STA 35C: Statistical Data Science III

Lecture 17: Regularization (cont'd) & Multiple Testing

Dogyoon Song

Fall 2026, UC Davis

Announcements

Midterm 2: Mon, Nov 9, 9:00–9:50 AM (in class)

- Arrive early: The exam ends at 9:50 AM sharp.
- One hand-written, letter-sized (8.5×11 inches), double-sided sheet of formulas/notes.
- A simple non-graphing scientific calculator is allowed; no other reference materials or electronic devices.
- **SDC accommodations:** Confirm scheduling with AES online.

Preparation and support

- Primary coverage: Lectures 12–19 (including this Friday)
- Core concepts from earlier may be assumed, especially regression and model assessment
- Check Canvas and the course webpage for practice materials.
- Office hours this week:
  - Instructor: Wed, 1–3 pm
  - TA: Thu, 12–1 pm

Agenda

- **Regularization: ridge vs. lasso**
  - Recap: Regularization, ℓ_2 penalty (ridge regression), and selecting λ
  - Complete Lecture 16: lasso coefficient paths, selecting λ, and an R example
  - Geometric intuition and comparison between ridge vs. lasso
- **Multiple hypothesis testing**: Motivation
  - Why ordinary single-hypothesis testing may fail in large-scale settings
  - Preview: Type-I error inflation and how to control it
    - Why per-test error control does not control errors across many tests
  - Preview: family-wise error and false discovery rate

Recap: Why regularization?

Challenges: Least squares estimates can

- be unstable when predictors are highly correlated or the data are noisy;
- be non-unique when $p > n$ or the design matrix is rank-deficient;
- use all predictors even when the true relationship is sparse.

Regularization can stabilize estimation by adding a penalty term: with $\lambda \geq 0$,

$$\hat{\beta}_\lambda \in \arg\min_{(\beta_0,\beta_1,\ldots,\beta_p)} \left\{ \underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{\text{RSS}} + \lambda \underbrace{R(\beta_1, \ldots, \beta_p)}_{\text{penalty}} \right\}$$

- The penalty shrinks coefficients, trading some bias for lower variance

Two popular choices:

- **Ridge**: $R(\beta_1, \ldots, \beta_p) = \sum_{j=1}^{p} \beta_j^2$
- **Lasso**: $R(\beta_1, \ldots, \beta_p) = \sum_{j=1}^{p} |\beta_j|$

Recap of ridge: Regularization reduces variance via shrinkage

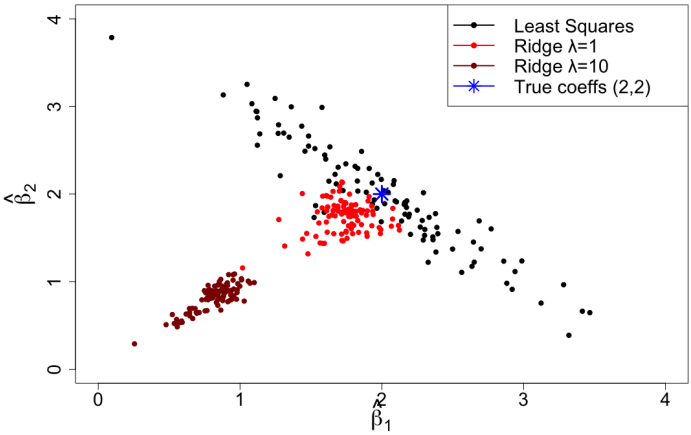


Figure: Scatter plots of 100 least squares estimates (**black**) vs. ridge estimates for $\lambda = 1$ in **red** and $\lambda = 10$ in **dark red**. As λ grows, the estimates cluster more tightly (lower variance) but shift away from the true value (blue star), indicating increased bias.

Recap of Lasso: Regularization reduces variance, but...

