Part.02
회귀분석

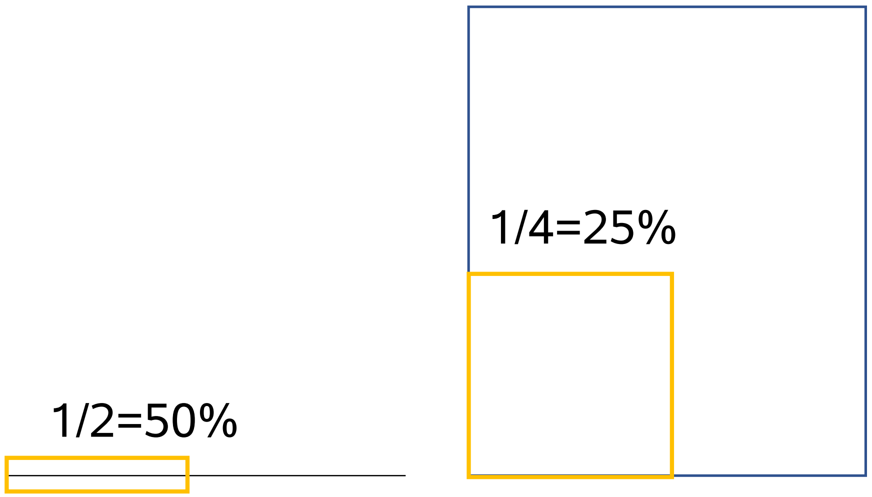
PCA - 차원축소

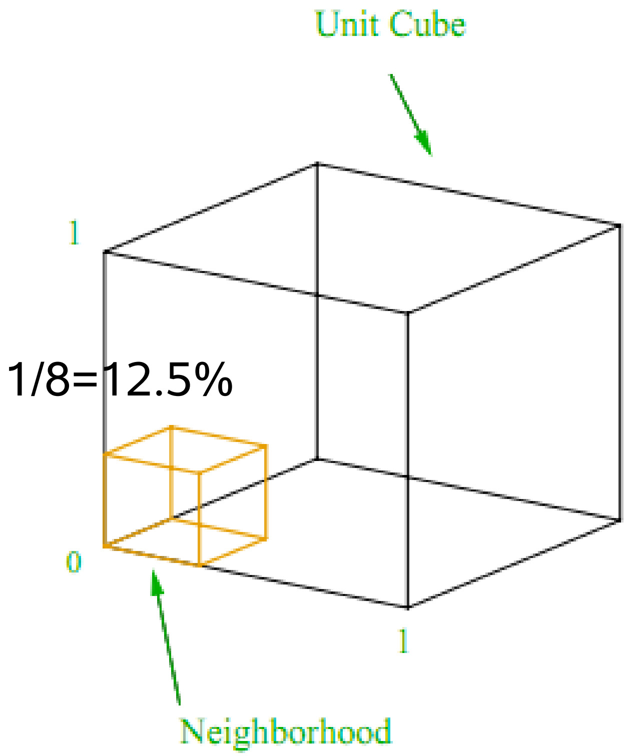
FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 김강진

I 차원의 저주

- 각 변수의 50%영역에 해당하는 자료를 가지고 있다고 할때,
  - 전체 자료의 얼마만큼을 확보할 수 있는가?





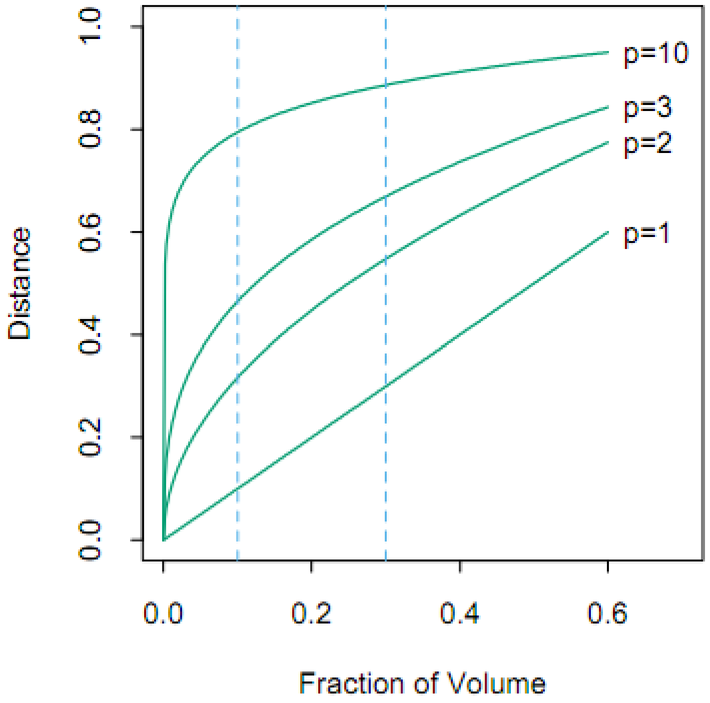


그림 발췌: “The Elements of Statistical Learning”, Tibshirani and Friedman

I 차원의 저주

- 관측치의 수는 한정되어 있음.
  - 차원이 커질 수록 한정된 자료는 커진 차원의 패턴을 잘 설명하지 못한다.
  - 차원이 증가함에 따라 model complexity가 기하급수적으로 높아짐.

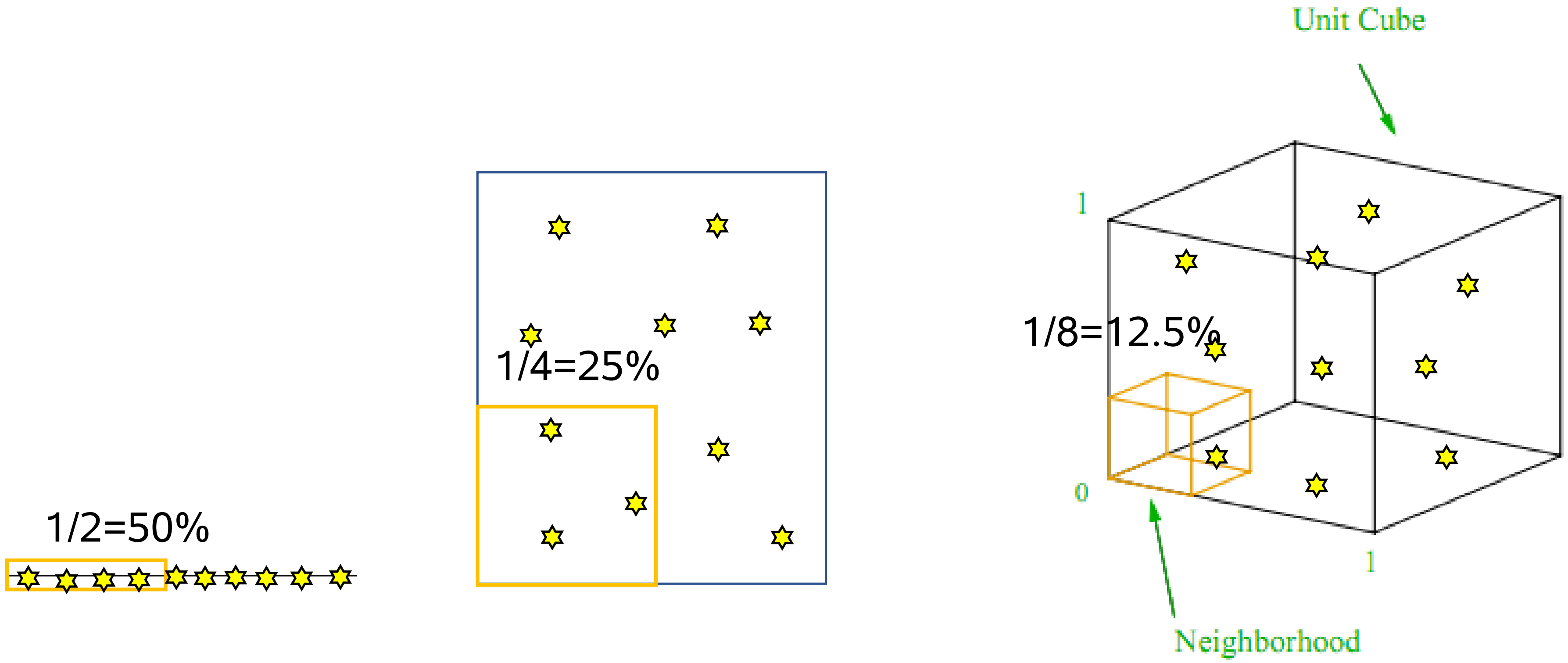


그림 발췌: “The Elements of Statistical Learning”, Tibshirani and Friedman

I 차원 축소의 필요성

- 가까이에 있는 변수가 가지는 값을 예측 값으로 하는 모델이 있다고 하자. (k-Nearest Neighborhood)
  - 쓸데 없는 변수가 추가되는 것은 모델의 성능에 매우 악영향을 끼침.
    - 상관계수가 매우 큰 서로 다른 독립 변수
    - 예측하고자 하는 변수와 관련이 없는 변수

