Chapter 6: Linear Model Selection and Regularization

Comprehensive Study Guide

Exam Preparation Materials

1. Introduction and Motivation

The standard linear regression model relates a response (Y) to p predictors (X_j):

```
Y = beta_0 + beta_1*X_1 + ... + beta_p*X_p + epsilon (Equation 6.1)
```

Why Use Alternatives to Least Squares?

1. Prediction Accuracy

Least squares often has **high variance**, especially when p is large relative to n. Shrinking or constraining coefficient estimates reduces variance and improves prediction accuracy.

2. Model Interpretability

When many predictors are irrelevant, selecting a subset yields a simpler and more interpretable model.

Three main classes of methods address these issues:

- **Subset Selection**: Identify and use only a subset of predictors
- **Shrinkage (Regularization)**: Fit all predictors but constrain coefficients
- **Dimension Reduction**: Project predictors into lower-dimensional space

2. Subset Selection Methods

Subset selection involves fitting models on reduced sets of predictors.

2.1 Best Subset Selection

Idea: Search all 2^p possible subsets.

Algorithm:

- Start with the null model (M_0) with no predictors
- For each subset size k = 1, ..., p, fit all C(p,k) models
- Select the best model (M_k) with smallest RSS or largest R^2
- Choose the best model among {M_0, ..., M_p} using validation criteria

Limitation: Computationally infeasible when p >= 40. For p=10, there are 1,024 models; for p=20, over 1 million models!

2.2 Stepwise Selection

Faster but approximate search procedures.

Forward Stepwise Selection

- Start with M■ (null model)
- Add the predictor that gives greatest improvement (smallest RSS)
- Fits only 1 + p(p+1)/2 models
- Can be used when n < p

Backward Stepwise Selection

- Start with full model M_p (requires n > p)
- Remove the least useful predictor at each step
- Greedy algorithm - may not find optimal solution

2.3 Choosing the Optimal Model

Training RSS and R^2 always improve with more predictors → **not useful for model selection**.

Use criteria that account for model complexity:

Criterion	Formula/Description	Selection Rule
C_p	RSS/n + 2*d*sigma^2/n	Choose lowest C_p
BIC	RSS/n + log(n)*d*sigma^2/n	Heavier penalty; smaller models
Adjusted R^2	Penalizes unnecessary predictors	Choose largest adjusted R^2
Cross-Validation	Direct estimate of test error	Choose lowest CV error

Key points: BIC has a heavier penalty than C_p when n > 7, leading to smaller models. Cross-validation makes fewer assumptions and works even when p > n.

3. Shrinkage Methods (Regularization)

Shrinkage methods fit all predictors but **constrain or shrink** coefficients toward zero. This reduces variance at the cost of a small increase in bias.

3.1 Ridge Regression

Ridge minimizes:

```
RSS + lambda * Sum(beta_j^2) (L2 penalty)
```

Key Points:

- **lambda >= 0** controls shrinkage: lambda=0 gives least squares; lambda→infinity gives all coefficients → 0
- **L2 penalty:** ||beta||_2^2 = Sum(beta_j^2)
- **Must standardize** predictors before applying ridge (not scale invariant)
- **Does NOT perform variable selection** — coefficients shrink but never exactly zero
- **Works when p > n** — unlike least squares
- **Reduces variance** especially effective with correlated predictors

3.2 The Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) minimizes:

```
RSS + lambda * Sum(|beta_j|) (L1 penalty)
```

Key Points:

- **L1 penalty:** ||beta||_1 = Sum(|beta_j|)
- **Produces sparse solutions** — some coefficients are **exactly zero**
- **Performs variable selection** automatically
- **Geometric intuition:** L1 constraint region (diamond) has sharp corners → zeros
- **Soft-thresholding:** Small coefficients get set to exactly zero
- **No closed-form solution** — requires numerical algorithms

3.3 Ridge vs. Lasso Comparison

Feature	Ridge (L2)	Lasso (L1)
Penalty	Sum(beta^2)	Sum(|beta|)
Sparsity	No (beta != 0)	Yes (beta = 0)
Variable Selection	No	Yes
Closed Form	Yes	No
Best When	Many small effects	Few large effects
Correlated Predictors	Uses all	Picks one

3.4 Choosing the Tuning Parameter lambda

Both ridge and lasso choose lambda using **cross-validation**:

- Create a grid of lambda values (often on log scale)
- For each lambda, perform k-fold CV (typically k=5 or 10)
- Select lambda that minimizes CV error
- Refit model on full dataset using chosen lambda

Bias-Variance Trade-off: As lambda increases, variance decreases quickly while bias increases slowly. Test MSE follows a U-shaped curve with optimal lambda at the minimum.

4. Dimension Reduction Methods

These methods transform predictors into new variables (Z_m), then fit linear regression on the transformed variables.

4.1 Principal Components Regression (PCR)

Steps:

- Compute principal components Z_m (unsupervised - doesn't use Y)
- Each component is a linear combination of all X_j
- Components are orthogonal and ordered by variance explained
- Fit model using first M components (choose M via CV)

Key Properties:

- M chosen via cross-validation
- Closely related to ridge regression
- Does NOT perform variable selection (all X_j used)
- Works best when high-variance directions relate to Y
- Potential issue: May discard low-variance components important for Y

4.2 Partial Least Squares (PLS)

PLS is a **supervised** alternative to PCR that uses Y when constructing components.

