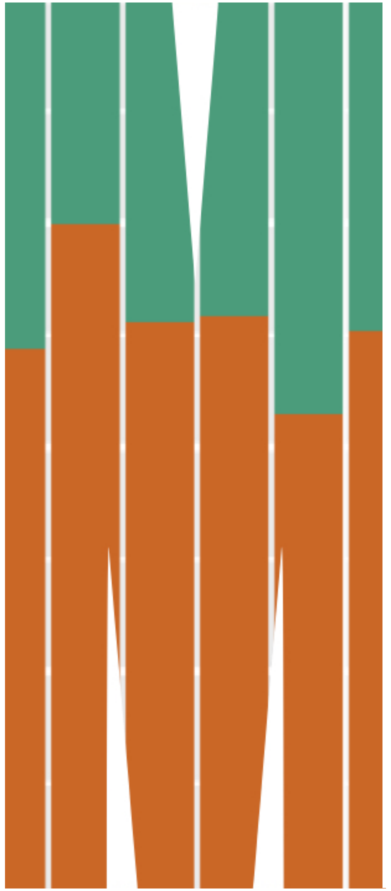


ETC3250/5250: Regularization

Semester 1, 2020

Professor Di Cook

Econometrics and Business Statistics
Monash University

Week 9 (a)

Too many variables

Fitting a linear regression model requires:

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}$$
$$\equiv \underset{\beta\in\mathbb{R}^p}{\text{minimize}}\ (y-X\beta)'(y-X\beta)$$

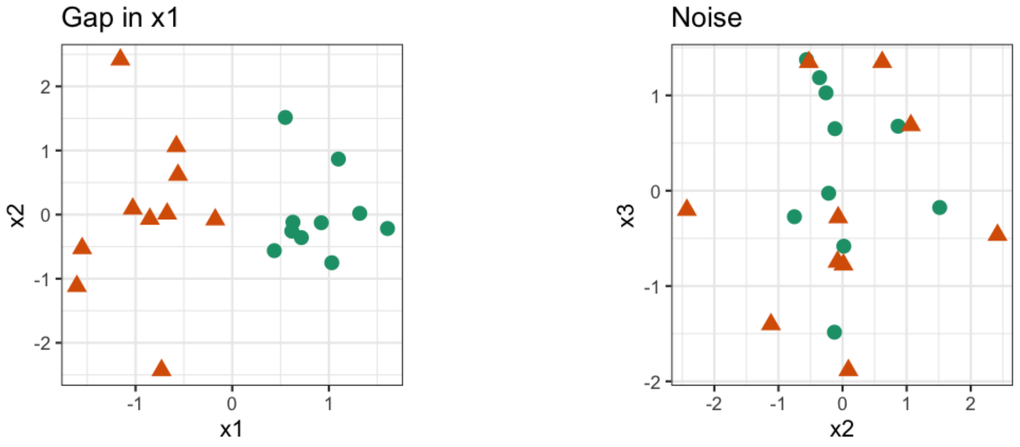
The least square solution for β is

$$\hat{\beta}=(X'X)^{-1}X'y$$

To invert a matrix, requires it to be full rank.

Example: Using simulation

- 20 observations
- 2 classes: A, B
- One variable with separation, 99 noise variables



00:23

What will be the optimal LDA coefficients?

Fit linear discriminant analysis on first two variables.

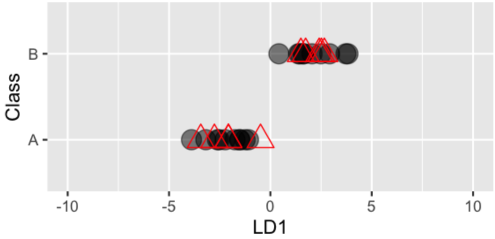
```
## Call:
## lda(cl ~ ., data = tr[, c(1:2, 101)], prior = c(0.5, 0.5))
##
## Prior probabilities of groups:
##   A   B
## 0.5 0.5
##
## Group means:
##           x1            x2
## A  0.8918346  0.0009586256
## B -0.8918346 -0.0009586256
##
## Coefficients of linear discriminants:
##            LD1
## x1 -2.41606038
## x2  0.05224863
```

Coefficient for `x1` MUCH higher than `x2`. As expected!

