Workshop 3

Empirical Spectral Distributions

Visualising the ESD

Sample a covariance based on $\mathbb{S}_n = \frac{1}{n}\mathbb{X}\mathbb{X}^*$ where $\mathbb{X}$ is a $p \times n$ matrix with Gaussian entries with mean zero and variance 1. This gives a matrix $\mathbb{S}_n$ of size $p \times p$.

```
p <- 50
n <- 500
X <- matrix(rnorm(p*n), p, n)
Sn <- X %*% t(X) / n
dim(Sn)
```

```
## [1] 50 50
```

Calculate the ratio $y = p/n$

```
p/n
```

```
## [1] 0.1
```

Calculate the eigenvalues.

```
e<-eigen(Sn)
L<-e$values
```

We have p eigenvalues.

```
length(L)
```

```
## [1] 50
```

Print out the eigenvalues, notice that are all real numbers.

```
L
```

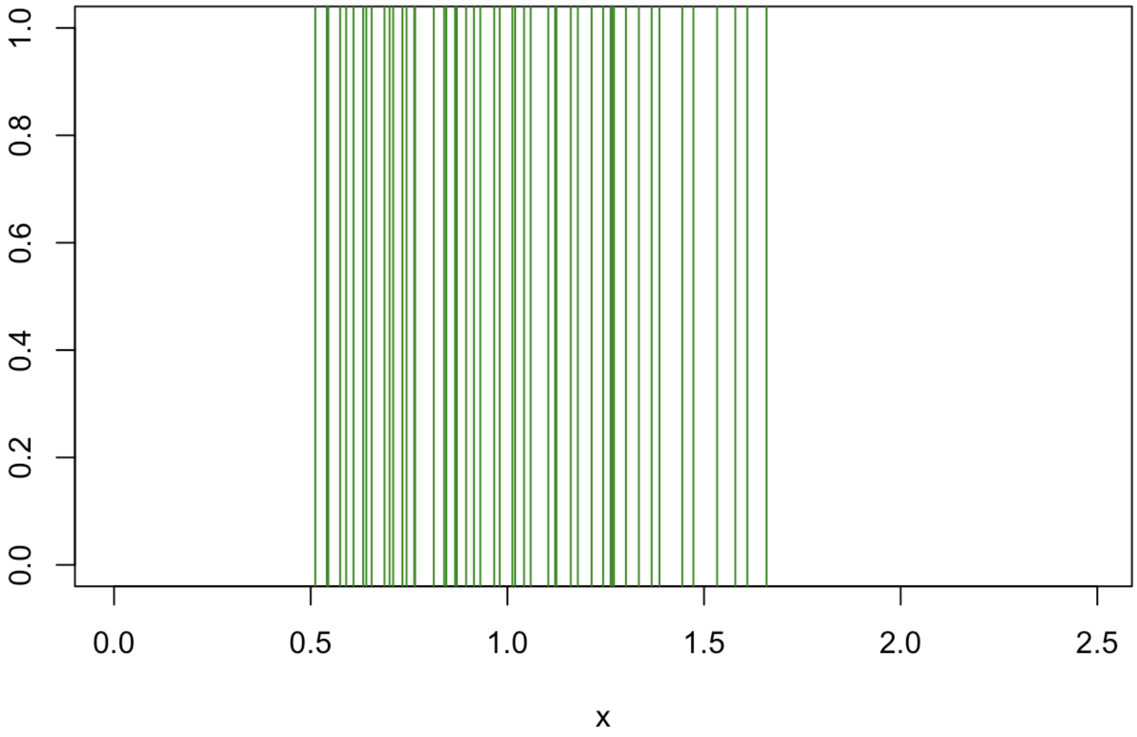
```
##  [1] 1.6590227 1.6099626 1.5796172 1.5334019 1.4730266 1.4446727 1.3866674
##  [8] 1.3669401 1.3343549 1.3014481 1.2711151 1.2670333 1.2625589 1.2436115
## [15] 1.2143434 1.1791325 1.1613370 1.1247311 1.1214396 1.1040174 1.0593323
## [22] 1.0425655 1.0197482 1.0128362 0.9805162 0.9666159 0.9311748 0.9149896
## [29] 0.8950640 0.8714392 0.8671514 0.8446229 0.8392360 0.8127923 0.7654460
## [36] 0.7632831 0.7436032 0.7331083 0.7097188 0.7005091 0.6876130 0.6548005
## [43] 0.6413234 0.6333489 0.6087335 0.5901800 0.5746486 0.5447642 0.5408793
## [50] 0.5114876
```

