I.11 Norms of Vectors and Functions and Matrices

- Norm이란?
  - 벡터의 norm: 벡터의 크기
    - $\|\boldsymbol{x}\|_2 = (|x_1|^2 + \cdots + |x_n|^2)^{1/2}$ ➔ Euclid norm 또는 l^2-norm
    - $\|\boldsymbol{x}\|_\infty = \max|x_i|$ ➔ l^∞-norm (max norm)
    - $\|\boldsymbol{x}\|_1 = |x_i| + \cdots + |x_n|$ ➔ l^1-norm
    - Preference: $\|\boldsymbol{x}\|_2$, $\|\boldsymbol{x}\|_1$, $\|\boldsymbol{x}\|_\infty$
    - $\|\boldsymbol{x}\|_2$의 문제점: 작은 요소가 너무 작아지는 문제가 있음
  - 행렬의 norm
    - $\|A\|_F = (\sigma_1^2 + \cdots + \sigma_r^2)^{1/2}$ ➔ Frobenius norm
    - $\|A\| = \max \sigma_i$
    - $\|A\| = \sigma_1 + \cdots + \sigma_r$

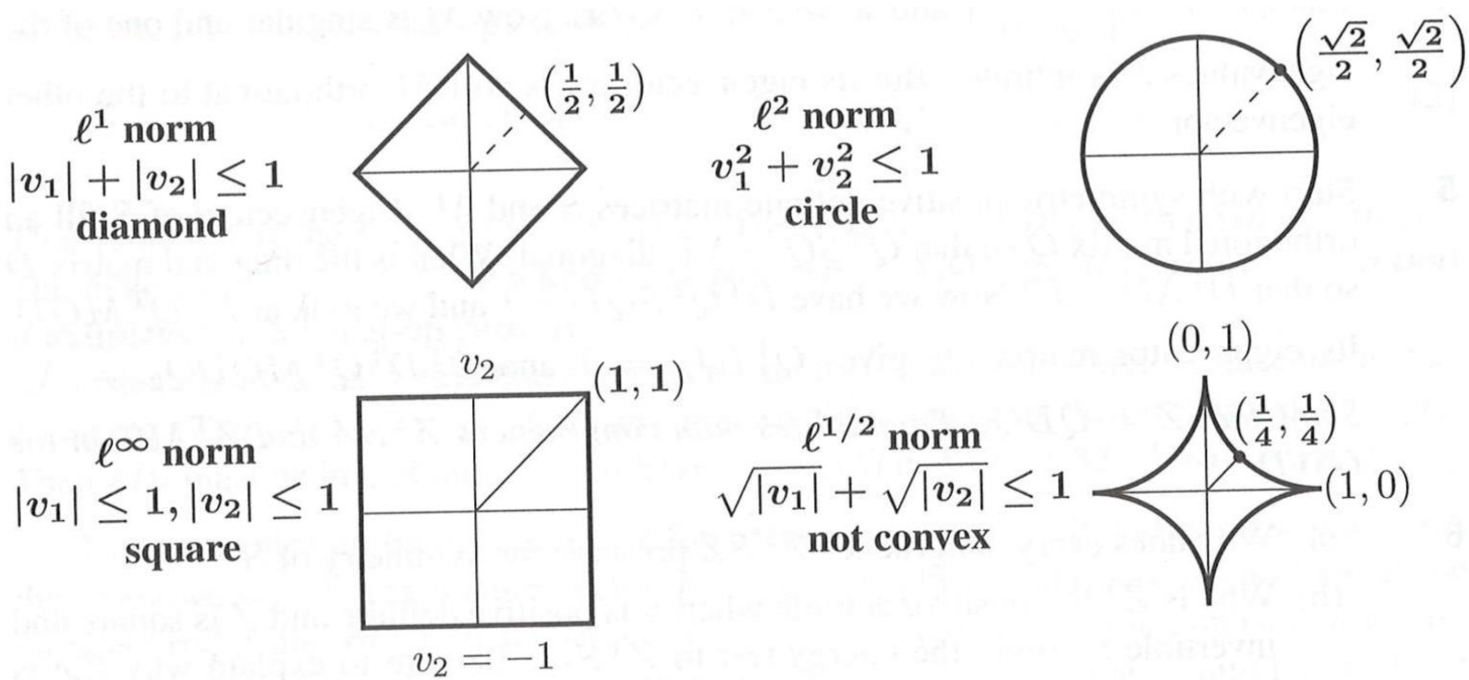
I.11 Norms of Vectors and Functions and Matrices

