SDSC 3006: Fundamentals of Machine Learning I

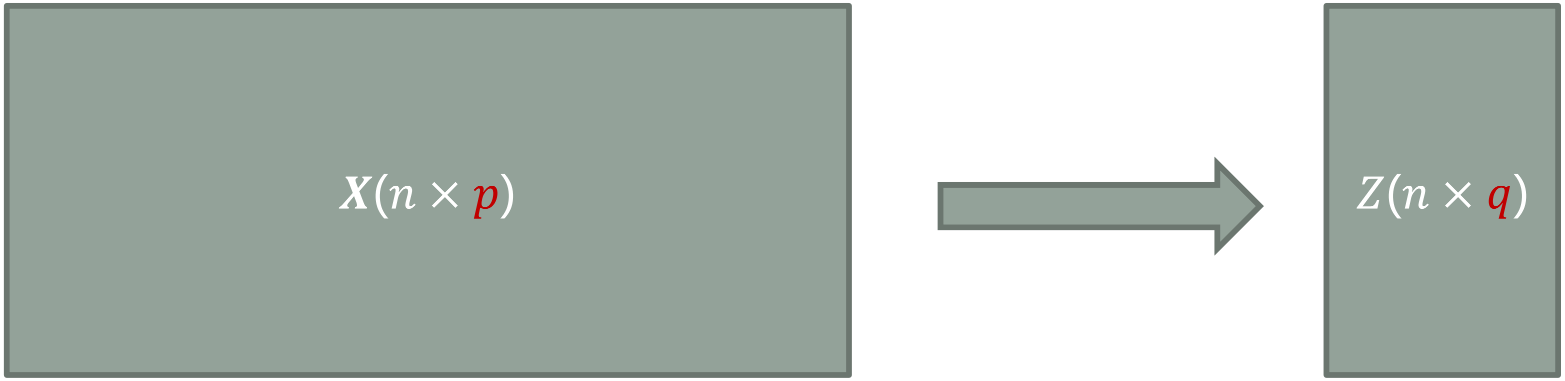
Topic 8. Principal Components Regression

Outline

- Principal components analysis (PCA)
- Principal components regression (PCR)

What Does PCA Do?

