

STAT5200_computing_project

STAT 5200 Computing Project

Repo organization

I initially ran the simulation with a different data generating process before I found the one in Hastie's paper. The plots from the initial simulation are in `sim1_plots/` while the plots in the presentaiton are in `sim2_plots/`

Code organization

The initial data generating process is implemented in the `generate_data` function, while Hastie's is `generate_data2`. The code has a functional-style, and the only direct interaction should be with the `complete_run` function at the bottom. To use the initial simulation simply set `SNR == 0` otherwise hastie's setup will be run. Furthermore, if `train` is equal to 1 then the simulation will include results for the training data. If `train` is not equal to 1 then we only get the test results. The code takes around 15 minutes to run.

Signal to Noise Ratio Clarification

The hastie simulation allows you to set a desired signal to noise ratio, SNR. To achieve this in the paper they give the following formuatl

$$SNR = \frac{\|\beta\|_2^2}{\sigma^2}$$

This implies that

$$\|\beta\|_2 = \sqrt{SNR}\sigma = r$$

Then we want to contstruct a vector that exactly has this norm. First we generate our original betas,

$$\beta_{\text{raw}} \sim N(0, 1)$$

Then, we set

$$\beta = \beta_{\text{raw}} \frac{r}{\|\beta_{\text{raw}}\|}$$

This fraction is a constant so if we take the L2 norm of both sides we see that

$$\|\beta\| = r$$

which is what we desire, implying our scaling is valid.