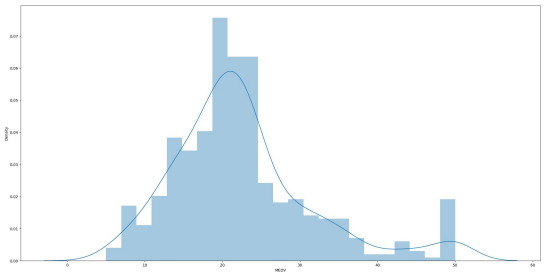
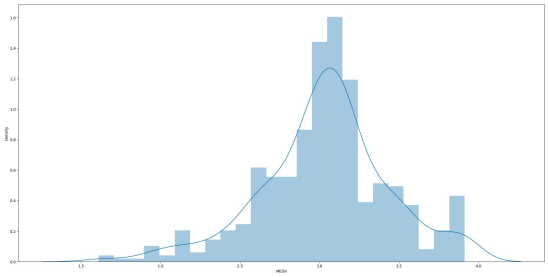
**Overview of Data:** What I did to process the Boston House Prices Data set:



distribution of response variable



distribution of log of response variable

1. Take the log of the response variable MEDV
    a. as seen in the figures above, the distribution of the response is slighly skewed, by taking the log of y we can account for this skewness.
2. Split the data into training and testing sets.
3. Standardize the data (train, test) according to the training data

## 1. Statistical Significance in Linear Regression

| feature | pval |
|---|---|
| CRIM | 1.3779e-11 |
| ZN | 0.2370 |
| INDUS | 0.2564 |
| CHAS | 0.0121 |
| NOX | 2.1684e-05 |
| RM | 2.5002e-09 |
| AGE | 0.8116 |
| DIS | 1.0227e-07 |
| RAD | 0.0003 |
| TAX | 0.00307 |
| PTRATIO | 3.6778e-10 |
| B | 1.4549e-06 |
| LSTAT | 2.6046e-31 |

| Top Features |
|---|
| CRIM |
| RM |
| PTRATIO |
| LSTAT |

## 2. Best Subsets

| Top Features |
|---|
| CRIM |
| RM |
| B |
| LSTAT |

## 3. RFE

| Features selected |
|---|
| CHAS |
| NOX |
| RM |
| PTRATIO |

## 4. Lasso

*larger lambda -> lower model complexity -> more features to 0*

| lambda = 0.1 | lambda = 0.5 | lambda = 1 |
|---|---|---|
| -0.037 LSTAT | -0.0261 LSTAT | -0.0123 LSTAT |
| -0.0131 PTRATIO | -0.0006 TAX | -0.0009 TAX |
| -0.0061 CRIM | -0.0 CRIM | -0.0 CRIM |
| -0.0003 TAX | -0.0 INDUS | -0.0 INDUS |
| -0.0 INDUS | 0.0 CHAS | 0.0 CHAS |
| 0.0 CHAS | -0.0 NOX | -0.0 NOX |
| -0.0 NOX | 0.0 RM | 0.0 RM |
| 0.0 RM | -0.0 AGE | -0.0 AGE |
| -0.0 DIS | -0.0 DIS | -0.0 DIS |
| 0.0002 ZN | 0.0 RAD | 0.0 RAD |
| 0.0005 B | -0.0 PTRATIO | -0.0 PTRATIO |
| 0.0009 RAD | 0.0001 ZN | 0.0002 ZN |
| 0.0011 AGE | 0.0006 B | 0.0007 B |

**Top Features**

**LSTAT**
**B**
**TAX**
**ZN**

## 5. Elastic Net

*lambda = 1.0* *alpha = .5*

| | |
|---|---|
| -0.02572 LSTAT | -0.0 AGE |
| -0.000623 TAX | -0.0 DIS |
| -0.0 CRIM | 0.0 RAD |
| -0.0 INDUS | -0.0 PTRATIO |
| 0.0 CHAS | 8.6e-05 ZN |
| -0.0 NOX | 0.000607 B |
| 0.0 RM | |

