

# Intensive course of Bayesian methods: requirements and examination for undergraduate students

*By S. Andreon, stefano.andreon@brera.inaf.it*

Since attenders will solve by themselves problems using some software, they should come with some routines working on their computer, because there will be no time during the course to write them from scratch.

1. Each attender should have its own computer (with an internet connection), or one provided by the organizers, and acquainted with it.
2. Each attender should have installed JAGS<sup>1</sup> which in turn may demand the installation of some additional libraries.

- In order to check having properly installed JAGS attenders should run the model (name the file as model.bug and make you sure only ASCII is in it)

```
# Bayesian Methods for the Physical Science. Learning from
# Examples in Astronomy and Physics. By S. Andreon and B. Weaver.
model {
  for (i in 1:length(nrec)) {
    nrec[i] ~ dbin(eff[i],ninj[i])
    nrec.rep[i] ~ dbin(eff[i],ninj[i])
    eff[i] <- A + (B-A)*phi((E[i]-mu)/sigma)
  }
  A~dunif(0,1)
  B~dunif(0,1)
  mu~dunif(0,100)
  sigma~dunif(0,100)
}
```

re-addressing the input from the file (save it as model.cmd and check that only ASCII is in it)

```
model in model.bug
data in data.dat.R
compile,nchains(1)
initialize
update 3000
monitor set A, thin(10)
monitor set B, thin(10)
monitor set mu, thin(10)
update 100000
coda *
data to testata
```

---

<sup>1</sup><https://sourceforge.net/projects/mcmc-jags/>

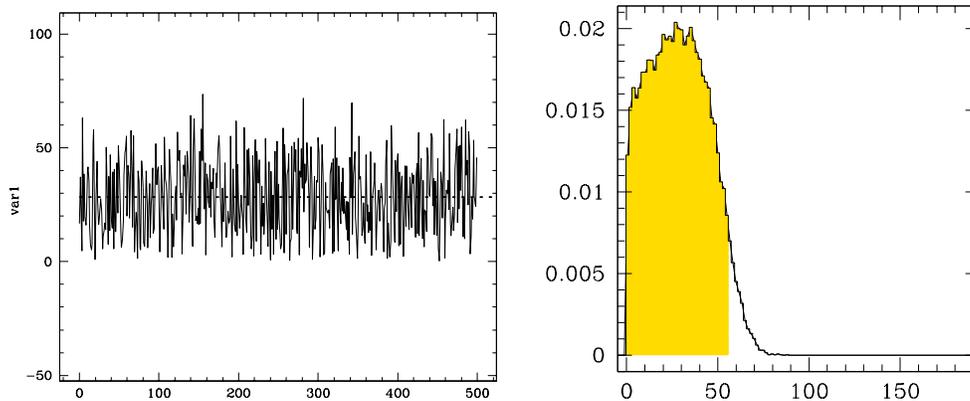


Figure 1: Left panel: Trace plot. Right panel: Marginal distribution (histogram)

```
samplers to testsamplers
exit
```

using the data listed at this URL<sup>2</sup> (save the file as data.dat.R)

Under linux the command to do all that is just

```
jags < model.cmd
```

provided that the jags executable is in the path.

- Each attendee should be able to make plots and simple data manipulation using his/her preferred environment. In particular, the attendee should have already written routines for:

- properly reading files in the following format: CODAindex.txt<sup>3</sup> describes the content of the CODAchain1.txt<sup>4</sup> file by listing the variable names, where they start and where they end. For example (inspect CODAindex.txt) the variable  $s$  starts at line 1 and ends at 50000 (and it is on the 2nd column) of CODAchain1.txt. The reading routine should also work for a different number of variables (e.g. 10) or of samplings (e.g. 30000).
- compute mean and standard deviation (check that  $s$  has mean 28.5 and standard deviation 16.5)
- compute the shortest interval including  $x$  % of the samplings (check that the 95% interval of  $s$  is [0,56])
- produce a trace plot i.e a plot that gives the variable value as a function of its rank, as in Fig. 1. This plotting routine should properly works also if CODAindex.txt contains, say, 10 variables.

