Bayesian inference and data cloning in population projection matrices

Julián de la Horra[1], Juan Miguel Marín[2] and María Teresa Rodríguez-Bernal[3]

[1]Department of Mathematics, Universidad Autónoma de Madrid, Spain [2]Department of Statistics, Universidad Carlos III de Madrid, Spain
[3]Department of Statistics, Universidad Complutense de Madrid, Spain

Introduction

- We consider a discrete time model for describing the evolution of an age-structured population, which is divided into k groups or intervals of age.
- For each group or interval of age, we need to specify two rates:
  - The survival rate, s_i (for $i = 1, \ldots, k-1$), namely, the proportion of individuals of group i which will survive to the next period of time (becoming individuals of group $i+1$).
  - The reproductivity or fertility rate, f_i (for $i = 1, \ldots, k$), namely, the average number of surviving offsprings of each individual of group i.
- Let us denote by $N_i(t)$ (for $i = 1, ..., k$) the number of individuals of group i in a given period of time, t.
- The relationship between consecutive periods of times can be expressed by means of the following *Leslie matrix*:

$$\begin{pmatrix} N_1(t) \\ N_2(t) \\ N_3(t) \\ \vdots \\ N_k(t) \end{pmatrix} = \begin{pmatrix} f_1 & f_2 & \cdots & f_k \\ s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & s_{k-1} & 0 \end{pmatrix} \begin{pmatrix} N_1(t-1) \\ N_2(t-1) \\ N_3(t-1) \\ \vdots \\ N_k(t-1) \end{pmatrix}$$

The statistical problem

- $N_1(t)$ must be understood as a random variable with sampling density

$$N_1(t) \sim N(f_1 N_1(t-1) + \cdots + f_k N_k(t-1); \sigma_1),$$

where $f_1, \ldots, f_k$ and σ_1 are unknown parameters.

- In the same way, $N_j(t)$ (for $j = 2, \ldots, k$) must be understood as a random variable with sampling density

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

where s_{j-1} and σ_j (for $j = 2, \ldots, k$) are unknown parameters.

Bayesian approach

- Let us assume that we have observed $\mathbf{n}(t) = (n_1(t), \ldots, n_k(t))$ for $t = 1, \ldots, m$. We will use Bayesian MCMC algorithm for making inferences on the parameters, $f_1, \ldots, f_k$, $\sigma_1^2, \ldots, \sigma_k^2$ and $s_1, \ldots, s_{k-1}$.
- We take as prior distributions for the parameters:

$$f_j \sim \log N(\mu_j, \tau_j^2),$$
$$\sigma_j^2 \sim IGamma(\alpha_j, \beta_j),$$

for $j = 1, \ldots, k$ and

$$s_j \sim U(0, 1)$$

for $j = 1, \ldots, k-1$.

Application to real data

- In Holmes et al. (2007), the population of the Steller sea lions (*Eumetopias jubatus*) located in the Alaska coast is studied with an age-structured model from a frequentist point of view. It is observed a significant decline in the population of sea lions. Data were collected along 27 years since 1978 to 2004, although there are several years with partial or complete missing observations. Data consist of two groups of age: pup and adult classes.
- We apply the Bayesian MCMC algorithm in order to analyze these data.
- The original deterministic equations are:

$$N_1(t) = f_2 N_2(t-1)$$
$$N_2(t) = s_1 N_1(t-1)$$

where f_2 and s_1 are the parameters of the models.

- We assign vaguely informative prior distributions: log-normal distribution with mean equal to 0 and variance equal to 100, for f_2; uniform distribution between 0 and 1, for s_1; inverse-gamma distribution with mean equal to 1 and variance equal to 10 for σ_2^2.
  Then, we run 3 chains with a total number of 20000 iterations (10000 to burn-in) and thinning equal to 5.
- The posterior means, standard deviations and quantiles of the corresponding chains of each parameter are shown in Table 1.

