Assessment 5

Due by Thursday 2 November 2023 09:00

[Total Marks: 26 (STAT6017) / 20 (STAT3017)]

Question 1 [12 marks]

Consider two p-dimensional populations with covariance matrices I_p and $(I_p + \Delta)$ where

$$\Delta := \text{diag}(\delta_1, \delta_2, 0, \dots, 0)$$

with $\delta_1, \delta_2 \in \mathbb{R}$. Suppose we had p-dimensional random samples $\mathbb{x}_1, \dots, \mathbb{x}_{m+1} \sim N_p(0, I_p)$ from the first population and p-dimensional random samples $\mathbb{z}_1, \dots, \mathbb{z}_{n+1} \sim N(0, I_p + \Delta)$ from the second. We stack these random samples to obtain the data matrices **X** and **Z** and sample covariance matrices

$$\mathbb{S}_1 := \frac{1}{m}\mathbf{X}\mathbf{X}^T, \qquad \mathbb{S}_2 := \frac{1}{n}\mathbf{Z}\mathbf{Z}^T, \quad \mathbb{S} := \mathbb{S}_2^{-1}\mathbb{S}_1.$$

[2] (a) Assume $n, m, p \to \infty$ such that $y_n := p/n \to y \in (0, 1)$ and $c_m := p/m \to c > 0$. Take $\delta_1 = \delta_2 = 0$, $y = 1/4$, and $c = 3/4$, what is the lower bound a and the upper bound b of the limiting spectral distribution of $\mathbb{S}$? For each, give a formula in terms of c and y. Also give a numerical value.

[2] (b) Suppose that $\delta_1 = -\varepsilon$ and $\delta_2 = +\varepsilon$ for $\varepsilon = 1/10$. Would you expect $\mathbb{S}$ to have eigenvalues smaller than a and larger than b in that case?

[2] (c) In the paper **[A]** (see also **[B]**) it is suggested that the largest eigenvalue λ_1 of $\mathbb{S}$, scaled as $(\lambda_1 - b)/s_p$ where b is from question (a) and $s_p := (\frac{1}{m}(\sqrt{m} + \sqrt{p})(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{p}}))^{1/3}$, behaves like a Tracy-Widom distribution of order 1. Show this using a simulation in the case $n = 400$, $y_n = 1/4$ and $c_m = 3/4$. Plot the histogram and compare it against the Tracy-Widom distribution of order 1.

[2] (d) Considering **[B]**, suppose that $\delta_1 < \ell$ and $\delta_2 > \kappa$ for some choice of ℓ and κ. What would be the critical values of ℓ and κ that would ensure you would have a large fundamental spike and a small fundamental spike? Give a formula for ℓ and κ and also give a numerical value in the case $y = 1/4$ and $c = 3/4$.

[2] (e) Suppose that $\delta_1 = \ell - 1/100$ and $\delta_2 = \kappa + 1/100$ for your critical values of κ and ℓ you found in (d), then give a formula for each of the two locations where you think the spike eigenvalues will cluster around and also a numerical value for each. [1 mark] Also, perform a simulation experiment to illustrate this phenomena. That is, sample data and plot a histogram of eigenvalues of $\mathbb{S}$, compare it to the theoretical density expected if $\delta_1 = \delta_2 = 0$, and plot the location where you expect spike eigenvalues to cluster around. Take $n = 400$, $y_n = 1/4$, and $c_n = 3/4$. [1 mark]

[2] (f) Consider the signal detection problem where we are trying to *determine the number of signals* in observations of the form

$$x_i = Us_i + \varepsilon_i, \qquad i = 1, \ldots, m, \tag{SD}$$

where the x_i's are p-dimensional observations, s_i is a $k \times 1$ low dimensional signal ($k \ll p$) with covariance I_k, U is a $p \times k$ mixing matrix, and (ε_i) is an i.i.d. noise with covariance matrix Σ_2. None of the quantities on the right hand side of (SD) are observed. In Section 7.2 of **[B]**, they propose to estimate the number of signals k by

$$\hat{k} := \max\{i : \lambda_i \geq \beta + \log(p/p^{2/3})\},$$

where (λ_i) are the eigenvalues of $\mathbb{S}$. Reproduce Case 1 in Table 1 of **[B]** for the Gaussian case for values $p = 25, 75, 125, 175, 225, 275$. Fix $y = 1/10$ and $c = 9/10$, further parameters and setup can be found at the bottom of p.436 and on p.437.