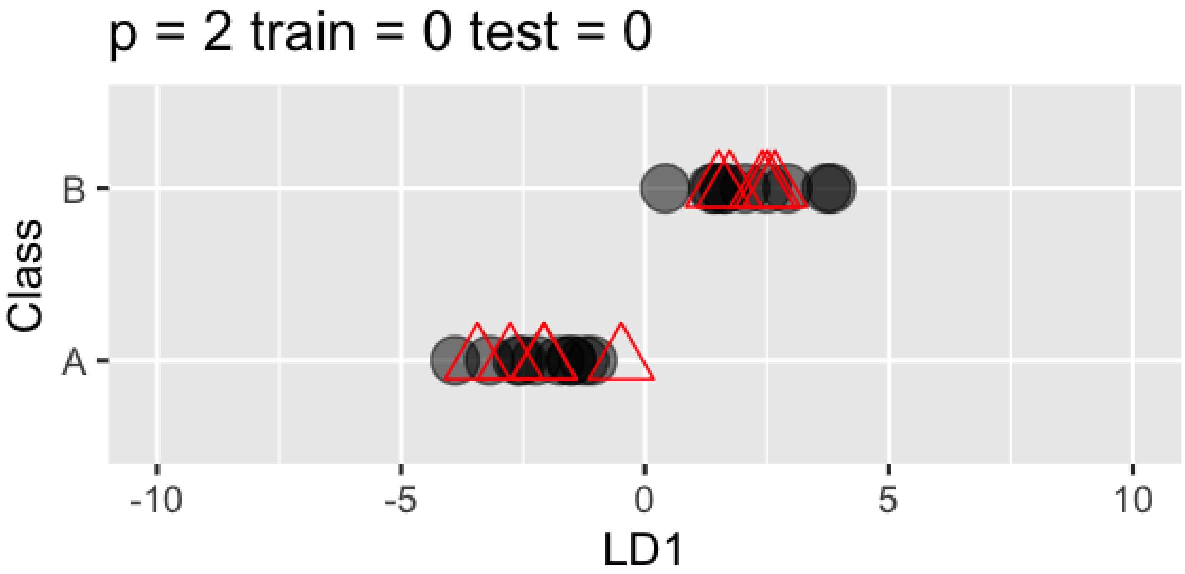
Predict the training and test sets

```
##
##      A  B
##   A 10  0
##   B  0 10
```

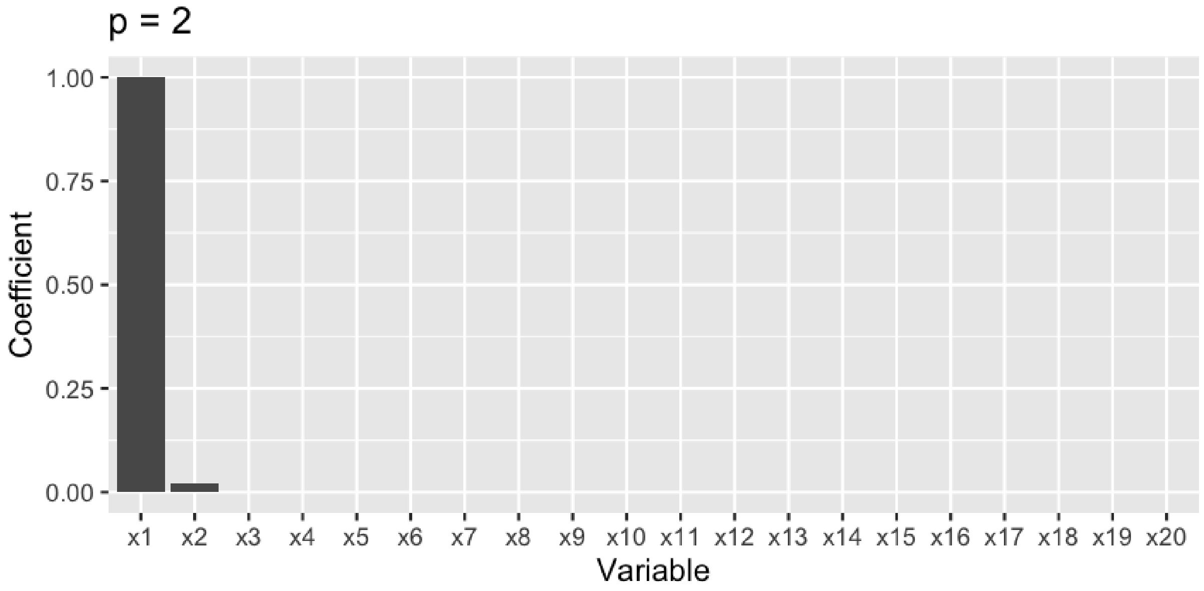
```
##
##     A B
##   A 5 0
##   B 0 5
```



What happens to test set (and predicted training values) as number of noise variables increases:



Estimated coefficients as dimensions of noise increase:



How do we tackle high-dimension, low sample size problems?

