Section 9: Model regularization

STA 35C – Statistical Data Science III

Instructor: Akira Horiguchi

Fall Quarter 2025 (Sep 24 – Dec 12)
MWF, 12:10 PM – 1:00 PM, Olson 158
University of California, Davis

Overview

Based on Chapter 6 of ISL book James et al. (2021).

Section 7: Model selection and regularization

1. What goes wrong in high dimensions?

2. Shrinkage methods
   - Ridge regression
   - LASSO

What goes wrong in high dimensions?

When might $p \gg n$?

n is often limited due to cost, sample availability, or other considerations. Examples:

- Rather than predicting blood pressure on the basis of just age, sex, and BMI, one might also collect measurements for half a million single nucleotide polymorphisms (SNPs; these are individual DNA mutations that are relatively common in the population) for inclusion in the predictive model. Then $n \approx 200$ and $p \approx 500,000$.
- A marketing analyst interested in understanding online shopping patterns could treat as features all of the search terms entered by users of a search engine. This is sometimes known as the "bag-of-words" model. The same researcher might have access to the search histories of only a few hundred or a few thousand search engine users who have consented to share their information with the researcher. For a given user, each of the p search terms is scored present (0) or absent (1), creating a large binary feature vector. Then $n \approx 1,000$ and p is much larger.

Motivation

Linear regression models the relationship between the response and predictors as:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon. \tag{1}$$

Recall: the OLS estimator $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$ is the vector of parameters $(\beta_0, \beta_1, \ldots, \beta_p)^\top$ that minimizes the residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad \text{where} \quad \hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{2}$$

- If the linear relationship is reasonable, the OLS estimates have low bias.
- If $n \gg p$, the OLS estimates tend to also have a low variance, and hence perform well on test observations.
- *If n is not much larger than p*, there can be a lot of variability in the least squares fit, resulting in overfitting/poor predictions. *If $p > n$*, there is no longer a unique OLS estimate, and the variance is infinite.
- *Try to reduce p?* Some predictors barely influence the response (if at all).
- *Feature selection*/*variable selection*, i.e. excluding irrelevant predictors from the linear model can lead to a more easily interpretable model.
- Least squares rarely ever yields any coefficient estimates that are exactly zero.

Here we explore fitting procedures other than least squares fit.

Alternative: Best subset selection

$\{a,b,c\}$ ϕ $\{b\}$ $\{c\}$ $\{bc\}$ $\{a\}$ $\{ab\}$ $\{ac\}$ $\{abc\}$

Given a positive integer $k \leq p$, what regression-coefficient estimators minimize the RSS *if at most k of the estimators are allowed to be non-zero?*

- *Subset selection:* identifies a subset of all p relevant predictors, then fits a model using least squares on the subset of predictors.
- For a long time, this problem was considered to be computationally intractable. How many possible subsets are there? 2^p $2^{10} = 1024$ $2^{20} \approx 1$ million
- Can approximate the solution using stepwise selection. (We did this in STA 35B.) $\approx p^2$ $10^2 = 100$ $20^2 = 400$
- 2016 paper[1]: "we demonstrate that our approach solves problems with n in the 1000s and p in the 100s in minutes to provable optimality"
- This is all we will say about this in STA 35C.

[1] https://projecteuclid.org/journals/annals-of-statistics/volume-44/issue-2/Best-subset-selection-via-a-modern-optimization-lens/10.1214/15-AOS1388.pdf

Shrinkage methods

Alternative: Shrinkage

Alternative: we can fit a model containing all p predictors using a technique that *constrains* or *regularizes* the coefficient estimates.

- I.e., a technique that shrinks the coefficient estimates towards zero (relative to the least squares estimate).
- This approach can also significantly reduce the variance of the coefficient estimates.

Two best-known techniques: *ridge regression* and *lasso (or LASSO)*.

Scaling predictor values

The OLS estimates are *scale equivariant*: multiplying values of predictors X_j by a constant c leads to scaling of the least squares coefficient estimates by a factor $\frac{1}{c}$.

