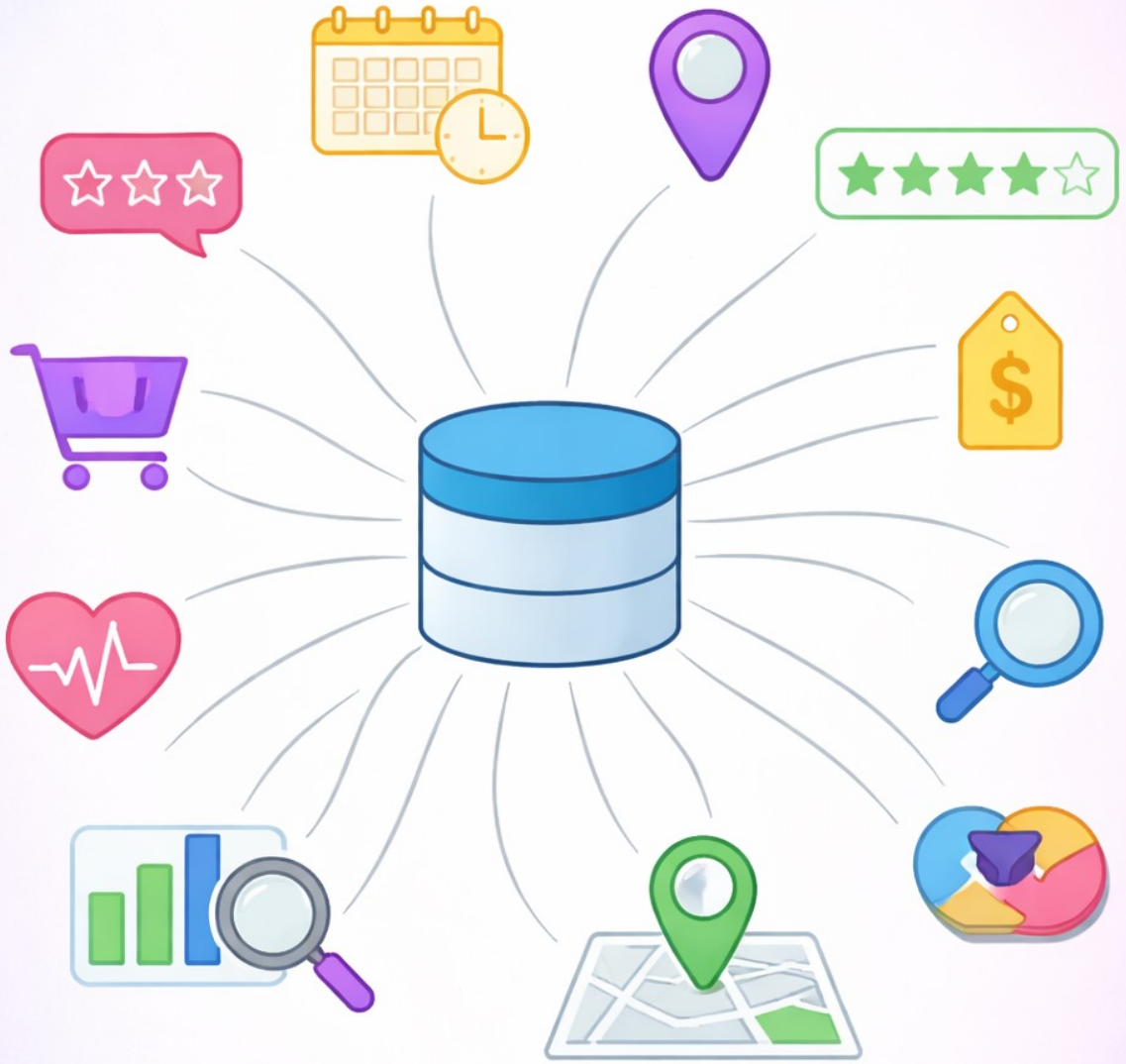
Georgia Tech

Feature Selection



Mohsen Moghaddam, Ph.D.

Gary C. Butler Family Associate Professor
H. Milton Stewart School of Industrial and Systems Engineering
George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

SAIL
Symbiotic & Augmented Intelligence Laboratory

Spring 2026

Learning Outcomes

- Explain how feature selection improves generalization performance and model interpretability by controlling model complexity
- Distinguish feature selection from feature extraction (e.g., PCA), and understand the advantages of each approach
- Identify scenarios where feature selection is essential, including high-dimensional, noisy, or correlated feature spaces
- Explain the role of sparsity in feature selection and how it connects statistical modeling, optimization, and interpretability
- Compare common evaluation metrics and search techniques used in feature selection, including subset selection, ridge regression, and LASSO
- Apply regularization-based feature selection methods and interpret the resulting models in real data examples

Motivation & Problem Setup

Example: Apartment Hunting

What is a reasonable rent (the **response variable**)? Can we predict rent accurately **without** using all available features?

