Chapter 6: Linear Model Selection and Regularization

Yonghyun Kwon

Department of Mathematics, Korea Military Academy

Subset Selection

Why Not Just Use Ordinary Least Squares?

Recall the linear model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

Two Motivations for Alternatives to OLS

- **Prediction Accuracy**: OLS estimates tend to have low bias but can have high variance, especially when p is large relative to n. Shrinking or setting some coefficients to zero can substantially reduce variance with a small increase in bias.
- **Model Interpretability**: Automatically performing *variable selection*—setting irrelevant coefficients to zero—yields a simpler, more interpretable model.

Three Classes of Methods

1. *Subset Selection*: Identify a subset of the p predictors believed to be related to the response. Fit OLS on the reduced set.

2. *Shrinkage (Regularization)*: Fit a model using all p predictors, but constrain/shrink the coefficient estimates toward zero. Reduces variance and can perform variable selection.

3. *Dimension Reduction*: Project the p predictors into an M-dimensional subspace ($M < p$) via linear combinations, then fit OLS using these M projections.

Best Subset Selection

Algorithm

1. Let $\mathcal{M}_0$ be the *null model* (intercept only, predicts $\bar{y}$).
2. For $k = 1, 2, \ldots, p$:
   (a) Fit all $\binom{p}{k}$ models with exactly k predictors.
   (b) Pick the best among these—call it $\mathcal{M}_k$—defined by smallest RSS (equivalently, largest R^2).
3. Select the single best model from $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ using *cross-validated prediction error*, C_p (AIC), BIC, or adjusted R^2.

Note: Step 2 identifies the best model of each size. Step 3 selects among models of different sizes.

Best Subset Selection: Limitations

- Must evaluate 2^p total models.
- For $p = 10$: $2^{10} = 1{,}024$ models — manageable.
- For $p = 40$: $2^{40} \approx 10^{12}$ models — *computationally infeasible*.

Statistical Concern

With a large search space, we are likely to find models that appear to fit the training data well by chance, even with no true predictive power. This leads to *overfitting* and high variance of coefficient estimates.

Forward Stepwise Selection

Algorithm

1. Let $\mathcal{M}_0$ be the null model (no predictors).
2. For $k = 0, 1, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the best (smallest RSS or largest R^2) and call it $\mathcal{M}_{k+1}$.
3. Select the best model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using CV, C_p, BIC, or adjusted R^2.

Note: Forward stepwise is not guaranteed to find the globally best model.

Backward Stepwise Selection

Algorithm

1. Let $\mathcal{M}_p$ be the full model (all p predictors).
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that remove one predictor from $\mathcal{M}_k$.
   (b) Choose the best (smallest RSS) and call it $\mathcal{M}_{k-1}$.
3. Select the best from $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using CV, C_p, BIC, or adjusted R^2.

Key Constraint

Backward stepwise requires $n > p$ (so the full model can be fit).

Forward stepwise can be used even when $n < p$.

Choosing the Optimal Model

- The model with all p predictors always has the smallest RSS and largest R^2 — these measure *training error*.
- We want to minimize *test error*, not training error.

Two Approaches to Estimate Test Error:

Indirect: Adjust Training Error

C_p, AIC, BIC, Adjusted R^2 add a penalty for model complexity.

Direct: Estimate Test Error

Validation set or k-fold cross-validation applied to each candidate model.

Model Selection Criteria

Mallow's C_p

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right),$$

where d = number of predictors and $\hat{\sigma}^2$ = estimate of $\mathrm{Var}(\varepsilon)$.

AIC (Akaike Information Criterion)

$$\text{AIC} = -2\log L + 2d,$$

where L is the maximized likelihood. For Gaussian errors, AIC $\propto C_p$.

BIC (Bayesian Information Criterion)

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n)\cdot d\hat{\sigma}^2\right).$$

Since $\log n > 2$ for $n > 7$, *BIC penalizes larger models more heavily* than C_p, tending to select smaller models.

Adjusted R^2

For a least squares model with d predictors:

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}.$$

- A *large* value of adjusted R^2 indicates a model with low test error. (Opposite direction from C_p, AIC, BIC.)
- Maximizing adjusted R^2 is equivalent to minimizing $\dfrac{\text{RSS}}{n-d-1}$.
- As d increases, RSS decreases but $n-d-1$ also decreases, so the ratio may increase.
- Unlike R^2, adjusted R^2 *penalizes unnecessary variables*.

