Sparse regression modelling

Interactive Session # 2, Bayesian Statistics

26 Nov. 2021

In this practical class, we analyze the Stanford HIV Drug Resistance Database, publicly available at `https://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006/`. Data consists of a [num_patients × num_covariates] matrix stored in the file `X.csv`, and a [num_patients × num_drugs] matrix stored in the file `Y.csv`.

In the design matrix X, for each patient the presence or absence of specific gene mutations is reported (i.e., for each patient we have a vector of zeros and ones). The features (columns of X) are given by mutation/position pairs:

$X_{i,j} = 1$ if the ith patient has the jth mutation/position pair and 0 otherwise

For example, in the dataset, three different mutations (A, C, and D) are observed at position 63 in the protease, and so three columns of X (named P63.A, P63.C, and P63.D) indicate the presence or absence of each mutation at this position.

Rows of the response matrix Y report the fold increase of lab-tested resistance to the k-th drug for each patient.

The goal is to make inference on Y starting from X using a linear model. In particular we want to make predictions on new patients as well as identify important genomic variations that might significantly affect the resistance to this kind of drugs.

For the moment, consider only the 1st drug, i.e. the **first column** of Y, and let the response y_i be a **suitable transformation** of such values.

We assume a linear model:

$$y_i = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

and consider several possible prior specifications for $\boldsymbol{\beta}$. In the following we denote with p the dimension of each $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$, i.e., $\boldsymbol{x}_i, \boldsymbol{\beta} \in \mathbb{R}^p$.

a) Normal prior

$$\boldsymbol{\beta} \sim \mathcal{N}_p(0, \tau^2 I_p)$$

b) The Bayesian Lasso

$$\beta_j \overset{\text{iid}}{\sim} \mathcal{DE}(0, \tau), \qquad j = 1, \ldots, p$$

where $\mathcal{DE}$ denotes the Double Exponential or Laplace distribution.

c) Ishwaran and Rao's spike-and-slab prior (SSVS prior in the lessons):

$$\begin{aligned} \beta_j \mid p_j &\overset{\text{ind}}{\sim} p_j \mathcal{N}(0, \tau_1^2) + (1 - p_j)\mathcal{N}(0, \tau_2^2) \qquad j = 1, \ldots, p \\ p_j &\overset{\text{iid}}{\sim} \text{Bern}(0.5) \qquad j = 1, \ldots, p \end{aligned}$$

Here $\tau_1 \ll \tau_2$.

d) The spike-and-slab Lasso.

$$\begin{aligned} \beta_j \mid p_j &\overset{\text{ind}}{\sim} p_j \mathcal{DE}(0, \tau_1) + (1 - p_j)\mathcal{DE}(0, \tau_2) \qquad j = 1, \ldots, p \\ p_j &\overset{\text{iid}}{\sim} \text{Bern}(0.5) \qquad j = 1, \ldots, p \end{aligned}$$

$\tau_1 \ll \tau_2$.

e) The horseshoe

$$\begin{aligned} \beta_j \mid \lambda_j, \tau &\overset{\text{ind}}{\sim} \mathcal{N}(0, \lambda_j^2 \tau^2) \qquad j = 1, \ldots, p \\ \lambda_j &\overset{\text{iid}}{\sim} \mathcal{HC}(0, 1) \qquad j = 1, \ldots, p \end{aligned}$$

Where $\mathcal{HC}$ denotes the half-Cauchy distribution, i.e. the Cauchy distribution truncated on $[0, +\infty)$.

f) The finnish (or regularised) horseshoe

$$\begin{aligned} \beta_j \mid \lambda_j, \tau, c &\overset{\text{ind}}{\sim} \mathcal{N}(0, \widetilde{\lambda}_j^2 \tau^2), \quad \widetilde{\lambda}_j^2 = \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2} \\ \lambda_j &\overset{\text{iid}}{\sim} \mathcal{HC}(0, 1) \qquad j = 1, \ldots, p \\ c &\sim \text{Inv-Gamma}(1.5, 1.5) \end{aligned}$$

Exercise 1: prior modelling

1.1.) Plot the marginal distribution of β_j under the different priors. [Hint: if you cannot marginalize out latent variable analytically, you can use Monte Carlo simulation].

Remember the usual analogy between the penalized MLE (maximum log likelihood estimator) and the MAP (maximum a posteriori). In fact the log posterior is proportional to

$$\log \pi(\boldsymbol{\beta} \mid \mathbf{y}) \propto \sum_{i=1}^{n} \ell(y_i \mid \beta) + \log \pi(\beta)$$

where $\ell(y_i \mid \beta)$ it the log-likelihood for the i-th observation.

1.2) Some priors can be seen as analogous to "famous" penalties in the frequentist literature. Can you recognize them?

