Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

M1586.002500 Information Engineering for CE Engineers
In-Class Material: Class 23
Support Vector Machine (ISL Chapter 9)

Review of support vector machine (SVM): Separate the class by a maximal margin hyperplane using tuning parameter and kernel to deal with non-separable and non-linear cases. An optimization algorithm is performed to obtain the hyperplane and only support vectors affect the classifier among dataset

1. **Support Vector Machine with More than Two Classes (K > 2 classes)**

   (a) One-versus-one approach

   Construct all-pairs of SVMs $\binom{K}{2}$

   SVM might compare the kth class, coded as $+1$, to the k'th class, coded as -1

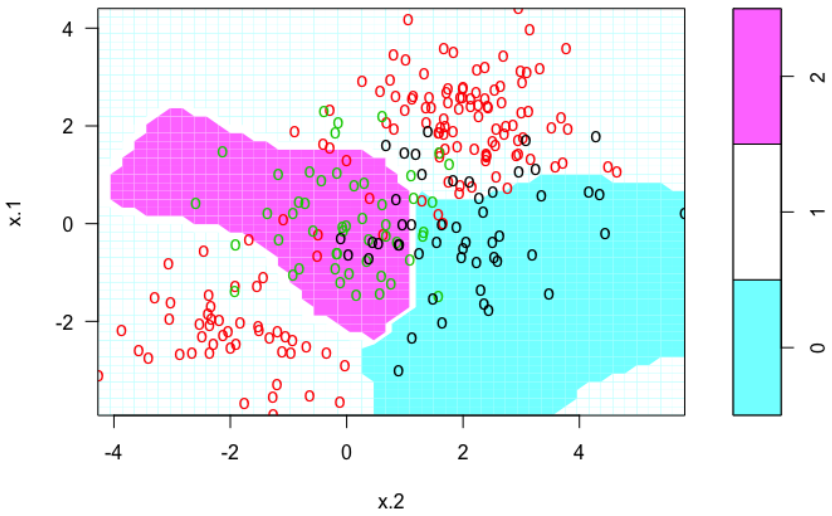
   The final classification is performed by assigning the test observation to the class to which it was most frequently assigned in these $\binom{K}{2}$ pairwise classification

   (b) One-versus-all approach

   Construct K SVMs

   SVM might compare the kth class to remaining $K-1$ classes (i.e. coded as $+1$ to kth class and -1 to others)

   Assign the observation x^* to the class for which $\beta_{0k} + \beta_{1k}x_1^* + \beta_{2k}x_2^* + \cdots + \beta_{pk}x_p^*$ is largest



Instructor: Junho Song
junhosong@snu.ac.kr

```
set.seed (1) # Assign seed for random generation

# Random generation for two cases X and corresponding y
x=matrix(rnorm(200*2), ncol=2)
x[1:100,]=x[1:100,]+2
x[101:150,]=x[101:150,]-2
y=c(rep(1,150),rep(2,50)) # 150 num of 1st class and 50 num of 2nd class

# Add 3rd class
x = rbind(x, matrix(rnorm(50*2), ncol=2))
y = c(y, rep(0,50))
x[y==0,2]=x[y==0,2]+2
dat=data.frame(x=x, y=as.factor(y))
par(mfrow=c(1,1))

# Plot
x1min = min(x[,1]); x1max = max(x[,1]);
x2min = min(x[,2]); x2max = max(x[,2])
plot(x,col=(y+1), xlab="X1", ylab="X2", xlim=c(x1min,x1max),
   ylim=c(x2min,x2max))

# Fit an SVM to the data (one-versus-one approach)
library(e1071)
svmfit=svm(y~., data=dat, kernel="radial", cost=10, gamma=1)

# Plot
plot(svmfit , dat, svSymbol = "o", dataSymbol = "o",  xlim=c(x2min-
   0.01,x2max), ylim=c(x1min-0.01,x1max))
```

2. Other Issues Regarding SVM

(a) SVM in "Loss+Penalty" form

Optimization problem for fitting the support vector classifier $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$ can be rewritten as

$$\text{minimize}\left\{\sum_{i=1}^{n} \max[0, 1 - y_i f(x_i)] + \lambda \sum_{j=1}^{p} \beta_j^2\right\}$$

"Hinge Loss" | Penalty

λ is corresponding to the nonnegative tuning parameter C

Thus, SVM takes the "Loss+Penalty" form which was shown in Ridge and Lasso relgressions, etc.

