IMA205 TP3 Theoretical Questions

Felipe Vicentin

March 14, 2025

OLS

Questions

1. Demonstrate that OLS is the estimator with the smallest variance: compute $\mathbf{E}[\tilde{\beta}]$ and $\mathrm{Var}(\tilde{\beta}) = \mathbf{E}[(\tilde{\beta} - \mathbf{E}[\tilde{\beta}])(\tilde{\beta} - \mathbf{E}[\tilde{\beta}])^\top]$ and show when and why $\mathrm{Var}(\beta^*) < \mathrm{Var}(\tilde{\beta})$. Which assumption of OLS do we need to use?

Answers

1. Let $\tilde{\beta} = (H + D)\mathbf{y}$. Then, using the fact that $\mathrm{Var}(A\mathbf{x}) = A\mathrm{Var}(\mathbf{x})A^\top$ for unbiased $\mathbf{x}$ and that $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, we have:

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \mathrm{Var}((H + D)\mathbf{y}) \\
&= \mathrm{Var}(H\mathbf{y}) + \mathrm{Var}(D\mathbf{y}) \\
&= \mathrm{Var}(\beta^\star) + D\mathrm{Var}(\mathbf{y})D^\top \\
&= \mathrm{Var}(\beta^\star) + D(\mathbf{y}\mathbf{y}^\top)D^\top \\
&= \mathrm{Var}(\beta^\star) + \|D\mathbf{y}\|_2^2
\end{aligned}$$

It is evident that $\mathrm{Var}(\tilde{\beta}) \geq \mathrm{Var}(\beta^\star)$. Strict inequality is obtained when the spectral radius of D is non-zero (i.e., $D \neq 0$). Of course, we are assuming the OLS estimator is unbiased. This is easily verifiable when considering that $(X^\top X)^{-1}$ exists and that the error in the data has 0 mean:

$$\begin{aligned}
\mathbf{E}[\beta^\star] &= \mathbf{E}[(X^\top X)^{-1} X^\top y] \\
&= \mathbf{E}[(X^\top X)^{-1} X^\top (X\beta + \varepsilon)] \\
&= (X^\top X)^{-1} X^\top (\mathbf{E}[X\beta] + \mathbf{E}[\varepsilon]) \\
&= \beta
\end{aligned}$$

Ridge regression

Questions

2. Show that the estimator of ridge regression is biased (that is $\mathbf{E}[\beta^*_{\text{ridge}}] \neq \beta$).

3. Recall that the SVD decomposition is $\mathbf{x_c} = UDV^T$. Write down by hand the solution β^*_{ridge} using the SVD decomposition. When is it useful using this decomposition? Hint: do you need to invert a matrix?

4. Remember that $\mathrm{Var}(\beta^*_{\text{OLS}}) = \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}$. Show that $\mathrm{Var}(\beta^*_{\text{OLS}}) \geq \mathrm{Var}(\beta^*_{\text{ridge}})$.

5. When λ increases what happens to the bias and to the variance? Hint: Compute $\mathrm{MSE} = \mathbf{E}[(y_0 - x_0^T\beta^*_{ridge})^2]$ at the test point (x_0, y_0) with $y_0 = x_0^T\beta + \epsilon_0$ being the true model and $x_0^T\beta^*_{\text{ridge}}$ the ridge estimate.

6. Show that $\beta^*_{\text{ridge}} = \frac{\beta^*_{\text{OLS}}}{1+\lambda}$ when $\mathbf{x_c}^T\mathbf{x_c} = I_d$.

Answers

2. As seen in class, $\beta^\star_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y}$. Then, it is easy to see that

$$\begin{aligned}\mathbf{E}[\beta^\star_{\text{ridge}}] &= \mathbf{E}[(X^\top X + \lambda I)^{-1} X^\top \mathbf{y}] \\ &= (X^\top X + \lambda I)^{-1} X^\top (\mathbf{E}[X\beta] + \mathbf{E}[\varepsilon]) \\ &= (X^\top X + \lambda I)^{-1} X^\top X\beta\end{aligned}$$

   Since $(X^\top X + \lambda I)^{-1} X^\top X \neq I$ for $\lambda > 0$, the estimator is biased.

3. If we use $X = UDV^\top$, we have

$$\begin{aligned}\beta^\star_{\text{ridge}} &= (X^\top X + \lambda I)^{-1} X^\top \mathbf{y} \\ &= ((UDV^\top)^\top UDV^\top + \lambda I)^{-1} (UDV^\top)^\top \mathbf{y} \\ &= (VDU^\top UDV^\top + \lambda I)^{-1} VDU^\top \mathbf{y} \\ &= (V(D^2 + \lambda I)V^\top)^{-1} VDU^\top \mathbf{y} \\ &= V(D^2 + \lambda I)^{-1} V^\top VDU^\top \mathbf{y} \\ &= V(D^2 + \lambda I)^{-1} DU^\top \mathbf{y}\end{aligned}$$

   We see that $D^2 + \lambda I$ is diagonal and no inversion is needed. This speeds up the algorithm substantially. The form above is particularly useful when testing many λ to find the best hyperparameter.

