다변량 분석

Jinseog Kim

September 18, 2007

1 Introduction

1.1 확률벡터

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = (x_1, x_2, \ldots, x_p)^T$$

기대값(평균벡터):

$$E\mathbf{x} = (Ex_1, Ex_2, \ldots, Ex_p)^T = (\mu_1, \mu_2, \ldots, \mu_p)^T$$

분산(공분산행렬, variance-covariance matrix:

$$\Sigma = V(\mathbf{x}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

여기서 $\sigma_{ii} = Var(x_i), i = 1, \ldots, p,$
$\sigma_{ij} = Cov(x_i, x_j), i \neq j$

So, using the vector-matrix form:

$$\Sigma = \mathrm{E}\left[(\mathbf{x} - \mathrm{E}[\mathbf{x}])\,(\mathbf{x} - \mathrm{E}[\mathbf{x}])^{\top}\right]$$

and

$$\mu = \mathrm{E}(\mathbf{x})$$

Note: The matrix Σ is "positive semi definite(양정치): For all non-zero vectors $\mathbf{z} \in C^p$,

$$\mathbf{z}^* \Sigma \mathbf{z} \geq 0$$

.

All eigenvalues λ_i of Σ are positive.

1.2 공분산행렬의 성질

Assume that
$\mathbf{x}, \mathbf{x_1}$ and $\mathbf{x_2}$ are a random $(p \times 1)$ vectors, $\mathbf{y}$ is a random $(q \times 1)$ vector,
$\mathbf{a}$ is $(\mathbf{p} \times \mathbf{1})$ vector, A and B are $(p \times q)$ matrices.

- $\Sigma = \mathrm{E}(\mathbf{x}\mathbf{x}^\top) - \mu\mu^\top$
- $\boldsymbol{\Sigma}$ is positive-definite matrix (positive semi-definite)
- $\mathrm{var}(A\mathbf{x} + \mathbf{a}) = A\ \mathrm{var}(\mathbf{x})\, A^\top$
- $\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \mathrm{cov}(\mathbf{y}, \mathbf{x})$
- $\mathrm{cov}(\mathbf{x_1} + \mathbf{x_2}, \mathbf{y}) = \mathrm{cov}(\mathbf{x_1}, \mathbf{y}) + \mathrm{cov}(\mathbf{x_2}, \mathbf{y})$
- If $p = q$, then
  $\mathrm{var}(\mathbf{x} + \mathbf{y}) = \mathrm{var}(\mathbf{x}) + 2\,\mathrm{cov}(\mathbf{x}, \mathbf{y}) + \mathrm{var}(\mathbf{y})$
- $\mathrm{cov}(A\mathbf{x}, B\mathbf{y}) = A\ \mathrm{cov}(\mathbf{x}, \mathbf{y})\, B^\top$
- If $\mathbf{x}$ and $\mathbf{y}$ are independent, then

1.3 다변량 확률표본

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ 이 평균이 $\boldsymbol{\mu}_{1\times p}$이고 공분산 행렬이 $\Sigma_{p\times p}$인 다변량 분포에서 얻어진 확률표본이라고 하자. 여기서 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, 즉,

$$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} = X^\top$$

- 표본평균:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \frac{1}{n}X^\top \mathbf{1}_n = (\bar{x}_1, \ldots, \bar{x}_p)^\top, \bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ki}, k = 1, \ldots, p.$$

- 표본공분산행렬:

$$S = (s_{kl}), s_{kl} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ki} - \bar{x}_k)(x_{li} - \bar{x}_l)$$

or

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^\top,$$

or

$$\frac{1}{n-1}X'(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}')X = \frac{1}{n-1}X'\left(I_n - \frac{1}{n}J\right)X.$$

다음과 같은 다변량 정규분포를 고려하자, 이 분포에서 20개의 난수를 발생시키고 발생된 난수에서 표본평균 및 표본 공분산 행렬을 구하라.

