Machine Learning I

Lecture 4: Parameter Shrinking Methods

Nathaniel Bade

Northeastern University Department of Mathematics

Table of contents

Variance Minimizing Methods

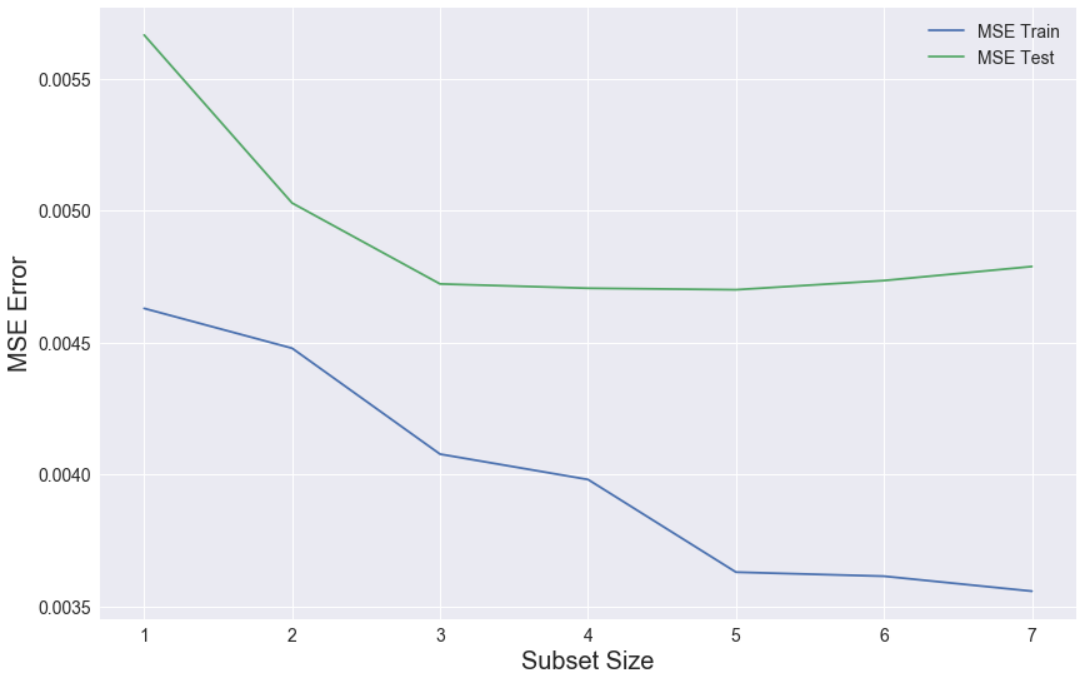
Outline

In Lecture 3, we looked at using feature selection to tune the variables. We noticed a few kinds of pathologies:

- Parameters strongly correlated with the output feature having statistically insignificant fit parameters $\hat{\beta}_i$.
- Parameters weakly correlated with the output feature having statistically significant fit parameters $\hat{\beta}_i$.

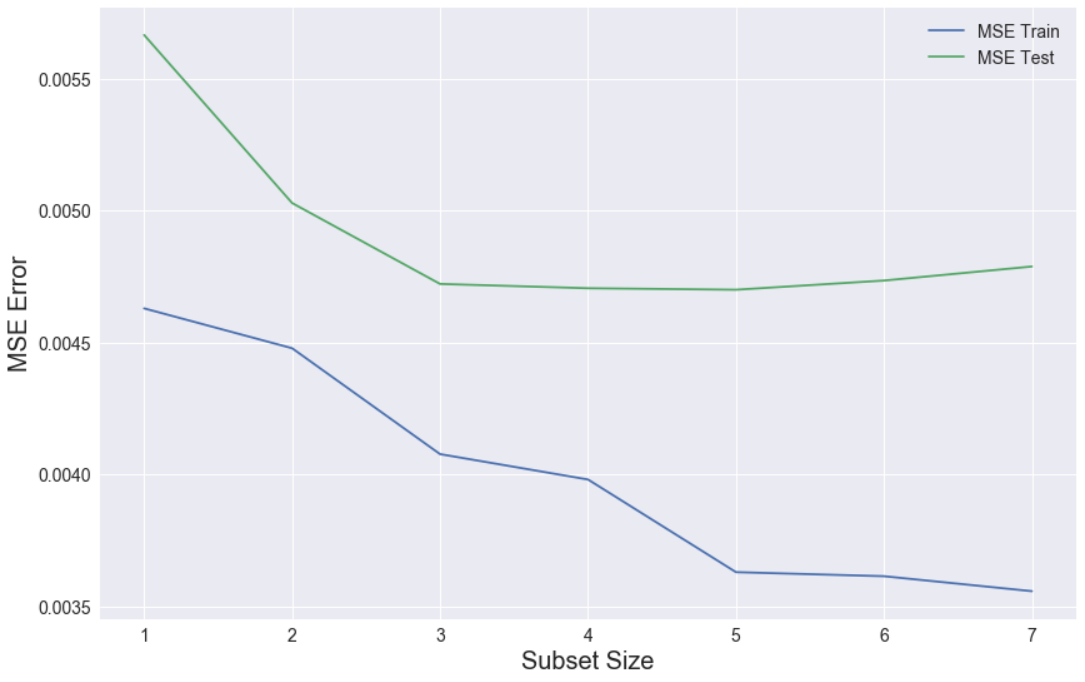
Both of these are indicators of overfitting, that the linear predictor is using insignificant or highly correlated features to boost the performance on the training set at the cost of higher variance.

Outline



We also discussed one family of fixes: **subset selection methods**. We saw that by either canvasing all possible subsets of variables (**best subset selection**) or adding or removing the best variable from the model one step at a time (**forward/backward selection**) we could improve performance on testing data.

Outline



In general, subset selection models are modifying the linear regression

$$y = X^T \beta$$

by enforcing the fact $\beta_j = 0$, for j in some subset of $\{1, \ldots, p\}$. We will see shortly that in the class of all linear models, this is known as increasing the **bias**, with the hope of lowing the **variance**.

Outline

How else can we improve our fit, given that we're restricting ourselves to the hypothesis class of linear models?

- Modify the loss function on the training set. Changing the loss functions will always make the RSS loss with respect to the training data worse, but could make the RSS loss with respect to new test data better.
- Construct linear composite features from the dataset and perform feature selection on these new features.
- Construct nonlinear features from the dataset.

We will spend this lecture discussing the first method.

Outline

We will start by discussing the **Gauss-Markov Theorem**. The theorem tells us that *there is no unbiased linear estimator with smaller variance than the minimum of the RSS loss function*. On some level this result is unsurprising, the bias variance tradeoff for RSS splits error into a irreducible part and a variance

$$\mathbf{Err}(x_0) = \sigma^2 + \cancelto{0}{\mathbf{Bias}^2(\hat{f}(x_0))} + \mathbf{Var}(\hat{f}(x_0)).$$

So any unbiased estimator with minimum error must raise the variance. Note this is also telling us the RSS is heavily tied variance minimization: For unbiased estimators, if you want to minimize variance across a hypothesis class, you need to minimize RSS.

Gauss-Markov Theorem

Bias and Variance

Let us recall some facts about **point estimators** on out way to defining the Bias and Variance.

Assume that w_i are drawn from a distribution $\mathcal{D}_\theta$ depending on some parameters θ. Another way to say this is that the probability of drawing w_i given θ is $w_i \sim P(w|\theta) = \mathcal{D}_\theta$. One job of statistics is to estimate the parameters θ given the data w_i.

For example, if w_i are drawn from a distribution with mean μ,

$$\hat{\mu}_N(w_1, \ldots, w_N) = \frac{1}{N} \sum_{i=1}^{N} w_i \,,$$

is a point estimator of μ for each N. However, especially for parameters with complicated dependence there many be many estimators. Even for μ, $\hat{\mu} = 4$ is *an* estimator, it's just a very bad one. So is $\hat{\mu} = w$, where w is a randomly sampled element of uniform distribution on w_i.

Bias and Variance

A sequence of point estimators $\hat{\theta}_N(w_1, \ldots, w_N)$ for $w \sim D_\theta$ is called **consistent** if for all true parameter values θ,

$$\lim_{N \to \infty} P_X[\, |\hat{\theta}_N - \theta| < \epsilon\,] = 1\,,$$

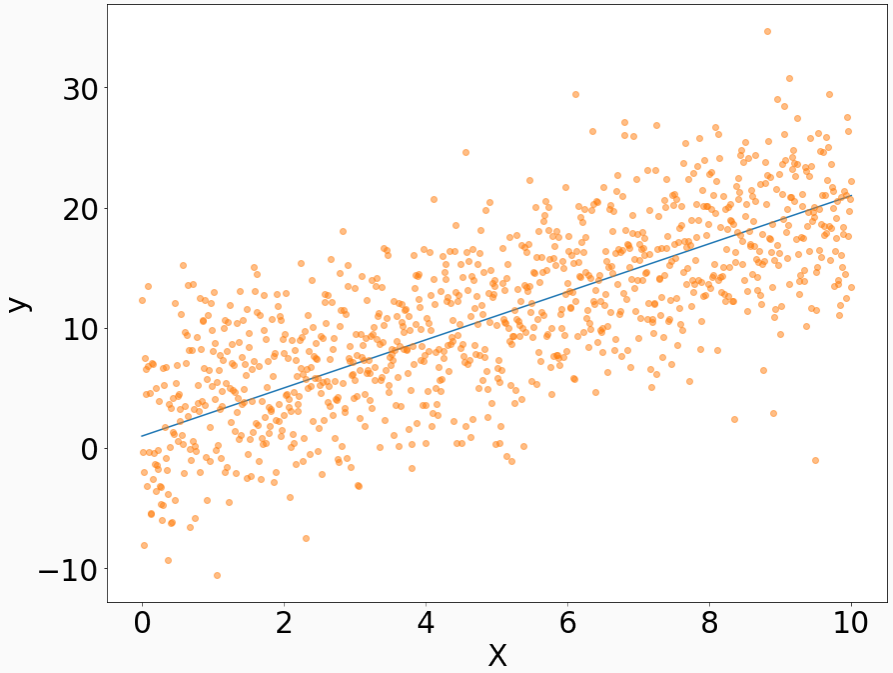
That is if the expected value of $\hat{\theta}_N(w_1, \ldots, w_N)$ converges to the actual value of θ as the number of i.i.d. sampled data points N goes to infinity.

For example,

$$\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} w_i\,,$$

is a consistent estimator of the mean of the distribution.On the other hand, $\hat{\mu} = 4$ is not, and neither is $\hat{\mu} = w$, for w is a randomly sampled element of uniform distribution on w_i.

Bias and Variance



As an example, if (y_i) are drawn from $y = X^T\beta + \epsilon$, minimizing the RSS function has given us *an* estimator of the parameter β:

$$\hat{\beta} = \beta \approx \hat{\beta}_N(\mathbf{X}, \mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\,.$$

However, there are other ways we could come to an estimate of β. Consistency gives us one evaluation of the point estimator, bias and variance give us others.

Bias and Variance

Definition: Given $P(w|\theta)$ and an estimator $\hat{\theta}(w)$ of θ, the **bias** of $\hat{\theta}$ relative to θ is

$$\text{Bias}_\theta[\hat{\theta}] = E_w[\hat{\theta}(w)] - \theta\,.$$

In words, it is the difference between the expectation value over i.i.d. samples of $P(w|\theta)$ of the estimator $\hat{\theta}$ and the true value θ. Not, that the bias is not taken in the large N limit, it is the expectation over a specific number of draws.

An estimator is said to be **unbiased** if the bias is 0 for all values of θ. For example,

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} w_i\,,$$

is an unbiased estimator for all N. In particular,

$$\hat{\mu} = w_1$$

is an unbiased estimator if w_i are i.i.d. even though it is not constant!

Bias and Variance

Definition: Given $P(w|\theta)$ and an estimator $\hat{\theta}$ of θ, the **variance** of en estimator is the expected value of the squared sampling deviations:

$$\mathbf{Var}(\hat{\theta}) = E_w[\,(\hat{\theta}(w) - E_w[\hat{\theta}(w)]\,)^2]\,.$$

We have already seen one computation of the variance in Lecture 3, where we showed that the variance of the parameters β for a model

$$y = X^T\beta + \epsilon\,,$$

with ϵ drawn from a distribution with variance σ^2 could be computed to by the diagonal elements of

$$\mathbf{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\,.$$

Probabilistic Models

Lets now move to modeling. Given a set of data $(\mathbf{X}, \mathbf{y})$, a probabilistic model is probability distribution for the values of y at each point X, usually written as the conditional probability $P_\theta(y|X)$. Such a distribution usually depends on parameters θ, and all distributions together give a hypothesis class of models.

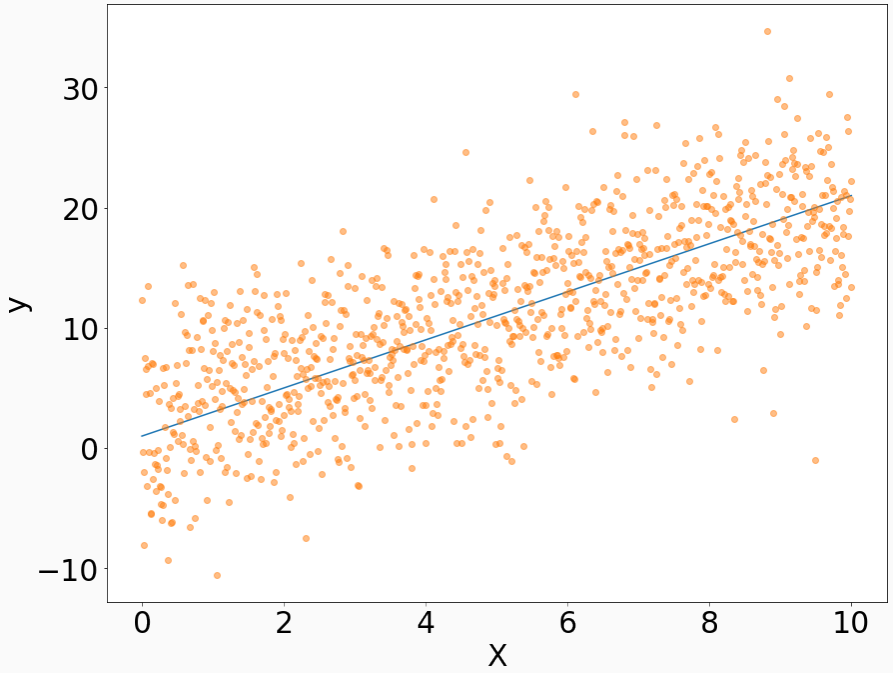
For example, a linear model with Gaussian noise $\epsilon \in \mathcal{N}(\mu, \sigma)$ is a hypothesis class of models parameterized by $\theta = (\beta, \mu, \sigma)$:

$$P_\theta(y|X) = X^T\beta + \epsilon\,.$$

Here, the probability density of y given that X have the value X is $X^T\beta + \epsilon$.

