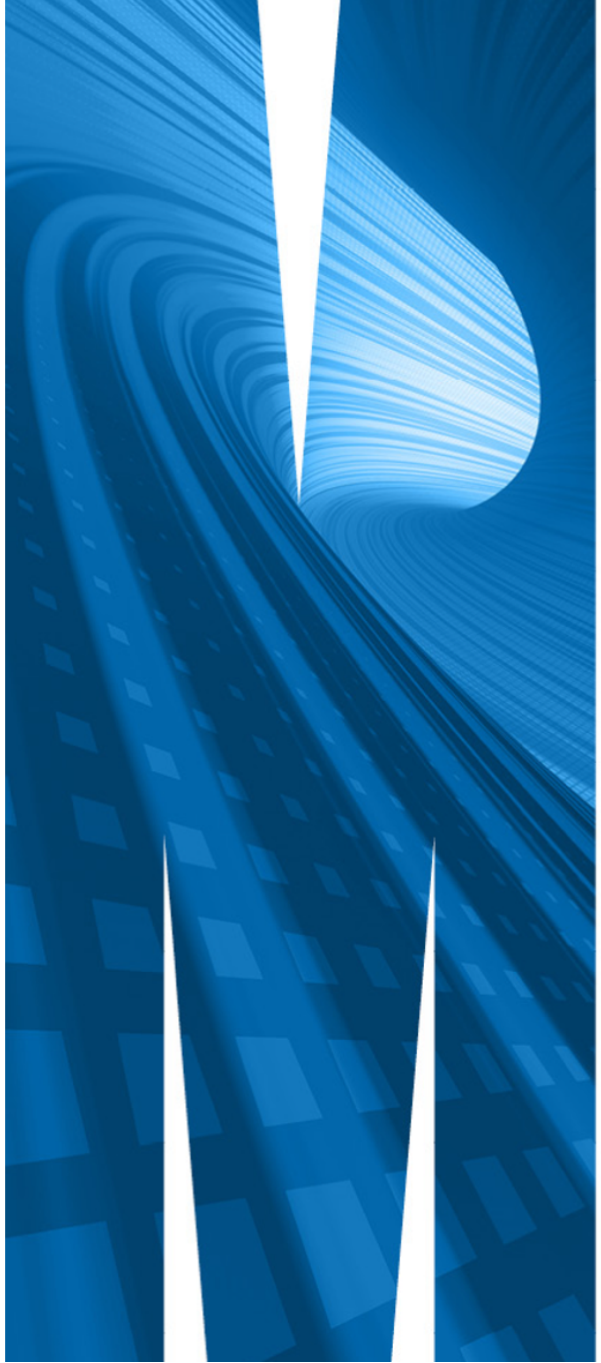
MONASH University

ETC3250/5250: Introduction to Machine Learning

Regularisation



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 8b

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 βs

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

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

Fit linear discriminant analysis on first two variables.

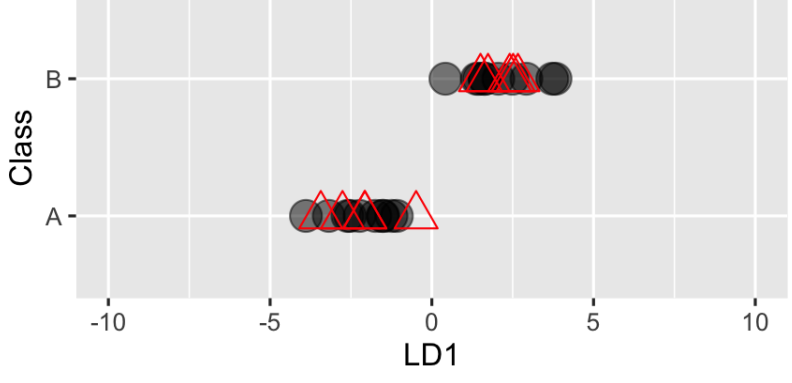
```
## Call:
## lda(cl ~ ., data = tr[, c(1:2, 101)], prior = c(0.5, 0.5))
##
## Prior probabilities of groups:
##   A   B
## 0.5 0.5
##
## Group means:
##           x1            x2
## A  0.8918346  0.0009586256
## B -0.8918346 -0.0009586256
##
## Coefficients of linear discriminants:
##           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

Coefficient for x1 MUCH higher than x2. As expected!