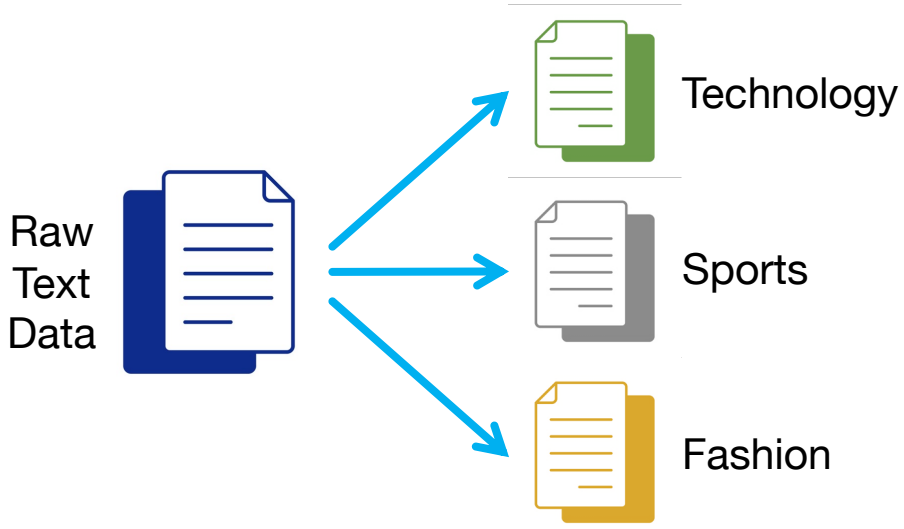
Redundant information

Rent ($)	Location	Living area (ft^2)	# bath	# bedroom
600	Midtown	230	1	1
1000	Buckhead	506	2	2
1100	Midtown	433	1	2
500	Downtown	109	1	1
⋮	⋮	⋮	⋮	⋮
?	**Midtown**	**150**	**2**	**1**
?	**Downtown**	**270**	**1**	**1.5**

Redundant or irrelevant features inflate variance and obscure interpretation

Example: Document Classification

Select the most important features before building a classifier to reduce overfitting caused by **rare or accidental correlations**



Spurious Correlation:

Suppose a rare term, such as "arachnocentric", carries no true semantic information about a class like "China." However, if this term appears only in China-related documents in the training data, the learning algorithm may incorrectly treat it as a strong indicator for that class. As a result, test documents containing "arachnocentric" may be misclassified as China-related, despite the association being spurious.

— Manning, Raghavan, Schutze, 2008

Why is this problem worse in text data than in apartment data?

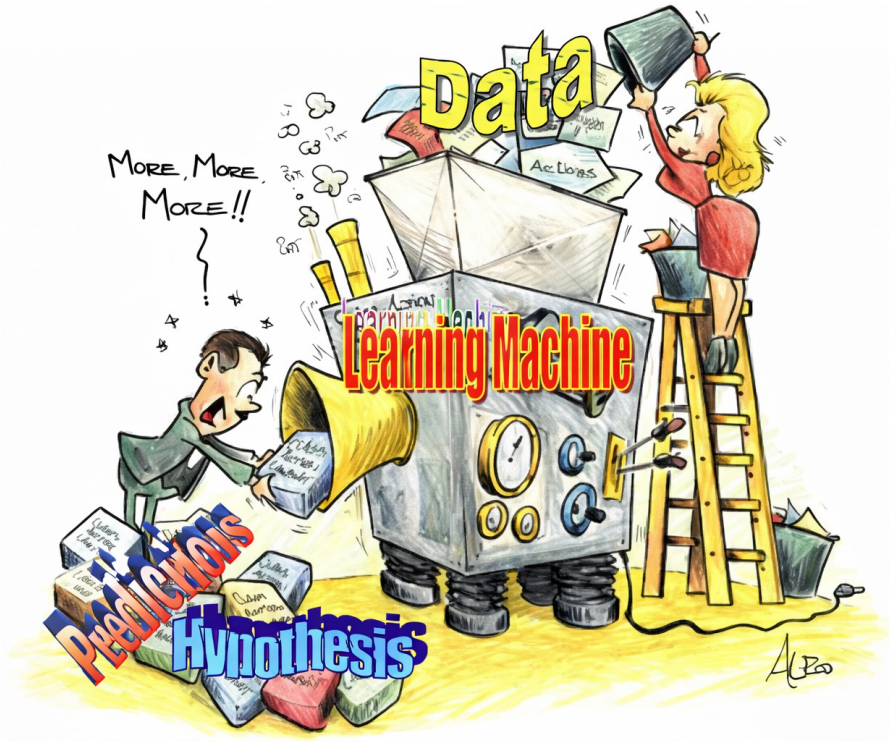
What Is Feature Selection?

Selecting a subset of **relevant features** to

- Simplify models by removing irrelevant or redundant variables
- Improve data efficiency by training smaller models
- Enhance interpretability by identifying which variables matter most
- Improve accuracy by eliminating noisy features
- Enhance generalization by reducing overfitting (bias–variance tradeoff)

Applications: NLP, image analysis, genetic data

"Garbage in, garbage out"



Keeps original variables ← **Feature Selection ≠ Feature Extraction** → Creates new variables (e.g., PCA)

Major Approaches

Feature selection = Evaluation metrics (**What makes a feature "good"?**)
+ Search techniques (**How do we find the features?**)

Evaluation Metrics

- Heuristics: e.g., eliminate variables with small variance
- Information theory: e.g., statistical dependence metrics
- Prediction error: e.g., cross validation
- Bias–variance tradeoffs (complexity control; e.g., LASSO)

Search Techniques

- Subset selection: enumeration
- ℓ_1-regularized methods: LASSO and extensions

How do different metrics and search strategies shape the features selected?

