서울시립대학교
UNIVERSITY OF SEOUL

6주차: 내적공간

서울시립대학교 통계학과
전종준

Contents

- 벡터공간과 내적
- 내적과 기하
- 함수공간
- 내적과 데이터 표현형

01. 벡터공간과 내적

벡터공간과 내적

벡터공간의 정의

- 벡터공간(vector space)은 스칼라 체(field) F 위에서 정의된 구조로, 두 가지 연산(벡터 덧셈, 스칼라 곱셈)에 대해 다음의 공리를 만족하는 벡터들의 집합 벡터
- 스칼라 필드 (예: Q, R, C)
  - 벡터 덧셈: $+: V \times V \mapsto V$
  - 스칼라 곱셈: $\cdot: F \times V \mapsto V$

벡터공간의 정의

- 모든 $u,v,w \in V$ $a,b \in F$에 대해:
  1) **덧셈의 교환법칙**: $u+v=v+u$
  2) **덧셈의 결합법칙**: $(u+v)+w=u+(v+w)$
  3) **영벡터 존재**: $\exists 0 \in V$ such that $v+0=v$
  4) **역벡터 존재**: $\forall v \in V, \exists -v \in V$ such that $v+(-v)=0$
  5) **스칼라 곱의 결합법칙**: $a(bv)=(ab)v$
  6) **1에 대한 항등성**: $1 \cdot v=v1$
  7) **스칼라에 대한 분배법칙**: $(a+b)v=av+bv$
  8) **벡터에 대한 분배법칙**: $a(u+v)=au+av$

벡터공간과 내적

내적의 정의

- 내적(inner product) 은 벡터공간에서 두 벡터 u,v∈V에 대해 실수 또는 복소수 값을 반환하는 함수로, 다음의 공리를 만족합니다: ⟨·,·⟩:V×V→F모든 u,v,w∈V, 스칼라 a∈Fa 에 대해:
  - **양의 정부호성 (Positive-definiteness)**:⟨v,v⟩≥0,and⟨v,v⟩=0 ⟺ v=0
  - **선형성 (Linearity in the first argument)**:
    ⟨au+v,w⟩=a⟨u,w⟩+⟨v,w⟩
  - **켤레 대칭성 (Conjugate symmetry)** (실수 공간에서는 단순 대칭성):
    $< u, v > = < v, \bar{u} >$

벡터공간과 내적

내적이 정의된 벡터공간의 예

- 실수를 원소로 가지는 유한차원 벡터공간

  벡터공간: $\boldsymbol{a} = (a_1, \dots, a_n) \in \mathbb{R}^n$

  내적공간: $< \boldsymbol{a}, \boldsymbol{b} > = \sum_{i=1}^{n} a_i b_i$ 으로 내적을 정의

- l_2 공간

  벡터공간: $\boldsymbol{a} \in l_2, \boldsymbol{a} = (a_1, a_2, \dots)$ 이면서 $\sum_{i=1}^{\infty} a_i^2 < \infty$ 인 원소로 가지는 벡터공간

  내적공간: $\boldsymbol{a}, \boldsymbol{b} \in l_2$ 에 대해서 $< \boldsymbol{a}, \boldsymbol{b} > = \sum_{i=1}^{\infty} a_i b_i$ 로 정의

  1) $\boldsymbol{a}, \boldsymbol{b} \in l_2$ 면 $\sum_{i=1}^{\infty} a_i b_i < \infty$

  2) $< k\boldsymbol{a}, \boldsymbol{b} > = k < \boldsymbol{a}, \boldsymbol{b} >$

  3) 내적의 정의를 만족하는지 확인

벡터공간과 내적

내적이 정의된 벡터공간의 예

- $L_2(\mu)$ 공간 (여기서 μ 는 measurable space $(\mathcal{X}, \mathcal{F})$위에서 정의된 유한측도)

  벡터공간: $f \in L_2(\mu)$, $(f: \mathcal{X} \mapsto \mathbb{R}$, measurable function) 에 대해서 $\int f(x)^2 d\mu(x) < \infty$ 를 만족

  내적공간: $f, g \in L_2(\mu)$에 대해서 $< f, g > = \int f(x)g(x)d\mu(x)$

  1) $f, g \in L_2(\mu)$면$< f, g > < \infty$

  2) $< kf, g > = k < f, g >$

  3) 내적의 정의를 만족하는지 확인

벡터공간과 내적

기저벡터

- l_∞ 공간

  어떤 임의의 원소 $\boldsymbol{a} \in l_2$ 를 선형결합으로 표현할 수 있는 원소들의 집합을 생각해보자

- $L_2(\mu)$ 공간

  어떤 임의의 원소 $f \in L_2(\mu)$, $(f: \mathcal{X} \mapsto \mathbb{R}$, measurable function) 를 선형결합으로 표현할 수 있는 원소들의 집합을 생각해보자.