The job of statistical inference is to estimate the parameters θ given $(\mathbf{X}, \mathbf{y})$. In this setting, a **point estimator** $\hat{\theta} = \hat{\theta}(\mathbf{X}, \mathbf{y})$ is a function of the data that returns an estimate for the data.

Bias and Variance



Again, minimizing the RSS function has given us *an* estimator of the linear coefficients, in $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. It is natural to ask if we can say anything about the bias and variance of this estimate.

In particular, it would be nice to find a class of unbiased estimators and see if we can minimize the variance among the model of that class.

Gauss-Markov Theorem

Theorem (Gauss-Markov Theorem)
For data $\mathbf{X}, \mathbf{y}$ *generated by a linear model*

$$y = X^T \beta_* + \epsilon \,,$$

with $E[\epsilon] = 0$ *and* σ_ϵ^2 *finite, the least squares estimate of parameters* $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ *has the smallest variance among all linear, unbiased estimators* $\widetilde{\beta}$.

Here, a **linear estimator** of β_* is a function linear in the target variable:

$$\widetilde{\beta} = \mathbf{B}\mathbf{y} \,, \qquad \widetilde{\beta}_j = \sum_i B_{ji} y_i \,.$$

An **unbiased** linear estimator is $\widetilde{\beta} = \mathbf{B}\mathbf{y}$ such that (for fixed $\mathbf{X}$),

$$E_\epsilon[\widetilde{\beta}] = \beta_* \,.$$

Proof setup

We want to write an equation comparing the variance of $\hat{\beta}$ with the variance of an arbitrary linear estimator $\widetilde{\beta}$. Since all of the label variance comes from $\mathbf{y}$ (and variance and expectation play nicely with addition and multiplication) we write

$$\widetilde{\beta} = \hat{\beta} + \widetilde{B}\mathbf{y} = \hat{B}\mathbf{y} + \widetilde{B}\mathbf{y}\,,$$

where $\hat{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Then $\widetilde{\beta} = (\hat{B} + \widetilde{B})\mathbf{y}$ parameterizes the space of arbitrary linear predictors.

In this notation, the variance of $\hat{\beta}$ from Lecture 3 can be written

$$\text{Var}(\hat{\beta}) = \sigma_\epsilon^2 \hat{B}\hat{B}^T = (\mathbf{X}^T\mathbf{X})^{-1}\,.$$

Proof setup

As an aside, the intuition here is that we would like to analyze

$$\text{Var}(\hat{\beta}) - \text{Var}(\tilde{\beta}) \stackrel{?}{=} \text{Var}(\hat{\beta} - \tilde{\beta}) = \text{Var}((B_1 - B_2)\mathbf{y}) \,,$$

since we can then peal all of the ϵ-variance off and just deal with the terms leftover. But of course, the first equality doesn't hold. Writing

$$\widetilde{\beta} = \hat{\beta} + \widetilde{B}\mathbf{y} = \hat{B}\mathbf{y} + \widetilde{B}\mathbf{y} \,,$$

is a workable substitute.

Proof of the Gauss-Markov Theorem

Proof: Let

$$\widetilde{\beta} = \hat{B}\mathbf{y} + \widetilde{B}\mathbf{y}\,,$$

be an unbiased estimator, where $\hat{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Expanding expectation value,

$$\begin{aligned}
E_\epsilon[\widetilde{\beta}] &= E_\epsilon[\hat{B}\mathbf{y} + \widetilde{B}\mathbf{y}] \\
&= \beta_* + E_\epsilon[\widetilde{B}\mathbf{y}] && \hat{\beta} \text{ is an unbiased estimator,} \\
&= \beta_* + E_\epsilon[\widetilde{B}(\mathbf{X}^T\hat{\beta} + \epsilon)] && \text{Definition of } \mathbf{y}, \\
&= \beta_* + E_\epsilon[\widetilde{B}\mathbf{X}^T\hat{\beta} + \widetilde{B}\epsilon] \\
&= \beta_* + \widetilde{B}\mathbf{X}^T\hat{\beta} && E[\epsilon] = 0.
\end{aligned}$$

Since $\widetilde{\beta}$ is unbiased we must have $E_\epsilon[\widetilde{\beta}] = \beta_*$ for any possible β_*. This implies that $\widetilde{B}\mathbf{X}^T = 0$.

Proof of the Gauss-Markov Theorem

Proof (cont.): We can make a direct computation of the variance:

$$\begin{aligned}
\mathrm{Var}_\epsilon[\widetilde{\beta}] &= \mathrm{Var}_\epsilon[(\hat{B}+\widetilde{B})\mathbf{y}]\,, \\
&= \sigma_\epsilon^2(\hat{B}+\widetilde{B})(\hat{B}+\widetilde{B})^T && \mathrm{Var}(A\epsilon)\,, = \sigma_\epsilon^2 AA^T\,, \\
&= \sigma_\epsilon^2(\hat{B}\hat{B}^T + \widetilde{B}\hat{B}^T + \hat{B}\widetilde{B}^T + \widetilde{B}\widetilde{B}^T) && \text{Expanding.}
\end{aligned}$$

Since $\widetilde{\beta} = \hat{B}\mathbf{y} + \widetilde{B}\mathbf{y}$ is unbiased $\widetilde{B}\mathbf{X}^T = 0$. So

$$\widetilde{B}\hat{B}^T = \widetilde{B}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = 0 = (\mathbf{X}^T\mathbf{X})^{-1}(\widetilde{B}\mathbf{X})^T = \hat{B}\widetilde{B}^T\,.$$

Since $\mathrm{Var}(\hat{\beta}) = \sigma_\epsilon^2\hat{B}\hat{B}^T$, the variance can be written

$$\mathrm{Var}_\epsilon[\widetilde{\beta}] = \mathrm{Var}(\hat{\beta}) + \sigma_\epsilon^2\widetilde{B}\widetilde{B}^T\,.$$

Proof of the Gauss-Markov Theorem

Proof: We have shown that we can write

$$\text{Var}_\epsilon[\widetilde{\beta}] = \text{Var}(\hat{\beta}) + \sigma_\epsilon^2 \widetilde{B}\widetilde{B}^T \,,$$

where $\widetilde{B}\widetilde{B}^T$ is a positive semidefinite matrix, so $\text{Var}_\epsilon[\widetilde{\beta}] \geq \text{Var}_\epsilon[\hat{\beta}]$ as claimed. $\square$

Implications of the Gauss-Markov Theorem

The Gauss-Markov Theorem implies that the least squares estimator has the smallest mean squared error of any unbiased linear estimator. But there may still exists **biased** estimators with a smaller mean squared error. That is we may be able to trade a small increase in bias for a large reduction in variance.

We should note that subset selection methods are one way of doing this. If your selection procedure drops coefficients whose true value is nonzero, you will incur a error due to bias. However, subset selection is a discrete process and so often exhibits high variance.

Ridge Regression

Discrete vs Continuous

We want to look at more continuous (and smooth) methods of tuning, bounding and turning off coefficients in a linear model. One way to proceed is to continuously (or smoothly) modify the loss function to control the coefficients of the linear estimator more carefully.

There continuous operations tend to be more stable than discrete ones. In a certain sense, information about one state in a discrete object gives you no information about another state. By contrast, in a continuous object information about a point gives you information about a whole neighborhood around it. Concretely, best subset selection as a discrete operation is requires summing over all possible subsets, there is no smoothly varying the fit.

In general, discrete structures are plagued with problems, most famously **Godel's Incompleteness Theorem**, the **Halting Problem**, and for us the **No Free Lunch Theorem**.

Ridge Regression

Ridge Regression modifies the loss function to penalize coefficients β that are too large:

$$\text{Ridge}(\hat{\beta}) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\,.$$

Here, $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage in the coefficients. Notice that this is the Lagrangian version of the problem

$$\hat{\beta}^{ridge} = \underset{\beta}{\text{argmin}}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2\,,$$

subject to

$$\sum_{j=1}^{p}\beta_j^2 \leq t\,,$$

which makes the size explicit. There is a one to one correspondence between λ and t.

Scaling and Ridge Regression

Notice that ridge regression is not equivariant under scaling.

$$\text{Ridge}(\hat{\beta}) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \,.$$

That is, for fixed λ, the error is not invariant under a scaling of the training data. If instead λ is 0, then any scaling of x_i or y_i is absorbed into a rescaling of β.

Practically, this means that we often standardize the data to have sample variance $\bar{s}_j = 1$ for each feature.

Re-centering for Ridge Regression

In addition, notice that we have not included β_0 in the Lagrange term

$$\text{Ridge}(\hat{\beta}) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \,.$$

This is because we don't want to restrict the intercept at all. It can be shown (**exercise**) that ridge regression is equivalent to ridge regression on the shifted system

$$x_{ij} \mapsto x_{ij} - \bar{x}_j \,, \qquad y_i = y_i - \bar{y} \,,$$

under which the intercept β_0 is best estimated by 0. We will assume form here on out that the data has been normalized and shifted, so β is a p vector, not a $p+1$ vector.

Solution for Ridge Regression

For standardized data, writing the ridge loss in vector form

$$\text{Ridge}(\hat{\beta}) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta\,,$$

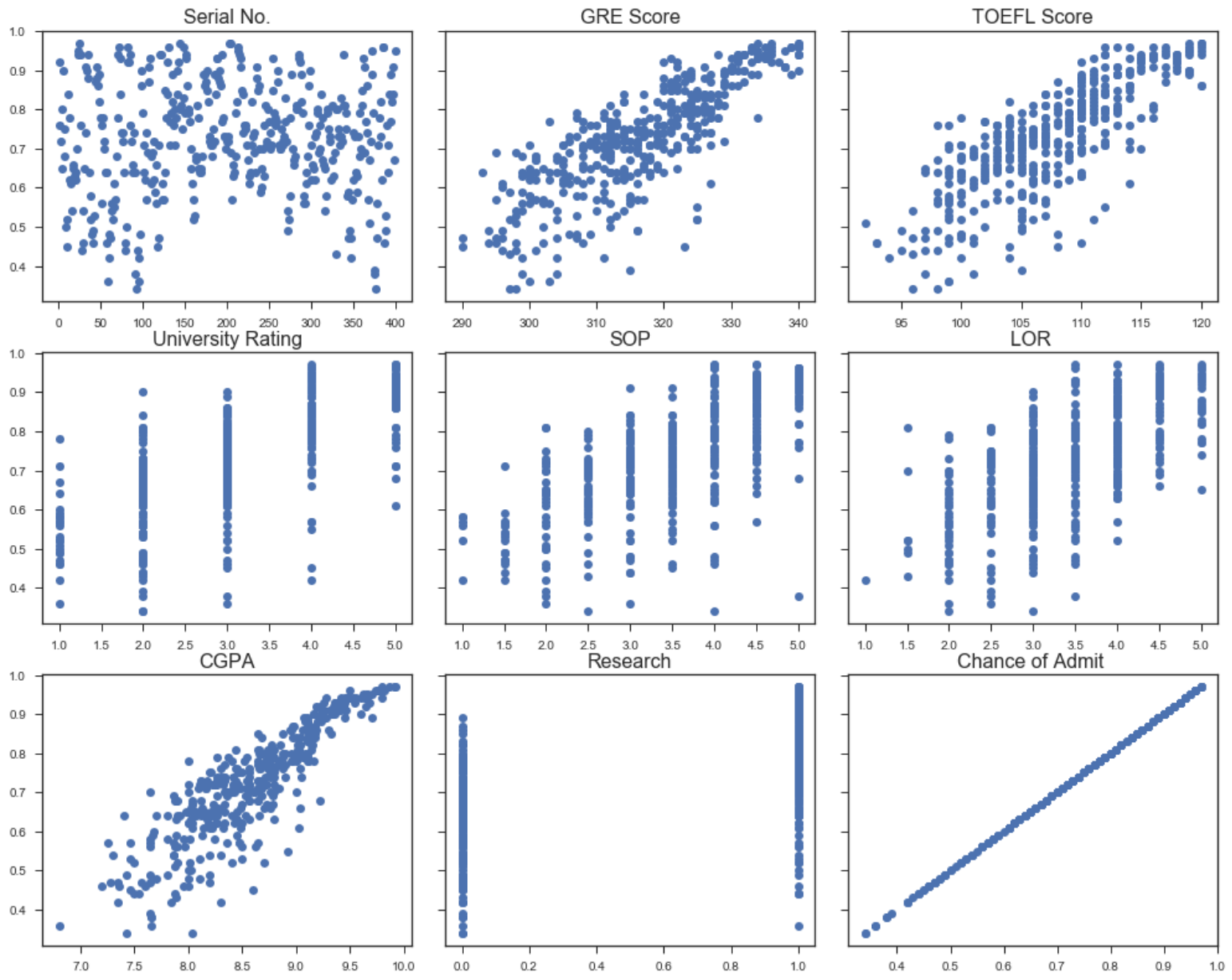
we can show that it is minimized by

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}\,.$$

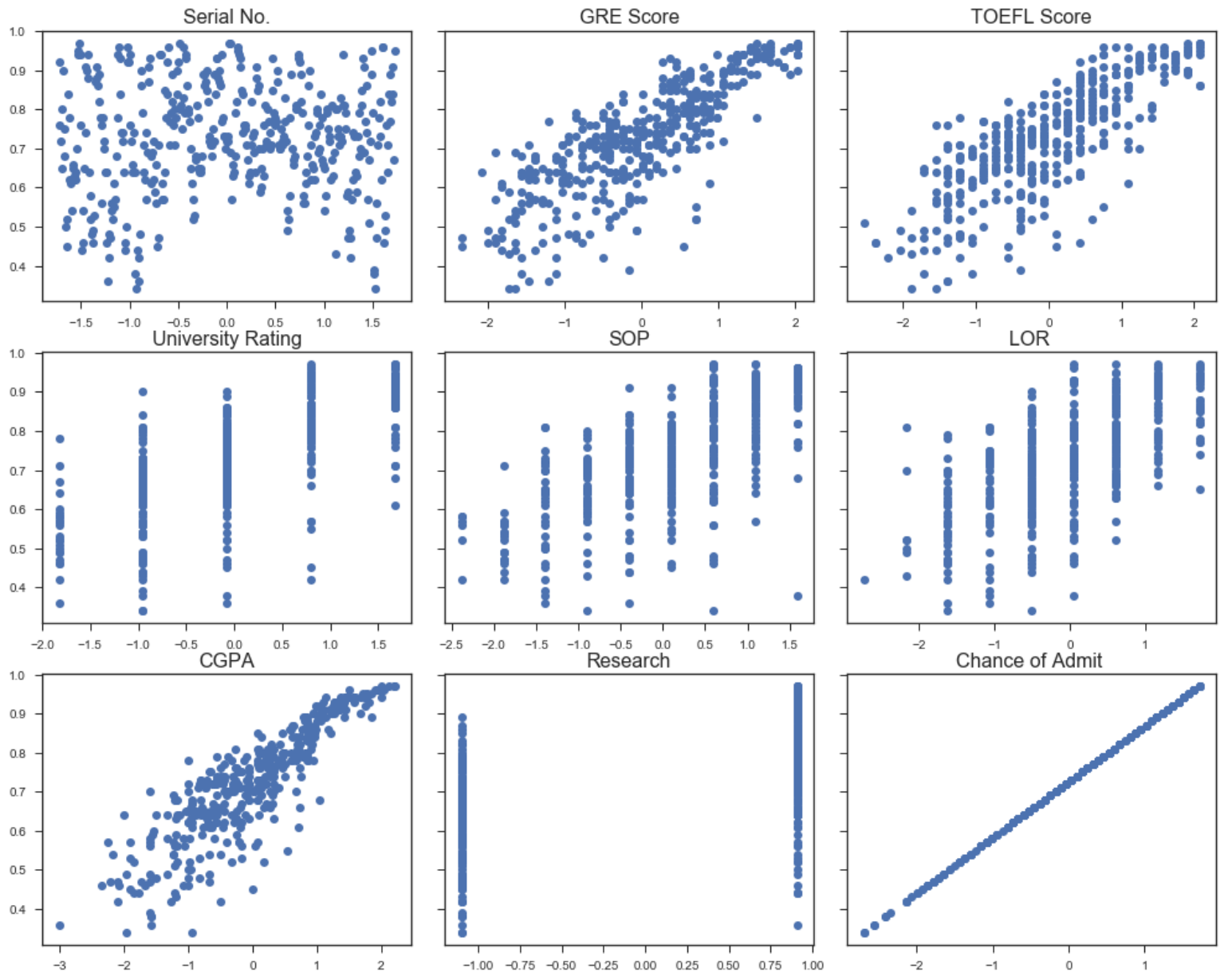
Notice that for $\lambda = 0$, $\hat{\beta}^{ridge} = \hat{\beta}$, so we actually have an entire family of solutions depending on λ.

