Multivariate Statistics (I)

2. Principal Component Analysis (PCA)

Contents

2.1 Comprehension of PCA

- **Definition of Principal Components (PCs) :** $\boldsymbol{x} = (x_1, \ldots, x_p)^t \sim (\mu, \Sigma)$

  Algebraic Def.: Particular linear combinations of the original p random variables

  kth PC : $\boxed{p_k = v_{k1}x_1 + v_{k2}x_2 + \cdots + v_{kp}x_p = \boldsymbol{v}_k^t\boldsymbol{x},\ k = 1, \ldots, p.}$

  Geometric Def.: Selection of a new coordinate system obtained by rotating the original system with as the coordinate axes.

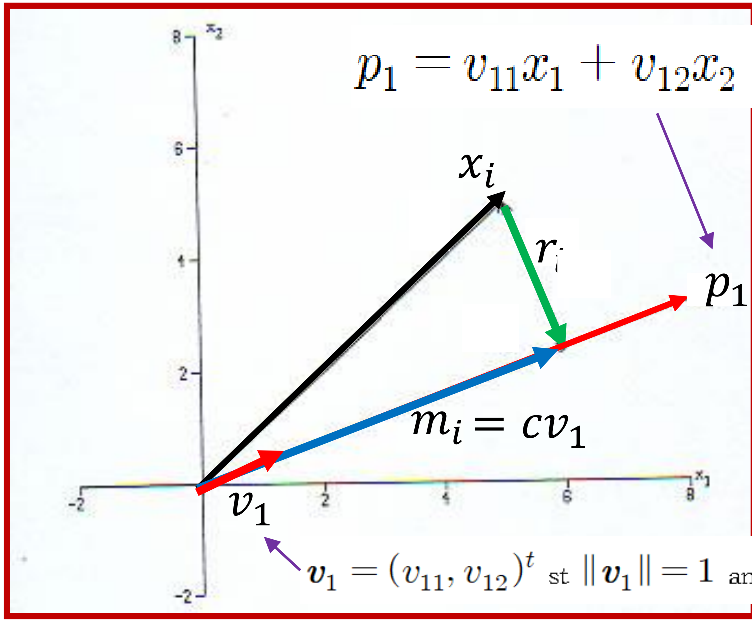
- **PCA** : technique for concerning with explaining the variance-covariance structure through PCs.

Objectives:

- ✓ Data (dimension) reduction: p variables -> k PCs ($k \ll p$)
- ✓ Interpretation of variables
- ✓ Checking the normality and outliers
- ✓ PCs scores can be used as a new data

2.1 Comprehension of PCA : History

$$\boldsymbol{x}_i = (x_{i1}, x_{i2})^t \sim \left(\boldsymbol{\mu} = (\mu_1, \mu_2)^t, Cov(\boldsymbol{x}) = \Sigma\right)$$



$$p_1 = v_{11}x_1 + v_{12}x_2 = \boldsymbol{v}_1^t \boldsymbol{x}$$

$\boldsymbol{v}_1 = (v_{11}, v_{12})^t$ st $\|\boldsymbol{v}_1\| = 1$ and $\boldsymbol{v}_1^t \boldsymbol{v}_k = 0, (1 \neq k)$

$$\begin{aligned}\boldsymbol{x}_i^t \boldsymbol{x}_i &= (\boldsymbol{m}_i + \boldsymbol{r}_i)^t (\boldsymbol{m}_i + \boldsymbol{r}_i) \\ &= \boldsymbol{m}_i^t \boldsymbol{m}_i + \boldsymbol{r}_i^t \boldsymbol{r}_i + 2\boldsymbol{r}_i^t \boldsymbol{m}_i \\ &= \boldsymbol{m}_i^t \boldsymbol{m}_i + \boldsymbol{r}_i^t \boldsymbol{r}_i\end{aligned}$$

$$\sum_{i=1}^{n} \boldsymbol{r}_i^t \boldsymbol{r}_i = \sum_{i=1}^{n} \boldsymbol{x}_i^t \boldsymbol{x}_i - \sum_{i=1}^{n} \boldsymbol{m}_i^t \boldsymbol{m}_i$$

$$Min_{\boldsymbol{v}_1} \sum_{i=1}^{n} \boldsymbol{r}_i^t \boldsymbol{r}_i \Leftrightarrow Max_{\boldsymbol{v}_1} \sum_{i=1}^{n} \boldsymbol{m}_i^t \boldsymbol{m}_i \Leftrightarrow Max_{\boldsymbol{v}_1} nc^2 \boldsymbol{v}_1^t \boldsymbol{v}_1$$

1857-1936

Karl Pearson (1901): best fitting subspace based on the orthogonal projection of a two-dimensional vector onto a one–dimensional subspace

Harold Hotelling (1933): approach for finding the PCs maximizing

1895-1973

$$Var(p_1) = \boldsymbol{v}_1^t \Sigma \boldsymbol{v}_1 \Longrightarrow Max_{\boldsymbol{v}_1} \boldsymbol{v}_1^t \Sigma \boldsymbol{v}_1 \Leftrightarrow Max_{\boldsymbol{v}_1} l_1 \boldsymbol{v}_1^t \boldsymbol{v}_1$$

$$\boxed{Max_{\boldsymbol{v}_k} \sum_{i=1}^{n} \boldsymbol{m}_i^t \boldsymbol{m}_i \Leftrightarrow Max_{\boldsymbol{v}_k} nc^2 \boldsymbol{v}_k^t \boldsymbol{v}_k \cong Max_{\boldsymbol{v}_k} \boldsymbol{v}_k^t \Sigma \boldsymbol{v}_k \Leftrightarrow Max_{\boldsymbol{v}_k} l_k \boldsymbol{v}_k^t \boldsymbol{v}_k}$$

2.1 Comprehension of PCA : Process Steps for PCA

[Step 1] Prepare a multivariate data matrix X

[Step 2] From the X, calculate S (or R)

[Step 3] Obtain the eigenvalues and eigenvectors of S (or R) based on the Spectral Decomposition

$$t_m = \frac{\sum_{k=1}^{m} l_k}{l_1 + l_2 + \cdots + l_p} \times 100$$

$$t_m = \frac{\sum_{k=1}^{m} l_k}{p} \times 100$$

[Step 4] Choose the first m($\leq p$) eigenvalues which are greater than 70% of total sum of eigenvalues

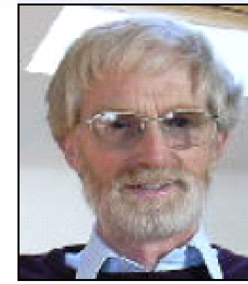
[Step 5] Obtain the PCs with eigenvectors corresponding the selected eigenvalues in [Step 4] and raw variables.

[Step 6] Calculate PCs scores from the centred data(or standardised data)

[Step 7] Consider PCs scores as a new multivariate data which are dimension reduction

2.2 Concepts of PCs

Figure 1 gives a plot of 50 observations on two highly correlated variables .
If we transform to PCs p_1, p_2, we obtain the plot given Figure 2 wrt PCs.
(Jolliffe(2002). *Principal Component Analysis*, Spring-Verlag, New York)