Evaluation Metrics

Example: Apartment Hunting

Features: Living area, distance to campus, # bedroom, denoted as

$$x = (x_1, x_2, \ldots, x_n)^\top$$

Target: Rent, denoted as y

Training set:

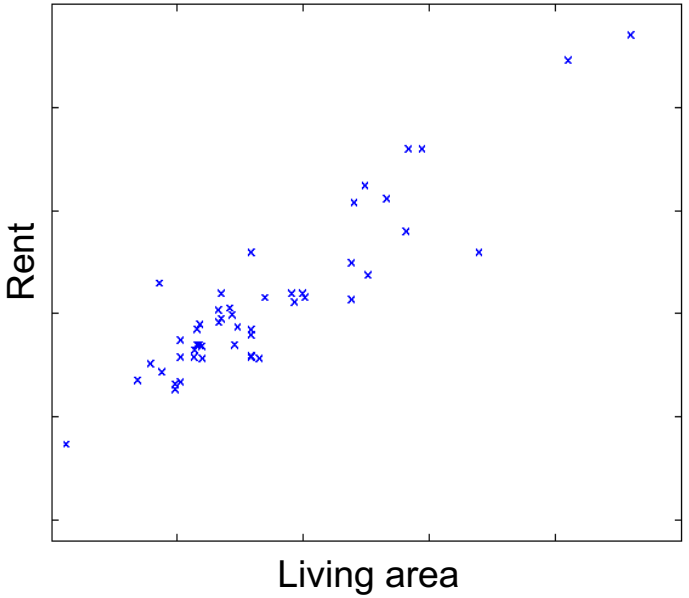
$$X = [x^1 \cdots x^m] \in \mathbb{R}^{(n \times m)}$$

$$y = (y^1, y^2, \ldots, y^m)^\top \in \mathbb{R}^m$$

Problem: Which subset of features should we select to predict rent?

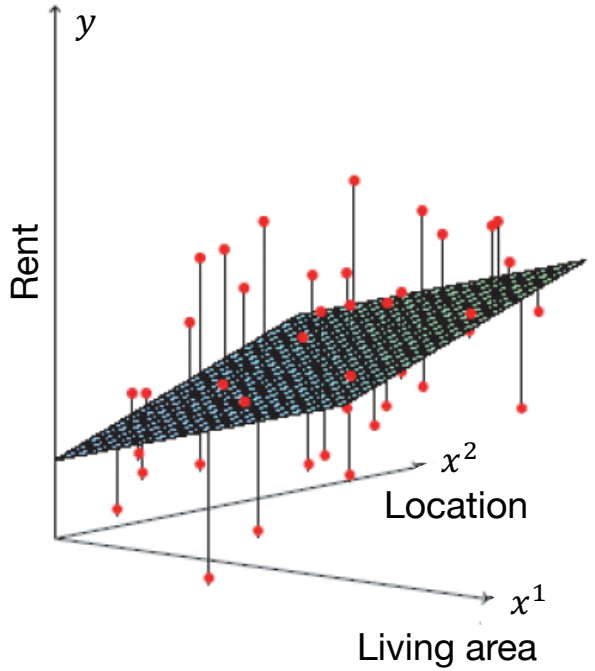
One Feature

Too simple? Underfitting?



Two Features

Enough features?



If two features are highly correlated here, what happens in a linear fit?

Linear Regression Model

- Assume the response variable y is a linear function of the input features x, plus unmodeled effects or random noise ϵ:

Bias Noise

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n + \epsilon$$

- Define the parameter vector:

$$\theta = (\theta_0, \theta_1, \ldots, \theta_n)^\top$$

- Augment the feature vector by one dimension (for bias θ_0):

$$x \leftarrow (1, x^\top)^\top$$

- Then the feature selection model can be written compactly as:

$$y = \theta^\top x + \epsilon$$

Feature selection: Finding a vector θ with many zero entries; if $\theta_j = 0$, then feature x_j contributes nothing to the prediction

Least Squares Method

Given m data points, find θ that minimizes the MSE:

$$\hat{\theta} = \arg\min_{\theta} L(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(y^i - \theta^\top x^i\right)^2$$

Set gradient to 0 and find the parameter:

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m}\sum_{i=1}^{m}\left(y^i - \theta^\top x^i\right)x^i = 0 \quad \Rightarrow \quad -\frac{2}{m}\sum_{i=1}^{m} y^i x^i + \frac{2}{m}\sum_{i=1}^{m} x^i x^{i\top}\theta = 0$$

$$\Rightarrow \quad \hat{\theta} = (XX^\top)^{-1}Xy$$

Poor evaluation metric—optimizes data fit only and doesn't **penalize complexity**

Ridge Regression

What if we cannot invert $XX^\top$ (e.g., when variables are correlated)?

