21/01/21 dec-10

Covariance :

$$cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

for sample covariance : $cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Relation b/w correlation & covariance : $\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$

→ Covariance is a measure of relationship b/w 2 random var.
→ This metric evaluates to what extent the variables change together.

* Positive covariance: indicates that 2 variables tend to move in same direction
* Negative covariance: reveals that 2 variable move in opposite direction.

where : x_i : values of x variable
y_i : " " y variable
$\bar{x}$: mean / avg. of x var
$\bar{y}$: mean / avg " y var.
σ_x : std. dev. of x
σ_y : " " " y.

data	x TCS	y Infosys	a $(x_i - \bar{x})$	b $(y_i - \bar{y})$	a * b
2013	1692	68	-352.8	-41.20	14,535.36
2014	1978	102	-66.8	-7.2	480.96
2015	1884	110	-160.8	0.8	-128.64
2016	2151	112	106.2	2.8	297.36
2017	2519	154	474.2	44.8	21,244.16

Mean (TCS) $\bar{x}$: $\frac{1692 + \cdots + 2519}{5} = 2044.8$

Mean (Infosys) $\bar{y}$: $\frac{68 + 102 + \cdots + 154}{5} = 109.2$

Cov{x, y}: $\frac{36429.20}{5} = 7285.4$

It can be concluded that, price of stock of TCS & Infosys move in same direction.

3 ⇒ Tuple duplication:

Redundant tuples are identified and removed during integration

4 Data value conflict detection and resolution:

for Eg: wt. in metric unit
wt in British Imperial unit

→ The semantic integrity and str. of data are great challenge during data integration.

3.4 Data reduction

- (a) dimensionality redn
  1. wavelet transform
  2. PCA
  3. Attribute subset seln.
- (b) numerosity redn
  - parametric (Regression, log linear models)
  - non-parametric: Histogram, clustering, sampling, data cube aggregn
- (c) data compression
  - lossless
  - lossy

defn: Data redn. is applied to obtain a reduced rep. of data set that is much smaller in volume, yet closely maintains the integrity of original data.

a) dimensionality reduction :

(Reducing attributes across dimension)

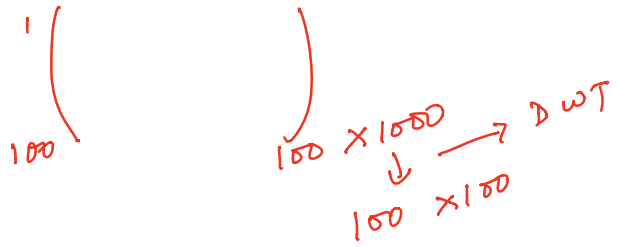
b) wavelet transform :

DWT is a linear signal processing technique that is when applied over a data vector X, transforms it to a numerically different vector X' of wavelet coefficients.

→ Fetching the strongest wavelet coefficients helps in lower dimension mapping.

$X = \langle 1 \ldots \ldots 1000 \rangle$ (0.5, 0.7) $\xrightarrow{DWT}$ $X' = \langle 1 \ldots \ldots 1000 \rangle$ (0.225, 0.618)