- Dimension reduction



$\boldsymbol{X}(n \times p)$

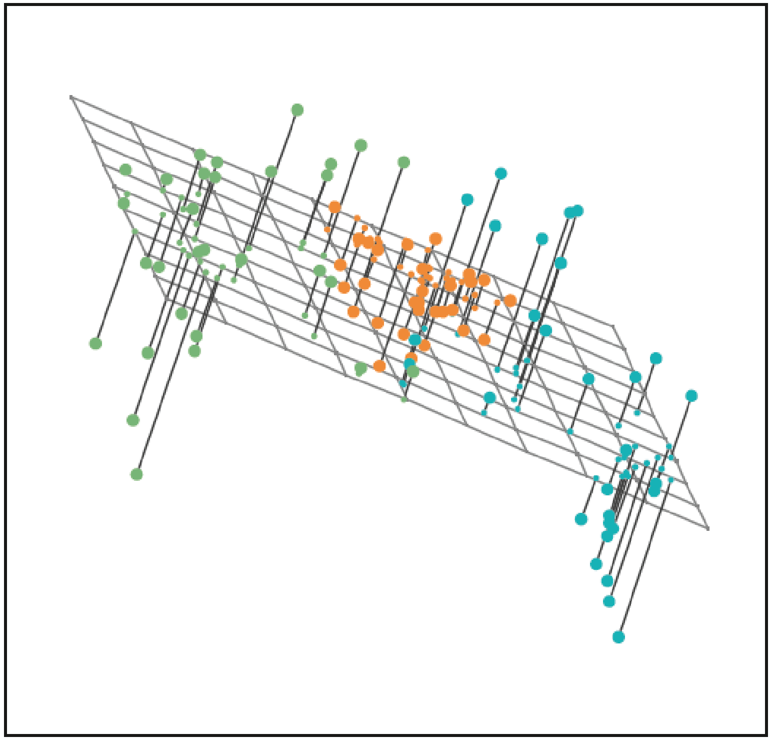
$Z(n \times q)$

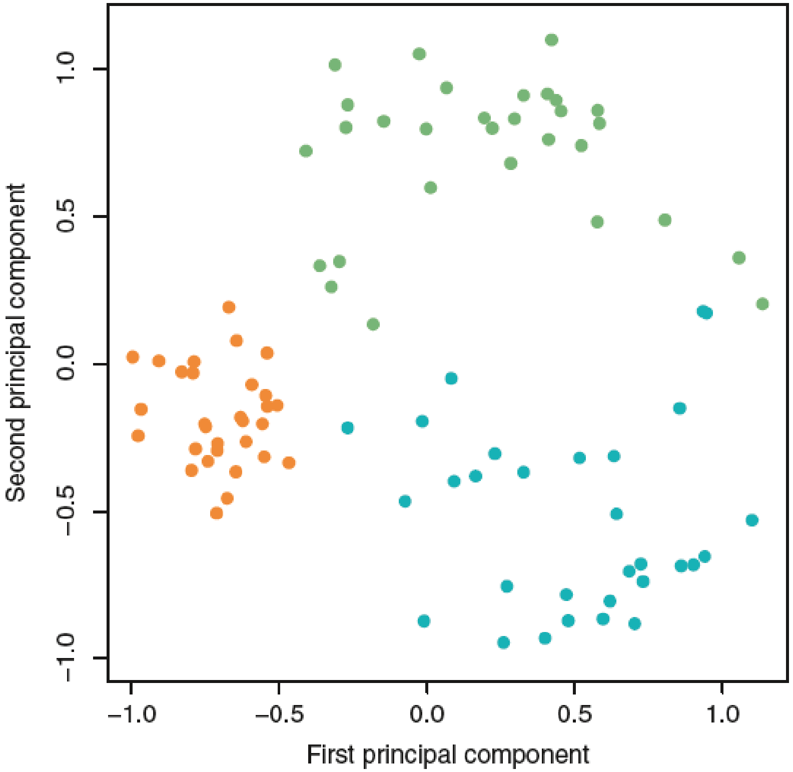
PCA finds a low-dimensional representation of the data that captures as much of the information as possible.

What Does PCA Do?

