Intensive course of Bayesian methods: aims, methods and requirements.

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

The course consists in one week-long, 3 h a day, laboratory on Bayesian (statistical) methods. It will not assume (almost) any previous knowledge in statistics. It is, by large, a laboratory course (i.e. attenders will do something, not just hear someone), and, for this reason, assumes a familiarity with the use of a plotting environment (at the attender choice).

Course Syllabus:

- Probability axioms. Computation of the posterior: analytical vs numerical sampling. Upper limits. Initial discussion on the role of the prior. Importance of checking numerical convergence. A glimpse on sensitivity analysis.
- Single parameters models. Combining information coming from multiple data. The prior (and the Malmquist-like effect). Prior sensitivity. Two-parameters models. Joint probability contours. Comparison of the performances of state-of-the-art methods to measure a dispersion.
- Introduction to regression. Comparison of regression fitters. Regressions (of increasing difficulty): non-linear regression with non-gaussian errors of different sizes (but no error on predictor and no intrinsic scatter). Allowing systematics (intrinsic scatter). Allowing errors on x. Regressions with two (or more) predictors. A glimpse on other important issues such as mixture of regressions, non-random data collection, model checking.

The preliminary program (based on the one given at the Max Planck and Torino Universities in 2016) is at this URL[1].

Rationale and Methods:

The purpose of the course is not to teach a content. Attenders will hear the instructor for a tiny fraction of the time, and then spend most of their time solving by themselves (with the teacher help) problems of increasing (statistical) complexity, often using real data (to be downloaded during the course). This poses constraints on requirements, and demands that the attenders attend *all* lectures.

Requirements for organisers.

Internet connection for all people (including teacher) and electrical power (and sockets) for all computers.

[1] http://www.brera.mi.astro.it/∼andreon/corso_metodi_bayesiani/CorsoMetodiBayesiani1516.html

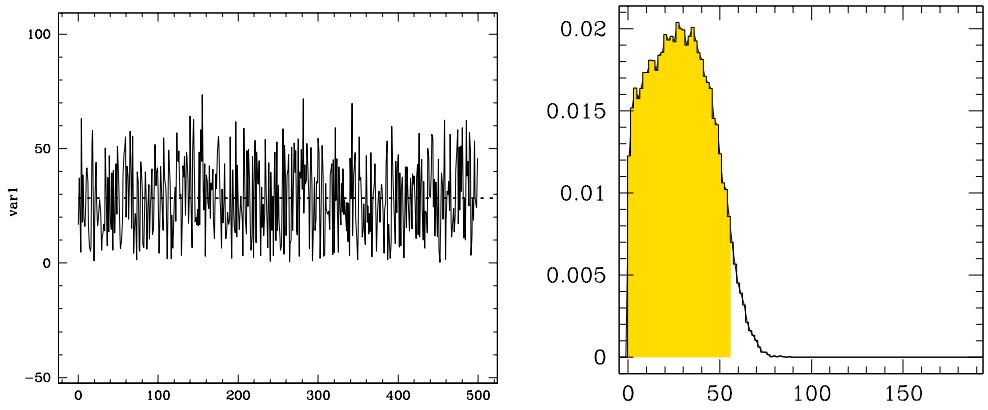

Figure 1: Left panel: Trace plot. Right panel: Marginal distribution (histogram)

Requirements for attenders.

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), and acquainted with it.

2. Each attender should have installed JAGS[2] which in turn may demand the installation of some additional libraries.

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

   - properly reading files in the following format: CODAindex.txt[3] describes the content of the CODAchain1.txt[4] 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])

[2]https://sourceforge.net/projects/mcmc-jags/
[3]http://www.brera.mi.astro.it/∼andreon/corso_metodi_bayesiani/CODAindex.txt
[4]http://www.brera.mi.astro.it/∼andreon/corso_metodi_bayesiani/CODAchain1.txt

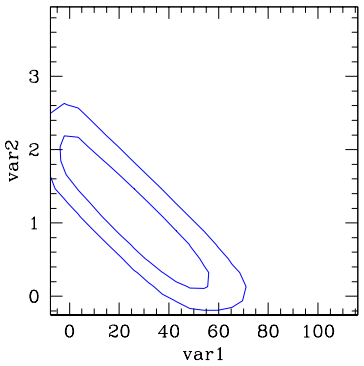

Figure 2: Contours. In this plot, contours are imprecisely determined close to $var1 = 0$ e $var2 = 0$ boundaries, and not corrected for smoothing effects. You are not asked to do better (but nothing precludes you from doing it).

- 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.
- 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 divide 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).

The attender should came 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 attenders 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.
If you have a statistical problem to address and you want to discuss it with me, do not esitate to do so.
See you soon.

stefano.andreon@brera.inaf.it