$$\boldsymbol{\mu} = (0,1,2)', \Sigma = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 1 \\ 1 & 1 & 5 \end{pmatrix}$$

```
library(MSBVAR)
## 다변량 정규분포에서 난수의 발생

## Calculate Covariance matrix
m<-c(0,1,2)
```

```
V<-matrix(c(3,2,1,2,4,1,1,1,5), ncol=3, nrow=3)

## generate errors from multivariate normal distribution
X<-rmultnorm(20, mu=m, vmat=V)
X<-t(X)

## 표본평균


## 표본분산
```

1.4 다변량 상관계수

확률벡터를 구성하는 변수들 사이의 관계를 측정하는 방법

- 상관계수(correlation coefficients)
- 다중상관계수(multiple correlation coefficients): 한 변수와 다른변수의 그룹간의 상관관계
- 부분상관계수(partial correlation coefficients):여러 개의 확률변수로 이루어진 두 개의 변수 그룹에서 한 그룹의 변수 값이 주어진 경우 다른 그룹내의 변수들간의 관계
- 정준상관계수(canonical correlation coefficients): 여러개의 확률변수로 이루어진 두 변수그룹간의 관계

1.5 상관계수(correlation coefficients)

$$R = D^{-1/2} S D^{-1/2}$$

여기서, $D = \begin{pmatrix} s_{11} & 0 & \dots & 0 \\ 0 & s_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_{pp} \end{pmatrix} = \mathrm{diag}(s_{11}, s_{22}, \dots, s_{pp})$,이고, S는 공분산 행렬 (covariance matrix)이다.

1.6 다중상관계수(multiple correlation coefficients)

다변량 자료가 p개의 확률변수로 이루어져 있다고 가정하자.
p개 중에서 임의로 하나의 변수를 선택하고(이를 x_0)
나머지 $p-1$개에서 $q \le p-1$개를 선택하자(이를 $\mathbf{x}_{(q)}$로 표현하자). 여기서 x_0와 $\mathbf{x}_{(q)}$의 관계를 하나의 값으로 표현하기 위해서 x_0와 $\mathbf{x}_{(q)}$의 선형결합(linear combination), 즉

$$\boldsymbol{\beta}^\top \mathbf{x}_{(q)} = \beta_1 x_{i_1} + \beta_2 x_{i_2} + \ldots + \beta_q x_{i_q},$$

로 표현하면 두 변수간의 상관계수를 구할 수 있다.

이렇게 모든 가능한 경우의 상관관계를 고려하고, 그 중에서 최대치를 다중상관계수라고 부른다.

2 행렬연산의 기초

2.1 행렬의 종류

- 정방행렬(square matrix) 행수와 열수가 같은 행렬 $A_{n\times n}$
- 대각행렬(diagonal matrix) 대각선원소를 제외한 모든 원소가 0인 행렬
- 항등행렬(Identity matrix): 대각행렬 중 대각선 원소가 모두 1인 행렬

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

- 정방 행렬의 원소가 모두 1인 행렬

$$J = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \end{pmatrix} = \mathbf{1}\mathbf{1}'$$

- 전치행렬(Transpose matrix): 행렬 A의 원소 a_{ij}의 행번호와 열번호를 바꾼 행렬.

$$A' = A^{\top} = (a_{ji})$$

- 대칭행렬(Symmetric matrix):

$$A' = A.$$

- 역행렬(Inverse matrix): 행렬 A에 대하여, 어떤 행렬 B존재하여, 다음을 만족할 때, B를 역행렬이라고 하고 $B = A^{-1}$로 표현한다.

$$AB = BA = I.$$

- 직교행렬(Orthogonal matrix):

$$AA' = A'A = I, \text{혹은 } A' = A^{-1}.$$

2.2 선형변환

행렬 A, 열벡터 $\mathbf{b}$,

$$\mathbf{y} = A\mathbf{x} + \mathbf{b}$$

를 $\mathbf{x}$의 $\mathbf{y}$로의 선형변환(linear transformation)이라고 한다.

- If $E(\mathbf{x}) = \boldsymbol{\mu}$, $Cov(\mathbf{x}) = \Sigma_{\mathbf{x}}$, then

$$E(\mathbf{y}) = A\boldsymbol{\mu}_x + \mathbf{b},\ Cov(\mathbf{y}) = A\Sigma_{\mathbf{x}}A^{\top}.$$

- **Affine transformation**
  만일 A가 non-singular matrix(정칙행렬, 역행렬이 존재하는 경우)
  이면, $\mathbf{y} = A\mathbf{x} + \mathbf{b}$를 Affine transformation이라고 부른다.

