2018년 자연과학 체험캠프 : 수리통계

차원축소를 통한 빅데이터분석

서울대학교 통계학과
정성규

최근 화두

- 인공지능, 딥러닝, 4차 산업혁명, 데이터 사이언스
- 모두 데이터(자료)와 관련이 있음

데이터를 다루는 학문?

- 통계학-데이터(자료)를 수집하고, 분석하고, 분석한 결과를 통해 의사결정을 내리는 과정을 다루는 학문
- 컴퓨터 사이언스-컴퓨터를 만들거나 사용하는 것과 관련한 이론, 실험, 공학을 다루는 학문

  - 출처: 위키피디아

- 데이터 사이언스- 자료로 부터 의미 있는 정보 또는 통찰력을 얻기 위한 방법론, 시스템을 다루는 분야
  - 자료가 관측된 분야의 지식도 필요 – Interdisciplinary Science
- 통계학의 역할?
- 수학의 역할?

빅 데이터란?

- 지구상에 현존하는 자료의 90%는 지난 2년 사이에 생성됨
- 자료의 단위 : gigabytes (1000^3), terabytes, petabytes, exabytes, zettabytes, yottabytes
- 매일, $2.5x1000^6$ (2.5 exabytes) 의 자료생성

- 자료의 형태
  - 전통적 형태 : 숫자
  - 새로운 형태 : 비디오, 오디오, 페이스북, 트위터
- 자료의 속도
  - 주기적 $->$ 실시간
- 빅데이터의 유형
  - 대용량
  - 고차원

빅데이터 예시

- 이미지 데이터
- 이미지 한개: $250 \times 250 = 62,500$ 픽셀
- 칼라 이미지: RGB $250 \times 250 \times 3 = 187,500$ 개의 값
- 1000명: $150 \times 180 \times 3 \times 1000 = 187,500,000$ 개의 값

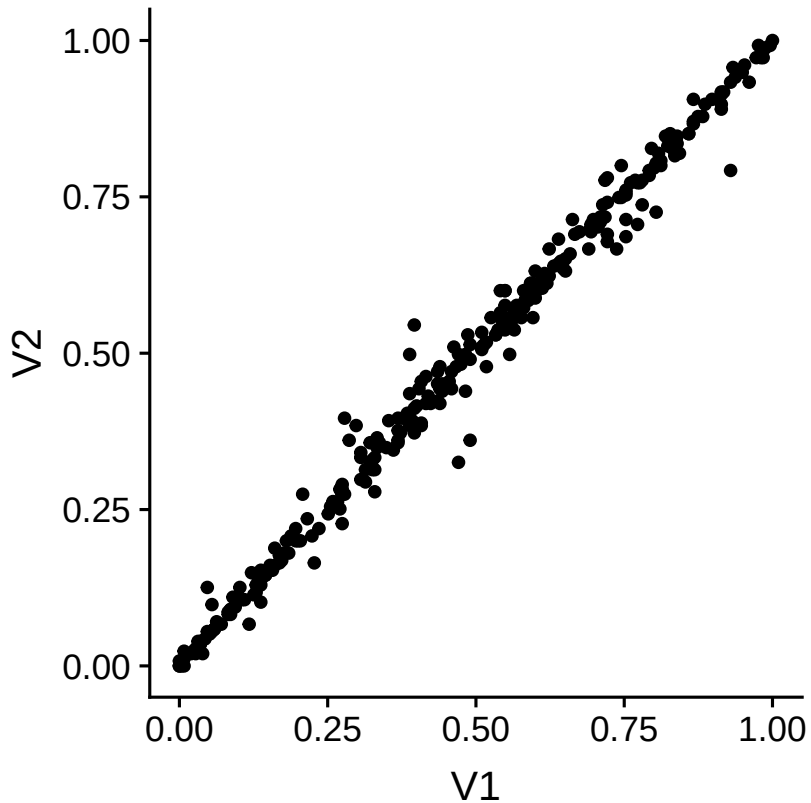
데이터의 이해

- 1개의 칼라 이미지 $=250 \times 250 \times 3$ 개의 변수값
- 1000 개의 이미지
- 한 이미지의 변수값을 가로로 늘여놓고,
- 각 이미지를 세로로 나열하면,
- 크기 $1000 \times (187500)$의 데이터 행렬이 된다.
- 아래는 처음 5개의 이미지와, 각 이미지 변수 중 첫 4개만을 본 데이터 행렬(의 일부)이다.

```
## # A tibble: 5 x 5
##   names                V1      V2      V3      V4
##   <chr>             <dbl>   <dbl>   <dbl>   <dbl>
## 1 Coco_dEste      0.0863  0.0824  0.0784  0.0745
## 2 Cole_Chapman    0.102   0.106   0.110   0.110
## 3 Coleen_Rowley   0.325   0.329   0.345   0.369
## 4 Colin_Campbell  0.616   0.627   0.631   0.631
## 5 Colin_Cowie     0.722   0.780   0.824   0.847
```

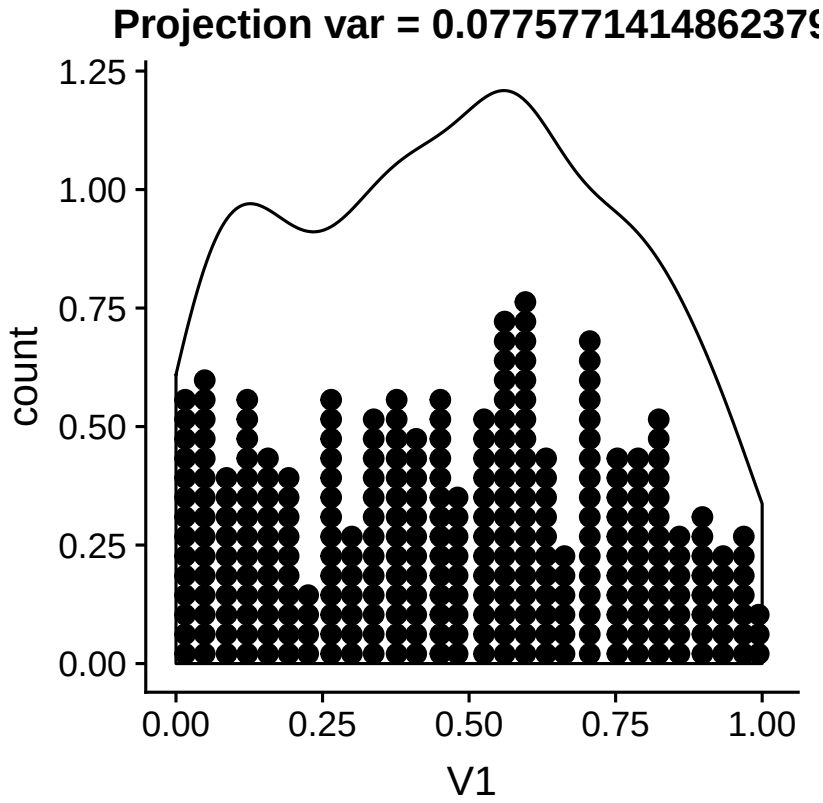
데이터의 기하적 이해

- 편의를 위해 첫 두 개의 변수만을 본다
- 한 이미지에서 나온 두 변수의 값을 (x,y) 쌍으로 이해하면, 2 차원 평면에서의 한 점이다.
- 이미지의 갯수가 1000개이므로 1000개의 점이 자료를 나타낸다.



