ECON426 Applied Econometrics

Introduction to

# LASSO Estimator

Samet Efe Keskin

April 27, 2026

# Table of contents

# Why do we even need LASSO?

# Why do we even need LASSO?

- Too many regressors ⇒ high variance and unstable estimates
- Multicollinearity ⇒ leads to noisy coefficients
- High-dimensional settings ⇒ weak interpretability
- Prediction and inference should not be the same.

# Why does OLS fail when p > n?

$$Y = X\beta + \varepsilon, \qquad X \in \mathbb{R}^{n \times p}$$

$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y$$

If $p > n$, then

$$\mathsf{rank}(X) \leq n < p$$

so the columns of X are linearly dependent. Therefore,

$$X'X \text{ is singular} \quad \Rightarrow \quad (X'X)^{-1} \text{ does not exist}$$

**Interpretation**

When the number of candidate regressors exceeds the number of observations, OLS cannot uniquely estimate the coefficient vector.

# High-Dimensionality problem: Growth Example

Suppose we want to explain growth using:

- initial income
- schooling
- investment
- population growth
- inflation

But we only observe 3 countries.

$$X = \begin{bmatrix} 1 & \text{income}_1 & \text{schooling}_1 & \text{investment}_1 & \text{popgrowth}_1 & \text{inflation}_1 \\ 1 & \text{income}_2 & \text{schooling}_2 & \text{investment}_2 & \text{popgrowth}_2 & \text{inflation}_2 \\ 1 & \text{income}_3 & \text{schooling}_3 & \text{investment}_3 & \text{popgrowth}_3 & \text{inflation}_3 \end{bmatrix}$$

$$X \in \mathbb{R}^{3\times 6} \quad \Rightarrow \quad n = 3, \;\; p = 6, \;\; p > n$$

## High-dimensionality problem: Labor Market example

Suppose we want to explain wages using:

- education
- experience
- tenure
- gender
- region dummies

But we only observe 4 individuals.

$$X \in \mathbb{R}^{4\times 6} \quad \Rightarrow \quad n = 4, \;\; p = 6, \;\; p > n$$

$$X = \begin{bmatrix} 1 & \text{educ}_1 & \text{exp}_1 & \text{tenure}_1 & \text{occ}_1 & \text{region}_1 \\ 1 & \text{educ}_2 & \text{exp}_2 & \text{tenure}_2 & \text{occ}_2 & \text{region}_2 \\ 1 & \text{educ}_3 & \text{exp}_3 & \text{tenure}_3 & \text{occ}_3 & \text{region}_3 \\ 1 & \text{educ}_4 & \text{exp}_4 & \text{tenure}_4 & \text{occ}_4 & \text{region}_4 \end{bmatrix}$$

Ridge regression replaces OLS:

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Let

$$L(\beta) = \|Y - X\beta\|_2^2 + \lambda\beta'\beta$$

$$\frac{\partial L(\beta)}{\partial \beta} = -2X'Y + 2X'X\beta + 2\lambda\beta$$

Setting the derivative equal to zero:

$$(X'X + \lambda I)\beta = X'Y$$

So the ridge estimator is:

$$\boxed{\hat{\beta}_{\text{ridge}} = (X'X + \lambda I)^{-1}X'Y}$$

Adding $\lambda I$ makes the problem well-defined and stabilizes estimation.

# The LASSO Objective Function

LASSO solves:

$$\min_{\beta} \left\{ \underbrace{\sum_{i=1}^{n} (y_i - x_i'\beta)^2}_{\text{Residual Sum of Squares (RSS)}} + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- First term: fit to the data (RSS)
- Second term: penalty on coefficient size
- $\lambda$: controls sparsity

## Statistical properties of LASSO

LASSO has different goals than OLS:

- Not unbiased:

$$\mathrm{E}[\hat{\beta}^{\text{lasso}}] \neq \beta$$

- Introduces bias intentionally (shrinkage)

Correctly identifies non-zero coefficients

- Consistent for prediction:

$$\sum_{i=1}^{n} (x_i'(\hat{\beta} - \beta))^2 \to 0$$

**Key idea**

LASSO trades bias for lower variance and variable selection.