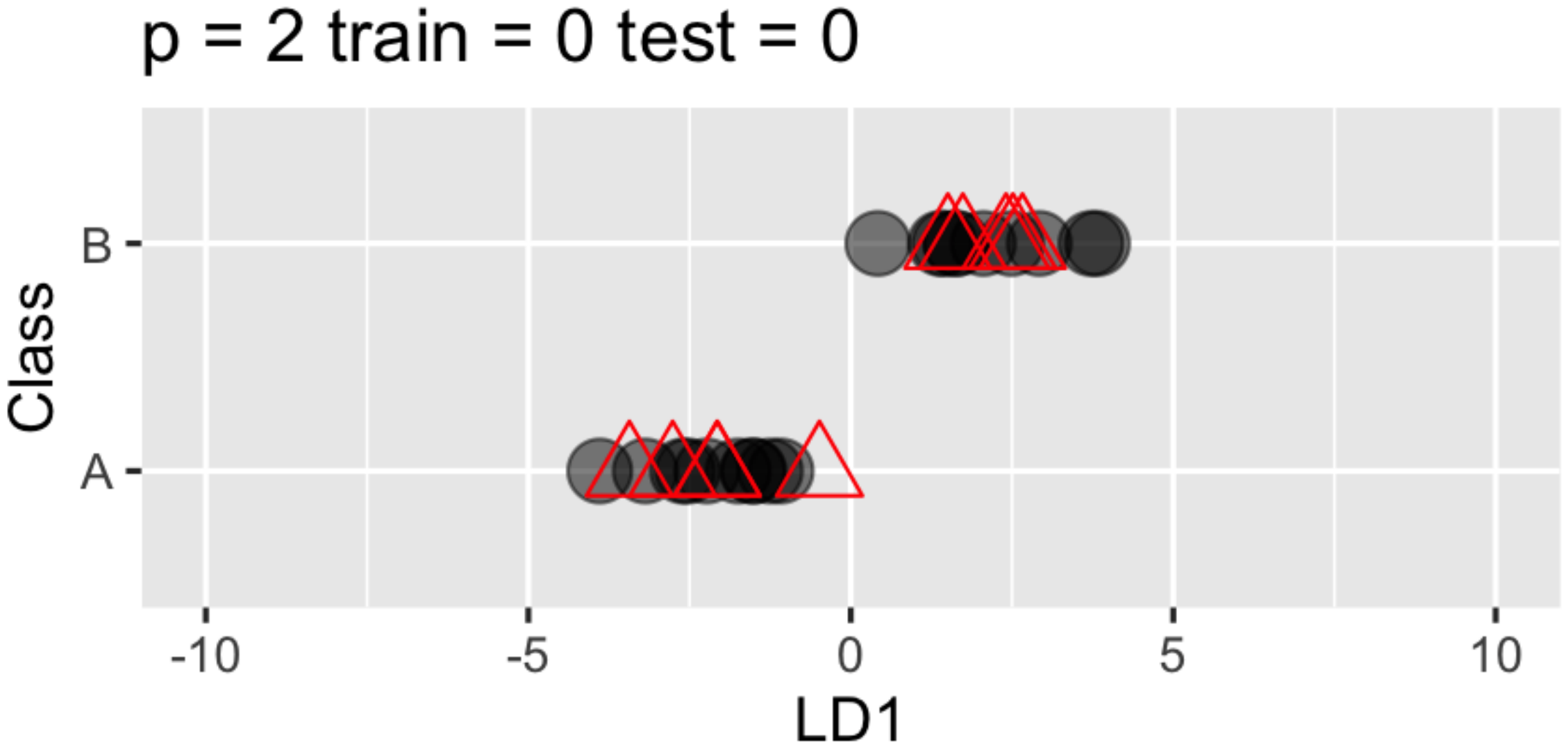
Predict the training and test sets

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

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

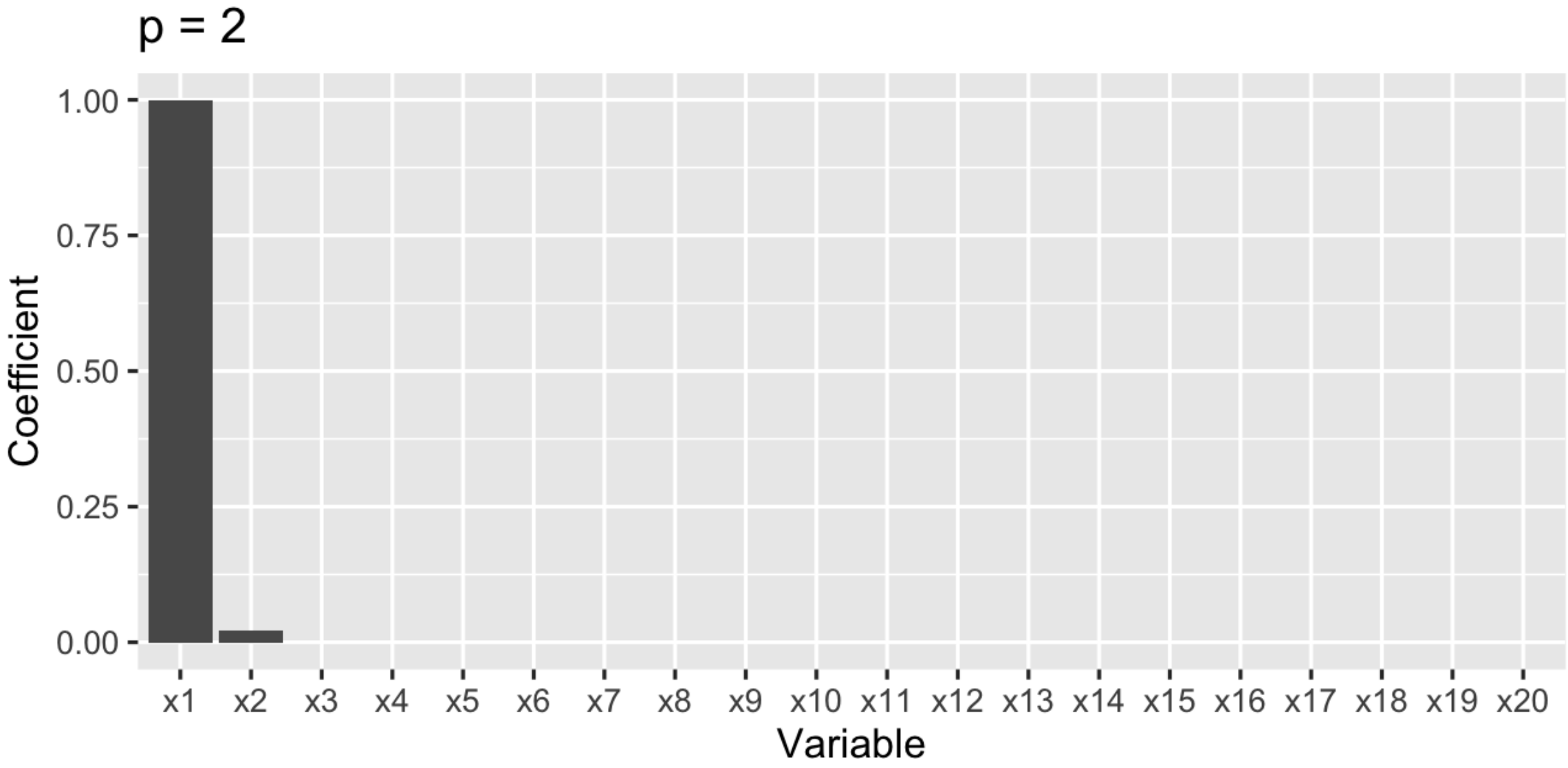


What happens to test set (and predicted training values) as number of noise variables increases


p = 2 train = 0 test = 0
Class
B
A
-10
-5
0
5
10
LD1

What happens to the estimated coefficients as dimensions of noise increase?

Remember, the noise variables should have coefficient = ZERO.


