# Random Matrix Approach to analyzing the Big Data

Tung Dang

Laboratory of Biometrics and Bioinformatics,
University of Tokyo

**Corresponding author:**

dangthanhtung91@vn-bml.com

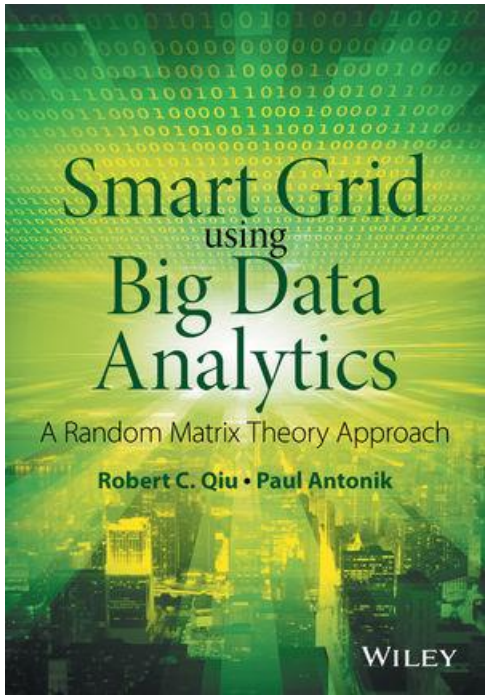
# Why are Random Matrices Cool?

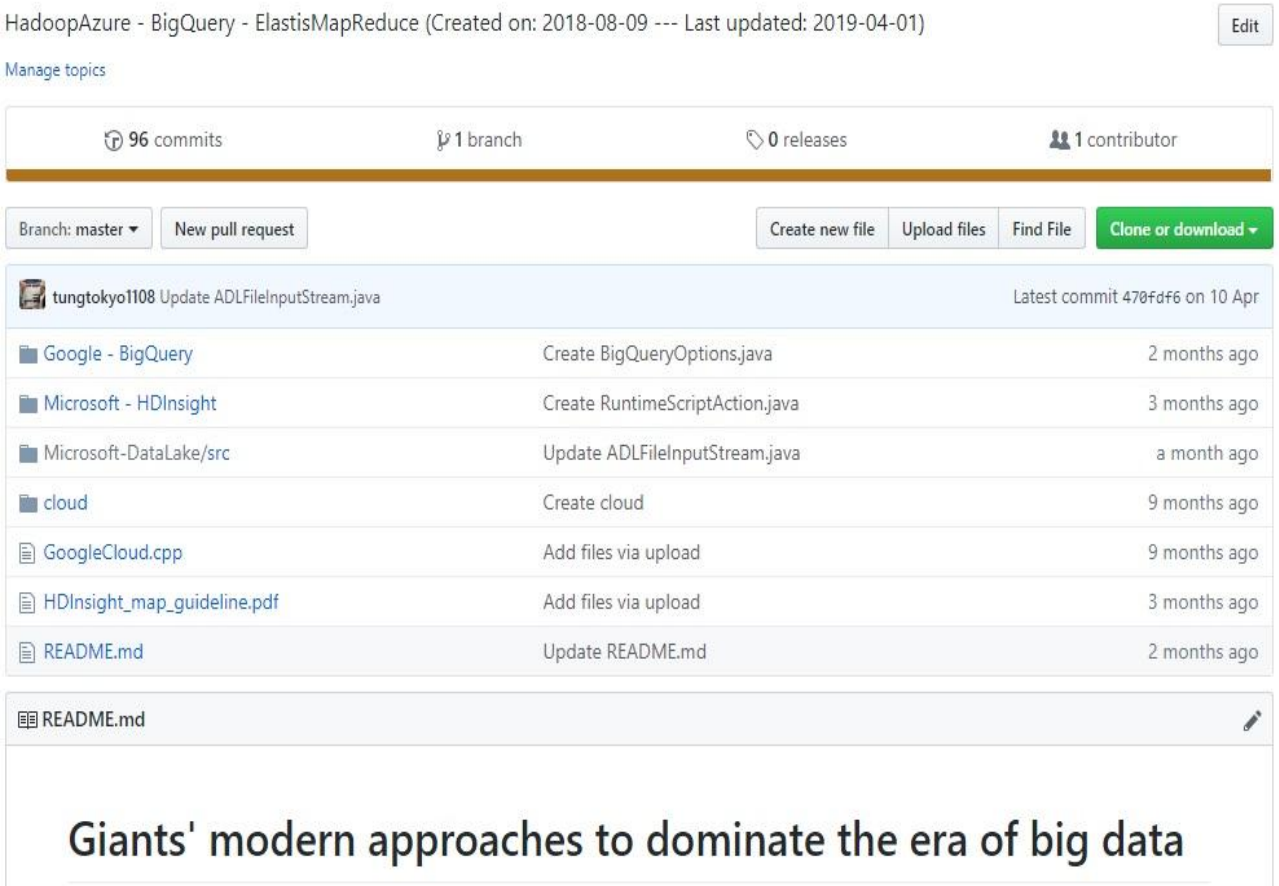

New Scientist (April 10 2010) cover story entitled *Entering the matrix: the simple idea at the root of reality*. Quoting Raj Rao Nadakuditi: "It really does feel like the ideas of random matrix theory are somehow buried deep in the heart of nature."

# Why are Random Matrices Cool?


Random matrix techniques in a recent study featured in the Wall Street Journal

# Why are Random Matrices Cool?



# Outline

- Foundational ideas
  - The Birth of Random Matrix Theory
  - The Wigner Matrix
  - The density of eigenvalues of Random Matrix
  - The Marchenko-Pastur law
- Application to real data
  - Random matrix theory to analyze noise dressing of financial data
  - Random matrix approach to cross correlations
  - Random matrix approach to elements of coevolution in biological sequences
- https://github.com/tungtokyo1108/My-Project--A-new-era-of-modern-analysis-for-Big-Data/blob/master/Technical_Revision/Random_Matrix_for_Big_Data/

# The Birth of Random Matrix Theory

- Pearson's methods to infer the distribution of standard deviations, in samples from a normal population,

$$dp = \frac{1}{\Gamma\left(\frac{N-1}{2}\right)} A^{\frac{N-1}{2}} . e^{-Aa} . a^{\frac{N-3}{2}} da \qquad A = \frac{N}{2\sigma^2}, \quad a = s^2,$$

$$N\bar{x} = \sum_1^N (x),$$

$$Ns^2 = \sum_1^N (x - \bar{x})^2.$$

- Bi-variate populations were considered,

$$A = \frac{N}{2\sigma_1^2 (1-\rho^2)}, \quad B = \frac{N}{2\sigma_2^2 (1-\rho^2)}, \quad H = -\frac{N\rho}{2\sigma_1\sigma_2 (1-\rho^2)},$$

$$a = s_1^2, \qquad b = s_2^2, \qquad h = s_1 s_2 r,$$

$$dp = \frac{1}{\sqrt{\pi}\,\Gamma\left(\frac{N-1}{2}\right)\Gamma\left(\frac{N-2}{2}\right)} \left| \begin{matrix} A & H \\ H & B \end{matrix} \right|^{\frac{N-1}{2}} . e^{-Aa-Bb-2Hh} . \left| \begin{matrix} a & h \\ h & b \end{matrix} \right|^{\frac{N-4}{2}} da\,db\,dh$$

The distribution of the correlation coefficient was deduced by direct integration

J. Wishart. (1928) Biometrika, vol. 20 A, pp. 32–52.

# The Birth of Random Matrix Theory

- Tri-variate populations were considered

$$dp = \frac{1}{(2\pi)^{\frac{3N}{2}} (\sigma_1\sigma_2\sigma_3)^N \Delta^{\frac{N}{2}}}$$

$$\times e^{-\frac{1}{2\Delta}\sum_1^N \left[\frac{(x-m_1)^2}{\sigma_1^2}\Delta_{11} + \frac{(y-m_2)^2}{\sigma_2^2}\Delta_{22} + \frac{(z-m_3)^2}{\sigma_3^2}\Delta_{33} + 2\frac{(y-m_2)(z-m_3)}{\sigma_2\sigma_3}\Delta_{23} + 2\frac{(z-m_3)(x-m_1)}{\sigma_3\sigma_1}\Delta_{31} + 2\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2}\Delta_{12}\right]}$$

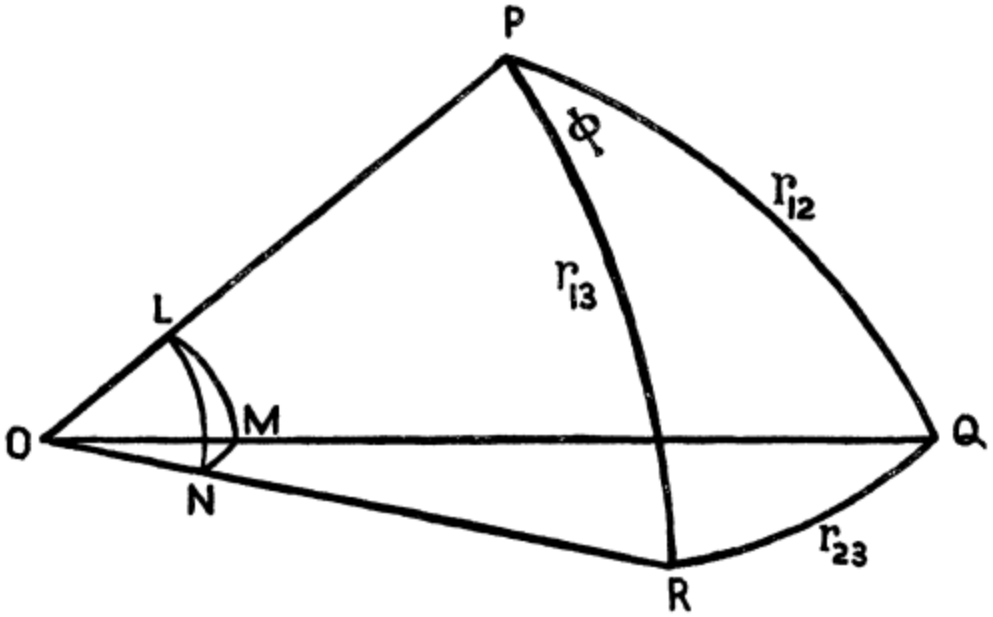
$$dx_1 dy_1 dz_1 dx_2 dy_2 dz_2 \ldots\ldots dx_N dy_N dz_N \ldots\ldots\ldots\ldots (5).$$

where $\Delta$ is the determinant $|\rho_{st}|$ $s, t = 1, 2, 3$, and $\Delta_{st}$ is the minor of $\rho_{st}$ in $\Delta$.