데이터의 기하적 이해

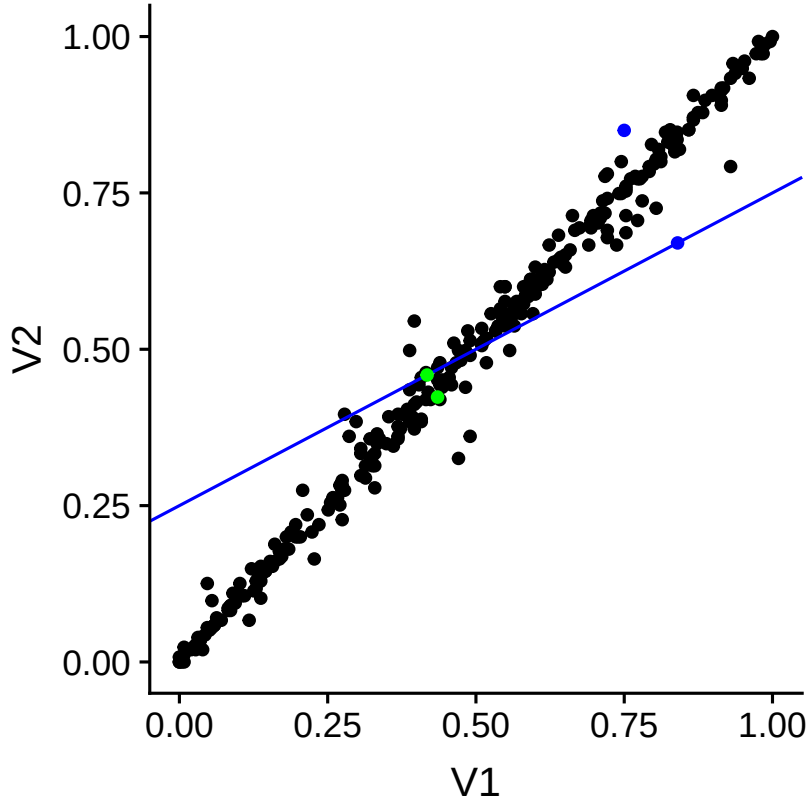
- 위 산점도에서 각 점을 수직으로 사영시켜 쌓으면 아래와 같은 점도표가 된다.
- 첫번째 변수(V1)값의 위치(평균)와 산포(분산)를 가늠해보자.



평균 $= \frac{1}{n}\sum_{i=1}^{n} x_i$. 분산 $= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

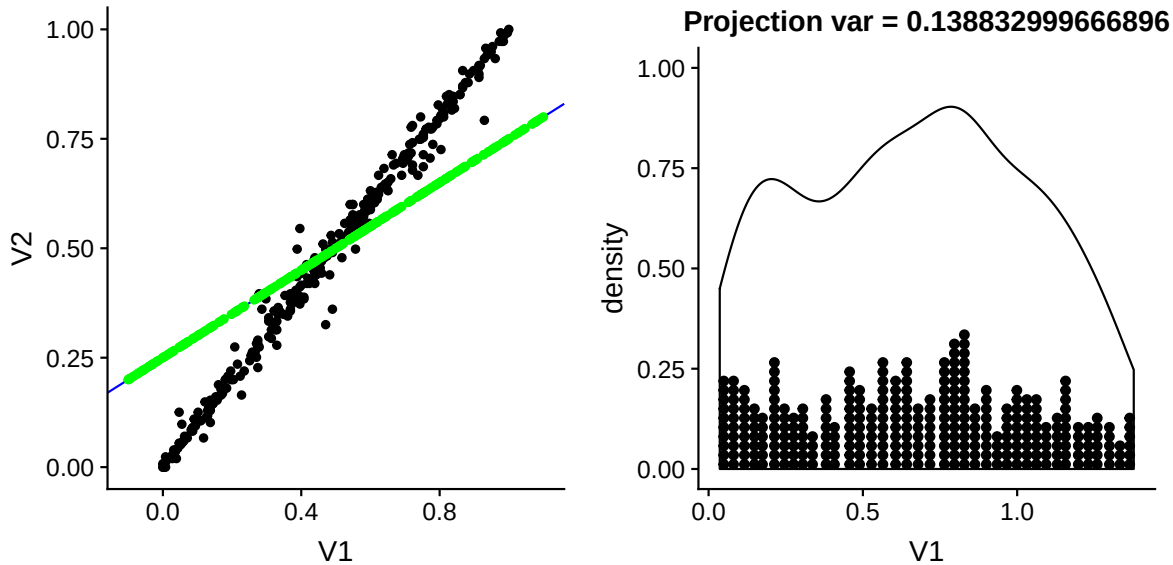
사영한 자료 (1/2)

- 위 산점도에서 각 점을 (1/2,1/2) 점을 지나면서 기울기가 1/2 인 직선으로 직교사영.



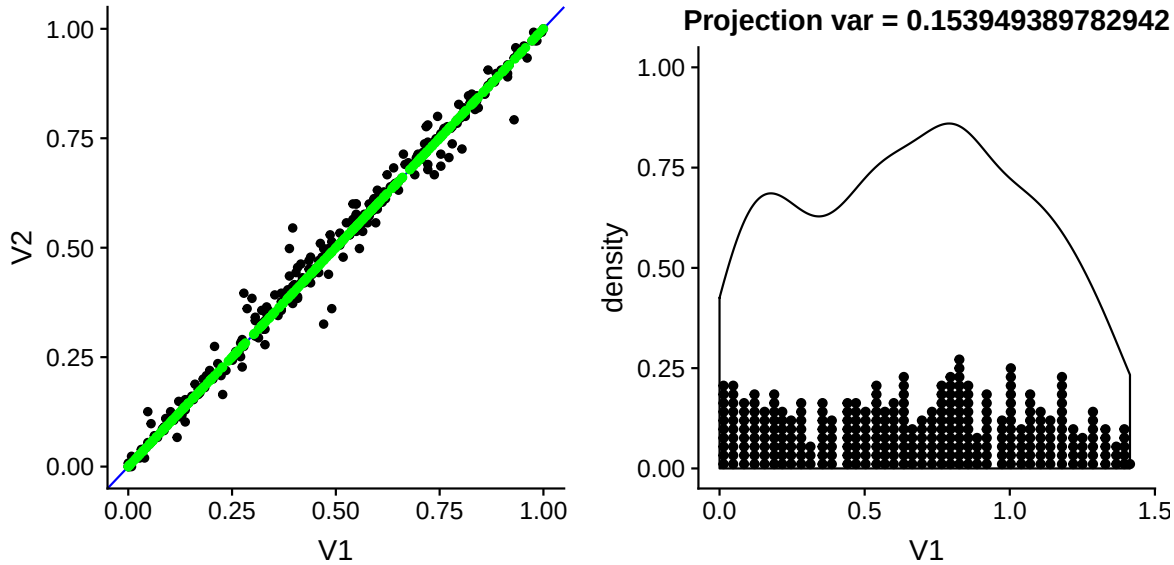
사영한 자료 (2/2)

- 위 산점도에서 각 점을 (1/2,1/2) 점을 지나면서 기울기가 1/2 인 직선으로 직교사영.



사영의 방향?

