Tutorial - Week 6

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

Question 1

The Fisher limiting spectral distribution (LSD), denoted $F_{s,t}$, has the density function

$$f_{s,t}(x) := \frac{1-t}{2\pi x(s+tx)}\sqrt{(b-x)(x-a)}, \quad a \le x \le b,$$

where

$$a := a(s,t) := \frac{(1-h)^2}{(1-t)^2}, \quad b := b(s,t) := \frac{(1+h)^2}{(1-t)^2}, \quad h := h(s,t) := (s+t-st)^{1/2}.$$

Suppose we had two independent p-dimensional vector samples $\mathbb{X} := \{\mathbb{x}_1, \ldots, \mathbb{x}_{n_1}\}$ and $\mathbb{Y} := \{\mathbb{y}_1, \ldots, \mathbb{y}_{n_2}\}$ where $p \le n_2$. We assume that each sample comes from a (possibly different) population distribution with i.i.d. components and finite second moment.

(a) How is the Fisher LSD related to the two vector samples $\mathbb{X}$ and $\mathbb{Y}$? What is the relationship between the two parameters (s,t) of $F_{s,t}$ and the three values (p, n_1, n_2) describing the dimensionality and sizes of $\mathbb{X}$ and $\mathbb{Y}$?

(b) Now that you explained the relationship between $\mathbb{X}$, $\mathbb{Y}$, the values (p, n_1, n_2) and the parameters (s,t) of $F_{s,t}$ in part (a), what would you expect the empirical density of eigenvalues be for the following three choices of triplets (p, n_1, n_2):

$$(50, 100, 100), \qquad (75, 100, 200), \qquad (25, 100, 200).$$

Plot the three densities on the same figure with an appropriate legend.

(c) Perform a simulation study for the same choices of the triplets (p, n_1, n_2) that are given in part (b). Setup your experiment correctly to demonstrate a histogram of eigenvalues and compare them to the appropriately parametrised densities from part (b). For each triplet (p, n_1, n_2), plot the histogram of eigenvalues, overlay the appropriate density, and ensure the plot is appropriately titled. Show the code for your simulation study.

(d) The first moment of the Fisher LSD in terms of its parameters s and t is given by

$$\int_a^b x\, f_{s,t}(x)\, dx = \frac{1}{1-t}.$$

Perform a numerical experiment to confirm this formula. That is, choose 3 values of (s,t) then use your simulation study code (from part c) to generate empirical eigenvalues for those values and calculate their sample mean. Compare the sample means to the formula.

(e) Describe what happens to the first moment when $t \to 0$ and $t \to 1$. Explain the $t \to 1$ case in terms of p, n_1, and n_2.

Question 2

The Bartlett statistic, see **[A]** page 413 Eq. (10)[1], is for $g = 2$ given by

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

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

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

where $c_g = N_g/N$ and I_p is the identity matrix of size $p \times p$. Notice we are in the Fisher regime $\mathbb{S}_1\mathbb{S}_2^{-1}$.

(a) Sample the distribution of V_1^* for $p = 3$, $n_1 = 100$, $n_2 = 100$. Plot the histogram of the distribution when you sample $m = 500$ times.

(b) Show that if the observations from each $\mathbb{X}_1$ and $\mathbb{X}_2$ are transformed by

$$\mathbb{x}_i^* = C\mathbb{x}_i + \mu_i, \qquad i = 1, 2$$

where $\mu_1, \mu_2 \in \mathbb{R}^p$ and C is a $p \times p$ matrix, the distribution of V_1^* remains unchanged. You can do this with a simulation.

(c) Now consider the distribution of $-2\log(V_1^*)$ and compare it to the χ^2_ν distribution where $\nu = \frac{1}{2}p(p+1)$. Is it a good fit?

(d) What happens if you repeat (c) with $p = 30$? Is it still a good fit?

(e) In **[B]** they show that[2] in the high-dimensional setting

Theorem. *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.

Perform this correction to the distribution of V_1^* in the case where $p = 30$. Does this look like a normal distribution?

[1] See `2003-Anderson-Book-Chapter10` in Readings on Wattle.
[2] This is Theorem 4.1 in **[B]**.

(f) Empirically determine the correct μ_2 and σ_2^2 for the corrected V_1^* and use this to plot a normal distribution over the histogram. Does this fit better?

References

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

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