4. Let us first compute the variance of the ridge estimator.

$$\begin{aligned}\text{Var}(\beta^\star_{\text{ridge}}) &= \text{Var}((X^\top X + \lambda I)^{-1} X^\top \mathbf{y}) \\ &= (X^\top X + \lambda I)^{-1} X^\top \text{Var}(\mathbf{y}) X (X^\top X + \lambda I)^{-1} \\ &= (X^\top X + \lambda I)^{-1} X^\top \text{Var}(\varepsilon) X (X^\top X + \lambda I)^{-1} \\ &= \sigma^2 (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1}\end{aligned}$$

   We know that the eigenvalues of $X^\top X + \lambda I$ are equal to the eigenvalues of $X^\top X$ plus $\lambda \geq 0$, so $X^\top X \preccurlyeq X^\top X + \lambda I$. In turn, this implies that:

$$\begin{aligned}(X^\top X + \lambda I)^{-1} &\preccurlyeq (X^\top X)^{-1} \\ (X^\top X + \lambda I)^{-1} (X^\top X) &\preccurlyeq (X^\top X)^{-1} (X^\top X) \\ (X^\top X + \lambda I)^{-1} (X^\top X) (X^\top X + \lambda I)^{-1} &\preccurlyeq (X^\top X)^{-1} (X^\top X) (X^\top X + \lambda I)^{-1} \\ (X^\top X + \lambda I)^{-1} (X^\top X) (X^\top X + \lambda I)^{-1} &\preccurlyeq (X^\top X)^{-1} (X^\top X) (X^\top X)^{-1} \\ \sigma^2 (X^\top X + \lambda I)^{-1} (X^\top X) (X^\top X + \lambda I)^{-1} &\preccurlyeq \sigma^2 (X^\top X)^{-1} \\ \text{Var}(\beta^\star_{\text{ridge}}) &\preccurlyeq \text{Var}(\beta^\star_{\text{OLS}})\end{aligned}$$

5. The bias of the Ridge estimator is given by

$$\begin{aligned}b(\beta^\star_{\text{ridge}}) &= \mathbf{E}(\beta^\star_{\text{ridge}}) - \beta \\ &= ((X^\top X + \lambda I)^{-1} X^\top X - I)\beta\end{aligned}$$

   We see that as λ increases, the bias gets greater in absolute value. The variance, on the other hand, gets smaller.

6. If $X^\top X = I$, then $\beta^\star_{\text{OLS}} = (X^\top X)^{-1} X^\top \mathbf{y} = X^\top \mathbf{y}$. The ridge estimator is $\beta^\star_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y} = (1 + \lambda) I^{-1} X^\top \mathbf{y} = \frac{1}{\lambda+1} X^\top \mathbf{y}$. So, we can conclude that $\beta^\star_{\text{ridge}} = \frac{\beta^\star_{\text{OLS}}}{1+\lambda}$.

Elastic Net

Questions

7. Compute by hand the solution of Eq.2 supposing that $\mathbf{x_c}^T\mathbf{x_c} = I_d$ and show that the solution is:

$$(\beta^*_{\text{ElNet}})_j = \frac{(\beta^*_{\text{OLS}})_j \pm \frac{\lambda_1}{2}}{1+\lambda_2}$$

Answers

7.

$$\begin{aligned}
f(\beta) &= (\mathbf{y}_c - \mathbf{x}_c\beta)^\top(\mathbf{y}_c - \mathbf{x}_c\beta) + \lambda_2\|\beta\|_2^2\lambda_1\|\beta\|_1 \\
&= \mathbf{y}_c^\top\mathbf{y}_c - 2\mathbf{y}_c^\top\mathbf{x}_c\beta + \beta^\top\mathbf{x}_c^\top\mathbf{x}_c\beta + \lambda_2\beta^\top\beta + \lambda_1\|\beta\|_1 \\
&= \mathbf{y}_c^\top\mathbf{y}_c - 2\mathbf{y}_c^\top\mathbf{x}_c\beta + (1+\lambda_2)\beta^\top\beta + \lambda_1\|\beta\|_1 \\
\implies \partial f(\beta) &= -2\mathbf{x}_c^\top\mathbf{y}_c + 2(1+\lambda_2)\beta + \lambda_1\partial(\|\cdot\|_1)(\beta) \\
\implies -\mathbf{x}_c^\top\mathbf{y}_c + (1+\lambda_2)\beta^\star &\in -\frac{\lambda_1}{2}\partial(\|\cdot\|_1)(\beta^\star) \\
\implies \beta_j^\star &\in \begin{cases} \{\frac{\mathbf{x}_c^\top\mathbf{y}_c + \frac{\lambda_1}{2}}{\lambda_2+1}\} & \text{if } \beta_j^\star < 0 \\ \{\frac{\mathbf{x}_c^\top\mathbf{y}_c - \frac{\lambda_1}{2}}{\lambda_2+1}\} & \text{if } \beta_j^\star > 0 \\ [\frac{\mathbf{x}_c^\top\mathbf{y}_c - \frac{\lambda_1}{2}}{\lambda_2+1}, \frac{\mathbf{x}_c^\top\mathbf{y}_c + \frac{\lambda_1}{2}}{\lambda_2+1}] & \text{if } \beta_j^\star = 0 \end{cases}
\end{aligned}$$

Since $\beta^\star_{\text{OLS}} = \mathbf{x}_c^\top\mathbf{y}_c$, we have that:

$$(\beta^\star_{\text{ElNet}})_j = \begin{cases} \frac{(\beta^\star_{\text{OLS}})_j - \frac{\lambda_1}{2}}{\lambda_2+1} & (\beta^\star_{\text{OLS}})_j > \frac{\lambda_1}{2} \\ \frac{(\beta^\star_{\text{OLS}})_j + \frac{\lambda_1}{2}}{\lambda_2+1} & (\beta^\star_{\text{OLS}})_j < \frac{\lambda_1}{2} \\ 0 & |(\beta^\star_{\text{OLS}})_j| \leq \frac{\lambda_1}{2} \end{cases}.$$