Given m data points, find θ that minimizes the regularized mean square error:

$$\theta^r = \arg\min_\theta L(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(y^i - \theta^\top x^i\right)^2 + \boxed{\lambda\|\theta\|_2^2}$$

← ℓ_2 **Regularizer** (penalizing large coefficients = complexity)

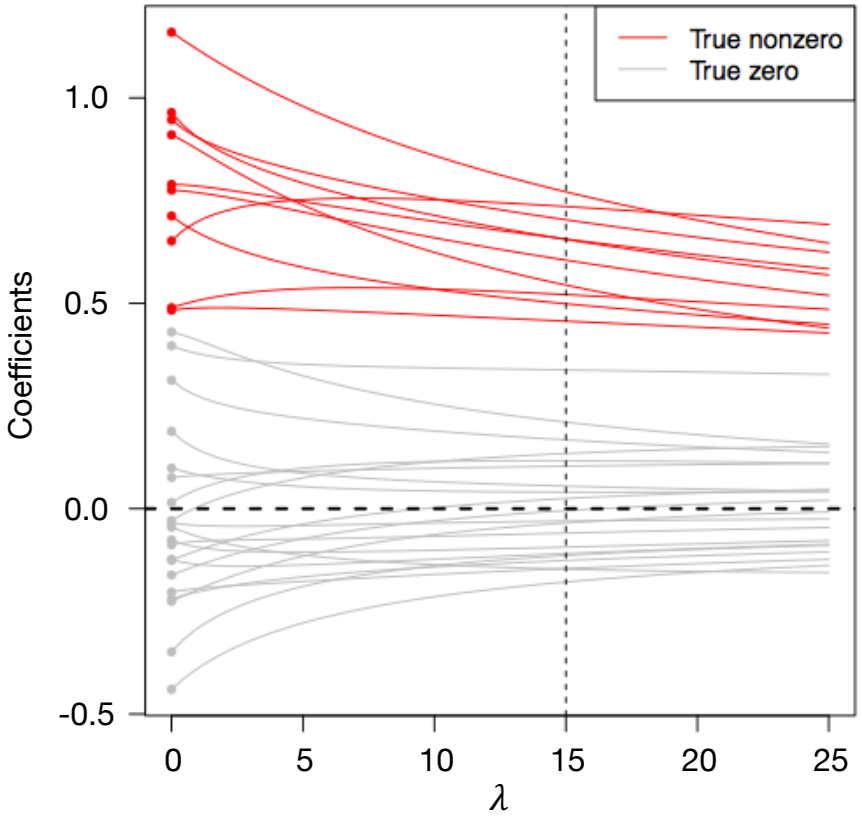
Gradient becomes (different solution for different λ):

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m}Xy + \frac{2}{m}XX^\top\theta + 2\lambda\theta = 0 \quad \Rightarrow \quad \theta^r = \left(\frac{1}{m}XX^\top + \lambda I\right)^{-1}\left(\frac{1}{m}Xy\right)$$

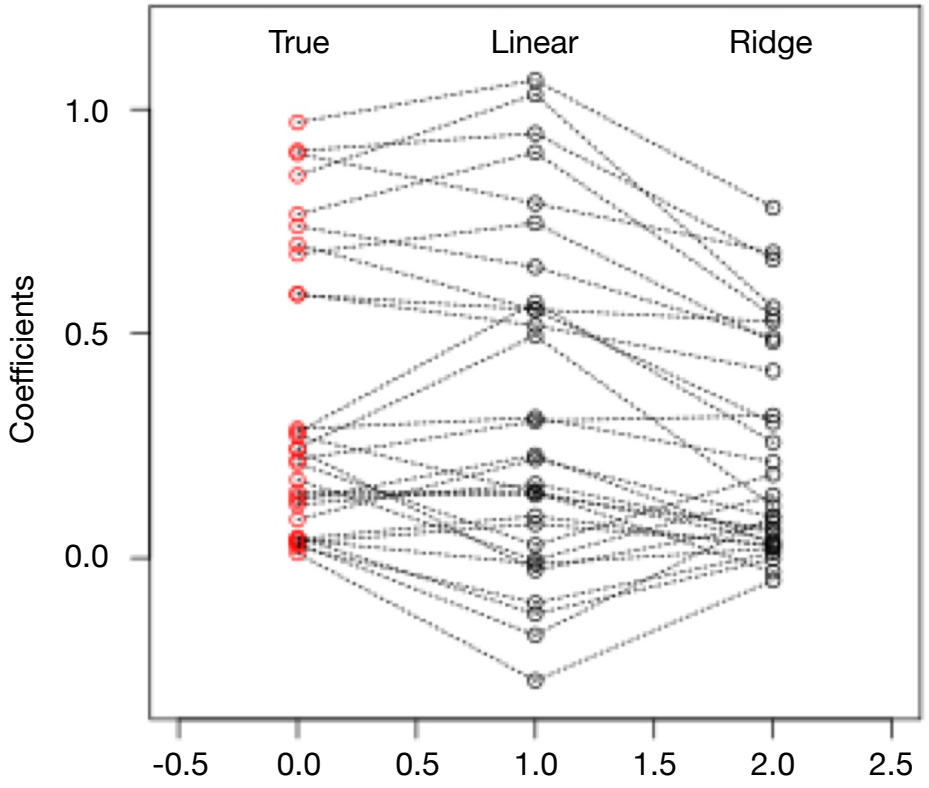
Ridge provides a stable ordering of feature importance—no feature selection

Numerical Comparison

- Ridge stabilizes estimates but doesn't perform feature selection
- **Intuition:** Use all features, but don't trust any single one too much



$m = 50, n = 20, \lambda = 25$



Search Techniques

Feature Selection for Regression

Selecting a subset of the **most "important" features** (variables) involves finding θ and their locations of "zero-entries":

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \ldots + \theta_n x_n + \epsilon$$