J. Wishart. (1928) Biometrika, vol. 20 A, pp. 32–52.

# The Birth of Random Matrix Theory

- The N values of x may be regarded geometrically as specifying a point P in an N-dimensional space, N values of y and N values of z specify points Q, R




$\theta_1$ being the angle $\frac{\pi}{2} - L\hat{O}M$

$$L\hat{O}N = \frac{\pi}{2} - \theta_2$$

- the points M and N do not vary independently

$$\cos\phi = r_{23\cdot 1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}}.$$

J. Wishart. (1928) Biometrika, vol. 20 A, pp. 32–52.

# The Birth of Random Matrix Theory

- The fundamental frequency distribution for the three variate case

$$dp = \frac{1}{\pi^{\frac{3}{2}}\,\Gamma\left(\frac{N-1}{2}\right)\Gamma\left(\frac{N-2}{2}\right)\Gamma\left(\frac{N-3}{2}\right)} \begin{vmatrix} A & H & G \\ H & B & F \\ G & F & C \end{vmatrix}^{\frac{N-1}{2}} . e^{-Aa-Bb-Cc-2Ff-2Gg-2Hh} \times \begin{vmatrix} a & h & g \\ h & b & f \\ g & f & c \end{vmatrix}^{\frac{N-5}{2}} dadbdcdfdgdh \quad \ldots\ldots\ldots\ldots(8).$$

$$A = \frac{N}{2\sigma_1^2}\cdot\frac{\Delta_{11}}{\Delta}, \quad B = \frac{N}{2\sigma_2^2}\cdot\frac{\Delta_{22}}{\Delta}, \quad C = \frac{N}{2\sigma_3^2}\cdot\frac{\Delta_{33}}{\Delta},$$

$$F = \frac{N}{2\sigma_2\sigma_3}\cdot\frac{\Delta_{23}}{\Delta}, \quad G = \frac{N}{2\sigma_3\sigma_1}\cdot\frac{\Delta_{31}}{\Delta}, \quad H = \frac{N}{2\sigma_1\sigma_2}\cdot\frac{\Delta_{12}}{\Delta},$$

J. Wishart. (1928) Biometrika, vol. 20 A, pp. 32–52.

# The Birth of Random Matrix Theory

- The fundamental frequency distribution for the Multi-variate case

$$dp = \frac{\left|\begin{matrix} A_{11} & A_{12} \dots A_{1n} \\ A_{21} & A_{22} \dots A_{2n} \\ \vdots & \vdots \quad\quad \vdots \\ A_{n1} & A_{n2} \dots A_{nn} \end{matrix}\right|^{\frac{N-1}{2}}}{(\sqrt{\pi})^{\frac{1}{2}n(n-1)}\,\Gamma\left(\frac{N-1}{2}\right)\Gamma\left(\frac{N-2}{2}\right)\dots\Gamma\left(\frac{N-n}{2}\right)}$$

$$\times e^{-A_{11}a_{11}-A_{22}a_{22}-\dots-A_{nn}a_{nn}-2A_{12}a_{12}-2A_{13}a_{13}-\dots-2A_{n-1\,n}a_{n-1\,n}}$$

$$\times \left|\begin{matrix} a_{11} & a_{12} \dots a_{1n} \\ a_{21} & a_{22} \dots a_{2n} \\ \vdots & \vdots \quad\quad \vdots \\ a_{n1} & a_{n2} \dots a_{nn} \end{matrix}\right|^{\frac{N-n-2}{2}} da_{11}da_{12}\dots da_{nn} \quad \dots\dots\dots\dots\dots\dots(9),$$

where $a_{pq} = s_p s_q r_{pq}$, and $A_{pq} = \frac{N}{2\sigma_p\sigma_q} \cdot \frac{\Delta_{pq}}{\Delta}$, $\Delta$ being the determinant

$$|\rho_{pq}|, p, q = 1, 2, 3, \dots n,$$

and $\Delta_{pq}$ the minor of $\rho_{pq}$ in $\Delta$.

J. Wishart. (1928) Biometrika, vol. 20 A, pp. 32–52.

# The Wigner Matrix

- Start with two independent families of i.i.d., zero mean, real-valued random variables $\{Z_{i,j}\}_{1\le i\le j}$ *and* $\{Y_i\}_{1\le i}$ such that $\mathrm{E}\left(Z_{1,2}^2\right)=1$ for all integers $k \ge 1$,

$$r_k := \max\left(E|Z_{1,2}|^k, E|Y_1|^k\right) < \infty$$

## Definition 1. Wigner matrix

The (symmetric) $N \times N$ matrix $X_N$ *is the Wigner matrix if its entries*

$$X_N(j,i) = X_N(i,j) = \begin{cases} Z_{i,j}/\sqrt{N}, & \text{if } i < j, \\ Y_i/\sqrt{N}, & \text{if } i = j. \end{cases}$$

E. Wigner. (1958) The Annals of Mathematics, vol. 67, pp. 325–327

# The Wigner's semicircle law

- $\lambda_i^N$ denote the (real) eigenvalues of $X_N$, with $\lambda_1^N \leq \lambda_2^N \leq \ldots \leq \lambda_N^N$ and define the *empirical distribution* of the eigenvalues as the (random) probability measure on R defined by

$$\mu(x) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i^N}$$

- **Theorem 1. Wigner's semicircle law**

*For a Wigner matrix, the empirical measure $\mu(x)$ converges weakly, in probability, to the standard semicircle distribution.*

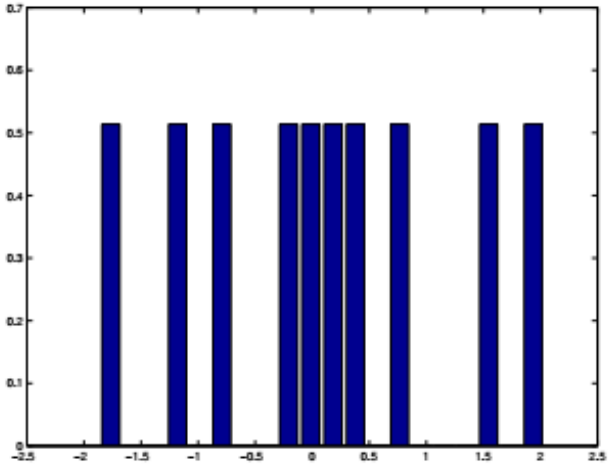
$$d\mu_{SC}(x) = \frac{1}{2\pi} \sqrt{4 - x^2}\, 1_{|x| \leq 2}\, dx$$

E. Wigner. (1958) The Annals of Mathematics, vol. 67, pp. 325–327

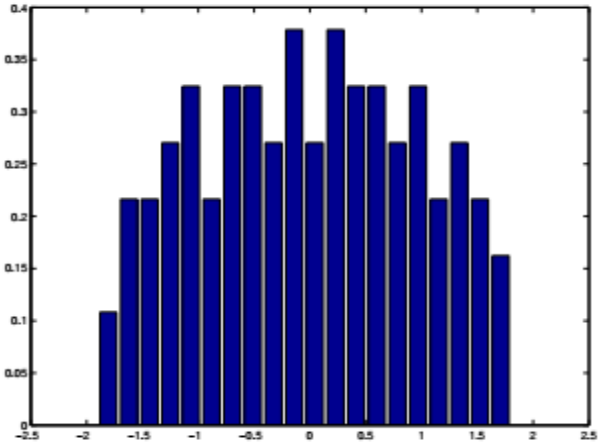
# The Wigner's semicircle law

- Here is one randomly generated 10 × 10 matrix and its eigenvalue histogram

$$\begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 \\ -1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \\ -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 \end{pmatrix}$$



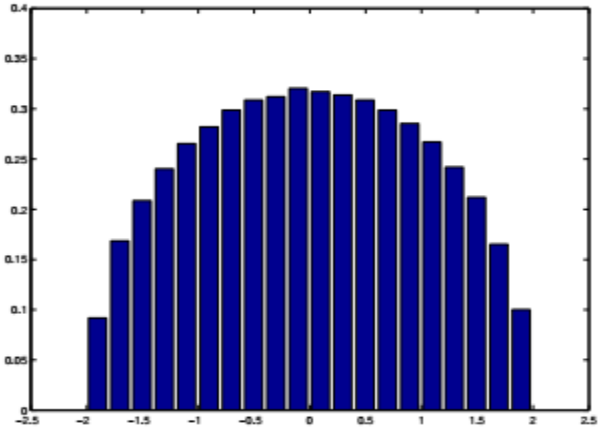
- The eigenvalue histograms 100 × 100 matrices



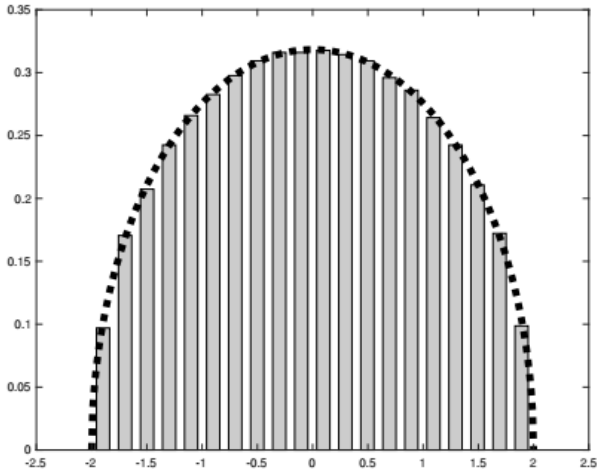
E. Wigner. (1958) The Annals of Mathematics, vol. 67, pp. 325–327