Question 2 [8 marks]

In this question, we shall consider high-dimensional sample covariance matrices of data that is sampled from an elliptical distribution. We say that a random vector $\mathbb{x}$ with zero mean follows an *elliptical distribution* if (and only if) it has the stochastic representation

$$\mathbb{x} = \xi A \mathbb{u}, \tag{$\bigstar$}$$

where the matrix $A \in \mathbb{R}^{p \times p}$ is nonrandom and $\text{rank}(A) = p$, $\xi \geq 0$ is a random variable representing the radius of $\mathbb{x}$, and $u \in \mathbb{R}^p$ is the random direction, which is independent of ξ and uniformly distributed on the unit sphere S^{p-1} in $\mathbb{R}^p$, denoted by $\mathbb{u} \sim \textbf{Unif}(S^{p-1})$. The class of elliptical distributions is a natural generalization of the multivariate normal distribution, and contains many widely used distributions as special cases including the multivariate t-distribution, the symmetric multivariate Laplace distribution and the symmetric multivariate stable distribution.

[2] (a) Write a function `runifsphere(n,p)` that samples n observations from the distribution $\textbf{Unif}(S^{p-1})$ using the fact that if $\mathbb{z} \sim N_p(0, I_p)$ then $\mathbb{z}/\|\mathbb{z}\| \sim \textbf{Unif}(S^{p-1})$. Check your results by: (1) set $p = 25, n = 50$ and show that the (Euclidean) norm of each observation is equal to 1, (2) generate a scatter plot in the case $p = 2, n = 500$ to show that the samples lie on a circle. [1 mark]

Show that you can simulate a multivariate t-distribution $t_\nu(0, I_p)$ by setting $\xi \sim \sqrt{\nu/C}$ in ($\bigstar$) with $A = I_p$ and $C \sim \chi^2_\nu$. Do this by sampling observations $\mathbb{x}_1, \ldots, \mathbb{x}_n$ and comparing the two marginal histograms of the observations against the density of the univariate t_ν distribution. Take $p = 2$, $n = 1000$, $\nu = 2$. [1 mark]

[2] (b) Suppose that $\mathbb{x}_1, \mathbb{x}_2, \ldots, \mathbb{x}_n$ are p-dimensional observations sampled from an elliptic distribution ($\bigstar$). We stack these observations into the data matrix $\mathbb{X}$ and calculate the sample covariance matrix $\mathbb{S}_n := \mathbb{X}\mathbb{X}^T/n$. Theorem 2.2 of the recent paper **[C]** is a central limit theorem for linear spectral statistics (LSS) of $\mathbb{S}_n$. For example, Eq. (2.10) in **[C]** provides the case of the joint distribution of the LSS $\phi_1(x) = x$

and $\phi_2(x) = x^2$. Following the notation used there (for all the following terms in this question). Perform a simulation experiment to examine the fluctuations of $\hat{\beta}_{n1}$ and $\hat{\beta}_{n2}$. In the experiment, take $H_p = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ and choose the distribution of $\xi \sim k_1 \mathbf{Gamma}(p, 1)$ with $k_1 = 1/\sqrt{p+1}$. Set the dimensions to be $p = 200$ and $n = 400$. Choose the number of simulations based on the computational power of your machine. Similar to Figure 1 in **[C]**, use a QQ-plot to show normality.

[2] (c) In the recent paper **[E]**, it is shown that if $\mathbb{x}_1, \mathbb{x}_2, \ldots, \mathbb{x}_n$ are p-dimensional observations sampled from an elliptic distribution (★) then the largest eigenvalue λ_1 of the sample covariance $\mathbb{S}_n$ (appropriately scaled) converges to the Tracy-Widom distribution as long as a certain condition on the tail of the distribution holds (Condition 2.7 in the paper). Perform a simulation to show that this holds true in the case of a double exponential distribution but not in the case of a multivariate student-t distribution. That is, simulate the largest eigenvalue of the sample covariance matrix and compare it to the Tracy-Widom distribution in each case. In the first case it should match (double exponential) and in the second case it shouldn't (multivariate student-t).

[2] (d) A nice property of elliptic distributions (★) is that the mixture coefficient ξ can feature heteroskedasticity and the overall distribution of $\mathbb{x}$ can exhibit heavy tails. Both are properties that are widely observed in financial and economic data, for example. In the recent paper **[F]**, they proposed a more generalised setting whereby the observations

$$\mathbb{x}_i = \xi_i A \mathbb{u}_i, \qquad i = 1, \ldots, n.$$

may exhibit the situation that

- ξ_i's can depend on each other and on $\{\mathbb{u}_i : i = 1, \ldots, n\}$ in an arbitrary way, and
- ξ_i's do *not* need to be stationary.