(b) Loss functions used for classification

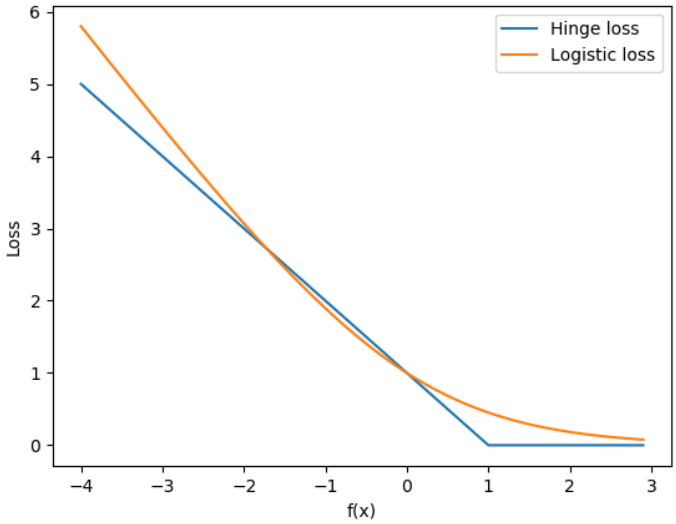
Definition: representing the cost paid for inaccuracy of predictions in classification problems

Instructor: Junho Song
junhosong@snu.ac.kr

Assumption: two classes with $y \in \langle 1, -1 \rangle$

Types of loss function

- Square loss: $\left(1 - yf(x)\right)^2$
- **Hinge loss**: $\max[0, 1 - yf(x)]$
- Logistic loss: $\frac{1}{\ln 2} \cdot \ln(1 + \exp(-yf(x)))$
- Cross entropy loss: $-t \cdot \ln\left(f(x)\right) - (1 - t) \cdot \ln\left(1 - f(x)\right)$, where $t = (1 + y)/2$



Hine loss: only support vectors play a role in the classifier obtained, i.e. observations on the correct side of the margin do not affect it. The hinge loss is zero if $y_i f(x_i) > 1$

Logistic loss: loss function for logistic regression is not exactly zero anywhere, but it is very small for observations that are far from the decision boundary

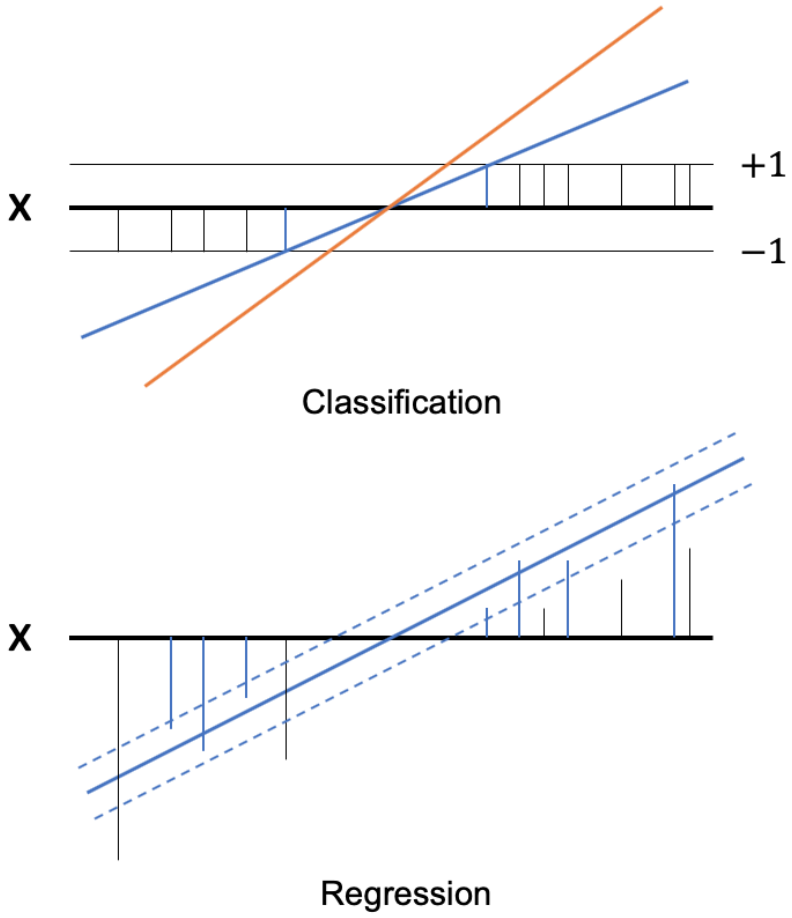
➔ Due to the similarities between their loss functions, logistic regression and the SVMs often give very similar results. When the classes are well separated, SVMs tend to behave better than logistic regression; in more overlapping regimes, logistic regression is often preferred