<sup>2</sup><http://www.brera.mi.astro.it/%7Eandreon/BayesianMethodsForThePhysicalSciences/data8.2.dat.R>

<sup>3</sup>[http://www.brera.mi.astro.it/~andreon/corso\\_metodi\\_bayesiani/CODAindex.txt](http://www.brera.mi.astro.it/~andreon/corso_metodi_bayesiani/CODAindex.txt)

<sup>4</sup>[http://www.brera.mi.astro.it/~andreon/corso\\_metodi\\_bayesiani/CODAchain1.txt](http://www.brera.mi.astro.it/~andreon/corso_metodi_bayesiani/CODAchain1.txt)

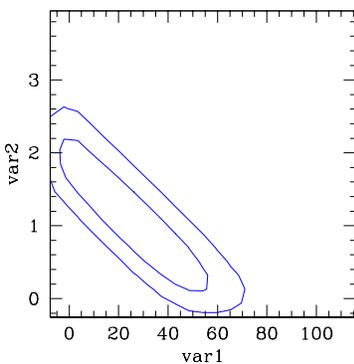


Figure 2: Contours. In this plot, contours are imprecisely determined close to  $var1 = 0$  e  $var2 = 0$  boundaries, and not corrected for smoothing effects. You are not asked to do better (but nothing precludes you from doing it).

- make histograms, as in Fig. 1, right panel (note that the integral must be 1 and independently on the adopted step).
- draw contours. The routine should work for non-elliptical contours, for example when one has two separate “islands”. The contours should include about 68% and about 95% of the samplings. A small margin of imprecision is allowed (i.e. 70% in place of 68% is fine). It is instead not allowed to drawing contours at pre-defined thresholds (say the peak value divided a magic number). Check your contours against those in Fig. 2 with the sampling in CODAchain1.txt. The latter contours are somewhat approximated (and I do not ask anything better).
- plot data with error bars.

The attendee should come with these routines working on his/her computer and with some ability to slightly modify them when needed, because this course, on statistical methods, has no time to address non-statistical issues. Past experience tell that attendees able to make the above routines, but coming without them, failed to attend the lectures (spent all time in writing the routines). Therefore, attending the lectures without these working routines is, by large, a loss of time.

*All non-graduate students willing to attend the course should come with their version of figures 1 and 2 printed on paper and give them to the instructor at the start of the course. Lacking to fulfill this requirement, they will not be admitted to the course.*

## Examination

Attendees will be asked to write a relation describing how they solved a statistical problem of their own choice similar to, but different from, those addressed during the course. Each attendee must solve a problem different from those chosen by other attendees. Attendees have to prepare a short relation (consider a soft limit of about 4 pages) with:

- Title
- Your name
- A few line of background (why the analysis of these data are interesting). A minimal background information about topic to understand what follows. Identify the problem you will address.
- Present data: plot data; list them if just a few, otherwise only plot them.
- Which errors? (here you are describing the likelihood: gaussian? Poisson?, a combination of, etc.).
- Description of the model: describe link between quantities (if any). With extra scatter? Describe priors used. Conclude with "therefore the JAGS model reads" followed by the JAGS model.
- Check convergence. Are results independent of chains? Check trace plots, insert one (and mention the other being similar).
- Plot posterior parameter distributions for the key parameters. If more than one parameter: plot joint probability contours. Overplot the prior. Optional: shade the 68% or 90% interval on individual parameters.
- Quote in the text (or in a table) posterior mean and standard deviation (or other summaries of the posterior distribution).
- Plot the fitted model on the data, with 68% bounds (when this applies).
- If you made a  $y$  vs  $x$  fit, write down the math expression with inserted values (and errors) for the parameters.
- Comment about prior sensitivity: are your parameters strongly affected by the prior choice? Optional: adopt a slightly different prior to prove little sensitivity (the mean/stddev should not change).
- Conclusion: so what?

A template relation (only useful for understanding the amount of blurb to be put in there) is given at the URL<sup>5</sup>

Attendees should complete their evaluation within 4 weeks after the end of the course, and can submit their document for evaluation at most twice. The second submission, if any, should account for all comments given to the first submission.

---

<sup>5</sup>[http://www.brera.mi.astro.it/%7Eandreon/corso\\_metodi\\_bayesiani/template\\_relation.pdf](http://www.brera.mi.astro.it/%7Eandreon/corso_metodi_bayesiani/template_relation.pdf)

See you soon.

stefano.andreon@brera.inaf.it