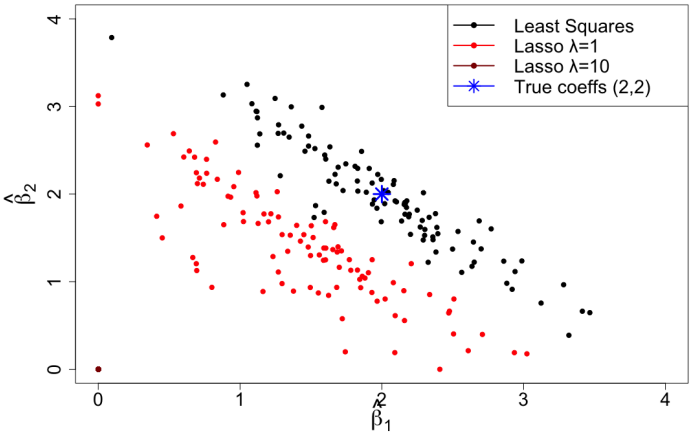


Figure: Scatter plots of 100 least squares estimates (**black**) vs. lasso estimates for $\lambda = 1$ in **red** and $\lambda = 10$ in **dark red**. Lasso can aggressively shrink or zero-out coefficients, but the variance reduction is less uniform than ridge. The shift from the true (blue star) may or may not be worth it.

Recap of Lasso: Regularization enables variable selection

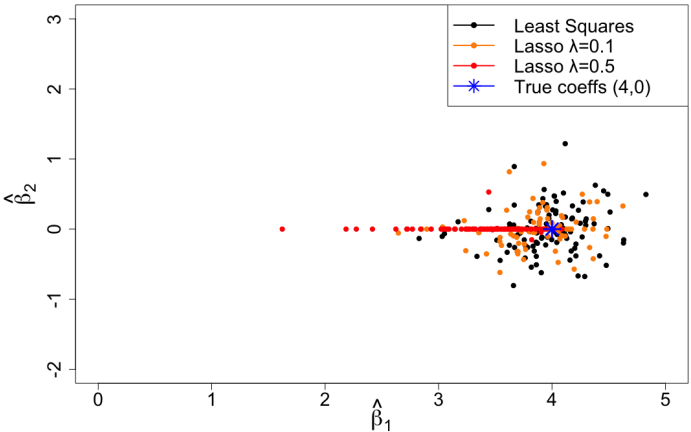


Figure: Scatter plots of 100 least squares estimates (**black**) vs. lasso estimates for $\lambda = 0.1$ in **orange** and $\lambda = 0.5$ in **red**. Here, the true $\beta_2 = 0$ (blue star), and lasso can correctly select the significant variable (X_1), while suppressing noise and driving estimates to zero for X_2, capturing the "sparse" true associations.

The lasso: `Credit` dataset example

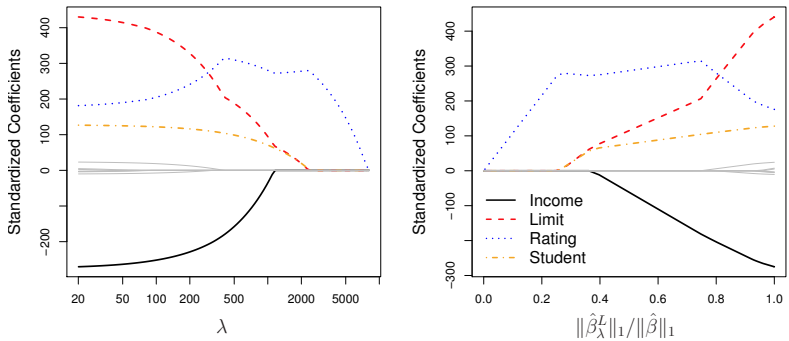


Figure: Standardized lasso coefficients for `Credit`, plotted vs. λ and the shrinkage level $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$ [JWHT21, Figure 6.6].

