PCA :- used to reduce the dimensionality of data set.

1) It is a statistical procedure that uses an orthogonal trans formation which converts a set of correlated variables to a set of uncorrelated variables.

2) PCA is most widely tool used in explo-ratory data analysis & in M/C learning for predictive models.

3) Moreover, PCA is an unsupervised statistical technique used to examine interrelations among set of variables.

4) It is a way of identifying patterns in data., & expressing the data in such a way as to highlight their similarities & differences.

5) This technique used in Image Processing Data may be compressed, without much loss of information.

formulas: 1) Mean
2) Std. Dev. / variance $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
3) Covariance $Cov(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$
4) Eigenvalues.
5) Eigen vector

Principle Component Analysis:

⇒ PCA, is a way of identifying patterns in data and expressing the data in such a way as to highlight their similarities & differences.

1) Step-1. Data:

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Data adjust $x - \bar{x}$, $y - \bar{y}$

$\bar{x} = 1.81$

$\bar{y} = 1.91$

X	Y
0.69	0.49
−1.31	−1.21
.39	0.99
0.09	1.09
1.29	0.79
0.49	−0.31
0.19	−0.81
−0.81	−0.31
−0.31	−0.71
−0.71	−1.01

$(x - \bar{x})$ $(y - \bar{y})$ 0.29

Step-3: Calculate the covariance matrix:

$$Cov(x,y) = \frac{\sum_{i=1}^{n} (x-\bar{x})(y-\bar{y})}{n}$$

$$C = \begin{pmatrix} Cov(x,x) & Cov(x,y) \\ Cov(y,x) & Cov(y,y) \end{pmatrix}$$

$$= \begin{pmatrix} Cov(x,x) & Cov(x,y) & Cov(x,3) \\ (y,x) & (y,y) & (y,3) \\ (3,x) & (3,y) & (3,3) \end{pmatrix}$$

$$Cov = \begin{pmatrix} 0.616 & 0.615 \\ 0.615 & 0.716 \end{pmatrix}$$

$\Rightarrow$ So, since non-diagonal elements in this covariance matrix is positive, we should expect that both x & y variable increases together.

Step-4: Calculate the eigenvectors & eigenvalues of the covariance matrix:

Finding eigenvalues: $\det(A - \lambda I) = 0$

finding eigenvectors: $(A - \lambda I)\, x = 0$

$\lambda \rightarrow$ eigenvalue

$x \rightarrow$ eigenvector

eigenvalues = $\begin{pmatrix} 0.0490 \\ 1.281 \end{pmatrix}$

eigenvectors = $\begin{pmatrix} -0.733 & -0.677 \\ 0.6778 & -0.735 \end{pmatrix}$

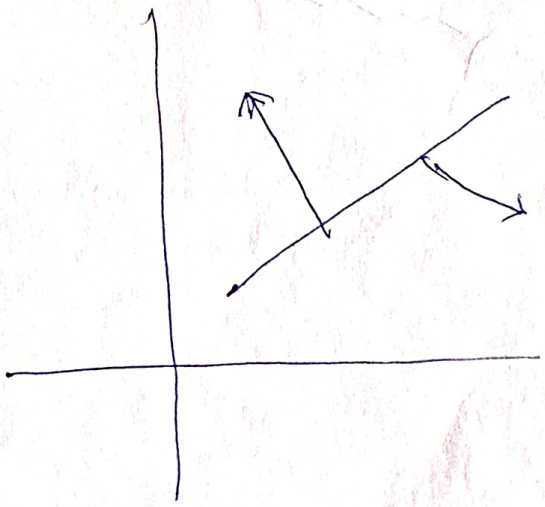
Step-5 : Calculate of new data set

New (final) data = Row feature vector * Row Data adjust

$$NFD = \begin{bmatrix} -0.733 & -0.677 \\ 0.677 & -0.735 \end{bmatrix} \begin{bmatrix} 0.69, -1.31, \ldots & -0.71 \\ 0.49, -1.21, \ldots & -1.01 \end{bmatrix}$$

$(2\times2)\ (2\times10) = 2\times10$

$$\begin{matrix} PCA_1 \\ PCA_2 \end{matrix} \begin{bmatrix} x_1 & \ldots & x_{10} \\ y_1 & \ldots & y_{10} \end{bmatrix}$$



