Lecture 4: PCA in high dimensions, random matrix theory and financial applications

Foundations of Data Science: Algorithms and Mathematical Foundations

Mihai Cucuringu
mihai.cucuringu@stats.ox.ac.uk

CDT in Mathematics of Random System
University of Oxford

September 26, 2019

Slides by Jim Gatheral, ”Random Matrix Theory and Covariance Estimation” (with minor changes)

Motivation

- Some of the state-of-the-art optimal liquidation portfolio algorithms (that balance risk against impact cost) involve inverting the covariance matrix
- Eigenvalues of the covariance matrix that are small (or even zero) correspond to portfolios of stocks that have
  - nonzero returns
  - extremely low or vanishing risk
- such portfolios are invariably related to estimation errors resulting from insuffient data
- use random matrix theory to alleviate the problem of small eigenvalues in the estimated covariance matrix

Goal is to understand:

- the basis of random matrix theory (RMT)
- how to apply RMT estimating covariance matrices
- whether the resulting covariance matrix performs better than (for example) the Barra covariance matrix

”Barra”: 3^{rd}-party company providing daily covariance matrices

Roadmap

- Random matrix theory
  - Random matrix examples
  - Wigner's semicircle law
  - The Marčenko-Pastur density
  - The Tracy-Widom law
  - Impact of fat tails
- Estimating correlations
  - Uncertainty in correlation estimates
  - Example with SPX stocks
  - Recipe for filtering the sample correlation matrix
- Comparison with Barra
  - Comparison of eigenvectors
  - The minimum variance portfolio
    - Comparison of weights
    - In-sample and out-of-sample performance

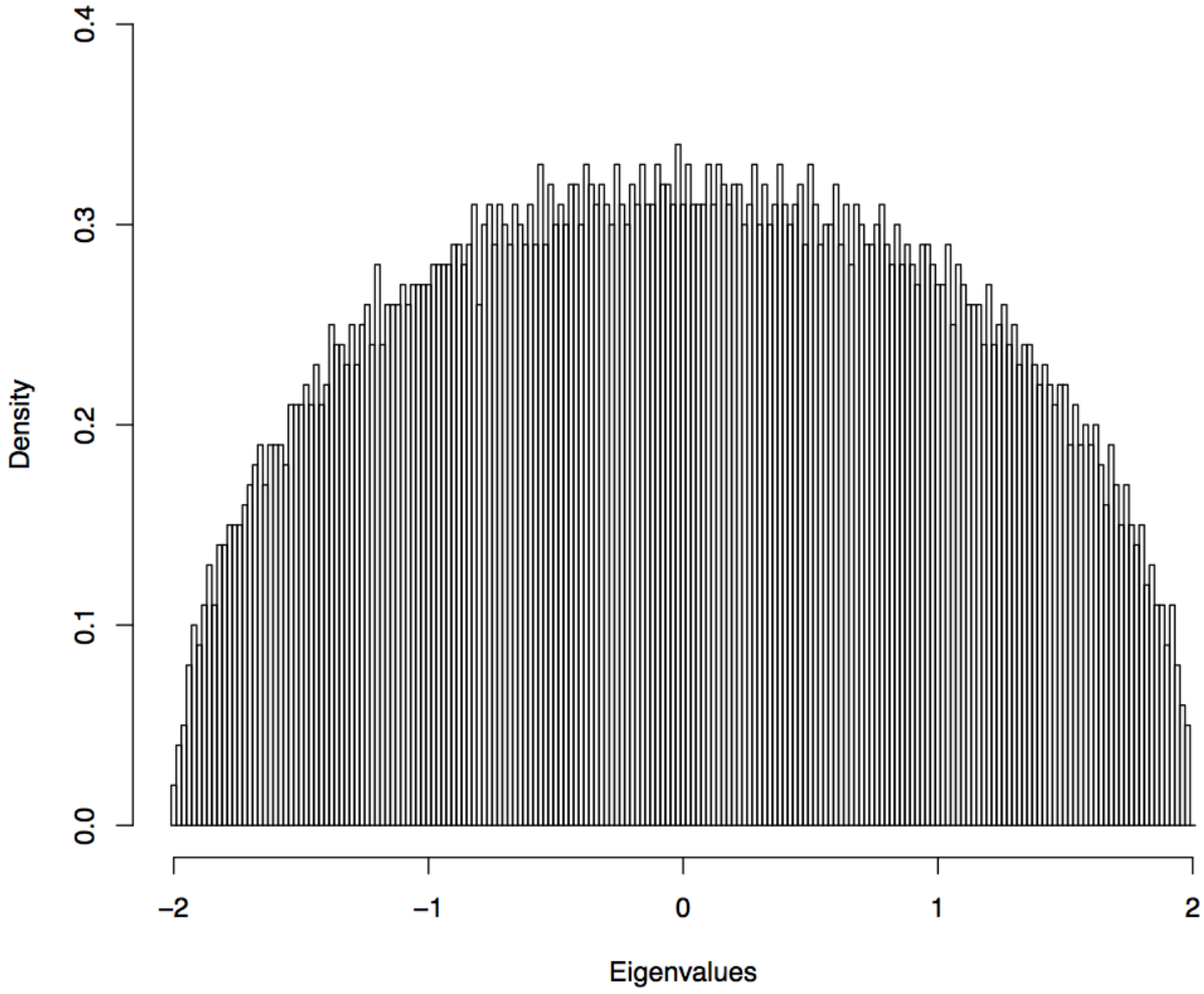
Example 1: Normal random symmetric matrix

- Generate a $5,000 \times 5,000$ random symmetric matrix with entries $A_{ij} \sim N(0,1)$
- Compute eigenvalues
- Draw the histogram of all eigenvalues.

R-code to generate a symmetric random matrix whose off-diagonal elements have variance $\frac{1}{N}$:

- n = 5000
- m = array(rnorm(n^2),c(n,n));
- m2 = (m+t(m))/sqrt(2*n); # Make m symmetric
- lambda = eigen(m2, symmetric=T, only.values = T);
- ev = lambda$values;
- hist(ev, breaks=seq(-2.01,2.01,0.02),main=NA, xlab="Eigenvalues",freq=F)