- The ridge regression or LASSO coefficient estimates, however, can substantially change when multiplying predictors with a constant.
- Here it is better to work with predictor values which are all on the same scale. This can be achieved by *standardizing* the observations x_{ij} of predictors X_j:

$$\tilde{x}_{ij} = \frac{x_{ij}}{s_j}\,, \tag{3}$$

where $s_j := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}$ is the standard deviation of the jth predictor, and where $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ is the sample mean of the observations of the jth predictor.

Remainder of slide deck will assume that predictors are standardized.

Shrinkage methods

Ridge regression

Ridge regression – Idea

For a user-chosen value $\lambda \geq 0$, the *ridge regression* coefficient estimates are defined as

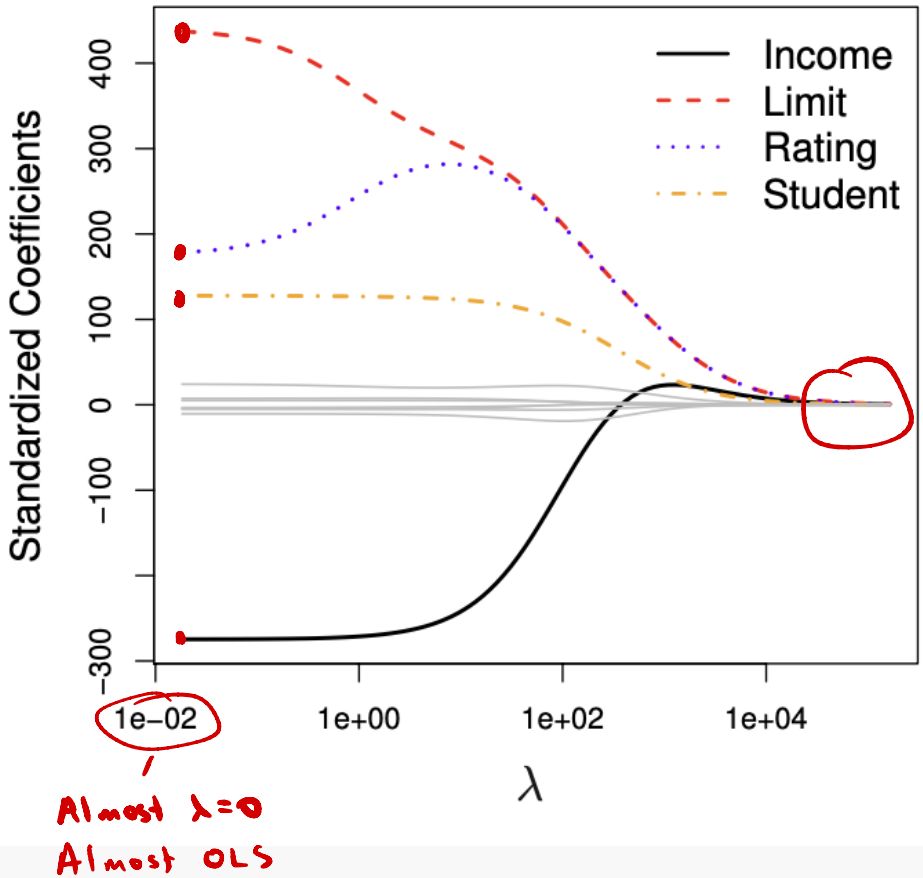
$$\hat{\beta}^R_{0,\lambda}, \hat{\beta}^R_{1,\lambda}, \ldots, \hat{\beta}^R_{p,\lambda} := \underset{\beta_0,\beta_1,\ldots,\beta_p}{\arg\min} \left\{ RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{4}$$

- Each value of λ results in a different set of ridge-regression coefficient estimates.

What does adding the *shrinkage penalty* $\lambda \sum_{j=1}^{p} \beta_j^2$ do?

