Case Study 4

1. Recap on correlation matrices and their properties

2. Principal Component Analysis

3. Application to cryptocurrency data (portfolio selection)

4. Basics of Random Matrix Theory

Case Study 3 - Useful formulas

Correlation coefficient

$$c_{ij} = \frac{1}{T}\sum_{t=1}^{T} \frac{(r_{it}-m_i)(r_{jt}-m_j)}{\sigma_i \sigma_j} = \frac{1}{T}\sum_{t=1}^{T} x_{it}x_{jt}$$

Standardized returns

$$x_{it} = \frac{r_{it}-m_i}{\sigma_i}$$

(zero mean & unit std. deviation)

Correlation matrix

$$C \to \text{matrix with entries } c_{ij}$$

Principal Component Analysis (PCA) I

GOAL: mapping a set of correlated variables onto a new set of completely uncorrelated variables (called principal components)

WHY: principal components can be ranked in order of importance based on the fraction of the variance of the original data that they can explain

Principal Component Analysis (PCA) II

Eigenvalue problem

$$C\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i \qquad i = 1, \ldots, N$$

Normalized eigenvectors

$$\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iN})$$

Eigenvalues

$$0 \leq \lambda_N \leq \ldots \leq \lambda_1$$

Principal Component Analysis (PCA) III

Definition of PCs

$$e_{it} = \frac{1}{\sqrt{\lambda_i}} \sum_{k=1}^{N} v_{ki} x_{kt}$$

PCs have zero mean, unit std. deviation and are **exactly** uncorrelated:

$$\frac{1}{T} \sum_{t=1}^{T} e_{it} e_{jt} = 0$$

Relationship with original variables

$$x_{it} = \sum_{k=1}^{N} \sqrt{\lambda_k} v_{ik} e_{kt}$$

PCA - Practical use

$$x_{it} = \sum_{k=1}^{N} \sqrt{\lambda_k} v_{ik} e_{kt}$$

$$\sum_{k=1}^{N} \lambda_k = N$$

Numbers of order 1

Whenever a few eigenvalues dominate, we can approximate original variables in terms of a few PCs and use them to gain intuition

