

R 교육 세미나

ToBig's 7기 박현진

Principal Component Analysis

; PCA
주성분 분석

Contents

Unit 01 | 들어가기 전에

Unit 02 | 차원 축소

Unit 03 | 주성분 분석

Unit 04 | 요약

Unit 01 | 들어가기 전에

Mean , Variance, Covariance, Correlation

	V_1	V_2	...	V_M
1	X(1,1)	X(1,2)		X(1,M)
2	X(2,1)	X(2,2)		X(2,M)
:				
N	X(N,1)	X(N,2)		X(N,M)

$$\text{Mean}(V_1) = \frac{(X_{1,1} + X_{2,1} + \dots + X_{N,1})}{N}$$

$$\text{Var}(V_1) = \frac{\sum (V_1 - \bar{V}_1)^2}{N-1}$$

$$\text{Cov}(V_1, V_2) = E(V_1 V_2) - E(V_1)E(V_2)$$

$$\text{Corr}(V_1, V_2) = \frac{\text{Cov}(V_1, V_2)}{\sqrt{\sum (V_1 - \bar{V}_1)^2} \sqrt{\sum (V_2 - \bar{V}_2)^2}}$$

Unit 01 | 들어가기 전에

Covariance matrix, Correlation matrix

	1	2	...	M
1	Var(V1)			
2	Cov(V1,V2)	Var(V2)		
:	:	:		
M	Cov(V1,VM)	Cov(V2,VM)		Var(VM)

	1	2	...	M
1	1			
2	Corr(V1,V2)	1		
:	:	:		
M	Corr(V1,VM)	Corr(V2,VM)		1

Unit 02 | 차원축소

Problem

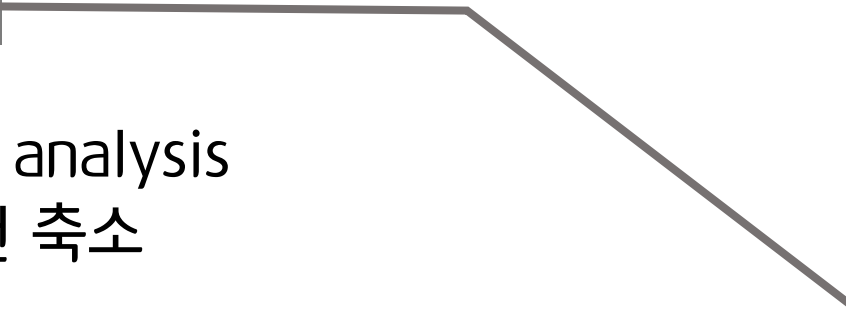
변수(설명변수)의 수 多

- ➡ 변수 간의 correlation
- ➡ 관심있는 대상(종속변수)에 무관한 변수 포함
- ➡ 계산량, data 저장공간, 비용
- ➡ overfitting

Unit 02 | 차원축소

방법 1

Correlation analysis
를 통한 차원 축소



	1	2	...	M
1	1			
2	Corr(V1,V2)	1		
:	:	:		
M	Corr(V1,VM)	Corr(V2,VM)	...	1

방법 2

categorical variables
에서의 차원 축소

- 상관관계가 높은 변수 둘 중 하나 제거
- 다른 DB에서 data 를 구했을 경우 변수의 중복을 막을 수 있음

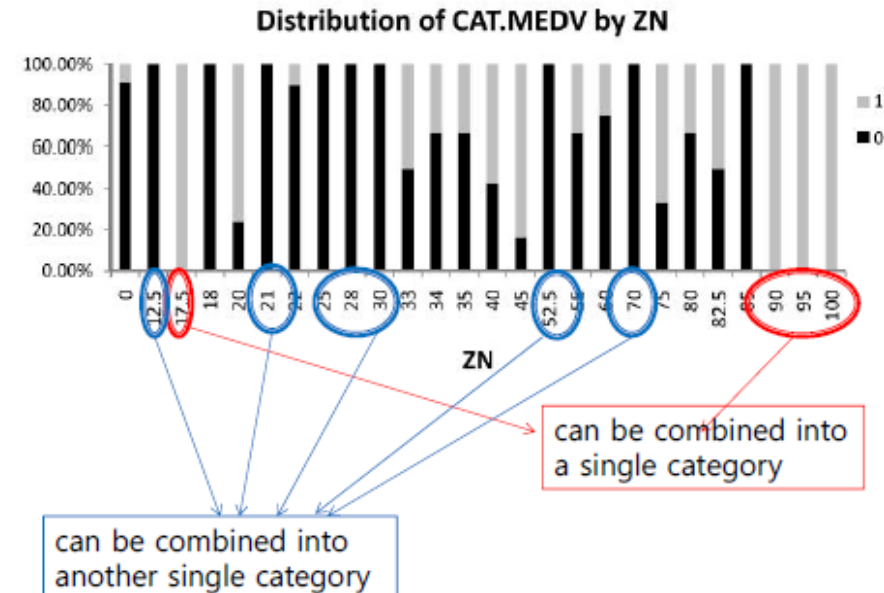
Unit 02 | 차원축소

방법 1

Correlation analysis
를 통한 차원 축소

방법 2

categorical variables
에서의 차원 축소



- 관측치가 적은 변수를 다른 것과 묶기, 제거
- 관련있는 변수 끼리 묶어주기
- 의미있는 변수만 사용하고 나머지는 “기타”

Unit 03 | 주성분 분석

Problem

변수 제거

- ➔ 정보 손실 多
- ➔ 변수간 correlation을 완전히 제거하지 못함
- ➔ 분석자의 주관적 판단

Unit 03 | 주성분 분석

solution

Principal Component analysis ;PCA

Unit 03 | 주성분 분석

idea

변수들 간의 **information**이 중첩 되어 있는 부분을 없애자

Goal

가장 많은 **information**을 포함하고 있는 적은 변수 생성

Unit 03 | 주성분 분석

PCA

1. 정량적 데이터(quantative data) only
2. 기존 Data 의 선형 결합인 새로운 변수 생성
3. 새로운 변수는 uncorrelated (information 중복 없음)
4. 새로 생성된 변수 : principal component

$$\begin{aligned}\vec{z}_1 &= \alpha_{11}\vec{x}_1 + \alpha_{12}\vec{x}_2 + \dots + \alpha_{1p}\vec{x}_p = \vec{\alpha}_1^T X \\ \vec{z}_2 &= \alpha_{21}\vec{x}_1 + \alpha_{22}\vec{x}_2 + \dots + \alpha_{2p}\vec{x}_p = \vec{\alpha}_2^T X \\ &\dots \\ \vec{z}_p &= \alpha_{p1}\vec{x}_1 + \alpha_{p2}\vec{x}_2 + \dots + \alpha_{pp}\vec{x}_p = \vec{\alpha}_p^T X\end{aligned}$$