Lets take a second to understand this on the student admissions data.

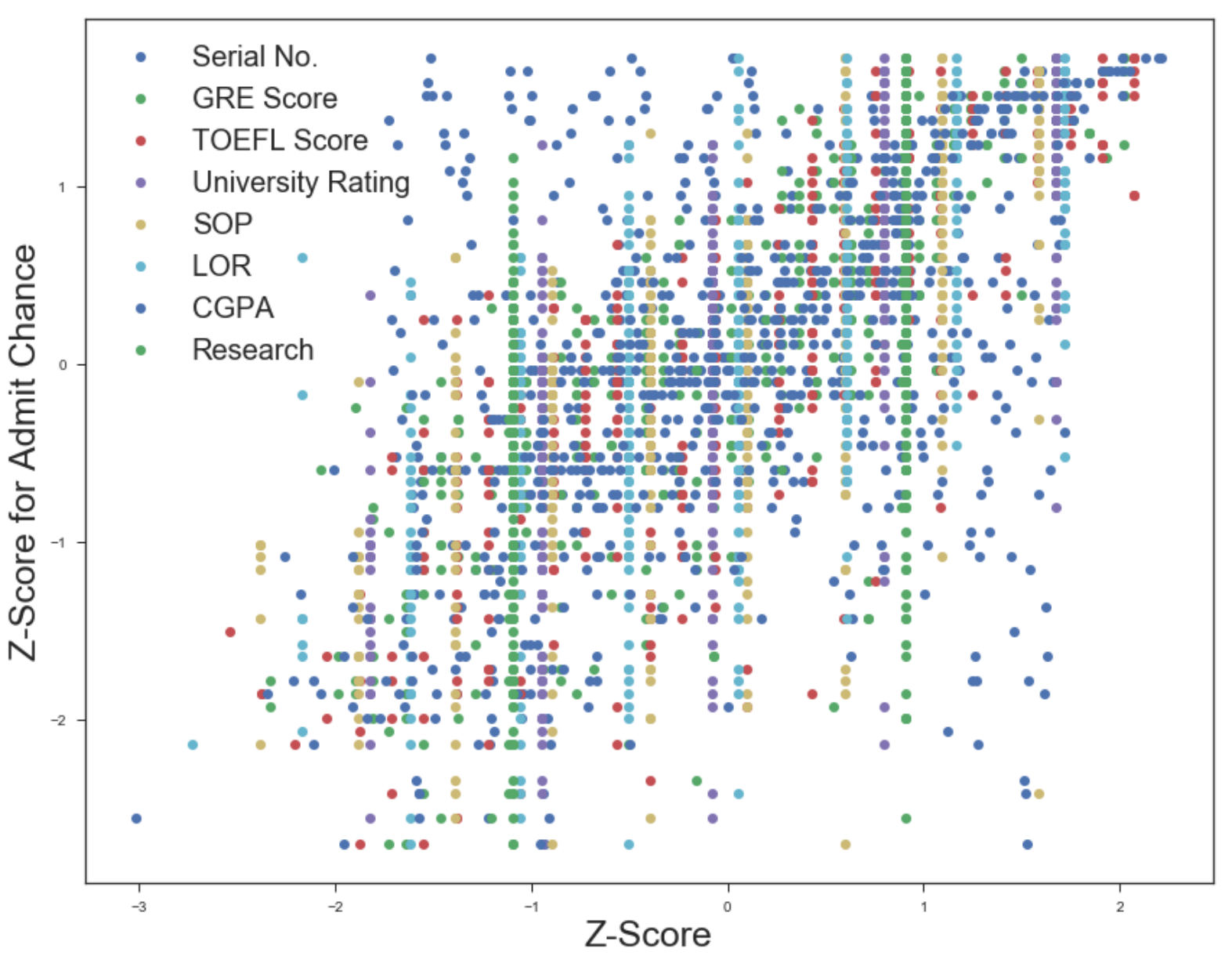
Admissions Data: Un-standardized



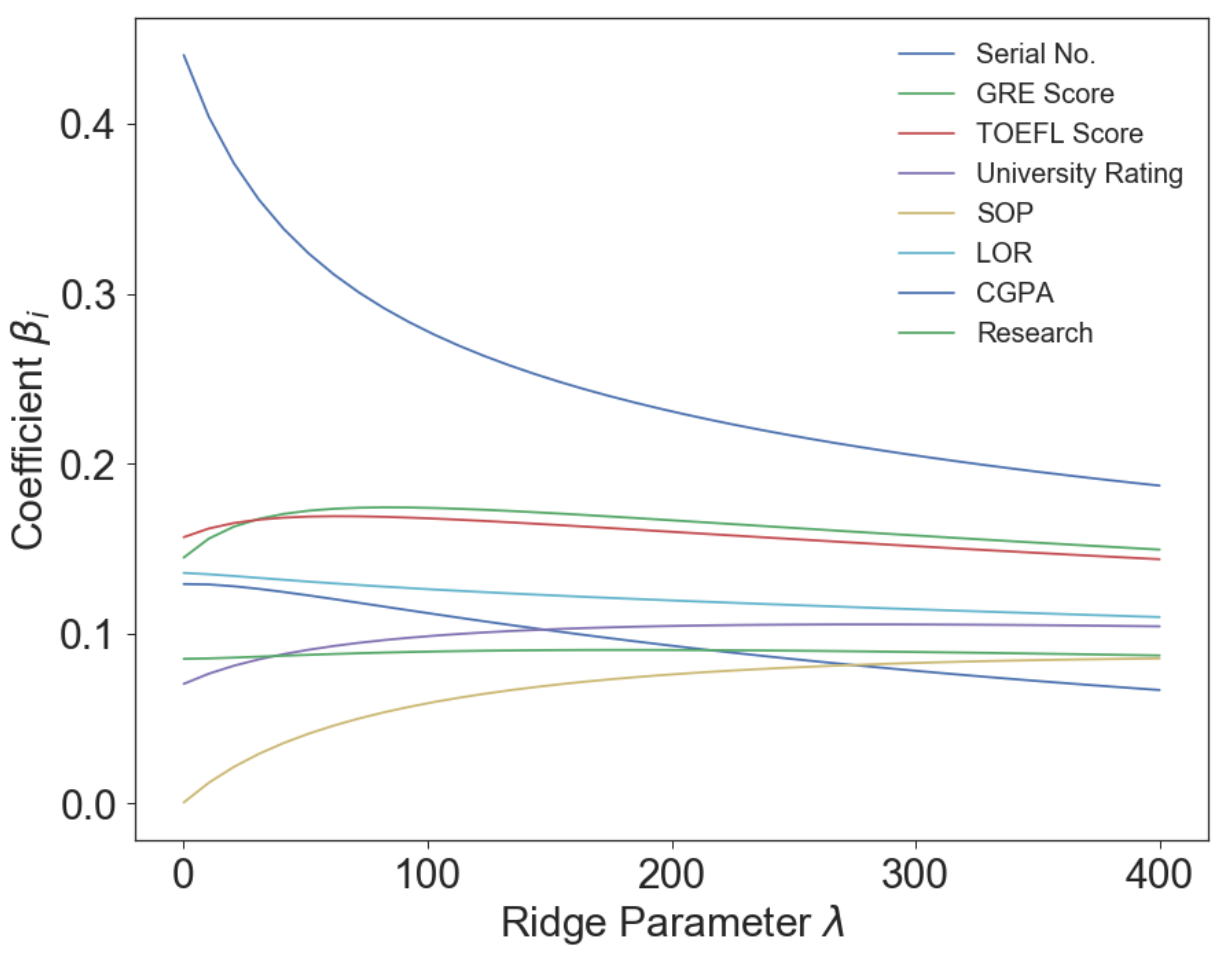
Admissions Data: Standardized



Admissions Data: Standardized

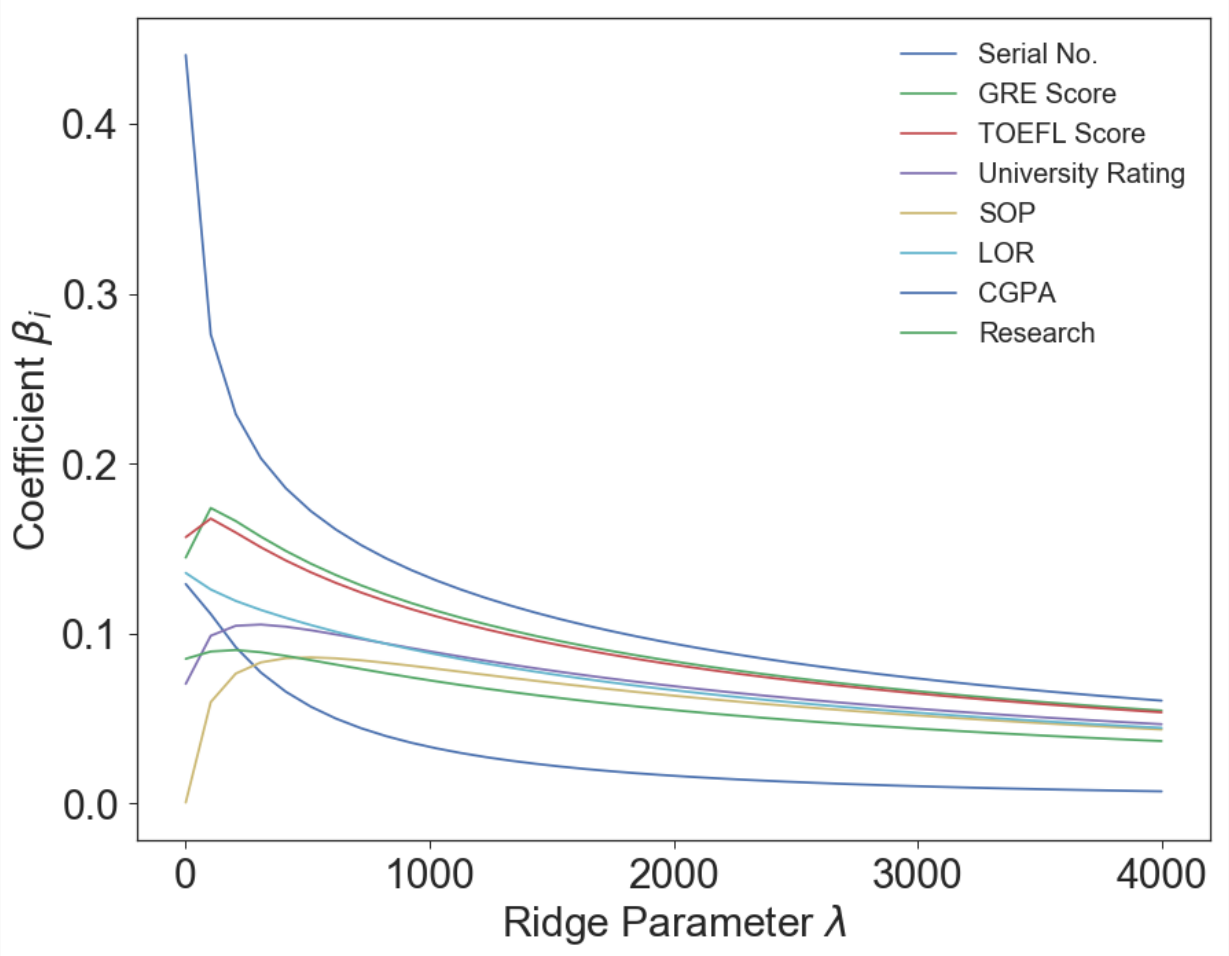


Solution for Ridge Regression



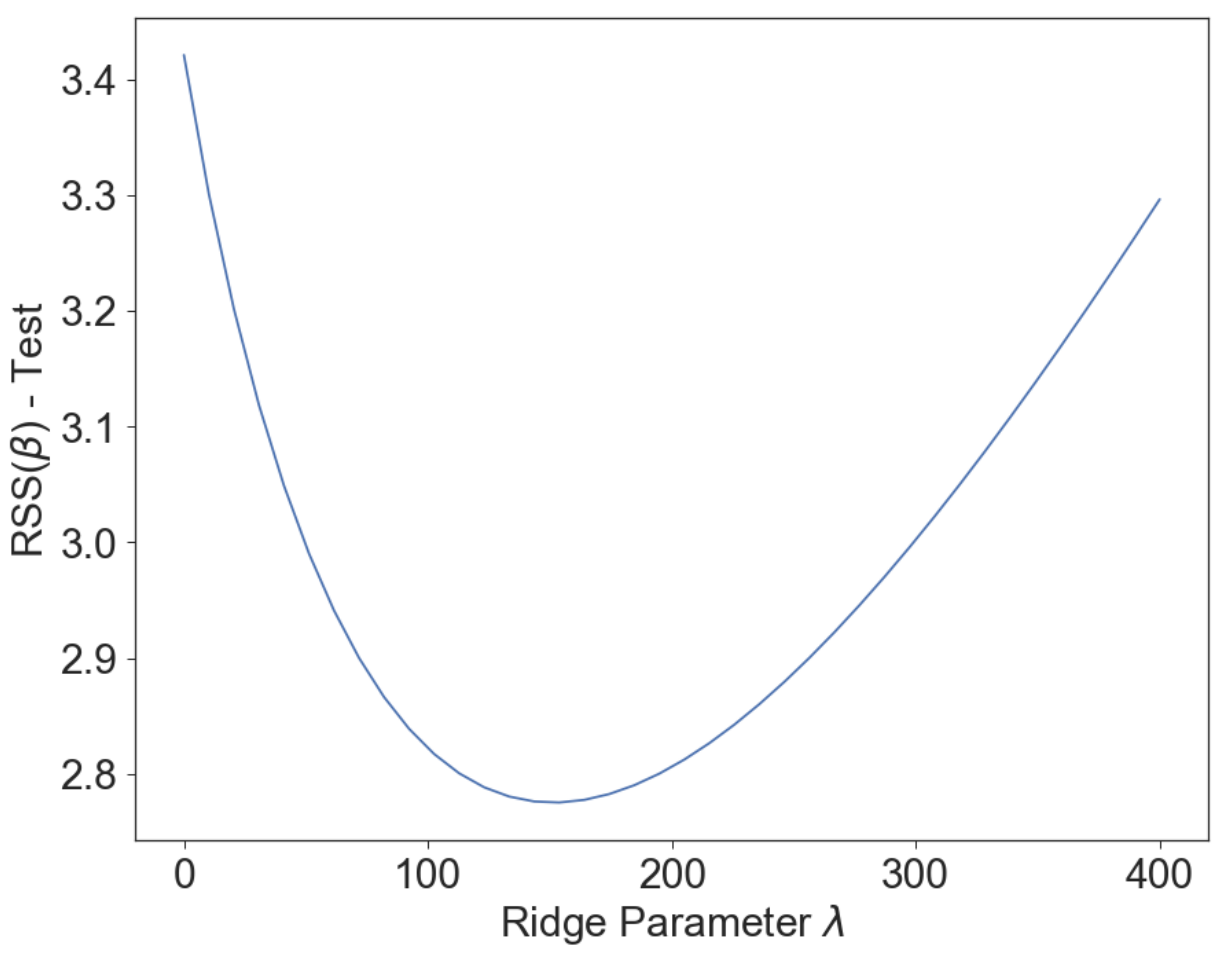
Applying ridge regression for λ between 0 and 400, we see the values fall off at different speeds.