2.3 선형독립(linearly independent)

- n개의 열벡터 $\mathbf{x}_1, \ldots, \mathbf{x}_n$가 0이 아닌 상수($a_i$)와의 선형결합

$$\mathbf{a}'\mathbf{x} = a_1\mathbf{x}_1 + \ldots + a_n\mathbf{x}_n = 0$$

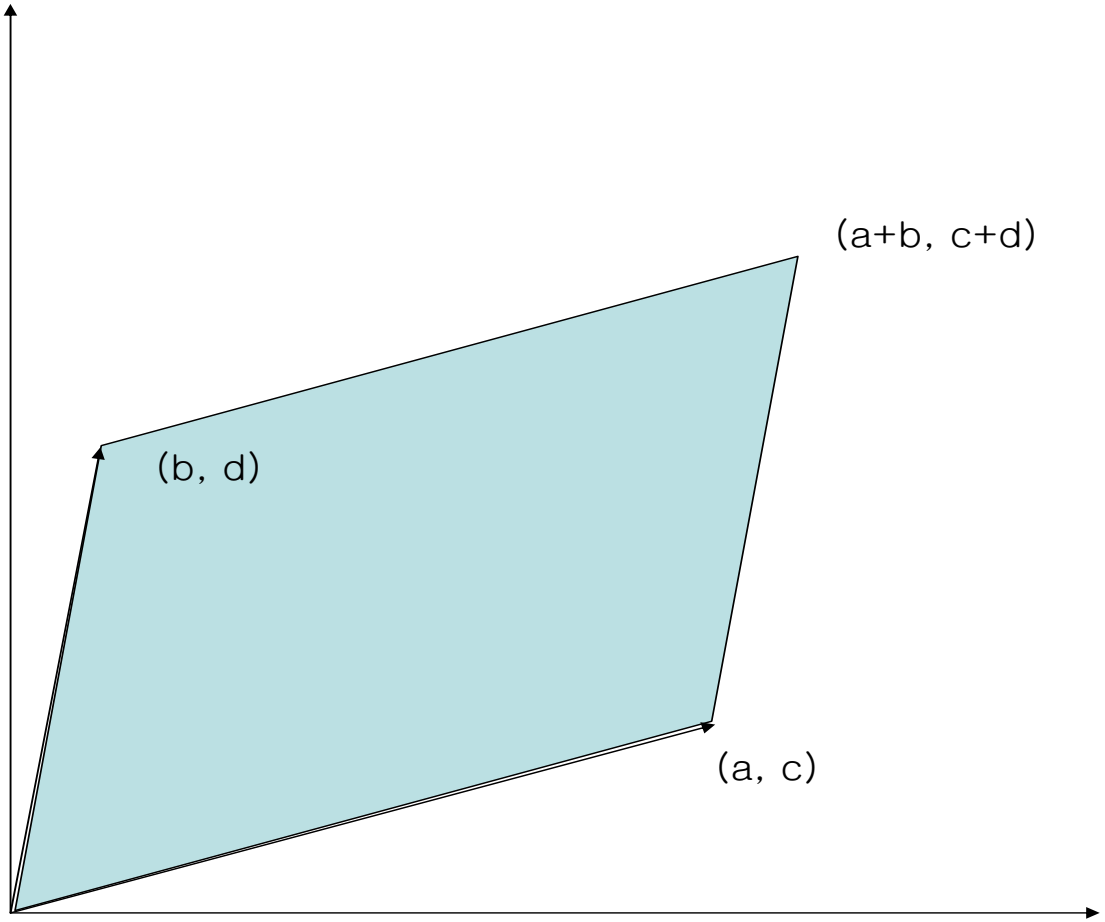
  을 만족하면 $\mathbf{x}_1, \ldots, \mathbf{x}_n$를 선형종속(linearly dependent)라고 부르며, $\mathbf{a} = \mathbf{0}$이면 선형독립이라고 부른다.