Direct enumeration is expensive

- Each of the n features has **2 independent choices** (selected/not selected)
- By the rule of product, there are $O(2^n)$ possibilities:

$$\underbrace{2 \times 2 \times \cdots \times 2}_{n \text{ times}} = 2^n$$

LASSO address this issue by convex relaxation

LASSO

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator replaces combinatorial search with convex optimization

$$\theta = \arg\min_{\theta} L(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(y^i - \theta^\top x^i\right)^2 + \boxed{\lambda\|\theta\|_1} \leftarrow$$

Convex relaxation

- This is a **convex optimization problem** and can be solved efficiently
- LASSO implicitly select features by shrinking unimportant coefficients **exactly** to zero
- Regularizer λ controls model complexity: larger $\lambda \rightarrow$ fewer parameters will be selected
- ℓ_1 penalty can be used for to encourage sparsity in solution

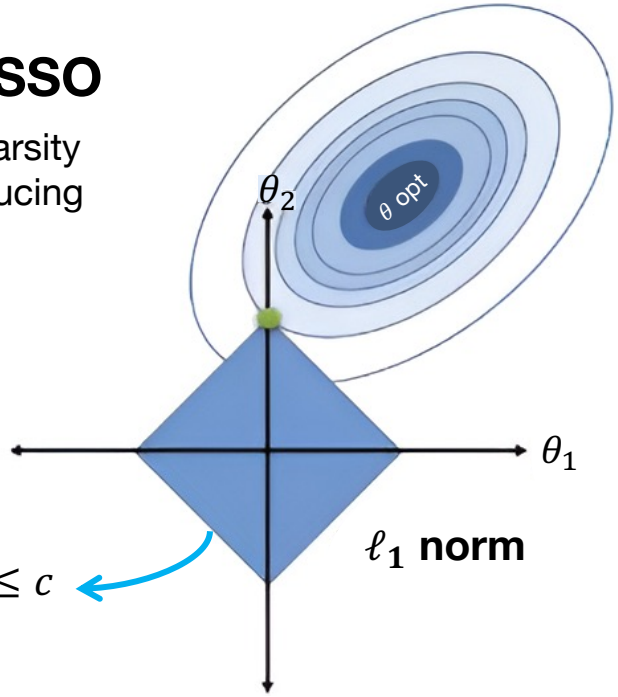
Why might penalizing absolute values push coefficients exactly to zero?

Why LASSO Selects Variables

Corners of the ℓ_1 constraint align with coordinate axes, producing sparsity

LASSO

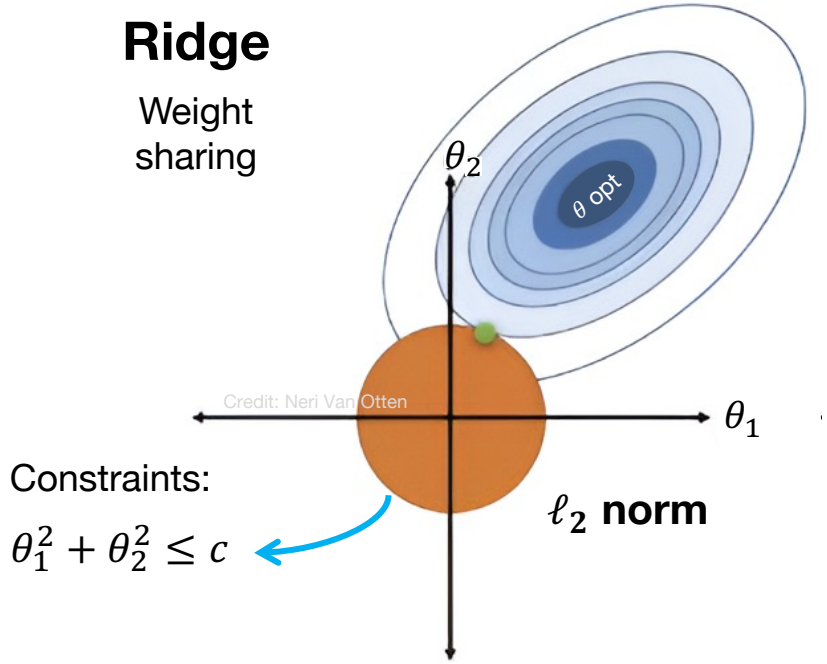
Sparsity inducing



$$\min \frac{1}{m} \sum_{i=1}^{m} \left(y^i - \theta^\top x^i\right)^2 + \lambda \|\theta\|_1$$

Ridge

Weight sharing



$$\min \frac{1}{m} \sum_{i=1}^{m} \left(y^i - \theta^\top x^i\right)^2 + \lambda \|\theta\|_2^2$$

Why LASSO Selects Variables (cont.)

Let's compare 1D ridge and LASSO, so we only solve for one parameter θ

Loss Function

$$\min_{\theta} \frac{1}{2}\|y - x\theta\|^2 + \text{regularization}$$

constant

$x^\top x = 1$ (ℓ_2-normalized feature column: $x \leftarrow \frac{x}{\|x\|_2}$)

$$\|y - x\theta\|^2 = (y - x\theta)^\top (y - x\theta) = y^\top y - 2\theta x^\top y + \theta^2 x^\top x$$

$$z = x^\top y$$

constant

$$\|y - x\theta\|^2 = \theta^2 - 2\theta z = (\theta - z)^2 - z^2$$

Loss Function

$$\min_{\theta} \frac{1}{2}(\theta - z)^2 + \text{regularization}$$

Why LASSO Selects Variables (cont.)