Solution for Ridge Regression



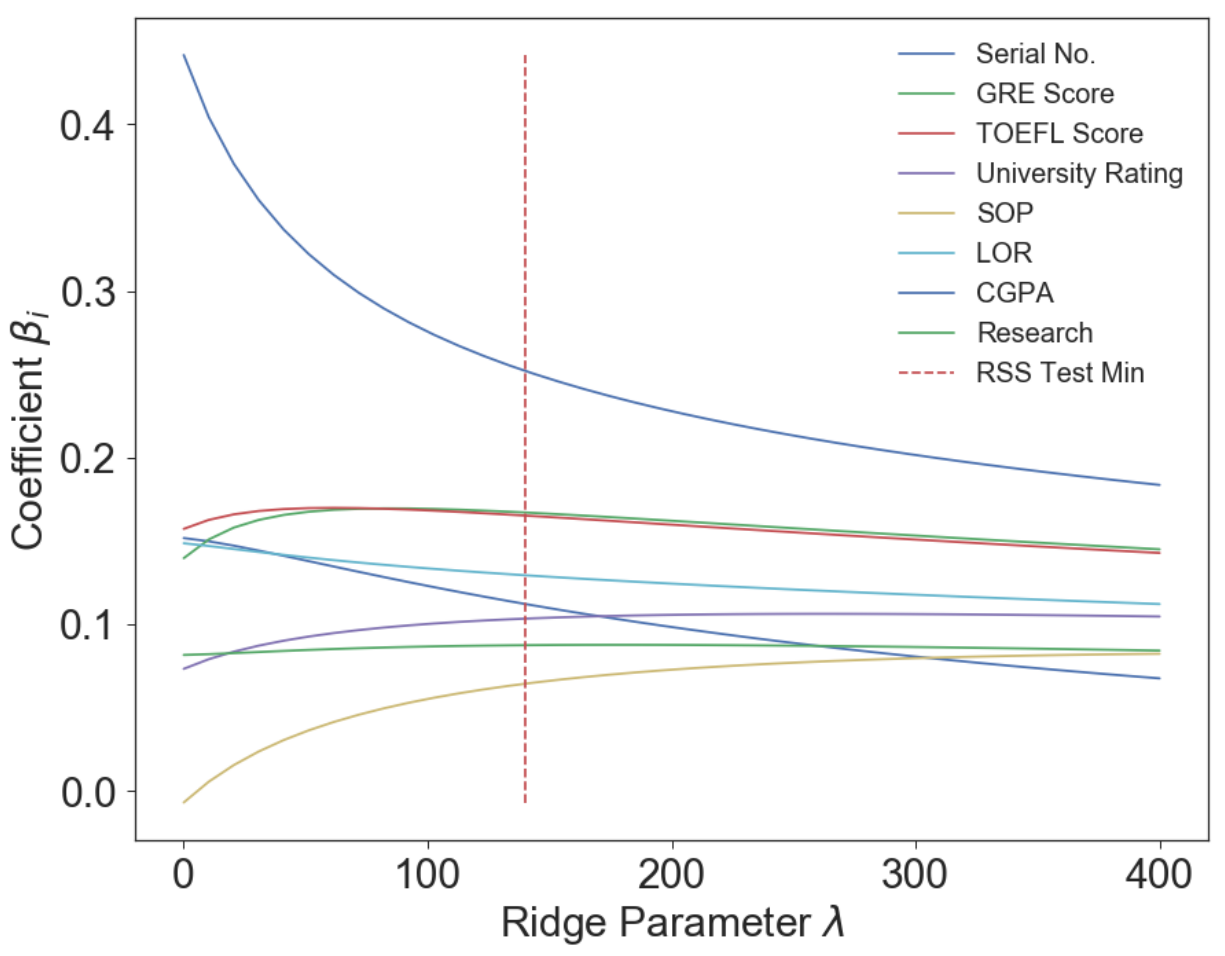
On a longer time frame, the all decrease to 0, as expected.

Solution for Ridge Regression



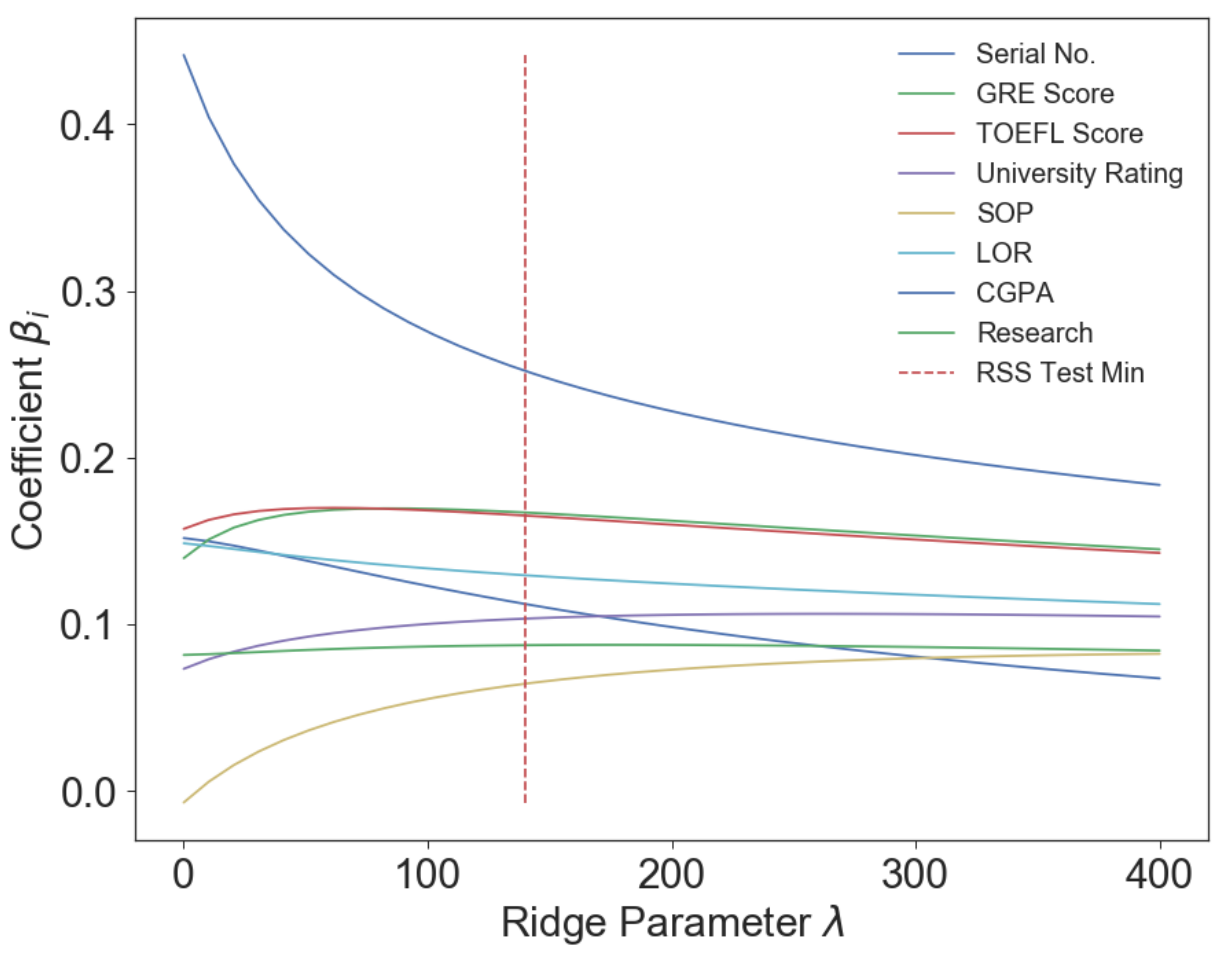
Computing *RSS* on a separate test set, *RSS* has a minimum around $\lambda = 150$, and the error is mush lower than the unbiased error at $\lambda = 0$.

Solution for Ridge Regression



Computing *RSS* on a separate test set, *RSS* has a minimum around $\lambda = 150$, and the error is mush lower than the unbiased error at $\lambda = 0$.

Solution for Ridge Regression



Lets take a moment and try to give some meaning to the horizontal axis.

Lasso Regression

Lasso Regression

Lasso Regression is similar to ridge regression, expect that we use the absolute value instead of the β^2. It turns out this slight non-smoothness leads to very different phenomena.

$$\text{Ridge}(\hat{\beta}) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\,.$$

Again, $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage in the coefficients. This is the Lagrangian version of the problem

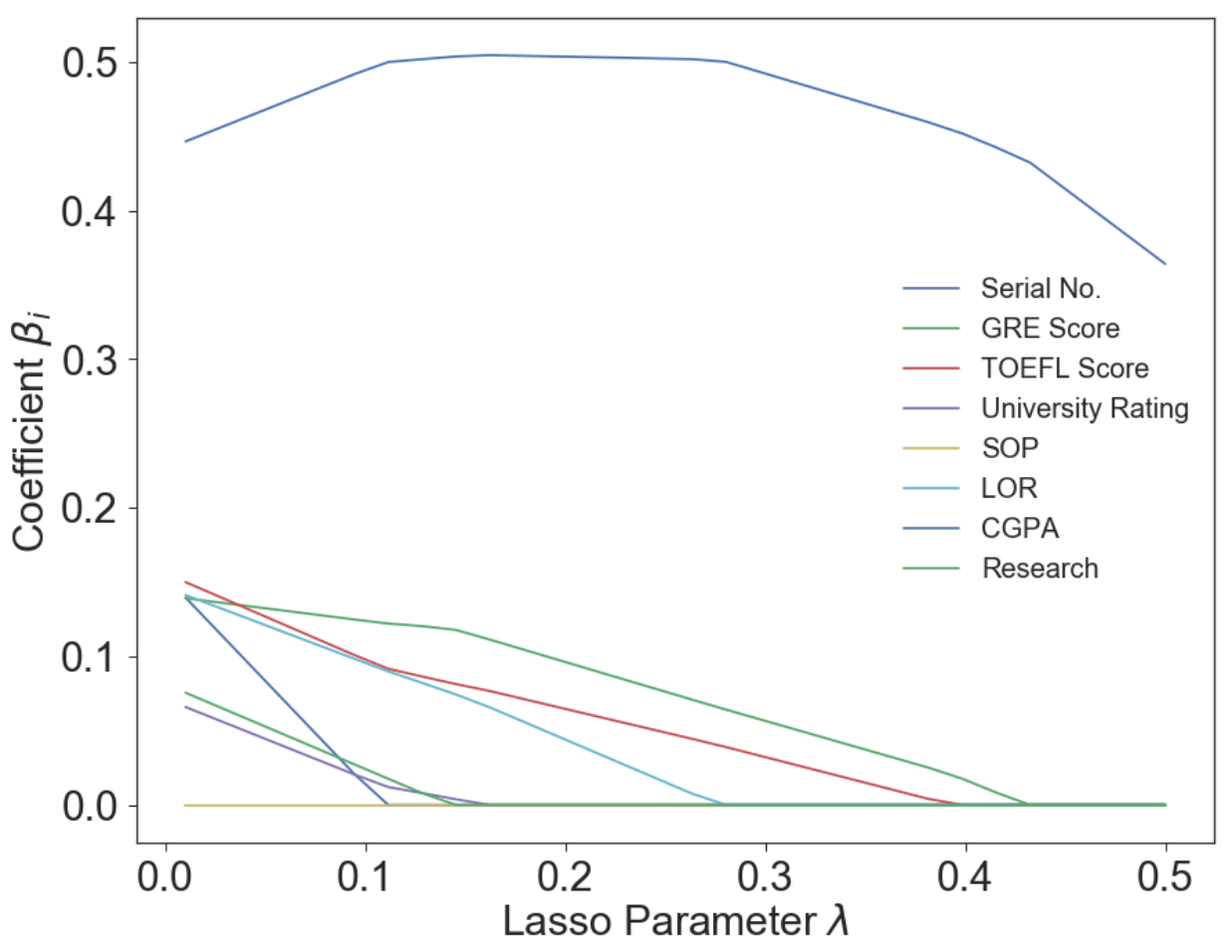
$$\hat{\beta}^{ridge} = \underset{\beta}{\text{argmin}}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2\,,$$

