# From Predictive to Generative

Qiang Sun

January 2026

# Roadmap

1. High-dimensional setting: why OLS breaks
2. Ridge regression: constraint vs penalty, geometry
3. Model selection: data splitting and $J$-fold CV
4. Bridge family $\ell_p$: ridge, lasso, sparsity vs convexity
5. Elastic net: combines ridge + lasso
6. From discriminative to generative modeling
7. QDA / LDA / Naive Bayes / DLDA

# High Dimensional Data and Regularization

# High Dimensional Data

## Definition

High dimensional data: "a lot of" features. More precisely, when the dimensionality $d$ is comparable to (or much larger than) the sample size $n$, we are in the high-dimensional regime.

## Regression notation

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n, \qquad \mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1d} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nd} \end{bmatrix} \in \mathbb{R}^{n \times d}.$$

# Why OLS Breaks When $d > n$

**Question**

What happens to

$$\widehat{\beta}^{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \quad \text{when } d > n?$$

**Key Idea (Non-invertibility)**

When $d > n$, $\mathbf{X}^\top \mathbf{X}$ is not invertible (rank at most $n$), so OLS is ill-defined.

**Two common remedies**

1. **Two-step:** reduce dimension first (e.g., PCA), then regress.
2. **Single-step:** regularize (e.g., ridge).

# Ridge Regression

# Ridge Estimator: Primal vs Dual

## Primal (constraint form)

$$\widehat{\beta}^{\,t} = \underset{\|\beta\|_2^2 \le t}{\arg\min}\ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2.$$

## Dual (penalty form)

$$\widehat{\beta}^{\,\lambda} = \underset{\beta \in \mathbb{R}^d}{\arg\min}\ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

## Key Idea (Equivalence)

For each $\lambda > 0$, there is a one-to-one mapping $t = t(\lambda)$ so that $\widehat{\beta}^{\,\lambda} = \widehat{\beta}^{\,t}$.

# Closed Form and Interpretation

## Closed form

$$\widehat{\beta}^{\lambda} = (\mathbf{X}^{\top}\mathbf{X} + \lambda I)^{-1}\mathbf{X}^{\top}\mathbf{Y}.$$

## Why ridge always works

$\mathbf{X}^{\top}\mathbf{X} + \lambda I$ is always invertible for $\lambda > 0$.

## Extreme cases

- If $t > \|\widehat{\beta}^{\text{OLS}}\|_2^2$ then $\lambda = 0$ (no shrinkage).
- If $t = 0$ then $\lambda = \infty$ (shrink everything to 0).

# Example: $\mathbf{X}^\top \mathbf{X} = I$

## Computation

If $\mathbf{X}^\top \mathbf{X} = I$, then

$$\widehat{\beta}^{\lambda} = (I + \lambda I)^{-1}\mathbf{X}^\top \mathbf{Y} = \frac{1}{1+\lambda}\mathbf{X}^\top \mathbf{Y} = \frac{1}{1+\lambda}\widehat{\beta}^{\text{OLS}}.$$

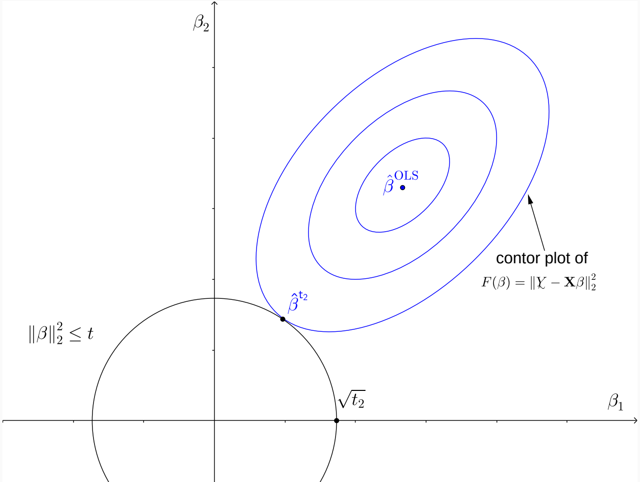
## Key Idea (Shrinkage)

Ridge shrinks coefficients continuously toward 0 as $\lambda$ increases.

