removed state over time**

**Fig. B.5.** Shows three sample paths (contrasted with red, blue and purple lines) each for the states $S, E, I, R$ of the SEIR Markov chain model $\{X(t_k) : k \geq 0\}$ with transition probabilities in Theorem 3.2, whenever $p = 0.0055$, $\lambda = 10$, the average incubation $T_1$ and infectious $T_2$ periods are respectively, $E(T_1) = 10$ and $E(T_2) = 20$. In addition, the following initial conditions are used $S(t_0) = 200,000$, $E(t_0) = 500$, $I(t_0) = 1000$, and $R(t_0) = 0$.

$$R_{Z_{k_3}}(\Theta_m|\Theta_m) = -\mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\log P(Z_{k_3}|Y,\Theta_m)\right],$$

is the entropy term representing uncertainty in $Z_3$, given the observed information $Y$, and the estimate of the unknown parameter $\Theta_m$ at the $m^{th}$ iteration.

To avoid repetition and minimize space, the steps of (C.1) will be automatically employed, wherever necessary, to add the effects of $C^{(2)}$ and $C^{(1)}$ in (4.7), into the Q-function (C.2), without explicitly expressing all steps in the process as it is done in (C.1).

From (C.2), observe that adding the effects of $C^{(2)}$, we obtain

$$
\begin{aligned}
Q_{Z_{k_3}}(\Theta|\Theta_m) &= \mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\log\left[P(Y, Z_{k_3}|\Theta)\right]\right] \\
&= \mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\log\left[\sum_{Z_{k_2}} P(Y, Z_{k_3}, Z_{k_2}|\Theta)\right]\right] \\
&= \mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\log\left[\sum_{Z_{k_2}}\left(P(Z_{k_2}|Y, Z_{k_3}, \Theta_m) \times \frac{P(Y, Z_{k_3}, Z_{k_2}|\Theta)}{P(Z_{k_2}|Y, Z_{k_3}, \Theta_m)}\right)\right]\right] \\
&= \mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\log\left[\mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m}\left(\frac{P(Y, Z_{k_3}, Z_{k_2}|\Theta)}{P(Z_{k_2}|Y, Z_{k_3}, \Theta_m)}\right)\right]\right].
\end{aligned}
\tag{C.3}
$$

Applying Jensen's inequality to (C.3), it is easy to see that

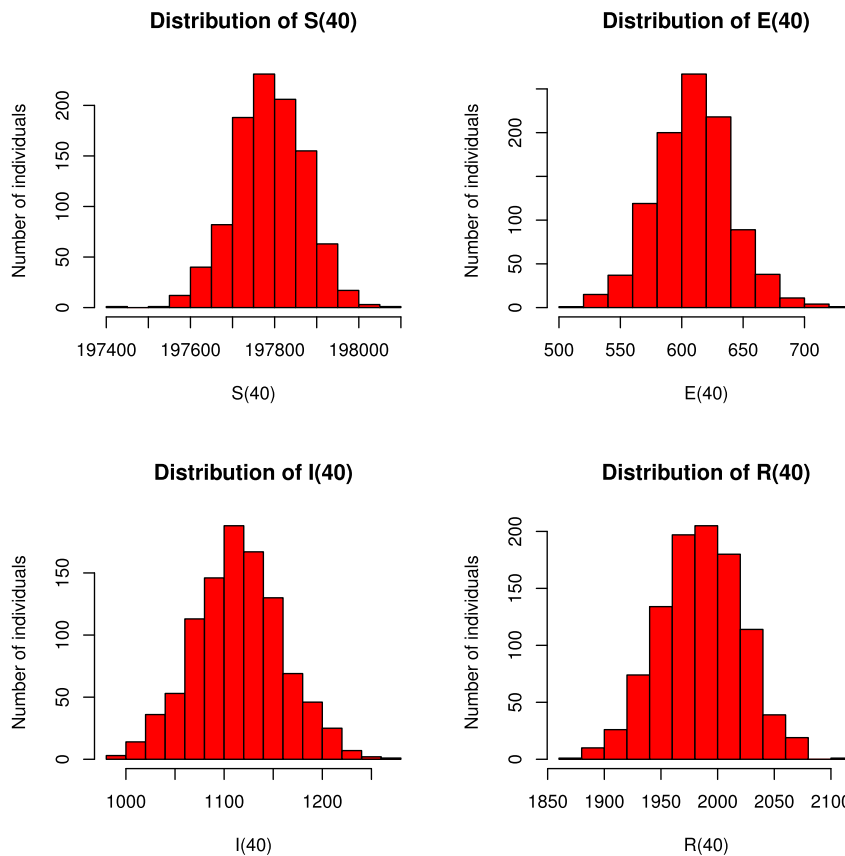$$Q_{Z_{k_3}}(\Theta|\Theta_m) \geq Q_{Z_{k_3}, Z_{k_2}}(\Theta|\Theta_m) + R_{Z_{k_3}, Z_{k_2}}(\Theta_m|\Theta_m),$$

where

$$Q_{Z_{k_3}, Z_{k_2}}(\Theta|\Theta_m) = \mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m}\left[\log\left(P(Y, Z_{k_3}, Z_{k_2}|\Theta)\right)\right]\right] \tag{C.4}$$

and

$$R_{Z_{k_3}, Z_{k_2}}(\Theta_m|\Theta_m) = -\mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m}\left[\log\left(P(Z_{k_2}|Y, Z_{k_3}, \Theta_m)\right)\right]\right].$$

**Distribution of S(40)**

**Distribution of E(40)**

**Distribution of I(40)**

**Distribution of R(40)**

**Fig. B.6.** Shows the approximate distributions for the states $S, E, I, R$ of the SEIR Markov chain model $\{X(t_k) : k \geq 0\}$ with transition probabilities in Theorem 3.2, whenever $p = 0.0055$, $\lambda = 10$, the average incubation $T_1$ and infectious $T_2$ periods are respectively, $E(T_1) = 10$ and $E(T_2) = 20$. In addition, the following initial conditions are used $S(t_0) = 200,000$, $E(t_0) = 500$, $I(t_0) = 1000$, and $R(t_0) = 0$. The histograms are based on 1000 sample realizations of the states $S, E, I, R$ at time $t_{40}$. Furthermore, the 95% confidence intervals for the populations means of the states $S, E, I, R$ at time $t_{40}$ are respectively, $197783.9 < \mathbb{E}[S(t_{40})] < 197794.2$, $607.8549 < \mathbb{E}[E(t_{40})] < 611.7331$, $1112.502 < \mathbb{E}[I(t_{40})] < 1118.106$ and $1983.614 < \mathbb{E}[R(t_{40})] < 1988.112$, where $\mathbb{E}[.]$ is the expectation operator.

