WED NOV 1   Multivariate Statistics

{ Cov of random vectors

Sample covariance

PCA: Examples, Theory

FINAL ASSIGNMENT DUE FRI OF FINALS WEEK

Random variables X

$$X_1, X_2, ---, X_n$$
$$\overline{X}, \quad \frac{1}{n-1} \sum_i (X_i - \overline{X})^2$$

MODEL

pdf/pmf

$$E(X) = \int x f(x) dx$$
$$Var(X) = \int (x - EX)^2 f(x) dx$$

Random Vector

$$X = \begin{bmatrix} X_1 \\ | \\ | \\ | \\ X_n \end{bmatrix} \qquad E(X) = \begin{bmatrix} E(X_1) \\ | \\ | \\ | \\ E(X_n) \end{bmatrix}$$

$$COV(X)_{i,j} = Cov(X_i, X_j) = E\left[(X_i - EX_i)(X_j - EX_j) \right]$$
$$\underset{n \times n}{}$$
$$Cov(X_i, X_i) = E\left[(X_i - EX_i)(X_i - EX_i) \right]$$
$$E[(X_i - EX_i)^2]$$
$$= Var(X_i)$$

$$\underline{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1K} \\ & & & \\ & & & \\ X_{n1} & - & - & -X_{nK} \end{bmatrix}$$
$$n \times k$$

rows are observations
columns as variables

$$X_{\bullet j} = j^{th} \text{ column}$$

$$X_{i \bullet} = i^{th} \text{ rows}$$

X, Y are random vectors $(n \times 1)$
$$COV_s(X,Y) = \frac{1}{n-1} \sum_i (X_i - \overline{X})(Y_i - \overline{Y})$$

S
A
M
P
L
E

$$X = \begin{bmatrix} X_1 \\ | \\ X_n \end{bmatrix}$$

$$VAR_s(X) = \frac{1}{n-1} \sum_i (X_i - \overline{X})^2$$

S
A
M
P
L
E

$$\text{cov}(X_i, Y_i) = \mathbb{E}[(X_i - \mathbb{E}X_i)(Y_i - \mathbb{E}Y_i)]$$

$(X_1, Y_1) \quad (X_2, Y_2) \quad --- \quad (X_n, Y_n) \quad$ are $\quad$ IID

$$\text{cov}(X_i, Y_i) = c_{i,j}$$

$$\text{cov}_s(X, Y) \longrightarrow \text{cov}(X_i, Y_i)$$

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \qquad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

$$\underbrace{\text{cov}_s(\underset{n \times k}{\overset{\uparrow}{X}})}_{k \times k}\; i,j \;=\; \text{cov}_s(X_{\cdot i}, X_{\cdot j})$$

height $\quad$ weight $\quad$ GPA

$$\begin{array}{cc} h_1 & w_1 \\ \vdots & \vdots \\ h_6 & w_6 \end{array}$$

IDEA OF PCA: START w/ K VARIABLES

GET NEW VARIABLES $\qquad X_{\cdot 1} \quad X_{\cdot 2} \quad --- \quad X_{\cdot k}$

$$Y_1, Y_2, --, Y_k$$

Y_i is a linear combination of old variables

Y_1, Y_2 capture "most information" about all $X_1, ---, X_{\cdot k}$