Normal random symmetric matrix



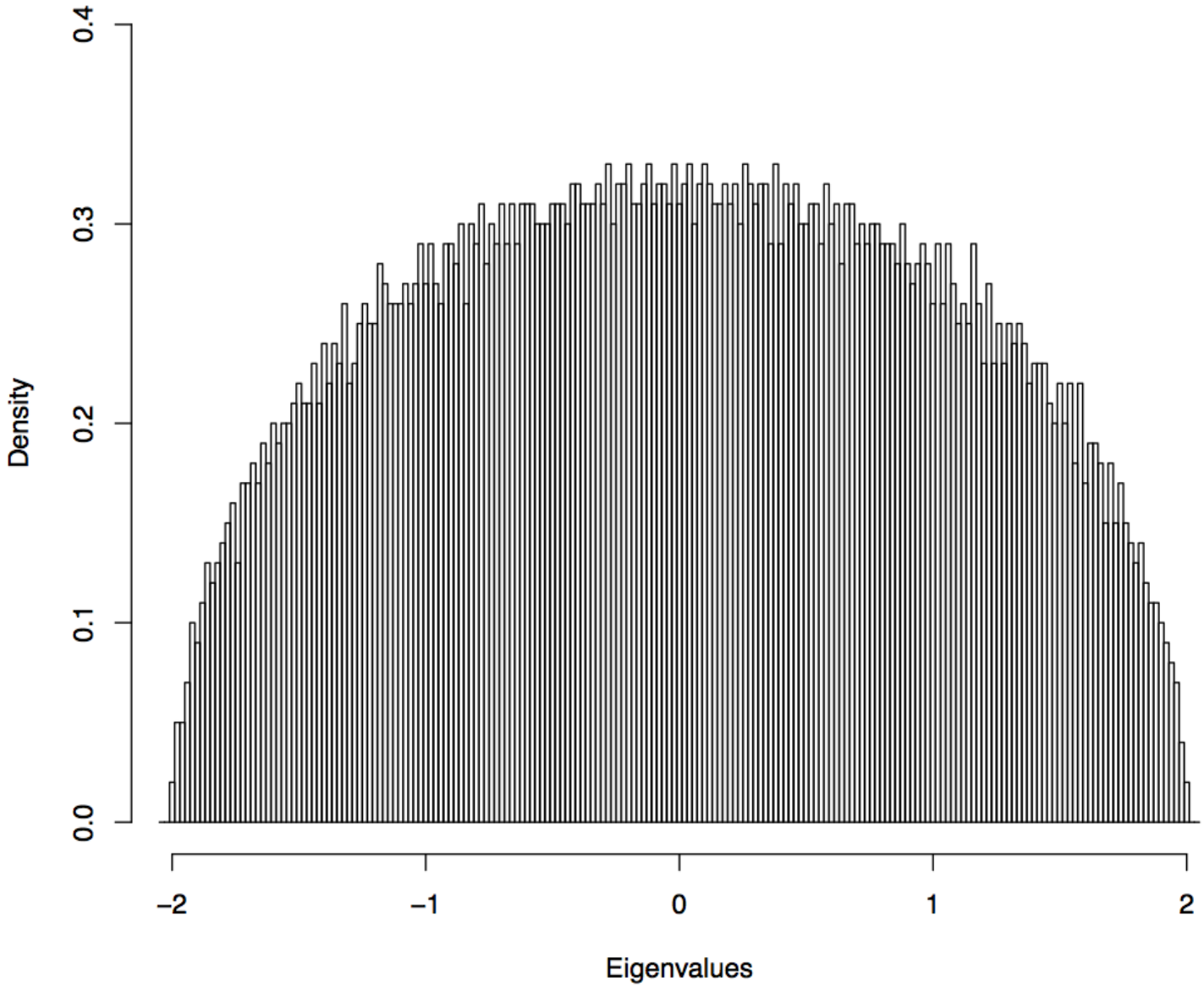
Uniform random symmetric matrix

- Generate a 5,000 x 5,000 random symmetric matrix with entries $A_{ij} \sim \text{Uniform}(0, 1)$
- Compute eigenvalues
- Draw the histogram of all eigenvalues

Here's some R-code again:

- n = 5000;
- mu = array(runif(n^2),c(n,n))
- mu2 = sqrt(12)*(mu+t(mu)-1)/sqrt(2*n)
- lambdau = eigen(mu2, symmetric=T, only.values = T)
- ev = lambdau$values;
- hist(ev, breaks=seq(-2.01,2.01,0.02), main=NA, xlab="Eigenvalues",freq=F)

Uniform random symmetric matrix



What can we conclude?

Note the striking pattern: the density of eigenvalues is a semicircle!

Wigner's semicircle law

Let $\tilde{A}$ by an $N \times N$ matrix with entries $\tilde{A}_{ij} \sim N(0, \sigma^2)$. Define

$$A_N = \frac{1}{\sqrt{N}} \left(\frac{A + A^T}{2} \right)$$

- A_N is symmetric with variance

$$\text{Var}[a_{ij}] = \begin{cases} \sigma^2/N & \text{if } i \neq j \\ 2\sigma^2/N & \text{if } i = j \end{cases} \tag{1}$$

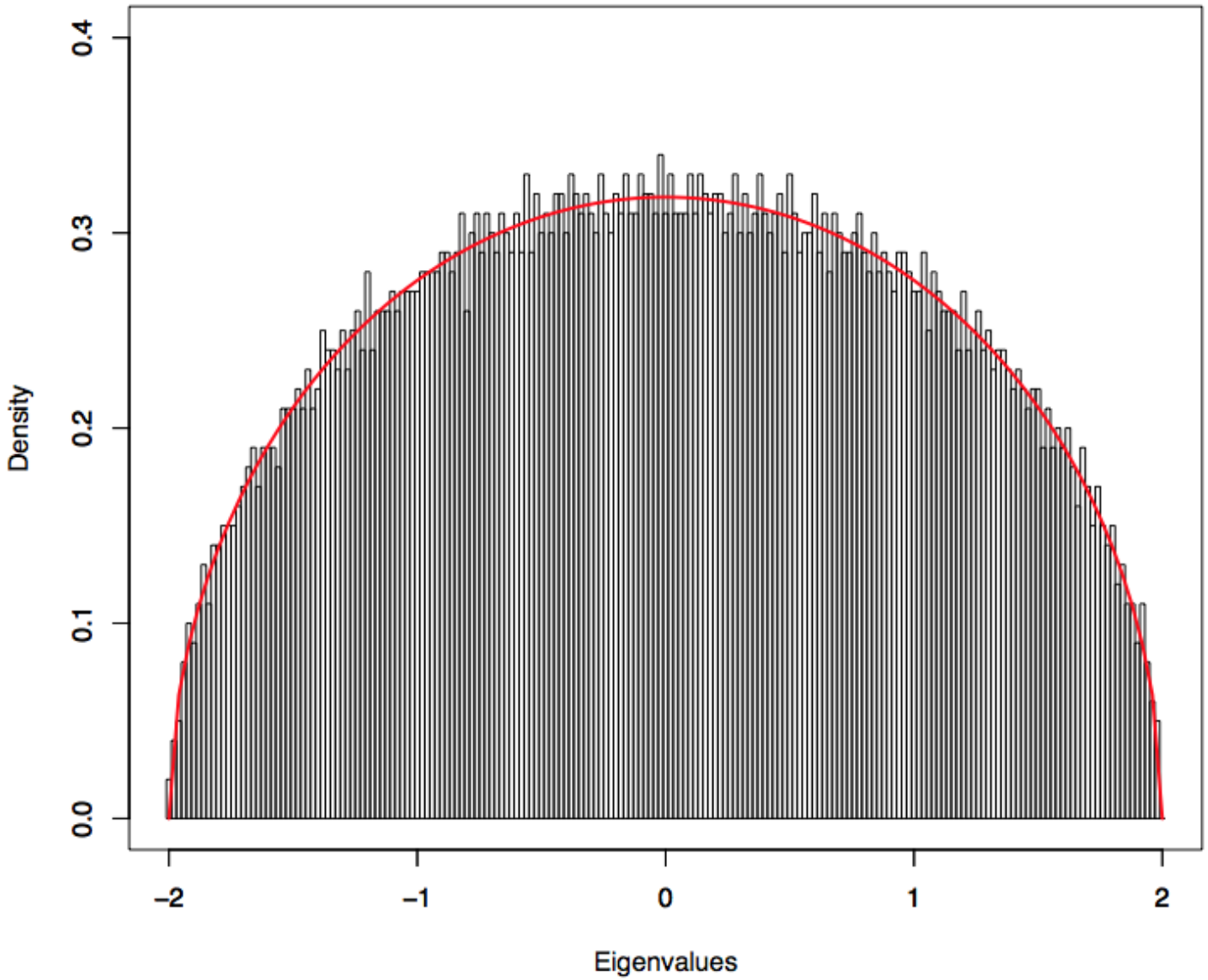
- the density of eigenvalues of A_N is given by

$$\rho_N(\lambda) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \delta(\lambda - \lambda_i)$$