p = 2
1.00
0.75
0.50
0.25
0.00
Coefficient
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10
x11
x12
x13
x14
x15
x16
x17
x18
x19
x20
Variable

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

Subset selection

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

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

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

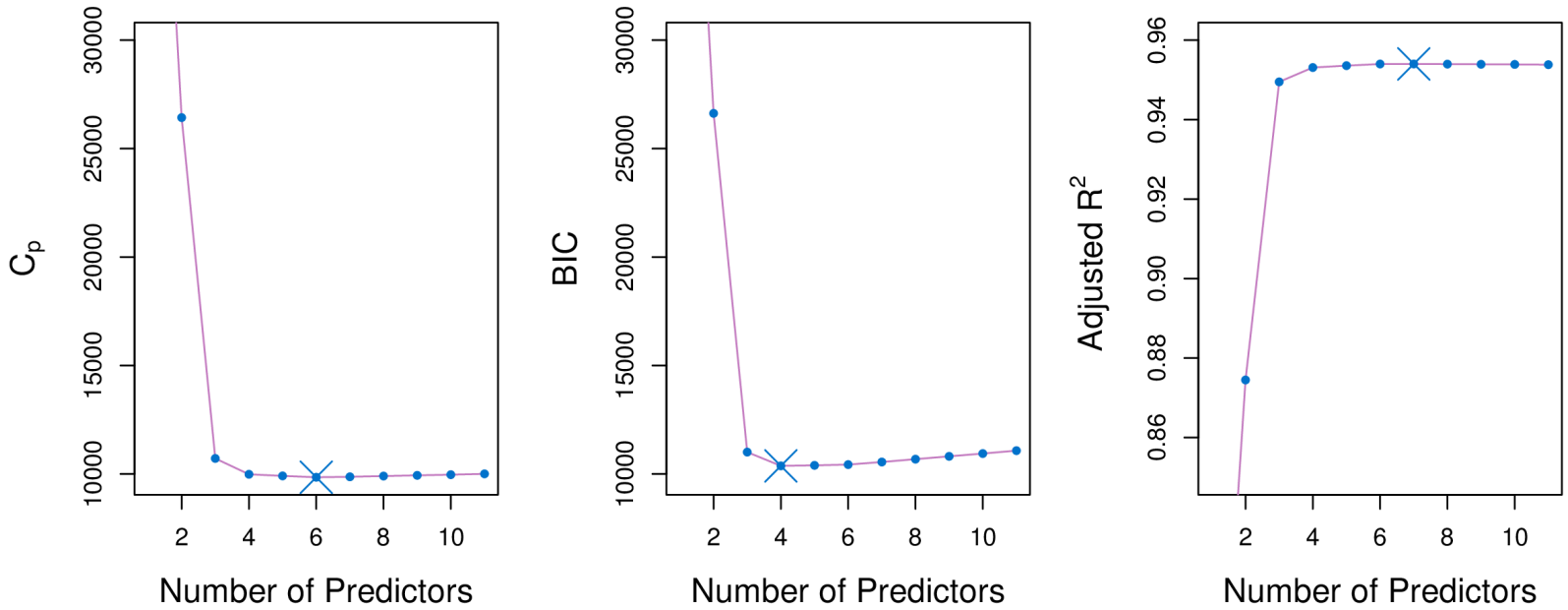
where $k \geq 0$ is a tuning parameter.

