Chapter 6: Linear Model Selection and Regularization (Lecture 2)

Thiago G. Martins | NTNU & Verizon Media
Spring 2020

Previous lecture

Subset selection and shrinkage methods

- Subset selection and shrinkage methods have controlled variance in two ways:
  - Using a subset of the original predictors.
  - Shrinking their coefficients towards zero.
- Those methods use the original (possibly standardized) predictors $X_1, \ldots, X_p$.

Dimension reduction methods

Dimension reduction methods

- Transform the original predictors

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

for $m = 1, \ldots, M, j = 1, \ldots, p$ and $M < p$

Dimension reduction methods

- Transform the original predictors

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

for $m = 1, \ldots, M, j = 1, \ldots, p$ and $M < p$

- Fit least square using the transformed predictors

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n$$

Constrained interpretation

- It can be shown that

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$$

Constrained interpretation

- It can be shown that

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$$

- So dimension reduction serves to constrain the coefficients of a standard linear regression
- This constrain increase the bias but is useful in situations where the variance is high

Outline

- We will cover two approaches to dimensionality reduction:
  - Principal Components
  - Partial Least Squares

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- Discussed in greater detail in Chapter 10 about unsupervised learning
- Focus in this lecture is how it can be applied for regression.
  - That is, in a supervised setting.

Principal Component Analysis (PCA)

- Discussed in greater detail in Chapter 10 about unsupervised learning
- Focus in this lecture is how it can be applied for regression.
    - That is, in a supervised setting.
- PCA is a (unsupervised) technique for reducing the dimension of a $n \times p$ data matrix X.

Principal Component Analysis (PCA)

- We want to create a $n \times M$ matrix Z, with $M < p$.
- The column Z_m of Z is the m-th principal component.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \quad \text{subject to} \quad \sum_{j=1}^{p} \phi_{jm}^2 = 1$$

Principal Component Analysis (PCA)

- We want to create a $n \times M$ matrix Z, with $M < p$.
- The column Z_m of Z is the m-th principal component.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \quad \text{subject to} \quad \sum_{j=1}^{p} \phi_{jm}^2 = 1$$

- We want Z_1 to have the highest possible variance.
  - That is, take the direction of the data where the observations vary the most.
  - Without the constrain we could get higher variance by increasing $\boldsymbol{\phi}_j$

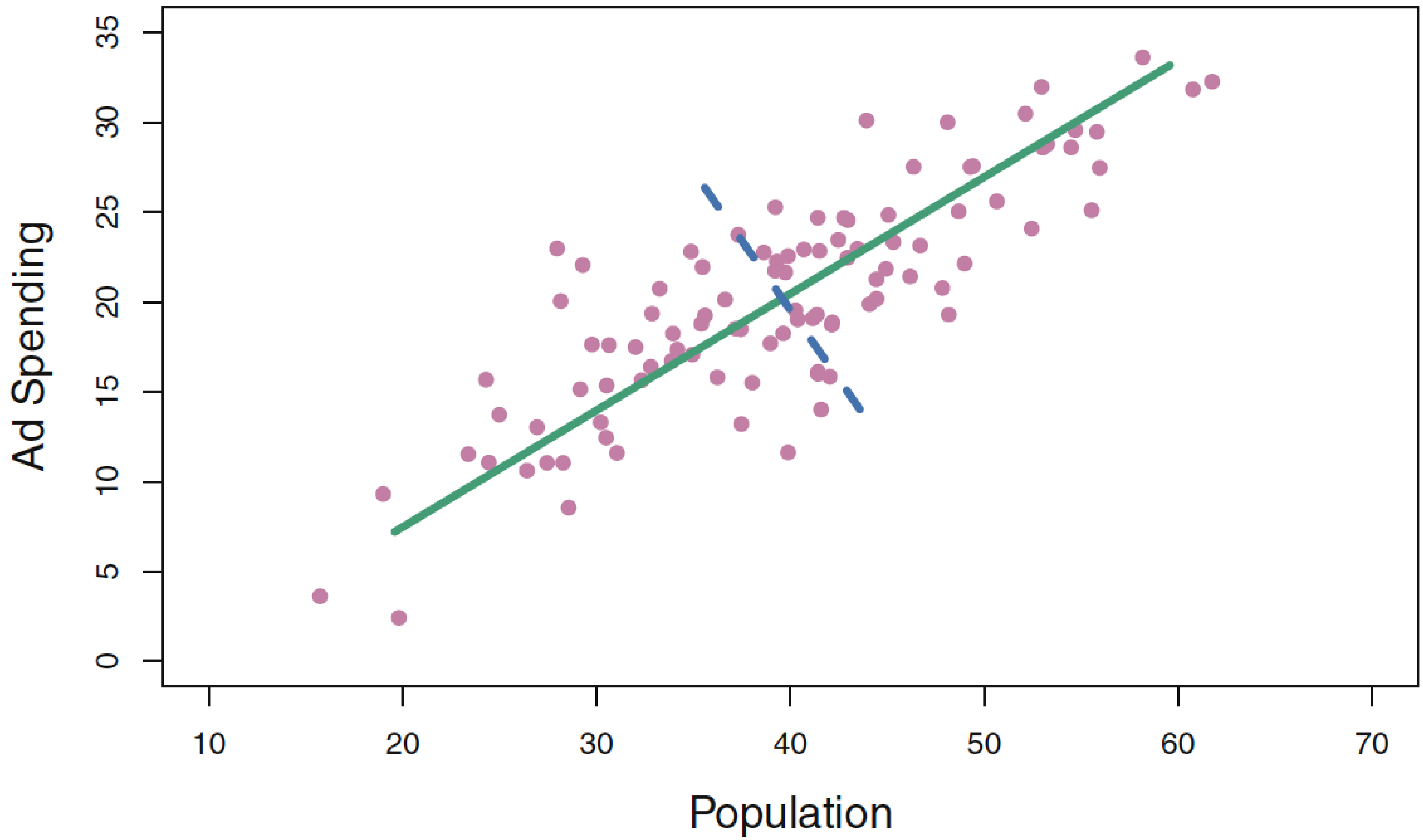
Principal Component Analysis (PCA)

