1 StatLearn Theoretical exercise 1

Let $\mathbf{X}$ denote an $n \times p$ matrix, $p \leq n$, and let $y \in \mathbb{R}^n$. We assume that the columns of $\mathbf{X}$ have empirical mean zero and assume that y has empirical mean zero. We further assume that $\mathbf{X}$ is not identically zero. Define, for any $\lambda > 0$, $\mathrm{RSS}_\lambda(\beta) = \|y - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$ and $t = \min\{\|\beta\|_2^2 | \mathbf{X}^t\mathbf{X}\beta = \mathbf{X}^t y\}$.

Exercise 1.1. *With $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^t$ being the Singular Value Decomposition of $\mathbf{X}$, $\mathbf{U}$ being $n \times p$ with orthonormal columns, $\mathbf{D}$ being $p \times p$ diagonal and $\mathbf{V}$ being $p \times p$ orthogonal, and $\hat{\beta}(\lambda)$ being the unique minimizer of RSS_λ, show that*

$$\|\hat{\beta}_\lambda\|_2^2 = \sum_{i=1}^{p} \frac{\mathbf{D}_{ii}^2}{(\mathbf{D}_{ii}^2 + \lambda)^2}(y^t\mathbf{U}_i)^2,$$

where $\mathbf{U}_i$ is the i'th column of $\mathbf{U}$.

Solution. We know from our results on ridge regression that

$$\begin{aligned}
\hat{\beta}_\lambda &= (\mathbf{X}^t\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^t y \\
&= ((\mathbf{U}\mathbf{D}\mathbf{V}^t)^t\mathbf{U}\mathbf{D}\mathbf{V}^t + \lambda I_p)^{-1}(\mathbf{U}\mathbf{D}\mathbf{V}^t)^t y \\
&= (\mathbf{V}\mathbf{D}\mathbf{U}^t\mathbf{U}\mathbf{D}\mathbf{V}^t + \lambda I_p)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^t y \\
&= \mathbf{V}\mathbf{V}^t(\mathbf{V}\mathbf{D}^2\mathbf{V}^t + \lambda I_p)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^t y \\
&= \mathbf{V}\mathbf{V}^{-1}(\mathbf{V}\mathbf{D}^2\mathbf{V}^t + \lambda I_p)^{-1}(\mathbf{V}^t)^{-1}\mathbf{D}\mathbf{U}^t y \\
&= \mathbf{V}(\mathbf{V}^t(\mathbf{V}\mathbf{D}^2\mathbf{V}^t + \lambda I_p)\mathbf{V})^{-1}\mathbf{D}\mathbf{U}^t y \\
&= \mathbf{V}(\mathbf{D}^2 + \lambda I_p)^{-1}\mathbf{D}\mathbf{U}^t y.
\end{aligned}$$

Now, for any vector $x \in \mathbb{R}^p$, we have $\|\mathbf{V}x\|_2^2 = (\mathbf{V}x)^t\mathbf{V}x = x^t\mathbf{V}^t\mathbf{V}x = x^t x = \|x\|_2^2$ from orthogonality of $\mathbf{V}$. In particular,

$$\begin{aligned}
\|\hat{\beta}_\lambda\|_2^2 &= \|(\mathbf{D}^2 + \lambda I_p)^{-1}\mathbf{D}\mathbf{U}^t y\|_2^2 \\
&= ((\mathbf{D}^2 + \lambda I_p)^{-1}\mathbf{D}\mathbf{U}^t y)^t(\mathbf{D}^2 + \lambda I_p)^{-1}\mathbf{D}\mathbf{U}^t y \\
&= y^t\mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda I_p)^{-2}\mathbf{D}\mathbf{U}^t y.
\end{aligned}$$

