PCA

Taejjing

PCA 란..?

Principle 주된

Component 성분을

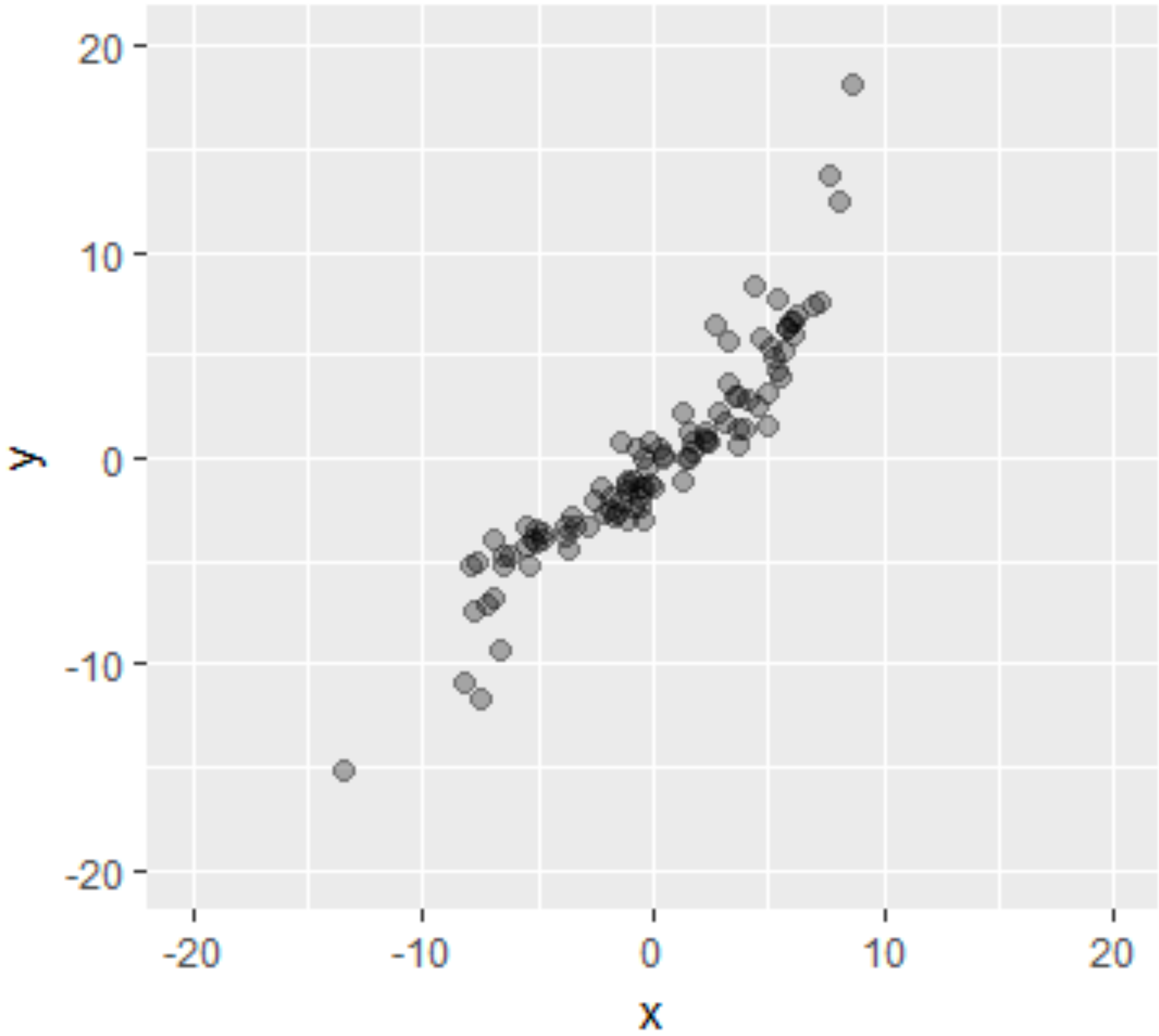
Analysis 분석하자

Taejjing

Thinking..

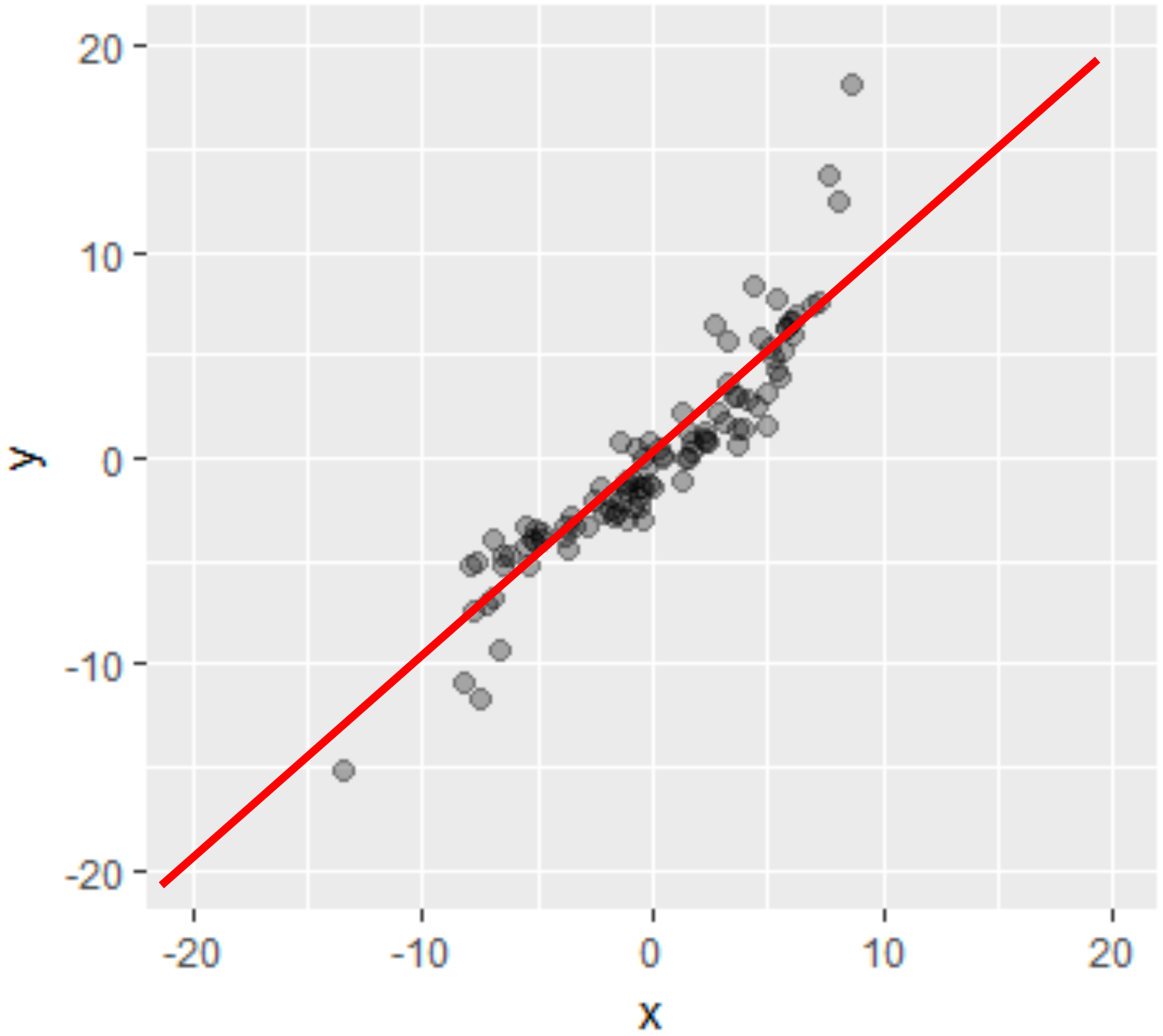
Taejjing

무엇을 하고 싶으신가요?