# The Wigner's semicircle law

- Here for two 3000 × 3000 matrices




- The eigenvalue distribution of such a random matrix converges to Wigner's semicircle for $N \to \infty$




E. Wigner. (1958) The Annals of Mathematics, vol. 67, pp. 325–327

# The density of eigenvalues of Random Matrix Theory

- $\left\{\xi_{i,j},\eta_{i,j}\right\}_{i,j=1}^{\infty} \sim \mathrm{N}(0,1)$
- Define $P_2^{(1)}, P_3^{(1)}, \ldots$ to be the laws of the random matrices

$$\begin{bmatrix} \sqrt{2}\xi_{1,1} & \xi_{1,2} \\ \xi_{1,2} & \sqrt{2}\xi_{2,2} \end{bmatrix} \in \mathscr{H}_2^{(1)}, \begin{bmatrix} \sqrt{2}\xi_{1,1} & \xi_{1,2} & \xi_{1,3} \\ \xi_{1,2} & \sqrt{2}\xi_{2,2} & \xi_{2,3} \\ \xi_{1,3} & \xi_{2,3} & \sqrt{2}\xi_{3,3} \end{bmatrix} \in \mathscr{H}_3^{(1)}, \ldots,$$

- Define $P_2^{(2)}, P_3^{(2)}, \ldots$ to be the laws of the random matrices

$$\begin{bmatrix} \xi_{1,1} & \frac{\xi_{1,2}+i\eta_{1,2}}{\sqrt{2}} \\ \frac{\xi_{1,2}-i\eta_{1,2}}{\sqrt{2}} & \xi_{2,2} \end{bmatrix} \in \mathscr{H}_2^{(2)}, \begin{bmatrix} \xi_{11} & \frac{\xi_{1,2}+i\eta_{1,2}}{\sqrt{2}} & \frac{\xi_{1,3}+i\eta_{1,3}}{\sqrt{2}} \\ \frac{\xi_{1,2}-i\eta_{1,2}}{\sqrt{2}} & \xi_{2,2} & \frac{\xi_{2,3}+i\eta_{2,3}}{\sqrt{2}} \\ \frac{\xi_{1,3}-i\eta_{1,3}}{\sqrt{2}} & \frac{\xi_{2,3}-i\eta_{2,3}}{\sqrt{2}} & \xi_{3,3} \end{bmatrix} \in \mathscr{H}_3^{(2)}, \ldots,$$

M. L. Mehta and M. Gaudin. (1960) Nuclear Physics, vol. 18, pp. 420–427.

# The density of eigenvalues of Random Matrix Theory

**Definition 2. Gaussian orthogonal ensemble-GOE or the Gaussian unitary ensemble-GUE**

A random matrix $X \in \mathscr{H}_N^{(\beta)}$ with the law $P_N^{(\beta)}$ is said to belong to the *Gaussian orthogonal ensemble (GOE)* or the *Gaussian unitary ensemble (GUE)* according as β = 1 or β = 2, respectively.

- Let calculate the density of $P_N^{(\beta)}$ with respect to Lebesgue measure $\ell_N^{(\beta)}$
- Let *Hi,j* denote the entry of $H \in \mathscr{H}_N^{(\beta)}$ in row *i* and column *j*.

$$tr\left(H^2\right) = tr\left(HH^T\right) = \sum_{i=1}^{N} H_{i,i}^2 + 2\sum_{1\le i<j\le N} H_{i,j}^2$$

M. L. Mehta and M. Gaudin. (1960) Nuclear Physics, vol. 18, pp. 420–427.

# The density of eigenvalues of Random Matrix Theory

## Lemma 1. The joint distribution of the elements $H_{ij}$

$$\frac{dP_N^{(\beta)}}{d\ell_N^{(\beta)}}(H) = \begin{cases} 2^{-N/2}(2\pi)^{-N(N+1)/4}\exp(-\mathrm{tr}H^2/4) & \text{if } \beta = 1, \\ 2^{-N/2}\pi^{-N^2/2}\exp(-\mathrm{tr}H^2/2) & \text{if } \beta = 2. \end{cases}$$

## Lemma 2.

Let a random matrix $X \in \mathscr{H}_N^{(\beta)}$ with law $P_N^{(\beta)}$ then for any non-random $N \times N$ orthogonal matrix ($\beta = 1$) or non-random $N \times N$ unitary matrix ($\beta = 2$) U, again $UXU^T$ has law $P_N^{(\beta)}$

M. L. Mehta and M. Gaudin. (1960) Nuclear Physics, vol. 18, pp. 420–427.

# The density of eigenvalues of Random Matrix Theory

## Theorem 2. Joint distribution of eigenvalues: GOE and GUE

Let $X \in \mathscr{H}_N^{(\beta)}$ *random with law* $P_N^{(\beta)}$ with β = 1 or β = 2. *The joint distribution of the eigenvalues* $\lambda_1(X) \le \lambda_2(X) \le \ldots \le \lambda_N(X)$ *has density with respect to Lebesgue measure which equals*

$$N!\overline{C}_N^{(\beta)} \prod_{1 \le i < j \le N} |\lambda_i - \lambda_j|^\beta \prod_{i=1}^N e^{-\beta \frac{\lambda_i^2}{4}}$$

$$N!\overline{C}_N^{(\beta)} = N! \left( \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \prod_{1 \le i < j \le N} |\lambda_i - \lambda_j|^\beta \prod_{i=1}^N e^{-\beta \frac{\lambda_i^2}{4}} \right)^{-1}$$

$$= (2\pi)^{-N/2} \left( \frac{\beta}{2} \right)^{\beta N(N-1)/4 + N/2} \prod_{j=1}^N \frac{\Gamma\left(\beta/2\right)}{\Gamma\left(j\beta/2\right)} \qquad \Gamma(s) = \int_0^\infty x^{s-1} e^{-x} dx$$

M. L. Mehta and M. Gaudin. (1960) Nuclear Physics, vol. 18, pp. 420–427.

# The density of eigenvalues of Random Matrix Theory

**Lemma 3. Distribution of nearest-neighbor eigenvalue spacings (Wigner surmise)**

Consider the eigenvalue spacing distribution, which reflects two point as well as eigenvalue correlation functions of all orders. For GOE matrices, the distribution of ‘‘nearest-neighbor’’ eigenvalue spacing $s \equiv \lambda_{k+1}(X) - \lambda_k(X)$

$$P_{GOE}(s) = \frac{\pi s}{2} \exp\left(-\frac{\pi}{4} s^2\right)$$

M. L. Mehta and M. Gaudin. (1960) Nuclear Physics, vol. 18, pp. 420–427.

# The Marchenko-Pastur law

- $X_1,\ldots,X_p \in \mathbb{R}^N; X_i = \left(X_{1,i},\ldots,X_{N,i}\right)^T$ be independently identically distributed such that

$$E\left[X_i\right] = \begin{pmatrix} E\left[X_{1,i}\right] \\ \vdots \\ E\left[X_{N,i}\right] \end{pmatrix} = 0 \qquad E\left[X_i X_i^T\right] = \Sigma_N$$

- Consider the matrix

$$B_N = B_{N,p} = \frac{1}{p}\sum_{k=1}^{p} X_k X_k^T$$

Law of large numbers with N is fixed

$$\lim_{p\to\infty} B_N = \lim_{p\to\infty} \frac{1}{p}\sum_{k=1}^{p} X_k X_k^T = \Sigma_N$$

- Then a natural question to ask is what would be the behavior of $B_N$ when both $N$ and $p$ go to infinity?

V. A. Marcenko and L. A. Pastur (1967) Math USSR-Sbornik, vol. 1, pp. 457–483.

# The Marchenko-Pastur law

- A **Wigner matrix** $B_{N,p}$ is an $N \times N$ matrix of the form

$$B_{N,p} = \frac{1}{p} X_{N,p} X_{N,p}^T = \frac{1}{p} \sum_{k=1}^{p} X_k X_k^T$$

- where $X_{N,p}$ is an $N \times p$ matrix with independently identically distributed centered entries of variance 1.
- The empirical spectral measure with $\lambda_1, \ldots, \lambda_N$ the eigenvalues of $B_N$ is given by

$$\mu_{B_N} = \frac{1}{N} \sum_{k=1}^{N} \delta_{\lambda_k}$$

- Let consider

$$c = \frac{N}{p} \xrightarrow{N, p \to \infty} c \in (0, \infty)$$

V. A. Marcenko and L. A. Pastur (1967) Math USSR-Sbornik, vol. 1, pp. 457–483.

# The Marchenko-Pastur law

- **Theorem 3. The Marchenko-Pastur Law**

*Let* $(X_{ij})_{i,j}$ *be a family of independently identically distributed random variables such that* $E[X_{11}]=0; E[X_{11}^2]=\sigma^2<\infty$. *Provided that* $N, p \to \infty$ *such that* $\frac{N}{p} \to c \in (0,\infty)$

$$\mu_{B_N} \to \mu_{MP}$$

*whose density is given by*

$$\left(1-\frac{1}{c}\right)_+ \delta_0 + \frac{1}{2\pi c\sigma^2 x}\sqrt{(\lambda^+ - x)(x-\lambda^-)}\,1_{[\lambda^-,\lambda^+]}(x)\,dx$$

*where* $(\cdot)_+ = \max(0, \cdot)$ *and* $\lambda^{\pm} = \sigma^2\left(1 \pm \sqrt{c}\right)^2$