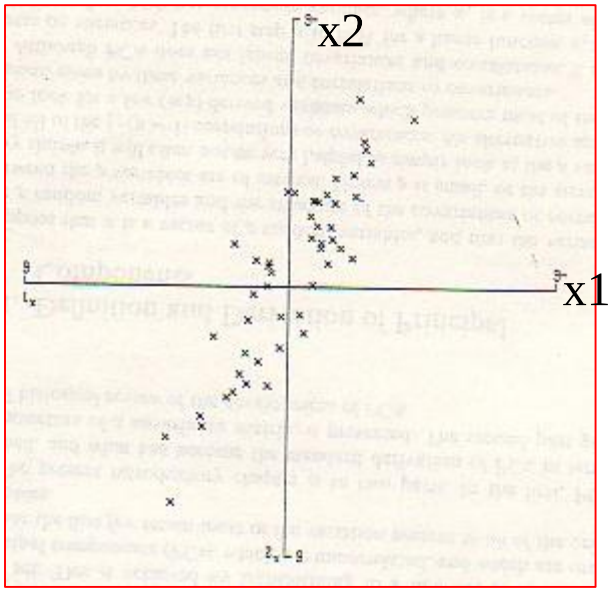


Figure 1: Plot of 50 observations on x1 and x2

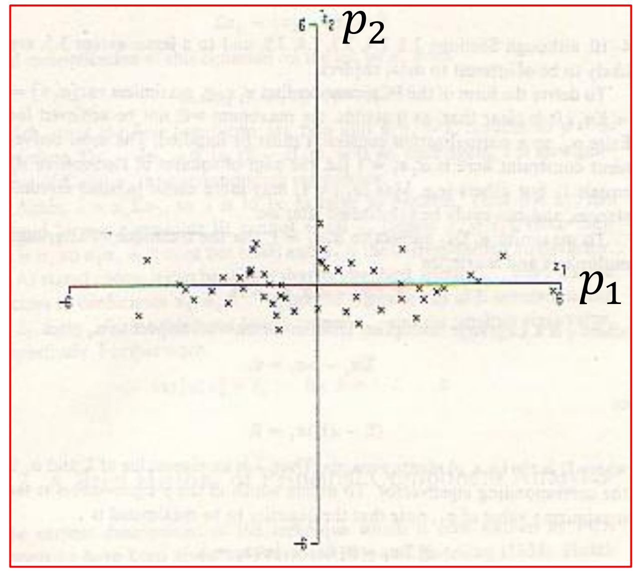


Figure 2: Plot of 50 observations on PCs y1 and y2

Consideration of variations in both x1 and x2:

- Rather more variation in the direction of x2 than x1.
- Clearly there is greater variation in the direction of p_1 but very little variation in the direction of p_2.

2.2 Concepts of PCs

[Example 2.2.1]3 subjects data

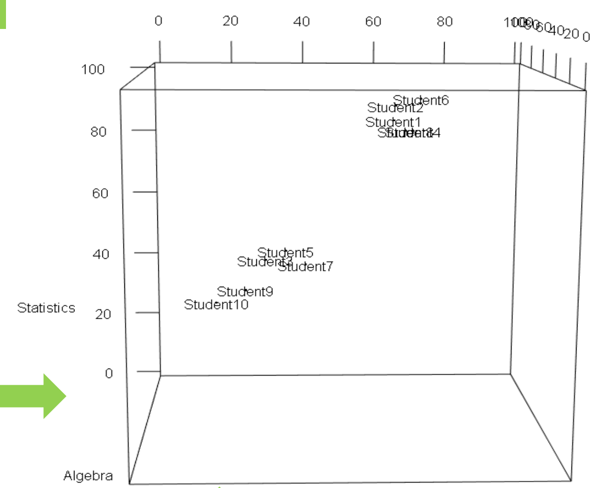
Students	math	Algebra	Statistics
Student1	65	85	85
Student2	65	80	90
Student3	30	40	50
Student4	70	83	82
Student5	35	43	52
Student6	72	82	92
Student7	40	43	48
Student8	68	83	82
Student9	25	32	43
Student10	17	51	35

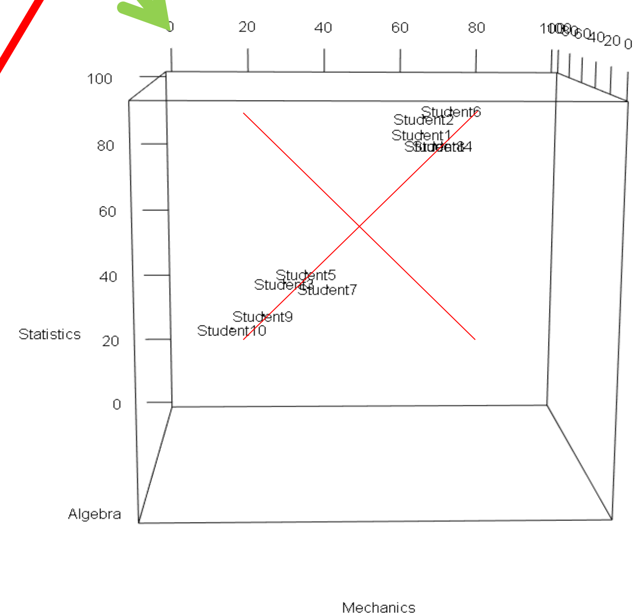
$$x_1 = (65,\ 85,\ 85)^t$$

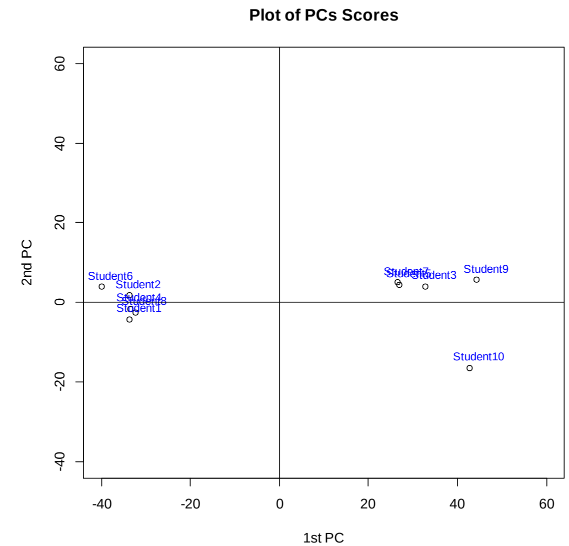
$$x_2 = (65,\ 80,\ 90)^t$$

$$\cdots$$

$$x_{10} = (10,\ 17,\ 35)^t$$







$$p_1 = (-33.6248, -4.2539)^t$$

$$p_2 = (-33.6898, 1.7761)^t$$

$$p_{10} = (42.6252, -16.3739)^t.$$

2.2 Concepts of PCs

◆ **PCA steps for 3 subjects data**

[step 1] Prepare a multivariate data matrix X

[step 2] From the X, calculate S

$S =$

	Mechanics	Algebra	Statistics
Mechanics	453.344	431.511	459.078
Algebra	431.511	484.622	450.244
Statistics	459.078	450.244	487.878

[step 3] Obtain the eigenvalues and eigenvectors of S (or R) based on the Spectral Decomposition

- $S = VDV^t : \quad V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p), \quad V^t V = VV^t = I$

  $D = diag(l_1, \ldots, l_p), \; l_1 \geq \cdots \geq l_p > 0$
- Eigenvalues: $(l_1, l_2, l_3) = (1369.521, 45.161, 11.162)$
- Eigenvectors: $V = (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3)$

$\boldsymbol{v}_1$	$\boldsymbol{v}_2$	$\boldsymbol{v}_3$
-0.567	0.426	0.705
-0.576	-0.817	0.030
-0.589	0.389	-0.708

2.2 Concepts of PCs

[step 4] Choose the first m($\leq p$) eigenvalues

$$t_m = \frac{\sum_{k=1}^{m} l_k}{l_1 + l_2 + \cdots + l_p} \times 100$$

which are greater than 70% of total sum of eigenvalues

- Explanatoy ratios : 96.05%, 3.17%, 0.78%
- $m = 1$, explanatory ratios of eigenvalues(l_1 =1369.521) is 96.05%

[step 5] Obtain the PCs with eigenvectors corresponding the selected eigenvalues in [Step 4] and raw variables.

