UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 02
Statistics and Econometrics Series 01
January 2013

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

TITULO

Bayesian inference and data cloning in population projection matrices

AUTOR / AUTORES*

Horra, Julián de la, Departamento de Matemáticas , Universidad Autónoma de Madrid, e-mail: julian.delahorra@uam.es.
Marín, J.Miguel, Departamento de Estadística, Universidad Carlos III de Madrid, e-mail: jmmarin@est-econ.uc3m.es.
Rodríguez-Bernal, María Teresa, Universidad Complutense de Madrid, e-mail: mayter@mat.ucm.es.

Abstract
Discrete time models are used in Ecology for describing the evolution of an age-structured population. Usually, they are considered from a deterministic viewpoint but, in practice, this is not very realistic.
The statistical model we propose in this article is a reasonable model for the case in which the evolution of the population is described by means of a projection matrix.
In this statistical model, fertility rates and survival rates are unknown parameters and they are estimated by using a Bayesian approach.

Usual Bayesian and data cloning methods (based on Bayesian methodology) are applied to real data from the population of the Steller sea lions located in the Alaska coast since 1978 to 2004. The estimates obtained from these methods show a good behavior when they are compared to the actual values.

Keywords: Population projection matrices, data cloning, age-structured population, Leslie matrix, Bayesian MCMC algorithm.

Bayesian inference and data cloning in population projection matrices

Julián de la Horra
Universidad Autónoma de Madrid

J. Miguel Marín
Universidad Carlos III de Madrid

María Teresa Rodríguez-Bernal*
Universidad Complutense de Madrid

ABSTRACT

Discrete time models are used in Ecology for describing the evolution of an age-structured population. Usually, they are considered from a deterministic viewpoint but, in practice, this is not very realistic. The statistical model we propose in this article is a reasonable model for the case in which the evolution of the population is described by means of a projection matrix. In this statistical model, fertility rates and survival rates are unknown parameters and they are estimated by using a Bayesian approach.

Usual Bayesian and data cloning methods (based on Bayesian methodology) are applied to real data from the population of the Steller sea lions located in the Alaska coast since 1978 to 2004. The estimates obtained from these methods show a good behavior when they are compared to the actual values.

Keywords: population projection matrices, data cloning, age-structured population, Leslie matrix, Bayesian MCMC algorithm.

AMS subject classification: 92D25, 62F15.

* Corresponding author.
e-mail: mayter@mat.ucm.es

1 Introduction

In this article, we consider discrete time models for describing the evolution of an age-structured population, which 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*. In each population, the length of all the intervals of age is the same but, of course, this common length depends on the population we are studying: one week, six months, one year, 15 years,... The key idea is to choose a suitable length for the age intervals in each population, depending on the reproductive cycles (see Caswell (2001) and Tuljapurkar et al. (2012)).

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 $i+1$). Notice that s_k is zero.
- 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 equations:

$$\begin{aligned} N_1(t) &= f_1 N_1(t-1) + \cdots + f_k N_k(t-1) \\ N_2(t) &= s_1 N_1(t-1) \\ N_3(t) &= s_2 N_2(t-1) \\ &\cdots \\ N_k(t) &= s_{k-1} N_{k-1}(t-1) \end{aligned}$$

These equations can also be formulated in a matrix notation:

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

This is a special square matrix in which all the elements are zero, except possibly those in the first row and those in the first subdiagonal below the diagonal. This type of matrix is usually called population projection matrix and it was firstly introduced by Leslie (1945) for studying age-structured populations. The matrix is also called *Leslie matrix* and, in the original formulation, usually only females are included because only females produce offspring. When all the transitions (including reverse) between stages are possible, the population projection matrix is called the *Lefkovich matrix* (see Lefkovitch (1965)).

Population projection matrices are very useful for studying, in an easy way, the evolution of a population (see e.g. Kajin et al. (2012)). In particular, they are very interesting for describing the long term evolution, in those cases in which the composition or the population achieves an equilibrium:

– The larger eigenvalue, λ, of the Leslie matrix determines the growth rate of the population, in the sense that the size of the whole population, and the size of each group of age is multiplied by the value of λ in each period of time (in the long term). Of course, when $0 < \lambda < 1$, the population size decreases over time, and when $\lambda > 1$, the population size increases over time.