V. A. Marcenko and L. A. Pastur (1967) Math USSR-Sbornik, vol. 1, pp. 457–483.

# The Marchenko-Pastur law

- **Remark**
  - Observe that the non-zero eigenvalues of $XX^T$ and $X^T X$ are the same so that we have $\mu = \left(1 - \frac{1}{c}\right)\delta_0 + \widetilde{\mu}$

    $\widetilde{\mu}$ is the limiting distribution of $X^T X$
  - Apart from the Dirac measure at 0, the support of $\mu_{MP}$ is compact and is spread on an interval of length $4\sigma^2\sqrt{c}$ around the variance
  - The Marchenko-Pastur theorem is a universality result in the sense that the limiting distribution depends on the distribution of the entries only through the variance
  - The mean and variance of the Marchenko-Pastur distribution are

$$\int_{\mathbb{R}} x d\mu_{MP}(x) = \sigma^2$$

$$\int_{\mathbb{R}} x^2 d\mu_{MP}(x) - \left(\int_{\mathbb{R}} x d\mu_{MP}(x)\right)^2 = \frac{\sigma^4}{c}$$

V. A. Marcenko and L. A. Pastur (1967) Math USSR-Sbornik, vol. 1, pp. 457–483.

# Random matrix theory to analyze noise dressing of financial data

- Study numerically the density of eigenvalues of the correlation matrix of N = 406 assets of the S&P 500, based on daily variations during the years 1991–1996, for a total of T = 1309 days.
- The price $S_i(t)$ of asset $i$ at time $t$
- The average return $R_P$ of a portfolio $P$ of $N$ assets

$$R_P = \sum_{i=1}^{N} p_i R_i$$

$p_i$ is the amount of capital invested in the asset $i$
$R_i$ are the expected returns of the individual assets.

- The empirical correlation matrix **C** is constructed from the time series of price changes $\delta x_i(t)$

$$C_{ij} = \frac{1}{T} MM^T = \frac{1}{T} \sum_{t=1}^{T} \delta x_i(t) \delta x_j(t)$$

**M** is a $N$ x $T$ rectangular matrix

L. Laloux et al. (1999) Phys. Rev. Lett. 83, 1467.

# Random matrix theory to analyze noise dressing of financial data

- The density of eigenvalues of **C**

$$P_C(\lambda) = \frac{1}{N}\frac{dn(\lambda)}{d\lambda} \xrightarrow[Q = T/N \geq 1]{N \to \infty, T \to \infty} P_{RM}(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda}$$

$$\lambda_{\min}^{\max} = \sigma^2\left(1 + 1/Q \pm 2\sqrt{1/Q}\right), \lambda \in [\lambda_{\min}, \lambda_{\max}]$$

$\sigma^2$ is equal to the variance of the elements of **M**, equal to 1 with our normalization.

- The most important features predicted by density of eigenvalues
  - The fact that the lower "edge" of the spectrum is strictly positive (except for Q=1); there is therefore no eigenvalues $[0, \lambda_{\min}]$. Near this edge, the density of eigenvalues exhibits a sharp maximum, except in the limit Q=1, where it diverges as $1/\lambda$
  - The density of eigenvalues also vanishes above a certain upper edge $\lambda_{\max}$

L. Laloux et al. (1999) Phys. Rev. Lett. 83, 1467.

# Random matrix theory to analyze noise dressing of financial data

- The low lying eigenvalues are essentially random can also be tested by studying the statistical structure of the corresponding eigenvectors
- The *i*th component of the eigenvector $\upsilon_{\alpha,i}$ corresponding to the eigenvalue $\lambda_\alpha$
- Porter-Thomas distribution in the theory of random matrices

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

$$\sum_{i=1}^{N} \upsilon_{\alpha,\mathrm{i}}^2 = N$$

$$u = \upsilon_{\alpha,i}$$

$\alpha$ is fixed, i is varied

L. Laloux et al. (1999) Phys. Rev. Lett. 83, 1467.

# Random matrix theory to analyze noise dressing of financial data

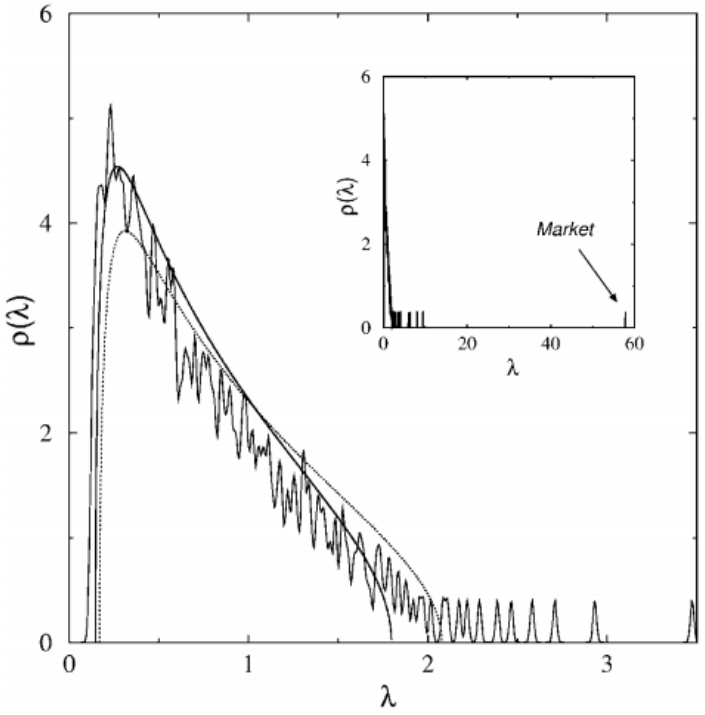



FIG. 1. Smoothed density of the eigenvalues of **C**, where the correlation matrix **C** is extracted from $N = 406$ assets of the S&P 500 during the years 1991–1996. For comparison we have plotted the density Eq. (3) for $Q = 3.22$ and $\sigma^2 = 0.85$: this is the theoretical value obtained assuming that the matrix is purely random except for its highest eigenvalue (dotted line). A better fit can be obtained with a smaller value of $\sigma^2 = 0.74$ (solid line), corresponding to 74% of the total variance. Inset: Same plot, but including the highest eigenvalue corresponding to the market, which is found to be 25 times greater than $\lambda_{max}$.

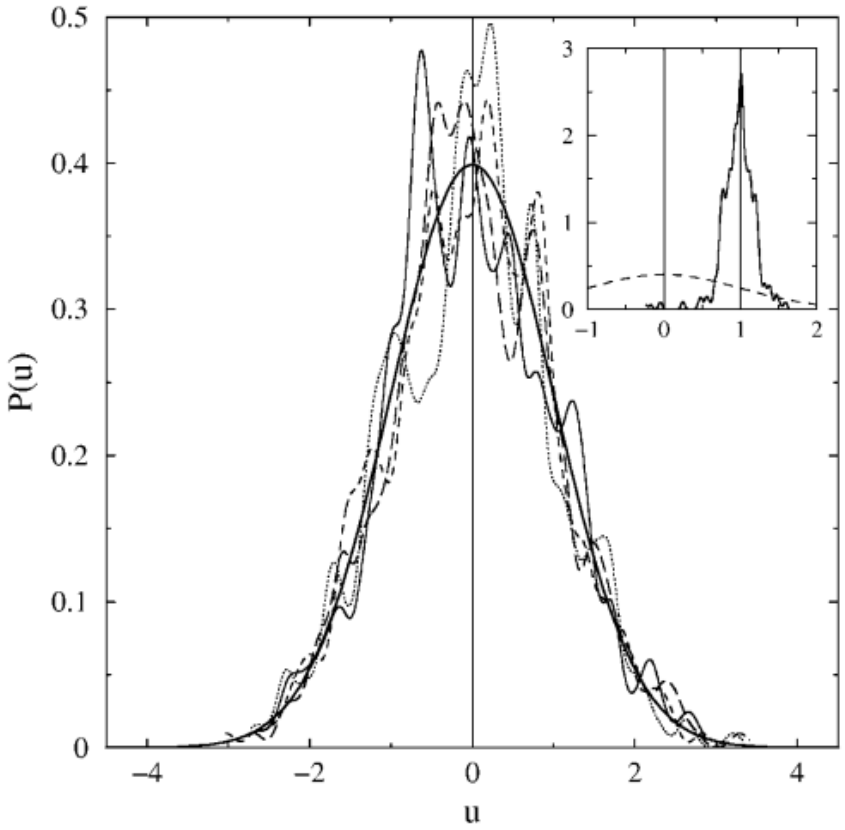



FIG. 2. Distribution of the eigenvector components $u = v_{\alpha,i}$, for five different eigenvectors well inside the interval $[\lambda_{min}, \lambda_{max}]$, and comparison with the no information assumption, Eq. (4). Note that there are *no* adjustable parameters. Inset: Plot of the same quantity for the highest eigenvalue, showing marked differences with the theoretical prediction (dashed line), which is indeed expected.