Subset selection

Identify a subset s of the p predictors, most related to response.

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} I(\beta_j \neq 0) \leq k, \quad k \geq 0.$$

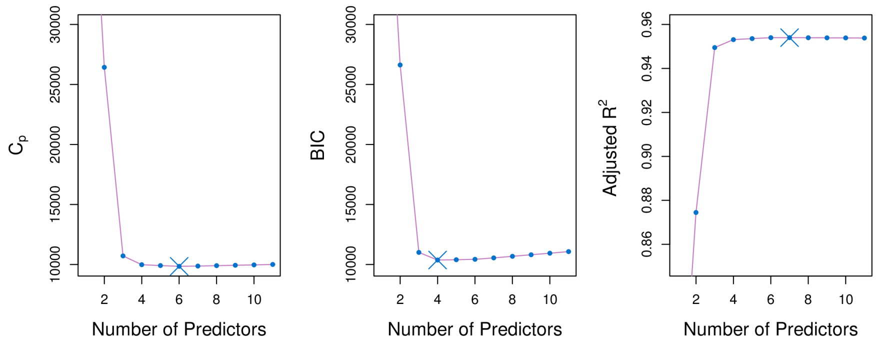
where $k \geq 0$ is a tuning parameter.

- Need to consider $\binom{p}{k}$ models containing s predictors computationally infeasible when p and s are large
- Stepwise procedures: forward, backward, etc.

Model fit statistics

These can be used to decide on choice of k.

- $MSE = RSS/n$, but the training MSE is an under-estimate of test MSE, and it will decrease with larger p.
- Methods for adjusting the training error for model size include Mallows C_p, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and adjusted R^2.



Mallows C_p

For a fitted least squares model containing d predictors, a reasonable estimate of the test MSE is:

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

where $\hat{\sigma}^2$ is an estimate of the variance of the error ε, computed from the full model containing all predictors.

The additional part penalises the training RSS to adjust for the under-estimation of test error.

AIC and BIC

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

and hence is $\propto C_p$.

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$

all tend to take on low values for models with small test error.

Adjusted R^2

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

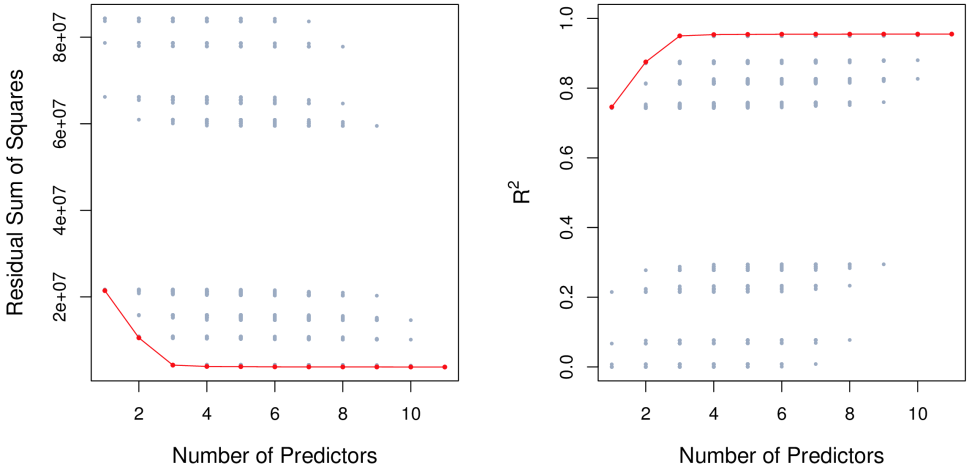
The intuition is that once all of the correct variables have been included in the model, adding additional *noise* variables will lead to only a very small decrease in RSS.

Best subset selection algorithm

1. Let $\mathcal{M}_o$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots, p$:
   a. Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   b. Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Best means smallest RSS (or largest R^2).
3. Select a single best model from among $\mathcal{M}_o, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Best subset selection algorithm

Best subset selection algorithm applied to the 11 predictors of the Credit data.



Foward stepwise selection

Forward stepwise selection is a computationally efficient alternative to best subset selection. It considers a much smaller set of models.