– The normalized eigenvector, corresponding to the larger eigenvalue of the Leslie matrix, gives the stabilized proportions of each group of age (in the long term).

The problem, in practice, is that it is not possible to know neither the exact values of the reproductivity rates, $f_1, \ldots, f_k$, nor the exact values of the survival rates, $s_1, \ldots, s_{k-1}$. This problem will be studied and solved in the following sections considering a stochastic version of the basic Leslie matrix models, by assuming additive random noise for each of the k classes of ages that we handle.

First, we have used a Bayesian linear model to estimate the posterior distribution of the parameters, by means of MCMC procedure (see e.g. Buckland et al. (2007), Clark (2005) and Clark et al. (2005)), using in this case the `Jags` software by means of the package `runjags` (Denwood M.J. (2011)) from the `R` project (R Core Team, R Foundation for Statistical Computing (2012)). Then, we have also applied a *data cloning* method to estimate the parameters. Data cloning methods have been developed to tackle with ecological complex models (see Lele et al. (2007) and Lele et al. (2010)). These methods give estimations of the parameters, by simulating the posterior distribution of them with a MCMC algorithm, which converge to those obtained by maximum likelihood (*ML*) method. With data cloning, it is not so relevant, in practice, which type of prior distribution is considered.

The article is organized as follows:

In Section 2, we pose a statistical model in which fertility rates and survival rates are unknown parameters. In Section 3, the Bayesian approach is considered and the chosen prior distributions are introduced and explained. We compute the conditional posterior distributions of the parameters and we consider a MCMC algorithm in order to simulate them. In Section 4, the Bayesian approach is applied to real data from the population of the Steller sea lions in the Alaska coast since 1978 to 2004. First, in Subsection 4.1, usual Bayesian estimates are obtained and analyzed; then, in Subsection 4.2, the technique of data cloning is introduced an apply to these real data. In Section 5, we give some final conclusions.

2 The statistical problem

We consider an interpretation of the deterministic equations (1) by means of two sampling statistical models:

Statistical model for estimating fertility rates

Let us assume that, in a determined period of time, there are $N_j(t-1)$ individuals in the jth 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., $N_1(t)$ must be understood as an *expected size* for the following period of time. Moreover,

the reproductivity rates, $f_1, \ldots, f_k$ must be understood as unknown parameters, where $f_1, \ldots, f_k > 0$.

Each animal of the $N_j(t-1)$ individuals of group j (for $j = 1, \ldots, k$) produces A_i^j offsprings ($i = 1, \ldots, N_j(t-1)$), where A_i^j is a random variable with mean f_j.

The total number of offsprings of the $N_j(t-1)$ individuals of group j is the random variable $D_j(t-1) = A_1^j + \ldots + A_{N_j(t-1)}^j$. So, $D_j(t-1)$ is the sum of $N_j(t-1)$ independent and identically distributed random variables and, 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 $N_1(t) = D_1(t-1) + \cdots + D_k(t-1)$, where the distribution of $N_1(t)$ can be approximated by a Normal distribution, provided that $N_j(t-1)$, for $j = 1, \ldots, k$, are large enough.

In this way, $N_1(t)$ must be understood as a random variable. The sampling density for this random variable is

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

where $f_1, \ldots, f_k$ are unknown parameters (of interest), and σ_1 is a (nuisance) unknown parameter.

Statistical model for estimating survival rates

Let us assume that, 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., $N_2(t)$ must be understood as an *expected size* for the following period of time. Moreover, the survival rate, s_1, must be understood as an unknown parameter, where $0 < s_1 < 0$.

