Data, Environment and Society: Lecture 20: Model Selection and Regularization, continued

Instructor: Duncan Callaway
GSI: Seigi Karasaki

October 30, 2018

Recap Thursday's lecture objectives

- Refine our understanding of model identification as an optimization problem
- Continue thinking about how to adjust errors to compare models with different p
- Understand what "regularization" is and why we do it
- Understand tradeoffs between subset selection, ridge and lasso

Recap Thursday's lecture objectives

- Refine our understanding of model identification as an optimization problem

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot R(\beta)$$

- Continue thinking about how to adjust errors to compare models with different p
- Understand what "regularization" is and why we do it
- Understand tradeoffs between subset selection, ridge and lasso

Recap Thursday's lecture objectives

- Refine our understanding of model identification as an optimization problem

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot R(\beta)$$

- Continue thinking about how to adjust errors to compare models with different p
  - k-fold cross validation, AIC, BIC, adjusted R^2...
- Understand what "regularization" is and why we do it

- Understand tradeoffs between subset selection, ridge and lasso

Recap Thursday's lecture objectives

- Refine our understanding of model identification as an optimization problem

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot R(\beta)$$

- Continue thinking about how to adjust errors to compare models with different p
  - k-fold cross validation, AIC, BIC, adjusted R^2...
- Understand what "regularization" is and why we do it
  - A tool for adapting optimization problems to be well behaved
  - In statistical learning, a tool to tradeoff bias and variance
  - Something that causes you to solve a different problem than the original $\rightarrow$ parameter bias
- Understand tradeoffs between subset selection, ridge and lasso

Recap Thursday's lecture objectives

- Refine our understanding of model identification as an optimization problem

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot R(\beta)$$

- Continue thinking about how to adjust errors to compare models with different p
  - k-fold cross validation, AIC, BIC, adjusted R^2...
- Understand what "regularization" is and why we do it
  - A tool for adapting optimization problems to be well behaved
  - In statistical learning, a tool to tradeoff bias and variance
  - Something that causes you to solve a different problem than the original $\rightarrow$ parameter bias
- Understand tradeoffs between subset selection, ridge and lasso
  - Speed (fastest to slowest): Ridge, Lasso, Subset
  - Subset selection and Lasso do feature selection. Ridge does not.
  - You can naturally tune prediction bias-variance with Ridge and Lasso

Today's objectives

1. Quick review of the basic mechanics of Subset selection, Ridge and Lasso.
2. Build deeper intuition on how they work and how they differ.
3. Learn how the bias-variance tradeoff gets tuned with regularization term parameters.
4. Understand the tradeoffs between these methods in more detail
5. Understand the importance of normalizing your variables.
6. Epilogue: the elastic net, a machine learning mashup.

Objective 1

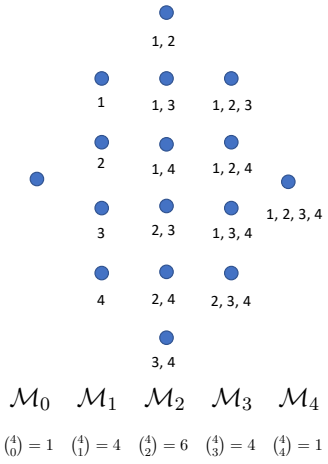
Some quick review.

Stepwise selection: Forward Selection

Forward selection: Start with $\mathcal{M}_0$. Then to choose the best model from each higher "level":

First, within each level, add one predictor at a time to the best model from the lower level. Use R^2 or other to find the best model from this set of $\mathcal{M}_{k-1}$ "$+1$"

Second, choose from your list $\{\mathcal{M}_0, \mathcal{M}_1, ...\mathcal{M}_p\}$ via cross validation or adjusted error metrics.

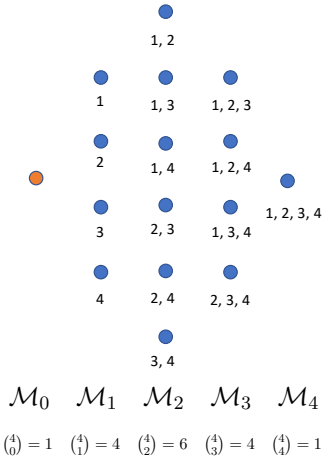


Stepwise selection: Forward Selection

Forward selection: Start with $\mathcal{M}_0$. Then to choose the best model from each higher "level":

First, within each level, add one predictor at a time to the best model from the lower level. Use R^2 or other to find the best model from this set of $\mathcal{M}_{k-1}$ "$+1$"

Second, choose from your list $\{\mathcal{M}_0, \mathcal{M}_1, ...\mathcal{M}_p\}$ via cross validation or adjusted error metrics.