$X' = \langle 1 \ldots \ldots 100 \rangle$

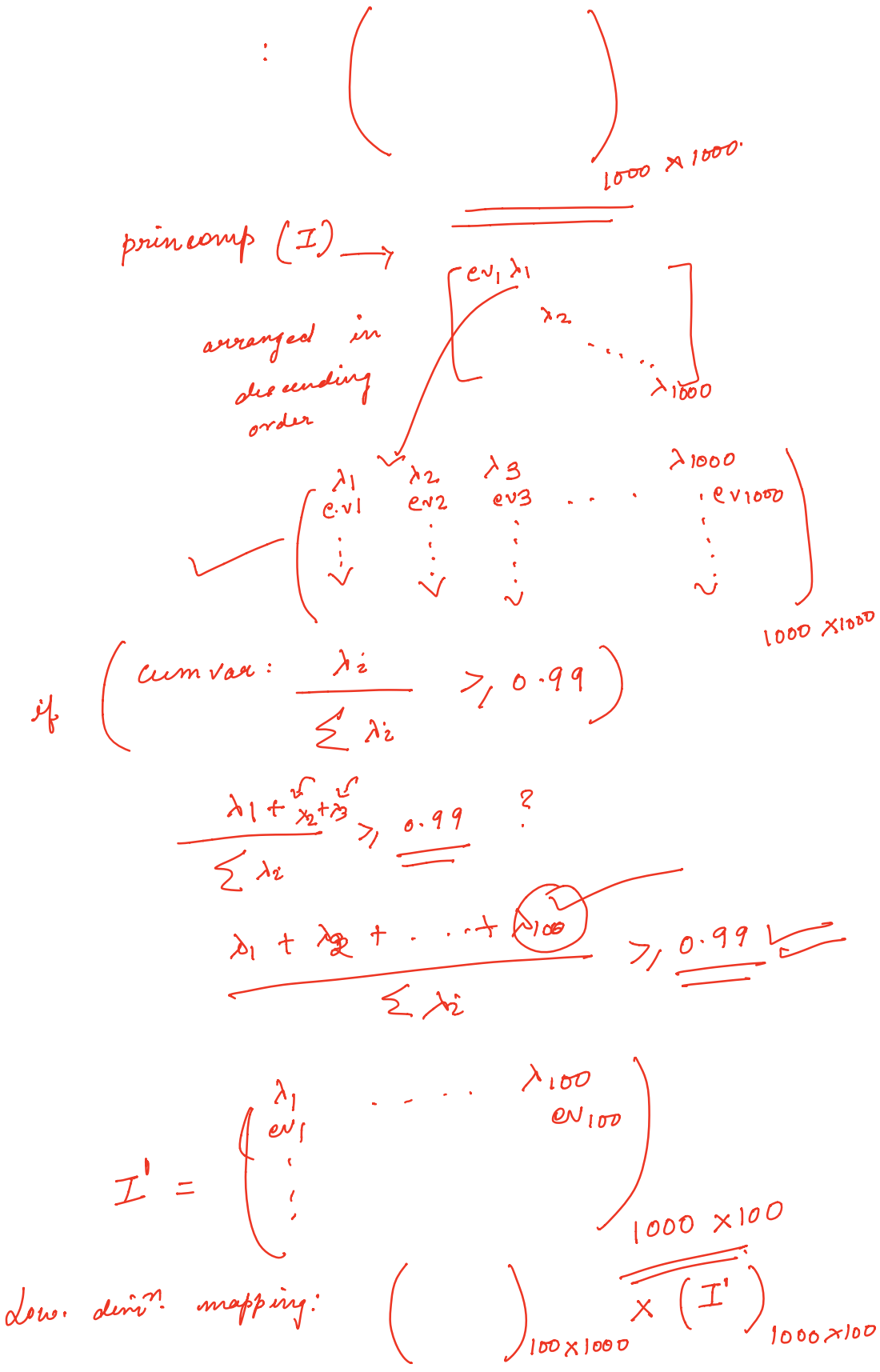


2) PCA :

$A =$ (100 × 1000)

↓

Cov. matrix $I = A^T A$: (1000 × 100) (100 × 1000)


1000 x 1000
princomp (I) →
arranged in descending order
ev1 λ1
λ2
λ1000
λ1 λ2 λ3 λ1000
ev1 ev2 ev3 ev1000
1000 x 1000
if (cumvar : λi / Σλi >= 0.99)
(λ1 + λ2 + λ3) / Σλi >= 0.99 ?
(λ1 + λ2 + . . . + λ100) / Σλi >= 0.99
I' = (λ1 . . . . λ100
ev1 ev100)
1000 x 100
low. dim^n. mapping: (100 x 1000) x (I') 1000 x 100

$= \left(\quad \right)_{100 \times 100}$

3) Attribute selection:

f.v = < f_1 ~~f_2~~ f_3 ~~f_4~~ f_5 >

{ }

{ f_1 } → ML → 90·/.

i) forward selection.

{ f_1 f_2 } → ML → 85·/.

{ f_1 f_3 } → 95·/.

{ f_1 f_3 f_4 } → 95·/.

{ f_1 f_3 f_5 } → 98·/. ✓

ii) backward selection:

~~{ f_1~~ f_2 f_3 f_4 f_5 } → 95·/.

~~{ f_2~~ f_3 f_4 f_5 } → ~~95~~·/. 90·/.

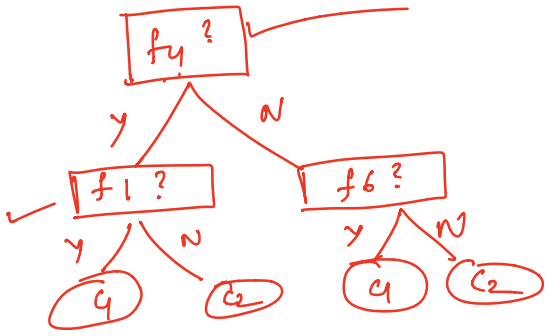
f_2 ~~f_3~~ f_4 f_5 → 95·/. 96·/.

{ f_1 f_3 f_5 } → 98·/.

(iii) Combination of forward selection & backward elimination

(iv) Decision Tree Induction :

Initial: { f_1 f_2 f_3 f_4 f_5 f_6 }



f_4 f_1 f_6 }