It’s pretty hard to draw the empirical spectral distribution

$$F^{\mathbb{S}_n}(x) = \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_k}(x).$$

One way is to think of this function as a plot where you have a horizontal line at every eigenvalue (of height $1/p$). Draw a horizontal line at every eigenvalue. Note that setting the height of this line to $1/p$ is tricky and can’t be done with the `abline` function.

```
plot(c(0,1.5*max(L)), c(0,1), type='n', xlab='x', ylab='')
abline(v=L, col='green4')
```



Linear spectral statistics

Remember in the lecture we didn't deal with the ESD $F^{\mathbb{S}_n}$ directly, but instead we considered

$$F^{\mathbb{S}_n}(\varphi) := \int \varphi(x) F^{\mathbb{S}_n}(dx),$$

for some choice of φ.

I gave two examples: $\varphi(x) = \log(x)$ for the generalised variance and $\varphi_z(x) = \frac{1}{x-z}$ for the Stieltjes transform (which depends on another variable z).

Generalised variance

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
```

Here `GV` is a (random) number, we need to perform lots of simulation to understand the distribution of this test statistic.

```
GV
```

```
## [1] -0.1389934
```

In an easy case

Consider the test statistic

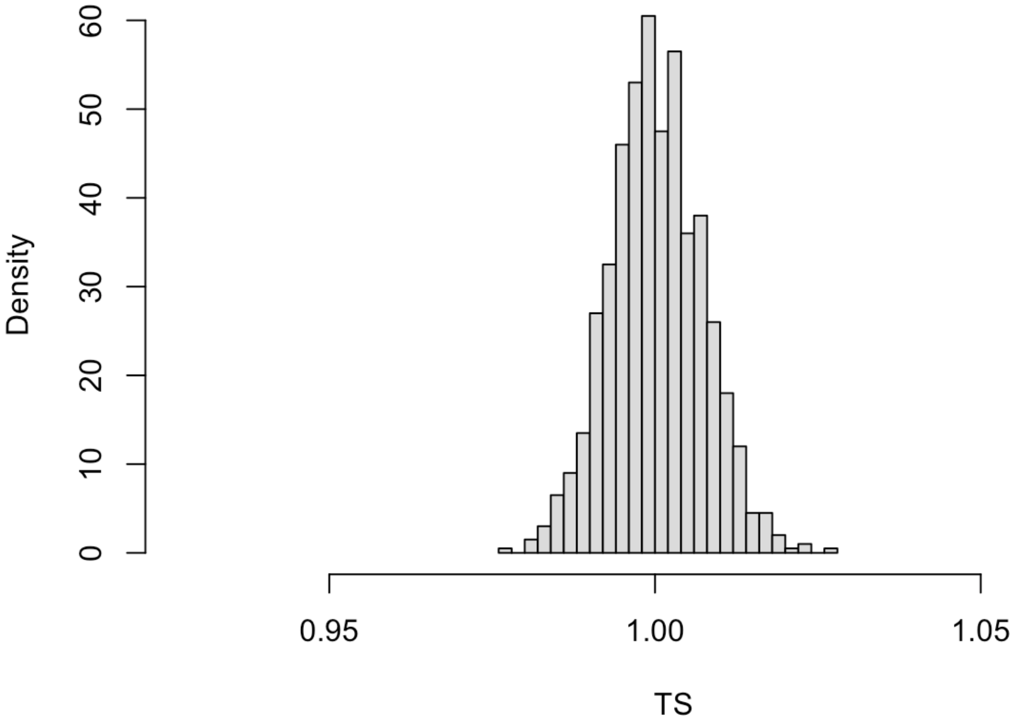
$$\mathbf{T}_n = \frac{1}{p}\sum_{k=1}^{p}\lambda_k.$$

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

Perform `N` simulations and collect all the test statistics.

