Multivariate Analysis — 1st Sem 2019/2020

1st Test – 5th Nov 2019 –

Group I

1. $(X,Y)^t \sim N_2(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$: $Var(X) = Var(Y)$

$$\underset{\sim}{Z} = \begin{bmatrix} X+Y \\ X-Y \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \underset{\sim}{A} \begin{bmatrix} X \\ Y \end{bmatrix}, \text{ where } \underset{\sim}{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$Var(Z) = A\,Var(X)\,A^t = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{11} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11}+\sigma_{12} & \sigma_{12}+\sigma_{11} \\ \sigma_{11}-\sigma_{12} & \sigma_{12}-\sigma_{11} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} 2(\sigma_{11}+\sigma_{12}) & 0 \\ 0 & 2(\sigma_{11}-\sigma_{12}) \end{bmatrix} = \underset{\sim}{\Sigma}_Z$$

Since (i) $Cov(X+Y, X-Y) = 0$ and

(ii) $\underset{\sim}{Z} = \underset{\sim}{A}\,\underset{\sim}{X}$ and $\underset{\sim}{X} \sim N_2(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$

then $\underset{\sim}{Z} \sim N_2(\underset{\sim}{\mu}_Z, \underset{\sim}{\Sigma}_Z)$ and

$Z_1 \perp\!\!\!\perp Z_2$ i.e. $X+Y \perp\!\!\!\perp X-Y$ QED

2. $Var(\underset{\sim}{X}_1 + \underset{\sim}{X}_2) \neq \underset{\sim}{\Sigma}_{11} + \underset{\sim}{\Sigma}_{22} + 2\,\underset{\sim}{\Sigma}_{12}$

$$E(\underset{\sim}{X}_1 + \underset{\sim}{X}_2) = E(\underset{\sim}{X}_1) + E(\underset{\sim}{X}_2) = \underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2$$

$$\mathrm{Var}(\underset{\sim}{X}_1 + \underset{\sim}{X}_2) = E\left((\underset{\sim}{X}_1+\underset{\sim}{X}_2)(\underset{\sim}{X}_1+\underset{\sim}{X}_2)^t\right) - (\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)^t$$

$$= E\left\{ \underset{\sim}{X}_1\underset{\sim}{X}_1^t + \underset{\sim}{X}_1\underset{\sim}{X}_2^t + \underset{\sim}{X}_2\underset{\sim}{X}_1^t + \underset{\sim}{X}_2\underset{\sim}{X}_2^t \right\} - \underset{\sim}{\mu}_1\underset{\sim}{\mu}_1^t - \underset{\sim}{\mu}_1\underset{\sim}{\mu}_2^t$$

$$- \underset{\sim}{\mu}_2\underset{\sim}{\mu}_1^t - \underset{\sim}{\mu}_2\underset{\sim}{\mu}_2^t =$$

$$= \left[E(\underset{\sim}{X}_1\underset{\sim}{X}_1^t) - \underset{\sim}{\mu}_1\underset{\sim}{\mu}_1^t\right] + \left[E(\underset{\sim}{X}_1\underset{\sim}{X}_2^t) - \underset{\sim}{\mu}_1\underset{\sim}{\mu}_2^t\right]$$

$$\left[E(\underset{\sim}{X}_2\underset{\sim}{X}_1^t) - \underset{\sim}{\mu}_2\underset{\sim}{\mu}_1^t\right] + \left[E(\underset{\sim}{X}_2\underset{\sim}{X}_2^t) - \underset{\sim}{\mu}_2\underset{\sim}{\mu}_2^t\right]$$

$$= \mathrm{Var}(\underset{\sim}{X}_1) + \mathrm{Cov}(\underset{\sim}{X}_1, \underset{\sim}{X}_2) + \mathrm{Cov}(\underset{\sim}{X}_2, \underset{\sim}{X}_1)$$

$$+ \mathrm{Var}(\underset{\sim}{X}_2) = \underset{\sim}{\Sigma}_{11} + \underset{\sim}{\Sigma}_{12} + \underset{\sim}{\Sigma}_{21} + \underset{\sim}{\Sigma}_{22}$$