**Top Features**

**LSTAT**
**TAX**
**B**
**ZN**

## 6. Adaptive Lasso

| iteration: 0 | iteration: 1 | iteration: 2 | iteration: 3 |
|---|---|---|---|
| -0.014908 LSTAT | -0.002215 CRIM | -0.0 CRIM | -0.0 CRIM |
| -0.00848 CRIM | -0.001331 LSTAT | 0.0 ZN | 0.0 ZN |
| -0.002926 RAD | 0.0 ZN | -0.0 INDUS | -0.0 INDUS |
| 0.0 INDUS | -0.0 INDUS | 0.0 CHAS | 0.0 CHAS |
| 0.0 CHAS | 0.0 CHAS | 0.0 NOX | 0.0 NOX |
| 0.0 NOX | 0.0 NOX | -0.0 AGE | -0.0 AGE |
| 0.0 DIS | -0.0 AGE | 0.0 DIS | 0.0 DIS |
| 0.000201 TAX | 0.0 DIS | -0.0 RAD | -0.0 RAD |
| 0.001821 AGE | -0.0 RAD | -0.0 TAX | -0.0 TAX |
| 0.001862 B | -0.0 TAX | 0.0 PTRATIO | 0.0 PTRATIO |
| 0.002319 ZN | 0.0 PTRATIO | -0.0 LSTAT | -0.0 LSTAT |
| 0.029367 PTRATIO | 0.002028 B | 0.00202 B | 0.002022 B |
| 0.290835 RM | 0.368847 RM | 0.365879 RM | 0.365766 RM |

**Top Features**

**RM**
**CRIM**
**B**
**LSTAT**

## SUMMARY

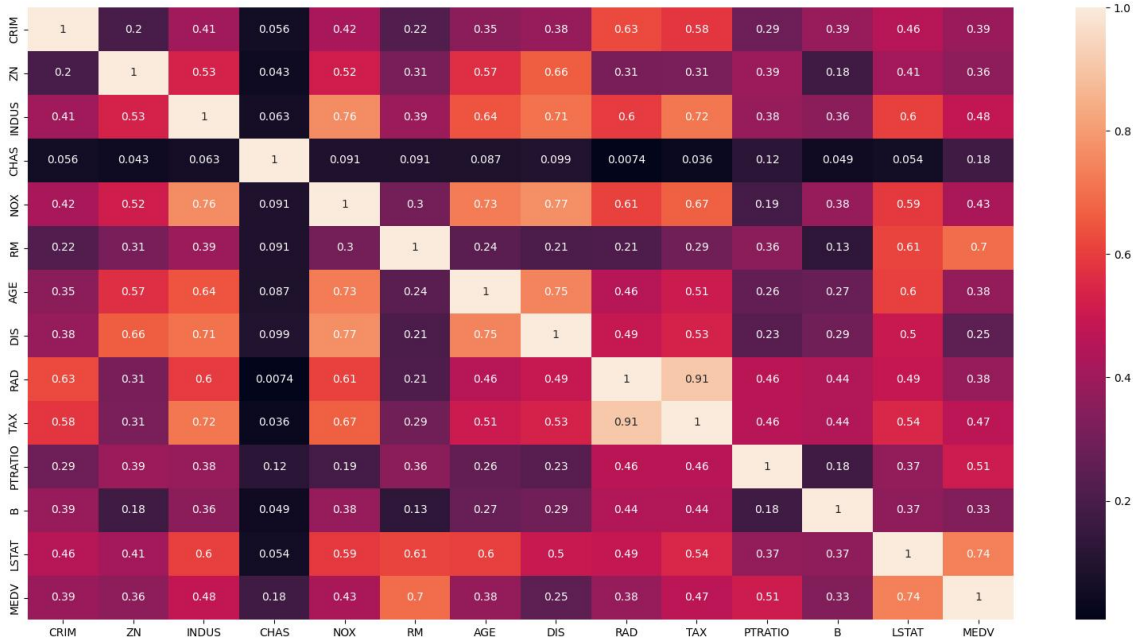
| Method | Top Features |
|---|---|
| Linear Regression | CRIM, RM, PTRATIO, LSTAT |
| Best Subsets | CRIM, RM, B, LSTAT |
| Recursive Feature Elimination | CHAS, NOX, RM, PRATIO |
| Lasso | LSTAT, B, TAX, ZN |
| Elastic Net | LSTAT, TAX, B, ZN |
| Adaptive Lasso | RM, CRIM, B, LSTAT |