```
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
}
```

```
hist(TS, breaks=p/3, xlim=c(0.95*min(TS),1.05*max(TS)), freq=FALSE, col='gray86
', main='')
```



Empirical CDF

Generate some data and calculate eigenvalues.

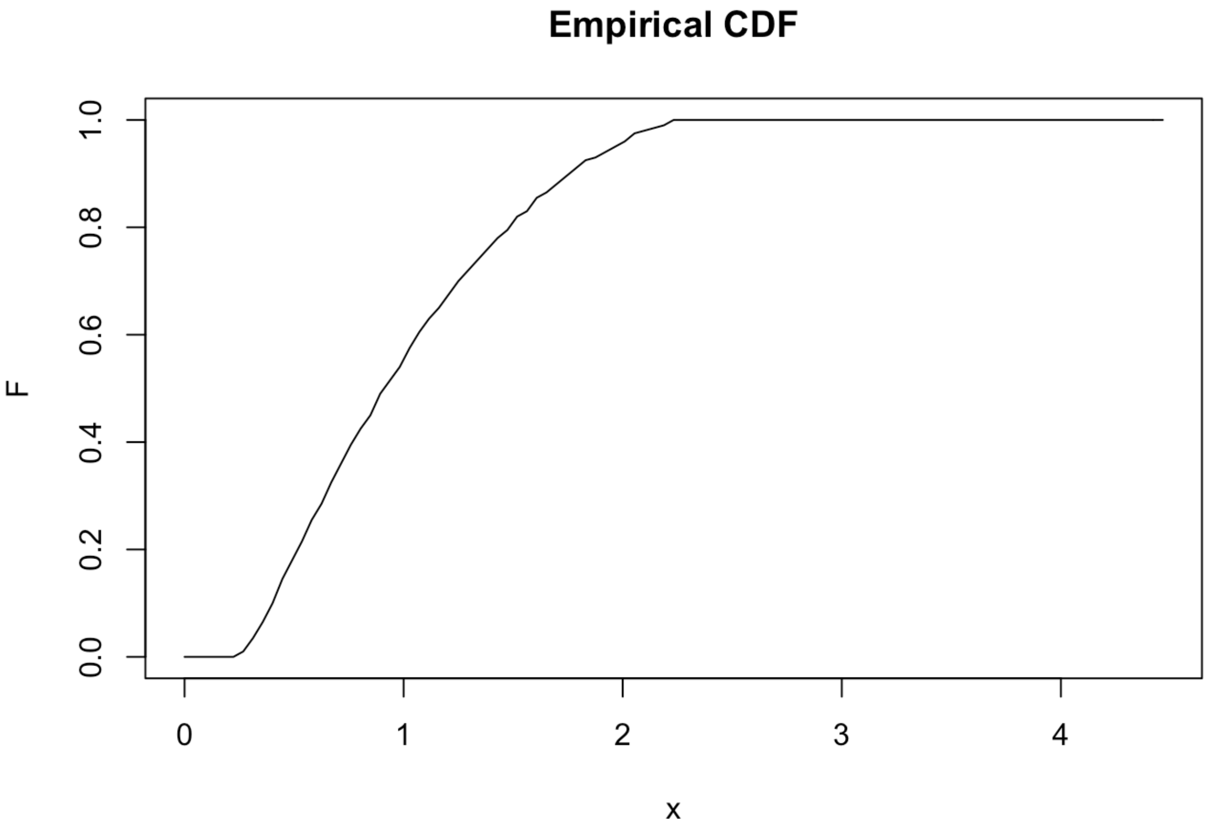
```
p <- 200
n <- 800
X <- matrix(rnorm(p*n), p, n)
Sn <- X %*% t(X) / n
L<-eigen(Sn)$values
```

Taking $\varphi_z(x) = \mathbb{1}(x \leq z)$ as the (z dependant) test function then I can get the empirical CDF.