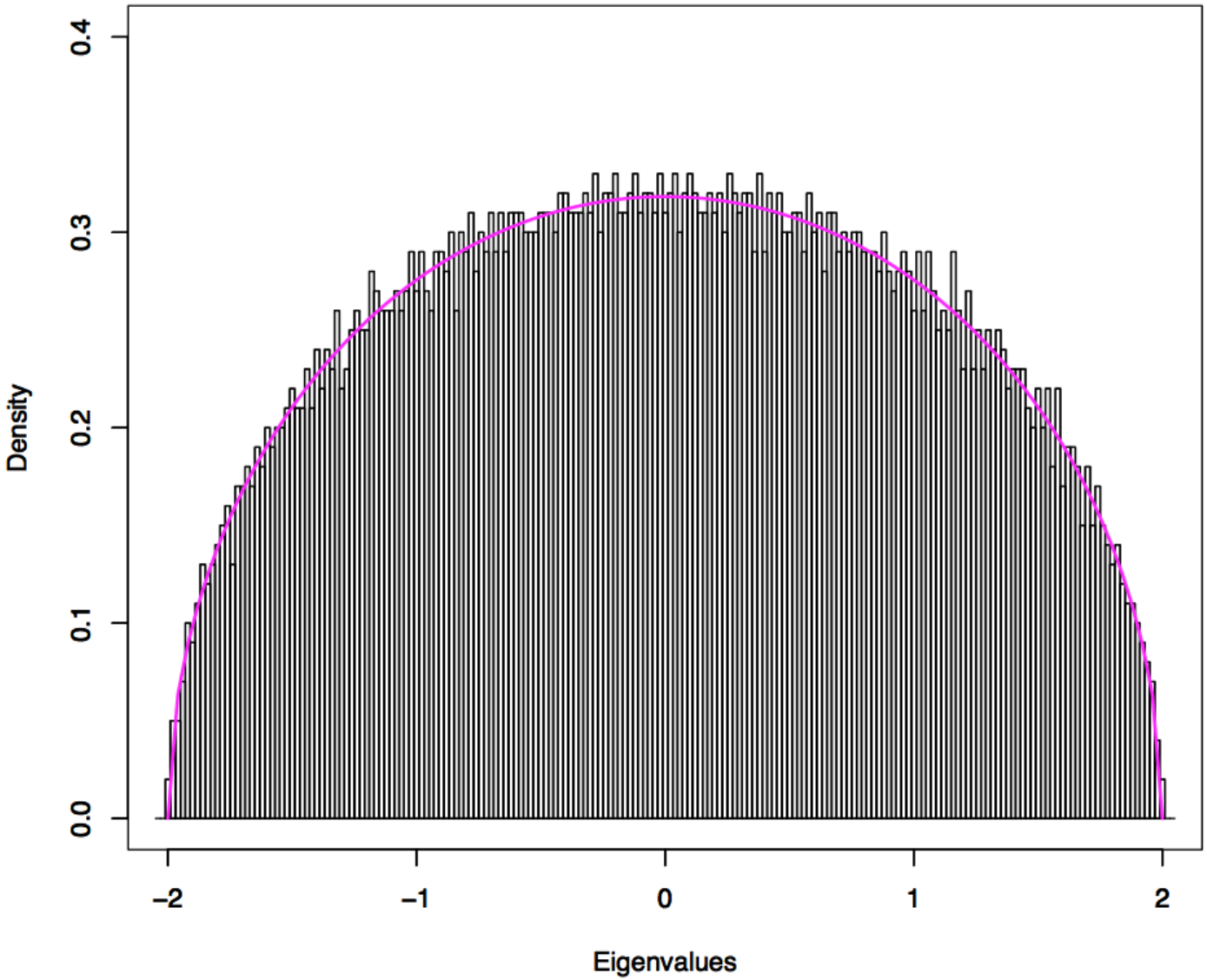
which, as shown by Wigner

$$\text{as } n \to \infty \longrightarrow \begin{cases} \frac{1}{2\pi\sigma^2} \sqrt{4\sigma^2 - \alpha^2} & \text{if } |\lambda| \leq 2\sigma \\ 0 & \text{otherwise} \end{cases} \stackrel{\text{def}}{=} \rho(\lambda) \tag{2}$$

Normal random matrix + Wigner semicircle density



Uniform random matrix + Wigner semicircle density



Random correlation matrices

- we have M stock return series with T elements each
- the elements of the $M \times M$ empirical correlation matrix E are given by

$$E_{ij} = \frac{1}{T} \sum_{t=1}^{T} x_{it} x_{jt}$$

where x_{it} denotes the return at time t of stock i, normalized by the standard deviation so that $\mathrm{Var}[x_{it}] = 1$

- in compact matrix form, this can be written as

$$E = HH^T$$

where H is the $M \times T$ matrix whose rows are the time series of returns, one for each stock

Eigenvalue spectrum of random correlation matrix

- Suppose the entries of H are random with variance σ^2
- Then, in the limit T, M $\to \infty$, while keeping the ratio $Q \stackrel{\text{def}}{=} \frac{T}{M} \geq 1$ constant, the density of eigenvalues of E is given by

$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda_- - \lambda)}}{\lambda}$$

where the max and min eigenvalues are given by

$$\lambda_\pm = \sigma^2 \left(1 \pm \sqrt{\frac{1}{Q}}\right)^2$$

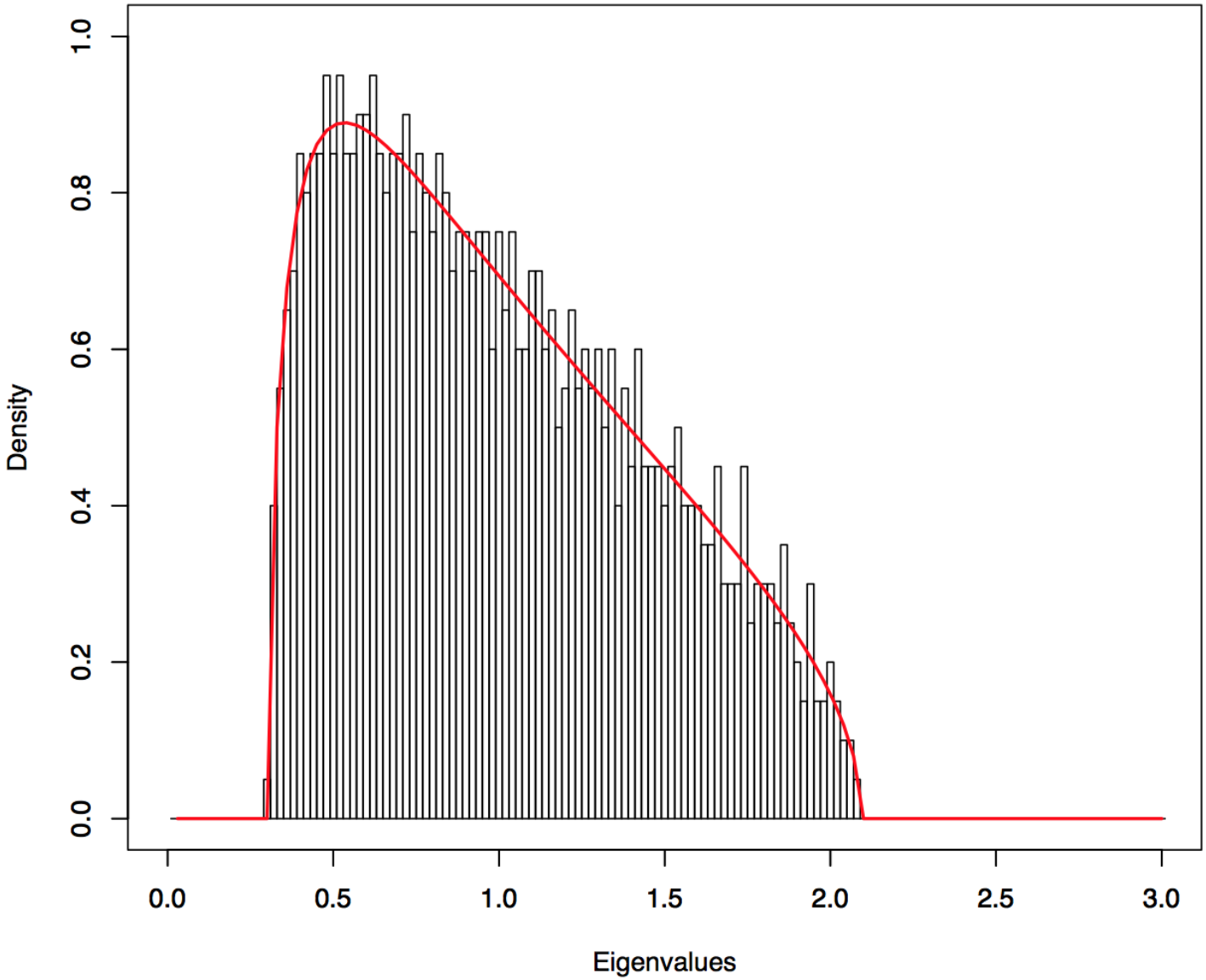
- $\rho(\lambda)$ is also known as the Marčhenko-Pastur distribution that describes the asymptotic behavior of eigenvalues of large random matrices

Example: IID random normal returns

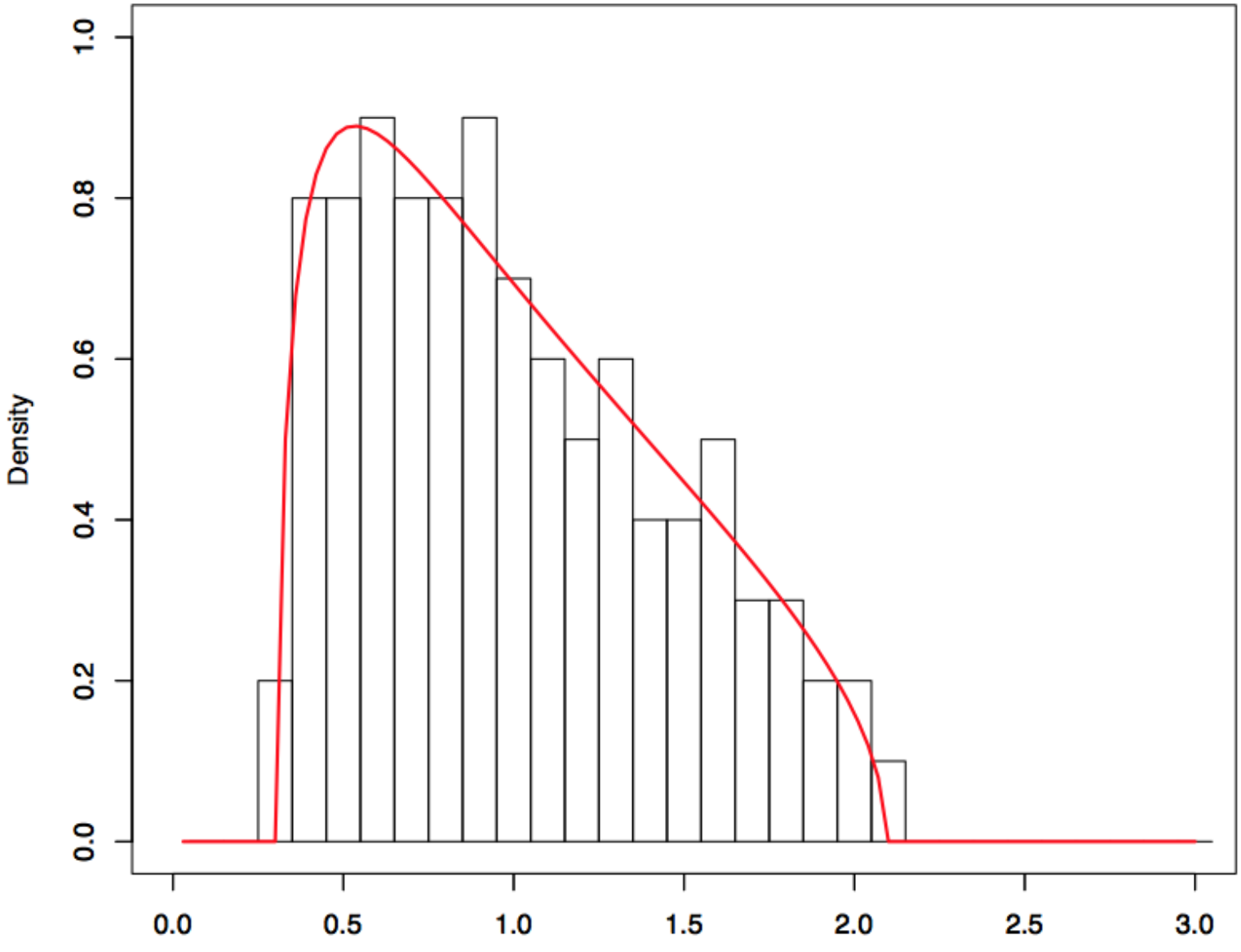
R-code:

- t = 5000;
- m = 1000;
- h = array(rnorm(m*t),c(m,t)); # Time series in rows
- e = h % * % t(h)/t; # Form the correlation matrix
- lambdae = eigen(e, symmetric=T, only.values = T);
- ee = lambdae$values;
- hist(ee, breaks =seq(0.01,3.01,.02), main=NA, xlab=”Eigenvalues”, freq=F)

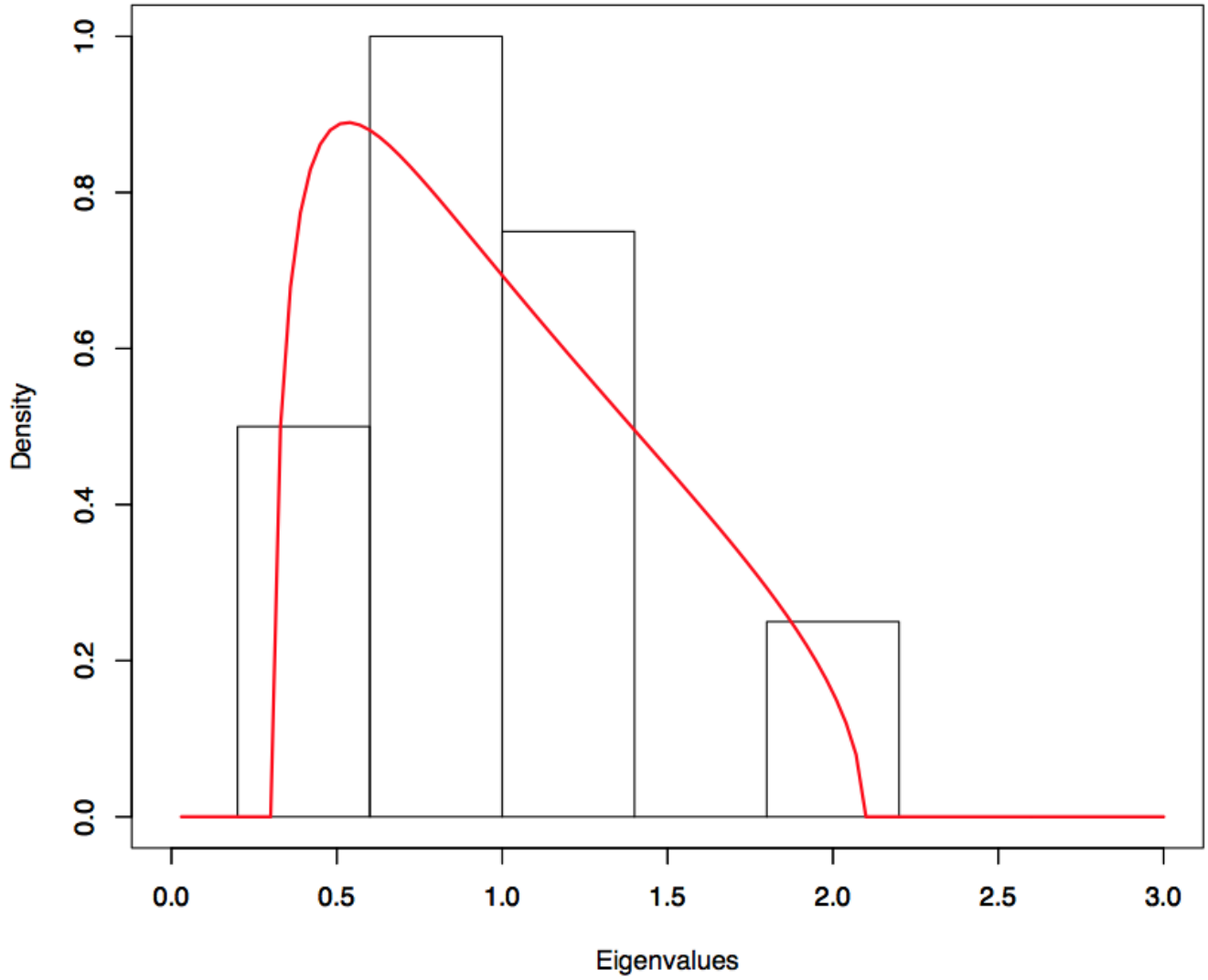
Empirical density with superimposed Marčhenko-Pastur density



Empirical density for $M = 100$, $T = 500$ (with Marčenko-Pastur density superimposed)

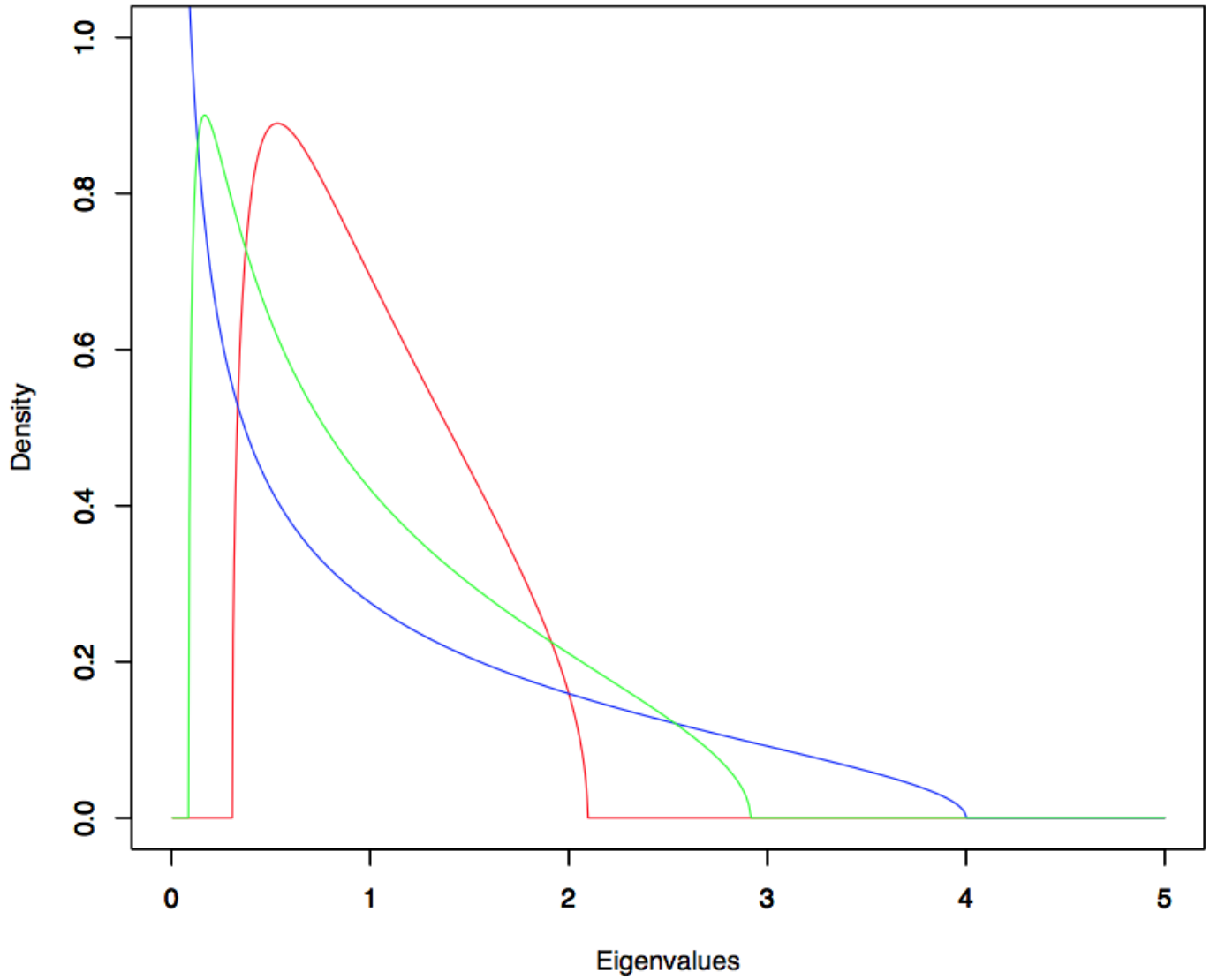


Empirical density for $M = 10$, $T = 50$ (with Marčenko-Pastur density superimposed)