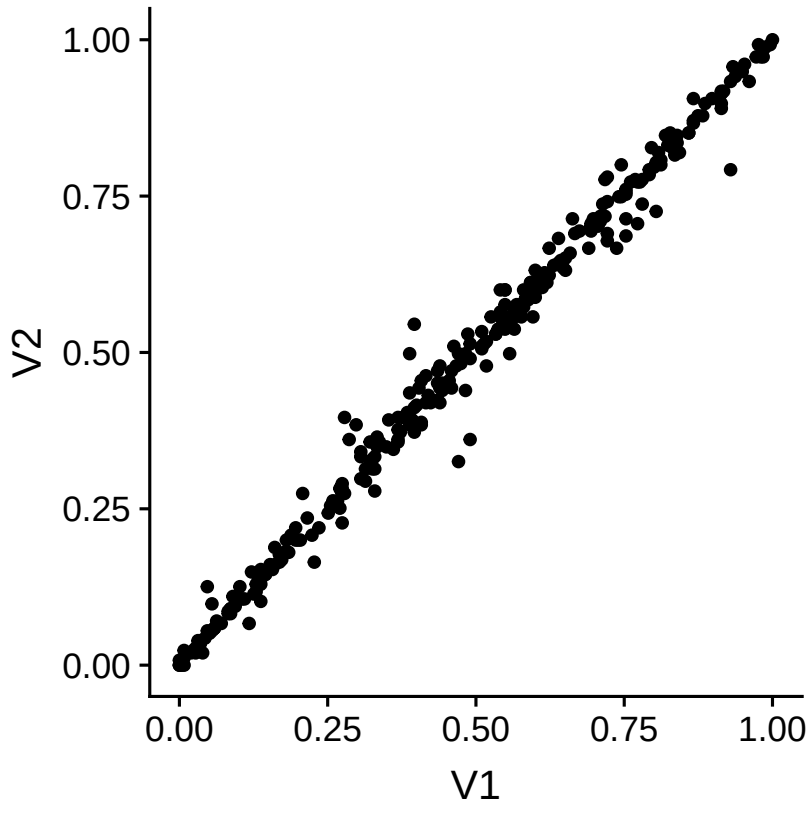
- 위 산점도에서 각 점을 (1/2,1/2) 점을 지나면서 기울기가 1인 직선으로 직교사영.



사영된 값들의 분산이 가장 큰 "직선"은?

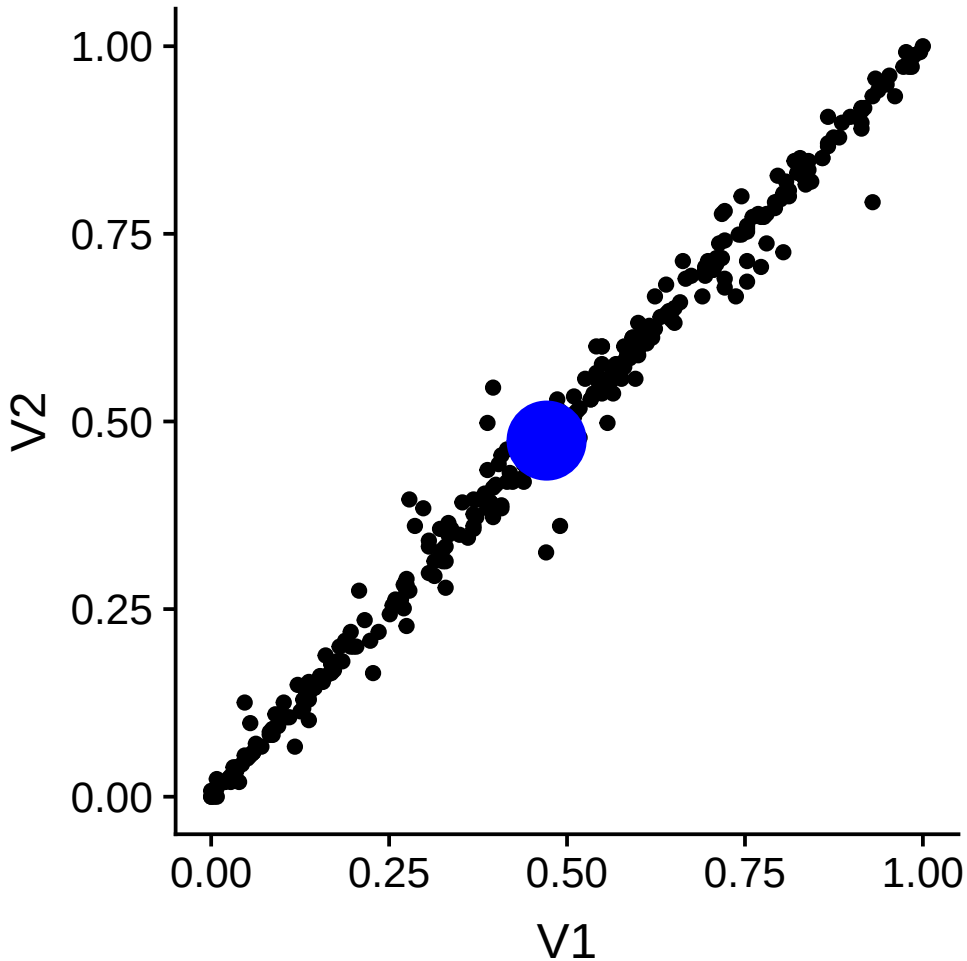
차원축소

- 위 이미지 데이터는 변수가 187500개가 있으므로, 각 이미지가 187500차원공간의 점.
- 차원축소: 정보를 많이 잃어버리지 않으면서 데이터의 차원을 줄여서 분석
- "주성분분석" 이용



차원축소: 0차원?

- 만약 0차원으로 줄인다면? 점!
- 이 경우, (V1의 평균, V2의 평균)
  $= (\bar{x}, \bar{y}) = (\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i)$



0차원 차원축소: 이미지 데이터

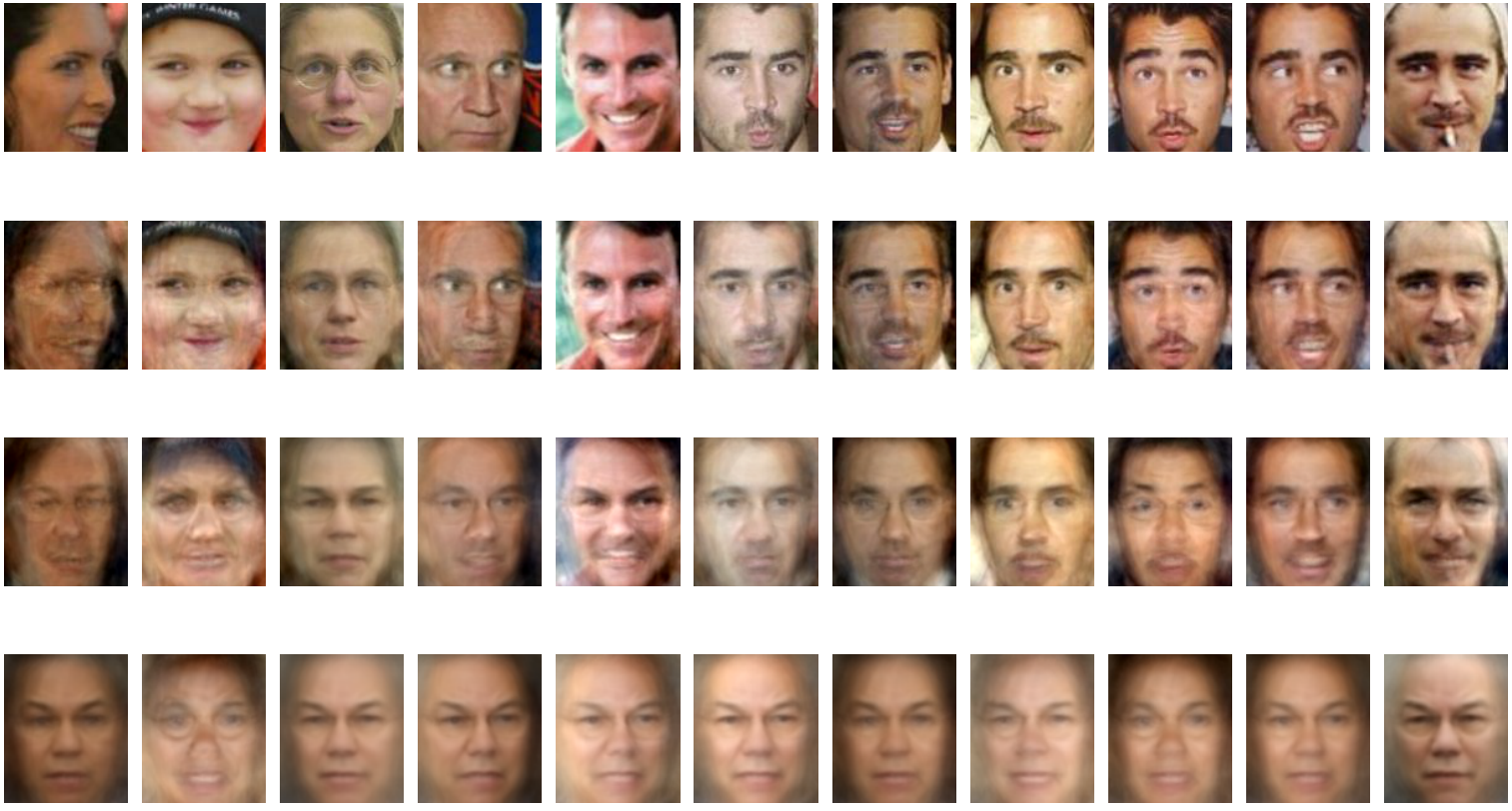
- 만약 0차원으로 줄인다면? 점!
- 이 경우, 각 187500 개 변수마다 관측값들의 평균으로 이루어진 “평균 이미지"

