Tutorial - Week 5

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

Question 1

We continue the analysis started in last week's tutorial where we were looking at the properties of the Hotelling T^2 statistic and testing means. First we are interested to understand what happens as p becomes large and how it compares to some other alternative methods.

(a) One of the problems when using the T^2 test statistic is that, as p becomes larger compared to the sample size n, the sample covariance matrix $\mathbb{S}$ becomes increasingly singular. A singular matrix is one that does not have an inverse and this causes problems as the T^2 test statistic contains the expression $\mathbb{S}^{-1}$. As a matrix is singular if and only if its determinant is zero, we can study $\det(\mathbb{S})$ as $p \to n$ to understand how quickly $\mathbb{S}$ becomes singular as p increases.
Perform a simulation experiment to show the behaviour of $\det(\mathbb{S})$ as $p \to n = 100$. That is, write a function that, given values of p and $n = 100$, will sample n observations from a multivariate normal distribution $N_p(0, I_p)$, calculate the sample covariance $\mathbb{S}$ of the data, then calculate the determinant of the sample covariance matrix $\det(\mathbb{S})$. The function should perform this simulation `nsims` number of times for a given p, calculate the mean and standard error of the simulation.

(b) Illustrate the behaviour of $\det(\mathbb{S})$ for $2 < p < n = 100$ by using an error bar plot. You can use the function `errbar` in the package `Hmisc` to do this.

(c) Now that we have identified there could be a problem with $\mathbb{S}^{-1}$ as p becomes large, let's consider what happens to the Hotelling T^2 test. First, similar to last week's tutorial, write a function that calculates the power of a Hotelling T^2 test. Take the case $n_1 = n_2 = 50$ with $\mu_2 - \mu_1 = (\delta, \dots, \delta)'$ with varying $0 \le \delta \le 0.5$. However, this time assume that the (population) covariance has an "AR(1) structure" given by

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho^p \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho^p & \rho^{p-1} & \cdots & 1 \end{pmatrix}, \tag{AR1}$$

for $\rho = 0.5$. This can be generated succinctly in R using the code

```
rho = 0.5
Sigma = rho^abs(outer(1:p, 1:p, "-"))
```

Plot the power curve for $0 \le \delta \le 0.5$, for the cases $p \in \{5, 25, 50, 75, 95\}$. What do you observe?

(d) Consider the one-sample version of Hotelling's T^2 test that uses the χ^2 distribution for testing

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad \mu \neq \mu_0,$$

where the test statistic is given by $T^2 = n(\bar{\mathbb{x}} - \mu_0)'\mathbb{S}^{-1}(\bar{\mathbb{x}} - \mu_0)$ and the distribution of $T^2 \sim \chi^2_p$ where n is the number of p-dimensional observations $\mathbb{x}_1, \dots, \mathbb{x}_n$ with sample mean $\bar{\mathbb{x}}$; see **[A]**[1]. Take the case $\mu_0 = (0, 0, \dots, 0)'$. Write a function to perform this test for a given data matrix $\mathbb{X}$.

(e) Use you one-sample version of Hotelling's T^2 test and perform a simulation study to understand whether the *size* of the test achieves the advertised α. Consider this for p varying between 2 and 50 and $\alpha = 0.05$. What does this show?

Question 2

Sometimes we might want to test whether two samples have the same amount of variation in the data. When your data is univariate (i.e., $p = 1$), this leads to a hypothesis testing problem concerning variances. That is, you have a random sample $Y_1, Y_2, \dots, Y_n$ from a (univariate) normal distribution with unknown mean μ and unknown variance σ^2. Then, you are interesting in testing $H_0 : \sigma^2 = \sigma_0^2$ for some fixed value σ_0^2 versus the alternative hypothesis. This was generalised to multivariate data and the two-sample case by Box **[B]**, where the hypotheses become

$$H_0 : \Sigma_1 = \Sigma_2 \quad \text{vs.} \quad H_1 : \Sigma_1 \neq \Sigma_2,$$

for two p-dimensional samples of size n_1 and n_2, respectively. To perform a test that seperates these two hypotheses, we calculate a statistic called *Box's M-test* which is given by

$$M := \frac{|\mathbb{S}_1|^{N_1/2}|\mathbb{S}_2|^{N_2/2}}{|\mathbb{S}_{\text{pl}}|^{N/2}},$$

where for $i \in \{1, 2\}$, $N_i = n_i - 1$, $\mathbb{S}_i$ is the sample covariance matrix of the ith sample, and $\mathbb{S}_{\text{pl}}$ is the pooled sample covariance matrix given by

$$\mathbb{S}_{\text{pl}} = \frac{N_1\mathbb{S}_1 + N_2\mathbb{S}_2}{N},$$

where $N := N_1 + N_2 = (n_1 + n_2 - 2)$. Box gave a χ^2-approximation for the distribution of M which first requires calculating

$$c_1 := \left[\frac{1}{N_1} + \frac{1}{N_2} - \frac{1}{N}\right]\left[\frac{2p^2 + 2p - 1}{6(p + 1)}\right],$$

then $u = -2(1 - c_1)\log(M)$ is approximately $\chi^2_{\frac{1}{2}p(p+1)}$-distributed. The null hypothesis H_0 is then rejected if $u > \chi^2_\alpha$ for your desired size α (e.g., $\alpha = 0.05$).

(a) Take $\Sigma_1 = \Sigma_2 = I_p$, $n_1 = 250$, $n_2 = 250$, $p = 3$, nsims=5000 and simulate the distribution of $-2(1 - c_1)\log(M)$ under the assumption that sample 1 is generated from $N_p(0, \Sigma_1)$ and sample 2 from $N_p(0, \Sigma_2)$. As the logarithm of a determinant may be numerically unstable, you may use the following R function to compute $\log(\det(x))$:

[1] See the "Motivation" section at **[A]**.

```
logdet = function(x) as.numeric(determinant(x, logarithm=TRUE)$modulus)
```

Compare the empirical distribution obtained from your simulation to the χ^2 distribution.

(b) Now consider what happens to the distribution if you redo the simulation with Σ_2 having the form of (AR1) with $\rho = 0.1$. What does this mean in relation to the hypothesis testing problem?

(c) Consider the `tibetskull` dataset from last week and perform a hypothesis testing problem to determine if the two samples have different (population) covariances.

References

[A] https://en.wikipedia.org/wiki/Hotelling%27s_T-squared_distribution

[B] Box (1949). A General Distribution Theory for a Class of Likelihood Criteria. Biometrica 36 (3-4), 317 – 346.