Now, with $\mathbf{A} = \mathbf{D}(\mathbf{D}^2 + \lambda I_p)^{-2}\mathbf{D}$, $\mathbf{A}$ is $p \times p$ diagonal with diagonal entries $\mathbf{D}_{ii}^2(\mathbf{D}_{ii}^2 + \lambda)^{-2}$, and we therefore obtain

$$\begin{aligned}
\|\hat{\beta}_\lambda\|_2^2 &= y^t\mathbf{U}\mathbf{A}\mathbf{U}^t y \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}(y^t\mathbf{U})_i\mathbf{A}_{ij}(\mathbf{U}^t y)_j \\
&= \sum_{i=1}^{p} \frac{\mathbf{D}_{ii}^2}{(\mathbf{D}_{ii}^2 + \lambda)^2}(y^t\mathbf{U})_i(\mathbf{U}^t y)_i \\
&= \sum_{i=1}^{p} \frac{\mathbf{D}_{ii}^2}{(\mathbf{D}_{ii}^2 + \lambda)^2}(y^t\mathbf{U}_i)^2.
\end{aligned}$$

□

Exercise 1.2. *Define $s : [0, \infty) \to \mathbb{R}$ by $s(\lambda) = \|\hat{\beta}_\lambda\|_2^2$. Prove that $s(\lambda) < t$ for all $\lambda > 0$ and that s is continuous and strictly decreasing, with $\lim_{\lambda\to\infty} s(\lambda) = 0$.*

Solution. First note that using the Singular Value Decomposition of $\mathbf{X}$, we find

$$\begin{aligned} t &= \min\{\|\beta\|_2^2 | \mathbf{X}^t\mathbf{X}\beta = \mathbf{X}^t y\} \\ &= \min\{\|\mathbf{V}^t\beta\|_2^2 | \mathbf{D}^2\mathbf{V}^t\beta = \mathbf{D}\mathbf{U}^t y\} \\ &= \min\{\|\beta\|_2^2 | \mathbf{D}^2\beta = \mathbf{D}\mathbf{U}^t y\} \\ &= \min\{\|\beta\|_2^2 | \text{for all } i \text{ such that } \mathbf{D}_{ii} \neq 0 : \beta_i = \mathbf{D}_{ii}^{-2}(\mathbf{D}\mathbf{U}^t y)_i\}. \end{aligned}$$

The requirements on β are now reduced to equations in β_i when $\mathbf{D}_{ii} \neq 0$. Therefore, the minimum is attained by putting the remaining coordinates of β equal to zero, yielding

$$t = \sum_{i:\mathbf{D}_{ii}\neq 0} (\mathbf{D}_{ii}^{-2}(\mathbf{D}\mathbf{U}^t y)_i)^2 = \sum_{i:\mathbf{D}_{ii}\neq 0} \frac{\mathbf{D}_{ii}^2}{\mathbf{D}_{ii}^4}(y^t\mathbf{U}_i)^2.$$

Now fix $\lambda > 0$. As $\mathbf{X}$ is not identically zero, $\mathbf{D}$ is not identically zero, and we then immediately obtain

$$s(\lambda) = \sum_{i=1}^{p} \frac{\mathbf{D}_{ii}^2}{(\mathbf{D}_{ii}^2 + \lambda)^2}(y^t\mathbf{U}_i)^2 < t.$$

Furthermore, as $\lambda \mapsto \frac{x^2}{x^2+\lambda}$ is continuous and strictly decreasing, s is continuous and strictly decreasing, and we have

$$\lim_{\lambda\to\infty} s(\lambda) = \lim_{\lambda\to\infty} \sum_{i=1}^{p} \frac{\mathbf{D}_{ii}^2}{(\mathbf{D}_{ii}^2 + \lambda)^2}(y^t\mathbf{U}_i)^2 = \sum_{i=1}^{p} \lim_{\lambda\to\infty} \frac{\mathbf{D}_{ii}^2}{(\mathbf{D}_{ii}^2 + \lambda)^2}(y^t\mathbf{U}_i)^2 = 0.$$

□

Exercise 1.3. *Show that the problem of minimizing $\beta \mapsto \|y - \mathbf{X}\beta\|_2^2$ subject to $\|\beta\|_2^2 \leq s(\lambda)$ has a unique solution given by $\hat{\beta}_\lambda$.*