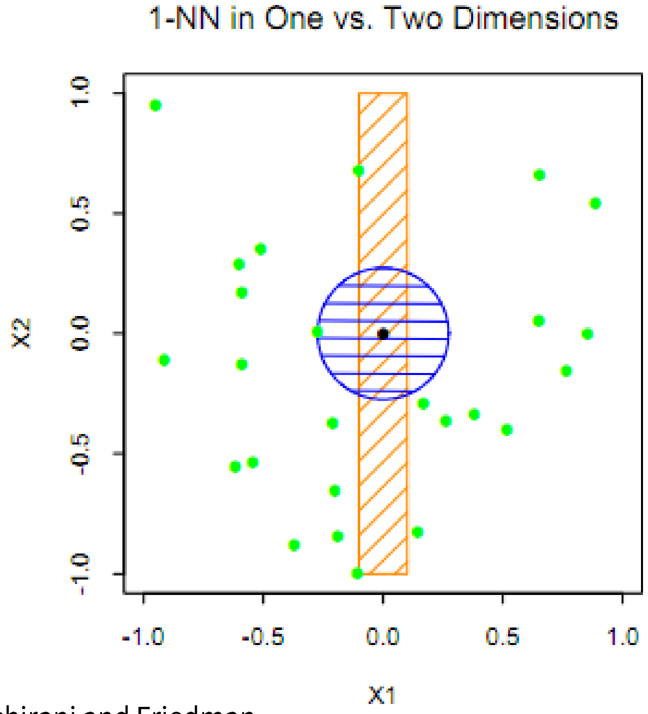


그림 발췌: “The Elements of Statistical Learning”, Tibshirani and Friedman

I 차원 축소법

- 상관계수가 높은 변수 중 일부를 분석에서 제외?
  - 정보의 손실 발생.
  - 상관계수가 0.8이라고 하면, 0.2에 해당하는 정보는 버려지게 됨.

- 차원을 줄이면서 정보의 손실을 최소화 하는 방법
  - Principal component을 활용

- 이외의 방법
  - 변수 선택법
  - penalty 기반 regression
  - convolutional neural network
  - drop out & bagging

- End of the clip.

Part.02

회귀분석

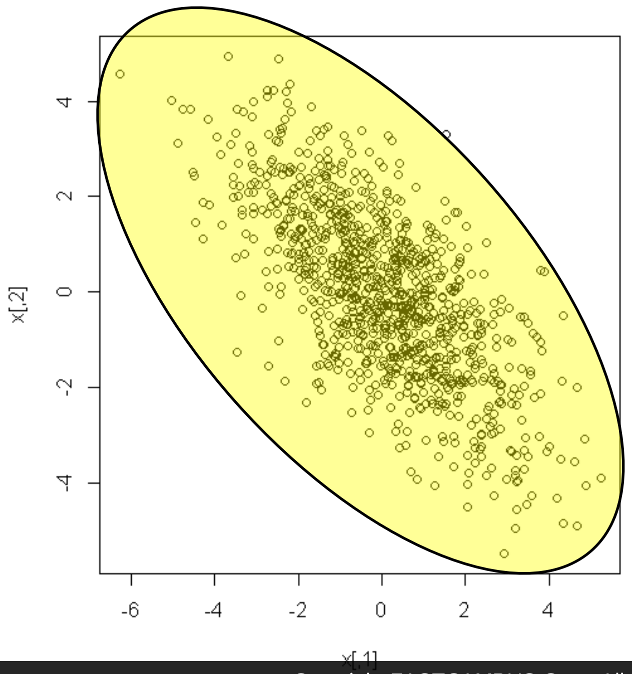
PCA – 공분산 행렬의 이해

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 김강진

I 공분산 행렬의 개념

- 공분산 행렬(Covariance matrix)의 정의
  - X_1, X_2이 음의 상관관계를 가지므로, 둘의 공분산은 음수일 것. X가 centering 되어 있다면,
  - $Cov\left(\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right) = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Var(X_2) \end{bmatrix} = \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}$
  - X가 centering 되어 있다면, $Cov(\mathbf{X})=(\mathbf{X}^T\mathbf{X})/(n-1)$

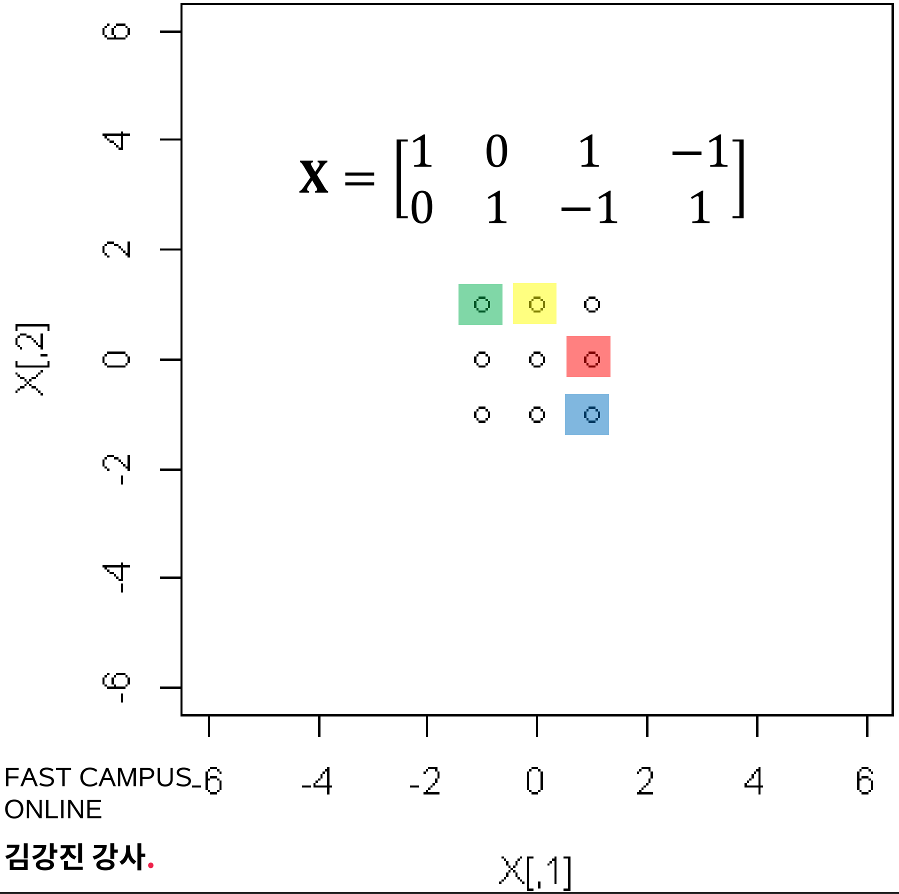


I 공분산 행렬의 개념

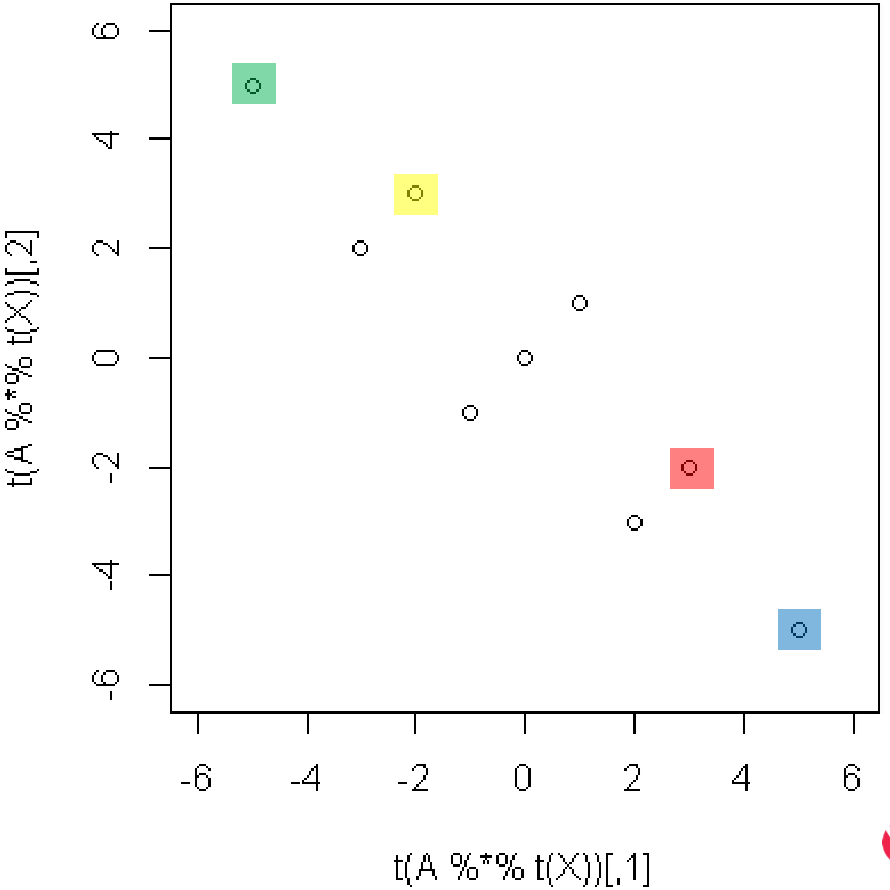
- 공분산의 형태 파악
  - 점과 내적연산을 할 경우, 점의 위치를 이동시켜
  - 해당 공분산 구조와 비슷한 형태를 가지게 된다.