Each animal of the $N_1(t-1)$ individuals of group 1 may survive to the next period of time (and become an individual of group 2) with probability s_1. So, the total number of individuals in group 2 for the next period of time, $N_2(t)$, is a random variable with Binomial distribution, $Bin(N_1(t-1); s_1)$. 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, $N_2(t)$ must be understood as a random variable. The sampling density for this random variable is

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

where s_1 is an unknown parameter (of interest), and σ_2 is a (nuisance) unknown parameter.

In the same way, $N_j(t)$ (for $j = 3, \ldots, k$) must be understood as a random variable. The sampling density for this random variable is

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

where s_{j-1} is an unknown parameter (of interest), and σ_j is a (nuisance) unknown parameter (for $j = 3, \ldots, k$).

3 Bayesian approach

In this section, we will use Bayesian statistical methods for making inferences on the parameters, $f_1, \ldots, f_k$, $\sigma_1^2, \ldots, \sigma_k^2$ and $s_1, \ldots, s_{k-1}$.

Let us assume that we have observed $\mathbf{n}(t) = (n_1(t), \ldots, n_k(t))$ for $t = 1, \ldots, m$. As we have seen in Section 2, the sampling density of $N_1(t)$ is (approximately) $N(f_1 N_1(t-1) + \ldots + f_k N_k(t-1); \sigma_1)$, where $f_1, \ldots, f_k$ are unknown parameters (of interest), and σ_1 is a (nuisance) unknown parameter.

In the same way, the random variable $N_j(t)$, for $j = 2, \ldots, k$ is (approximately) distributed as $N(s_{j-1} N_{j-1}(t-1); \sigma_j)$, where s_{j-1} are unknown parameters (of interest), and σ_j are (nuisance) unknown parameters.

Then, the likelihood functions for every $t = 2, \ldots, m$ are

$$\begin{aligned}
f(n_1(t)|f_1, \ldots, f_k, \sigma_1^2, \mathbf{n}(t-1)) &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{ -\frac{1}{2\sigma_1^2} \left(n_1(t) - n_1(t-1)f_1 - \cdots \right.\right. \\
&\quad \left.\left. \cdots - n_k(t-1)f_k\right)^2 \right\} \\
f(n_2(t)|s_1, \sigma_2^2, \mathbf{n}(t-1)) &= \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{ -\frac{1}{2\sigma_2^2} \left(n_2(t) - n_1(t-1)s_1\right)^2 \right\} \\
&\vdots \\
f(n_k(t)|s_{k-1}, \sigma_k^2, \mathbf{n}(t-1)) &= \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{ -\frac{1}{2\sigma_k^2} \left(n_k(t) - n_{k-1}(t-1)s_{k-1}\right)^2 \right\}.
\end{aligned}$$

Then, the likelihood function given $\mathbf{n}(t)$ is:

$$\begin{aligned}
f(\mathbf{n}(t)|\theta, \mathbf{n}(t-1)) &= \prod_{j=1}^{k} f(n_j(t)|\theta, \mathbf{n}(t-1)) \\
&= f(n_1(t)|f_1, \ldots, f_k, \sigma_1^2, \mathbf{n}(t-1)) \prod_{j=1}^{k-1} f(n_{j+1}(t)|s_j, \sigma_{j+1}^2, \mathbf{n}(t-1))
\end{aligned}$$

for $t = 2, \ldots, m$, and $\theta = (f_1, \ldots, f_k, \sigma_1^2, \ldots, \sigma_1^2, s_1, \ldots, s_{k-1})$.

As the process $\mathbf{N} = (N(t))_{t=1}^m$ is Markovian, we can obtain the likelihood function given that we have observed $\mathbf{n} = (\mathbf{n}(t))_{t=1}^m$:

$$L(\theta|\mathbf{n}) = \prod_{t=2}^{m} f(\mathbf{n}(t)|\theta, \mathbf{n}(t-1)).$$