- Z_2 should be uncorrelated to Z_1, and have the highest variance, subject to this constrain.
  - The direction of Z_1 must be perpendicular (or orthogonal) to the direction of Z_2

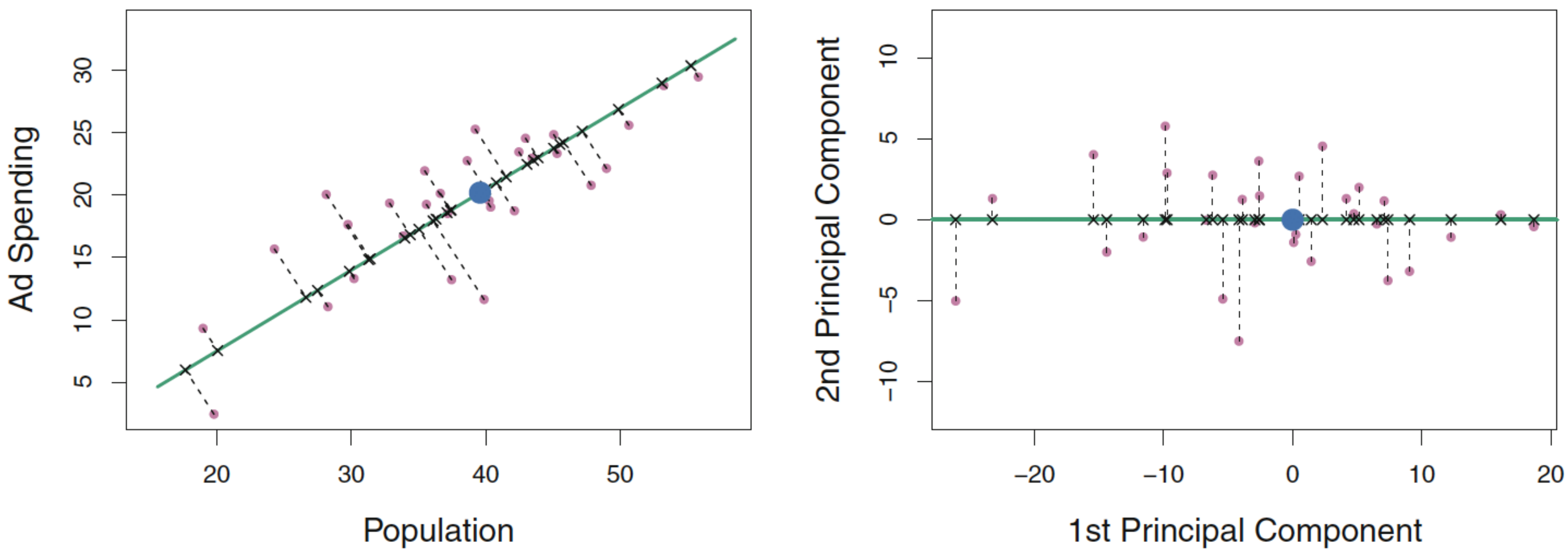
Principal Component Analysis (PCA)

- Z_2 should be uncorrelated to Z_1, and have the highest variance, subject to this constrain.
  - The direction of Z_1 must be perpendicular (or orthogonal) to the direction of Z_2
- And so on …
- We can construct up to p PCs that way.
  - In which case we have captured all the variability contained in the data
  - We have created a set of orthogonal predictors
  - But have **not** accomplished dimensionality reduction