# What does Least Absolute Shrinkage and Selection Operator do?

- It adds an $\ell_1$ penalty to the least-squares objective.
- Small coefficients are shrunk more aggressively.
- Some coefficients become exactly zero.
- So LASSO is both a shrinkage method and a variable selection method.

**Interpretation**

Large $\lambda \Rightarrow$ more shrinkage, fewer selected variables.
Small $\lambda \Rightarrow$ less shrinkage, more selected variables.

# LASSO as an optimization problem

LASSO is defined as:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- No closed-form solution (unlike OLS or Ridge)
- The objective function is convex
- The solution is obtained using numerical optimization algorithms

## Analogy

Similar to Maximum Likelihood Estimation:

- define an objective function
- solve using optimization methods

## How is LASSO computed?

Since there is no closed-form solution, LASSO is solved numerically.

Common algorithms:

- Coordinate Descent
- Gradient-based methods

Idea of coordinate descent:

- Update one coefficient at a time
- Apply soft-thresholding
- Repeat until convergence

Modern software (Stata, R, Python) implements these efficiently

# Solving LASSO: idea of coordinate descent

LASSO solves:

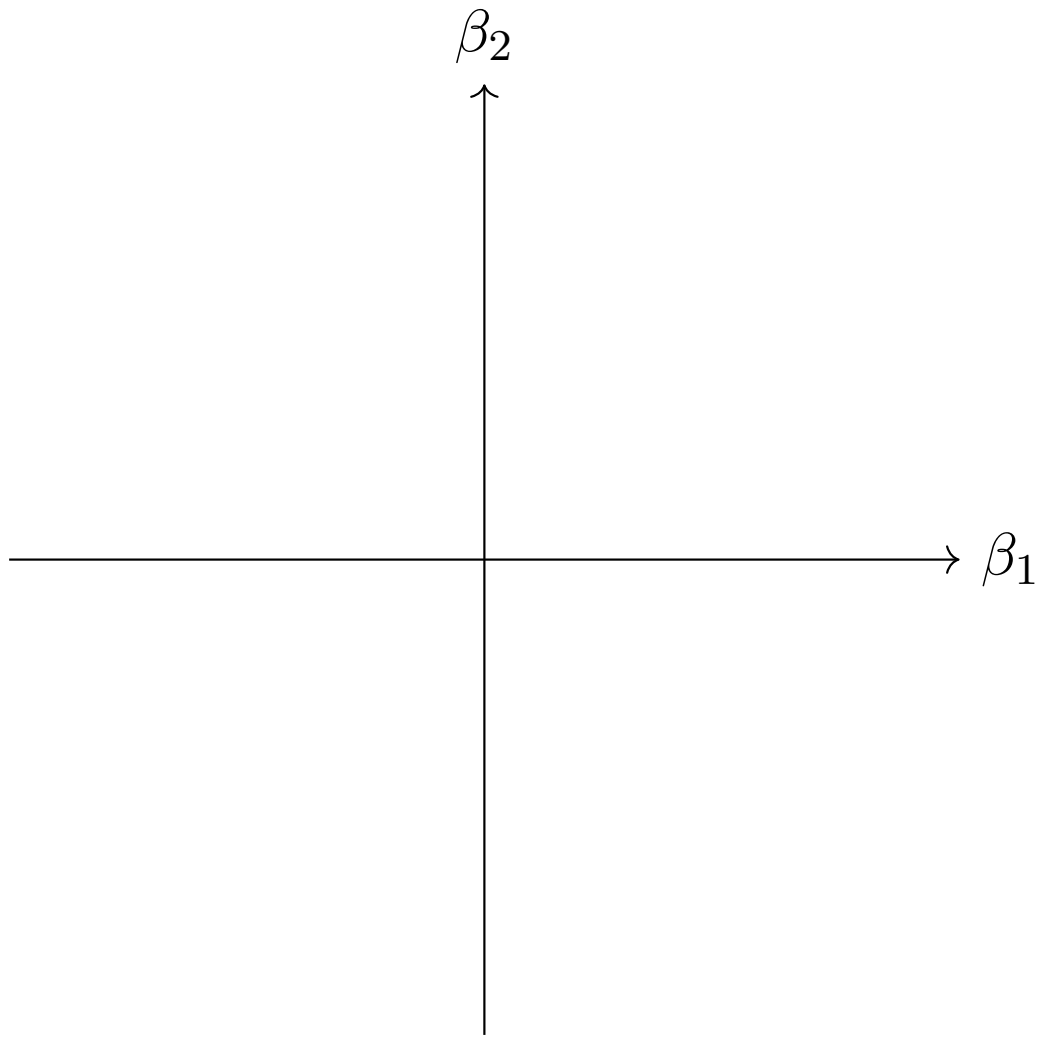
$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Instead of solving for all coefficients at once:

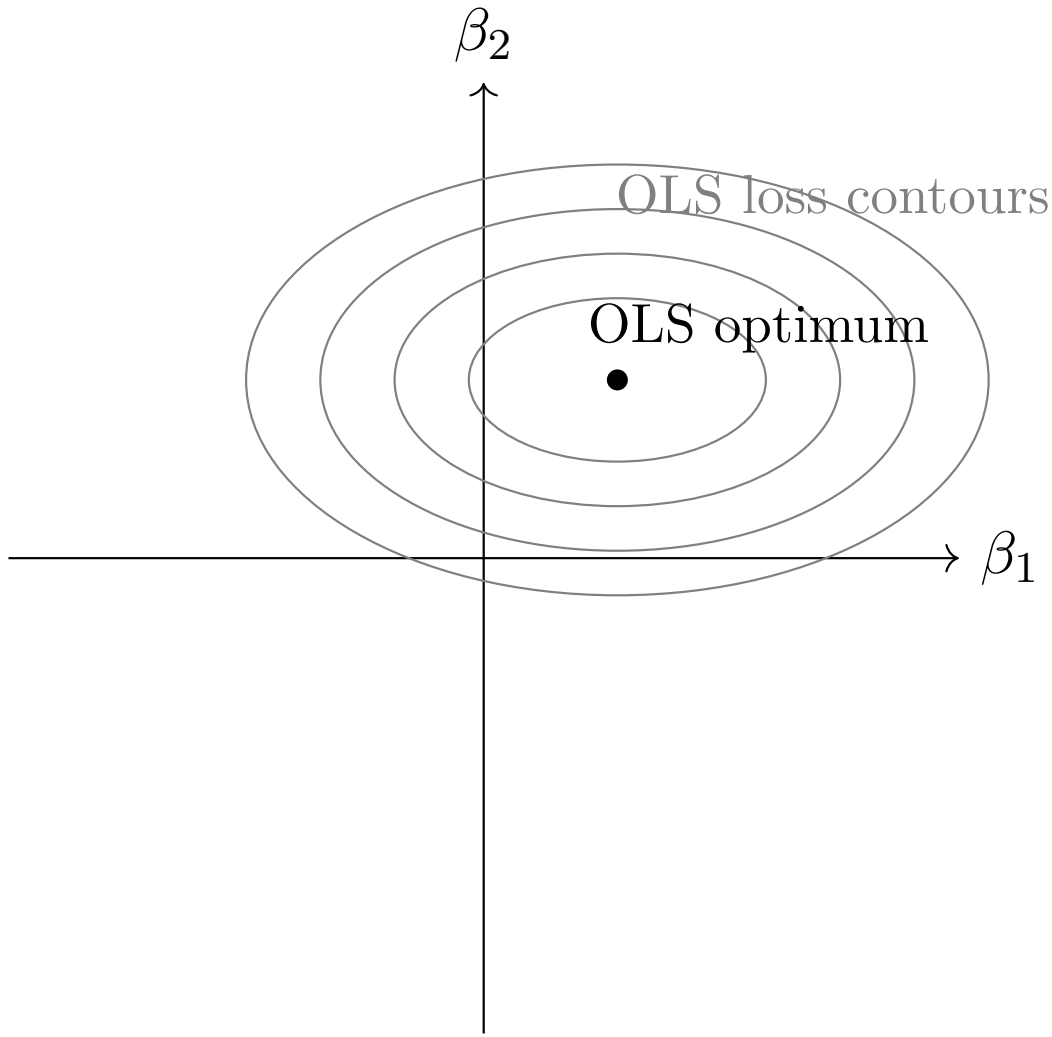
- Fix all coefficients except $\beta_j$
- Minimize with respect to $\beta_j$
- Repeat for all $j = 1, \ldots, p$

