Introduction to Linear Algebra
- Gilbert Strang

17

Linear Algebra in Probability and Statistics

ART NATURE

Note Book

sooraj ss1729@gmail.com +91-9400635788

INDEX

(17)

Name : SOORAJ. S. Subject : .....

Std. : ..... Div. : ..... Roll No. : .....

School / College : .....

S. No.	Date	Title	Page No.	Teacher's Sign / Remarks
		INTRODUCTION TO LINEAR ALGEBRA - Gilbert Strang, MIT (5th edition)		

Covariance Matrices & Joint Probabilities

Linear algebra enters when we run M different experiments at once. We might measure age, height and weight ($M = 3$ measurements of N people). Each expt. has its own mean value.

So we have a vector $m = (m_1, m_2, m_3)$ containing the m mean values. These could be sample mean of age & height and weight. Or m_1, m_2, m_3 could be expected values of age, height and weight based on known probabilities.

A matrix becomes involved when we look at variances. Each expt will have a sample variance S_i^2 or an expected $\sigma_i^2 = E\left[(x_i - m_i)^2\right]$ based on the squared distance from its mean. Those M numbers $\sigma_1^2, \ldots, \sigma_M^2$ will go on the main diagonal of the matrix.

Those M parallel experiments measure M different random variables, but the expts are not necessarily independent.

If we measure age, height & weight (a, h, w) for children, the results will be strongly correlated: Older children are generally taller and heavier.

Covariance:

$$\sigma_{ah} = E\left[(\text{age} - \text{mean age})(\text{height} - \text{mean height})\right]$$

To compute σ_{ah}, we have to know the joint probability of each pair (age & height). This is because age is connected to height.

P_{ah} : probability that a random child has age = a and height = h : both at once

P_{ij} : probability that expt 1 produces x_i and expt. 2 produces y_j

Suppose,
expt 1 (age) has mean m_1. Expt 2 (height) has mean m_2. The covariance b/w expts 1 and 2 looks at all pairs of ages x_i, heights y_j :

Covariance,

$$\sigma_{12} = \sum_{\text{all}} \sum_{i,j} P_{ij} (x_i - m_1)(y_j - m_2)$$

Ex:1. Flip 2 coins separately. With 1 for heads and 0 for tails, the results can be (1,1) or (1,0) or (0,1) or (0,0). Those 4 outcomes all have probability $P_{11} = P_{10} = P_{01} = P_{00} = \frac{1}{4}$

> Independent experiments have,
>
> By $P(i,j) = P(i)\,P(j)$

Prob. matrix, $P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix}$

Ex:2. Glue the coins together, facing the same way. The only possibilities are (1,1) and (0,0). Those have probabilities $\frac{1}{2}$ and $\frac{1}{2}$.

$P_{01} = P_{10} = 0$

(1,0) & (0,1) won't happen because the coins stick together:

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$$

Probability matrix,

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \qquad \begin{matrix} P_{11}+P_{12} = P_1 \\ P_{21}+P_{22} = P_2 \end{matrix} \quad (1^{st} \text{ coin})$$

$\quad P_1 \quad P_2$

$(2^{nd}$ coin) columns sum

These numbers P_1, P_2 and P_1, P_2 are called the marginals of the matrix P.

$P_1 = P_{11} + P_{12}$ = chance of heads from coin 1.
(coin 2 can be heads or tails)

$P_1 = P_{11} + P_{21}$ = chance of heads from coin 2
(coin 1 can be heads or tails).

Zero covariance σ_{12}

for independent trials.

$$V = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} = \text{diagonal covariance matrix}$$

* Independent experiments have $\sigma_{12} = 0$ because

$$P_{ij} = (P_i)(P_j)$$

$$\sigma_{12} = \sum_i \sum_j P_{ij} (x_i - m_1)(y_j - m_2)$$

$$= \sum_i \sum_j (P_i)(P_j)(x_i - m_1)(y_j - m_2)$$

$$= \left[\sum_i P_i (x_i - m_1)\right]\left[\sum_j P_j (y_j - m_2)\right]$$

$$= [0][0] = 0$$

since sum of deviations from the mean is zero.

The glued coins show perfect correlation. Heads on one means heads on the other.