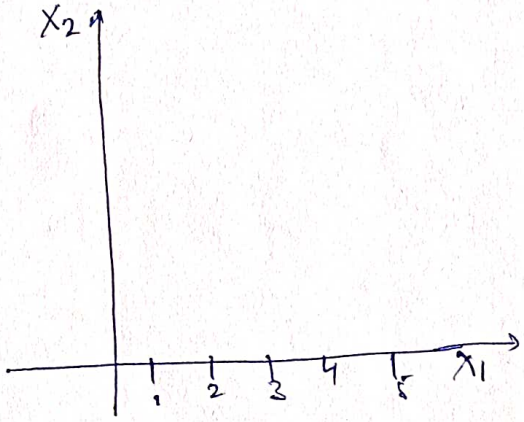
LDA:

1) Samples for class 1: (ω_1)

$$X_1 = \begin{bmatrix} 4,2; \\ 2,4; \\ 2,3; \\ 3,6; \\ 4,4 \end{bmatrix}; \quad N_1 = 5$$

Class 2 → (ω_2)

$$X_2 = \begin{bmatrix} 9,10; \\ 6,8; \\ 9,5; \\ 8,7; \\ 10,8 \end{bmatrix}; \quad N_2 = 5$$



2) Calculate class mean:

$$\mu_1 = \frac{1}{N_1} \sum_{x \in \omega_1} x$$

$$\mu_1 = \frac{1}{5}\left[\begin{bmatrix}4\\2\end{bmatrix} + \begin{bmatrix}2\\4\end{bmatrix} + \begin{bmatrix}2\\3\end{bmatrix} + \begin{bmatrix}3\\6\end{bmatrix} + \begin{bmatrix}4\\4\end{bmatrix}\right] = \begin{bmatrix}3\\3.8\end{bmatrix}$$

$$\mu_2 = \frac{1}{N_2} \sum_{x \in \omega_2} x = \frac{1}{5}\left[\begin{bmatrix}9\\10\end{bmatrix} + \begin{bmatrix}6\\8\end{bmatrix} + \begin{bmatrix}9\\5\end{bmatrix} + \begin{bmatrix}8\\7\end{bmatrix} + \begin{bmatrix}10\\8\end{bmatrix}\right]$$

$$= \begin{bmatrix}8.4\\7.6\end{bmatrix}$$

3) Calculate Covariance of first class: (6)

$S_1 = Cov(X, Y)$

$Cov(x,x) = \frac{1}{n}\sum_{i=1}^{5}(x-\bar{x})(x-\bar{x})$

for class 1

4	2
2	4
2	3
3	6
4	4
$\frac{15}{5}$	$\frac{19}{5} = 3.8 = \mu_2$

$\mu = 3$

$= \frac{1}{4}[(4-3)(4-3)+(2-3)(2-3)+(2-3)(2-3)+(3-3)(3-3)+(4-3)(4-3)]$

$= \frac{1}{4}[4] = 1$

$\Rightarrow Cov(x,y) = \frac{1}{4}[(4-3)(2-3.8)+(2-3)(4-3.8)+(2-3)(3-3.8)+(3-3)(6-3.8)+(4-3)(4-3.8)]$

$= \frac{1}{4}\sum_{i=1}^{5}(x-\bar{x})(y-\bar{y})$

$= \frac{1}{4}[-1] = -0.25$

$\Rightarrow Cov(x,y) = Cov(y,x) = -0.25$

$\rightarrow Cov(y,y) = \frac{1}{4}\sum_{i=1}^{5}(y-\bar{y})(y-\bar{y}) = [(2-3.8)(2-3.8)+(4-3.8)(4-3.8)+(3-3.8)(3-3.8)+(6-3.8)(6-3.8)+(4-3.8)(4-3.8)]$

$= 2.2$

$= \frac{1}{4}[8.8] = 1.88$

$S_1 = Cov(x,y) = \begin{bmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{bmatrix}$

4) Similarly $S_2 = Cov(x,y)$ for class 2

9	10
6	8
9	5
8	7
10	8

$S_2 = \begin{bmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{bmatrix}$

5) Calculate with-in-class scatter matrix:

$$S_W = S_1 + S_2 = \begin{bmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{bmatrix}$$