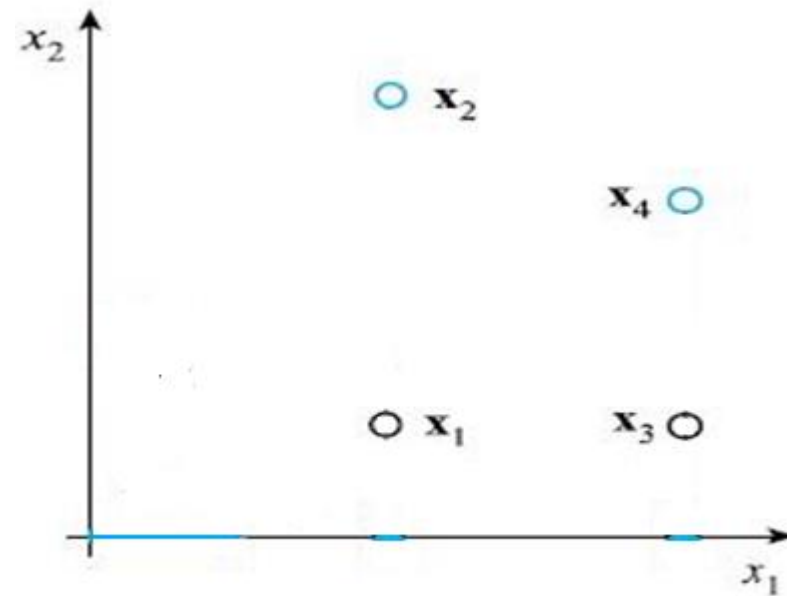
Unit 03 | 주성분 분석

Information loss

Information?

해당 feature(축)으로
data를 투영시켰을 때(차원 축소)
data의 흩어짐이 유지되는 정도

즉, information은 해당 축의 분산



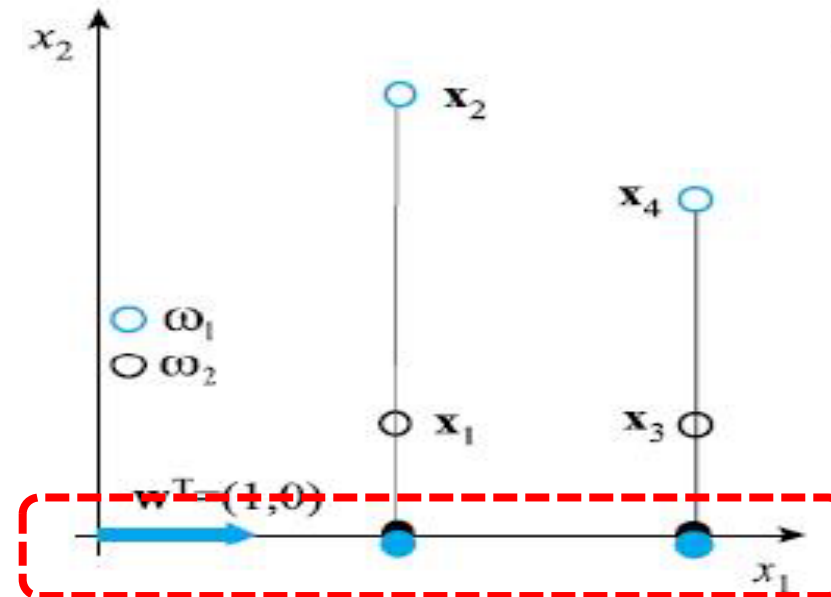
Unit 03 | 주성분 분석

Information loss

Information?

해당 feature(축)으로
data를 투영시켰을 때(차원 축소)
data의 흩어짐이 유지되는 정도

즉, information은 해당 축의 분산



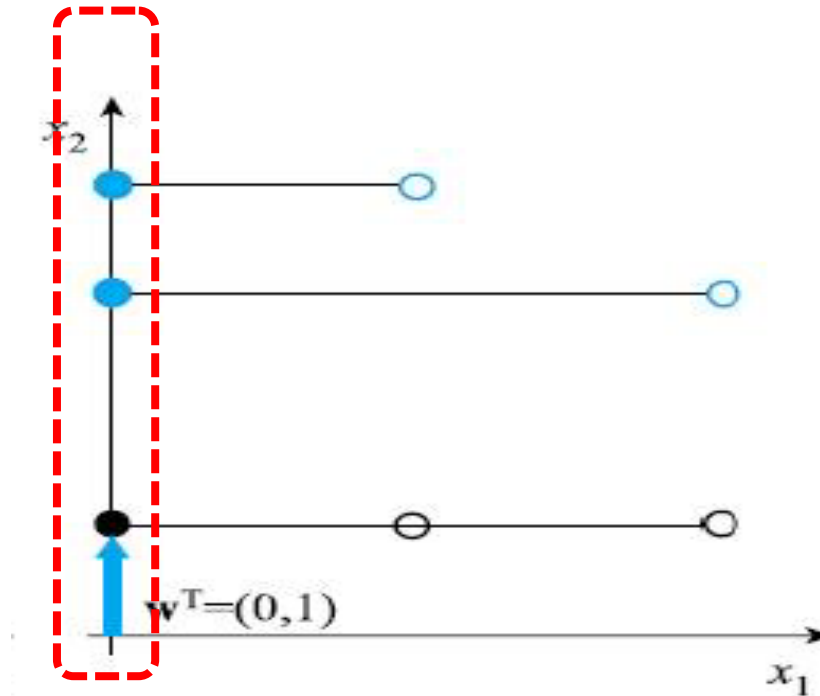
Unit 03 | 주성분 분석

Information loss

Information?

해당 feature(축)으로
data를 투영시켰을 때(차원 축소)
data의 흩어짐이 유지되는 정도

즉, information은 해당 축의 분산



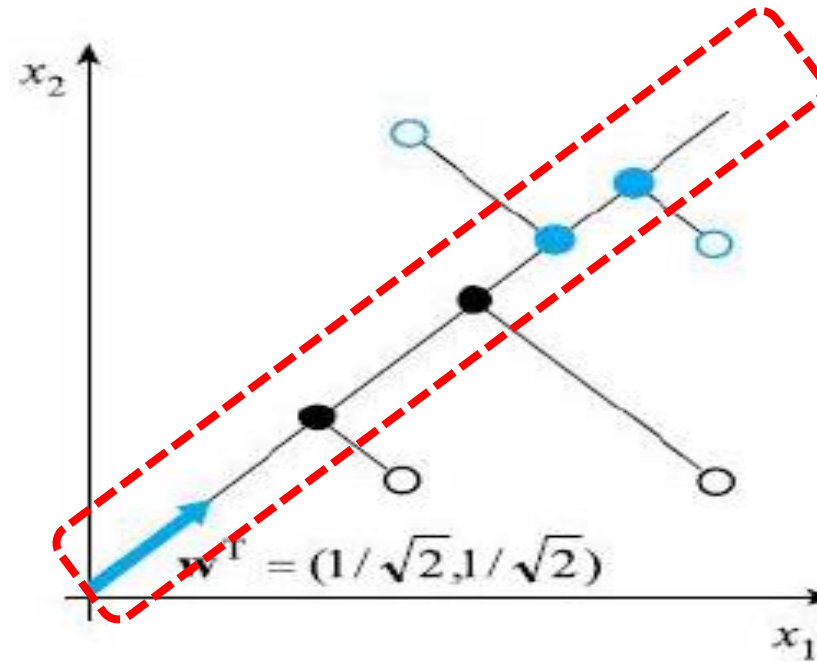
Unit 03 | 주성분 분석

Information loss

Information?