벡터공간과 내적

내적공간의 예

- $L_2(\mu)$ 공간

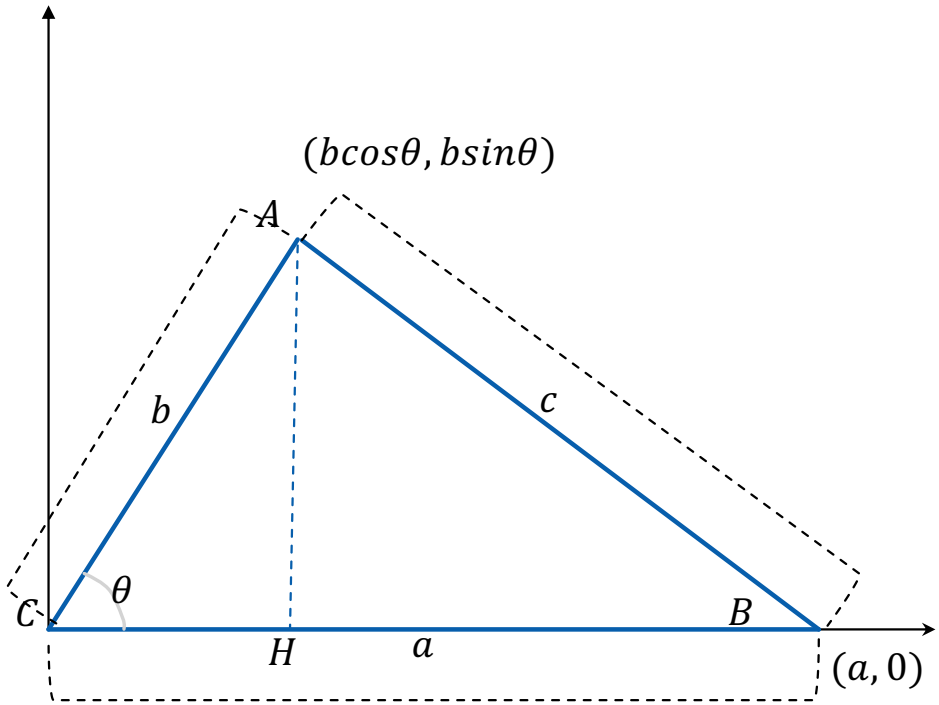
  어떤 임의의 원소 $f \in L_2(\mu)$, ($f: \mathcal{X} \mapsto \mathbb{R}^p$, measurable function) 를 선형결합으로 표현할 수 있는 원소들의 집합을 생각해보자.

  $\int \|f\|^2 d\mu(x) < \infty$ 를 만족하는 f 의 모임을 생각한다.

  $f, g \in L_2(\mu)$에 대해서 $< f, g > = \int f(x)^T g(x) d\mu(x)$ 으로 정의하면 내적이 되는가?

02. 내적과 기하

cosine law

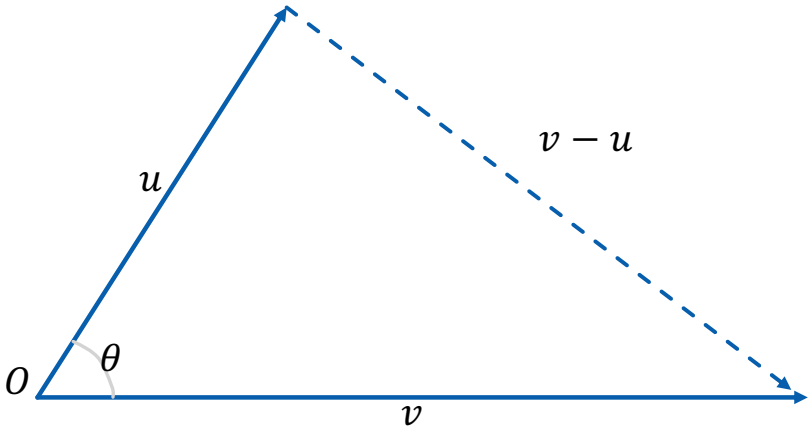


- 피타고라스 정리를 적용하면 $\overline{AH}^2 + \overline{BH}^2 = \overline{AB}^2$ 이므로 좌표계를 이용하여 다음을 알 수 있음

