Biostat 285: Homework 1

Instructions

You are allowed to consult the textbooks, your notes, the material on Canvas, the original articles/material of the methods decribed and referred to in the HW.

Submit your .Rmd/.Rnw file as a separate file together with the pdf version on Canvas

Problem 1 - Spike and Slab vs Horseshoe Regression

In this problem, we will see how to implement a Bayesian regression with variable selection via a spike-and-slab prior. Please, download the following .csv data file from

http://138.68.227.229/ mguindani/teaching/bio285/dataHW1.csv

(follow the link). The dataset contains a dependent variable y and 50 predictors. Only a subset of the predictors is believed to be associated with the outcome.

First, we consider a discrete spike and slab prior on the regression coefficients of the linear regression between y and a predictor x_j:

$$\beta_j|\pi_0, \tau \sim \pi_0\, \delta_0 + (1-\pi_0)N(0, \tau^2),$$

$j = 1, \ldots, p$, with large variance τ^2, and $\pi_0 \in (0, 1)$.

1. Implement (i.e. write the code and run) a *stochastic search variable selection* (SSVS) method via Gibbs sampler to identify a subset of relevant predictors. For implementing the code, in addition to the slides on Canvas, you can find the following resources useful (follow the link):

   George & McCullogh (1993) Variable Selection via Gibbs Sampling, JASA.

   Dellaportas et al. (2000) Bayesian Variable Selection Using the Gibbs Sampler.

   Write a pseudo-code (i.e., in *Latex!*) with the relevant steps (including full conditionals) of the algorithm. Then, write up the algorithm in a code using your favorite programming language.

   Comment on the mixing and convergence of the chain.

Describe posterior inference by reporting posterior means and 95% credible intervals of the regression coefficients and corresponding posterior probabilities of inclusion (PPIs) in the model. Find the optimal threshold that controls a false discovery rate (FDR) of 5%.

Optional, bonus Implement a SSVS using Jags (or Stan). In addition to the slides on Canvas, you can find the following resources useful (follow the link):

Ntzoufras, Gibbs Variable Selection Using BUGS. Compare the inference from the Jags model with the one obtained under (1).

2.(a) Now consider the *horseshoe prior* by Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). Write the linear regression model under the horseshoe prior. You can conduct inference with the horseshoe prior by downloading and installing the R package *horseshoe*. Compare the inference from the horseshoe prior with the one obtained under (1), (2) and (3). What can you conclude about the relevant β's in the regression? How do you take decisions about these parameters?

2 (b) How would you set up the prior for the global shrinkage parameter using the Horseshoe prior? You may find the following manuscript of interest:

Piironen and Vehtari (2017) On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior, especially Section 3.1.

Problem 2 - A Simulation comparison

Simulate data Y_i, $i = 1, \ldots, n$, obtained as follows:

- First generate a $n \times p$ design matrix, where each element is obtained as a draw from a standard normal
- Fix $p^* = 10$ true non-zero coefficients defined as $\boldsymbol{\beta^*} = [-5, -3, -2, -1, -.5, 0.5, 1, 2, 3, 5]$. Fix $p - p^*$ coefficients equal to zero, then generate data

$$Y_i = x_i \times \boldsymbol{\beta} + \varepsilon_i$$

  where $\boldsymbol{\beta} = [\boldsymbol{\beta^*}, \mathbf{0}]$ being a $p \times 1$ vector of coefficients, and $\varepsilon_i \sim N(0, \sigma^2)$
- Using the package BoomSpikeSlab, run a Bayesian variable selection model with a (gaussian) spike-and-slab prior under the following scenarios:
  - Scenario 1: $n = 100$, $p = 100$ and $\sigma = .5$
  - Scenario 2: $n = 100$, $p = 100$ and $\sigma = 3$
  - Scenario 3: $n = 100$, $p = 1000$ and $\sigma = .5$
  - Scenario 4: $n = 100$, $p = 1000$ and $\sigma = 3$
  - Scenario 5: $n = 100$, $p = 10000$ and $\sigma = .5$

– Scenario 6: $n = 100$, $p = 10000$ and $\sigma = 3$

- Comment on the results obtained out of this simulation exercise, in particular the effect of larger p (lower p^*/p) and larger σ (separately).

  Note that I am not interested in seeing pages of code and default figures. Choose how to describe the results carefully.

- Now, let's consider the use of non-local priors. We are going to use the package *mombf*. You can also look at the package manual here. Use the command

```
library(mombf)
fit <- modelSelection(y, x)
```

  to run a model selection with default non-local prior specifications. Then run

```
head(postProb(fit1))
```

  to identify the models that receive higher probability a posteriori. Run Scenarios 1, 2, 3, 4 above, and

  – Scenario 7: $n = 500$, $p = 1000$ and $\sigma = 3$ and comment on the results.