In the final step, we add the effects of $C^{(1)}$ into the Q-function in (C.4) similarly as above. That is, from (C.4), adding $C^{(1)}$, we obtain

$$Q_{Z_{k_3}, Z_{k_2}}(\Theta|\Theta_m) = \mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m}\left(\log\left(P(Y, Z_{k_3}, Z_{k_2}|\Theta)\right)\right)\right]$$

$$= \mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m}\left(\log\left[\sum_{Z_{k_1}} P(Y, Z_{k_3}, Z_{k_2}, Z_{k_1}|\Theta)\right]\right)\right]$$

$$= \mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m}\left(\log\left(E_{Z_{k_1}|Y,Z_{k_3},Z_{k_2},\Theta_m}\left[\frac{P(Y, Z_{k_3}, Z_{k_2}, Z_{k_1}|\Theta)}{P(Z_{k_1}|Y, Z_{k_3}, Z_{k_2}, \Theta_m)}\right]\right)\right)\right]. \quad (C.5)$$

Applying Jensen's inequality to (C.5), it is easy to see that

$$Q_{Z_{k_3}, Z_{k_2}}(\Theta|\Theta_m) \geq Q_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta|\Theta_m) + R_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta_m|\Theta_m),$$

where

$$Q_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta|\Theta_m) = \mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m}\left[\mathbb{E}_{Z_{k_1}|Y,Z_{k_3},Z_{k_2},\Theta_m}\left[\log\left[P(Y, Z_{k_3}, Z_{k_2}, Z_{k_1}|\Theta)\right]\right]\right]\right] \quad (C.6)$$

and

$$R_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta_m|\Theta_m) = -\mathbb{E}_{Z_{k_3}|Y,\Theta_m}\left[\mathbb{E}_{Z_{k_2}|Y,Z_{k_3},\Theta_m}\left[\mathbb{E}_{Z_{k_1}|Y,Z_{k_3},Z_{k_2},\Theta_m}\left[\log\left[P(Z_{k_1}|Y, Z_{k_3}, Z_{k_2}, \Theta_m)\right]\right]\right]\right]. \quad (C.7)$$

Thus, from (C.1)-(C.7), it is easy to see that the log-likelihood function of the incomplete data satisfies

$$l(\Theta|Y) = \log\left[L(\Theta|Y)\right] \geq Q_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta|\Theta_m) + R_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta_m|\Theta_m) + R_{Z_{k_3}, Z_{k_2}}(\Theta_m|\Theta_m) + R_{Z_{k_3}}(\Theta_m|\Theta_m).$$

Hence, clearly the Q-function with the effects of all the missing hierarchical information $C^{(3)}$, $C^{(2)}$ and $C^{(1)}$ is $Q_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta|\Theta_m)$ in (C.6), and the terms $R_{Z_{k_3}, Z_{k_2}, Z_{k_1}}(\Theta_m|\Theta_m)$, $R_{Z_{k_3}, Z_{k_2}}(\Theta_m|\Theta_m)$, and $R_{Z_{k_3}}(\Theta_m|\Theta_m)$ are entropy terms. $\square$

## Appendix D. Proof of Theorem 5.1

**Proof.** Given that $\Theta^2 = (p, \lambda, \delta_e, \delta_r)$ is the unknown parameter for the SEIR model in Theorem 3.2, it is easy to see that applying the method in (5.5)-(5.7) to find the log-likelihood function of the incomplete data, $l(\Theta^2|Y) = l(\Theta^2|\hat{\mathfrak{E}}_T, \hat{\mathfrak{I}}_T)$, it follows from Definition 5.1 and Definition 4.1[5.], that

$$l(\Theta^2|Y) = l(\Theta^2|\hat{\mathfrak{E}}_T, \hat{\mathfrak{I}}_T) = \log\left[P(\mathfrak{E}_T = \hat{\mathfrak{E}}_T, \mathfrak{I}_T = \hat{\mathfrak{I}}_T|\Theta^2)\right]$$

$$= \log\left[\, P\left(\, S(t_T) = \hat{s}_T, I(t_T) = \hat{i}_T; S(t_{T-1}) = \hat{s}_{T-1}, I(t_{T-1}) = \hat{i}_{T-1}; \right.\right.$$

$$\left.\left.\cdots; S(t_0) = \hat{s}_0, I(t_0) = \hat{i}_0|\Theta^2\,\right)\,\right]$$

$$= \log\left[\prod_{k=1}^{T} P\left(S(t_k) = \hat{s}_k, I(t_k) = \hat{i}_k|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right)\right]$$

$$= \sum_{k=1}^{T} \log\left[P\left(S(t_k) = \hat{s}_k, I(t_k) = \hat{i}_k|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right)\right] \tag{D.1}$$

Applying a similar method in (C.1) to (D.1), we impute the missing the missing data for $E(t_k)$ over $k = 1, 2, \ldots, T$ into (D.2), and apply Jensen's inequality as shown below.

$$l(\Theta^2|Y) = \sum_{k=1}^{T} \log\left[\sum_{\Omega^1_{E(t_k)}} P\left(S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right)\right]$$

$$= \sum_{k=1}^{T} \log\left[\sum_{\Omega^1_{E(t_k)}} P\left(S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right) \times \right.$$

$$\left. \times \frac{P\left(E(t_k) = \hat{e}_k|S(t_k) = \hat{s}_k, I(t_k) = \hat{i}_k, ; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}, \Theta^2_m\right)}{P\left(E(t_k) = \hat{e}_k|S(t_k) = \hat{s}_k, I(t_k) = \hat{i}_k, ; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}, \Theta^2_m\right)}\right], \tag{D.2}$$

where $\Omega^1_{E_k}$, $k \geq 1$ is given in (5.10)-(5.11).

