**STA 561: Homework 2 (Due January 31 at midnight)**

Reminder: work together! Share ideas, brainstorm, explain/verify your answers but write up your own work. Your homework should be submitted as pdf file generated using either latex or an python notebook.

1. (Hard-thresholding) Suppose that $X_1, \ldots, X_n \sim_{i.i.d.}$ Normal$(\mu, \sigma^2)$. Consider the hard-thresholding estimator $\widehat{\mu}_n^H = \overline{X}_n 1_{|\overline{X}_n| \geq \alpha}$ of $\mu$. Our goal in this problem is to derive the value $\alpha^{\text{opt}}$ which minimizes $\mathbb{E}(X - \widehat{\mu}_n^H)^2$.

   (a) Suppose $Z \sim N(\omega, \tau^2)$. Derive a closed form expression for $\mathbb{E}Z1_{|Z| \leq \gamma}$, where $\gamma$ is a constant.

   (b) Derive a closed form expression for $\mathbb{E}\left(X - \widehat{\mu}_n^H\right)^2$. Explain how you would solve for $\alpha^{\text{opt}}$. You do not need a closed form expression for $\alpha^{\text{opt}}$ [1].

   (c) Your answer depends on unknown parameters $\mu$ and $\sigma^2$. Suggest an estimator of $\alpha^{\text{opt}}$, say $\widehat{\alpha}_n$.

[1] Note that $\mathbb{E}Z1_{|Z| \geq \gamma} = \mathbb{E}Z - \mathbb{E}Z1_{|Z| \leq \gamma}$, which you derived in part (a).

2. (Great British Bake-off) In this problem you will be comparing Lasso and ridge regression in terms of MSE. You should implement the following methods:

| Estimator | Tuning method |
|---|---|
| Lasso | AIC |
| Lasso | BIC |
| Lasso | LOO-CV |
| Adaptive lasso | AIC |
| Adaptive lasso | BIC |
| Adaptive lasso | LOO-CV |
| Ridge | AIC |
| Ridge | BIC |
| Ridge | LOO-CV |
| Adaptive ridge | AIC |
| Adaptive ridge | BIC |
| Adaptive ridge | LOO-CV. |

Throughout you will assume that $Y = \boldsymbol{X}^{\intercal}\boldsymbol{\beta}^{*} + \epsilon$, where $\epsilon \sim \text{Normal}(0, \sigma^2)$ and $\boldsymbol{X} \sim \text{Normal}_p\left\{0, \boldsymbol{\Sigma}(\rho)\right\}$, where $\left\{\boldsymbol{\Sigma}(\rho)\right\}_{i,j} = \rho^{|i-j|}$. We will be varying the sample size $n$, the dimension $p$, and the correlation parameter $\rho \in (0, 1)$.

(a) Recall that the theoretical $R^2$ is given by $1 - P(Y - \boldsymbol{X}^{\intercal}\boldsymbol{\beta}^{*})^2/Var(Y)$. Let $r \in (0, 1)$ be a constant. Given $\rho$ and $\boldsymbol{\beta}^{*}$, derive expression for $\sigma^2$ so that the $R^2 = r$, i.e., how should we set $\sigma^2$ so that the $R^2$ is a equal to a pre-specified value.[2]

(b) For each of the parameter settings given below choose $\sigma^2$ so that the $R^2$ is equal to 0.8. Then generate 1000 datasets of size $n$ from the generative model, fit the models listed above to each dataset and compute the MSE, average the MSE's across the 1000 datasets to obtain an average error. Record your answers in a table.

Sparse signal: $n = 100$, $p = 10, 25, \text{and} 50$, $\rho = 0, 0.25, 0.5$, $\beta_j^{*} = (2/\sqrt{n})1_{j \le \sqrt{p}}$

Dense signal: $n = 100$, $p = 10, 25, \text{and} 50$, $\rho = 0, 0.25, 0.5$, $\beta_j^{*} = \{5/(j\sqrt{n})\}$.

(c) Did you notice any general patterns? Does one method perform better with sparse or dense signals?

[2] This is useful to know as it allows you to easily control the signal-to-noise ratio in your simulation experiments.