When $p = 20$, best subset selection requires fitting 1,048,576 models, whereas forward stepwise selection requires fitting only 211 models.

Foward stepwise selection - algorithm

1. Let $\mathcal{M}_o$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 0, 1, 2, \ldots, p - 1$:
   a. Consider all $p - k$ models that augment $\mathcal{M}_k$ with *one additional predictor*.
   b. Pick the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Best means smallest RSS (or largest R^2).
3. Select a single best model from among $\mathcal{M}_o, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Backwise stepwise selection

- Backward stepwise starts with all variables in the model, and removes the variable with smallest RSS.
- Forward and backwards stepwise procedures are not guaranteed to provide the best model.
- Backwards stepwise requires that $n > p$, but forward stepwise does not, and can stop adding variables once $n(< p)$ is reached.

Shrinkage methods

Shrinkage methods fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks some of the coefficient estimates towards zero.

There are two main methods: Ridge regression and Lasso.

Ridge regression

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Least squares:

$$\underset{\beta}{\mathrm{minimize}}\ \mathrm{RSS}$$

Ridge regression:

$$\underset{\beta}{\mathrm{minimize}}\ \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

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

Ridge regression

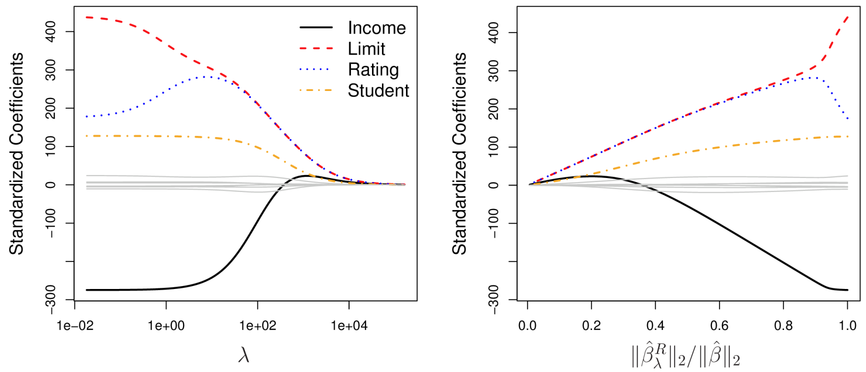
$$\lambda \sum_{j=1}^{p} \beta_j^2$$

is called a shrinkage penalty. It is small when $\beta_1, \ldots, \beta_p$ are close to 0.

λ serves as a tuning parameter, controlling the relative impact of these two terms on the regression coefficient estimates. When it is 0, the penalty term has no effect on the fit.

Ridge regression will produce a different set of coefficients for each λ, call them $\hat{\beta}_{\lambda}^{R}$. Tuning λ, typically by cross-validation, is critical component of fitting the model.

Standardized ridge regression coefficients for the Credit data set.



(Chapter6/6.4.pdf)

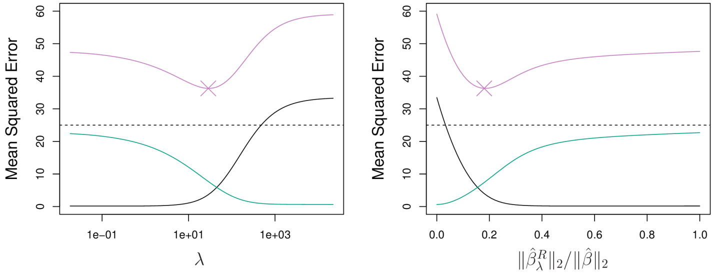
- $p = 10$
- Left side of plot corresponds to least squares.
- When λ is extremely large, then all of the ridge coefficient estimates are basically zero, which is the null model.
- 4 of 10 variables have larger coefficients, and one, Rating, initially increases with λ.
- Right-side plot, x-axis indicates amount the coefficients shrink to 0, value of 1 indicates LS.

The scale of variables can affect ridge regression performance.

