6017_A3

2023-09-11

Question 1

Suppose we had two independent p-dimensional vector samples $X := x_1, ..., x_{n_1}$ and $Y := y_1, ..., y_{n_2}$ where $p \leq n_2$. 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 X and 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 X and Y?

Denote the S_1 as the sample covariance matrix for the X and the S_2 as the sample covariance matrix for the Y. The random Fisher matrices take the form of $V_n = S_1 S_2^{-1}, n = (n_1, n_2)$. When the $p/n_1 \to y_1$ and $p/n_2 \to y_2$, the empirical spectral distribution (ESD) of $F_n^{V_n}$ of V_n converges to the a limiting spectral distribution (LSD) F_{y_1,y_2}, that is, $s = y_1, t = y_2$.

(b)Now that you explained the relationship between X, 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 $(p, n1, n2)$: $(50, 100, 100), (75, 100, 200), (25, 100, 200)$. Plot the three densities on the same figure with an appropriate legend.

Let's first calculate the s, t *and* h for each triple. The first one has $s = \frac{1}{2}, t = \frac{1}{2}, h = \sqrt{3}/2$, while the second has $s = \frac{3}{4}, t = \frac{3}{8}, h = 3\sqrt{3/2}/4$. The third has $s = \frac{1}{4}, t = \frac{1}{8}, h = \sqrt{11/2}/4$. We can compute each support for each triple. It is expected that the first triple has the longest tail, as $t \to 1, a \to \frac{1}{4}(1-s)^2, b \to \infty$. Also, it is clear that the second triple is more concentrated around 0 as $s \uparrow, t \to 1, a \to \frac{1}{4}(1-s)^2$, the left support is getting closer to 0. The third triple illustrates the case where the LSD of Fisher matrix is getting closer to the MP distribution with fixed s and $t \to 0$.

Firstly, we can write a function about LSD of the Fisher matrix.

```
pdf_Fisher <- function(x, p = p, n1 = n1, n2 = n2){
  s = p/n1
  t = p/n2
  h = (s + t - s*t)^{1/2}
  a = (1 - h)^2 / (1 - t)^2
  b = (1 + h)^2 / (1 - t)^2
  ifelse(x <= a | x >=b, 0, suppressWarnings((1-t)/ (2 * pi * x * (s + t*x)) *sqrt((b - x) * (x -a)), "
}
```

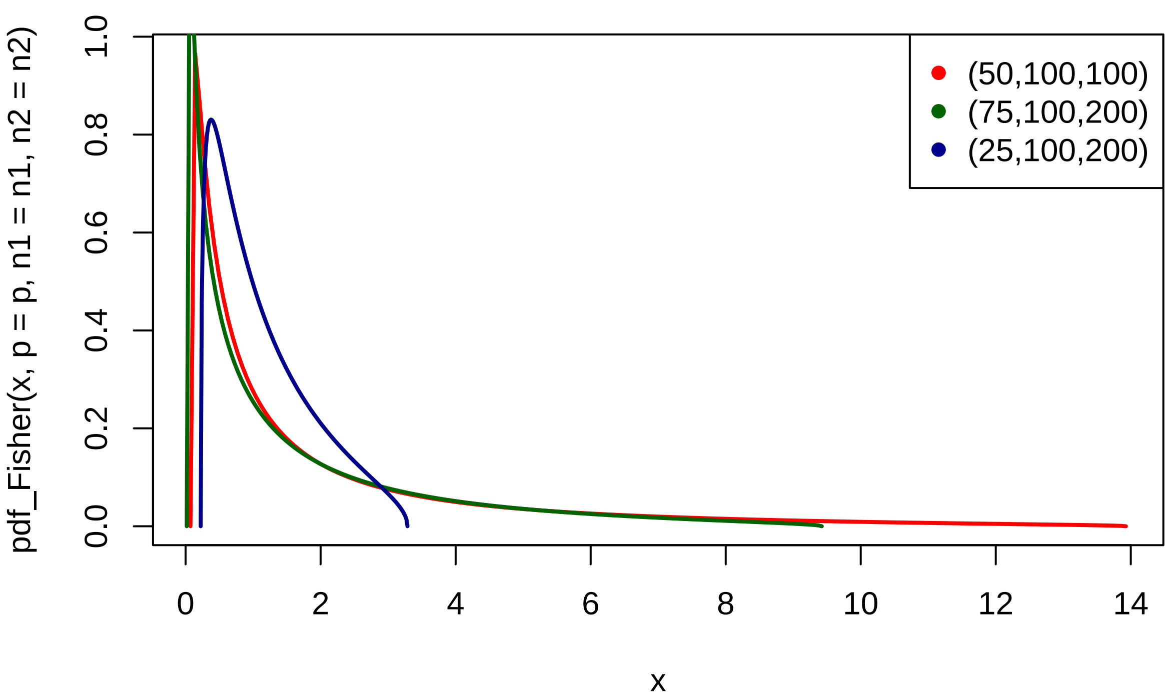
Then, we can write a function to create a equally spaced points inside the support of of the Fisher LSD.

```
fisher_support <- function(n_points = 200, p = p, n1 = n1, n2 = n2){
  s = p/n1
  t = p/n2
  h = (s + t - s*t)^{1/2}
```

```r
  a = (1 - h)^2 / (1 - t)^2
  b = (1 + h)^2 / (1 - t)^2
  seq(a, b, length.out = n_points)
}
```

