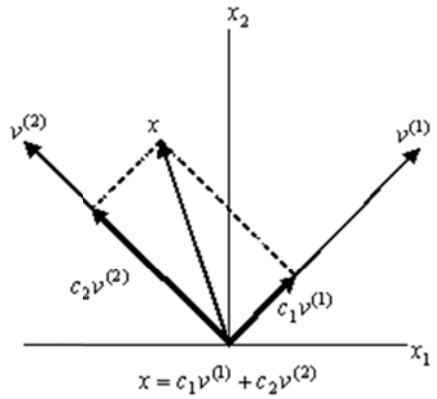
구하기

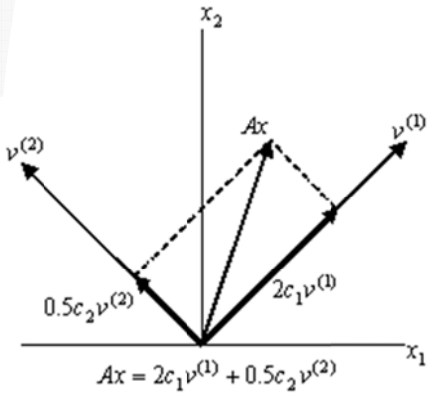
행렬 A의 고유 방정식(characteristic equation) $|A_{n\times n} - \lambda I_n| = \underline{0}$를 만족하는 $\lambda_1, \lambda_2, ..., \lambda_n$들을 고유치(eigen value, characteristic value, latent value)라 하고 각 고유치에 대해 $A\underline{e}_i = \lambda_i \underline{e}_i$를 만족하는 벡터를 고유 벡터(eigen vector)라 한다. 여기서는 고유치와 고유 벡터 계산 방법에 대해서는 다루지 않겠지만 언급하고 싶은 것은 고유치와는 달리 고유 벡터는 무수히 많이 존재한다는 것이다. 이 성질 때문에 요인 분석에서 요인 회전이 가능하다.

고유치가 사용되는 예제를 보면 $\underline{n}_t$를 시점 t 에서의 각 연령 분포 벡터라 하고 행렬 A 를 각 연령에서 시점 t 에서 (t+1)까지 살아 남을 확률에 대한 행렬이라면 $\underline{n}_{t+1} = A\underline{n}_t$ 이다. 만약 시간에 따라 인구의 분포가 stable 하다면 $\underline{n}_{t+1} = \lambda \underline{n}_t$ 가 될 것이고 $A\underline{n}_t = \lambda \underline{n}_t$ 이 될 것이다.

개념

또한 고유 벡터를 구하는 식에서 $A\underline{e}_i = \lambda_i \underline{e}_i$ 을 살펴보면 행렬 A 의 크기 고유치에 나타나 있다. 즉 고유치는 행렬 A 의 크기(변동)에 대한 정보이다. 이 개념이 다변량 분석에 이용된다.





Professor Kwon, Sehyug | Dept. of Statistics, HANNAM University

통계학 활용

대칭 행렬에 대해서는 다음이 성립한다. 우리가 다변량에서 이용하게 될 공분산 행렬이나 상관 행렬은 모두 대칭 행렬이므로 알아 두면 유용한 성질이다.

(1)고유치는 실수이다.

(2)대칭 행렬은 대각화가 가능하다(Diagnosable). $A = U^{-1}DU$ D 는 대각원소가 A 의 고유치인 대각 행렬이고 U 는 직교 행렬이다.

(3)고유 벡터는 orthogonal 하다. 즉 $e_i' e_j = 0 \; for \; i \neq j$

(4)행렬의 계수와 0 이 아닌 고유치의 수는 같다. 즉 0 인 고유치가 존재하는 행렬은 full-rank 가 아니며 역 행렬이 존재하지 않는다.

평균벡터

자료 행렬 $X_{n\times p} = \begin{bmatrix} x_{11}\, x_{12} \cdots x_{1p} \\ x_{21}\, x_{22} \cdots x_{2p} \\ \cdots\cdots x_{ij} \cdots\cdots \\ x_{n1}\, x_{n2} \cdots x_{np} \end{bmatrix} = [\underline{x}_1 \; \underline{x}_2 \; \underline{x}_3 \cdots \underline{x}_p]$, $\underline{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{bmatrix}_{n\times 1}$

II EXAMPLE II $X_{4\times 3} = \begin{bmatrix} 2 & 3 & 2 \\ 4 & 2 & 3 \\ 2 & 1 & 1 \\ 5 & 2 & 2 \end{bmatrix} = [\underline{x}_1 \; \underline{x}_2 \; \underline{x}_3]$

다음을 자료의 평균 벡터(mean vector)라 한다.

