Bayesian inference - Practical exercises
Guiding document

Elise Billoir, Marie Laure Delignette-Muller and Sandrine Charles
ebilloir@pole-ecotox.fr
marielaure.delignettemuller@vetagro-sup.fr
sandrine.charles@univ-lyon1.fr

SETAC Glasgow - May 12, 2013

Introduction

The purpose of this document is to guide you through the practical sessions. This document complements:

- the slides of Marie Laure Delignette-Muller's lecture which provides an introduction to Bayesian inference.
- the slides of Sandrine Charles's lecture which provides an overview of the practical exercises, with elements of context, particularities and modelling for the different case studies.
- the slides of Elise Billoir's lecture which provides information about the practice of Bayesian inference within the R/JAGS/rjags software combo.

This document is meant to help you run the first example described in Elise Billoir's lecture and other applications, basic ones then more advanced ones. The solution (R code) is in the documents:

- *Bayesian inference - Practical exercises - **Dose-response modelling of survival and growth data***,
- *Bayesian inference - Practical exercises - **Reproduction data***,
- *Bayesian inference - Practical exercises - **Survival data***.

According to the table of content of the present document, we propose to work on the following examples:

1. Dose-response modelling
    - Survival data (Binomial error model)
        - log-logistic concentration-response curve
        - Pires-Fox concentration-response curve
    - Growth data (Gaussian error model)
        - log-logistic concentration-response curve
        - Pires-Fox concentration-response curve
    - Reproduction data (Poisson error model)
        - with a simple Poisson error model
        - with an overdispersed Poisson error model*
2. Time-concentration-response modelling of survival data* (Conditional binomial and multinomial error model)

* these are advanced examples meant for the short course participants quickly comfortable with the basic applications.

1 Dose-response modelling

1.1 Survival data

1.1.1 Log-logistic model

This first example is detailed in the slides of Elise Billoir's lecture. The data file (`chlordan_survival_21day.txt`) is provided, as well as the model specification file (`logistic_binomial.txt`).

The first proposed exercice is to run this example by your own on your computer, using the R code shown on the slides.

Before the next examples, just a reminder: the successive steps for carrying out Bayesian inference are as follows:

1. Setting the data (including information to define priors)
2. Visualizing the data
3. Specifying the model according to BUGS/JAGS syntax

4. Initializing the model + data
5. Burn-in phase
6. Further running the algorithm + monitoring of parameters to generate samples
7. Convergence checking
8. Visualizing the results : summary statistics, sample trace and posterior distributions
9. View of the joint posterior distribution
10. Comparison of prior and posterior statistics
11. Visualizing the fitting (model and observed data)
12. Visualizing the fitting (model, observed data and predicted data)
13. Calculation of the deviance information criterion (DIC)

1.1.2 Threshold model

The idea with this second example is to try fitting another concentration-response curve to the survival data set. After the log-logistic model, let's use the so-called Pires-Fox model (Pires et al. 1997, Fox 2010). This is a **threshold** model that writes as follows:

$$Y = d\exp(-b(x - nec)I(x - nec))$$

$$I(x - nec) = \begin{cases} 1 & \text{if} \quad x > nec \\ 0 & \text{if} \quad x \leq nec \end{cases}$$

There are three parameters: nec, the threshold, (a.k.a. No Effect Concentration in ecotoxicology), d, the basal (*i.e.*, control) response and b, reflecting the slope (*i.e.*, effect intensity).

The main changes concern Step 3 (model specification), though some of next steps have to be modified accordingly.

Hint 1 We suggest that you name the R objects related to this second example `M2, M2.MCMC, M2.su`, etc..

Hint 2 Lines 5 and 10 to 13 of the below specification (which is the previously used log-logistic model) have to be changed. We suggest that you save this new model specification file as `PiresFox_binomial.txt`.

```
1  model
2  {
3  for (i in 1:n)
4 {
5 p[i]<-1/(1+pow((x[i]/LC50),b))
6 y[i]~ dbin(p[i],Ninit[i])
7 }
8
9  # specification of priors (may be changed if needed)
10 log10b ~ dunif(-2,2)
11 log10LC50 ~ dnorm(meanlog10LC50 , taulog10LC50)
12 b <- pow(10,log10b)
13 LC50 <- pow(10,log10LC50)
14 }
```

Hint 3 In the BUGS/JAGS syntax, there is a so-called `step(x)` function that tests for $x \geq 0$, returning 1 if true, else 0. This is equivalent to the $I()$ notation used in the model description.

Hint 4 We suggest that you use the following priors:

- $d \sim Unif(inf = 0, sup = dmax)$ with $dmax = 1$ since its a survival probability.
- $\log 10(b) \sim Unif(inf = -2, sup = 2)$ (same as in the first example)
- use the same strategy for nec as the one used in the first example for $LC50$

Chapters 6 and 7 of the JAGS manual (provided as Appendix) can help you with the BUGS/JAGS syntax.