- Orthogonal invariance (대각 행렬에 대한 불변성)
  - 직교 행렬 Q로 기저 변환 (또는 좌표축 변환)을 했을 때 변하지 않는 값들
  - 벡터의 길이: $(Q\boldsymbol{x})^T Q\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{x}$
  - 행렬 $A = U\Sigma V^T$의 singular value σ: $Q_1 A Q_2 = Q_1 U\Sigma V^T Q_2$
    - 직교 행렬의 곱은 직교 행렬임
  - Spectral norm $\|A\|_2 = \max \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \sigma_1$
  - Nuclear norm $\|A\|_N = \sum \sigma_i$
  - Frobenius norm $\|A\|_F = (\sigma_1^2 + \cdots + \sigma_r^2)^{1/2} = \sqrt{\mathrm{tr}(A^T A)}$
  - 단위 행렬(I)의 경우
    - $\|I\|_2 = 1$
    - $\|I\|_N = n$
    - $\|I\|_F = \sqrt{n}$

I.11 Norms of Vectors and Functions and Matrices

- 벡터 norm의 성질
  - 아래 3가지 종류의 벡터 norm에 적용되는 공통적인 성질은?
    - $\|\boldsymbol{x}\|_2 = (|x_1|^2 + \cdots + |x_n|^2)^{1/2}$ ➔ Euclid norm 또는 l^2-norm
    - $\|\boldsymbol{x}\|_\infty = \max|x_i|$ ➔ l^∞-norm (max norm)
    - $\|\boldsymbol{x}\|_1 = |x_i| + \cdots + |x_n|$ ➔ l^1-norm
  - **0**-벡터를 제외한 모든 벡터 $\boldsymbol{x}$에 대해 $\|\boldsymbol{x}\| > 0$
  - $\|c\boldsymbol{x}\| = |c|\|\boldsymbol{x}\|$
  - $\|\boldsymbol{x} + \boldsymbol{y}\| \le \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$: triangular inequality
  - $\|\boldsymbol{x}\|_{1/2} = \left(\sqrt{|x_1|} + \sqrt{|x_2|}\right)^2$은 벡터 norm인가?
  - 좋은 norm의 기준
    - 평면내 $\|\boldsymbol{x}\| \le 1$를 만족하는 영역을 찾았을 때 영역의 모양이 Convex인 경우

I.11 Norms of Vectors and Functions and Matrices



I.9 PCA and the Best Low Rank Matrix

- 행렬 A의 A_k
  - 행렬 A가 주어졌을 때 A_k는 A를 SVD를 이용하여 rank-1 행렬들의 합으로 나타냈을 때, σ_1부터 σ_k까지 해당하는 항들만의 합을 나타냄