It is often useful to interpret the frequentist penalties in the penalized MLE as a the norm of the vector $\boldsymbol{\beta}$. For instance, for ridge regression the penalty is

$$\lambda \|\boldsymbol{\beta}\|_2^2$$

Given a particular penalty, it is interesting to look at the "unit ball" of the corresponding norm when the dimension of β is equal to 2, i.e. plotting the set

$$\{\boldsymbol{\beta} = (\beta_1, \beta_2), \beta_i \in \mathbb{R} : \|\boldsymbol{\beta}\|_\pi = 1\}$$

where $\|\cdot\|_\pi$ is the norm associated to the prior / penalty in the penalized MLE.

1.3) First of all, let us fix $\tau = 1$ in all the priors. Approximate the log density of β_j under the Horseshoe with the lower bound from Carvalho et al (2010)

$$\log \pi(\beta_j \mid \tau) \geq -\log \log \left(1 + \frac{2\tau^2}{\beta_j^2}\right).$$

Then, plot the unit-ball according to the penalties associated to the priors (a), (b) and (e).

Exercise 2: Prediction

We now focus on the predictive performance. Consider the likelihood as in Equation (1) and assume that $\sigma \sim logN(0, 3)$ where $logN$ denotes the log-Normal distribution.

2.1) Code the linear regression model and all different priors in Stan and run the MCMC algorithms on the dataset.

For (a)-(b)-(e)-(f) add another level of hierarchy by assuming

$$\tau \sim \mathcal{HC}(0, 1).$$

For the spike-and-slab priors (c)-(d) select τ_1 such that $P(|\beta_j| < 0.1) > 0.25$ and τ_2 such that $P(|\beta_j| > 10) > 0.25$.

Hint:

$$\begin{aligned} P(|\beta_j| < 0.1) &= P(|\beta_j| < 0.1 | p_j = 1) P(p_j = 1) + \\ &\quad P(|\beta_j| < 0.1 | p_j = 0) P(p_j = 0) \\ &\geq 0.5 P(|\beta_j| < 0.1 | p_j = 1) \end{aligned}$$

2.2) Comment on the different results obtained. Focus in particular on:

1- Which model gives the best predictive performance?

2- Are there some sampling difficulties specific to some models?

Exercise 3: Variable Selection

Given that all the priors are absolutely continuous, $P(\beta_j = 0 \mid \mathbf{y}) = 0$ for all j.

Hence, we perform variable selection using hard-shrinkage (HS): we keep only the variables whose 95% posterior credible interval does not contain zero.

3.1) Implement the hard-shrinkage criterion (to get the posterior credible interval, you can use the arviz.hdi function)

3.2) Which variables are selected from the different models? Do they agree?

Exercise 4: going hierarchical

We move on to considering the dataset with all the responses. By considering a multiple-linear model we assume

$$y_{ik} = \alpha + x_i^T \beta_k + \varepsilon_{ik}, \quad \varepsilon_{ik} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_k^2), \qquad k = 1, \ldots, K \tag{2}$$

where i is the index of the patient and k is the index of the drug.

Of course, different models (one for each k) could be fitted independently, but a smarter alternative would be to model all the responses jointly using a hierarchical model.

Let $\beta_k = (\beta_{k,1}, \ldots, \beta_{k,p})$. We want to encourage a priori the following idea: if a variable (genetic mutation) is responsible for an increase (resp. decrease) of the resistance to drug k, it is likely that it will be responsible for the resistance to other drugs as well.

Statistically, this means that if $\beta_{k,j}$ is substantially different from zero, also $\beta_{k',j}$ should be. Analogously, we can assume the opposite: if $\beta_{k,j} \approx 0$ we want to encourage $\beta_{k',j} \approx 0$. This later idea can be easily built in a hierarchical model.

Consider for instance the spike-and-slab model

$$\begin{aligned} \beta_{jk} \mid p_{jk} &\overset{\text{ind}}{\sim} p_{jk}\mathcal{N}(0,\tau_1) + (1-p_{jk})\mathcal{N}(0,\tau_2) \quad j = 1,\dots,p,\ k = 1,\dots,K \\ p_{j1},\dots p_{jK} \mid p_{0j} &\overset{\text{iid}}{\sim} \text{Bern}(p_{j0}) \\ p_{0j} &\overset{\text{iid}}{\sim} \text{Beta}(a,b) \quad j = 1,\dots,p \end{aligned}$$

4.1) Code model (2) and the hierarchical spike-and-slab prior in Stan and fit the whole dataset to it, considering only the responses for which the number of missing values is at most 100. Remove from the dataset all subjects with missing values.

4.2) Build a prior similar to the hierarchical spike-and-slab one, but by starting from the horseshoe prior or its regularised version (HINT: think about what is the role of the parameter λ_j)