Let's compare 1D ridge and LASSO, so we only solve for one parameter θ

Ridge Regression

$$\min_{\theta} \frac{1}{2}(\theta - z)^2 + \lambda\theta^2$$

Step 1: Differentiate

$$(\theta - z) + 2\lambda\theta = 0$$

Step 2: Solve

$$\theta = \frac{z}{1 + 2\lambda}$$

→ *Does ridge shrink θ? When is $\theta = 0$?*

Ridge almost never produces exact zeros

Why LASSO Selects Variables (cont.)

Let's compare 1D ridge and LASSO, so we only solve for one parameter θ

LASSO Regression

$$\min_{\theta} \frac{1}{2}(\theta - z)^2 + \lambda|\theta|$$

Case 1: $\theta > 0$ ($|\theta| = \theta$) $\quad \theta = z - \lambda$

Case 2: $\theta < 0$ ($|\theta| = -\theta$) $\quad \theta = z + \lambda$

Case 3: $\theta \approx 0$; if $|z| \leq \lambda$, $\theta \neq 0$ increases the objective

$$\Longrightarrow \quad \theta = \begin{cases} 0, & |z| \leq \lambda \\ z - \lambda \text{sign}(z), & |z| > \lambda \end{cases}$$

→ *Does this create a whole interval of zeros? Does ridge have such an interval?*

LASSO performs feature selection—ridge does not

LASSO: Compressive Sensing

You can't directly observe every pixel, only indirect/aggregate measurements

$$y = Ax + \epsilon$$

- $y \in \mathbb{R}^m$: observed measurements
- $A \in \mathbb{R}^{m \times p}$: measurement matrix ($m \ll p$)
- $x \in \mathbb{R}^p$: pixels (unknown)

→ Infinite solutions satisfy $y = Ax + \epsilon$

Ridge Reconstruction (ℓ_2)

→ Stabilizes solution & spreads energy

$$\hat{x} = \arg\min_x \|y - Ax\|_2^2 + \lambda \|x\|_2^2$$

LASSO Reconstruction (ℓ_1)

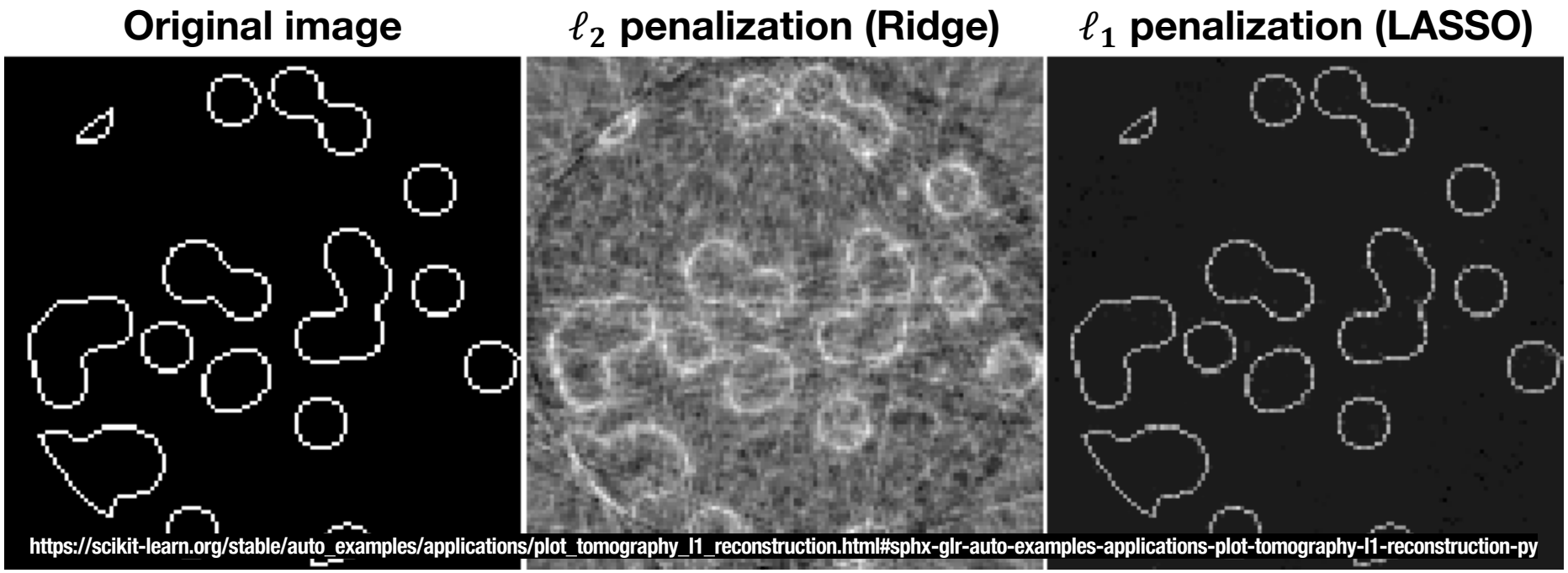
→ High sparsity & many zero coefficients

$$\hat{x} = \arg\min_x \|y - Ax\|_2^2 + \lambda \|x\|_1$$

Treat pixels as parameters and solve a regularized linear inverse problem

LASSO: Compressive Sensing (cont.)