The trick to dealing with these kind of observations is to *self-normalise* them. That is, we consider the new observations $\tilde{\mathbb{x}}_1, \ldots, \tilde{x}_n$ where

$$\tilde{\mathbb{x}}_i := \frac{\mathbb{x}_i}{\|\mathbb{x}_i\|}.$$

The paper introduces two tests (LR-SN and JHN-SN) to consider the *sphericity test*

$$H_0 : \Sigma \propto I_p \qquad \text{v.s.} \qquad \Sigma \not\propto I_p$$

where $\propto$ means "proportional to". Reproduce the simulation experiment shown in Table 5 of **[F]** for the case $p/n = 0.5$ and only for LR-SN and JHN-SN for $p = 100, 200, 500$. Do this in the case of 1,000 replications.

Question 3 [6 marks]

We will consider some additional tasks relating to the above questions. *These are for STAT6017 students only.*

[2] (a) Unfortunately, the results of **[C]** do not cover all elliptic distributions due to a moment condition on the distribution, see Table 1 in **[C]**. The results in **[D]** extend their results to more general elliptic distributions such as multivariate Gaussian mixtures[1]. A p-dimensional vector $\mathbb{x} \in \mathbb{R}^p$ is a multivariate Gaussian mixture with k subpopulations if its density function has the form

$$f(\mathbb{x}) = \sum_{j=1}^{k} p_j \phi(\mathbb{x}; \mu_j, \Sigma_j)$$

where (p_j) are the k mixing weights and $\phi(\cdot; \mu_j, \Sigma_j)$ denote the density function of the jth subpopulation with mean vector μ_j and covariance Σ_j. In the case where $\mu_1 = \mu_2 = \cdots \mu_k = 0 \in \mathbb{R}^p$ and $\Sigma_j = v_j \Sigma$ for some $v_j > 0$ with $j = 1, \ldots, k$. Write an R function to sample from such a distribution using the representation from Eq. (11) in **[D]**.

[2] (b) Using your code from (a), perform a simulation experiment to simulate fluctations of $\hat{\beta}_2 := \int x^2 dF^{\mathbb{S}_n}(x)$ under a Gaussian scale mixture model where the variable ξ has a discrete distribution with two mass points $\mathbb{P}(\xi = 1.8\sqrt{p}) = 0.8$ and $\mathbb{P}(\xi = 1.5\sqrt{p}) = 0.2$. Consider the cases: (i) $p = 100$, $n = 150$, (ii) $p = 600, n = 900$. In each case, plot a histogram of the distribution of $\hat{\beta}_2$ against the theoretical limiting density and also a QQ-plot similar to Figure 1 in **[D]**. Note: this is the experiment just above Section 3 in **[D]**.

[2] (c) In addition to Question 2 (d), also reproduce the simulation experiment shown in Table 7 of **[F]** for the case $p/n = 0.5$ and only for LR-SN and JHN-SN for $p = 100, 200, 500$. Do this in the case of 1,000 replications.

References

[A] Han, Pan, Zhang (2016). The Tracy-Widom law for the largest eigenvalue of F-type matrices. Annals of Statistics, Vol. 44.

[B] Wang, Yao (2017). Extreme eigenvalues of large-dimensional spiked Fisher matrices with application. Annals of Statistics, Vol 45, No. 1.

[C] Hu, Li, Liu, Zhou (2019). *High-dimensional covariance matrices in elliptical distributions with application to spherical test*. Annals of Statistics.

[D] Zhang, Hu, Li (2022). *CLT for linear spectral statistics of high-dimensional sample covariance matrices in elliptical distributions*. Journal of Multivariate Analysis.

[E] Jun, Jiahui, Long, Wang (2022). *Tracy-Widom limit for the largest eigenvalue of high-dimensional covariance matrices in elliptical distributions*. Bernouilli.

[F] Yang, Zheng, Chen (2021). *Testing high-dimensional covariance matrices under the elliptical distribution and beyond*. Journal of Econometrics.

Note: I have placed these references in the 'Readings' folder on Wattle.

[1] Recall I mentioned in Lecture 1 that one difficulty in big datasets is the presence of multiple subpopulations.