Regularization and Shrinkage: Why do They Matter?

💻 Content covered:

- Motivation
- From linear regression to Ridge and Lasso
- Connections with Singular Value Decomposition and Principle Component Analysis
- Example

👾 This lecture will be held online on Microsoft Teams.

🔴 The event will be recorded and will be publicly available.

🎉 Attendance is FREE for members! Whether you are a student at Imperial College or not, sign up to be a member at www.icdss.club/joinus

⭐ We encourage participants of this workshop to have looked at our previous sessions on YouTube. Prerequisites: basic understanding of Bayesian statistics

📖 A schedule of our lecture series is currently available

Motivation and Setup

Given covariates $x_i = (x_{i1}, x_{i2}, \dots, x_{id})^\intercal \in \mathbb{R}^d$ and response variables $y_i, i = 1, \dots, n$, so that

$$y_i = f(x_i) + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise.

For convenience, we write everything in matrix form: for $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$,

$$Y = f(X) + \epsilon,$$

where $Y = (y1, \dots, y_n)^\intercal$, and $X \in \mathbb{R}^{n \times d}$ whose i-th row is x_i.

E.g. When $f(X) = \alpha_0 \mathbf{1} + X\beta_0, Y = X\beta_0 + \epsilon$, and $(\alpha_0, \beta_0) \in \mathbb{R}^{d+1}$ are the true parameters.

Want to learn f.

Linear regression

Assume $Y = \alpha_0 \mathbf{1} + x\beta_0$.

Least-square (LS) solution: Find the optimal $(\hat{\alpha}, \hat{\beta}) \in \mathbb{R}^{d+1}$ by minimizing the least-square error:

$$(\hat{\alpha}, \hat{\beta}) = \operatorname{argmin}_{(\alpha,\beta) \in \mathbb{R}^{d+1}} \|Y - \alpha \mathbf{1} - X\beta\|_2^2.$$

Let $\tilde{X} = [\mathbf{1}\ X]$ be the design matrix with an intercept. Simple linear algebra gives the LS solution:

$$(\hat{\alpha}, \hat{\beta}^\intercal)^\intercal = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^\intercal Y.$$

Caveat: For the LS solution to be well-defined, the $d \times d$ matrix $\tilde{X}^T \tilde{X}$ needs to be invertible, or equivalently, $\tilde{X}$ need to have full-rank ($n \geq d$).

Problem: What about when $d >> n$?

E.g. 1. In cancer prediction problems, it is common to have thousands of gene expressions as your covariates, but only a few hundreds of patients' record.

E.g. 2. In a customer behaviour analysis where you are given whether a customer has purchased a product from Amazon as the response, and some features (type of product, price, time of visit etc.) as covariates. You don't really want to *predict* buy/not buy, but to understand which features, amongst a handful of them, are most correlated to the purchase behaviour.

Summary

LS is good, but does not give us an answer when:

1. The problem is high-dimensional and we have more features than cases.
2. We are interested in selecting features that are most "important".

Ridge Regression

Ridge solution: Find the optimal $(\hat{\alpha}, \hat{\beta}) \in \mathbb{R}^{d+1}$ by minimizing the least-square error **with L_2-penalty**:

$$(\hat{\alpha}, \hat{\beta}) = \operatorname{argmin}_{(\alpha,\beta)\in\mathbb{R}^{d+1}} \|Y - \alpha\mathbf{1} - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

where $\|\beta\|_2^2 = \sum_{j=1}^{d} \beta_j^2$ is called the penalty/regularization term, and $\lambda > 0$ is a hyperparameter we need to choose (often by cross-validation).

Why adding in a regularization helps?

Under some conditions (Y is centred, columns of X are standardized), the Ridge solution can be derived analytically:

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} Y_i = 0,$$

$$\hat{\beta} = (X^\intercal X + \lambda I_n)^{-1} X^\intercal Y,$$

Key observations: Comparing with LS solution $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$,