- Raw Variable: $\boldsymbol{x} = (x_1, x_2, x_3)^t = (\text{Mechanics, Algebra, Statistics})^t$
- Centred variables: $\boldsymbol{y} = (y_1, y_2, y_3)^t$
- First PC : $p_1 = \boldsymbol{v}_1^t \boldsymbol{y} = -0.567 y_1 - 0.576 y_2 - 0.589 y_3$
- Second PC : $p_2 = \boldsymbol{v}_2^t \boldsymbol{y} = 0.426 y_1 - 0.817 y_2 + 0.389 y_3$

2.2 Concepts of PCs

[step 6] Calculate PCs scores from the centred data(or standardised data)

$$Y = \begin{array}{l} \boldsymbol{y}_1 = (16.3,\ 22.8,\ 19.1)^t \\ \boldsymbol{y}_2 = (16.3,\ 17.8,\ 24.1)^t \\ \qquad \cdots\cdots \\ \boldsymbol{y}_{10} = (-31.7,\ -11.2,\ -30.9)^t \end{array}$$

$$P = YV_{(2)} \qquad V_{(2)} = (\boldsymbol{v}_1,\ \boldsymbol{v}_2)$$

[step 7] Consider PCs scores as a new multivariate data which are dimension reduction

$$P = \begin{array}{l} \boldsymbol{p}_1 = (-33.6248,\ -4.2539)^t \\ \boldsymbol{p}_2 = (-33.6898,\ 1.7761)^t \\ \qquad \cdots\cdots \\ \boldsymbol{p}_{10} = (42.6252,\ -16.3739)^t \end{array}$$

2.3 Algebraic derivation of PCs : [Table 2.3.1]

- Random vector: $\boldsymbol{x} = (x_1, \ldots, x_p)^t \sim (\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^t, \; Cov(\boldsymbol{x}) = \Sigma > 0)$
- Spectral decomposition: $\Sigma = VDV^t = \sum_{k=1}^{p} l_k \boldsymbol{v}_k \boldsymbol{v}_k^t$

$$V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p) : V^t V = VV^t = I$$ Orthogonal matrix

$$D = diag(l_1, \ldots, l_p) : \; l_1 \geq \cdots \geq l_p > 0$$ Diagonal matrix with eigenvalues

- *k*th PC: $p_k = v_{k1}x_1 + v_{k2}x_2 + \cdots + v_{kp}x_p = \boldsymbol{v}_k^t \boldsymbol{x}, \; k = 1, \ldots, p$
- PC vector: $\boldsymbol{p} = \begin{bmatrix} p_1 \\ \vdots \\ p_p \end{bmatrix} = V^t \boldsymbol{x} \;\Rightarrow\; Cov(\boldsymbol{p}) = V^t \Sigma V = D$

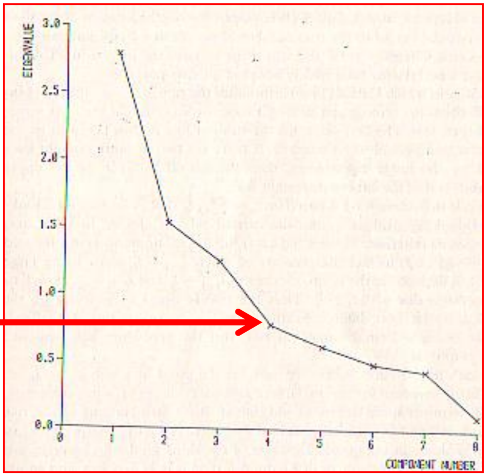
2.4 Selection and goodness-of-fit of PCs

- **How many PCs ?**
  - ✓ Retain only the components whose variances (eigenvalues) are greater than or equal to 1 for R, or 0 for S(Rule of thumb= Kaiser(1960)'s rule)
  - ✓ Percentage of variation accounted for by the first *m* PCs

    (Goodness-of-fit):

$$t_m = \frac{\sum_{k=1}^{m} l_k}{p} \times 100$$ for *R*

$$t_m = \frac{\sum_{k=1}^{m} l_k}{l_1 + l_2 + \cdots + l_p} \times 100$$ for *S*



The slope actually increases bt 3 and 4, but then falls sharply

Plot of eigenvalues against components number

2.4 Selection and goodness-of-fit of PCs

[Table 2.4.1] PCs coefficients for interpretations of PCs

- Random vector : $\boldsymbol{x} = (x_1, \ldots, x_p)^t \sim Cov(\boldsymbol{x}) = \Sigma$

- *kth* PC : $p_k = v_{k1}x_1 + \cdots + v_{kl}x_l + \cdots + v_{kp}x_p$

Correlation coefficient between the *kth* PC and *lth* variable :

$$\boxed{Corr(x_l, p_k) = \frac{v_{kl}\sqrt{l_k}}{\sqrt{\sigma_{ll}}} = \gamma_{lk},\ l = 1, \ldots, p,\ k = 1, \ldots, p.} \quad (2.4.3)$$

Importance(contribution) of x_l to p_k

(Proof) $Corr(x_l, p_k) = \dfrac{Cov(x_l, p_k)}{\sqrt{Var(x_l)}\sqrt{Var(p_k)}}$

- $p_k = \boldsymbol{v}_k^t \boldsymbol{x}$,
- $x_l = e_l^t \boldsymbol{x}$ where $e_l = (0, \ldots, 1, \ldots, 0)^t$,
- $\Sigma \boldsymbol{v}_k = l_k \boldsymbol{v}_k$.

$$\Rightarrow Cov(x_l, p_k) = Cov(e_l^t \boldsymbol{x}, \boldsymbol{v}_k^t \boldsymbol{x}) = e_l^t \Sigma \boldsymbol{v}_k = e_l^t l_k \boldsymbol{v}_k = l_k e_l^t \boldsymbol{v}_k = l_k v_{kl}$$

- $Var(x_l) = \sigma_{ll}$
- $Var(p_k) = l_k$

$$\Rightarrow Corr(x_l, p_k) = \frac{l_k v_{kl}}{\sqrt{\sigma_{ll}}\sqrt{l_k}} = \frac{\sqrt{l_k}\, v_{kl}}{\sqrt{\sigma_{ll}}}$$

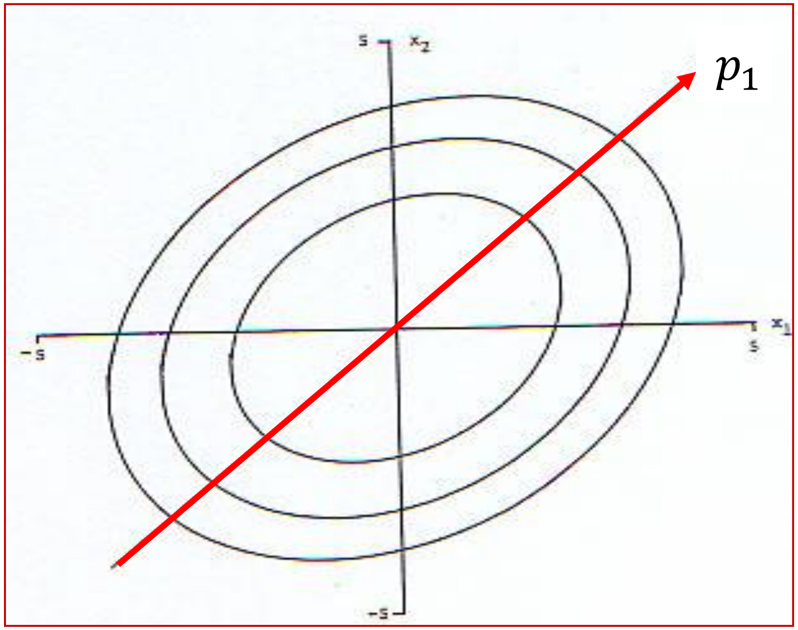
2.4 Selection and goodness-of-fit of PCs

[Example 2.4.1] Selection of PCs in PCA by Spectral Decomposition of Covariance Matrix

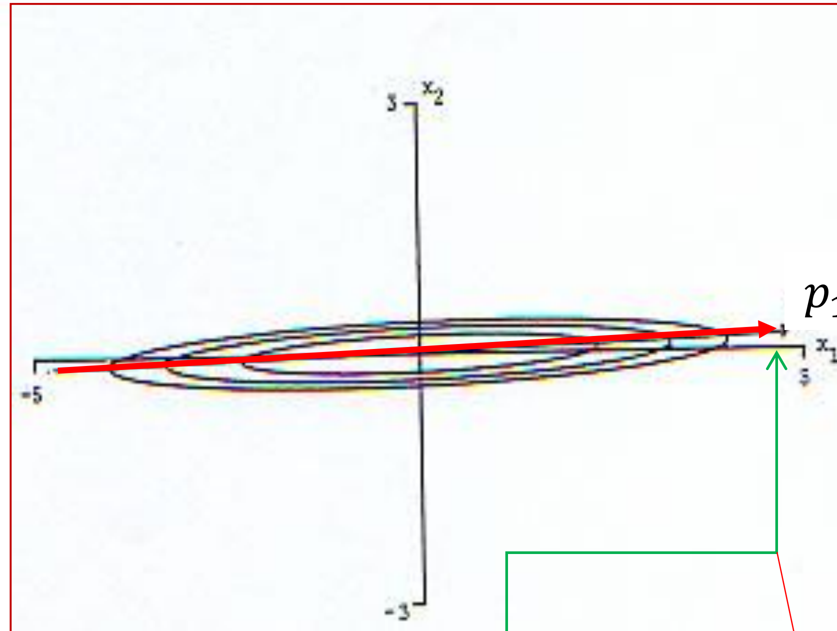
Children height and weights : $x_1: cm\ ,\ x_2: g$ | $x_1: mm\ ,\ x_2: g$

