Computational Statistics and Data Analysis 70 (2014) 19–34

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

An *EM* algorithm for the model fitting of Markovian binary trees

Sophie Hautphenne *, Mark Fackrell

University of Melbourne, Australia

ARTICLE INFO

Article history:
Received 12 December 2012
Received in revised form 28 August 2013
Accepted 29 August 2013
Available online 5 September 2013

Keywords:
Branching process
Markovian binary tree
Parameter estimation
EM algorithm

ABSTRACT

Markovian binary trees form a class of continuous-time branching processes where the lifetime and reproduction epochs of individuals are controlled by an underlying Markov process. An *Expectation–Maximization* (*EM*) algorithm is developed to estimate the parameters of the Markov process from the continuous observation of some populations, first with information about which individuals reproduce or die (the *distinguishable* case), and second without this information (the *indistinguishable* case). The performance of the *EM* algorithm is illustrated with some numerical examples. Fits resulting from the distinguishable case are shown not to be significantly better than fits resulting from the indistinguishable case using some goodness of fit measures.

1. Introduction

A *Markovian binary tree* (*MBT*) is a continuous-time branching process in which individuals' lifetime and reproduction times are controlled by an underlying Markov phase process called a transient *Markovian arrival process* (*MAP*). The underlying phases may be purely fictitious, or may have some physical interpretation such as the actual age of an individual (see Hautphenne and Latouche, 2012), or some physiological state of the individual (see Lin and Liu, 2007 and Caswell, 2001). Reproduction and death rates usually depend on the current phase of the individual, and newborns start in a phase that may depend on their parent's phase at the time of birth. This allows for some dependence between the parent's and child's lifetimes. From an observer's point of view, the phase transitions are hidden, only the birth and death events are visible. For the sake of simplicity we assume that individuals give birth to one child at a time.

Since *MAP*s are dense in the class of counting processes (see Asmussen and Koole, 1993), *MBT*s offer considerable modeling versatility. However, they can only be applied to real systems if there is some reliable method of model fitting. That is, we need to be able to estimate the parameters of the *MBT* from measurements taken from real populations. In our case, the parameters that need to be estimated are the phase transition rates associated with the underlying Markov process when no birth or death occurs, and those transition rates when a birth or a death occurs. The problem is that these phase transitions are not observed, so we have a problem of estimation from incomplete data. For these types of statistical problems, the *Expectation–Maximization* (*EM*) algorithm has proven to be a good means of estimating the maximum likelihood estimator.

Since the 1970s a number of authors have worked on parameter estimation for branching processes, although very few have considered estimation with incomplete data. They have instead focused mainly on the properties of the maximum likelihood estimator for the offspring distribution and mean of a Galton–Watson branching process, and the distribution and mean of the number of immigrants when immigration is allowed; to cite but a few, see for instance Dion (1974, 1975), Heyde and Seneta (1972), Heyde (1974, 1975), Athreya and Keiding (1977), and Sankaranarayanan (1989). Veen and Schoenberg

* Correspondence to: Department of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia. Tel.: +61 3 8344 9073; fax: +61 3 8344 4599.

E-mail address: sophiemh@unimelb.edu.au (S. Hautphenne).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.08.015

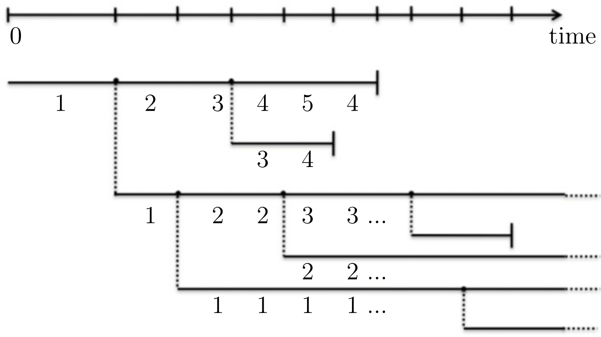

Fig. 1. Example of a trajectory of an *MBT*. Each plain horizontal line represents the lifetime of an individual, and each dotted vertical line represents the birth of a new individual. Individuals' labels are updated at each population size change.

(2008) used the *EM* algorithm to estimate the parameters of a subcritical branching process with immigration applied to seismology. In their case, the information about which earthquake event triggers each other event is unobservable and can only be described probabilistically.

