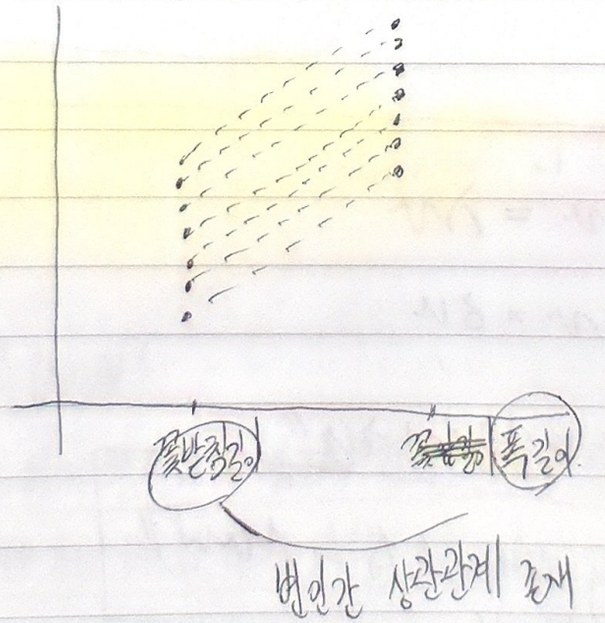
3.5 PCA

N개의 M차원 데이터가 있으면 보통 그 데이터들은 서로 다른 값을 가진다. 하지만 이러한 데이터간의 변이(variation)는 무작위가 아니라 특정한 규칙에 따라 만들어지는 경우가 있다. 예를 들어, 붓꽃의 꽃받침 길이는 꽃마다 다르지만, 꽃받침 길이가 약 2배 커지면 꽃받침 폭도 약 2배 커지는 것이 일반적이다. 이러한 데이터 간의 변이 규칙을 찾아낼 때 PCA를 이용할 수 있다.

PCA (Principal Component Analysis)
= 주성분분석
= 고차원 정보가 주어졌을 때 원래의 고차원 데이터와 가장 비슷하면서 더 낮은 차원의 데이터를 찾아내는 방법 (=차원축소)

"더 낮은 차원의 데이터가 더 높은 차원의 데이터를 설명할 수 있다"
→ 몇 가지 원인으로 데이터 변인을 설명할 수 있다.



→ "꽃의 크기"라는 근본적인 데이터가 "꽃받침길이", "폭길이"로 표현된 것
→ 측정되지는 않지만 측정 데이터의 저변에 숨어서 측정데이터를 결정짓는 데이터
→ 잠재변수 (latent variable)

PCA의 가정

- 잠재변수와 측정 데이터가 선형적인 관계로 연결되어 있다고 가정함.
  → i번째 표본의 측정 데이터 벡터 x_i의 각 원소를 선형결합하면 그 뒤에 숨은 i번째 표본의 잠재변수 u_i의 값을 계산할 수 있다고 가정한다.

$$(u_i = w^T x_i)$$

→ 이 식에서 w는 측정 데이터 벡터의 각 원소를 고려한 가중치 벡터다. 붓꽃의 예에서는 꽃받침길이와 꽃받침 폭을 선형결합하여 꽃의 크기를 나타내는 어떤 값을 찾은 것처럼 생각할 수 있다.

$$u_i = w_1 x_{i,1} + w_2 x_{i,2}$$

(latent)

ex) 미식축구 쿼터백의 패서 레이트 (passer rate)
→ 쿼터백의 "실력"이라는 잠재변수

$$\text{passing rate} = 5 \cdot \frac{\text{completions}}{\text{attempts}} + 0.25 \cdot \frac{\text{passing yards}}{\text{attempts}} + 20 \cdot \frac{\text{touchdowns}}{\text{attempts}} - 25 \cdot \frac{\text{interceptions}}{\text{attempts}} + 0.125$$

⇒ 차원축소 관점에서 보면 4차원 데이터를 1차원으로 축소한 것.

⇒ 고차원 데이터에서 latent variable을 찾아 간단하게 만들자
= PCA.