subject to

$$\sum_{j=1}^{p}|\beta_j| \leq t\,.$$

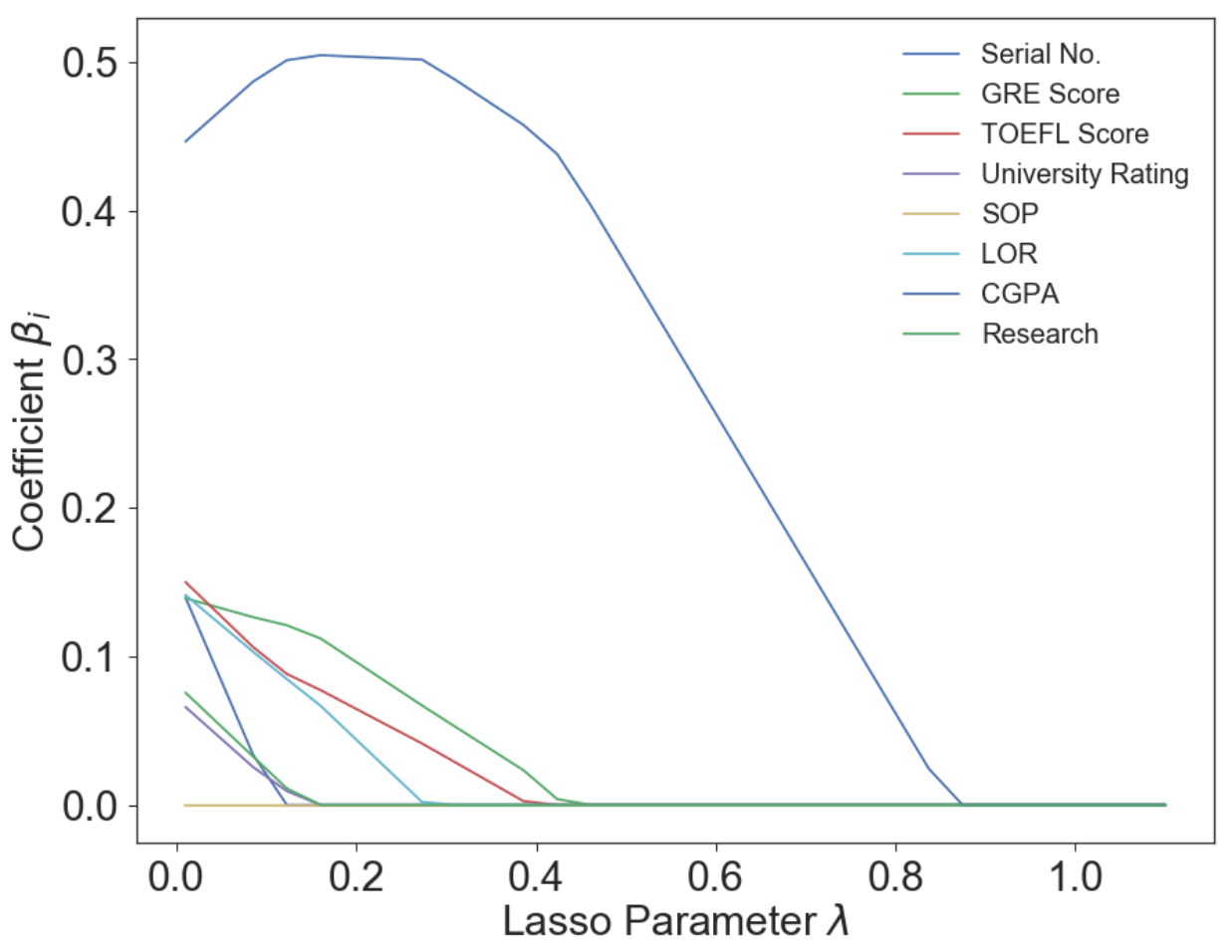
There is a one to one correspondence between λ and t.

Lasso Regression



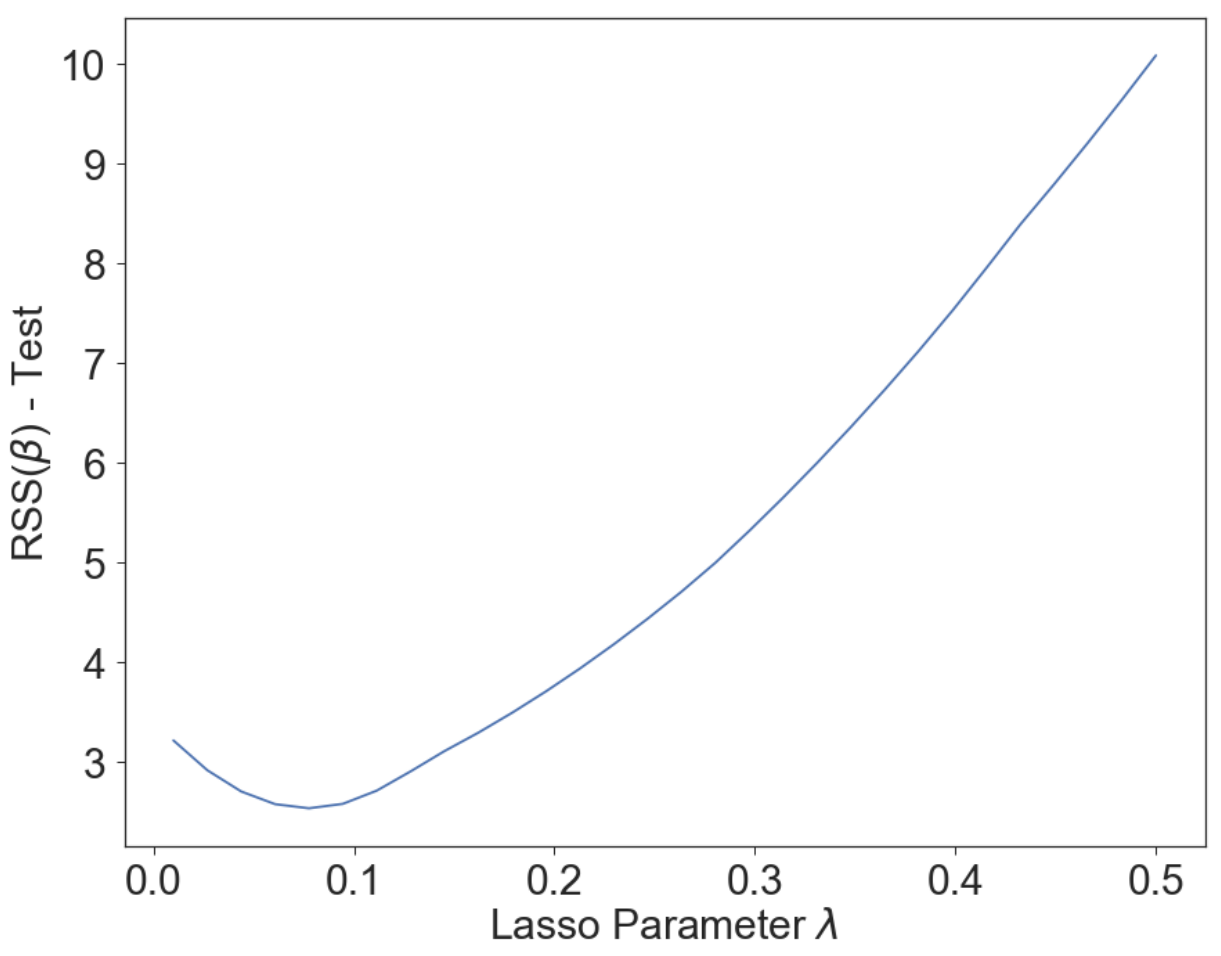
Lasso Regression differs from ridge regression in that it is able to set parameters to zero. As λ increases, the beta values are continuously (an apparently piecewise linearly) shrunk to zero.

Lasso Regression



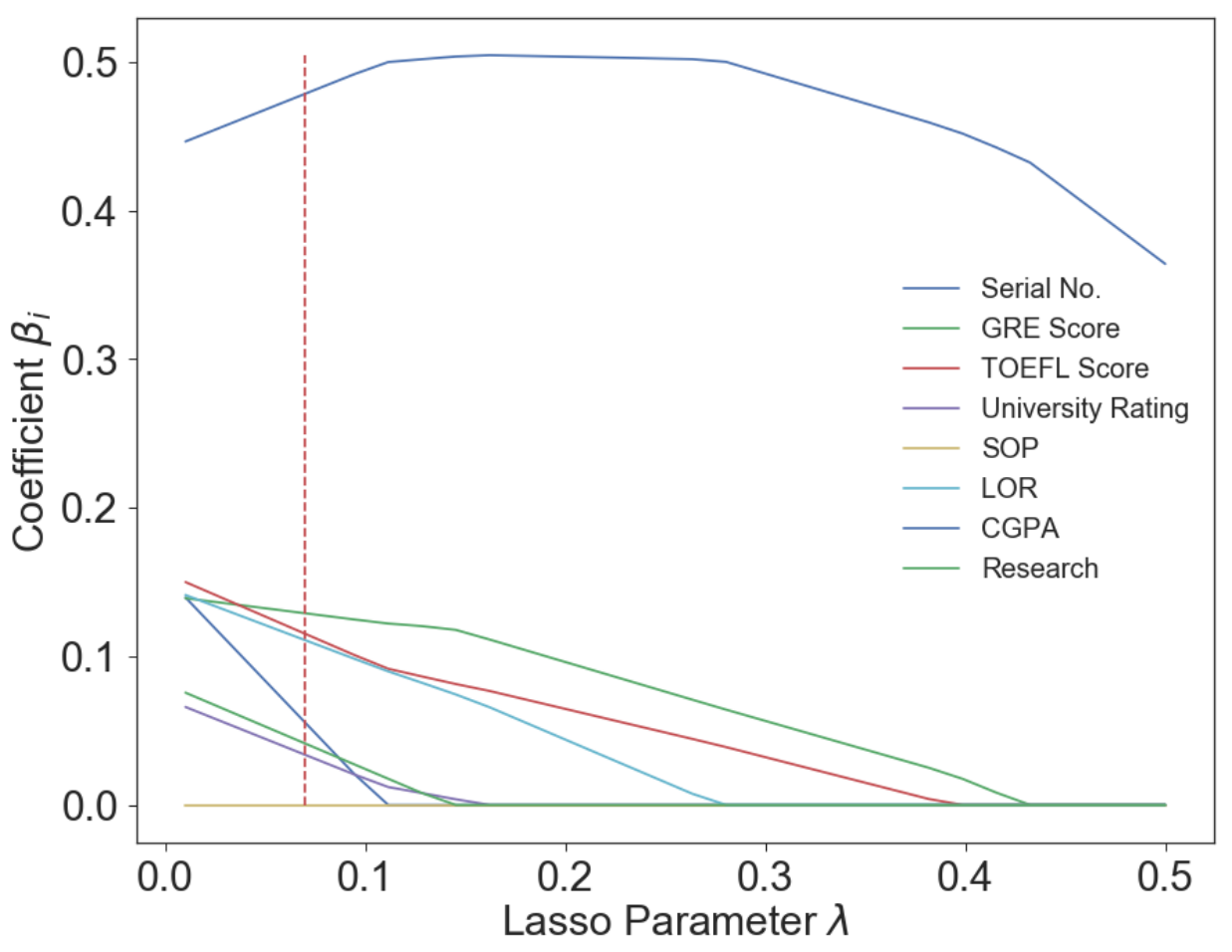
Lasso Regression differs from ridge regression in that it is able to set parameters to zero. As λ increases, the beta values are continuously (an apparently piecewise linearly) shrunk to zero.

Lasso Regression



We see the minimum *RSS* value is around $\lambda = .08$, plotting this against the coefficients we find that all of the parameters are nonzero, and **CGPA** is approaching its max.

Lasso Regression



We see the minimum *RSS* value is around $\lambda = .08$, plotting this against the coefficients we find that all of the parameters are nonzero, and **CGPA** is approaching its max.

Lasso Regression: Analytic

How does lasso regression set coefficients to 0? Analytically, assume $p = 1$ and assume the least squares solution has $\hat{\beta} > 0$. Then the derivative of the loss is

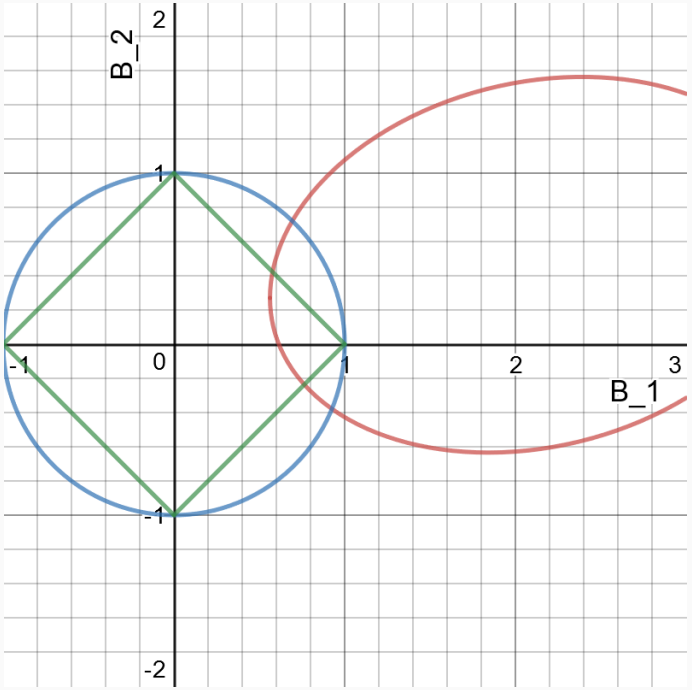
$$\frac{\partial}{\partial \beta}\textbf{Lasso}(\beta) = \frac{\partial}{\partial \beta}[(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta] = \lambda - 2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)\,,$$

which has the solution $\tilde{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}(2\mathbf{X}^T\mathbf{y} - \lambda)$. We can push $\tilde{\beta}$ to zero by pushing λ to $2\mathbf{X}^T\mathbf{y}$, but as soon as we cross 0 the derivative changes to

$$\frac{\partial}{\partial \beta}\textbf{Lasso}(\beta) = -\lambda - 2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)\,.$$

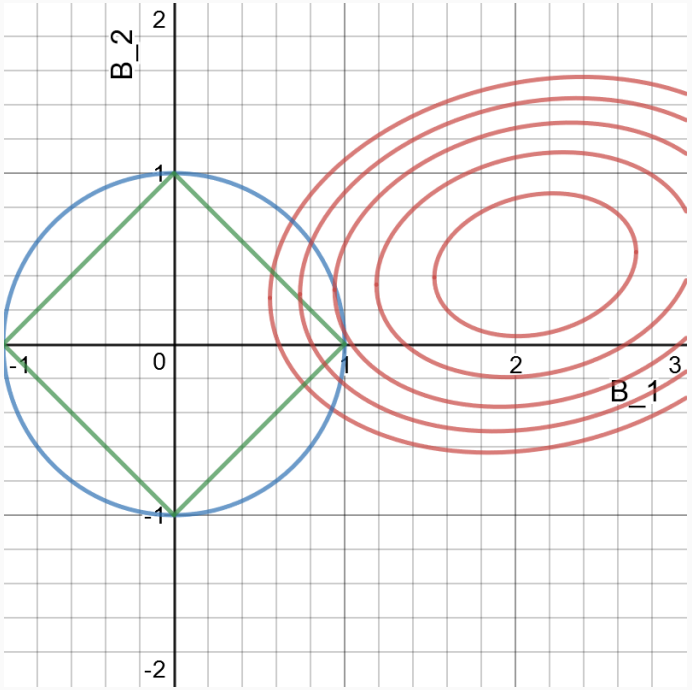
The solution $\tilde{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}(2\mathbf{X}^T\mathbf{y} + \lambda)$ is clearly positive, a contradiction. So 0 is the minimum of $\tilde{\beta}$ for $\lambda > 2\mathbf{X}^T\mathbf{y}$. A similar argument holds if $\hat{\beta} < 0$.