- Lasso can force some coefficients to zero as λ increases
- Achieves *variable selection* directly (predictors with $\hat{\beta}^L_j = 0$ are excluded)

The lasso: Selecting λ

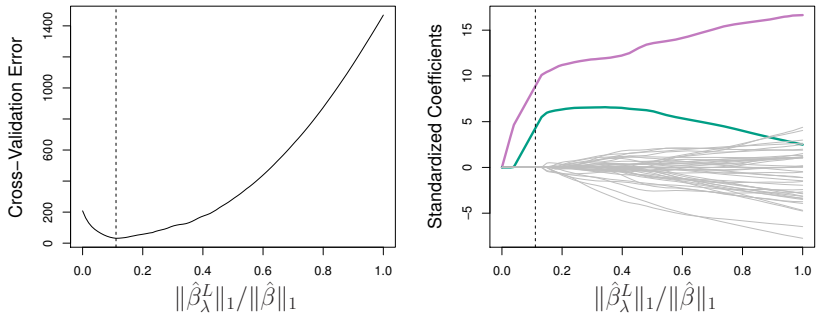


Figure: **Left:** Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data. **Right:** The corresponding lasso coefficient estimates, with the two *signal* variables in color and the *noise* variables in gray. The vertical dashed line indicates the fit that minimizes the cross-validation error [JWHT21, Figure 6.13].

- The lasso cleanly separates two *signal* variables from *noise* variables
- In contrast, standard least squares (far right, with $\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1 = 1$) only identifies the purple variable without discarding the noise predictors

The lasso: Example with a toy dataset (R script)

- **Setup:**
  - $n = 5$, $p = 2$
  - Compare lasso vs. OLS at $\lambda = 1$
- **Objective:** See how lasso can shrink coefficients to zero

```
# Toy data: 5 obs, 2 predictors
df <- data.frame(
  x1 = c(1,2,3,4,5),
  x2 = c(2,1,3,1,2),
  y  = c(2,2.5,6,4,6.5)
)
```

```
# OLS fit
ols_fit <- lm(y ~ x1 + x2, data=df)
cat("OLS Coeffs:\n", coef(ols_fit), "\n")

# Run install.packages("glmnet") once if needed
library(glmnet)

# Prepare X, y
X <- as.matrix(df[, c("x1","x2")])
y <- df$y

# Lasso with alpha=1, lambda=1
lasso_fit <- glmnet(X, y,
                    alpha=1,
                    lambda=1,
                    intercept=TRUE,
                    standardize=TRUE)
cat("Lasso Coeffs (lambda=1):\n",
    as.matrix(coef(lasso_fit)),"\n")
```

Pop-up quiz: Lasso

Suppose a lasso model is fit for several values of λ.

Question: What generally happens as λ increases?

A) The penalty weakens, so coefficients move closer to least squares.

B) Coefficients are increasingly shrunk, and some may become exactly zero.

C) Training RSS must decrease because the model becomes more flexible.

D) The intercept is penalized more strongly than the other coefficients.

Answer: B. Increasing λ strengthens the ℓ_1 penalty; lasso shrinks coefficients and can set some exactly to zero, enabling variable selection.

(Optional[1]) Alternative formulation: Constrained form

Ridge and lasso can be expressed as equivalent *constrained* optimization problems:

$$\text{(Ridge)} \quad \text{minimize}_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s_\lambda$$

$$\text{(Lasso)} \quad \text{minimize}_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s'_\lambda$$

- For each $\lambda \geq 0$, there exist corresponding s_λ, s'_λ such that solving the above problems yield the same ridge/lasso regression coefficient estimates
- Geometrically: feasible region is an ℓ_2-ball for ridge or ℓ_1-ball for lasso

[1] That is, it is good to know for intuition, but its mathematical details are beyond the scope of STA 35C