The covariance σ_{12} moves from 0 to $\sigma_1 \sigma_2 = \frac{1}{4}$ – which is the largest possible value of σ_{12}:

Means $= \frac{1}{2} + \cdot$ (for both)

$$\sigma_{12} = \frac{1}{2}\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right) + 0 + 0 + \frac{1}{2}\left(0-\frac{1}{2}\right)\left(0-\frac{1}{2}\right) = \frac{1}{4}$$

Heads or tails from coin 1 gives complete information about heads or tails from coin 2:

Glued coins give largest possible covariances

Singular covariance matrix : determinant = 0

$$V_{glue} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

* $\sigma_1^2\sigma_2^2 \geq \sigma_{12}^2 \Longrightarrow -\sigma_1\sigma_2 \leq \sigma_{12} \leq \sigma_1\sigma_2$

Proof

Cauchy - Schwarz inequality

$$\left(\sum_i (x_i - m_1)^2\right)\left(\sum_j (y_j - m_2)^2\right) \geq \left(\sum_{i,j} (x_i - m_1)(y_j - m_2)\right)^2$$

□ The Covariance Matrix V is +ve definite

Expected covariance σ_{12} b/w 2 experiments 1 and 2 (2 coins):

σ_{12} = expected value of [(o/p 1 − mean 1) times (o/p 2 − mean 2)]

$$= \sum_{\text{all } i,j} \sum P_{ij} (x_i - m_1)(y_j - m_2)$$

$P_{ij} \geq 0$ is the probability of seeing o/p x_i in expt 1 and y_j in expt. 2.

Total probability (all pairs) is 1

$$\sum_{\text{all } i,j} \sum P_{ij} = 1$$

Fix on one particular o/p x_i in expt 1.

Allow all o/p y_j in expt 2.

Add the probabilities of $(x_i, y_1), (x_i, y_2), \ldots, (x_i, y_n)$:

Row sum P_i of P,

$$\sum_{j=1}^{n} P_{ij} = \text{probability } P_i \text{ of } x_i \text{ in expt. 1}$$

Some y_j must happen in expt 2.

Whether the 2 coins are completely separate or glued together, we still get $\frac{1}{2}$ for the probability $P_H = P_{HH} + P_{HT}$ that coin 1 is heads:

(separate) $P_{HH} + P_{HT} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

(glued) $P_{HH} + P_{HT} = \frac{1}{2} + 0 = \frac{1}{2}$

Covariance matrix,

$$V = \sum\sum V_{ij}$$

$$= \sum_{\text{all } i,j}\sum P_{ij} \begin{bmatrix} (x_i - m_1)^2 & (x_i - m_1)(y_j - m_2) \\ (x_i - m_1)(y_j - m_2) & (y_j - m_2)^2 \end{bmatrix}$$

$$= \sum_{\text{all } i,j}\sum P_{ij} \begin{bmatrix} x_i - m_1 \\ y_j - m_2 \end{bmatrix} \begin{bmatrix} x_i - m_1 & y_j - m_2 \end{bmatrix}$$

$$(X - \bar{X}) \quad (X - \bar{X})^T$$

$$V_{11} = \sum_{\text{all } i,j}\sum P_{ij} (x_i - m_1)^2$$

$$= \sum_{\text{all } i} (\text{probability of } x_i)(x_i - m_1)^2 = \sigma_1^2$$

V_{ij} has diagonal entries $P_{ij}(x_i - m_1)^2 \geq 0$ and $P_{ij}(y_j - m_2)^2 \geq 0$ and $\det(V_{ij}) = 0$.

$\Rightarrow \operatorname{rank}(V_{ij}) = 1$

$$\begin{bmatrix} (x_i-m_1)^2 & (x_i-m_1)(y_j-m_2) \\ (x_i-m_1)(y_j-m_2) & (y_j-m_2)^2 \end{bmatrix} = \begin{bmatrix} x_i-m_1 \\ y_j-m_2 \end{bmatrix} \begin{bmatrix} x_i-m_1 & y_j-m_2 \end{bmatrix}$$

$$= UU^T$$

$$x^T(UU^T)x = (U^Tx)^T(U^Tx) = \|U^Tx\|^2 \geq 0$$

ILA (8)

$\Rightarrow$ Every matrix UU^T is positive semidefinite.

& also, if S & T are symmetric positive definite, so is $S+T$

$\therefore$

The whole matrix V (combining these matrices UU^T with weights $P_{ij} \geq 0$) is at least semidefinite

* The covariance matrix V is positive definite unless the experiments are dependent.

The o/p from each trial is a vector X with M components.

The covariance matrix is now M by M.

V is created from the o/p vectors X and their average $\bar{X} = E[X]$:

Covariance matrix,

$$V = E\left[(X - \bar{X})(X - \bar{X})^T\right]$$

XX^T & $\bar{X}\bar{X}^T$ = (column)(row) are M by M matrices.

Take any linear combination $c^T X = c_1 X_1 + \cdots + c_M X_M$.