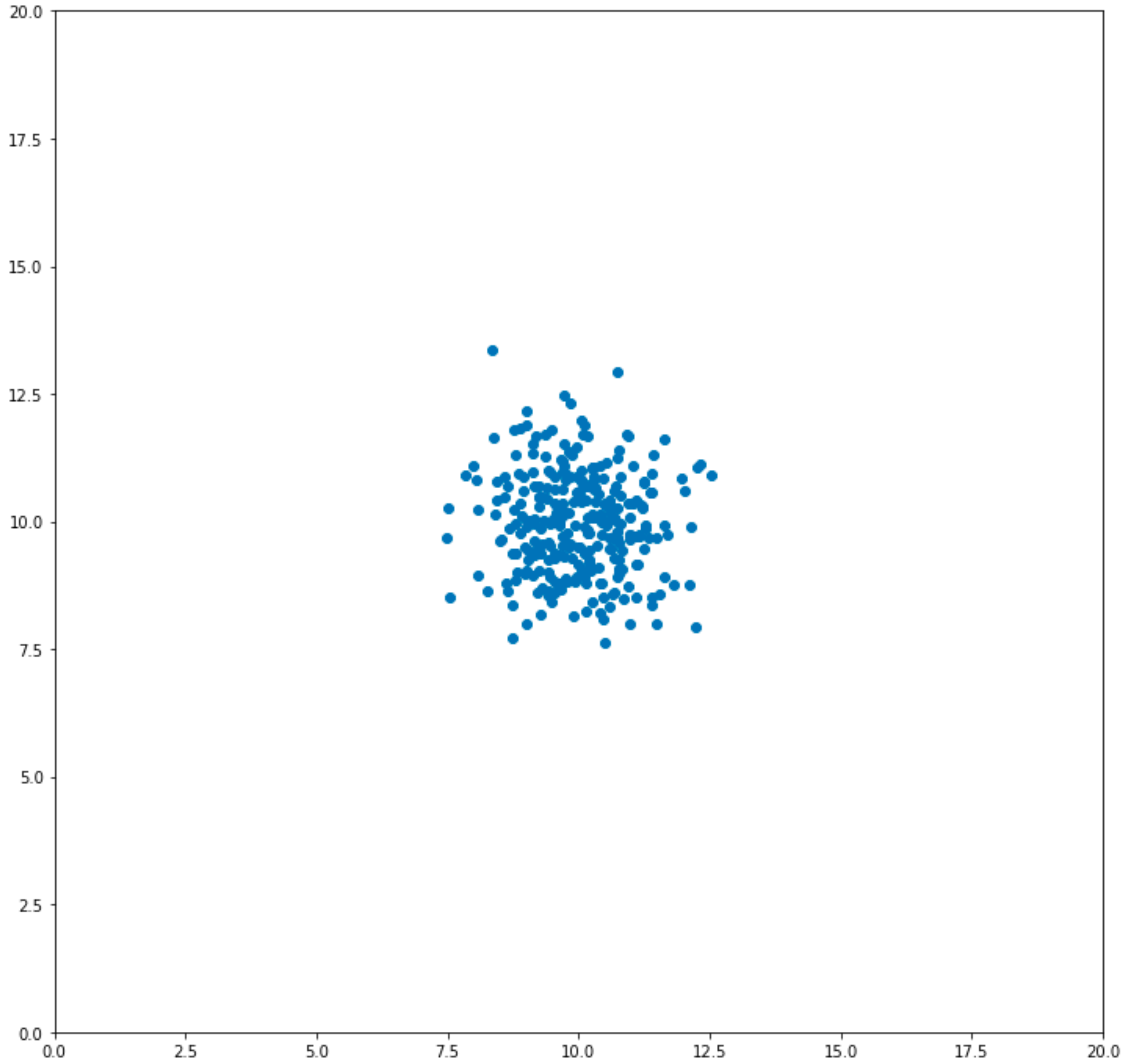
Taejjing

왜 그렇게 생각했나요?

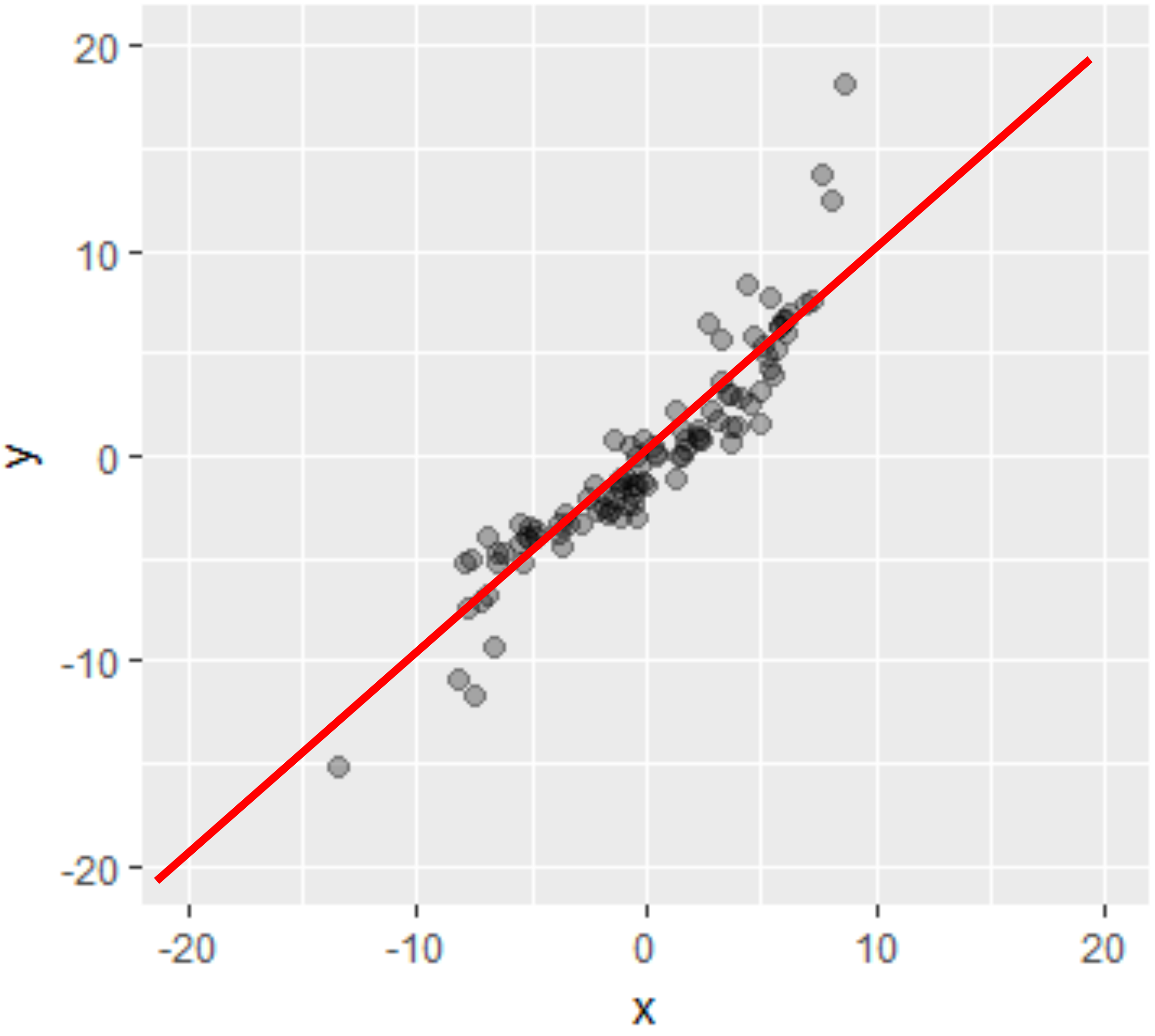


Taejjing

그럼, 이런 경우는 어떨까?