The lasso prefers “spiky” solutions

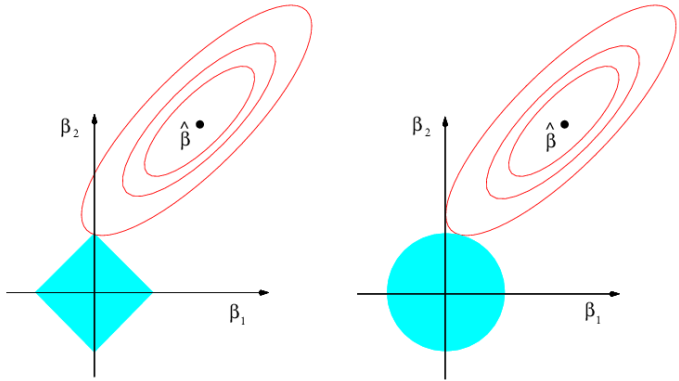


Figure: Contours of the RSS (**red** ellipses) and the feasible sets (**cyan** areas). **Left:** For lasso, the constraint $\|\beta\|_1 \leq s$ (a diamond shape) can yield corner solutions having exact zeros. **Left:** For ridge, the constraint $\|\beta\|_2^2 \leq s'$ is round, so typically yielding no exact zeros [JWHT21, Figure 6.7].

Comparison of ridge vs. lasso

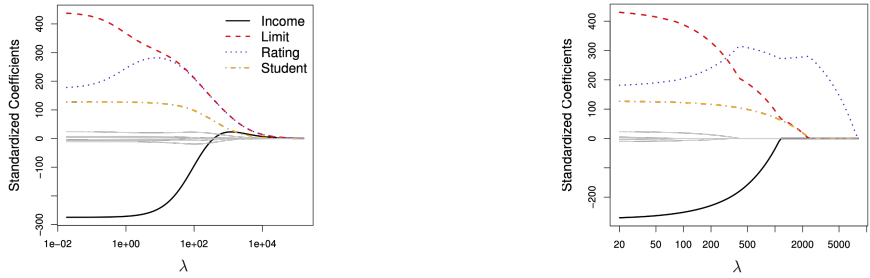


Figure: Standardized ridge (left) and lasso (right) coefficients on `Credit` dataset, plotted vs. λ [JWHT21, excerpted from Figures 6.4 & 6.6].

Ridge:

- More stable under collinearity
- Typically no exact zeros
- Often simpler closed-form solution

Lasso:

- Possibly less stable under correlated predictors
- Produces zero coefficients (variable selection)
- More interpretable if many X_j are irrelevant

Regularization: Summary

- **Why regularization?**
  - Remedy high variance or ill-posedness, especially when $p \approx n$ or $p > n$
  - Potentially yield simpler, more interpretable models (esp. lasso)
- **How?** Add a penalty
  - Ridge: $\sum_{j=1}^{p} \beta_j^2$ shrinks all β_j stably, rarely yielding exact zeros
  - Lasso: $\sum_{j=1}^{p} |\beta_j|$ can drive some β_j to 0, enabling variable selection
  - Tuning parameter λ typically selected via cross-validation
- **Ridge vs. Lasso**:
  - Ridge is stable under collinearity and has simpler closed-form solutions
  - Lasso can yield sparse solutions (some $\beta_j = 0$)
  - Neither strictly dominates: test performance depends on the data
    $\rightarrow$ usually do cross-validation to choose

Pop-up quiz: Regularization

Question: Which statement is *false* regarding ridge and lasso?

A) Ridge often works well when many predictors have modest effects or are highly correlated.

B) Lasso can set some coefficients exactly to zero, giving built-in variable selection.

C) Once λ is chosen by cross-validation, ridge will always outperform lasso in test MSE.

D) Both ridge and lasso use a tuning parameter λ that controls the amount of shrinkage.

Answer: C. Neither method uniformly dominates; after tuning, the better test performance depends on the data-generating structure.

