Presentation on

Data Cloning

Jennifer La Rosa
Laxman Ghimire

12 November

Outline of Presentation

- What is Data Cloning
- Why Data Cloning
- Where can it be used?
- Example: Estimation of unknown parameters from the Leslie's projection matrix

  1. Introduction
  2. Statistical Models
  3. Data Cloning based on Bayesian Approach
  4. Using Real Data
  5. Algorithm
  6. Result
- Conclusion

What is Data Cloning ?

Cloning: Literal Meaning

- In biology, cloning is the process of producing similar populations of genetically identical individuals

- The term also refers to the production of multiple copies of a product

Cloning : In statistical sense

The data cloning method is a general technique to compute maximum likelihood estimates along with their asymptotic variances by means of the computation of the posterior distributions by using a MCMC methodology (Lele et al. (2007) and Lele et al. (2010)).

$$\pi^k(\xi, \theta|y) \overset{k\to\infty, D}{\longrightarrow} N\left(\begin{pmatrix} \hat{\xi} \\ \hat{\theta} \end{pmatrix}, \frac{1}{k} I^{-1}\left(\hat{\xi}, \hat{\theta}\right) \right)$$

What is Data Cloning ? (Contd.)

- Data Cloning applies Bayesian prior distributions and MCMC simulations to k copies (clones) of the data.

- If the number of clones is large, the sample mean vector of the resulting simulated posterior distribution corresponds to the maximum likelihood (ML) estimates of the parameters

Data cloning methods have been developed to tackle with ecological complex models (Lele et al. (2007) and Lele et al. (2010)).

Estimation two Unknown Parameters ‘Fertility Rates’ and ‘Survival Rates’ using Data Cloning

Introduction I

- We consider discrete time models for describing the evolution of an age-structured population
- The population is divided into k groups or intervals of age, each interval of age having the same length
- We assume that the unit of time is the same as the age class width, and it is called the projection interval.
- The length of all the intervals of age depends on the population we are studying: one week, six months, one year, 15 years,... (Depending on the reproductive cycles)

Let,
s_i = The survival rate (i = 1, . . . , k-1)
= the proportion of individuals of group i which will survive to the next period of time (becoming individuals of group i + 1).

f_i= The reproductively or fertility rate (for i = 1, . . . , k)
= the average number of surviving offsprings of each individual of group i.

Introduction II

And,

Let N_i (t) (for i = 1, ..., k) be the number of individuals of group i in a given period of time, t.

Then,

The relationship between consecutive periods of times can be expressed as

$$
\begin{array}{lll}
N_1(t) & = & f_1N_1(t-1) + \cdot\cdot\cdot + f_k\, N_k\, (t-1) \\
N_2(t) & = & s_1N_1(t-1) \\
N_3(t) & = & s_2N_2(t-1) \\
\dots & \dots & \dots\ \dots\ \dots \\
\dots & \dots & \dots\ \dots\ \dots \\
N_k(t) & = & s_{k-1}\, N_{k-1}\, (t-1)
\end{array}
$$

- **These equations can also be formulated in the matrix form.**
- **The matrix is usually called population projection matrix.**
- **The matrix is also called Leslie matrix.**

The statistical Models

We consider two statistical models

1.Statistical model for estimating fertility rates

Let,

In a determined period of time, there are $N_j(t-1)$ individuals in the j th group of age (for $j = 1, \ldots, k$).

For the next period of time, we expect to have about

$N_1(t) = f_1 N_1(t-1) + \cdots + f_k N_k(t-1)$ individuals in the first group of age,

i.e., $N_1(t)$ is as an expected size for the following period of time

The total number of offsprings of the $N_j(t-1)$ individuals of group j is the randomvariable $D_j(t-1)$

So, $D_j(t-1)$ is the sum of $N_j(t-1)$ independent and identically distributed random variables

1.Statistical model for estimating fertility rate (contd)

Therefore, the distribution of D_j (t − 1) can be approximated by a Normal distribution with expectation $f_j N_j (t-1)$ (provided that N_j (t − 1) is large enough).

So, the total number of offsprings for the next period of time is

$N_1(t) = D_1(t-1) + \cdot\cdot\cdot + D_k (t-1)$,

Where, the distribution of $N_1(t)$ can be approximated by a Normal distribution, provided that N_j (t − 1), for j = 1, . . . , k, are large enough.

In this way, $N_1(t)$ can be taken as a random variable.

The sampling density for this random variable is

$$N_1 (t) \sim N (f_1N_1(t-1) + \cdot\cdot\cdot + f_k N_k (t-1); \sigma_1),$$

Where,

$f_1 , . . . , f_k$ are unknown parameters (of interest)

and σ_1 is A (nuisance) unknown parameter.