	Mean	SD	2.5%	50%	97.5%
f_2	0.6911	0.0274	0.6403	0.6900	0.7490
s_1	0.9753	0.0261	0.9031	0.9837	0.9994
σ_1^2	1.0446	0.2679	0.6632	0.9993	1.7091
σ_2^2	3.5695	0.6870	2.4960	3.4756	5.1909
λ	0.8208	0.0199	0.7789	0.8215	0.8579
eigen1	0.4570	0.0059	0.4464	0.4566	0.4701
eigen2	0.5430	0.0059	0.5299	0.5434	0.5536

Table 1: Statistics of the simulated posterior distributions of parameters.

- In this model, the estimated kernel densities from the MCMC samples of the posterior distributions are unimodal, and a *post hoc* analysis of the chains did not show a significant departure from convergence.

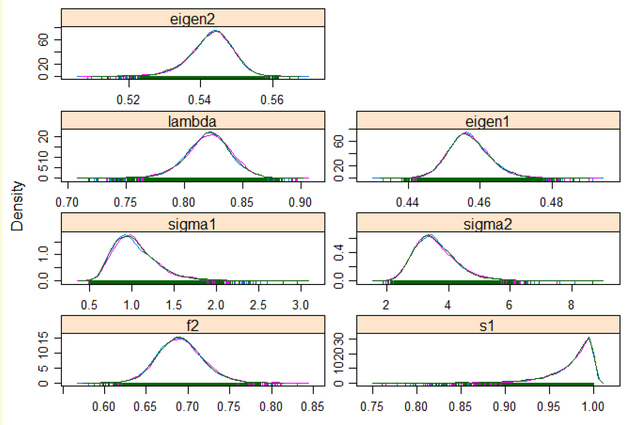

Figure : Density plots of the posterior distributions of parameters

Data Cloning

- 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 (see Lele et al. (2007) and Lele et al. (2010)).
- The data cloning algorithm can be summarized in the following steps:
  - **Step 1**: Create k-cloned data set $\mathbf{n}^{(k)} = (\mathbf{n}, \mathbf{n}, \ldots, \mathbf{n})$, where the observed data vector is repeated k times.
  - **Step 2**: Using an MCMC algorithm, generate random numbers from the posterior distribution that is based on a prior $\pi(\theta)$ and the cloned data vector $\mathbf{n}^{(k)} = (\mathbf{n}, \mathbf{n}, \ldots, \mathbf{n})$, where the k copies of $\mathbf{n}$ are assumed to be independent of each other. In practice, any proper prior distribution can be used.
  - **Step 3**: Compute the sample mean and variances of the values $(\theta)_j$, $j = 1, \ldots, M$ (for M iterations of the MCMC run) generated from the marginal 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.
- We complete the analysis of the Steller sea lions data by applying the data cloning technique. The confidence intervals (95%) for the parameters, based on the Wald approximation, are shown in Table 2.

	2.5%	97.5%
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

Table 2: Confidence intervals (95%) for parameters

References

- Holmes E.E., Fritz L.W., York A.E., and Sweeney K. (2007). Age-Structured Modeling Reveals Long-Term Declines in the Natality of Western Steller Sea Lions. *Ecological Applications* 17(8), 2214–2232.
- Lele, S.R., Dennis, B., and Lutscher F. (2007). Data Cloning: Easy Maximum Likelihood Estimation for Complex Ecological Models Using Bayesian Markov Chain Monte Carlo Methods. *Ecology Letters* 10, 551–563.
- Lele S.R., Nadeem K., and Schmuland B. (2010). Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning. *Journal of the American Statistical Association* 105, N. 492, 1617–1625.
- Leslie, P. (1945). On the Use of Matrices in Certain Population Mathematics. *Biometrika* 33, 183–212.
- Plummer M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. *DSC 2003 Working Papers*. `http://www-fis.iarc.fr/~martyn/software/jags`
- Robert, C.P., and Casella, G. (2004). Monte Carlo Statistical Methods. 2nd ed. New York: Springer.