  $(bsin\theta)^2 + (a - bcos\theta)^2 = c^2$

  즉, $a^2 + b^2 - 2abcos\theta = c^2$

cosine law



Cosine law 를 내적이 정의된 벡터공간에 적용

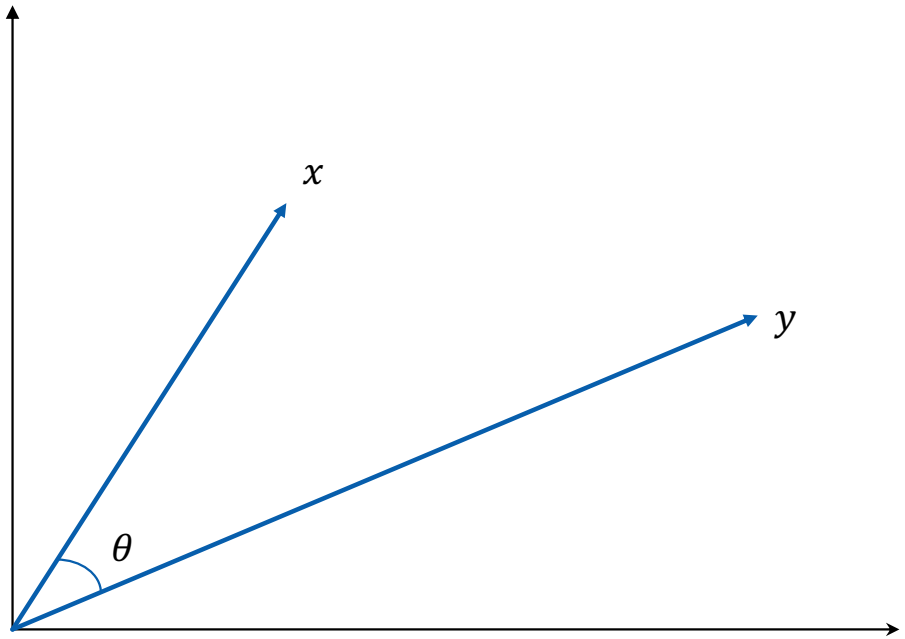
- $a^2 + b^2 - 2abcos\theta = c^2$ 에서

  $a^2 = \|v\|^2, b^2 = \|u\|^2, c^2 = \|v-u\|^2$ 에

  대응되므로 이를 정리하면 $cos\theta = \frac{v^T u}{\|v\|\|u\|}$

- 두 벡터의 내적이 0이라는 것은 두 벡터가 직교한다고 해석할 수 있음

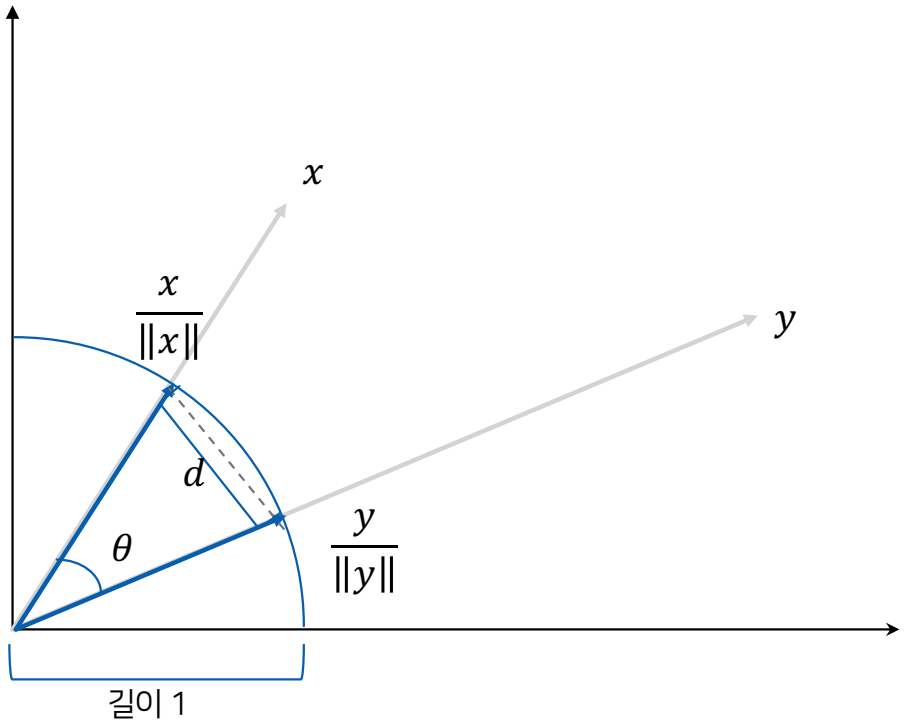
*cosine simlarity



Cosine similary 는 두 데이터의 유사도 측도

- $\cos(x, y) = \frac{x^T y}{\|x\|\|y\|}$
- 벡터공간에서 cosine law 를 통해서 $\cos(x, y)$가 두 벡터가 이루는 각도임을 알 수 있음

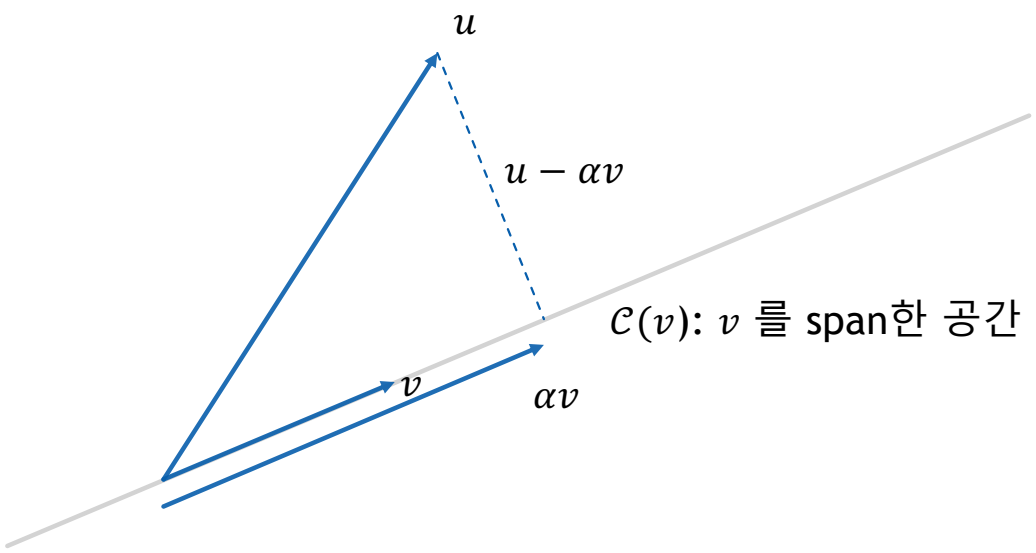
*cosine simlarity



Cosine similary 는 두 데이터의 유사도 측도

- sin 반각공식을 이용하면 $\cos\theta = 1 - \frac{d^2}{2}$
- 즉, cosine simarlity 단위원으로 표준화된 데이터의 유클리디안 거리의 함수임.
- 어떤 텍스트 토큰과 가까운 다른 텍스트 토큰을 찾기 위해 텍스트 임베딩 벡터간 cosine 유사도의 계산 -> ***"단위원 위에서 거리가 가까운 토큰 임베딩 벡터를 찾는 것임"***

정사영 (Projection)



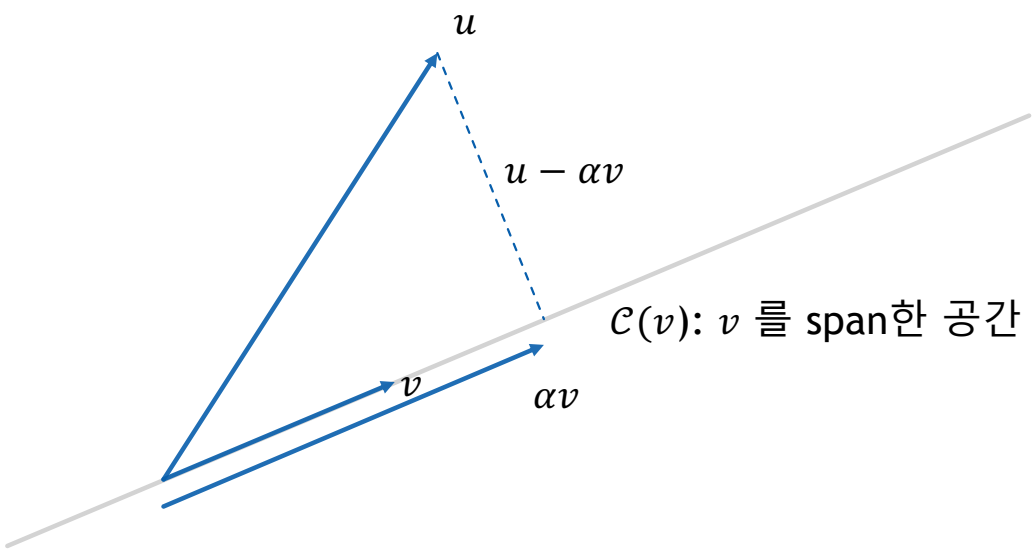
$\mathcal{C}(v)$ 위에서 u와 직교하는 벡터

- $\mathcal{C}(v)$ 의 임의의 벡터는 αv $(\alpha \in \mathbb{R})$로 표시할 수 있음
- $< u - \alpha v, \alpha v > = 0$ 만족하는 0이 아닌 α 는 $\alpha = u^T v / \|v\|^2$