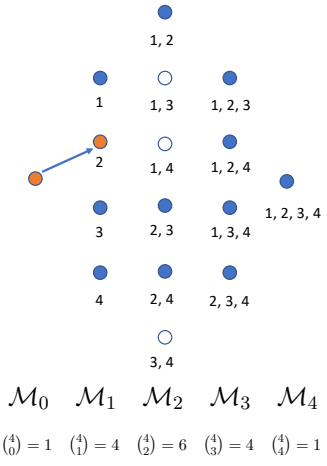


Stepwise selection: Forward Selection

Forward selection: Start with $\mathcal{M}_0$. Then to choose the best model from each higher "level":

First, within each level, add one predictor at a time to the best model from the lower level. Use R^2 or other to find the best model from this set of $\mathcal{M}_{k-1}$ "$+1$"

Second, choose from your list $\{\mathcal{M}_0, \mathcal{M}_1, ...\mathcal{M}_p\}$ via cross validation or adjusted error metrics.

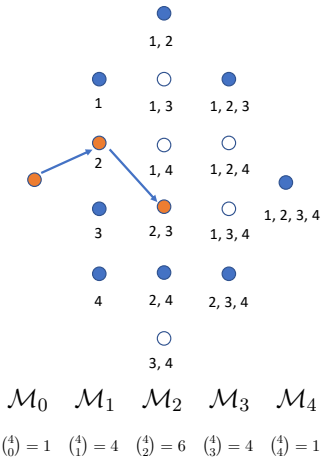


Stepwise selection: Forward Selection

Forward selection: Start with $\mathcal{M}_0$. Then to choose the best model from each higher "level":

First, within each level, add one predictor at a time to the best model from the lower level. Use R^2 or other to find the best model from this set of $\mathcal{M}_{k-1}$ "$+1$"

Second, choose from your list $\{\mathcal{M}_0, \mathcal{M}_1, ...\mathcal{M}_p\}$ via cross validation or adjusted error metrics.

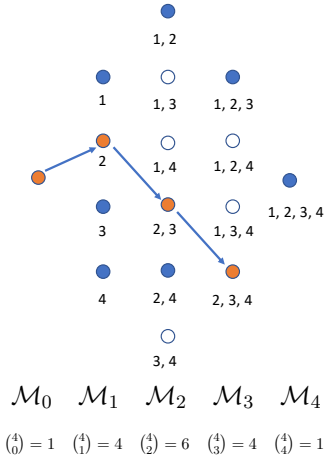


Stepwise selection: Forward Selection

Forward selection: Start with $\mathcal{M}_0$. Then to choose the best model from each higher "level":

First, within each level, add one predictor at a time to the best model from the lower level. Use R^2 or other to find the best model from this set of $\mathcal{M}_{k-1}$ "$+1$"

Second, choose from your list $\{\mathcal{M}_0, \mathcal{M}_1, ...\mathcal{M}_p\}$ via cross validation or adjusted error metrics.

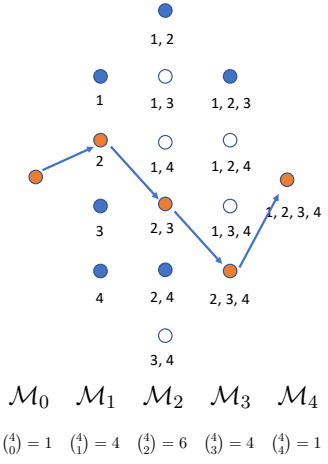


Stepwise selection: Forward Selection

Forward selection: Start with $\mathcal{M}_0$. Then to choose the best model from each higher "level":

First, within each level, add one predictor at a time to the best model from the lower level. Use R^2 or other to find the best model from this set of $\mathcal{M}_{k-1}$ "$+1$"

Second, choose from your list $\{\mathcal{M}_0, \mathcal{M}_1, ...\mathcal{M}_p\}$ via cross validation or adjusted error metrics.