Linearity $\Longrightarrow$ $E[c^T X] = c^T E[X] = c^T \bar{X}$

$$\mathrm{Var}[c^T X] = E\left[(c^T X - c^T \bar{X})(c^T X - c^T \bar{X})^T\right]$$

$$= c^T E\left[(X - \bar{X})(X - \bar{X})^T\right] c = c^T V c$$

$\mathrm{Var}[c^T X] \geq 0 \Longrightarrow c^T V c \geq 0$

$\therefore$ The covariance matrix V is therefore positive semidefinite. By the energy test $c^T V c \geq 0$.

$$V = \sum_{\text{all } x,y,z} P_{x,y,z} \, U U^T$$

$$= \sum_{\text{all } x,y,z} P_{x,y,z} \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \\ z-\bar{z} \end{bmatrix} \begin{bmatrix} x-\bar{x} & y-\bar{y} & z-\bar{z} \end{bmatrix}$$

$$= \sum_{\text{all } x,y,z} P_{x,y,z} \begin{bmatrix} (x_i-\bar{x})^2 & (x-\bar{x})(y-\bar{y}) & (x-\bar{x})(z-\bar{z}) \\ (x_i-\bar{x})(y-\bar{y}) & (y-\bar{y})^2 & (y-\bar{y})(z-\bar{z}) \\ (x-\bar{x})(z-\bar{z}) & (y-\bar{y})(z-\bar{z}) & (z-\bar{z})^2 \end{bmatrix}$$

$$= E\left[(X-\bar{X})(X-\bar{X})^T\right] = E\left[U U^T\right]$$

where,

$$U = X - \bar{X} = \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \\ z-\bar{z} \end{bmatrix}$$

The value of the expectation symbol E is that it also allows pdf's (probability density functions like $p(x,y,z)$ for continuous random variables x, y and z).

i.e., if we allow all numbers as ages, heights, and weights, instead of age $i = 0, 1, 2, \ldots$ then we need $p(x,y,z)$ instead of $P_{i,j,k}$.

Covariance matrix,

$$V = E[U U^T]$$

$$= \iiint p(x,y,z)\, U U^T \, dx\, dy\, dz$$

with $$U = X - \bar{X} = \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \\ z - \bar{z} \end{bmatrix}$$

where, $\iiint P(x,y,z)\,dx\,dy\,dz = 1$

Independent variables x, y, z

$$P(x,y,z) = P_1(x)\,P_2(y)\,P_3(z)$$

Dependent variables x, y, z

$P(x,y,z) = 0$ except when $ax + dy + ez = 0$

□ Diagonalization of the Covariance Matrix

Covariance matrices V opens up the link b/w probability and linear algebra.

For problems with many simultaneous random variables, put them into vectors:

$$\vec{X} = \begin{bmatrix} R \\ S \end{bmatrix}, \quad \vec{Y} = \begin{bmatrix} T \\ U \\ V \end{bmatrix}$$

the covariance matrix,

$$Cov(\vec{X}, \vec{Y}) = E\left[(\vec{X} - E(\vec{X}))(\vec{Y} - E(\vec{Y}))^T\right]$$

$$= \begin{bmatrix} cov(R,T) & cov(R,U) & cov(R,V) \\ cov(S,T) & cov(S,U) & cov(S,V) \end{bmatrix}$$

Often there is one vector with all the variables:

$$\vec{X} = \begin{bmatrix} R \\ S \\ T \end{bmatrix}$$

$$Cov(\vec{X}) = Cov(\vec{X}, \vec{X})$$

$$= E\left[\left(\vec{X} - E(\vec{X})\right)\left(\vec{X} - E(\vec{X})\right)^T\right]$$

$$= \begin{bmatrix} Cov(R,R) & Cov(R,S) & Cov(R,T) \\ Cov(S,R) & Cov(S,S) & Cov(S,T) \\ Cov(T,R) & Cov(T,S) & Cov(T,T) \end{bmatrix}$$

$$= \begin{bmatrix} Var(R) & Cov(R,S) & Cov(R,T) \\ Cov(S,R) & Var(S) & Cov(S,T) \\ Cov(T,R) & Cov(T,S) & Var(T) \end{bmatrix}$$

$$Cov(\vec{X}, \vec{Y}) = Cov(\vec{Y}, \vec{X})^T$$

$$cov(A\vec{X}+B, \vec{Y}) = A\, cov(\vec{X}, \vec{Y})$$

$$cov(\vec{X}, C\vec{Y}+D) = Cov(\vec{X}, \vec{Y})\, C^T$$

$$cov(A\vec{X}+B) = A\, cov(\vec{X})\, A^T$$

$$Cov(\vec{X}_1+\vec{X}_2, \vec{Y}) = Cov(\vec{X}_1, \vec{Y}) + Cov(\vec{X}_2, \vec{Y})$$

$$Cov(\vec{X}, \vec{Y}_1+\vec{Y}_2) = Cov(\vec{X}, \vec{Y}_1) + Cov(\vec{X}, \vec{Y}_2)$$

A, C : constant matrices

~~$A = C$~~

$\vec{B}, \vec{D}$: constant vectors

A key question in analyzing data analysis is to figure out which variables are responsible for most of the variation in the data.

These may not be the variables that are measured, but may instead be some linear combination of the measured variables.

⇒ Mathematically, this corresponds to diagonalizing the covariance (or correlation) matrix.

ie.,

The diagonalization of covariance matrices is used to transform correlated random variables into uncorrelated random variables.