- Data visualization

input variables $X_1, X_2, X_3 \rightarrow$ principal components PC_1, PC_2

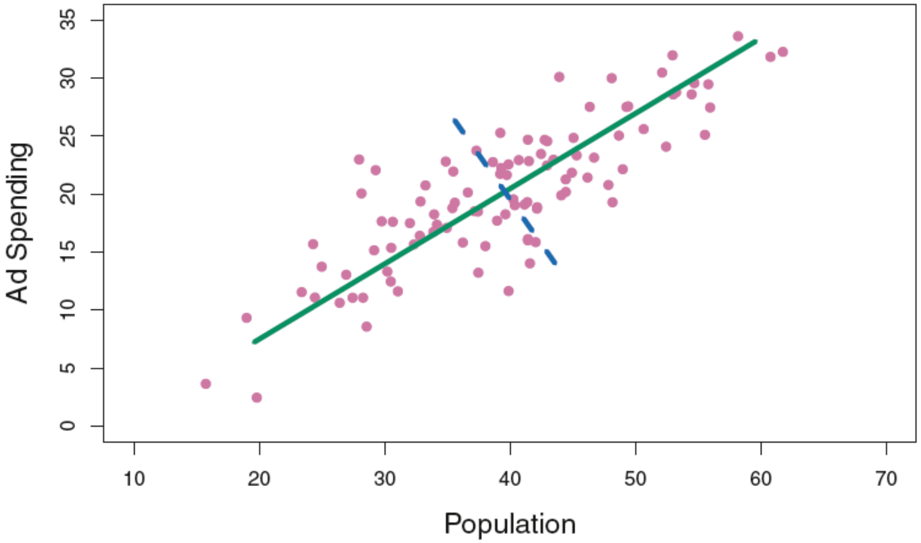


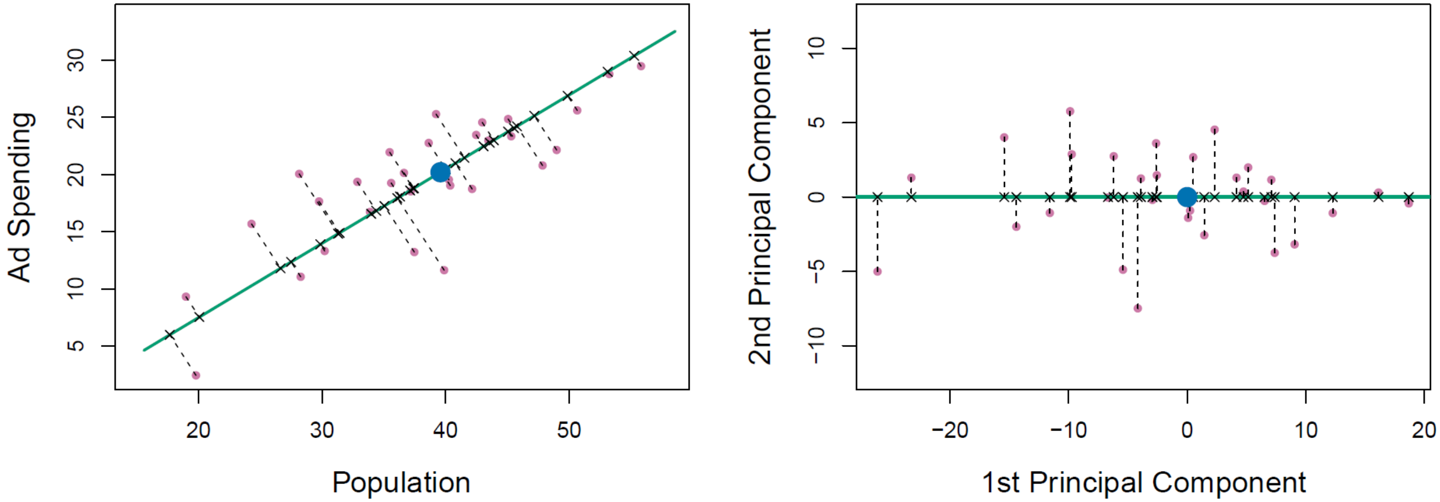


What Are Principal Components?

- PCA is an unsupervised approach because it involves only the predictors.
- Assume there are p predictors/features. In the p-dimensional feature space, not all directions are equally interesting. PCA seeks a small number of dimensions that are most interesting. Those dimensions are called principal components (PCs).
- “interesting” is measured by variance, i.e., the amount that the observations vary along each dimension.

Principal Components





Principal Components

- Each PC is a linear combination of the p features.

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

$$\begin{bmatrix} z_{11} \\ \cdots \\ z_{n1} \end{bmatrix} = \phi_{11}\begin{bmatrix} x_{11} \\ \cdots \\ x_{n1} \end{bmatrix} + \phi_{21}\begin{bmatrix} x_{12} \\ \cdots \\ x_{n2} \end{bmatrix} + \ldots + \phi_{p1}\begin{bmatrix} x_{1p} \\ \cdots \\ x_{np} \end{bmatrix}$$

Scores of the 1st PC

Loadings of the 1st PC

Algorithm to Find the First PC

- Center the individual columns in X matrix to have mean zero before PCA
- We look for the linear combination

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip}$$

that has maximal variance. That is,

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Algorithm to Find the Second PC

- After the first PC Z_1 is determined, we can find the second PC.
- We look for the linear combination

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \ldots + \phi_{p2}x_{ip}$$

that has maximal variance among all linear combinations that are uncorrelated with Z_1.