Let's plot the three cases together.

```r
p = 50
n1 = 100
n2 = 100
x = fisher_support(p = p, n1 = n1, n2 = n2)
plot(x, pdf_Fisher(x, p = p, n1 = n1, n2 = n2), type='l', lwd=2, col="red")
p = 75
n1 = 100
n2 = 200
x = fisher_support(p = p, n1 = n1, n2 = n2)
lines(x, pdf_Fisher(x, p = p, n1 = n1, n2 = n2), type='l', lwd=2, col="darkgreen", ylab="")
p = 25
n1 = 100
n2 = 200
x = fisher_support(p = p, n1 = n1, n2 = n2)
lines(x, pdf_Fisher(x, p = p, n1 = n1, n2 = n2), type='l', lwd=2, col="darkblue", ylab="")
legend("topright", c("(50,100,100)", "(75,100,200)","(25,100,200)"),
       col = c("red", "darkgreen","darkblue"),
       pch = c(16,16,16))
```



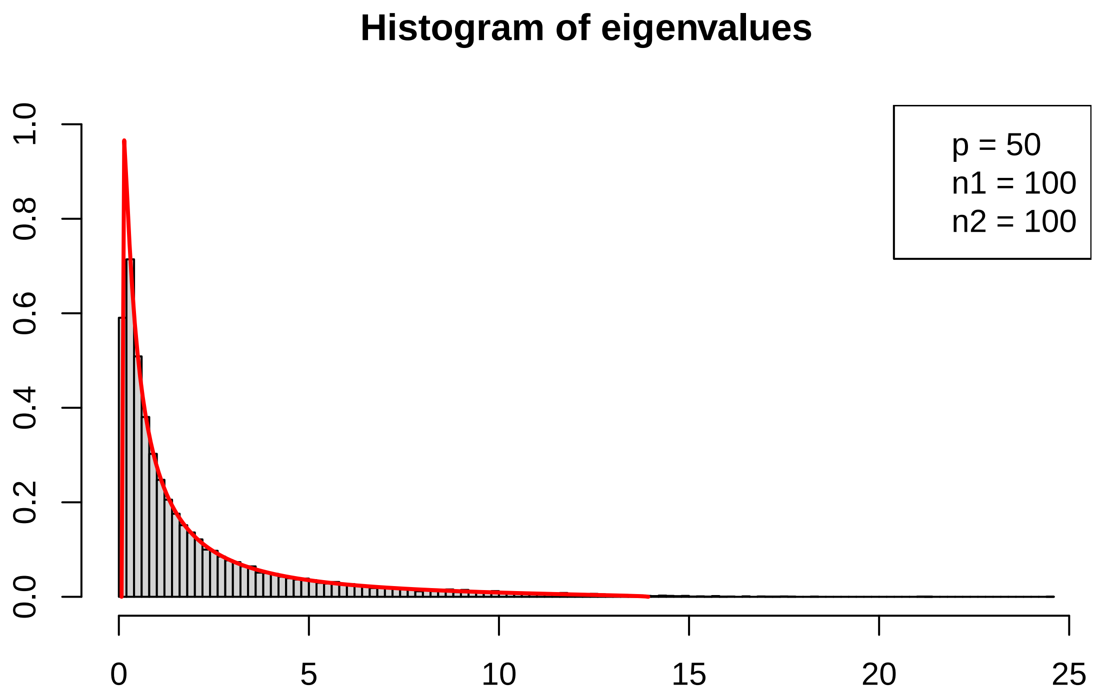
(c) Perform a simulation study for the same choices of the triplets $(p, n1, n2)$ that are given in part (b). Setup your experiment correctly to demonstrate a histogram of eigenvalues and compare them to the appropriately parametrised densities from part (b). For each triplet $(p, n1, n2)$, plot the histogram of eigenvalues, overlay the appropriate density, and ensure the plot is appropriately titled. Show the code for your simulation study.