* Covariance matrix V is real-symmetric & +ve definite (or atleast semidefinite), so diagonalization $V = Q \Lambda Q^T$ finds real, +ve eigenvalues $\lambda_k = \sigma_k^2 \geq 0$ and an orthonormal basis of eigenvectors q_1 to q_M.

* The eigenvectors form a coordinate system in which the covariance matrix V becomes diagonal.

ie., coordinates in which the variables are uncorrelated.

* The diagonal entries in this coordinate system, the eigenvalues, are the variances of these uncorrelated components.

* The process of diagonalizing the covariance matrix is called principal component analysis (PCA).

Ex:- (Decorrelation of random vectors)

A random vector $X=(X_1,X_2,X_3)^T$ has covariance matrix

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

Design a non-trivial transformation that will generate from X a new random vector Y whose components are uncorrelated.

Ans: $|K_x-\lambda I|=0=\begin{vmatrix}2-\lambda & -1 & 1\\ -1 & 2-\lambda & 0\\ 1 & 0 & 2-\lambda\end{vmatrix}$

$$(2-\lambda)^3+1(\lambda-2)+(\lambda-2)=0.$$

$$\Rightarrow (2-\lambda)\left[\lambda^2-4\lambda+4-2\right]=(2-\lambda)(\lambda^2-4\lambda+2)=0$$

$\Delta=16-8=8$

$\frac{4\pm 2\sqrt{2}}{2}$

$\lambda_1=2,\ \lambda_2=2+\sqrt{2},\ \lambda_3=2-\sqrt{2}$

$$V_1=\begin{bmatrix}0\\ 1/\sqrt{2}\\ 1/\sqrt{2}\end{bmatrix},\ V_2=\begin{bmatrix}1/\sqrt{2}\\ -1/2\\ 1/2\end{bmatrix},\ V_3=\begin{bmatrix}1/\sqrt{2}\\ 1/2\\ -1/2\end{bmatrix}$$

$$V = Q \Lambda Q^T$$

$$A = Q^{-1} = Q^T = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}^T$$

$$= \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/2 & 1/2 \\ 1/\sqrt{2} & 1/2 & -1/2 \end{bmatrix}^T$$

$$= \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/2 & 1/2 \\ 1/\sqrt{2} & 1/2 & -1/2 \end{bmatrix}$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = AX = \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/2 & 1/2 \\ 1/\sqrt{2} & 1/2 & -1/2 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}}X_2 + \frac{1}{\sqrt{2}}X_3 \\ \frac{1}{\sqrt{2}}X_1 - \frac{1}{2}X_2 + \frac{1}{2}X_3 \\ \frac{1}{\sqrt{2}}X_1 + \frac{1}{2}X_2 - \frac{1}{2}X_3 \end{bmatrix}$$

The covariance of Y is given by,

$$K_y = Q^T K_x Q = \frac{1}{4} \begin{bmatrix} 0 & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -1 & 1 \\ \sqrt{2} & 1 & -1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0 & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -1 & 1 \\ \sqrt{2} & 1 & -1 \end{bmatrix}$$

$$= \frac{1}{4} \begin{bmatrix} 0 & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -1 & 1 \\ \sqrt{2} & 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 2\sqrt{2}+2 & 2\sqrt{2}-2 \\ 2\sqrt{2} & -\sqrt{2}-2 & -\sqrt{2}+2 \\ 2\sqrt{2} & \sqrt{2}+2 & \sqrt{2}-2 \end{bmatrix}$$

$$= \frac{1}{4} \begin{bmatrix} 8 & 0 & 0 \\ 0 & 8+4\sqrt{2} & 0 \\ 0 & 0 & 8-4\sqrt{2} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2+\sqrt{2} & 0 \\ 0 & 0 & 2-\sqrt{2} \end{bmatrix}$$

Mean & Variance of $z = x + y$

Sample mean

We have N samples of x.
Their mean (= average) is m_x.
We also have N samples of y and their mean is m_y.

The sample mean of $z = x + y$ is $m_z = m_x + m_y$

Mean of sums = Sum of means

$$\frac{1}{N}\sum_{i=1}^{N}(x_i + y_i) = \frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N} y_i$$

Expected mean

P_{ij} : joint probability of the pair (x_i, y_i)

its value depends on whether the expts are independent, which we don't know.

$$E[x+y] = \sum_i \sum_j P_{ij} (x_i + y_j)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} P_{ij}\, x_i + \sum_{i=1}^{N} \sum_{j=1}^{N} P_{ij}\, y_j$$

$$= \sum_{i=1}^{N} \left(P_{i1} + P_{i2} + \cdots + P_{iN}\right) x_i + \sum_{j=1}^{N} \left(P_{1j} + P_{2j} + \cdots + P_{Nj}\right) y_j$$

$$= \sum_{i=1}^{N} P_i\, x_i + \sum_{j=1}^{N} P_j\, y_j$$

$$= E[x] + E[y]$$

$$E[x+y] = E[x] + E[y]$$

Variance.