1.2 Growth data

The idea is now to work on another type of data. Therefore, the stochastic part of the model (i.e. the error model) has to be adapted accordingly. The data are provided in the file `chlordan_growth_21day.txt`.

1.2.1 Log-logistic model

The main changes concern Step 1 (data setting) and Step 3 (model specification), but the other steps have to be modified accordingly.

Hint 1 We suggest that you name the R objects related to this third example `M3, M3.MCMC, M3.su, etc.`.

Hint 2 For setting the growth data, you now have to read the data table in `chlordan_growth_21day.txt` and adapt to the column names.

Hint 3 Again, the model specification has to be changed. We suggest that you save the new model specification file as `logistic_normal.txt`.

Hint 4 This example uses the log-logistic equation for the deterministic part of the model, as it was the case in the first example with survival data (Elise Billoir's lecture). However in the first example with survival data, we have used a 2-parameter log-logistic model, fixing $d = 1$ and $c = 0$ because we considered that the survival probability decreased from 1 to 0 with increasing concentration. For the growth endpoint, it does not make sense to fix those parameters, therefore we use the 4-parameter log-logistic model which writes as follows :

$$Y = c + \frac{d - c}{1 + (\frac{x}{e})^b}$$

Hint 5 We suggest that you describe the data variability using a Gaussian error model. Table 7.1 (of the JAGS manual) indicates how to specify a normal (i.e. Gaussian) error model according to the BUGS/JAGS syntax. Be careful, the parameter τ used in this syntax is not the standard deviation σ but the precision ($\tau = \frac{1}{\sigma^2}$).

Hint 6 We suggest that you use the following priors:

- $d \sim Unif(inf = 0, sup = dmax)$ with $dmax = 5$
- $c \sim Unif(inf = 0, sup = cmax)$ with $cmax = 5$
- $\log 10(b) \sim Unif(inf = -2, sup = 2)$ (same as in the previous examples)
- $\sigma \sim Unif(inf = 0, sup = 2)$, also specifying in the model that `tau <- 1/pow(sigma,2)`

- use the same strategy for e as the one used in the first example for $LC50$

Table 7.1 (of the JAGS manual) will help you write them according to the BUGS/JAGS syntax.

1.2.2 Threshold model

The idea with this last example is to try fitting another concentration-effect curve to the growth data set: the Pires-Fox model given above. The main changes concern Step 3 (model specification), but the next steps have to be modified accordingly.

Hint 1 We suggest that you name the R objects related to this fourth example `M4, M4.MCMC, M4.su, etc.`.

Hint 2 Again, the model specification has to be changed. We suggest that you save the new model specification file as `PiresFox_normal.txt`.

Hint 3 We suggest that you use the same priors as for survival data since it is the same concentration range, though adjusting the value of $dmax$ and adding a prior for σ as in the previous example.

1.3 Reproduction data

In this part we will work on a third type of data, concerning reproduction. The idea is to work on the stochastic part of the model, always using the 3-parameter log-logistic model (with $c = 0$) for the deterministic part. The data are provided in the file `snails_repro.txt`. Two variables can be modelled as a function of concentration (`conc`):

- the number of eggs (`Neggcumul`)
- or the number of clutches (`Nclutchcumul`)

produced by snails after 56 days of exposure to a toxicant. As this exposure also induced mortality among individuals, we will take as a covariable in the model, the number of days each animal stayed alive during the experiment. For this purpose, the cumulated number of individual-days (`Nindtime`) was calculated for each replicate. In this experiment, a replicate corresponds to a pool of 5 animals raised in the same beaker. The cumulated number of individual-days corresponds to the sum of the numbers of days each animal of the replicate stayed alive during the experiment.

1.3.1 Poisson error model

Count data are often modelled using a Poisson distribution. Suppose the mean number of eggs (or clutches) per individual-day (Y) is described as follows for an exposure concentration x:

$$Y = f(x) = \frac{d}{1 + (\frac{x}{e})^b}$$

The Poisson model then describes the observed number of eggs (or clutches) for replicate j exposed to concentration x_j by

$$N_j \sim Poisson(f(x_j) \times nid_j)$$

with nid_j the cumulated number of individual-days in replicate j.

Hint 1 We suggest that you try to fit this model first to egg data. Concerning the definition of priors we suggest that you model following information:

- a maximum value of d at 100
- $\log 10(b) \sim Unif(inf = -2, sup = 2)$
- a lognormal distribution of e (which here corresponds to $EC50$) assuming equal to 95% the probability that e lies between the smallest and the highest tested concentrations.

Hint 2 After a classical exploration of your results, we suggest that you focus on the stochastic part by performing posterior predictive check. We suggest that you use the model and the joint posterior distribution of parameters to simulate cumulated numbers of eggs for each replicate. Then we suggest you plot predicted values as 95% credible intervals, against observed values.