It is important to standardise the scale of predictors prior to ridge regression.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sigma_{x_j}}$$

Simulation scenario! Ridge regression improves on least squares, for large number of variables, in the bias-variance tradeoff. It sacrifices some bias for the benefit of decreased variance.



bias variance test error

(Chapter6/6.5.pdf)

The Lasso

Ridge regression:

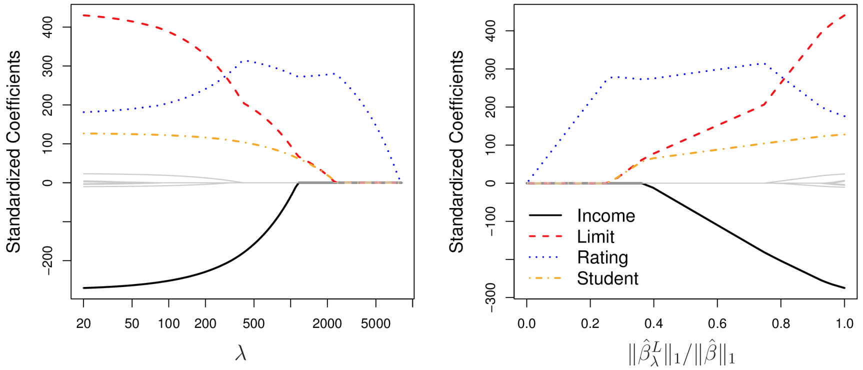
$$\underset{\beta}{\text{minimize}}\ \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso:

$$\underset{\beta}{\text{minimize}}\ \mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

and same $\lambda \geq 0$ is a tuning parameter.

Standardized lasso coefficients for the Credit data set.

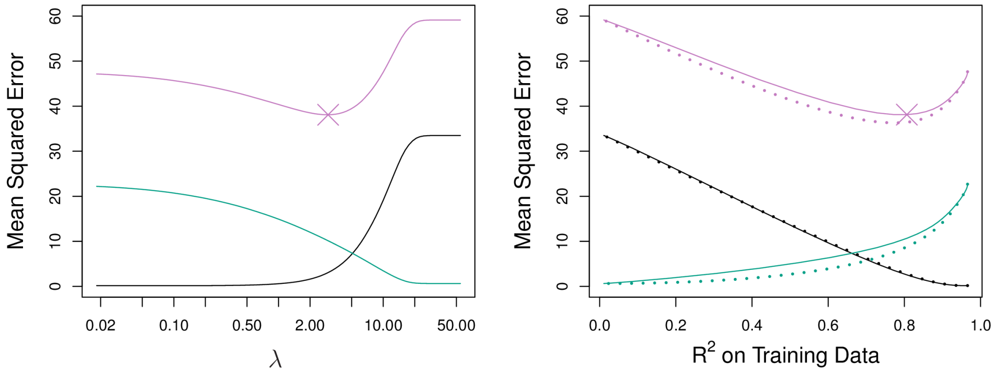


(Chapter6/6.6.pdf)

- $p = 10$
- Has the effect of forcing some variables exactly to 0.
- Cleaner solution than ridge regression.

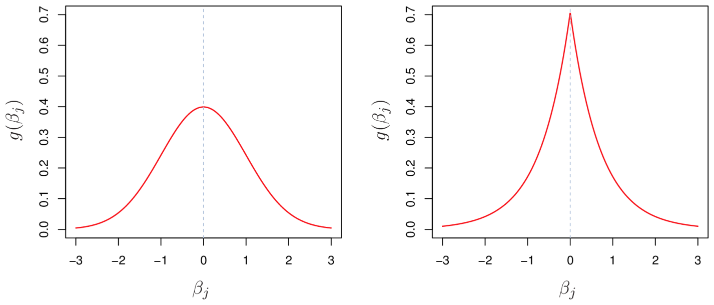
Simulation scenario!

Bias-variance tradeoff with lasso, and comparison against ridge regression.



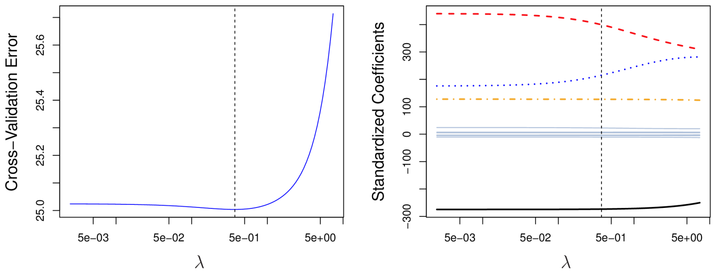
Bias Variance Test error

Bayesian interpretation: Ridge regression is the posterior mode for β under a Gaussian prior (left); The lasso is the posterior mode for β under a double-exponential prior (right).



(Chapter6/6.11.pdf)

Cross-validation on the Credit example, yields a suggestion to use $\lambda = 0.5$ for ridge regression model.