$$\sigma_z^2 = \sum_i \sum_j P_{ij} (x_i + y_j - m_x - m_y)^2$$

$$= \sum_i \sum_j P_{ij} (x_i - m_x)^2 + \sum_i \sum_j P_{ij} (y_j - m_y)^2$$

$$+ 2 \sum_i \sum_j P_{ij} (x_i - m_x)(y_j - m_y)$$

$$= \sum_i P_i (x_i - m_x)^2 + \sum_j P_j (y_j - m_y)$$

$$+ 2 \sum_i \sum_j P_{ij} (x_i - m_x)(y_j - m_y)$$

$$= \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$$

$$\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$$

□ Covariance Matrix for $Z = AX$

Let $X = \begin{bmatrix} x \\ y \end{bmatrix}$ and $A = \begin{bmatrix} 1 & 1 \end{bmatrix}$

$$AX = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x+y \end{bmatrix} = \begin{bmatrix} z \end{bmatrix}$$

$$\sigma_z^2 = \sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$$

$$\Updownarrow$$

$$\sigma_z^2 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \iff \underline{\sigma_z^2 = AVA^T}$$

The covariance matrix of $Z = AX$ is:

$$V_z = A V_x A^T$$

X: vector

$$E[AX] = A\,E[X]$$

$$Cov[AX] = A\,cov[X]\,A^T$$

$$E[X] = E[X_1] + \cdots + E[X_n]$$

Proof

$$E[AX] = E\left(\begin{bmatrix} \sum_{j=1}^{n} a_{1j} X_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj} X_j \end{bmatrix} \right)$$

$$= E\left[\sum_{j=1}^{n} a_{1j} X_j\right] + \cdots + E\left[\sum_{j=1}^{n} a_{mj} X_j\right]$$

$$= \sum_{j=1}^{n} a_{1j}\, E[X_j] + \cdots + \sum_{j=1}^{n} a_{mj}\, E[X_j]$$

$$= A\,E[X]$$

$$\text{Cov}[AX] = E\left[(AX - E[AX])(AX - E[AX])^T\right]$$

$$= E\left[A(X - E[X])(X - E[X])^T A^T\right]$$

$$= A\, E\left[(X - E[X])(X - E[X])^T\right] A^T$$

$$= A\, \text{Cov}(X)\, A^T$$

The Correlation ρ

Rescaling or standardizing the random variables x and y:

$$X = \frac{x}{\sigma_x} \quad , \quad Y = \frac{y}{\sigma_y}$$

$\sigma_X^2 = 1$ & $\sigma_Y^2 = 1$ ← variances of the new variables X & Y.

This is just like dividing a vector 'v' by its length to produce a unit vector $\frac{v}{\|v\|}$ of length 1.

The correlation of x & y is the covariance of X and Y.

Correlation,

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \text{covariance of } X = \frac{x}{\sigma_x} \text{ & } Y = \frac{y}{\sigma_y}$$

$$\sigma_x^2 \sigma_y^2 \geq \sigma_{xy}^2$$

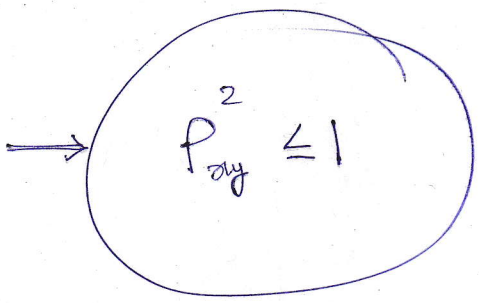
$$-\sigma_x \sigma_y \leq \sigma_{xy} \leq \sigma_x \sigma_y$$

$$-1 \leq \rho_{xy} \leq 1$$

Zero covariance gives zero correlation.

Independent random variables produce $\rho_{xy} = 0$.

always $\sigma_{xy}^2 \leq \sigma_x^2 \sigma_y^2$ ⟹ the covariance matrix V is atleast positive semidefinite.



Correlation near $\rho = +1$ means strong dependence in the same direction.

ie.,

y tends to be above its mean when x is also above its mean and coise versa.

⟹ often voting the same.

-ve correlation means that y tends to be below its mean when x is above its mean

⟹ Voting in opposite directions.

Correlation matrix, $R = \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix}$

* R always has 1's on the diagonal because we normalized to $\sigma_x = \sigma_y = 1$.

ie.,

R is the correlation matrix for x and y, and the covariance matrix for $X = \frac{x}{\sigma_x}$ and $Y = \frac{y}{\sigma_y}$.

The number ρ_{xy} is also called the Pearson coefficient.

* The correlation matrix R comes from the covariance matrix V, when we rescale every row and every column. Divide each row i and column i by the i^{th} standard deviation σ_i.