A bit more recap

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot R(\beta)$$

The regularizing term, $R(\beta)$ can take a lot of forms. We looked at

$$R(\beta) = \sum_{i=1}^{K} I(\beta_i) \qquad = \|\beta\|_0 \quad \text{(subset selection)}$$

$$R(\beta) = \sum_{i=1}^{K} |\beta_i| \qquad = \|\beta\|_1 \quad \text{(lasso)}$$

$$R(\beta) = \sum_{i=1}^{K} \beta_i^2 \qquad = \|\beta\|_2^2 \quad \text{(ridge)}$$

Note, last time I referred to $\sum_{i=1}^{K} \beta^2$ as the 2-norm. It's not! ($\|\beta\|_2 = \sqrt{\sum_{k=1}^{K} \beta^2}$ is.)

Side note: "Regularization"

Regularization Refers to the process of adding a term to the objective function of a problem that

- Makes the problem "well behaved" (easier to solve)
- Solves a different problem from the one you originally wanted.

In our case, the sum of squared coefficients in Ridge makes the problem very simple to solve, but we get coefficients that are biased.

In forecasting, the reduction in variance outweighs the growth in variance. But what about inference?

Side note: "Regularization"

Regularization Refers to the process of adding a term to the objective function of a problem that

- Makes the problem "well behaved" (easier to solve)
- Solves a different problem from the one you originally wanted.

In our case, the sum of squared coefficients in Ridge makes the problem very simple to solve, but we get coefficients that are biased.

In forecasting, the reduction in variance outweighs the growth in variance. But what about inference?

Vapnik: "*Regularization theory was one of the first signs of the existence of intelligent inference*"

Why? You avoid coefficients you shouldn't have considered.

vapnik
Professor of Columbia, Fellow of NEC Labs America, machine learning, statistics, computer science
Verified email at nec-labs.com

Citation indices	All	Since 2012
Citations	198895	90510
h-index	118	77
i10-index	410	304

Objective 2

Building deeper intuition on how these methods work and how they differ.

Identifying parameters

$$\hat{\beta}_{\text{ols}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{Y}\right)$$
$$\hat{\beta}_{\text{ridge}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_k\right)^{-1}\left(\mathbf{X}^T\mathbf{Y}\right)$$
$$\hat{\beta}_{\text{lasso}} = \text{Something you need to solve numerically (with something like gradient descent)}$$

Here

- λ is a tuning parameter – it is not unique.
- $\mathbf{I}_k$ is the $k \times k$ identity matrix

Important! Ridge and Lasso produce different β estimates for different choices of λ.

Regularized problems can be converted to constrained problems

Subset selection: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} I(\beta_k) \quad \Leftrightarrow$

Regularized problems can be converted to constrained problems

Subset selection: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} I(\beta_k) \quad \Leftrightarrow \quad \begin{cases} \min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 \\ \text{subject to} \\ \sum_{k=1}^{K} I(\beta_k) \leq s \end{cases}$

Important:

- λ and s are parameters that need to be tuned.
- Increasing λ has the same effect as decreasing s. (Forces selection of fewer features.)
- λ and s are not independent.

Regularized problems can be converted to constrained problems

Subset selection: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} I(\beta_k) \quad \Leftrightarrow \begin{cases} \min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 \\ \text{subject to} \\ \sum_{k=1}^{K} I(\beta_k) \leq s \end{cases}$

Ridge: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} |\beta_k| \quad \Leftrightarrow$

Lasso: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} \beta_k^2 \quad \Leftrightarrow$

Regularized problems can be converted to constrained problems

Subset selection: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} I(\beta_k) \quad \Leftrightarrow \begin{cases} \min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 \\ \text{subject to} \\ \sum_{k=1}^{K} I(\beta_k) \leq s \end{cases}$

Ridge: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} |\beta_k| \quad \Leftrightarrow \begin{cases} \min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 \\ \text{subject to} \\ \sum_{k=1}^{K} |\beta_k| \leq s \end{cases}$

Lasso: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} \beta_k^2 \quad \Leftrightarrow$

Regularized problems can be converted to constrained problems

Subset selection: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} I(\beta_k) \quad \Leftrightarrow \begin{cases} \min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 \\ \text{subject to} \\ \sum_{k=1}^{K} I(\beta_k) \leq s \end{cases}$