We take as prior distributions for the parameters, log-normal distributions for $f_1, \ldots, f_k$, uniform distributions on $(0, 1)$ for $s_1, \ldots, s_{k-1}$ and inverse-gamma distributions for $\sigma_1^2, \ldots, \sigma_k^2$. That is,

$$\begin{aligned}
f_j &\sim \log N(\mu_j, \tau_j^2), \\
\sigma_j^2 &\sim IGamma(\alpha_j, \beta_j),
\end{aligned}$$

with density functions given by

$$\pi(f_j) = \frac{1}{f_j \tau_j \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\tau_j^2} (\ln f_j - \mu_j)^2 \right\},$$

$$\pi(\sigma_j^2) = \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)}(\sigma_j^2)^{-\alpha_j-1}\exp\left\{-\frac{\beta_j}{\sigma_j^2}\right\},$$

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

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

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

This gives as prior distribution on θ :

$$\pi(\theta) = \prod_{j=1}^{k}\left[\pi(f_j)\pi(\sigma_j^2)\right]\prod_{j=1}^{k-1}\pi(s_j).$$

Then, the corresponding conditional posterior distributions for f_j are:

$$\begin{aligned}\pi(f_j|\theta_{-\mathbf{f_j}}, \mathbf{n}) \quad \propto \quad & \exp\left\{-\frac{1}{2}\frac{1}{\sigma_1^2}\sum_{t=2}^{m}\left(n_1(t) - n_1(t-1)f_1 - \cdots - n_k(t-1)f_k\right)^2\right\} \\ & \frac{1}{f_j}\exp\left\{-\frac{1}{2\tau_j^2}(\ln f_j - \mu_j)^2\right\}\end{aligned}$$

for $j = 1, \ldots, k$, and $\theta_{-\mathbf{f_j}}$ denotes all the parameters but f_j.

The conditional posterior distributions for σ_j^2 are the following inverse-gamma distributions:

$$\begin{aligned}(\sigma_1^2|\theta_{-\boldsymbol{\sigma_1^2}}, \mathbf{n}) &\sim IGamma(\tilde{\alpha}_1, \tilde{\beta}_1) \\ \tilde{\alpha}_1 &= \alpha_1 + (m-1)/2 \\ \tilde{\beta}_1 &= \beta_1 + \frac{1}{2}\sum_{t=2}^{m}\left(n_1(t) - n_1(t-1)f_1 - \cdots - n_k(t-1)f_k\right)^2\end{aligned}$$

$$\begin{aligned}(\sigma_j^2|\theta_{-\boldsymbol{\sigma_j^2}}, \mathbf{n}) &\sim IGamma(\tilde{\alpha}_j, \tilde{\beta}_j) \\ \tilde{\alpha}_j &= \alpha_j + (m-1)/2 \\ \tilde{\beta}_j &= \beta_j + \frac{1}{2}\sum_{t=2}^{m}\left(n_j(t) - n_{j-1}(t-1)s_{j-1}\right)^2\end{aligned}$$

for $j = 2, \ldots, k$, and $\theta_{-\boldsymbol{\sigma_j^2}}$ denotes all the parameters but σ_j^2.

Finally,

$$\pi(s_j|\theta_{-\mathbf{s_j}}, \mathbf{n}) \propto \exp\left\{-\frac{1}{2}\frac{1}{\sigma_{j+1}^2}\sum_{t=2}^{m}\left(n_{j+1}(t) - n_j(t-1)s_j\right)^2\right\} I_{(0,1)}(s_j)$$

for $j = 1, \ldots, k-1$, and $\theta_{-\mathbf{s_j}}$ denotes all the parameters but s_j.

Therefore, MCMC sampling procedure consists of simulating from the previous conditional distributions in order to obtain samples from the joint posterior distribution of the parameters (see e.g. Robert and Casella (2004)).

4 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.

4.1 Bayesian estimates

In this subsection, we apply the Bayesian approach described before in order to analyze these data. The algorithms have been programmed using `Jags` (see Plummer (2003)) software in all cases by means of the package `runjags` (Denwood M.J. (2011)) from the `R` project (R Core Team, R Foundation for Statistical Computing (2012)). One advantage of using `Jags` is that it constructs the full conditional distributions and it carries out the Gibbs sampling from the model specifications. All codes are available from the authors, upon request.