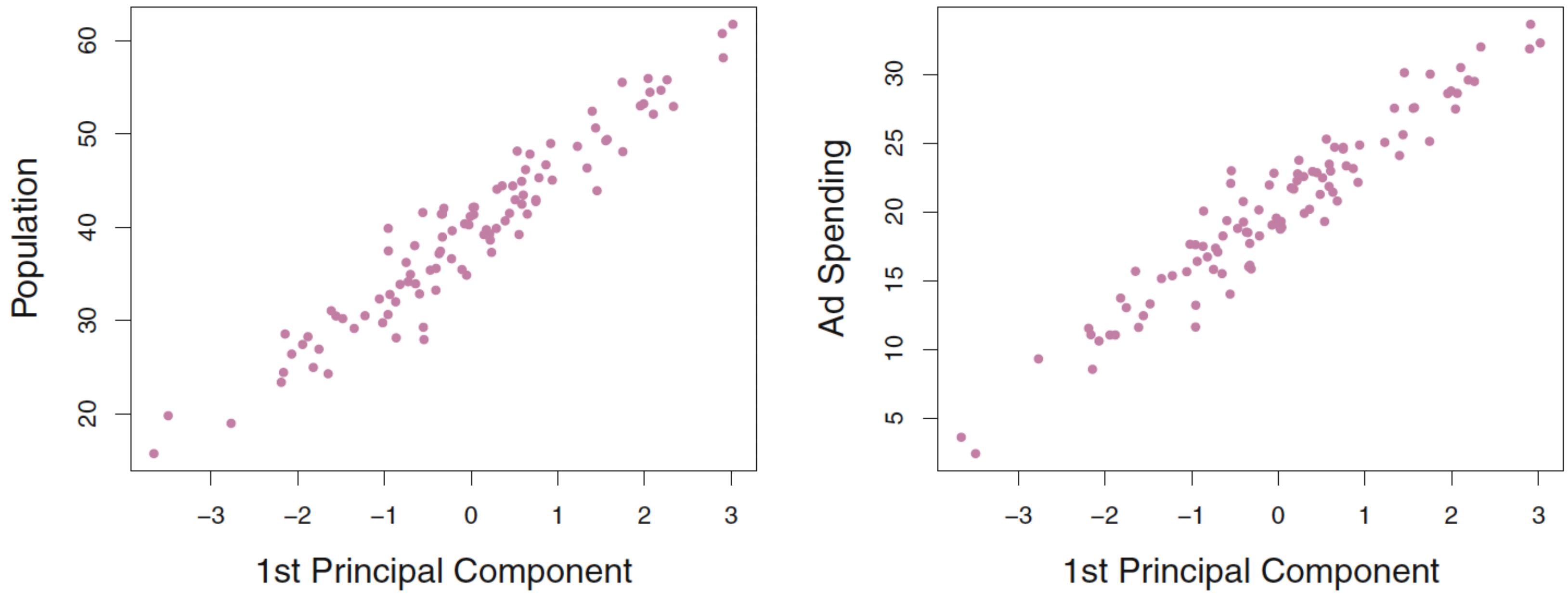
PCA Example - Ad spending



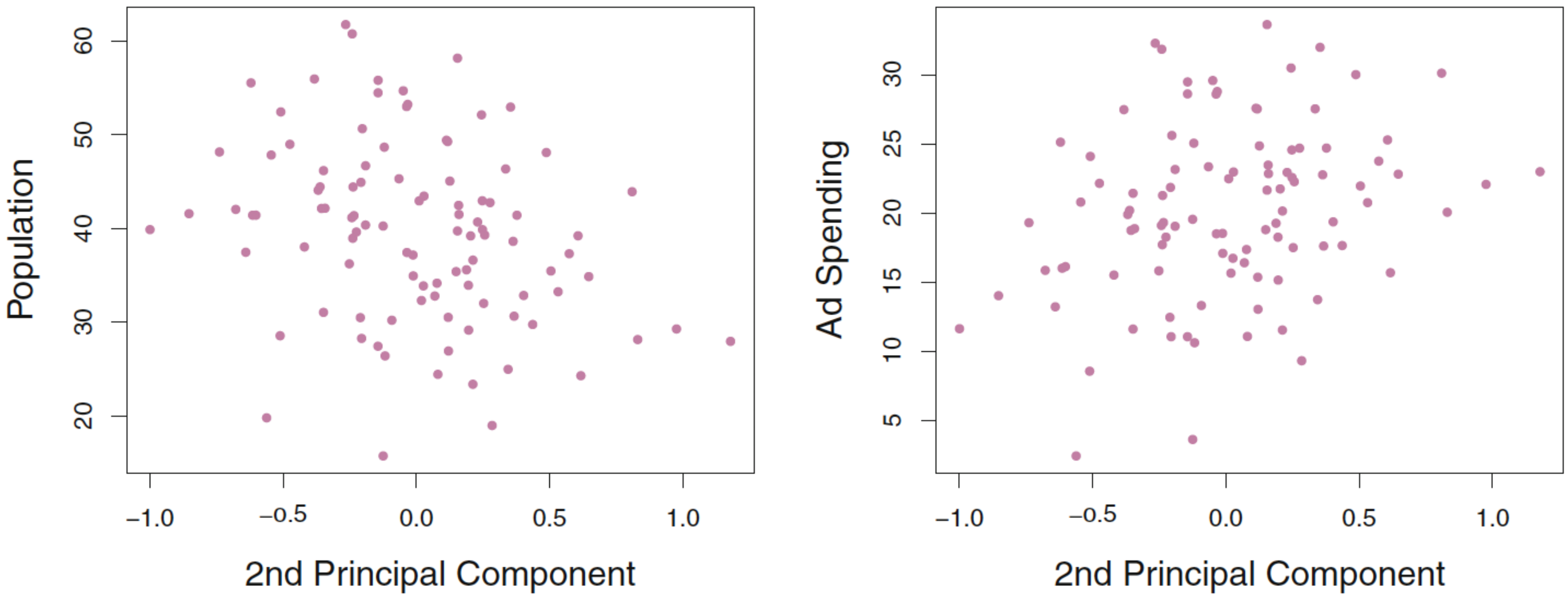
PCA Example - Ad spending (II)



PCA Example - Ad spending (III)



PCA Example - Ad spending (IV)



PCA - Overview

- Principal component analysis (PCA) is a dimensionality reduction technique
    - Our ability to visualize data is limited to 2 or 3 dimensions.