- It is small if $\beta_1, \ldots, \beta_p$ are close to zero, and so the ridge regression estimates are shrunk toward zero compared to the OLS estimates.
- The tuning parameter λ controls the impact of the shrinkage penalty. The larger the λ, the stronger the shrinkage.
  - If $\lambda = 0$, then (4) is RSS, and we get back the OLS estimates (no shrinkage).
  - As $\lambda \to \infty$, RSS's impact on (4) becomes increasingly smaller. Thus minimizing (4) requires shrinking $\beta_1, \ldots, \beta_p$ increasingly strongly.
- The shrinkage is only applied to $\beta_1, \ldots, \beta_p$, but not to β_0, because we want to shrink the estimated association of each predictor X_j with the response Y.

Ridge regression – Illustration



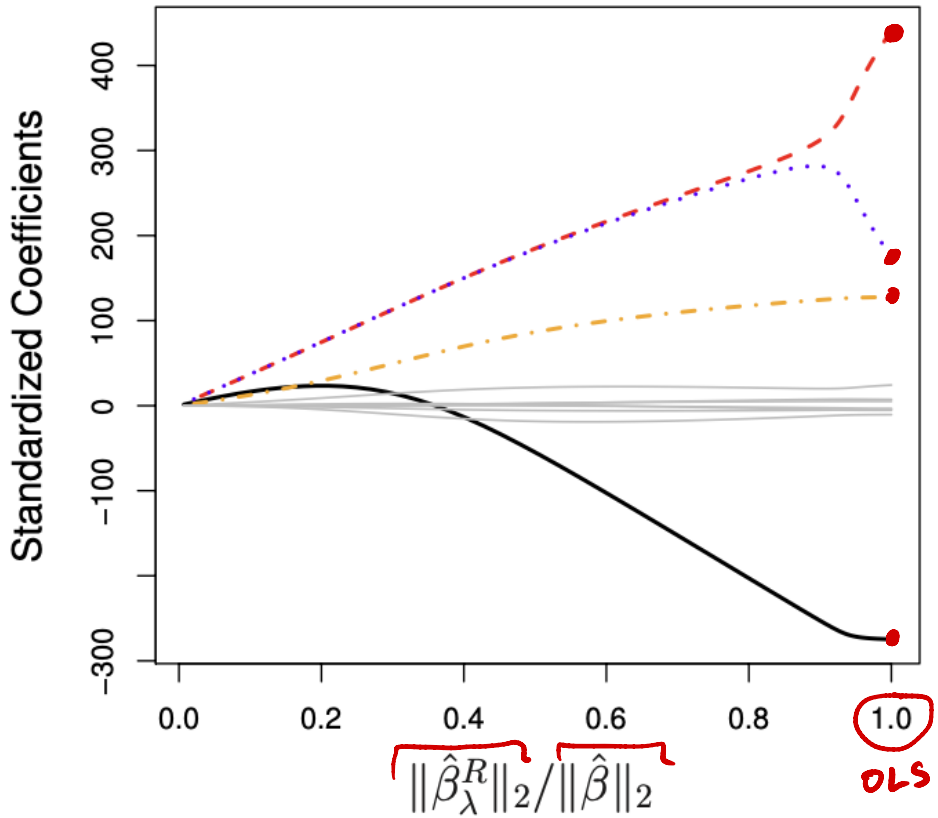


Figure 1: From James et al. (2021). Ridge regression coefficients (performed on standardized predictors) for the Credit data set in R.

Ridge regression – Advantages over least squares

1. For least squares (i.e., ridge regression with $\lambda = 0$), the variance is large but there is no bias. As λ increases, the flexibility of the ridge regression fit decreases, which reduces variance but increases bias.