L. Laloux et al. (1999) Phys. Rev. Lett. 83, 1467.

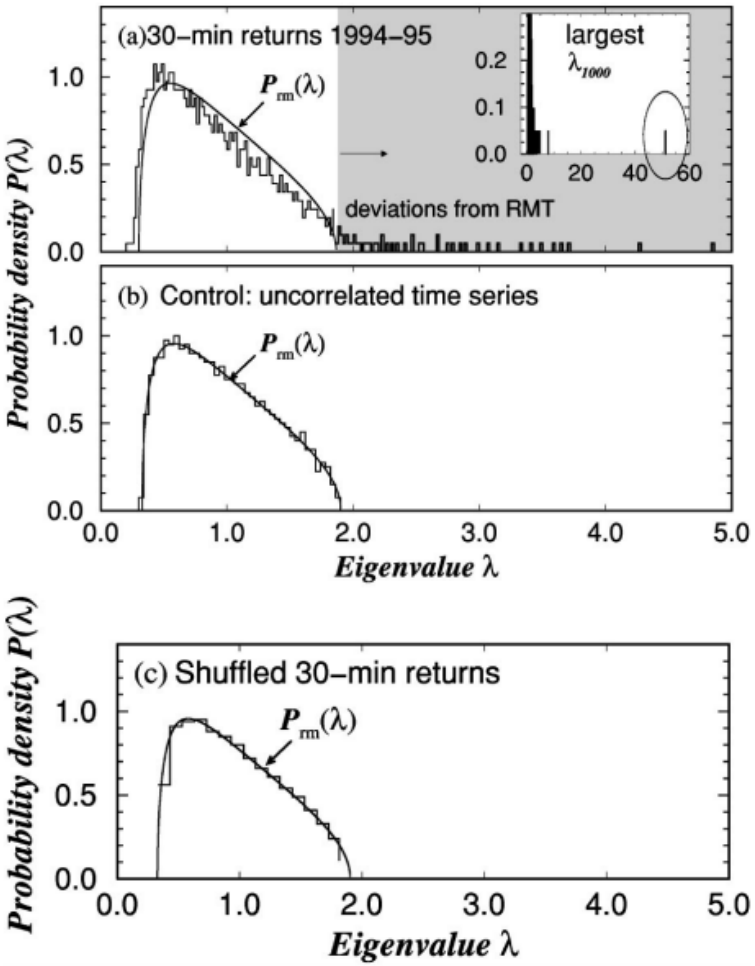
# Random matrix approach to cross correlations

- ***Database I*. The Trades and Quotes (TAQ) database**
  - The database forms $T = 6448$ records of 30-min returns of $N = 1000$ US stocks for the 2-yr period 1994–1995.
  - Analyze the prices of a subset comprising 881 stocks of those 1000 we analyze for 1994–1995. This data extract $T = 6448$ records of 30-min returns of $N = 881$ US stocks for the 2-yr period 1996–1997.

- ***Database II*. The Center for Research in Security Prices (CRSP) database**
  - Analyze daily returns for the stocks that survive for the 35-yr period 1962–1996 and extract $T = 8685$ records of 1-day returns for $N = 422$ stocks.

V. Plerou et al. (2002) Phys. Rev. E. 65, 066126-1–066126-18.

# Random matrix approach to cross correlations



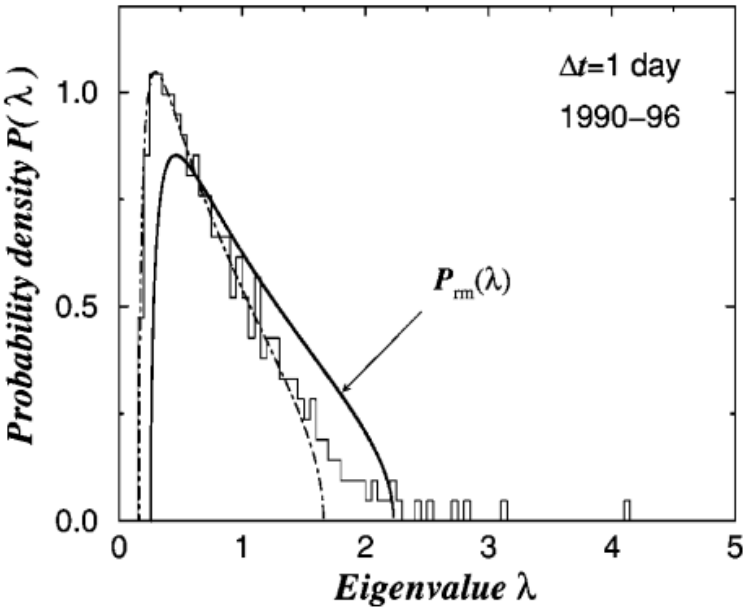




FIG. 4. $P(\lambda)$ for C constructed from daily returns of 422 stocks for the 7-yr period 1990–1996. The solid curve shows the RMT result $P_{rm}(\lambda)$ of Eq. (6]) using $N=422$ and $L=1737$. The dot-dashed curve shows a fit to $P(\lambda)$ using $P_{rm}(\lambda)$ with $\lambda_+$ and $\lambda_-$ as free parameters. We find similar results as found in Fig. 3(a) for 30-min returns. The largest eigenvalue (not shown) has the value $\lambda_{422}=46.3$.

V. Plerou et al. (2002) Phys. Rev. E. 65, 066126-1–066126-18.

# Random matrix approach to cross correlations

- **Problems**
  - The presence of a well-defined ‘‘bulk’’ of eigenvalues which fall within the bounds $\lambda_{\min}^{\max} = \sigma^2\left(1+1/Q \pm 2\sqrt{1/Q}\right), \lambda \in [\lambda_{\min}, \lambda_{\max}]$
  - Note deviations for a few (=20) largest and smallest eigenvalues. In particular, the largest eigenvalue $\lambda_{1000} \approx 50$ for the 2-yr period, which is 25 times larger than $\lambda_{\max} = 1.94$
  - Having demonstrated that the eigenvalue statistics satisfy the RMT predictions, we now proceed to analyze the eigenvectors. RMT predicts that the components of the normalized eigenvectors of a GOE matrix are distributed according to a Gaussian probability distribution with mean zero and unit variance.

$$\rho_{\mathrm{rm}}(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right).$$

V. Plerou et al. (2002) Phys. Rev. E. 65, 066126-1–066126-18.

# Random matrix approach to cross correlations

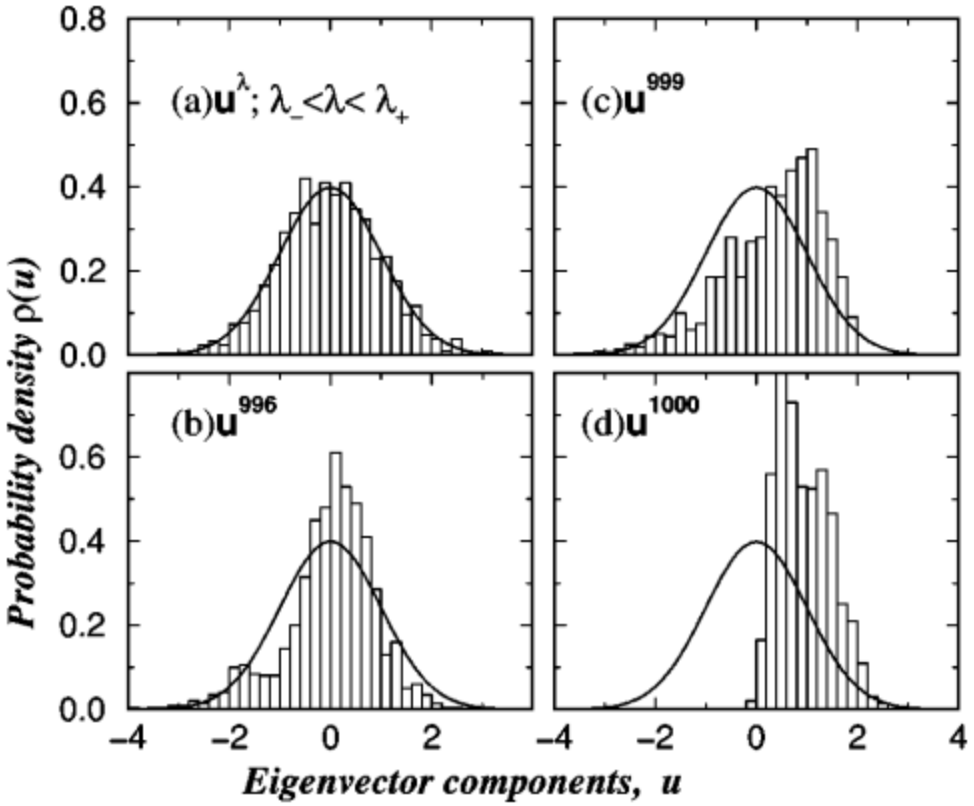



FIG. 8. (a) Distribution $\rho(u)$ of eigenvector components for one eigenvalue in the bulk $\lambda_- < \lambda < \lambda_+$ shows good agreement with the RMT prediction of Eq. (17) (solid curve). Similar results are obtained for other eigenvalues in the bulk. $\rho(u)$ for (b) $u^{996}$ and (c) $u^{999}$, corresponding to eigenvalues larger than the RMT upper bound $\lambda_+$ (shaded region in Fig. 3). (d) $\rho(u)$ for $u^{1000}$ deviates significantly from the Gaussian prediction of RMT. The above plots are for C constructed from 30-min returns for the 2-yr period 1994–1995. We also obtain similar results for C constructed from daily returns.

