Workshop 5

- Linear spectral statistics
- CLT for Linear Spectral Statistics
  - Sampling the test statistic
  - Histogram of the test statistic distribution
  - Moment of the MP distribution
  - CLT
- Task for this week

Linear spectral statistics

We can generate one realisation of the sample covariance matrix $\mathbf{S}_n$.

```
p <- 200
n <- 800
X <- matrix(rnorm(p*n), p, n)
Sn <- X %*% t(X) / n
```

Linear spectral statistics are function of the eigenvalues of the sample covariance matrix $\mathbf{S}_n$. They are easy to obtain by using the `eigen` function in R. For example, we can calculate the *generalised variance* statistics.

```
L<-eigen(Sn)$values
GV <- sum(log(L))/p
```

CLT for Linear Spectral Statistics

Sampling the test statistic

To look at the CLT, we need to simulate a large number of sample covariance matrices and then calculate the sum of their eigenvalues divided by p. To make things easier, we are going to study the test statistic

$$\mathbf{T}_n = \frac{1}{p}\sum_{k=1}^{p}\lambda_k.$$

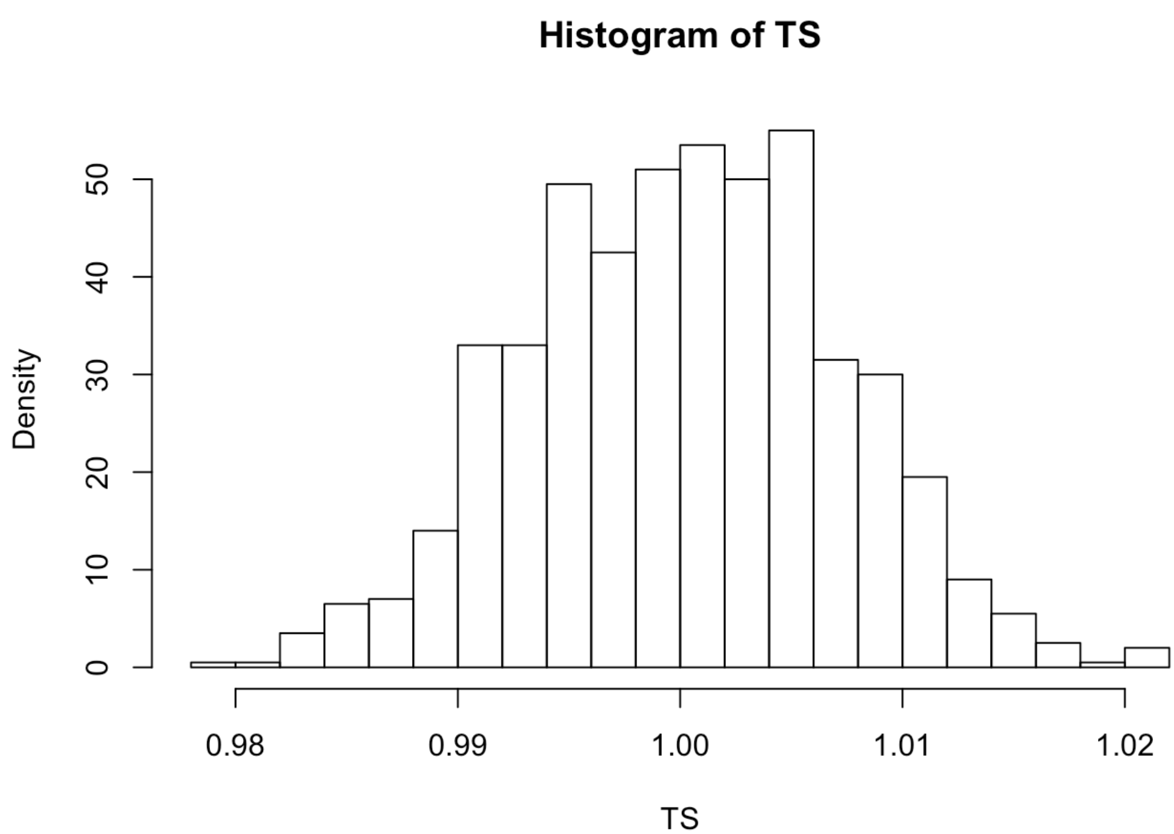
Notice here that the test function $\varphi(x) = x$.

```
N <- 1000
TS <- numeric(N)
for (i in 1:N) {
  p <- 100
  n <- 400
  X <- matrix(rnorm(p*n), p, n)
  Sn <- X %*% t(X) / n
  L<-eigen(Sn)$values
  TS[i] <- sum(L)/p
}
```

Histogram of the test statistic distribution

We can now plot a histogram of the fluctuations on this test statistic.

```
hist(TS, breaks="FD", freq=FALSE)
```



Moment of the MP distribution

We want to check the CLT by look at the deviation from $F_y(\varphi)$. We know from Homework 2 that

$$F_{s,t}(\phi) = \int x \, dF_{s,t}(x) = \frac{1}{1-t}$$

where $F_{s,t}$ is the LSD for the random Fisher matrix and remember that $F_{y,0}$ is the Marchenko-Pastur distribution so that means

$$F_y(\phi) = \int x \, dF_{y,0}(x) = 1.$$

CLT

The theorem that we looked at this week told us that the quantity

$$p\left(F^{\mathbf{S}_n}(\varphi) - F_{y_n}(\varphi)\right)$$

is Normally distributed with a mean and variance that we can calculate explicitly.

Since

$$F^{\mathbf{S}_n} = \frac{1}{p}\sum_{k=1}^{p} \delta_{\lambda_k}$$

we have, in the case $\varphi(x) = x$ that

$$F^{\mathbf{S}_n}(\varphi) = \frac{1}{p}\sum_{k=1}^{p} \varphi(\lambda_k) = \frac{1}{p}\sum_{k=1}^{p} \lambda_k$$

And we see from the histogram above that $F^{\mathbf{S}_n}(\varphi)$ should fluctuate around 1. In other words, the mean difference between $F^{\mathbf{S}_n}(\varphi)$ and $F_{y_n}(\varphi)$ should be zero.

The CLT we calculated this week also gave us the variance

$$(\beta + \kappa)y.$$

Since the entries of the data matrix are Gaussian $\beta = 0$ and also real numbers so $\kappa = 2$. This means that the variance of the difference between $F^{\mathbf{S}_n}(\varphi)$ and $F_{y_n}(\varphi)$ should be equal to

$$2y_n = 2\frac{p}{n}.$$

```
var(p*TS-p)
```

```
## [1] 0.4854432
```

```
2 * p / n
```

```
## [1] 0.5
```

Task for this week

Redo the above calculations in the case $\varphi(x) = x^2$.