주성분 분석 (Principal Component Analysis, PCA)

- 여러 개의 다변량 양적변수를 선형결합으로 표시되는 중요한 몇개의 주성분으로 표현하여 전체의 변동을 설명하는 것
- 앞의 얼굴 이미지 데이터는 187500차원 점이지만, 차원축소를 하여 200개 차원으로 충분한 정보가 제공이 될까?

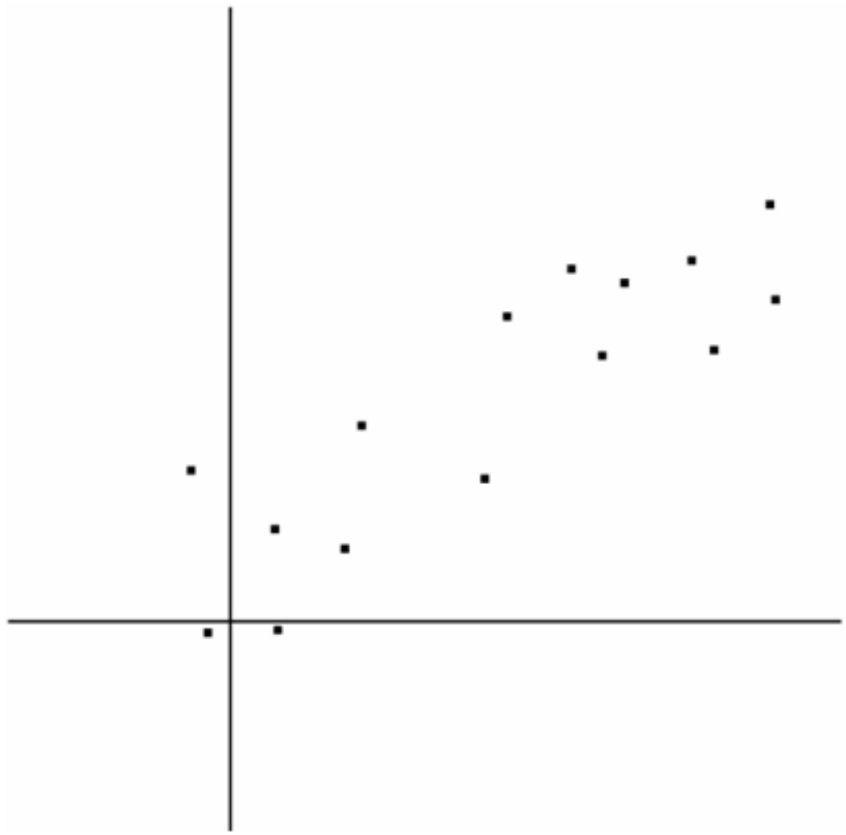
주성분 분석

원자료, 주성분 100개, 주성분 30개, 주성분 5개.



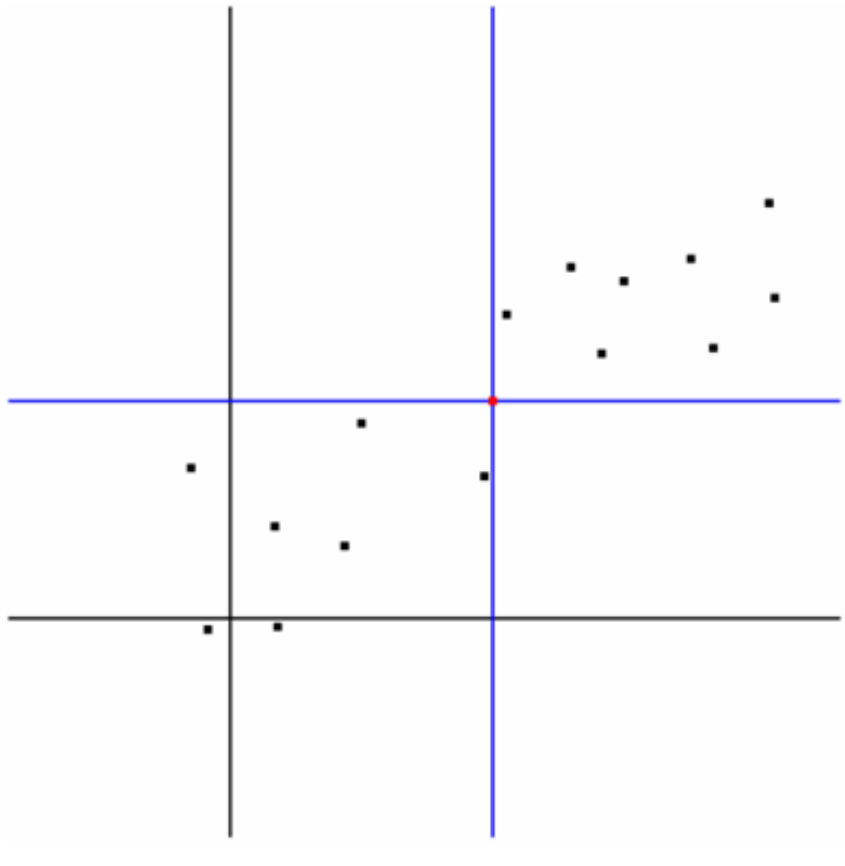
주성분 분석의 개념

(x_i, y_i)