Marčenko-Pastur densities depends on $Q = T/M$

density for $Q = 1$ (blue), 2 (green) and 5 (red).



Tracy-Widom: of the largest eigenvalue

- For certain applications we would like to know
  - where the random bulk of eigenvalues ends
  - where the spectrum of eigenvalues corresponding to true information begins
- $\Rightarrow$ need to know the distribution of the largest eigenvalue
- The distribution of the largest eigenvalue of a random correlation matrix is given by the Tracy-Widom law

$$P(T\lambda_{max} < \mu_{TM} + s\sigma_{TM}) = F_1(s)$$

where

$$\mu_{TM} = \left(\sqrt{T - \frac{1}{2}} + \sqrt{M - \frac{1}{2}}\right)^2$$

$$\sigma_{TM} = \left(\sqrt{T - \frac{1}{2}} + \sqrt{M - \frac{1}{2}}\right)\left(\frac{1}{\sqrt{T - \frac{1}{2}}} + \frac{1}{\sqrt{M - \frac{1}{2}}}\right)^{1/3}$$

Fat-tailed random matrices

So far, we have considered matrices whose entries are

- Gaussian
- uniformly distributed

But, in practice: stock returns exhibit a fat-tailed distribution

- Bouchaud et al.: fat tails can massively increase the maximum eigenvalue in the theoretical limiting spectrum of the random matrix
- ”Financial Applications of Random Matrix Theory: a short review”, J.P. Bouchaud and M. Potters
  `http://arxiv.org/abs/0910.1205`
- For extremely fat-tailed distributions (Cauchy for example), the semi-circle law no longer holds

Sampling error

- suppose we compute the sample correlation matrix of M stocks with T returns in each time series.
- assume the true correlation were the identity matrix
- Q: expected value of the greatest sample correlation?
- for $N(0,1)$ distributed returns, the median maximum correlation ρ_{max} should satisfy:

$$\log 2 \approx \frac{M(M-1)}{2} N(-\rho_{max}\sqrt{T})$$

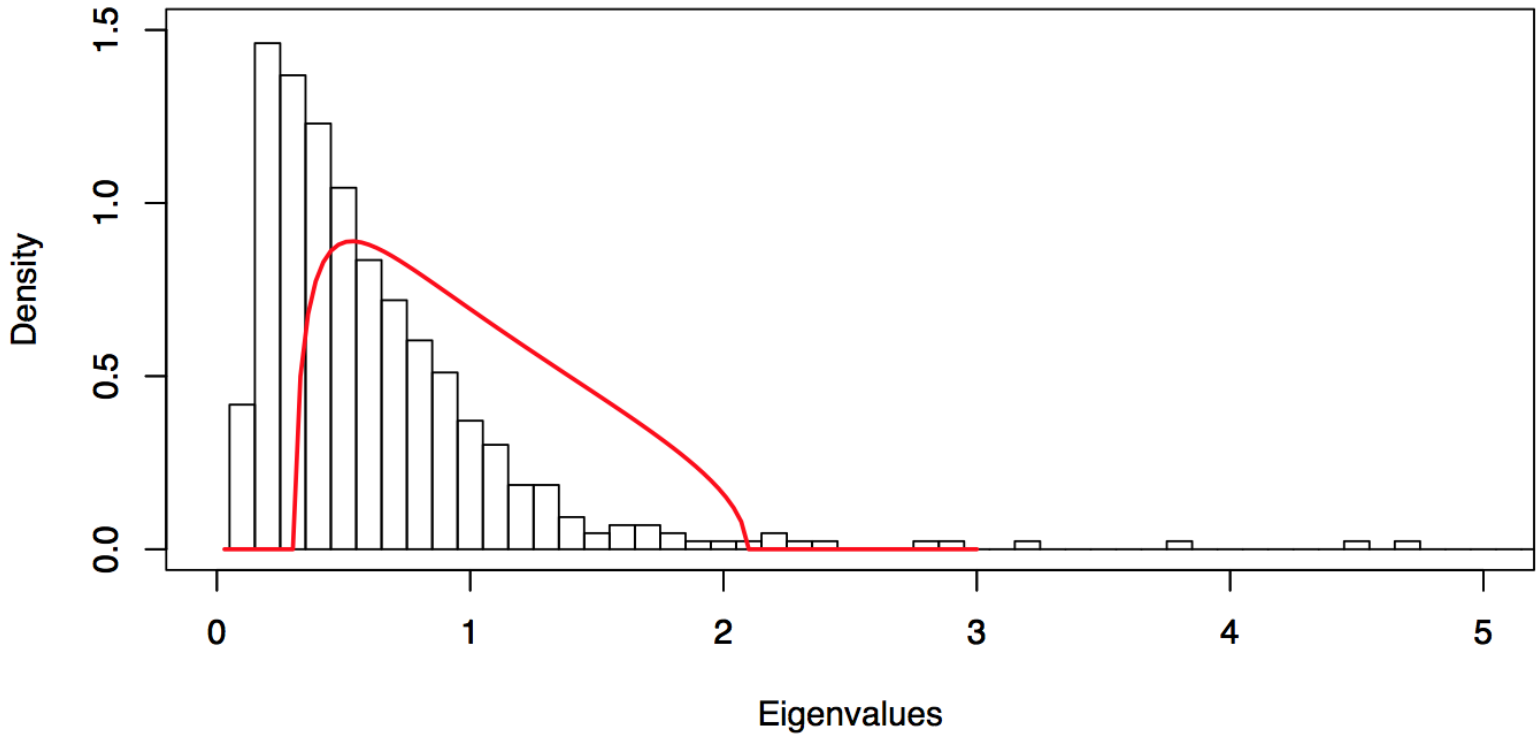
- with M = 500,T = 1000, we obtain $\rho_{max} \approx 0.14$
- sampling error induces spurious (and potentially significant) correlations between stocks!

An experiment with real data

- $M = 431$ stocks in the S&P 500 index for which we have $T = 5 \times 431 = 2155$ consecutive daily returns
- $Q = T/M = 5$
- There are $M(M-1)/2 = 92,665$ distinct entries in the correlation matrix to be estimated from $2,155 \times 431 = 928,805$ data points

First, compute the eigenvalue spectrum and superimpose the Marčenko-Pastur density with $Q = 5$.