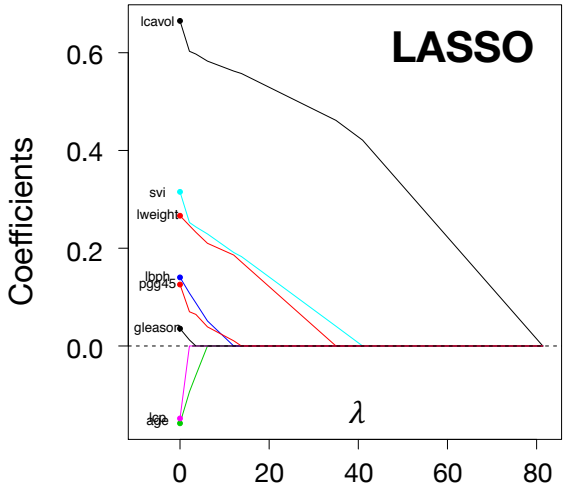
Exploiting sparsity to recover signals from limited measurements

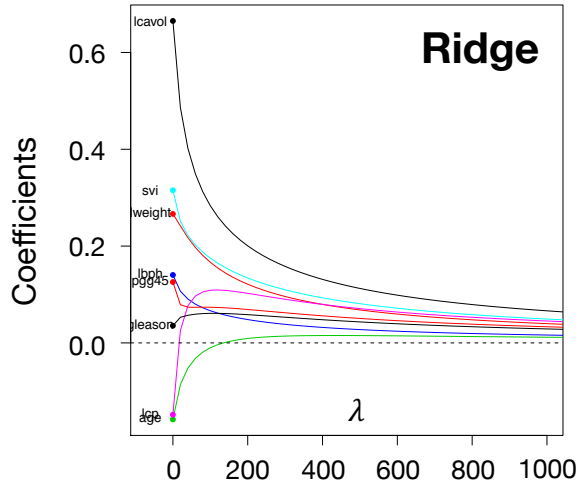


LASSO succeeds if sparsity is a valid assumption about underlying structure

LASSO: Medical Diagnosis

- **Clinical question:** We study the level of prostate-specific antigen (PSA) in men with prostate cancer
- **Data:** 97 patients, 8 clinical predictors
- **Comparison:** Ridge regression shrinks coefficients but does not perform variable selection—LASSO produces a sparse, interpretable model
- **LASSO's interpretation:** If we restrict the model to 3 leading factors, LASSO selects *cancer volume*, *seminal vesicle invasion*, and *prostate weight*
- **Model selection:** Regularization parameter λ is chosen using cross-validation





Elastic Net

Ridge: useful when features are correlated, doesn't select features

LASSO: feature selection, not stable when features are correlated

Elastic net: For correlated features, combines

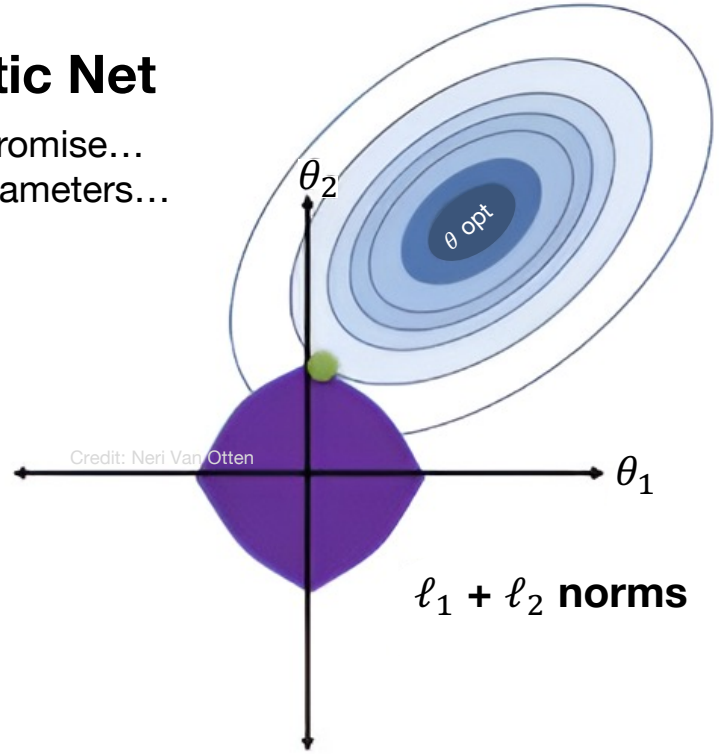
- **Sparsity** (ℓ_1) of LASSO
- **Stability** (ℓ_2) of Ridge

$$\min \frac{1}{m}\sum_{i=1}^{m}\left(y^i - \theta^\top x^i\right)^2 + \lambda(\alpha\|\theta\|_2^2 + (1-\alpha)\|\theta\|_1),$$

$$\alpha \in [0,1]$$

Elastic Net

Compromise…
Two parameters…



ℓ_1 + ℓ_2 **norms**

Comparing Ridge, LASSO, & Elastic Net

Ridge stabilizes coefficients, **LASSO** selects features, and **Elastic Net** selects features while respecting correlation structure