PCA - Overview

- Principal component analysis (PCA) is a dimensionality reduction technique
  - Our ability to visualize data is limited to 2 or 3 dimensions.
  - Lower dimension can reduce numerical algorithms computational time.

PCA - Overview

- Principal component analysis (PCA) is a dimensionality reduction technique
    - Our ability to visualize data is limited to 2 or 3 dimensions.
    - Lower dimension can reduce numerical algorithms computational time.
    - Many statistical models suffer from high correlation between covariates

PCA - Overview

- Principal component analysis (PCA) is a dimensionality reduction technique
  - Our ability to visualize data is limited to 2 or 3 dimensions.
  - Lower dimension can reduce numerical algorithms computational time.
  - Many statistical models suffer from high correlation between covariates
- PCA is not scale invariant,
  - standardize all the p variables before applying PCA.

PCA - Overview

- Principal component analysis (PCA) is a dimensionality reduction technique
  - Our ability to visualize data is limited to 2 or 3 dimensions.
  - Lower dimension can reduce numerical algorithms computational time.
  - Many statistical models suffer from high correlation between covariates
- PCA is not scale invariant,
  - standardize all the p variables before applying PCA.
- Assume $\mathbf{\Sigma}$ to be the covariance matrix associated with $\boldsymbol{X}$.
  - The fraction of the original variance kept by the M principal component

$$R^2 = \frac{\sum_{i=1}^{M} \lambda_i}{\sum_{j=1}^{p} \lambda_j}, \quad \lambda_i' s \text{ eigenvalues of } \mathbf{\Sigma}$$

Recommended exercise 7

How many principal components should we use for the Credit Dataset? Justify?

Principal Components Regression (PCR)

Principal Components Regression (PCR)

- Principal Components Regression involves:
  - Constructing the first M principal components $Z_1, \ldots, Z_M$
  - Using these components as the predictors in a standard linear regression model

Principal Components Regression (PCR)

- Principal Components Regression involves:
  - Constructing the first M principal components $Z_1, \ldots, Z_M$
  - Using these components as the predictors in a standard linear regression model
- Key assumptions: A small number of principal components suffice to explain:

1. Most of the variability in the data.
2. The relationship with the response.

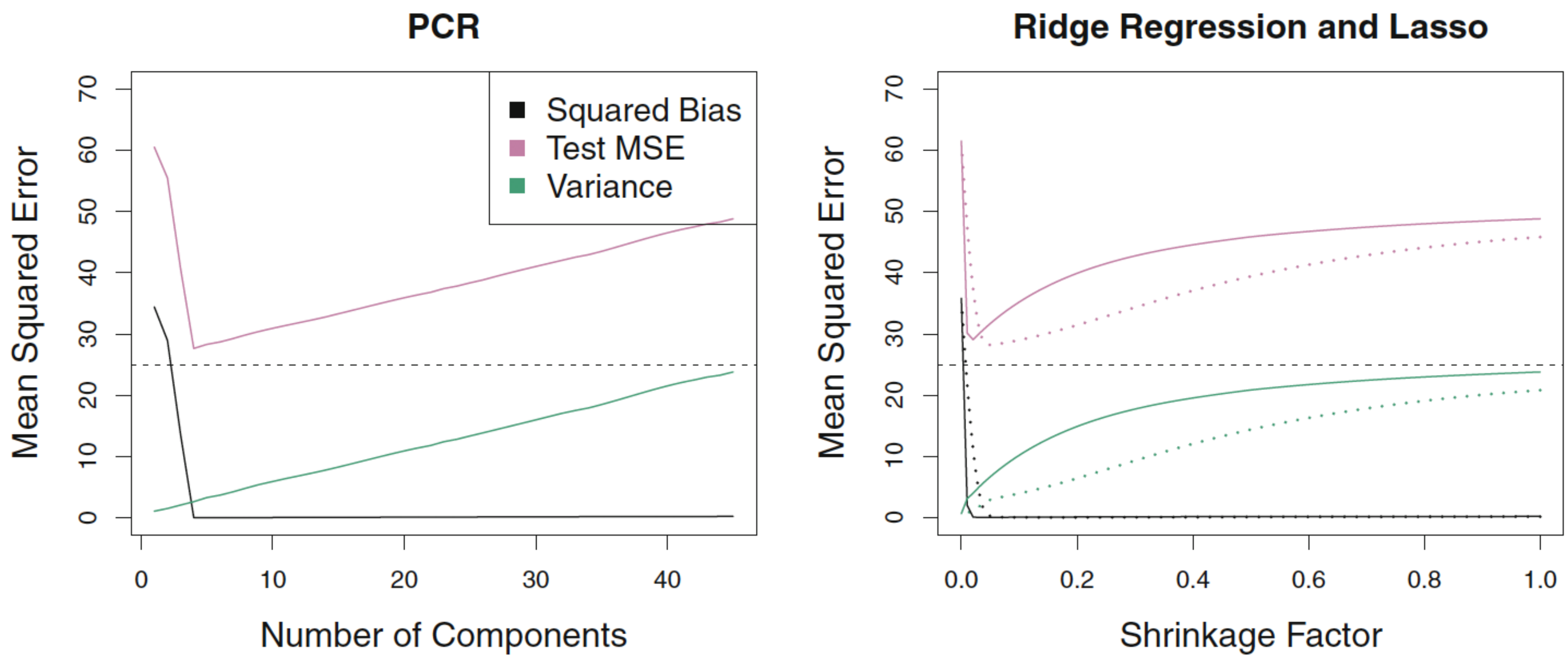
Principal Components Regression (PCR)

