BAYESIAN STATISTICAL MODELS - University of Milano-Bicocca

Case study - reach data

Lecturer: Riccardo Corradin

We conside again the dataset *frogs.csv*, available at the module page. The data consist in a set of measures for different amphibious. Specifically, for each observed animal we measure

- The scientific *family*.
- The *body weight*.
- The *nose-to-tail length*.
- The *brain weight*.

We are interested study the body weight, here playing the role of the response variable, as function of the other observed quantities. We first transform on suitable scales the positive real-valued observed quantities, then we consider a linear model with the informative prior assumptions, i.e.

$$\begin{aligned} Y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 &\sim N(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \sigma^2), \qquad i = 1, \dots, n, \\ \boldsymbol{\beta} \mid \sigma^2 &\sim N(\boldsymbol{b}_0, \sigma^2 \Sigma_0), \\ \sigma^2 &\sim IG(a_0, b_0). \end{aligned}$$

- Produce a code that estimate all the possible models, computing also the WAIC and LPML indices.
- Test the best model you identify against the saturated model and the model including only the intercept term.

We consider a set of $n = 681$ observations from the Rotterdam Early Arthritis Cohort (REACH) dataset. The study was initiated in 2004 to investigate the development of rheumatoid arthritis in patients with early manifestations of joint impairment. Information regarding basic patient characteristics, serological measurements, and patterns of disease involvement at baseline has been gathered for the 681 recruited patients. It is of interest to know which of the following 12 factors are potentially associated with the development of rheumatoid arthritis considered as a binary (yes/no) outcome.

- ACCP (cyclic citrullinated peptide antibody)
- age

- ESR (erythrocyte sedimentation rate, is the rate at which red blood cells sediment in a period of one hour)
- DC (duration of complaints in days)
- stiffness (duration of morning stiffness in minutes)
- RF (rheumatoid factor)
- gender
- Sym (symmetrical pattern of joint inflammation; yes/no)
- SJC (swollen joint count)
- TJC (tender joint count)
- BCPH (bilateral compression pain in hands; yes/no)
- BCPF (bilateral compression pain in feet; yes/no)

We consider a probit model of the form

$$\begin{aligned} Y_i \mid \theta_i &\sim Be(\theta_i), \qquad i = 1, \ldots, n, \\ \theta_i &= \Phi(\beta_0 + \boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}), \\ \beta_0 &\sim N(0, \tau^2), \qquad \beta_j \sim \pi N(0, \tau^2) + (1-\pi)N(0, c\tau^2),\ j = 1, \ldots 12, \\ \pi &\sim Beta(\alpha_1, \alpha_2), \end{aligned}$$

where $\tau^2 = 10^3$, $c = 0.001$, $\alpha_1 = 1$ and $\alpha_2 = 1$.

- Implement the model in STAN.
- Produce a sample of $10\,000$ realizations from the posterior distribution of interest. Check the diagnostic of the model.
- Find the best model according to HPD, MPM and HS criteria.