Seminar in Statistics
Synopsis 2: *The linear birth–death process: an inferential retrospective*

Anders Gantzhorn Kristensen - tzk942

3rd November 2023

Paper Setup

In the paper: *"The linear birth–death process: an inferential retrospective"* [6], inference for the linear birth-death process is introduced. The author: Simon Tavaré initially emphasizes the model's ubiquity in probability theory. In particular, it is mentioned as an example of a Markov process, where several closed-form result exists. Tavaré continues with the special case of the constant rate linear birth-death process with birth- and death rate λ and μ respectively. The model assumptions and parameters are specified starting with one individual at time zero. After that a couple of results for the process given. He extends the model to other starting values and then provides formulas for calculation of transition probabilities, $p_{nm}(t)$ for number of individuals n and m in the beginning and end respectively over a time period, t. These formulas are compared for $t = 1$ for some values of n, m, λ, μ. While the formulas in theory yield the same answer, the answers are vastly different, due to computational difficulties that are explained. Additionally, bounds on the error one makes in the computation is discussed and how one could control this in `R`[5]. The paper then has a brief section, where it introduces two methods to simulate the process and a few estimators based in frequentism. However, the marjority of the inference is based on Bayesian methods. Firstly, the author introduces rejection sampling, but quickly realizes that the method performs poorly in this framework. Instead, Tavaré uses more sophisticated MCMC-methods and a method called "Approximate Bayesian computation" (ABC). Finally, the model is generalized to non-constant rates. In this context the paper briefly summarizes analogous concepts to previous sections, i.e. simulation, bayesian inference etc. Tavaré concludes the paper by putting the model into a more general perspective by contrasting it to other continuous time Markov chains.

Main Results

Tavaré denotes the number of individuals alive at time, t, by $Z(t)$ and the number of families $F(t)$. The latter is introduced to help derivations. For the constant rate linear birth-death process with birth- and death rate λ and μ respectively, he defines

$$\alpha(t) = \frac{\mu\left(e^{(\lambda-\mu)t} - 1\right)}{\lambda e^{(\lambda-\mu)t} - \mu}, \qquad \beta(t) = \frac{\lambda}{\mu}\alpha(t), \qquad \lambda \neq \mu. \tag{1}$$

$$\alpha(t) = \frac{\lambda t}{1 + \lambda t} = \beta(t), \qquad \lambda = \mu. \tag{2}$$

The paper gives examples of how to compute the distribution of $\{p_{nm}(t), m = 0, 1, \dots\}$ with n starting individuals. For instance the result from [1]

$$p_{nm}(t) = \sum_{j=0}^{\min(m,n)} \binom{n}{j}\binom{n+m-j-1}{n-1}\alpha(t)^{n-j}\beta(t)^{m-j}\left(1 - \alpha(t) - \beta(t)\right)^{j}. \tag{3}$$

However, as illustrated in table 1 [6], the computed values from (3) can yield subpar results, especially for larger values of n and m, where the results were nonsensical. The two other mentioned methods agree- and are a bit more stable as they use more sophisticated techniques to overcome the numerical difficulties. We refrain from writing the formulas here, due to limited space; but it is formula (11) and (17) in [6] we refer to. Hereafter Tavaré states that the MLE of λ and μ are

$$\hat{\lambda} = \frac{B_t}{S_t}, \qquad \hat{\mu} = \frac{D_t}{S_t}, \tag{4}$$

with B_t, D_t the number of births- and deaths in $[0, t]$ respectively, whereas $S_t = \int_0^t Z(u)\mathrm{d}u$. We return to these estimators in our own simulation study, however the paper only briefly applies these frequentist methods.
Instead, Tavaré lets $\theta = (\lambda, \mu)$ and $\mathcal{D} = (F_1, \ldots, F_m, Z_1, \ldots, Z_m)$ and continues with bayesian inference. He proposes three methods for simulating from the posterior, $f(\theta|\mathcal{D})$. Yet, he only goes with MCMC and the ABC (approximate bayesian computation) method. In his simulation study Tavaré finds that the ABC methods performs worse than MCMC. While both methods on average and in median estimates the parameters quite well; the posterior distribution from the ABC method is overdispersed when compared to MCMC's (see table 4 and figure 2 and -4 [6]). Finally, the paper considers the theory of time-dependent rates and provides inference in an example where

$$\mu(t) \equiv \mu, \qquad \lambda(t) = \lambda + \frac{\gamma}{t+1}, \tag{5}$$

for some parameters (μ, λ, γ) and does bayesian inference on these with the ABC method.