6) Between class scatter matrix:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

$$= \left[\begin{bmatrix} 3 \\ 3.8 \end{bmatrix} - \begin{bmatrix} 8.4 \\ 7.6 \end{bmatrix}\right] \cdot \left[\begin{bmatrix} 3 \\ 3.8 \end{bmatrix} - \begin{bmatrix} 8.4 \\ 7.6 \end{bmatrix}\right]^T$$

$$= \begin{bmatrix} -5.4 \\ -3.8 \end{bmatrix} \begin{bmatrix} -5.4 & -3.8 \end{bmatrix}$$

$$= \begin{bmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{bmatrix}$$

Calculate

1) LDA Projection:

$$S_W^{-1} S_B \, w = \lambda w$$

$$\Rightarrow |S_W^{-1} S_B - \lambda I| = 0$$

$$\Rightarrow \left| \begin{bmatrix} 3.3 & -0.3 \\ 0.3 & 5.5 \end{bmatrix}^{-1} \begin{bmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\Rightarrow \begin{vmatrix} 9.22 - \lambda & 6.48 \\ 4.23 & 2.97 - \lambda \end{vmatrix}$$

$$\Rightarrow \lambda_1 = 0, \quad \lambda_2 = 12.2007$$

$\Rightarrow$ when $\lambda_2 = 12.2007$, then eigenvector

optimal $\longrightarrow$ $w_2 = \begin{bmatrix} 0.908 \\ 0.417 \end{bmatrix}$

$\Rightarrow$ optimal projection is one that gives maximum ($J = T/w$)

SVD:-

1) SVD states that any matrix A can be factorized as:

$$A = USV^T$$

⇒ Where U & V are orthogonal matrices with orthonormal eigenvectors chosen from AA^T & A^TA resp.

⇒ S → is a diagonal matrix with r elements equal to the root of the positive eigenvalues of AA^T & A^TA. The diagonal elements are composed of singular values.

$$\begin{bmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} = \begin{matrix} S \\ \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \end{matrix}$$

singular value

⇒ $A_{m \times n} = U_{m \times m} S_{m \times n} V^T_{n \times n}$

$$\begin{matrix} A \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & & & \\ x_{m1} & & & x_{mn} \end{bmatrix} \end{matrix} = \begin{matrix} U \\ \begin{bmatrix} u_{11} & & u_{m1} \\ & & \\ u_{1m} & & u_{mm} \end{bmatrix} \end{matrix}$$

$m \times n$

$$\begin{matrix} S \\ (m \times n) \\ \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{bmatrix} \end{matrix} \quad \begin{matrix} V \\ \begin{bmatrix} v_{11} & & v_{n1} \\ & & \\ v_{1n} & & v_{nn} \end{bmatrix} \end{matrix}$$

Sol:-

$$A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$$

$$\Rightarrow AA^T = \begin{bmatrix} 17 & 8 \\ 8 & 17 \end{bmatrix} \Rightarrow \text{eigenvalues } \lambda_1 = 25, \lambda_2 = 9$$

$$A^T A = \begin{bmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{bmatrix} \Rightarrow \text{eigenvalues } \lambda_1 = 25, \lambda_2 = 9, \lambda_3 = 0$$

eigenvectors

$$u_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \quad u_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

$$v_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix} \quad v_2 = \begin{bmatrix} 1/\sqrt{18} \\ -1/\sqrt{18} \\ 4/\sqrt{18} \end{bmatrix}$$

$$v_3 = \begin{bmatrix} 2/3 \\ -2/3 \\ -1/3 \end{bmatrix}$$

$\Rightarrow$ Singular values are square root of positive eigenvalues. $\sqrt{25} = 5 \quad \sqrt{9} = 3$

$\Rightarrow$ Therefore the SVD composition is

$$A = USV^T = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{18} & -1/\sqrt{18} & 4/\sqrt{18} \\ 2/3 & -2/3 & -1/3 \end{bmatrix}$$

Unit- IV

⇒ Feature selection:

1) Filter Method:

1a.) Correlation Analysis

↳ Pearson's correlation

$$r_{x,y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, \quad -1 < r < 1$$

if $r = 0 \rightarrow$ No relation

1b.) Covariance Analysis $cov(X,Y)$

2a) ANOVA

3a) Chi-Square

4) LDA

Feature/Response	Continuous	Categorical
Continuous	Pearson	LDA
Categorical	ANOVA	Chi-Square

(1) Filter Method:

set of all features → selecting the best subset → learning method → Performance.

ANOVA [Analysis of Variance].

↳ Statistical technique that is used to compare the means of more than two populations is known as ANOVA.