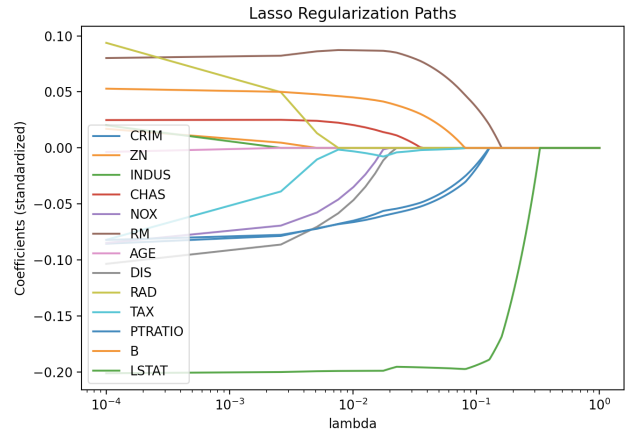
## DISCUSSION

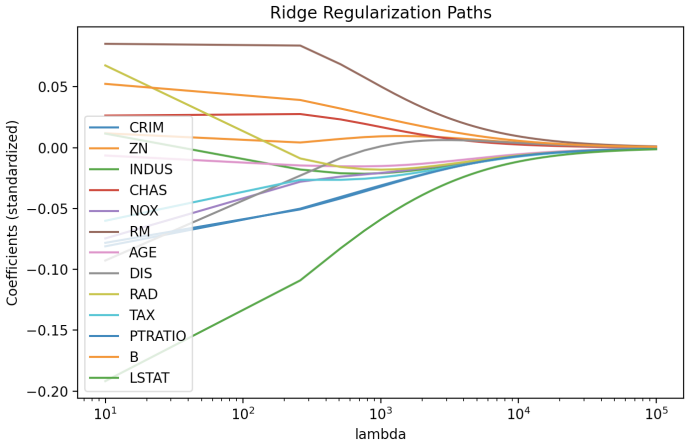
Across all methods, we see that the features LSTAT, RM, CRIM, and B were most often selected as important methods. If we look back at our corelation plot on page 1, we see that LSTAT and MEDV were highly correlated as well as RM and MEDV; this gives some intuition as to why these features were chosen (although it's important to note that correlation is univariate).

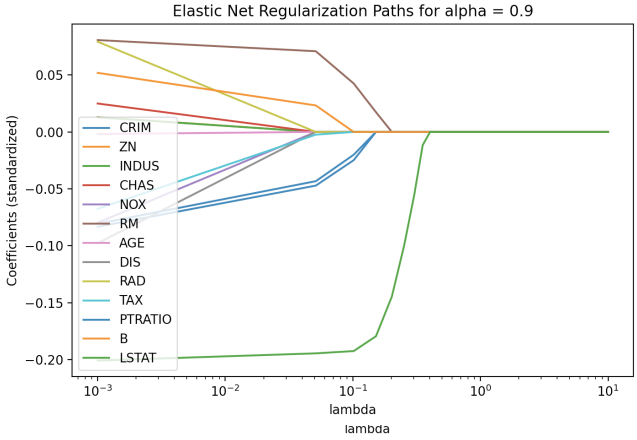
Additionally, Elastic Net and Lasso chose the same features which makes sense as Elastic Net is a combination of l1 and l2 norm regularization. We also know that LASSO tends to choose one out of highly correlated features. This could explain now Linear regression chose both RM and LSTAT (which are more correlated) wheras Lasso only chose LSTAT.

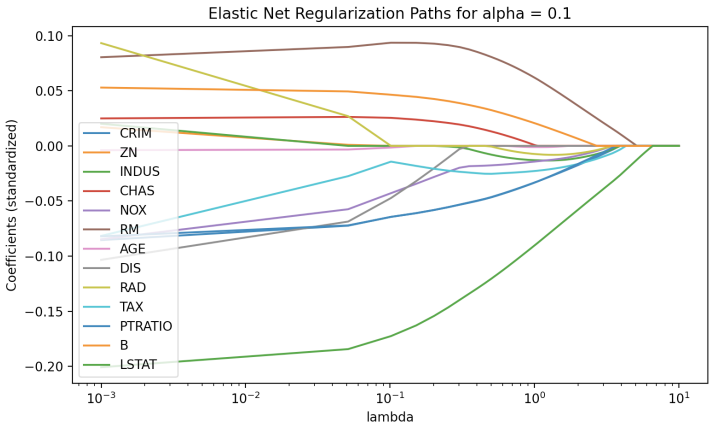
## FEATURE CORRELATION PLOT






## (iii) Regularization Analysis

Across all methods, we see that LSTAT, RM, CRIM, PTRATIO, ZN, and B were the most **consistently selected**. This is because they were their **coefficients were the last to be sent to 0** compared to other features.
Moreover, LSTAT and RM were the top features across all models as they had **larger coefficients**/ weights in the model compared to other features. This supports LSTAT and RM being consistently selected as they were the last to go to 0 across most models. This also makes sense when we think about our correlation plot from earlier since CRIM and LSTAT had high correlation against the response variable MEDV (this is just a useful insight as the pairwise correlation is under the assumption that all the other features are in the model/ would be a univariate analysis to choose features this way). Another interesting observation is that in our correlation plot, RAD and TAX have a correlation of 0.91 (highly correlated), and in both plots, these features miror eachother accross the x-axis. In Lasso, once RAD goes to 0, TAX steadies our more which supports the idea of sparsity in Lasso and how Lasso will end up picking one of a group of correlated features.