$$\mathbf{AX} = \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & -1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & -2 & 5 & -5 \\ -2 & 3 & -5 & 5 \end{bmatrix}$$



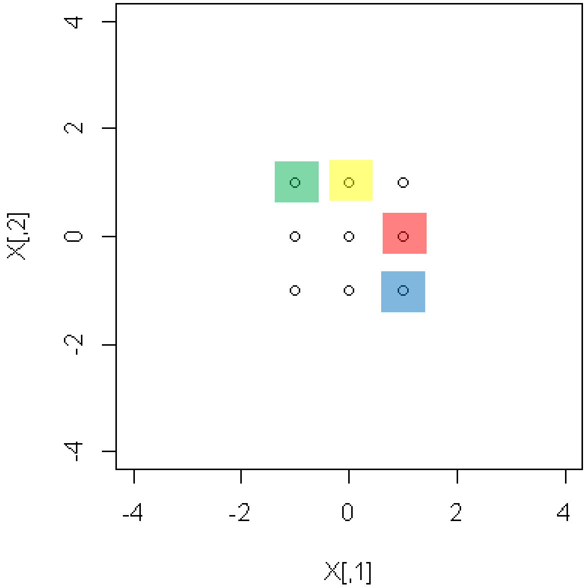
$$\mathbf{A} \equiv \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}$$



I 공분산 행렬의 개념

- 대칭행렬이지만, 일반적인 공분산이 아닌 경우
  - Positive definite이 아님.

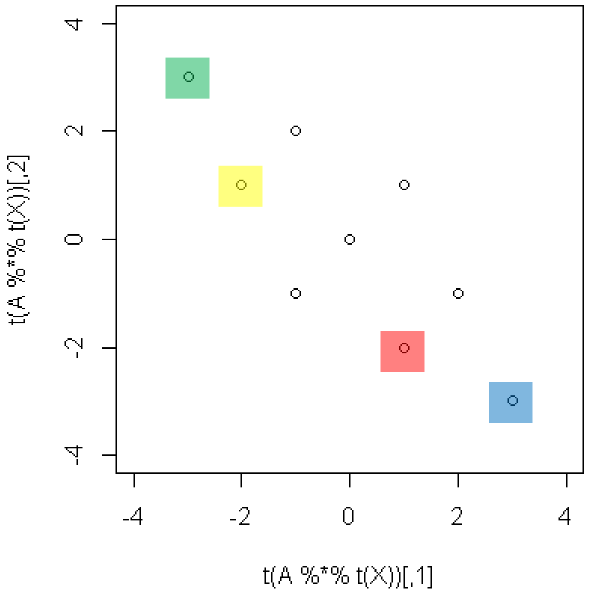
$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & -1 & 1 \end{bmatrix}$$



$$\mathbf{A} \equiv \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}$$

$$\mathbf{AX} = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & -1 & 1 \end{bmatrix}$$

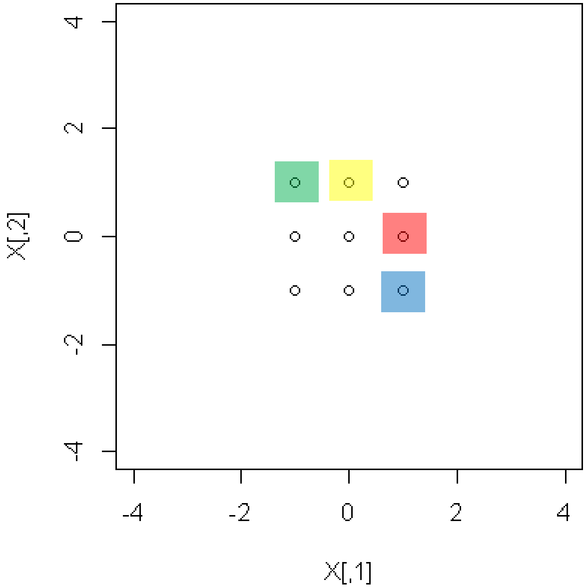
$$= \begin{bmatrix} 1 & -2 & 3 & -3 \\ -2 & 1 & -3 & 3 \end{bmatrix}$$



| 공분산 행렬의 개념

- 대칭행렬이지만, 일반적인 공분산이 아닌 경우
  - 행렬식이 0인 경우.

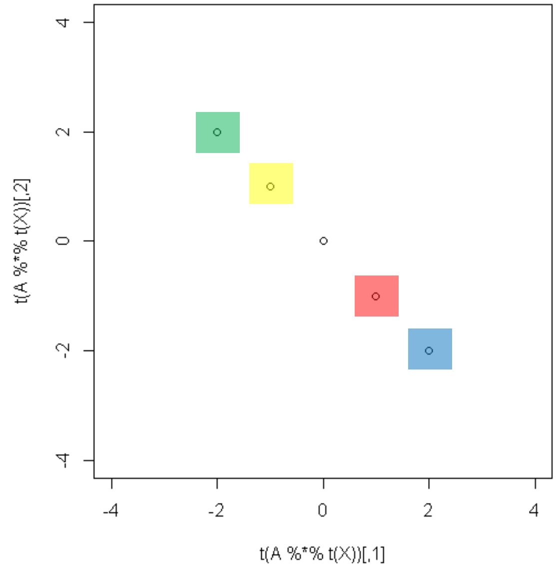
$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & -1 & 1 \end{bmatrix}$$



$$\mathbf{A} \equiv \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\mathbf{AX} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & -1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -1 & 2 & -2 \\ -1 & 1 & -2 & 2 \end{bmatrix}$$



- End of the clip.