Applying Jensen's inequality to (D.2), i.e. to the average $\mathbb{E}_{E(t_k)|S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta^2_m}[\cdots]$ with respect to the conditional mass $P(E(t_k) = \hat{e}_k|S(t_k) = \hat{s}_k, I(t_k) = \hat{i}_k, ; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}, \Theta^2_m)$, it is easy to see

$$l(\Theta^2|Y) \geq \sum_{k=1}^{T} \mathbb{E}_{E(t_k)|S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta^2_m}\left[\right.$$

$$\log\left\{P\left(S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right)\right\}\right]$$

$$- \sum_{k=1}^{T} \mathbb{E}_{E(t_k)|S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta^2_m}\left[\right.$$

$$\log\left[P\left(E(t_k) = \hat{e}_k|S(t_k) = \hat{s}_k, I(t_k) = \hat{i}_k, ; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}, \Theta^2_m\right)\right]\right]. \tag{D.3}$$

For each $k = 1, 2, 3, \ldots, T$, define

$$Q_{E(t_k)}(\Theta^2|\Theta^2_m) = \mathbb{E}_{E(t_k)|S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta^2_m}\left[\right.$$

$$\log\left\{P\left(S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right)\right\}\right], \tag{D.4}$$

and

$$R_{E(t_k)}(\Theta^2_m|\Theta^2_m) = -\mathbb{E}_{E(t_k)|S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta^2_m}\left[\right.$$

$$\log\left\{P\left(E(t_k) = \hat{e}_k|S(t_k) = \hat{s}_k, I(t_k) = \hat{i}_k, ; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}, \Theta^2_m\right)\right\}\right].$$

It follows from (D.3) that

$$l(\Theta^2|Y) \geq \sum_{k=1}^{T} Q_{E(t_k)}(\Theta^2|\Theta^2_m) + \sum_{k=1}^{T} R_{E(t_k)}(\Theta^2_m|\Theta^2_m). \tag{D.5}$$

Similarly to the proof of Theorem 4.1, it is easy to see that $Q_{E(t_k)}(\Theta^2|\Theta^2_m)$ is the corresponding Q-function with the effects of only the non-hierarchical missing data $E(t_k)$ over $k = 1, 2, \ldots, T$. Also, $R_{E(t_k)}(\Theta^2_m|\Theta^2_m)$ is the corresponding entropy term, similarly defined as in the proof of Theorem 4.1.

Observe from (D.4) and (5.8), that the Q-function $Q_{E(t_k)}(\Theta^2|\Theta^2_m)$ is the average over the domain $(\Omega^1_{E_k})$ of missing values for $E(t_k)$ over $k = 1, 2, \ldots, T$, of the log-likelihood function of the incomplete data contained in the sum $l_1(\Theta^2|\hat{H}_T)$ in (5.8). Thus, for each $k = 1, 2, \ldots, T$, define by

$$l_{1,k}(\Theta^2|S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}))$$

$$= \log\left\{P\left(S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right)\right\}, \tag{D.6}$$

the $k^{th}$ component of the log-likelihood of the incomplete data contained in the sum $l_1(\Theta^2|\hat{H}_T)$ in (5.8). Thus, from (D.4) and (D.6), it follows that

$$Q_{E(t_k)}(\Theta^2|\Theta_m^2) = \mathbb{E}_{E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\left[l_{1,k}(\Theta^2|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}))\right]. \tag{D.7}$$

Thus, we are only left to add the hidden hierarchical data $Z = \{C^{(4)}, C^{(3)}, C^{(2)}, C^{(1)}\}$ given in (5.12), into the log-likelihood function $l_{1,k}(\Theta^2|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}))$ in (D.6), and substitute the results back into (D.7) to obtain the final Q-function with the effects of the missing non-hierarchical data and the missing hierarchical hidden data.

Since the hidden hierarchical data $Z$ is a four level hierarchical data, we apply Theorem 4.1 & Proposition 4.1 to add $Z$ into $l_{1,k}(\Theta^2|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}))$ in (D.6). Also, since this process is similar to the proof of Theorem 4.1, we omit the step-by-step process, and just present the result of the final step, and define the notations used.

It is easy to see from Theorem 4.1 & Proposition 4.1, that for each $k = 1, 2, \ldots, T$, it follows from (D.4)-(D.7) that $l(\Theta^2|Y)$ in (D.5)

$$
\begin{aligned}
l(\Theta^2|Y) \geq & \sum_{k=1}^{T} Q_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta_m^2) \\
& + \sum_{k=1}^{T} R_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta_m^2|\Theta_m^2) \\
& + \sum_{k=1}^{T} R_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j}}(\Theta_m^2|\Theta_m^2) + \sum_{k=1}^{T} R_{E(t_k),Z_{t_k},Z_{t_k,n}}(\Theta_m^2|\Theta_m^2) \\
& + \sum_{k=1}^{T} R_{E(t_k),Z_{t_k}}(\Theta_m^2|\Theta_m^2) + \sum_{k=1}^{T} R_{E(t_k)}(\Theta_m^2|\Theta_m^2),
\end{aligned}
\tag{D.8}
$$

where

$$
\begin{aligned}
Q_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta_m^2) = & \mathbb{E}_{E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \mathbb{E}_{Z_{t_k}|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \mathbb{E}_{Z_{t_k,n}|S(t_k),E(t_k),Z_{t_k},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \mathbb{E}_{Z_{t_k,n,j}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \mathbb{E}_{Z_{t_k,n,j,l}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& l_{1,k}(\Theta^2|Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l},S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}))\big]\big]\big]\big]\big]
\end{aligned}
$$

and

$$
\begin{aligned}
& l_{1,k}(\Theta^2|Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l},S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1})) \\
& = \log\left\{ P\left(Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l},S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k | S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right)\right\}.
\end{aligned}
$$

Also, the entropy terms are given as follows.

$$
\begin{aligned}
R_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta_m^2|\Theta_m^2) = & -\mathbb{E}_{E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \mathbb{E}_{Z_{t_k}|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \mathbb{E}_{Z_{t_k,n}|S(t_k),E(t_k),Z_{t_k},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \mathbb{E}_{Z_{t_k,n,j}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \mathbb{E}_{Z_{t_k,n,j,l}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \log \mathbb{P}(Z_{t_k,n,j,l}|Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2)\big]\big]\big]\big]\big] \\
R_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j}}(\Theta_m^2|\Theta_m^2) = & -\mathbb{E}_{E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \mathbb{E}_{Z_{t_k}|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \mathbb{E}_{Z_{t_k,n}|S(t_k),E(t_k),Z_{t_k},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \mathbb{E}_{Z_{t_k,n,j}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \log \mathbb{P}(Z_{t_k,n,j}|Z_{t_k},Z_{t_k,n},S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2)\big]\big]\big]\big] \\
R_{E(t_k),Z_{t_k},Z_{t_k,n}}(\Theta_m^2|\Theta_m^2) = & -\mathbb{E}_{E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \mathbb{E}_{Z_{t_k}|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \mathbb{E}_{Z_{t_k,n}|S(t_k),E(t_k),Z_{t_k},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \log \mathbb{P}(Z_{t_k,n}|Z_{t_k},S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2)\big]\big]\big] \\
R_{E(t_k),Z_{t_k}}(\Theta_m^2|\Theta_m^2) = & -\mathbb{E}_{E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \mathbb{E}_{Z_{t_k}|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \log \mathbb{P}(Z_{t_k}|S(t_k),E(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2)\big]\big]
\end{aligned}
$$

and

$$
\begin{aligned}
R_{E(t_k)}(\Theta_m^2|\Theta_m^2) = & -\mathbb{E}_{E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2}\big[ \\
& \log \mathbb{P}(E(t_k)|S(t_k),I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2)\big]. \quad \square
\end{aligned}
$$

## Appendix E. Proof of Theorem 5.2

**Proof.** The E-step of the EM-algorithm consists of finding the Q-function in (5.16) given by

$$Q(\Theta^2|\Theta_m^2) = \sum_{k=1}^{T} Q_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta_m^2), \tag{E.1}$$

where $Q_{E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},Z_{t_k,n,j,l}}(\Theta^2|\Theta_m^2)$ is defined in (5.14). We compute the Q-function in a step-by-step manner as follows.