  즉, $\frac{u^T v}{\|v\|^2} v$ 는 $\mathcal{C}(v)$ 위의 벡터로

  $u - \frac{u^T v}{\|v\|^2} v$ 와 직교

정사영 (Projection)



최단거리 벡터

- $\mathcal{C}(v)$ 의 벡터로서 u와 유클리디안 거리가 가장 가까운 벡터
- $f(\alpha) = \|u - \alpha v\|^2$ 를 최소화 하는 α는

$$\alpha = u^T v / \|v\|^2$$

즉, $\frac{u^T v}{\|v\|^2} v$ 는 $\mathcal{C}(v)$ 위의 벡터로 u와 가장 거리가 가까운 벡터

- *정사영의 두 가지 의미:*

1) 직교, 2)최단거리

Gram-Schmidt 직교화

Gram-Schimidt 직교화

- 벡터공간 $\mathcal{C}$ 를 생성하는 선형독립인 벡터 (기저) 의 벡터 $u_1, \dots, u_n$ 이 있을때 동일한 벡터공간 벡터공간 $\mathcal{C}$ 를 생성하는 직교기저 $v_1, \dots, v_n$ 를 어떻게 만드나?
- 직교화 과정
  - $v_1 = u_1$
  - $v_2 = u_2 - (\frac{u_2^T v_1}{\|v_1\|^2}) v_1$
  - $v_3 = u_3 - \frac{u_3^T v_2}{\|v_2\|^2}\ v_2 - \frac{u_3^T v_1}{\|v_1\|^2}\ v_1$
  - $\dots$
  - $v_n = u_n - \frac{u_n^T v_{n-1}}{\|v_{n-1}\|^2}\ v_{n-1} - \dots - \frac{u_n^T v_2}{\|v_2\|^2}\ v_2 - \frac{u_n^T v_1}{\|v_1\|^2}\ v_1$

직교기저를 통한 내적의 해석

정사영으로서 내적

- 벡터공간 $\mathcal{C}$ 를 생성하는 직교기저 $v_1, \dots, v_n$ 가 있고 $\|v_k\| = 1$ all k 이라고 가정해보자.
- 어떤 벡터 $u \in \mathcal{C}$ 에 대해서 $u^T v_k$ 는 u를 v_k에 정사영을 나타내는 계수다. 즉, $(u^T v_k)v_k$ 는 v_k방향으로 u의 정사영이다. -> $u^T v_k$ 은 u가 가진 v_k방향의 성분값이라고 볼 수 있음.
- 단위직교벡터에 대해서 어떤 벡터의 내적은 그 벡터가 가지고 있는 단위직교벡터 방향성분으로 해석할 수 있음

$$u = (u^T v_1)v_1 + (u^T v_2)v_2 + \cdots + (u^T v_n)v_n$$

- 만약 $(u^T v_k) = 0$ 이라면 u는 v_k방향의 성분이 없는 것임. $(u^T v_k) \approx 0$ 이라면 u는 v_k방향의 성분이 매우 작은 것임 -> u 정보를 압축할 수 있음

정사영

열생성공간으로 벡터의 정사영

- 생성공간: 어떤 벡터가 주어진 경우 그 벡터들의 선형결합으로 만들수 있는 모든 벡터들을 모아놓은 공간

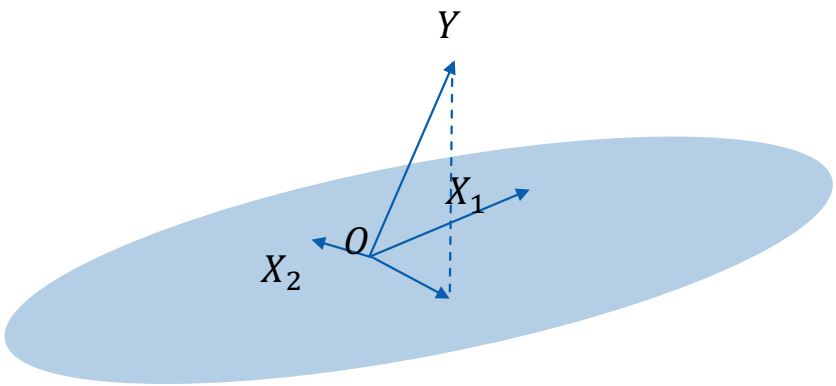
  $X_1, \dots, X_p$ 를 벡터라고 할 때, $\mathcal{C}(X_1, \dots, X_p)$ 로 표시

- 벡터 Y 의 생성공간 $\mathcal{C}(X_1, \dots, X_p)$로 정사영

  (Projection of Y onto $\mathcal{C}(X_1, \dots, X_p)$)

- Y 와 **최소거리**를 갖는 생성공간 $\mathcal{C}(X_1, \dots, X_p)$위의 벡터를 찾는 방법으로 정사영을 정의할 수 있음