Taejjing



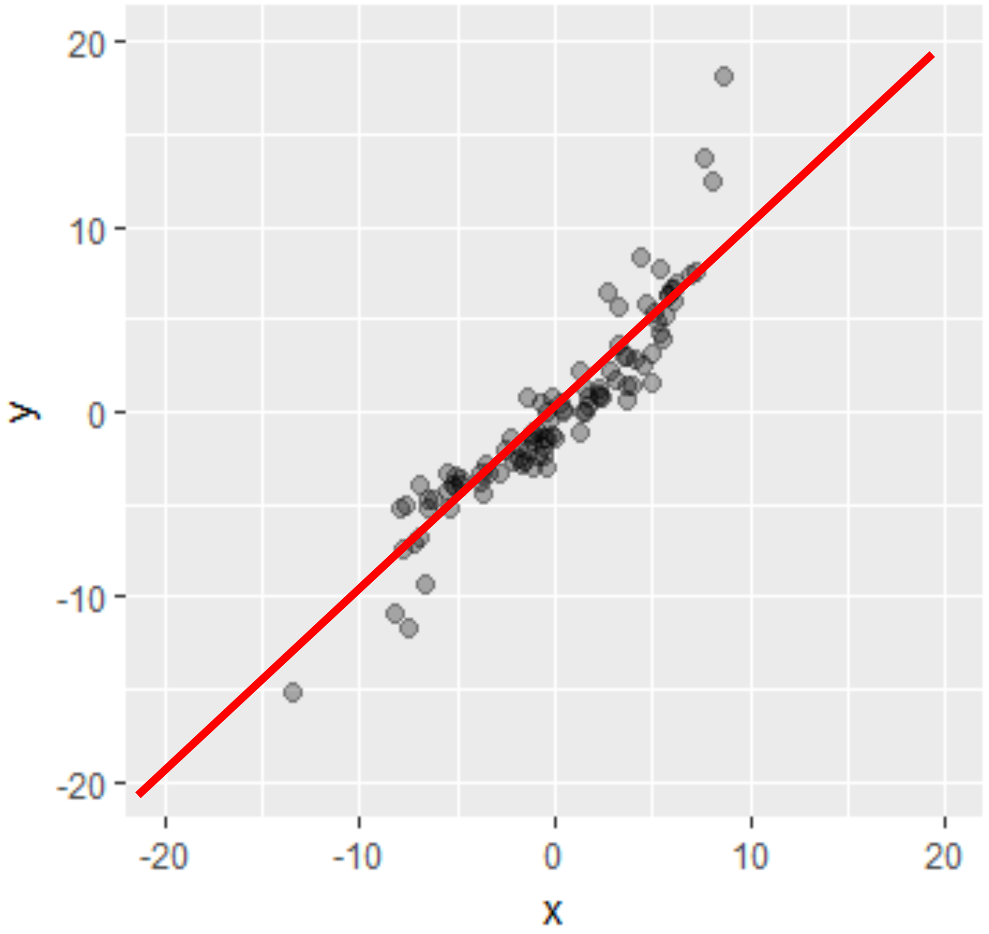
어떻게 선을 그을 수 있었을까요?