Part.02

회귀분석

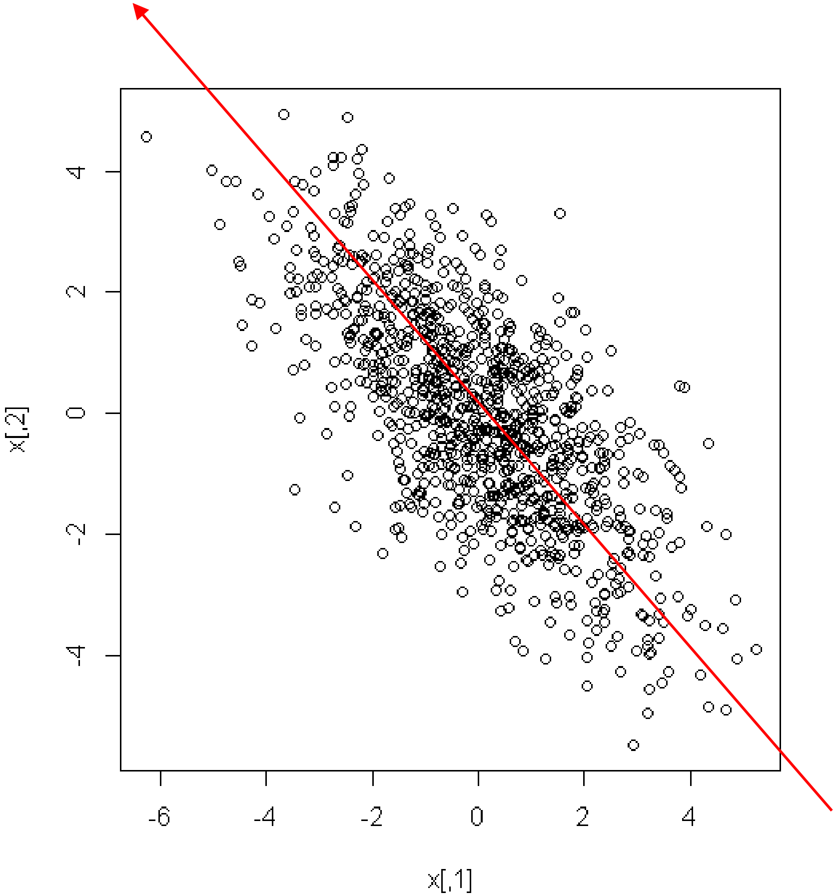
PCA – Principal Conponents 의 이해

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 김강진

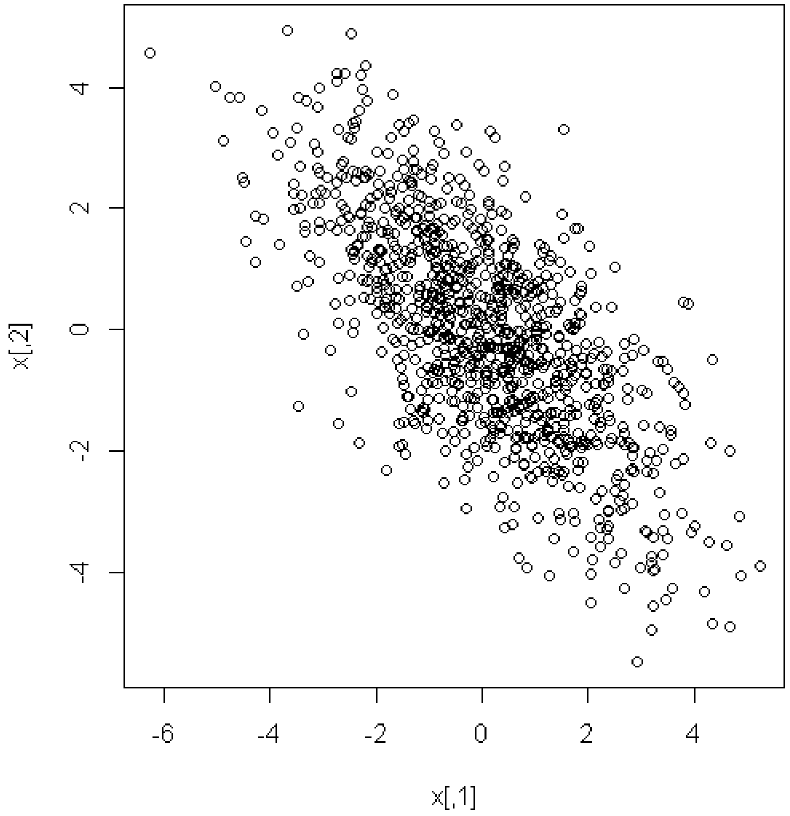
Principal Components의 개념

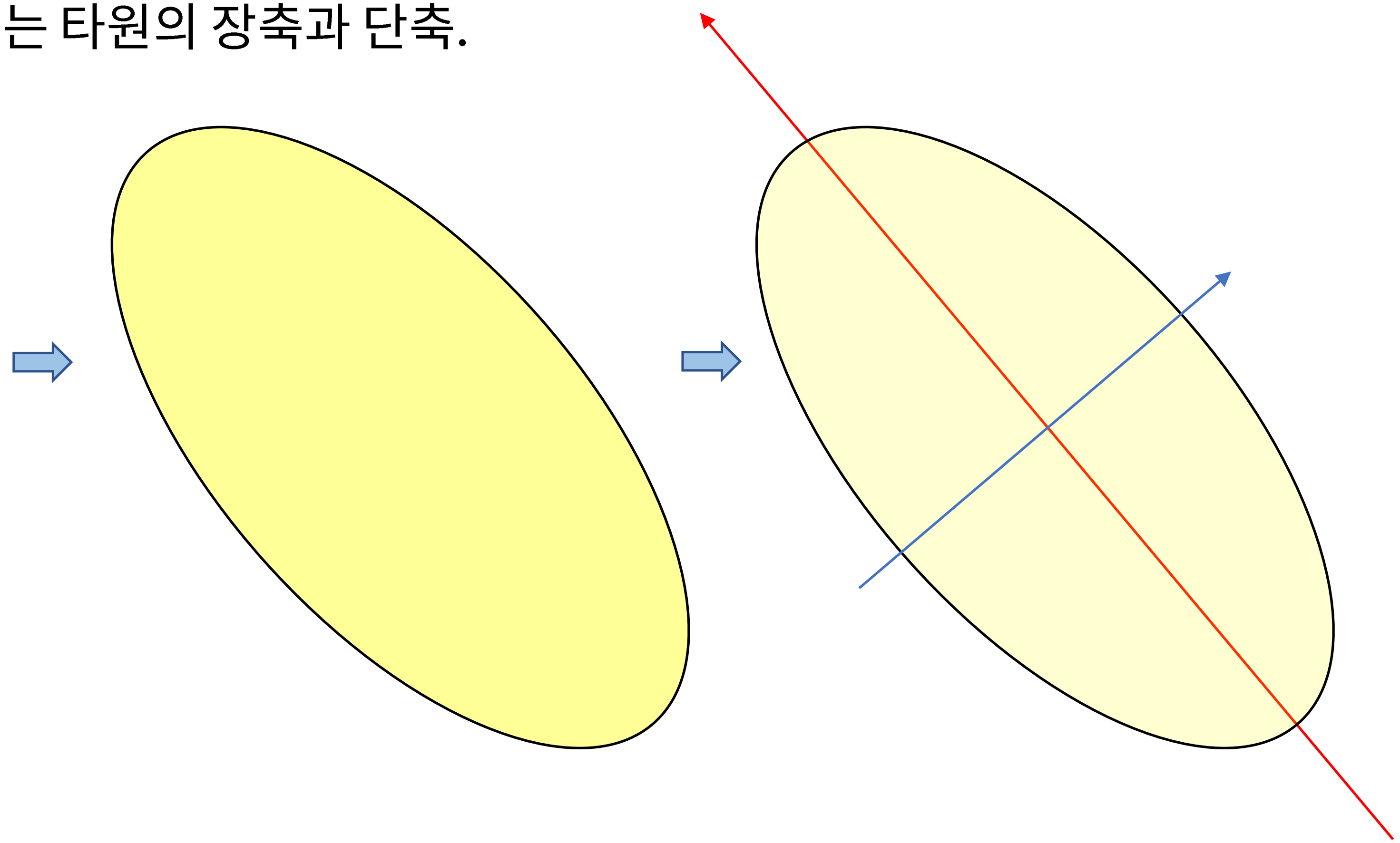
- 차원을 줄이면서 정보의 손실을 최소화 하는 방법
    - 더 적은 개수로 데이터를 충분히 잘 설명할 수 있는 새로운 축을 찾아냄.



I Principal Components의 개념

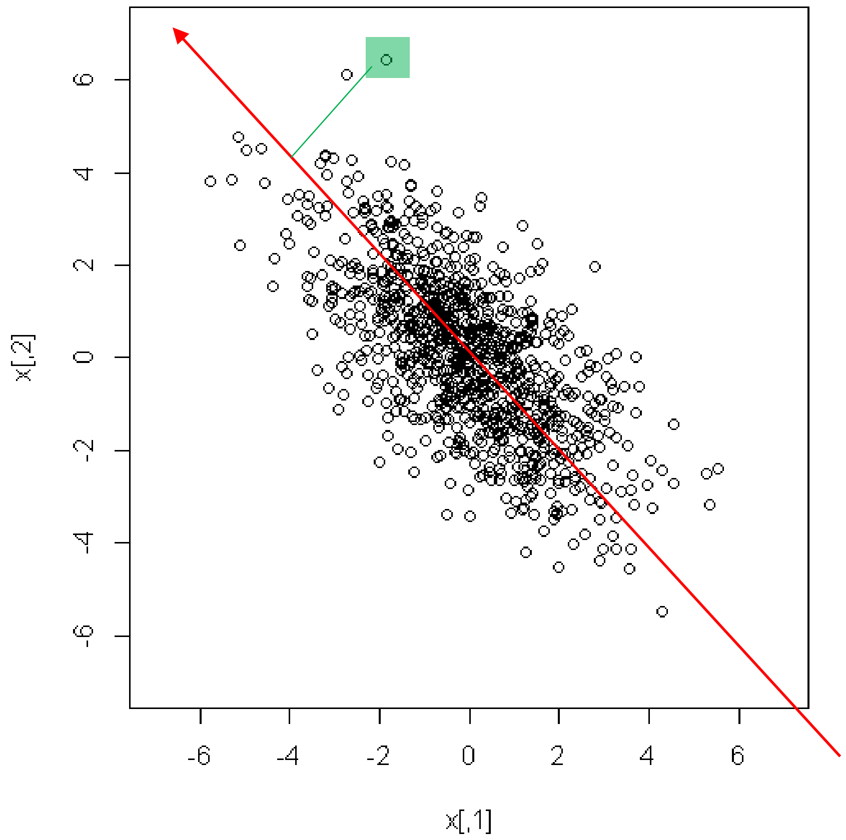
- Principal Components (PC)를 얻어내는 것
  - 공분산이 데이터의 형태를 변형시키는 방향의 축과 그것에 직교하는 축을 찾아내는 과정.
  - 2차원의 경우 공분산이 나타내는 타원의 장축과 단축.

