Week 12. 주성분 분석

- 고차원의 데이터는 계산과 시각화가 어려워 분석하기가 쉽지 않다.
- 따라서 원래 데이터의 분포를 가능한 유지하면서 데이터의 차원을 줄이는 것이 필요하다.
  → 이를 차원 축소 (dimensionality dimension) 라 한다.

- 주성분 분석 (Principal Component Analysis)은 가장 널리 사용되는 차원축소 기법 중 하나로, 원 데이터의 분포를 최대한 보존하면서 고차원 공간의 데이터들을 저차원 공간 데이터로 변환한다.

12.1 차원축소

1974년 Motor Trend Magazine, 자동차 특성 기술하고 정량화하는 11개 변수.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

자동차 1대 → 11차원 벡터 정보 고차원 데이터로 계산, 시각화의 어렵다.
원 데이터의 분포를 가능한 유지하면서 데이터 차원을 줄이는 것이 필요하다. ⇒ 차원축소

? 몇몇 일부 변수의 데이터만 뽑아서 사용할수도 있지만 (feature selection)
변수들 사이에 어떠한 일정한 관계가 있는데 미리 알기 어려워, 원분포 보존을 보장할 수 없다.

12.2 주성분 분석 (PCA)

Principal Component Analysis

원 분포를 최대한 보존하면서, 고차원 공간의 데이터를 저차원 공간으로 변환한다.

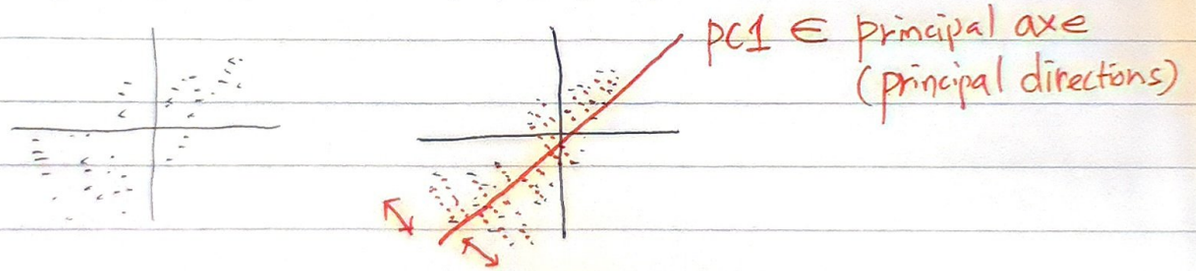
→ PCA는 기존 변수를 조합하여 서로 연관성이 없는 새로운 변수, 즉 주성분을 생성

→ 첫번째 주성분 PC1이 원데이터 분포를 가장 많이 보존하고, 두번째 PC2가 그 다음으로 원 데이터 분포를 많이 보존하는 식

→ 11차원 데이터를 PCA하여 11개의 주성분을 생성할 수 있으며, PC1, PC2, PC3이 원데이터의 분포(성질)의 약 90%를 보존한다면 10%의 손실을 감안해도 합리적 분석이 큰 무리가 없이 3차원으로 줄일수 있다.
~ 계산/시각화 용이!

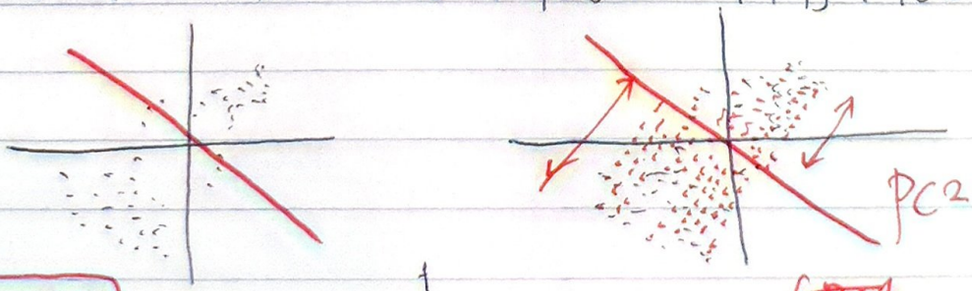
주성분을 구하기 위해 "새로운 축"을 찾아야 함 → 주축 (principal axes)
→ 주방향 (principal directions)

ex) 2차원 데이터의 첫번째 주축을 찾아 정사영
→ 정사영된 데이터 (성분)들이 PC1을 이루게된다.



ex2) 두번째 축 — 원데이터를 정사영할때, 새 데이터 (PC2) 분포가 PC1 다음으로 큰 축.

* PC1과 PC2가 서로 관계가 없도록, 첫번째 축 ⊥ 두번째 축 (수직)
⇒ 예상치 못한 변수 간 영향을 방지하기 위하여
⇒ 서로 관계가 있는 변수끼리는 각 주성분에 모이도록 하되, 각 주성분들(축) 사이 관계성 제거



PC1 ax = x축
PC2 ax = y축 ⇒

(~~PC3~~

PC1 ax가 PC2 ax보다 분포의 정보 크다.