We first consider the complete model shown in Section 3. The original deterministic equations are:

$$\begin{aligned} N_1(t) &= f_1 N_1(t-1) + f_2 N_2(t-1) \\ N_2(t) &= s_1 N_1(t-1) \end{aligned}$$

where f_1, 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_1 and f_2; uniform distribution between 0 and 1, for s_1; gamma distribution with mean equal to 1 and variance equal to 10 for σ_1^2 and σ_2^2.

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_1	0.6470	0.2405	0.0044	0.6773	0.9548
f_2	0.2333	0.1845	0.0000	0.2185	0.6706
s_1	0.9594	0.0438	0.8432	0.9734	0.9991
σ_1^2	1.0125	0.2383	0.6399	0.9829	1.5633
σ_2^2	3.7224	0.8168	2.5283	3.5923	5.7181

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

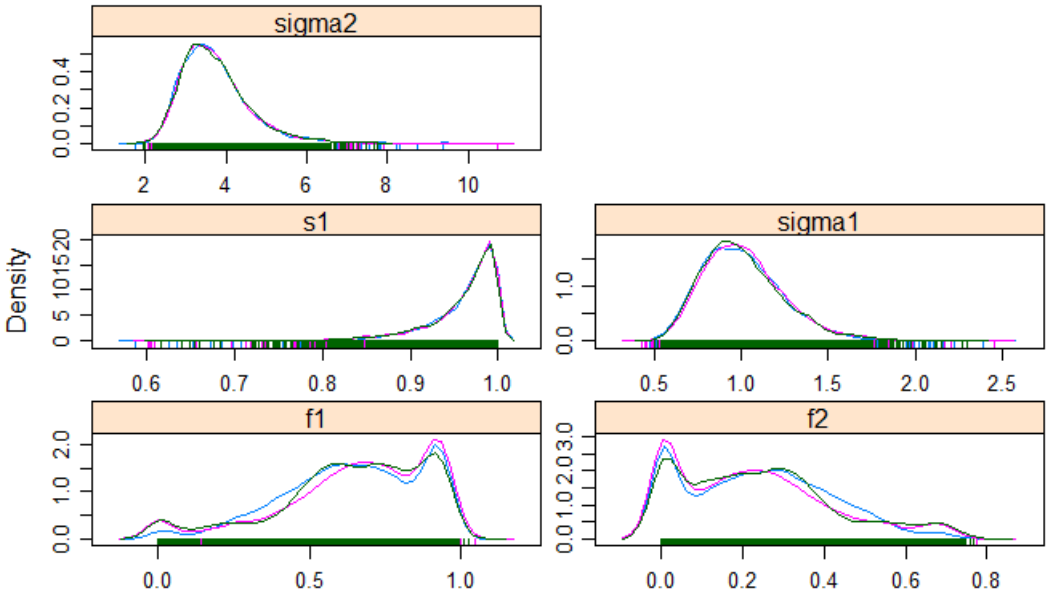

Figure 1: Density plots of the posterior distributions of parameters

Nevertheless, we find some drawbacks with these estimates. The estimated kernel densities from the MCMC samples of the posterior distributions of f_1 and f_2 are multimodal and there are parameters convergence problems (see Figure 1). A possible explanation for this problem is that we have assumed that the first group of age (pups) can be reproductive, and this assumption is not very realistic.

Therefore, we next consider a simpler model, where we assume that pups cannot be reproductive, namely, $f_1 = 0$. In this model, we will also compute the posterior distribution of the largest eigenvalue, λ, and the corresponding normalized eigenvector, (eigen1, eigen2).

We assign the same vaguely informative prior distributions as in the previous model. 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 2.

	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 2: 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 (see Figure 2).