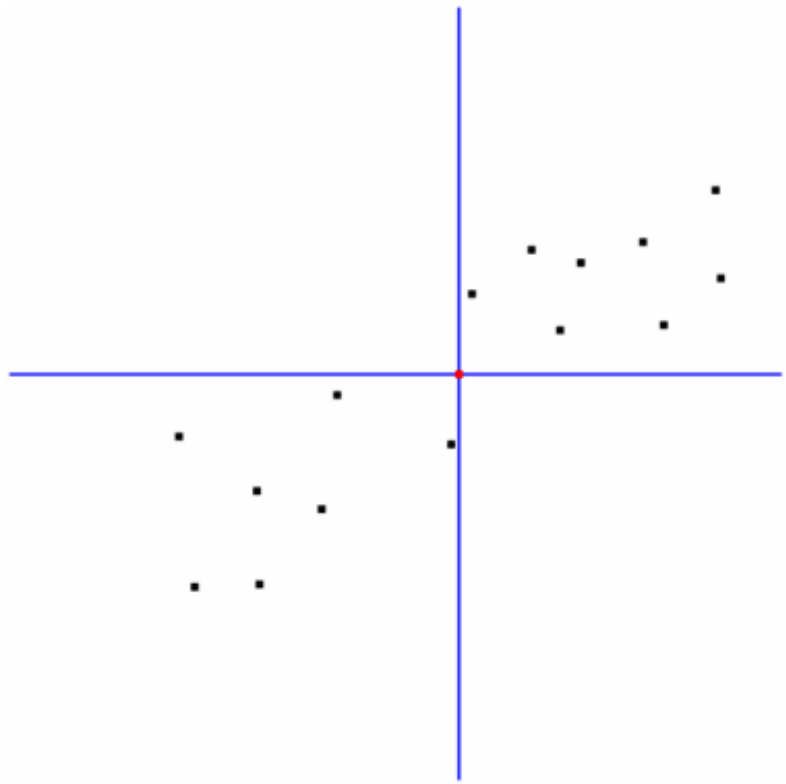
Step 1 : 주어진 자료를

$$(x_i - \bar{x}, y_i - \bar{y})$$



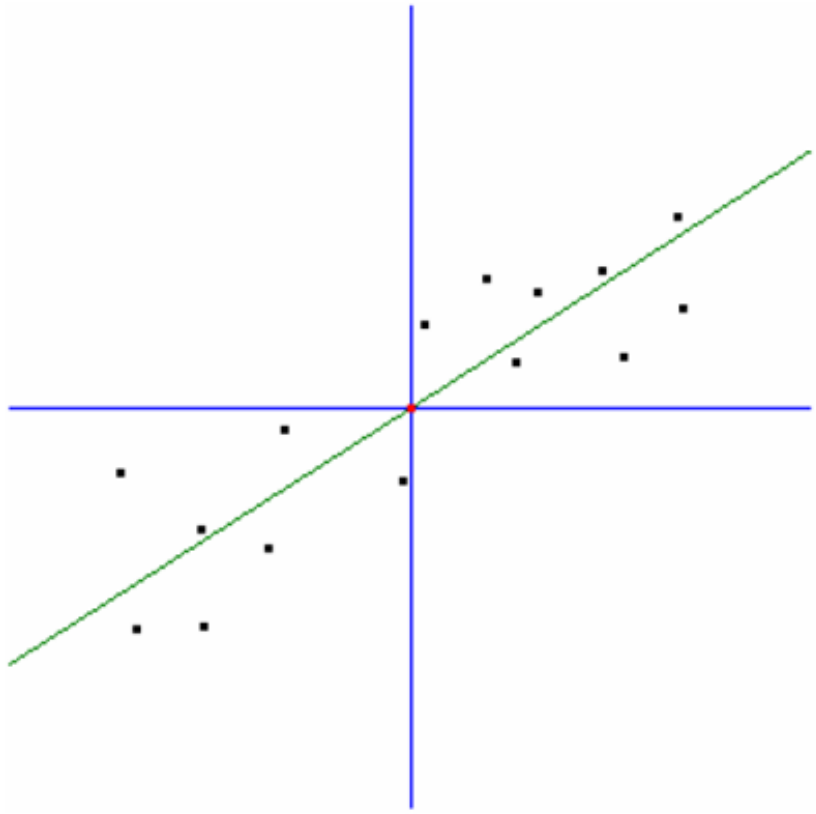
Step 2 : 좌표를 중심으로 이동시킨다.

$$(x_i', y_i') = (x_i - \bar{x}, y_i - \bar{y})$$



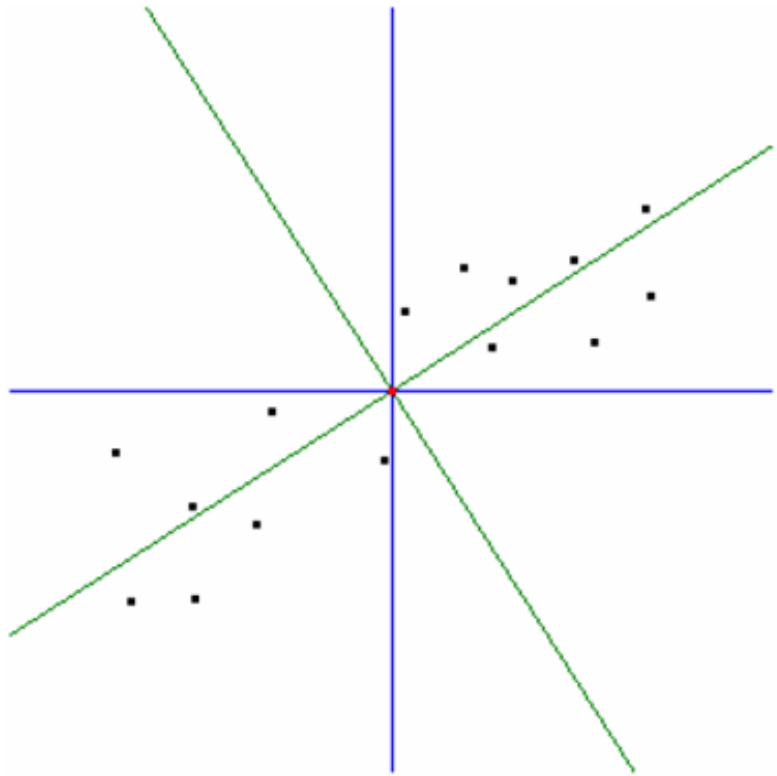
Step 3 : 이것을 새로운 좌표로 이용한다.

$$a_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$$

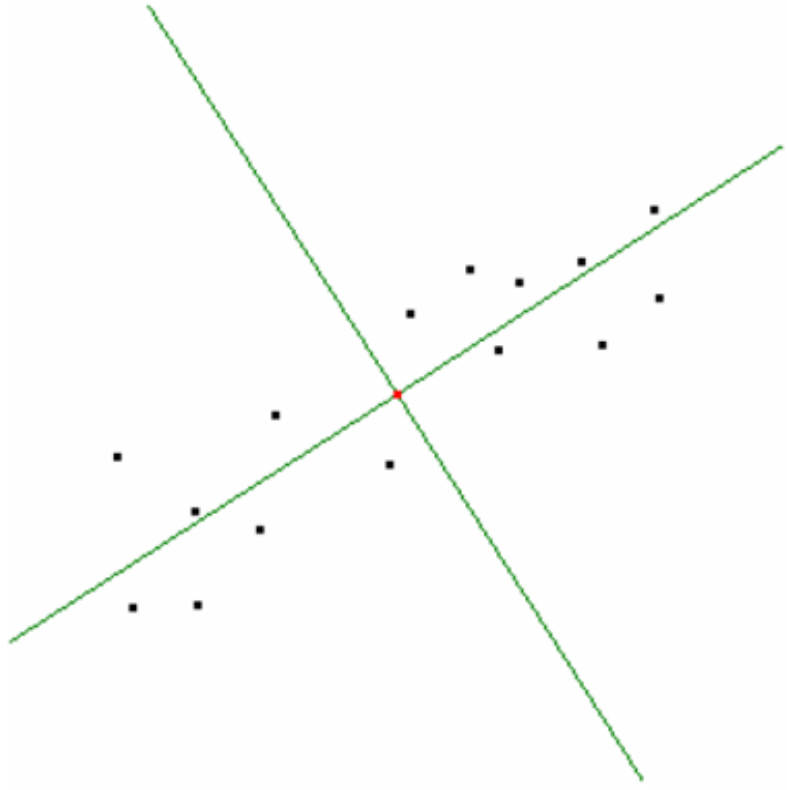


Step 4 : 산포가 최대가 되는 방향을 찾는다.