- Directions chosen to explain BOTH predictor variance AND correlation with Y
- First PLS direction places highest weight on predictors most correlated with Y
- Often needs fewer components than PCR
- Better than PCR when response relates to low-variance directions

5. Considerations in High Dimensions

High-dimensional setting: p >= n or even p >> n

Examples: Genomics (p=500,000 SNPs, n=100 patients), text analysis (p=millions of words)

5.1 Problems When p >= n

- **Perfect training fit:** Can achieve RSS=0, but useless for prediction (extreme overfitting)
- **Least squares fails:** Infinitely many solutions or no unique solution
- **Traditional criteria break:** C_p, AIC, BIC formulas invalid
- **R^2=1 on training:** Meaningless as quality measure
- **Multiple testing issues:** Many predictors significant by chance alone

5.2 Solutions

Methods with regularization are essential:

- **Lasso** - Often best when few predictors truly matter (sparsity)
- **Ridge regression** - Good when many predictors with small effects
- **PCR/PLS** - Effective for dimension reduction
- **Elastic net** - Combines ridge + lasso penalties

5.3 Interpretation Challenges

- **Multicollinearity:** Many models fit data equally well
- **Selected variables may be proxies** for true predictors
- **Focus on prediction,** not individual coefficient interpretation
- **Must use CV or test sets** — training error is meaningless
- **Never report training error** as measure of performance

6. Bias-Variance Trade-Off

```
Expected Test MSE = Variance + Bias^2 + sigma^2 (irreducible error)
```

For regularization methods as lambda increases:

lambda	Flexibility	Variance	Bias	Test MSE
0	High	High	Low	High
Small	High	High	Low	Decreasing
Optimal	Medium	Medium	Medium	MINIMUM
Large	Low	Low	High	Increasing
infinity	None	0	High	High

The sweet spot: Optimal lambda balances flexibility and stability. Variance decreases faster than bias increases, leading to improved test MSE.

7. Method Comparison Summary

Method	Variable Selection?	Best Use Case
Best Subset	Yes	Small p (<=30), need optimal
Forward Stepwise	Yes	Large p, works when p>n
Ridge	No	Many correlated predictors
Lasso	Yes	Few large effects, need sparsity
PCR	No	Correlated predictors, unsupervised
PLS	No	Weak relationships with Y

8. Key Formulas for Exams

RSS (Residual Sum of Squares):

```
RSS = Sum[(y_i - y_hat_i)^2]
```

TSS (Total Sum of Squares):

```
TSS = Sum[(y_i - y_bar)^2]
```

R^2:

```
R^2 = 1 - RSS/TSS
```

Adjusted R^2:

```
Adj-R^2 = 1 - [RSS/(n-d-1)] / [TSS/(n-1)]
```

Mallow's C_p:

```
C_p = (RSS + 2*d*sigma_hat^2) / n
```

BIC:

```
BIC = [RSS + log(n)*d*sigma_hat^2] / n
```

Ridge Objective:

```
min { RSS + lambda * Sum(beta_j^2) }
```

Lasso Objective:

```
min { RSS + lambda * Sum(|beta_j|) }
```

Test MSE Decomposition:

```
E[Test MSE] = Var(y_hat) + Bias(y_hat)^2 + sigma^2
```

9. Common Exam Pitfalls to Avoid

X Using training R^2 or RSS for model selection

- ✓ Use CV, C_p, BIC, or adjusted R^2

X Forgetting to standardize before ridge/lasso

- ✓ Always standardize predictors first

X Confusing ridge and lasso

- ✓ Ridge: L2, no selection; Lasso: L1, sparse

X Using test data for model selection

- ✓ Test data ONLY for final evaluation

X Expecting PCR to always beat least squares

- ✓ PCR only helps when variance relates to Y

X Interpreting coefficients literally in high-D

- ✓ Focus on prediction, not interpretation

X Forgetting BIC penalizes more than C_p

- ✓ BIC → smaller models when n > 7

X Thinking selection finds causal relationships

- ✓ Only finds predictive associations

10. Practice Questions

1. Why can't we use R^2 to select the best subset of predictors?

2. Under what circumstances would you expect lasso to outperform ridge regression?

3. Explain why the lasso penalty can set coefficients exactly to zero, but ridge cannot.

4. What is the difference between supervised and unsupervised dimension reduction?

5. When $p > n$, why does least squares fail, but ridge regression can still be computed?

6. If two predictors are highly correlated, how do ridge and lasso handle them differently?

7. Explain the bias-variance trade-off in the context of choosing λ for ridge regression.

8. Why is standardization essential before applying ridge or lasso?

9. What is the one-standard-error rule in cross-validation, and why use it?

10. In high dimensions (p ■ n), why be cautious about interpreting selected variables?

Study Tips for Success

- Focus on understanding the **why** behind each method, not just formulas
- Practice comparing methods and knowing **when to use** each one
- Understand the **bias-variance trade-off** for all regularization methods
- Know the **geometric intuition** for why lasso creates sparse solutions
- Remember: **Training error is NOT test error** — especially in high dimensions
- Be able to explain the **differences** between similar methods (ridge vs. lasso, PCR vs. PLS)
- Understand **computational complexity** (best subset vs. forward stepwise)
- Practice explaining concepts to others — teaching solidifies understanding

Good luck with your exam!