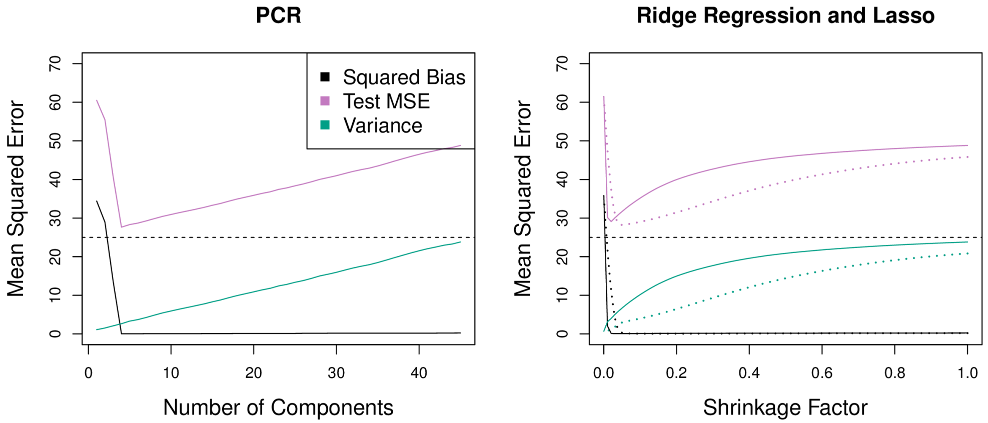


(Chapter6/6.12.pdf)

Principal component regression

The principal components regression (PCR) approach involves constructing the first M principal components, $Z_1, \ldots, Z_M$, and then using these components as the predictors in a linear regression model, that is fit using least squares.

PCR, ridge regression, and the lasso compared on simulated data. PCR does well when the response is related to few PCs.

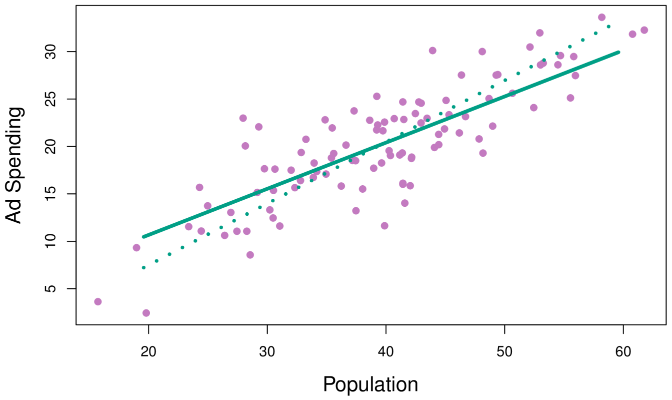


Bias Variance Test error

(Chapter6/6.19.pdf)

Partial least squares

Partial least squares (PLS), a supervised alternative to PCR.



Two predictors are shown: Solid line is PLS, dashed line is PCR.

(Chapter 6 / 6.21)

Partial least squares

1. Standardise all variables
2. Find $Z_1 = \phi_{1j} X_j$ by setting ϕ_{1j} to be the coefficient from a simple linear regression model $Y \sim X_j$.
3. To find Z_2, first regress each variable on Z_1 and use the residuals, call these X_j^r. Then find $Z_2 = \phi_{2j} X_j^r$ by setting ϕ_{2j} to be the coefficient from a simple linear regression model $Y \sim X_j^r$.
4. Repeat steps 2-3 until we have $Z_1, \dots, Z_M$.

Final model fitted for Y using $Z_1, \dots, Z_M$.

Performance is no better than ridge regression or PCR. Can reduce bias, has potential to increase variance. PLS is similar to partial regression, where new variables are first regressed on predictors that are already in the model, and it is the residuals that are used.

Penalised LDA

Recall: LDA involves the eigen decomposition of $\Sigma^{-1}\Sigma_B$, where

$$\Sigma_B = \frac{1}{K}\sum_{i=1}^{K}(\mu_i - \mu)(\mu_i - \mu)'$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_i)(x_i - \mu_i)'$$

The eigen-decomposition is an analytical solution to a sequential optimisation problem:

$$\underset{\beta_k}{\text{maximize}}\beta_k^T\hat{\Sigma}_B\beta_k$$

$$\text{subject to } \beta_k^T\hat{\Sigma}\beta_k \leq 1, \beta_k^T\hat{\Sigma}\beta_j = 0 \;\forall i < k$$

👩‍💻 Made by a human with a computer

Slides at https://iml.numbat.space.

Code and data at https://github.com/numbats/iml.

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.