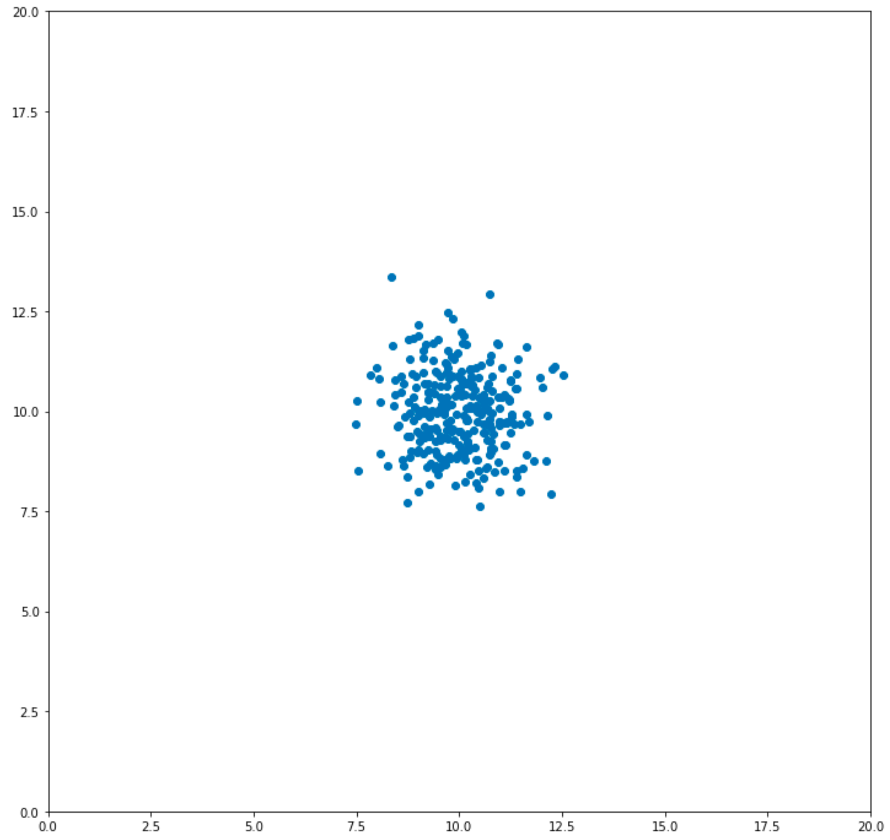
그냥 그렇게 해야 할 것 같았어요
그래야 마음이 편하거든요

그냥 머리 속에서 이렇게 하래요

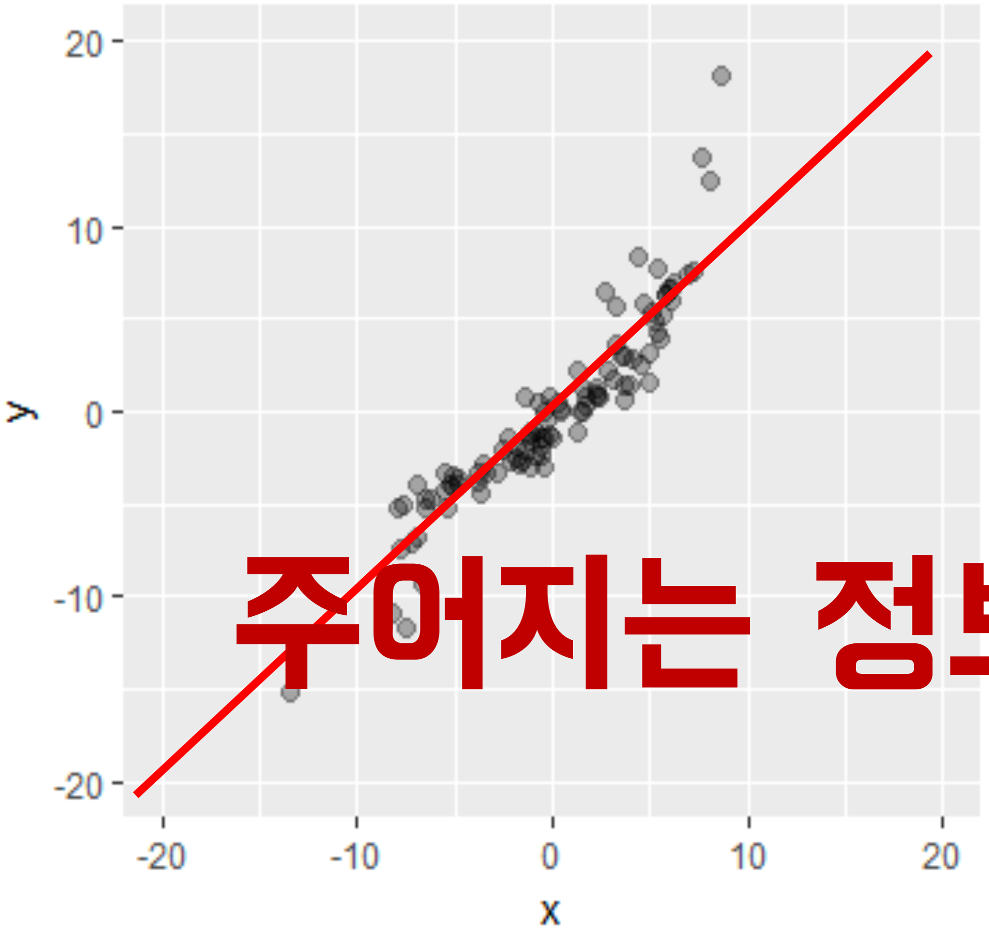
Taejjing

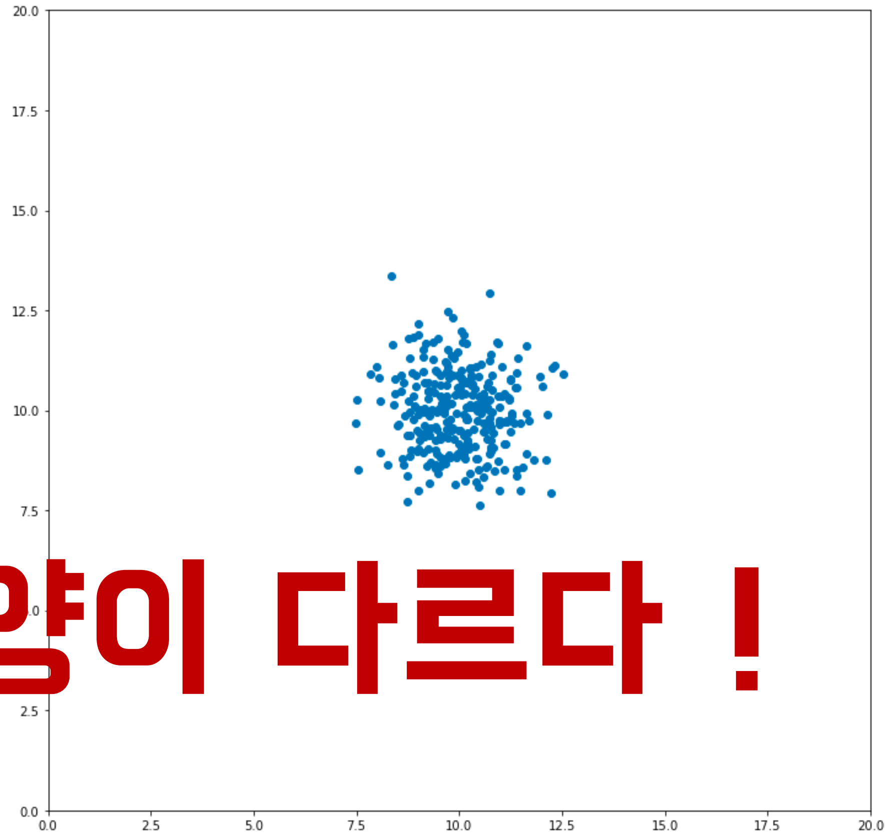
두 그림의 차이는 무엇일까요?





Taejjing





주어지는 정보의 양이 다르다 !