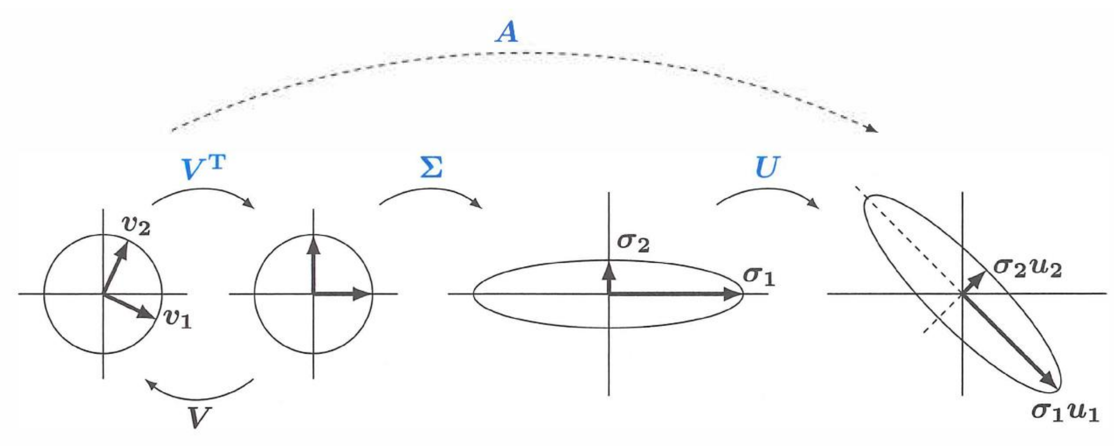
$$A_k = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T + \cdots + \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^T$$

  - A_k의 rank는 당연히 k임

- Eckart-Young 정리
  - 만약 행렬 B의 rank가 k이면, $\|A - B\| \geq \|A - A_k\|$이 항상 성립한다.
  - Rank가 k 인 행렬 중 행렬 A에 가장 가까운 행렬은 A_k이다.

I.8 Singular Values and Singular Vectors in SVD

- The singular vectors $\boldsymbol{v}_i$
- $\frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$ 의 최대값
  - $\frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$은 벡터 $\boldsymbol{x} = \boldsymbol{v}_1$일 때 최대가 되고, 이 때$\frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sigma_1$값을 가진다.
  - SVD를 아래 그림과 같이 입력 벡터의 회전 ➔ scaling ➔ 회전의 관점으로 보면 특이값(singular value)이 최대일 때, 비율이 최대가 되는 것은 당연해 보임.



I.8 Singular Values and Singular Vectors in SVD

- (증명) $\frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$은 벡터 $\boldsymbol{x} = \boldsymbol{v}_1$일 때 최대가 되고, 이 때 $\frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sigma_1$값을 가진다.
- 어떤 함수의 최대값 또는 최소값은 그 함수의 기울기가 모든 변수들에 대해 0이 되는 지점에서 일어남
- $\frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \frac{\boldsymbol{x}^T A^T A\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \frac{\boldsymbol{x}^T S\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}$ ➔ 대칭행렬의 경우 이런 형태의 비율을 Rayleigh quotient라고 부름
- 위 함수는 $x_1, \dots, x_n$의 함수이므로, 각각의 변수들에 대한 미분을 계산.

$$\frac{\partial}{\partial x_i}(\boldsymbol{x}^T\boldsymbol{x}) = \frac{\partial}{\partial x_i}(x_1^2 + \cdots + x_n^2) = 2(\boldsymbol{x})_i$$

$$\frac{\partial}{\partial x_i}(\boldsymbol{x}^T S\boldsymbol{x}) = \frac{\partial}{\partial x_i}\left(\sum_j \sum_k S_{jk} x_j x_k\right) = 2\sum_k S_{ik} x_k = 2(S\boldsymbol{x})_i$$

- $\frac{\partial}{\partial x_i}\left(\frac{\boldsymbol{x}^T S\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}\right)$ 의 분모는 $\boldsymbol{x} \neq \boldsymbol{0}$인 경우 0보다 크므로, 분자가 0이 되는 경우만 고려➔ $\left\{\frac{\partial}{\partial x_i}(\boldsymbol{x}^T S\boldsymbol{x})\right\}(\boldsymbol{x}^T\boldsymbol{x}) - (\boldsymbol{x}^T S\boldsymbol{x})\frac{\partial}{\partial x_i}(\boldsymbol{x}^T\boldsymbol{x}) = 2(S\boldsymbol{x})_i(\boldsymbol{x}^T\boldsymbol{x}) - (\boldsymbol{x}^T S\boldsymbol{x})2(\boldsymbol{x})_i = \boldsymbol{0}$
- $\{(\boldsymbol{x}^T\boldsymbol{x})S\}\boldsymbol{x} - (\boldsymbol{x}^T S\boldsymbol{x})\boldsymbol{x} = \boldsymbol{0}$ ➔ $S\boldsymbol{x} = \frac{(\boldsymbol{x}^T S\boldsymbol{x})}{(\boldsymbol{x}^T\boldsymbol{x})}\boldsymbol{x}$ ➔ 벡터 $\boldsymbol{x}$가 eigenvector일 때 기울기가 0이 됨.
- 벡터 $\boldsymbol{x}$가 eigenvector $\boldsymbol{v_i}$ 일 때는 $\frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \sqrt{\lambda_i} = \sigma_i$이므로 특이값이 최대인 σ_1이 $\frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$의 최대값이 됨.