Comparing the Lasso and Ridge regularization plots, we see that the Ridge coefficients more consistently slope to 0 wheras in Lasso the features drop our more separately. Ex: after 10^-1, LSTAT is clearly the dominating feature and all others seem to be sent to 0 wheras in Ridge, although LSTAT has a large coefficient, the other features are still present in the model.
This again displays the idea of **sparsity in Lasso** that comes from the l1 norm versus the l2 norm regulation in Ridge.

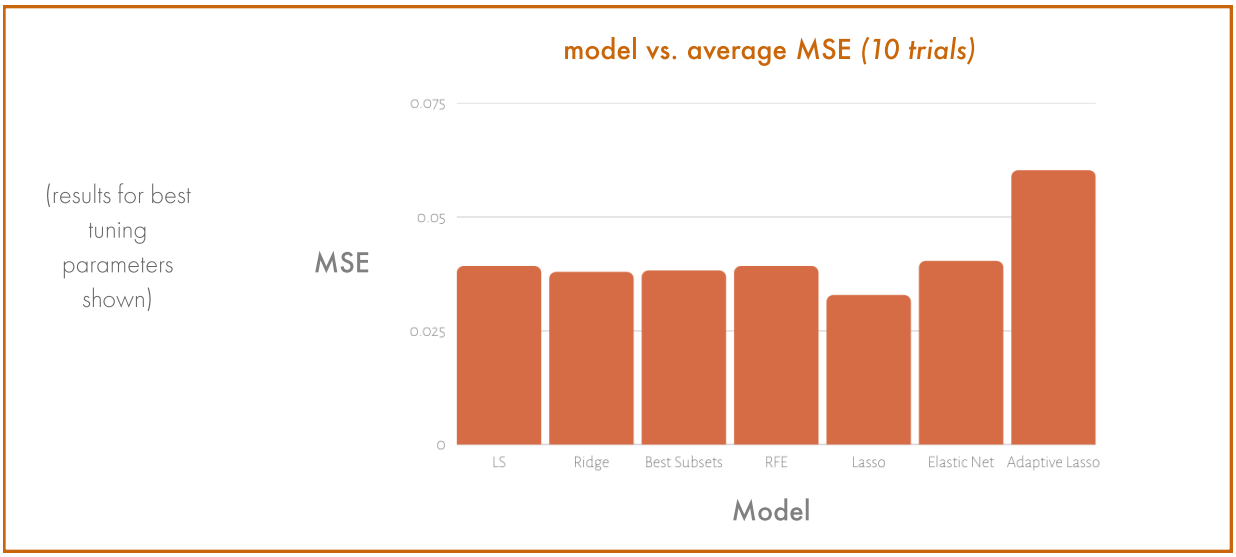
Additionally, for Elastic Net, when alpha is closer to 1, it's more like Lasso (alpha favors the l1 norm), but when alpha is closer to 0, it's more like Ridge (favors the l2 norm more). This matches our plots since when alpha = 0.1 the plot is more similar to Ridge and when alpha = 0.8 the plot is more similar to Lasso. We see this as features drop off less uniformly/ more quickly when alpha favors the l1 norm versus the l2 norm.

**Data Pre-processing:** removed entries with missing data, split data, took the log of standardized the data by centering / normalizing the features according to the training data

**B. i**

| Method | average MSE | tuned parameters (chosen each of 10 trials) |
|---|---|---|
| Least Squares | 0.0381 | none |
| Ridge | 0.0356 | ***lambda*** = 16.1618, 0.001, 3.435, 0.001,13.131, 20.0, 0.001, 0.001, 0.001, 3.233 |
| Best Subsets | 0.0382 | **Selected Features** (for some iterations): ['CRIM' 'ZN' 'CHAS' 'NOX' 'DIS' 'RAD' 'TAX' 'PTRATIO' 'LSTAT'] , ['CRIM' 'ZN' 'CHAS' 'RM' 'DIS' 'RAD' 'TAX' 'PTRATIO' 'B' 'LSTAT'] , ['CRIM' 'CHAS' 'NOX' 'RM' 'AGE' 'DIS' 'PTRATIO' 'B' 'LSTAT'] |
| RFE | 0.0392 | ***number of features to recursively eliminate*** = 9, 6, 9, 12, 12, 10, 11, 12, 8, 11 |
| Lasso | 0.0396 | all trials chose ***lambda*** = 0.001 |
| Elastic Net | 0.0392 | ***lambda, alpha*** = (0.001, 0.999), (0.001, 0.001), (0.001, 0.001), (0.001, 0.3840), (0.001, 0.0816), (0.001, 0.999), (0.001, 0.999) ,(0.708, 0.001), (0.102, 0.1623), (0.001, 0.001) |
| Adaptive Lasso | 0.0602 | ***lambda*** = 0.0211, 0.0211, 0.0110, 0.02118, 0.001, 0.0110, 0.02118, 0.0110, 0.0312, 0.02118 |