Break a high-dimensional problem into simple one-dimensional problems

# Geometry of Ridge vs LASSO

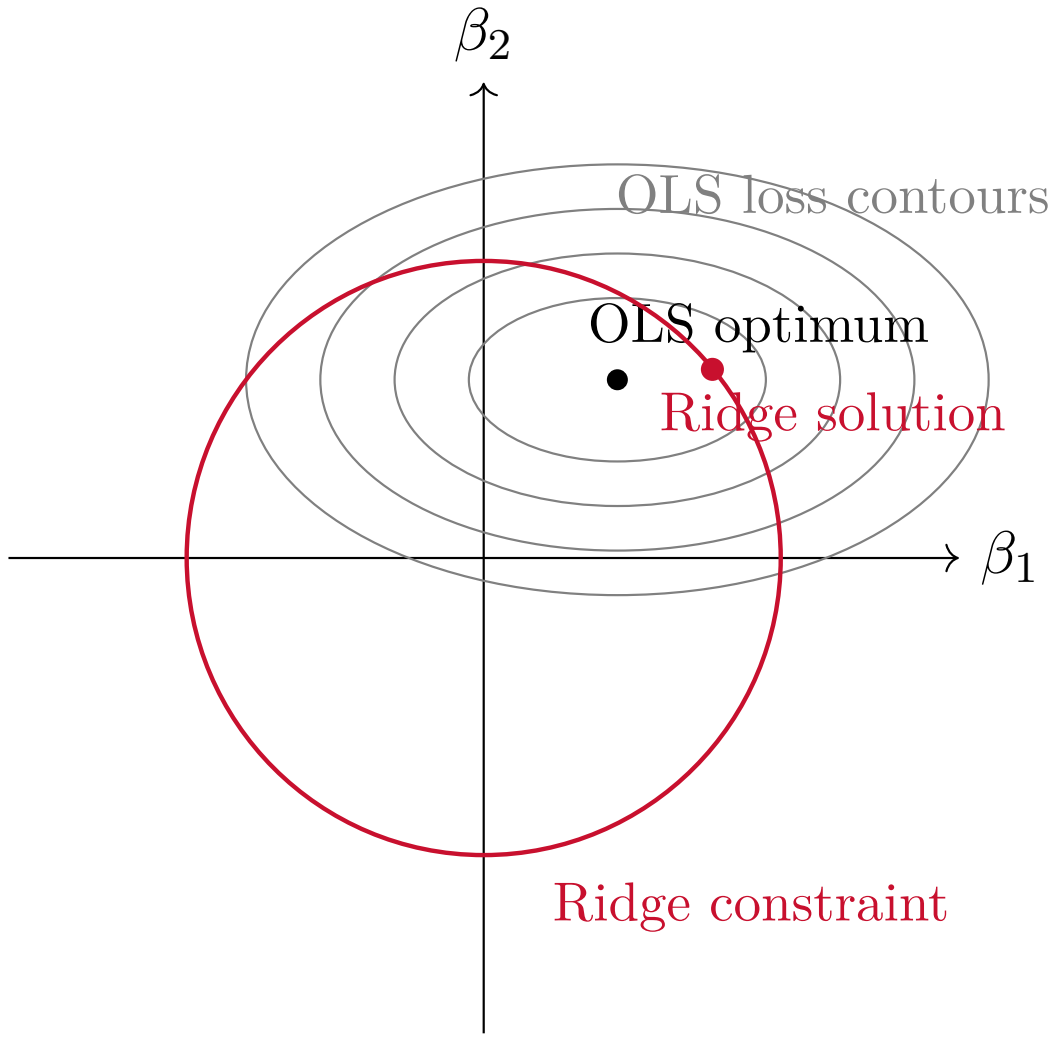