차원축소와 투영

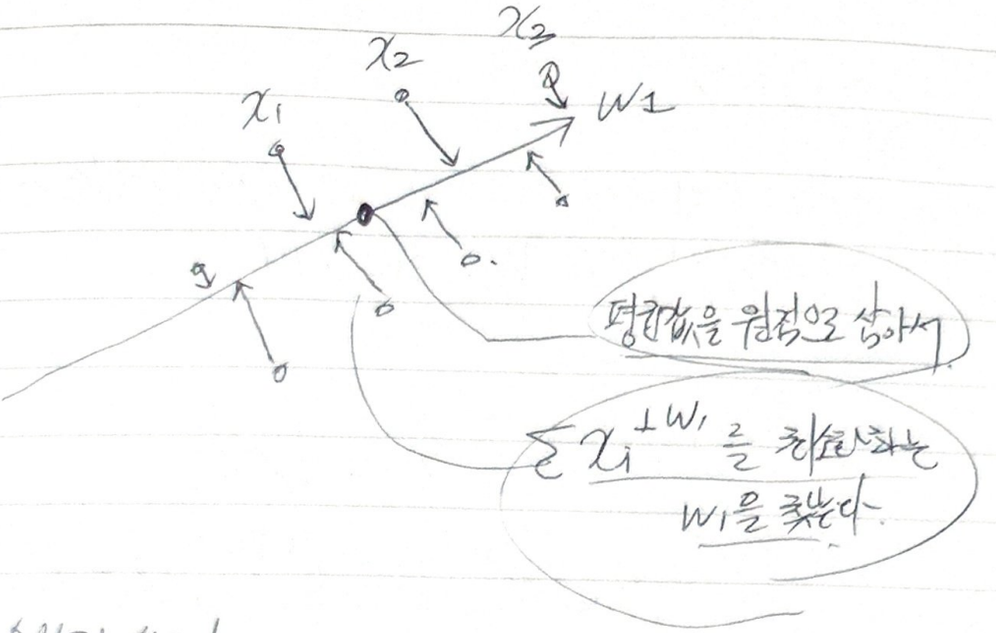
차원축소문제는 고차원 벡터를 더 낮은 차원의 벡터공간에 투영하는 문제로 생각하여 풀 수 있다. 즉, 투영해서 살펴볼 근우랭크 근사(low-rank approximation) 문제가 된다. 이 문제는 다음과 같이 서술할 수 있다.

> N개의 M차원 데이터 벡터 $x_1, x_2, \cdots, x_N$을 정규직교인 기저벡터 $w_1, w_2, \cdots, w_K$로 이루어진 K차원 벡터공간으로 투영하여 가장 비슷한 N개의 K차원 벡터 $x_1'', x_2'', \cdots, x_N''$을 만들기 위한 정규직교 기저벡터 $w_1, w_2, \cdots, w_K$를 찾는다.

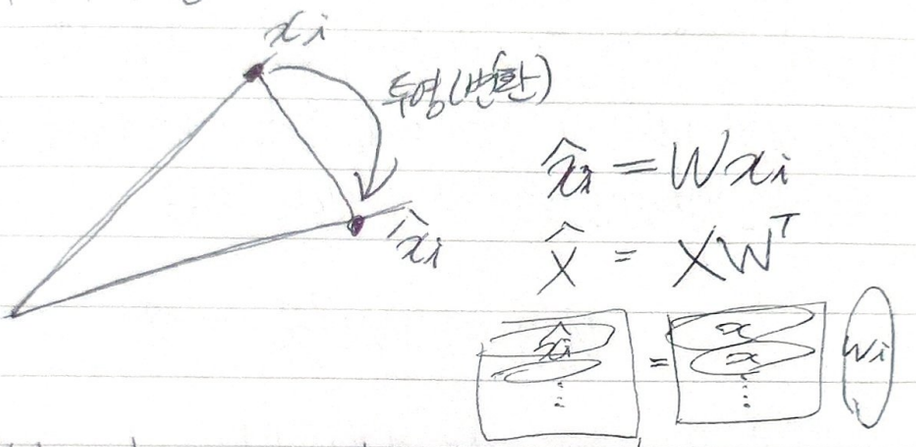
하지만, 위에서 근우랭크 근사 문제라 말하기 근사 성능을 높이기 위해 적어도 원점을 지나야 한다는 제한조건을 없애야 한다. 따라서 문제는 다음과 같이 바뀐다.

> N개의 M차원 벡터 $x_1, x_2, \cdots x_N$이 대해 어떤 상수 벡터 x_0을 뺀 데이터 벡터 $x_1-x_0, x_2-x_0, \cdots, x_N-x_0$을 정규직교인 기저벡터 $w_1, w_2, \cdots, w_K$로 이루어진 K차원 벡터공간으로 투영하여 가장 비슷한 N개의 K차원 벡터 $x_1'', x_2'', \cdots, x_N''$을 만들기 위한 정규직교 기저벡터 $w_1, w_2, \cdots, w_K$와 상수 x_0을 찾는다.

