Principal Component Analysis

team E : 안동현 신민경 여인언 정지윤

Contents

주성분 분석이란?

데이터의 **분포**를 설명하기 위해,

분포를 가장 잘 설명해주는 **주성분**을 이용하는 방법

주성분이란?

변수들의 **변동(variation)**을 가장 잘 설명하는 성분

주성분이란?

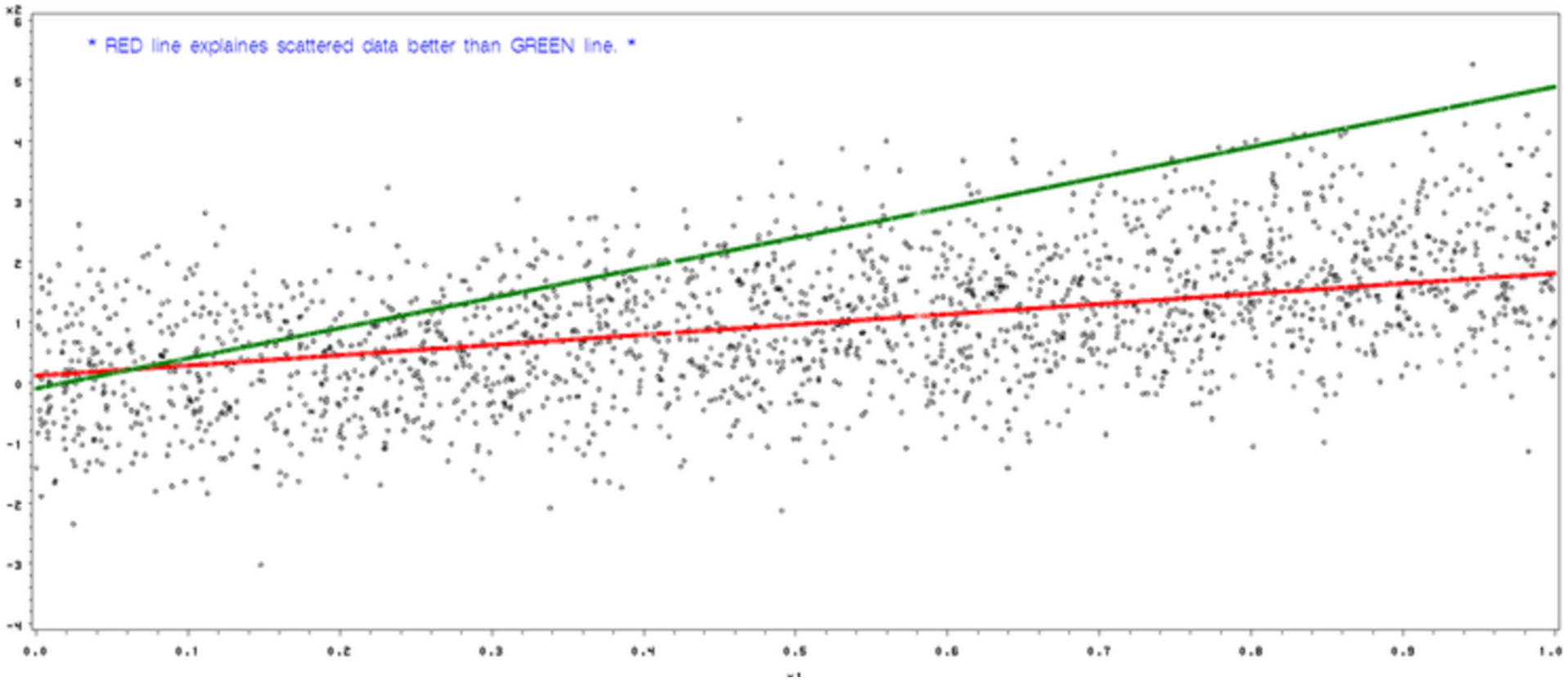
주성분 = 변수들을 설명하는 새로운 **축(axis)**

= 변수들을 **대표**하는 개념

ex) 키, 몸무게, 팔굽혀 펴기 횟수, 달리기 기록 …. > ‘체력’

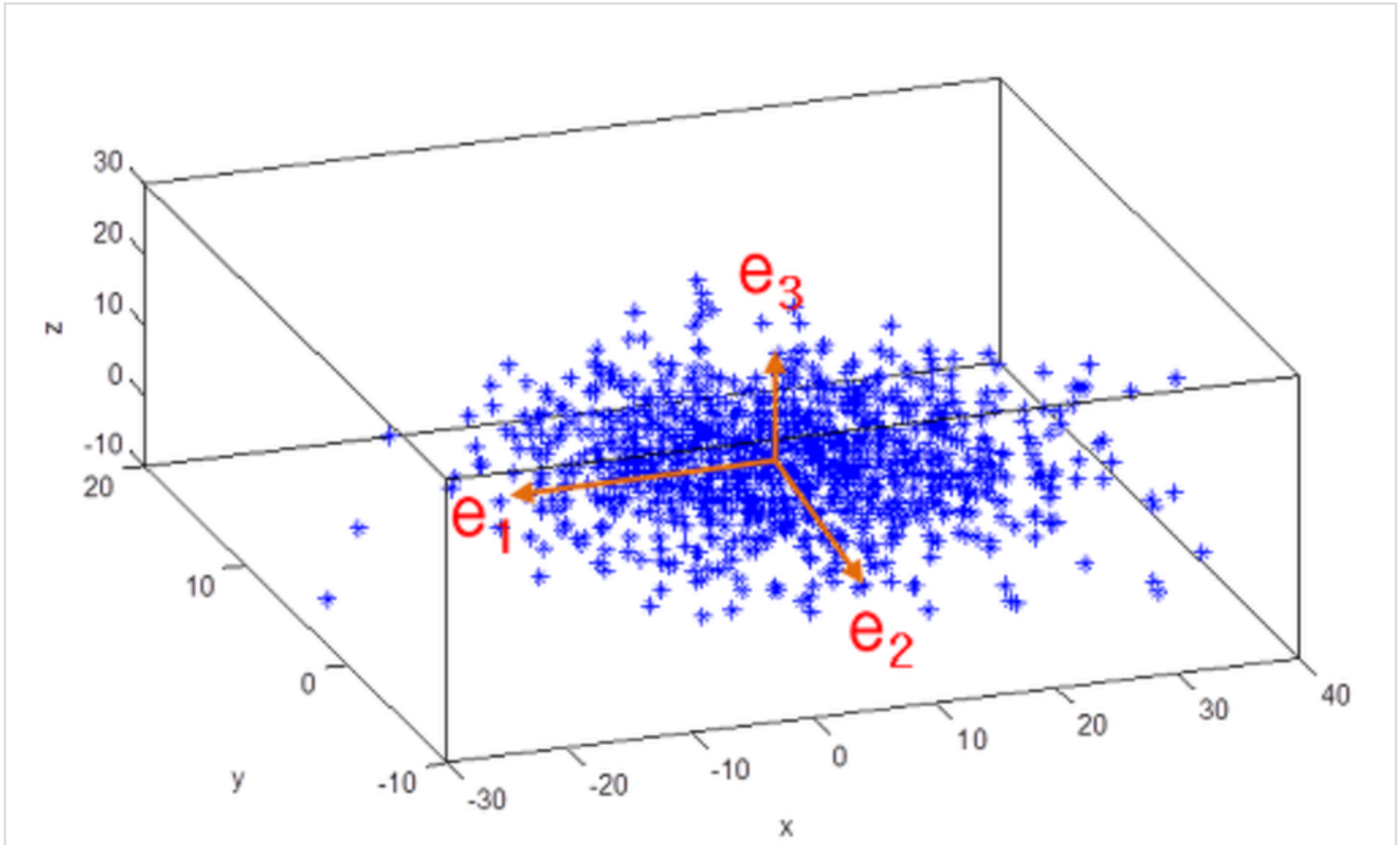
자동차 배기량, 크기, 엔진 기능, 브랜드 …. > ‘가격’

