1 Ridge Regression

1.1 Equivalence between two forms

$$\begin{aligned}\hat{\beta}_{\text{ridge},t} &= \underset{\tilde{\beta}}{\operatorname{argmin}}\, ||Y - \mathbb{X}\tilde{\beta}||_2^2 \\ &\text{with } ||\tilde{\beta}||_2^2 - t \leqslant 0\end{aligned} \tag{1}$$

And its Lagrangian form:

$$\hat{\beta}_{\text{ridge},\lambda} = \underset{\beta}{\operatorname{argmin}}\, ||Y - \mathbb{X}\beta||_2^2 + \lambda||\beta||_2^2. \tag{2}$$

1 and 2 are equivalent because of the following arguments:

In 1, $\hat{\beta}_{\text{ridge},t}$ satisfies $||\tilde{\beta}||_2^2 - t = 0$ because the minimum is stained at the edge of the constraints. Notice by Lagrangian, it's equivalent to minimizing :

$$L(\beta, \alpha) = ||Y - \mathbb{X}\beta||_2^2 + \alpha(||\beta||_2^2 - t)$$

Suppose the solution for 2 is β_λ, then let $t = ||\beta_\lambda||_2^2$, $\alpha = \lambda$, $\beta = \beta_\lambda$. We'll have $\dfrac{\partial L}{\partial \beta} = 0$, $\dfrac{\partial L}{\partial \alpha} = 0$. Since the solution of Lagrangian is achieved at $\nabla L = 0$. So when $t = ||\beta_\alpha||_2^2$, $\beta = \beta_\lambda$ are exactly the solution to 1. Thus the equivalence holds.

1.2 Ridge Estimator

Take derivative of 2, in the form of $(Y - \mathbb{X}\beta))^\top(Y - \mathbb{X}\beta) + \lambda\beta^\top\beta$ w.r.t β, we get:

$$\begin{aligned}&\frac{\partial(Y^\top Y - 2Y^\top\mathbb{X}\beta + \beta^\top\mathbb{X}^\top\mathbb{X}\beta + \lambda\beta^\top\beta)}{\partial\beta} \\ &= -2Y^\top\mathbb{X} + 2\beta^\top\mathbb{X}^\top\mathbb{X} + 2\lambda\beta^\top = 0\end{aligned} \tag{3}$$

Use 3, we get $\hat{\beta}_{\text{ridge},\lambda} = (\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top Y$

1.3 SVD interpretation

Assume $\mathbf{rank}(\mathbb{X}_{n\times p}) = k \leqslant p, n$. By SVD decomposition, $\mathbb{X} = U_{n\times k}D_{k\times k}V_{p\times k}^\top$, or $\mathbb{X} = \tilde{U}_{n\times n}\tilde{D}_{n\times p}\tilde{V}_{p\times p}^\top$. The two forms are equivalent and first k columns of U (V) are the same as $\tilde{U}$ $(\tilde{V})$, nonzero diagonal elements of $\tilde{D}$ are the same as diagonal elements of D. A more careful comparison of LS and ridge regressions are given below:

$$\begin{aligned}
\hat{\beta}_{\text{LS}} &= VD^{-1}U^\top Y & &= \sum_{j=1}^{k} \mathbf{v}_j \left(\frac{1}{d_j}\right) \mathbf{u}_j^\top Y \\
\hat{\beta}_{\text{ridge},\lambda} &= \tilde{V}(\tilde{D}^\top \tilde{D} + \lambda I)^{-1} \tilde{D}^\top \tilde{U}^\top Y & &= \sum_{j=1}^{p} \mathbf{v}_j \left(\frac{d_j}{d_j^2 + \lambda}\right) \mathbf{u}_j^\top Y.
\end{aligned}$$

Similarly

$$\begin{aligned}
\mathbb{X}\hat{\beta}_{\text{LS}} &= UU^\top Y & &= \sum_{j=1}^{k} \mathbf{u}_j \mathbf{u}_j^\top Y \\
\mathbb{X}\hat{\beta}_{\text{ridge},\lambda} &= \tilde{U}\tilde{D}(\tilde{D}^\top \tilde{D} + \lambda I)^{-1} \tilde{D}^\top \tilde{U}^\top Y & &= \sum_{j=1}^{p} \mathbf{u}_j \left(\frac{d_j^2}{d_j^2 + \lambda}\right) \mathbf{u}_j^\top Y.
\end{aligned}$$

We can see the LS estimation is equivalent to projecting Y to the column space of U, while ridge regression shrinks the projection to each direction by a factor of $\frac{d_j^2}{d_j^2 + \lambda}$.