(c) SVM regression → Support Vector Regression

Seek coefficients that minimize a different type of loss, where only residuals larger in absolute value than some positive constant contribute to the loss function

The value y of all examples deviates less than the certain value

Instructor: Junho Song
junhosong@snu.ac.kr



	Classification	Regression
Goal	Minimize the loss	
Condition	All observations are classified correctly	The value y of all observations deviates less than the certain constant

Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

M1586.002500 Information Engineering for CE Engineers

In-Class Material: Class 24

Unsupervised Learning (ISL Chapter 10)

Unsupervised Learning: only a set of features, $X_1, \ldots, X_p$ measured, i.e. no response

1) Is there an informative way to visualize the data?
2) Can we discover subgroups among the variables or among the observations?

This chapter deals with

1) **Principal Components Analysis (PCA)**: data visualization and data pre-processing
2) **Clustering**: methods for discovering unknown subgroups

1. Principal Components Analysis (PCA)

(a) The process by which *principal* components are identified/computed, and the subsequent use of these components in understanding the data.

(b) The first principal component Z_1: projection on the direction of the largest variance

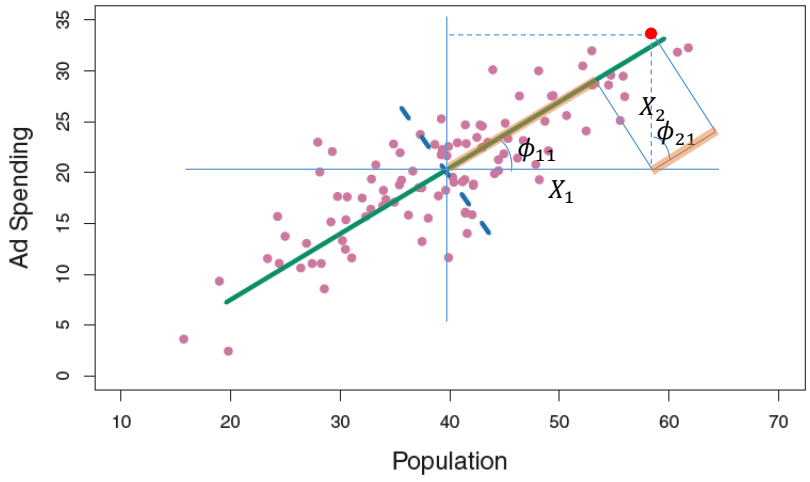
- Consider the normalized linear combination of the features $X_1, X_2, \cdots, X_p$.