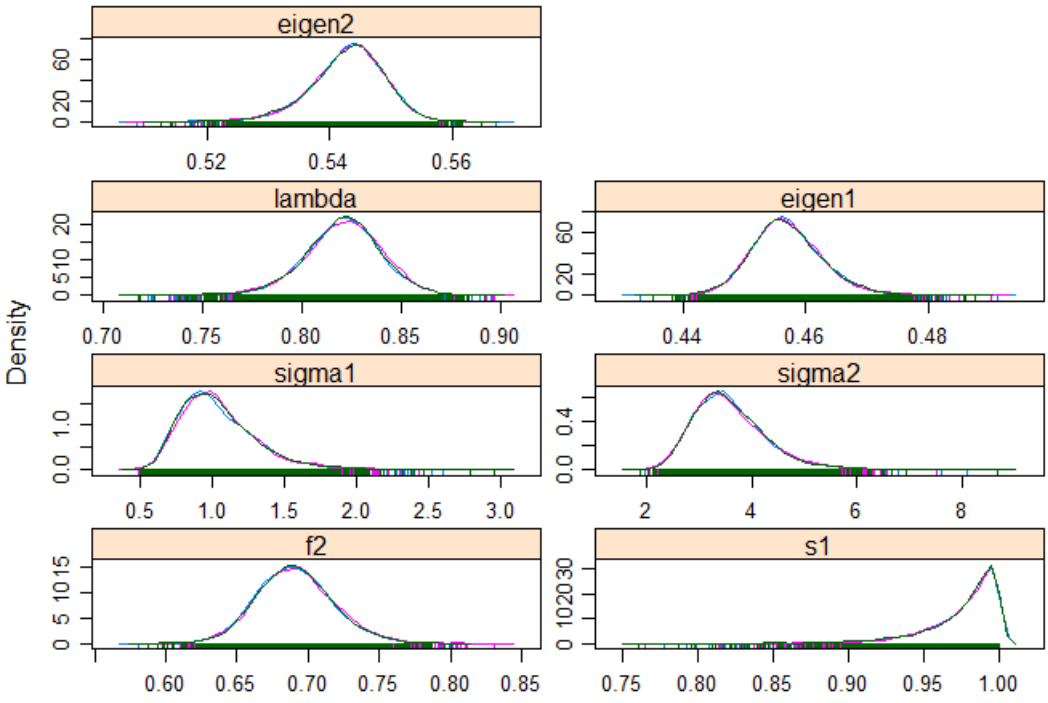

Figure 2: Density plots of the posterior distributions of parameters

The main conclusions we can obtain from the results in Table 2 are:

1. The posterior mean and the posterior median of the parameter f_2 are very similar, approximately, 0,69. The practical meaning of this value is that the fertility rate is estimated to be 0,69 for each adult (each period of time or projection interval).

2. The posterior mean and the posterior median of the parameter s_1 are very similar, approximately, 0,98. The practical meaning of this value is that the survival rate is estimated to be a 98% for pups (each period of time or projection interval).

3. The posterior mean and the posterior median of the largest eigenvalue, λ, are very similar, approximately, 0,82. The practical meaning of this value is that (in the long term) the size of the population is estimated to decline a 18% each year.

4. The normalized eigenvector corresponding to λ is estimated to be, approximately, (0,46 ; 0,54). The practical meaning of this vector is that (in the long term) a 46% of the population will be pups and a 54% will be adults.

We consider now the predictive distributions of $N_1(t)$ and $N_2(t)$ for all these years. We can estimate the predictive posterior means for each year with missing data and, on the other hand, we can check the adjustment of the model to the real data by comparing predicted values to actual values.

In Tables 3 and 4, the first column shows the real data for $N_1(t)$ and $N_2(t)$, where the notation *NA* denotes a missing observation. The second column shows the predictive posterior mean for each year, and the remaining columns show the posterior median and the estimated HPD interval (with probability 0.95) for each year.

Notice that the predicted values are close to the actual values and the estimated HPD intervals include the actual values in all these cases.