```
F_ <- function(z) {
  total <- 0.
  for (i in 1:p) {
    if (L[i] <= z) {
      total <- total + 1
    }
  }
  return (total/p)
}
F <- Vectorize(F_)
```

And now plot it.

```
plot(F, from=0., to=2*max(L), main="Empirical CDF")
```



Marcenko-Pastur distribution

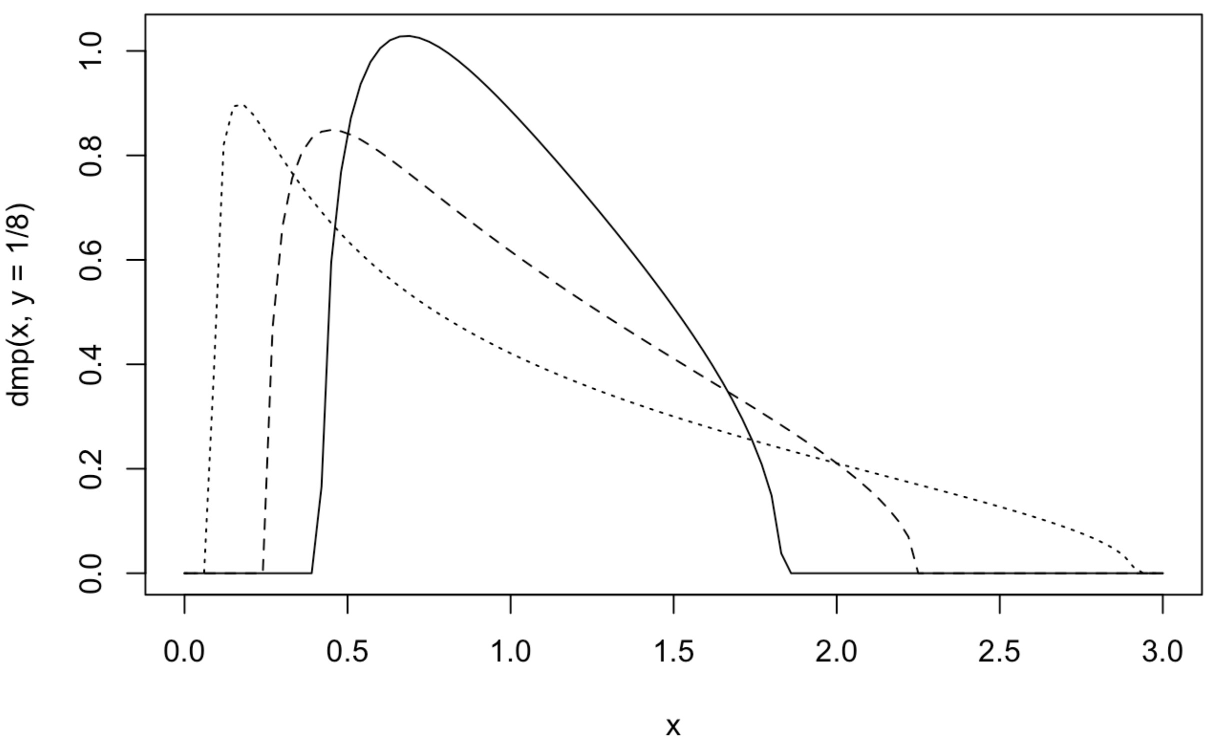
Density

We can implement the density function.

```
dmp <- function(x, y, sigma=1) {
  a <- (1-sqrt(y))^2
  b <- (1+sqrt(y))^2
  ifelse(x <= a | x >= b, 0, suppressWarnings(sqrt((x - a) * (b - x))/(2 * pi *
sigma * x * y)))
}
```

Plot curves for various valued of y.