I.9 PCA and the Best Low Rank Matrix

- Spectral norm: $\|A\|_2 = \max \frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sigma_1$ (또는 l^2-norm으로도 불림)
- Eckart-Young 정리의 예 (l^2-norm의 경우)
  - 주어진 행렬 A 가 대각 행렬인 경우, rank-2를 가진 행렬 B 중 $\|A-B\|_2$를 최소로 하려면 A_2 가 당연함
  - $A = \begin{bmatrix} \mathbf{4} & 0 & 0 & 0 \\ 0 & \mathbf{3} & 0 & 0 \\ 0 & 0 & \mathbf{2} & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix}$과 $A_2 = \begin{bmatrix} \mathbf{4} & 0 & 0 & 0 \\ 0 & \mathbf{3} & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 \\ 0 & 0 & 0 & \mathbf{0} \end{bmatrix}$의 비교
  - l^2-norm의 값은 임의의 대각 행렬 Q_1, Q_2에 의한 변환에 불변하므로, 위의 예는 4,3,2,1의 특이값을 가지는 임의의 4x4행렬에 대해서 적용됨

I.9 PCA and the Best Low Rank Matrix

- Eckart-Young 정리 (l^2-norm의 경우)
  - $\text{rank}(B) \le k$인 임의의 행렬 B가 주어지면, $\|A-B\|_2 = \max \frac{\|(A-B)\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \ge \sigma_{k+1}$ 이다.
  - (증명) $\|A-A_k\|_2 = \sigma_{k+1}$임을 이용. 증명의 핵심은 $\boldsymbol{x} \ne \boldsymbol{0}$이면서 $B\boldsymbol{x} = \boldsymbol{0}$ 와 $\boldsymbol{x} = \sum_{i=1}^{k+1} c_i \boldsymbol{v}_i$를 동시에 만족하는 벡터 $\boldsymbol{x}$를 찾는 것임.
  - 이러한 조건을 만족하는 벡터 $\boldsymbol{x}$는 다음의 관계가 성립

$$\|(A-B)\boldsymbol{x}\|_2^2 = \|A\boldsymbol{x}\|_2^2 = \left\|\sum_{i=1}^{k+1} c_i \sigma_i \boldsymbol{u}_i\right\|_2^2 = \sum_{i=1}^{k+1} c_i^2 \sigma_i^2$$

$$\sum_{i=1}^{k+1} c_i^2 \sigma_i^2 \ge \left(\sum_{i=1}^{k+1} c_i^2\right) \sigma_{k+1}^2 = \|\boldsymbol{x}\|_2^2 \sigma_{k+1}^2$$

  - 따라서 $\|(A-B)\boldsymbol{x}\|_2^2 \ge \|\boldsymbol{x}\|_2^2 \sigma_{k+1}^2$ 이므로, $\frac{\|(A-B)\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \ge \sigma_{k+1}$이 성립
  - 위의 조건을 만족하는 벡터 $\boldsymbol{x}$가 항상 존재한다고 말할 수 있는가?

