Mattias Villani
Department of Statistics
Stockholm University

Computer Lab 4 - Regularization and Variable Selection

The labs are the only examination, so you should do the labs **individually**.
You can use any programming language you prefer, but do **submit the code**.
Submit a readable report in **PDF** (no Word documents!) or a **JuPyteR notebook**

In this lab you will use the prostate cancer dataset from the book Elements of Statistical Learning (ESLII, see Section 3.2.1 for a description of the dataset and the regression model setup). The dataset can be downloaded here: prostate cancer data. Use the same model as in ESLII:

$$\texttt{lpsa} = \beta_0 + \beta_1\texttt{lcavol} + \beta_2\texttt{lweight} + \beta_3\texttt{age} + \beta_4\texttt{lbph} + \beta_5\texttt{svi} + \beta_6\texttt{lcp} + \beta_7\texttt{gleason} + \beta_8\texttt{pgg45}$$

with $\varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Use all 97 observations in the dataset (note that the ESLII book uses a random sample of 60 observations). Standardize the covariates to have zero mean and unit variance.

1. **Bayesian regularization**. The linear regression model with a iid Gaussian (L2-regularization) prior is

$$\begin{aligned}
y_i &= \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \\
\boldsymbol{\beta}|\sigma^2, \lambda &\sim N\left(\mathbf{0}, \frac{\sigma^2}{\lambda} I_p\right) \\
\sigma^2 &\sim \mathsf{Inv} - \chi^2\left(\nu_0, \sigma_0^2\right) \\
\lambda^{-1} = \psi^2 &\sim \mathsf{Inv} - \chi^2\left(\omega_0, \psi_0^2\right)
\end{aligned}$$

   and we use a non-informative $\beta_0 \sim N(0, 100^2)$ prior for the intercept.

   (a) You can use my implementations of the Gibbs sampler in Julia and R for sampling from the posterior $p(\beta_0, \boldsymbol{\beta}, \sigma^2, \lambda|\boldsymbol{y}, \boldsymbol{X})$ [If you are a Pythonista, ask chatGPT to translate the code and check for correctness.]. Use the sampler to analyze the prostate cancer dataset. Set the prior hyperparameters to $\nu_0 = 0.01$ and $\sigma_0^2 = 1$, $\omega_0 = 0.01$ and $\psi_0^2 = 1$. Draw a posterior sample of 10000 draws (after a burn-in of 1000 draws) and present summaries of the results.

   (b) Explore if the posterior distribution of λ and the elements of $\boldsymbol{\beta}$ are sensitive to the prior on ψ^2 by trying out at least two other values for ω_0.

(c) Now use the horseshoe prior

$$\begin{aligned}\beta_j|\sigma^2, \lambda_j^2, \tau^2 &\sim N\left(0, \sigma^2\tau^2\lambda_j^2\right)\\ \lambda_j &\overset{\text{iid}}{\sim} C^+(0,1)\\ \tau &\sim C^+(0,1)\\ \sigma^2 &\sim \mathsf{Inv}-\chi^2\left(\nu_0, \sigma_0^2\right)\end{aligned}$$

and the same non-informative $\beta_0 \sim N(0, 100^2)$ prior for the intercept. Implement the Gibbs sampling algorithm for the horseshoe prior described in the Section *Global-local regularization and Horseshoe* in the Bayesian Learning book. Use this implementation to sample 10000 iterations (after a burn-in of 1000 iterations) from the posterior $p(\beta_0, \boldsymbol{\beta}, \sigma^2, \tau|\boldsymbol{y}, \boldsymbol{X})$ for the prostate cancer dataset. Compare with the results from the L2-prior in 1a) and to the least squares estimate.

(d) [**Bonus question** if you feel up to it, and know RStan or Turing.jl well (so that this is a quick thing for you). Sample the posterior for the regression with a horseshoe prior using RStan or Turing.jl. Compare the resulting posteriors from the Gibbs sampler and RStan/Turing.jl's HMC sampler. Compare effective sample size per second of computing time. Since RStan is coded in C++, your Python/R code for the Gibbs sampler will be slower. Julia is closer to C++ speed and my Julia implementation makes 10000 draws in 0.3 seconds. Note also that my prior for β_j has a variance scaled by σ^2, which is not always how other people do it (e.g. the original Horseshoe paper)].

2. **Bayesian variable selection**

(a) Implement Bayesian variable selection using the spike-and-slab prior, **or** find a package in your favorite language that does it for you (I leave that choice up to you, depending on how much time you spent on Problem 1 above and how useful variable selection is for your research). Analyze the prostate cancer data with prior hyperparameters $\tau = 10$ and $\omega = 0.5$. Explore how the posterior inclusion probabilities for the variables depend on the prior hyperparameter τ. Try to explain the Bayesian logic behind these results.

(b) The standard spike-and-slab prior uses the following prior for the binary selection indicators for the p covariates:

$$z_1, \ldots, z_p|\omega \sim \text{Bernoulli}(\omega)$$

and it is common to set $\omega = 0.5$. What is the distribution on the number of covariates with non-zero regression coefficients? Would this always be a good prior?

Good luck!