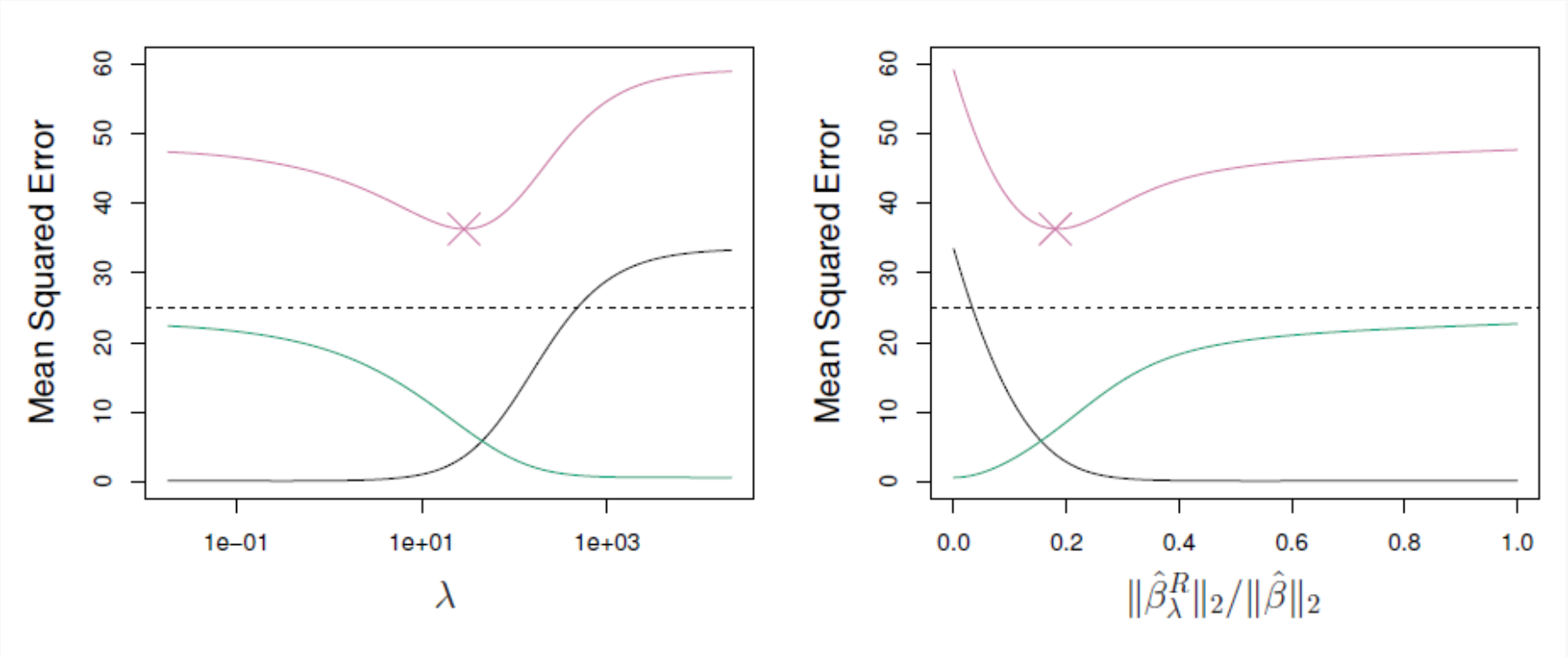


Figure 2: From James et al. (2021). Squared bias (in black), variance (in green), and test mean squared error (in purple) for the ridge regression predictions on a simulated data set. Horizontal dashed lines indicate the minimum possible MSE. Purple crosses indicate the ridge regression models for which the MSE is smallest.

Ridge regression works best when the OLS estimates have high variance.

Ridge regression – Advantages over least squares

2. Ridge regression has also a computational advantage over best subset selection, which requires searching through 2^p models.

End of 11/3 lecture

Shrinkage methods

LASSO

LASSO

The ridge regression penalty will shrink all of the cofficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$).

- This property is fine for prediction accuracy, but does not simplify model interpretation (an issue when the number of variables p is large).
- The *LASSO* coefficient estimates $\hat{\beta}_\lambda^L$ are the values that minimize

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \,. \tag{5}$$

- Instead of minimizing the sum of RSS and an ℓ^2 penalty term as in ridge regression, LASSO minimizes the sum of RSS and an ℓ^1 penalty term $\lambda \|\beta\|_1$.
- LASSO also shrinks the coefficients toward zero depending on λ, but *usually forces some of the coefficients to be exactly zero.*
- Thus LASSO performs *variable selection*, and models generated from LASSO are usually easier to interpret than those produced by ridge regression.
- We say a model is *sparse* if many of the coefficients are exactly zero.

