Simulation Based Estimation for Birth and Death Processes

Katherine B. Ensor
Eileen Bridges
Martin Lawera

UMSA Presentation March 19, 2002
Motivation and Background

- Difficult to obtain likelihood equation from well understood axioms derived from stochastic models.
- Deterministic approximations often used which can be solved more easily.
- What if a stochastic component is still necessary?
- SIMEST - alternative to using deterministic models to simplify.
  - Motivated by the work of other scientists in modeling cancer progression.
  - Premise of previous work was to bypass likelihood equations by estimating directly from model assumptions.
- **PROBLEM**: Representing birth-death models.
Examples of Birth-Death Processes

- **Measles Epidemic**
  - Birth - indicates an increase in infective persons. Increase is proportional to number of infective persons and susceptible persons.
  - Death - indicates a decrease in infective persons due to recovery or death.

- **Political Party**
  - Birth - increase in number spreading campaign doctrine. Increase is proportional to number of “spreaders” and number of “susceptibles”
  - Death - decrease in active spreaders. **NOTE:** Different from the measles model in that death can be temporary.

- **Marketing**
  - Birth - entry of a new product to the market. Proportional to advertising and potential customers.
  - Death - withdrawal of product from the market.
SIMEST - Criterion Function

- How should we compare simulated and observed data?

- Estimate \( \theta \in \Theta \) such that for \( m \) realizations of a process, \( S_n(\theta) \) which denotes the difference between simulated and actual results is minimized.

- Consider the sample \( t_1, \ldots, t_n \) from the stochastic process \( \{W(s), s \geq 0\} \) which represent the waiting time until the \( s^{th} \) event.

- Simulate \( m \) observations of this process.

- Divide the time axis into bins and let \( \hat{p}_1, \ldots, \hat{p}_k \) represent the proportion \( n \) observations falling into a given bin.

- Let \( \tilde{p}_1(\theta), \ldots, \tilde{p}_k(\theta) \) denote the proportion of simulated data points in each of the bins.

- Use Pearson goodness of fit statistic:

\[
S_n(\theta) = \sum_{j=1}^{k} \frac{(\tilde{p}_j(\theta) - \hat{p}_j)^2}{\tilde{p}_j(\theta)}
\]

- Estimator of \( \theta \) is the value \( \hat{\theta} \in \Theta \) which minimizes \( S_n(\theta) \)
SIMEST - Single Realization

• Instead of $n$ different values of $N(t)$, consider one value each for $N(t_1), \ldots, N(t_n)$.

• Consider the following Birth-Death process:
  
  - Process $N(t)$ has parameters $\lambda_n$ and $\mu_n$
  
  - $P(N(t + \Delta t) = n + 1|N(t) = n) = \lambda_n \Delta t + o(\Delta t)$
  
  - $P(N(t + \Delta t) = n - 1|N(t) = n) = \mu_n \Delta t + o(\Delta t)$

• We can derive the following distributions of the next birth and death from the above:
  
  - $F_B(t) = 1 - P\{0 \text{ births in } (t, t + \Delta t)\} = 1 - e^{-\lambda_n t}$
  
  - $F_D(t) = 1 - P\{0 \text{ deaths in } (t, t + \Delta t)\} = 1 - e^{-\mu_n t}$

• Using the inverse cdf transformation we obtain time until next birth or death:
  
  - $t_B = \frac{\log(\lambda_n)}{U_1}$
  
  - $t_D = \frac{\log(\mu_n)}{U_2}$

• With $U_1$ and $U_2$ independent, uniformly distributed random variables.
SIMEST - Simulation of $N(t)$ and Goodness of Fit

- To simulate $N(t)$ we use the following algorithm:
  
  1. Generate $U_1 and U_2$
  2. Compute $t_B$ and $t_D$
  3. Set $t = t + \min(t_B, t_D)$
  4. If $t_D < t_B$ then $N(t) = N(t) - 1$, else $N(t) = N(t) + 1$
  5. If $t < \max t$ and if $N(t) > 0$ go to 1, otherwise stop

- Determining goodness of fit: extend previous function.

- Bin the time access as discussed before, but this time define $\hat{n}_1, \ldots, \hat{n}_k$ as the observed value of $N(t)$ at the right endpoint of the bin.

- Let $\tilde{n}_1 (\theta), \ldots, \tilde{n}_k$ denote the average value of the $m$ simulated realations at the respective times.

- Goodness of fit function:

\[
S_n(\theta) = \sum_{j=1}^{k} \frac{(\tilde{n}_j (\theta) - \hat{n}_j)^2}{\hat{n}_j (\theta)}
\]
SIMEST - Alternate Goodness of Fit

- Possibly the number of observed births and deaths by time $t$ is a better measure than total number at time $t$.

- Let $\hat{n}_{b1}, \ldots, \hat{n}_{bk}$ and $\hat{n}_{d1}, \ldots, \hat{n}_{dk}$ indicate the number of observed births and deaths.

- Let $\tilde{n}_{b1}, \ldots, \tilde{n}_{bk}$ and $\tilde{n}_{d1}, \ldots, \tilde{n}_{dk}$ indicate the number of simulated births and deaths.

- Goodness of fit function:

$$S_n(\theta) = w \sum_{j=1}^{k} \left( \frac{\tilde{n}_{b_j}(\theta) - \hat{n}_{d_j}}{\tilde{n}_{b_j}(\theta)} \right)^2 + (1 - w) \sum_{j=1}^{k} \left( \frac{\tilde{n}_{s_j}(\theta) - \hat{n}_{d_j}}{\tilde{n}_{d_j}(\theta)} \right)^2$$

- With $w$ being some appropriate weight function.

- **NOTE:** Separating births and deaths avoids a cancelling effect, and is crucial to the estimation process.
Advantages to using SIMEST

• SIMEST leads to strongly consistent estimators of the parameters when estimating from \( n \) independent and identically distributed observations.

• Fairly easy to develop confidence intervals.

• Fairly easy to obtain information on the mean and variance.

• Easily used on parallel systems.

• Allows implementation of stochastic models without solving differential or difference equations.

• Can recover correct parameters when mean path of birth and death process is used as input.
Disadvantages to using SIMEST

- For varying $N$, the estimates provided by SIMEST are suspect.