- Need to consider $\binom{p}{s}$ 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

- $MSE = RSS/n$ under-estimate of test MSE 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

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

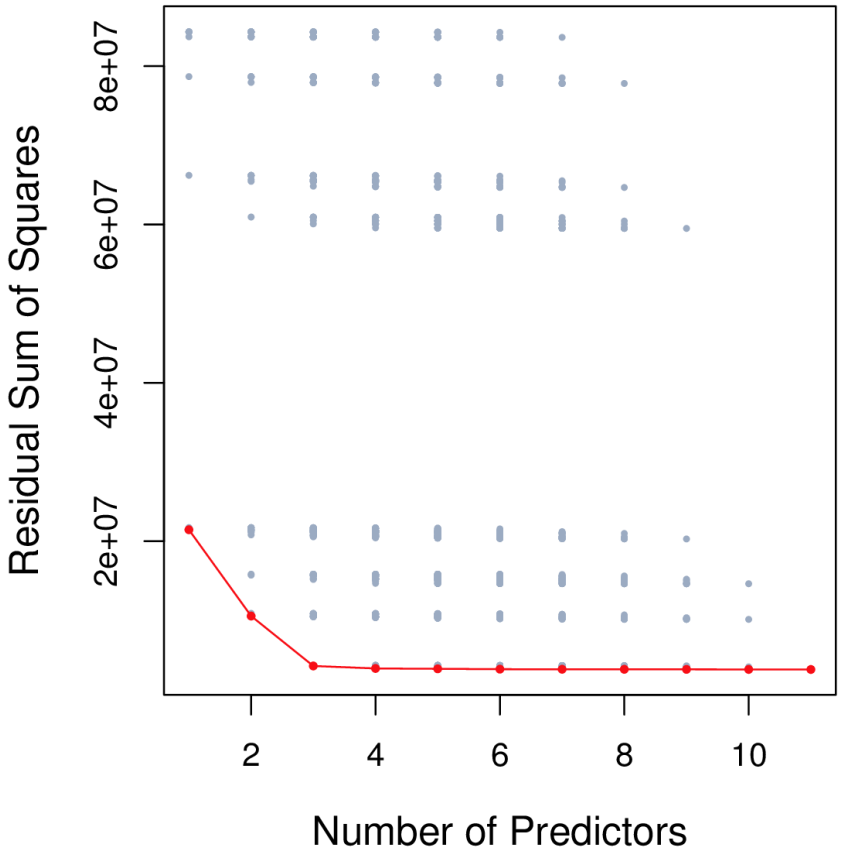
The intuition is that once all of the correct variables have been included in the model, adding additional *noise* variables will lead to only a very small decrease in RSS.

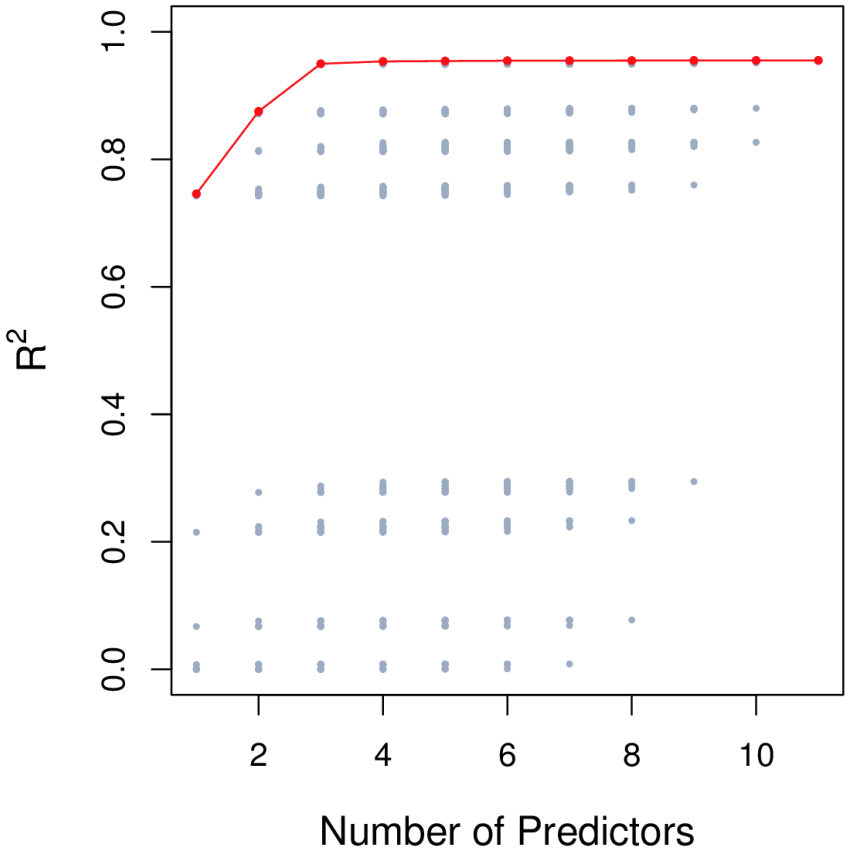
Best subset selection algorithm

1. Let $\mathcal{M}_0$ 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 $\mathcal{M}_0, \ldots, \mathcal{M}_p$ 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.