# Geometry of Ridge

## Contour lines

A contour line of a function is a curve along which the function has a constant value.



# Model Selection

# Model Selection via Data Splitting

Split data $\mathcal{D}$ into $\mathcal{D}_1$ (train) and $\mathcal{D}_2$ (validation), with sizes $n_1$ and $n_2$.

## Candidate tuning parameters

$$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}.$$

Fit $\widehat{\beta}^{\lambda_k}$ on $\mathcal{D}_1$.

## Data-splitting score

$$\mathcal{DS}(k) = \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} (Y_i - X_i^\top \widehat{\beta}^{\lambda_k})^2.$$

Pick the $\lambda_k$ with smallest $\mathcal{DS}(k)$.

# Pros / Cons of Data Splitting

**Advantages**

- Simple (conceptually + computationally).
- Good generalization performance in practice.
- Conditionally (on $\mathcal{D}_1$), $\mathcal{DS}(k)$ is an unbiased estimator of risk $R(\widehat{\beta}^{\lambda_k})$.

**Disadvantage**

- "Wastes" data: $\mathcal{D}_2$ is not used for training at all.

# $J$-Fold Cross Validation

## Definition

Split $\mathcal{D}$ into $J$ equal subsets $\mathcal{D}_1, \ldots, \mathcal{D}_J$. For each fold $j$, train on $\mathcal{D} \setminus \mathcal{D}_j$ and validate on $\mathcal{D}_j$.

## CV score

For each $\lambda_k$ compute $\mathcal{DS}_j(k)$ and average:

$$\mathrm{CV}(k) = \frac{1}{J} \sum_{j=1}^{J} \mathcal{DS}_j(k).$$

Pick the $\lambda_k$ minimizing $\mathrm{CV}(k)$.

## Remark

After selecting $\lambda$, refit on the full dataset $\mathcal{D}$.

# Bridge Family, Lasso, Elastic Net

# Bridge Regression: $\ell_p$ Regularization

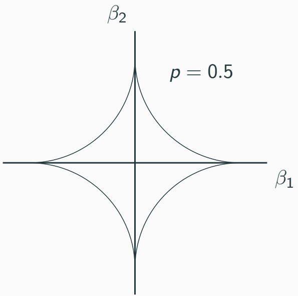
For $x = (x_1, \ldots, x_d)^\top$, define

$$\|x\|_p = \Big( \sum_{j=1}^{d} |x_j|^p \Big)^{1/p} \quad (p \geq 1).$$

**Two observations**

- If $1 \leq p < \infty$: $\|\cdot\|_p$ is a norm and $\{x : \|x\|_p \leq t\}$ is convex.
- If $0 < p < 1$: $\|\cdot\|_p$ is not a norm and the constraint set is non-convex.

# Geometric Intuition: $\ell_p$ Balls




**Takeaway**

As $p$ decreases, the $\ell_p$ ball becomes "pointier" along coordinate axes $\Rightarrow$ encourages sparsity.

# Bridge Estimator Family

**Definition (penalized form)**

For $0 < p < \infty$ and $\lambda > 0$,

$$\widehat{\beta}^{\text{bridge}} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \; \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_p^p.$$

**Special cases**

- $p = 2 \quad \Rightarrow$ ridge regression.
- $p = 1 \quad \Rightarrow$ lasso (most important case).

