3.5

Consider the ridge regression problem (3.41). Show that this problem is equivalent to the problem

$$\hat{\beta}^c = \arg\min_{\beta^c} \left\{ \sum_{i=1}^{N} [y_i - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j^c]^2 + \lambda \sum_{j=1}^{p} (\beta_j^c)^2 \right\}.$$

Given the correspondence between β^c and the original β in (3.41). Characterize the solution to this modified criterion. Show that a similar result holds for the lasso.

We center each x_{ij} by replacing x_{ij} with $x_{ij} - \bar{x}_j$, then (3.41) becomes

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} [y_i - \beta_0 - \sum_{j=1}^{p} \bar{x}_j\beta_j - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j]^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Looking at the two problems, we can see β^c can be transformed from original β as

$$\begin{aligned} \beta_0^c &= \beta_0 + \sum_{j=1}^{p} \bar{x}_j\beta_j \\ \beta_j^c &= \beta_j \ \text{ for } j = 1, \ldots, p. \end{aligned}$$

It's easy to see that exact same centering technique applies to the lasso.

To characterize the solution, we first take derivative w.r.t β_0^c and set it equal to 0, which yields

$$\sum_{i=1}^{N} \left(y_i - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j \right) = 0,$$

which further implies that $\beta_0^c = \bar{y}$. Next we set

$$\begin{aligned} \tilde{y}_i &= y_i - \beta_0^c, \\ \tilde{x}_{ij} &= x_{ij} - \bar{x}_j, \end{aligned}$$

the problem, in matrix form, becomes

$$\min_{\beta^c} (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta^c)^T (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta^c) + \lambda \beta_c^T \beta_c.$$

It's easy to see the solution is

$$\hat{\beta}_c = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} + \lambda \mathbf{I})^{-1} \tilde{\mathbf{X}}^T \mathbf{y}.$$

3.6

Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim N(0, \tau\mathbf{I})$, and Gaussian sampling model $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ and σ^2.

By Bayes' theorem we have

$$\begin{aligned} P(\beta|\mathbf{y}) &= \frac{P(\mathbf{y}|\beta)P(\beta)}{P(\mathbf{y})} \\ &= \frac{N(\mathbf{X}\beta, \sigma^2\mathbf{I})N(0, \tau\mathbf{I})}{P(\mathbf{y})}. \end{aligned}$$

Taking logarithm on both sides we get

$$\ln(P(\beta|\mathbf{y})) = -\frac{1}{2}\left(\frac{(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)}{\sigma^2} + \frac{\beta^T\beta}{\tau}\right) + C,$$

where C is a constant independent of β. Therefore, if we let $\lambda = \frac{\sigma^2}{\tau}$, then maximizing over β for $P(\beta|\mathbf{y})$ is equivalent to

$$\min_{\beta}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) + \lambda\beta^T\beta,$$

which is essentially solving a ridge regression.

3.7

Assume $y_i \sim N(\beta_0 + x_i^T\beta, \sigma^2)$, $i = 1, 2, \ldots, N$, and the parameters β_j are each distributed as $N(0, \tau^2)$, independently of one another. Assuming σ^2 and τ^2 are known, show that the (minus) log-posterior density of β is proportional to $\sum_{i=1}^{N}(y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$ where $\lambda = \sigma^2/\tau^2$.

By Bayes' theorem we have

$$P(\beta|\mathbf{y}) = \frac{P(\mathbf{y}|\beta)P(\beta)}{P(\mathbf{y})}. \quad (1)$$

By assumptions here we have

$$\begin{aligned} P(\mathbf{y}|\beta) &= \frac{1}{(2\pi)^{N/2}\sigma^N}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2\right), \\ P(\beta) &= \frac{1}{(2\pi)^{p/2}\sigma^p}\exp\left(-\frac{1}{\tau^2}\sum_{j=1}^{p}\beta_j^2\right). \end{aligned}$$

Therefore, with $\lambda = \sigma^2/\tau^2$, from (1) we have

$$-\ln(P(\beta|\mathbf{y})) = \frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}(y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right) + C,$$

where C is a constant independent of β.

The claim is true if and only if $C = 0$, which is not the case here.