LASSO regression – Illustration

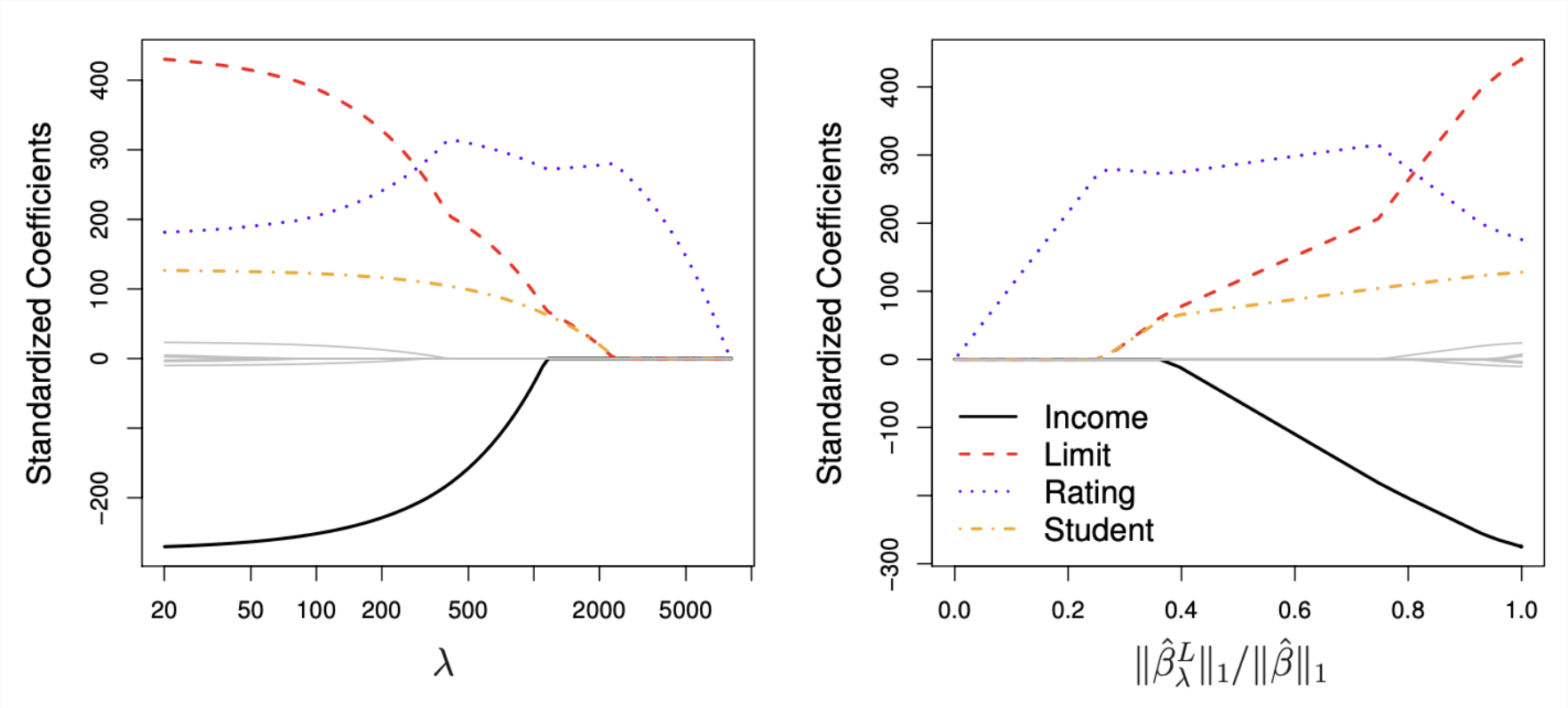


Figure 3: LASSO coefficients (performed on standardized predictors) for the Credit data set in R.

Alternate formulations

- LASSO, ridge regression, and subset selection solve the following minimization problems for some $s \geq 0$:

$$\text{LASSO:} \quad \min_{\beta} RSS \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s, \tag{6}$$

$$\text{ridge regression:} \quad \min_{\beta} RSS \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s, \tag{7}$$

$$\text{subset selection:} \quad \min_{\beta} RSS \quad \text{subject to} \quad \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}} \leq s. \tag{8}$$

- Hence, we can consider ridge regression and LASSO as computationally feasible alternatives to best subset selection.
- Also helps explain why LASSO often forces some predictors to be exactly zero, and why ridge regression does not.