```
curve(dmp(x, y=1/8), from = 0, to = 3, lty=1)
curve(dmp(x, y=1/4), from = 0, to = 3, lty=2, add=TRUE)
curve(dmp(x, y=1/2), from = 0, to = 3, lty=3, add=TRUE)
```



Eigenvalues of sample covariance matrix

Sample a covariance based on $\mathbb{S}_n = \frac{1}{n}\mathbb{X}\mathbb{X}^*$ where $\mathbb{X}$ is a $p \times n$ matrix with Gaussian entries with mean zero and variance 1. This gives a matrix $\mathbb{S}_n$ of size $p \times p$.

```
p <- 100
n <- 500
X <- matrix(rnorm(p*n), p, n)
Sn <- X %*% t(X) / n
dim(Sn)
```

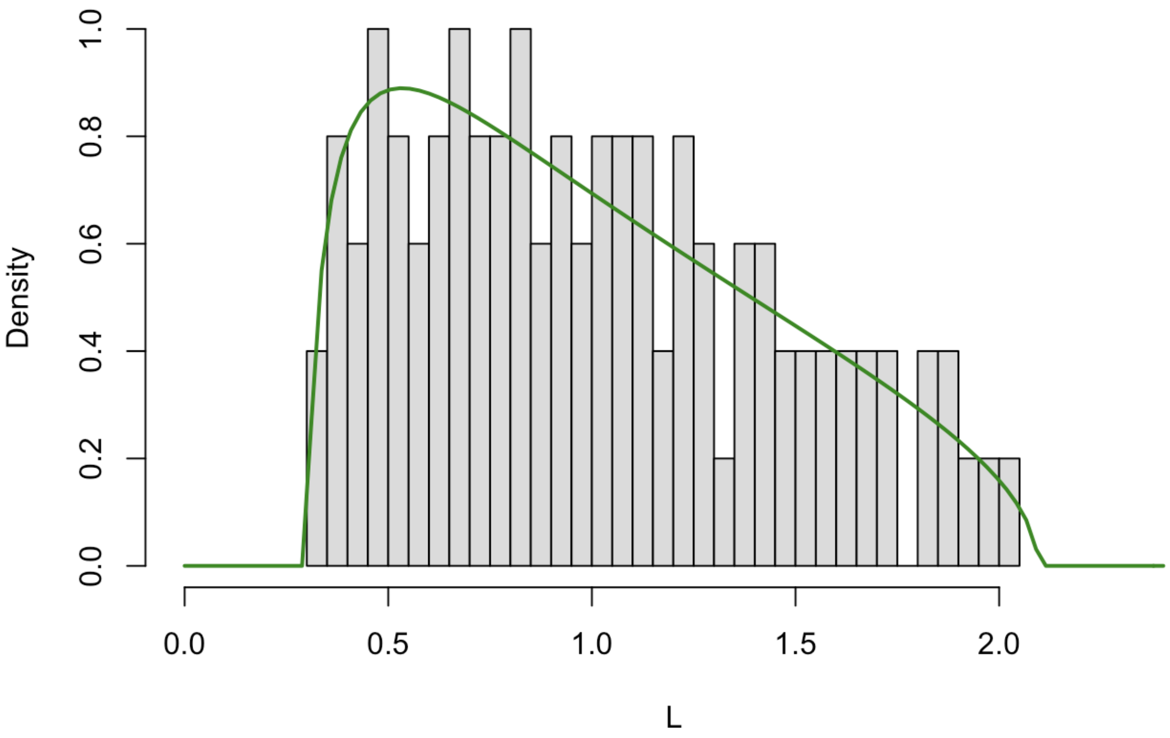
```
## [1] 100 100
```

Calculate the eigenvalues.

```
e<-eigen(Sn)
L<-e$values
```

Plot a histogram of the eigenvalues against the MP density.

```
hist(L, breaks=p/3, xlim=c(0,1.2*max(L)), freq=FALSE, main='', col='gray86')
curve(dmp(x, y=p/n), from = 0, to = 1.2*max(L), lty=1, lw=2, col='green4', add=
TRUE)
```



Now try sampling again (i.e., run the code of this section again) and also vary p and n to see how the shape changes.