Method	Structural Assumption	Regularization	Feature Selection	Behavior	Strengths	Limitations
OLS*	All features matter	None	❌	Unstable	Simple, unbiased	Overfits, high variance
Ridge	Many small effects	ℓ_2	❌	Shares weight	Stable, good prediction	Not interpretable
LASSO	Few important features	ℓ_1	✅	Picks one arbitrarily	Sparse, interpretable	Unstable w/ correlation
Elastic Net	Sparse but grouped effects	$\ell_1 + \ell_2$	✅	Keeps groups	Stable + sparse	Extra tuning

* Ordinary Least Squares

Demo: Diabetes Dataset

Data:

- 442 diabetes patients
- 10 baseline predictors
- Age, sex, body mass index
- Average blood pressure
- Six blood serum measurements

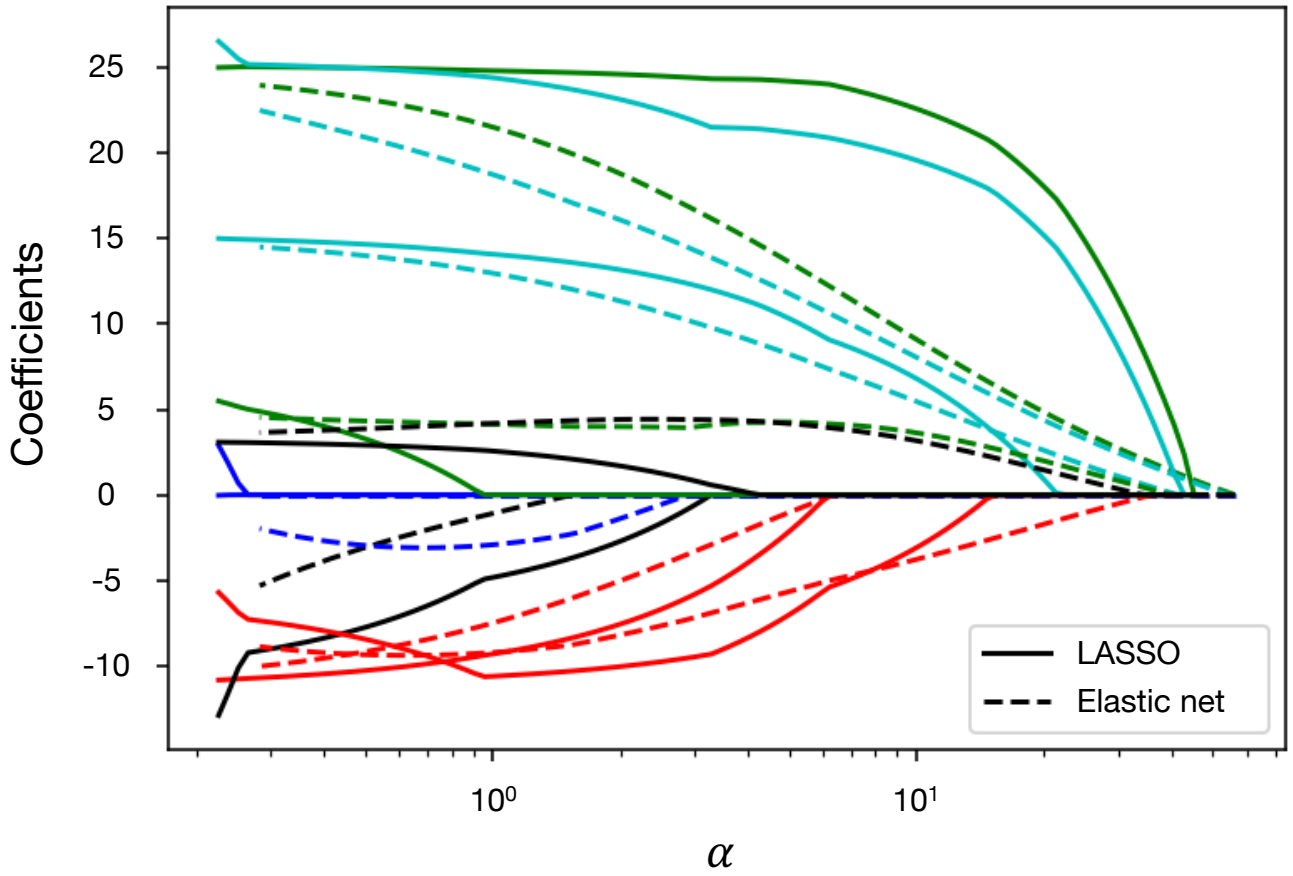
Response variable: Quantitative measure of disease progression one year after baseline

Model selection:

- Cross-validation to select regularization parameter λ
- Controls model sparsity and prediction performance

Solution Path

Feature selection traces how models evolve as sparsity constraints change



Key Takeaways

What We Learned This Week

- Feature selection improves generalization and interpretability by controlling model complexity
- Least squares uses all features and is sensitive to noise and multicollinearity
- Ridge regression stabilizes estimates via ℓ_2 regularization but does not perform variable selection
- LASSO uses ℓ_1 regularization to induce sparsity and automatically select features
- Elastic Net combines ℓ_1 and ℓ_2 penalties to handle correlated predictors effectively
- The regularization parameter λ governs the bias–variance tradeoff and sparsity level
- Cross-validation is essential for selecting λ and achieving good out-of-sample performance
- Feature selection plays a critical role in high-dimensional data and real-world applications such as healthcare and signal reconstruction