# Geometry of Ridge vs LASSO



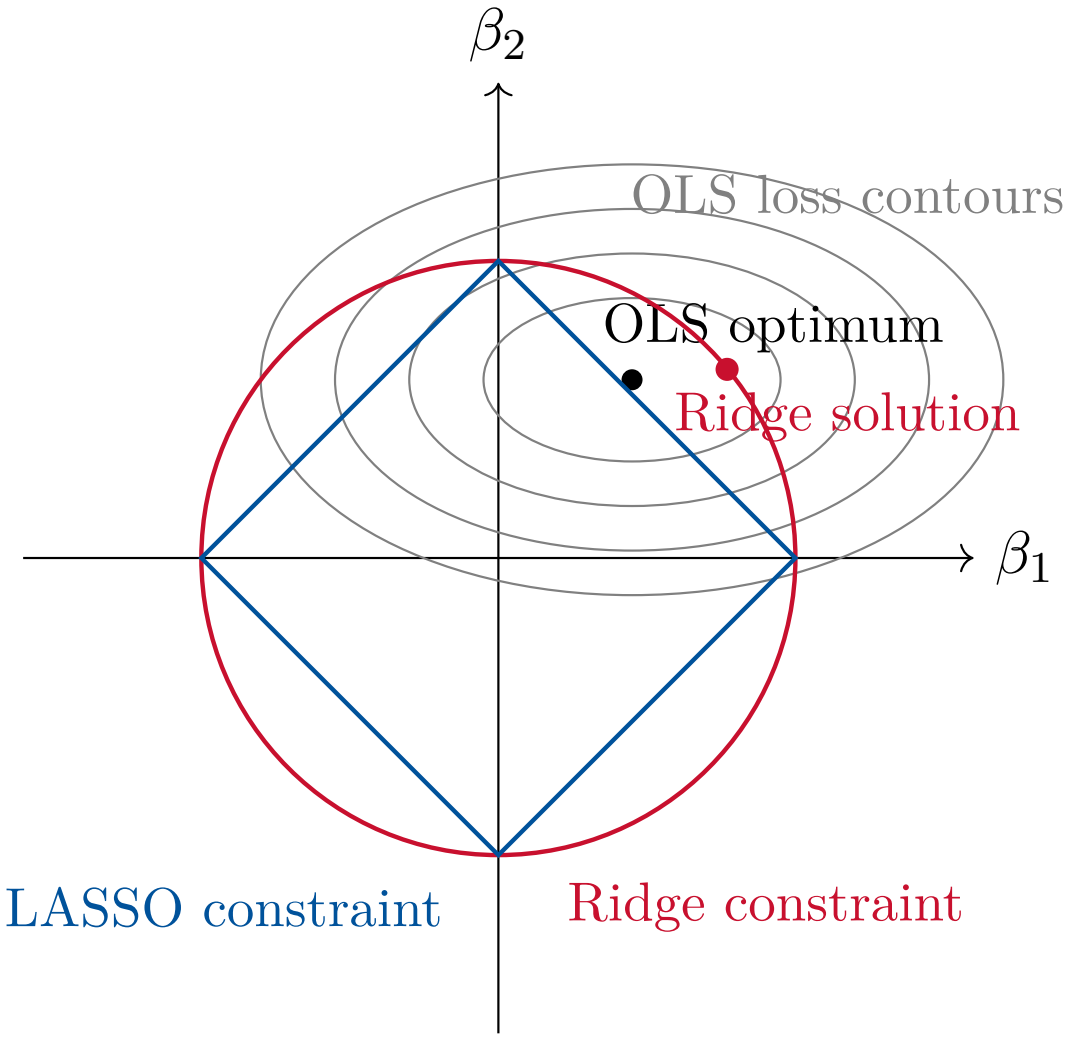

The ellipses show combinations of coefficients with the same residual sum of squares.