**B.ii**



# Data Analysis B.iii Reflection

**1. Which types of methods give the best prediction error? Why do these methods perform well?**

Lease squares, Ridge, Lasso, and Best Subsets give the best prediction errors. Lasso had the lowest average MSE, but it often chose a very small lambda value making it more similar to least squares.

As discussed in class, we also know that regularization is very helpful where p>>n (we have more features than observations). However, in our data, n>p, which could be part of the reason why methods like Lasso actually did worse than our least squares estimate and why lasso and ridge tended to choose smaller lambda (since those models closer match least squares).

Also, we see that the Elastic net MSE is between the Lasso and the Ridge Mse's, this makes sense as it's a combination of both of these models. As seen through the MSE's of Ridge and Lasso, Ridge performed better than Lasso, indicating that the l2 norm regularization could be a better fit for our data. However, since in our data we have n>>p, its possible that using Least squares is sufficient enough. (By the MSE Existence theorem,

**2. Do any methods seem to overfit to the training set? If so, why?**

It's possible that adaptive lasso overfit our training data. It has the highest MSE our of all the models and eliminated the most features. It's possible that this model over emphasized / eliminated features when running evaluations on our validation set; after the many iterations on our validation set it's possible that the model didn't generalize well to our testing data.

**3. Do all the methods that give similar predictions choose the same subset of variables?**

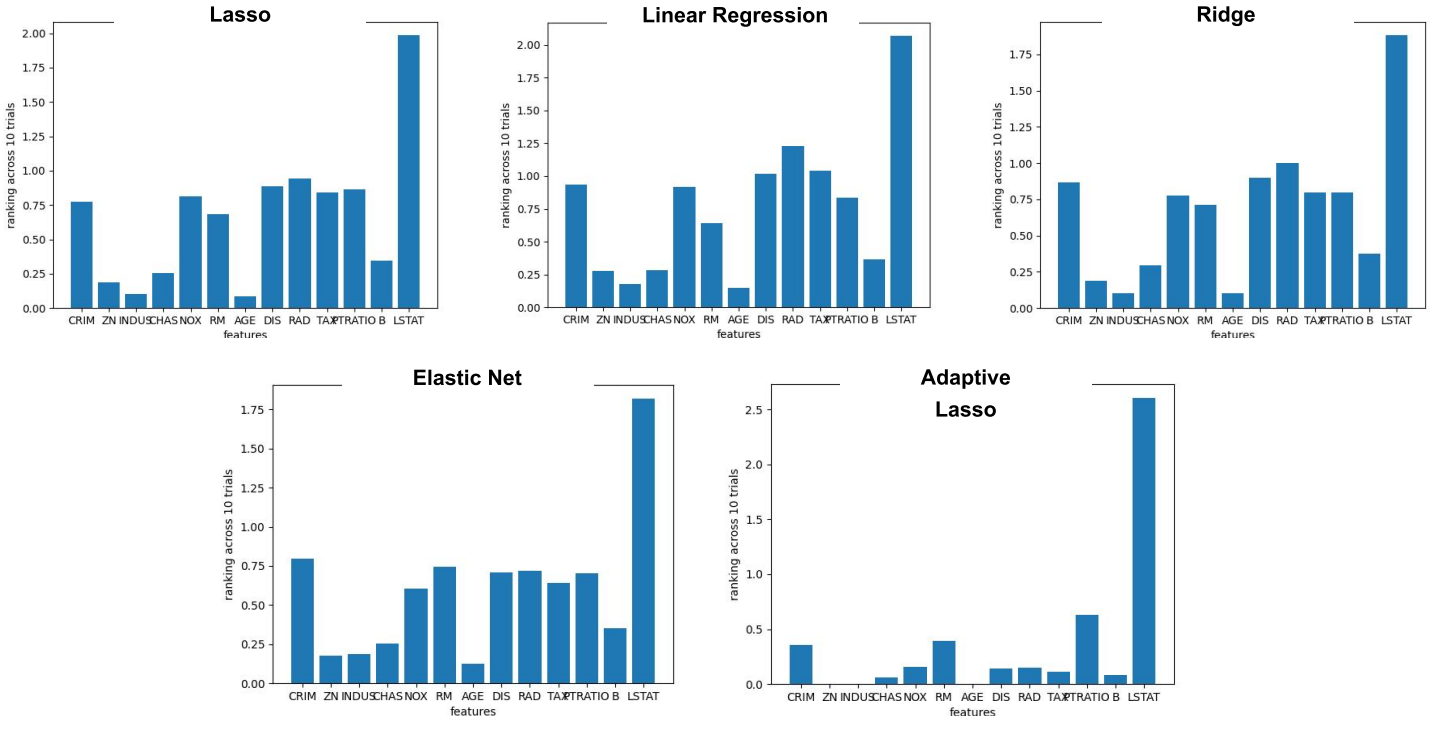
To get some insight on the variables/features chosen, I summed the absolute value of each features coefficients for the chosen model (if parameters needed tuning these are only for the best models) accross the 10 trials, and them plotted these sums. I just did it for 5 of the methods, to get some insight on how these models chose features relative to eachother. Based on our results, we got similar predictions from Least squares, lasso, and ridge, all of which chose features similarly. Additionally, Adaptive lasso is the sparsest which makes sense as it performs an iterative, more sparse lasso.