Lasso Regression: Geometric



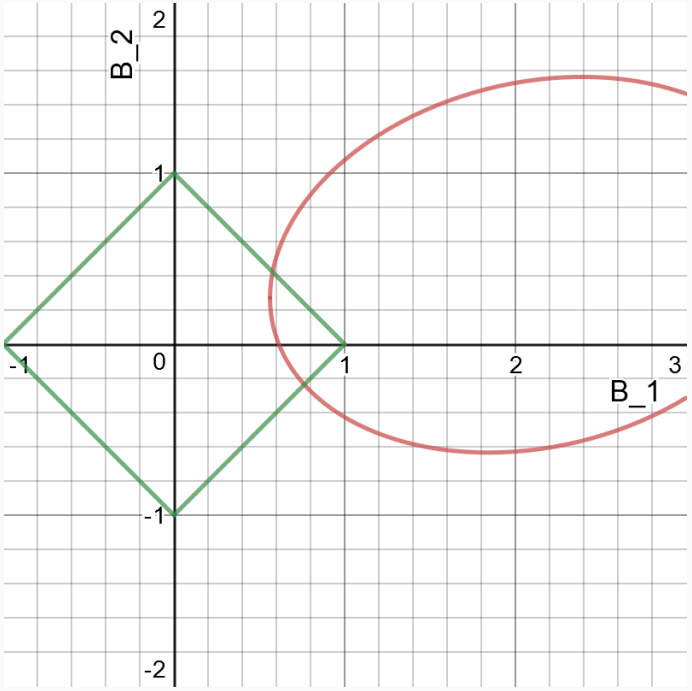
Geometrically, these are Lagrange multiplier problems with unconstrained λ. That means that we're trying to find the minimum of RSS subject to lying inside some region.

Lasso Regression: Geometric



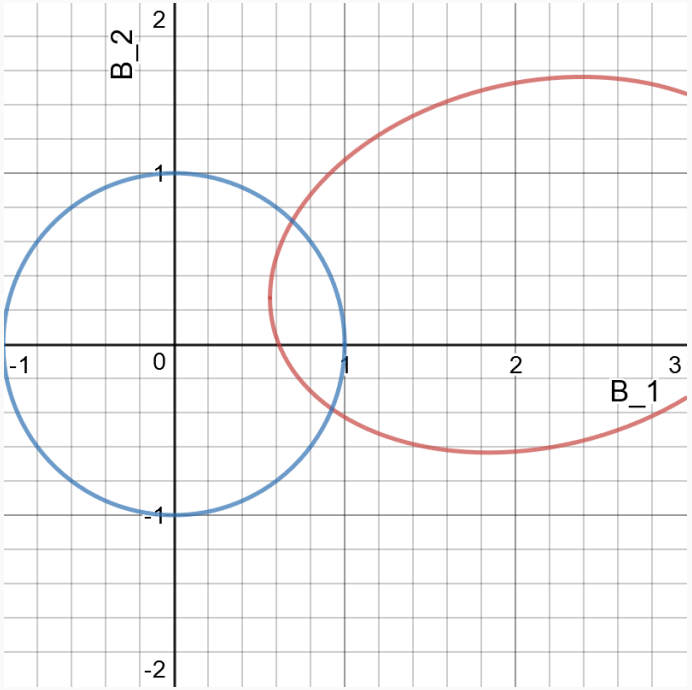
Since RSS is a quadratic function, the level curves will be elliptical.

Lasso Regression: Geometric



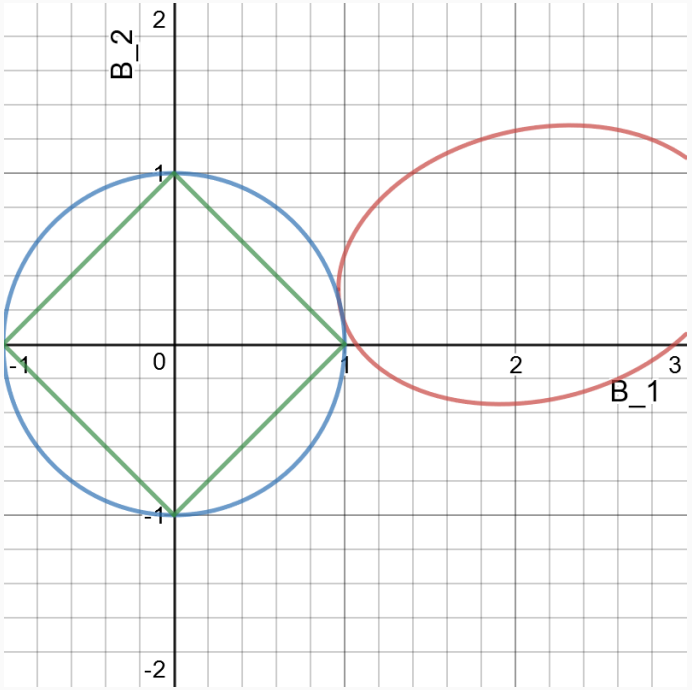
Since RSS is a quadratic function, the level curves will be elliptical. Similarly, lasso regression imposes a square condition $\sum_i |\beta_i| < \lambda^{-1}$,

Lasso Regression: Geometric



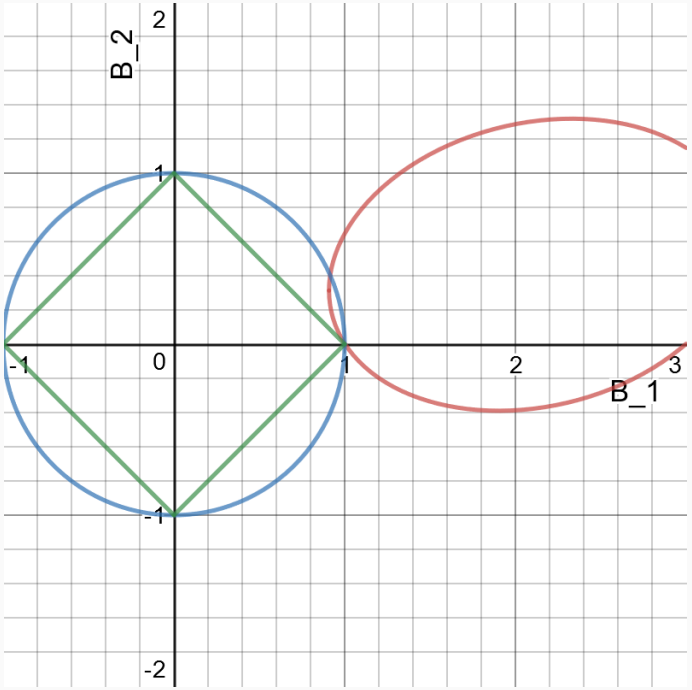
Since RSS is a quadratic function, the level curves will be elliptical. Similarly, lasso regression imposes a square condition $\sum_i |\beta_i| < \lambda^{-1}$, while ridge regression imposes a circular condition $\sum_i \beta_i^2 < \lambda^{-1}$.

Lasso Regression: Geometric



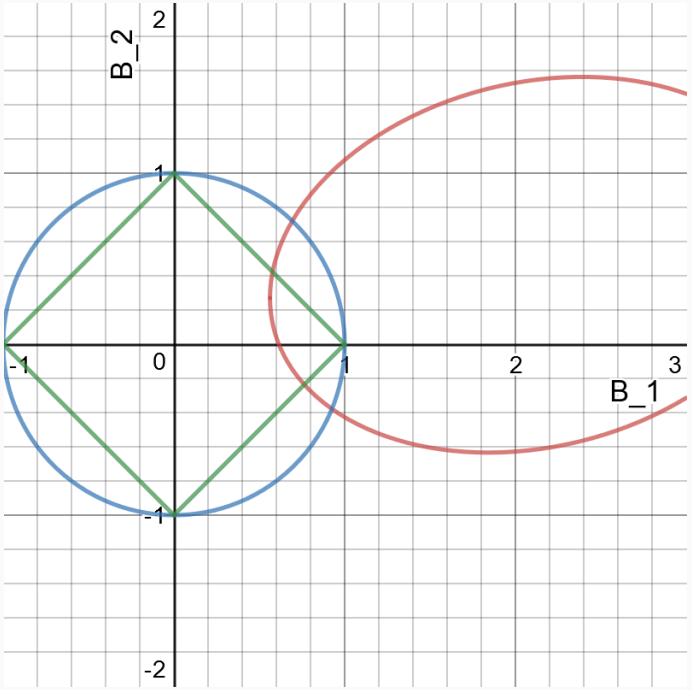
If $\lambda \gg 1$, the minimum of RSS will not be contained in the bounded region, and so the contained minimum will occur on the boundary. For a smooth boundary this can occur anywhere.

Lasso Regression: Geometric



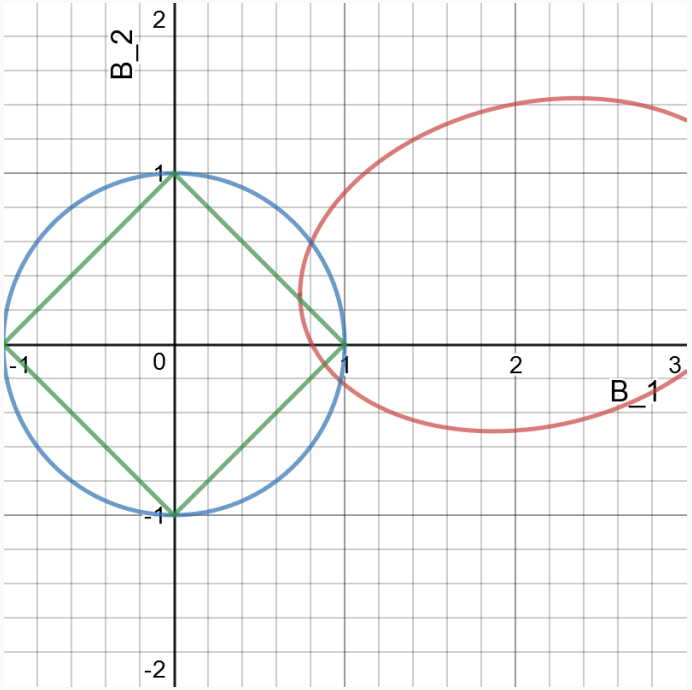
If $\lambda \gg 1$, the minimum of RSS will not be contained in the bounded region, and so the contained minimum will occur on the boundary. For a smooth boundary this can occur anywhere. For a singular boundary, it is much more likely to occur at a corner.

Lasso Regression: Geometric



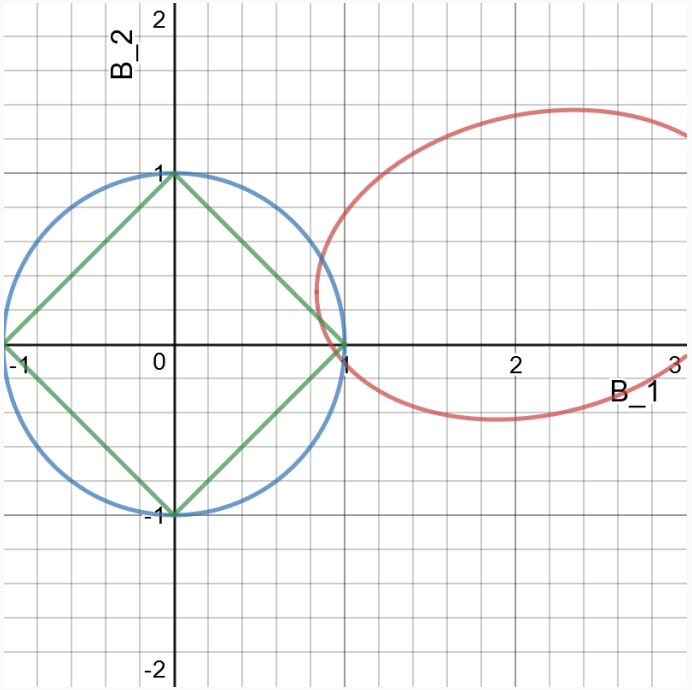
If $\lambda \gg 1$, the minimum of RSS will not be contained in the bounded region, and so the contained minimum will occur on the boundary. For a smooth boundary this can occur anywhere. For a singular boundary, it is much more likely to occur at a corner. This tightening to a corner gives the lasso its name.

Lasso Regression: Geometric



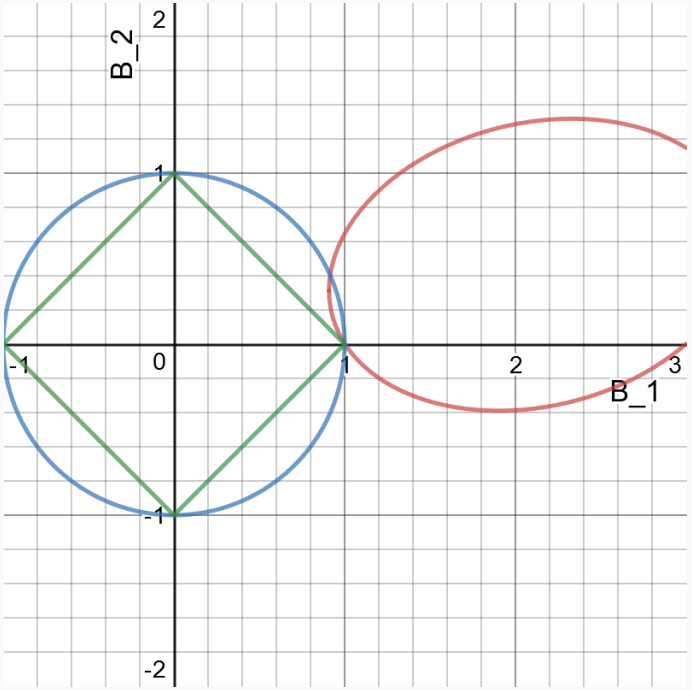
If $\lambda \gg 1$, the minimum of RSS will not be contained in the bounded region, and so the contained minimum will occur on the boundary. For a smooth boundary this can occur anywhere. For a singular boundary, it is much more likely to occur at a corner. This tightening to a corner gives the lasso its name.

Lasso Regression: Geometric



If $\lambda \gg 1$, the minimum of RSS will not be contained in the bounded region, and so the contained minimum will occur on the boundary. For a smooth boundary this can occur anywhere. For a singular boundary, it is much more likely to occur at a corner. This tightening to a corner gives the lasso its name.