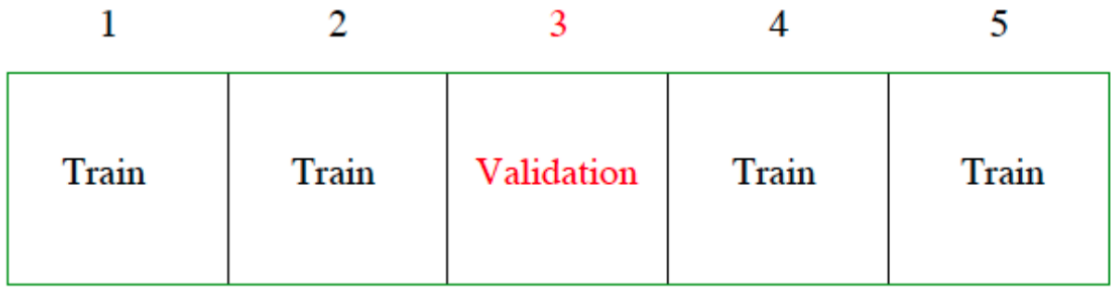
- $(X^\intercal X + \lambda I_n)$ is always invertible for **any** X, as long as $\lambda > 0$. So the Ridge solution is always well-defined.
- Adding the penalty term **shrinks** the fitted coefficients in $\hat{\beta}$ towards zero (more on this later).
- $\hat{\beta}$ is now biased, but always has a smaller variance for a judicious choice of λ (bias-variance trade-off):

Mean-Square Error = Variance + Bias^2.

Choosing λ

The optimal λ is often chosen by cross-validation:

1. Split training data into various subsets, called **folds**.
2. For each λ over a pre-defined grid of values $\lambda_1, \ldots, \lambda_k$, calculate Ridge solution from all but one folds, compute out-of-sample error on the rest fold, and repeat to get an averaged loss.
3. Pick the λ that gave the smallest averaged cross-validation loss over all folds.



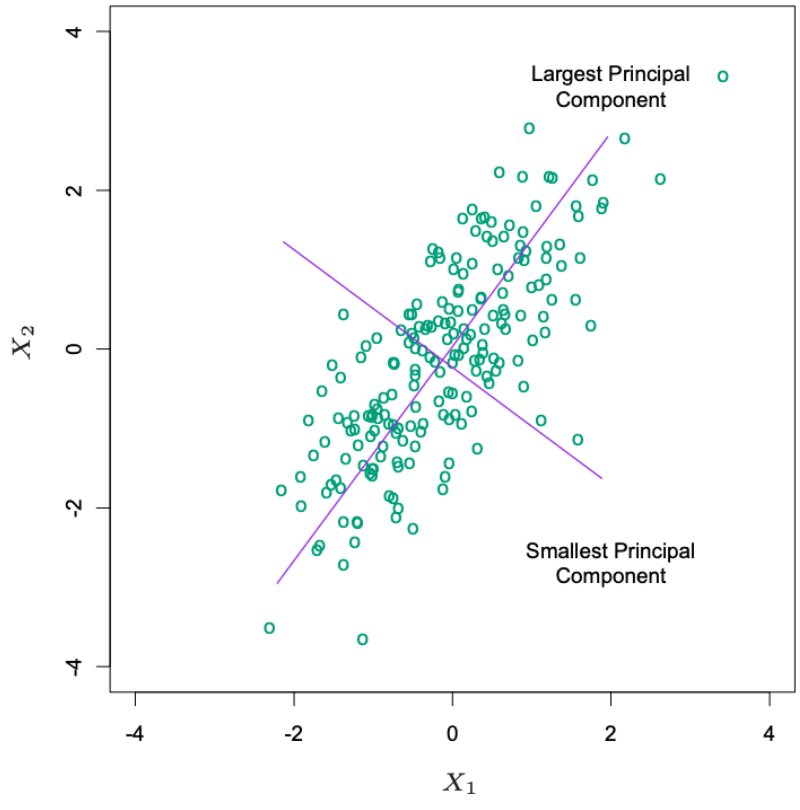
Connections with PCA

Idea of Principal Value Analaysis (PCA):

Find the directions along which the **features** X with the largest variance (i.e. most informative), and only look at the first few of them.

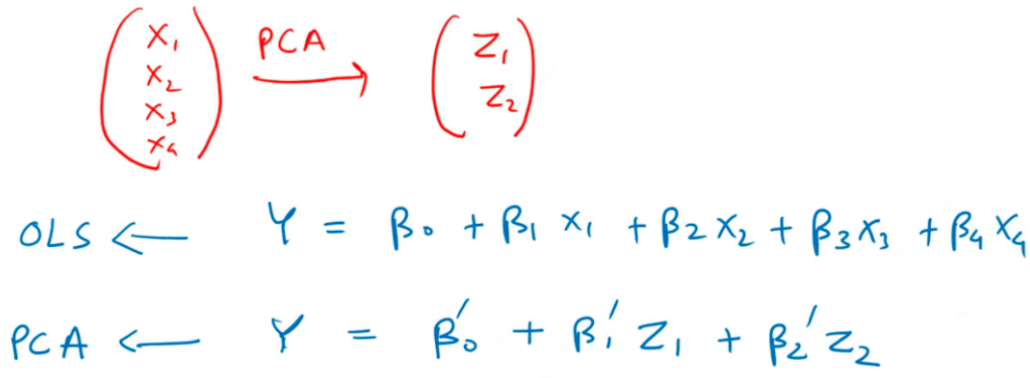
The variance is quantified by the **eigenvalues** of the matrix $X^\intercal X$, and the directions are given by their **eigenvectors**, called principal components (PCs).