(Chapter 6/6.1)

Forward 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.

Forward stepwise selection - algorithm

1. Let $\mathcal{M}_0$ 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 $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross- validated prediction error, C_p (AIC), BIC, or adjusted R^2

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

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

Least squares:

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

Ridge regression:

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

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

Ridge regression

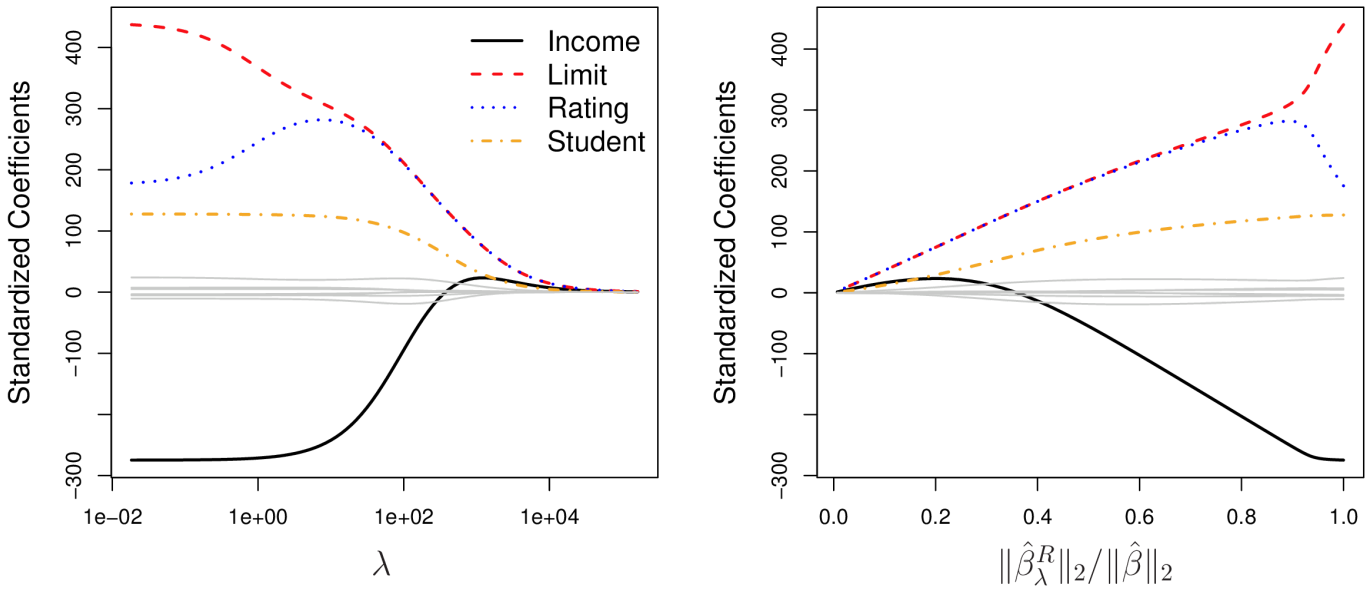
$$\lambda \sum_{j=1}^{p} \beta_j^2$$

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)

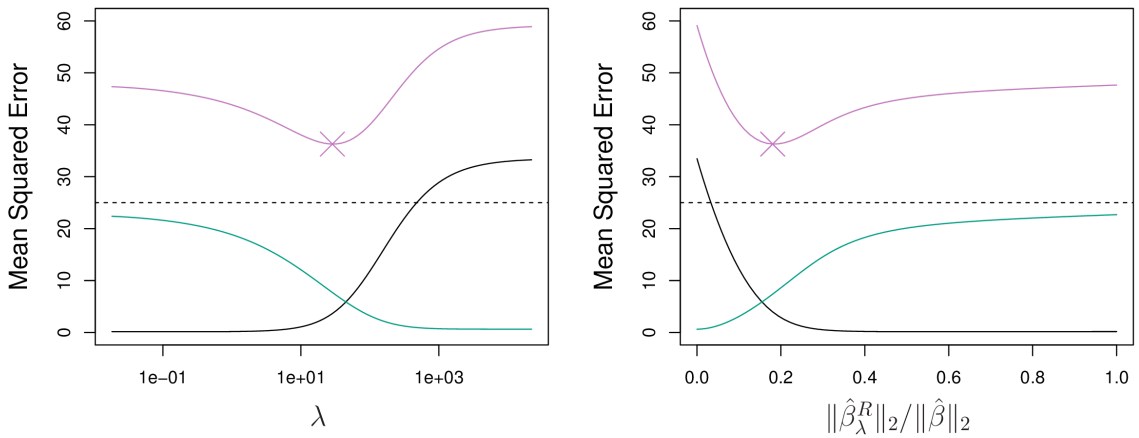
- $p = 10$
- Left side of plot corresponds to least squares.