주성분 분석의 목적



변수들의 **변동(variation)**을 가장 잘 설명하는 '축'을 찾음

주성분 분석의 목적



<그림 2> 3D에서의 PCA

http://setosa.io/ev/principal-component-analysis/

주성분 분석의 목적

고차원 데이터의

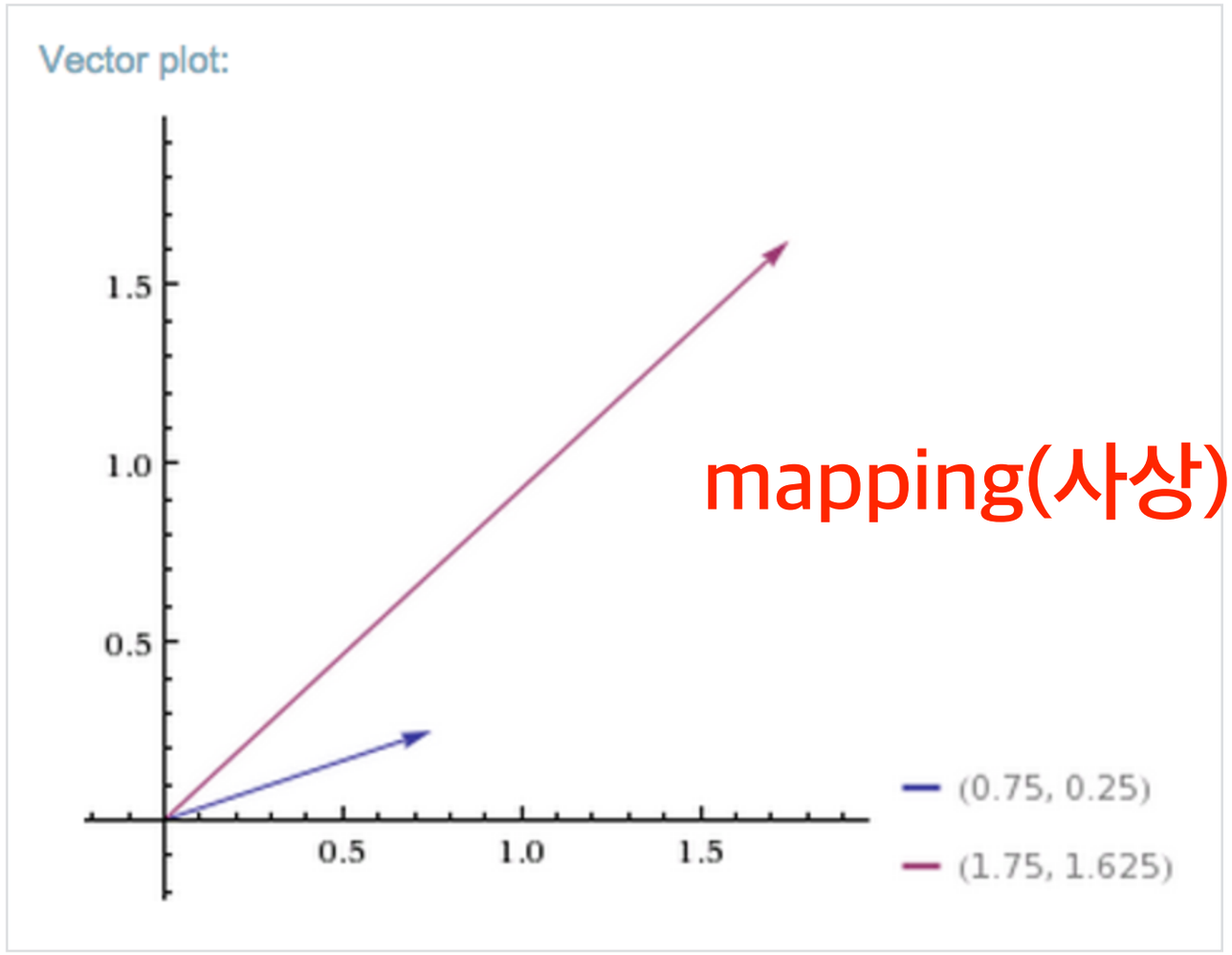
-> **차원**을 **축소**한다.

-> 정보를 **축약**해 보여준다.