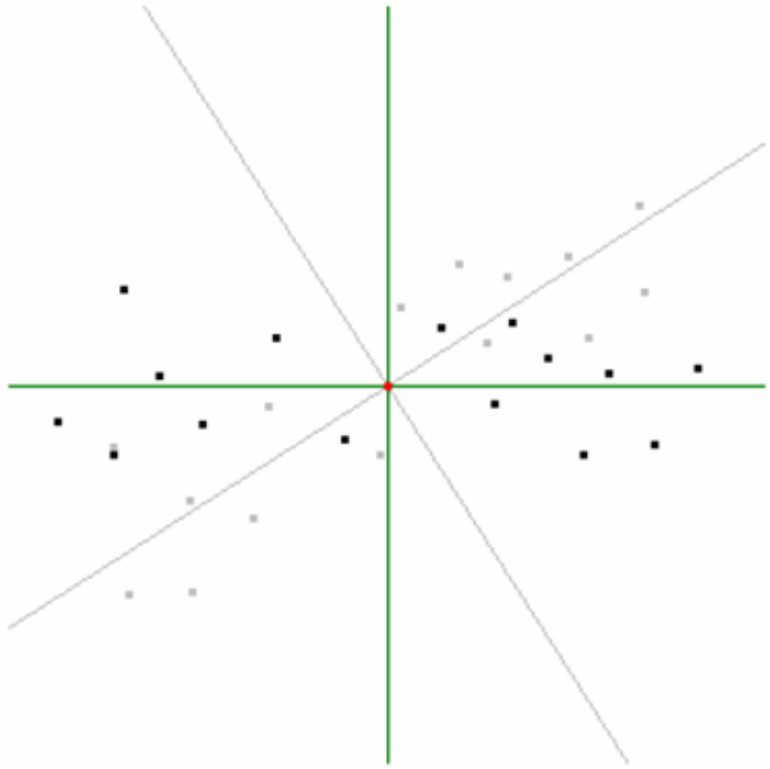
$$a_2 = (\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})$$



Step 5 : 다음의 수직방향을 찾는다.

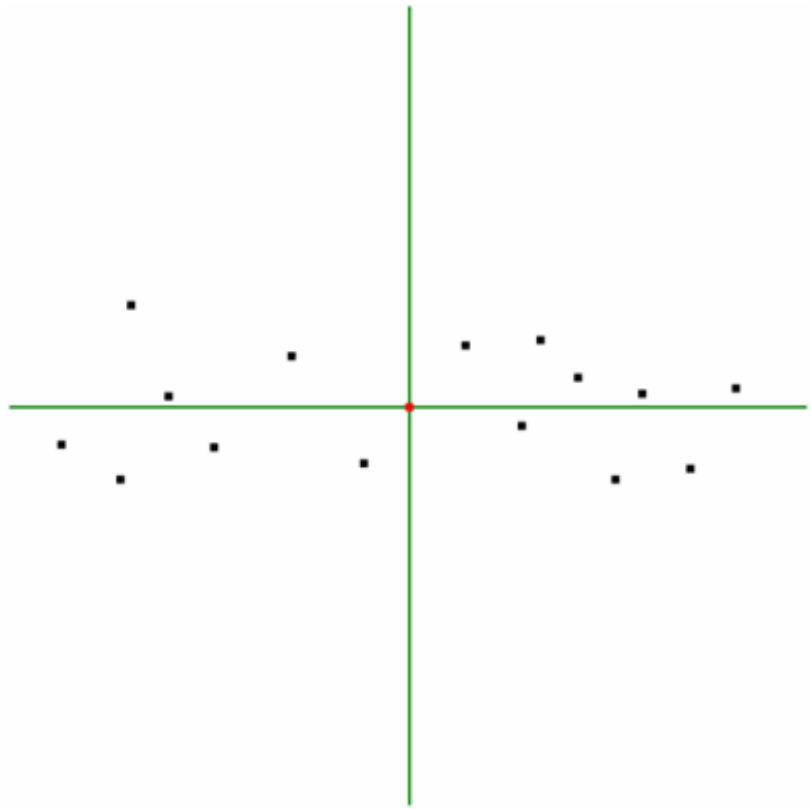


Step 6 : 회전된 축만 남겨 논다.

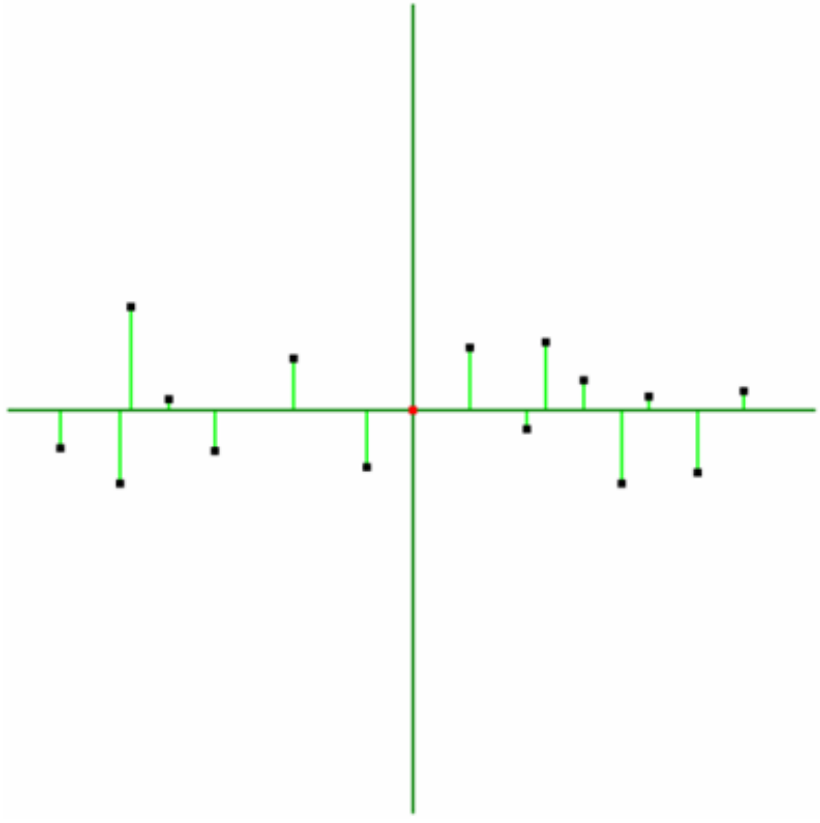


Step 7 : 축을 회전시켜 원래 자리로 이동한다.