Validation and Cross-Validation for Model Selection

- Subset selection procedures produce a sequence $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$. We select $\hat{k}$ that minimizes estimated test error.
- *Advantages over C_p/AIC/BIC*:
  - Provides a direct estimate of test error.
  - Does not require estimation of σ^2.
  - Applicable even when model degrees of freedom are hard to define.

One-Standard-Error Rule

When several models have similar cross-validation errors, select the simplest model whose CV error is within one standard error of the minimum. This guards against over-selection.

Shrinkage Methods

Motivation for Shrinkage

- Subset selection methods fit OLS on a chosen subset of predictors.
- As an alternative, fit using all p predictors but *constrain or shrink* the coefficient estimates toward zero.
- Shrinking coefficients reduces their variance. The slight increase in bias can be more than offset, leading to a lower MSE overall.

We study two main shrinkage methods:

Ridge Regression

ℓ_2 penalty: $\lambda \sum_j \beta_j^2$

All coefficients shrunk toward 0; none exactly 0.

Lasso

ℓ_1 penalty: $\lambda \sum_j |\beta_j|$

Some coefficients can be set exactly to 0 (variable selection).

Ridge Regression: Formulation

OLS minimizes:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right)^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}).$$

Ridge regression minimizes:

$$\ell(\boldsymbol{\beta}) = \underbrace{(\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta})}_{\text{RSS}} + \underbrace{\lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}}_{\text{shrinkage penalty}},$$

where $\lambda \geq 0$ is a *tuning parameter*.

- $\lambda = 0$: reduces to OLS.
- $\lambda \to \infty$: all $\hat{\beta}_j \to 0$.
- Selecting a good λ is critical — use *cross-validation*.

Note: Ridge regression does not penalize the intercept β_0. Predictors should be standardized before fitting.

Derivation of the Ridge Estimator

Result (HW 3, Problem 1A)

The ridge estimator minimizing $\ell(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ is:

$$\boxed{\hat{\boldsymbol{\beta}}_\lambda = (X^\top X + \lambda I_p)^{-1} X^\top \boldsymbol{y}.}$$

Derivation. Differentiate and set to zero:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = -2X^\top(\boldsymbol{y} - X\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = \mathbf{0} \implies (X^\top X + \lambda I_p)\boldsymbol{\beta} = X^\top \boldsymbol{y}.$$

Since $X^\top X + \lambda I_p$ is positive definite for any $\lambda > 0$, it is invertible.

Bias of the Ridge Estimator

Result (HW 3, Problem 1B)

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_\lambda) = (X^\top X + \lambda I_p)^{-1} X^\top X \boldsymbol{\beta}.$$

The estimator is *biased* for $\lambda > 0$; it is unbiased only when $\lambda = 0$.

Derivation. Since $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_\lambda) = (X^\top X + \lambda I_p)^{-1} X^\top \mathrm{E}(\boldsymbol{y}) = (X^\top X + \lambda I_p)^{-1} X^\top X \boldsymbol{\beta}.$$

The *bias* is:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_\lambda) - \boldsymbol{\beta} = \left[(X^\top X + \lambda I_p)^{-1} X^\top X - I_p \right] \boldsymbol{\beta} = -\lambda (X^\top X + \lambda I_p)^{-1} \boldsymbol{\beta}.$$

As $\lambda \uparrow$, the bias increases in magnitude.

Variance of the Ridge Estimator

Result (HW 3, Problem 1C)

$$\mathrm{Var}(\hat{\beta}_\lambda) = \sigma^2 (X^\top X + \lambda I_p)^{-1} X^\top X (X^\top X + \lambda I_p)^{-1}.$$

Furthermore, $\mathrm{Var}(\hat{\beta}_\lambda) \preceq \mathrm{Var}(\hat{\beta}_0) = \sigma^2 (X^\top X)^{-1}$.

Proof sketch. Let $X^\top X = P \Lambda P^\top$ be the eigendecomposition with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_j > 0$. Then:

$$D := \mathrm{Var}(\hat{\beta}_0) - \mathrm{Var}(\hat{\beta}_\lambda) = \sigma^2 P \Delta P^\top, \quad \Delta_{jj} = \frac{\lambda(\lambda + 2\lambda_j)}{\lambda_j (\lambda_j + \lambda)^2} \geq 0.$$

Since $\Delta \succeq 0$, we have $D \succeq 0$, i.e., $\mathrm{Var}(\hat{\beta}_\lambda) \preceq \mathrm{Var}(\hat{\beta}_0)$. □

Ridge *reduces variance* relative to OLS, at the cost of introducing bias.