[step 2] covariance matrix	$\Sigma_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix}$	$\Sigma_2 = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}$
[step 3] eigenvlaues : (l_1, l_2) eigenvetors : $V=(v_1,\ v_2)$	$(l_1, l_2) = (124,\ 36)$ v_1 v_2 0.707 −0.707 explain rate 0.707 0.707	$(l_1, l_2) =$ (8024.369, 55.631) v_1 v_2 -0.998 0.055 explain rate 055 -0.998
[step 4] select eigenvalues	explain rate : 77.5%, 22.5% description ratio of $m=1$ eigenvalue $(l_1 = 124)$ is 77.5%	explain rate: 99.31%, 0.69% description ratio of $m=1$ eigenvalue$(l_1 = 8024.369)$ is 99.34%
[step 5] PC coefficeint and PC	first PC : $p_1 = 0.707x_1 + 0.707x_2$	first PC : $p_1 = -0.998x_1 - 0.055x_2$
interpretation and explnatory power of PC	The explanatory power of the first principal component is 77.5%.	The explanatory power of the first principal component is 99.31%.

2.4 Selection and goodness-of-fit of PCs



fit = 77.5%



fit = 99.31%

Contours of constant probability based on

$$\Sigma_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix}$$

$$p_1 = 0.707x_1 + 0.707x_2$$

$$\Sigma_2 = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}$$

$$p_1 = 0.998x_1 + 0.055x_2$$

Difference between PCs for the two scales of measurement in x1

Both variables have the same degree of variation

Most of the variation is the direction of x1

2.5 Algebraic derivation of sample PCs

[Table 2.5.1] Algebraic Representation of Covariance Matrix and Correlation Matrix

Centered data matrix : $Y = HX$	Standardized data matrix : $Z = HXD_s^{-1/2}$
Covariance matrix : $S = \frac{1}{n-1} Y^t Y = \frac{1}{n-1} X^t HX$	Correlation matrix : $R = \frac{1}{n-1} Z^t Z = \frac{1}{n-1} D_s^{-1/2} X^t HXD_s^{-1/2}$
$R = D_s^{-1/2} S \, D_s^{-1/2}$	

Note:
PCA based on the S in the sensitivity of PCS to the measurement units of variances.

2.5 Algebraic derivation of sample PCs

[Table 2.5.2] for S

[Table 2.5.3] Derivation & Properties of sample PCs based on the Spectral decomposition of R

- $n \times p$ data matrix : $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^t$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^t$, $i = 1, \ldots, n$.
- $n \times p$ standardized data matrix : $Z = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^t$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})^t$, $i = 1, \ldots, n$.
- spectral decomposition :

$$\boxed{Z^t Z/(n-1) = R = VDV^t = \sum_{k=1}^{p} l_k \boldsymbol{v}_k \boldsymbol{v}_k^t}$$

  - $V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$: orthogonal matrix satisfying $V^t V = V V^t = I$
  - $D = diag(l_1, \ldots, l_p)$: diagonal matrix of eigenvalues satisfying $l_1 \geq \cdots \geq l_p > 0$

- $k-th$ PC : $\boxed{p_k = v_{k1} z_1 + v_{k2} z_2 + \cdots + v_{kp} z_p = \boldsymbol{v}_k^t \boldsymbol{z}, \ k = 1, \ldots, p.}$ (2.5.2)

- PC score : $\boldsymbol{p}_i = \begin{bmatrix} p_{i1} \\ \vdots \\ p_{ip} \end{bmatrix} = V^t \boldsymbol{z}_i$, $i = 1, \ldots, n$.

  $\Rightarrow \boxed{P = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n]^t = ZV}$: $n \times p$ PC score matrix.

- Goodness-of-fit : $\boxed{t_m = \dfrac{\sum_{k=1}^{m} l_k}{p} \times 100}$

- Correlation coefficient matrix between principal component and variable :

  $\Gamma = VD^{1/2}$

  - $D^{1/2} = diag(\sqrt{l_1}, \ldots, \sqrt{l_p})$: $l_1 \geq \cdots \geq l_p > 0$

2.5 Algebraic inducement of sample pc

- **[Example 2.5.1] KLPGA player's grades (www.klpga.com, 2006)**

[Data 1.3.2] KLPGA player scores (klpga.txt)

- Raw variable : $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5, x_6)^t$

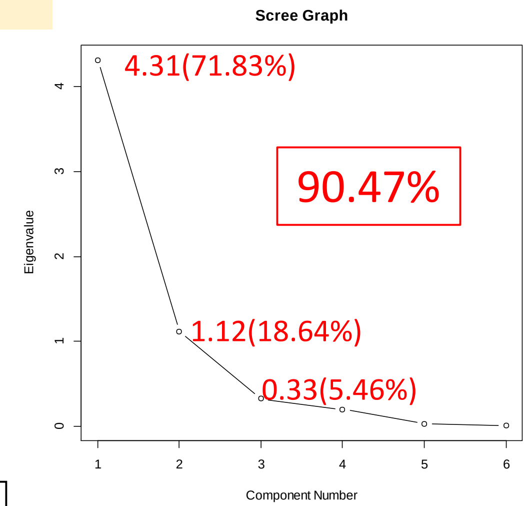
$$= (\boxed{평균퍼팅수,\ 그린적중율,\ 파세이브율,\ 파브레이크율},\ \boxed{평균타수,\ 상금율})^t$$

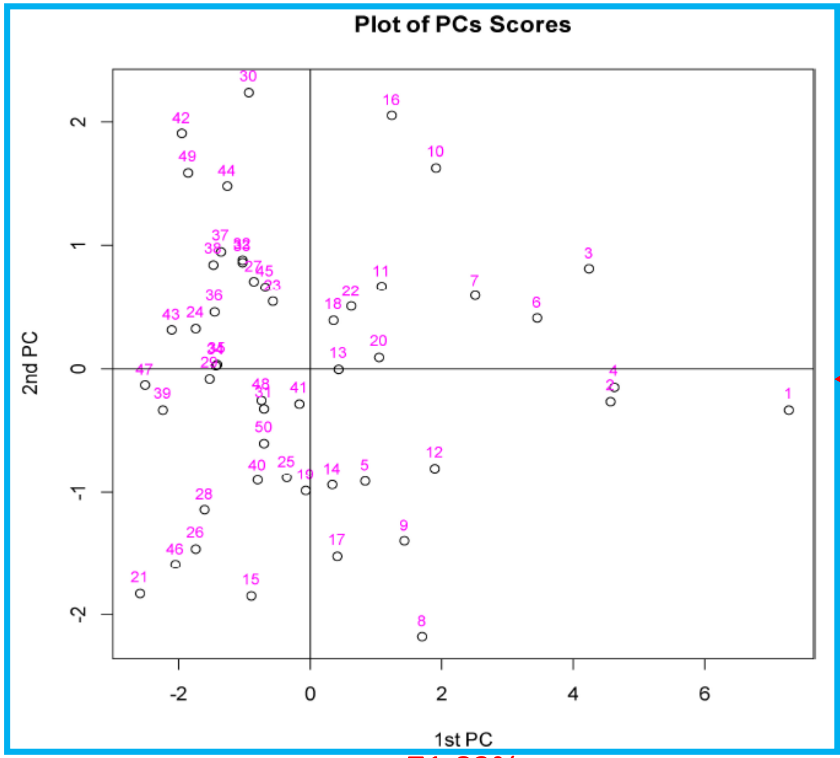
Putting average, Green in regulation % , Par save % , Par break % , Scoring average, Prize rate