I.9 PCA and the Best Low Rank Matrix

- Eckart-Young 정리 (l^2-norm의 경우)
  - (증명 계속) $\boldsymbol{x} \neq \mathbf{0}$이면서 $B\boldsymbol{x} = \mathbf{0}$ 와 $\boldsymbol{x} = \sum_{i=1}^{k+1} c_i \boldsymbol{v}_i$를 동시에 만족하는 벡터 $\boldsymbol{x}$가 항상 존재한다고 말할 수 있는가?
  - 행렬 B의 rank는 최대 k임. 따라서 행렬 B의 nullspace의 차원은 적어도 $n-k$임.
  - $\boldsymbol{x} = \sum_{i=1}^{k+1} c_i \boldsymbol{v}_i$를 만들 때 사용된 $\boldsymbol{v}_i$가 생성하는 공간의 차원은 $k+1$임. 이 벡터 $\boldsymbol{x}$가 행렬 B의 nullspace에 포함되지 않는다면 벡터 $\boldsymbol{x}$내에 nullspace의 성분이 없어야 함.
  - 하지만, nullspace의 차원과 $\boldsymbol{v}_i$가 생성하는 공간의 차원의 산술적인 합은 적어도 $(n-k)+(k+1) = n+1$ 임.
  - 두 부분공간이 공통으로 가지고 있는 차원이 있어야 하고, 벡터 $\boldsymbol{x}$를 그 차원내에 선택하면 위의 두 조건을 동시에 만족가능함.
  - 따라서 $\|A-B\|_2 = \max \frac{\|(A-B)\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \geq \sigma_{k+1}$은 성립해야 함

I.9 PCA and the Best Low Rank Matrix

- Frobenius Norm의 몇가지 다른 형태
  1. $\|A\|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2$
  2. $\|A\|_F^2 = \mathrm{tr}(A^T A) = (A^T A)_{11} + \cdots + (A^T A)_{nn}$
  3. $\|A\|_F^2 = |a_{11}|^2 + |a_{12}|^2 + \cdots + |a_{mn}|^2$ (모든 $|a_{ij}|^2$성분들의 합)

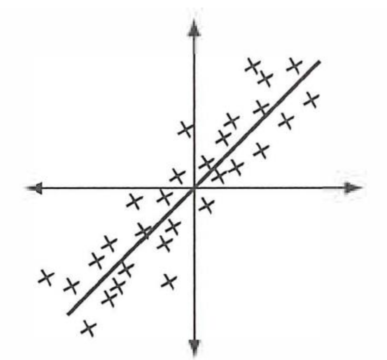
  - 2의 정의로부터 3의 유도
    - $(A^T A)_{11}$는 A의 첫번째 열의 성분들의 제곱의 합 $|a_{11}|^2 + |a_{21}|^2 + \cdots + |a_{m1}|^2$임
    - 마찬가지로 $A^T A$의 대각성분 $(A^T A)_{ii}$는 A의 i번째 열의 성분들의 제곱의 합 $|a_{1i}|^2 + |a_{2i}|^2 + \cdots + |a_{mi}|^2$이므로, 모든 대각 성분의 합은 행렬 A의 모든 성분들의 제곱의 합이 됨.
- Frobenius norm에 대해서도 Eckart-Young 정리가 성립함을 보일 수 있음 ➔ 교재 74페이지 참고.

I.9 PCA and the Best Low Rank Matrix

- Principal component analysis (PCA)
  - N명의 사람을 측정하는데 매번 측정마다 (나이, 키, 몸무게)같은 m 개의 측정값을 얻었다고 가정
  - 이 경우 행렬 A_0는 각 사람의 데이터를 열벡터로 가진 $m \times n$ 행렬로 가정