Eigenvalues and eigenvectors, which deviates significantly from the distribution predicted by RMT, have been useful to detect the changes of market

V. Plerou et al. (2002) Phys. Rev. E. 65, 066126-1–066126-18.

# Random matrix approach to cross correlations

- The largest eigenvalue and the corresponding eigenvector fall out the prediction of RMT
  - The largest eigenvalue and its corresponding eigenvector as the collective ‘‘response’’ of the entire market to stimuli.
  - Compare the projection (scalar product) of the time series G on the eigenvector u1000, with a standard measure of US stock market performance—the returns $G_{SP}(t)$ of the S&P 500 index. The price $S_i(t)$ of stock $i$ at time $t$

$$G^{1000}(t) \equiv \sum_{j=1}^{1000} u_j^{1000} G_j(t).$$

$$G_i(t) \equiv \ln S_i(t+\Delta t) - \ln S_i(t),$$

V. Plerou et al. (2002) Phys. Rev. E. 65, 066126-1–066126-18.

# Random matrix approach to cross correlations

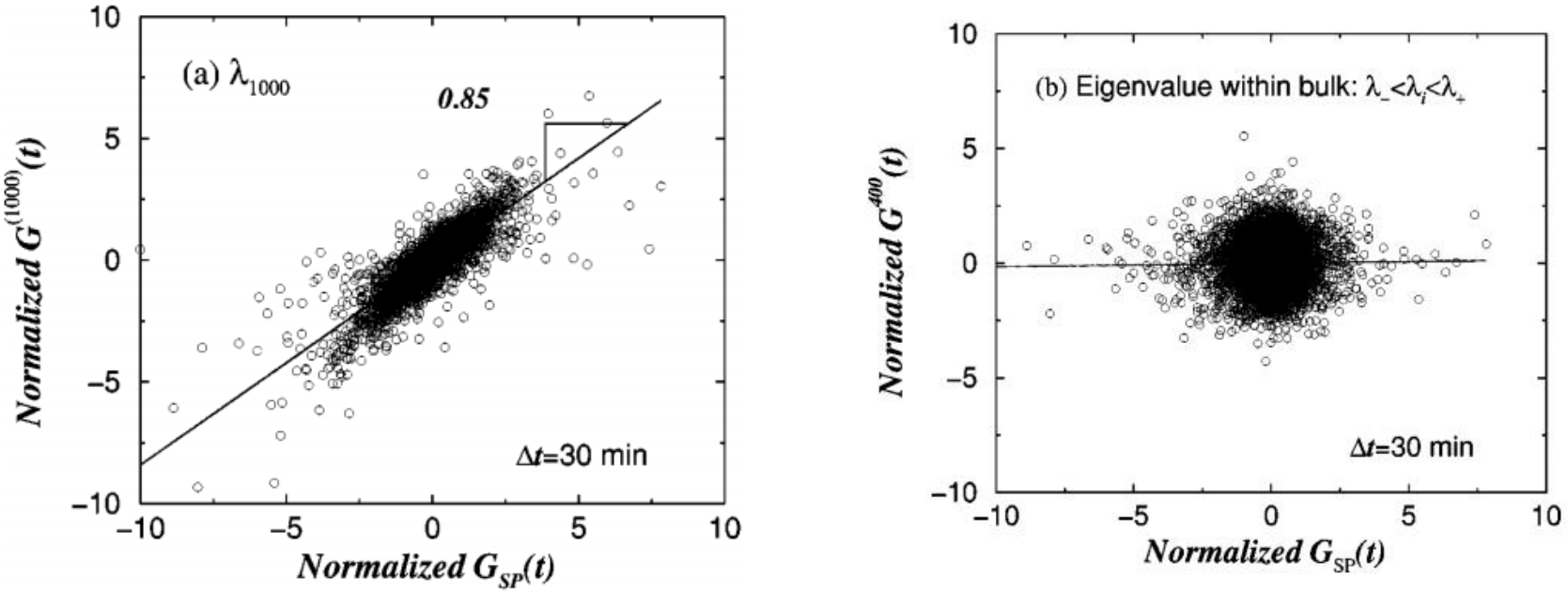



FIG. 9. (a) S&P 500 returns at $\Delta t = 30$ min regressed against the 30-min return on the portfolio $G^{1000}$ [Eq. (18)] defined by the eigenvector $u^{1000}$, for the 2-yr period 1994–1995. Both axes are scaled by their respective standard deviations. A linear regression yields a slope $0.85 \pm 0.09$. (b) Return (in units of standard deviations) on the portfolio defined by an eigenvector corresponding to an eigenvalue $\lambda_{400}$ within the RMT bounds regressed against the normalized returns of the S&P 500 index shows no significant dependence. Both axes are scaled by their respective standard deviations. The slope of the linear fit is $0.014 \pm 0.011$, close to 0.

V. Plerou et al. (2002) Phys. Rev. E. 65, 066126-1–066126-18.

# Random matrix approach to cross correlations

- Eigenvalue and corresponding eigenvector $U^{999}$ fall out the standard interval of RMT which contain all stocks with large values of market capitalization.

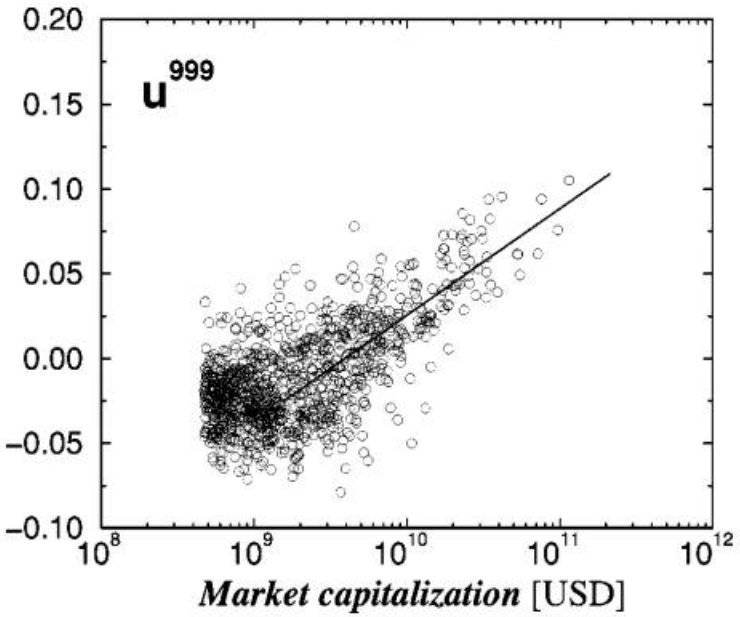



FIG. 12. All $10^3$ eigenvector components of $u^{999}$ plotted against market capitalization (in units of U.S. dollars) shows that firms with large market capitalization contribute significantly. The straight line, which shows a logarithmic fit, is a guide to the eye.

| Ticker | Industry | Industry code |
|---|---|---|
| | $u^{999}$ | |
| XON | Oil & gas equipment/services | 2911 |
| PG | Cleaning products | 2840 |
| JNJ | Drug manufacturers/major | 2834 |
| KO | Beverages-soft drinks | 2080 |
| PFE | Drug manufacturers/major | 2834 |
| BEL | Telecom services/domestic | 4813 |
| MOB | Oil & gas equipment/services | 2911 |
| BEN | Asset management | 6282 |
| UN | Food—major diversified | 2000 |
| AIG | Property/casualty insurance | 6331 |

V. Plerou et al. (2002) Phys. Rev. E. 65, 066126-1–066126-18.

# Random matrix approach to cross correlations

•

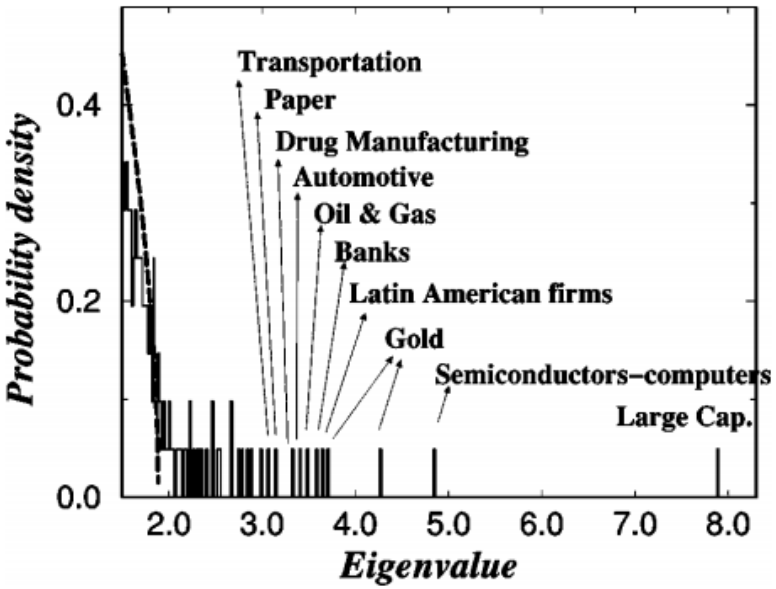



FIG. 13. Schematic illustration of the interpretation of the eigenvectors corresponding to the eigenvalues that deviate from the RMT upper bound. The dashed curve shows the RMT result of Eq. (6).

- Stocks of firms in the electronics and computer industry ($U^{998}$)
- A combination of gold mining and investment firms ($U^{996}$ and $U^{997}$)
- A mixture of three industry groups—telecommunications, metal mining, and banking ($U^{995}$)
- Banking firms ($U^{994}$)
- Oil and gas refining and equipment ($U^{993}$)
- Auto manufacturing firms ($U^{992}$)
- Drug manufacturing firms ($U^{991}$)
- Paper manufacturing ($U^{990}$)

V. Plerou et al. (2002) Phys. Rev. E. 65, 066126-1–066126-18.

# Random matrix approach to elements of coevolution in biological sequences

- There are some challenges that we apply standard statistical method to analyze the statistical correlations between pairs of positions in the alignment
  1. Proteins are gathered in an alignment based on sequence similarity, with no guarantee to have been subject to common selective constraints
  2. Sequences are not sampled independently during evolution but through a branching process, which introduces a sampling bias
  3. The information content of the alignment, $\sim ML\log_2 A \sim 10^5 - 10^7$ bits, is small compared to the number $\sim A^2 L^2/2 \sim 10^6 - 10^8$ of continuous parameters defining the correlations between every pair of amino acids, which implies a severe under-sampling. A = 20 natural amino acids is present at position i in sequence s; some positions contain a gap, inserted to ensure an optimal alignment and represented as a 21st amino acid. Typical numbers are $M=10^2$–$10^4$ for the number of sequences and $L=10^2 - 10^3$ for the length of the alignment.
  4. Two positions may be correlated while not directly interacting, reflecting a fundamental difference between interactions and correlations.

Rivoire O. (2013) Phys Rev Lett. 110(17):178102.