**4. Which is the overall best method for prediction on this dataset?**

Based on our data, Ridge is the best method for prediction on this data set with the lowest average MSE of 0.0356.

## Feature Importance across 10 Trials

***(sum of the absolute values of feature coefficients)***



1. Theory: Ridge Regression. Prove the MSE Existence Theorem for the ridge regression estimator. *Recall that the MSE Existence Theorem states that there exists a value of $\lambda$ for which $MSE(\mathbf{X}\hat{\beta}^{Ridge(\lambda)}) < MSE(\mathbf{X}\hat{\beta}^{LS})$.*

## A. Calculating $tr\left(Var\left(X\hat{\beta}^{r}\right)\right)$

Let the covariance matrix have the eigen decomposition $X^TX = P^TDP$ • P is orthogonal $\Rightarrow P^{-1} = P^T$

$$(X^TX + \lambda I)^{-1} = (P^TDP + \lambda P^TP)^{-1} = (P^T(D+\lambda I)P)^{-1} = P^{-1}(D+\lambda I)^{-1}(P^T)^{-1} = P^T(D+\lambda I)^{-1}P$$

• $tr\left(Var\left(X\hat{\beta}^{r}\right)\right) = tr\left(X\,Var\left(\hat{\beta}^{r}\right)X^T\right)$

$= tr\left(\sigma^2 \underset{n\times n}{(X^TX+\lambda I)^{-1}} X^TX \underset{n\times n}{(X^TX+\lambda I)^{-1}} X^T\right)$

$= tr\left(\sigma^2 P^T(D+\lambda I)^{-1}PP^TDP\,P^T(D+\lambda I)^{-1}PP^T\right)$

$= tr\left(\sigma^2 P^T(D+\lambda I)^{-1}D(D+\lambda I)^{-1}DP\right)$

properties of trace $= tr\left(\sigma^2 (D+\lambda I)^{-1}D(D+\lambda I)^{-1}DPP^T\right)$

$= tr\left(\sigma^2 (D+\lambda I)^{-1}D(D+\lambda I)^{-1}D\right)$

$= tr\left(\sigma^2 (D+\lambda I)^{-2}D^2\right)$

$= \sum_{i=1}^{n} \frac{d_{ii}\sigma^2}{(d_{ii}+\lambda)^2} = \boxed{\sum_{i=1}^{n} \frac{d_i^2\sigma^2}{(d_i+\lambda)^2}}$

$$D = \begin{bmatrix} d_{11} & & \\ & d_{22} & \\ & & \ddots \\ & & & d_{nn} \end{bmatrix}$$

$$(D+\lambda I)^{-1}D(D+\lambda I)^{-1} = \begin{bmatrix} \frac{d_{11}}{(d_{11}+\lambda)^2} & & \\ & \ddots & \\ & & \frac{d_{nn}}{(d_{nn}+\lambda)^2} \end{bmatrix}$$

$$\frac{d\,tr\left(Var\left(X\hat{\beta}^{r}\right)\right)}{d\lambda} = \sum_{i=1}^{n} \frac{-2d_i^2\sigma^2}{(d_i+\lambda)^3} = -2\sigma^2 \sum_{i=1}^{n} \frac{d_i^2}{(d_i+\lambda)^3}$$