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

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \ \text{(Normalized)}$$

where, $\phi_{11}, \cdots, \phi_{p1}$: the **loadings** of the first principal component.
The principal component loading vector: $\phi_1 = \begin{pmatrix} \phi_{11} & \phi_{21} & \cdots & \phi_{p1} \end{pmatrix}^{\mathrm{T}}$

Note: Loading ϕ_{j1} is the cosine of the angle between X_j and Z_1



- Note that, before PCA, the features $X_1, X_2, \cdots, X_p$ are centered to have mean zero (how?)

Instructor: Junho Song
junhosong@snu.ac.kr

- Find the loading vector that maximizes the variance of the samples $z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}, i = 1, \ldots, n$

  This is achieved by solving the optimization problem $\underset{\phi_{11},\cdots,\phi_{p1}}{\text{maximize}}\left\{\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2\right\}$, i.e.

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

  Note that $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$ is the sample variance because $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$, and thus the average of the $z_{11}, \cdots, z_{n1}$ is zero.

- $z_{11}, \cdots, z_{n1}$: **Scores** of the first principal component

  Geometric interpretation : The loading vector ϕ_1 with elements $\phi_{11}, \phi_{21}, \cdots, \phi_{p1}$ defines the direction in feature space along which the data vary most.

(c) The second principal component Z_2

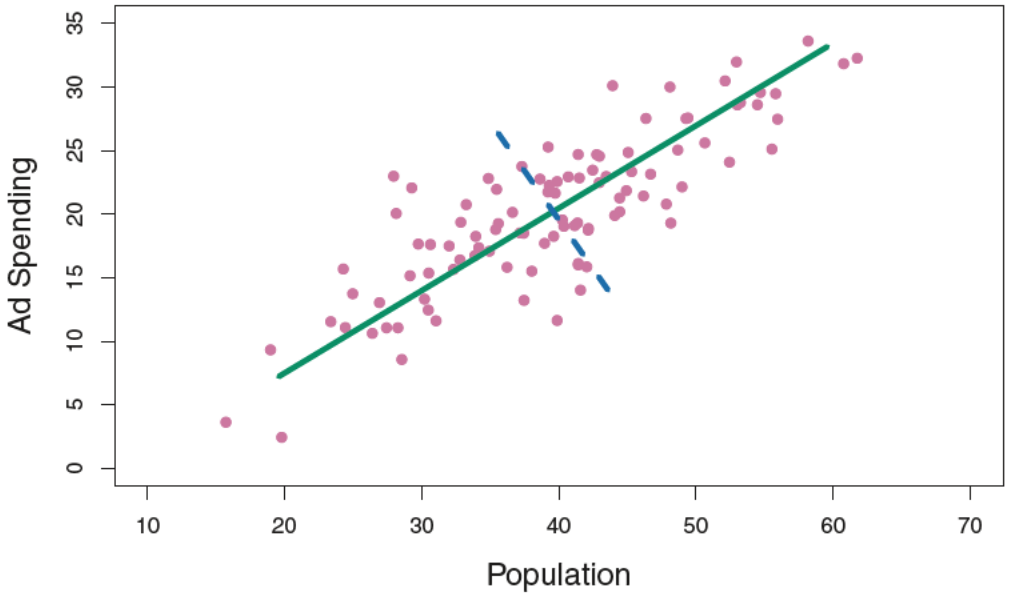
- The linear combination of $X_1, X_2, \cdots, X_p$ that has maximal variance out of all linear combinations that are *uncorrelated* with Z_1

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

  where, ϕ_2 is the second principal component loading vector, with elements $\phi_{12}, \phi_{22}, \cdots, \phi_{p2}$.