평균 대비	사람1	사람2	사람3	사람4	사람5	사람6
나이 차이	3	-4	7	1	-4	-3
키 차이	7	-6	8	-1	-1	7

  - 우선 행렬 A_0의 각 행마다의 평균값을 구한 다음, 해당하는 행에서 뺀 행렬을 A라고 함
  - ➔ 행렬 A의 각 열은 $\mathbf{R}^m$공간내 n개의 점에 해당
  - ➔ 행렬 A의 각 행의 평균은 0
  - ➔ 행렬 A의 모든 열의 합과 평균도 0임
  - ➔ n개의 점들이 원점을 중심으로 분포되어 있는데, 특정한 선이나 평면같은 $\mathbf{R}^m$공간의 낮은 차원을 가진 부분공간일 가능성이 있음
  - 원점을 지나는 직선은 행렬 A의 $\boldsymbol{u}_1$벡터임



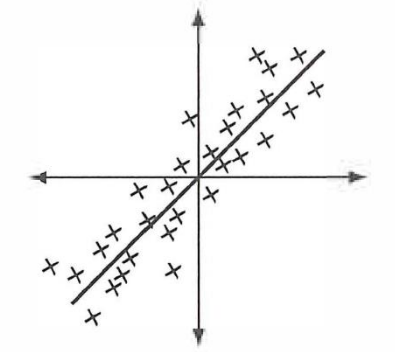
I.9 PCA and the Best Low Rank Matrix

- Principal component analysis (PCA)의 통계적 관점
  - 행렬 A_0의 각각의 행에서 평균을 제거하였으므로 행렬 A의 모든 값은 평균으로부터 떨어진 거리임
  - 분산
    - 평균으로부터 떨어진 거리의 제곱의 합
    - 주어진 행렬 A의 AA^T의 대각 원소들의 합임
  - 공분산
    - 측정 데이터들간의 상관관계를 나타냄
    - 주어진 행렬 A의 AA^T의 대각이 아닌 원소들과 관련됨
    - ➔ (A의 i번째 행)과 (A의 j번째 행)간의 내적임
  - 샘플들의 공분산 행렬은 샘플의 개수가 n일 때, 다음과 같이 나타남
    - $S = \frac{AA^T}{n-1}$

I.9 PCA and the Best Low Rank Matrix

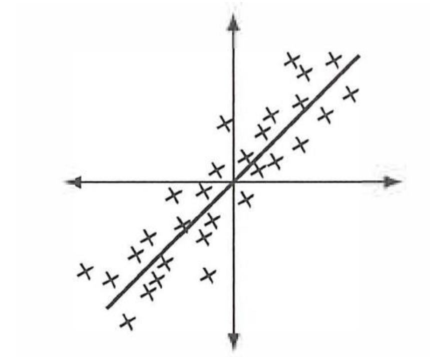
- Principal component analysis (PCA)의 통계적 관점
  - 공분산 행렬 $S = \frac{AA^T}{n-1}$
  - $A = \begin{bmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & 7 \end{bmatrix}$
  - $S = \frac{AA^T}{6-1} = \begin{bmatrix} 20 & 25 \\ 25 & 40 \end{bmatrix}$
  - ➔ eigenvalue: 57, 3
  - eigenvalue 57에 해당하는 eigenvector $\boldsymbol{u}_1$는 $\boldsymbol{u}_1 \approx (0.6, 0.8)$임

I.9 PCA and the Best Low Rank Matrix



- PCA의 기하학적 의미
  - 벡터 $\boldsymbol{u}_1$이 나타내는 선과 점들간에는 perpendicular least square의 관계가 존재함.
  - 일반적인 최소 자승법(least square)
    - 다른 종류의 해석임
    - 는 $A\boldsymbol{x} = \boldsymbol{b}$를 푸는 해의 근사값을 얻기 위해 $|A\boldsymbol{x} - \boldsymbol{b}|^2$를 최소화하는 $\boldsymbol{x}$를 구하는 방법임
    - $A^T A\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$ 형태로 해를 구함
  - PCA는 특이값과 해당하는 특이벡터로 해를 구함
  - 비슷한 결과를 얻는 경우도 많으나, 최적화하는 목표가 다름
  - $\boldsymbol{u}_1$이 나타내는 선으로부터 거리의 제곱의 합이 최소가 됨