12.3 주성분 분석의 계산

X = data matrix, $(n \times p)$ (p: # data features. 여기서는 확률변수 $x_1, x_2, \cdots x_p$의 갯수; n: #sample 갯수)

(통계학에서는 일반적으로 데이터 행렬 X와 구분하기 위하여, 확률변수 $X_1, X_2, \cdots$ 대신에 소문자 $x_1, x_2, \cdots$를 쓴다.)

* 데이터행렬 X의 (i,j) 성분 x_{ij}는 i번째 표본의 j번째 확률변수 x_j에 대한 하나의 데이터를 의미하고, j열 $X^{(j)}$는 확률변수 x_j가 갖는 모든 데이터를 의미한다.

확률변수 x_1 x_2 $\cdots$ x_p

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n \times p}$$

$X^{(1)}$, x_1의 모든 데이터

X를 센터링(centering, 확률변수의 평균을 0으로 조정)하여, $\tilde{X}$를 정의하자. 이를 위하여

① X의 각 열의 평균을 구한다.
확률변수 x_j의 평균은 $\overline{x_j}$로 표기한다.

$$\overline{x_1} = \frac{1}{n}(x_{11} + x_{21} + \cdots x_{n1})$$

② 각 열별로 데이터에서 열의 평균 $\overline{x_j}$을 뺀다. 이 행렬은 센터링된 행렬 $\tilde{X}$과 한다.

$$\tilde{X} = X - \begin{bmatrix} \overline{x_1} & \cdots & \overline{x_p} \\ \vdots & & \vdots \\ \overline{x_1} & & \overline{x_p} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} - \begin{bmatrix} \overline{x_1} & \overline{x_2} & \cdots & \overline{x_p} \\ \vdots & \vdots & & \vdots \\ \overline{x_1} & \overline{x_2} & \cdots & \overline{x_p} \end{bmatrix}$$

여기서 센터링된 행렬 $\tilde{X}$의 각 열의 평균은 0이 된다. → **mean-centered matrix**

앞으로 주어진 행렬 X는 '이미 센터링이 된' 행렬이라고 가정하자.

③ X의 특잇값 분해(SVD)를 구한다.

$$X = [u_1\ u_2 \cdots u_k\ u_{k+1} \cdots u_n] \begin{bmatrix} s_1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_k & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & s_{k+1} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & s_p \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_p^T \end{bmatrix}$$

($\Sigma_1 = \mathrm{diag}(s_1, \cdots s_p)$)

U, V — 직교행렬 ($U \perp V$)
S_1 = 크기순 배열된 특잇값 $s_1 \geq s_2 \geq \cdots \geq 0$을 주대각선 성분으로 하는 $p \times p$ 대각선행렬 $S_1 = \mathrm{diag}(s_1, \cdots s_p)$

$$= USV^T = U \begin{bmatrix} S_1 \\ 0 \end{bmatrix} V^T = \sum_{i=1}^{p} s_i u_i v_i^T$$

④ V의 열벡터 $\begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_p^T \end{bmatrix}$ 가 주축 (principal axes)이 된다.
→ p개만큼의 주축이 생성 가능하다

⑤ $Z = US$의 열벡터들이 원 데이터를 주축에 정사영하여 얻어진 주성분 점수 (principal component score, PC score)가 된다.

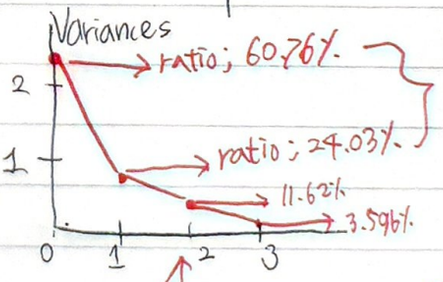
⑥ 특잇값 S_i를 이용하여 계산한 $\frac{S_i^2}{n}$은 i번째 PC의 분산이 되고,
$\frac{S_i^2}{S_1^2 + S_2^2 + \cdots + S_p^2}$은 i번째 PC가 원 데이터의 분포를 보존하는 비율이다.
각 PC별로 원 데이터의 분포를 보존하는 비율을 계산하여, 몇 차원으로 축소할지 결정한다.

⑦ 데이터를 p차원에서 k ($\ll p$)차원으로 줄이기 위하여, U의 처음 k개의 열벡터 (U_k)와 S의 k번째 선행 주 부분행렬 (leading principal submatrix)(S_k)을 택하면, $U_k S_k$는 처음 k개의 PC를 포함하는 $n \times k$ 행렬이 된다.

$$Z_k = U_k S_k = [u_1, u_2 \cdots u_k] \begin{bmatrix} s_1 & & \\ & s_2 & \\ & & \ddots \\ & & & s_k \end{bmatrix}$$

12.4 주성분 분석 사례 ; 4차원 데이터 → 2차원 데이터

```
// 센터링 진행                          X.ncols()
for col in range(p):
    v = np.array(X.column(col))
    m = mean(v)
    st = std(v)                        X.nrows()
    for row in range(n):
        X_ctr[row,col] = (X[row,col] - m) / st
```



