Optimal Sampling for Population Models

Daniel E. Pagendam

University of Queensland, Brisbane, Australia

Introduction: *Birth-death processes* are useful statistical models for a range of natural phenomena, such as the dynamics of populations or the spread of epidemics. In most practical situations, population data is collected at a discrete number of points in time and there are practical constraints governing the number of samples that can be taken and the time frame for the sampling. Given these constraints, a pertinent question emerges: at what times should we sample the population in order to obtain the most precise estimates of model parameters (e.g. infection rates, per capita birth and death rates)?

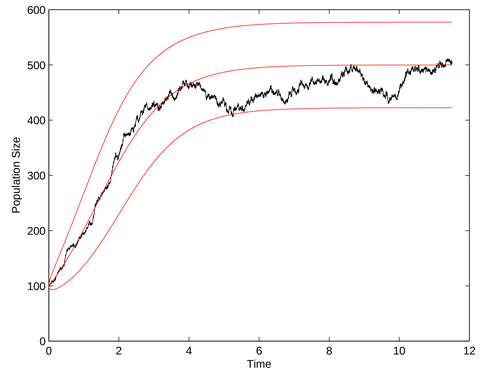

Figure 1: *An example of the stochastic logistic (SL) process, showing the deterministic trajectory ± 2 standard deviations (red).*

Approach:

- We use a *Gaussian diffusion approximation* of a stochastic logistic (SL) process to estimate the likelihood of the collected data.
- The approximation models the fluctuations of the population about a deterministic trajectory using a multivariate Gaussian probability distribution.
- Numerical methods (such as the Cross-Entropy method) can be used to find the sampling times that maximise the amount of information (Fisher Information) about the unknown parameters.
- The more information that is obtained about the parameters, the smaller the area of confidence regions about the parameter estimates.

Methods: In order to estimate parameters and design efficient sampling schedules, it is necessary to calculate the likelihood of data. However, the likelihood for many stochastic processes is difficult to calculate analytically. Many birth-death processes can be approximated by a *Gaussian diffusion* about a deterministic trajectory so that the likelihood for n samples is n-variate Gaussian, which is easy to work with numerically and analytically.

For the SL process $Y(t)$ with population ceiling N (i.e. maximum population size), the transition rates q_N when the population size is i are

$$q_N(i,i+1) = \frac{\lambda}{N} i(N-i)$$
$$q_N(i,i-1) = \mu i$$

However, it is convenient to model the population "density" $X_N(t) = \frac{Y(t)}{N}$ which describes the proportion of individuals in the population relative to N. The rationale is that as $N \to \infty$, the density process $X_N(t)$ tracks the deterministic trajectory (see [2]), given by

$$x(t) = \frac{x_0 x_{eq}}{x_{eq} + (x_0 - x_{eq})e^{-\lambda x_0 t}} \quad (t \geq 0)$$

where x_0 is the population density at $t = 0$ and $x_{eq} = 1 - \frac{\mu}{\lambda}$ is the stable equilibrium population density (also known as the carrying capacity).

Furthermore, the fluctuation $Z_N(t) = \sqrt{N}(X_N(t) - x(t))$ of the population about the deterministic trajectory converges weakly to a Gaussian diffusion $Z(\cdot)$ (see [1, 3]), such that $Z(t) \sim \mathrm{N}(0, V(t))$, where

$$V(t) = \frac{x_0 x_{eq}}{\alpha(x_0 + (x_{eq} - x_0)e^{-\alpha t})^4}\Big(\mu x_0^3(1 - e^{-2\alpha t}) + x_0^2(x_{eq} - x_0)e^{-\alpha t}(1 - e^{-\alpha t})(\lambda + 5\mu) + 2x_0\alpha t e^{-2\alpha t}(x_{eq} - x_0)^2(\lambda + 2\mu) + \beta(x_{eq} - x_0)^3 e^{-2\alpha t}(1 - e^{-\alpha t})\Big)$$

Our diffusion approximation allows us to model a finite set of observations $\mathbf{y} = (y_1, \ldots, y_n)$ of the SL process $Y(t)$ at times $(t_1, \ldots, t_n)$ as a random vector with an approximate multivariate normal (MVN) distribution. We write $\mathbf{y} \sim \mathrm{MVN}(\mathbf{m}, \Sigma)$, with

$$m_i \simeq N x(t_i) \qquad \text{and} \qquad \Sigma_{ij} \simeq \begin{cases} NV(t_i)e^{2(A(t_j) - A(t_i))} & \forall i < j \\ NV(t_i) & \forall i = j \end{cases}$$

where $A(t_i) = 2\log(x_{eq}) - \lambda x_{eq} t_i - 2\log(x_0 + (x_{eq} - x_0)e^{-(\lambda - \mu)t_i})$. Figure 1 shows a simulated SL process with the deterministic trajectory ($Nx(t)$) and lines corresponding to ± 2 standard deviations (i.e. $\pm 2\sqrt{NV(t)}$) superimposed.

Results: For comparative purposes we demonstrate the benefits of using the optimal sampling schedule by comparing it to an equidistant sampling schedule (which is likely to be simpler to implement in practice), using simulated data.

- The optimal sampling schedule can result in much greater precision when estimating parameters.
- Use of the optimal sampling schedule can overcome problematic likelihood surfaces that can be encountered in population models.
- The relative benefit (efficiency) of the optimal sampling schedule over the equidistant schedule is often greatest when the number of samples is relatively small.

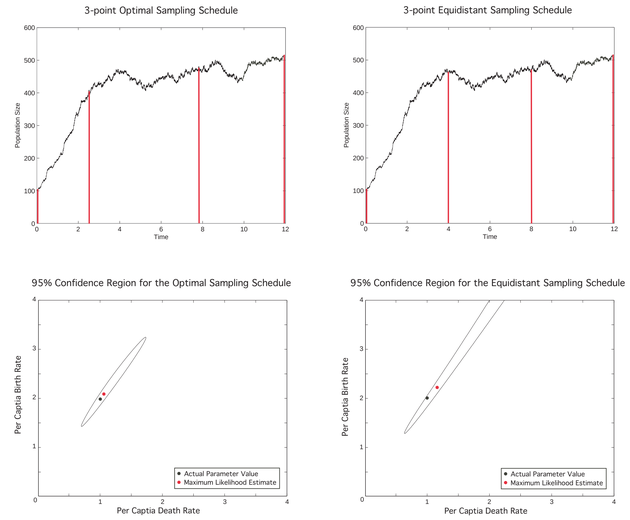

Figure 2: *Comparison of 95% confidence regions for optimal and equidistant sampling schedules with 3 sampling points.*

Conclusions:

- Gaussian diffusion approximations coupled with numerical optimisation methods such as the Cross-Entropy method allow us to easily determine optimal sampling schedules for a wide range of population models.
- Optimal sampling schedules can significantly increase the precision of parameter estimates.
- Optimal sampling schedules can be used to design efficient experiments involving the observation of birth-death processes over time (e.g. experiments in cell biology, epidemiology and chemistry).

References:

[1] Barbour, A. 1974. On a functional central limit theorem for Markov population processes. *Adv. Appl. Probab.* **6(1)**, 21-39.

[2] Kurtz, T. 1970. Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Probab.* **7(1)**, 49-58.

[3] Kurtz, T. 1971. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab.* **8(2)**, 344-356.