[Linear Algebra] Linear Transformation

$$A \qquad v \qquad b$$

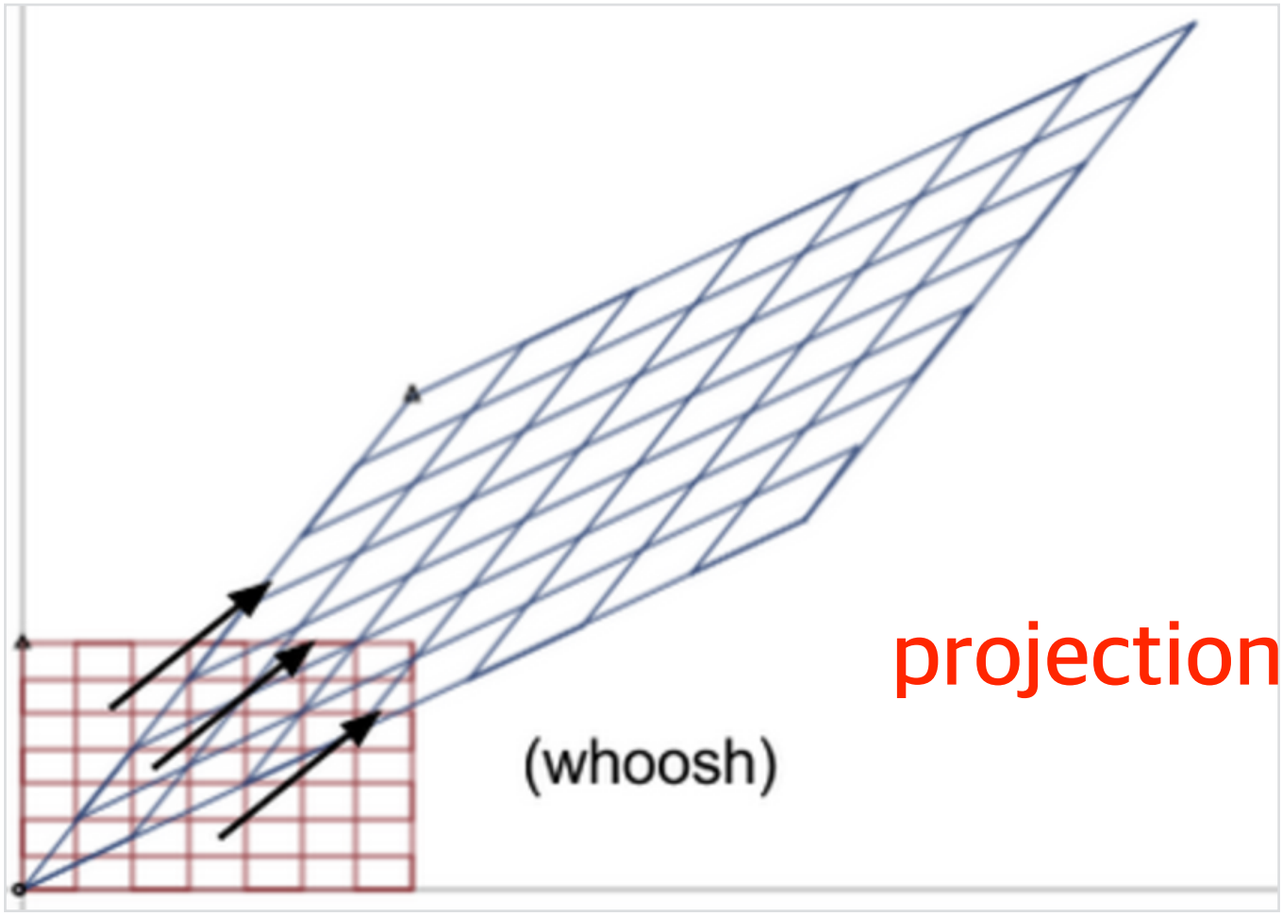
$$\begin{bmatrix} 2 & 1 \\ 1.5 & 2 \end{bmatrix} * \begin{matrix} 0.75 \\ 0.25 \end{matrix} = \begin{matrix} 1.75 \\ 1.625 \end{matrix}$$



https://deeplearning4j.org/kr/kr-eigenvector

[Linear Algebra] Linear Transformation

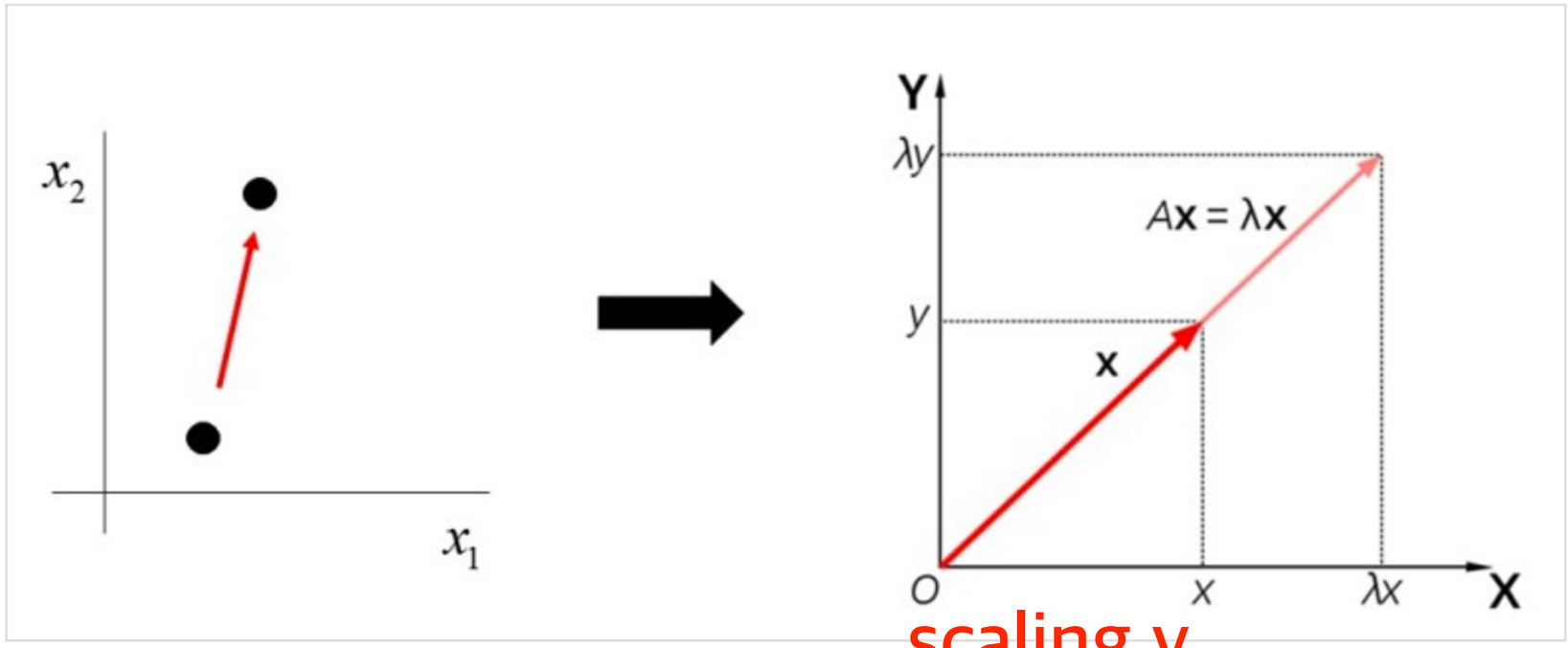
$$\underbrace{\begin{bmatrix} 2 & 1 \\ 1.5 & 2 \end{bmatrix}}_{A} * \underbrace{\begin{matrix} 0.75 \\ 0.25 \end{matrix}}_{v} = \underbrace{\begin{matrix} 1.75 \\ 1.625 \end{matrix}}_{b}$$