- Principal Components Regression involves:
  - Constructing the first M principal components $Z_1, \ldots, Z_M$
  - Using these components as the predictors in a standard linear regression model
- Key assumptions: A small number of principal components suffice to explain:

1. Most of the variability in the data.
2. The relationship with the response.

- The assumptions above are not guaranteed to hold in every case.
  - This is true specially for assumption 2 above.
  - Since the PCs are selected via unsupervised learning.

Example: PCR vs. Lasso and Ridge (Simulated data)



Example: PCR vs. Lasso and Ridge (Simulated data)

- PCR performed well on simulated data, recovering the need for $M = 5$
  - However, results are only slightly better than lasso and very similar to Ridge.

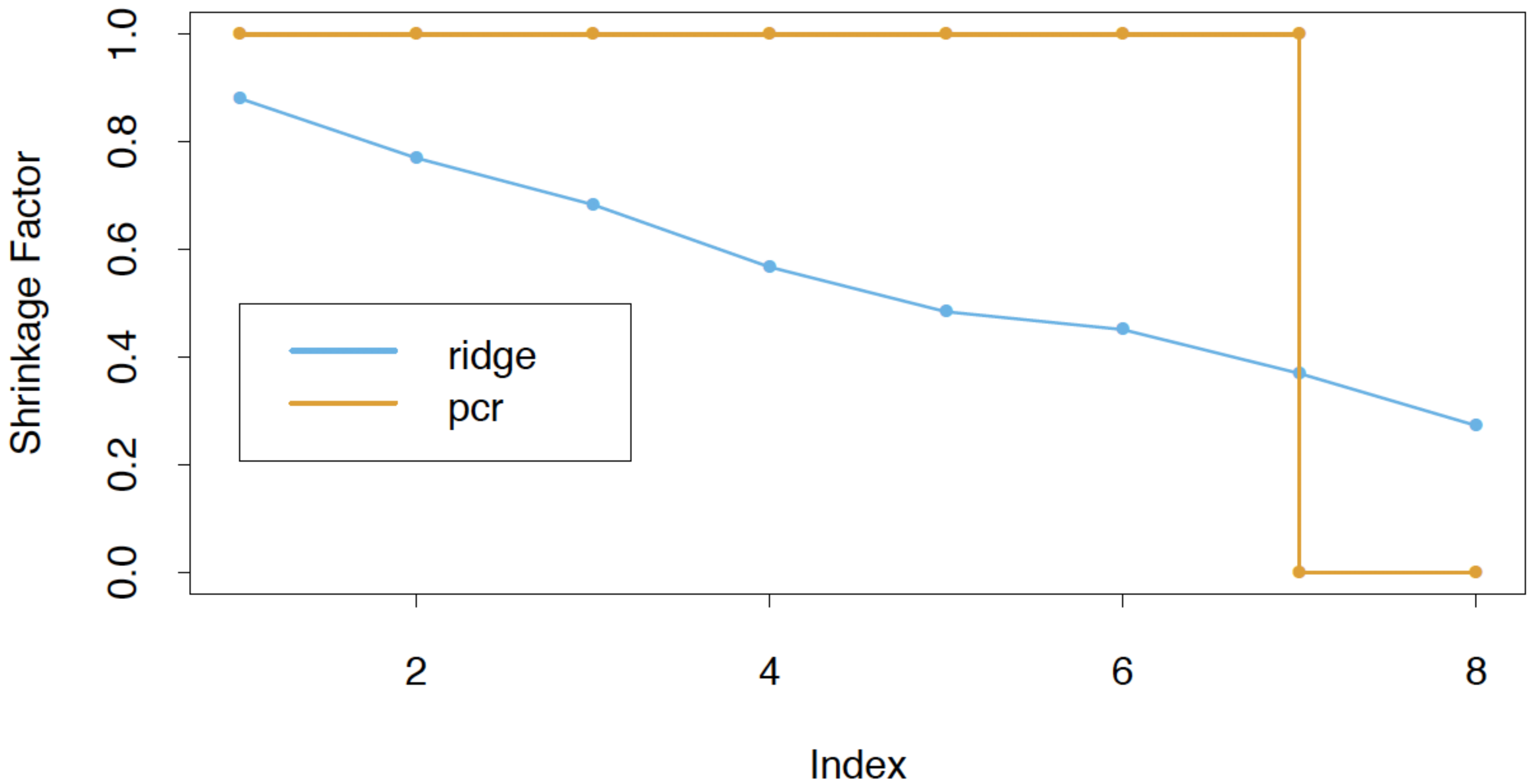
Example: PCR vs. Lasso and Ridge (Simulated data)

- PCR performed well on simulated data, recovering the need for $M = 5$
    - However, results are only slightly better than lasso and very similar to Ridge.
- Similar to Ridge, PCR does not perform feature selection
    - PCs are linear combination of all predictors