가로로 퍼져 있음

(분포가) 넓어 보임

다음 점의 위치를 대략 알 수 있음

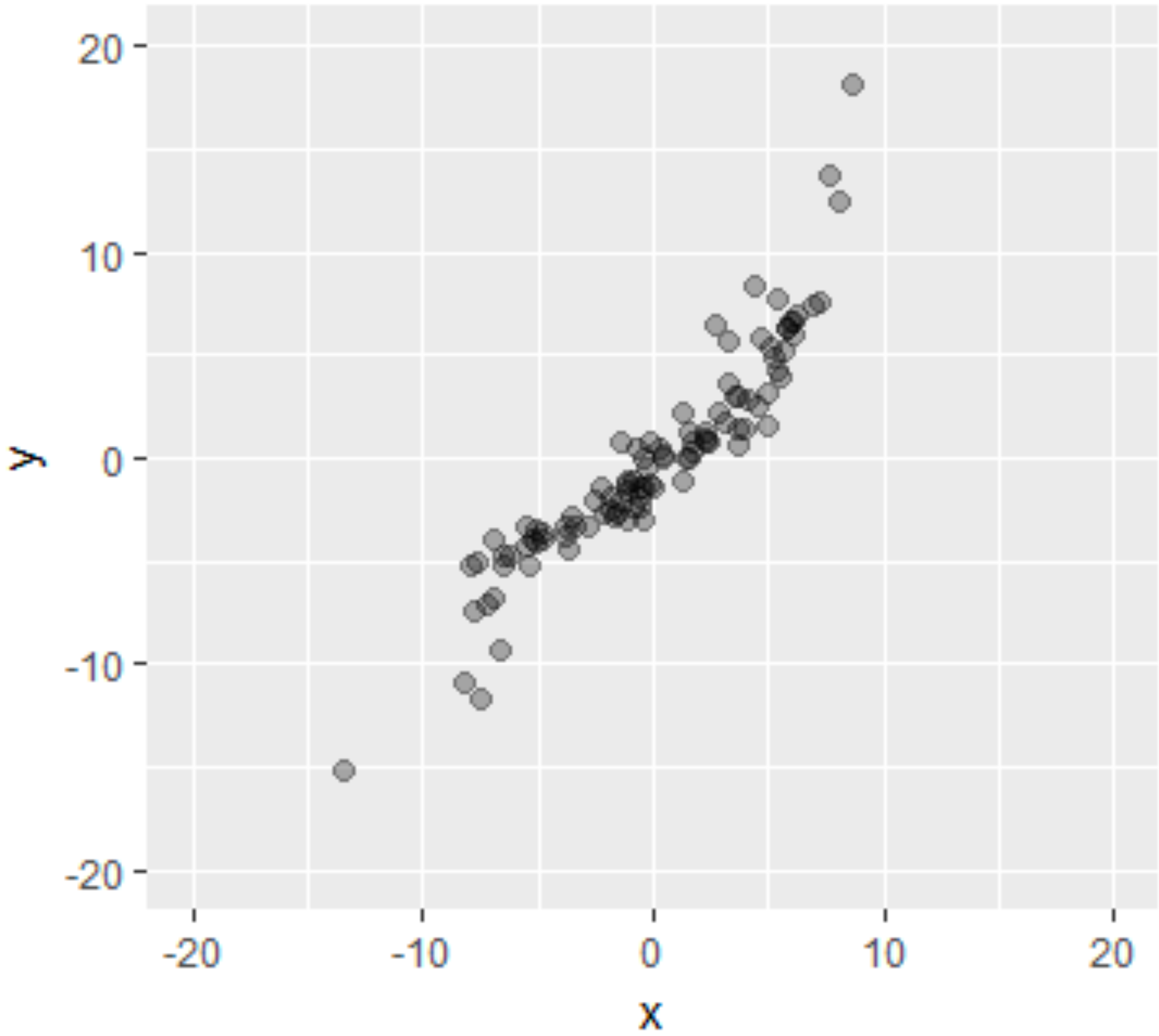
모여 있음

좁아 보임

어디로 튈 지 모르겠음

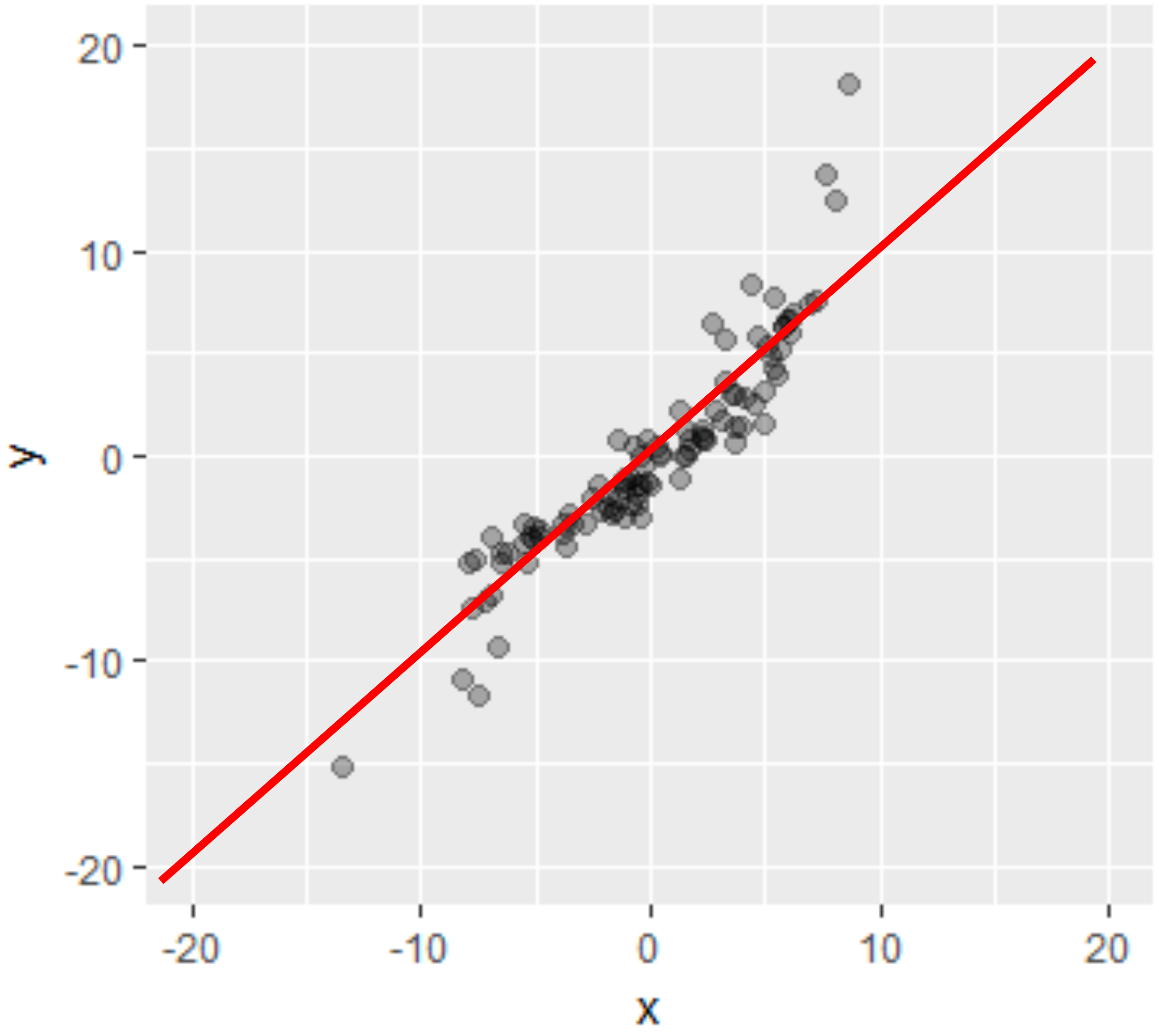
Taejjing

임의의 그래프에서



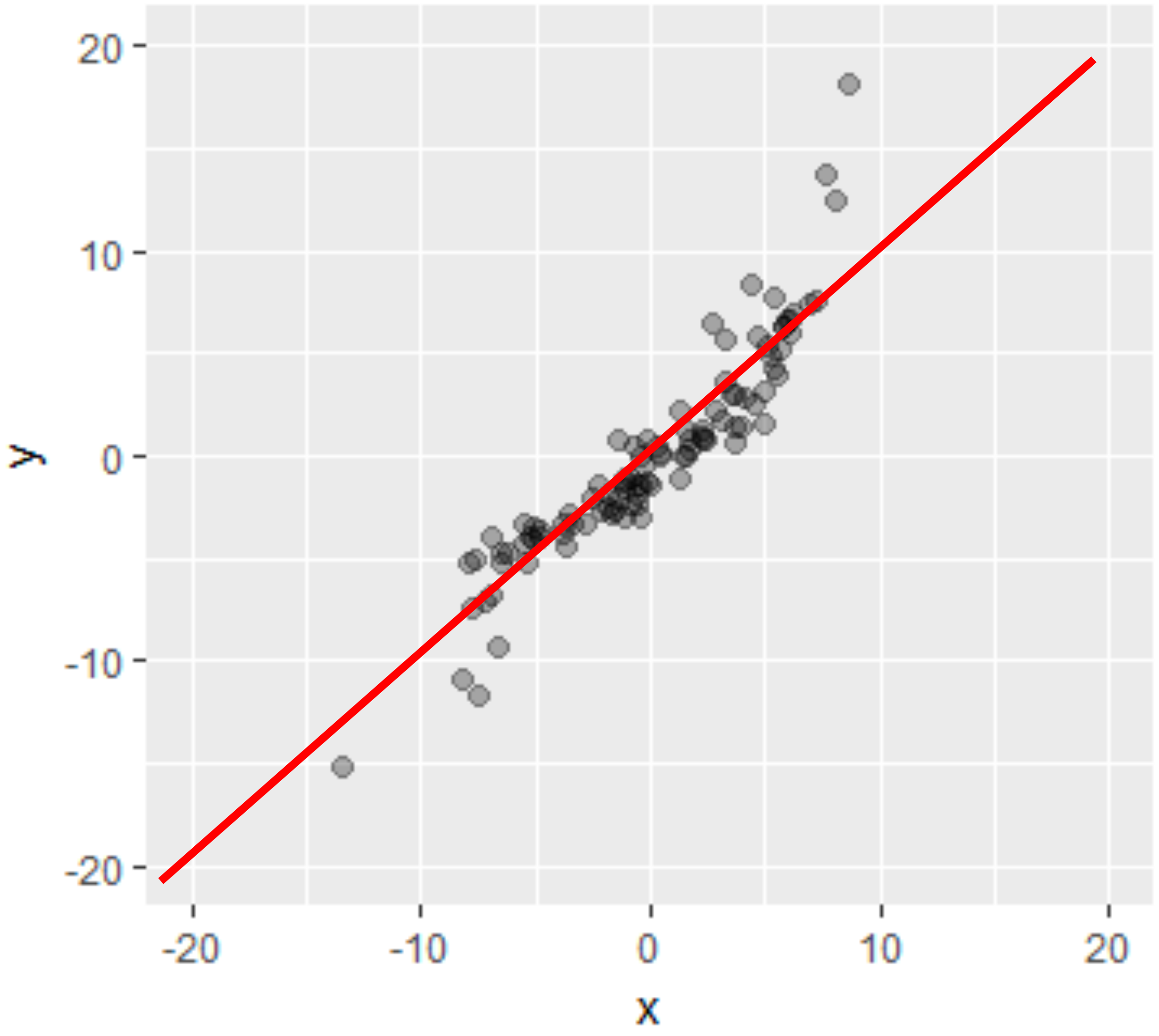
Taejjing

분산이 가장 큰 방향으로 선을 잡으면



Taejjing

이 방향(벡터)는 주어진 데이터에 대한
정보를 최대한 살릴 수 있으니