Hint 3 We suggest that you calculate the deviance information criterion of this model fitted to egg data.

1.3.2 An overdispersed Poisson error model

Several methods may be used to model an overdispersion based on the Poisson model. We suggest that you try a hierarchical gamma-Poisson model described as follows:

$$N_j \sim Poisson(F_j \times nid_j)$$

with F_j from a gamma distribution of mean $f(x_j)$ and of dispersion parameter β. Such a gamma distribution is parameterized in JAGS/BUGS by its shape (α) and its rate (β). Its mean is equal to $\frac{\alpha}{\beta}$ and its variance to $\frac{\alpha}{\beta^2}$. F_j is then defined as follows:

$$F_j \sim gamma(\alpha_j, \beta)$$

with $\alpha_j = f(x_j) \times \beta$

Hint 1 We suggest that you try to fit this model first to egg data. Concerning the definition of priors, we suggest that you use the same information as in the Poisson model for d, b and e, and a uniform distribution between -2 and 2 for $log10(\beta)$.

Hint 2 After a classical exploration of your results, as with the Poisson model, we suggest that you focus on the stochastic part by plotting the predicted values as 95% credible intervals against observed values and compare the results to those obtained with the Poisson model fitted to the same data.

Hint 3 We suggest that you calculate the deviance information criterion of this model fitted to egg data and compare it to the one obtained with the Poisson model fitted to the same data.

Hint 4 At last, you could compare both models to fit clutch data.

2 Time-concentration-response modelling

In this part, we will work with survival data both concentration- and time-dependent:

1. A survival dataset on *Daphnia magna* exposed to a range of zinc concentrations (file `daphnids_survival.txt`).

2. A survival dataset on snails exposed to a range of concentrations of some toxicant as provided in the file `snails_survival.txt`

The deterministic part of the model will be a threshold time-dependent survival model written as follows:

$$S(x,t) = e^{-h(x)\times t}$$

with $S(x,t)$ the survival probability at concentration x and time t, and $h(x)$ the instantaneous mortality rate at concentration x.

Function $h(x)$ writes as follows:

$$h(x) = \begin{cases} h_0 & \text{if} \quad x \leq necS \\ h_0 + k_S(x - necS) & \text{else} \end{cases}$$

with $necS$ the No Effect Concentration, k_S a coefficient of the effect intensity and h_0 the instantaneous natural mortality rate.

More details about this model can be found in Forfait-Dubuc *et al.* (2012)[1].

2.1 A conditional binomial stochastic part

Given the nature of data (binary time-dependent survival data), the stochastic part can be thought as a conditional binomial law:

$$Y(x,t) \sim \mathcal{B}(p(x,t), Y(x,t-1))$$

with $Y(x,t)$ the number of survivors at concentration x and time t, and $p(x,t)$ the probability to survive at concentration x between $t-1$ and t for organisms which are still alive at time $t-1$:

$$p(x,t) = 1 - \frac{S(x,t-1) - S(x,t)}{S(x,t-1)}$$

Hint 1 First, upload *D. magna* data file. Then, vizualize the data according to both time and concentration. For this purpose, we provide you with the file `plot_functions.R`. In particular, look at the function `plotTandC`. The vizualization may suggest to you that fixing $h_0 = 0$ is reasonable for this dataset.

Hint 2 Try to fit a two-parameter threshold model to *D. magna* data with the following prior definition:

- $k_S \sim Unif(inf = -1.5, sup = 0.5)$
- a lognormal distribution of $necS$ assuming equal to 95% the probability that nec_S lies between the smallest and the highest tested concentrations in log scale.

[1]Forfait-Dubuc, C., Charles, S., Billoir, E., & Delignette-Muller, M. (2012). Survival data analyses in ecotoxicology: critical effect concentrations, methods and models. What should we use? *Ecotoxicology*, 1–12.

Hint 3 Check for convergence and explore your fit results. In particular, superimpose the fitted curve with median parameter values to the data. Also compare priors and posteriors, on the basis of either the associated statistics or their distributions.

Hint 4 Switch now to the snail dataset (file `snail_survival.txt`). Would you fit the same model as with *D. magna* data? If not, on which criterion(s) would you compare both models?

2.2 A multinomial stochastic part

The stochastic part may equivalently be written as a multinomial law (a generalization of the binomial law). See Jager *et al.* (2011) for more details[2].

Hint Rewrite the previous model with a multinomial stochastic part instead of the conditional binomial one. Check for equivalence. Do you see any reason to prefer this modelling way?

[2]Jager, T., Albert, C., Preuss, T., & Ashauer, R. (2011). General Unified Threshold Model of Survival-a Toxicokinetic-Toxicodynamic Framework for Ecotoxicology. *Environmental Science & Technology*, 45, 2529–2540.