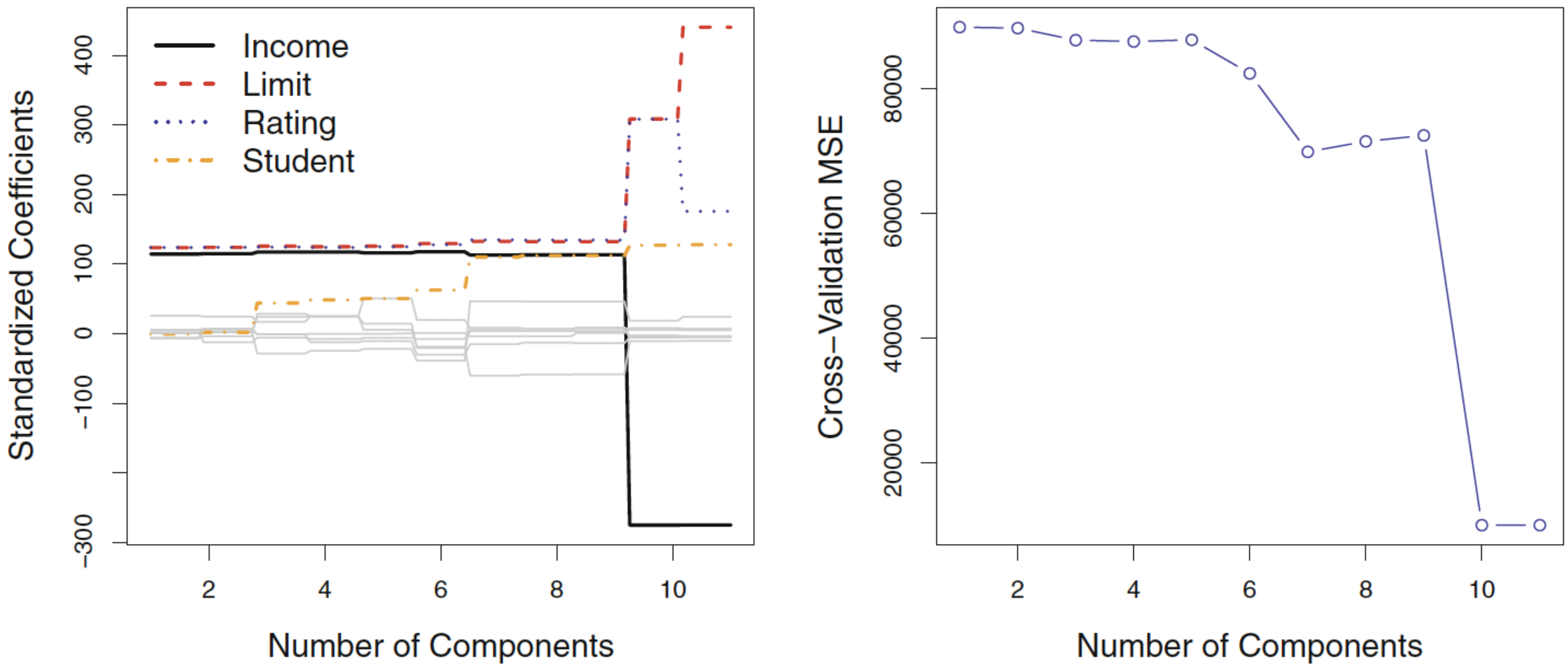
Example: PCR vs. Lasso and Ridge (Simulated data)

- PCR performed well on simulated data, recovering the need for $M = 5$
  - However, results are only slightly better than lasso and very similar to Ridge.
- Similar to Ridge, PCR does not perform feature selection
  - PCs are linear combination of all predictors
- PCR can be seen as discretized version of Ridge regression.
  - Ridge shrinks coefs. of the PCs by $\lambda_j^2/(\lambda_j^2 + \lambda)$
  - Higher pressure on less important PCs
  - PCR discards the $p - M$ smallest eigenvalue components.

Example: Shrinkage Factor



Example: PCR (Credit Data)



Recommended exercise 8

Apply PCR on the Credit dataset and compare the results with the methods covered in Lecture 1.

PCR (Drawback)

- Dimensionality reduction is done via an unsurpevised method (PCA)
    - No guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

Partial Least Squares (PLS)

Partial Least Squares (PLS)

- PLS works similar to PCR
  - Dimension reduction: $Z_1, \ldots, Z_M$, $M < p$
  - Z_i linear combination of original predictors.
  - Apply standard linear model using $Z_1, \ldots, Z_M$ as predictors.

Partial Least Squares (PLS)

- PLS works similar to PCR
  - Dimension reduction: $Z_1, \ldots, Z_M, M < p$
  - Z_i linear combination of original predictors.
  - Apply standard linear model using $Z_1, \ldots, Z_M$ as predictors.
- But it uses the response Y in order to identify new features
  - attempts to find directions that help explain both the response and the predictors.

Partial Least Squares (Algorithm)

- $Z_1 = \sum_{i=1}^{p} \phi_{j1} X_j$
  - ϕ_{j1} is the coefficient from the simple linear regression of Y onto X_j.
  - this coefficient is proportional to the correlation between Y and X_j.
  - PLS puts highest weight on the variables that are most strongly related to the response.