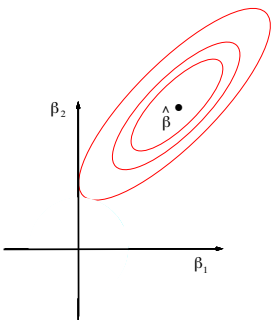
Ridge: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} |\beta_k| \quad \Leftrightarrow \begin{cases} \min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 \\ \text{subject to} \\ \sum_{k=1}^{K} |\beta_k| \leq s \end{cases}$

Lasso: $\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \sum_{k=1}^{K} \beta_k^2 \quad \Leftrightarrow \begin{cases} \min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 \\ \text{subject to} \\ \sum_{k=1}^{K} \beta_k^2 \leq s \end{cases}$

Setting intuition about constrained problems

Imagine a simple two-feature problem.

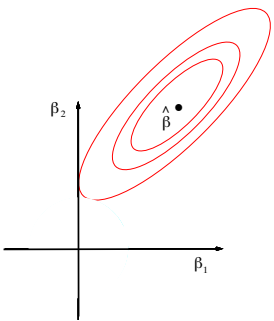
No constraints. Red lines are "constant RSS" contours



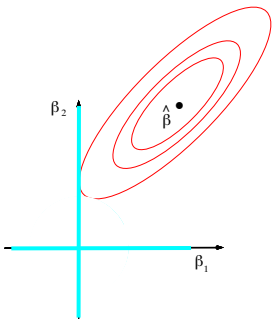
Setting intuition about constrained problems

Imagine a simple two-feature problem.

No constraints. Red lines are "constant RSS" contours



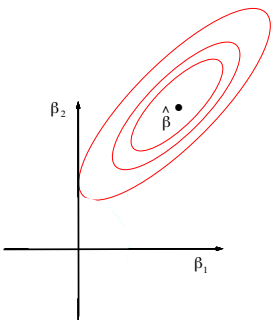
Subset selection. If $\sum_{k=1}^{K} I(\beta_k) \leq 1$, solutions on blue lines.



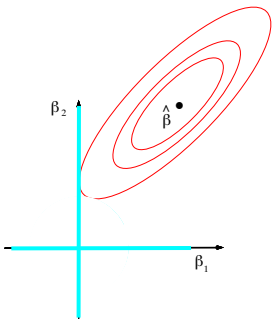
Setting intuition about constrained problems

Imagine a simple two-feature problem.

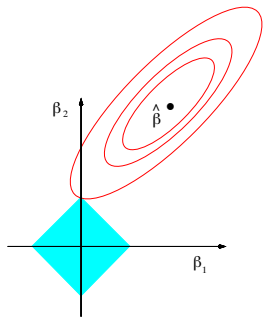
No constraints. Red lines are "constant RSS" contours



Subset selection. If $\sum_{k=1}^{K} I(\beta_k) \leq 1$, solutions on blue lines.



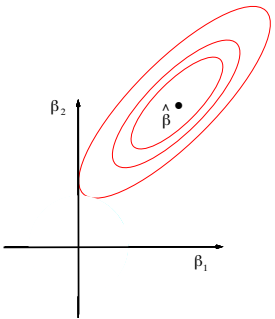
Lasso. Solutions must be in blue diamond.



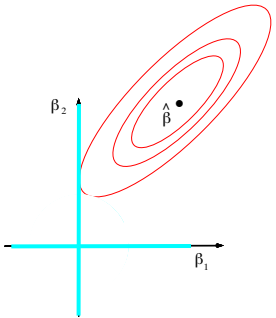
Setting intuition about constrained problems

Imagine a simple two-feature problem.

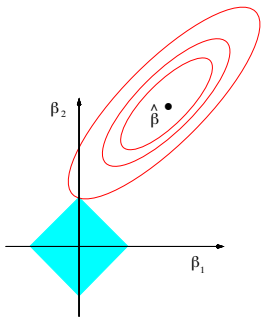
No constraints. Red lines are "constant RSS" contours



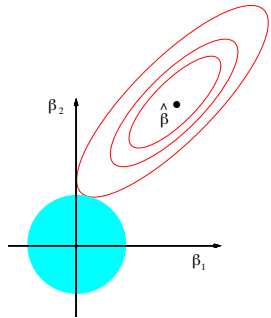
Subset selection. If $\sum_{k=1}^{K} I(\beta_k) \leq 1$, solutions on blue lines.



Lasso. Solutions must be in blue diamond.