Eg: Anti-Anxiety Medicaleiw:

$\alpha = 0.05$

0 mg	50 mg	100 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2
57	47	21

1) Define Null & Alternative hypothesis

H_0 ; $\mu_{0mg} = \mu_{50mg} = \mu_{100mg}$

H_1 : not all μ's are equal

$\alpha = 0.05$

2) Calculate Degree of freedom

$N = 21, \quad n = 7$

$df_{between} = a - 1 = 3 - 1 = 2$

$df_{within} = N - a = 21 - 3 = 18$

$df_{total} = N - 1 = 21 - 1 = 20$

$F(2, 18) = ?$

From table

$F(2, 18) = 3.55$

$\therefore F(2, 18) > 3.55$

ANOVA : →

⇒ In statistics two-way ANOVA examines the influence of two different categorical independent variables on one continuous dependent variable.

⇒ The two-way ANOVA not only aims at assessing the main effect of each independent variable but also if there is any interaction b/w them.

⇒ Calculate F

⇒ $F = \frac{MS_{between}}{MS_{within}}$

⇒ $MS_{between} = \frac{98.67}{2} = 49.34$

⇒ $MS_{within} = \frac{10.29}{18} = 0.57$

⇒ $F = \frac{49.34}{0.57} = 86.56$

⇒ $F(2, 18) = 86.56$ ⇒ Reject Null Hypothesis

Conclusion: ⇒ The three conditions differed significantly on anxiety level.

3) Calculate Test Statistics: (14)

	SS	df	MS	F
between		2		
within		18		
total		20		

$$\Rightarrow SS_{between} = \frac{\sum(\sum a_i)^2}{n} - \frac{T^2}{N}$$

$$= \frac{57^2+47^2+21^2}{7} - \frac{125^2}{21} = \boxed{98.67}$$

Omg Group: $9+8+7+8+8+9+8 = 57$

50mg : $7+6+6+7+8+7+6 = 47$

100mg : $4+3+2+3+4+3+2 = 21$

125

$$\Rightarrow SS_{within} = \sum y^2 - \frac{\sum(\sum a_i)^2}{n}$$

$$= 853 - \frac{57^2+47^2+21^2}{7} = \boxed{10.29}$$

$$\sum y^2 = 9^2+8^2+7^2+8^2+8^2+9^2+8^2+7^2+6^2+6^2$$
$$+7^2+8^2+7^2+6^2+4^2+3^2+2^2+3^2+4^2+3^2+2^2$$
$$= \boxed{853}$$

$$\Rightarrow SS_{total} = \sum y^2 - \frac{T^2}{N} \Rightarrow 853 - \frac{125^2}{21}$$

$$= \boxed{108.95}$$

Chi-Square Test (Independence Test)

χ^2-test

↳ is used to test whether or not two variables are independent.

Steps:

1) H_0 is that the two variables are independent.
2) H_a is that the two variables are not independent.
3) The expected frequency $E_{R,C}$, for the entry in row R, column C, is calculated using $E_{R,C} = \frac{(\text{Sum of row R}) \times (\text{sum of col C})}{\text{Sample size}}$
4) Degree of freedom: (no. of rows − 1) × (no. of col. − 1)

Example:

Result \ Treatment	Medicine 1	2	3	Total
Significant Improvement	58	81	61	200
Slight Improvement	42	19	39	100
Total	100	100	100	300

at $\alpha = 0.01$, is there enough evidence to conclude that the treatment & result are independent.

Soln:

1) The Null hypothesis H_0 : the treatment & response are independent.

2) The Alternative hypothesis: H_a → The treatment & response are dependent.

3) $\alpha = 0.10$

4) $df = (row-1) \times (col-1)$
$= (2-1) \times (3-1) = 2$

5) The test statistic can be calculated as:

Row, Col.	E	O	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
1, 1	$\frac{200 \times 100}{300} = 66.7$	58	-8.7	75.69	1.135
1, 2	$\frac{200 \times 100}{300} = 66.7$	81	14.3	204.49	3.067
1, 3	$\frac{200 \times 100}{300} = 66.7$	61	-5.7	32.49	0.487
2, 1	$\frac{100 \times 100}{300} = 33.3$	42	8.7	75.69	2.271
2, 2	$\frac{100 \times 100}{300} = 33.3$	19	-14.3	204.69	6.135
2, 3	$\frac{100 \times 100}{300} = 33.3$	39	5.7	32.49	0.975

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 14.07$$

⇒ from $\alpha = 0.10$ & $df = 2$, the critical value is 4.605

⇒ since $\chi^2 > 4.605$ ⇒ Reject Null hypothesis.

⇒ & to believe that there is a relationship b/w treatment & response.