Properties of Principal Components

- The principal components are orthogonal (uncorrelated).
- They are ordered according to the decreasing variance in the data they capture: Z_1 has the largest variance, Z_2 has the second largest variance, etc.
- The principal component scores $Z_1, Z_2, \dots, Z_q$ can be used in further supervised learning (e.g., as predictors in regression)

Example

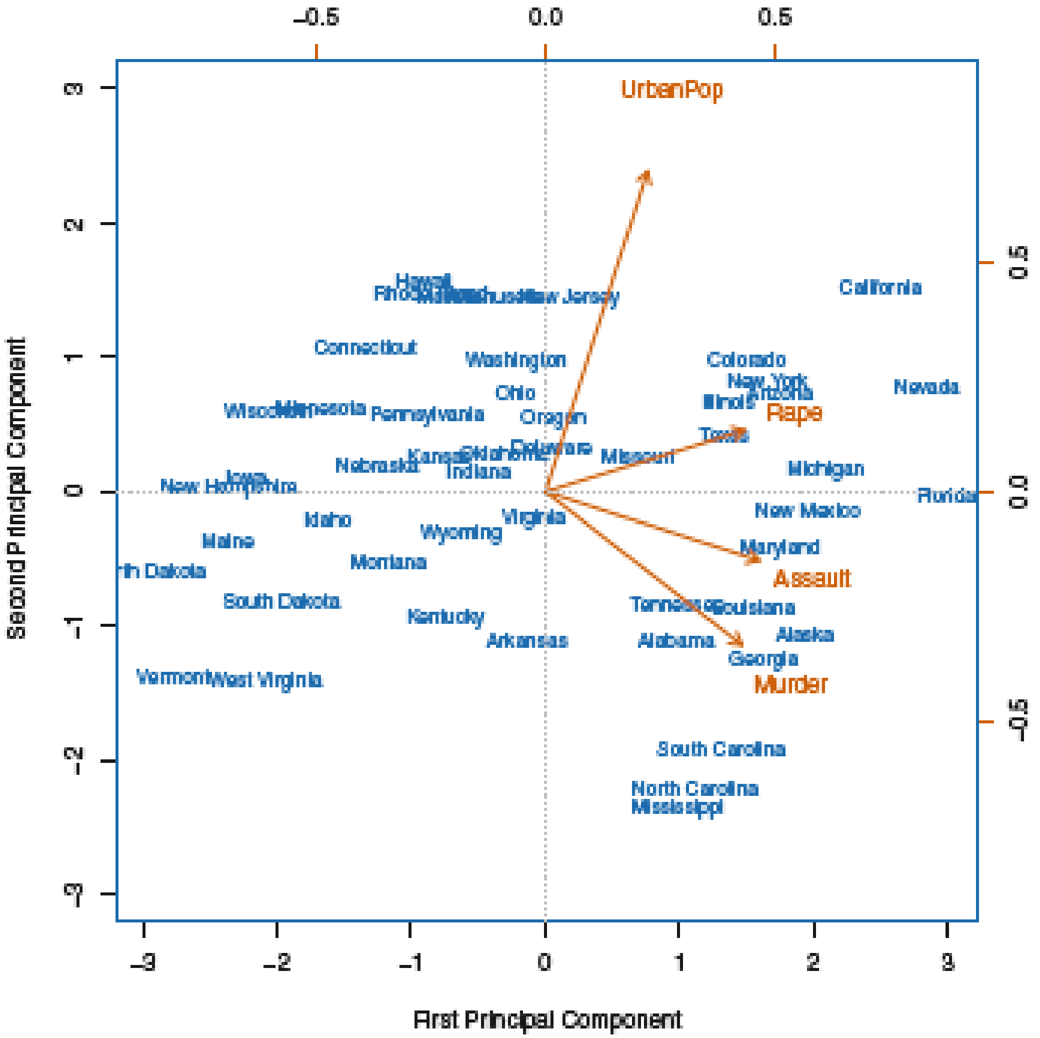
- USArrests dataset: for each of the 50 states in the US, the data set contains the number of arrests per 100,000 residents for each of three crimes: **Assault**, **Murder**, and **Rape**. **UrbanPop** (the percent of the population in each state living in urban areas) is also recorded.