N1.Original	**mean.Pred**	**2.5%.Pred**	**50%.Pred**	**97.5%.Pred**
17835	17825.9129	15613.6900	17826.7500	19983.6625
19886	18774.1027	16293.1500	18738.3000	21451.2075
NA	19667.5289	17085.1675	19635.2000	22382.2200
NA	13789.7102	8256.4633	13803.4500	19282.5225
NA	15796.6545	10673.1925	15810.3000	20872.3400
NA	9860.1999	3305.4065	9863.8000	16499.4100
15019	14375.4613	11267.2875	14422.6000	17266.5525
NA	7514.4115	859.6350	7550.7600	14182.5775
11598	13121.8950	10812.3975	13097.7000	15560.2525
NA	6286.2519	499.4133	6279.3950	11968.4425
NA	8156.2473	3471.3123	8182.0750	12804.1025
6394	6059.4277	3085.0633	6069.0250	8974.5238
4648	5926.1801	3810.6958	5911.4700	8172.8977
4057	4875.5646	2773.2155	4853.7350	7103.5117
3646	4336.3730	2166.3773	4327.8800	6541.9532
3176	3952.7268	1845.4098	3944.8200	6105.2165
2831	2767.1512	-173.8028	2788.8850	5661.7325
NA	3122.3494	960.8739	3104.1050	5338.8713
NA	2354.1419	-2407.7405	2344.7800	7113.6865
2056	2692.3802	515.5751	2692.0050	4832.8192
1876	2326.4589	177.5042	2314.3500	4532.9227
NA	2392.8985	250.4562	2390.9000	4593.1505
1675	1608.5363	-1251.2535	1595.5050	4477.8692
1540	2190.9913	127.7382	2185.6050	4331.1110
1608	1555.2286	-1315.9493	1546.7250	4448.8035
NA	2338.9151	249.7018	2336.2200	4489.4268
1578	1507.9984	-1355.9712	1518.4200	4414.1917

Table 3: Real data $N_1(t)$, predictive posterior mean, median and HPD intervals

N2.Original	mean.Pred	2.5.Pred	50.Pred	97.5.Pred
27155	27185.1199	19977.1600	27208.3000	34334.0100
28460	17405.0576	10014.2850	17439.3500	24741.5775
NA	19392.6618	12030.3200	19409.6000	26631.5925
NA	19469.8959	11614.1500	19557.0500	27043.8000
NA	13590.0600	4164.4053	13654.5000	22450.0450
NA	15802.6567	6925.9250	15872.8000	24444.4050
NA	9772.1244	-69.0556	9861.0600	19270.9350
19002	14670.3281	7265.2410	14684.4500	21937.6400
NA	7472.3022	-2283.1920	7532.4550	17003.2325
NA	11329.2117	4121.4423	11376.2000	18639.6450
NA	6420.2901	-2567.5992	6483.9300	15232.2050
8552	8019.0728	-420.7472	8087.0300	16240.7050
7050	6217.6741	-877.7903	6235.5700	13337.7150
6273	4566.7273	-2718.2838	4563.3350	11758.1350
5721	3929.1436	-3219.5595	3886.4600	11159.7925
NA	3554.7445	-3812.2867	3539.5450	10807.4675
4520	3030.0880	-4246.2618	3036.8050	10321.0750
NA	2743.7408	-4571.9108	2717.3850	10084.4875
3915	3100.3327	-4570.2440	3083.4100	10752.7200
3352	2333.2751	-6049.4407	2386.5700	10433.4775
3467	2000.5593	-5373.0212	2048.5100	9273.8912
NA	1876.0020	-5401.0065	1895.6400	9152.3500
3180	2324.5061	-5180.4608	2348.6550	9834.0530
NA	1625.3428	-5681.5315	1605.5000	9133.8677
3366	1501.4333	-5871.3358	1529.3800	8627.5825
NA	1582.2276	-5682.2018	1628.2950	8817.5312
3055	2319.0093	-5042.2410	2287.8750	9853.0598

Table 4: Real data $N_2(t)$, predictive posterior mean, median and HPD intervals

4.2 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)).

In a MCMC procedure, we generate samples from the posterior distribution $\pi(\theta|\mathbf{n})$ that is proportional to the product of the likelihood function $L(\theta|\mathbf{n})$ and a given proper prior distribution $\pi(\theta)$, but it is not necessary to calculate the likelihood function therein.