For each $k = 1, 2, \ldots T$, assume that $S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, Z_{t_k} = n, Z_{t_k,n} = j, Z_{t_k,n,j} = l, I(t_k) = \hat{i}_k; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}, \Theta_m^2$ are given, where the values of $n, j, l$ are given in Definitions 3.2-3.5&5.3. It follows from the distribution of $Z_{t_k,n,j,l}$ in (3.10)-(3.11) that

$$\mathbb{P}(Z_{t_k,n,j,l} = 1|S(t_k), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2) = p^m.$$

It is easy to see from (5.15) that

$$\mathbb{E}_{Z_{t_k,n,j,l}|S(t_k),E(t_k),Z_{t_k},Z_{t_k,n},Z_{t_k,n,j},I(t_k);S(t_{k-1}),I(t_{k-1}),\Theta_m^2} \Big[ l_{1,k}(\Theta^2|Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}, S(t_k), E(t_k), I(t_k);$$

$$S(t_{k-1}), I(t_{k-1})) \Big]$$

$$= \log \Big\{ \mathbb{P}\Big( Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big) \Big\} \times$$

$$\times \mathbb{P}(Z_{t_k,n,j,l} = 1|S(t_k), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2)$$

$$= \log \Big\{ \mathbb{P}\Big( Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big) \Big\} \times$$

$$\times p^m. \tag{E.2}$$

Also, observe from (E.2) that

$$\mathbb{P}\Big( Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big)$$

$$= \mathbb{P}\Big( Z_{t_k}|S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big) \times$$

$$\times \mathbb{P}\Big( Z_{t_k,n}|Z_{t_k}; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big) \times$$

$$\times \mathbb{P}\Big( Z_{t_k,n,j}|Z_{t_k}, Z_{t_k,n}; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big) \times$$

$$\times \mathbb{P}\Big( Z_{t_k,n,j,l} = 1|Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big) \times$$

$$\times \mathbb{P}\Big( S(t_k) = \hat{s}_k|Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big) \times$$

$$\times \mathbb{P}\Big( E(t_k) = \hat{e}_k|Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big) \times$$

$$\times \mathbb{P}\Big( I(t_k) = \hat{i}_k|Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big). \tag{E.3}$$

We find expressions for every term in (E.3). From (2.1)-(2.4), observe that given the observed data $Y$ in Definitions 5.1-5.2, it is easy to see that for $k = 1, 2, \ldots, T$

$$S(t_k) = \hat{s}_k \Leftrightarrow C_{SE}(t_{k-1}) = \hat{s}_{k-1} - \hat{s}_k,$$

$$E(t_k) = \hat{e}_k \Leftrightarrow C_{EI}(t_{k-1}) = E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k,$$

$$I(t_k) = \hat{i}_k \Leftrightarrow C_{IR}(t_{k-1}) = \hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k,$$

where $E(t_k)$ is random, and the support is given by $E(t_k) = \hat{e}_k \in \Omega_{E_k}^1$.

From (E.3),

$$\mathbb{P}\Big( S(t_k) = \hat{s}_k|Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big)$$

$$= \mathbb{P}\Big( C_{SE}(t_{k-1}) = \hat{s}_{k-1} - \hat{s}_k|Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big)$$

$$= \binom{\hat{s}_{k-1}}{\hat{s}_{k-1} - \hat{s}_k} p^{\hat{s}_{k-1} - \hat{s}_k} (1-p)^{\hat{s}_k}. \tag{E.4}$$

The equation (E.4) follows because, given that the $Z_{t_k,n,j} = l$ and $Z_{t_k,n,j,l} = 1$, that is, the $l^{th}$ infectious contacts with susceptible person leads to infection, then $C_{SE}(t_{k-1})$ is binomial with parameters $p$ and $\hat{s}_{k-1}$.

Also,

$$\mathbb{P}\Big( E(t_k) = \hat{e}_k|Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big)$$

$$= \mathbb{P}\Big( C_{EI}(t_{k-1}) = E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k|Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k;$$

$$S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \Big)$$

$$= \binom{E(t_{k-1})}{E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k} \Big( P_{EI}(t_{k-1}) \Big)^{E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k} \Big( 1 - P_{EI}(t_{k-1}) \Big)^{E(t_k) - [\hat{s}_{k-1} - \hat{s}_k]}, \tag{E.5}$$

where $P_{EI}(t_{k-1})$ is defined in (3.15), and (E.5) also follows from (3.14). In addition,

$$\mathbb{P}\left(I(t_k) = \hat{i}_k | Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right)$$

$$= \mathbb{P}\left(C_{IR}(t_{k-1}) = \hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k | Z_{t_k}, Z_{t_k,n},\right.$$

$$\left. Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right)$$

$$= \binom{\hat{i}_{k-1}}{\hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k} \left(P_{IR}(t_{k-1})\right)^{\hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k}$$

$$\times \left(1 - P_{IR}(t_{k-1})\right)^{\hat{i}_k - [E(t_{k-1}) - E(t_k)] - [\hat{s}_{k-1} - \hat{s}_k]}, \tag{E.6}$$

where $P_{IR}(t_{k-1})$ is defined in (3.16), and (E.6) also follows from (3.14).