- When λs extremely large, then all of the ridge coefficient estimates are basically zero, which is the null model.
- 4 of 10 variables have larger coefficients, and one, Rating, initially increases with λ
- Right-side plot, x-axis indicates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

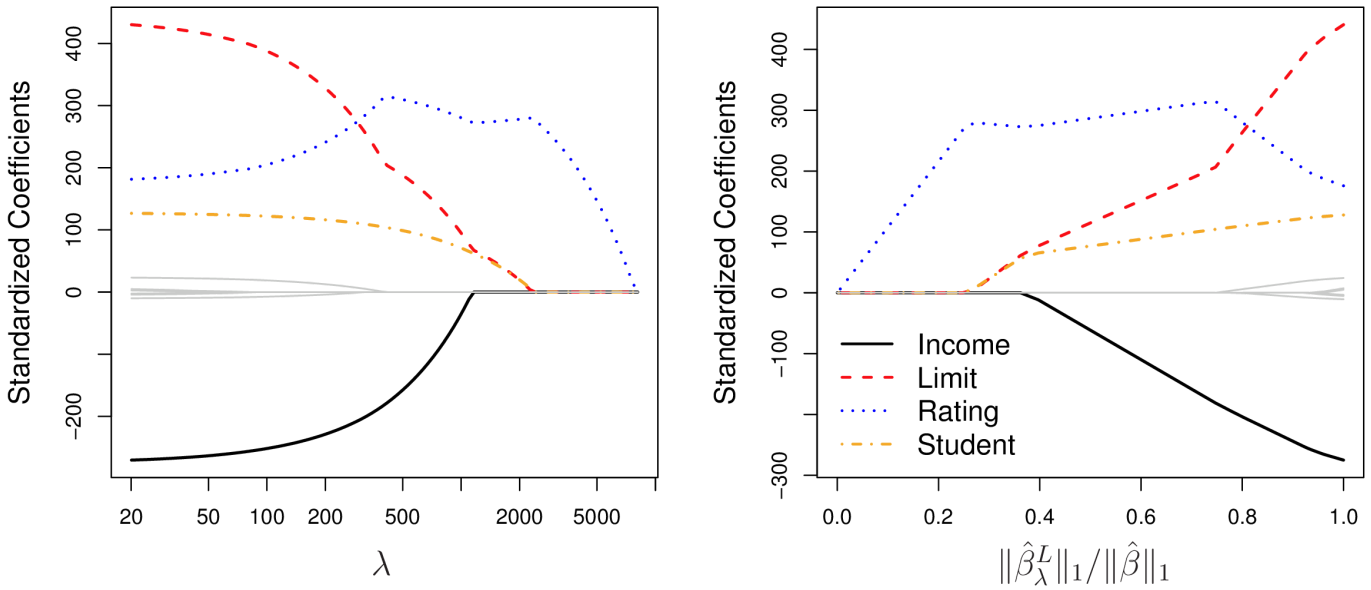
$$\underset{\beta}{\text{minimize}}\ \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso:

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

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

Standardized lasso coefficients for the Credit data set.