$$\underline{\bar{x}} = (1/n)\underline{1}'_n X_{n\times p} = (1/n)[1\ 1 \ldots 1] \begin{bmatrix} x_{11}\, x_{12} \cdots x_{1p} \\ x_{21}\, x_{22} \cdots x_{2p} \\ \cdots\cdots x_{ij} \cdots\cdots \\ x_{n1}\, x_{n2} \cdots x_{np} \end{bmatrix} = \begin{bmatrix} \sum x_{i1}/n \\ \sum x_{i2}/n \\ \cdot \\ \sum x_{ip}/n \end{bmatrix}$$

II EXAMPLE II 가족 수(X_1), 학벌 (X_2), 하루 용돈(X_3) 변수의 평균 벡터를 구하시오.

$$\underline{\bar{x}} = (1/4)[1\ 1\ 1\ 1] \begin{bmatrix} 2 & 3 & 2 \\ 4 & 2 & 3 \\ 2 & 1 & 1 \\ 5 & 2 & 2 \end{bmatrix} = (1/4)[13\ 8\ 8] = [3.25\ 2\ 2]$$

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM University

공분산행렬과 상관행렬

j 번째 변수와 k 번째 변수의 공분산은 $\sigma_{jk} = \mathrm{cov}(x_j\ x_k) = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_{ij} - \bar{x_i})(x_{ik} - \bar{x_k})$

$j = k$ 이면 공분산은 분산이므로 $\sigma_{kk} = \mathrm{var}(x_k) = \mathrm{cov}(x_k, x_k) = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_{ik} - \bar{x_k})^2$

그러므로 공분산 행렬(covariance matrix)은 $\Sigma = \begin{bmatrix} \sigma_{11}\ \sigma_{12} \cdots \sigma_{1p} \\ \sigma_{21}\ \sigma_{22} \cdots \sigma_{2p} \\ \vdots \\ \sigma_{p1}\ \sigma_{p2} \cdots \sigma_{pp} \end{bmatrix}$ 이다.

만약 표본 데이터인 경우에는 σ 대신 s 을 사용하고 Σ 기호 대신 S 사용한다. j 번째 변수와 k 번째 변수의 상관 계수는 $r_{jk} = \mathrm{cov}(x_j\ x_k) / \mathrm{var}(x_j)\,\mathrm{var}(x_k)$ 이므로 상관 계수 행렬(correlation matrix) $R = \begin{bmatrix} r_{11}\ r_{12} \cdots r_{1p} \\ r_{21}\ r_{22} \cdots r_{2p} \\ \vdots \\ r_{p1}\ r_{p2} \cdots r_{pp} \end{bmatrix}$ 이다.

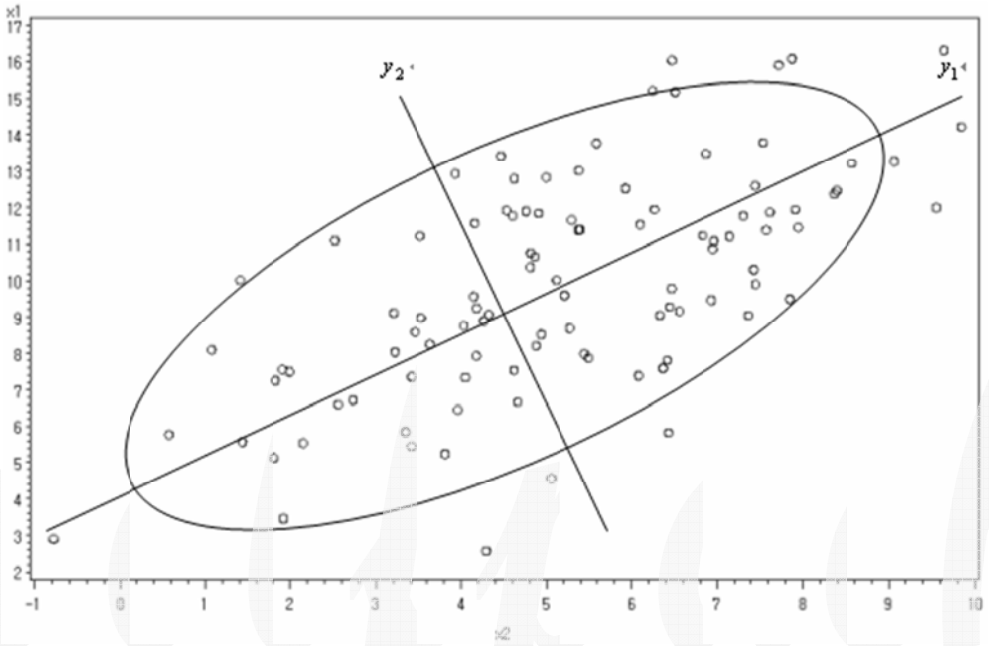
공분산 행렬 Σ 의 고유치는 $|\Sigma - \lambda I| = 0$ 의 해이다. 공분산 행렬은 대칭이므로 모든 고유치 값은 실수이고 행렬 차수만큼의 고유치가 존재한다. 행렬 Σ 의 고유치를 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ 하면 행렬 Σ 의 고유치 λ_i 대해 $\Sigma \underline{u}_i = \lambda_i \underline{u}_i$ 을 만족하는 u_i 을 고유벡터라 한다. 고유 벡터는 수없이 많이 존재하는데 다변량에서는 고유치 λ_j 에 대응하는 고유 벡터를 $\underline{u}_j$ 라 하면 $\underline{u}_i' \underline{u}_i = 1$ 이고 $\underline{u}_i' \underline{u}_j = 0$ for $\lambda_i \neq \lambda_j$ 을 만족하는 고유벡터를 사용한다.