@ $R = DVD$ for the diagonal matrix D,

where $D = \text{diag}\left[\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_M}\right]$

* ⓑ If covariance V is positive definite, correlation $R = DVD$ is also positive definite.

$$V = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \; ; \; R = \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix}$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Ex:3. Suppose that y is just $-x$. A coin flip has o/p $x=0$ or 1. The same flip has o/p $y=0$ or -1.

Ans: The mean, $m_x = \frac{1}{2}$

$$m_y = -\frac{1}{2}$$

The covariance is, $\sigma_{xy} = -\sigma_x \sigma_y$

$$\Longrightarrow \rho_{xy} = -1$$

$$R = \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad \text{when } y = -x$$

The correlation matrix has determinant zero (singular & only semidefinite).

Ex: 4 Suppose the random variables x, y, z are independent. What matrix is R?

Ans: $\rho_{xx} = \rho_{yy} = \rho_{zz} = 1$

All 3 cross correlations, $\rho_{xy} = \rho_{yz} = \rho_{xz} = 0$

$$R = \begin{bmatrix} \rho_{xx} & \rho_{xy} & \rho_{xz} \\ \rho_{yx} & \rho_{yy} & \rho_{yz} \\ \rho_{zx} & \rho_{zy} & \rho_{zz} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

is the identity matrix.

12.2(A) Suppose x & y are independent random variables with mean 0 and variance 1. Then the covariance matrix V_x for $X = (x, y)$ is 2×2 identity matrix. What are the mean m_z & the covariance matrix V_z for the 3-comp. vector $Z = (x, y, ax+by)$?

Ans:

$$Z = \begin{bmatrix} x \\ y \\ ax+by \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = AX$$

$$m_z = A m_x$$

mean of $ax+by$ is $a m_x + b m_y$.

$$V_z = A V_x A^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} 1 & 0 & a \\ 0 & 1 & b \end{bmatrix} = \begin{bmatrix} 1 & 0 & a \\ 0 & 1 & b \\ a & b & a^2+b^2 \end{bmatrix}$$

$$\sigma^2_{ax+by} = \sigma^2_{ax} + \sigma^2_{by} + 2\sigma_{ax,by} = a^2 + b^2 + 0 = a^2 + b^2$$

$$\det(V_2) = a^2 + b^2 - b^2 - a^2 = 0$$

The 3rd component $z = ax + by$ is completely dependent on x and y. The rank of V_2 is only 2.

Multivariate Gaussian & Weighted Least Squares

The normal probability density $p(x)$ [the Gaussian] depends on only 2 numbers:

— Mean m & Variance σ^2

$$p(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \times e^{-(x-m)^2/2\sigma^2}$$

The graph of $p(x)$ is a bell-shaped curve centered at $x = m$. The cont. variable x can be anywhere b/w $-\infty$ and $+\infty$. With probability close to $2/3$, that random x will lie b/w $m-\sigma$ and $m+\sigma$.

$$\int_{-\infty}^{+\infty} p(x)\,dx = 1 \quad \& \quad \int_{m-\sigma}^{m+\sigma} p(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-1}^{+1} e^{-X^2/2}\,dX \approx \frac{2}{3}$$

$\because X = \frac{x-m}{\sigma} \longrightarrow dX = \frac{dx}{\sigma}$

Every Gaussian turns into a standard Gaussian $p(x)$ with mean $m=0$ and variance $\sigma^2=1$.

The standard normal distribution $N(0,1)$ has $p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Integrating $p(x)$ from $-\infty$ to x gives the cumulative distribution $F(x)$: the prob. that a random ~~variable~~ sample is below x.

$F = \frac{1}{2}$ at $x=0$ (mean).

Two-dimensional Gaussians

We have $M = 2$ Gaussian random variables x and y. They have means m_1 & m_2. They have variances σ_1^2 and σ_2^2.

Independent x and y

If they are independent, then their probability density $p(x,y)$ is just $p_1(x)$ times $p_2(x)$. Multiply probabilities when variables are independent:

Independent x & y:

$$p(x,y) = p_1(x)\,p_2(x) = \frac{1}{2\pi\sigma_1\sigma_2}\, e^{-(x-m_1)/2\sigma_1^2} \times e^{-(x-m_2)/2\sigma_2^2}$$

The covariance of x & y will be $\sigma_{12} = 0$.
The covariance matrix will be diagonal.
The variances σ_1^2 and σ_2^2 are always on the main diagonal of V.