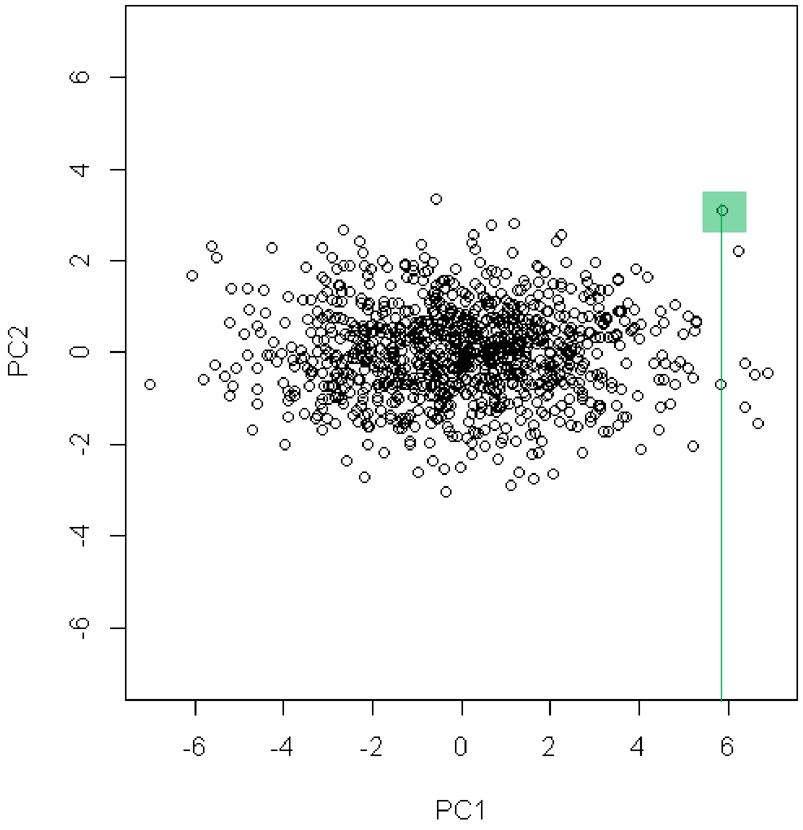




I Principal Components의 개념

- Principal components = PC = PC score
  - 찾아낸 새로운 축에서의 좌표값을 의미.
  - 새로운 축에 내린 정사영





- End of the clip.

Part.02

회귀분석

PCA 수학적 개념이해 – 행렬연산, 행렬식, 특성방정식

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 김강진

I Matrix

- 행렬식 (determinant) $|A|$, det(A) 구하기
- 2 by 2 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \Rightarrow |A| = a_{11}a_{22} - a_{12}a_{21}$$

- In general,

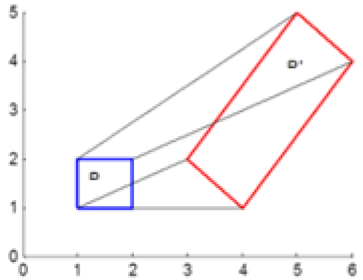
$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$\Rightarrow |A| = +a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$= a_{11}\left(a_{22}a_{33} - a_{23}a_{32}\right) - a_{12}\left(a_{21}a_{33} - a_{23}a_{31}\right) + a_{13}\left(a_{21}a_{32} - a_{22}a_{31}\right)$$

I Matrix

- 행렬식의 기하학적 의미
  - $|A|$는 A 선형변화의 스케일 성분을 의미.
  - $D' = AD$일 때,
    - $area(D') = |A|area(D)$
  - $A = \begin{bmatrix} 1 & 2 \\ -1 & 3 \end{bmatrix}$, $|A|$=5인 경우.



    - P'의 면적은 5가 된다.
  - $|A|$=0인 경우, D'는 선분이 된다.

이미지출처: https://darkpgmr.tistory.com/104

I 행렬식의 활용

- A 의 역행렬이 존재할 경우, 다음과 같이 계산할 수 있다.
  - $C = AB$
  - $B = A^{-1}C$
- 따라서 아래와 같이 계산할 수 있다.
  - $AB = 0$
  - $A^{-1}B = A^{-1}0$
  - $B = 0$
- 행렬식과 역행렬의 존재성의 관계
  - 행렬식 $|A| = 0$인 경우, 역행렬은 존재하지 않는다.
  - 행렬식은 역행렬 존재성에 대한 판별식 역할을 한다.

| 행렬식의 활용

- (예제)
  - $A = \begin{bmatrix} 4 & 2 \\ 3 & 5 \end{bmatrix}$
  - A의 역행렬은 존재하는가?

Part.02

회귀분석

PCA 수학적 개념이해 – Eigen vector, eigen value

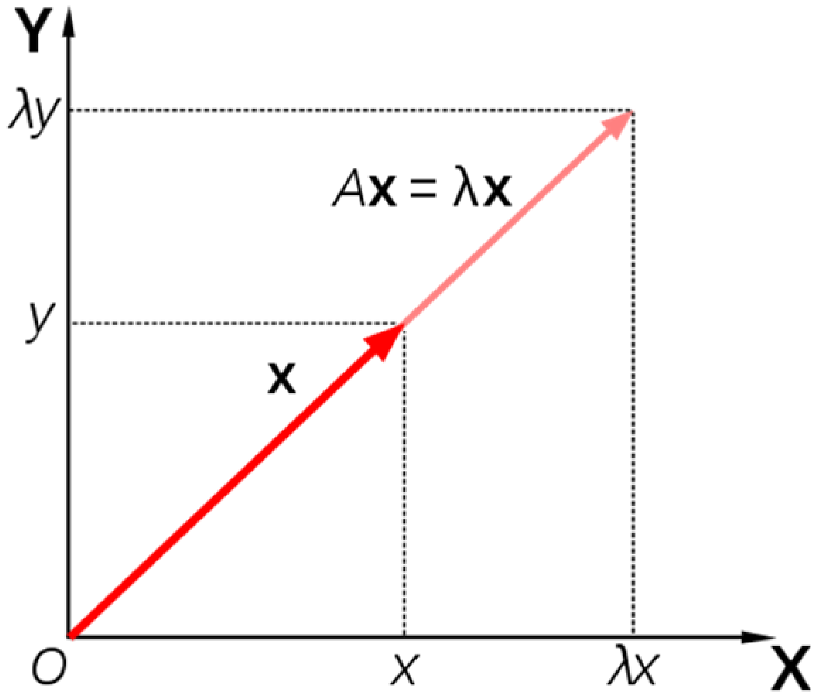
FASTCAMPUS ONLINE

머신러닝과 데이터분석 A-Z

강사. 김강진

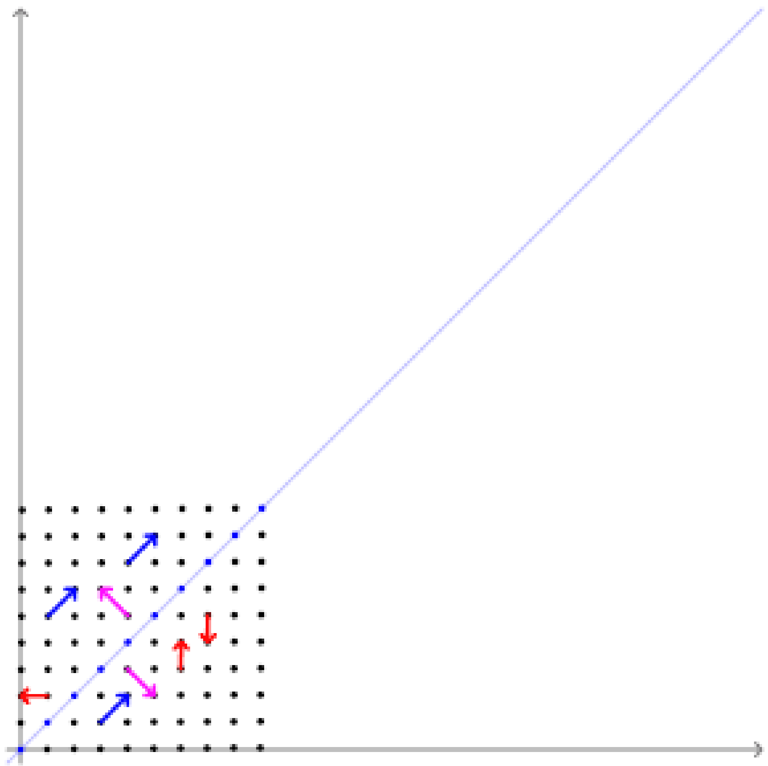
I Eigen value, eigen vector

- 정방행렬 A에 대하여,
- 아래를 만족할 경우 v 는 고유 벡터 (eigen vector) 이고 and λ 는 고유값 (eigen value)이다.
    - $Av = \lambda v$
    - 아래와 같이 A는 v를 선형변환한다.



I Eigen value, eigen vector

- 기하학적으로는,
  - 임의의 점에 대하여 A라는 transformation을 행할 때 고유 벡터는 방향이 바뀌지 않는다는 의미.
  - 고유값은 그 변화되는 스케일의 정도.