Let's construct a function to simulate the eigenvalues for a number of Fisher matrices. Let's adhere to the traditional assumption that the entries in the F matrix have mean 0 and variance 1.

```
simulate_Fisher <- function(p, n1, n2){
  mu1 = rep(0,p)
  mu2 = rep(0,p)
  Sigma1 = diag(p)
  Sigma2 = diag(p)
  X = rmvn(n1, mu1, Sigma1)
  Y = rmvn(n2, mu2, Sigma1)
  S1 = cov(X)
  S2 = cov(Y)
  Fisher_matrix = S1 %*% solve(S2)
  eigen(Fisher_matrix, only.values = TRUE)$values
}
```

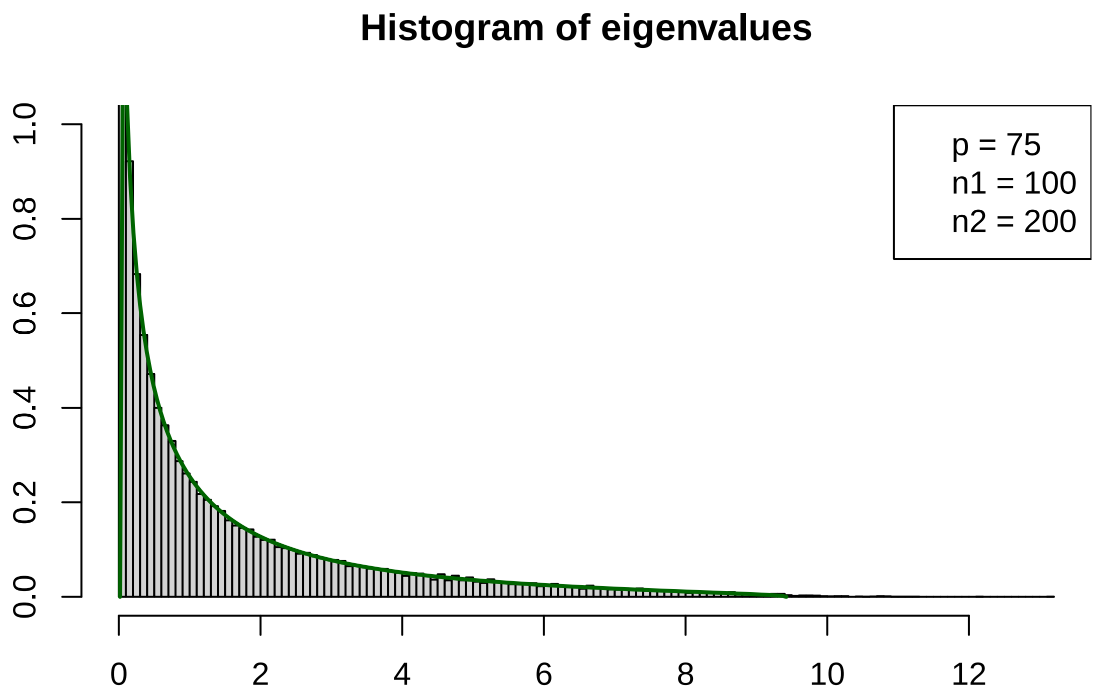
Now, we are ready to plot the each triplet. Let's first plot the case of $(50, 100, 100)$. To make the result smooth, let's make the a number of Fisher matrix large, say 500.

```
eigenvalues = c(replicate(500, simulate_Fisher(50, 100, 100)))
hist(eigenvalues, breaks = "FD",
     freq = FALSE, ylim = c(0,1),
     main = "Histogram of eigenvalues",
     xlab = "",
     ylab = "")
x = fisher_support(p = 50, n1 = 100, n2 = 100)
lines(x, pdf_Fisher(x, p = 50, n1 = 100, n2 = 100),
      type='l', lwd=2, col="red", ylab="")
legend("topright", c("p = 50", "n1 = 100", "n2 = 100"))
```