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

The expression $[L(\theta|\mathbf{n})]^k$ is the likelihood for k copies of the original data and, for k large enough, $\pi^{(k)}(\theta|\mathbf{n})$ 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 (see 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.

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. 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.

We have programmed the algorithm using package `dclone` from the `R` project (R Core Team, R Foundation for Statistical Computing (2012)). We consider, after checking a set of possible values, that 50 is a suitable number of clones. Then, we have used the same prior distributions as in Section 3, but taking the option of updating them as each clone is introduced in the calculation of the posterior distributions (see examples in Sólymos (2010)).

The confidence intervals (95%) for the parameters, based on the Wald approximation, are shown in Table 5.

	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 5: Confidence intervals (95%) for parameters

Notice that intervals shown in Table 5 (obtained from the cloning method) are quite similar to intervals shown in Table 2 (obtained from the usual Bayesian methods). Remember that intervals obtained from the cloning method have to be interpreted in a frequentist sense (although the cloning method is based on Bayesian methodology).

5 Final conclusions

Discrete time models are used in Ecology for describing the evolution of an age-structured population. Usually, they are considered from a deterministic viewpoint but, in practice, this is not very realistic. The statistical model we propose in this article is a reasonable model for the case in which the evolution of the population is described by means of a Leslie matrix. In this statistical model, fertility rates and survival rates are unknown parameters and they are estimated by using the Bayesian approach.

Real data from the population of the Steller sea lions located in the Alaska coast since 1978 to 2004 are analyzed.

First, the usual Bayesian methods are applied and the main results are:

(1) Either the posterior mean or the posterior median can be used for estimating the fertility and survival rates because they are very similar.

(2) The largest eigenvalue of the Leslie matrix and its normalized eigenvector are very interesting to estimate because they have a clear practical meaning as explained in the introduction. Either the posterior mean or the posterior median can be used for estimating them because they are very similar.

(3) The predicted values for the number of pups and adults are close to the actual values, and the estimated HPD intervals include the actual values in all the cases.

Therefore, Bayesian methods seem to be quite suitable for analyzing these type of problems.

Then, data cloning method is used. This 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. The data cloning method is applied to the same real data, and the results we obtain are very reasonable when they are compared to the actual values.

References

Buckland S.T., Newman K.B., Fernández C., Thomas L., and Harwood J. (2007). Embedding Population Dynamics Models in Inference. *Statistical Science* 22(1), 44–58.

Caswell, H. (2001). *Matrix Population Models*. Sinauer Associates, Sunderland, MA.

Clark J.S. (2005). Why environmental scientists are becoming Bayesians. *Ecology Letters* 8(1), 2–14.

Clark J.S., Ferraz G., Oguge N., Hays H., and Dicostanzo J. (2005). Hierarchical Bayes for Structured, Variable Populations: From Recapture Data to Life-History Prediction. *Ecology* 86(8), 2232–2244.

Denwood M.J. (2011). runjags: Run Bayesian MCMC Models in the BUGS syntax from Within R.
`http://cran.r-project.org/web/packages/runjags`

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.

Kajin M., Almeida P.J.A.L., Vieira M.V., and Cerqueira R. (2012). The State of the Art of Population Projection Models from the Leslie Matrix to Evolutionary Demography. *Oecologia Australis* 16(1), 13–22.

Lefkovitch, L.P. (1965). The Study of Population Growth in Organisms Grouped by Stages. *Biometrics*, 21, 1–18.

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`

R Core Team, R Foundation for Statistical Computing (2012). R: A Language and Environment for Statistical Computing (2012), Vienna, Austria, ISBN 3-900051-07-0. `http://www.R-project.org`

Robert, C.P., and Casella, G. (2004). Monte Carlo Statistical Methods. 2nd ed. New York: Springer.

Sólymos P. (2010). Dclone: Data Cloning in R. *The R Journal*, Vol. 2/2.

Tuljapurkar S., Coulson T., and Steiner U.K. (2012). Structured Population Models: Introduction. *Theoretical Population Biology* 82(4), 241–243.