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM University

고유치의 기하학적 해석

평균 $\underline{\mu}=\begin{bmatrix}10\\5\end{bmatrix}$이고 공분산-분산 행렬 $\Sigma=\begin{bmatrix}9 & 2\\2 & 4\end{bmatrix}$인 이변량 정규분포를 고려하자.

$\Sigma=\begin{bmatrix}9 & 2\\2 & 4\end{bmatrix}$의 고유치는 $\lambda_1=9.7$, $\lambda_2=3.2$이고 각 고유치에 대응하는 고유 벡터는 $\underline{a}_1=\begin{bmatrix}0.94\\0.33\end{bmatrix}$, $\underline{a}_2=\begin{bmatrix}-0.33\\0.94\end{bmatrix}$이다.




```
data one;
    do i=1 to 100;
    z1=rannor(0);
    z2=rannor(0);
    x1=10+3*z1;
    x2=5+2/3*2*z1+2*sqrt(1-(2/3)*(2/3))*z2;
    output;
    end;
run;

proc gplot data=one;
    symbol v=circle;
    plot x1*x2;
run;
```

이변량 자료를 추출하여 산점도를 그리면 위 그림과 같다. 타원(ellipse)의 방정식은 $\frac{y_1^2}{\lambda_1}+\frac{y_2^2}{\lambda_2}=3(3=p+1)$이고 타원 긴 쪽 길이는 $2\sqrt{3\lambda_1}=10.8$, 타원 짧은 쪽 길이는 $2\sqrt{3\lambda_2}=6.2$이다. 또한 y_1의 방향에서 분산은 $\lambda_1=9$이고 y_2의 방향에서 분산은 $\lambda_2=4$

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM University

이다. 그러므로 만약 λ_2가 0 이 가까우면 자료는 직선 상에 모이게 된다. 0 이 되면 자료가 모두 직선 상에 모이므로 두 변수의 상관 계수는 0 이다. $\lambda_1 = \lambda_2 = 0.5$ 이면 자료 산점도의 형태는 원이 되고 상관 계수는 0 이다. 이처럼 고유치의 값은 두 변수 간의 상관 관계를 나타내는 지표가 되며 다변량은 이를 일반화한 것이다. y_1과 y_2는 주성분이며 타원의 길이는 주성분의 원 변수 변동에 대한 설명력이다.

```
data one;
    input x1 x2;
    cards;
    1 3
    3 7
    5 9
    7 13
;
run;
proc corr data=one cov outp=out1;
run;
proc print data=out1;
run;
```

데이터에는 변수가 2 개 관측치가 4 개이다. COV 옵션에 의해 공분산을 구한다. 그 결과를 OUT1 에 저장한다.

TYPE	_NAME_	x1	x2
COV	x1	6.6667	10.6667
COV	x2	10.6667	17.3333
MEAN		4.0000	8.0000
STD		2.5820	4.1633
N		4.0000	4.0000
CORR	x1	1.0000	0.9923
CORR	x2	0.9923	1.0000

```
data out2;
    set out1;
    if _type_="COV";
    keep x1 x2;
run;

proc print data=out2;
run;
```

자동 생성 변수 _TYPE_에서 "COV"인 관측치만 선택하고 변수는 X1, X2 만 남긴다.

x1	x2
6.6667	10.6667
10.6667	17.3333

```
proc iml;
reset print;
use out2;
read all into x;
call eigen(m,e,x);
run;
```

USE 옵션은 SAS data 를 사용하여 행렬을 만들 때 사용한다. READ 는 SAS data 데이터를 into 행렬 X 로 만들라는 의미이다. M 은 고유치, E 는 고유벡터를 출력한다.

```
X     2 rows     2 cols     (numeric)
        6.6666667 10.666667
        10.666667 17.333333

M     2 rows     1 col      (numeric)
              23.925696
              0.0743041

E     2 rows     2 cols     (numeric)
        0.5257311 0.8506508
        0.8506508 -0.525731
```


Professor Kwon, Sehyug | Dept. of Statistics, HANNAM University