projection(사영)

https://deeplearning4j.org/kr/kr-eigenvector

[Linear Algebra] Eigenvalue & Eigenvector



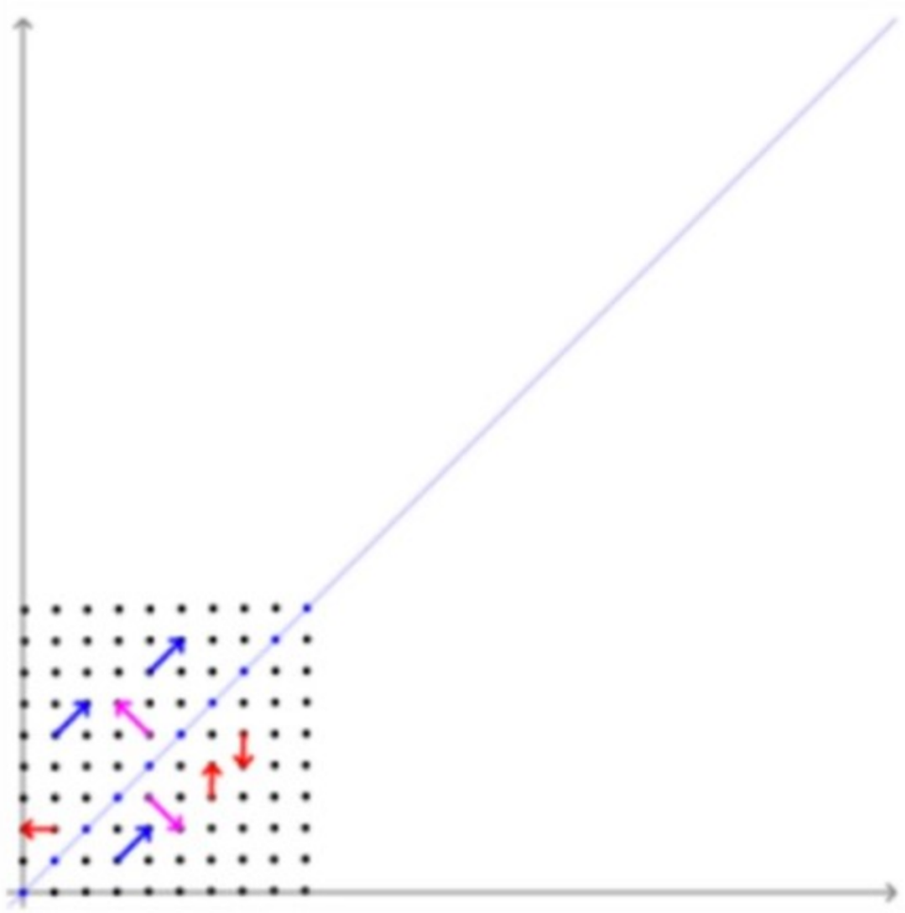
scaling v

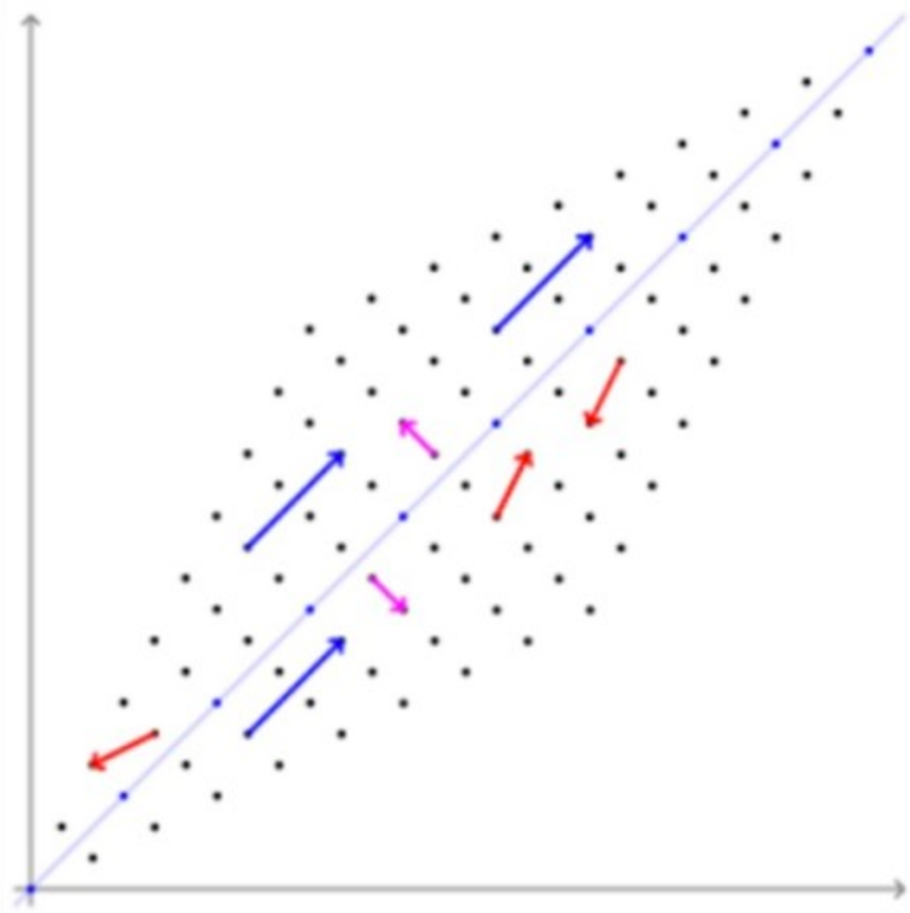
$$A\mathbf{v}=\lambda\mathbf{v}=b$$

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}=\lambda\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