  (정사영의 두 가지 의미) 1) 직교, **2) 최단거리**



내적공간의 예시

열생성공간으로 벡터의 정사영

- 생성공간 생성공간 $\mathcal{C}(X_1, \dots, X_p)$ 위의 임의의 벡터는 $\alpha_1 X_1 + \cdots + \alpha_p X_p$ $(\alpha_j \in \mathbb{R})$ 로 표시할 수 있음
- Y 와 $\alpha_1 X_1 + \cdots + \alpha_p X_p$ 의 유클리디안 거리의 제곱을 다음과 같이 함수로 표시함

  $\mathrm{f}(\alpha) = \|Y - X\alpha\|^2$ 단, $X = [X_1, \dots, X_p] \in \mathbb{R}^{n \times p}$, $\alpha = (\alpha_1, \dots, \alpha_p)^T \in \mathbb{R}^p$

- $\mathrm{f}(\alpha)$ 를 최소화하는 α 는$\alpha = (X^T X)^{-1} X^T Y$ 로 주어지므로 $X(X^T X)^{-1} X^T Y$ 는 생성공간 $\mathcal{C}(X_1, \dots, X_p)$ 위의 벡터로 Y 와 거리가 가장 가까움.
- $\boldsymbol{X(X^T X)^{-1} X^T Y}$ 를 $\mathcal{C}(X_1, \dots, X_p)$에 대한 벡터 Y 의 정사영임.

정사영

정사영 연산자 (Projection operator)

- $X_1, \dots, X_p$ 를 열벡터라고 하고 , $X = [X_1, \dots X_p]$ 라 두었을때 생성공간 $\mathcal{C}(X)$에 대한 정사영 연산자

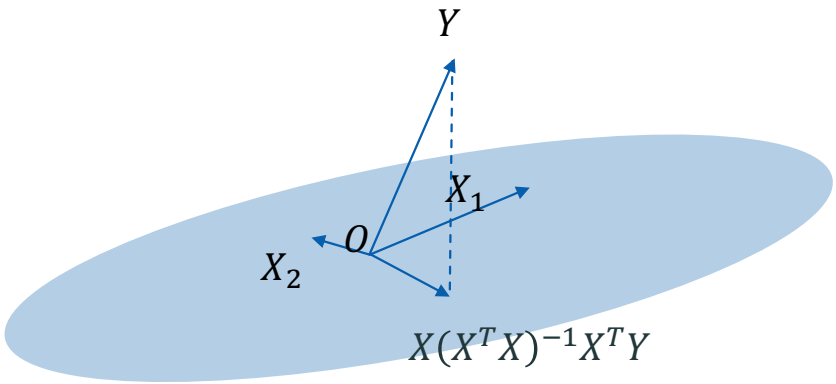
  $\Pi_{\mathcal{C}(X)} = X(X^TX)^{-1}X^T$ ($(X^TX)^{-1}$ 가 존재함을 가정)

- 정사영 연산자의 성질

  $\Pi_{\mathcal{C}(X)}\mathrm{Y} \in \mathcal{C}(X)$, $(\mathrm{I} - \Pi_{\mathcal{C}(X)})\mathrm{Y} \in \mathcal{C}^{\perp}(X)$

  $\Pi_{\mathcal{C}(X)} = \Pi_{\mathcal{C}(X)}\Pi_{\mathcal{C}(X)}$

  $\Pi_{\mathcal{C}(X)}\left(I - \Pi_{\mathcal{C}(X)}\right) = 0$



단, $X = [X_1, X_2]$

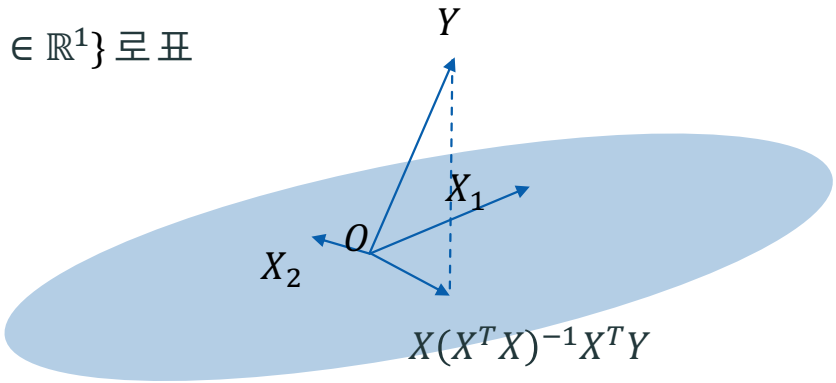
회귀모형과 정사영

정사영 연산자 (Projection operator)

- X 는 $n \times p$ 의 설명변수 데이터 행렬, Y 는 n차원 반응변수 열벡터를 생각하자
- X_j 를 X 의 j번째 열벡터라고 하면 $X = [X_1, \dots X_p]$ 라 표시할 수 있다.
- 생성공간 $\mathcal{C}(X)$은 $\{U \in \mathbb{R}^n : \hat{Y} = \beta_1 X_1 + \cdots + \beta_p X_p, \beta_1, \dots, \beta_p \in \mathbb{R}^1\}$ 로 표시할 수 있다.
- $\hat{\mathrm{Y}} = \Pi_{\mathcal{C}(X)}\mathrm{Y} = X(X^T X)^{-1} X^T Y = X\hat{\beta}$

  $\mathrm{SSE} = \left\|\mathrm{Y} - \Pi_{\mathcal{C}(X)}\mathrm{Y}\right\|^2,$

  $\mathrm{R}^2 = \left\|\Pi_{\mathcal{C}(X)}\mathrm{Y}\right\|^2 / \|\mathrm{Y}\|^2$ (R은 $|\cos(\Pi_{\mathcal{C}(X)}\mathrm{Y},\ Y)|$로 볼 수 있음)