In this paper we develop an *EM* algorithm for the parameter estimation of an *MBT* from the continuous monitoring of a population during a fixed time interval. The most complete information we could expect to record from one observation consists of the times at which the size of the population changes due to a birth or a death, the type of event (birth or death) associated with each size change, and the identity of the individual responsible for the size change (that is, the parent for a birth, or the individual who died) if individuals are labeled and hence distinguishable. If individuals are not labeled then they are indistinguishable. We apply the *EM* algorithm to both the distinguishable and indistinguishable cases. It turns out that the parameter estimation is not significantly better in the distinguishable case when compared to the indistinguishable case.

The paper is organized as follows. In Section 2 we formally define *MBT*s and give some background to the *EM* algorithm. Section 3 contains a detailed description of the *EM* algorithm for *MBT*s, first applied when individuals are distinguishable, and then when individuals are indistinguishable. In Section 4 we present some numerical examples and demonstrate that there is no statistically significant advantage in having distinguishable individuals. The paper concludes with some directions for further research.

2. Background

2.1. Markovian binary trees

The underlying process of a *Markovian binary tree* (*MBT*) is a Markovian counting process called a transient *Markovian arrival process* (*MAP*). This transient *MAP* represents the manner in which an individual passes through different stages in its life, until it eventually dies, and it counts how many times the individual gives birth to children (each birth corresponding to an arrival in the *MAP*). Whenever a child is born, its life follows a random path which is an independent replica of its parent's *MAP*.

Transient *MAP*s are described in Latouche et al. (2003). They are two-dimensional Markovian processes $\{(M(t), \varphi(t)) : t \in \mathbb{R}^+\}$ on the state space $\mathbb{N} \times \{0, 1, \ldots, n\}$, where n is finite. The states $(k, 0)$ are absorbing for all $k \geq 0$; the other states are transient. The process $M(t)$ counts the number of arrivals in $[0, t]$ and is called the *level* process. The process $\varphi(t)$ is a continuous-time Markov chain, called the *phase* process.

A transient *MAP* is characterized by two $n \times n$ rate matrices D_0 and D_1 and a nonnegative $n \times 1$ rate vector $\boldsymbol{d}$. Feasible transitions are from (k, i) to (k, j), for $k \geq 0$ and $1 \leq i \neq j \leq n$ at the rate $(D_0)_{ij}$, or from (k, i) to $(k+1, j)$ for $1 \leq i, j \leq n$ at the rate $(D_1)_{ij}$, or from (k, i) to $(k, 0)$ at rate d_i. The first transitions (at rate $(D_0)_{ij}$) are *hidden*: the phase of the individual changes but not the count. The second transitions (at rate $(D_1)_{ij}$) are *observable*: a birth (arrival) is recorded, at which time the state of the individual may or may not change. Finally, the third transitions (at rate d_i) indicate the termination of the individual's life.

The matrix D_1 and the vector $\boldsymbol{d}$ are nonnegative, D_0 has nonnegative off-diagonal elements and strictly negative elements on the diagonal such that $D_0\,\mathbf{1} + D_1\,\mathbf{1} + \boldsymbol{d} = \mathbf{0}$, where $\mathbf{1}$ is an $n \times 1$ vector of ones. One also defines the initial probability vector $\boldsymbol{\alpha} = (\alpha_i)_{1 \leq i \leq n}$, and we assume that $\boldsymbol{\alpha}\mathbf{1} = 1$, so that $\varphi(0) \neq 0$ a.s.

*MBT*s are continuous-time branching processes described by random collections of independent transient *MAP*s with the same dynamics. One starts at time 0 with one individual controlled by a *MAP* with parameters $(\boldsymbol{\alpha}, D_0, D_1, \boldsymbol{d})$. At the first birth, a new independent *MAP* appears and begins to evolve, while the parent *MAP* continues, possibly giving birth again, until it eventually makes a transition to its absorbing phase and dies; at that time it is removed from the system. Child *MAP*s themselves may spawn new *MAP*s. See Fig. 1 for an example of a trajectory of an *MBT*.

Each newborn starts in a phase that may depend on its parent's phase transition at the time of birth. The $n^2 \times n$ matrix P specifies how the initial phase of a child *MAP* is chosen: $P_{j,ik}$ is the conditional probability that a child starts its life in