I.9 PCA and the Best Low Rank Matrix



- PCA의 기하학적 의미
  - $\boldsymbol{u}_1$이 나타내는 선으로부터 거리의 제곱의 합이 최소가 됨
    - 행렬 A의 j번째열의 열벡터 $\boldsymbol{a}_j$를 직교하는 두 기저 $\boldsymbol{u}_1, \boldsymbol{u}_2$와 평행한 성분으로 나눠서 생각해보면,

$$\sum_{1}^{n} \|\boldsymbol{a}_j\|^2 = \sum_{1}^{n} |\boldsymbol{a}_j^T \boldsymbol{u}_1|^2 + \sum_{1}^{n} |\boldsymbol{a}_j^T \boldsymbol{u}_2|^2$$

    - 등식의 오른쪽 첫번째 항은 $\boldsymbol{u}_1^T \boldsymbol{a}_j \boldsymbol{a}_j^T \boldsymbol{u}_1$의 합이므로, $\boldsymbol{u}_1^T (AA^T)\, \boldsymbol{u}_1$에 해당
    - 등식의 왼쪽 항이 샘플 데이터에 의해 고정된 값일 경우에는, 오른쪽 첫번째 항이 최대가 됨에 따라 나머지 거리가 최소로 됨
- 선형 대수적 관점
  - 모든 데이터의 분산의 합: $T = \|A\|_F^2/(n-1) = (\|\boldsymbol{a}_1\|^2 + \cdots + \|\boldsymbol{a}_n\|^2)/(n-1)$
  - $T = (\sigma_1^2 + \cdots + \sigma_r^2)/(n-1)$: 각각의 principal component가 전체 데이터의 분산에 끼치는 영향을 볼 수 있음

II.2 Least Squares: Four Ways

- 최소 자승법 (least-squares)
  - 예를 들어 3개의 데이터$(t, b) = \{(1,1), (2,2), (3,2)\}$를 가지고 있다고 가정
  - 이 데이터에 fitting할 수 있는 직선 $b = C + Dt$를 찾고자 하면 다음과 같은 연립방정식이 가능함
  - $\begin{cases} C + D = 1 \\ C + 2D = 2 \\ C + 3D = 2 \end{cases} \Rightarrow A\boldsymbol{x} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \boldsymbol{b}$
  - 일반적으로 해가 존재하는 문제인가?
  - 역행렬과 유사한 것을 만드는 방법은?

II.2 Least Squares: Four Ways

- 최소 자승법 (least-squares)
  - 예를 들어 3개의 데이터$(t, b) = \{(1,1), (2,2), (3,2)\}$를 가지고 있다고 가정
  - 이 데이터에 fitting할 수 있는 직선 $b = C + Dt$를 찾고자 하면 다음과 같은 연립방정식이 가능함
  - $\begin{cases} C + D = 1 \\ C + 2D = 2 \\ C + 3D = 2 \end{cases} \rightarrow A\boldsymbol{x} = \begin{bmatrix} & \\ & \\ & \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \end{bmatrix} = \boldsymbol{b}$
  - 일반적으로 해가 존재하는 문제인가?
  - 역행렬과 유사한 것을 만드는 방법은?

II.2 Least Squares: Four Ways

- Projection 행렬
  - $P = A(A^TA)^{-1}A^T$ 는 행렬 A의 열벡터들로 생성된 부분공간으로 주어진 입력 벡터를 projection시킴
  - $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p} = \boldsymbol{b} - P\boldsymbol{b} = (I - P)\boldsymbol{b}$
  - $A^T\boldsymbol{e} = A^T(A\hat{\boldsymbol{x}} - b) = 0 \Rightarrow A^TA\hat{\boldsymbol{x}} = A^Tb$
  - $\boldsymbol{b}$가 입력벡터인 경우, $A^T\boldsymbol{b}$는 행렬 A의 행벡터 공간으로 mapping이 됨. Why?
  - $A^TA\boldsymbol{x} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}\begin{bmatrix} C \\ D \end{bmatrix}$

$$= A^T\boldsymbol{b} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 11 \end{bmatrix}$$