# Lasso: Primal and Dual

## Primal (constraint)

$$\widehat{\beta}^{\,t} = \underset{\|\beta\|_1 \le t}{\arg\min} \; \|\mathbf{Y} - \mathbf{X}\beta\|_2^2.$$

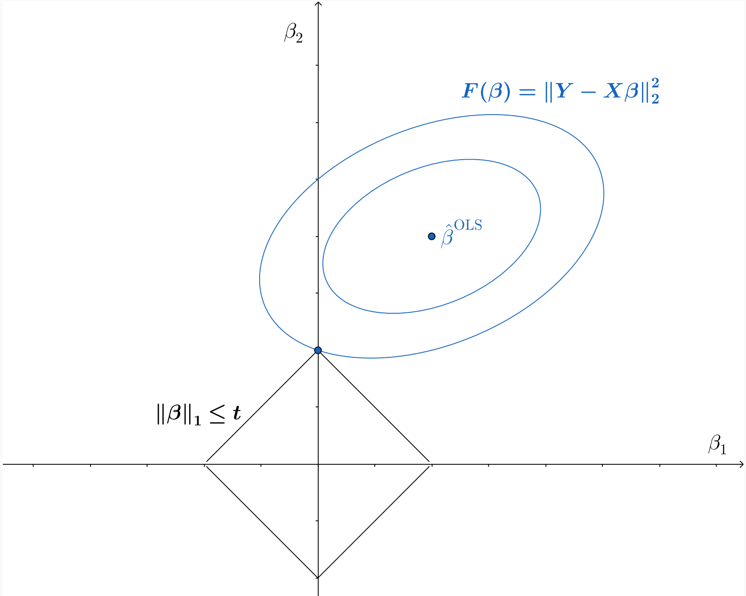
## Dual (penalty)

$$\widehat{\beta}^{\,\lambda} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \; \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1.$$

## Key Idea (Sparsity)

Lasso often yields many coefficients exactly equal to 0, enabling variable selection.

# Geometric Picture of Lasso



# Example: Lasso as Variable Selection

## Setup

Suppose the true regression function is

$$f(X) = \beta_1 X_1 + \cdots + \beta_d X_d, \quad \text{with } \beta_2 = 0.$$

## Takeaway

With a suitable $\lambda > 0$, lasso can yield $\widehat{\beta}_2^{\lambda} = 0$ and thus select variables.

## Remark

Larger $\lambda$ typically produces a sparser solution.

# Elastic Net

## Motivation

- Ridge: handles collinearity well, strongly convex, but not sparse.
- Lasso: sparse and convex, but can struggle with collinearity.

## Definition

For $\lambda > 0$ and $0 \leq \alpha \leq 1$,

$$\widehat{\beta}^{\text{Elastic}} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \; \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\Big(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2\Big).$$

## Remark

$\alpha = 1$ gives lasso; $\alpha = 0$ gives ridge.

# From Discriminative to Generative Modeling

## Discriminative vs Generative

By Bayes factorization,

$$p(y, x) = p(y \mid x)\, p(x).$$

**Discriminative modeling**

Model $p(y \mid x)$ directly (e.g., logistic regression), ignore $p(x)$.

**Generative modeling**

Model the joint mechanism via $p(x \mid y)$ and $p(y)$ (hence also $p(x)$), enabling generation of new $x$.

**Remark**

If the LDA model is correct, LDA can be more statistically efficient.

# Generative Route to Classification via Bayes

We can write

$$\mathbb{P}(Y=+1 \mid X=x) = \frac{\mathbb{P}(x \mid Y=+1)\mathbb{P}(Y=+1)}{\mathbb{P}(x \mid Y=+1)\mathbb{P}(Y=+1) + \mathbb{P}(x \mid Y=-1)\mathbb{P}(Y=-1)}.$$

Define

$$\eta \triangleq \mathbb{P}(Y=+1), \qquad p_+(x) \triangleq \mathbb{P}(x \mid Y=+1), \qquad p_-(x) \triangleq \mathbb{P}(x \mid Y=-1).$$

Then

$$\mathbb{P}(Y=+1 \mid X=x) = \frac{p_+(x)\eta}{p_+(x)\eta + p_-(x)(1-\eta)}.$$

**Key Idea (What to estimate)**

To implement Bayes classification, estimate $\eta$, $p_+(x)$, and $p_-(x)$ from data.