Multiple hypothesis testing: Motivation

Recall single-hypothesis testing:

- For each predictor X_j, test $H_0 : \beta_j = 0$
- Reject H_0 if $p < \alpha$ (e.g., $\alpha = 0.05$); Type I error rate $= \alpha$ for *one* test
  - **Type I** (False positive): Null is true, but we reject
  - **Type II** (False negative): Null is false, but we fail to reject

Modern data analysis often tests **many variables** (or features) simultaneously

- We want to identify which predictors are "significant" among many candidates

Examples:

- Testing thousands of genes/biomarkers for disease association
- Testing many (possibly high-dimensional $p > n$) predictors for stock price forecasting

Problem: Merely applying ordinary tests to each predictor can yield many false positives

$\rightarrow$ **Type-1 error inflation!**

Multiple hypothesis testing: Illustration

Coin-flip analogy:

- Testing *fairness* of a coin: $H_0 : p = 0.5$
- Suppose 1,024 fair coins are each flipped 10 times, with all tosses independent
- For any one coin,

$$P(\text{10 heads}) = \left(\frac{1}{2}\right)^{10} = \frac{1}{1024}.$$

- For a prespecified coin, 10 heads gives the exact two-sided p-value $2/1024 \approx 0.00195$ $\rightarrow$ **likely a biased coin**
- However, seeing one coin with 10 heads (across 1024 tried) is **NOT surprising!**
  $\because$ across 1,024 coins, the chance of at least one all-heads is $1 - (1 - 1/1024)^{1024} \approx 0.632$

Key points:

- With many tests, extreme results can happen just by chance
- We must account for that when claiming "significance"

Multiple hypothesis testing: Challenges

Setting:

- Suppose we have m predictors to test simultaneously
- Suppose each test has a per-hypothesis Type I error rate $\alpha > 0$

Problem:

- With m tests, we have m chances for false positives
- If all m nulls are true and tests are independent,

$$P(\text{at least one false rejection}) = 1 - (1 - \alpha)^m,$$

which can be large as m grows

- For $m = 20$ and $\alpha = 0.05$, $P(\text{at least one false positive}) = 1 - 0.95^{20} \approx 0.64$, and the expected number of false positives is $20 \times 0.05 = 1$

How to address this?

- Requiring $p \leq 0.05$ for each *does not* guarantee a $\leq 5\%$ chance of *any* false positive
- We need **multiple-comparison corrections** (next Lecture)
  - *Family-Wise Error Rate (FWER)* ensures probability of *any* false positive is $\leq \alpha$
  - *False Discovery Rate (FDR)* limits the *proportion* of false positives among all rejections

Pop-up quiz: Multiple testing

Suppose we test $m = 20$ independent null hypotheses at level $\alpha = 0.05$, and suppose all null hypotheses are actually true.

Question: Which statement is most accurate?

A) The probability of at least one false rejection is exactly 0.05.

B) The expected number of false rejections is $20 \cdot 0.05 = 1$.

C) The probability of at least one false rejection decreases as m increases.

D) Multiple testing is not a concern if each individual test uses $\alpha = 0.05$.

Answer: B. If all nulls are true, each test has false-rejection probability 0.05, so the expected number of false positives is $m\alpha = 20 \times 0.05 = 1$.

Wrap-up & next steps

Regularization:

- Ridge (ℓ_2 penalty) is stable under correlated predictors
- Lasso (ℓ_1 penalty) can set some coefficients exactly to zero (variable selection)
- Typically pick λ via cross-validation

	Ridge	**Lasso**
Penalty	$\sum_{j=1}^{p} \beta_j^2$	$\sum_{j=1}^{p} \lvert\beta_j\rvert$
Effect	shrinks coefficients	shrinks and can set to zero
Variable selection	No	Yes
Often useful when	many small/moderate effects	only some predictors matter

Multiple hypothesis testing:

- Per-test error control does not control the chance of any false positive
- We need corrections for controlling false positives $\rightarrow$ FWER and FDR (next lecture)

Suggested reading: [JWHT21, Ch. 6.2 and 13.1–13.2]

Next lecture: Multiple testing: FWER and FDR control

References


Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.
Springer, New York, NY, 2nd edition, 2021.