- $p = 4, n = 50$

- Plot the first two principal components

Plot of the First Two PCs

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

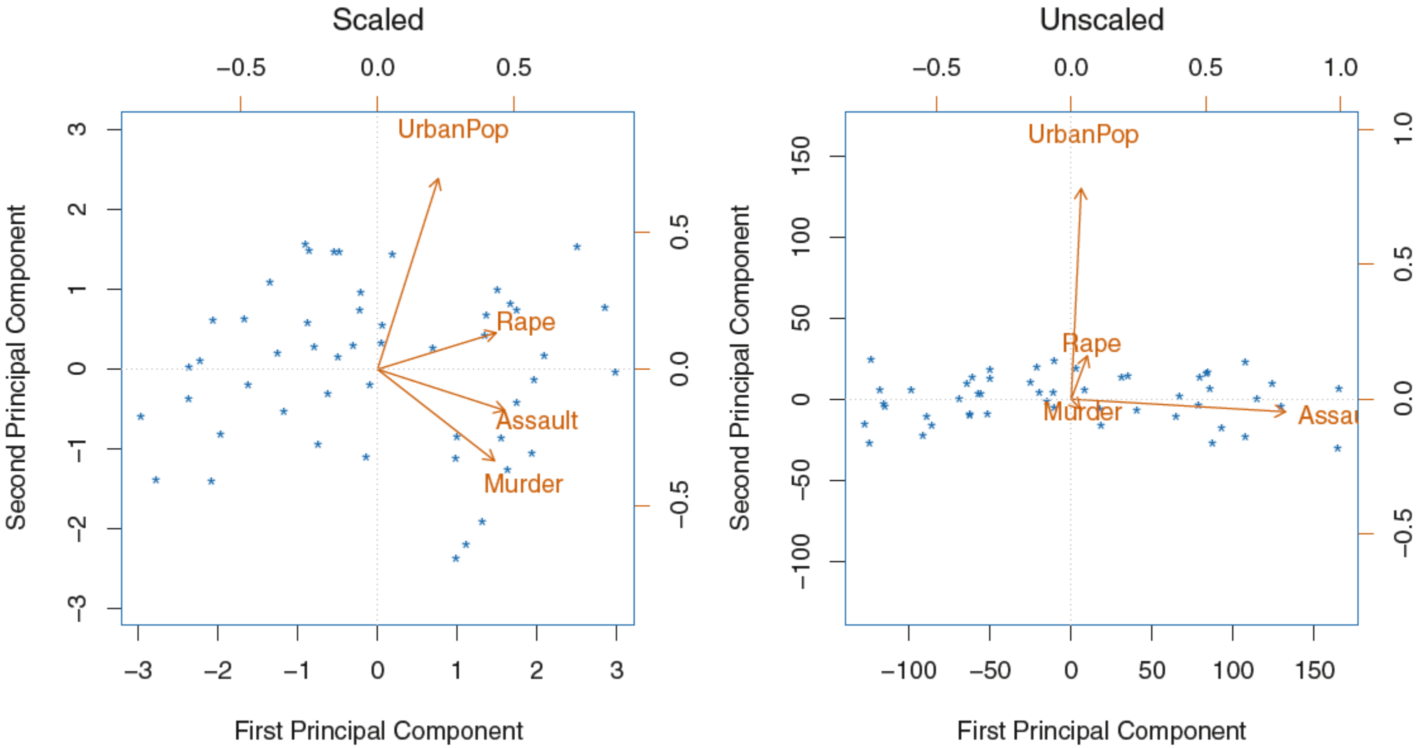


Interpretation:

- **PC1** places similar weights on Assault, Murder, Rape, with much less weight on UrbanPop. Hence, this component roughly corresponds to a measure of overall rates of serious crimes.
- **PC2** places most of its weight on UrbanPop. Hence, this component roughly corresponds to the level of urbanization of the state.
- **Crime rates:** States with large positive scores on PC1 have high crime rates, while those with negative scores on PC1 have low crime rates.
- **Urbanization:** States with large positive scores on PC2 have a high level of urbanization, while those with negative scores on PC2 have low level of urbanization.