1.4 Bayesian interpretation

$$Y_i \sim N(X_i^\top \beta, \sigma^2)$$

and put a prior distribution of $\beta \sim N(\mathbf{0}, \tau^2 I)$.

Then we have the following posterior (making some conditional independence assumptions)

$$p(\beta|Y, \mathbb{X}, \sigma^2, \tau^2) \propto p(Y|\mathbb{X}, \beta, \sigma^2) p(\beta|\tau^2).$$

The kernel of exponential function is $-\frac{2}{\sigma^2}\beta^\top \mathbb{X}^\top Y + \beta^\top (\frac{\mathbb{X}^\top \mathbb{X}}{\sigma^2} + \frac{1}{\tau^2})\beta$, that is the quadratic form of β, so the posterior distribution of β is normal and its mean is ridge estimator with $\lambda = \sigma^2/\tau^2$.

1.5 Generalized Cross Validation

Similar to LS regression, for ridge regression, define hat matrix $H^\lambda = \mathbb{X}(\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1} \mathbb{X}^\top$, so $\hat{Y} = H^\lambda Y$. Let $H^\lambda = \left(\mathbf{h}_1^\lambda, \ldots, \mathbf{h}_n^\lambda\right)^\top$, and define the usual leave-one-out cross validation as:

$$\begin{aligned}
\mathbb{V}_0(\lambda) &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbb{X}\beta_\lambda^{[i]})^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \mathbf{h}_k^{\lambda\top} Y)^2}{(1 - H_{kk}^\lambda)^2}
\end{aligned}$$

where

$$\beta_\lambda^{[k]} = \text{argmin} \sum_{\substack{i=1 \\ i \neq k}}^{n} (Y_i - X\beta_i)^2 + \lambda ||\beta||_2^2$$

So λ can be chosen to minimize $\mathbb{V}_0(\lambda)$.
And the GCV (Generalized Cross Validation) is defined as:

$$\begin{aligned} \mathbb{V}(\lambda) &= \frac{\frac{1}{n}||(I-H)Y||_2^2}{\left[\frac{1}{n}\text{trace}(I-H)\right]^2} \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{X}\beta_\lambda^{[i]})^2 w_k(\lambda) \end{aligned}$$

where:

$$w_k(\lambda) = \frac{(1-H_{kk})^2}{\left[\frac{1}{n}\text{trace}(I-H)\right]^2}$$

This definition is the same as what we've learned in class, since $tr(H) = df(\hat{\beta})$. And the GCV is a weighted version of usual CV in the case of ridge regression [1].

1.6 Ridge in practice

Ridge regression can be simply implemented using **lm** by data augmentation, thus using

$$\tilde{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{n+p} \text{ and } \tilde{\mathbb{X}} = \begin{bmatrix} \mathbb{X} \\ \sqrt{\lambda} I \end{bmatrix}$$

or use **glmnet**.

2 Lasso

2.1 why Lasso?

In short, ℓ_1 constraint is both convex and able to do model selection (in the sense it forces some parameters to be 0). The estimator satisfies

$$\hat{\beta}_{lasso}(t) = \underset{||\beta||_1 \leq t}{\text{argmin}} ||\mathbb{Y} - \mathbb{X}\beta||_2^2$$

In its corresponding Lagrangian dual form:

$$\hat{\beta}_{lasso}(\lambda) = \underset{\beta}{\text{argmin}} ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda||\beta||_1$$

2.2 Lasso in practice

glmnet uses gradient descent to quickly fit the lasso solution. **lars** is another option, which exploits the fact that the coefficient profiles are piecewise linear and leads to an algorithm with the same computational cost as the full least-squares fit on the data.

The tuning parameter is often chosen by cross-validation.

References

[1] G. H. Golub, M. Heath, G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics 21 (2) (1979) 215–223.