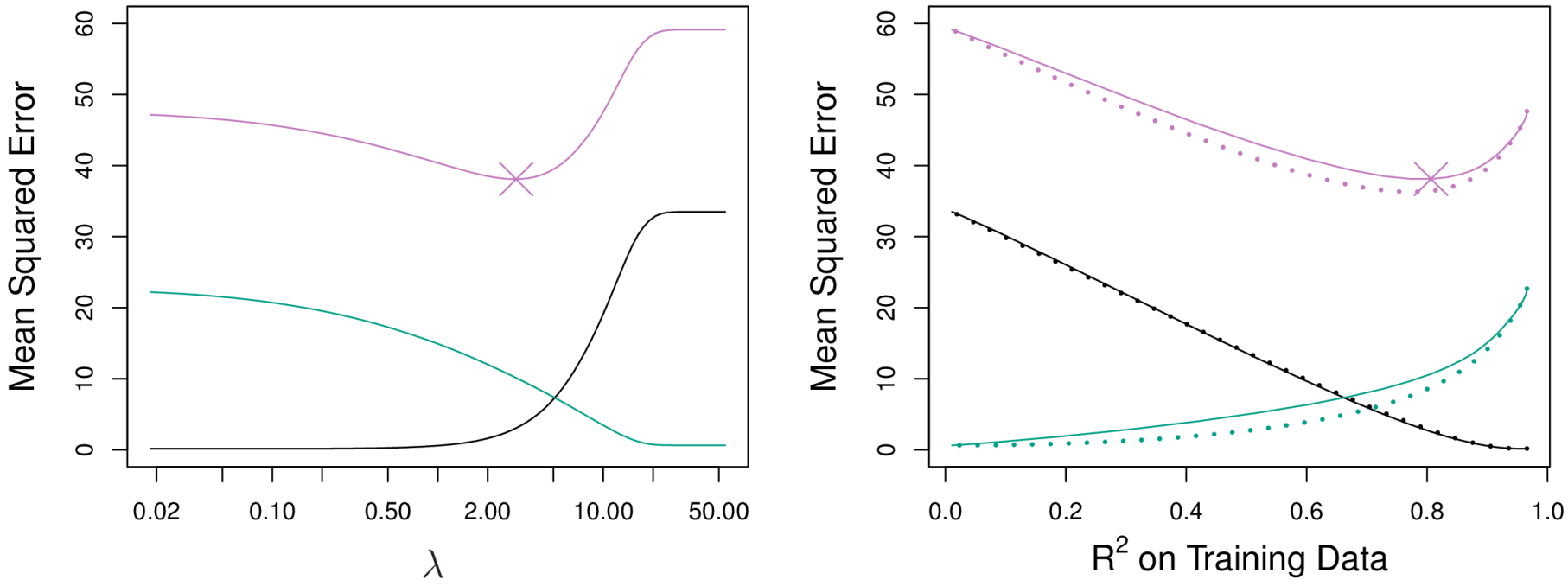


(Chapter6/6.6.pdf)

- $p = 10$
- 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

(Chapter6/6.5.pdf)

Examples of regularised techniques

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:

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

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

Penalised LDA

The problem is inverting Σ, fix it by

$$\underset{\beta_k}{\text{maximize}} \left(\beta_k^T \hat{\Sigma}_B \beta_k + \lambda_k \sum_{j=1}^{p} |\hat{\sigma}_j \beta_{kj}| \right)$$
$$\text{subject to } \beta_k^T \tilde{\Sigma} \beta_k \leq 1$$