$$x_{it} \approx \sum_{k=1}^{M<N} \sqrt{\lambda_k} v_{ik} e_{kt}$$

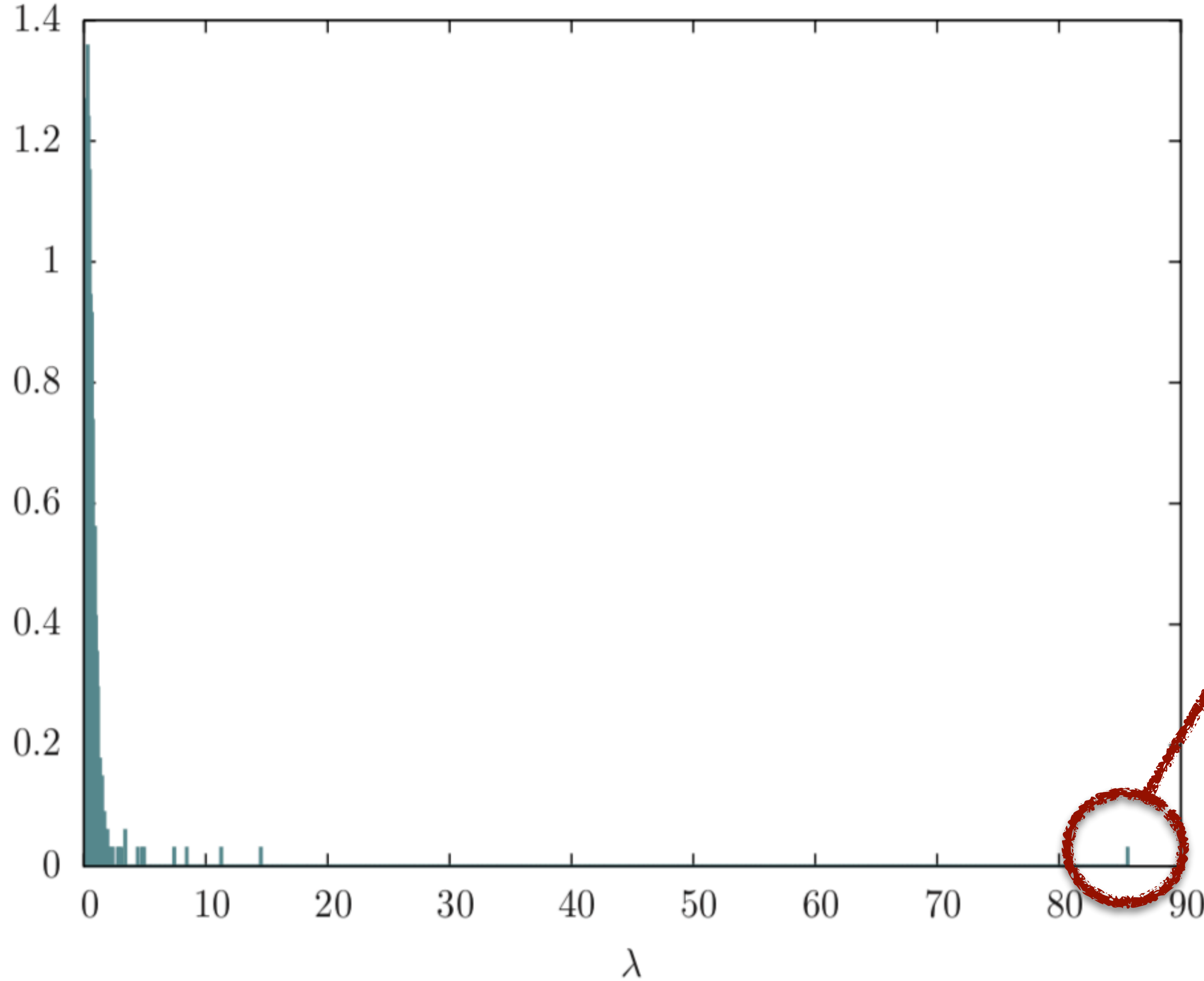
PCA - Practical use

PCs **“explain”** the variance of the data

$$1 = \mathrm{Var}[x_{it}] = \sum_{k=1}^{N} \lambda_k (v_{ki})^2$$

The presence of large eigenvalues means that the corresponding PCs account for a large fraction of the overall variance in the data