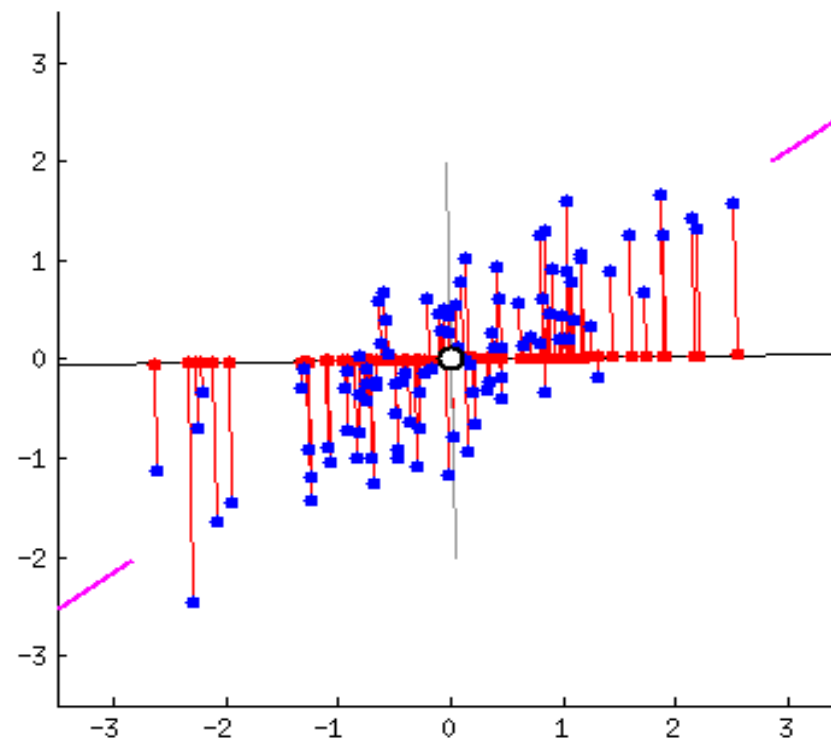
해당 feature(축)으로
data를 투영시켰을 때(차원 축소)
data의 흩어짐이 유지되는 정도

즉, information은 해당 축의 분산



Unit 03 | 주성분 분석

information



Unit 03 | 주성분 분석

예제

Name	칼로리	지방	단백질	...	비타민
콘푸라이트	70	5	4		
чекс	120	7	3		
아몬드 푸레이크	80	2	4		
:	:	:	:		
:	:	:	:		

Description of Variables

Name: name of cereal**mfr:** manufacturer**type:** cold or hot**calories:** calories per serving**protein:** grams**fat:** grams**sodium:** mg.**fiber:** grams**carbo:** grams complex carbohydrates**sugars:** grams**potass:** mg.**vitamins:** % FDA rec**shelf:** display shelf**weight:** oz. 1 serving**cups:** in one serving

Unit 03 | 주성분 분석

예제

Name	칼로리	지방
콘푸라이트	70	5
체크스	120	7
아몬드 푸레이크	80	2
:	:	:
:	:	:

 $\text{VAR}(\text{칼로리})=400$ $\text{VAR}(\text{지방})=200$ $\text{Cov}(\text{칼로리}, \text{지방})=180$ $\text{Corr}(\text{칼로리}, \text{지방})=0.7$

- 강한 양의 상관관계
- 70%의 변동성(information)이 두 변수에서 중첩 된다

 $\text{총분산} = \text{VAR}(\text{칼로리}) + \text{VAR}(\text{지방}) = 600$

칼로리(변수1)의 정보:

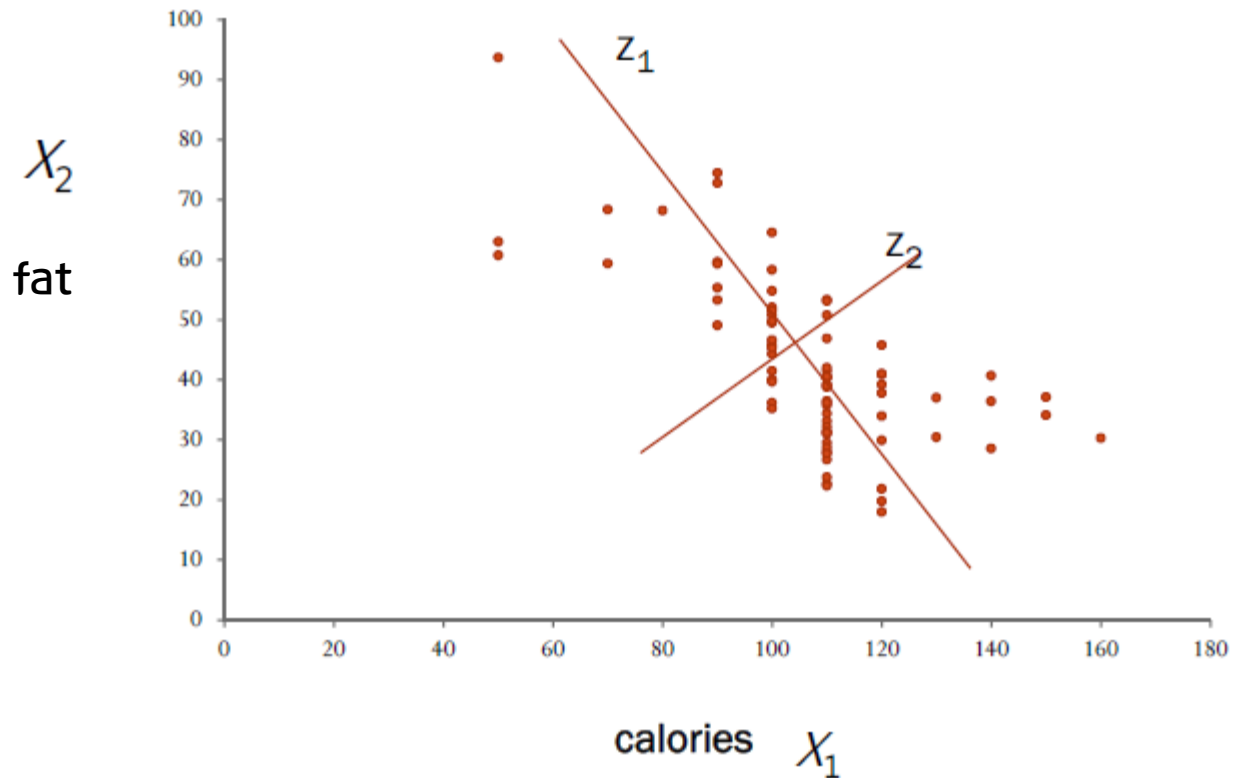
$$\begin{aligned} &= \text{VAR}(\text{칼로리}) / \{\text{VAR}(\text{칼로리}) + \text{VAR}(\text{지방})\} \\ &= 400 / 600 = 0.66 \end{aligned}$$