In general, $\mathrm{Cov}(X_{1i}, X_{2j}) \neq \mathrm{Cov}(X_{1j}, X_{2i})$

and being so $\underset{\sim}{\Sigma}_{12} \neq \underset{\sim}{\Sigma}_{21}$

3. $\underset{\sim}{X} \sim N_3(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$, $\underset{\sim}{\mu} = (-3, 1, 4)^t$, $\underset{\sim}{\Sigma} = \underset{\sim}{\Gamma}\underset{\sim}{\Lambda}\underset{\sim}{\Gamma}^t$

$$\underset{\sim}{\Sigma} = \begin{bmatrix} 1 & 1 & 0 \\ & 2 & 0 \\ & & 3 \end{bmatrix}, \quad \underset{\sim}{\Gamma} = \begin{bmatrix} 0 & 0.53 & 0.85 \\ 0 & 0.85 & -0.53 \\ 1 & 0.00 & 0.00 \end{bmatrix}, \quad \underset{\sim}{\Lambda} = \begin{bmatrix} 3 & & \\ & 2.62 & \\ & & 0.38 \end{bmatrix}$$

(a) $P(MD(\underset{\sim}{X}, \underset{\sim}{\mu}) > 3.0574) = 1 - P(MD^2(\underset{\sim}{X}, \underset{\sim}{\mu}) \leq 9.348)$

$$= 1 - F_{\chi^2_{(3)}}(9.348) = 1 - 0.9707 = 0.0293$$

$$0.95 = F_{\chi^2_{(3)}}(7.815) \leq F_{\chi^2_{(3)}}(9) \leq F_{\chi^2_{(3)}}(9.348) = 0.975$$

$$0.025 \leq 1 - F_{\chi^2_{(3)}}(9) \leq 0.05$$

(b) If $\underset{\sim}{Y} = \underset{\sim}{\Sigma}^{-1/2}(\underset{\sim}{X} - \underset{\sim}{\mu})$ then

(i) $\underset{\sim}{Y} \sim N_3(\cdot,\cdot)$ since $\underset{\sim}{Y} = \underset{\sim}{A}\underset{\sim}{X}$ and $\underset{\sim}{X}$ has $N_3(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$

(ii) $E(\underset{\sim}{Y}) = \underset{\sim}{\Sigma}^{-1/2} E(\underset{\sim}{X} - \underset{\sim}{\mu}) = 0$

$$Var(\underset{\sim}{Y}) = Var\left(\underset{\sim}{\Sigma}^{-1/2}(\underset{\sim}{X} - \underset{\sim}{\mu})\right) =$$

$$= \underset{\sim}{\Sigma}^{-1/2} \underset{\sim}{\Sigma} \left(\underset{\sim}{\Sigma}^{-1/2}\right)^t =$$

$$= \left(\underset{\sim}{\Gamma}\underset{\sim}{\Lambda}^{-1/2}\underset{\sim}{\Gamma}^t\right)\underset{\sim}{\Gamma}\underset{\sim}{\Lambda}\underset{\sim}{\Gamma}^t\left(\underset{\sim}{\Gamma}\underset{\sim}{\Lambda}^{-1/2}\underset{\sim}{\Gamma}^t\right)$$

$$= \underset{\sim}{\Gamma}\underset{\sim}{\Lambda}^{-1/2}\underset{\sim}{\Lambda}\underset{\sim}{\Lambda}^{-1/2}\underset{\sim}{\Gamma}^t = \underset{\sim}{\Gamma}\underset{\sim}{\Gamma}^t = \underset{\sim}{I}$$

where $\underset{\sim}{\Sigma}^{-1/2} = \underset{\sim}{\Gamma}\underset{\sim}{\Lambda}^{-1/2}\underset{\sim}{\Gamma}^t =$

$$= \begin{bmatrix} 0 & 0.53 & 0.85 \\ 0 & 0.85 & -0.53 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & & \\ & 1/\sqrt{2.62} & \\ & & 1/\sqrt{0.38} \end{bmatrix} *$$

$$* \begin{bmatrix} 0 & 0 & 1 \\ 0.53 & 0.85 & 0 \\ 0.85 & -0.53 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0.53 & 0.85 \\ 0 & 0.85 & -0.53 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0.577 \\ 0.325 & 0.526 & 0 \\ 1.376 & -0.851 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 1.342 & -0.447 & 0 \\ -0.447 & 0.894 & 0 \\ 0 & 0 & 0.577 \end{bmatrix}$$