(참) x_0은 데이터벡터 $x_1, x_2, \cdots x_N$의 평균벡터이고, $w_1, w_2, \cdots w_K$는 가장 큰 K개의 특잇값에 대응하는 오른쪽 특이벡터 $v_1, v_2, \cdots v_K$이다.



PCA의 수학적 설명



- 데이터가 원점을 중심으로 조제화된 경우에는 벡터의 변환행렬을 곱하는 연산으로 투영벡터를 계산할 수 있다. (설명 편의를 위해 데이터가 원점에 모여있다고 가정)
- ~~데이터셋의~~ 다음처럼 데이터 x에 변환행렬 $W \in \mathbb{R}^{K \times M}$을 곱해서 새로운 데이터 $\hat{x}$을 구하는 연산을 생각하자.

$$\hat{x}_i = W x_i$$
$$x \in \mathbb{R}^M,\ W = \mathbb{R}^{K \times M},\ \hat{x} \in \mathbb{R}^K$$

모든 데이터 $x_i\ (i = 1, \dots, N)$에 대해 변환을 하면 벡터가 아닌 행렬로 표현 가능하다.

$$\hat{X} = X W^T$$
$$X \in \mathbb{R}^{N \times M},\ \hat{X} \in \mathbb{R}^{N \times K},\ W^T \in \mathbb{R}^{M \times K}$$

<PCA 특징>

차원축소 벡터 $\hat{x}_i$가 원래의 벡터 x_i가 지녔던 정보와 가장 유사하게 만드는 행렬 W를 찾는 것이다.

변환

$\hat{x}_i$는,

$$\hat{\tilde{x}} = U\hat{x}$$

$$\hat{x} \in R^k,\ U \in R^{M \times k},\ \hat{\tilde{x}} \in R^M$$

이렇게 변환과 역변환을 통해 원래의 차원을 되돌린 벡터 $U\hat{x}$는 원래의 벡터 x와 비슷할뿐 정확히 같지는 않다. 여기서 이 값을 다시 한번 차원축소 변환하면 $\hat{x}$가 된다. 즉,

$$W\hat{\tilde{x}} = WU\hat{x} = \hat{x}$$

따라서 W와 U는 다음 관계가 있다.

$$WU = I$$

특변환행렬 U를 알고 있다고 가정하고 역변환을 했을 때, 원래 벡터 x와 가장 비슷해지는 차원축소 벡터 $\hat{x}$를 다음과 같이 최적화를 이용하여 찾는다.

$$\arg\min_{\hat{x}} \|x - U\hat{x}\|^2$$

목적함수는 다음과 같이 바꿀수있다.

$$\|x - U\hat{x}\|^2 = (x - U\hat{x})^T(x - U\hat{x})$$
$$= x^Tx - \hat{x}^TU^Tx - x^TU\hat{x} + \hat{x}^TU^TU\hat{x}$$
$$= \boxed{x^Tx - 2x^TU\hat{x} + \hat{x}^T\hat{x}}$$

이 목적함수를 최소화하려면 $\hat{x}$로 미분한식이 0벡터가 되는 값을찾아야 한다. 이부분은 행렬의 미분과 관련한 부분에서 배우게된다. 위 목적함수를 미분한 식은 다음과 같다.

$$\boxed{-2U^Tx + 2\hat{x} = 0}$$

이 식을 정리하면

$$\hat{x} = U^Tx$$

가 된다. 우리에의 변환식

$$\hat{x} = Wx$$

과 비교하면,

$$U = W^T \quad (\because U^Tx = Wx)$$

임을 알수 있다. 따라서 다음식이 성립한다.

$$WW^T = I \quad (\because WU = I)$$

방금 문제는 원점의 변환행렬 W를 찾는 것이다. 이경우의 최적화 문제는 다음과 같이 된다.

$$\boxed{\arg\min_W \sum_{i=1}^{N} \|x_i - W^TWx_i\|^2}$$

모든 데이터에 대해 적용하면 목적함수는 다음처럼 된다.

$$\arg\min_{x} \|X - XW^TW\|^2$$

→ 이 문제는 rank-k 근사문제이고,
W는 제일 큰 k개의 특잇값에 대응하는
고윳값 특이벡터로 만들어진 행렬이다.

사이킷런의 PCA 기능

사이킷런의 decomposition 서브패키지는 PCA 분석을 위한 PCA 클래스를 제공한다.

input parameter

- n_components (정수)

method

- fit, transform() : 특징행렬을 낮은 차원의 근사행렬로
- inverse_transform() : 변환된 근사행렬을 원래 차원으로

attribute

- mean_ : 평균벡터
- components_ : 주성분벡터