On the Use of PCA

- **Scaling the variables**: scale each variable to have standard deviation 1 before performing PCA if variables are measured in different units



On the Use of PCA

- **Uniqueness of PCs**: Each PC loading vector is unique, up to a sign flip. Two different softwares may yield the same PC loadings with different signs.
- Each PC loading vector specifies a direction in the p-dimensional space. Flipping the sign has no effect as the direction does not change.
- Similarly, the score vectors are unique up to a sign flip, since the variance of Z is the same as the variance of $-Z$.

On the Use of PCA

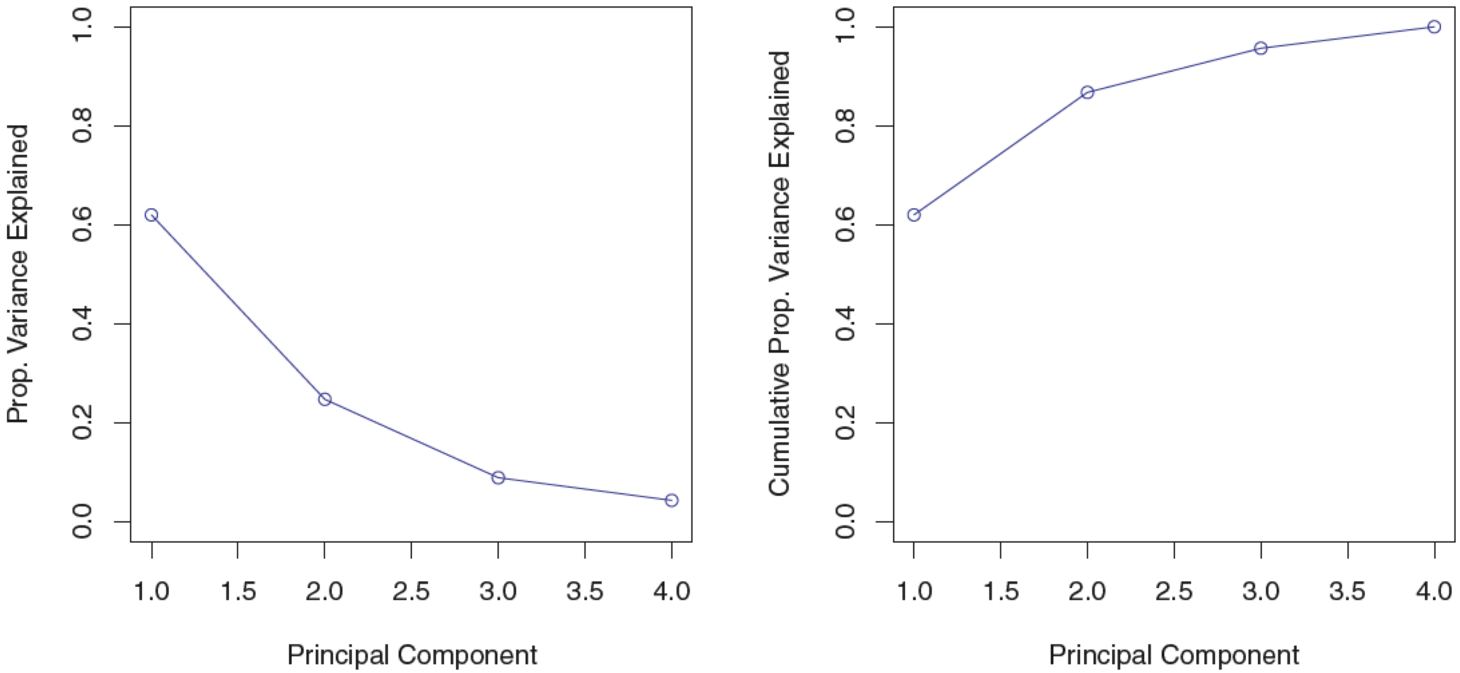
- How much of the information in a given data set is lost by projecting the observations onto the first few PC?
- The **proportion of variance explained (PVE)** by each PC

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}$$

- The total of PCs = $\min(n-1, p)$.
- PVEs of all PCs sum to 1.