Taejjing

이것을 Principle Component 이라고 합시다

principle

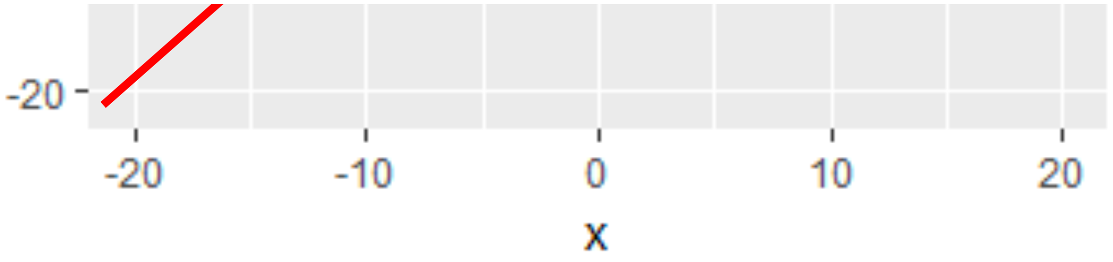
noun

UK /ˈprɪn.sə.pəl/ **US** /ˈprɪn.sə.pəl/

principle ***noun*** **(IDEA)**

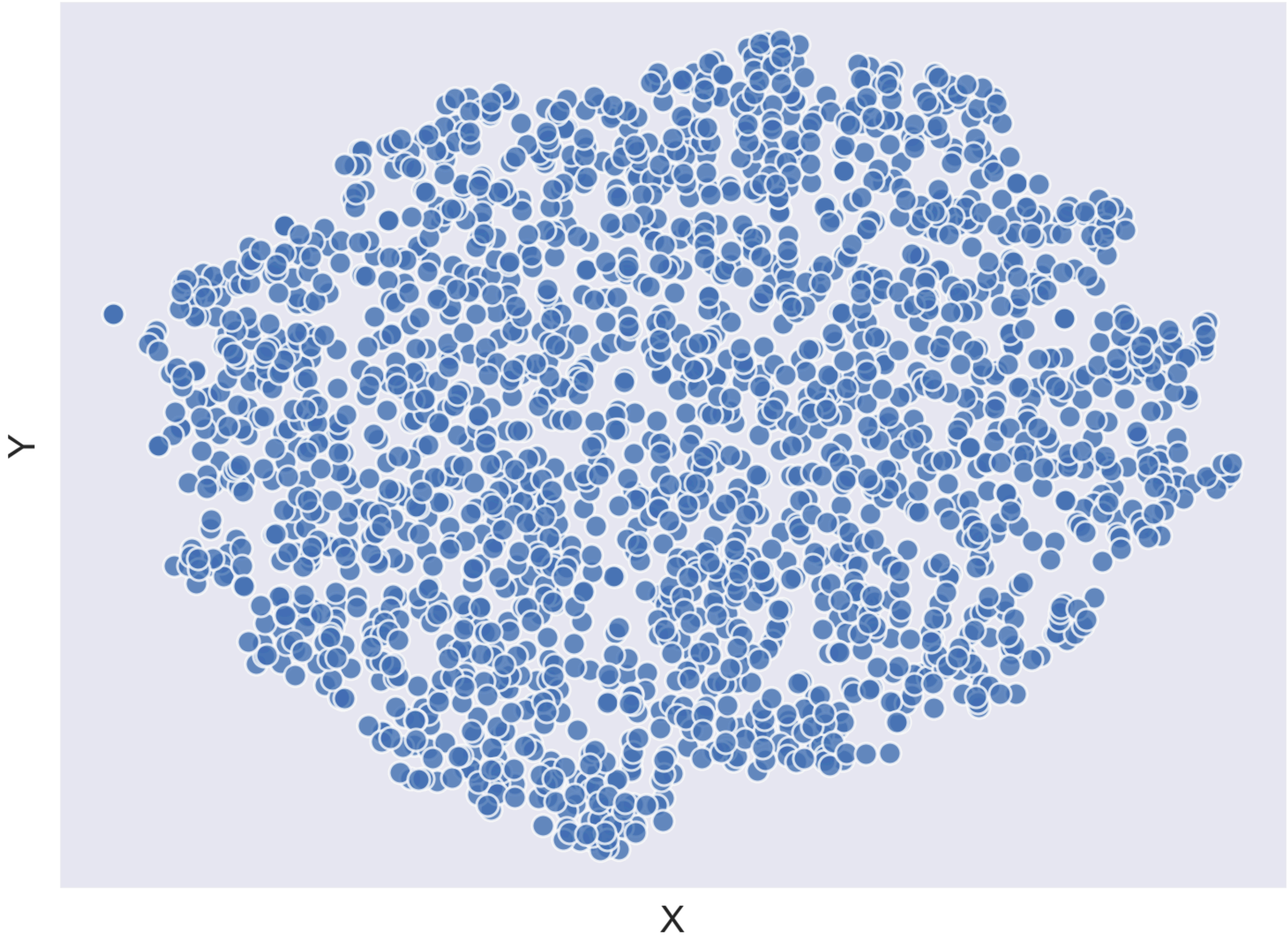
C1 [C]

a basic idea or rule that explains or controls how something happens or works:



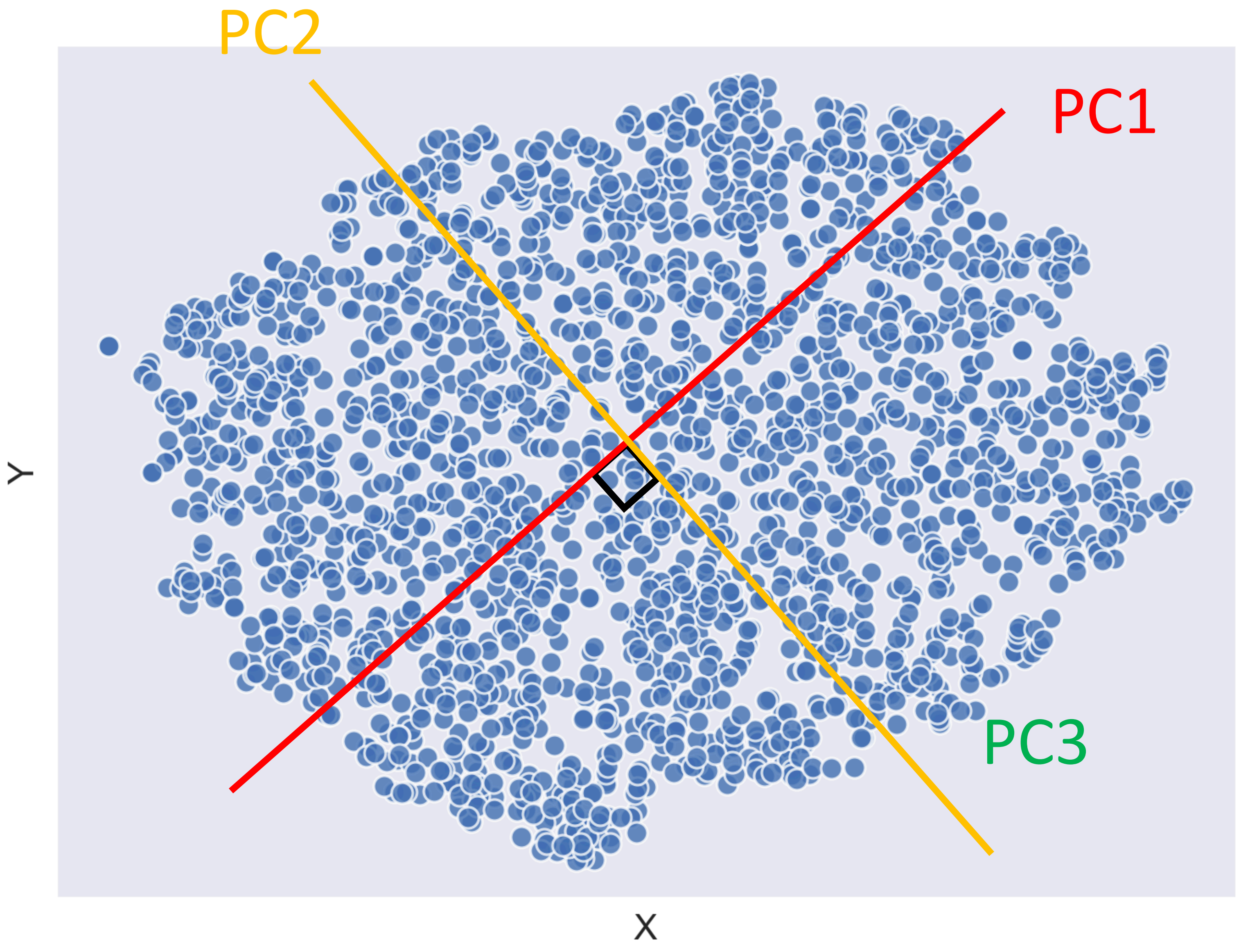
Taejjing

그럼 이건 ?



Taejjing

여러가지 경우를 생각해 볼 수 있죠.



Taejjing

임의의 데이터가

여러(많은) Principle Component로

구성되어 있다고 가정

데이터를 구성하고 있는 Principle들을 찾아서
그것을 토대로 데이터의 특징을 알아보는 것 !

Principle 주된

Component 성분을

Analysis 분석하자

Taejjing

그런데, 차원 축소라면서요..?

Taejjing

만약, 다 뽑을 수 없는 경우라면...

특출한 몇명만 뽑아야 합니다.

Taejjing

데이터를 구성하고 있는 Principle들을 찾아서
그것을 토대로 **데이터의 특징을 알아보는 것** !

그 중에서 가장
두드러진 Principle들 선택!

How..?

두드러진 Principle을

데이터의 정보를 많이 가지고 있을

분산이 최대인 방향을 가진 벡터를

n차원의 데이터

n x n 공분산 행렬
(Covariance Matrix)

고유 분해
(Eigen Decomposition)

Eigen Vector들 가운데서
Eigen Value가 가장 큰 벡터

분산을 가장 크게 만드는 벡터

Taejjing

공분산 행렬 (Covariance Matrix)

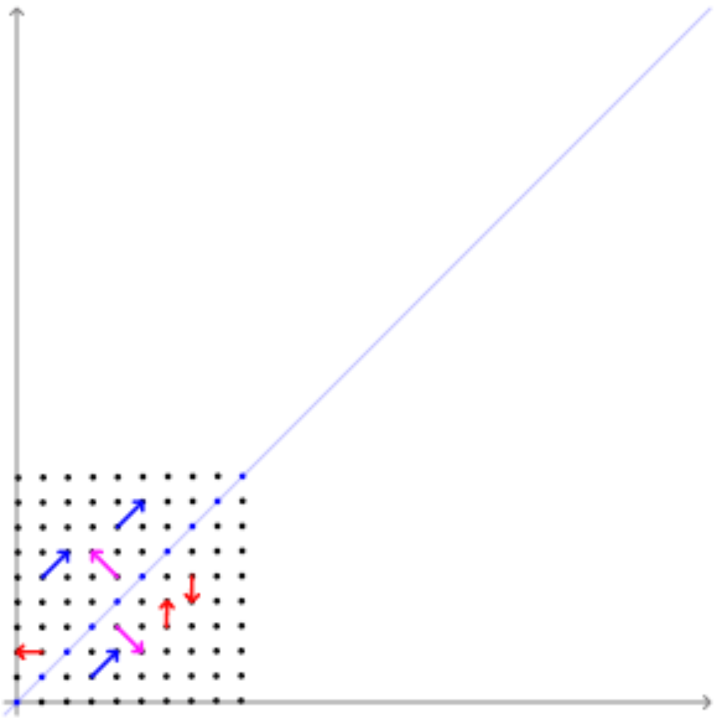
$$cov(x, y) = E[(x - m_x)(y - m_y)]$$

$$cov(X, Y) = \begin{bmatrix} cov(x_1, y_1) & \cdots & cov(x_1, y_n) \\ \vdots & & \vdots \\ cov(x_n, y_1) & \cdots & cov(x_n, y_n) \end{bmatrix}$$

두 변수의 퍼짐의 정도에 얼마나 상관관계가 있나?

Taejjing

고유 분해 (Eigen Decomposition)



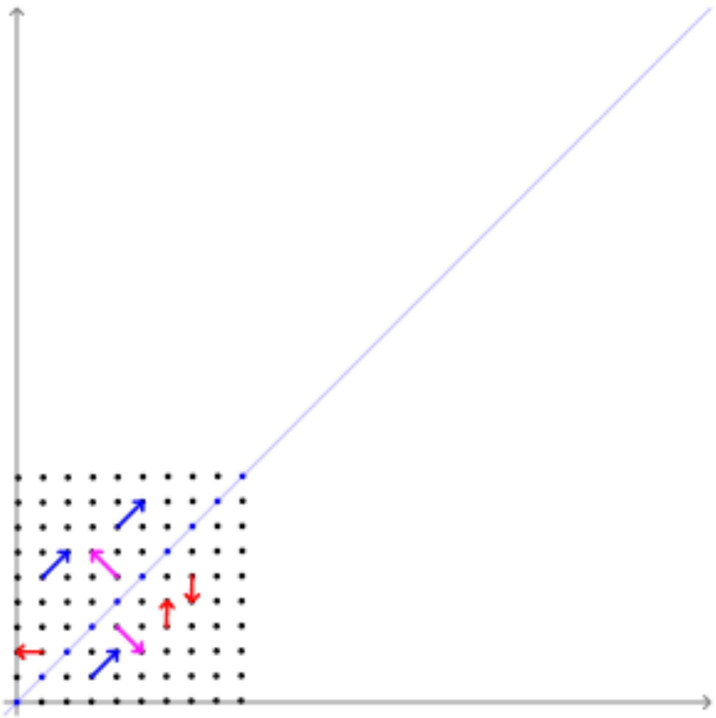
임의의 벡터를 행렬 A로 변환한다고 했을 때

어떤 벡터는 크기에 변화는 있을 뿐
그 방향은 전혀 변하지 않음

즉, 행렬이 변환하려는 방향과
벡터의 방향이 같음

고유 벡터
(Eigen Vector)