Bias-Variance Decomposition of MSE

Result (HW 3, Problem 1D)

Let $\boldsymbol{\mu} = \mathrm{E}[\hat{\boldsymbol{\beta}}_\lambda]$. Then:

$$\mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^\top\right] = \underbrace{\mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda)}_{\text{variance}} + \underbrace{(\boldsymbol{\mu} - \boldsymbol{\beta})(\boldsymbol{\mu} - \boldsymbol{\beta})^\top}_{\text{squared bias}}.$$

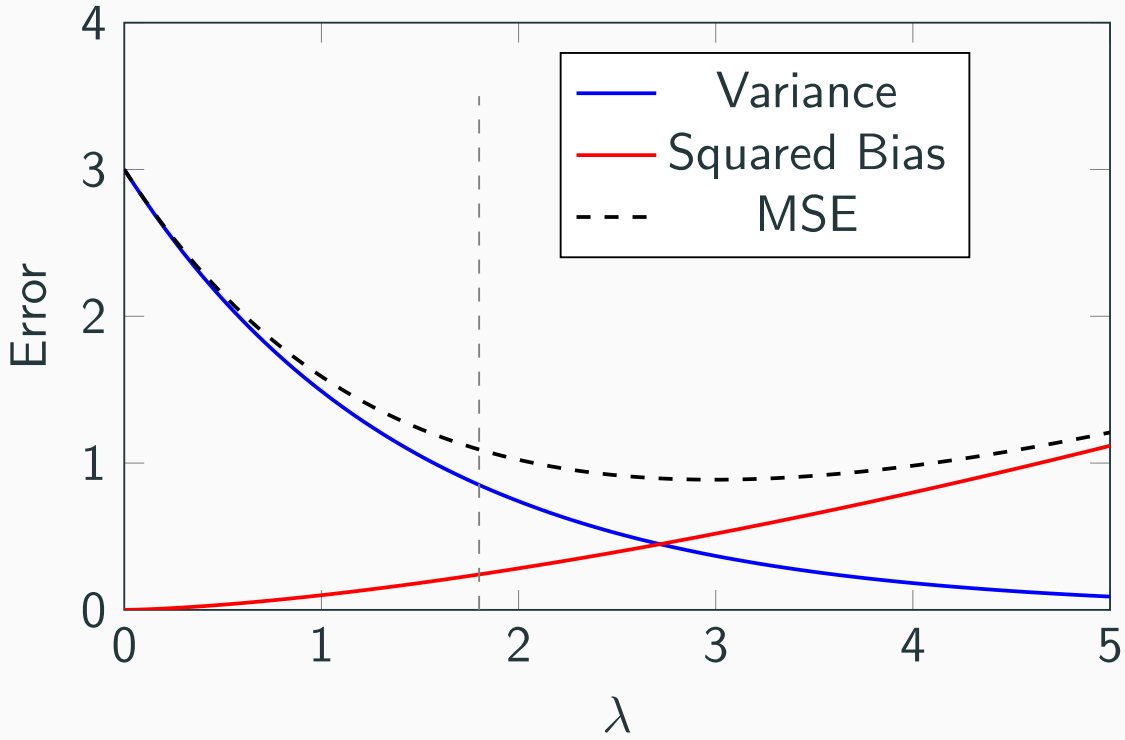
Proof. Write $\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta} = (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{\beta})$ and expand:

$$\mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^\top\right] = \mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu})^\top\right] + (\boldsymbol{\mu}-\boldsymbol{\beta})(\boldsymbol{\mu}-\boldsymbol{\beta})^\top.$$

Cross terms vanish since $\mathrm{E}[\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\mu}] = \mathbf{0}$. □

Scalar form: $\mathrm{MSE}(\hat{\beta}) = \mathrm{Var}(\hat{\beta}) + \mathrm{Bias}(\hat{\beta})^2$.

The Bias-Variance Tradeoff in Ridge Regression



As λ increases: variance $\downarrow$, bias^2 $\uparrow$. The optimal λ^* minimizes MSE.

Ridge Regression: Scaling of Predictors

Important!

OLS estimates are *scale equivariant*: multiplying X_j by c scales $\hat{\beta}_j$ by $1/c$, so $X_j\hat{\beta}_j$ is unchanged.