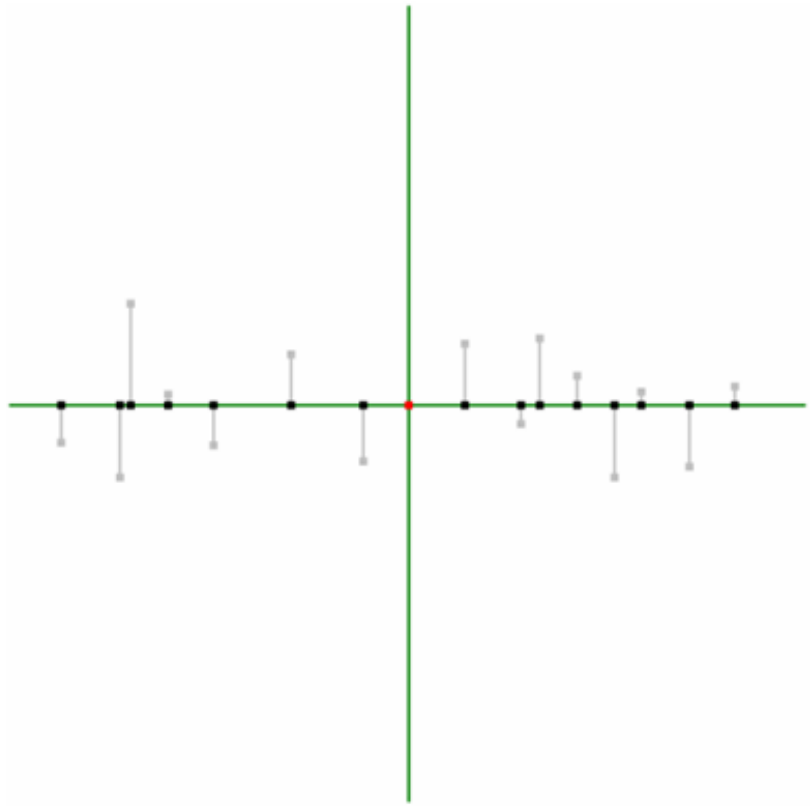
$$(z'_{1i}, z'_{2i}) = \left(\frac{1}{\sqrt{2}} x'_i + \frac{1}{\sqrt{2}} y'_i, \frac{1}{\sqrt{2}} x'_i - \frac{1}{\sqrt{2}} y'_i \right)$$



Step 8 : 새로운 좌표계 완성. X축 분산이 매우 (제일) 크다.



Step 9 : 자료를 X축으로 내려서 차원 축소 가능



Step 10 : 차원축소 완성 (1차원으로 차원 축소)

주성분의 정의 및 계산

주성분의 정의

- 첫 번째 주성분 : 변수(X_j)들의 전체 분산 중 가장 큰 부분을 설명할 수 있는 선형결합

$$C_1 = \boldsymbol{a}_1' X = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p$$

- 두 번째 주성분 : 첫 번째 주성분과 독립이면서 첫 번째 주성분에 의해 설명되지 않는 잔여 분산을 최대한 설명하는 선형결합

$$C_2 = \boldsymbol{a}_2' X = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p$$

- 세 번째 주성분 : $\cdots$
- 일반적으로 주성분의 수는 변수의 수만큼 유도가능하나, 자료 요약을 위해 전체 분산의 대부분을 설명할 수 있는 소수의 주성분만을 고려한다.

첫 번째 주성분의 정의

- 각 관측값은 $(X_1, \ldots, X_p)$의 좌표를 가진 p차원 공간의 점.
- 변수 $X_1, \ldots, X_p$의 **선형 결합**은 계수 $a_{11}, \ldots, a_{p1}$을 이용한 $C_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p$.

  선형? 두 미지수 C_1과 X_1의 관계가 선형 (linear).
- 각 관측값을 $X = (X_1, \ldots, X_p)$ 라는 벡터로, 계수들의 벡터를 $\boldsymbol{a}_1 = a_{11}, \ldots, a_{p1}$로 쓸 때, C_1은 X를 $\boldsymbol{a}_1$ **방향**으로 사영한 값

$$C_1 = \boldsymbol{a}_1' X$$

  .
- $\boldsymbol{a}_1' \boldsymbol{a}_1 = 1$이라는 조건하에서 $C_1 = \boldsymbol{a}_1' X$의 분산을 최대화하는 선형결합이 첫번째 주성분.

주성분의 성질

- 주성분은 **주성분방향** $\boldsymbol{a}_1 = (a_{11}, \ldots, a_{p1})$으로 정의됨.
- 관측값들을 주성분방향에 사영한 값을 **주성분점수**이라고 함.
- 주성분점수의 **분산**이 가장 큰 방향이 첫번째 주성분 방향.
- 첫번째 주성분 방향과 직교하면서, 주성분점수의 분산을 가장 크게 해주는 방향이 두번째 주성분 방향.

주성분분석의 계산

- 기하최적화 문제인 주성분분석의 해답은 행렬대수를 이용하면 쉽게 얻을 수 있다.