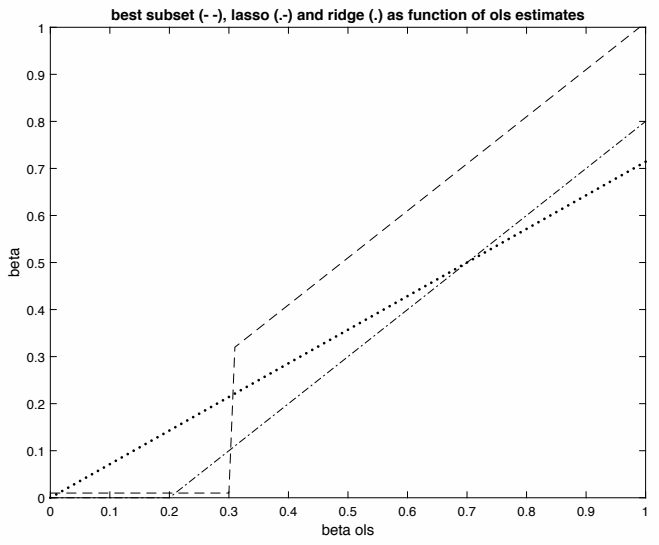


Ridge. Solutions must be in blue circle.



Figures adapted from Elements of Statistical Learning

What do the parameter estimates look like?



We often call regularization approaches "shrinkage" methods because, in general, they smoosh parameters closer to zero.

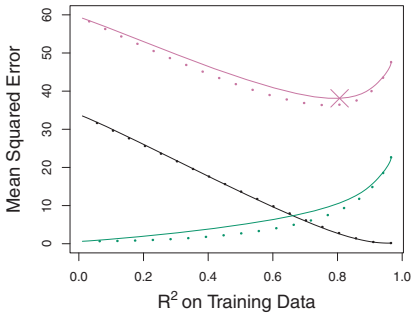
Figure taken from Jackson 2015 NBER lecture

Objective 3

Learn how the bias-variance tradeoff gets tuned with regularization term parameters.

Lasso or ridge?

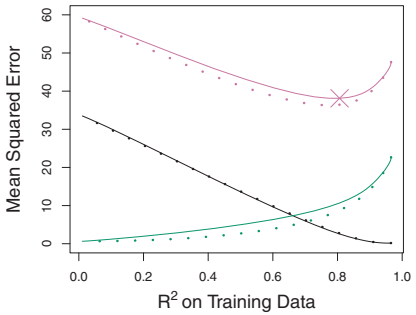
Simulated data set from $n = 50$ and $p = 45$.
The response is a function of all 45 predictors.
Figure shows test MSE on y-axis.



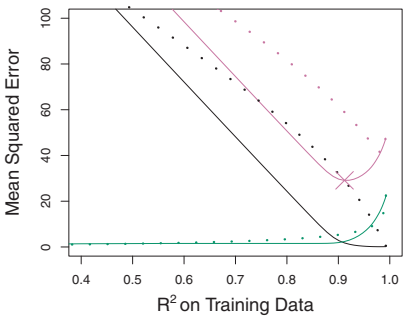
Red show total MSE. (Black and green are variance and bias.)
Dashed line – best Ridge. Solid line – best lasso.

Lasso or ridge?

Simulated data set from $n = 50$ and $p = 45$. The response is a function of all 45 predictors. Figure shows test MSE on y-axis.



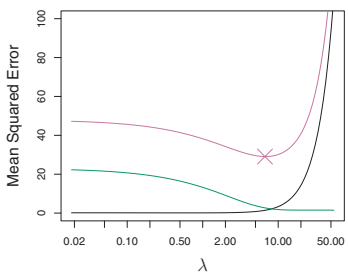
Simulated data set from $n = 50$ and $p = 45$. The response is a function of only 2 of the 45 predictors. Figure shows test MSE on y-axis.

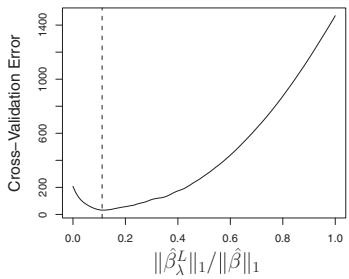


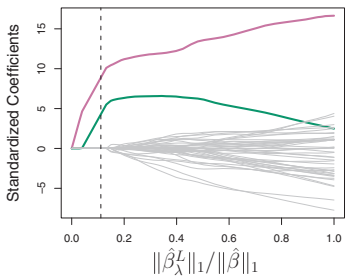
Red show total MSE. (Black and green are variance and bias.)
Dashed line – best Ridge. Solid line – best lasso.

