# Data-driven statistical modelling with optimisation VT21
# HW4

Fanny Bergström

March 7 2021

## Exercise 4.1

We suppose that we have two identical variables $X_1 = X_2$ and a response variable $Y$. We perform ridge regression with penalty $\lambda > 0$. The rigde regression problem in this setting is given by

$$\underset{\beta \in \mathbb{R}^2}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{2} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{2} \beta_j^2 \right\}.$$

We further assume that data is centered such that $\beta_0 = 0$. Taking the derivative of above expression with respect to $\beta_1$ and $\beta_2$ respectively, and setting it equal to zero yields

$$\beta_1 (\frac{1}{N} \sum_{i=1}^{N} x_{i1}^2 + 2\lambda) + \beta_2 (\sum_{i=1}^{N} x_{i1}x_{i2}) = \frac{1}{N} \sum_{i=1}^{N} y_i x_{i1}$$

and

$$\beta_1 (\sum_{i=1}^{N} x_{i1}x_{i2}) + \beta_2 (\frac{1}{N} \sum_{i=1}^{N} x_{i2}^2 + 2\lambda) = \frac{1}{N} \sum_{i=1}^{N} y_i x_{i2}.$$

Solving this linear equation by subtracting one from the other yields

$$\beta_1 = \beta_2,$$

and we can conclude that the beta estimates are equal.

## Exercise 4.2

We consider a slightly noisy version of the identical twins example in the beginning of Section 4.2, where the two variables are strongly positively correlated. Figure 1 is showing is a schematic drawing of the contours of the loss function (green) and the penalty function (red and blue). We see that in the left side of the figure with illustrates the lasso, $\beta_1$ will dominate. On the right side, illustrating the elastic net, we can see that $\beta_1$ is less dominating because of the convexity of the penalty function. How convex the penalty is depends on the hyperparameter $\alpha \in [0, 1]$. When $\alpha = 1$ the elastic net becomes the ridge penalty and the coefficient sharing is at its max.

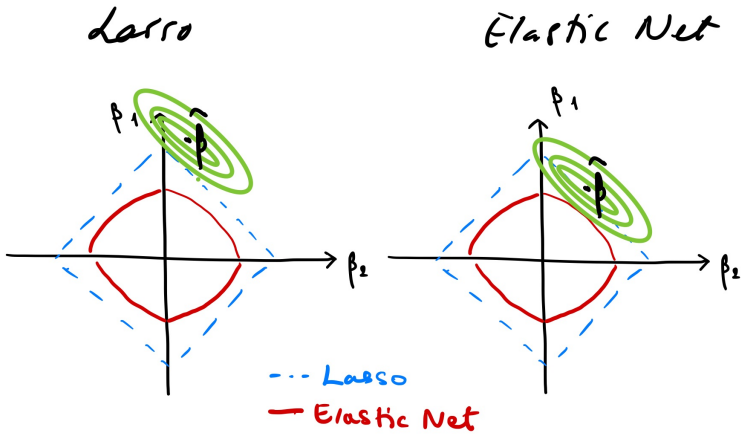


Figure 1: Estimation picture for the lasso (left) and elastic net (right). The dashed blue blue line and the solid red line are the constraint regions. The green ellipses represent the contours of the ML loss function, with the dot at the center being the ML estimate.

## Exercise 4.3

For this exercise, we consider the elastic-net problem (eq (4.2) in the SLS book).

### a)

Minimizing (4.2) with respect to $\beta_0$, we get

$$\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta).$$

Putting this expression equal to zero and solving for $\beta_0$ yields

$$\frac{1}{N}\beta_0 = \sum_{i=1}^{N}(y_i - x_i^T\beta). \tag{1}$$

If each predictor $x$ is centered at 0 then eq (1) becomes

$$\beta_0^c = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y},$$

which is what we wanted to show in this exercise.

The centered regression is given by

$$\begin{aligned}\hat{Y} - \bar{x} &= \beta_0^c + \beta^T(x - \bar{x}) \\ &= \beta_0^c - \beta^T\bar{x} + \beta^T x.\end{aligned}$$

Thus, if we want to convert back to the estimate of $\beta_0$ for the original (uncetered) data we subtract the mean of each $x$ times $\beta$, which then means that the uncentered $\beta_0$ would be given by

$$\beta_0 = \beta_0^c - \beta^T\bar{x}.$$

## b)

We will verify that the update of $\beta_j$ by coordinate descent is given by the following soft-thesholding expression (eq (4.4) in SLS book)

$$\hat{\beta}_j = \frac{S_{\lambda\alpha}(\sum_{i=1}^{N} r_{ij}x_{ij})}{\sum_{i=1}^{N} x_{ij}^2 + \lambda(1-\alpha)}, \tag{2}$$

where $S$ is the soft-thresholding function defined $S_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$ and $r_{ij} = y_i - \hat{\beta}_0 - \sum_{k\neq j} x_{ik}\hat{\beta}_k$ is the partial residual.

Considering the objective function of the elastic net, we want to calculate the gradient at $\beta_j$, which only exists if $\beta_j \neq 0$. We will in the following consider the three cases; $b_j > 0$, $\beta_j < 0$ and $\beta_j = 0$. Supposing that that we have the estimates of $\beta_0$ and $\beta_{k\neq j}$.