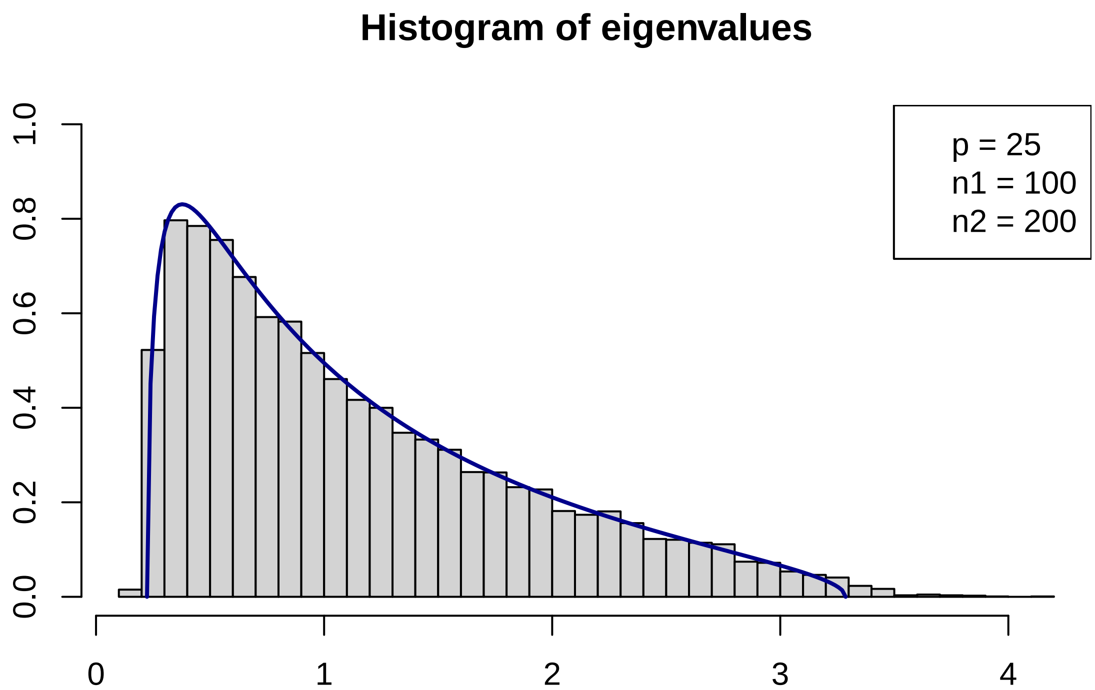
Let's first plot the case of $(75, 100, 200)$ with the number of Fish matrix to be 500.

```
eigenvalues = c(replicate(500, simulate_Fisher(75, 100, 200)))
hist(eigenvalues, breaks = "FD",
     probability = TRUE,
     ylim = c(0,1),
     main = "Histogram of eigenvalues",
     xlab = "",
     ylab = "")
x = fisher_support(p = 75, n1 = 100, n2 = 200)
lines(x, pdf_Fisher(x, p = 75, n1 = 100, n2 = 200),
      type='l', lwd=2, col="darkgreen", ylab="")
legend("topright", c("p = 75", "n1 = 100", "n2 = 200"))
```