Own Contribution

As the paper focuses on bayesian methods, we tried delving into estimation with the MLE from (4). The simulation study is done in `R` [5] and we simulate using the wait–jump–wait construction [6]. Though, it quickly became apparent that is was necessary to use `C++` [2] for these simulations. As a precision metric we use the absolute relative error (ARE); this is the center for the study. More precisely, we consider the median and the 2.5%- and 97.5%-quantiles of the ARE over the grid of starting values, $n \in \{1, 2, 3, 5, 10, 25\}$ and max observation time, $T \in \{5, 7.5, 10, 15, 20, 25, 30, 40\}$; we run each combination 100 times and calculate the MLE and ARE. We also do a small benchmark with the `microbenchmark`-method [4]. The illustrations are made with ggplot2 [7] and shown in the appendix apart from one plot. We consider the case when $(\lambda_0, \mu_0) = (0.6, 0.3)$. As is evident from figure 2 we have quite accurate estimates in the median- and better cases. However, due to rapid extinction the worse cases have no way of providing sensible estimates. This is also clear from the fact that the worse cases does not become better with increasing T in the same way as the median- and better cases. This is expected as the asymptotic results are also only provided conditional on non-extinction [3]. To illustrate this problem, we consider the case where the parameters swap values, i.e. $(\lambda_0, \mu_0) = (0.3, 0.6)$
It is evident from figure 1 that with a larger death rate early extinction becomes more likely; we get a bad worse case even for $n = 5$. What is true in both cases, though, is

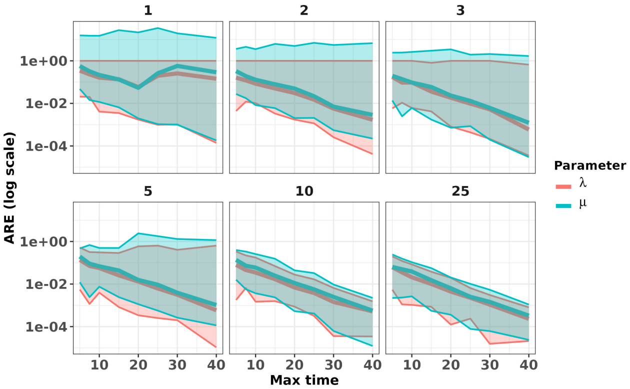

Figure 1: Median, 2.5%- and 97.5%-quantiles for ARE stratified after n and parameter with $(\lambda_0, \mu_0) = (0.6, 0.3)$.

that we do not gain much precision going above $n = 10$ for fixed T. Conversely, we gain orders of magnitude of precision, when increasing T. However, as is clear in figure 3, this increases the computation time dramatically. In addition, we see that it generally is harder to estimate μ than λ. This might be, because we need many deaths without extinction for this, and that can be hard to achieve.

Strengths & Weaknesses

We briefly conclude this synopsis by touching on strengths and weaknesses of the paper. For starters, the paper's main shortfall is that it provides no novel results; it merely reiterates well-known findings. On the other hand, the paper is quite readable even without any prior knowledge about the subject. In addition, the process in the simulation studies is clearly explained, and upon request one can acquire the source code, which is considered good scientific practise. On the contrary, the paper seems to be overly focused closed-form solutions, using it as the primary mean of motivating the model. *"Explicit availability of transition functions makes these (linear birth-death ed.) models a useful calibration for other computational approaches"*[6]. While mathematically nice to look at, the intractability of formula (9)[6], should illustrate that in practise, approximations can be more reliable than estimates from explicit formulas and there is no reason to a priori prefer one over the other. Additionally, the paper can be criticised for choosing the ABC method over MCMC in the inference for the general model. It performed worse than MCMC in the constant rate case, and many of the choices involved in the method seem somewhat arbitrary. In particular, is it not clear how well the metric from formula (26) does generally; especially when the dimension of the problems grows. On a similar note the author uses much space on simple model and difficulties with estimation in it. Though, one would expect that estimation in the non-homogeneous model can be much more difficult than in the homogeneous - and the simple non-homegenous cases shown in the paper.

References

[1] N. T. J. Bailey. *The Elements of Stochastic Processes with Applications to the Natural Sciences.* John Wiley, New York, 1964.

[2] Dirk Eddelbuettel and James Joseph Balamuta. Extending ir/i with c: A brief introduction to rcpp. *The American Statistician*, 72(1):28–36, January 2018.