지방(변수2)의 정보:

$$\begin{aligned} &= \text{VAR}(\text{지방}) / \{\text{VAR}(\text{칼로리}) + \text{VAR}(\text{지방})\} \\ &= 200 / 600 = 0.34 \end{aligned}$$

Unit 03 | 주성분 분석

First & Second Principal Components

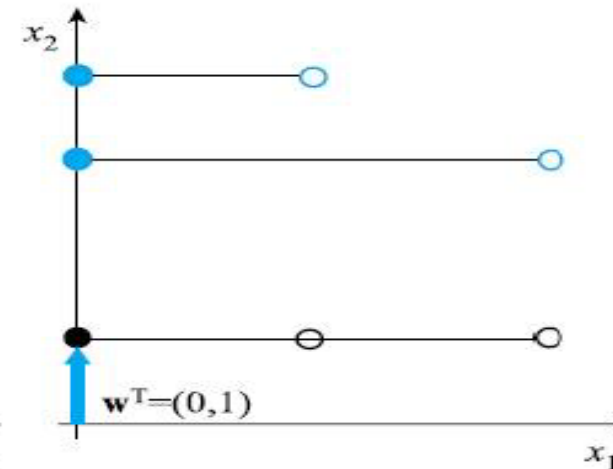


$$\text{총분산} = \text{VAR}(\text{칼로리}) + \text{VAR}(\text{지방}) = 600$$

$$\begin{aligned} \text{칼로리(변수1)의 정보} &: \text{VAR}(\text{칼로리}) / \{\text{VAR}(\text{칼로리}) + \text{VAR}(\text{지방})\} \\ &= 400 / 600 = 0.66 \end{aligned}$$

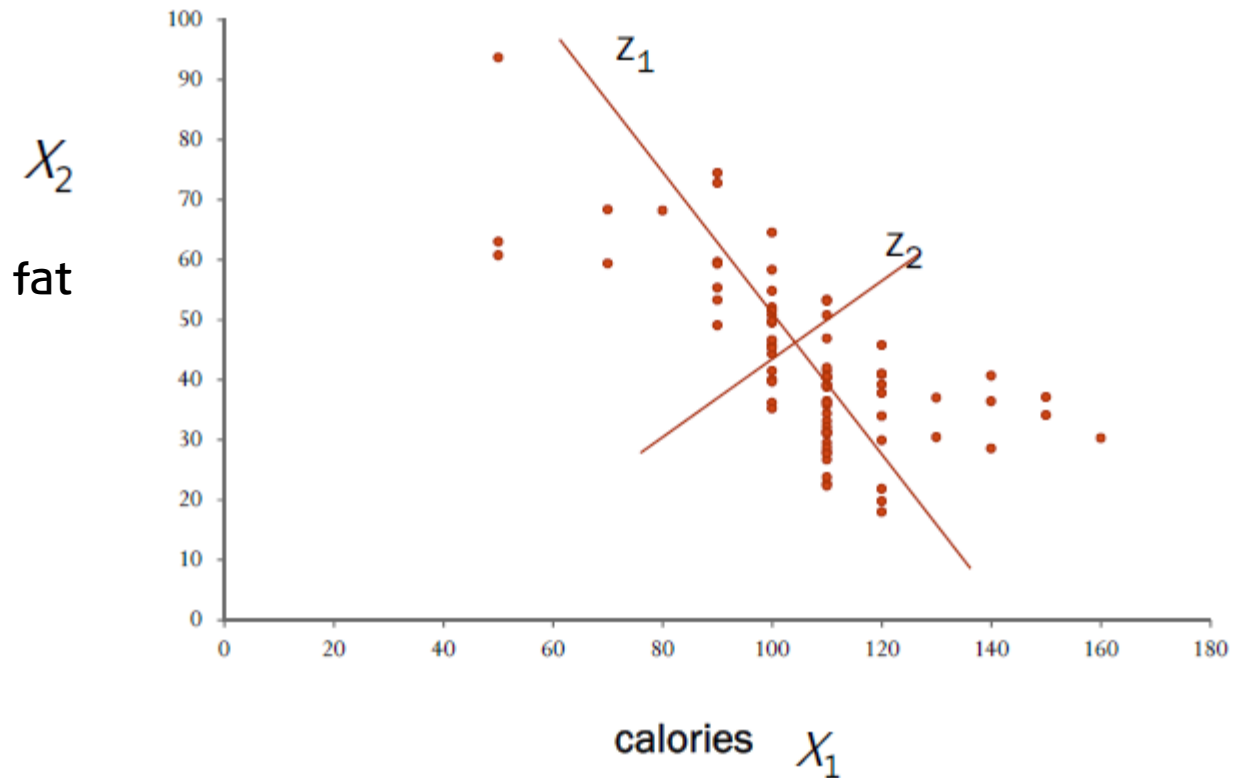
$$\begin{aligned} \text{지방(변수2)의 정보} &: \text{VAR}(\text{지방}) / \{\text{VAR}(\text{칼로리}) + \text{VAR}(\text{지방})\} \\ &= 200 / 600 = 0.34 \end{aligned}$$

정보 손실



Unit 03 | 주성분 분석

First & Second Principal Components

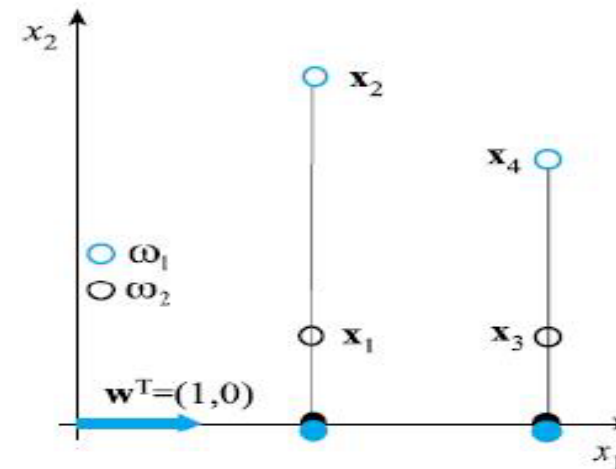


$$\text{총분산} = \text{VAR}(\text{칼로리}) + \text{VAR}(\text{지방}) = 600$$

$$\begin{aligned} \text{칼로리(변수1)의 정보} &: \text{VAR}(\text{칼로리}) / \{\text{VAR}(\text{칼로리}) + \text{VAR}(\text{지방})\} \\ &= 400 / 600 = 0.66 \end{aligned}$$