2. Statistical model for estimating survival rates

Let,

In a determined period of time, there are $N_1(t-1)$

individuals in the first group of age.

For the next period of time,

We expect to have about $N_2(t)$ individuals in the second group of age,

i.e., $N_2(t)$ can be considered as an expected size for the following period of time.

So, the survival rate, s_1, is an unknown parameter,

Each animal of the $N_1(t-1)$ individuals of group 1 may survive to the next period of time with probability $0<s_1<1$.

$N_2(t)$, is a random variable with Binomial distribution,

$B(N_1(t-1); s_1)$.

The distribution of $N_2(t)$ can be approximated by a Normal distribution, provided that $N_1(t-1)$ is large enough.

In this way, $N_2(t)$ is a random variable.

The sampling density for this random variable is

$$N_2(t) \sim N(s_1 N_1(t-1); \sigma_2)$$

2. Statistical model for estimating survival rates

Where,

s_{j-1} is an unknown parameter (of interest)

and σj is a (nuisance) unknown parameter (for j = 3, . . . , k).

In the same way, $N_j(t)$ (for j = 3, . . . , k) is a random variable.

The sampling density for this random variable is

$$N_j\,(t) \sim N\,(s_{j-1}\,N_{j-1}(t-1);\,\sigma j\,)$$

Data Cloning Based on Bayesian Approach

The prior distribution unknown parameters θ consists of

($f_1, . . . , f_k , \sigma_1, . . . , \sigma_k, s_1, . . . , s_{k - 1}$).

1. The prior distributions considered for the parameters,

- **Log normal distributions for $f_1, . . . , f_k$**
- **Uniform distributions on (0, 1) for $s_1, . . . , s_{k-1}$**
- **Inverse-gamma distributions for $\sigma_1 , . . . , \sigma_k$.**

2. Posterior distributions

We generate samples from the posterior distribution, $\pi^{(k)}(\theta|n)$ that is proportional to the k th power of the likelihood, $[L(\theta|n)]^k$, multiplied by a proper prior distribution, $\pi(\theta)$.

Using Real Data

Real data from the population of the Steller sea lions located in the Alaska coast

The data applied since 1978 to 2004. (Holmes et al. (2007)

Data were collected along 27 years

There are several years with partial or complete missing observations.

Algorithms

- The expression $[L(\theta|n)]^k$ is the likelihood for k copies of the original data
- For large k , $\pi^{(k)}(\theta|n)$ converges to a multivariate normal

  distribution with mean equal to the ML estimate of the parameters
- And covariance matrix equal to 1/k times the inverse of the Fisher information matrix for the ML estimates (Lele et al. (2007))

➢In this way, after obtaining samples from the posterior distribution from a MCMC procedure, we compute the sample means, and they provide an approximation of the maximum likelihood estimates of the parameters

·

Algorithms

Step 1

Create k-cloned data set $n^{(k)} = (n, n, . . . , n)$, where the observed data vector is repeated k times.

Step 2

Using an MCMC algorithm, generate random numbers from the posterior $n^{(k)} = (n, n, . . . , n)$, distribution that is based on a prior π(θ.) and the cloned data vector

Step 3

Compute the sample mean and variances of the values $(\theta)_j$, $j = 1, . . . , M$ (for M iterations of the MCMC run) generated from the posterior distribution.

The ML estimates of $(\theta)_j$ correspond to the posterior mean values and the approximate variances of the ML estimates correspond to k times the posterior variances.

Result from the Steller Sea Lions Data

As there is an important number of missing data, classical techniques do not work well.

But by means of data cloning we can use the Bayesian approach to compute the predictive distributions of the missing observations in a natural way.

Then, we obtain the ML estimators derived from the posterior distributions of the parameters.

Employed the programmed the algorithm using package dclone from the Rproject

The number of clones used 50.

The confidence intervals (95%) for the parameters, based on, are shown in table

Parameters	95 % confidence interval
f_2	0.6423, 0.7375
s_1	0.9934, 1.0057
σ_1^2	0.4956, 1.3405
σ_2^2	2.1266, 4.3699
λ	0.8017, 0.8591
eigen1	0.4452 , 0.4624
eigen2	0.5376, 0.5548

Conclusion

1. Discrete time models are used in Ecology for describing the evolution of an age-structured population.

2. The statistical model is a reasonable model for the case in which the evolution of the population is described by means of a Leslie matrix.

3. Fertility rates and survival rates are unknown parameters and they were estimated by using Data Cloning based on the Bayesian approach.

4. Data cloning is a general technique to approximate maximum likelihood estimates along with their asymptotic variances by means of the computation of the posterior distributions by using a MCMC methodology.

5. The data cloning method is applied to the real data.

Thank you