# Geometry of Ridge vs LASSO



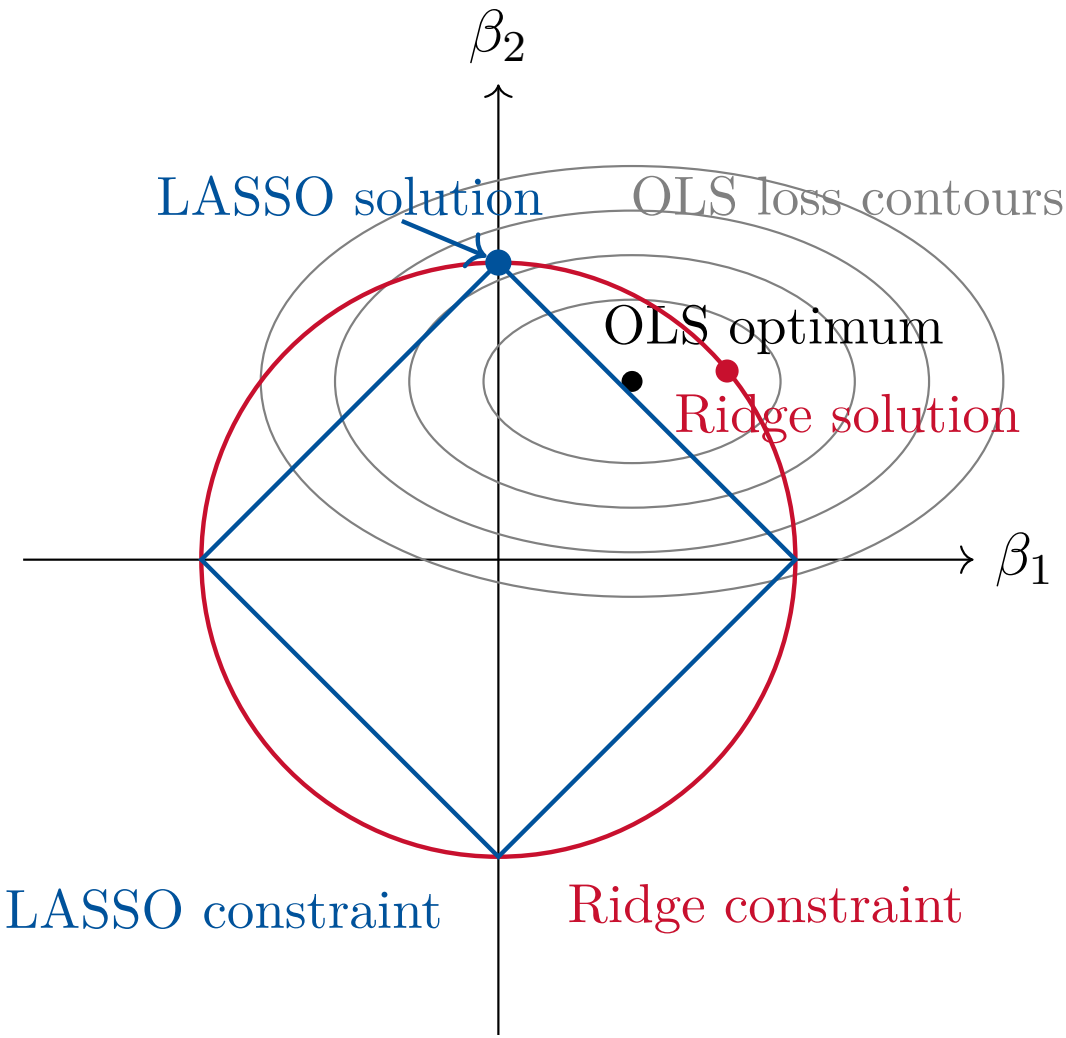

Ridge adds an $\ell_2$ constraint: coefficients are shrunk, but usually not set exactly to zero.

# Geometry of Ridge vs LASSO




LASSO uses an $\ell_1$ constraint, which has corners instead of a smooth boundary.

# Geometry of Ridge vs LASSO




Corners make zero coefficients likely $\Rightarrow$ LASSO performs variable selection.

# Why the Diamond Matters: Geometry

- Ridge uses an $\ell_2$ ball: smooth, circular.
- LASSO uses an $\ell_1$ ball: diamond-shaped.
- The least-squares contours are ellipses.
- The first contact point with the diamond often occurs at a corner.
- Corners correspond to coefficients equal to zero.