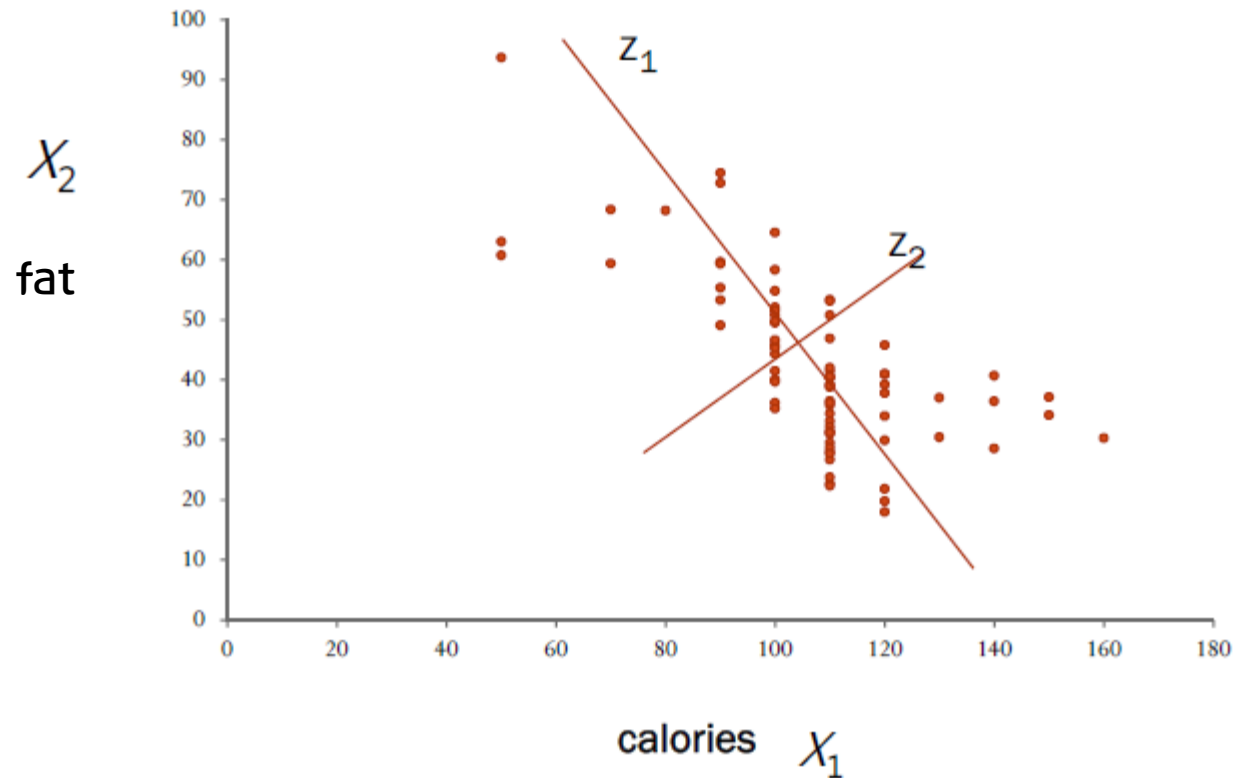
$$\begin{aligned} \text{지방(변수2)의 정보} &: \text{VAR}(\text{지방}) / \{\text{VAR}(\text{칼로리}) + \text{VAR}(\text{지방})\} \\ &= 200 / 600 = 0.34 \end{aligned}$$

