1 The relationship between ridge regression and Bayesian regression

Ridge regression is a regularised form for linear regression. In linear regression we want to write the response variable, y, as a linear sum of the independent variables, x_i

$$y_i = \theta_0 + \sum_{j=1}^{p-1} \theta_j x_{ij} = \sum_{j=0}^{p-1} \theta_j x_{ij}, \tag{1}$$

where I have introduced $x_0 = 1$ for simplicity of notation. It would also be ok to do this problem writing this simply for 2 variables.

We saw in the lectures that ridge regression can be written as an optimisation problem where we minimise

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \theta_j^2, \tag{2}$$

where RSS is given by

$$\mathrm{RSS} = \sum_i \frac{\left(y_i - \sum_j \theta_j x_{ij}\right)^2}{\sigma_i^2}. \tag{3}$$

Problem a) At the same time we know that for Bayesian regression we need to calculate the posterior likelihood which is the product of the likelihood of the data and the prior on the parameters. There is also a normalising constant which we will ignore here.

The likelihood is given in the lectures and can be written

$$L = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\mathrm{RSS}/2}. \tag{4}$$

In the problem set we were told to take the parameters $\theta \sim N(0, \tau^2)$, thus the prior will be

$$\mathrm{prior}(\theta_j) = \frac{1}{\sqrt{2\pi}\tau} e^{-\theta_j^2/2\tau^2}. \tag{5}$$

We can multiply the priors of each parameter together and this all gives us a posterior

$$\boldsymbol{P(\cdot;\theta) = \left(\prod_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\mathbf{RSS}/2}\right) \times \left(\prod_j \frac{1}{\sqrt{2\pi}\tau} e^{-\theta_j^2/2\tau^2}\right)} \tag{6}$$

We need to maximise this or, more conveniently, minimise the log posterior. The log posterior then looks like

$$\boldsymbol{\ln P(\cdot;\theta) = -\frac{1}{2}\mathbf{RSS} - \frac{1}{2}\sum_j \frac{\theta_j^2}{\tau^2}} + \text{terms independent of } \boldsymbol{\theta.} \tag{7}$$

We can ignore the terms that are independent of θ because for minimisation we are only interested in derivatives with respect to θ. We can then compare equation (7) with equation (8) and we see that if we multiply by two and identify $\lambda = 1/\tau^2$ we have the same result.

Problem b) With the LASSO we know that the loss function to optimise is

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\theta_j|^2 \tag{8}$$

If we look back at the derivation of equation (7) we see that modulo terms that do not depend on θ, RSS is the log likelihood and if we then interpret the second term as the log prior, we see that we need a prior of form

$$\mathrm{prior}(\theta_j) \propto e^{-|\theta_j|}, \tag{9}$$

which is a so-called Laplace prior. Thus a Bayesian regression with a Laplace prior on the model parameters leads to the LASSO.