# MLE for the Class Prior $\eta$

With i.i.d. data $(X_i, Y_i)_{i=1}^n$, let

$$n_+ = \sum_{i=1}^n \mathbb{I}(Y_i = +1), \qquad n_- = n - n_+.$$

## Log-likelihood terms involving $\eta$

$$\sum_{i:Y_i=+1} \log \eta + \sum_{i:Y_i=-1} \log(1-\eta).$$

## MLE

$$\widehat{\eta}^{\text{MLE}} = \frac{n_+}{n}.$$

# Discriminant Analysis: QDA and LDA

# Quadratic Discriminant Analysis (QDA)

## QDA (Gaussian class-conditional densities)

Assume

$$p_+(x) = \mathcal{N}(\mu_+, \Sigma_+), \qquad p_-(x) = \mathcal{N}(\mu_-, \Sigma_-).$$

## Decision rule idea

Classify by comparing $\mathbb{P}(Y = +1 \mid X = x)$ and $\mathbb{P}(Y = -1 \mid X = x)$, equivalently compare the log-likelihood ratio plus prior term.

## Key Idea (Why "quadratic"?)

The boundary involves quadratic forms $(x - \mu_\pm)^\top \Sigma_\pm^{-1} (x - \mu_\pm)$.

# QDA Decision Boundary (from your derivation)

Under QDA, $\mathbb{P}(Y = +1 \mid X = x) > \mathbb{P}(Y = -1 \mid X = x)$ is equivalent to

$$\frac{1}{2}\log\frac{|\Sigma_-|}{|\Sigma_+|} + \frac{1}{2}(x-\mu_-)^\top\Sigma_-^{-1}(x-\mu_-) - \frac{1}{2}(x-\mu_+)^\top\Sigma_+^{-1}(x-\mu_+) + \log\frac{\eta}{1-\eta} > 0.$$

Define Mahalanobis distances

$$r_-(x) = \sqrt{(x-\mu_-)^\top\Sigma_-^{-1}(x-\mu_-)}, \quad r_+(x) = \sqrt{(x-\mu_+)^\top\Sigma_+^{-1}(x-\mu_+)}.$$

**Bayes rule in QDA form**

$$h^*(x) = \begin{cases} +1, & \frac{1}{2}r_-^2(x) - \frac{1}{2}r_+^2(x) + \frac{1}{2}\log\frac{|\Sigma_-|}{|\Sigma_+|} + \log\frac{\eta}{1-\eta} > 0, \\ -1, & \text{otherwise.} \end{cases}$$

# QDA Parameter MLEs

Let $n_+ = \sum_{i=1}^{n} \mathbb{I}(Y_i = +1)$ and $n_- = \sum_{i=1}^{n} \mathbb{I}(Y_i = -1)$.

## MLEs

$$\widehat{\mu}_+^{\mathrm{MLE}} = \frac{1}{n_+} \sum_{i:Y_i=+1} X_i, \qquad \widehat{\mu}_-^{\mathrm{MLE}} = \frac{1}{n_-} \sum_{i:Y_i=-1} X_i,$$

$$\widehat{\Sigma}_+^{\mathrm{MLE}} = \frac{1}{n_+} \sum_{i:Y_i=+1} (X_i - \widehat{\mu}_+)(X_i - \widehat{\mu}_+)^\top,$$

$$\widehat{\Sigma}_-^{\mathrm{MLE}} = \frac{1}{n_-} \sum_{i:Y_i=-1} (X_i - \widehat{\mu}_-)(X_i - \widehat{\mu}_-)^\top.$$

## Remark

(QDA has many parameters; this matters a lot in high dimension.)