정보 손실



Unit 03 | 주성분 분석

First & Second Principal Components

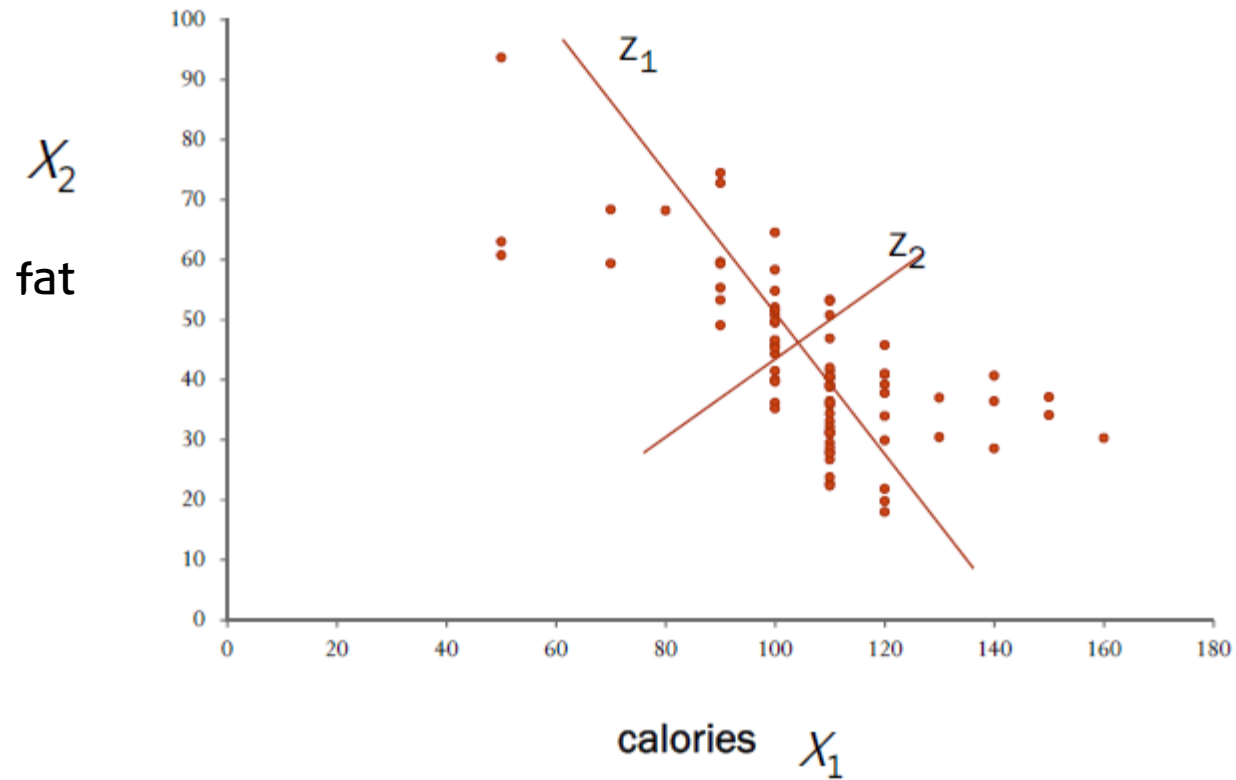


새로운 변수(축) Z_1 , Z_2 를 만들자

1. 새로운 변수 Z 중 하나를 제거하는 것은 최소한의 정보(variance)를 잃는 것
2. Z_1 과 Z_2 는 uncorrelated

Unit 03 | 주성분 분석

First & Second Principal Components

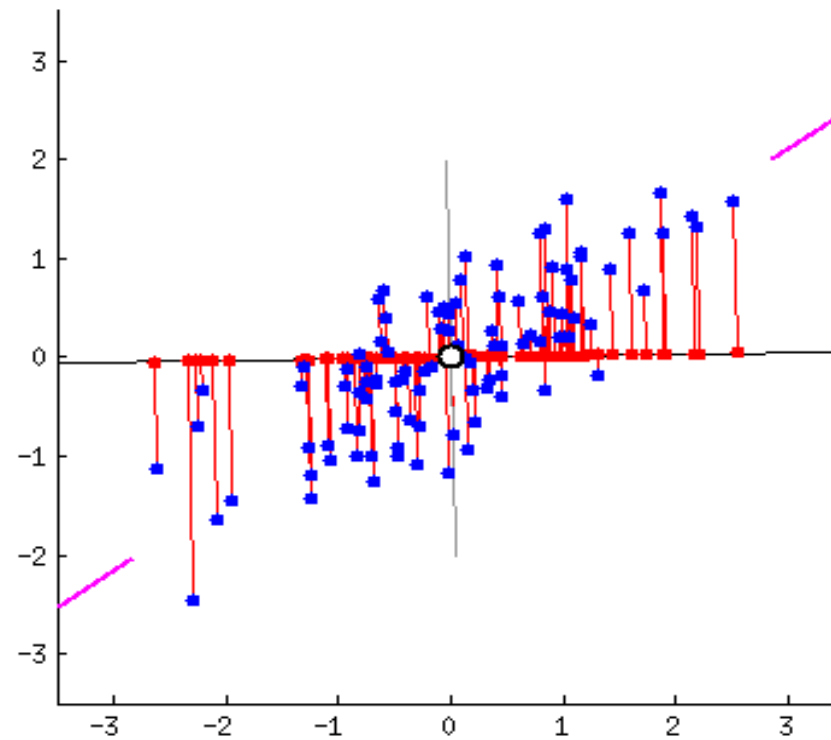


새로운 변수(축) Z_1 , Z_2 를 만들자

1. $\text{VAR}(Z_1)$ 을 max로 만드는 Z_1
2. $\text{Corr}(Z_1, Z_2) = 0$

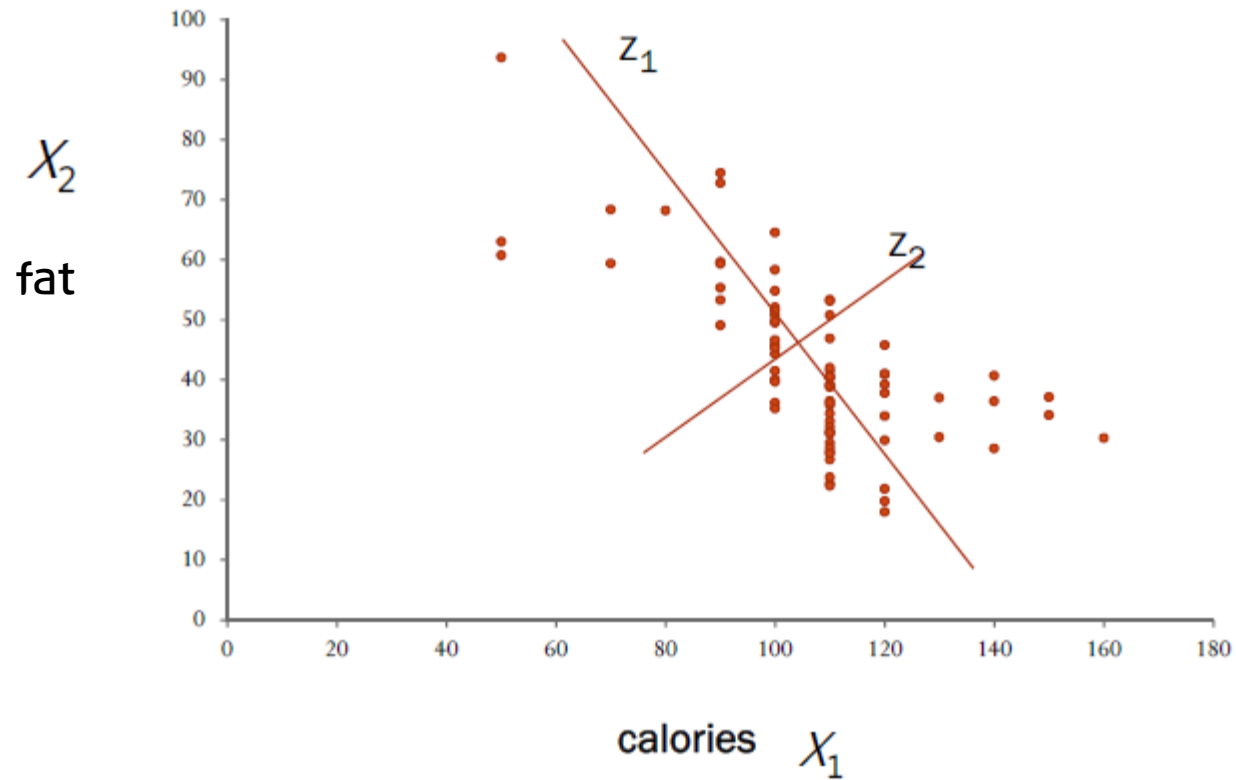
Unit 03 | 주성분 분석

information



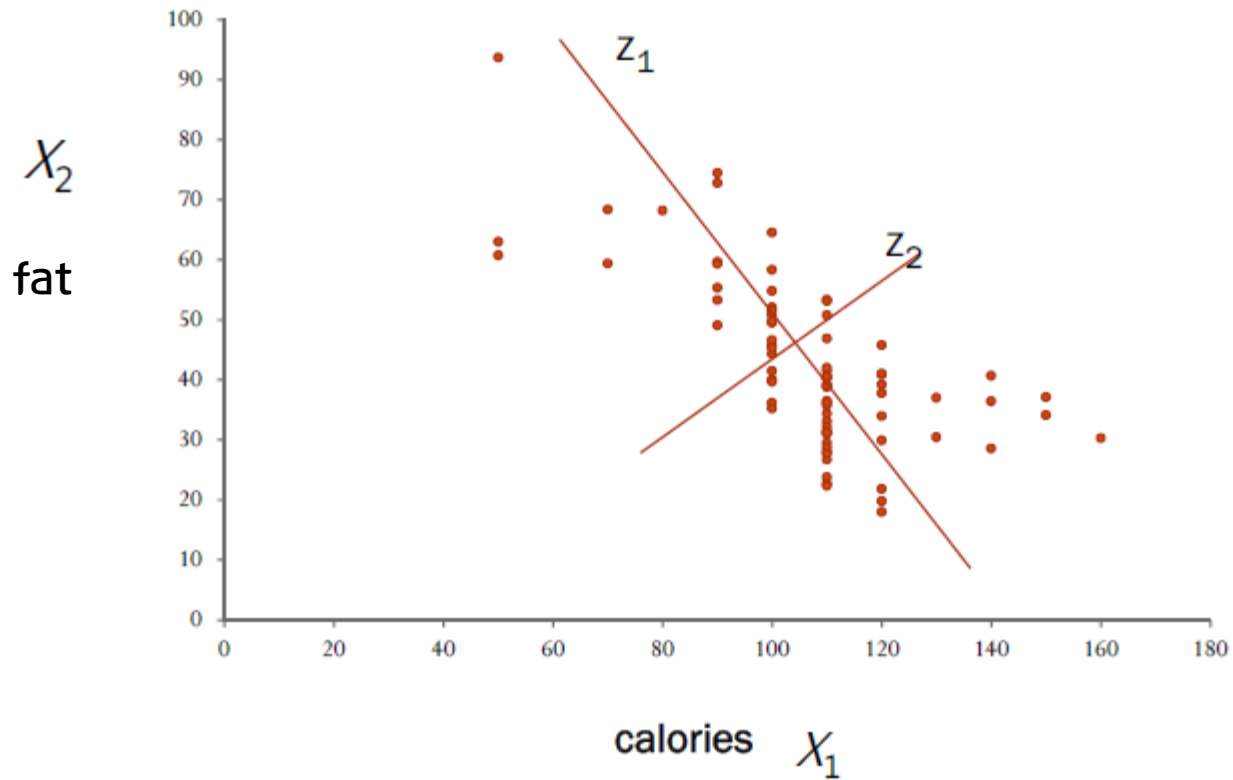
Unit 03 | 주성분 분석

First & Second Principal Components

 $\text{Var}(Z_1) \geq \text{Var}(\text{칼로리}), \text{Var}(\text{지방})$ $\text{Var}(Z_1) = 500$ $\text{Var}(Z_2) = 100$

Unit 03 | 주성분 분석

First & Second Principal Components

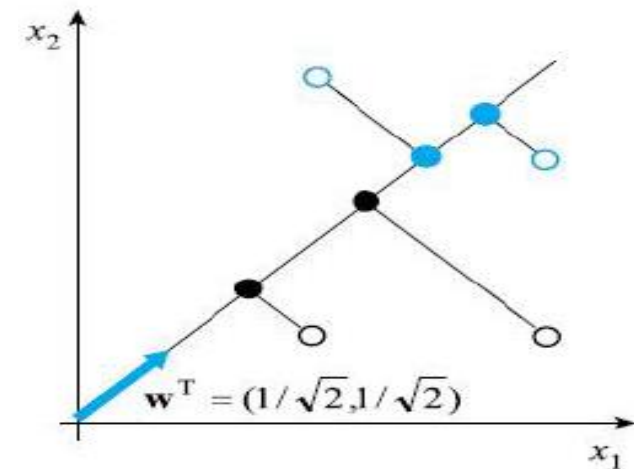


$$\text{총분산} = \text{VAR}(Z_1) + \text{VAR}(Z_2) = 600$$

$$\begin{aligned} \text{Z1의 정보} &= \text{VAR}(Z_1) / \{\text{VAR}(Z_1) + \text{VAR}(Z_2)\} \\ &= 500 / 600 = 0.83 \end{aligned}$$

$$\begin{aligned} \text{Z2의 정보} &= \text{VAR}(Z_2) / \{\text{VAR}(Z_1) + \text{VAR}(Z_2)\} \\ &= 100 / 600 = 0.17 \end{aligned}$$

정보 손실



Unit 03 | 주성분 분석

결과 비교

	Variance(%)	correlation
원래 변수(X1,X2)	66% : 34%	0.7
주성분 분석을 통해 새로 생성한 변수(Z1,Z2)	83% : 17%	0

➡ 정보 손실 : 34% → 17%

➡ Correlation: 70% → 0%

Unit 03 | 주성분 분석

주성분 찾기

$$\begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} \quad \begin{matrix} \nearrow \\ \nearrow \end{matrix} \quad \boxed{A X = \lambda X} \quad \begin{matrix} \nwarrow \\ \nwarrow \end{matrix}$$

$n \times n$ 행렬

Eigenvalue

Eigenvector