where $\hat{\sigma}_j$ is the within-class standard deviation for variable j This is penalised LDA, and see reference, and the R package.

PDA Index

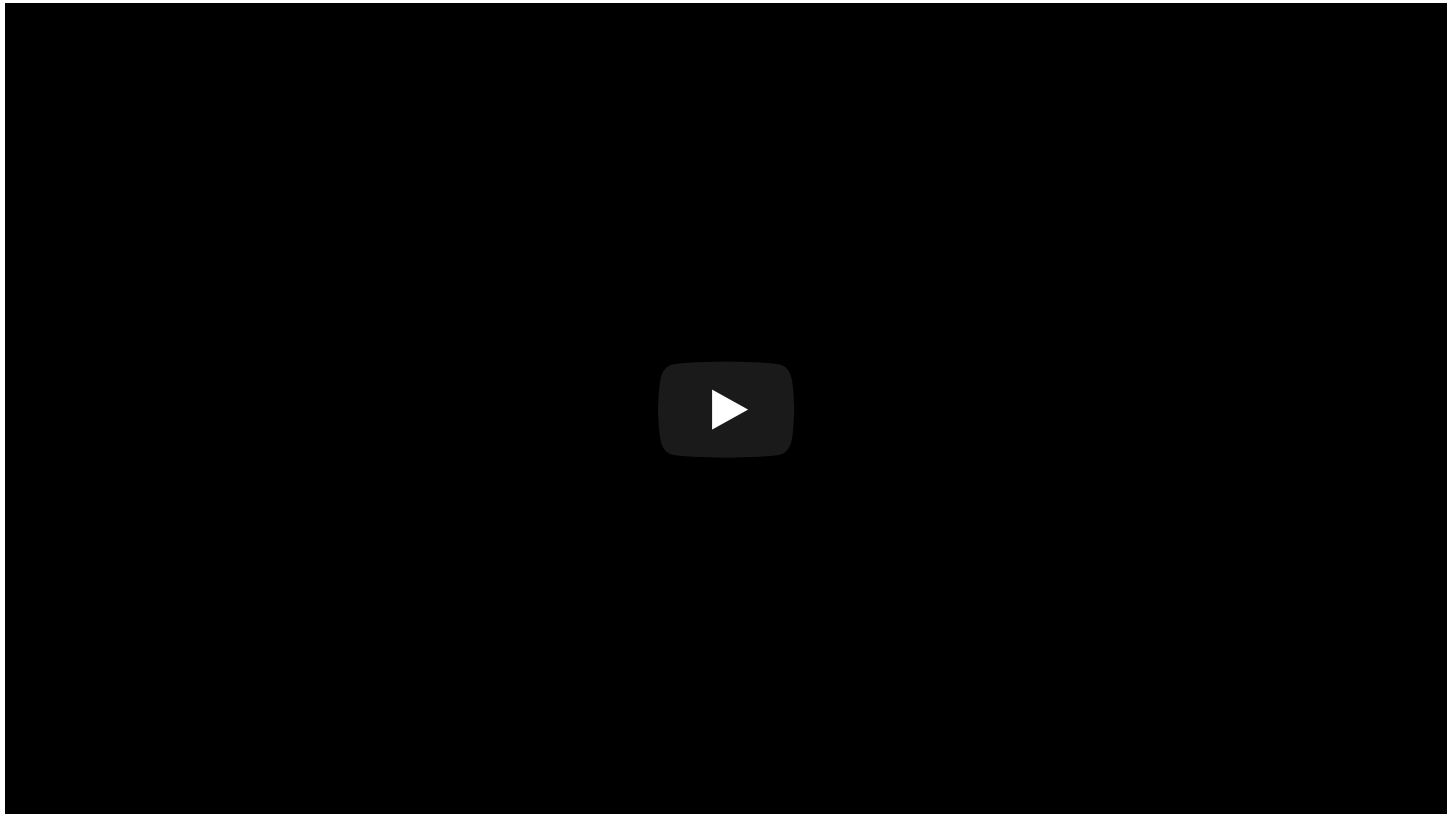
Penalised LDA projection pursuit index. Available in the `tourr` package.

$$I_{PDA}(A, \lambda) = 1 - \frac{\left|A' \left\{(1-\lambda)\hat{\Sigma} + n\lambda I_p\right\} A\right|}{\left|A' \left\{(1-\lambda)(\hat{\Sigma}_B + \hat{\Sigma}) + n\lambda I_p\right\} A\right|}$$

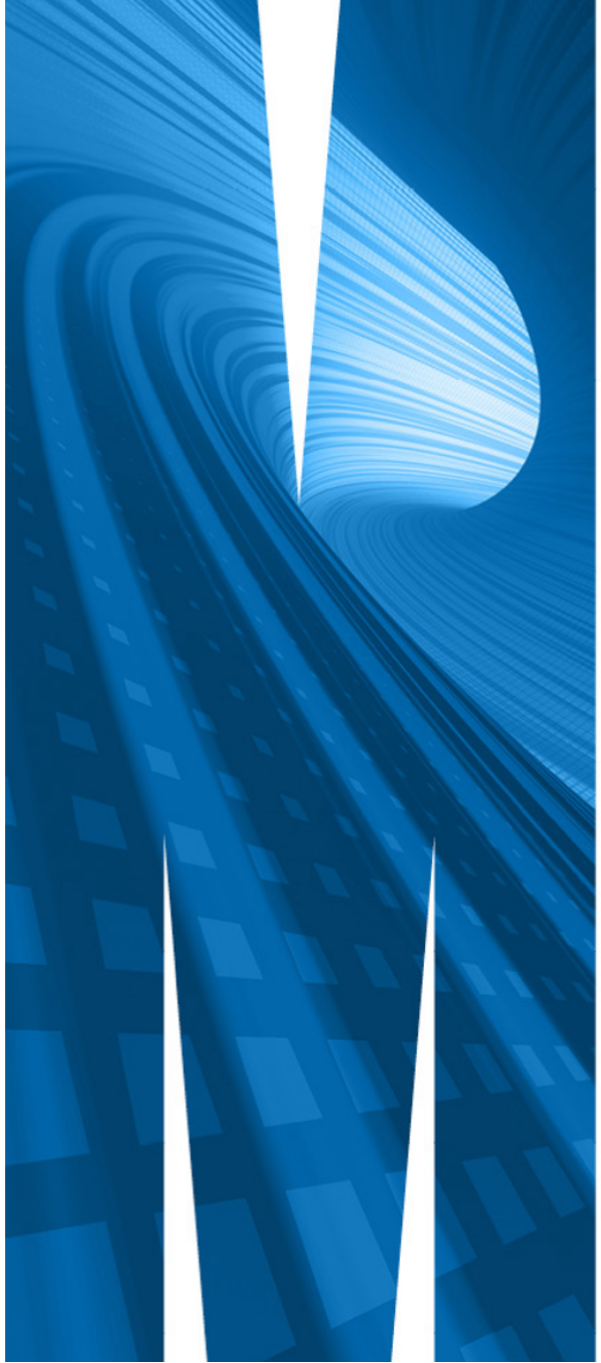
Optimising this function over $p \times d$ projection matrix A

Lasso regression

Read the example of lasso regression or watch the screencast by Julia Silge here



MONASH University



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 8b