I Eigen value, eigen vector

- Eigen value, eigen vector의 표현
  - $(A - \lambda I)v = 0$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \lambda\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix} - \begin{bmatrix} \lambda x_1 \\ \lambda x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} a_{11}x_1 + a_{12}x_2 - \lambda x_1 \\ a_{21}x_1 + a_{22}x_2 - \lambda x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} (a_{11} - \lambda)x_1 + a_{12}x_2 \\ a_{21}x_1 + (a_{22} - \lambda)x_2 \end{bmatrix} = 0$$

  - 위의 식에서, x=0이 아닌 다른 해가 존재하려면,
  - $\begin{bmatrix} (a_{11} - \lambda) & a_{12} \\ a_{21} & (a_{22} - \lambda) \end{bmatrix} = A - \lambda I$ 가 역행렬이 존재하지 않아야 한다.
  - 행렬식 $|A| = 0$이게 하는 λ를 계산.

Eigen value, eigen vector

- Eigen value, eigen vector의 계산
  - (예제) $A = \begin{bmatrix} 4 & 2 \\ 3 & 5 \end{bmatrix}$
  - λ = 7, 2
  - $\begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Part.02

회귀분석

PCA 수학적 개념이해 – Singular Value Decomposition (SVD)

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 김강진

I Singular-value decomposition

- Singular-value decomposition (SVD)
- $\mathrm{n} \times p$ Matrix $\mathbf{X}$ 를 아래와 같은 요소로 나누는 것을 SVD라 한다.
  - $\mathbf{X} = \mathbf{UDV^T}$
  - $\mathbf{U}$: $\mathrm{n} \times p$, $\mathbf{D}$: $\mathrm{p} \times p$, $\mathbf{V}$: $\mathrm{p} \times p$
  - $\mathbf{V^T V} = \mathbf{I}_p$, $\mathbf{U^T U} = \mathbf{I}_p$, $\mathbf{D}$: diagonal matrix
  - Column vectors of $\mathbf{V}$: eigen vectors of $\mathbf{X^T X}$
  - Diagonal entries of $\mathbf{D}$: eigen values of $\mathbf{X^T X}$

- 위와 같이 SVD를 통하여, 임의의 matrix의 공분산 구조 행렬의 eigen vector, eigen value를 얻을 수 있다.
  - $\mathbf{X}$이 centered 되어있다면, $\mathbf{X^T X}$ 는 $\mathbf{X}$의 공분산 구조임.

I SVD와 eigen vector, eigen value와의 연관성

- $\mathbf{V} = [v_1 \quad \cdots \quad v_p]$에서, $v_1 \quad \cdots \quad v_p$가 eigen vectors인 이유
  - SVD에 의해, $\mathrm{X} = \mathrm{UDV}^{\mathrm{T}}$이므로,
    - $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$

      $(\mathbf{X}^T\mathbf{X})\mathbf{V} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{D}^2$

      $(\mathbf{X}^T\mathbf{X})[v_1 \quad \cdots \quad v_p] = [v_1 \quad \cdots \quad v_p]\mathbf{D}^2 = [d_1^2 v_1 \quad \cdots \quad d_p^2 v_p]$
    - $(\mathbf{X}^T\mathbf{X})v_i = d_i^2 v_i,\ i = 1,\ldots,p$
      - v_i: eigen vectors
      - d_i^2: eigen values

Part.02

회귀분석

PCA – PCA 수행과정 및 수학적 개념 적용

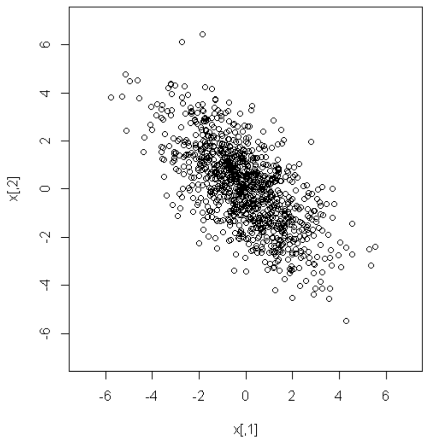
FASTCAMPUS
ONLINE

머신러닝과 데이터분석 A-Z

강사. 김강진

I PCA 수행 과정

- 1. Mean centering
  - 1000개의 관측치
  - 평균값: Z1: -0.0289, Z2: 0.0288



	Z1	Z2
0	2.0046	-1.0718
1	0.1823	0.2250
2	-0.5197	0.2620
3	-0.6507	-1.7142
4	0.3236	-3.0723

	X1	X2
0	2.0334	-1.1006
1	0.2112	0.1962
2	-0.4909	0.2332
3	-0.6219	-1.7430
4	0.3525	-3.1011

I PCA 수행 과정

- 2. SVD 수행
  - 3개 관측치에 대해서만 나타냄.

$$\mathbf{X} = \mathbf{UDV}^T = \begin{bmatrix} -0.0309 & -0.0216 \\ -0.0001 & -0.0092 \\ 0.0071 & 0.0060 \end{bmatrix} \begin{bmatrix} 71.54 & 0 \\ 0 & 31.15 \end{bmatrix} \begin{bmatrix} -0.70 & 0.71 \\ 0.71 & 0.70 \end{bmatrix}^T$$

I PCA 수행 과정

- 3. SVD 결과를 활용하여 공분산의 eigen vector, eigen value 구하기

$$\mathbf{U}=\begin{bmatrix}-0.0309 & -0.0216\\ -0.0001 & -0.0092\\ 0.0071 & 0.0060\end{bmatrix}, \mathbf{D}=\begin{bmatrix}71.54 & 0\\ 0 & 31.15\end{bmatrix}, \mathbf{V}=\begin{bmatrix}-0.70 & 0.71\\ 0.71 & 0.70\end{bmatrix}$$

- SVD와 공분산 행렬의 eigen vector, eigen value의 관계

(SVD 복습) $(\mathbf{X}^T\mathbf{X})v_i = d_i^2 v_i$

$$Cov(\mathbf{X})=\frac{(X^TX)}{n-1}$$

$$Cov(\mathbf{X})v_i = \frac{(X^TX)}{n-1}v_i = \frac{d_i^2}{n-1}v_i$$

- $\lambda_1 = \frac{(71.54)^2}{1000-1} = 5.12, \lambda_2 = \frac{(31.15)^2}{1000-1} = 0.97, v_1 = \begin{bmatrix}-0.70\\ 0.71\end{bmatrix}, v_2 = \begin{bmatrix}0.71\\ -0.70\end{bmatrix}$

I PCA 수행 과정

- 4. PC score 구하기
  - $\mathbf{X} = \mathbf{UDV}^T$
  - $\mathbf{XV} = \mathbf{UD} = \begin{bmatrix} -0.0309 & -0.0216 \\ -0.0001 & -0.0092 \\ 0.0071 & 0.0060 \end{bmatrix} \begin{bmatrix} 71.54 & 0 \\ 0 & 31.15 \end{bmatrix} = \begin{bmatrix} -2.21 & -0.67 \\ -0.01 & -0.29 \\ 0.51 & 0.19 \end{bmatrix}$
- 5. PC score를 활용하여 분석 진행
  - PC score를 설명변수로 활용하여 분석 진행.
    - $\mathbf{Y} = [\mathbf{1} \quad PC_1 \quad \cdots \quad PC_q] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_q \end{bmatrix} + \varepsilon$

- End of the clip.

Part.02

회귀분석

| PCA – PCA의 심화적 이해

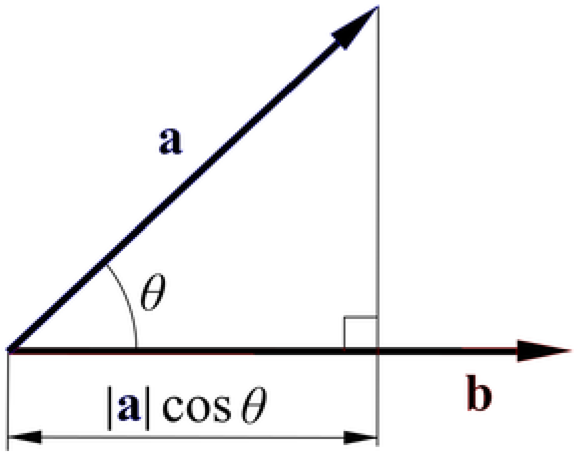
FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z|