0은 반드시 아님!

Unit 03 | 주성분 분석

주성분 찾기

$$A x = \lambda x$$

$$[\lambda I - A] * x = 0$$

$$\det(\lambda I - A) = 0 \quad \leftarrow \text{특성방정식}$$

Unit 03 | 주성분 분석

주성분 찾기

$$\begin{aligned}\vec{z}_1 &= \alpha_{11}\vec{x}_1 + \alpha_{12}\vec{x}_2 + \dots + \alpha_{1p}\vec{x}_p = \vec{\alpha}_1^T X \\ \vec{z}_2 &= \alpha_{21}\vec{x}_1 + \alpha_{22}\vec{x}_2 + \dots + \alpha_{2p}\vec{x}_p = \vec{\alpha}_2^T X \\ &\dots \\ \vec{z}_p &= \alpha_{p1}\vec{x}_1 + \alpha_{p2}\vec{x}_2 + \dots + \alpha_{pp}\vec{x}_p = \vec{\alpha}_p^T X\end{aligned}$$

$$Z = \begin{bmatrix} \vec{z}_1 \\ \vec{z}_2 \\ \dots \\ \vec{z}_p \end{bmatrix} = \begin{bmatrix} \vec{\alpha}_1^T X \\ \vec{\alpha}_2^T X \\ \dots \\ \vec{\alpha}_p^T X \end{bmatrix} = \begin{bmatrix} \vec{\alpha}_1^T \\ \vec{\alpha}_2^T \\ \dots \\ \vec{\alpha}_p^T \end{bmatrix} X = \vec{A}^T X$$

Unit 03 | 주성분 분석

주성분 찾기

$$\begin{aligned}\max_{\alpha} \{Var(Z)\} &= \max_{\alpha} \{Var(\vec{\alpha}^T X)\} \\ &= \max_{\alpha} \{ \vec{\alpha}^T Var(X) \vec{\alpha} \} \\ &= \max_{\alpha} \{ \vec{\alpha}^T \Sigma \vec{\alpha} \}\end{aligned}$$

$$\|\alpha\| = \vec{\alpha}^T \vec{\alpha} = 1$$

$$\begin{aligned}\vec{z}_1 &= \alpha_{11}\vec{x}_1 + \alpha_{12}\vec{x}_2 + \dots + \alpha_{1p}\vec{x}_p = \underline{\vec{\alpha}_1}^T X \\ \vec{z}_2 &= \alpha_{21}\vec{x}_1 + \alpha_{22}\vec{x}_2 + \dots + \alpha_{2p}\vec{x}_p = \underline{\vec{\alpha}_2}^T X \\ &\dots \\ \vec{z}_p &= \alpha_{p1}\vec{x}_1 + \alpha_{p2}\vec{x}_2 + \dots + \alpha_{pp}\vec{x}_p = \underline{\vec{\alpha}_p}^T X\end{aligned}$$