Partial Least Squares (Algorithm)

- $Z_1 = \sum_{i=1}^{p} \phi_{j1} X_j$
  - ϕ_{j1} is the coefficient from the simple linear regression of Y onto X_j.
  - this coefficient is proportional to the correlation between Y and X_j.
  - PLS puts highest weight on the variables that are most strongly related to the response.
- To obtain the second PLS direction, Z_2:
  - We regress each variable on Z_1 and take the residuals
  - The residuals are remained info not explained by Z_1
  - We the compute Z_2 using this orthogonalized data, similarly to Z_1.

Partial Least Squares (Algorithm)

- $Z_1 = \sum_{i=1}^{p} \phi_{j1} X_j$
  - ϕ_{j1} is the coefficient from the simple linear regression of Y onto X_j.
  - this coefficient is proportional to the correlation between Y and X_j.
  - PLS puts highest weight on the variables that are most strongly related to the response.
- To obtain the second PLS direction, Z_2:
  - We regress each variable on Z_1 and take the residuals
  - The residuals are remained info not explained by Z_1
  - We the compute Z_2 using this orthogonalized data, similarly to Z_1.
- We can repeat this iteration process M times to get $Z_1, \ldots, Z_M$.

Recommended exercise 9

Apply PLS on the Credit dataset and compare the results with the methods covered in Lecture 1 and PCR.

Partial Least Squares (Performance)

- In practice, PLS often performs no better than ridge regression or PCR.
  - Supervised dimension reduction of PLS can reduce bias.
  - It also has the potential to increase variance.

In summary

- PLS, PCR and ridge regression tend to behave similarly.

In summary

- PLS, PCR and ridge regression tend to behave similarly.
- Ridge regression may be preferred because it shrinks smoothly, rather than in discrete steps.

In summary

- PLS, PCR and ridge regression tend to behave similarly.
- Ridge regression may be preferred because it shrinks smoothly, rather than in discrete steps.
- Lasso falls somewhere between ridge regression and best subset regression, and enjoys some of the properties of each.

Considerations in high dimensions

High dimension

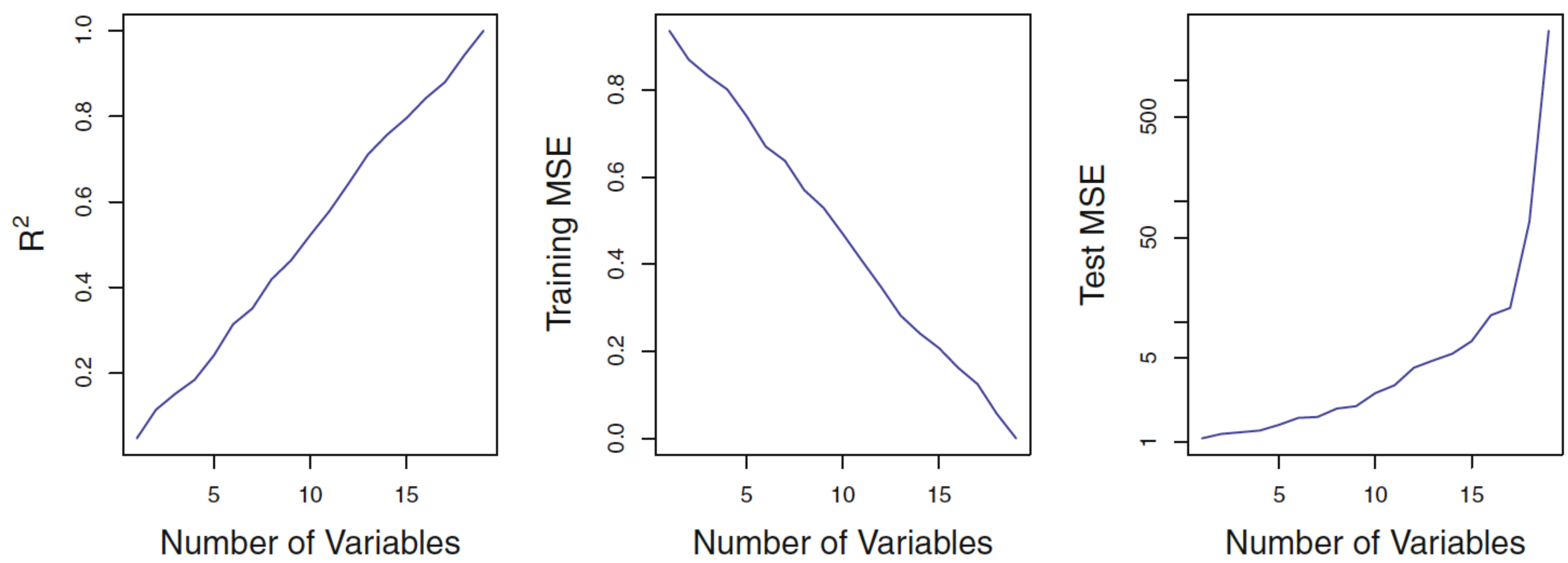
- High dimension problems: $n < p$
- More common nowadays

High dimension issues (Example)

- Standard linear regression cannot be applied.
  - Perfect fit to the data, regardless of relationship
  - Unfortunately, the C_p, AIC, and BIC approaches are problematic (hard to estimate σ^2)

