Tutorial - Week 9

Please read the related material and attempt these questions before attending your allocated tutorial. Solutions are released on Friday 4pm.

Question 1

(a) Sample $n = 300$ observations from a $\mathcal{N}(0_p, I_p)$ distribution where $p = 10$. Calculate the multiple correlation coefficient using the sample covariance matrix and the sample correlation matrix. Are the results the same?

(b) Establish analytically the formula on page 8 of Week 8 Lecture Notes:

$$S_{21}^T S_{22}^{-1} S_{21} / S_{11} = R_{21}^T R_{22}^{-1} R_{21}.$$

Question 2

(a) Let R^2 be the sample Multiple Correlation Coefficient between X_1 and $(X_2, \ldots, X_p)^T$ based on an iid sample of size n from $N_p(\mu, \Sigma)$ where $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ is a positive definite matrix. Assuming that the population Multuple Correlation Coefficient, ρ, is zero, page 10 of Week 8 Lecture Notes claims that

$$\frac{n-p}{p-1} \frac{R^2}{1-R^2} \sim F_{p-1, n-p}.$$

Perform a simulation study with $N = 2500$ replicates to check this when $p = 50$, $n = 300$, $\Sigma = I_p$ and $\mu = \mathbf{0}_p$.

(b) Empirically examine the size of the classical test for $\rho = 0$ under the same assumptions as Question (2a).

(c) Empirically examine the power of the classical test with $\alpha = 0.05$ for $\rho = 0$ where the data is generated under the same assumptions as Question (2a) with the exception that the population covariance matrix has the structure

$$\Sigma = \begin{bmatrix} A_\rho & 0 \\ 0 & I_{p-2} \end{bmatrix}$$

for

$$A_\rho = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

and $\rho \in [-1, 1]$. Note that $\rho^2 = \dfrac{\sigma_{(1)}^T \Sigma_{22}^{-1} \sigma_{(1)}}{\sigma_{11}} = A_{12} \times 1 \times A_{21}/1$ where Σ is now decomposed as

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{(1)}^T \\ \sigma_{(1)} & \Sigma_{22} \end{bmatrix}$$

with $1 = \sigma_{11}$ a scalar.

Question 3

(a) In @zheng2014inference, the authors claim that when $n \to \infty$ and $p \to \infty$ with $p/n \to y \in (0, 1)$ and under some technical assumptions that

$$\sqrt{n}(R^2 - y_n) \to N(0, 2y(1-y))$$

where $y_n = p/n$, $R^2 = S_{21}^T S_{22}^{-1} S_{21}/S_{11}$ and assuming that $\rho = 0$. Check this result by a simulation study. You may assume that the technical assumptions are satisfied when $X_1, \ldots, X_n \overset{iid}{\sim} N_p(0, I_p)$ with $(n > p > 0)$.

(b) Empirically examine the size of the test based on Zheng's method for the null hypothesis that $\rho = 0$ vs $\rho \neq 0$.

(c) Empirically examine the power of the test based on Zheng's method for $\rho = 0$ vs $\rho \neq 0$ under the same assumptions as Question 2(c).

Question 4

In this question, we consider a dataset presented in (Anderson 2003) . Available for analysis is the sample correlation matrix

$$\mathbf{R}^X := \begin{bmatrix} 1.00 & 0.4248 & 0.0420 & 0.0215 & 0.0573 \\ 0.4248 & 1.0000 & 0.1487 & 0.2489 & 0.2843 \\ 0.0420 & 0.1487 & 1.0000 & 0.6693 & 0.4662 \\ 0.0215 & 0.2489 & 0.6693 & 1.0000 & 0.6915 \\ 0.0573 & 0.2843 & 0.4662 & 0.6915 & 1.000 \end{bmatrix}$$

The underlying data matrix $\mathbf{X}_{n \times p}$ has $p = 5$ columns corresponding to scores on tests in arithmetic speed, X_1, arithmetic power X_2, memory for words X_3, memory for meaningful symbols X_4, and memory for meaningless symbols X_5.

Of interest is the hypothesis at the 1% significance level that arithmetic speed is independent of the three memory scores.

(a) Write the underlying data matrix as

$$\mathbf{X} = \mathbf{X}_{n \times p} = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 & X_5 \end{bmatrix}$$

and let

$$\mathbf{Y} = \mathbf{Y}_{n \times (p-1)} = \begin{bmatrix} X_1 & X_3 & X_4 & X_5 \end{bmatrix}$$

which is the data matrix $\mathbf{X}$ including only the columns for arithmetic speed and the three memory scores.

Check whether the sample correlation matrix formed from $\mathbf{Y}$, say $\mathbf{R}^Y$, is of the form

$$\mathbf{R}^Y = \begin{bmatrix} \mathbf{R}_{11}^X & \mathbf{R}_{13}^X & \mathbf{R}_{14}^X & \mathbf{R}_{15}^X \\ \mathbf{R}_{31}^X & \mathbf{R}_{33}^X & \mathbf{R}_{34}^X & \mathbf{R}_{35}^X \\ \mathbf{R}_{41}^X & \mathbf{R}_{43}^X & \mathbf{R}_{44}^X & \mathbf{R}_{45}^X \\ \mathbf{R}_{51}^X & \mathbf{R}_{53}^X & \mathbf{R}_{54}^X & \mathbf{R}_{55}^X \end{bmatrix},$$

so that the sample correlation matrix for $\mathbf{Y}$ is just the sample correlation matrix for $\mathbf{X}$ with 1 row and 1 column removed.

(b) Can you use the correlation matrix $\mathbf{R}^X$ to analyze the hypothesis of interest?
(c) Formulate the hypothesis of interest statistically.
(d) Test the hypothesis of interest using the classical method. Is this a reasonable thing to do? Why or why not? What is the conclusion?
(e) Test the hypothesis of interest using Zheng et al.'s method. What is the conclusion?
(f) Which of these two analyses (in Question 3(d) and Question 3(e)) do you prefer? Why?