The exponent in $p(x,y)$ is just the sum of the x-exponent & the y-exponent.
Good to notice that the 2 exponents can be combined into $\frac{-1}{2}(x-m)^T V^{-1}(x-m)$:

$$\frac{-(x-m_1)^2}{2\sigma_1^2} - \frac{(y-m_2)^2}{2\sigma_2^2} = \frac{-1}{2}\begin{bmatrix} x-m_1 & y-m_2 \end{bmatrix}\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1}\begin{bmatrix} x-m_1 \\ y-m_2 \end{bmatrix}$$

$$= \frac{-1}{2}(x-m)^T V^{-1}(x-m)$$

Non-independent x and y

When $M = 2$, the first variable x may give partial information about the second variable y (and vice versa). Maybe a part of y is decided by x and part is truly independent. It is the $M \times M$ covariance matrix V that accounts for dependencies b/w the M variables $x = x_1, x_2, \ldots, x_M$.

Multivariate Gaussian probability distribution :

$$p(x) = \frac{1}{(\sqrt{2\pi})^M \sqrt{\det V}} e^{-(x-m)^T V^{-1} (x-m)/2}$$

$x = (x_1, \ldots, x_M)$ & $m = (m_1, \ldots, m_M)$

The M square roots of 2π and the determinant of V are included to make the total probability equal to 1.

We'll use the eigenvalues λ and orthonormal eigenvectors q of the symmetric matrix $V = Q \Lambda Q^T$.

$$\therefore \quad V^{-1} = Q \Lambda^{-1} Q^T$$

$$X = x - m$$

$$(x-m)^T V^{-1} (x-m) = X^T V^{-1} X = X^T Q \Lambda^{-1} Q^T X$$

$$= Y^T \Lambda^{-1} Y$$

where, $Y = Q^T X = Q^T (x-m)$ are statistically independent & their covariance matrix Λ is diagonal.

⇒ This step of diagonalizing V by its eigenvector matrix Q is the same as "uncorrelating" the random variables. Covariances are zero for the new variables $X_1, \ldots, X_m$. ?

- This is the point where linear algebra helps calculus to compute multidimensional integrals.

$$-\frac{1}{2}(x-m)^T V^{-1}(x-m) = X^T Q \Lambda^{-1} Q^T X = Y^T \Lambda^{-1} Y$$

$$= -\frac{1}{2}[y_1 \cdots y_M]\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_M\end{bmatrix}^{-1}\begin{bmatrix}y_1 \\ \vdots \\ y_M\end{bmatrix}$$

$$= -\frac{y_1^2}{2\lambda_1} - \frac{y_2^2}{2\lambda_2} - \cdots - \frac{y_M^2}{2\lambda_M}$$

$$\int \cdots \int e^{-Y^T \Lambda^{-1} Y/2}\, dY = \int_{-\infty}^{+\infty} e^{-y_1^2/2\lambda_1}\, dy_1 \cdots \int_{-\infty}^{+\infty} e^{-y_M^2/2\lambda_M}\, dy_M$$

$$= \left(\sqrt{2\pi\lambda_1}\right) \cdots \left(\sqrt{2\pi\lambda_M}\right)$$

$$= \left(\sqrt{2\pi}\right)^M \sqrt{\det V}$$

Since, $\int_{-\infty}^{+\infty} e^{-at^2}\, dt = \sqrt{\pi/a}$

Vector m of means:

$$\int\int \cdots \int x\, p(x)\, dx = (m_1, m_2, \dots) = m$$

Covariance matrix V:

$$\int \cdots \int (x-m)\, p(x)\, (x-m)^T\, dx = V$$

Weighted Least Squares

Least squares started from an unsolvable system $Ax=b$. We chose $\hat{x}$ to minimize the error $\|b-Ax\|^2$. This lead to the least square equation $A^TA\hat{x} = A^Tb$. The best $A\hat{x}$ is the projection of b onto the $C(A)$. But, is this squared distance $E=\|b-Ax\|^2$ the right error measure to minimize ?

If the measurement errors in b are independent random variables, with mean $= 0$ and variance $\sigma^2 = 1$ and a normal distribution.

→ use least squares.

If the errors are not independent or their variances are not equal

→ use weighted least squares.

Suppose, there is an $m \times n$ matrix A and a vector $x \in \mathbb{R}^n$ such that

$$Y = Ax + \epsilon$$

Here A and x are non-random but ϵ is normally distributed random vector with mean 0 and covariance matrix V.

$\therefore$ Y is normally distributed with mean Ax and covariance matrix V.

Let b be the observed value of Y.

Our goal is to estimate $\hat{x}$, given A & b.

Let f_y be the probability density function for a normally distributed random vector with mean Ax and covariance matrix V.

* The argument of the maxima (arg max) are the points, or elements, of the domain of some function at which the function values are maximized.

In contrast to globa maxima, which refers to the largest o/p of a function, arg max refers to the inputs, at which the function o/p are as large as possible.