## B. Calculating $Bias\left(X\hat{\beta}^{r}\right)^2$

• $E\left(X\hat{\beta}^{ridge}\right) = X(X^TX+\lambda I)^{-1}X^TX\beta$

$= X(X^TX+\lambda I)^{-1}(X^TX+\lambda I-\lambda I)\beta$

$= X\left[(X^TX+\lambda I)^{-1}(X^TX+\lambda I) - (X^TX+\lambda I)^{-1}\lambda I\right]\beta$

$= X\left[I - \lambda I(X^TX+\lambda I)^{-1}\right]\beta$

$= X\beta - X\lambda I(X^TX+\lambda I)^{-1}\beta$

• $Bias\left(X\hat{\beta}^{ridge}\right)^2 = \left(E\left(X\hat{\beta}^{ridge}\right) - X\beta\right)^T\left(E\left(X\hat{\beta}^{ridge}\right) - X\beta\right)$

$= \left(-\lambda X(X^TX+\lambda I)^{-1}\beta\right)^T\left(-\lambda X(X^TX+\lambda I)^{-1}\beta\right)$

$= \lambda^2\beta^T(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}\beta$

$= \lambda^2\beta^T(P^TDP+\lambda I)^{-1}P^TDP(P^TDP+\lambda I)^{-1}\beta$

$= \lambda^2\beta^T(P^TDP+\lambda P^TP)^{-1}P^TDP(P^TDP+\lambda P^TP)^{-1}\beta$

$= \lambda^2\beta^TP^T(D+\lambda I)^{-1}PP^TDPP^T(D+\lambda I)^{-1}P\beta$

$= \lambda^2\beta^TP^T(D+\lambda I)^{-1}D(D+\lambda I)^{-1}P\beta$

$= \boxed{\sum_{i=1}^{p} \beta_i^2 p_{i2}^2 \frac{d_i\lambda^2}{(d_i+\lambda)^2}}$

Let $X^TX = P^TDP$ the eigen decomposition of $X^TX$ where $d_1, d_2, \ldots, d_p$ are the eigen values of $X^TX$.

$$\frac{d}{d\lambda} Bias\left(X\hat{\beta}^{r}\right)\Big|_{\lambda=0} = \sum_{i=1}^{p} \beta_i^2 p_{i2}^2 \left.\frac{(d_i+\lambda)^2 2\lambda d_i - d_i\lambda^2 2(d_i+\lambda)}{(d_i+\lambda)^2}\right|_{\lambda=0} = 0$$

## C. Proving that $\exists \lambda$ s.t. $MSE(X\hat{\beta}^r) < MSE(X\hat{\beta}^{ls})$

- Combining A, B, we get that:

$$MSE(X\hat{\beta}^r) = \underbrace{\sum_{i=1}^{n} \frac{d_i \sigma^2}{(d_i+\lambda)^2}}_{\text{tr}(\text{Var}(X\hat{\beta}^r))} + \underbrace{\sum_{i=1}^{p} \beta_i^2 p_{ii}^2 \frac{d_i \lambda^2}{(d_i+\lambda)^2}}_{\text{Bias}(X\hat{\beta}^r)^2}$$

- Taking the derivative at $\lambda = 0$, we see that,

$$\frac{d\, MSE(X\hat{\beta}^r)}{d\lambda}\Big|_{\lambda=0} = \left( \sum_{i=1}^{n} \frac{-2 d_i^2 \sigma^2}{(d_i+\lambda)^3} + \sum_{i=1}^{p} \beta_i^2 p_{ii}^2 \frac{(d_i+\lambda)^2 2\lambda d_i - d_i \lambda^2 2(d_i+\lambda)}{(d_i+\lambda)^2} \right)\Big|_{\lambda=0}$$

$$= -2 \underbrace{\sigma^2 \sum_{i=1}^{n} \frac{1}{d_i}}_{\text{always positive since } X^TX \text{ is positive semi def} \Rightarrow d_i \geq 0. \text{ If a } d_i = 0, \text{ then } \lim_{\lambda \to 0} \frac{d MSE(X\hat{\beta}^r)}{d\lambda} = -\infty} + 0 < 0$$

in either case, the derivative is negative.

(i) Thus, we see that the $\frac{d\, MSE(X\hat{\beta}^r)}{d\lambda}\Big|_{\lambda=0} < 0 \Rightarrow MSE(X\hat{\beta}^r)$ is decreasing at $\lambda = 0$.

(ii) Additionally, we know that when $\lambda = 0$, $MSE(X\hat{\beta}^r) = MSE(X\hat{\beta}^{ls})$ because ls is r without regularization.