Lasso Regression: Geometric



If $\lambda \gg 1$, the minimum of RSS will not be contained in the bounded region, and so the contained minimum will occur on the boundary. For a smooth boundary this can occur anywhere. For a singular boundary, it is much more likely to occur at a corner. This tightening to a corner gives the lasso its name.

Generalizations

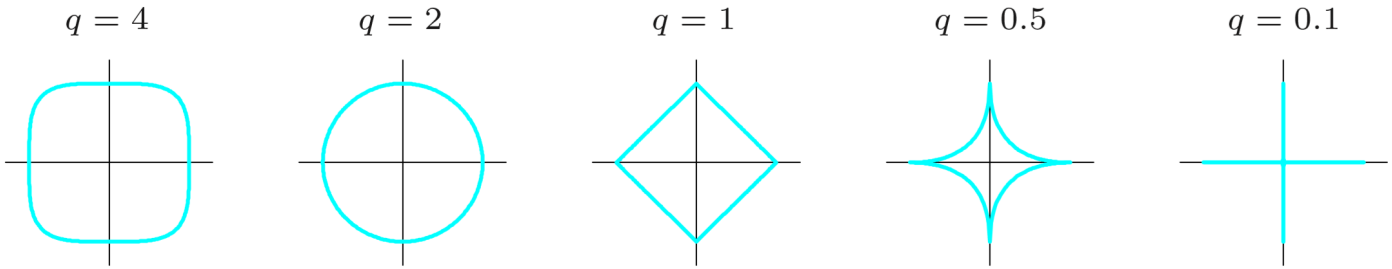


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

As a final word, we can generalize ridge and lasso regression by using the loss function

$$\text{Ridge}(\hat{\beta}) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \,,$$

where q is a positive. Although this unifies the framework nicely, HTF express skepticism about it's usefulness.

Degrees of Freedom

Degrees of Freedom

As we move between methods of estimating the underlying regression function for a learning problem, we want to compare estimates of test error between different methods. We can use cross validation to split our training set into a training and test set, but in practice comparing cross validation curves between methods isn't straight forward.

For example, what does it mean to pick Ridge Regression with $\lambda = 150$ over the linear regression on three variables? Or k-nearest neighbors for $k = 5$? We would like some sort of measure of the relative complexity between estimators.

Degrees of Freedom

The notion of **degrees of freedom** is often used to provide an abstraction of the number of "effective" parameters used to fit a model.

Though conceptually quite broad, degrees of freedom have a concrete definition for noisy predictors: Suppose data is generated by

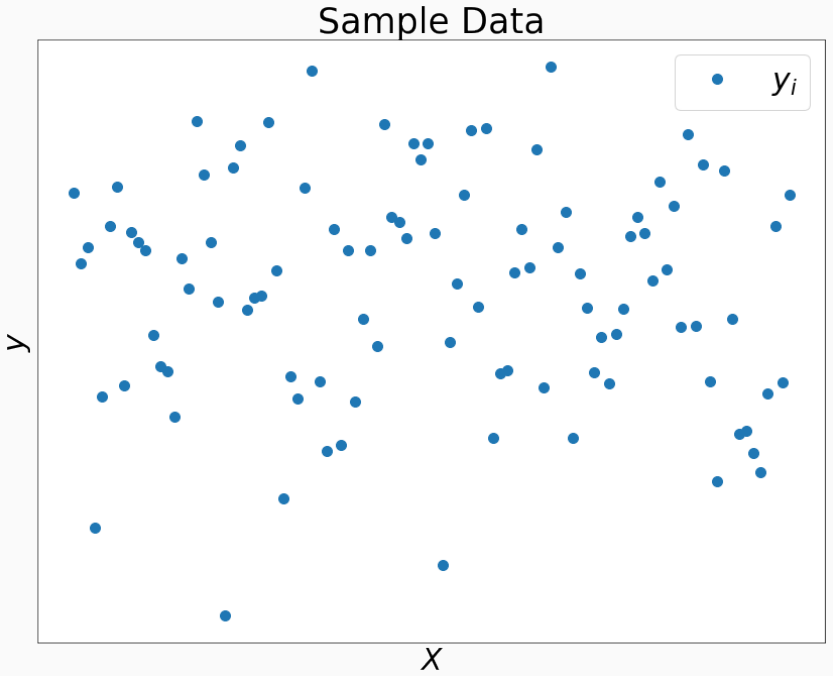
$$y = f(x) + \epsilon\,, \qquad E[\epsilon] = 0\,,\ \mathrm{Var}(\epsilon) = \sigma_\epsilon^2$$

and suppose we have fit some $\hat{y}_i = \hat{f}(x_i)$ to it a training sample of size N. The **number of degrees of freedom** of $\hat{f}$ is

$$\mathrm{df}(\hat{f}) = \frac{1}{\sigma_\epsilon^2}\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y_i) = \frac{1}{\sigma_\epsilon^2}\sum_{i=1}^{N}\mathrm{Cov}(\hat{f}(x_i), y_i)\,.$$

I will justify this definition by examples.

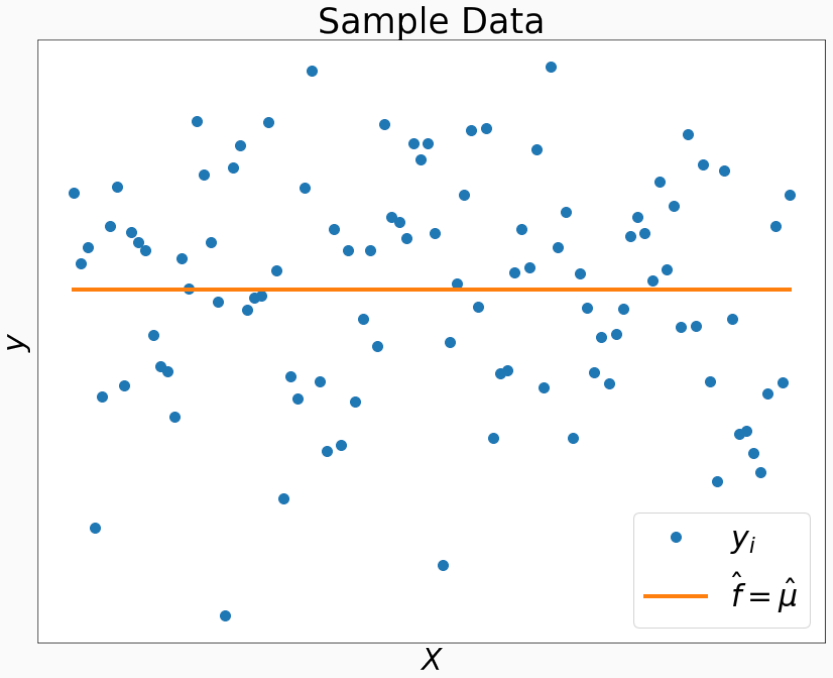
Degrees of Freedom: Examples



For example, assume that we have drawn a sample $y_i \in P(y|X)$ from some distribution.

If a predictor $\hat{f}$ has one degree of freedom, we would expect $\hat{f}$ with a single parameter, ie the constant predictor. For example, we would expect the mean predictor $\hat{f}(x_i) = \bar{y} = \frac{1}{N}(y_1 + \ldots + y_N)$ to have a single degree of freedom.

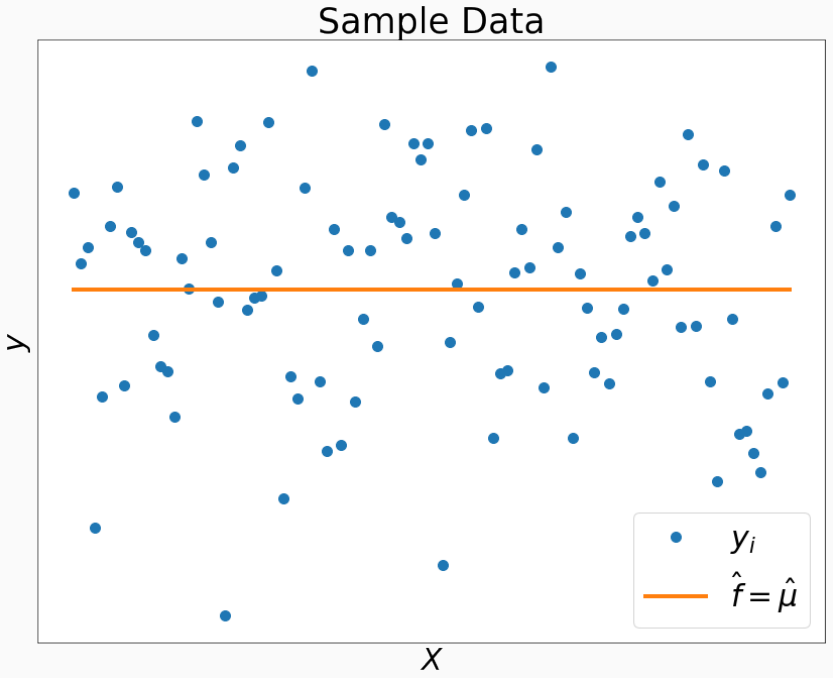
Degrees of Freedom: Examples



For example, assume that we have drawn a sample $y_i \in P(y|X)$ from some distribution.

If a predictor $\hat{f}$ has one degree of freedom, we would expect $\hat{f}$ with a single parameter, ie the constant predictor. For example, we would expect the mean predictor $\hat{f}(x_i) = \bar{y} = \frac{1}{N}(y_1 + \ldots + y_N)$ to have a single degree of freedom.

Degrees of Freedom: Examples

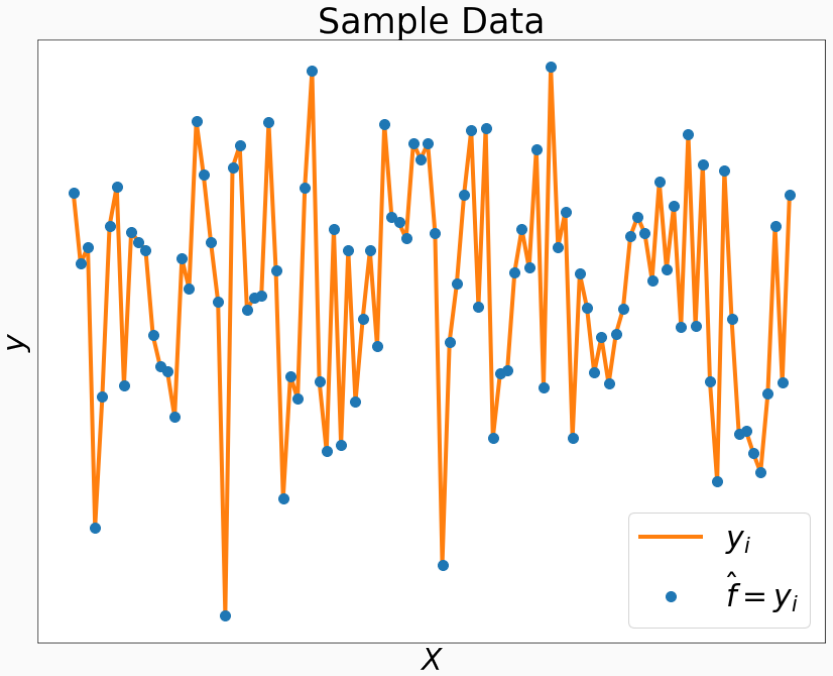


Indeed, since the ϵ_i are i.i.d., the mean predictor $\hat{f} = \frac{1}{N}(y_1 + \ldots + y_N)$ has

$$\mathrm{df}(\hat{f}) = \frac{1}{\sigma_\epsilon^2} \sum_{i=1}^{N} \mathrm{Cov}(\hat{f}, y_i) = \frac{1}{N\sigma_\epsilon^2} \sum_{i=1}^{N} \mathrm{Cov}(y_1 + \ldots + y_n, y_i) = 1\,.$$

By i.i.d., $\mathrm{Cov}(y_1 + \ldots + y_n, y_i) = \mathrm{Cov}(y_i, y_i) = \sigma_\epsilon^2$.

Degrees of Freedom: Examples

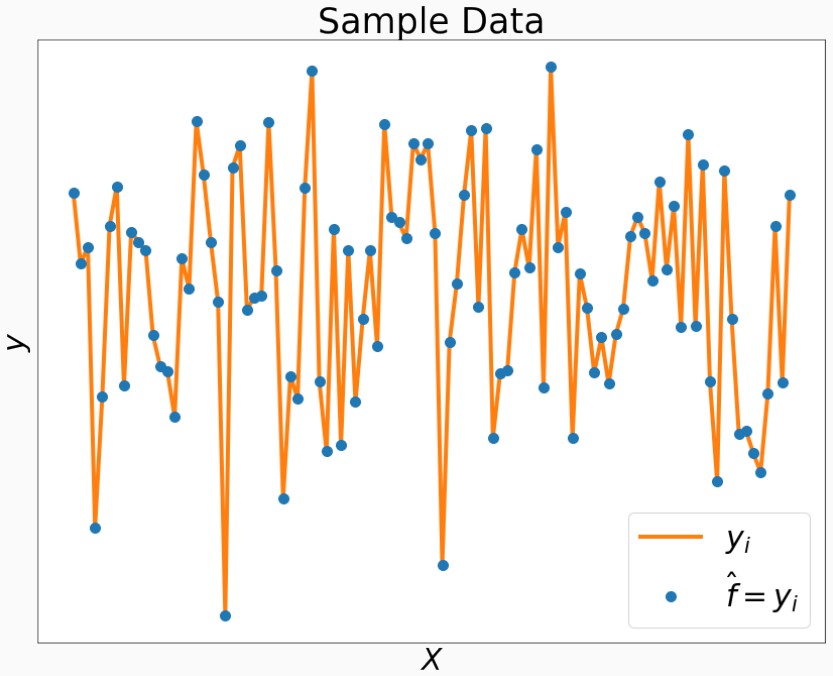


On the other hand, the identity estimator $f(x_i) = y_i$ has N degrees of freedom:

$$\mathrm{df}(\hat{f}) = \frac{1}{\sigma_\epsilon^2} \sum_{i=1}^{N} \mathrm{Cov}(y_i, y_i) = \frac{\sigma_\epsilon^2 N}{\sigma_\epsilon^2} = N\,.$$

Again, this intuitively makes sense: we would need at least N parameters to consistently make such a fit.