From (E.3), the distribution of the hierarchical missing data is given as follows. Note that from Definitions 3.2-3.5, we deduce that

$$\mathbb{P}\left(Z_{t_k} | S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right) = (\lambda \Delta t)^{Z_{t_k}} e^{-\lambda \Delta t} \frac{1}{(Z_{t_k})!},$$

$$\mathbb{P}\left(Z_{t_k,n} | Z_{t_k}; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right) = \binom{Z_{t_k} = n}{Z_{t_k,n} = j} \left(\alpha_{t_k}\right)^{Z_{t_k,n}} \left(1 - \alpha_{t_k}\right)^{Z_{t_k} - Z_{t_k,n}},$$

where $\alpha_{t_k}$ is defined in Definitions 3.2-3.5. Furthermore, observe from Definitions 3.2-3.5 that

$$\mathbb{P}\left(Z_{t_k,n,j} | Z_{t_k}, Z_{t_k,n}; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right) = \frac{1}{Z_{t_k,n}},$$

and

$$\mathbb{P}\left(Z_{t_k,n,j,l} = 1 | Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}; S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right) = p. \tag{E.7}$$

It follows from (E.4)-(E.7) that

$$\log\left\{\mathbb{P}\left(Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k | S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2\right)\right\}$$

$$= \log\left[\binom{\hat{s}_{k-1}}{\hat{s}_{k-1} - \hat{s}_k}\right] + \log\left[\binom{E(t_{k-1})}{E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k}\right]$$

$$+ \log\left[\binom{\hat{i}_{k-1}}{\hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k}\right] + \log\left(\frac{1}{(Z_{t_k})!}\right)$$

$$+ \log\left[\binom{Z_{t_k} = n}{Z_{t_k,n} = j} \left(\alpha_{t_k}\right)^{Z_{t_k,n}} \left(1 - \alpha_{t_k}\right)^{Z_{t_k} - Z_{t_k,n}}\right] + \log\left(\frac{1}{Z_{t_k,n}}\right)$$

$$+ \left(\hat{s}_{k-1} - \hat{s}_k\right) \log p + \left(\hat{s}_k\right) \log(1 - p)$$

$$+ \left(E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k\right) \log\left(P_{EI}(t_{k-1})\right) + \left(E(t_k) - [\hat{s}_{k-1} - \hat{s}_k]\right) \log\left(1 - P_{EI}(t_{k-1})\right)$$

$$+ \left(\hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k\right) \log\left(P_{IR}(t_{k-1})\right)$$

$$+ \left(\hat{i}_k - [E(t_{k-1}) - E(t_k)] - [\hat{s}_{k-1} - \hat{s}_k]\right) \log\left(1 - P_{IR}(t_{k-1})\right)$$

$$+ Z_{t_k} \log(\lambda \Delta t) - \lambda \Delta t + \log(p). \tag{E.8}$$

To reduce unnecessary terms, we define the following random variable that is independent of the parameter $\Theta^2$, but depends on the other random variable $E(t_{k-1}), E(t_k) Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1$

$$\mathbb{K}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1)$$

$$= \log\left[\binom{\hat{s}_{k-1}}{\hat{s}_{k-1} - \hat{s}_k}\right] + \log\left[\binom{E(t_{k-1})}{E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k}\right]$$

$$+ \log\left[\binom{\hat{i}_{k-1}}{\hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k}\right] + \log\left(\frac{1}{(Z_{t_k})!}\right)$$

$$+ \log\left[\binom{Z_{t_k} = n}{Z_{t_k,n} = j} \left(\alpha_{t_k}\right)^{Z_{t_k,n}} \left(1 - \alpha_{t_k}\right)^{Z_{t_k} - Z_{t_k,n}}\right] + \log\left(\frac{1}{Z_{t_k,n}}\right). \tag{E.9}$$

Also define the following random variable that dependents on both the parameter $\Theta^2$, and the other random variable $E(t_{k-1}), E(t_k) Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1$

$$\mathbb{J}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; \Theta^2)$$

$$= \left(\hat{s}_{k-1} - \hat{s}_k\right) \log p + \left(\hat{s}_k\right) \log(1 - p)$$

$$+ \left(E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k\right) \log\left(P_{EI}(t_{k-1})\right) + \left(E(t_k) - [\hat{s}_{k-1} - \hat{s}_k]\right) \log\left(1 - P_{EI}(t_{k-1})\right)$$

$$+ \left(\hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k\right) \log\left(P_{IR}(t_{k-1})\right)$$

$$+ \left( \hat{i}_k - [E(t_{k-1}) - E(t_k)] - [\hat{s}_{k-1} - \hat{s}_k] \right) \log \left( 1 - P_{IR}(t_{k-1}) \right)$$

$$+ Z_{t_k} \log (\lambda \Delta t) - \lambda \Delta t + \log (p). \tag{E.10}$$

Thus, from (E.9), (E.11) and (E.8), we obtain the following.

$$\log \left\{ \mathbb{P} \left( Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1, S(t_k) = \hat{s}_k, E(t_k) = \hat{e}_k, I(t_k) = \hat{i}_k | S(t_{k-1}) = \hat{s}_{k-1}, I(t_{k-1}) = \hat{i}_{k-1}; \Theta^2 \right) \right\}$$

$$= \mathbb{K}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1) + \mathbb{J}(E(t_{k-1}), E(t_k) Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; \Theta^2). \tag{E.11}$$

Therefore, from (E.2) we obtain

$$\mathbb{E}_{Z_{t_k,n,j,l} | S(t_k), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} \left[ l_{1,k}(\Theta^2 | Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}, S(t_k), E(t_k), I(t_k); \right.$$

$$\left. S(t_{k-1}), I(t_{k-1})) \right]$$

$$= \left[ \mathbb{K}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1) + \mathbb{J}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; \Theta^2) \right] \times$$

$$\times p^m. \tag{E.12}$$

The result in (5.21) follows immediately by substituting (E.12) into (5.14) and (E.1). $\square$

## Appendix F. Proof of Theorem 5.3

**Proof.** Given the Q-function in (5.21), observe that the gradient operator $\nabla_{\Theta^2} = \left( \frac{\partial(\cdots)}{\partial p}, \frac{\partial(\cdots)}{\partial \lambda}, \frac{\partial(\cdots)}{\partial \delta_e}, \frac{\partial(\cdots)}{\partial \delta_r} \right)^T$ applied to $Q(\Theta^2 | \Theta_m^2)$ leads to

$$\nabla_{\Theta^2} Q(\Theta^2 | \Theta_m^2) = \sum_{k=1}^T \nabla_{\Theta^2} Q_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}^2 (\Theta^2 | \Theta_m^2), \tag{F.1}$$

since for each $k = 1, 2, \ldots, T$,

$$\nabla_{\Theta^2} Q_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}^1 (\Theta^2 | \Theta_m^2) = \vec{0}.$$