Solution. We first show that $\hat{\beta}_\lambda$ is an argument solution to the constrained minization problem. We have $\|\hat{\beta}_\lambda\|_2^2 = s(\lambda)$, and for any β with $\|\beta\|_2^2 \leq s(\lambda)$, we obtain the two inequalities

$$\begin{aligned} \|y - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 &\geq \|y - \mathbf{X}\hat{\beta}_\lambda\|_2^2 + \lambda\|\hat{\beta}_\lambda\|_2^2 \\ \|\beta\|_2^2 &\leq \|\hat{\beta}_\lambda\|_2^2, \end{aligned}$$

which together imply $\|y - \mathbf{X}\hat{\beta}_\lambda\|_2^2 \leq \|y - \mathbf{X}\beta\|_2^2$, so that $\hat{\beta}_\lambda$ is a solution of the constrained minimization problem. Conversely, assume that β_λ^* is a solution to the constrained minimization problem, we want to show that β_λ^* is equal to $\hat{\beta}_\lambda$, so that $\hat{\beta}_\lambda$ is the unique solution

argument to the constrained minimization problem. As $\|\beta_\lambda^*\|_2^2 \leq s(\lambda)$ and $\|\hat{\beta}_\lambda\|_2^2 = s(\lambda)$, we have

$$\begin{aligned} \|y - \mathbf{X}\hat{\beta}_\lambda\|_2^2 &\geq \|y - \mathbf{X}\beta_\lambda^*\|_2^2 \\ \|\beta_\lambda^*\|_2^2 &\leq \|\hat{\beta}_\lambda\|_2^2, \end{aligned}$$

so that $\|y-\mathbf{X}\beta_\lambda^*\|_2^2+\lambda\|\beta_\lambda^*\|_2^2 \leq \|y-\mathbf{X}\hat{\beta}_\lambda\|_2^2+\lambda\|\hat{\beta}_\lambda\|_2^2$. Therefore, β_λ^* is a solution argument of the penalized minimization problem. As the penalized minimization problem has the unique solution $\hat{\beta}_\lambda$, we conclude $\beta_\lambda^* = \hat{\beta}_\lambda$. Finally, we conclude that the constrained minimization problem has the unique solution argument $\hat{\beta}_\lambda$. □

Exercise 1.4. *Show that*

$$\text{tr } \mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^t = \sum_{i=1}^{p} \frac{\mathbf{D}_{ii}^2}{\mathbf{D}_{ii}^2 + \lambda},$$

and show that the trace is always less than or equal to p, and strictly less than p if $\lambda > 0$.

Solution. We have

$$\begin{aligned} \text{tr } \mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^t &= \text{tr } \mathbf{U}\mathbf{D}\mathbf{V}^t((\mathbf{U}\mathbf{D}\mathbf{V}^t)^t\mathbf{U}\mathbf{D}\mathbf{V}^t + \lambda I_p)^{-1}(\mathbf{U}\mathbf{D}\mathbf{V}^t)t \\ &= \text{tr } \mathbf{U}\mathbf{D}\mathbf{V}^t(\mathbf{V}\mathbf{D}\mathbf{U}^t\mathbf{U}\mathbf{D}\mathbf{V}^t + \lambda I_p)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^t \\ &= \text{tr } \mathbf{U}\mathbf{D}\mathbf{V}^{-1}(\mathbf{V}\mathbf{D}^2\mathbf{V}^t + \lambda I_p)^{-1}(\mathbf{V}^t)^{-1}\mathbf{D}\mathbf{U}^t \\ &= \text{tr } \mathbf{U}\mathbf{D}(\mathbf{V}^t(\mathbf{V}\mathbf{D}^2\mathbf{V}^t + \lambda I_p)\mathbf{V})^{-1}\mathbf{D}\mathbf{U}^t \\ &= \text{tr } \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda I_p)^{-1}\mathbf{D}\mathbf{U}^t \\ &= \text{tr } \mathbf{D}\mathbf{U}^t\mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda I_p)^{-1} \\ &= \text{tr } \mathbf{D}^2(\mathbf{D}^2 + \lambda I_p)^{-1}, \end{aligned}$$

and as the matrix in the trace is diagonal with diagonal entries $\mathbf{D}_{ii}^2(\mathbf{D}_{ii}^2 + \lambda)^{-1}$, this proves the expression for the trace. As $\lambda \mapsto \frac{x^2}{x^2+\lambda}$ is decreasing, the trace is less than p. As the mapping is strictly decreasing whenever $x \neq 0$, and $\mathbf{D}$ is nonzero, the trace is strictly less than p if $\lambda > 0$. □