IMA205 - Introduction to Supervised Learning

Marina Sangineto Jucá

Theoretical Questions

Ordinary Least Squares (OLS)

(a) **Variance of OLS vs. Unbiased Estimators:**
Let $\tilde{\beta} = C\mathbf{y} = (H + D)\mathbf{y}$ be another linear unbiased estimator. Since $\mathbf{E}[\tilde{\beta}] = C\mathbf{X}\beta = \beta$ (unbiasedness), we require $C\mathbf{X} = I$. For OLS, $H = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, so $H\mathbf{X} = I$. For $\tilde{\beta}$:

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2(H + D)(H + D)^T.$$

Subtracting $\mathrm{Var}(\beta^*) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, we get:

$$\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\beta^*) = \sigma^2 DD^T \succeq 0.$$

Thus, $\mathrm{Var}(\beta^*) < \mathrm{Var}(\tilde{\beta})$ unless $D = 0$. This relies on the **Gauss-Markov assumptions** (homoscedasticity, no autocorrelation).

Ridge Regression

(a) **Bias of Ridge Estimator:**
The Ridge estimator $\beta^*_{\text{ridge}} = (\mathbf{X}_c^T\mathbf{X}_c + \lambda I)^{-1}\mathbf{X}_c^T\mathbf{y}_c$. Its expectation:

$$\mathbf{E}[\beta^*_{\text{ridge}}] = (\mathbf{X}_c^T\mathbf{X}_c + \lambda I)^{-1}\mathbf{X}_c^T\mathbf{X}_c\beta \neq \beta.$$

Hence, it is biased.

(b) **SVD Decomposition:**
Let $\mathbf{X}_c = UDV^T$. Then:

$$\beta^*_{\text{ridge}} = V(D^2 + \lambda I)^{-1}DU^T\mathbf{y}_c.$$

SVD avoids direct inversion of $\mathbf{X}_c^T\mathbf{X}_c + \lambda I$, useful for ill-conditioned matrices.

(c) **Variance Comparison:**
$\mathrm{Var}(\beta^*_{\text{ridge}}) = \sigma^2 V(D^2 + \lambda I)^{-2}D^2V^T$. Since $(D^2 + \lambda I)^{-2}D^2 \preceq (D^2)^{-1}$, we have $\mathrm{Var}(\beta^*_{\text{OLS}}) \geq \mathrm{Var}(\beta^*_{\text{ridge}})$.

(d) **Bias-Variance Trade-off:**
As $\lambda \uparrow$, bias $\uparrow$ (deviation from true β), variance $\downarrow$ (shrinkage). The MSE at test point (x_0, y_0):

$$\text{MSE} = \text{Bias}^2 + \text{Variance} + \sigma^2.$$

Initially, MSE decreases (variance drops faster than bias increases), then increases.

(e) **Special Case ($\mathbf{X}_c^T \mathbf{X}_c = I_d$):**
Substituting $\mathbf{X}_c^T \mathbf{X}_c = I_d$ into Ridge:

$$\beta^*_{\text{ridge}} = (\mathbf{X}_c^T \mathbf{X}_c + \lambda I)^{-1} \mathbf{X}_c^T \mathbf{y}_c = \frac{\beta^*_{\text{OLS}}}{1+\lambda}.$$

Elastic Net

(a) **Advantages Over Lasso:**

- Removes Lasso's N-variable limit when $d > N$.
- Groups correlated variables instead of random selection.
- Stabilizes solution paths.
- Combines Ridge and Lasso for better prediction in high correlation.

(b) **Solution Under $\mathbf{X}_c^T \mathbf{X}_c = I_d$:**
The Elastic Net objective becomes:

$$\arg\min_{\beta} \|\mathbf{y}_c - \mathbf{X}_c \beta\|^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1.$$

Taking subgradients:

$$\beta_j = \frac{\text{sign}(\beta^*_{\text{OLS},j})(|\beta^*_{\text{OLS},j}| - \lambda_1/2)}{1+\lambda_2}.$$

This matches the given solution structure.