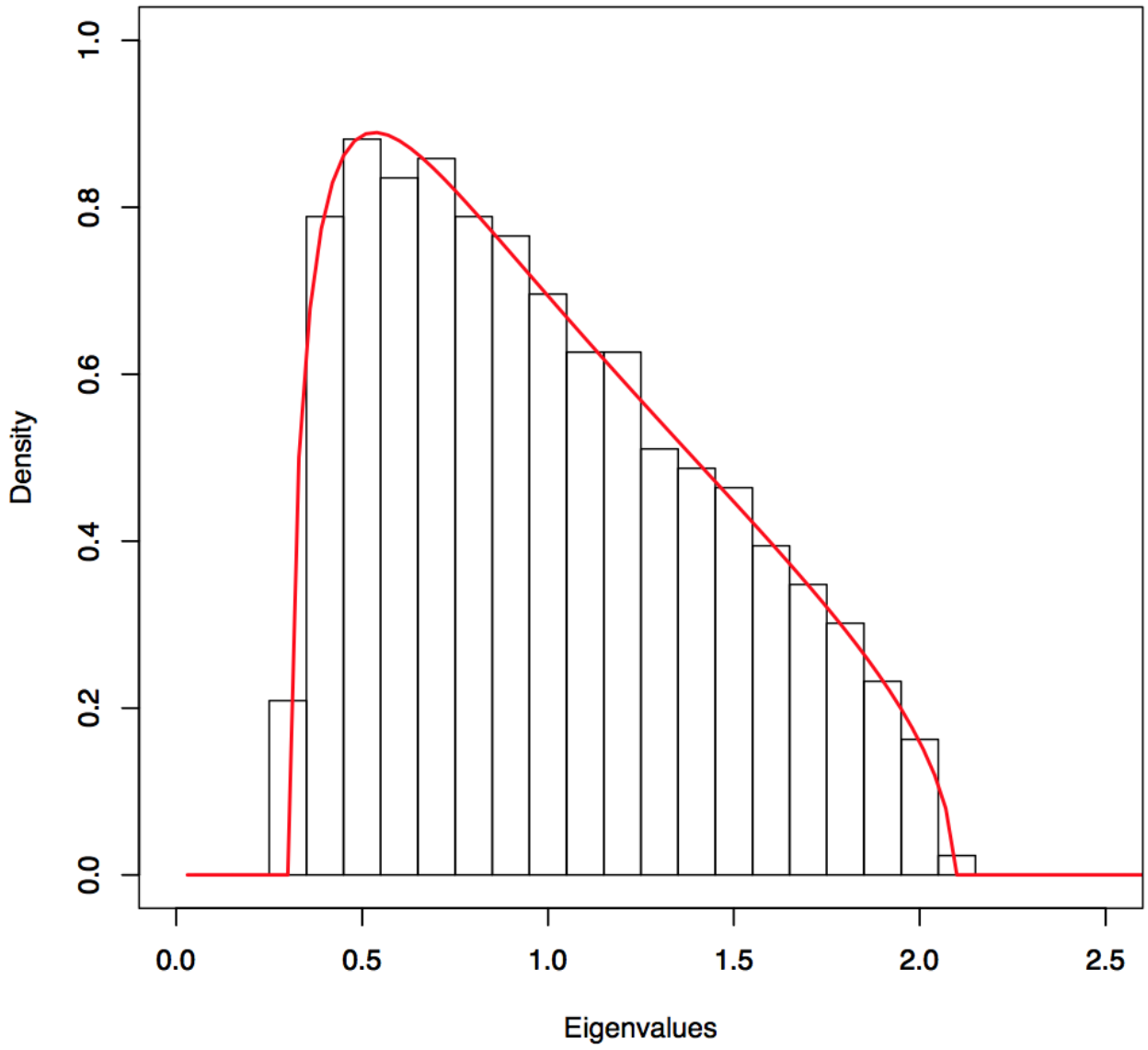
The eigenvalue spectrum of the sample correlation



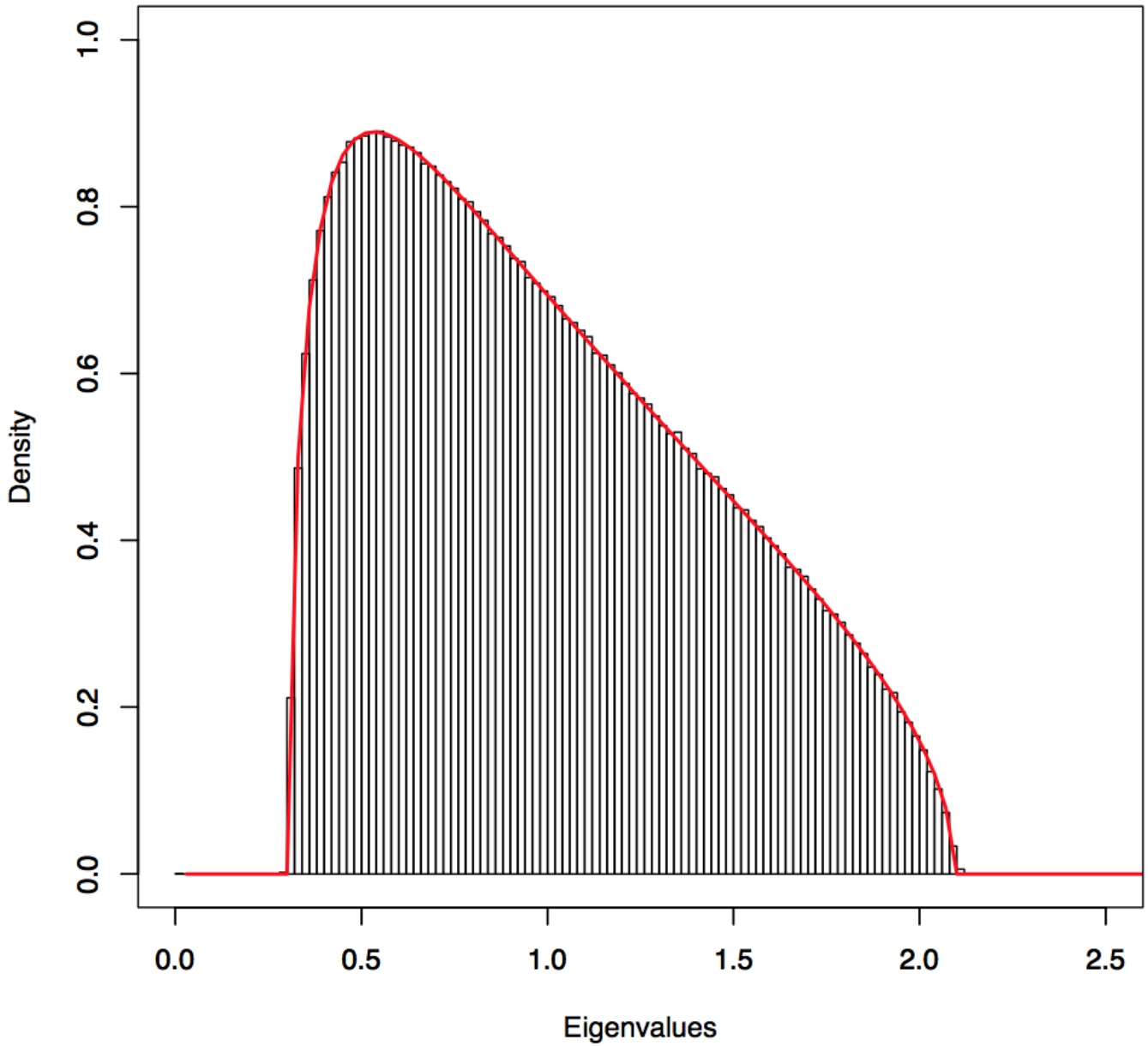
Note that the top eigenvalue is 105.37 – way off the end of the chart! The next biggest eigenvalue is 18.73

With randomized return data

If we shuffle the returns in each time series:

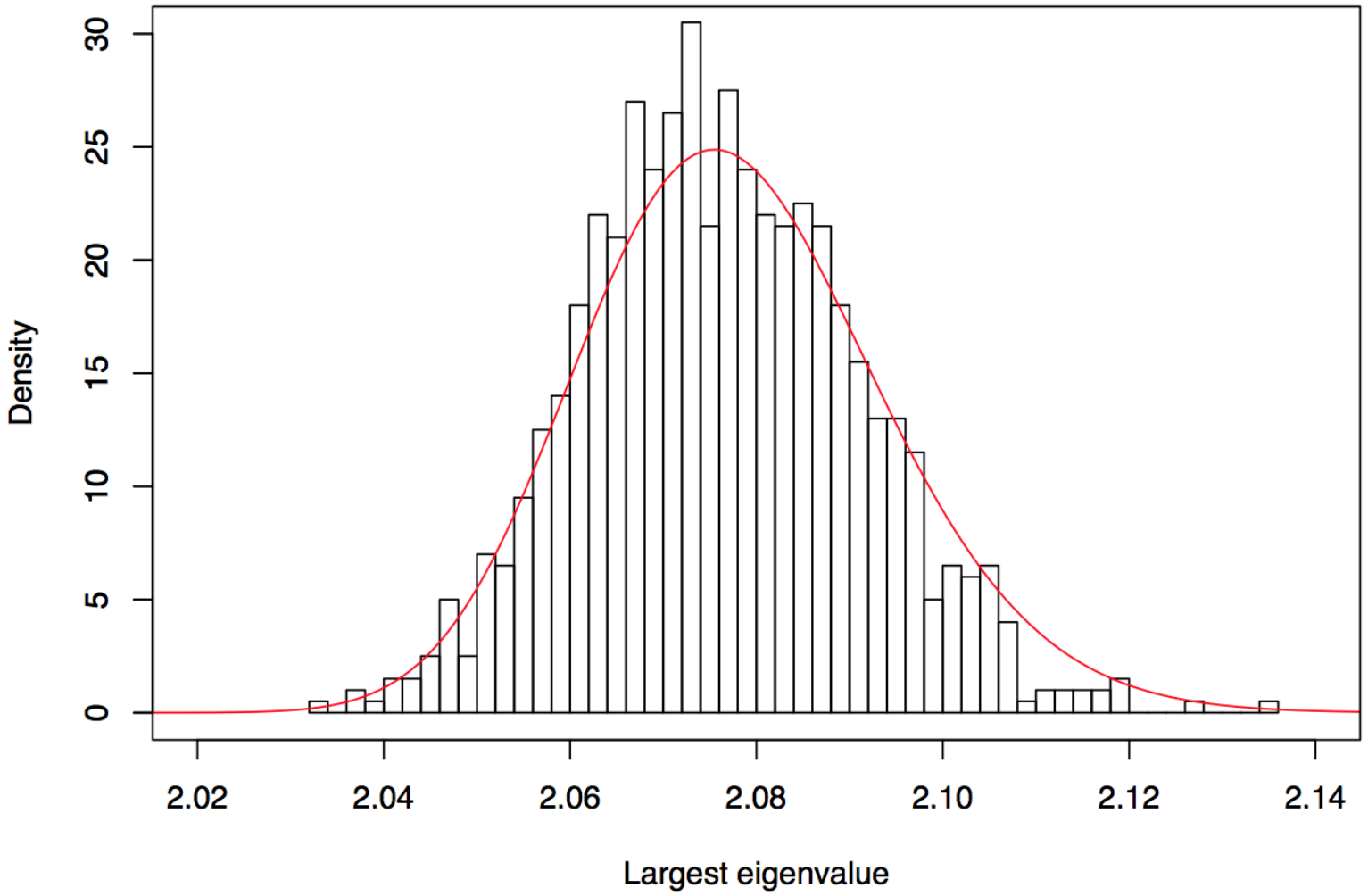


Repeating 1,000 times and average



Distribution of the largest eigenvalue

We can compare the empirical distribution of the largest eigenvalue with the Tracy-Widom density (in red):



Interim conclusions

Remarks:

- Even though return series are fat-tailed:
  - the Marčenko-Pastur density is a very good approximation to the density of eigenvalues of the correlation matrix of the randomized returns
  - the Tracy-Widom density is a good approximation to the density of the largest eigenvalue of the correlation matrix of the randomized returns
- the Marčenko-Pastur density does not remotely fit the eigenvalue spectrum of the sample correlation matrix
  - $\Rightarrow$ there is non-random structure in the return data
- can compute the theoretical spectrum arbitrarily accurately by performing numerical simulations

Problem formulation

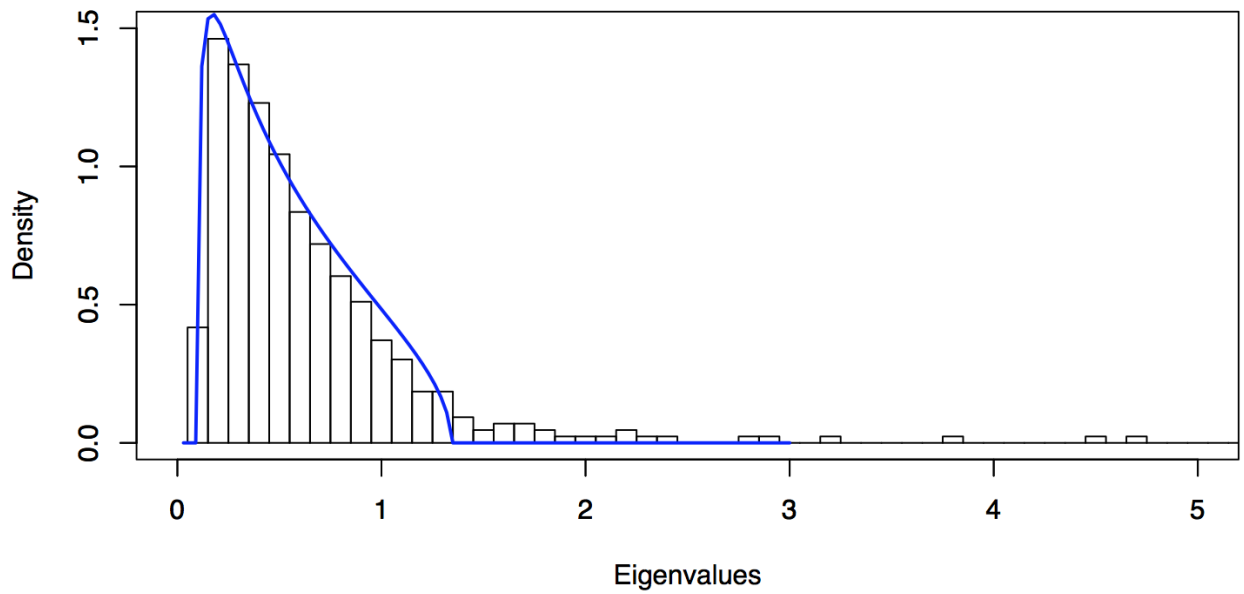
- Which eigenvalues are significant and how do we interpret their corresponding eigenvectors?

A hand-waving practical approach

Suppose we find the values of σ and Q that best fit the bulk of the eigenvalue spectrum. We find

$$\sigma = 0.73; Q = 2.9$$

and obtain the following plot:



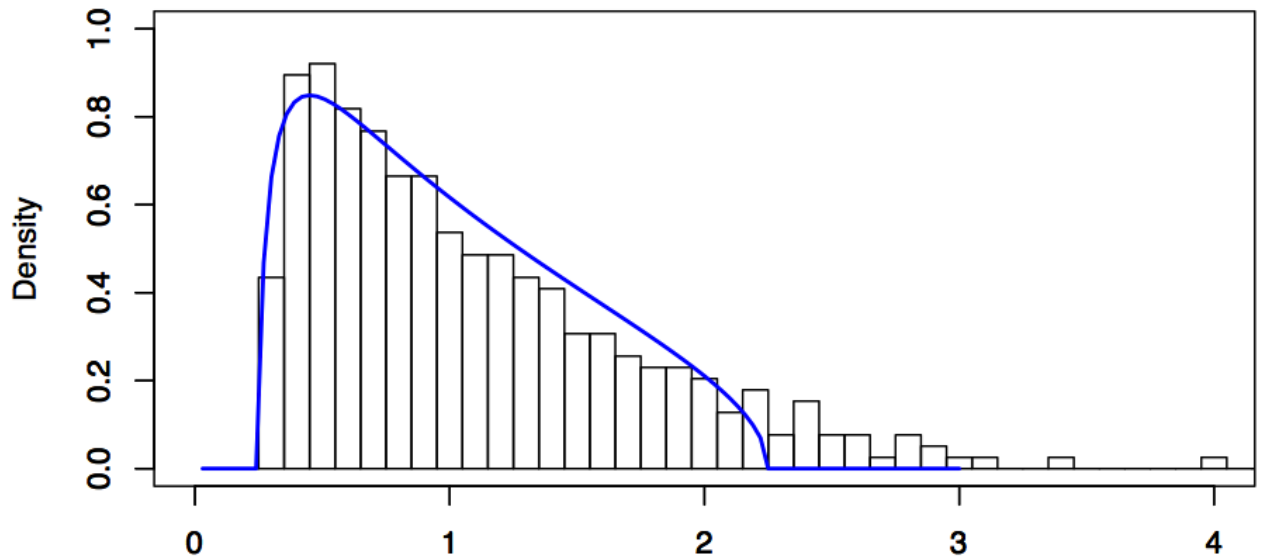
Max and min MP eigenvalues are 1.34 and 0.09 respectively.

Some analysis

- If we are to believe this estimate, a fraction $\sigma^2 = 0.53$ of the variance is explained by eigenvalues that correspond to random noise. The remaining fraction 0.47 has information
- From the plot, it looks as if we should cut off eigenvalues above 1.5 or so
- Summing the eigenvalues themselves, we find that 0.49 of the variance is explained by eigenvalues greater than 1.5

More carefully: correlation matrix of residual returns

- For each stock, subtract factor returns associated with the top 25 eigenvalues ($\lambda > 1.6$)
- For $\sigma = 1; Q = 4$ we get the best fit of the Marčenko-Pastur density and obtain the following plot:



- Maximum and minimum Marčenko-Pastur eigenvalues are 2.25 and 0.25 respectively.