**Visual message**

$$|\beta_1| + |\beta_2| \leq c$$

diamond $\Rightarrow$ sparsity

This is why LASSO selects variables while ridge usually only shrinks them.

| Method | Penalty | Main strength | Main weakness |
|---|---|---|---|
| OLS | none | Unbiased under classical assumptions | Fails or becomes unstable in high dimensions |
| Ridge | $\ell_2$ | Stabilizes coefficients, handles collinearity | Usually keeps all variables in the model |
| LASSO | $\ell_1$ | Shrinkage + variable selection | Introduces bias; selection can be unstable with correlated regressors |

# Applied example: returns to education

We want to estimate the causal effect of education on wages:

$$\text{wage}_i = \beta_1 \cdot \text{educ}_i + X_i'\gamma + \varepsilon_i$$

Dataset:

- Individuals from labor survey
- Outcome: wages
- Treatment: years of education
- Controls: demographics, experience, region, occupation, etc.

Based on Belloni et al. (2014)

Potentially hundreds of control variables

# LASSO selects relevant controls

Using LASSO to select controls:

- Start with 150 candidate variables
- LASSO selects a small subset

## Selected variables

- experience
- experience$^2$
- gender
- region dummies
- occupation dummies

Most variables are dropped (coefficients = 0)

# Results: effect of education

| Method | $\hat{\beta}_{\text{educ}}$ | Std. Error |
|---|---|---|
| Naive OLS | 0.02 | 0.015 |
| OLS with many controls | 0.05 | 0.030 |
| Post-LASSO / Double LASSO | 0.08 | 0.012 |

## Interpretation

- Naive OLS underestimates effect
- Too many controls $\rightarrow$ noisy estimates
- LASSO selects relevant controls and improves inference

## Applied case 2: portfolio allocation in high dimensions

In portfolio optimization, key quantities depend on the inverse covariance matrix:

$$\Theta = \Sigma^{-1}$$

Examples:

$$w_{GMV} = \frac{\Sigma^{-1}1}{1'\Sigma^{-1}1}$$

where $w_{GMV}$ is the global minimum variance portfolio.

- p = number of assets
- n = number of time observations

If $p > n$, the sample covariance matrix is singular, so direct inversion is problematic.

Based on Callot, Caner, Onder, and Ulasan (2019)

## Nodewise regression: estimating $\Sigma^{-1}$ with LASSO

For each asset $j = 1, \ldots, p$, regress its demeaned return on all the others:

$$r^*_{t,j} = (r^*_{t,-j})'\gamma_j + \eta_{t,j}$$

Estimate $\gamma_j$ by LASSO:

$$\hat{\gamma}_j = \arg\min_{\gamma \in \mathbb{R}^{p-1}} \left( \frac{1}{n} \sum_{t=1}^{n} (r^*_{t,j} - r^*_{t,-j}\gamma)^2 + 2\lambda_j \sum_{k \neq j} |\gamma_k| \right)$$

Repeat this for all $j = 1, \ldots, p$, then combine the estimated rows to build an estimate of the precision matrix.

Idea: estimate the inverse covariance matrix directly, without inverting a singular sample covariance matrix.

## Real high-dimensional example from the paper

Empirical application in the paper:

$$n = 293 \text{ monthly observations}, \qquad p = 304 \text{ assets}$$

So:

$$R \in \mathbb{R}^{293 \times 304}, \qquad p > n$$

The authors compare Nodewise LASSO, POET, and Ledoit–Wolf methods.
Main finding:

- Nodewise regression-based portfolios perform well out of sample
- especially in terms of Sharpe ratio, variance, and turnover on monthly data

Callot et al. (2019)

# Illustrative output: monthly portfolio performance

| Method | Return | Variance | Sharpe Ratio |
|---|---|---|---|
| POET | 0.02499 | 0.01953 | 0.1788 |
| Nodewise | 0.02644 | 0.01580 | 0.2104 |
| Ledoit–Wolf | 0.07140 | 0.18421 | 0.1664 |

Nodewise delivers a higher Sharpe ratio with lower variance in this monthly out-of-sample exercise.

# LASSO in Stata: Main Commands

- Prediction / model selection:

```
lasso linear y controls
elasticnet linear y controls
sqrtlasso y controls
```

- Inference:

```
dsregress y d, controls(controls)
poregress y d, controls(controls)
xporegress y d, controls(controls)
```

**Interpretation**

Use lasso for prediction; use dsregress/poregress/xporegress for inference.