Graphical comparison of ridge regression and LASSO

Why does LASSO often force some of the predictors to be exactly zero?

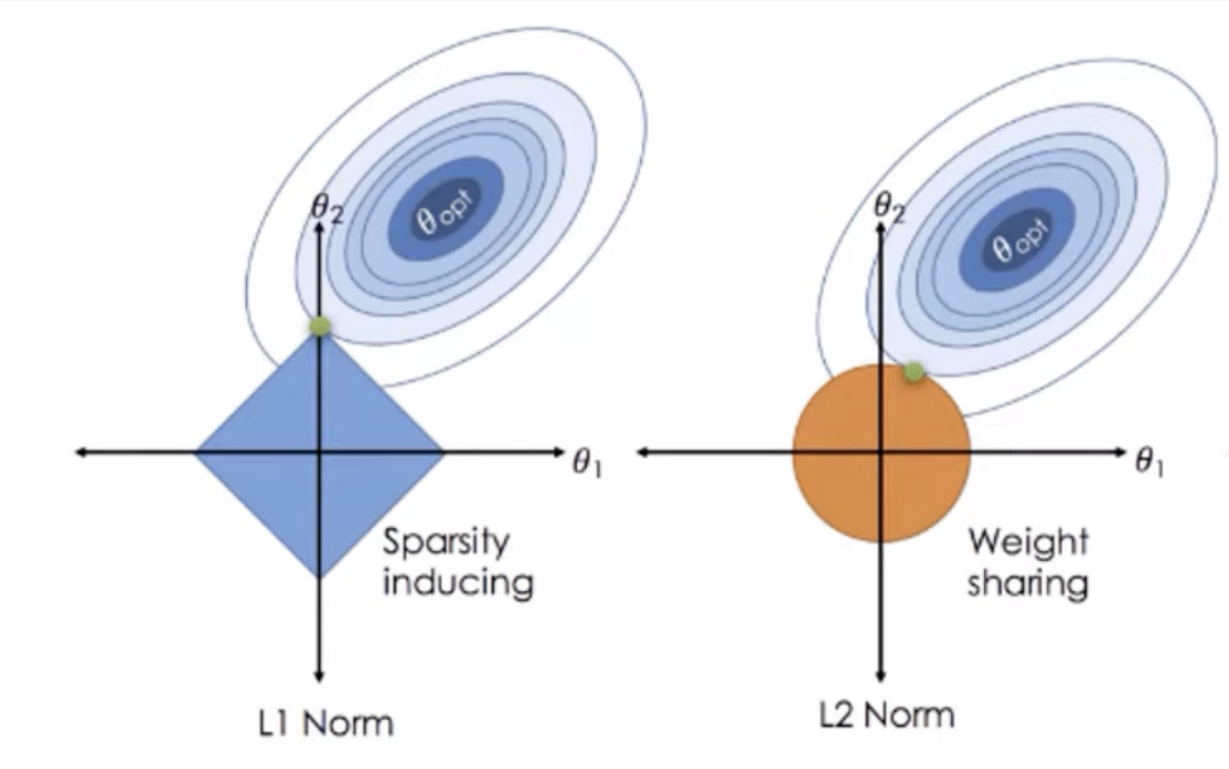


Figure 4: Here $p = 2$. The OLS solution is denoted as β_{opt}; ellipses are the contours of the RSS.
Left: Blue diamond is the constraint region $|\beta_1| + |\beta_2| \leq s$ for LASSO.
Right: Orange circle is the constraint region $\beta_1^2 + \beta_2^2 \leq s$ for ridge regression.
Source: `https://www.youtube.com/watch?app=desktop&v=iJE2fZcNPlA`

Performance of LASSO and ridge regression

- Typically, LASSO produces simpler and more interpretable models than does ridge regression, as only a subset of the predictors are involved.
- Regarding prediction accuracy, neither method universally dominates the other. Ridge regression might be better if e.g. the regression function truly depends on all p predictors, or vice versa if e.g. the regression function truly depends on only a small subset of the p predictors.
- Cross-validation can be performed to determine which approach is better on a particular data set.