On the Use of PCA

- How many PCs to use?
- Choose the smallest number of PCs required to explain a sizable amount of the variation in the data.
- Scree plot:



Principal Components Regression

- **Assumption:** the directions in which the predictors show the most variation are the directions that are associated with the response.

- Use the selected PCs as the predictors in a linear regression model fit using least squares.

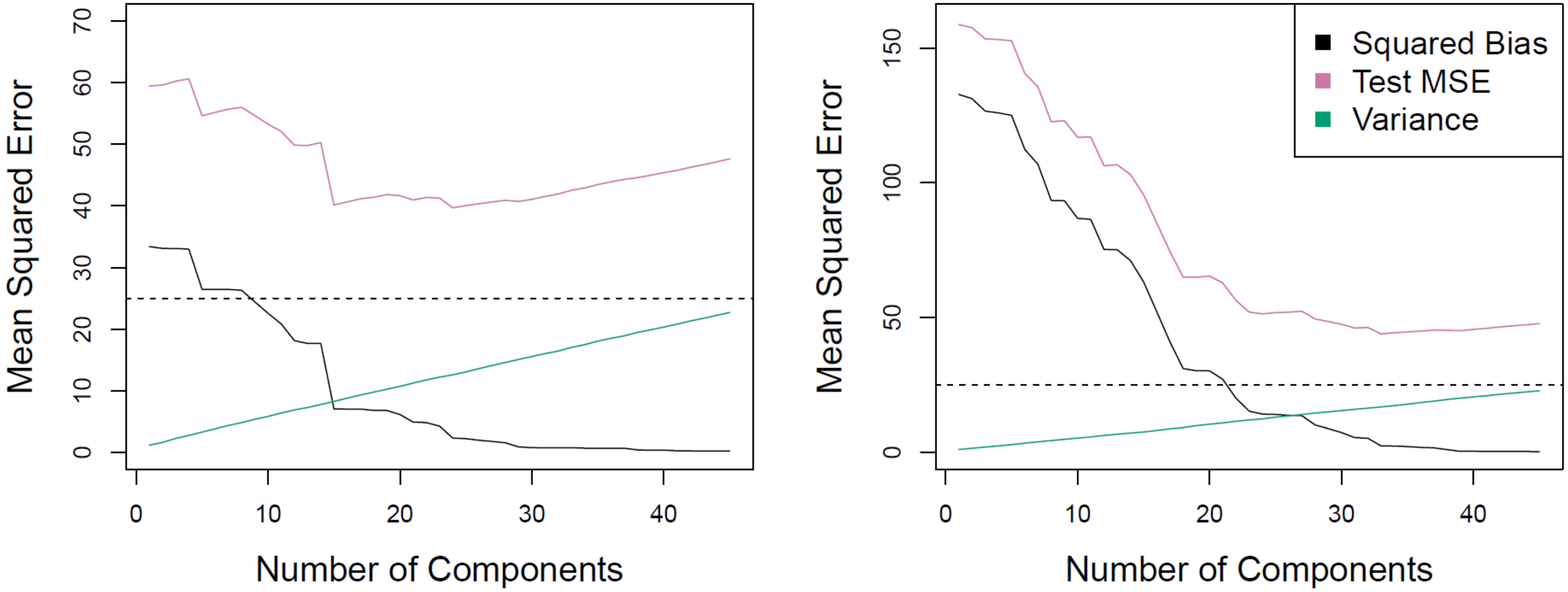
- PCR is not a feature selection method.

- Selecting PCs by cross validation.

- It works well when the first few PCs are sufficient to capture most of the variation in the predictors as well as the relationship with the response.