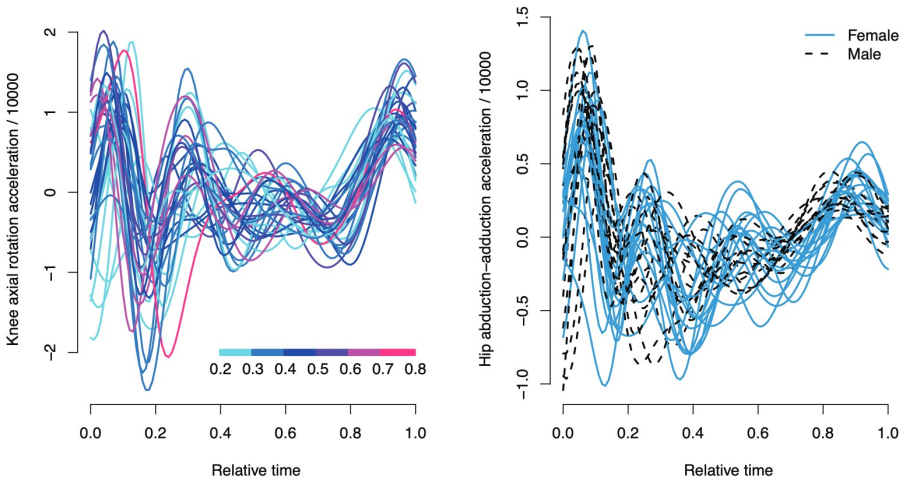
단, $X = [X_1, X_2]$

03. 함수공간

함수 공간

데이터로서 함수의 예시

- 인간의 운동 및 운동제어에 대한 분석:
  시간에 따른 관절의 움직임 (각도, 가속도 등)의 정보를 하나의 데이터로 간주
  예시) 관절 이상이 있는 사람과 향후 1년내 관절 이상이 생길 사람의 움직임 패턴 분석 -> 언제, 어떤 움직임에 차이가 있는가?



[Gertheiss et al., 2024] 왼쪽: 무릎 축 회전 가속도 측정값(색상은 최대 모멘트에 해당), 오른쪽: 엉덩이 외전-내전 가속도 프로파일(색상은 피험자의 성별에 해당)

Gertheiss, J., Rügamer, D., Liew, B. X., & Greven, S. (2024). Functional data analysis: An introduction and recent developments. *Biometrical Journal*, *66*(7), e202300363.

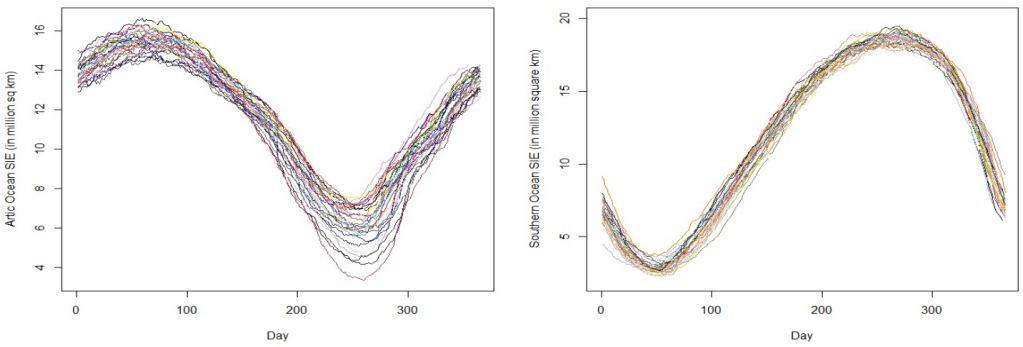
함수 공간

데이터로서 함수의 예시

- 북극/남극의 해빙 면적의 연중 변화

  위성사진을 통해 얻어진 1년동안 해빙의 면적 관측 정보를 하나의 데이터로 간주

  예시) 연중 해빙 면적의 변화에 대한 95% 함수 신뢰구간을 어떻게 만들 것인가?



[Das et al., 2018] 왼쪽: 북극 일별 해빙 면적의 시계열 데이터, 오른쪽: 남극 일별 해빙 면적의 시계열 데이터

Das, P., Lahiri, A., & Das, S. (2018). Understanding sea ice melting via functional data analysis. Current Science, 115(5), 920-929.

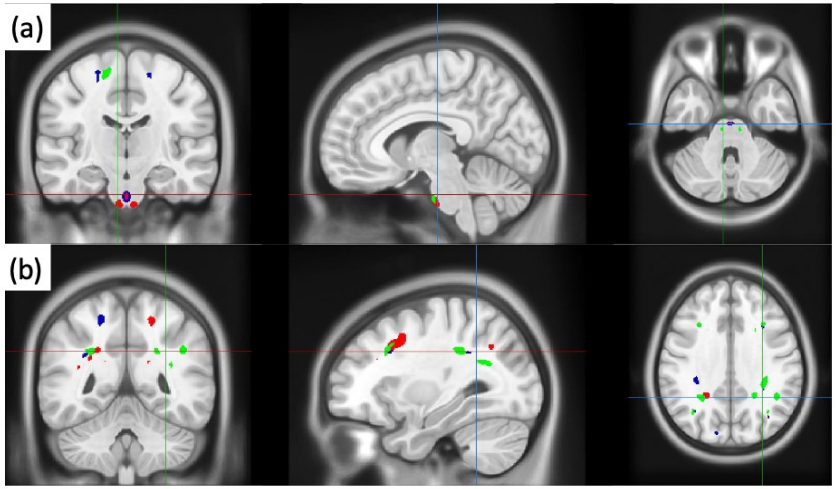
함수 공간

데이터로서 함수의 예시

- 뇌의 fMRI 데이터 분석

  한사람의 fMRI를 이용한 스캐닝 데이터를 하나의 데이터로 간주

  예시) 뇌의 기능적 차이를 발생시키는 부분을 어떻게 찾을 것인가?



[Consagra et al., 2024] 질병상태와 연관이 있다고 추측되는 fMRI 데이터상의 위치

Consagra, W., Venkataraman, A., & Qiu, X. (2024). Efficient Multidimensional Functional Data Analysis Using Marginal Product Basis Systems. Journal of Computational and Graphical Statistics, 33(2), 567-577.