Choosing λ

As we've seen, λ can take a range of values. Here we see the tradeoff for the 2-of-45 predictors example lasso example from the last slide.

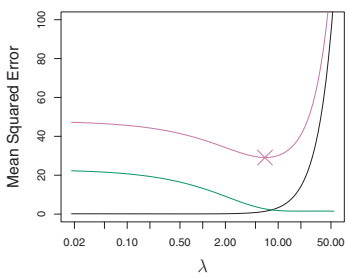


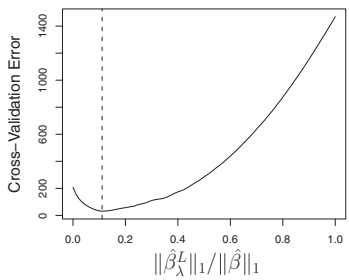


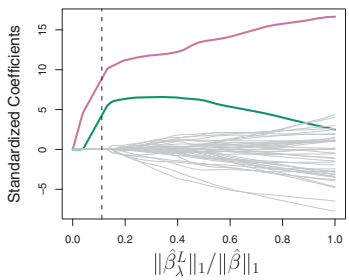


Choosing λ

As we've seen, λ can take a range of values. Here we see the tradeoff for the 2-of-45 predictors example lasso example from the last slide.







Simple λ selection strategy: use k-fold cross validation to test performance on a grid of λ values.

Objective 4

Understand the tradeoffs between these methods in more detail

Lasso or ridge?

Lasso performs better when the number of predictors truly is small, and otherwise it performs worse.

Since you can't know ahead of time if the response is a function of a few or all predictors, it makes sense to just try both.

Ridge regression advantages over OLS

Ridge regression advantages over OLS

First. Suppose some of your features are linear combinations of the others. That means you can write $x_{i,j} = Ax_{i,-j}$ for at least 1 value of j.

Then $\mathbf{X}^T\mathbf{X}$ is not "full rank" and you can't invert it. I.e., $(\mathbf{X}^T\mathbf{X})^{-1}$ doesn't exist.

e.g., $\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$

Ridge regression advantages over OLS

First. Suppose some of your features are linear combinations of the others. That means you can write $x_{i,j} = Ax_{i,-j}$ for at least 1 value of j.

Then $\mathbf{X}^T\mathbf{X}$ is not "full rank" and you can't invert it. I.e., $(\mathbf{X}^T\mathbf{X})^{-1}$ doesn't exist.

$$\text{e.g., } \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix} \quad \text{vs.} \quad \mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} = \begin{bmatrix} 1+\lambda & 2 & 3 \\ 2 & 4+\lambda & 6 \\ 3 & 6 & 9+\lambda \end{bmatrix}$$

But you *can* invert $\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_k\right)$. (Since you've added a little shift to the diagonals of the matrix, which restores linear independence)

Ridge regression advantages over OLS

First. Suppose some of your features are linear combinations of the others. That means you can write $x_{i,j} = Ax_{i,-j}$ for at least 1 value of j.

Then $\mathbf{X}^T\mathbf{X}$ is not "full rank" and you can't invert it. I.e., $(\mathbf{X}^T\mathbf{X})^{-1}$ doesn't exist.

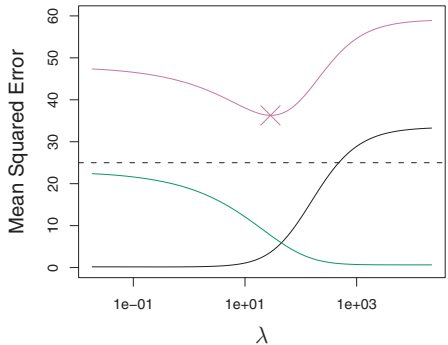
$$\text{e.g., } \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix} \quad \text{vs.} \quad \mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} = \begin{bmatrix} 1+\lambda & 2 & 3 \\ 2 & 4+\lambda & 6 \\ 3 & 6 & 9+\lambda \end{bmatrix}$$

But you *can* invert $\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_k\right)$. (Since you've added a little shift to the diagonals of the matrix, which restores linear independence)

Second. Computation! It's faster than subset selection (but solves a different problem if your objective is parameter interpretation).