- The direction ϕ_2 is orthogonal to the direction ϕ_1 ($\because Z_1$ and Z_2 are uncorrelated)

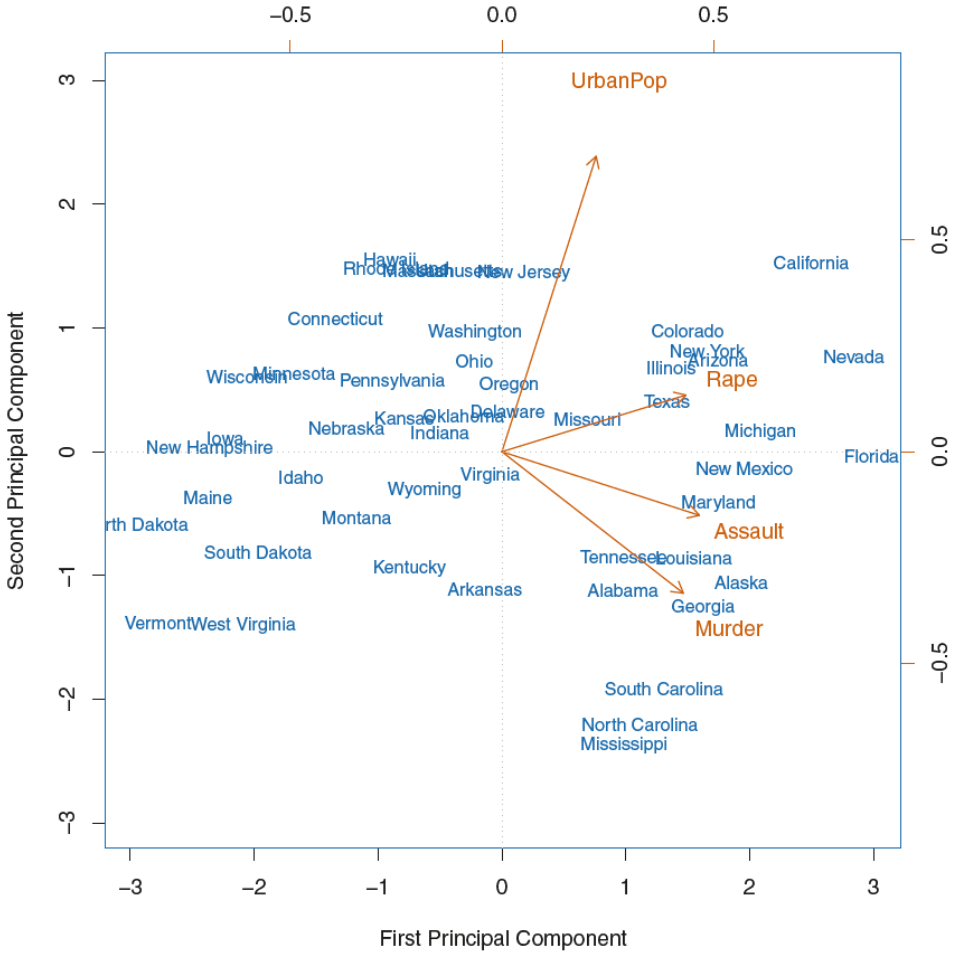
- Can be obtained by solving the optimization problem with the additional constraint



In the figure, green line: 1st principal component, and blue line: 2nd principal component

Instructor: Junho Song
junhosong@snu.ac.kr

(d) Visualization based on PCA: biplots



(e) Interpretation of principal components

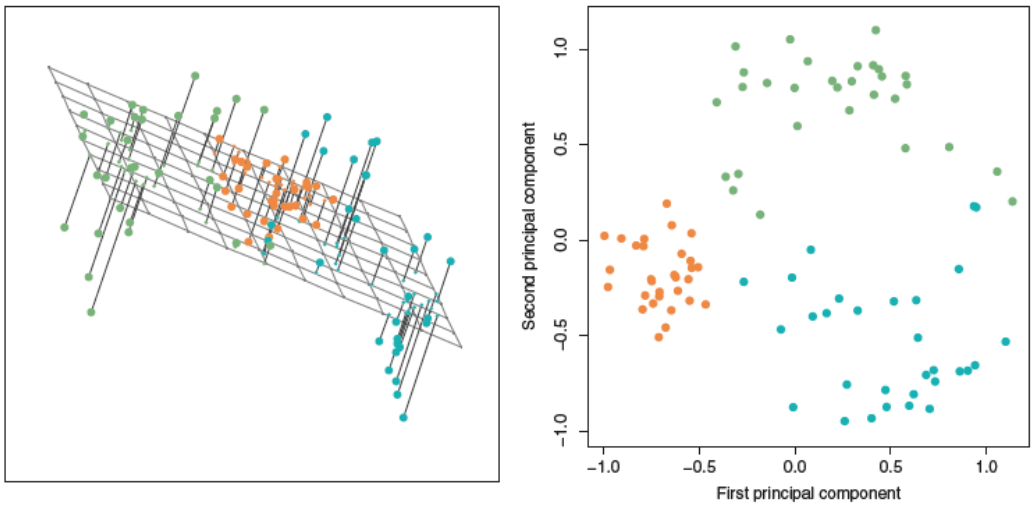
- Principal components provide low-dimensional linear surfaces that are closest to the observations
- A single dimension of the data that lies as close as possible to all of the data points → provide a good summary of the data
- The first M principal component score vectors and the first M principal component loading vectors provide the best M-dimensional approximation.