If $\beta_j > 0$, then gradient of the elastic net objective function, $R$, with respect to

$\beta_j$ is

$$\frac{\partial R}{\partial \beta_j} = -\sum_{i=1}^{N} x_{ij}(y_i - \beta_0 - x_i^T\beta) + \lambda(1-\alpha)\beta_j + \lambda\alpha \tag{3}$$

$$= -\sum_{i=1}^{N} x_{ij}(y_i - \tilde{\beta}_0 - \sum_{k} x_{ik}\beta_k) + \lambda(1-\alpha)\beta_j + \lambda\alpha \tag{4}$$

$$= -\sum_{i=1}^{N} x_{ij}(y_i - \tilde{\beta}_0 - \sum_{k\neq j} x_{ik}\beta_k) + \frac{1}{N}\sum_{i=1}^{N} x_{ij}^2\beta_j + \lambda(1-\alpha)\beta_j + \lambda\alpha. \tag{5}$$

Next, we put $\frac{\partial R}{\partial \beta_j} = 0$ and solve for $\beta_j$. From eq (5) we get that

$$\beta_j = \frac{\sum_{i=1}^{N} x_{ij}(y_i - \beta_0 - \sum_{k\neq j} x_{ik}\beta_k) - \lambda\alpha}{\sum_{i=1}^{N} x_{ij}^2 + \lambda(1-\alpha)} \tag{6}$$

$$= \frac{\sum_{i=1}^{N} x_{ij}r_{ij} - \lambda\alpha}{\sum_{i=1}^{N} x_{ij}^2 + \lambda(1-\alpha)}. \tag{7}$$

From eq. (7) we can see that $\beta_j$ is positive if $\sum_{i=1}^{N} x_{ij}r_{ij} > \lambda\alpha$.

Next, we consider the case when $\beta_j < 0$, then the derivative of the objective function becomes

$$\frac{\partial}{\partial \beta_j} = -\sum_{i=1}^{N} x_{ij}(y_i - \tilde{\beta}_0 - \sum_{k\neq j} x_{ik}\beta_k) + \frac{1}{N}\sum_{i=1}^{N} x_{ij}^2\beta_j + \lambda(1-\alpha)\beta_j + \lambda\alpha.$$

Solving for $\beta_j$, we get that

$$\beta_j = \frac{\sum_{i=1}^{N} x_{ij}r_{ij} - \lambda\alpha}{\sum_{i=1}^{N} x_{ij}^2 + \lambda(1-\alpha)}.$$

In this case, $\beta_j < 0$, when $\sum_{i=1}^{N} x_{ij}r_{ij} < -\lambda\alpha$.

In the third case when $\beta_j = 0$, we consider the derivative of $R$

$$\frac{\partial R}{\partial \beta_j} = -\sum_{i=1}^{N} x_{ij}(y_i - \tilde{\beta}_0 - \sum_{k\neq j} x_{ik}\beta_k) + \frac{1}{N}\sum_{i=1}^{N} x_{ij}^2\beta_j + \lambda(1-\alpha)\beta_j + s_j\lambda\alpha,$$

where $s_j \in [0,1]$ is the subgradient. Then the $b_j$ update is given by

$$\beta_j = \frac{\sum_{i=1}^{N} x_{ij}r_{ij} - s_j\lambda\alpha}{\sum_{i=1}^{N} x_{ij}^2 + \lambda(1-\alpha)},$$

where $b_j = 0$ if $|\sum_{i=1}^N x_{ij} r_{ij}| \leq \lambda\alpha$.

Finally, we can summarise the results from the three cases

$$\beta_j = \begin{cases} \frac{\sum_{i=1}^N x_{ij} r_{ij} + \lambda\alpha}{\sum_{i=1}^N x_{ij}^2 + \lambda(1-\alpha)} & \text{if} \quad \sum_{i=1}^N x_{ij} r_{ij} > \lambda\alpha \\ 0 & \text{if} \quad |\sum_{i=1}^N x_{ij} r_{ij}| \leq \lambda\alpha \\ \frac{\sum_{i=1}^N x_{ij} r_{ij} - \lambda\alpha}{\sum_{i=1}^N x_{ij}^2 + \lambda(1-\alpha)} & \text{if} \quad \sum_{i=1}^N x_{ij} r_{ij} < -\lambda\alpha, \end{cases}$$

which can also be written as

$$\begin{aligned} \beta_j =& \frac{\text{sign}(\sum_{i=1}^N x_{ij} r_{ij})(|\sum_{i=1}^N x_{ij} r_{ij}| - \lambda\alpha)_+}{\sum_{i=1}^N x_{ij}^2 + \lambda(1-\alpha)} \\ =& \frac{S_{\lambda\alpha}(\sum_{i=1}^N r_{ij} x_{ij})}{\sum_{i=1}^N x_{ij}^2 + \lambda(1-\alpha)}. \end{aligned}$$

This shows that the update of $\beta_j$ is given by the expression in eq. (2), which is what we wanted to do in this excersie.