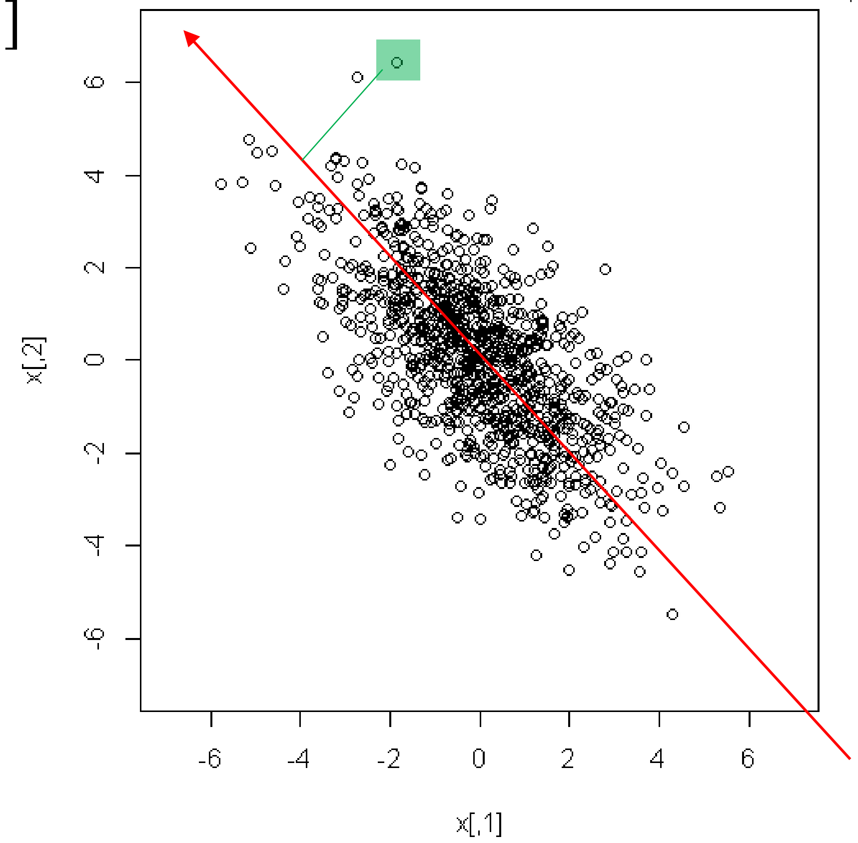
강사. 김강진

I PCA 수행 과정

- PC score 의 기하학적 해석

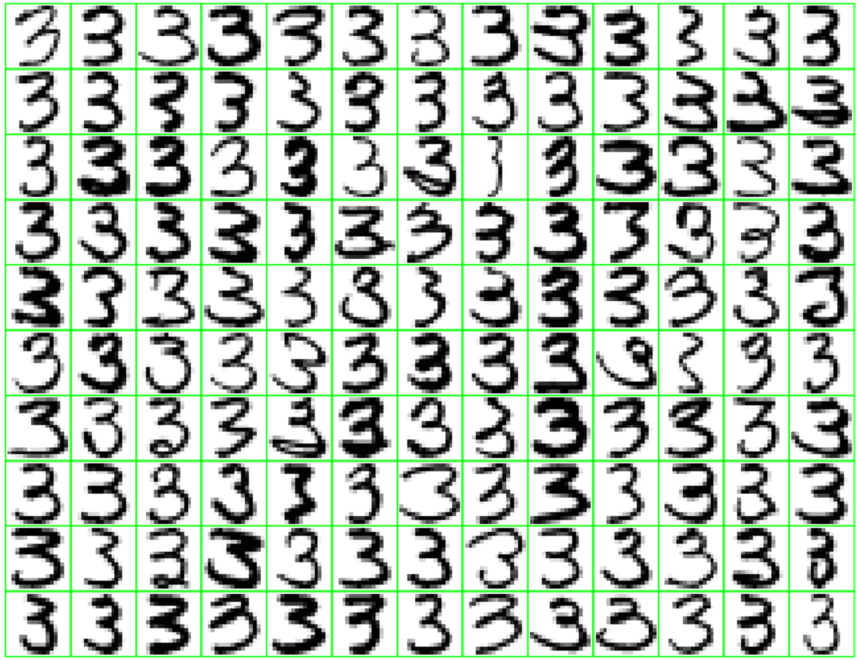
  - $\mathbf{XV} = \mathbf{UD}, \mathbf{x}_i^T = [x_{i1} \quad \cdots \quad x_{ip}], \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}, \mathbf{V} = [v_1 \quad \cdots \quad v_p]$
  - i 번째 관측치의 PC score $= \mathbf{x}_i^T \mathbf{V} = [\mathbf{x}_i^T v_1 \quad \cdots \quad \mathbf{x}_i^T v_p]$
  - I 번째 관측치의 1번째 PC score $= \mathbf{x}_i^T v_\mathbf{1}$
    - 관측치 $a = \mathbf{x}_i^T$와, eigen vector $\mathrm{b} = v_1$의 내적 값.





I PCA 수행 과정

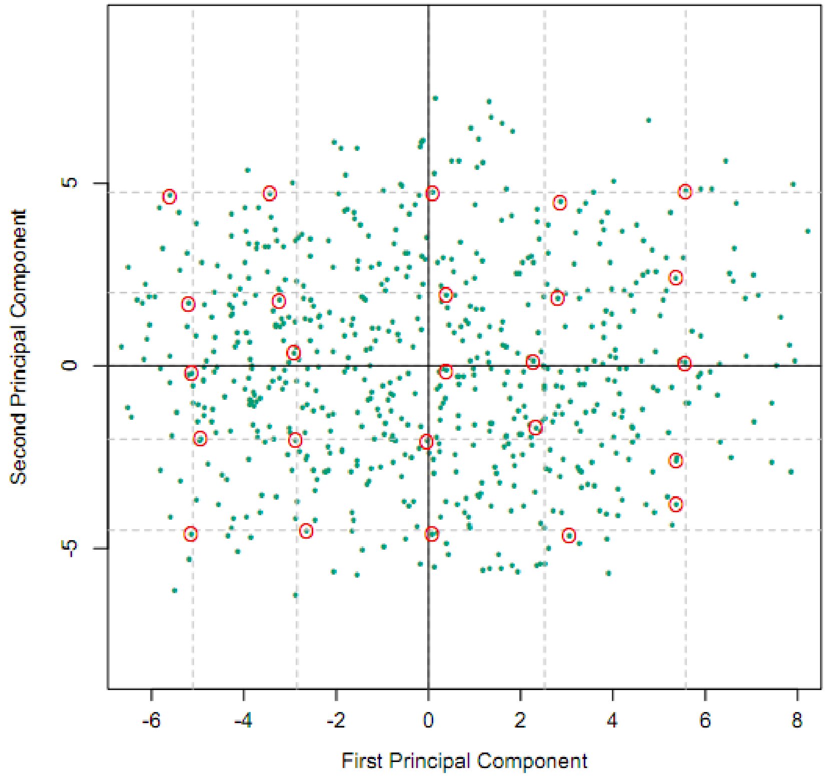
- PCA 활용 – 각 PC의 의미 파악
  - Hand writing of “3”, 130 samples
    (The Elements of Statistical Learning, Tibshirani and Friedman)

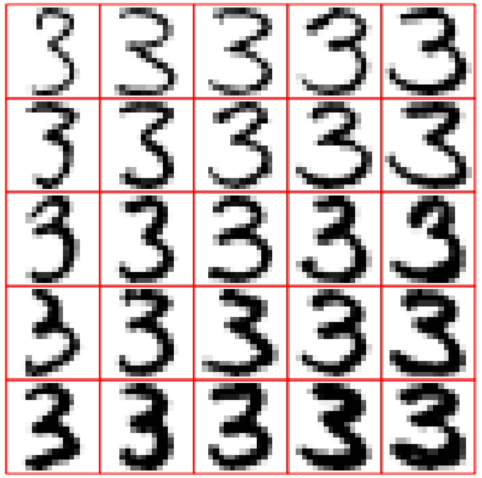


  - (16 x 16) grayscale image, X: (130 x 256) matrix

I PCA 수행 과정

- PCA 활용 – 각 PC의 의미 파악
  - Plot by PC1 and PC2 with quantile grid.





- PC1: 3의 아래쪽 꼬리의 길이, PC2: 글씨의 두께

- End of the clip.