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}$$

M is sufficiently large → can give a good approximation to the data.

When $M = \min(n-1, p)$ → $x_{ij} = \sum_{m=1}^{M} z_{im}\phi_{jm}$

Instructor: Junho Song
junhosong@snu.ac.kr



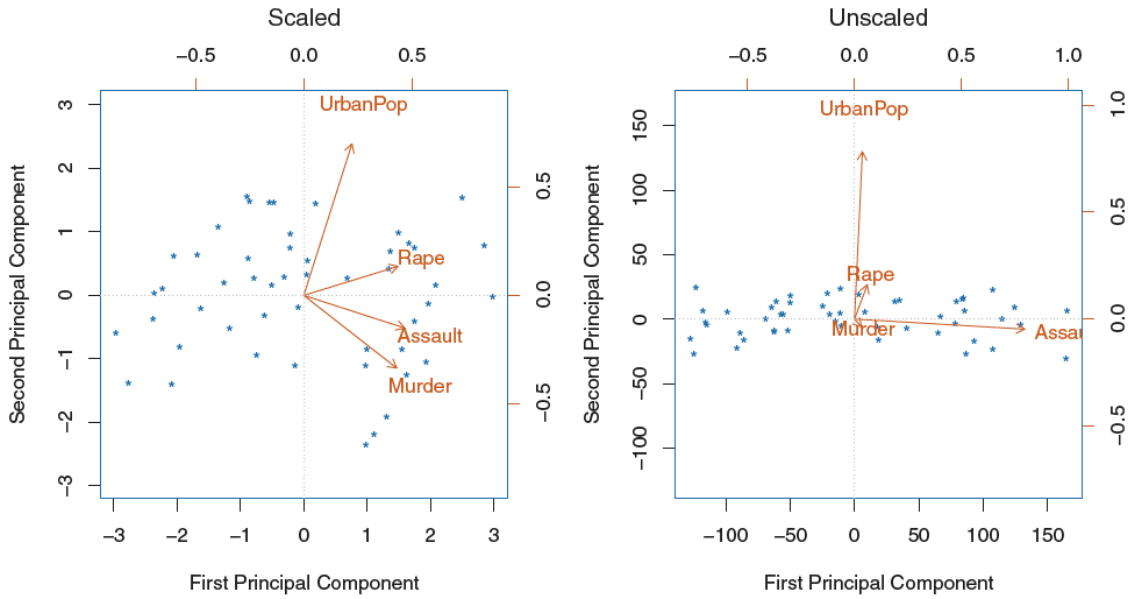
Left: The first two principal component directions span the plane that best fits the data
Right: The coordinates of the projection of the first two principal component score vectors

2. More on PCA

(a) Scaling the variables

Scale each of the variables to have standard deviation one.

