Generalized Ridge regression

Consider the regression model

$$Y = X\beta_* + \varepsilon\,,$$

where $X \in \mathbb{R}^{n\times d}$, β_* is an unknown vector in $\mathbb{R}^d$ and $\varepsilon \sim \mathcal{N}(0,\sigma^2 I_n)$. Define the generalized Ridge estimator by:

$$\widehat{\beta} \in \mathrm{Argmin}_{\beta\in\mathbb{R}^d}\left\{(Y-X\beta)^\top W(Y-X\beta) + (\beta-\beta_0)^\top\Delta(\beta-\beta_0)\right\}\,,$$

where $\beta_0 \in \mathbb{R}^d$, $W \in \mathbb{R}^{n\times n}$ is a diagonal matrix with elements in $[0,1]$, $\Delta \in \mathbb{R}^{d\times d}$ is a symmetric definite-positive matrix.

1. Provide the expression of $\widehat{\beta}$ when $\beta_0 = 0$, $W = I_n$ and $\Delta = \lambda I_d$ where $\lambda > 0$.

   Proof in lecture notes.

2. Solve the optimization problem in the general case.

   For all $\beta \in \mathbb{R}^d$, write

   $$\mathcal{L}(\beta) = (Y-X\beta)^\top W(Y-X\beta) + (\beta-\beta_0)^\top\Delta(\beta-\beta_0)\,.$$

   Therefore, for all $\beta \in \mathbb{R}^d$,

   $$\nabla\mathcal{L}(\beta) = 2\left(\left(X^\top WX + \Delta\right)\beta - \Delta\beta_0 - X^\top WY\right)\,.$$

   Note that $X^\top WX + \Delta$ is definite-positive so that $\nabla\mathcal{L}(\beta) = 0$ has a unique solution given by

   $$\widehat{\beta} = \left(X^\top WX + \Delta\right)^{-1}(\Delta\beta_0 + X^\top WY)\,.$$

3. Compute $\mathbb{E}[\widehat{\beta}]$ and show that the estimator is unbiased when $\beta_0 = \beta_*$.

   Assuming that the design is not random,

   $$\mathbb{E}[\widehat{\beta}] = \left(X^\top WX + \Delta\right)^{-1}(\Delta\beta_0 + X^\top W\mathbb{E}[Y])\,.$$

   This yields

   $$\mathbb{E}[\widehat{\beta}] = \left(X^\top WX + \Delta\right)^{-1}(\Delta\beta_0 + X^\top WX\beta_*)\,.$$

   In the case where $\beta_0 = \beta_*$,

   $$\mathbb{E}[\widehat{\beta}] = \left(X^\top WX + \Delta\right)^{-1}\left(X^\top WX + \Delta\right)\beta_* = \beta_*$$

   and the estimator is unbiased.

4. Compute $\mathbb{V}[\widehat{\beta}]$ and the mean squared error $\mathbb{E}[\|\widehat{\beta} - \beta_*\|_2^2]$ when $\beta_0 = \beta_*$.

   By definition of $\widehat{\beta}$,

   $$\begin{aligned}\mathbb{V}[\widehat{\beta}] &= \left(X^\top WX + \Delta\right)^{-1}X^\top W\mathbb{V}[Y]W^\top X\left(X^\top WX + \Delta\right)^{-1}\\ &= \sigma^2\left(X^\top WX + \Delta\right)^{-1}X^\top WW^\top X\left(X^\top WX + \Delta\right)^{-1}\\ &= \sigma^2\left(X^\top WX + \Delta\right)^{-1}X^\top W^2X\left(X^\top WX + \Delta\right)^{-1}\,.\end{aligned}$$

If $\beta_0 = \beta_*$, as the estimator is unbiased,

$$
\begin{aligned}
\mathbb{E}[\|\widehat{\beta} - \beta_*\|_2^2] &= \mathrm{Trace}\left(\mathbb{V}[\widehat{\beta}]\right) \\
&= \sigma^2 \mathrm{Trace}\left(\left(X^\top W X + \Delta\right)^{-1} X^\top W^2 X \left(X^\top W X + \Delta\right)^{-1}\right) \\
&= \sigma^2 \mathrm{Trace}\left(X^\top W^2 X \left(X^\top W X + \Delta\right)^{-2}\right) .
\end{aligned}
$$

5. Assume that $W = I_n$, $\beta_0 = 0$ and $\Delta = V\Lambda V^\top$ where $X = UDV^\top$ is a singular value decomposition of X and Λ is a diagonal matrix with positive diagonal components. Provide an expression of $\widehat{\beta}$ as a function of U, D, V, Λ and Y.

   In the proposed setting,
   $$\widehat{\beta} = \left(X^\top X + \Delta\right)^{-1} X^\top Y .$$
   Let $X = UDV^\top$ be a singular value decomposition of X and choose $\Delta = V\Lambda V^\top$. Then,
   $$
   \begin{aligned}
   \widehat{\beta} &= \left((UDV^\top)^\top UDV^\top + V\Lambda V^\top\right)^{-1} (UDV^\top)^\top Y \\
   &= \left(VD^\top U^\top UDV^\top + V\Lambda V^\top\right)^{-1} VD^\top U^\top Y \\
   &= V\left(D^\top D + \Lambda\right)^{-1} D^\top U^\top Y .
   \end{aligned}
   $$
   Contrary to the classical Ridge estimator, this estimator shrinks values of β with a different penalty for each component thanks to the matrix Λ.