v: Eigenvector, λ: Eigenvalue

[Linear Algebra] Eigenvalue & Eigenvector

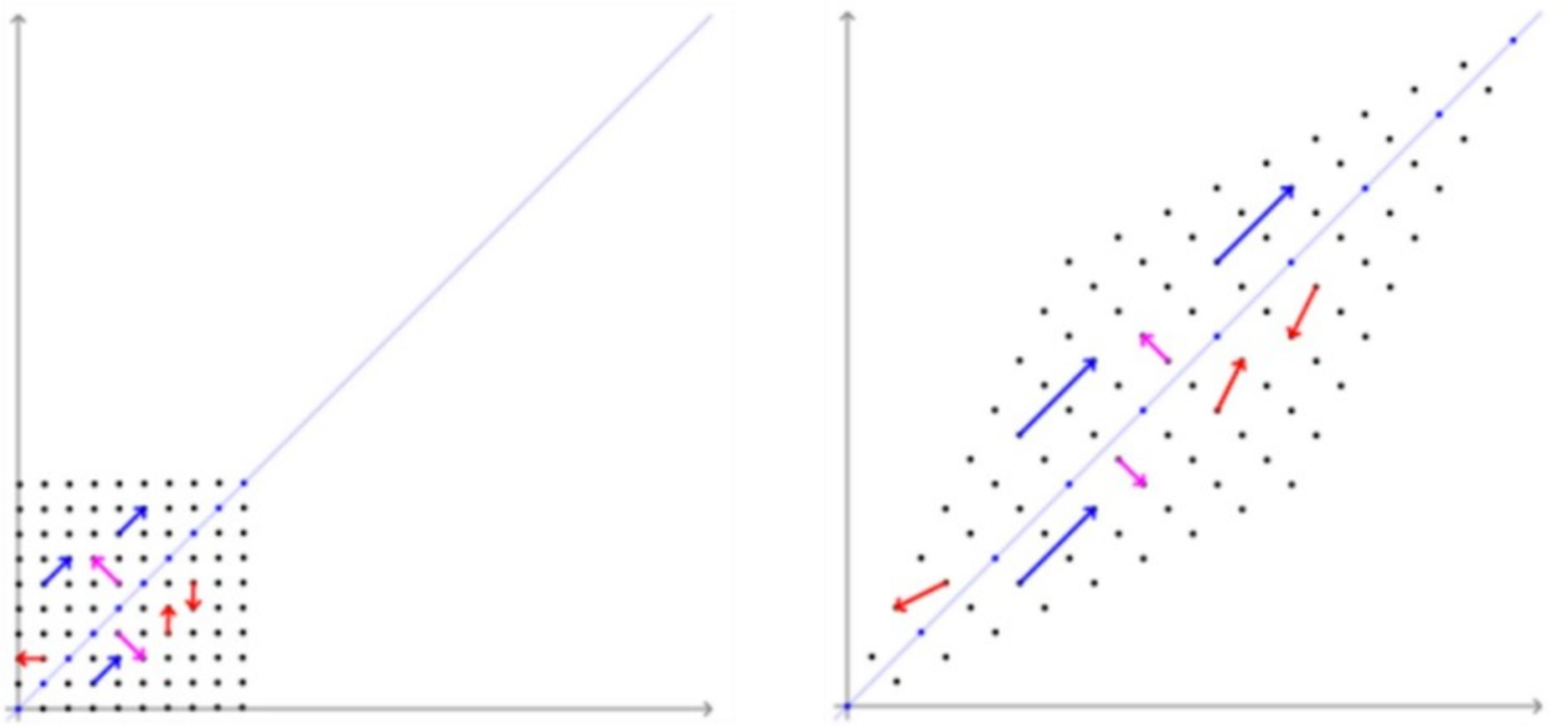




방향이 **변하지 않는** 벡터(파란색, 분홍색)

> 기준이 될 수 있다!

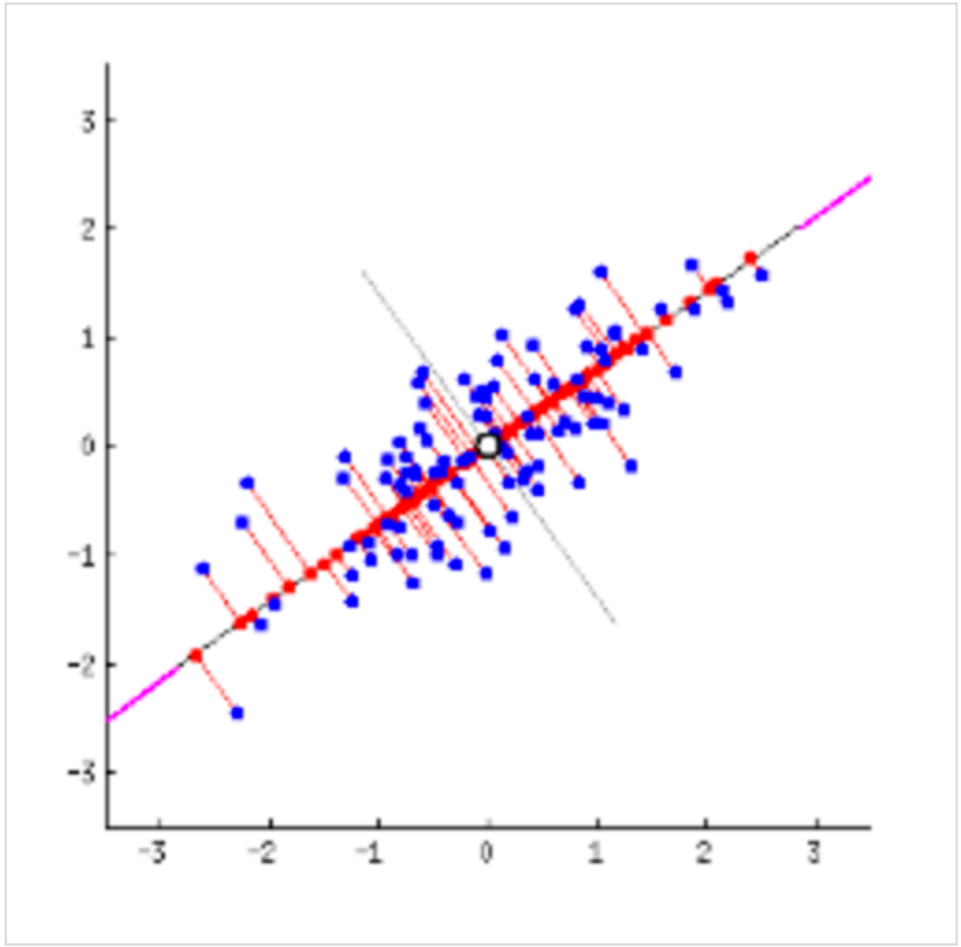
[Linear Algebra] Eigenvalue & Eigenvector