Principal component regression: Use the top, say s, eigenvectors as the covariates, and perform least-square fit to find $\hat{\beta}$.



Principal component regression:

- 1: Perform PCA to create PCs as our new input features
- 2: Use these PCs as input features to train our model for a least-square fit.
- 3: Transform these PCs back to the original input features, in order to make predictions on the actual dataset.



Connections with PCA

Adding the penalty term in Ridge regression effectively shrinks the Ridge solution $\hat{\beta}$ according to the **eigenvalues** of the matrix $X^\intercal X$.

- Let $\hat{\beta}^{(LS)}$ = LS solution, $\hat{\beta}_s^{(PC)}$ = PC regression with s PCs, $\hat{\beta}_\lambda^{(R)}$ = Ridge solution with regularization parameter $\lambda > 0$.
- Let $D_1 \geq D_2 \geq \ldots \geq D_d$ be the eigenvalues of $X^\intercal X$ with corresponding eigenvectors $u_1, \ldots, u_d$.

Fact: The fitted values can be rewritten as

$$X\hat{\beta}^{(LS)} = \sum_{j=1}^{d} (u_j^\intercal Y) u_j$$

$$X\hat{\beta}_s^{(PC)} = \sum_{j=1}^{s} (u_j^\intercal Y) u_j$$

$$X\hat{\beta}_\lambda^{(R)} = \sum_{j=1}^{d} \frac{D_j^2}{D_j^2 + \lambda} (u_j^\intercal Y) u_j$$

Connections with PCA

Key observations:

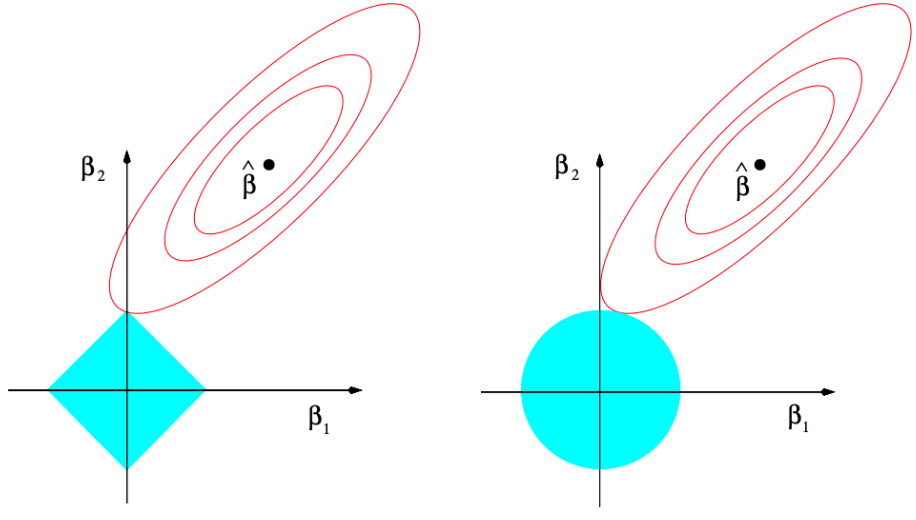
- Ridge shrinks the directions with the smallest eigenvalues the most.
- $\lambda \uparrow$, shrinkage $\uparrow$

Lasso Regression

Lasso solution: Find the optimal $(\hat{\alpha}, \hat{\beta}) \in \mathbb{R}^{d+1}$ by minimizing the least-square error **with L_1-penalty**:

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin}_{(\alpha,\beta)\in\mathbb{R}^{d+1}} \frac{1}{2}\|Y - \alpha\mathbf{1} - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

where $\|\beta\|_1 = \sum_{j=1}^{d} |\beta_j|$, $\lambda > 0$ is a hyperparameter we need to choose.



Key observations:

- Lasso is more likely to give rise to $\hat{\beta}$ that are **exactly** zero.
- Lasso solutions has nice theoretical properties: with judicious choice of λ and under regularity conditoins, $\hat{\beta} \approx \beta_0$ with high probability.
- Lasso can be combined with Ridge to give the *elastic net* penalty: for $\alpha \in [0, 1]$

$$\lambda\left(\alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1\right).$$