Group II

1. Given the classical estimates of the eigenvalue we conclude that the 1st three PCs explain 86.46% of the observed total variability

$$\frac{\sum_{i=1}^{3} \hat{d}_i}{\sum_{i=1}^{7} \hat{d}_i} = \frac{\sum_{i=1}^{3} \hat{d}_i}{7} = 0.8646 .$$

Note that the estimates were obtained from the standardized variables so:

$$tr(\underset{\sim}{R}) = 7 = \sum_{i=1}^{7} \hat{d}_i$$

If we use the criteria to select the number of PCs which $\hat{d}_i \geq \bar{\hat{d}}$ (=) $\hat{d}_i \geq 1$ then only one PC should be selected since

$$\hat{d}_i \geq 1 \ (=) \ \sqrt{\hat{d}_i} \geq 1$$

and the R output returns the value $\sqrt{\hat{d}_i}$, $i=1,\dots,7$.

2. $$\hat{y}_1 = +0.45\left(\frac{x_1 - 2.69}{0.51}\right) + 0.31\left(\frac{x_2 - 1.79}{0.05}\right) + 0.40\left(\frac{x_3 - 13.17}{1.50}\right)$$

$$+0.43\left(\frac{x_4 - 2.02}{0.94}\right) + 0.45\left(\frac{x_5 - 6.21}{0.40}\right)$$

$$+0.24\left(\frac{x_6 - 41.28}{3.47}\right) + 0.30\left(\frac{x_7 - 28.52}{6.15}\right)$$

3. The first PC is a weighted mean of all the events, and since large scores on the events means a better performance, competitors with the best marks have very hu values on the 1st PC and competitors with bad performances have the lowest possible values on PC1.

the second PC, PC_2, is essentially a contrast between high jump performance and the 800m run (all other weights are considered approximately zero). Being so, a competitor with very good result on high jump ($\hat{a}_{22} = -0.65$) and very poorly on 800 m ($\hat{a}_{27} = 0.66$) will have a very negative score. Contrarily, competitors with good results on 800 m run and bad on high jump will have very high scores on the PC_2.

4. $\underline{x}_1 = (3.73, 1.86, 15.8, 4.05, 7.27, 45.66, 34.92)^t$

while the maximum values (best scores on all events) are:

Max: 3.73, 1.86, 16,23, 4.05, 7.27, 47.50, 39.23

So, the USA competitor won the 100m hurdles, high jump, 200m, and the long jump.

So

$$\hat{y}_1 = +0.45\left(\frac{3.73 - 2.69}{0.51}\right) + 0.31\left(\frac{1.86 - 1.79}{0.05}\right) + 0.40\left(\frac{15.8 - 13.17}{1.50}\right)$$

$$+0.43\left(\frac{4.05 - 2.02}{0.94}\right) + 0.45\left(\frac{7.27 - 6.21}{0.40}\right)$$

$$+0.24\left(\frac{45.66 - 41.28}{3.47}\right) + 0.30\left(\frac{34.92 - 28.52}{6.15}\right) = \ldots$$

So, her score is expected to be very high, which is confirmed by the plot PC_1 vs Total score.

5. The robust estimates of PC lead to:

(i) higher % of variability explained by the first two PC:

	class	PCD
$\frac{\sum \hat{d}_i^2}{p}$	0.7461	0.7685

(ii) the interpretation of the 1st PC stays the same - overall measure of performance in both cases.

The 2nd PC interpretation changes:

(i) PC_2(PCD) is mainly the javelin results

PC_2(class) is a contrast between high jumps and run 800m

So for the point of view of building a overall index of the quality/performance of the competitors (and their order) there are no major differences.

6. Yes, plot PC1 vs Total Score clear indicates a high positive correlation between these two indexes.