⇒ Feature Selection: Wrapper Method:

1) Forward selection
2) Backward elimination
3) Recursive feature elimination → Greedy method

⇒ Embedded Method:
1) Lasso Reg. L1 Regularization
2) Ridge Reg. L2 Regularization
3) Elastic Net → both L1 & L2

⇒ LASSO (Least Absolute Shrinkage & Selection operator) → 1) Powerful feature selection tech. that is very useful for regression problem.

2) It is essentially a regularization method.

3) LASSO → is a regression analysis method that performs both variable selection & regularization in order to enhance the prediction accuracy & interpretability of statistical model it produces.

Ridge	LASSO
1) It includes all features (or none) in the model.	1) Along with shrinkage coefficients LASSO also performs feature selection.
2) Major Advantage → is coefficient shrinkage & reducing model complexity.	2) It arbitrarily selects any one feature among highly correlated features.
3) It works well in presence of highly correlated features.	

- LASSO Method:

1) Feature Selection (Why?).

1) make the model easier to interpret, removing variables that are redundant & do not add any information.
2) reduce the size of problem to enable algo. to work faster, making it possible to handle with high-dim data.
3) reduce overfitting.

LASSO: — 1) Powerful method to perform two tools

① Feature Selection
② Regularization (Shrinkage)

⇒ The Goal of this process is to minimize the prediction error.

Linear Regression	multiple Linear Reg.	LASSO Linear Reg.
$Y_i = X\beta + \epsilon$	$Y = B_0 + B_1 x_1 + B_2 x_2 \ldots B_k x_k + \epsilon_i$ $\epsilon \rightarrow$ Random error $B_0, B_1 \rightarrow \ldots B_k$ $\rightarrow$ Reg. Coefficients	

Goal of Linear Reg. Analysis to fit a straight line to no of points. as well as to minimize the error.

LASSO :→ 1) It minimizes the sum of square errors, with a upper bound on the sum of the absolute values of the model parameters.

2) The Lasso estimate is defined by the solution to the l_1 optimization problem

$$\text{minimize}\left(\frac{||Y-X\beta||_2^2}{n}\right) \text{ subject to } \sum_{j=1}^{k}||\beta||_1 < t$$

where $t \rightarrow$ is the upper bound for the sum of coefficients. The optimization problem is equivalent to the parameter estimation that follows:

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}}\left(\frac{||Y-X\beta||_2^2}{n} + \lambda||\beta||_1\right)$$

where $||Y-X\beta||_2^2 = \sum_{i=0}^{n}(Y_i - X\beta)_i^2$

& $||\beta||_1 = \sum_{j=1}^{k}|\beta_j|$ & $\lambda \geq 0$

⇒ The larger the value of λ, → the greater the amount of shrinkage.

⇒ The relationship b/w λ & t is a reverse relationship

when $t = 0$ λ will be infinite

when $t = \infty$ λ will be zero.

⇒ when to minimize optimization problems some coefficients are shrunk to zero i.e. $\hat{\beta}_j(\lambda) = 0$

⇒ for some value of γ (depending on the value of parameter λ).

⇒ In this way the feature with coefficient equal to zero are excluded from the model.

⇒ For this LASSO is powerful method for feature selection as compared to other (Ridge).

⇒ $\hat{\beta}^{ridge} = \underset{\beta \in R^p}{\operatorname{argmin}} \| \gamma - X\beta \|_2^2 + \lambda \| \beta \|_2^2$

$\lambda \| \beta \|_2^2$ ↳ L-2 Penalty

⇒ $\hat{\beta}^{lasso} = \underset{\beta \in R^p}{\operatorname{argmin}} \| \gamma - X\beta \|_2^2 + \lambda \| \beta \|_1$

$\| \gamma - X\beta \|_2^2$ ↳ Loss

$\lambda \| \beta \|_1$ ↳ L-1 Penalty

⇒ λ → tunning parameter → controls the strength of penalty

⇒ $\hat{\beta}^{lasso}$ ⇒ the Linear Reg. estimate when $\lambda = 0$ & $\hat{\beta}^{lasso} = 0$ when $\lambda = \infty$

⇒ when λ in b/w $[0 \& \infty]$

↳ 1) Fitting a linear model of γ on X. & shrinking the coefficients.

but the nature of l_1 penalty causes some coefficients to be shrunken to zero exactly.

⇒ This is what makes the lasso substantially different from ridge regression.

⇒ It is able to perform variable selection in the linear model.

⇒ As d - increases, more coefficients are set to zero (less variables are selected) & among the non-zero coefficients, more shrinkage is employed.