Homework 3

Due on Monday 8 October 17:00

Question 1 [2 marks]

Let S_1 and S_2 be sample covariance matrices for p-dimensional observations of size n_1 and n_2, respectively. Let $V_n = S_1 S_2^{-1}$ where $n = (n_1, n_2)$ and assume $n_2 > p$. Make an appropriate choice of parameters, sample from V_n and produce a plot showing the histogram of eigenvalues of V_n compared to the LSD F_{y_1,y_2} given by equation (14) in the paper.

Hint: See pages 1210 and 1211 of **[A]** *and Workshop 2, Section 2.2. You can either simulate the data matrix $\mathbb{X}$ using* `rnorm` *(and then construct the sample covariances) or draw the sample covariances directly using* `rWishart`.

Question 2 [3 marks]

In **[A]**, Theorem 3.1, it is proved that under the null hypothesis

$$T_n = \upsilon(f)^{-1/2}[-\log \Lambda_n - p\,F_{y_1,y_2}(f) - m(f)] \;\Rightarrow\; \mathcal{N}(0,1)$$

where $m(f)$, $\upsilon(f)$ and $F_{y_1,y_2}(f)$ are given in the paper in equations (26), (27), and (29), respectively.

Demonstrate numerically that this theorem works by making an appropriate choice of parameters, sampling a large number of T_n, and comparing the histogram of values of T_n against the density of a standard normal.

Hint: See page 1212 and notice that Λ_n is given in terms of $\mathbb{F}$ and the quantity $\mathbb{F}$ is given in terms of the ratio of two Wishart matrices. Therefore, for this task, sample $\mathbb{F}$ by posing

$$\mathbb{F} = \frac{n-q}{q_1} S_1^{-1} S_2, \qquad S_1 \sim W_p(\Sigma, n-q), \quad S_2 \sim W_p(\Sigma, q_1),$$

and using the `rWishart` *function in R. Now from $\mathbb{F}$ it should be straightforward to generate Λ_n. See Workshops 5, 6, and 7 where we have done similar CLT checks.*

Question 3 [10 marks]

The Bartlett statistic, see **[B]** page 413 Eq. (10), is for $g = 2$ given by

$$V_1 = \frac{|\mathbb{A}_1|^{N_1/2} |\mathbb{A}_2|^{N_2/2}}{|\mathbb{A}_1 + \mathbb{A}_2|^{N/2}}$$

c_1 c_2 N_g

where $N_g := n_g - 1$ and $N := N_1 + N_2$. Setting $\mathbb{S}_g = \mathbb{A}_g/N$, multiplying through the numerator and denominator by $|\mathbb{S}_2^{-1}|$ and using the fact that $|AB| = |A||B|$ for matrices A and B, we can instead consider

$$V_1^* = \frac{|\mathbb{S}_1 \mathbb{S}_2^{-1}|^{N_1/2}}{|c_1 \mathbb{S}_1 \mathbb{S}_2^{-1} + c_2|^{N/2}}$$

I_p

where $c_g = N_g/N$. Notice we are in the Fisher regime $\mathbb{S}_1 \mathbb{S}_2^{-1}$. A recent result, Theorem 4.1 in **[C]**, shows that

V=A/B, logV = logA - logB = ...

Theorem 1. *Assume $N_1 \to \infty$, $N_2 \to \infty$, and $p \to \infty$ such that $y_{N_1} = p/N_1 \to y_1 \in (0,1)$ and $y_{N_2} = p/N_2 \to y_2 \in (0,1)$. Then*

$$-\frac{2}{N} \log V_1^* - p\, F_{y_{N_1}, y_{N_2}}(f) \to N(\mu_2, \sigma_2^2),$$

where

$$F_{a,b}(f) := \frac{a+b-ab}{ab} \log\left(\frac{a+b}{a+b-ab}\right) + \frac{a(1-b)}{b(a+b)} \log(1-b) + \frac{b(1-a)}{a(a+b)} \log(1-a),$$

and μ_2 and σ_2 can be determined.

Read the paper to determine the appropriate μ_2 and σ_2 and use these constants to develop and algorithm in R to test the hypothesis $H_1 : \Sigma_1 = \Sigma_2$ for large p. Perform a simulation study to compare its performance (type I error and power) to Box's M-test for varying p.

Note that:

- Instead of calculating `log(det(A))` for some matrix `A`, it might be advisable in R to use the equivalent `determinant(A, logarithm=True)` as it is more numerically accurate for matrices with small determinant.
- If your observations are real-valued, then μ_1 and σ_2^2 are given in the paper by (4.8) and (4.9).

We note that the recent paper **[D]** provides some improvements on **[C]** in that it allows the Fisher matrix $\mathbb{F} = \mathbb{S}_1 \mathbb{S}_2^{-1}$ to be of the form $\mathbb{F} = \mathbb{S}_1 \mathbb{T}^* \mathbb{S}_2^{-1} \mathbb{T}$ where $\mathbb{T}$ is a deterministic matrix.

For some tips on simulation studies, see **[E]** and **[F]**.

see last year project ref.

References

week 9

[A] Bai, Jiang, Yao and Zheng (2013). Testing linear hypotheses in high-dimensional regressions. Statistics Vol 47, Issue 6.

[B] Anderson (2003). An introduction to Multivariate Statistical Analysis. Wiley. Ch10

[C] Bai, Jiang, Yao, Zheng (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. Annals of Statistics Vol 37, No. 6B, 3822–3840.

[D] Zheng, Bai, Yao (2017). CLT for eigenvalue statistics of large-dimensional general Fisher matrices with applications. Bernouilli 23(2), 1130–1178.

[E] http://www4.stat.ncsu.edu/~davidian/st810a/simulation_handout.pdf

[F] https://stats.stackexchange.com/a/40874