함수 공간

$L_2 - space$

- 내적값이 유한한 값을 가지기 위해서 생각할 수 있는 하나의 조건:

$$< u, v > \leq \|u\|^2 \|v\|^2$$

- 전체 집합 $\mathcal{F}_0$를 실수 위에서 정의된 연속함수라고 가정해보자. 다음과 같이 $\mathcal{F} \subset \mathcal{F}_0$을 정의함.

$$\mathcal{F} = \{f \in \mathcal{F}_0 : \int f^2(x)dx < \infty\}$$

- $\mathcal{F}$에 내적을 다음과 같이 정의할 수 있음. (내적의 정의를 만족하는지 확인!)

$$< f, g > = \int f(x)g(x)dx$$

바나하 공간 (Banach Space)

$L_2 - space\ \mathcal{F}$

- $\mathcal{F}_0$를 벡터공간(linear vector space)이라고 가정. (벡터공간 정의확인!)
  임의의 $f, g \in \mathcal{F}_0$ 에 대해서 $\alpha_1 f + \alpha_2 g \in \mathcal{F}_0$ 를 의미함.

- $\mathcal{F}$위에서 norm 을 $\|f\| = \sqrt{\int f^2(x)dx}$ 으로 가정하고 $\mathcal{F} = \{f \in \mathcal{F}_0 : \int f^2(x)dx < \infty\}$ 이라 놓자.
  이를 $L_2 - space$ 라 부른다.
  - $L_p - space$는 complete linear space 다. (Riesz-Fischer Theorem)
    complete space 의 의미: $\{f_n\} \in \mathcal{F}$ 에 대해서 만약 f가 존재하여 $\|f_n - f\| \to 0$ as $n \to \infty$ 라면 $f \in \mathcal{F}$. ($\mathcal{F}$는 완비되어 있음)
    (참고) 선형벡터공간 + norm 정의됨 + 완비공간 -> Banach Space 라 부름.

바나하 공간 (Banach Space)

Banach space

- 정의: 바나하 공간은 노옴(norm)이 정의된 완비된 선형벡터 공간이다.

  노옴(norm): 양의 정부호성 ($\|x\| \geq 0, \|x\| = 0 \Leftrightarrow x = 0$), 동차성 ($\|\alpha x\| = |\alpha|\|x\|$), 삼각부등식 ($\|x + y\| \leq \|x\| + \|y\|$)

  완비성: 모든 코시 수열이 수렴

바나하 공간 (Banach Space)

Banach space 공간의 예

- 유클리드 공간
- $L_p - space$
- $l_p - space$: 무한차원 벡터공간

  수열 $x = (x_1, x_2, \dots)$ 에 대해서 $l_p - space$ 의 norm 은 $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ 로 정의

힐버트 공간 (Banach Space)

힐버트 공간으로서 $L_2 - space\ \mathcal{F}$

- $\mathcal{F}$ 위에서 추가적으로 다음을 만족하는 map $<\cdot,\cdot>: \mathcal{F}\times\mathcal{F} \mapsto \mathbb{R}$ 을 정의

  ($\mathcal{F}$ 는 바나하 공간에 내적을 도입한 특별한 공간)

  1) $< \alpha_1 f + \alpha_2 g, h > = \alpha_1 < f, h > + \alpha_2 < g, h >$

  2) $< f, g > =< g, f >$

  3) $< f, f >= \|f\|^2 \geq 0$

- 예) $< f, g > = \int f(x)g(x)dx$ 는 내적이라고 두면 $L_2 - space\ \mathcal{F}$ 는 힐버트 공간이 된다.

04. 내적과 데이터 표현형

기저와 데이터 표현

벡터공간과 기저

- 벡터공간은 벡터들의 집합으로, 벡터 덧셈과 스칼라 곱셈 연산에 대해 닫혀 있는 구조
- 데이터 x 가 벡터공간의 원소라고 가정 (데이터의 공간을 벡터공간이라고 가정!)
- 벡터공간 V의 기저(basis)란, 선형 독립이면서 V의 모든 벡터를 선형 결합으로 표현할 수 있는 벡터들의 집합
- 즉, $\{v_1, \dots, v_n\}$ 이 V의 기저라면, 모든 $x \in V$ 는 다음과 같이 표현됨.

$$x = a_1 v_1 + a_2 v_2 + \cdots + a_n v_n$$

- 이 때 $(a_1, \dots, a_n)$ 은 기저 $\{v_1, \dots, v_n\}$ 에 대한 데이터 x 의 좌표벡터 -> $(\boldsymbol{a_1}, \dots, \boldsymbol{a_n})$ **은 데이터 $\boldsymbol{x}$의 표현형**

기저와 데이터 표현

예시

- 3차원 공간에 벡터 x=(2,3,1) 가 있다고 가정
- 이 벡터는 표준 기저를 기준으로 다음과 같이 표현됨:

  x=**2**·e1+**3**·e2+**1**·e3

  여기서 표준 기저는 :e1=(1,0,0),e2=(0,1,0),e3=(0,0,1)
- 이 때 x의 좌표는 (2,3,1)이며, 이것이 바로 **표준 기저에 대한 기저화 표현**.

기저와 데이터 표현

예시

- 표준 기저가 아닌 다음의 새로운 기저를 설정:
  $v1=(1,1,0), v2=(0,1,1), v3=(1,0,1)$
- 이 기저에 대해 벡터 $x=(2,3,1)$ 를 다음과 같이 표현할 수 있음:
  $x=a1v1+a2v2+a3v3,\ a1=1, a2=2, a3=1$
- 즉, x는 새로운 기저에 대해 좌표가 (1,2,1)이 됨.