행렬대수

특이값 분해 크기가 $n \times p$인 행렬 X를 세 행렬의 곱으로 표현

$$X = UDV^T.$$

- U : 왼쪽 특이벡터들의 행렬
- D : 특이값 행렬
- V : 오른쪽 특이벡터들의 행렬

특이값 분해와 주성분분석

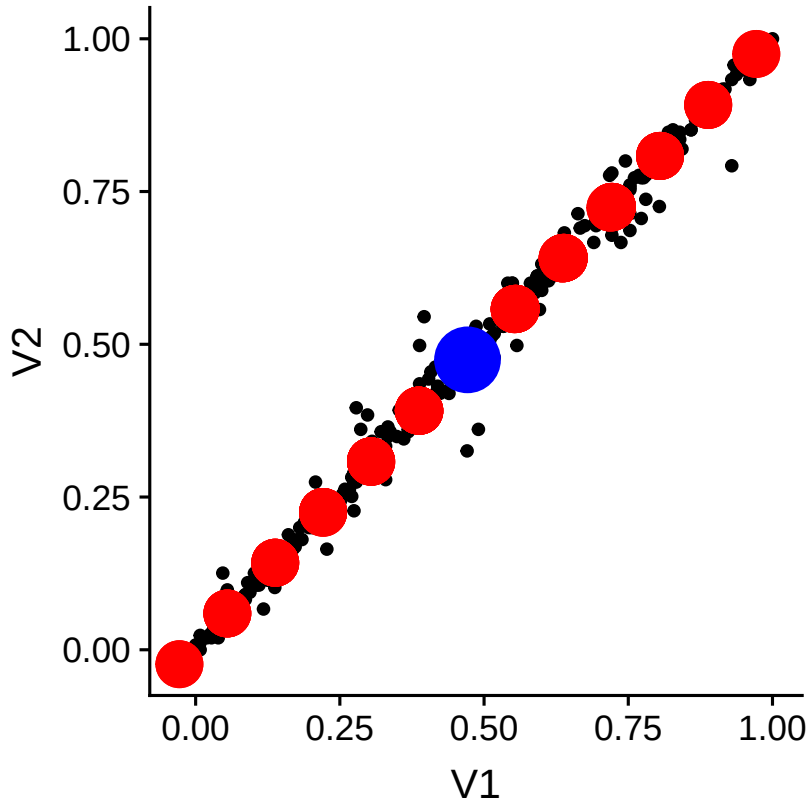
자료행렬 X에서 평균을 빼고 난 뒤,

$$X = UDV^T.$$

- U : 주성분 점수로 이루어진 행렬
- D : 주성분 분산(의 제곱근)으로 이루어진 행렬
- V : 주성분 방향벡터들로 이루어진 행렬

주성분의 해석

- 주성분방향은 자료의 주요한 변동을 설명

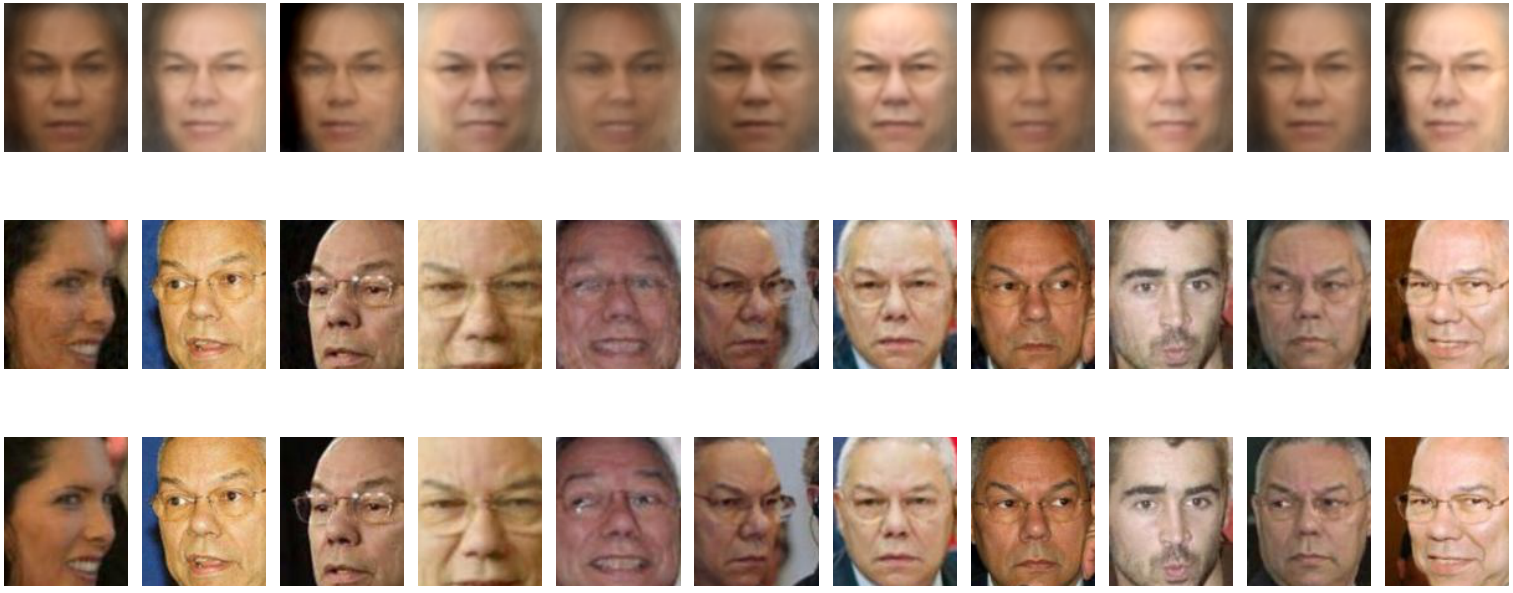


주성분의 해석


PC1: −3sd
PC1: −2sd
PC1: −1sd
PC1: 0sd
PC1: 1sd
PC1: 2sd
PC1: 3sd
PC2: −3sd
PC2: −2sd
PC2: −1sd
PC2: 0sd
PC2: 1sd
PC2: 2sd
PC2: 3sd
PC3: −3sd
PC3: −2sd
PC3: −1sd
PC3: 0sd
PC3: 1sd
PC3: 2sd
PC3: 3sd

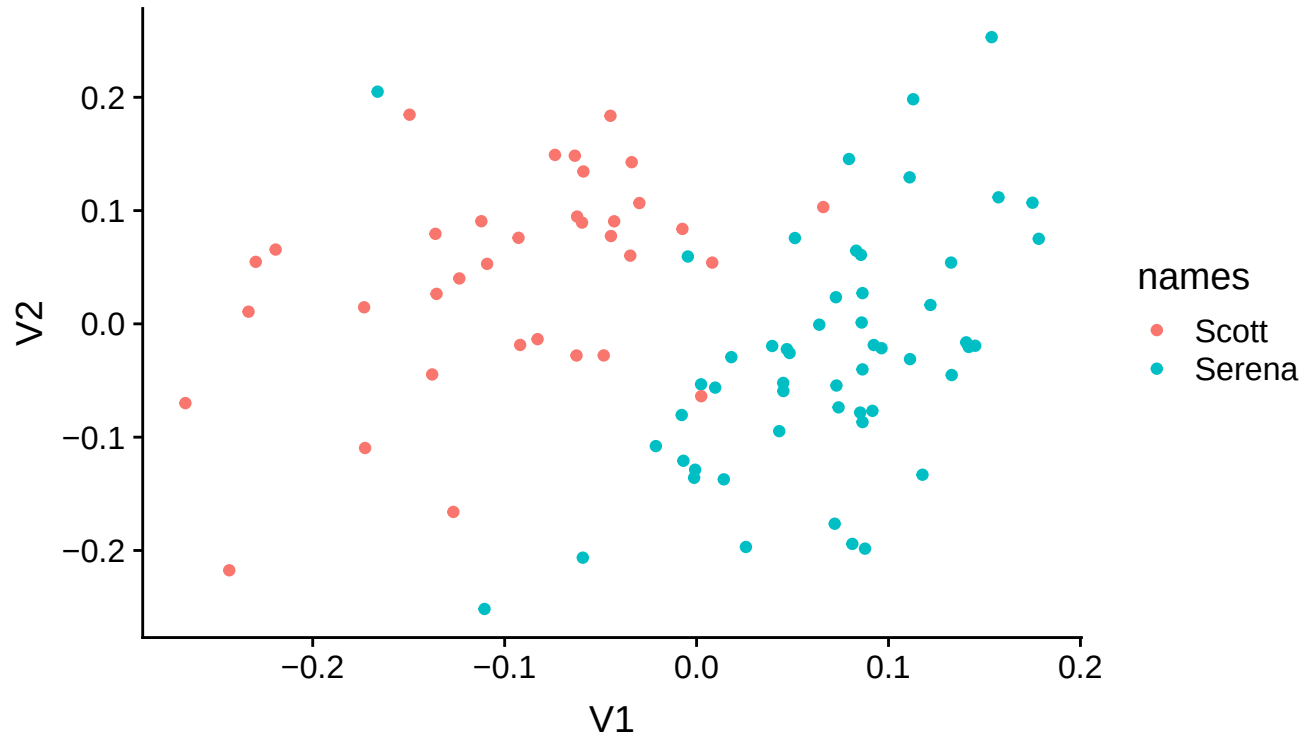
주성분 분석을 이용한 차원축소

- 임의의 벡터 X는 기저벡터 (basis vector)의 선형결합으로 표현된다
- 앞에서 구한 고유벡터들을 기저벡터로 쓰면 임의의 벡터는 $X = C_1\boldsymbol{a}_1 + C_2\boldsymbol{a}_2 + \cdots + C_p\boldsymbol{a}_p$로 표현할수 있다.
- 여기서 $X \approx C_1\boldsymbol{a}_1 + C_2\boldsymbol{a}_2 + \cdots + C_k\boldsymbol{a}_k$ $(k < n)$로 보고 이 정보를 이용할 수 있다. (아래: $K = 5,200$, 원자료)



주성분 분석을 이용한 차원축소

- k차원으로 축소된 자료만을 이용하면 자료변동의 시각화도 가능



실습

데이터

- 1000개의 이미지 파일
- 하나의 이미지는 250 $\times$ 250 픽셀 $\times$ 3 색상 채널.
- 얼굴 주인의 이름 정보도 역시 볼 수 있음.
- 인터넷에서 “labeled face in the wild"를 검색하면 원본을 찾을 수 있음.

실습 목표

통계처리 프로그램인 R을 이용하여,

1. 이미지파일을 읽어 자료행렬을 만든다.
2. 자료의 각 관측값이 이미지임을 이해하고, 그림을 그린다.
3. 특이값분해를 이용하여 자료의 주성분을 찾아, 자료의 주변동을 이미지의 변화로 해석한다.
4. 주성분분석을 이용한 차원축소를 행한다.

생각해 볼것들

- k 바꿔가며 그렸을때 나타나는 현상
- 주성분을 몇개까지 고려해야 차원이 축소되면서 여전히 정보를 충분히 가지고 있을까?
- 데이터가 훨씬 클때 나타날 수 있는 문제들?