Example - S&P500 data



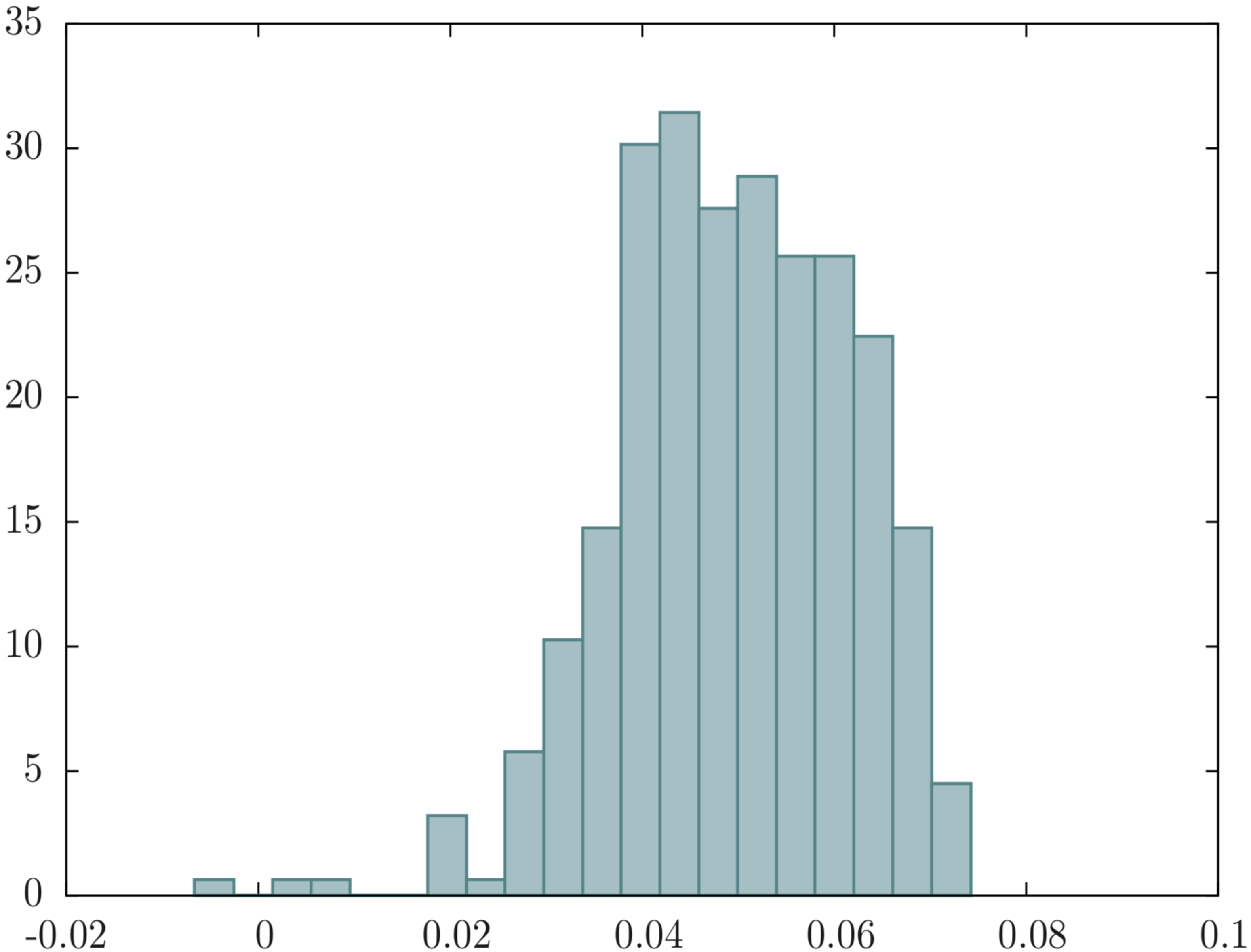
One huge and isolated eigenvalue

$$\lambda_1/N \approx 21\%$$

As a first approximation we can assume the whole market to be explained by the first PC:

$$x_{it} \approx \sqrt{\lambda_1} v_{1i} e_{1t}$$

The market mode

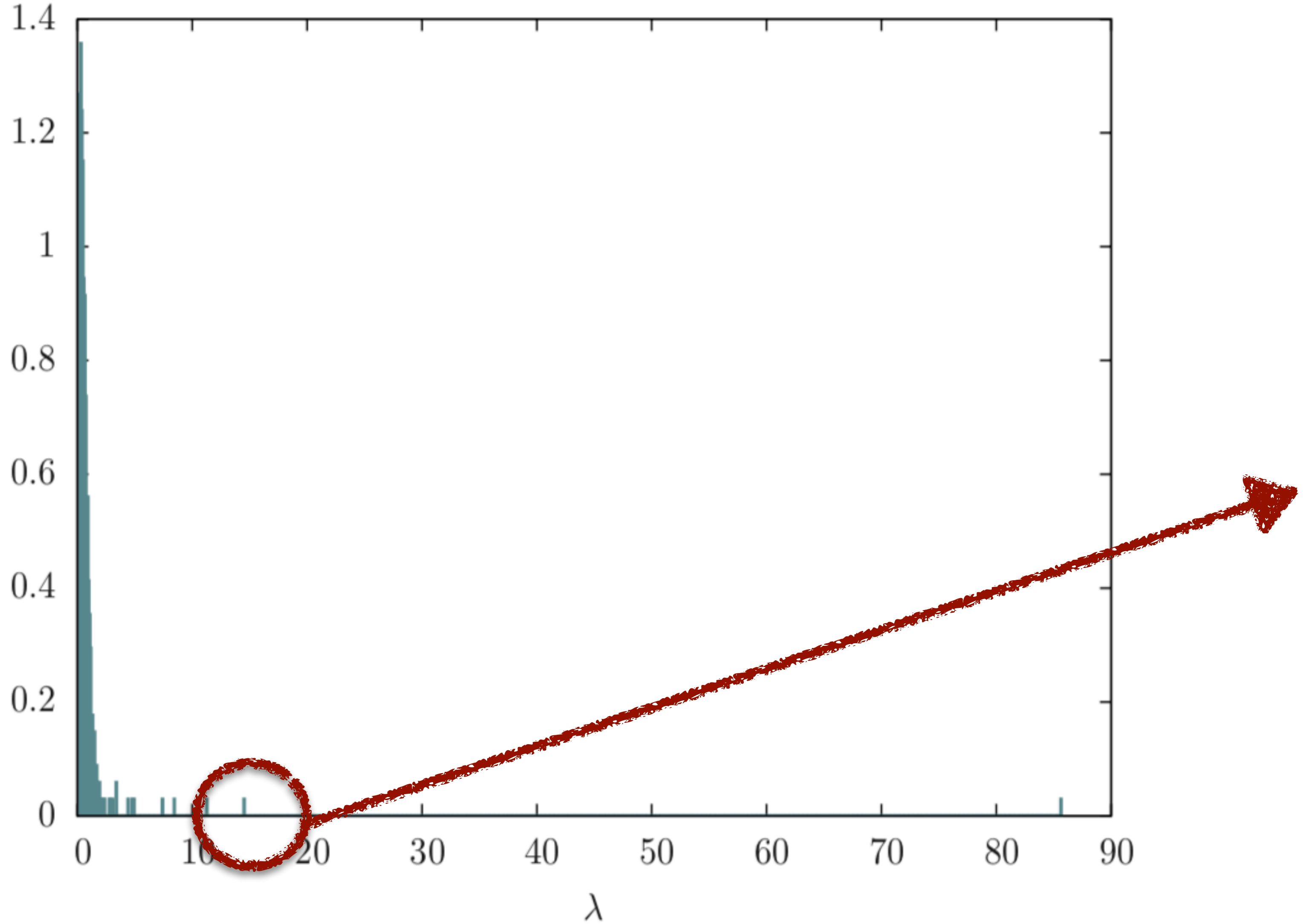


Almost all components share the same sign

Common factor driving all stocks

$$x_{it} \approx \sqrt{\lambda_1} v_{1i} e_{1t}$$

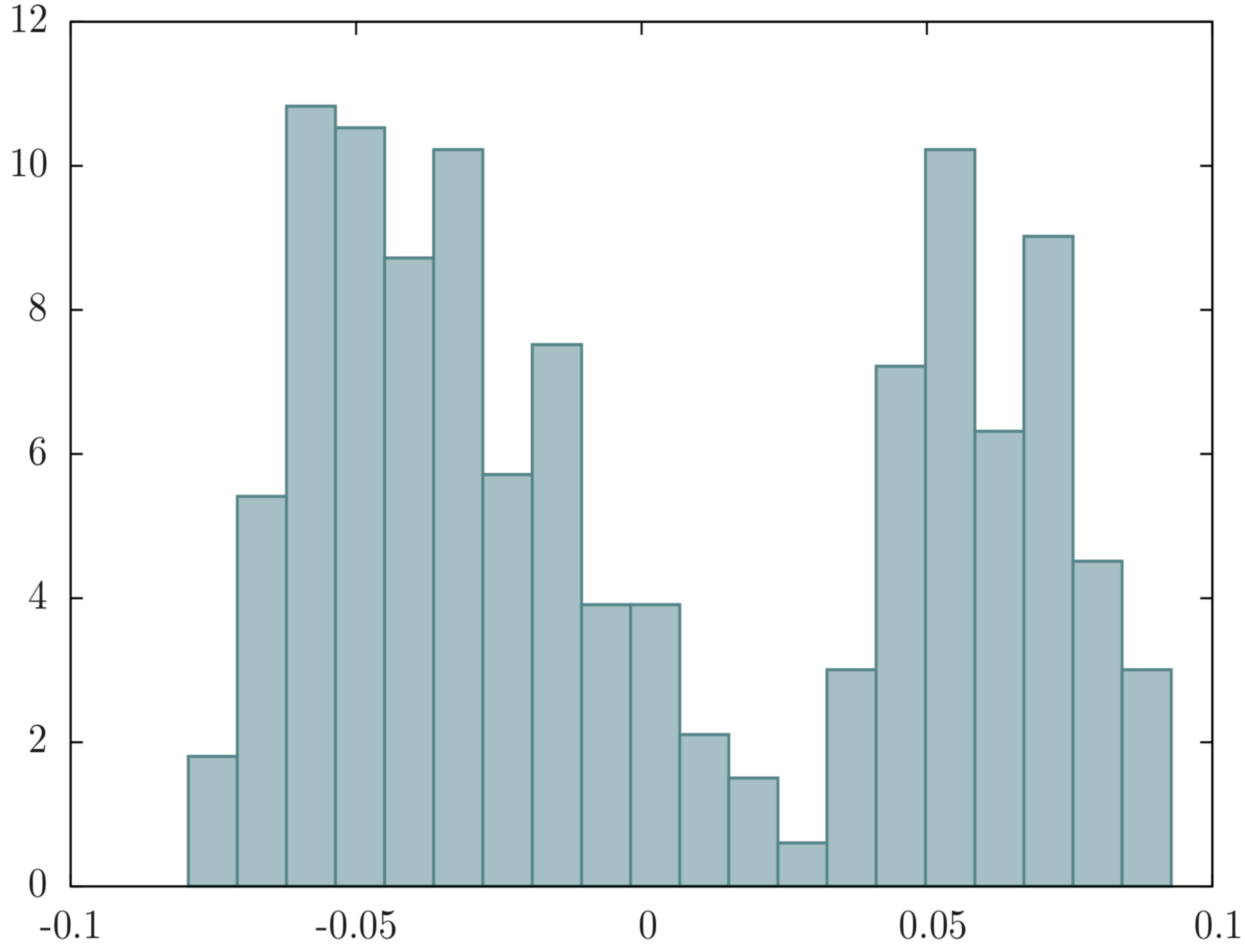
2nd eigenvalue



Considerably smaller than 1st one but still significantly larger than other eigenvalues

$$\lambda_2/N \approx 4\%$$

2nd eigenvector



Roughly 50-50 split between positive / negative components

Source of negative correlation between groups of stocks (typically belonging to different sectors)

$$x_{it} \approx \sqrt{\lambda_1} v_{1i} e_{1t} + \sqrt{\lambda_2} v_{2i} e_{2t}$$

Markowitz's optimal portfolio theory

Finding weights that minimise portfolio variance

Budget constraint

$$\sum_{i=1}^{N} w_i = 1$$

N assets, short selling allowed (weights can be negative)

Optimal weights

$$w_i^* = \frac{\sum_{j=1}^{N} (C^{-1})_{ij}}{\sum_{j,k=1}^{N} (C^{-1})_{jk}}$$

Dependence on inverse of correlation matrix

Fundamental concepts of Random Matrix Theory

A theory to **disentangle signal from noise** based on the spectral properties of matrices (not only correlation matrices)

In practice: measuring deviations from the spectral properties of an empirical correlation matrix and the average properties of a suitable ensemble of random matrices.

The Marcenko-Pastur distribution

Average eigenvalue density of a correlation matrix with "rectangularity ratio" $q = N/T$ (i.e., constructed from N time series of length T of **i.i.d. Gaussian random numbers** with zero mean and standard deviation σ)

$$p(\lambda) = \frac{1}{2\pi q\sigma^2}\frac{\sqrt{(\lambda-\lambda_-)(\lambda_+-\lambda)}}{\lambda}$$

$$\lambda_\pm = \sigma^2(1+\sqrt{q})^2$$

Porter-Thomas law

The eigenvector components of a random matrix follow a Gaussian distribution with zero mean and standard deviation $1/\sqrt{N}$

$$p(x) = \sqrt{\frac{N}{2\pi}}\,\exp\left(-\frac{x^2 N}{2}\right)$$