고유 분해 (Eigen Decomposition)



$$Av = \lambda v$$

따라서, 어떤 행렬 A의 변환에 대해서

스케일의 변화만 존재하는

행렬 A의 방향성을 잘 설명할 수 있는

특정 벡터를 찾는 과정

공분산 행렬 X 고유 분해

Covariance Matrix 데이터의 각 성분의 분산(퍼짐)의 정도

X

Eigen Decomposition 행렬의 방향성을 내포하는 특정 벡터를 찾는 과정

주어진 데이터에서
가장 큰 분산을 나타내는 벡터를 구할 수 있다!

Taejjing

그래서... PCA는요

방금의 과정을 통해서 만들어진 특출한 벡터 위로

데이터를 Projection하면

Principle Components로 만들어진
새로운 차원의 데이터 완성!

➡ 차원 감소
(Dimension Reduction)

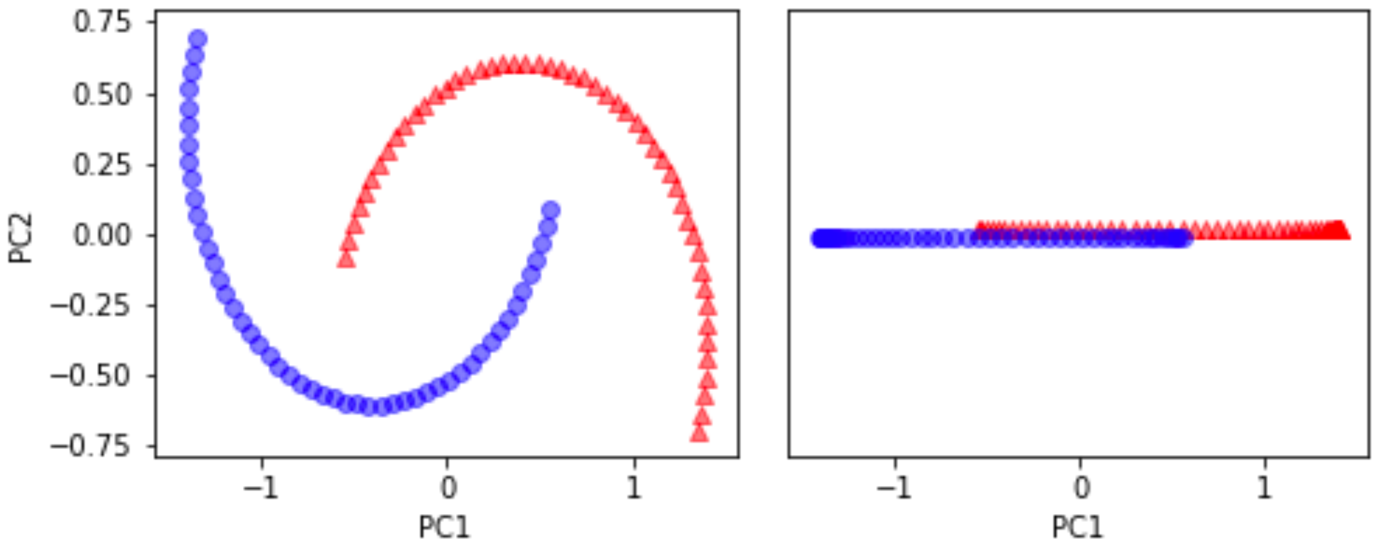
Taejjing

Kernel PCA ???

방금의 과정을 통해서 만들어진 특출한 벡터 위로

데이터를 Projection하려 했으나

선형적으로 해결할 수 있는 문제가 아니네

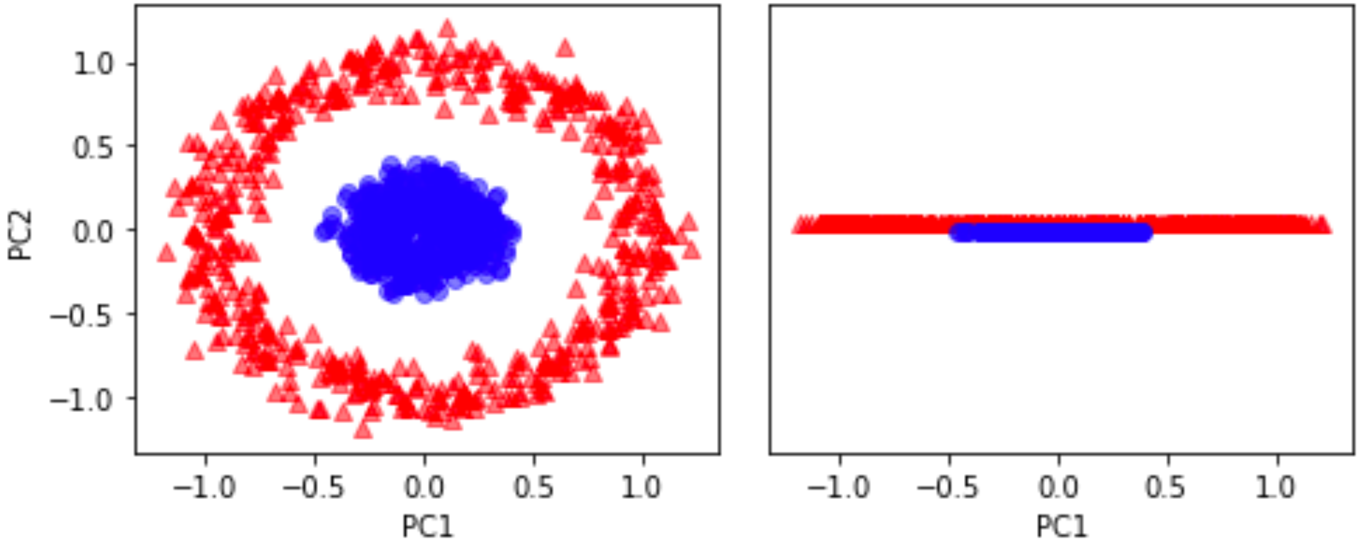


Kernel PCA ???

방금의 과정을 통해서 만들어진 특출한 벡터 위로

데이터를 Projection하려 했으나

선형적으로 해결할 수 있는 문제가 아니네



Taejjing

Kernel PCA ???

- Polynomial kernel

$$k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle^d$$

- RBF kernel

$$k(\mathbf{x}, \mathbf{y}) = \exp\left\{-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right\}$$

- Sigmoid kernel

$$k(\mathbf{x}, \mathbf{y}) = \tanh\left(\kappa \langle \mathbf{x}, \mathbf{y} \rangle + \theta\right),$$

for suitable values of gain κ and threshold θ.

비선형 함수를 빌려 아예 Base를 변경

Taejjing