Ridge regression coefficient estimates depend on the scale of each predictor, because of the ℓ_2 penalty.

Solution: Standardize each predictor before fitting:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}.$$

All standardized predictors have empirical standard deviation 1, so the penalty treats them equally.

The Lasso

Ridge regression keeps all p predictors in the final model (no exact zeros).

The Lasso (Least Absolute Shrinkage and Selection Operator) minimizes:

$$\ell(\boldsymbol{\beta}) = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1.$$

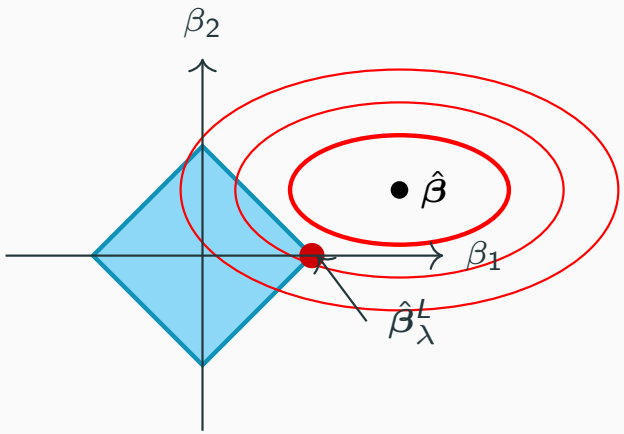
- Uses an ℓ_1 penalty ($\|\boldsymbol{\beta}\|_1 = \sum |\beta_j|$) instead of ℓ_2.
- The ℓ_1 penalty forces some coefficients to be *exactly zero* for sufficiently large λ.
- Hence lasso performs *variable selection* (sparse models), unlike ridge.

Why Does Lasso Give Sparse Solutions?

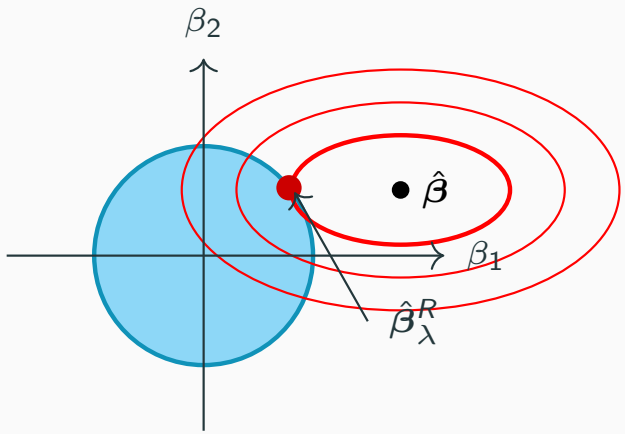
Both lasso and ridge can be written as constrained problems:

$$\min_{\boldsymbol{\beta}} \ \text{RSS} \quad \text{subject to} \quad \sum_j |\beta_j| \le s \quad \text{(Lasso)}$$

$$\min_{\boldsymbol{\beta}} \ \text{RSS} \quad \text{subject to} \quad \sum_j \beta_j^2 \le s \quad \text{(Ridge)}$$



Lasso (ℓ_1)



Ridge (ℓ_2)

The ℓ_1 ball has *corners on axes*: the RSS ellipse often first touches the constraint at a corner, giving $\hat{\beta}_j = 0$.

Comparing Ridge and Lasso

Ridge Regression

- Penalty: $\lambda\|\boldsymbol{\beta}\|_2^2$
- Coefficients shrunken toward 0, never exactly 0
- No variable selection
- Closed-form solution
- Better when many predictors have small but nonzero effects

Lasso

- Penalty: $\lambda\|\boldsymbol{\beta}\|_1$
- Some coefficients become exactly 0
- Automatic variable selection (sparse model)
- No closed-form (requires convex optimization)
- Better when only a few predictors are truly related to response

Neither universally dominates. Use cross-validation to choose.

Selecting the Tuning Parameter λ

Cross-Validation Procedure

1. Choose a grid of λ values: $\lambda_1 > \lambda_2 > \cdots > \lambda_K$.
2. For each λ_k, compute k-fold cross-validation error.
3. Select $\hat{\lambda}$ minimizing the cross-validation error.
4. Refit the model using all observations with $\hat{\lambda}$.