Given an arbitrary set X, a totally ordered set Y and a function $f: X \to Y$, the arg max over some subset S of X is defined by

$$\operatorname{arg\,max}_S f := \operatorname*{arg\,max}_{x \in S} f(x) := \{x \in S : f(s) \leq f(x) \text{ for all } s \in S\}$$

$:=$ — defined to be equal to

A natural estimate of x is the max. likelihood estimate

$$\hat{x} = \arg\max_{x} f_y(b)$$

$$= \arg\max_{x} \frac{1}{(2\pi)^m \det(V)} e^{-\frac{1}{2}(b-Ax)^T V^{-1}(b-Ax)}$$

maximizing $f_y(b)$ is equivalent to minimizing $\ln[f_y(x)]$.

ie., minimizing, $E(x) = (b-Ax)^T V^{-1}(b-Ax)$

* Let $c_1, c_2, \ldots, c_m > 0$ be a set of positive numbers. The corresp. weighted inner product and weighted norm on $\mathbb{R}^n$ are defined by

$$\langle v|w\rangle = \sum_{i=1}^{m} c_i v_i w_i$$

$$\|v\| = \sqrt{\langle v|v\rangle} = \sqrt{\sum_{i=1}^{m} c_i v_i^2}$$

The numbers c_i are the weights.

The larger the weights c_i, the more the i^{th} coordinate of v contributes to the norm.

$$\langle v|w\rangle = v^T C w,$$

where $$C = \begin{bmatrix} c_1 & 0 & 0 & \cdots & 0 \\ 0 & c_2 & 0 & \cdots & 0 \\ 0 & 0 & c_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & c_m \end{bmatrix}$$

C : diagonal weight matrix

* Weighted norms are particularly relevant in statistics and data fitting, where one wants to emphasise certain quantities and deemphasize others; this is done by assigning weights to the different components of the data vector V.

The weight matrix W is necessarily positive definite.

$W = B^T B$, for a matrix B with independent columns.

ILA(8)

$B = QR$ | Q has orthonormal columns

$B^T B = (QR)^T (QR) = R^T Q^T Q R = R^T R$

↳ Cholesky factorization

ILA⑥

For any vector x, the weighted norm satisfies:

$x^T W x = x^T R^T R x = (xR)^T (xR)$

$$\min_{x \in \mathbb{R}^m} E(x) = \min_{x \in \mathbb{R}^m} (b - Ax)^T W (b - Ax)$$

$$= \min_{x \in \mathbb{R}^m} (b - Ax)^T R^T R (b - Ax)$$

$$= \min_{x \in \mathbb{R}^m} \left[R(b - Ax)\right]^T \left[R(b - Ax)\right]$$

$$= \min_{x \in \mathbb{R}^m} \left[Rb - RAx\right]^T \left[Rb - RAx\right]$$

ie.,

the solution that we are looking for is the least square solution to the equation

$$RAx = Rb.$$

Using the ordinary least squares, the solution satisfies

$$(RA)^T (RA)\,\hat{x} = (RA)^T (Rb)$$

$$A^T (R^T R)\, A\,\hat{x} = A^T (R^T R)\, b$$

$$A^T W A\,\hat{x} = A^T W b$$

$$W = \begin{bmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & \cdots & 0 \\ 0 & 0 & w_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_m \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\sigma_3^2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sigma_m^2} \end{bmatrix} = V^{-1}$$

- Since each weight is inversely proportional to the error variance, it reflects the information in that observation. So an observation with small error variance has a large weight since it contains relatively more information than an observation with large error variance (small weight).

Weighted least squares

$$A^T V^{-1} A \hat{x} = A^T V^{-1} b$$

$$\hat{x} = \left(A^T V^{-1} A\right)^{-1} A^T V^{-1} b$$

□ The Variance in the Estimated $\hat{x}$

Often the important question is not the best $\hat{x}$ for one particular set of measurements b. This is only one sample.

The real goal is to know the reliability of the whole experiment. That is measured (as reliability always is) by the variance in the estimate $\hat{x}$.

$$E[AX] = A\,E[X]$$

$$\text{Cov}[AX] = A\,\text{cov}[X]\,A^T$$

If b has covariance matrix V, then $\hat{x} = Lb$ has covariance matrix LVL^T.

where, $$L = (A^T V^{-1} A)^{-1} A^T V^{-1}$$

$$\text{Cov}[\hat{x}] = \text{Cov}[Lb] = E\left[(\hat{x} - x)(\hat{x} - x)^T\right]$$

$$= L \, \text{Cov}(b) \, L^T$$

$$= L V L^T$$

$$= \left(A^T V^{-1} A\right)^{-1} A^T V^{-1} \; V \; V^{-1} A \left(A^T V^{-1} A\right)^{-1}$$

$$= \left(A^T V^{-1} A\right)^{-1}$$

ART NATURE

☎: 9446390009
KUNNAMKULAM STATIONERY
PATHANAMTHITTA

Pages with cover	160
Price ₹ Incl.of GST	50.00

Size : 18.4 X 24.8 cms

Mfrs: Ramsons Enterprises. Madurai. Ph : 0452-2671408