Part.02

회귀분석

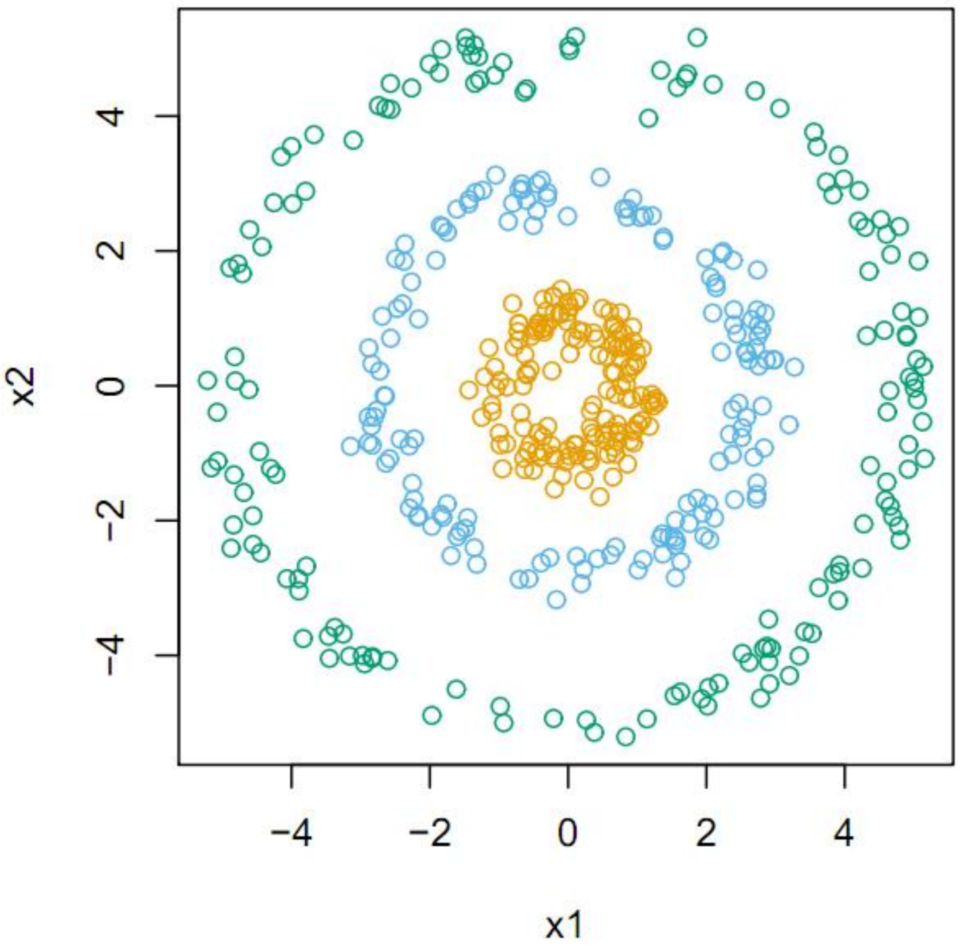
| PCA – Kernel PCA

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 김강진

Kernel PCA

- 다음과 같은 데이터는 어떻게 처리해야 할까?
  - X1, X2 사이의 공분산은 0일 것임. 비선형적 관계.



Kernel PCA

- 관측치 사이의 패턴이 존재하는 것으로 보이나, 변수간의 선형관계가 아닐 때
  - 관측치 사이의 패턴을 수치화하고, 이것의 PC를 구해냄.
    - K (Kernel matrix)는 관측치 사이의 유사도 개념.
      - 비슷한 관측치일수록 큰 값
      - 서로 이질적인 관측치일 수록 작은 값
      - $n \times n$ matrix

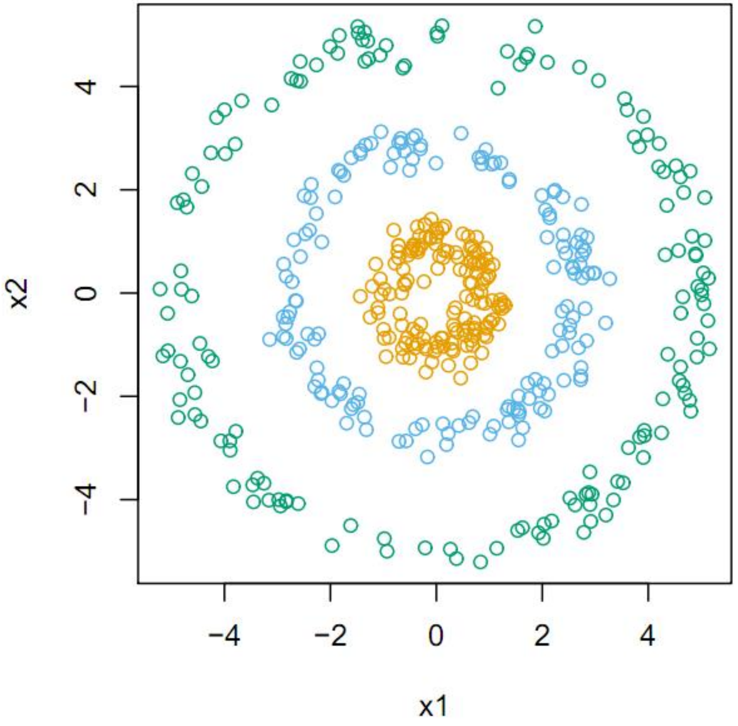
- K의 예시
  - X가 centering 되어 있을 때,
    - $K = XX^T$, 또는 $K(x, x') = \exp(-[x - x']^T[x - x'])$

Kernel PCA

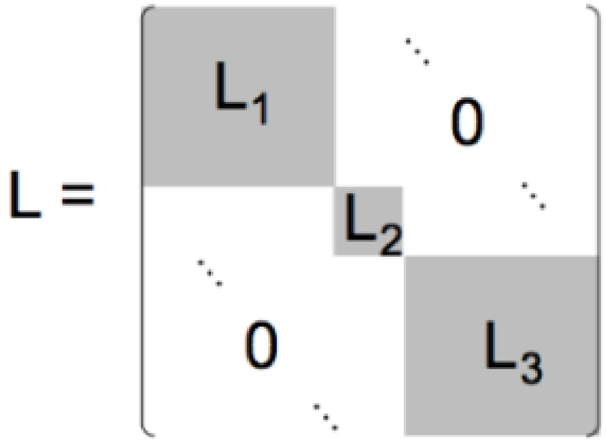
- $K = XX^T$인 경우,
  - $\mathbf{X} = \mathbf{UDV}^T$
  - $\mathbf{K} = \mathbf{UDV}^T\mathbf{VD}^T\mathbf{U}^T = \mathbf{UD}^2\mathbf{U}^T$
  - K의 Eigen vectors = $\mathbf{U}$의 칼럼벡터
  - K의 Eigen values = $\mathbf{D}^2$의 대각성분
  - PC score
    - $\mathbf{UD}$로 구함.
      - $\mathbf{X} = \mathbf{UDV}^T$
      - $\mathbf{XV} = \mathbf{UD}$

I Kernel PCA

- K의 형태는?

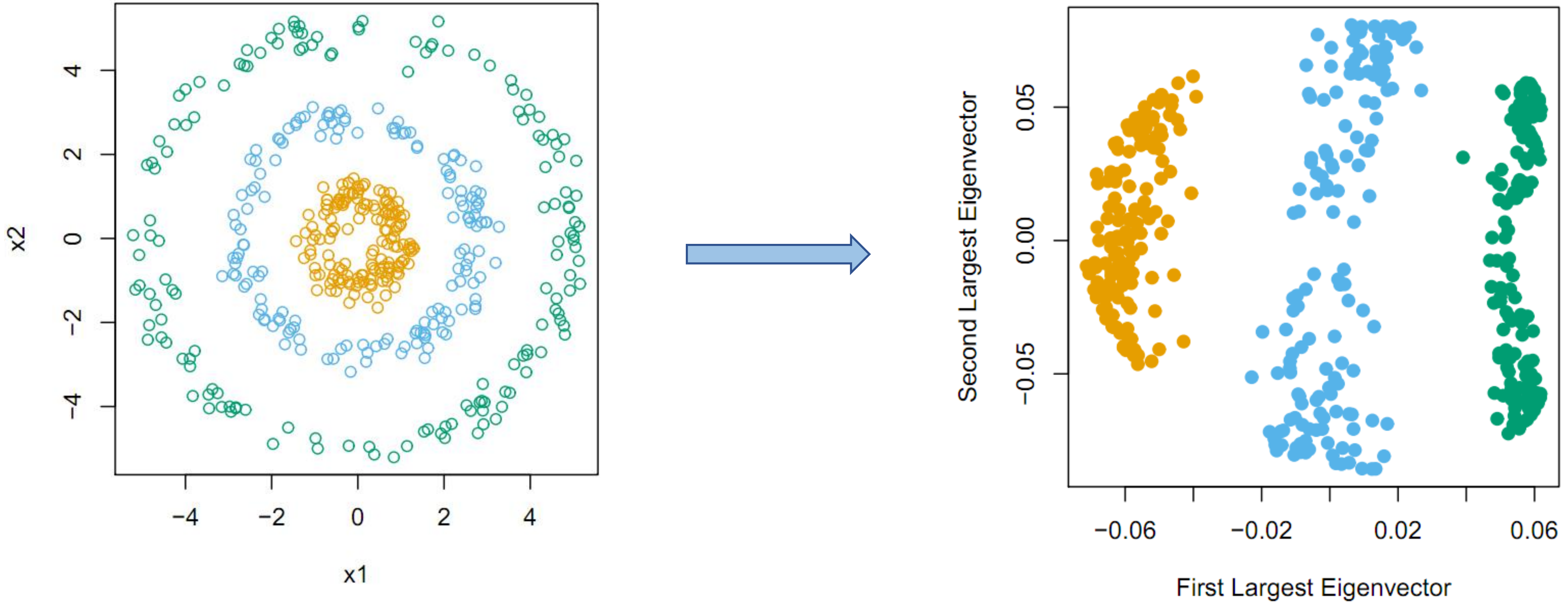


K matrix



I Kernel PCA

- Kernel PCA 결과



- End of the clip.