Unit 03 | 주성분 분석

주성분 찾기

$$L = \vec{\alpha}^T \Sigma \vec{\alpha} - \lambda(\vec{\alpha}^T \vec{\alpha} - 1)$$

$$\frac{\partial L}{\partial \vec{\alpha}} = \Sigma \vec{\alpha} - \lambda \vec{\alpha} = 0$$

$$(\Sigma - \lambda) \vec{\alpha} = 0$$

$$\Rightarrow \Sigma \vec{\alpha} = \lambda \vec{\alpha}$$

$$\Rightarrow \vec{\alpha} = \text{eigenvector}(\text{Cov}(X))$$

$$\begin{aligned} \vec{z}_1 &= \alpha_{11}\vec{x}_1 + \alpha_{12}\vec{x}_2 + \dots + \alpha_{1p}\vec{x}_p = \underline{\underline{\vec{\alpha}_1}}^T X \\ \vec{z}_2 &= \alpha_{21}\vec{x}_1 + \alpha_{22}\vec{x}_2 + \dots + \alpha_{2p}\vec{x}_p = \underline{\underline{\vec{\alpha}_2}}^T X \\ &\dots \\ \vec{z}_p &= \alpha_{p1}\vec{x}_1 + \alpha_{p2}\vec{x}_2 + \dots + \alpha_{pp}\vec{x}_p = \underline{\underline{\vec{\alpha}_p}}^T X \end{aligned}$$

Unit 03 | 주성분 분석

주성분 찾기

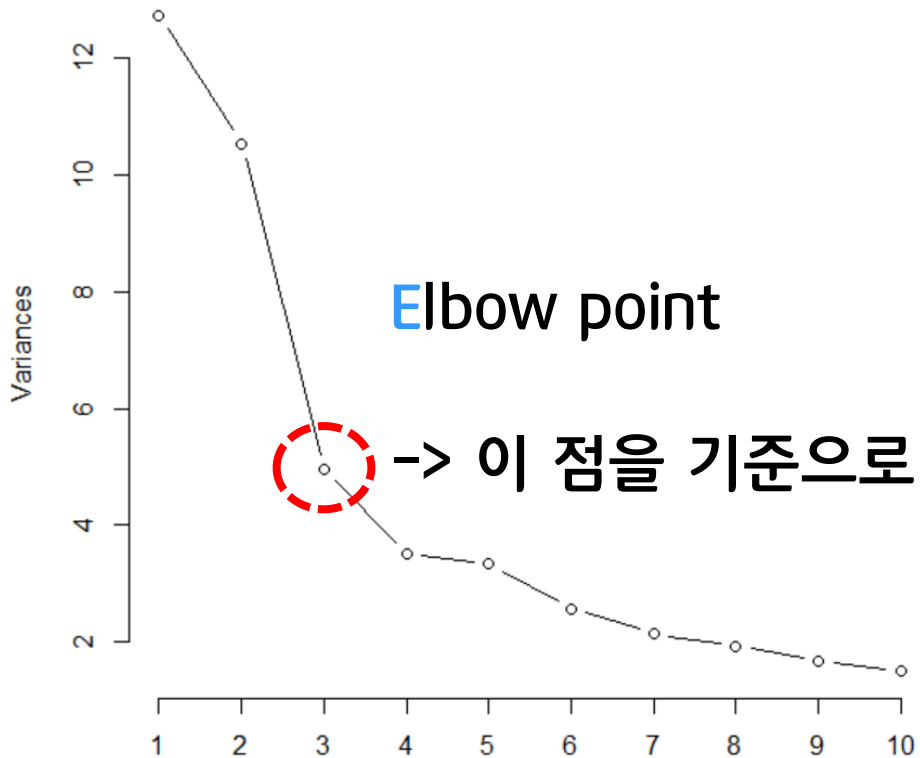
$$\begin{aligned}\vec{z}_1 &= \alpha_{11}\vec{x}_1 + \alpha_{12}\vec{x}_2 + \dots + \alpha_{1p}\vec{x}_p = \vec{\alpha}_1^T X \\ \vec{z}_2 &= \alpha_{21}\vec{x}_1 + \alpha_{22}\vec{x}_2 + \dots + \alpha_{2p}\vec{x}_p = \vec{\alpha}_2^T X \\ &\dots \\ \vec{z}_p &= \alpha_{p1}\vec{x}_1 + \alpha_{p2}\vec{x}_2 + \dots + \alpha_{pp}\vec{x}_p = \vec{\alpha}_p^T X\end{aligned}$$

$$Z = \begin{bmatrix} \vec{z}_1 \\ \vec{z}_2 \\ \dots \\ \vec{z}_p \end{bmatrix} = \begin{bmatrix} \vec{\alpha}_1^T X \\ \vec{\alpha}_2^T X \\ \dots \\ \vec{\alpha}_p^T X \end{bmatrix} = \begin{bmatrix} \vec{\alpha}_1^T \\ \vec{\alpha}_2^T \\ \dots \\ \vec{\alpha}_p^T \end{bmatrix} X = \underline{A^T} X$$

$A = \text{eigenvector}(\text{Cov}(X))$ 의 나열

Unit 03 | 주성분 분석

변수의 수 줄이기



$$\begin{aligned}\vec{z}_1 &= \alpha_{11}\vec{x}_1 + \alpha_{12}\vec{x}_2 + \dots + \alpha_{1p}\vec{x}_p = \underline{\underline{\vec{\alpha}_1^T}} X \\ \vec{z}_2 &= \alpha_{21}\vec{x}_1 + \alpha_{22}\vec{x}_2 + \dots + \alpha_{2p}\vec{x}_p = \underline{\underline{\vec{\alpha}_2^T}} X \\ &\dots \\ \vec{z}_p &= \alpha_{p1}\vec{x}_1 + \alpha_{p2}\vec{x}_2 + \dots + \alpha_{pp}\vec{x}_p = \underline{\underline{\vec{\alpha}_p^T}} X\end{aligned}$$

Unit 04 | 요약

PCA 요약

1. 기존 data matrix ($n \times m$)의 centered data matrix 생성
2. 위의 centered data matrix 의 covariance matrix 생성
3. 위의 covariance matrix에서 m 개(변수의 수)의 eigen value 와 eigen vector 계산, 나열
4. 정렬된 고유 벡터 가운데 일부 선택(elbow point)
5. 해당 고유 벡터와 기존 data matrix 내적

Q & A

들어주셔서 감사합니다.