(i) & (ii) $\Rightarrow$ as $\lambda$ increases in some neighborhood $> 0$, $MSE(X\hat{\beta}^r)$ becomes smaller than it was for $\lambda = 0$ because $\frac{d\, MSE(X\hat{\beta}^r)}{d\lambda} < 0$
i.e., it becomes smaller than $MSE(X\hat{\beta}^{ls})$ since $MSE(X\hat{\beta}^r) = MSE(X\hat{\beta}^{ls})$ at $\lambda = 0$ which is exactly what we want!

Therefore, there exists some $\lambda > 0$ such that $MSE(X\hat{\beta}^r) < MSE(X\hat{\beta}^{ls})$ ■

---

4. Properties: Ridge Regression. Derive an expression for the ridge estimator and ridge prediction, $\hat{Y}$, using the singular value decomposition (SVD) of $\mathbf{X}$. Use this expression to explain how ridge regression behaves based on patterns of variation in $\mathbf{X}$ and based on highly correlated groups of features.

(a) The SVD of $X$ is $X = UDV^T$ where $U_{n \times p}$, $V_{p \times p}$ are orthogonal, $D_{p \times p}$ is $\text{diag}(d_1, d_2 \cdots, d_p)$ $d_1 \geq d_2 \cdots \geq d_p$.

- The column vectors $u_i$ of $U$ span the column space of $X$
- The column vectors $v_i$ of $V$ span the row space of $X$.

(b) The RR solutions using this SVD of $X$ are:

$$\begin{aligned} X\hat{\beta}^r &= X(X^TX + \lambda I_p)^{-1} X^T y \\ &= UDV^T (VD^TU^TUDV^T + \lambda I_p)^{-1} VD^TU^T y \qquad I_p = VV^T,\ V^{-1} = V^T \\ &= UDV^T (VDDV^T + \lambda VV^T)^{-1} VD^TU^T y \qquad \text{distribute the inverse and substitute } V^{-1} = V^T,\ (V^T)^{-1} = V \\ &= UDV^T V (D^2 + \lambda I)^{-1} V^T V D^T U^T y \\ &= UD(D^2 + \lambda I)^{-1} D^T U^T y \\ &= \sum_{i=1}^{p} u_i \frac{d_i^2}{d_i^2 + \lambda} u_i^T y \end{aligned}$$

Meaning: Ridge transforms $y$ into the basis $U$ and shrinks the $i^{th}$ coordinate WRT $\frac{d_i^2}{d_i^2 + \lambda}$.

(c) The covariance matrix $X^TX$ expressed using this SVD of $X$ is:

$$X^TX = VDU^TUDV^T = VD^2V^T$$ (aka $X^TX$'s eigen decomposition with eigen vals $d_1^2, d_2^2, \ldots, d_p^2$ and corresponding vectors $v_1, \ldots, v_p$)

From this covariance matrix decomposition, we can analyze the sample variance of each $Xv_i$.
Why is this useful? As discussed in the textbook, each eigen vector of $X^TX$ is a principal component direction of $X$ (these are vectors that our data could be projected onto).

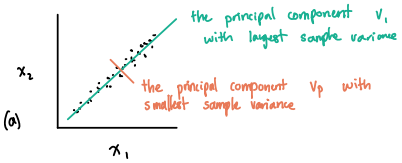
More specifically, since $d_1 \geq d_2 \geq \dots \geq d_p$, and $\text{Var}(Xv_i) = \frac{d_i^2}{n}$ ← sample variance

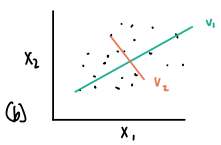
We see that $\text{Var}(Xv_1) \geq \text{Var}(Xv_2) \geq \dots \geq \text{Var}(Xv_p)$.

Since each $v_i$ represents a principal component of $X$, $v_i$'s with small sample variance correspond to projections of the data that have a smaller variance, i.e, Ridge will shrink coefficients of components with smaller $(\text{Var}\, X v_i)$ more than those with high variance.

Correlated Features

We can get some better intuition by visualizing this for $X$ with dimension $p = 2$.




In the first figure, $x_1$ and $x_2$ are highly correlated versus $x_1$ and $x_2$ in the second figure. In Ridge, for figure (a), the coefficients of $v_2$ will be less than $v_1$ (it will largely favor $v_1$ since it has high variance) whereas in (b), $v_1$ and $v_2$ will have more similar coefficients. This explains how Ridge deals with correlated features by choosing principal component directions which account for this correlation (vs. losing a lot of information).