LASSO

- Start with the simple case in which $\mathbf{X}$ is orthogonal.
  - How to derive the solution $\hat{\boldsymbol{\beta}}^{\text{lasso}}$?
  - Understand the selection/shrinkage effect of Lasso?
  - What's the difference between Lasso and Ridge?
- Coordinate Decent for general $\mathbf{X}$ (leave the computation to `R`).
- What if $p > n$?
- How to select the tuning parameter λ? (see R page)

The Lasso solution is define to be

$$\hat{\boldsymbol{\beta}}_{\mathsf{lasso}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda|\boldsymbol{\beta}|\right).$$

Suppose $\mathbf{X}_{n\times p}$ is orthogonal, i.e., $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$. Then

$$\begin{aligned}
\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 &= \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{LS}} + \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{LS}} - \mathbf{X}\boldsymbol{\beta}\|^2 \\
&= \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{LS}}\|^2 + \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathsf{LS}} - \mathbf{X}\boldsymbol{\beta}\|^2 \qquad (2)
\end{aligned}$$

where the cross-product term,

$$2(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{LS}})^T(\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{LS}} - \mathbf{X}\boldsymbol{\beta}) = 2\mathbf{r}^T(\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{LS}} - \mathbf{X}\boldsymbol{\beta}) = 0,$$

since the n-dim vector in red (which is a linear combination of columns of $\mathbf{X}$, no matter what value $\boldsymbol{\beta}$ takes) is in $C(\mathbf{X})$, therefore orthogonal to the residual vector $\mathbf{r}$. Also note that the 1st term in (2) is not a function of $\boldsymbol{\beta}$. Therefore

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathsf{lasso}} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \lambda|\boldsymbol{\beta}|\right) \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(\|\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{LS}} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda|\boldsymbol{\beta}|\right) \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left[(\hat{\boldsymbol{\beta}}^{\mathsf{LS}} - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}^{\mathsf{LS}} - \boldsymbol{\beta}) + \lambda|\boldsymbol{\beta}|\right] \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left[(\hat{\boldsymbol{\beta}}^{\mathsf{LS}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}^{\mathsf{LS}} - \boldsymbol{\beta}) + \lambda|\boldsymbol{\beta}|\right] \\
&= \arg\min_{\beta_1,\ldots,\beta_p} \sum_{j=1}^{p} \left[(\beta_j - \hat{\beta}_j^{\mathsf{LS}})^2 + \lambda|\beta_j|\right].
\end{aligned}
$$

So we can solve the optimal β_j for each of $j = 1, \ldots, p$ separately by solving the following generic problem:

$$
\arg\min_x (x-a)^2 + \lambda|x|, \quad \lambda > 0.
$$