행렬이 작용하는 **힘**과 방향이 같은 벡터

https://deeplearning4j.org/kr/kr-eigenvector

[Linear Algebra] Eigenvalue & Eigenvector

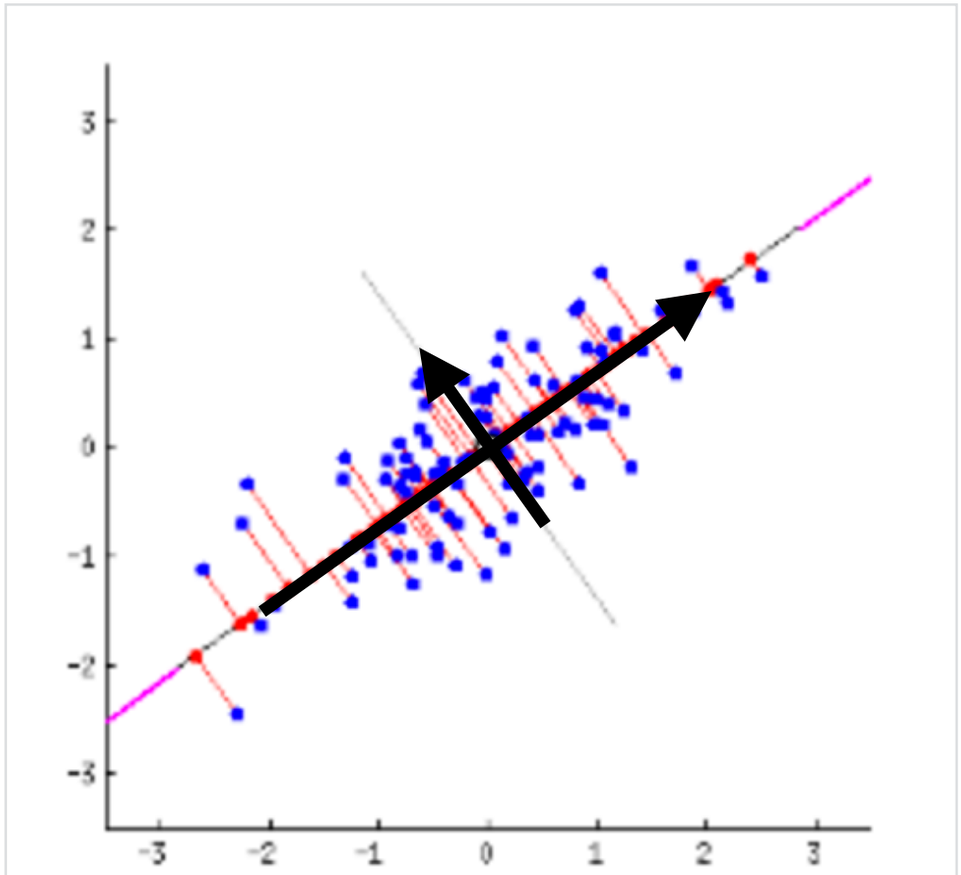


Eigenvector = **주성분 벡터**; 힘이 작용하는 **방향**

Eigenvalue = **분산**; 힘의 **크기**

https://deeplearning4j.org/kr/kr-eigenvector

[Linear Algebra] Eigenvalue & Eigenvector



1st P.C. : 분산이 가장 큰 벡터

2nd P.C. : 1st P.C.와 orthogonal하고, 분산이 가장 큰 벡터

https://deeplearning4j.org/kr/kr-eigenvector

[Linear Algebra] Eigenvalue & Eigenvector

- 모든 **정방행렬**은 eigenvector를 가짐(not unique)

- **N**-dimentional 데이터는 최대 **N개의 P.C.**를 가짐 (by vector space's definition)

Covariance Matrix

covariance

– $\mathrm{cov}(x,y) = E[(x-m_x)(y-m_y)]$

covariance matrix

– $x=[x_1,\ldots,x_n]^T$: sample data, n차원 열벡터

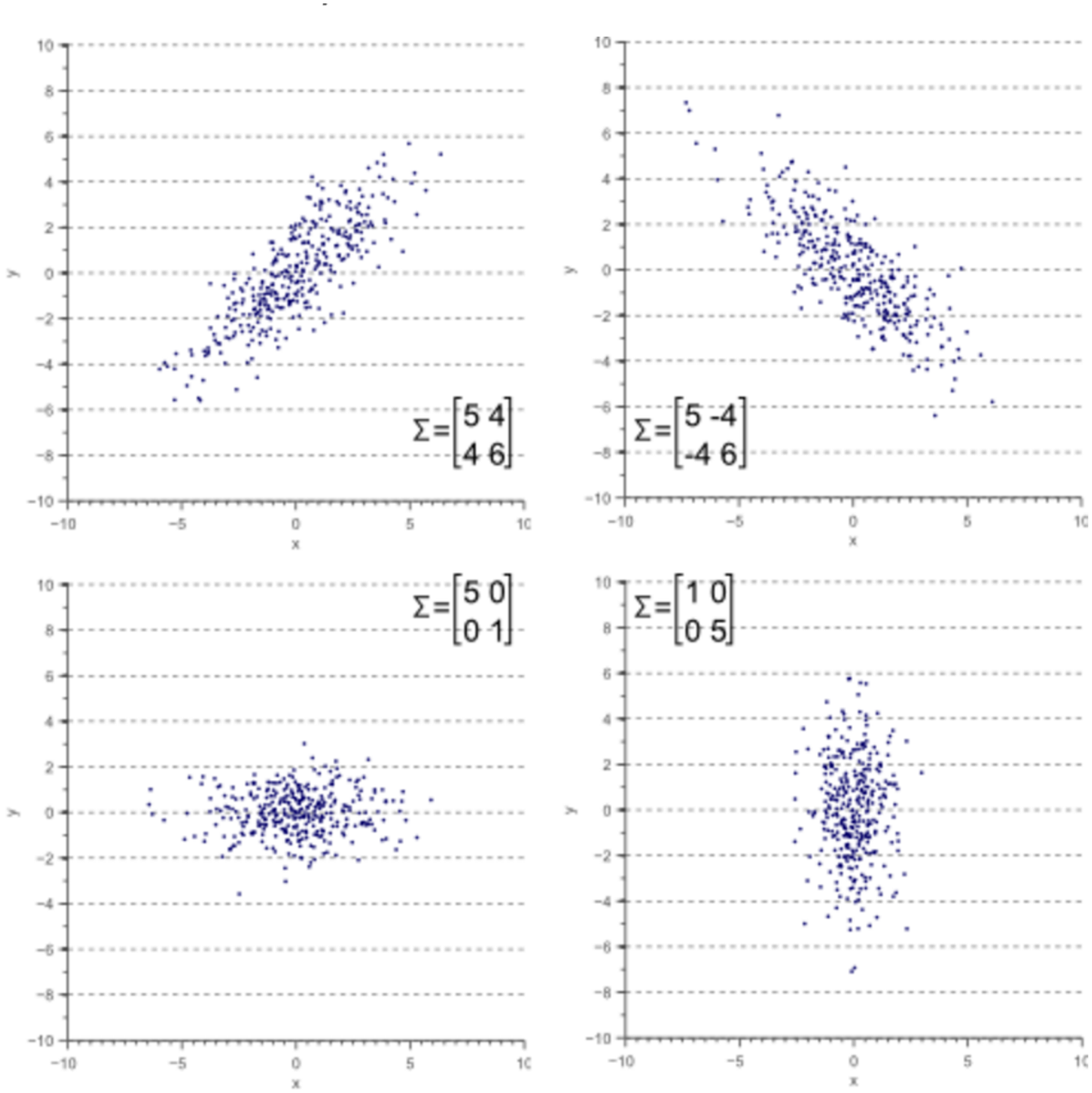
– $C = E[(x-m_X)(x-m_X)^T]$: n×n 행렬

– $\langle C \rangle_{ij} = E[(x_i-m_{xi})(x_j-m_{xj})^T]$: i번째 성분과 j번째 성분의 공분산

– C is real and symmetric

$$C = \begin{pmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nn} \end{pmatrix}$$