- standardization variable : $\boldsymbol{z} = (z_1, z_2, z_3, z_4, z_5, z_6)^t$

- first PC : $p_1 = \boldsymbol{v}_1^t \boldsymbol{z} = -0.215z_1 + 0.39z_2 + 0.44z_3 + 0.45z_4 - 0.48z_5 + 0.43z_6$

$p_2 = \boldsymbol{v}_2^t \boldsymbol{z} = 0.84z_1 + 0.53z_2 + 0.04z_3 - 0.05z_4 + 0.00z_5 - 0.07z_6$





18.63%

71.83%

player	p_1	p_2
1	7.264	-0.327
2	4.566	-0.262
3	4.230	0.810
4	4.612	-0.143
5	0.832	-0.906
6	3.448	0.407
7	2.500	0.592
8	1.689	-2.170
9	1.432	-1.392
10	1.906	1.633
......		
47	-2.507	-0.132
48	-0.742	-0.254
49	-1.852	1.597
50	-0.699	-0.595

KLPGA Korea Ladies Professional Golf Association

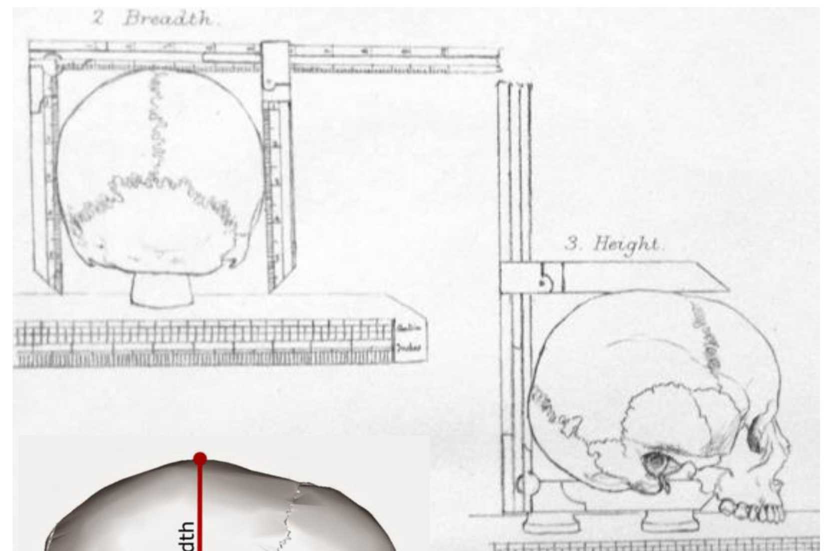
British Open: 2013 — Muirfield, Gullane, Scotland

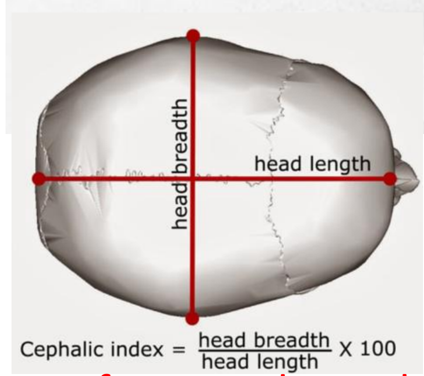
2.5 Algebraic inducement of sample PCs

- **[Example 2.5.2] [Data 2.5.1] Skull data(22 man, 18 woman of ancient race Naqada from Egypt)**

skull	L	B	H	OH	U	S	Q	FH	NB	NH	BL	HL
5F	-2.2700	-0.7810	-1.7400	-1.3300	-1.8900	-0.9020	-1.4800	-1.9000	-1.4100	-1.0600	1.4120	0.4960
7M	0.0364	0.3764	-0.9660	-0.6850	0.3166	-0.6520	-0.9220	0.1855	-0.4530	0.4895	0.2161	-1.1200
10F	-0.7070	1.5570	-0.3270	0.1476	-0.4180	-0.4010	1.1240	0.5110	0.6635	-0.8000	1.7180	0.3807
13M	0.9280	1.4390	-0.6750	-0.0960	1.2610	0.1838	1.5570	1.3570	2.5770	0.7475	0.2161	-1.6600

variable		
L	Greatest length	Head
B	Breadth	Head
H	Height	Head
OH	Auricular height	Head
U	Circumference above the superciliary ridger	Head
S	Sagittal circumferences	Head
Q	Cross-circumference	Head
FH	Upper face height	Face
NB	Nasal breadth	Face
NH	Nasal height	Face
BL	Cephalic index	Index
HL	Ratio of height to length	Index





BL shows the shape of head and classify the pattern of race and people

2.5 Algebraic inducement of sample PCs

	PC score		
	p_1	p_2	p_3
5F	4.925	-0.288	-0.671
7M	0.830	1.231	1.141
10F	0.258	1.768	-2.052
13M	-2.613	3.266	0.714
26M	-4.388	0.779	-1.067
32M	2.012	2.637	0.628
43F	3.066	1.197	-0.511
45F	3.539	0.612	1.645
46F	1.682	1.057	-2.648
52M	0.653	0.617	0.151
....			
145F	-1.904	-1.610	-1.400
146F	1.169	-0.933	0.997
148M	0.311	-2.109	1.109
151M	1.087	-1.148	-0.596
152M	-4.170	0.295	-0.575

	Head							Face			Index	
PCs	L z_1	B z_2	H z_3	OH z_4	U z_5	S z_6	Q z_7	FH z_8	NB z_9	NH z_{10}	BL z_{11}	HL z_{12}
$p_1 =$	-0.39	-0.16	-0.31	-0.34	-0.40	-0.38	-0.33	-0.25	-0.25	-0.17	0.21	0.07
$p_2 =$	-0.03	0.50	-0.38	-0.22	0.12	-0.20	0.15	0.18	0.32	0.20	0.39	0.38
$p_3 =$	0.25	-0.38	-0.21	-0.25	0.06	-0.08	-0.36	0.22	0.11	0.11	-0.48	-0.49

General PC of Size

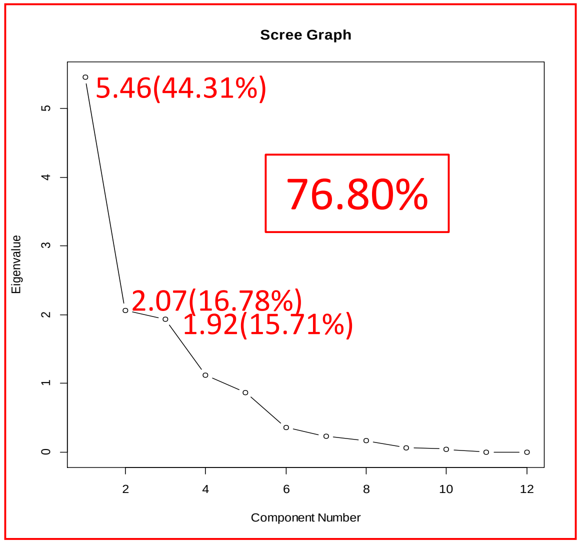
$$p_1 = v_1^t z = -0.39z_1 - 0.16z_2 + \cdots + 0.07z_{12}$$