Example: LASSO > ridge regression

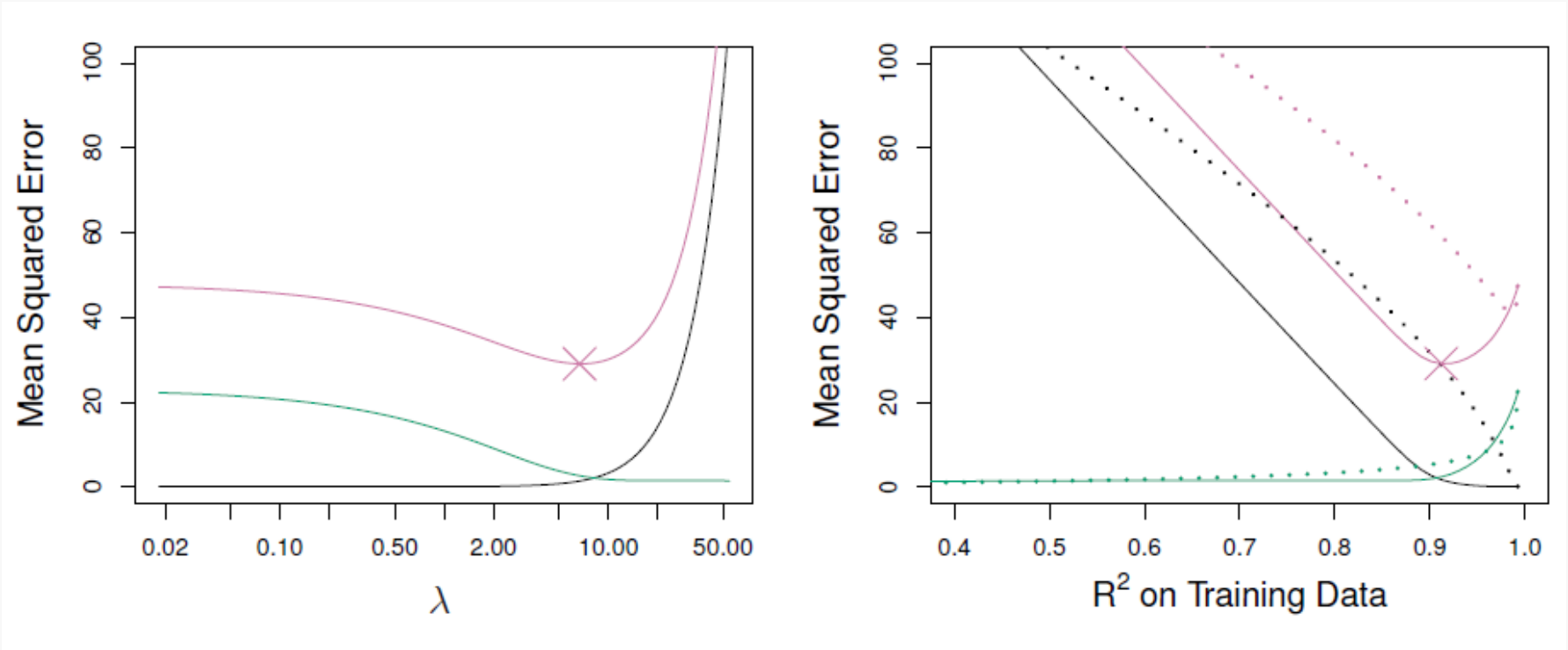


Figure 5: From James et al. (2021). Left: Plots of squared bias (in black), variance (in green), and test MSE (in purple) for LASSO on a simulated data set. Right: Comparison of squared bias, variance, and test MSE between LASSO (solid) and ridge (dotted). Both are plotted against their R^2 on the training data. The crosses in both plots indicate the LASSO model for which the MSE is smallest. (Plotting against R^2 can be used to compare models with different types of regularization.)