Covariance Matrix



https://deeplearning4j.org/kr/kr-eigenvector

주성분 분석 방법

1. covariance matrix를 구한다.

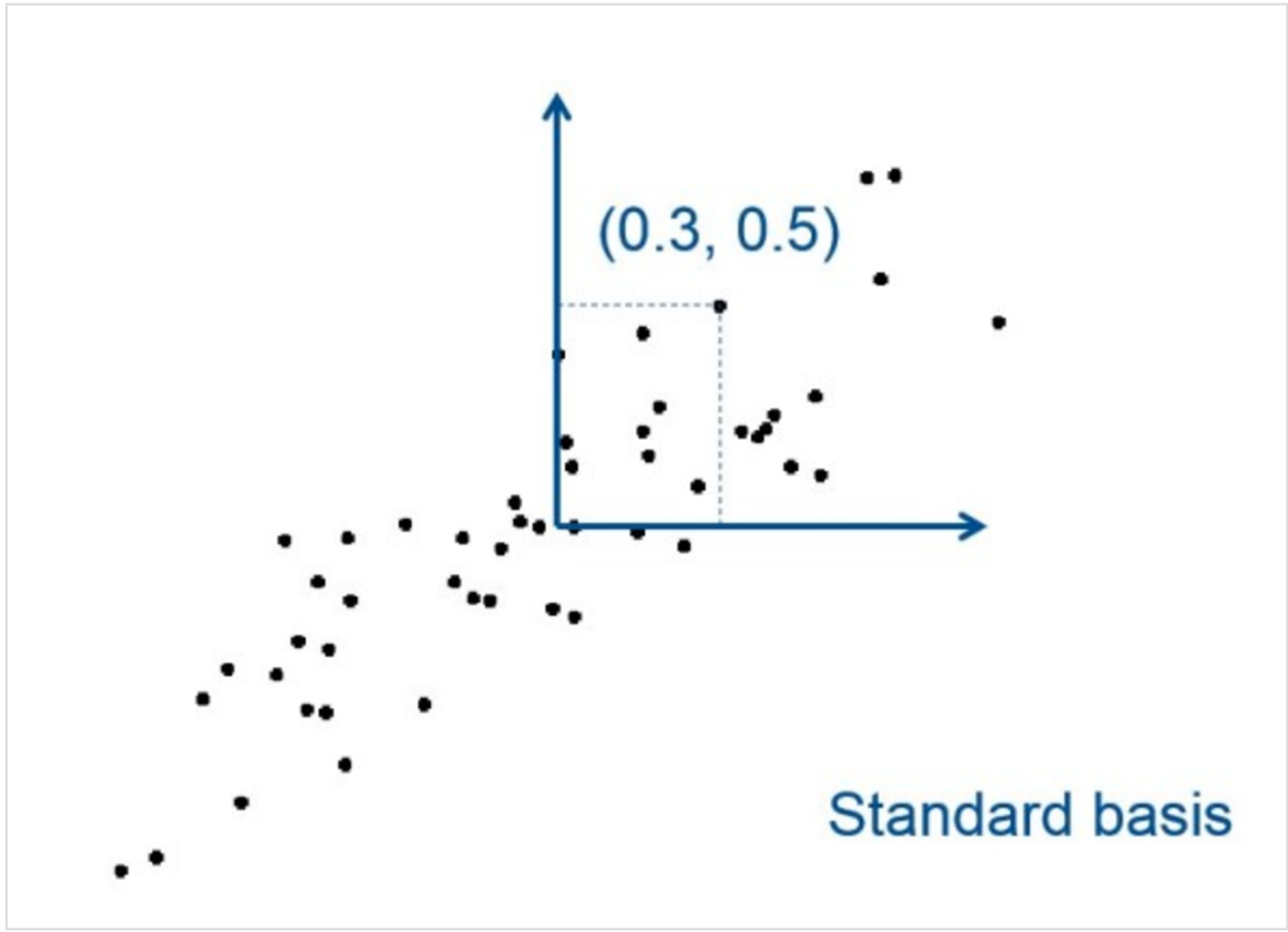
2. covariance matrix의 eigenvector를 구한다.

3. eigenvalue가 큰 순으로 principal component 채택

4. 각각의 P.C. 의 설명력을 구한다.