Volume

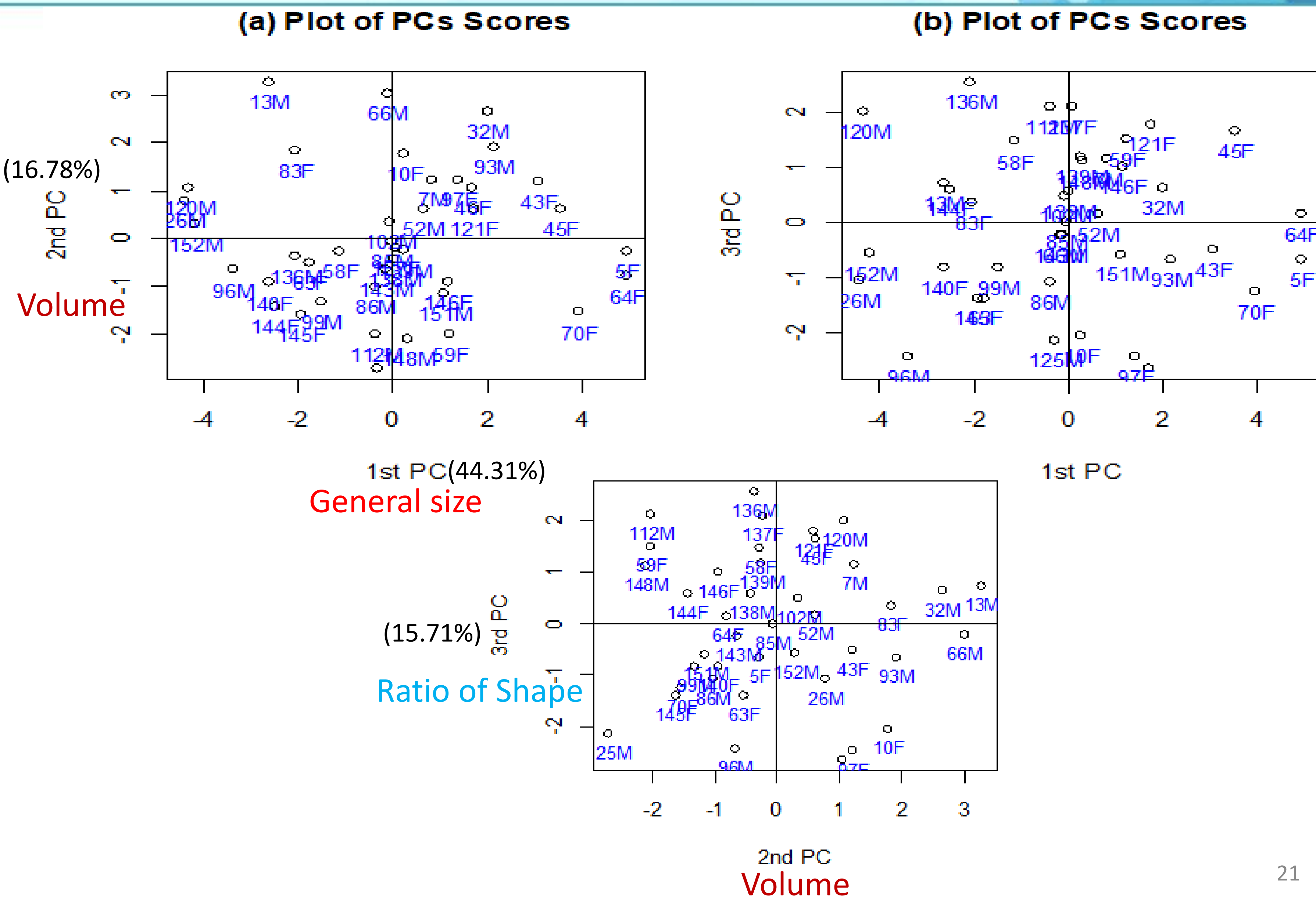
$$p_2 = v_2^t z = -0.03z_1 + 0.50z_2 + \cdots - 0.38z_{12}$$

$$p_3 = v_3^t z = 0.25z_1 - 0.38z_2 + \cdots - 0.49z_{12}$$

Ratio of Shape



2.5 Algebraic inducement of sample pc



2.6 Visualizations of PCA

- Biplot : Gabriel(1971)

plots of the *n* observations and *p* variables in 2-dimensional space with providing relationships between them.

$$Y = U\Lambda V^t = \sum_{k=1}^{p} \lambda_k u_k v_k^t$$

SVD: Singular Value Decomposition

$$Y = U(V\Lambda)^t = GH^t$$ Factorization

Geometric Properties : Choi and Shin(2013, Chapter 1)

(1) $h_j^t h_k \doteq s_{jk}$;

(2) $\| h_j \|^2 \doteq s_j^2$;

(3) $\cos(\theta_{jk}) \doteq r_{jk}$;

$$H = V\Lambda = (v_1\lambda_1, \ldots, v_p\lambda_p) = (h_1, \ldots, h_p)^t$$

$$(n-1)S = Y^tY = H^tH$$

$$G = U = (u_1, \ldots, u_p) = (g_1, \ldots, g_n)^t$$

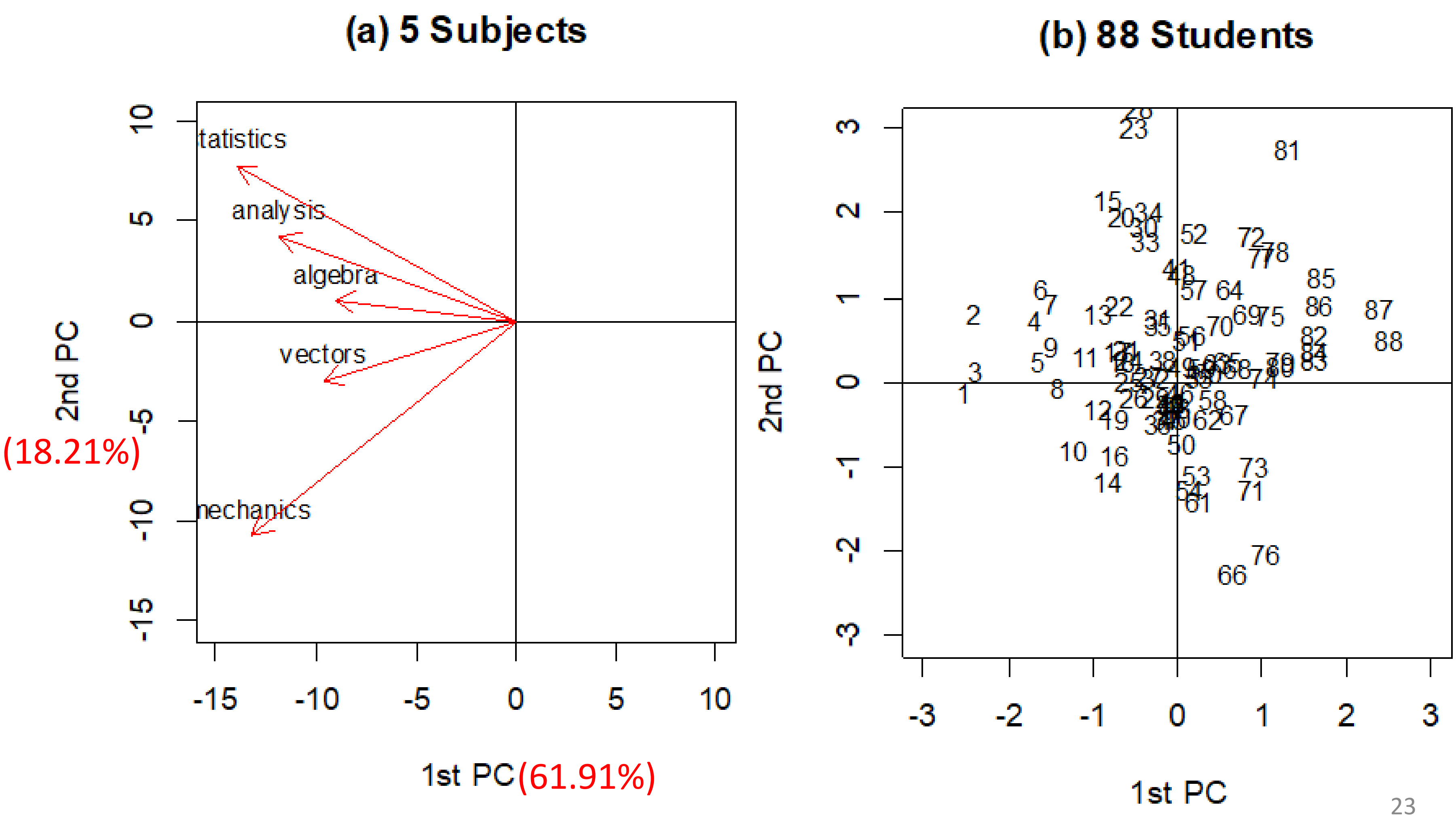
$$\|y_r - y_s\|^2_{S^{-1}} = [(y_r - y_s)^t S^{-1} (y_r - y_s)] = (n-1)\|g_r - g_s\|^2$$

Remark:

PC scores Matrix: P = YV = U$\Lambda V^t V$ = U Λ = G Λ $\Rightarrow$ P $\cong$ G

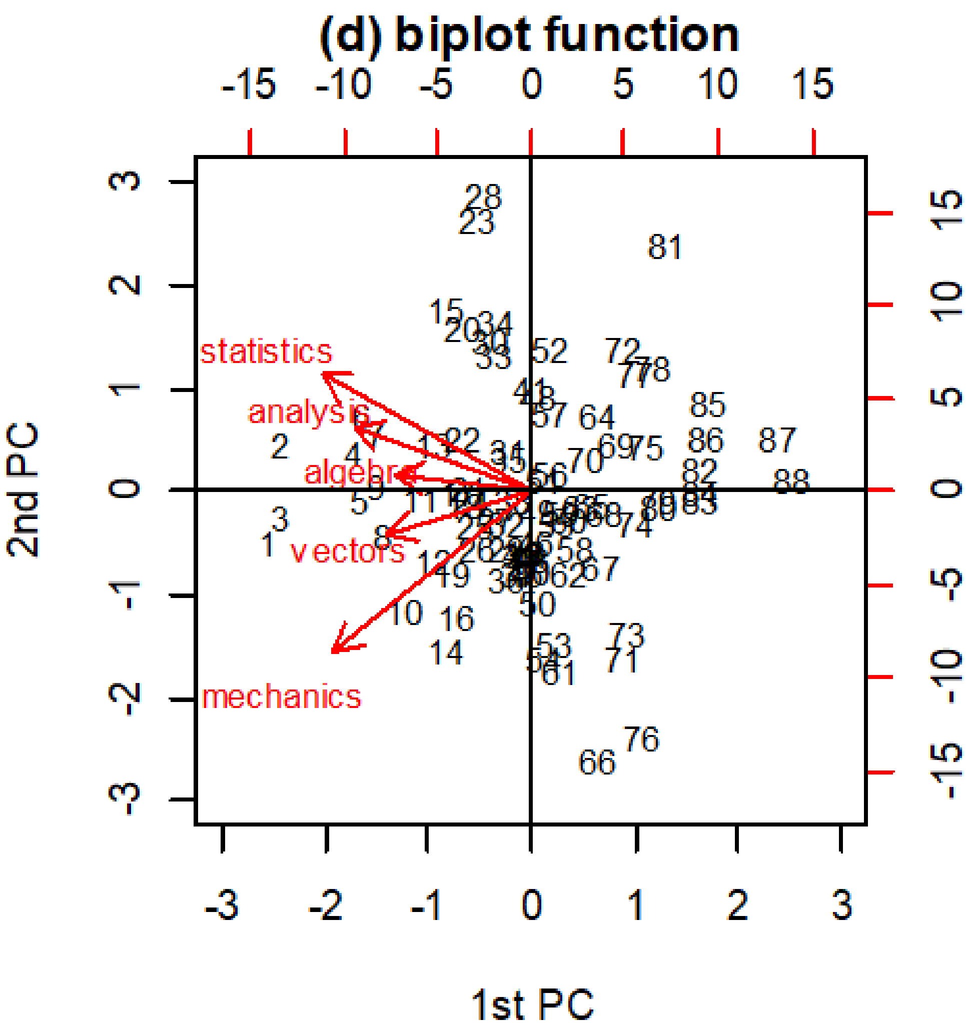
2.6 Visualizations of PCA

- [Example 2.6.1]5subjects-Pcbiplot



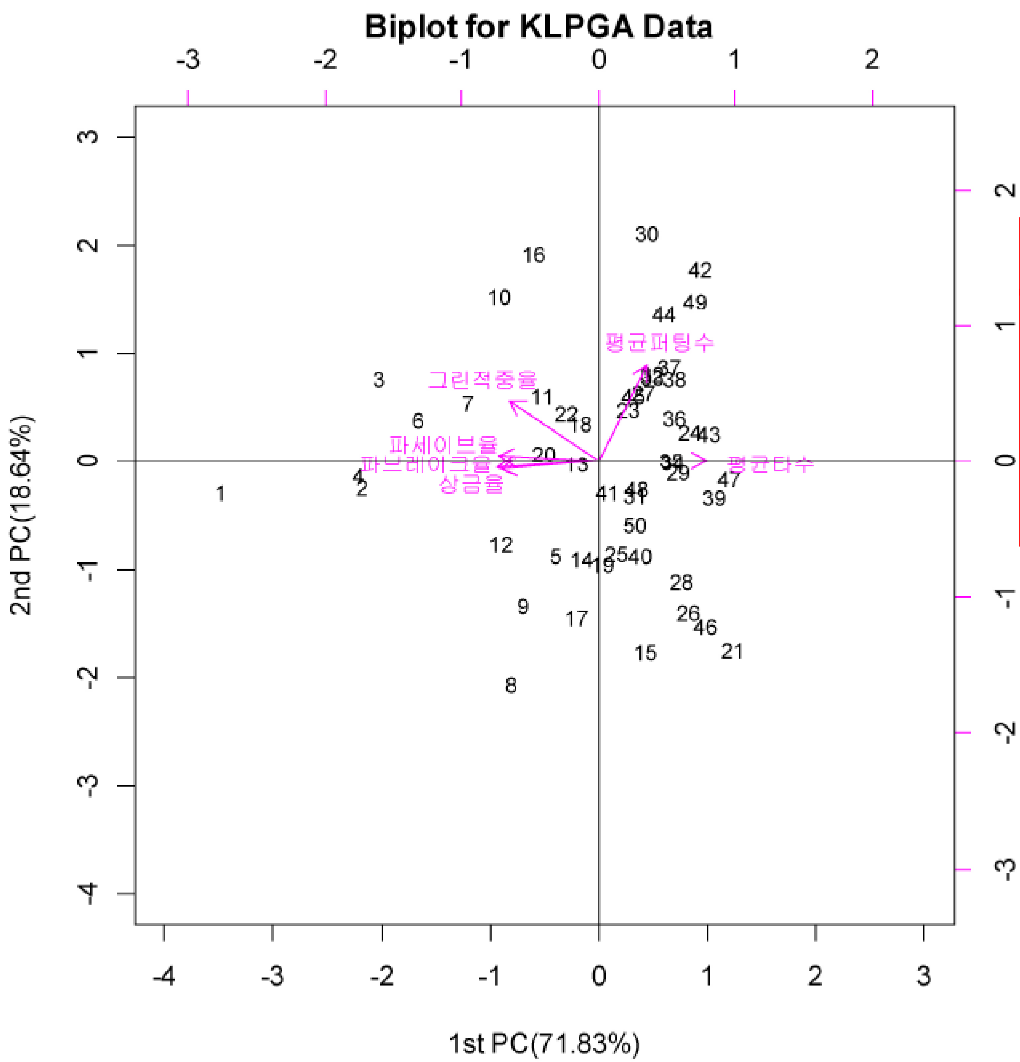
2.6 Visualizations of PCA

- [Example 2.6.1]5subjects-Pcbiplot



2.6 Visualizations of PCA

- [Example 2.6.2] klpga-PCbiplot



Goodness-of-fit

$$t_m = \frac{\sum_{k=1}^{m} \lambda_k^2}{\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_p^2} \times 100$$

2.7 R for PCA : Practice Time

R-code:

PCA and PC Biplot	
princomp()	Spectral Decomposition of S or R
prcomp()	SVD of Y and Z
PCA	klpga-PCAfunctions.R klpga-PCAsteps.R kull-PCAsvd.R
PC Biplot	5subjects-PCbiplot