# Random matrix approach to elements of coevolution in biological sequences

- An array $x^a_{s,i}$ where *s* labels the sequences (row in the alignment), *i* the positions (columns) and *a* is a number between 1 and 20; $x^a_{s,i} = 1$ indicates that sequence *i* has amino acid *a* at position *i*, and $x^a_{s,i} = 0$ otherwise.
- The distance between two sequences is the fraction of amino acids by which they differ

$$S_{rs} = \frac{1}{L} \sum_{i=1}^{L} \sum_{a=0}^{20} \tilde{x}^a_{ri} \tilde{x}^a_{si}$$

- Sequence weights

$$w_s \equiv \frac{\nu_s^{-1}}{\sum_r \nu_r^{-1}}, \quad \text{with} \quad \nu_s \equiv |\{r : S_{rs} > \delta\}|$$

- A covariance matrix

$$C^{ab}_{ij} = f^{ab}_{ij} - f^a_i f^b_j.$$

$$f^a_i \equiv \sum_s w_s x^a_{si}, \qquad f^{ab}_{ij} \equiv \sum_s w_s x^a_{si} x^b_{sj}.$$

Rivoire O. (2013) Phys Rev Lett. 110(17):178102.

# Random matrix approach to elements of coevolution in biological sequences

- **Direct Coupling Analysis (DCA)** aims at identifying structural contacts between positions by inferring direct interactions from indirect correlations.
- A regularized covariance matrix

$$\bar{C}_{ij}^{ab} = \bar{f}_{ij}^{ab} - \bar{f}_i^a \bar{f}_j^b, \qquad \bar{f}_i^a = (1-\mu) f_i^a + \mu \frac{1}{A+1}, \qquad \bar{f}_{ij}^{ab} = (1-\mu) f_{ij}^{ab} + \mu \frac{1}{(A+1)^2}$$

- The inverse matrix $J = -C^{-1}$
- A model for the distribution of amino acids at every pair of positions *ij*

$$g_{ij}^{ab} = \exp(J_{ij}^{ab} + h_i^a + h_j^b + h_0)$$

where $h_i^a$, $h_j^b$, $h_0$ are uniquely determined by requiring that $\sum_b g_{ij}^{ab} = \bar{f}_i^a$

- A matrix of direct information

$$\mathcal{D}_{ij} = \sum_{a,b=0}^{A} g_{ij}^{ab} \ln \frac{g_{ij}^{ab}}{\bar{f}_i^a \bar{f}_j^b}$$

Rivoire O. (2013) Phys Rev Lett. 110(17):178102.

# Random matrix approach to elements of coevolution in biological sequences

- **Statistical Coupling Analysis (SCA)** aims at identifying groups of positions under selection for a common functional property
- Two principles:
  1. the conservation of amino acids involved in the function
  2. their correlations induced by cooperative interactions
- A measure of amino acid conservation

$$W_i^a = \left| \ln \left( \frac{f_i^a (1 - q^a)}{(1 - f_i^a) q^a} \right) \right|$$

$q^a = \sum_{i=1}^{L} f_i^a / L$ is the mean frequency of amino acid *a*

- A conservation-weighted correlation matrix

$$C_{ij} = \sqrt{\sum_{a,b=1}^{A} \left( W_i^a W_j^a C_{ij}^{ab} \right)^2}.$$

- For the SCA matrix $C_{ij}$, this analysis leads to coevolving units called protein sectors

Rivoire O. (2013) Phys Rev Lett. 110(17):178102.

# Random matrix approach to elements of coevolution in biological sequences

- Framework of computation
  1. Compute the eigenvectors associated with the top $k_{top}$ eigenvalues inspired by previous applications of **random matrix theory (RMT)** to the study of covariance matrices
  2. Rotate these eigenvectors into maximally independent components, $V^{(1)},\ldots,V^{(k_{top})}$ using **independent component analysis (ICA)**
  3. Define coevolving units as sets of positions making largest contributions to a component $S_k = \left\{ i : V_i^{(k)} > \varepsilon \right\}$
  4. The analysis involves two cutoffs:
     - the number $k_{top}$ of modes that is retained
     - threshold $\varepsilon > 0$ of significance for the contribution of positions to the components

Rivoire O. (2013) Phys Rev Lett. 110(17):178102.

# Random matrix approach to elements of coevolution in biological sequences

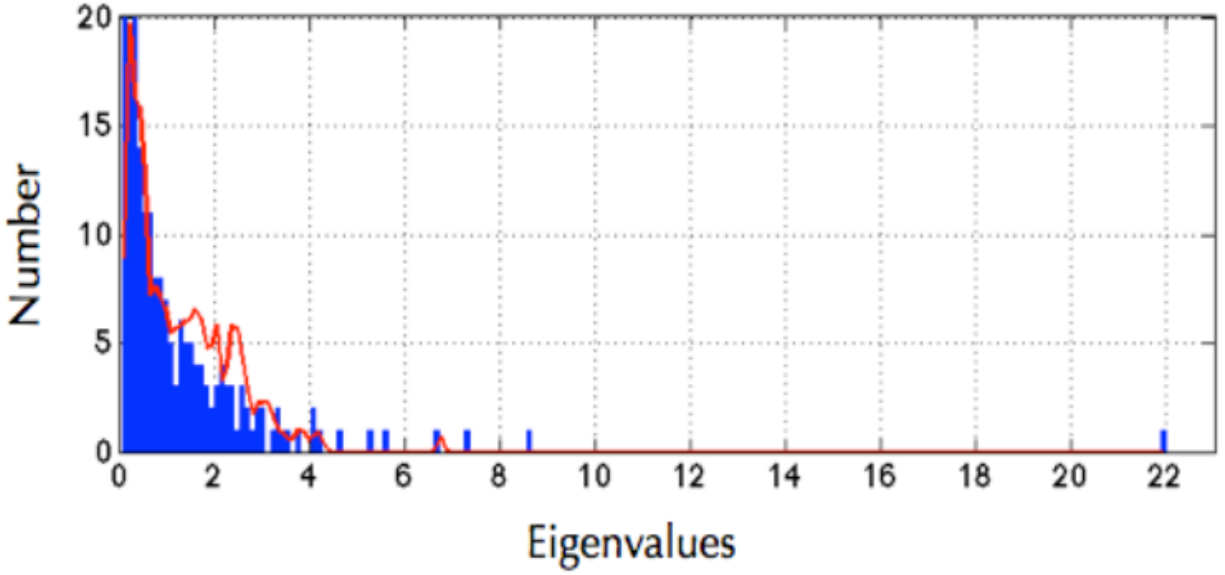



FIG. S 2: Spectrum of the SCA matrix $\mathcal{C}_{ij}$ – In blue, histogram of the $L$ eigenvalues of the matrix $\mathcal{C}_{ij}$ (truncated to 20 along the $y$-axis). In red, average spectrum over 100 randomized alignments, where the amino acids are drawn independently at each position $i$ according to the frequencies $f_i^a$. This shows that between 3 and 7 eigenvalues may considered as significant.

Rivoire O. (2013) Phys Rev Lett. 110(17):178102.

# Random matrix approach to elements of coevolution in biological sequences

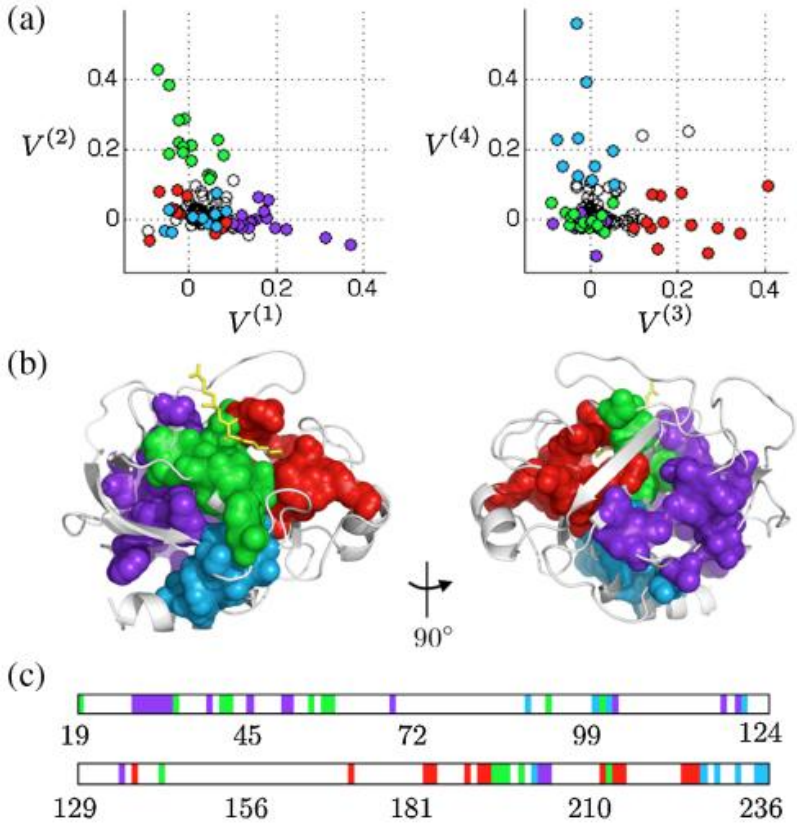



FIG. 1 (color online). Protein sectors in the trypsin family, as inferred from the Pfam alignment PF00089 [7]—(a) Projections of the positions $i$ along the vectors $V^{(k)}$ obtained by rotating by ICA the top $k_{\text{top}} = 4$ eigenvectors of the SCA matrix $\mathcal{C}_{ij}$ [2]: Each dot corresponds to a position $i$, with coordinates $(V_i^{(1)}, V_i^{(2)})$ in the first graph and $(V_i^{(3)}, V_i^{(4)})$ in the second. Sector $k$ is defined by the positions $i$ with $V_i^{(k)} > \epsilon$ and $V_i^{(\ell)} < \epsilon$ for $\ell \neq k$, with $\epsilon = 0.1$. The positions of each sector are represented with a different color: purple ($k = 1$), green ($k = 2$), red ($k = 3$), and cyan ($k = 4$). (b) Location of the sectors on a three-dimensional structure of trypsin [21]. (c) Location of the sectors along the sequence (cut in two for readability), with nonsector positions in white (numbering system of bovine chymotrypsin).

Top of eigenvalues and eigenvectors, which fall out the interval predicted by RMT, have been useful to detect the co-evolution elements

Rivoire O. (2013) Phys Rev Lett. 110(17):178102.