Distribution of eigenvector components

- If there is no information in an eigenvector, we expect the distribution of the components to be a maximum entropy distribution
- Specifically, if we normalized the eigenvector u such that its components u_i satisfy

$$\sum_{i=1}^{M} u_i^2 = M,$$

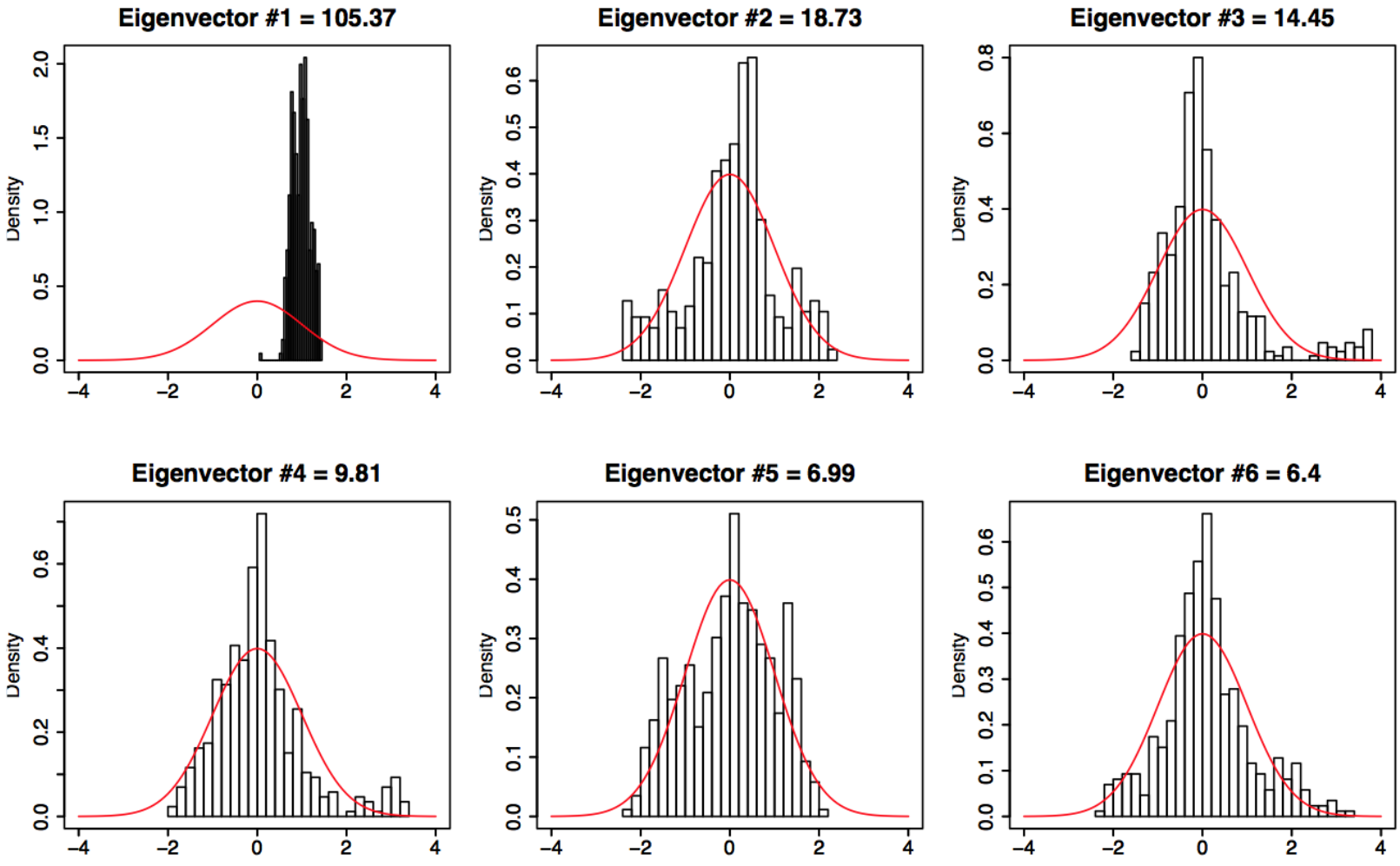
the distribution of the u_i should have the limiting density

$$p(u) = \sqrt{\frac{1}{2\pi}} e^{-\frac{u^2}{2}}$$

- Next, superimpose the empirical distribution of eigenvector components and the zero-information limiting density for various eigenvalues...

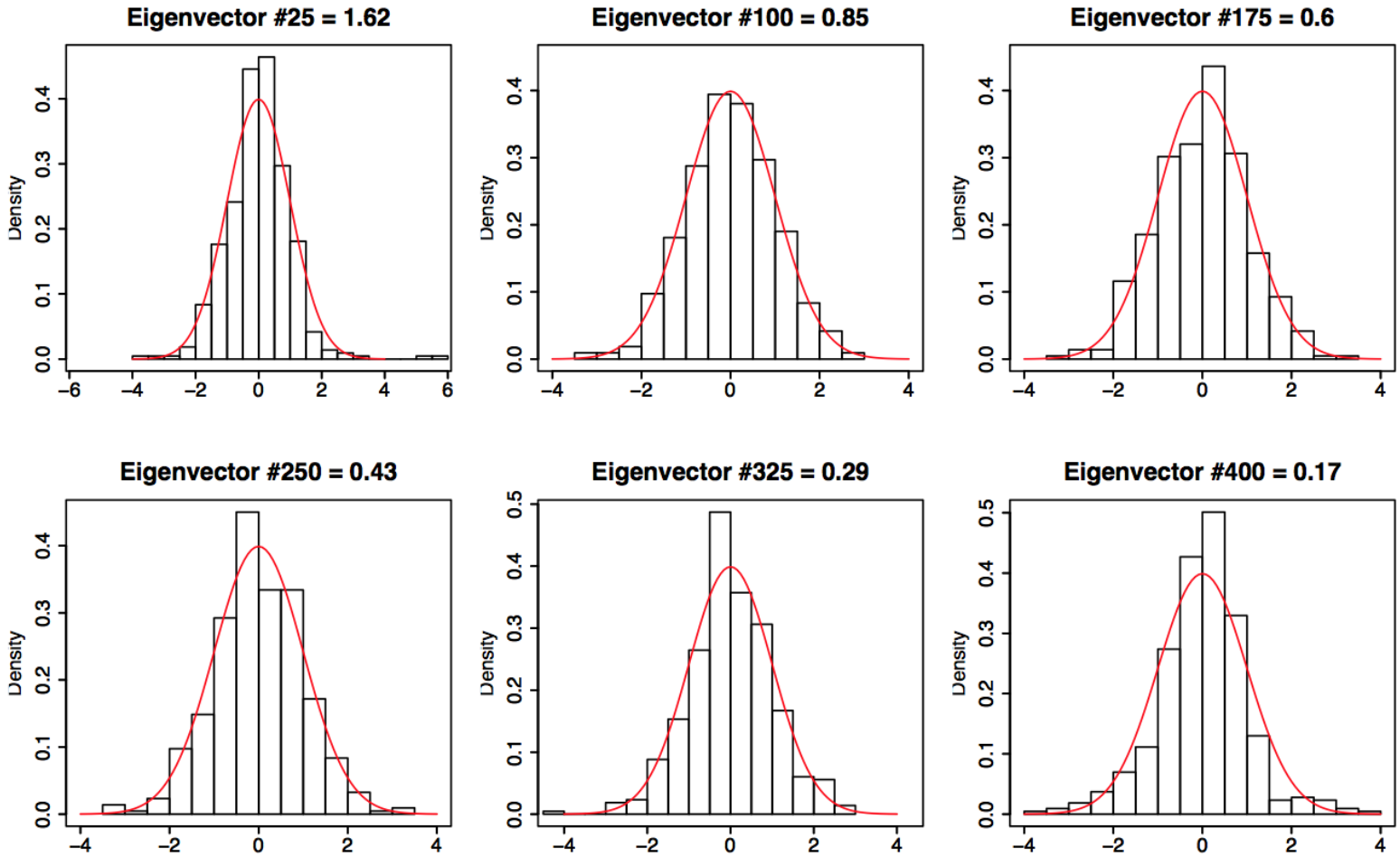
Informative eigenvalues (1)

Plots for the six largest eigenvalues:



Non-informative eigenvalues

Plots for six eigenvalues in the bulk of the distribution:



Resulting recipe

1. Fit the Marcenko-Pastur distribution to the empirical density to determine Q and σ
2. All eigenvalues above a threshold λ^* are considered informative; otherwise eigenvalues relate to noise
3. Replace all noise-related eigenvalues λ_i below λ^* with a constant and renormalize so that $\sum_{i=1}^{M} \lambda_i = M$
   - Recall that each eigenvalue relates to the variance of a portfolio of stocks
   - A very small eigenvalue means that there exists a portfolio of stocks with very small out-of-sample variance – something we probably don't believe
4. Undo the diagonalization of the sample correlation matrix C to obtain the denoised estimate C'
   - Remember to set diagonal elements of C' to 1

Interesting problem to consider:

- how many meaningful eigenvalues/eigenvectors (k) have there been in the US equity market over the last 10-15 years?
- sliding window based on previous m=6 months of data
- perform analysis every $s = 1$ month
- infer k using RMT
- end result: plot time versus k, for different values of m
- explore the interplay with mean reversion
- redo the analysis in daily traded volume space