Degrees of Freedom: Examples

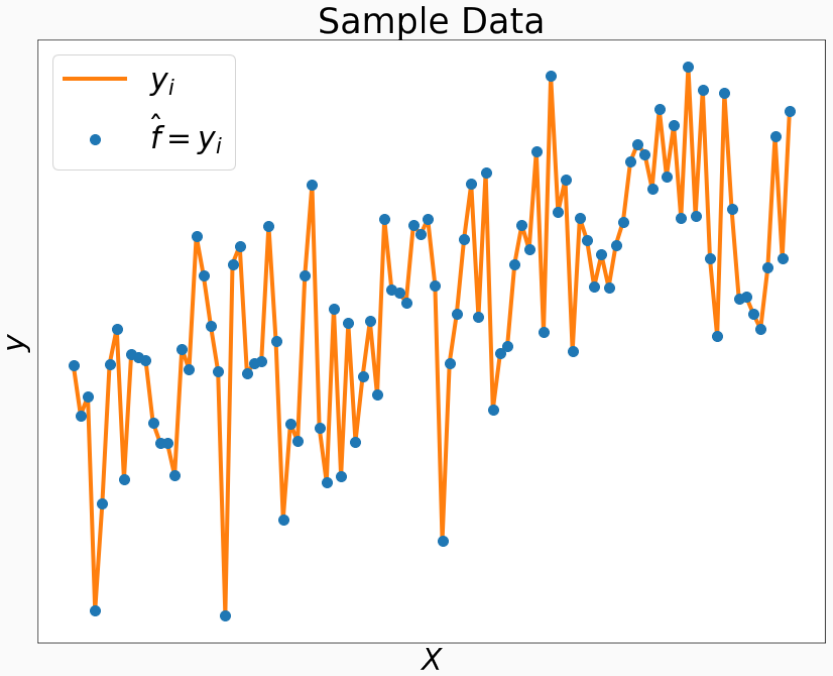


On the other hand, the identity estimator $f(x_i) = y_i$ has N degrees of freedom:

$$\mathrm{df}(\hat{f}) = \frac{1}{\sigma_\epsilon^2} \sum_{i=1}^{N} \mathrm{Cov}(y_i, y_i) = \frac{\sigma_\epsilon^2 N}{\sigma_\epsilon^2} = N\,.$$

Again, this intuitively makes sense: we would need at least N parameters to consistently make such a fit.

Degrees of Freedom: Examples



This notion of degrees of freedom gives a continuous measurement for the number of points we correctly guess, normalized by the number of standard deviations from the mean. We would not like to see that if the labels y_i depend linearly on p parameters, the the expected number of degrees of freedom will indeed be p.

Degrees of Freedom: Examples

We can write the degrees of freedom more compactly as

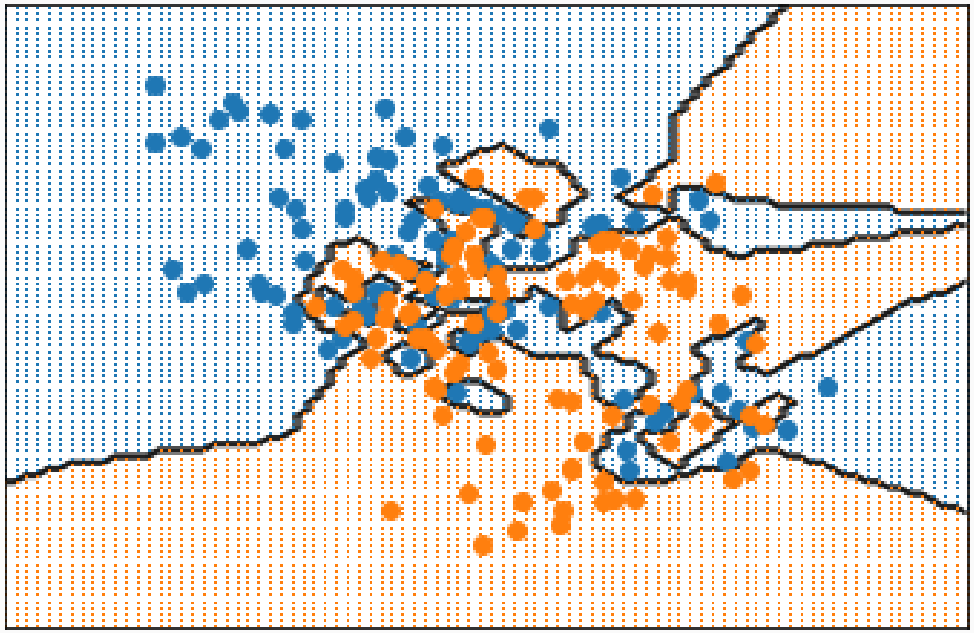
$$\mathrm{df}(\hat{f}) = \frac{1}{\sigma_\epsilon^2}\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y_i) = \frac{1}{\sigma_\epsilon^2}\mathrm{Tr}\big(\mathrm{Cov}(\hat{y}, y)\big)\,.$$

For a linear model with p inputs (we assume no β_0), the RSS solution has

$$\begin{aligned}
df(\hat{\beta}) &= \frac{1}{\sigma^2}\mathrm{Tr}\big(\mathrm{Cov}(\hat{y}, y)\big) \\
&= \frac{1}{\sigma^2}\mathrm{Tr}\big(\mathrm{Cov}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \mathbf{y})\big) && \text{Def. of } \hat{y}, \\
&= \frac{1}{\sigma^2}\mathrm{Tr}\big(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Cov}(\mathbf{y}, \mathbf{y})\big) && bfX \text{ fixed}, \\
&= \mathrm{Tr}\big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X})\big) && \mathrm{Tr}(AB) = \mathrm{Tr}(BA) \\
&= p\,.
\end{aligned}$$

Which again follows our intuition.

Degrees of Freedom: Examples

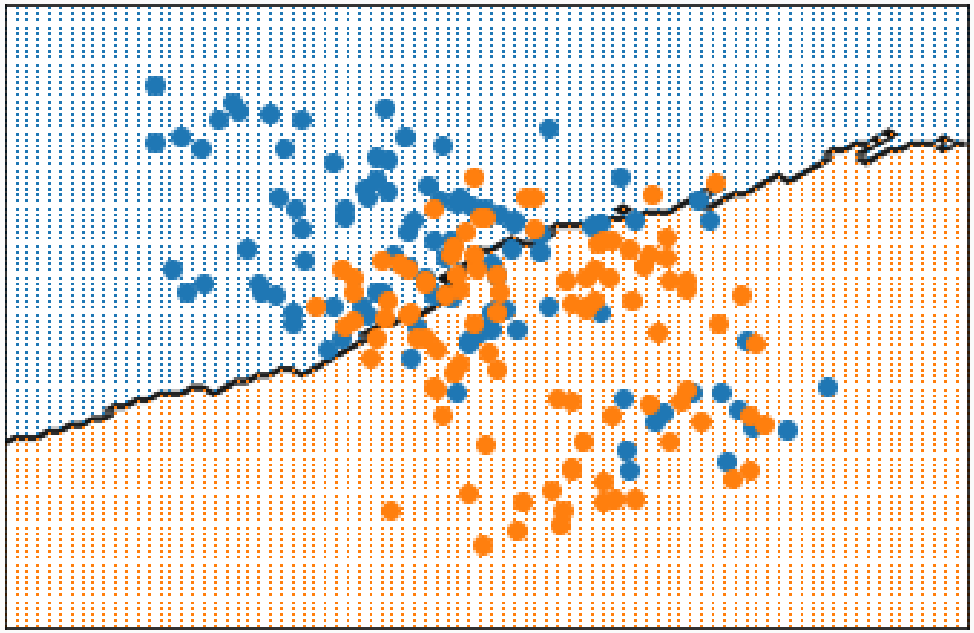


(**Exercise**) Show that k-nearest neighbors has

$$\mathrm{df}(\hat{f}^{knn}) = \frac{N}{k}\,.$$

Again, this follows our intuition that k-NN interpolates between high and low variance.

Degrees of Freedom: Examples



(**Exercise**) Show that k-nearest neighbors has

$$\mathrm{df}(\hat{f}^{knn}) = \frac{N}{k}\,.$$

Again, this follows our intuition that k-NN interpolates between high and low variance.

Singular Value Decomposition and Degrees of Freedom for Ridge Regression

Singular Value Decomposition

Let $N \geq p$. The **singular value decomposition** (**SVD**) of a real $N \times p$ matrix $\mathbf{X}$ is a factorization of $\mathbf{X}$ into a product

$$\mathbf{X} = UDV^T,$$

where

D is $N \times p$ matrix with $d_1 \geq d_2 \ldots \geq d_p \geq 0$ down the diagonal and 0's elsewhere.

U is an orthogonal $N \times N$ matrix whose columns span the column space of A.

V is an orthogonal $p \times p$ matrix with columns spanning the row space.

SVD is a generalization of the eigen-decomposition of a symmetric matrix, and is a very useful tool for analyzing linear algorithms.

Singular Value Decomposition

Singular value decomposition allows us to write

$$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = UU^T\,. \qquad \textbf{(Exercise)}$$

so the least squares solution becomes $\mathbf{X}\hat{\beta} = UU^T\mathbf{y}$.

Similarly, we can write the ridge solution as

$$\begin{aligned}\mathbf{X}\beta^{Ridge} &= \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} \\ &= UD(D^TD + \lambda I)^{-1}D^TU^T\mathbf{y}\,.\end{aligned}$$

Lets take a moment to look at the central term.

Singular Value Decomposition

Since D is diagonal $(N \times p)$ matrix, $D^T D$ is a diagonal $p \times p$ matrix with entries d_j^2. Then

$$D(D^T D + \lambda I)^{-1} D^T$$

is a diagonal matrix with j'th diagonal entry

$$\frac{d_j^2}{d_j^2 + \lambda}.$$

But that means we can write the ridge solution as a sum of the columns u_j of U:

$$\mathbf{X}\beta^{Ridge} = \sum_{j=1}^{p} u_j u_j^T \mathbf{y} \frac{d_j^2}{d_j^2 + \lambda}.$$

We see explicitly how the parameters shrink when $\lambda \to \infty$.

Eigenvalue Decomposition

There is still one mysterious set of terms, and those are the d_i^2's. It turns out the d_i^2's are accessing the principle components of the matrix $\mathbf{X}^T\mathbf{X}$. We will say more on this later, but for now notice that

$$\mathbf{X}^T\mathbf{X} = VD^TDV$$

gives an eigen-decomposition of $\mathbf{X}^T\mathbf{X}$.

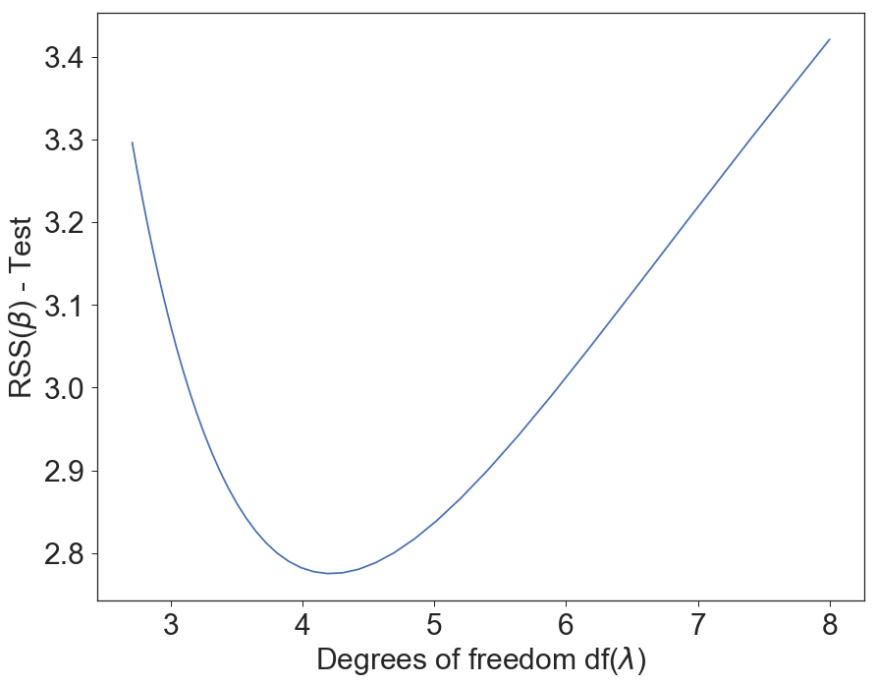
Degrees of Freedom

Finally, we give the proper interpretation of λ in terms of degrees of freedom. Recall, the number of the degrees of freedom of a classifier are given by

$$\begin{aligned} df(\hat{\beta}^{ridge}) &= \frac{1}{\sigma^2}\text{Tr}\big(\text{Cov}(\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda I)^{-1}\mathbf{X}^T\mathbf{y}, \mathbf{y})\big) \\ &= \text{Tr}\big(\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda I)^{-1}\mathbf{X}^T\big) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}\,. \end{aligned}$$

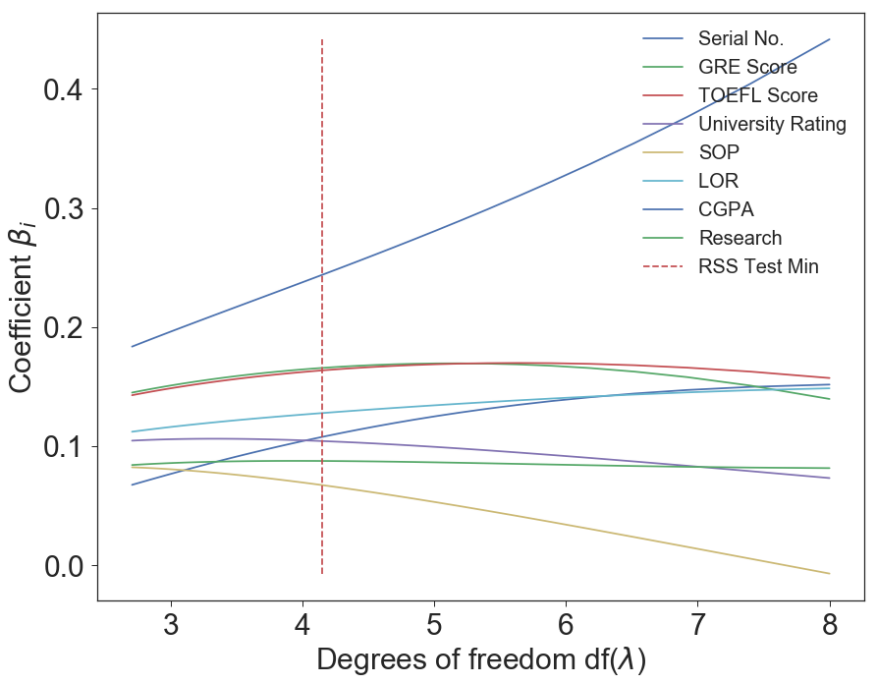
We see that for $\lambda = 0$ there are p degrees of freedom as before, but that the number monotonically decrease as λ gets large. So as we tune λ, we are lowering our degrees of freedom, possibly raising our bias, but lowering our overall test error.

Linear Models: Definition



This also allows us to properly chart the test error vs the degrees of freedom. Here we see the test error.

Linear Models: Definition



And here we see the coefficients plotted against the degrees of freedom. Notice that it has settled on slightly more degrees of freedom than best subset section did (3), although they are in the ball park.

References

References: This lecture is taken from the middle of HTF Chapter 3. For more information about linear predictors check there.

For an excellent discussion of degrees of freedom in regression models take a look at

`http://www.stat.cmu.edu/~ryantibs/advmethods/notes/df.pdf`