Example: Ridge regression > LASSO

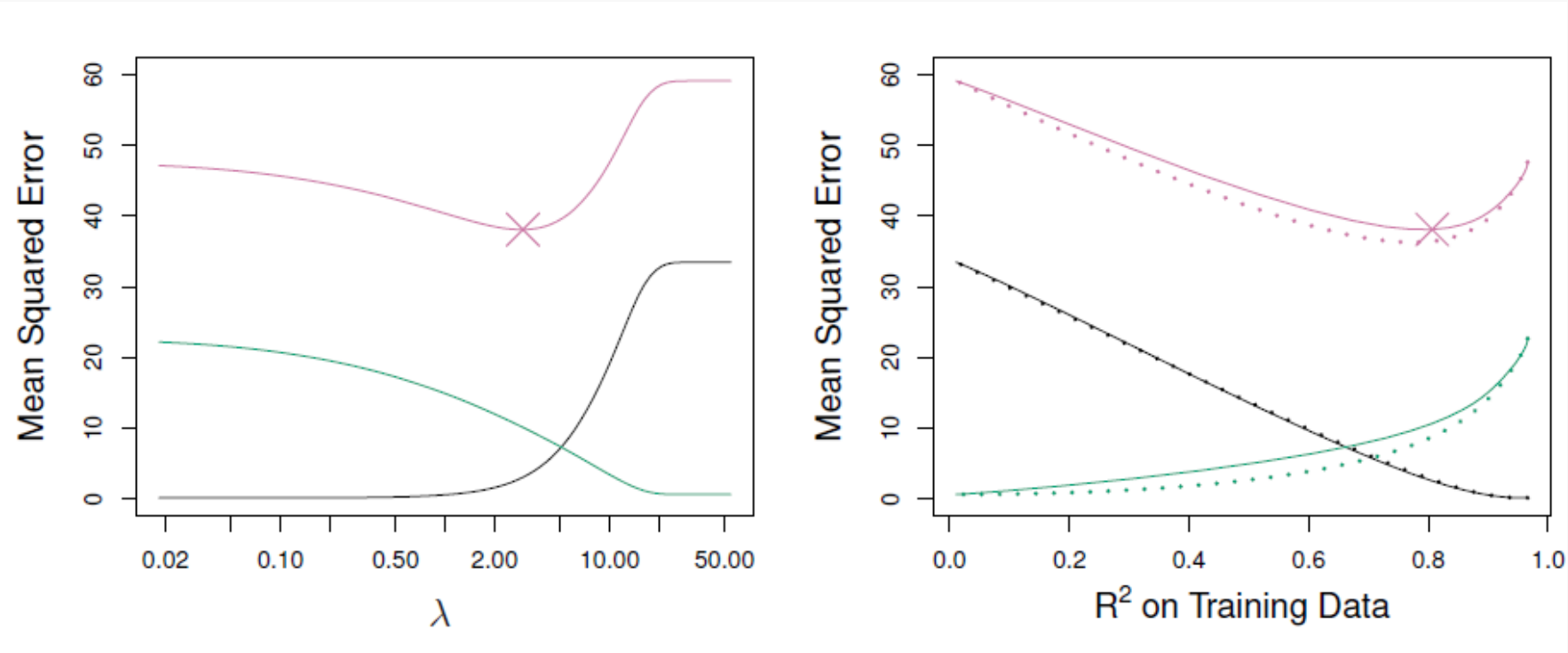


Figure 6: From James et al. (2021). Left: Plots of squared bias (in black), variance (in green), and test MSE (in purple) for LASSO on a simulated data set. Right: Comparison of squared bias, variance, and test MSE between LASSO (solid) and ridge (dotted). Both are plotted against their R^2 on the training data. The crosses in both plots indicate the LASSO model for which the MSE is smallest. (Plotting against R^2 can be used to compare models with different types of regularization.)

Selecting the tuning parameter

- Both LASSO and ridge regression require a parameter $\lambda \geq 0$, or equivalently a constraint s as described before, but how to choose value?
- Cross-validation can be used to tackle the problem of selecting the optimal tuning parameter: We choose a grid of λ values (as fine as possible), compute the CV error for each value of λ, and then select λ for which the CV error is the smallest.