PC1, PC2 분산비율 합 84.79%
=
PC1, PC2 원 분포 보존율 84.79%

특잇값분해
대각성분 특잇값 → PC 분산
i번째 PC가 원 분포 보존 비율
p개의 주성분 열 행렬

```
U, S, V = X_ctr.SVD()
Var_PC = [(S[i,i]^2/n).n(digits=4) for i in range(p)]
Prop_Var = [100*Var_PC[i]/sum(Var_PC) for i in range(p)]
PC_score = U*S
```

12.5 주성분 분석과 공분산행렬

주성분 분석은 원 데이터의 분포를 최대한 보존하면서 고차원 데이터 → 저차원 변환.
But 원 데이터 분포에 대한 정보는 공분산 행렬에 담겨 있다. 왜냐하면 공분산 행렬은 주대각선 상의 성분에 각 확률변수가 얼마나 퍼져 있는지를 나타내는 분산과 주대각선 이외의 성분에 확률변수 간 상관관계를 나타내는 공분산으로 되어 있기 때문이다.

데이터 행렬 X가 센터링 (mean-centered) 된 행렬이면, $p \times p$ 공분산행렬 Σ은 다음과 같이 계산한다.

$$\Sigma = \begin{bmatrix} Var(x_1) & Cov(x_1,x_2) & \cdots & Cov(x_1,x_p) \\ Cov(x_2,x_1) & Var(x_2) & \cdots & Cov(x_2,x_p) \\ \vdots & \vdots & & \vdots \\ Cov(x_p,x_1) & Cov(x_p,x_2) & \cdots & Var(x_p) \end{bmatrix}$$

$$= \frac{1}{n} \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1}^2 & \frac{1}{n}\sum_{i=1}^{n} x_{i1}x_{i2} & \cdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2}x_{i1} & \frac{1}{n}\sum_{i=1}^{n} x_{i2}^2 & \cdots \\ \vdots & & \frac{1}{n}\sum_{i=1}^{n} x_{ip}^2 \end{bmatrix}$$

비대각성분 = 각 확률변수 간 공분산 (상관관계)

대각성분 = 각 확률변수 x_{np}의 분산

$$= \frac{1}{n} X^T X$$

주성분 분석의 목표: "공분산행렬로부터 얻은 정보"를 최대한 보존하는 "더 적은 개수의 새 변수들" 찾기

- $X = USV^T = U\begin{bmatrix} S_1 \\ 0 \end{bmatrix} V^T$ 로부터 다음 관계를 얻는다.

공분산 행렬 $\Sigma = \frac{1}{n} X^T X = \frac{1}{n} V[S_1^T \; 0] U^T U \begin{bmatrix} S_1 \\ 0 \end{bmatrix} V^T = V\left(\frac{S_1^2}{n}\right) V^T$

(X^T, X)

따라서 $\Sigma V = V\left(\frac{S_1^2}{n}\right)$, 즉 $\Sigma V_i = \left(\frac{S_i^2}{n}\right) V_i \; (i=1,2,\cdots,p)$ 관계가 성립한다.

$= \lambda_i \cdot V_i$ = 고유값, 고유벡터

- $\lambda_i = \frac{S_i^2}{n}$ → Σ의 고유값 (eigenvalue)
- V_i → Σ의 고유벡터
  - → i번째 축 (Principal Axe)
- i번째 PC의 분산

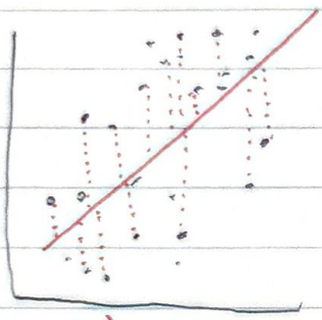
12.6 주성분 분석과 선형회귀

(관점1) 선형회귀에서는 각 데이터 (x_i, y_i)와 일차함수 $y = a + bx$에 의해 계산된 점 $(x_i, \hat{y}_i)$ 사이의 거리 (즉 y축 상에서의 직선거리)가 최소가 되도록 a, b를 결정

(관점2) 주성분 분석에서는 각 데이터 (x_i, y_i)에서 일차함수 $y = a + bx$에 이르는 거리가 최소가 되도록 a와 b를 결정 ··· 그래야 가장 넓은 분포를 얻을 수 있다.

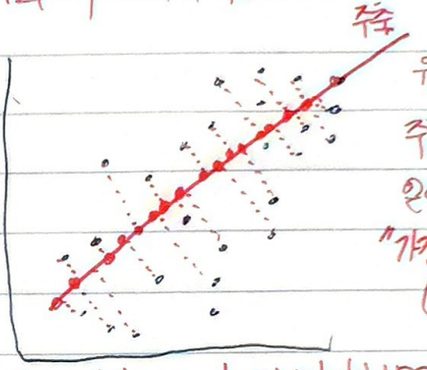
⇒ PCA로 얻어진 직선이 데이터와 더 가까이 있다.

그림1



$\min(y - \hat{y})$

그림2



원 데이터를 주축에 정사영할 때, 얻어진 분포가 "가장 퍼져있도록" (∵ 각각의 분산이 커)

* PCA는 data scale에 민감하므로 mean-centered, normalization (정규화)를 반드시 가정해야 한다.