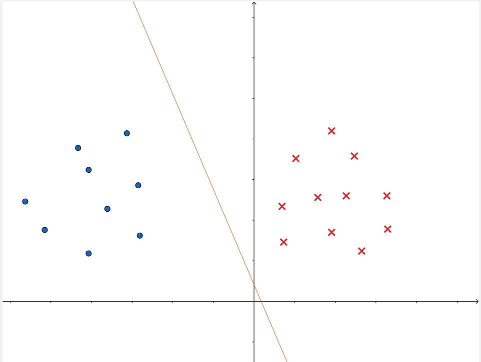
# Linear Discriminant Analysis (LDA)

## Definition

LDA is the special case of QDA with shared covariance: $\Sigma_+ = \Sigma_- = \Sigma$.

## Consequence

The decision boundary becomes linear in $x$.



# LDA Boundary is Linear

Under $\Sigma_+ = \Sigma_- = \Sigma$, the QDA condition simplifies to

$$(\mu_+ - \mu_-)^\top \Sigma^{-1} x + \frac{1}{2}\mu_-^\top \Sigma^{-1}\mu_- - \frac{1}{2}\mu_+^\top \Sigma^{-1}\mu_+ + \log\frac{\eta}{1-\eta} > 0.$$

## Linear form

This is $\beta^\top x + \beta_0 > 0$ with

$$\beta = (\mu_+ - \mu_-)^\top \Sigma^{-1}, \qquad \beta_0 = \frac{1}{2}\mu_-^\top \Sigma^{-1}\mu_- - \frac{1}{2}\mu_+^\top \Sigma^{-1}\mu_+ + \log\frac{\eta}{1-\eta}.$$

# LDA vs Linear Logistic Regression (Remark)

## Similarity

Linear logistic regression models log-odds by a linear score:

$$\log \frac{\mathbb{P}(Y = +1 \mid X = x)}{\mathbb{P}(Y = -1 \mid X = x)} = f(x), \quad f \in \{\beta^\top x + \beta_0\}.$$

LDA also implies a linear log-odds:

$$\log \frac{\mathbb{P}(Y = +1 \mid X = x)}{\mathbb{P}(Y = -1 \mid X = x)} = \beta^\top x + \beta_0.$$

## Remark

When comparing model spaces, compare *joint* distributions rather than only marginals.

# LDA Parameter MLEs

## MLEs

$$\widehat{\mu}_+^{\text{MLE}} = \frac{1}{n_+} \sum_{i: Y_i = +1} X_i, \qquad \widehat{\mu}_-^{\text{MLE}} = \frac{1}{n_-} \sum_{i: Y_i = -1} X_i,$$

$$\widehat{\Sigma}^{\text{MLE}} = \frac{n_+ \widehat{\Sigma}_+ + n_- \widehat{\Sigma}_-}{n_+ + n_-},$$

where

$$\widehat{\Sigma}_+ = \frac{1}{n_+} \sum_{i: Y_i = +1} (X_i - \widehat{\mu}_+)(X_i - \widehat{\mu}_+)^\top, \quad \widehat{\Sigma}_- = \frac{1}{n_-} \sum_{i: Y_i = -1} (X_i - \widehat{\mu}_-)(X_i - \widehat{\mu}_-)^\top.$$

# Naive Bayes and DLDA in High Dimension

# Naive Bayes Regularization

**Naive Bayes assumption (class-conditional independence)**

For $x = (x_1, \dots, x_d)^\top$,

$$\mathbb{P}(x \mid Y = +1) = \prod_{j=1}^{d} \mathbb{P}(x_j \mid Y = +1), \qquad \mathbb{P}(x \mid Y = -1) = \prod_{j=1}^{d} \mathbb{P}(x_j \mid Y = -1).$$

**Key Idea (Why it helps)**

Reduces the number of parameters dramatically by turning a $d$-dimensional density into $d$ univariate models.