In addition, the series of conditional expectation operators in (5.20) depend only on $\Theta_m^2$, therefore we can pass the gradient operator under the conditional expectation operators. That is, from (5.20),

$$\nabla_{\Theta^2} Q_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}^2 (\Theta^2 | \Theta_m^2) = \mathbb{E}_{E(t_k) | S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} \big[$$

$$\mathbb{E}_{Z_{t_k} | S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} \big[ \mathbb{E}_{Z_{t_k,n} | S(t_k), E(t_k), Z_{t_k}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} \big[$$

$$\mathbb{E}_{Z_{t_k,n,j} | S(t_k), E(t_k), Z_{t_k}, Z_{t_k,n}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} \big[$$

$$\nabla_{\Theta^2} \mathbb{J}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; \Theta^2) \big] \big] \big] \big]. \tag{F.2}$$

From (5.18) observe that for each $k = 1, 2, \ldots, T$,

$$\nabla_{\Theta^2} \mathbb{J}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; \Theta^2)$$

$$= \begin{pmatrix} \frac{(\hat{s}_{k-1} - \hat{s}_k)}{p} - \frac{\hat{s}_k}{1-p} + \frac{1}{p} \\ Z_{t_k} \frac{1}{\lambda} - \Delta t \\ \left( E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k \right) \left[ \frac{\Delta t e^{-\delta_e \Delta t}}{1 - e^{-\delta_e \Delta t}} \right] - \left( E(t_k) - [\hat{s}_{k-1} - \hat{s}_k] \right) \Delta t \\ \left( \hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k \right) \left[ \frac{\Delta t e^{-\delta_r \Delta t}}{1 - e^{-\delta_r \Delta t}} \right] - \left( \hat{i}_k - [E(t_{k-1}) - E(t_k)] - [\hat{s}_{k-1} - \hat{s}_k] \right) \Delta t \end{pmatrix} \tag{F.3}$$

To reduce complex notations, denote the complex conditional expectation operator in (F.3) by

$$\tilde{\mathbb{E}}_{k;\Theta_m^2} [\cdots] = \mathbb{E}_{E(t_k) | S(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} \big[$$

$$\mathbb{E}_{Z_{t_k} | S(t_k), E(t_k), I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} \big[ \mathbb{E}_{Z_{t_k,n} | S(t_k), E(t_k), Z_{t_k}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} \big[$$

$$\mathbb{E}_{Z_{t_k,n,j} | S(t_k), E(t_k), Z_{t_k}, Z_{t_k,n}, I(t_k); S(t_{k-1}), I(t_{k-1}), \Theta_m^2} \big[$$

$$\cdots ]]]]. \tag{F.4}$$

Thus, using (F.4) and (F.3), the expression (F.2) becomes

$$\nabla_{\Theta^2} Q_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}^2 (\Theta^2 | \Theta_m^2) = \tilde{\mathbb{E}}_{k;\Theta_m^2} \left[ \nabla_{\Theta^2} \mathbb{J}(E(t_{k-1}), E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1; \Theta^2) \right]$$

$$= \begin{pmatrix} \frac{1}{p} \tilde{\mathbb{E}}_{k;\Theta_m^2} [\hat{s}_{k-1} - \hat{s}_k] - \left( \frac{1}{1-p} \right) \tilde{\mathbb{E}}_{k;\Theta_m^2} [\hat{s}_k] + \frac{1}{p} \\ \frac{1}{\lambda} \tilde{\mathbb{E}}_{k;\Theta_m^2} [Z_{t_k}] - \Delta t \tilde{\mathbb{E}}_{k;\Theta_m^2} [1] \\ \left[ \frac{\Delta t e^{-\delta_e \Delta t}}{1 - e^{-\delta_e \Delta t}} \right] \tilde{\mathbb{E}}_{k;\Theta_m^2} \left[ \left( E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k \right) \right] - \Delta t \tilde{\mathbb{E}}_{k;\Theta_m^2} \left[ \left( E(t_k) - [\hat{s}_{k-1} - \hat{s}_k] \right) \right] \\ \left[ \frac{\Delta t e^{-\delta_r \Delta t}}{1 - e^{-\delta_r \Delta t}} \right] \tilde{\mathbb{E}}_{k;\Theta_m^2} \left[ \left( \hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k \right) \right] \\ - \Delta t \tilde{\mathbb{E}}_{k;\Theta_m^2} \left[ \left( \hat{i}_k - [E(t_{k-1}) - E(t_k)] - [\hat{s}_{k-1} - \hat{s}_k] \right) \right] \end{pmatrix} \tag{F.5}$$

Hence, from (F.5) and (F.1) it follows that the vector equation

$$\nabla_{\Theta^2} Q(\Theta^2 | \Theta_m^2) = \sum_{k=1}^{T} \nabla_{\Theta^2} Q^2_{E(t_k), Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l}}(\Theta^2 | \Theta_m^2) = \vec{0}, \tag{F.6}$$

leads to the following solution for the vector $\Theta^2 = (p, \lambda, \delta_e, \delta_r)^T$

$$\Theta^2 = \begin{pmatrix} \dfrac{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}[\hat{s}_{k-1} - \hat{s}_k] + \sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}[1]}{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}[\hat{s}_{k-1} - \hat{s}_k] + \sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}[1] + \sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}[\hat{s}_k]} \\ \dfrac{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}[Z_{t_k}]}{\Delta t \sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}[1]} \\ -\dfrac{1}{\Delta t} \log\left[\dfrac{1}{A_e^m + 1}\right] \\ -\dfrac{1}{\Delta t} \log\left[\dfrac{1}{B_r^m + 1}\right] \end{pmatrix} \tag{F.7}$$

where

$$A_e^m = \frac{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k\right)\right]}{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_k) - [\hat{s}_{k-1} - \hat{s}_k]\right)\right]},$$

and

$$B_r^m = \frac{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(\hat{i}_{k-1} - \hat{i}_k + E(t_{k-1}) - E(t_k) + \hat{s}_{k-1} - \hat{s}_k\right)\right]}{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(\hat{i}_k - [E(t_{k-1}) - E(t_k)] - [\hat{s}_{k-1} - \hat{s}_k]\right)\right]} \tag{F.8}$$