When the variables are measured in different units, unscaled variables can affect the size of variance → may affect the model obtained



Two principal component biplots (Left: Scaled, Right: Unscaled)

(b) Uniqueness of the principal components

Instructor: Junho Song
junhosong@snu.ac.kr

Each principal component loading vector is unique, up to a sign flip → This is OK because flipping the sign has no effect on the *direction*

(c) The Proportion of Variance Explained (PVE)

The total variance present in a data set is defined as

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

and the variance explained by mth principal component is

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

Therefore, the PVE of the mth principal component is given by

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

- The PVE of each principal component is a positive quantity
- The cumulative PVE of the first M principal components can be calculated by simply sum the PVE of the mth principal component over each of the first M PVEs.
- It has total $\min(n-1, p)$ principal components, and sum of PVEs is equal to one.

(d) Deciding how many principal components to use

Use the first few principal components in order to visualize or interpret the data.

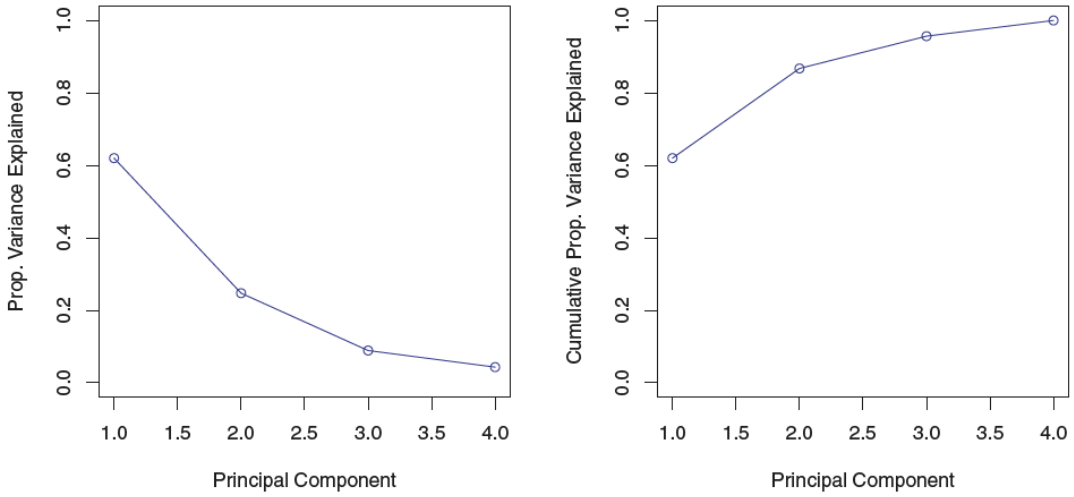
Usually, determine the number of principal components required to visualize the data by examining a *scree plot* (left figure below)

- Choose the smallest number of principal components that are required in order to explain a sizable amount of the variation in the data.
- Done by eyeballing the scree plot, and looking for a point at which the proportion of variance explained by each subsequent principal component drops off → *elbow* in the scree plot

But, this type of visual analysis is inherently *ad hoc*.

How many principal components are enough? → no well-accepted objective protocol

Instructor: Junho Song
junhosong@snu.ac.kr



Left: a scree plot depicting the proportion of variance explained
Right: the cumulative proportion of variance explained by the four principal components

```
?USArrests
states=row.names(USArrests)

# checking means and variances of features
apply(USArrests, 2, mean)
apply(USArrests, 2, var)

# PCA of the feature set
pr.out=prcomp(USArrests, scale=TRUE)
# scale = TRUE --> (1) centers the variables to have mean zero
# (2) scale the variables to have standard deviation one

dim(pr.out$x) # x: scores of each observation

#biplot of the first two PC
biplot(pr.out, scale=0)

#flipping
pr.out$rotation=-pr.out$rotation
pr.out$x=-pr.out$x
biplot(pr.out, scale=0)

pr.var=pr.out$sdev^2 # variance explained by each PC
pve=pr.var/sum(pr.var) # proportion of variance explained

# scree plot and cumulative PVE plot
plot(pve, xlab="Principal Component", ylab="Proportion of Variance
   Explained", ylim=c(0,1),type='b')
plot(cumsum(pve), xlab="Principal Component", ylab ="Cumulative Proportion
    of Variance Explained", ylim=c(0,1), type='b')
```