직교기저와 데이터 표현

직교기저

- 직교기저(orthogonal basis) 는 서로 직각(90도) 방향을 이루는 벡터들로 구성된 기저
- 기저 $\{v_1, \ldots, v_n\}$ 가 직교기저이면 다음이 성립함:

  $< v_i, v_j > = 0$, for $i \neq j$

- 특히, 모든 벡터의 길이까지 1이면 정규직교기저(orthonormal basis)라 부름:

$$||v_i|| = 1$$

직교기저와 데이터 표현

정규직교기저를 이용한 데이터의 표현형

- 데이터가 $x = a_1 v_1 + a_2 v_2 + \cdots + a_n v_n$ 로 표현되고 기저 $\{v_1, \ldots, v_n\}$ 가 정규직교기저면 다음이 성립함:

$$< x, v_j > = a_j$$

- 기저벡터를 이용한 데이터의 표현 ($< x, y > = x^T y$ 인 경우)

$$\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = a = \boldsymbol{F}x = \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix} x = \begin{bmatrix} v_1^T x \\ \vdots \\ v_n^T x \end{bmatrix}$$

- 기저와 데이터의 각 내적값을 표현형이라 볼 수 있음

직교기저와 데이터 표현

표현형으로 부터 원 데이터의 복원

- ($\boldsymbol{F}$:정규직교기저) $\boldsymbol{F^{-1} = F^T}$ **이 성립함.**
  - $\boldsymbol{FF^T = I}$ ($v_i^T v_j = 0$, for $i \neq j$, $||v_i|| = 1$), $\boldsymbol{F^T F}x = x$ 임을 보일 수 있으므로 $\boldsymbol{F^T F = I}$
  - 즉 $\boldsymbol{FF^T = F^T F = I}$ 이므로 $\boldsymbol{F^{-1} = F^T}$ (역행렬의 유일성)

- $x = \boldsymbol{F^T F x} = [v_1 \quad \cdots \quad v_n] \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix} x = [v_1 \quad \cdots \quad v_n] \begin{bmatrix} v_1^T x \\ \vdots \\ v_n^T x \end{bmatrix} = v_1 v_1^T x + \cdots + v_n v_n^T x$

  $= (v_1 v_1^T + \cdots + v_n v_n^T) x$

직교기저와 데이터 표현

데이터 압축(encoding)과 역변환(decoding)

- 정규직교 기저벡터의 구성행렬 $\boldsymbol{F}$ 가 $\boldsymbol{F^{-1}} = \boldsymbol{F^{T}}$ 다음을 만족하므로
  $x = \boldsymbol{F^{-1}}\,\boldsymbol{F}x = \boldsymbol{F^{T}}\boldsymbol{F}x = \boldsymbol{F^{T}}a$ *(데이터의 복원, 표현형의 역변환)*
- $x = a_1 v_1 + \cdots + a_n v_n$ 이고 어떤 a_i를 0으로 대치하고자 함 (표현형을 간단하게 하고자함)
- 일부의 a_i 가 0으로 대치시킨 데이터 표현형을 $\hat{a} = (\widehat{a_1}, \ldots, \widehat{a_n})$ 이라고 할때 $\hat{x} = \boldsymbol{F^{T}}\hat{a} = \widehat{a_1} v_1 + \cdots + \widehat{a_n} v_n$ 이라고 표시하자
- $\hat{x}$ 와 x 의 유클리디안 제곱거리는?

직교기저와 데이터 표현

데이터 압축(encoding)과 역변환(decoding)

- $\left|\left|\hat{x}-x\right|\right|^2 = (\hat{x}-x)^T(\hat{x}-x) = (\hat{a}-a)^T FF^T(\hat{a}-a) = ||\hat{a}-a||^2$
- 확인:
  - 표현형을 간단하게 하여 복원된 역변환 결과는 무엇인가?
  - 복원결과와 원본데이터의 차이는 무엇인가?
  - 표현형의 차이는 무엇인가?
  - 간단한 표현형에 대한 제약조건을 0이 아닌 원소의 개수로 하였을때, 최적의 $\hat{a}$ 의 선택에 대한 문제를 설정해보아라.

직교기저와 데이터 표현

함수 데이터의 기저

- L_2-공간은 실수 구간 [a,b] 위에서 정의된 **제곱 적분 가능 함수**들의 공간

$$L_2[a,b] = \left\{ f : [a,b] \mapsto R : \int_a^b f(t)^2 dt < \infty \right\}$$

- 이 공간은 내적(inner product) 이 정의된 Hilbert 공간으로, 내적은 다음과 같이 정의함.

$$< f, g >= \int_a^b f(t)g(t)dt$$

직교기저와 데이터 표현

함수 데이터의 기저

- L_2 -공간에서의 기저(basis) 란 다음을 만족하는 함수열임
  - $\varphi_k \epsilon L_2[a,b]$
  - $< \varphi_k, \varphi_j > = 0$ for $k \neq j$, $< \varphi_k, \varphi_k > = 1$
  - 임의의 $f \epsilon L_2[a,b]$ 에 대해서 $\sum_{k=1}^{\infty} a_k^2 < \infty$ 인 $a = (a_1, a_2, \dots)$ 가 존재하여

$$\lim_{N \to \infty} \left\| f - \sum_{k=1}^{N} a_k \varphi_k \right\|_{L_2} = 0$$

직교기저와 데이터 표현

Fourier 기저

- $L_2[-\pi, \pi]$ 의 정규직교기저:

  $\mathcal{B} = \{1, \sin(t), \cos(t), \sin(2t), \cos(2t), \dots\}$

- $\mathcal{B}$ 는 $L_2[-\pi, \pi]$ 의 정규직교 기저를 이룸.

- 즉, 임의의 $f \in L_2[-\pi, \pi]$ 는 $\sum_{k=1}^{\infty} a_k^2 + \sum_{k=1}^{\infty} b_k^2 < \infty$ 인 $a = (a_1, a_2, \dots)$, $b = (b_1, b_2, \dots)$ 가 존재하여

$$f(t) = a_0 + \sum_{k=1}^{\infty} a_k \sin(kt) + b_k \cos(kt)$$

직교기저와 데이터 표현

[질문]Fourier 기저를 이용한 데이터 표현형을 어떻게 구할 것인가?