# Naive Bayes Log-Odds Decomposition

Under naive Bayes,

$$\log \frac{\mathbb{P}(Y = +1 \mid X = x)}{\mathbb{P}(Y = -1 \mid X = x)} = \sum_{j=1}^{d} \log \frac{\mathbb{P}(x_j \mid Y = +1)}{\mathbb{P}(x_j \mid Y = -1)} + \log \frac{\eta}{1 - \eta}.$$

**Additive score**

Define $f_j(x_j) = \log \frac{\mathbb{P}(x_j|Y=+1)}{\mathbb{P}(x_j|Y=-1)}$. Then the classifier is based on

$$\sum_{j=1}^{d} f_j(x_j) + \log \frac{\eta}{1 - \eta}.$$

# Example: Diagonal LDA (DLDA)

## DLDA model

Assume LDA but with diagonal covariance

$$\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d^2).$$

Equivalently, coordinates are conditionally independent given the class.

## Coordinate-wise Gaussians

For each $j$,

$$X_j \mid Y = +1 \sim \mathcal{N}(\mu_{+j}, \sigma_j^2), \qquad X_j \mid Y = -1 \sim \mathcal{N}(\mu_{-j}, \sigma_j^2).$$

# DLDA MLEs (from your notes)

## Means

$$\widehat{\mu}_+^{\mathrm{MLE}} = \frac{1}{n_+} \sum_{i:Y_i=+1} X_i, \qquad \widehat{\mu}_-^{\mathrm{MLE}} = \frac{1}{n_-} \sum_{i:Y_i=-1} X_i.$$

## Diagonal covariance

$$\widehat{\Sigma}^{\mathrm{MLE}} = \mathrm{diag}(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_d^2), \qquad \widehat{\sigma}_j^2 = \frac{n_+ \widehat{S}_{+j}^2 + n_- \widehat{S}_{-j}^2}{n_+ + n_-},$$

where

$$\widehat{S}_{+j}^2 = \frac{1}{n_+} \sum_{i:Y_i=+1} (X_{ij} - \widehat{\mu}_{+j})^2, \quad \widehat{S}_{-j}^2 = \frac{1}{n_-} \sum_{i:Y_i=-1} (X_{ij} - \widehat{\mu}_{-j})^2.$$

# If Features Are Categorical

**Discrete generative models**

If $X_j$ is categorical, we can model $\mathbb{P}(X_j \mid Y)$ using a discrete distribution (e.g., Bernoulli for binary, multinomial for multi-category).

**Key Idea**

Naive Bayes naturally supports mixing continuous and categorical features by choosing appropriate univariate models for each coordinate.

# Number of Free Parameters (Model Complexity)

## Number of Free Paremters

- **Full QDA** $(\Sigma_+, \Sigma_-, \mu_+, \mu_-, \eta)$:

$$d(d+1) + 2d + 1$$

  (since each symmetric $\Sigma_\pm$ has $d(d+1)/2$ parameters).

- **Full LDA** $(\Sigma, \mu_+, \mu_-, \eta)$:

$$\frac{d(d+1)}{2} + 2d + 1$$

- **DLDA** $(\sigma_1^2, \ldots, \sigma_d^2, \mu_+, \mu_-, \eta)$:

$$3d + 1.$$

## Key Idea (High-dimensional lesson)

Regularization (e.g., diagonal/naive Bayes) reduces parameters and can improve performance when $d$ is large.

# Wrap-up

## Wrap-Up

- When $d > n$, OLS breaks: $\mathbf{X}^\top \mathbf{X}$ is not invertible.
- Ridge fixes this via $\ell_2$ regularization; closed form exists for $\lambda > 0$.
- Model selection: data splitting and $J$-fold cross-validation.
- Bridge family connects ridge ($p = 2$) and lasso ($p = 1$).
- Elastic net balances sparsity and stability under collinearity.
- Generative modeling estimates $p(x \mid y)$ and $\eta$; QDA/LDA are Gaussian instances.
- Naive Bayes/DLDA reduce parameter count dramatically in high dimension.