- 행렬 X는 ,열벡터들 $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$으로 이루어 졌다고 할때, 선형독립인 열의 개수를 행렬 X의 계수(rank)라고 한다.

2.4 행렬식(determinant)

행렬식(determinant)는 정방 행렬 $A_{n\times n}$를 scalar값으로 mapping(대응)시킴. ($\det(A)$ 혹은 $|A|$로 표현)

의미: 행렬 A를 n차원 공간에서의 n개의 벡터라고 할 때, 그 벡터들로 만들어지는 도형의 부피(Volumn)의 개념으로 이해할 수 있다.


(a+b, c+d)
(b, d)
(a, c)

2.5 trace

정방행렬의 대각선 원소의 합을 행렬의 trace라고 한다.

$$tr(A+B) = tr(A) + tr(B)$$

$$tr(aA) = a \times tr(A)$$

$$tr(A) = tr(A')$$

.

If $A_{n\times n}$ matrix and $B_{n\times n}$ matrix, then

$$tr(AB) = tr(BA).$$

$$\mathrm{tr}(AB) = \sum_{i=1}^{n}(AB)_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{n} B_{ji}A_{ij} = \sum_{j=1}^{n}(BA)_{jj} = \mathrm{tr}(BA)$$

$$tr(ABC) = tr(CAB) = tr(BCA)$$

2.6 이차형식(Quadratic form)

$A_{p\times p}$ 가 대칭행렬(symmetric)이고 $\mathbf{x}$ 가 p 열벡터 일때,

$$Q(\mathbf{x}) = \mathbf{x}'A\mathbf{x}$$

를 $\mathbf{x}$의 A에 대한 이차형식이라고 부른다.

- 만일 $\mathbf{x}'A\mathbf{x} > 0$ (resp. $\mathbf{x}'A\mathbf{x} < 0$) for every vector $\mathbf{x} \neq 0$ 이면 A를 "양정치(positive definite)" (resp. 음정치(negative definite))행렬 이라고 부른다.
- If we change the strict inequality into $\geq; \leq$, A를 "semi-definite"(즉, 양반정치, 음반정치) 행렬 이라고 부른다.
- If $Q(v) < 0$ for some v and $Q(v) > 0$ for some other v, Q is said to be "indefinite".

이차형식은 χ^2 분포, mahalanobis 거리와 같은 데서 이용된다.

2.7 고유치, 고유벡터-eigen value and eigen vector

주어진 행렬 A 에 대하여 어떤 벡터($\mathbf{x}$)의 선형변환이 자기 자신($\mathbf{x}$)과 어떤 상수(λ)의 곱으로 표현할 수 있다고 하자. 즉,

$$A\mathbf{x} = \lambda\mathbf{x}.$$

이는

$$A\mathbf{x} = (\lambda I)\mathbf{x},$$

또는

$$(A - \lambda I)\mathbf{x} = 0$$

로 표현될 수 있다.

이 때, eigenvector는 위 식을 만족하는 $\mathbf{0}$이 아닌 벡터, eigenvalue는 역시 0이 아닌 스칼라 값이다. 다음의 예를 살펴보자.

$$A = \begin{bmatrix} 0 & 0 \\ -\frac{1}{2} & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1-\lambda & 0 \\ -\frac{1}{2} & 1-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

여기서 eigen values와 eigen-vector는 $\lambda = 1, \mathbf{x} = \begin{bmatrix} 0 \\ c \end{bmatrix}$

See page 43–46 for the properties of eigen-value and eigen vectors.

2.8 멱등행렬(idempotent matrix)

$$A^2 = AA = A.$$

Example: 회귀분석에서 hat matrix $H = X(X'X)^{-1}X'$.

종속변수의 관측치를 행렬로 표현하면 $\mathbf{y}$이고, 회귀계수의 추정치는 $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$이다. 이 때, $\mathbf{y}$의 추정치는

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\mathbf{y} = H\mathbf{y}.$$

또한 잔차벡터는

$$\mathbf{y} - X\hat{\boldsymbol{\beta}} = [I - X(X'X)^{-1}X']\mathbf{y} = (I - H)\mathbf{y}.$$

$$H^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = H.$$

이므로 $X(X'X)^{-1}X'$는 멱등행렬이다.