Practical Notes

- Typical grid: log-scale from 10^{-3} to 10^{3}.
- Use `glmnet` package in R.
- `cv.glmnet` automates cross-validation.

One-SE Rule

Can also apply the *one-standard-error rule*: choose the largest λ whose CV error is within one SE of the minimum.

Dimension Reduction Methods

Dimension Reduction: Overview

- Previous methods (subset selection, ridge, lasso) fit a model using the original predictors $X_1, \ldots, X_p$.
- *Dimension reduction* transforms the predictors first, then fits OLS on the transformed variables.

General Setup

Let $Z_1, Z_2, \ldots, Z_M$ be $M < p$ *linear combinations* of $X_1, \ldots, X_p$:

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j, \quad m = 1, \ldots, M.$$

Fit the linear regression model by OLS:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i.$$

Principal Components Regression (PCR)

Idea: Use PCA to define the M directions $Z_1, \ldots, Z_M$.

PCA Recap

- *First principal component* Z_1: the (normalized) linear combination of $X_1, \ldots, X_p$ with the largest variance. Equivalently, it minimizes the sum of squared perpendicular distances to the data.
- *Second principal component* Z_2: largest variance among all directions uncorrelated with Z_1.
- And so on. The principal components are orthogonal.

PCR: Regress y onto $Z_1, \ldots, Z_M$ (the first M principal components).

PCR: Choosing M

- When $M = p$: PCR = OLS (same fit, different parameterization).
- When M is small: high bias, low variance.
- As M increases: bias decreases, variance increases.

Selecting M by Cross-Validation

1. Fit PCR for $M = 1, 2, \ldots, p$.
2. Compute cross-validation MSE for each M.
3. Select M^* minimizing CV-MSE (or apply one-SE rule).

When is PCR useful?

PCR works well when the first few PCs capture most of the variation in X and that variation is related to y. It is most effective when predictors are highly correlated.

Partial Least Squares (PLS)

Limitation of PCR: PCA directions are chosen without reference to y. The directions with the most variance in X may not be most predictive of y.

PLS: Supervised Dimension Reduction

1. **First PLS direction** Z_1: Set $\phi_{1j} \propto \mathrm{Corr}(y, X_j)$ — assign highest weight to variables most correlated with y.
2. **Subsequent directions**: Regress each X_j on Z_1 and take residuals. Then compute the next PLS direction from residuals (ensuring orthogonality). Repeat.

- PLS identifies directions that explain both X and y.
- Number of directions M is chosen by cross-validation.
- In practice, PLS often performs similarly to PCR and ridge.

Summary: Comparison of Methods

Method	Variable Selection	Shrinkage	Tuning Parameter
Best Subset	✓		k
Forward/Backward	✓		k
Ridge		✓	λ
Lasso	✓	✓	λ
PCR			M
PLS			M

Chapter Summary

Key Takeaways

- OLS can be improved by *constraining* or *selecting* predictors.
- The bias-variance tradeoff is central: reducing variance (at the cost of some bias) often lowers test MSE.
- **Subset selection**: explicit variable selection; computationally expensive for large p.
- **Ridge**: shrinks all coefficients, never to exactly zero. Variance $\leq$ OLS variance.
- **Lasso**: shrinks some coefficients to exactly zero; performs variable selection.
- **PCR/PLS**: reduce dimensionality via linear combinations.
- Lasso is an especially active research area; related methods include the *elastic net* (combining ℓ_1 and ℓ_2 penalties).

Exercises

1. Consider a ridge regression problem. Suppose $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$.

(A) Derive the ridge estimator $\hat{\boldsymbol{\beta}}_\lambda = (X^\top X + \lambda I_p)^{-1} X^\top \boldsymbol{y}$.

(B) Find $\mathrm{E}(\hat{\boldsymbol{\beta}}_\lambda)$. For what value of λ is the estimator unbiased?

(C) Show that $\mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda) \preceq \mathrm{Var}(\hat{\boldsymbol{\beta}}_0)$ using eigendecomposition of $X^\top X$.

(D) Prove the bias-variance decomposition:

$$\mathrm{E}\left[(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^\top\right] = \mathrm{Var}(\hat{\boldsymbol{\beta}}_\lambda) + \left(\boldsymbol{\beta} - \mathrm{E}[\hat{\boldsymbol{\beta}}_\lambda]\right)\left(\boldsymbol{\beta} - \mathrm{E}[\hat{\boldsymbol{\beta}}_\lambda]\right)^\top.$$