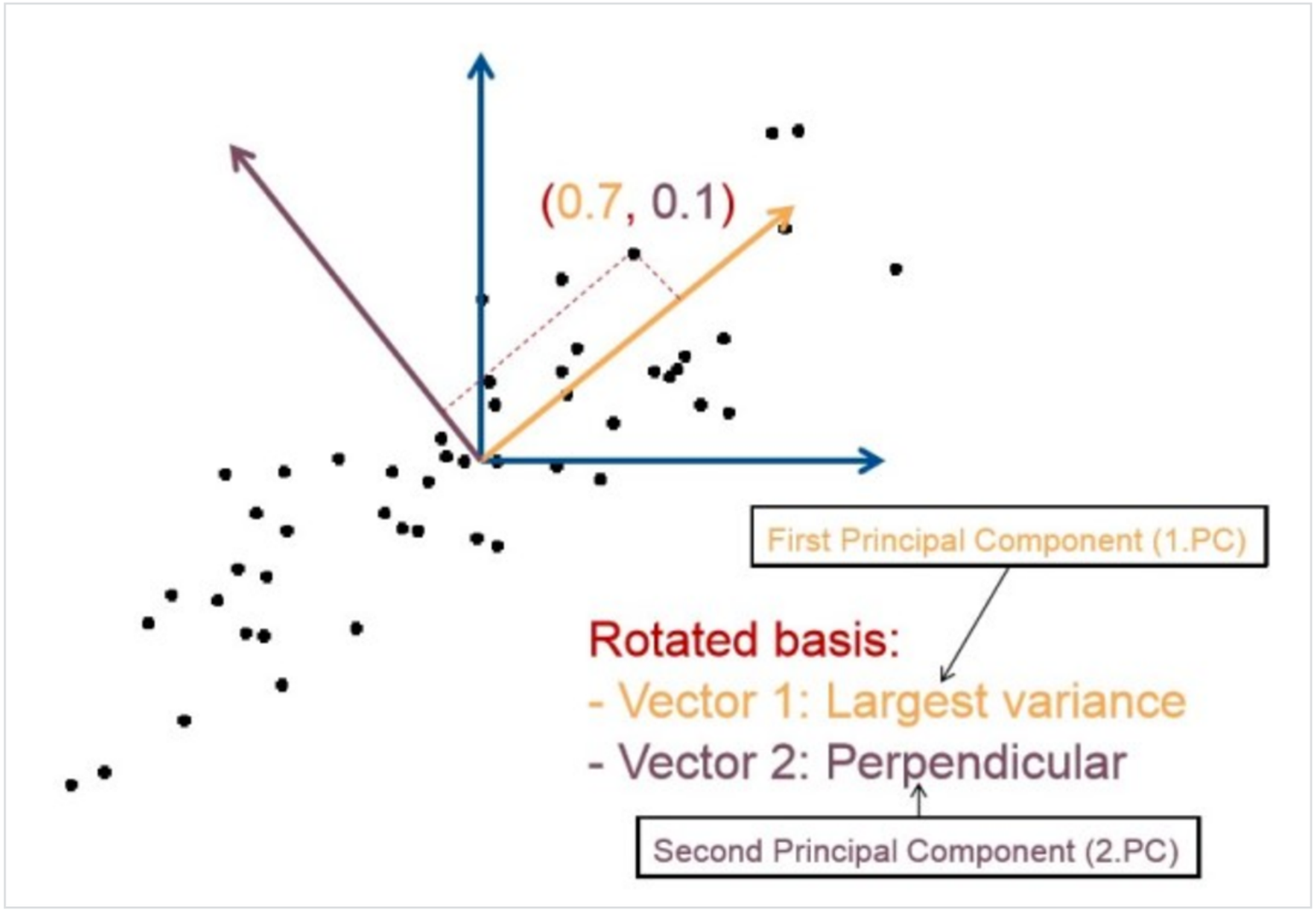
5. 설명력이 70~90%가 될 수 있는 만큼 P.C. 선택

주성분 분석 방법



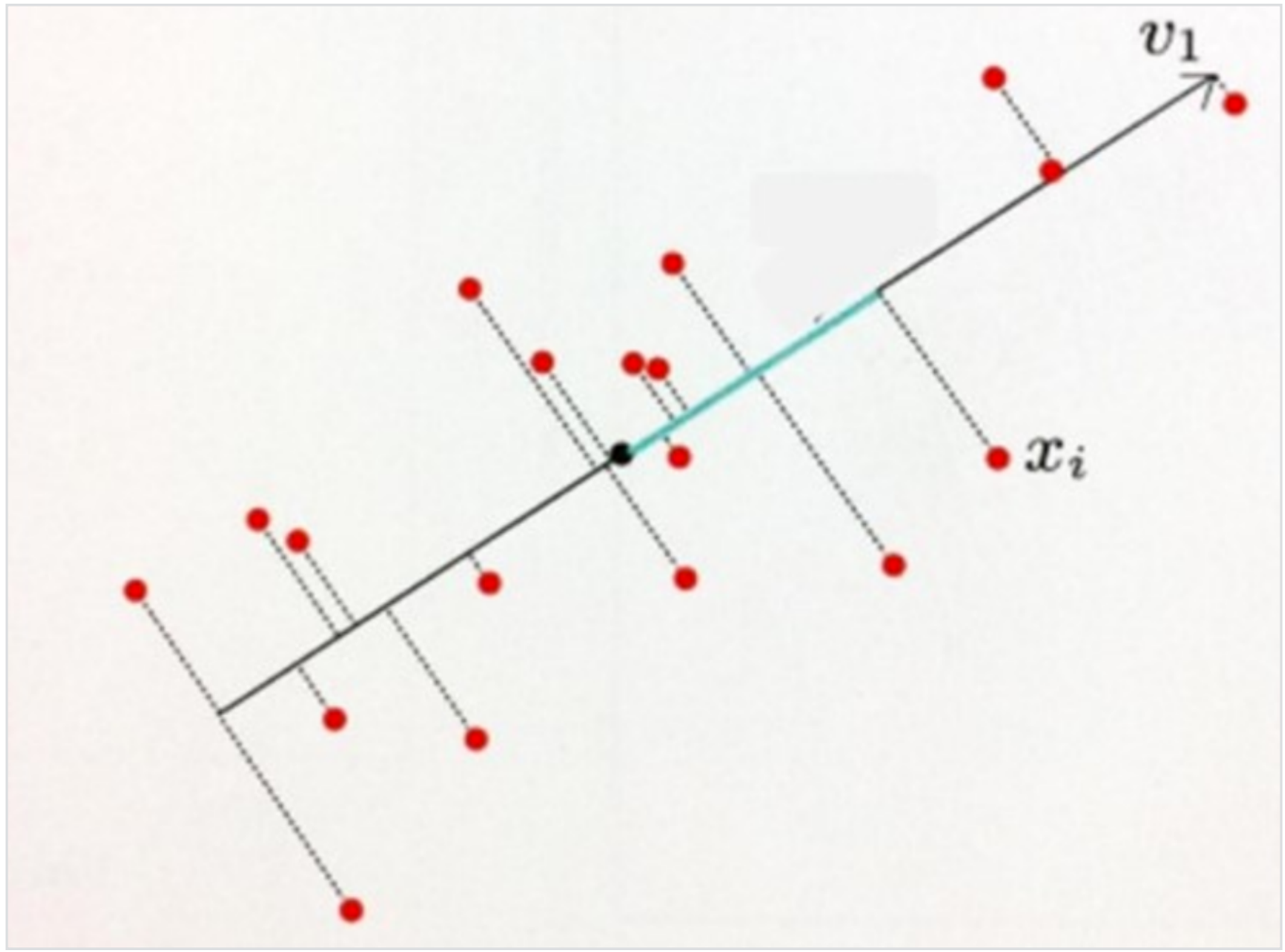
데이터의 **평균**을 기준으로 covarience matrix 구하기

주성분 분석 방법



eigen value => eigen vector를 구해 P.C. 추출

주성분 분석 방법



1st P.C.를 축(axis)으로 데이터 분포 설명