Recall, the only random variables are the missing hierarchical and non-hierarchical data are $E(t_{k-1}), E(t_k)$, and $Z_{t_k}, Z_{t_k,n}, Z_{t_k,n,j}, Z_{t_k,n,j,l} = 1$ defined in Definitions 5.3 & 5.2. Thus, (F.7)-(F.8) reduce to

$$\Theta^2 = \begin{pmatrix} \dfrac{\sum_{k=1}^{T}[\hat{s}_{k-1} - \hat{s}_k + 1]}{\sum_{k=1}^{T}[\hat{s}_{k-1} - \hat{s}_k + 1] + \sum_{k=1}^{T}[\hat{s}_k]} \\ \dfrac{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}[Z_{t_k}]}{\left(\sum_{k=1}^{T}[\Delta t]\right)} \\ \dfrac{1}{\Delta t} \log\left[A_e^m + 1\right] \\ \dfrac{1}{\Delta t} \log\left[B_r^m + 1\right] \end{pmatrix}$$

where

$$A_e^m = \frac{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1}) - E(t_k)\right)\right] + \sum_{k=1}^{T}\left[\left(\hat{s}_{k-1} - \hat{s}_k\right)\right]}{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[E(t_k)\right] - \sum_{k=1}^{T}\left[\left(\hat{s}_{k-1} - \hat{s}_k\right)\right]},$$

and

$$B_r^m = \frac{\sum_{k=1}^{T}\left[\left(\hat{i}_{k-1} - \hat{i}_k\right)\right] + \sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[\left(E(t_{k-1}) - E(t_k)\right)\right] + \sum_{k=1}^{T}\left[\left(\hat{s}_{k-1} - \hat{s}_k\right)\right]}{\sum_{k=1}^{T}\left[\hat{i}_k\right] - \sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}\left[E(t_{k-1}) - E(t_k)\right] - \sum_{k=1}^{T}\left[\hat{s}_{k-1} - \hat{s}_k\right]}.$$

Therefore, clearly from (5.17), given the $m^{th}$ step estimate $\Theta_m^2$ of the parameter $\Theta^2$, it is easy to see that the $(m+1)^{th}$ step estimate $\Theta_{m+1}^2$ of the parameter $\Theta^2$ is given by

$$\Theta_{m+1}^2 = \Theta_{max}^2 = \Theta^2 = \begin{pmatrix} \dfrac{\sum_{k=1}^{T}[\hat{s}_{k-1} - \hat{s}_k + 1]}{\sum_{k=1}^{T}[\hat{s}_{k-1} - \hat{s}_k + 1] + \sum_{k=1}^{T}[\hat{s}_k]} \\ \dfrac{\sum_{k=1}^{T} \tilde{\mathbb{E}}_{k;\Theta_m^2}[Z_{t_k}]}{\left(\sum_{k=1}^{T}[\Delta t]\right)} \\ \dfrac{1}{\Delta t} \log\left[A_e^m + 1\right] \\ \dfrac{1}{\Delta t} \log\left[B_r^m + 1\right] \end{pmatrix} \qquad \square$$

## References

[1] Cdc-fluview (1-17-2019): https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html; https://www.cdc.gov/flu/weekly/weeklyarchives2017-2018/Week39.htm.

[2] Cdc, influenza flu, how flu spreads: https://www.cdc.gov/flu/about/disease/spread.htm.

[3] Immunization action coalition, influenza: Questions and answers, Saint Paul, Minnesota, www.immunize.org/catg.d/p4208.pdf.

[4] Operational considerations for adapting a contact tracing program to respond to the Covid-19 pandemic in non-us settings, https://www.cdc.gov/coronavirus/2019-ncov/global-covid-19/operational-considerations-contact-tracing.html.

[5] Us census bureau, 2018 national and state population estimates, https://www.census.gov/newsroom/press-kits/2018/pop-estimates-national-state.html.

[6] H. Abbey, An examination of the Reed-frost theory of epidemics, Hum. Biol. 24 (1952) 201.

[7] L. Alkema, A.E. Raftery, S.J. Clark, Probabilistic projections of HIV prevalence using Bayesian melding, Ann. Appl. Stat. (2007) 229–248.

[8] L. Allen, An introduction to stochastic epidemic models, in: F. Brauer, P. van den Driessche, J. Wu (Eds.), Mathematical Epidemiology, in: Lecture Notes in Mathematics, vol. 1945, Springer, 2008, pp. 81–130.

[9] R.M. Anderson, M. Robert, Infectious Diseases of Humans: Dynamics and Control, Oxford University Press, 1992.

[10] H. Andersson, T. Britton, Stochastic Epidemic Models and Their Statistical Analysis, Lectures Notes in Statistics, vol. 151, Springer, 2000.

[11] N.T.J. Bailey, The Mathematical Theory of Infectious Diseases, Griffin & Co., New York, 1975.

[12] D. Bernoulli, Throwing foreign aid at HIV/AIDS in developing countries: missing the target?, Mem. Phys. Acad. R. Sci. 6 (1) (1760).

[13] J. Bilmes, A Gentle Tutorial of the EM Algorithm and It's Application to Parameter Estimation for Gausian Mixture and Hidden Markov Models, International Computer Science Institute, 1998, 15 pages.

[14] A.S. Bryke, S.W. Raudenbush, Hierarchical Linear Models: Applications and Data Analysis Methods, Sage Publications, Newbury Park, California, 1992.

[15] B. Canto, C. Coll, E. Sanchez, Estimation of parameters in a structured sir model, Adv. Differ. Equ. 2017 (2017) 33.

[16] A. Capaldi, S. Behrend, B. Berman, J. Smith, J. Wright, A.L. Lloyd, Parameter estimation and uncertainty quantification for an epidemic model, Math. Biosci. Eng. 9 (2012) 553–576.

[17] G. Casella, R.L. Berger, Statistical Inference, second edition, 2002.

[18] B. Choi, G.A. Rempala, Inference for discretely observed stochastic kinetic networks with applications to epidemic modelling, Biostatistics 13 (2012) 153–165.

[19] G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts, Infect. Dis. Model. 2 (2017) 379–398.

[20] A. Dempter, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. B 39 (1977) 1–38.

[21] O. Diekmann, J.A.P. Heesterbeek, J.A.J. Metz, On the definition and the computation of the basic reproduction ratio ro in models for infectious diseases in heterogeneous populations, J. Math. Biol. 28 (1990) 365–382.

[22] M. Draief, L. Massoulié, Galton-Watson branching processes, in: Epidemics and Rumours in Complex Networks, in: London Mathematical Society Lecture Note Series, Cambridge University Press, Cambridge, 2009, pp. 7–18.

[23] S. Duncan, M. Gyongy, Using the EM algorithm to estimate the disease parameters for smallpox in 17th century London, in: 2006 IEEE Conference on Computer Aided Control System Design, 2006 IEEE International Conference on Control Applications, 2006 IEEE International Symposium on Intelligent Control, 2006, pp. 3312–3317.

[24] F. Etbaigha, A.R. Willms, et al., An SEIR model of influenza A virus infection and reinfection within a farrow-to-finish swine farm, PLoS ONE 13 (2018) e0202493.

[25] R. Fierro, A class of stochastic epidemic models and its deterministic counterpart, J. Korean Stat. Soc. 39 (2010) 397–407.