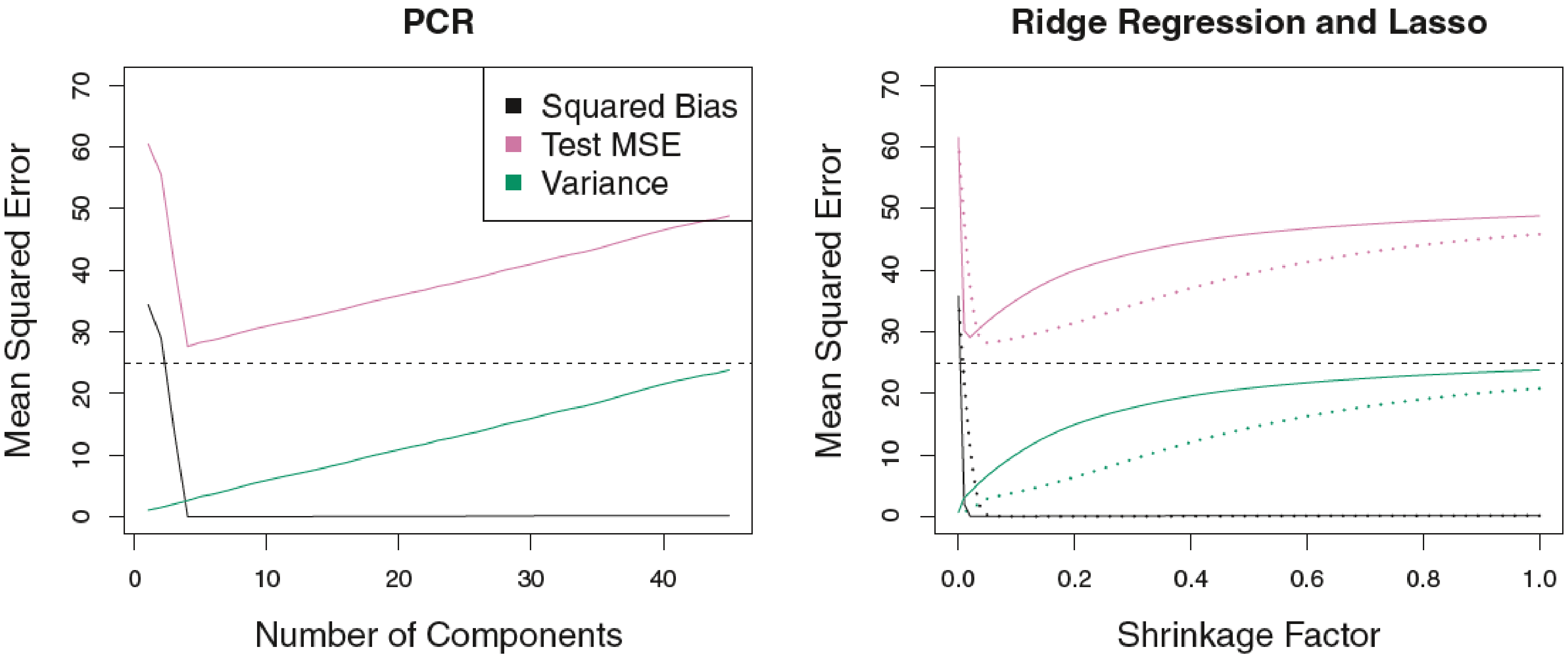
Simulation Study

- **Case 1:** response is a function of all the 45 predictors
- **Case 2:** response is a function of only 2 predictors



Simulation Study

- **Case 3:** response only depends on the first 5 PCs



Question (1)

Read slides of Topic 8. Then answer the following questions.

- What is the purpose of PCA?
- What are “scores” and “loadings” in PCA?
- Should we standardize each variable before PCA?
- If the data consist of p variables (i.e., X matrix has p columns), how many principal components will we obtain initially?
- How are the initial principal components ordered?
- How can we reduce the number of principal components?
- How can the final (reduced) principal components be used in statistical learning?

Question (2)

- What is PCR?
- Is PCR a feature selection method?
- When does PCR work well?