Let's first plot the case of $(25, 100, 200)$ with the number of Fish matrix to be large, say 500.

```
eigenvalues = c(replicate(500, simulate_Fisher(25, 100, 200)))
hist(eigenvalues, breaks = "FD",
     probability = TRUE,
     ylim = c(0,1),
     main = "Histogram of eigenvalues",
     xlab = "",
     ylab = "")
x = fisher_support(p = 25, n1 = 100, n2 = 200)
lines(x, pdf_Fisher(x, p = 25, n1 = 100, n2 = 200),
      type='l', lwd=2, col="darkblue", ylab="")
legend("topright", c("p = 25", "n1 = 100", "n2 = 200"))
```

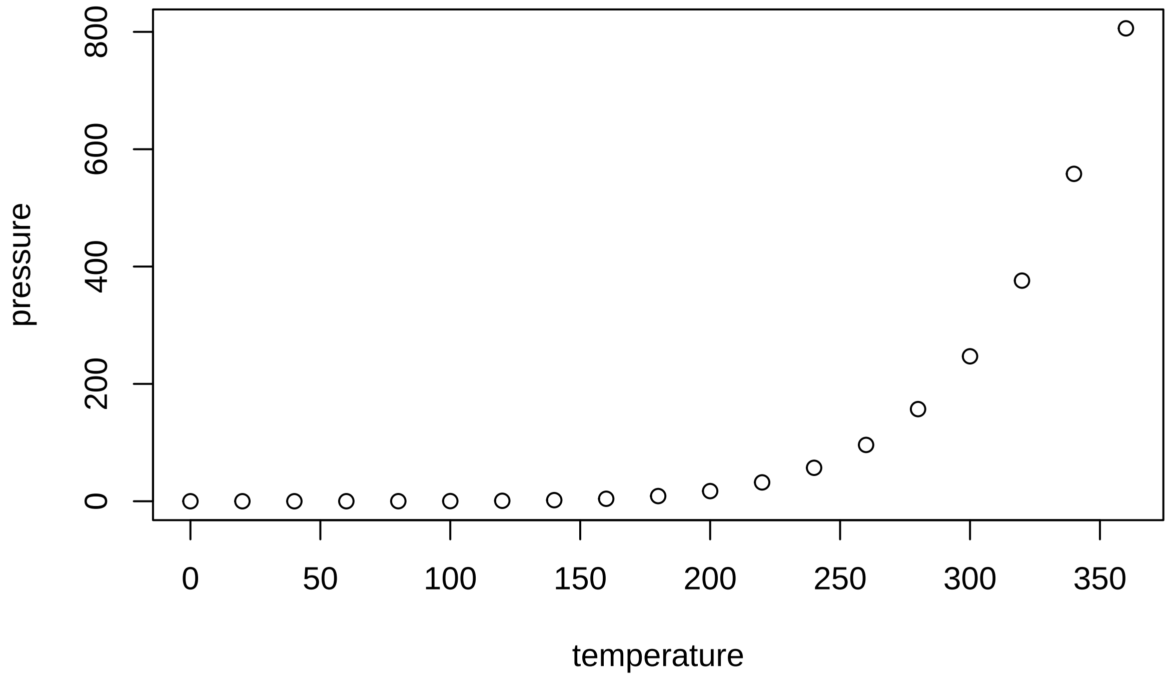


(d) Derive a formula for the first moment of the Fisher LSD in terms of its parameters s and t. You can assume that $h > t$ as this always holds by the definition of h.

Question 2

Show that the second moment

$$1 = 2$$



Question 3

Show that the variance equals $h2/(1-t)3$.

Reference