[26] R. Fierro, V. Leiva, N. Balakrishnan, Statistical inference on stochastic epidemic model, Commun. Stat., Simul. Comput. 44 (2015) 2297–2314.

[27] D. Gamerman, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Chapman & Hall, 1997.

[28] J. Gani, D. Jerwood, Markov chain methods in chain binomial epidemic models, Biometrics 27 (1971) 591–603.

[29] M. Greenwood, On the statistical measure of infectiousness, J. Hyg. Camb. 31 (1931) 336–351.

[30] M. Gupta, Y. Chen, Theory and use of the EM algorithm, Found. Trends Signal Process. 4 (2010) 223–296.

[31] S. He, S. Tang, L. Rong, A discrete stochastic model of the Covid-19 outbreak: forecast and control, Math. Biosci. Eng. 17 (2020) 2792–2804.

[32] H.W. Hethcote, The mathematics of infectious diseases, SIAM Rev. 42 (2000) 599.

[33] W. Huang, Y. Chen, The multiset em algorithm, Stat. Probab. Lett. 126 (2017) 41–48.

[34] V. Islam, Stochastic models for epidemics: current issues and developments, in: Celebrating Statistics: Papers in Honor of Sir David Cox on His 80th Birthday, Oxford University Press, 2005.

[35] M.j. Keeling, J.V. Ross, On methods for studying stochastic disease dynamics, J. R. Soc. Interface 5 (2008) 171–181.

[36] W.O. Kermack, A.G. McKendrick, A contribution to the mathematical theory of epidemics Part I, Proc. R. Soc. Lond. A 115 (1–2) (1927) 700.

[37] S.Y. Kung, M.W. Mak, S.H. Lin, Expectation-Maximization Theory. In Biometric Authentication: A Machine Learning Approach, Pearson, 2004.

[38] K.E. Lamb, D. Greenhalgh, et al., A simple mathematical model for genetic effects in pneumococcal carriage and transmission, J. Comput. Appl. Math. 235 (2011) 1812–1818.

[39] P.E. Lekone, B.F. Finkenstadt, Statistical inference in a stochastic epidemic SEIR model with control intervention: Eboloa as a case study, Biometrics 62 (2006) 1170–1177.

[40] M.Y. Li, J.R. Graef, et al., Global dynamics of SEIR model with varying total population size, Math. Biosci. 160 (1999) 191–213.

[41] C. Liu, D.B. Rubin, Y.N. Wu, Parameter expansion to accelerate em: the px-em algorithm, Biometrika 85 (1998) 653–666.

[42] E. Mochan, D. Swigon, et al., A mathematical model of intrahost pneumococcal pneumonia infection dynamics in murine strains, J. Theor. Biol. 353 (2014) 44–54.

[43] M. Otieno, O.J. Joseph, et al., Mathematical model for pneumonia dynamics with carriers, Int. J. Math. Anal. 7 (2013) 2457–2473.

[44] M. Phelan, Inference from censored Markov chains with applications to multiwave panel data, Stoch. Process. Appl. 29 (1988) 85–102.

[45] A. Porter, J. Oleson, A path specific SEIR model for use with general latent and infectious time distributions, Biometrics 69 (2013) 101–108.

[46] S. Riley, C. Fraser, C.A. Donnelly, et al., Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions, Science 300 (2003) 1961–1966.

[47] J.V. Ross, P.K. Pollett, On parameter estimation in population models, Theor. Popul. Biol. 70 (2006) 498–510.

[48] M.D.L. Sen, S. Alonso-Quesada, et al., On the stability of an SEIR epidemic model with distributed time-delay and a general class of feedback vaccination rules, Appl. Math. Comput. 270 (2015) 953–976.

[49] A.M. Smith, J.A. McCullers, et al., Mathematical model of a three-stage innate immune response to a pneumococcal lung infection, J. Theor. Biol. 276 (2011) 106–116.

[50] L.M. Sullivan, K.A. Dukes, E. Losina, Tutorial in biostatistics. An introduction to hierarchical linear modelling, Stat. Med. 18 (1999) 855–888.

[51] G. Teshome, O. Daniel, et al., Co-dynamics of pneumonia and typhoid diseases with cost effective optimal control analysis, Appl. Math. Comput. 316 (2018) 438–459.

[52] T. Tsutsui, N. Minamib, M. Koiwai, et al., A stochastic-modeling evaluation of the foot-and-mouth-disease survey conducted after the outbreak in Miyazaki, Japan in 2000, Prev. Vet. Med. 61 (2003) 45.

[53] H.C. Tuckwell, J.R. Williams, Some properties of a simple stochastic epidemic model of sir type, Math. Biosci. 28 (2007) 76–97.

[54] D. Wanduku, Global properties of a two-scale network stochastic delayed human epidemic dynamic model, Nonlinear Anal., Real World Appl. 13 (2012) 794–816.

[55] D. Wanduku, Complete global analysis of a two-scale network sirs epidemic dynamic model with distributed delay and random perturbations, Appl. Math. Comput. 294 (2017) 49–76.

[56] D. Wanduku, Threshold conditions for a family of epidemic dynamic models for malaria with distributed delays in a non-random environment, Int. J. Biomath. 11 (2018) 1850085.

[57] D. Wanduku, The stochastic extinction and stability conditions for nonlinear malaria epidemics, Math. Biosci. Eng. 16 (2019) 3771–3806.

[58] D. Wanduku, C. Newman, O. Jegede, B. Oluyede, Modeling the stochastic dynamics of influenza epidemics with vaccination control, and the maximum likelihood estimation of model parameters, in: H. Dutta (Ed.), Mathematical Modelling in Health, Social and Applied Sciences. Forum for Interdisciplinary Mathematics, Springer, Singapore, 2020.

[59] C. Wei, M.A. Tanner, A Monte Carlo implementation of the em algorithm and the poor man's data augmentation algorithm, J. Am. Stat. Assoc. 85 (1990) 699–704.

[60] H. Woltman, A. Feldstain, J. Christine MacKay, R. Meredith, An introduction to hierarchical linear modeling, Tutor. Quant. Methods Psychol. 8 (2012) 52–69.

[61] R. Yaesoubi, T. Cohen, Generalized Markov models of infectious disease spread: a novel framework for developing dynamic health policies, Eur. J. Oper. Res. 215 (2011) 679–687.

[62] W. Yang, A. Karspeck, J. Shaman, Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics, PLoS Comput. Biol. 10 (2014) e1003583.

[63] C. Zimmer, R. Yaesoubi, T. Chen, A likelihood approach for real time calibration of stochastic compartmental epidemic models, PLoS Comput. Biol. 13 (2017) e1005257.