[3] Niels Keiding. Maximum likelihood estimation in the birth-and-death process. *The Annals of Statistics*, 3(2), March 1975.

[4] Olaf Mersmann. *microbenchmark: Accurate Timing Functions*, 2023. R package version 1.4.10.

[5] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2023.

[6] Simon Tavaré. The linear birth-death process: an inferential retrospective. *Adv. Appl. Probab.*, 50(A):253–269, December 2018.

[7] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

Appendix: Plots from the simulation study

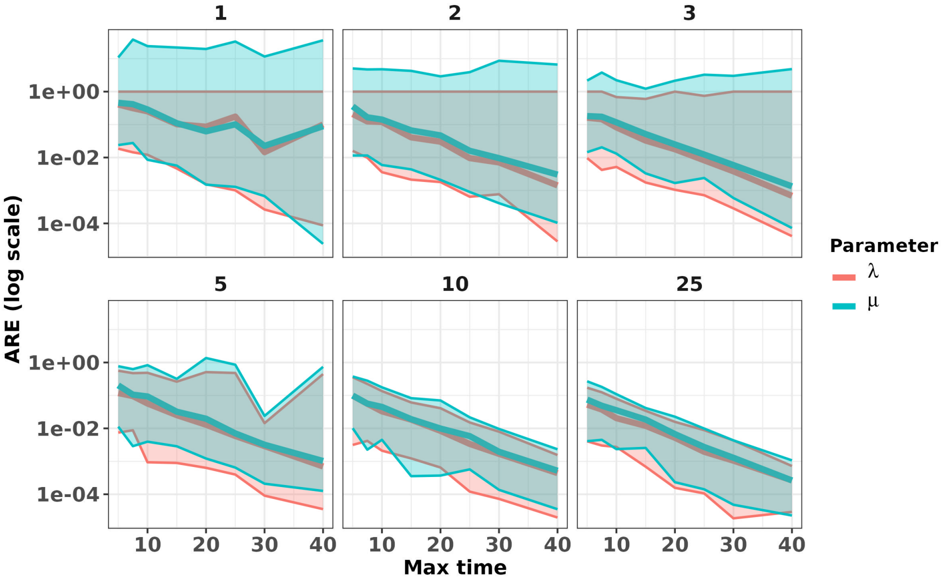

Figure 2: Median, 2.5%- and 97.5%-quantiles for ARE stratified after n and parameter with $(\lambda_0, \mu_0) = (0.3, 0.6)$.

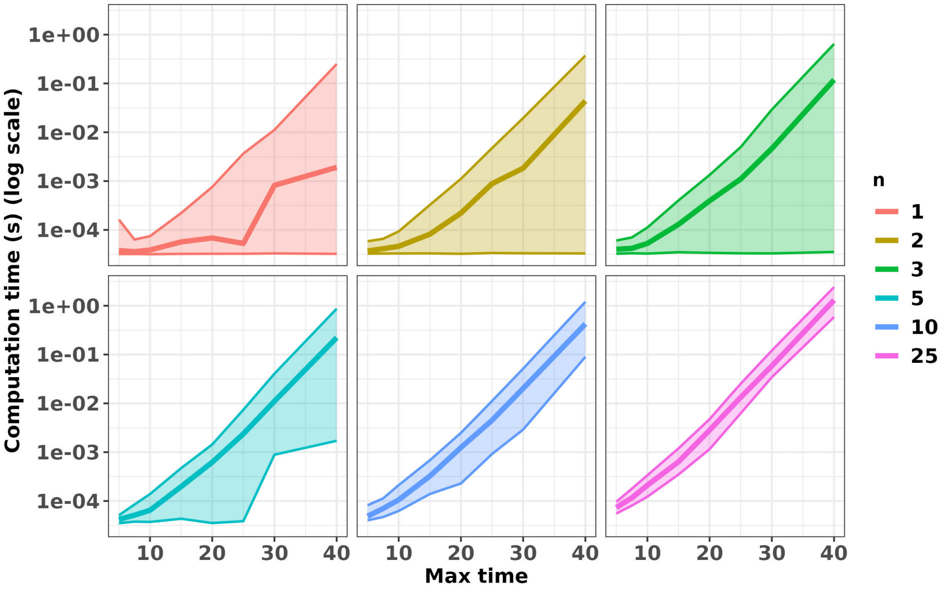

Figure 3: Median, 2.5%- and 97.5% - quantiles for simulation and computation of MLEs stratified after n.