Ridge regression advantages over OLS, ctd

Third. Bias-variance tradeoff! Figure from ISLR



green: variance

black: bias

red: total error

...but how can we choose λ? The short answer is k-fold cross validation.

Model selection tradeoffs

	Subset selection	Ridge	Lasso
Computing time			

Model selection tradeoffs

	Subset selection	Ridge	Lasso
Computing time	high	very low	low
Drives parameters to zero?			

Model selection tradeoffs

	Subset selection	Ridge	Lasso
Computing time	high	very low	low
Drives parameters to zero?	yes	no	yes
Parameters biased relative to "true" model?			

Model selection tradeoffs

	Subset selection	Ridge	Lasso
Computing time	high	very low	low
Drives parameters to zero?	yes	no	yes
Parameters biased relative to "true" model?	no	yes	yes
Easy to tune bias-variance?			

Model selection tradeoffs

	Subset selection	Ridge	Lasso
Computing time	high	very low	low
Drives parameters to zero?	yes	no	yes
Parameters biased relative to "true" model?	no	yes	yes
Easy to tune bias-variance?	no	yes	yes
Handles correlated features well?			

Model selection tradeoffs

	Subset selection	Ridge	Lasso
Computing time	high	very low	low
Drives parameters to zero?	yes	no	yes
Parameters biased relative to "true" model?	no	yes	yes
Easy to tune bias-variance?	no	yes	yes
Handles correlated features well?	yes	yes	no
Interpretability?			

Model selection tradeoffs

	Subset selection	Ridge	Lasso
Computing time	high	very low	low
Drives parameters to zero?	yes	no	yes
Parameters biased relative to "true" model?	no	yes	yes
Easy to tune bias-variance?	no	yes	yes
Handles correlated features well?	yes	yes	no
Interpretability?	size of parameters	not really	presence of parameters

Lasso and ridge:

+ less prediction variance than OLS, especially with many predictors
− more prediction bias than OLS

With highly correlated predictors, Lasso is unstable: indifferent between

- $\hat{\beta}_1 = 0$ and $\hat{\beta}_2 = \beta_1 + \beta_2$
- $\hat{\beta}_1 = \beta_1 + \beta_2$ and $\hat{\beta}_2 = 0$

Objective 5

Normalize your variables!

Normalizing variables

For ridge and lasso, it's important to "normalize" your variables:

$$x' = \frac{x - \bar{x}}{\sigma_x} \tag{1}$$

...and then fit the model to the normalized values.

Any guesses why?

Normalizing variables

For ridge and lasso, it's important to "normalize" your variables:

$$x' = \frac{x - \bar{x}}{\sigma_x} \tag{1}$$

...and then fit the model to the normalized values.

Any guesses why?

Reason: This way variables with large numeric values don't dominate the solution.

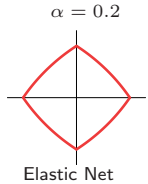
Hot topic: Elastic nets...

...are cool!

Elastic nets

- Drive parameters to zero like lasso
- Deals with correlated predictors well, like ridge (by shinking them together)
- Give you another !&@!%* parameter to tune
- Still aren't always best – good to try several shrinkage methods, not just this.

$$\lambda \sum_{k=1}^{K} \left(\alpha\beta_j^2 + (1-\alpha)|\beta_j|\right)$$



from Elements of Statistical Learning

Reading questions

- What are their main answers to the question posed in Section 1.2, "Why Find Spatial Predictions for Ozone?"
- Section 2.1: Why do the authors average their data? Could they be missing anything important by doing this?
- Section 4.1: Be prepared to describe in words the model the authors are fitting to the data. Is this a meaningful model to estimate? Why or why not?
- Section 4.2, paragraph 3 to end of Section 4.2: How does their formulation of the lasso procedure differ from what we learned in class? Can you see how the two formulations could get similar results? How does λ in the formulation we learned relate to t in the formulation presented in this paper?
- Section 4.3: Be prepared to comment on their model validation procedure. How might you improve on it? What problems are they exposed to by taking the approach they do?
- Interpretation. Think about ways in which this approach might be practically applied to improving our understanding of air quality. Can it be adapted to address equity issues in situations where monitoring stations are sparse?