R-code list of Chapter 2 Principal Component Analysis

3subjects-PCAsteps.R	[R-코드 2.2.1]	[자료 1.1.1]의 [PCA 수행단계]
5subjects-PCAsteps.R	[R-코드 2.2.2]	[자료 1.3.1]의 [PCA 수행단계]
5subjects-PCAsteps-scree.R	[보기 2.2.2]	[그림 2.4.1]의 스크리그림
example241-PCAsteps.R	[R-코드 2.4.1]	[보기 2.4.1]의 [PCA 수행단계]와 상관계수 γ_{lk} 계산
klpga-PCAsteps.R	[R-코드 2.5.1]	[자료 1.3.1]의 스펙트럼분해에 의한 [PCA 수행단계]
skull-PCAsteps.R	[R-코드 2.5.2]	두개골 자료의 스펙트럼분해에 의한 [PCA 수행단계]
skull-PCAsvd.R	[R-코드 2.5.3]	두개골 자료의 비정칙값분해에 의한 [PCA 수행단계]
5subjects-PCbiplot.R	[R-코드 2.6.1]	두 가지 시험성적자료에 대한 주성분행렬도
klpga-PCbilpot.R	[R-코드 2.6.2]	KLPGA 선수 성적의 주성분 행렬도
klpga-PCAfunctions.R	[R-코드 2.7.1]	KLPGA 선수 성적의 주성분분석을 위한 함수 princomp()와 prcomp()를 활용
censustract-PCbiplot.R	[연습문제 2.7]	61개 지역의 총인구조사 상관행렬의 주성분행렬도
turtle-PCbiplot.R	[연습문제 2.11]	거북이 등딱지 자료의 공분산행렬에 대한 주성분행렬도

2.7 R for PCA : Practice Time

[R-code 2.5.1] klpga-PCAsteps.R : Spectral Decomposition

```
# PCA Steps for KLPGA

#[Step 1] Data Matrix X
Data1.3.2<-read.table("klpga.txt", header=T)
X=Data1.3.2
rownames<-rownames(X)

#[Step 2] Covariance Matrix S(or Correlation Matix R)
R=round(cor(X),3)
R

#[Step 3] Spectral Decomposition
eigen.R=eigen(R)
round(eigen.R$values, 2) # Eigenvalues
V=round(eigen.R$vectors, 2) # Eigenvectors

#[Step 4] Choice of Eigenvalues and Eigenvectors
gof=eigen.R$values/sum(eigen.R$values)*100 # Goodness-of fit
round(gof, 2)

#[Step 5] PCs : liner combination of original variables
V2=V[,1:2]
V2
#[Step 6] PCS, PCs Scores and New Data Matrix P
Z=scale(X, scale=T) # Standardized Data Matrix
Z
P=Z%*%V2                # PCs Scores
round(P, 3)

#[Step 7] Plot of PCs Scores
plot(P[,1], P[, 2], main="Plot of PCs Scores", xlab="1st PC", ylab="2nd PC")
text(P[,1], P[, 2], labels=rownames, cex=0.8, col="blue", pos=3)
abline(v=0, h=0)

#Correlations bt PCs and variables
D=diag(sqrt(eigen.R$values[1:2]))
corr=V2%*%D
corr
```

2.7 R for PCA : Practice Time

[R-code 2.5.3] skull-PCAsvd.R : Singular Value Decomposition

```
# PCA Steps based on the SVD for Skull Data
#[Step 1] Data Matrix X
Data1.3.2<-read.table("skull.txt", header=T)
Z=as.matrix(Data1.3.2)
rownames<-rownames(Z)
colnames<-colnames(Z)
n=nrow(Z)

#[Step 2] Singular Values Decomposition
svd.Z=svd(Z)
U=svd.Z$u # Right singular vectors
V=svd.Z$v # Left singular vectors : Eigenvectors
round(V, 2)
D=diag(svd.Z$d)
#[Step 3] Choice of Singular Values and Eigenvectors
round(svd.Z$d, 2)
eigen=(svd.Z$d)^2
round(eigen/(n-1), 2)
gof=eigen/sum(eigen)*100 # Goodness-of fit
round(gof, 2)

#[Step 5] PCs : liner combination of original variables
V3=V[,1:3]
V3
round(t(V3), 2)
#[Step 6] PCS, PCs Scores and New Data Matrix P
Z # Standardized Data Matrix
P=U%*%D              # PCs Scores : P=Z%*%V3
round(P, 3)
```

```
#[Step 7] Plot of PCs Scores
par(mfrow=c(2,2))
plot(P[,1], P[, 2], main="(a) Plot of PCs Scores", xlab="1st PC", ylab="2nd PC")
text(P[,1], P[, 2], labels=rownames, cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

plot(P[,1], P[, 3], main="(b) Plot of PCs Scores", xlab="1st PC", ylab="3rd PC")
text(P[,1], P[, 3], labels=rownames, cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
plot(P[,2], P[, 3], main="(c) Plot of PCs Scores", xlab="2nd PC", ylab="3rd PC")
text(P[,2], P[, 3], labels=rownames, cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

#Correlations bt PCs and variables
D=diag(svd.Z$d[1:3]/sqrt(n-1))
corr=V3%*%D
round(corr, 3)
```

2.7 R for PCA : Practice Time

[R-code 2.6.1] 5subjects-pcbiplot.R : Biplot based on the SVD

```
# PC Biplots for 5 Subjects Exam
library("MVT")
data(examScor)
X=examScor
n <- nrow(X)
rownames(X)
colnames(X)
joinnames=c(rownames(X),colnames(X))

Y <- scale(X,scale=F)

# Biplot based on the Singular Value Decomposition
svd.Y <- svd(Y)
U <- svd.Y$u
V <- svd.Y$v
D <- diag(svd.Y$d)
G <- (sqrt(n-1)*U)[,1:2]
H <- (sqrt(1/(n-1))*V%*%D)[,1:2]
C<- rbind(G, H)
rownames(G)<-rownames(X)
rownames(H)<-colnames(X)
rownames(C)<-joinnames

# Godness-of-fit
eig <- (svd.Y$d)^2
per <- eig/sum(eig)*100
gof <- sum(per[1:2])
per
gof
```

```
# Biplots
par(mfrow=c(2,2))
par(pty="s")
lim1 <- range(pretty(H))
plot(H[,1],H[,2],xlab="1st PC",ylab="2nd PC", main="(a) 5 Subjects",
          xlim=lim1,ylim=lim1,pch=15,col=2, type="n")
abline(v=0,h=0)
text(H[,1], H[,2],colnames(X),cex=0.8,col=1,pos=3)
arrows(0,0,H[,1],H[,2],col=2,code=2, length=0.1)

lim2 <- range(pretty(G))
plot(G[,1],G[,2],xlab="1st PC",ylab="2nd PC", main="(b) 88 Students",
            xlim=lim2,ylim=lim2,pch=16, type="n")
abline(v=0,h=0)
text(G[,1],G[,2],rownames(X),cex=0.8,pos=3)

lim3 <- range(pretty(C))
plot(C[,1],C[,2],xlab="1st PC",ylab="2nd PC", main="(c) 5 Subjects and 88
Students",
           xlim=lim3,ylim=lim3,pch=16, type="n")
abline(v=0,h=0)
text(C[,1],C[,2],joinnames,cex=0.8,pos=3)
arrows(0,0,C[89:93,1],C[89:93,2],col=2,code=2, length=0.1)

biplot(G,H, xlab="1st PC",ylab="2nd PC", main="(d) biplot function",
                xlim=lim2,ylim=lim2,cex=0.8,pch=16)
abline(v=0,h=0)
```

2.7 R for PCA : Practice Time

[R-code 2.7.1] klpga-PCAfunctions.R : princomp(), prcomp()

```
Data1.3.2<-read.table("klpga.txt", header=T)
X<-Data1.3.2

# PCA based on the SD using princomp( )
pca.R<-princomp(X, cor=T)
summary(pca.R, loadings=T) # 설명력, 주성분계수
round(pca.R$scores, 3)  # 주성분점수
screeplot(pca.R, type="lines") # 스크리그림

# 주성분 행렬도
biplot(pca.R, scale=0, xlab="1st PC",ylab="2nd PC",
                    main="PC Biplot for KLPGA Data ")
abline(v=0, h=0)

# PCA on the SVD using prcomp( )
pcasvd.Z<-prcomp(X, scale=T)
summary(pcasvd.Z)  # 설명력
round(pcasvd.Z$rotation, 3) # 주성분계수
screeplot(pcasvd.Z, type="lines") #스크리그림

# 주성분 행렬도
biplot(pcasvd.Z, scale=0,  xlab="1st PC",ylab="2nd PC",
                    main="PC Biplot for KLPGA Data ")
abline(v=0, h=0)
```