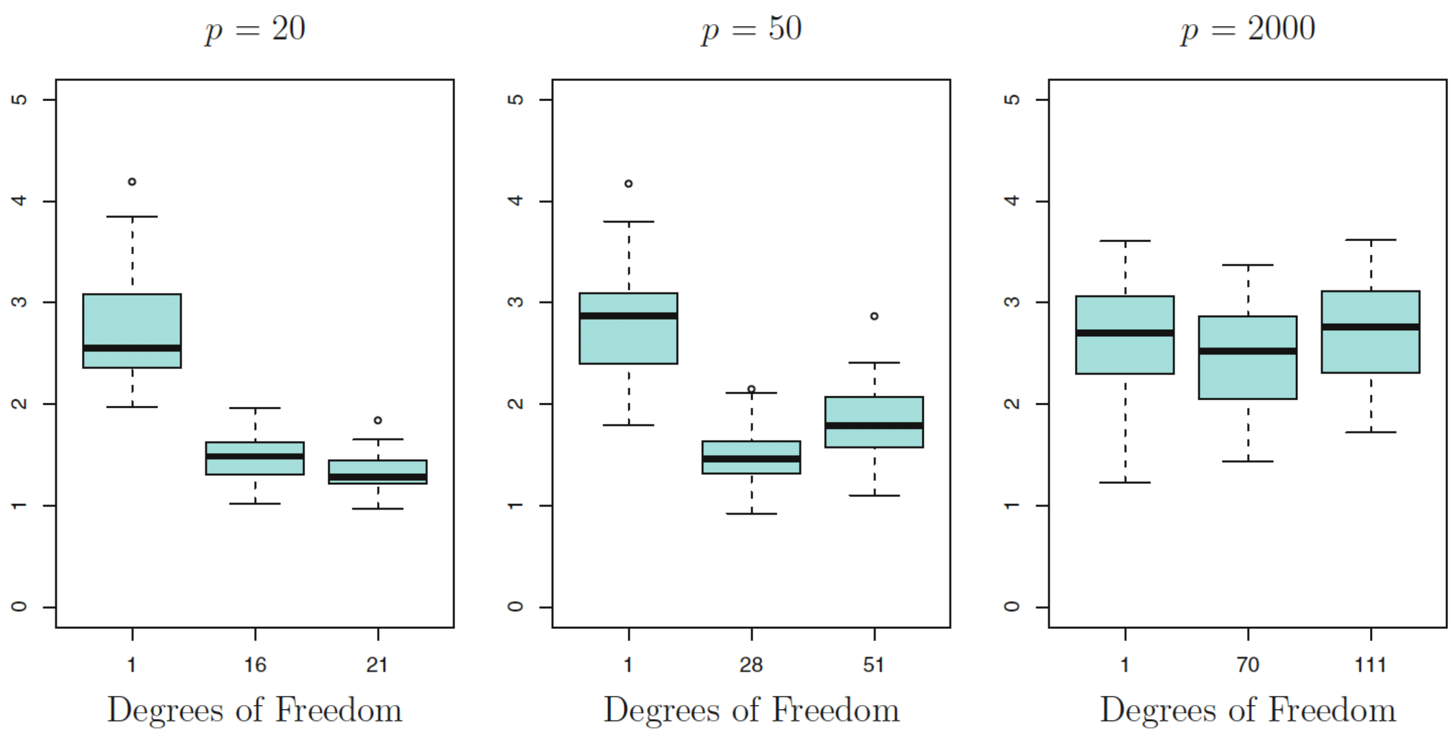
High dimension issues (Example)

- Standard linear regression cannot be applied.
  - Perfect fit to the data, regardless of relationship
  - Unfortunately, the C_p, AIC, and BIC approaches are problematic (hard to estimate σ^2)



Noise predictors

- The test error tends to increase as the dimensionality of the problem
  - Unless the additional features are truly associated with the response.



The danger of too many features

- In general, adding additional signal features helps (smaller test set errors)

The danger of too many features

- In general, adding additional signal features helps (smaller test set errors)
- However, adding noise features that are not truly associated with the response increases test set error.
    - Noise features exacerbating the risk of overfitting
    - Previous example shows that regularizations does not eliminate the problem

The danger of too many features

- In general, adding additional signal features helps (smaller test set errors)
- However, adding noise features that are not truly associated with the response increases test set error.
  - Noise features exacerbating the risk of overfitting
  - Previous example shows that regularizations does not eliminate the problem
- New technologies that allow for the collection of measurements for thousands or millions of features are a double-edged sword

Interpreting results in high dimension

- In the high-dimensional setting, the multicollinearity problem is extreme

Interpreting results in high dimension

- In the high-dimensional setting, the multicollinearity problem is extreme
- Essentially, this means:
  - We can never know exactly which variables (if any) truly are predictive of the outcome.
  - We can never identify the best coefficients for use in the regression.

Interpreting results in high dimension

- In the high-dimensional setting, the multicollinearity problem is extreme
- Essentially, this means:
    - We can never know exactly which variables (if any) truly are predictive of the outcome.
    - We can never identify the best coefficients for use in the regression.
    - At most, we can hope to assign large regression coefficients to variables that are correlated with the variables that truly arec predictive of the outcome.
    - We will find one of possibly many suitable predictive models.

The end

Thank you for showing up