# Random matrix approach to elements of coevolution in biological sequences

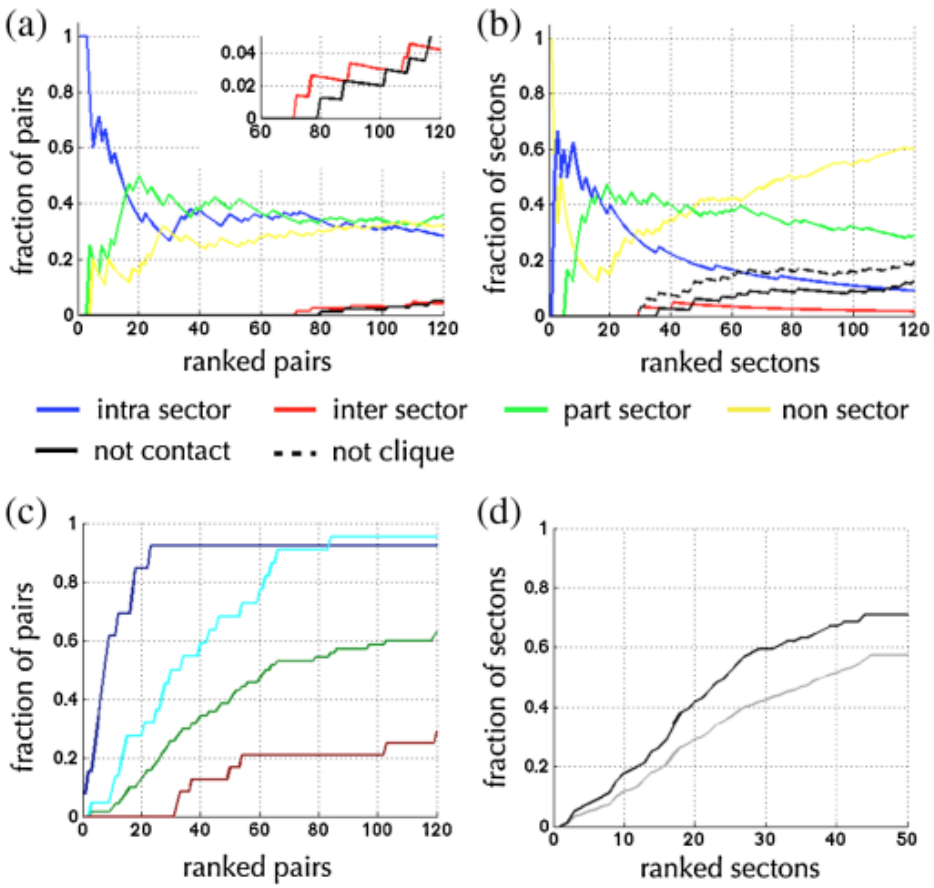



FIG. 2 (color online). Relations between top pairs of $\tilde{\mathcal{D}}_{ij}$, sectors and sectons—(a) Fraction of top pairs $ij$ of $\tilde{\mathcal{D}}_{ij}$, ranked by decreasing value of $\tilde{\mathcal{D}}_{ij}$, that are within a sector (blue curve), across two sectors (red), partly in a sector (green), and outside sectors (yellow). The fraction of pairs not in contact (black) becomes nonzero at rank 80 (zoom in inset). (b) Similar to (a), but for sectons instead of top pairs, and with an extra curve (dotted line) for the fraction of sectons that are not cliques, i.e., with two positions not directly in physical contact, but possibly contacting through other positions in the secton. As top pairs of $\tilde{\mathcal{D}}_{ij}$, sectons respect the decomposition into sectors. (c) Fraction of contacting pairs within sectons of size 2 (blue) or size $\geq 3$ (green) that are top pairs of $\tilde{\mathcal{D}}_{ij}$, for the top 35 sectons that are structurally connected. Contacts in sectons of size $\geq 3$ can be partitioned into contacts associated with the 2 positions contributing most to the secton (cyan), which are nearly all top pairs of $\tilde{\mathcal{D}}_{ij}$, and other contacts (red), of which only $\sim$20% are top pairs of $\tilde{\mathcal{D}}_{ij}$. (d) Fraction of the top 79 (black) or 120 (gray) pairs of $\tilde{\mathcal{D}}_{ij}$ contained in a secton: $\sim$30% of these top pairs are not in a secton.

Rivoire O. (2013) Phys Rev Lett. 110(17):178102.

# Random matrix approach to elements of coevolution in biological sequences

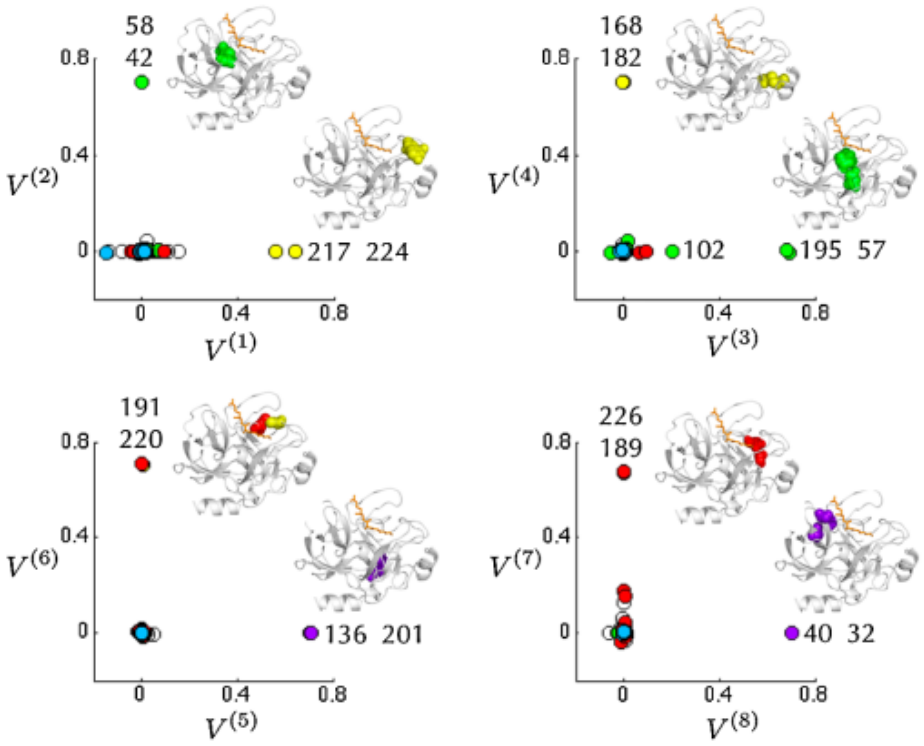



FIG. 3 (color online). Top protein sectons in the trypsin family—each graph is a projection of the positions along ($V^{(k)}$, $V^{(k+1)}$), the components of order $k$ and $k+1$ obtained by rotating by ICA the top eigenvectors of the truncated matrix of direct information $\tilde{\mathcal{D}}_{ij}$. Sectons are defined by $s_k = \{i: V_i^{(k)} > \epsilon\}$, with $\epsilon = 0.2$. The labeling of positions follows the numbering system of bovine chymotrypsin (in several instances, positions appear as superimposed), and the colors reflect the sectors as in Fig. 1, with yellow for nonsector positions. The location of the sectons on the three-dimensional structure is also indicated (more sectons are shown in Fig. S10 [8]). Sectons $s_2$, $s_4$, $s_5$, and $s_6$ are disulfide bonds, and $s_3$ is the catalytic triad.

Rivoire O. (2013) Phys Rev Lett. 110(17):178102.

# Population Structure and Eigenanalysis – software SMARTPCA

- Motivation: methods for inferring population structure from genetic data do not provide formal significance tests for population differentiation.
- Solution:
  - An approach - principal components analysis - to studying population structure
  - Tracy–Widom Theory – a solid statistical footing, using results from modern statistics to develop formal significance tests.
  - Propose BBP threshold to estimate for data size needed for significant.
- Approach: PCA has three major features
  - Runs extremely quickly on large datasets (within a few hours on datasets with hundreds of thousands of markers and thousands of samples)
  - PCA framework provides the first formal tests for the presence of population structure in genetic data.
  - PCA method does not attempt to classify all individuals into discrete populations or linear combinations of populations
  - PCA outputs each individual's coordinates along axes of variation

*Nick Patterson, et al(2006), PLoS Genetics, Volume 2 , Issue 12, e190*

# Population Structure and Eigenanalysis – software SMARTPCA

- A Test for Population Structure: This leads immediately to a formal test for the presence of population structure in a biallelic dataset
  1. Compute the matrix M

$$\mu(j) = \frac{\sum_{i=1}^{m} C(i,j)}{m} \qquad C(i,j) - \mu(j) \qquad M(i,j) = \frac{C(i,j) - \mu(j)}{\sqrt{p(j)(1-p(j))}}$$

  2. Compute X = MM'. X is mxm
  3. Order the eigenvalues of X so that $\lambda_1 > \lambda_2 ... > \lambda_{m'} > 0$ where m' = m-1
  4. Using the eigenvalues $\{\lambda_i\}_{1 \le i \le m}$ and estimate n' from

$$n' = \frac{(m+1)\left(\sum_i \lambda_i\right)^2}{\left((m-1)\sum_i \lambda_i^2\right) - \left(\sum_i \lambda_i\right)^2}$$

  5. The largest eigenvalue of M is $\lambda_1$. Set

$$l = \frac{(m')\lambda_1}{\sum_{i=1}^{m'} \lambda_i}$$

*Nick Patterson, et al(2006), PLoS Genetics, Volume 2 , Issue 12, e190*

# Population Structure and Eigenanalysis – software SMARTPCA

- Compute an eigenvector decomposition of X. Eigenvectors corresponding to ‘‘large’’ eigenvalues are exposing nonrandom population structure.

$$X = \frac{1}{n} MM'$$

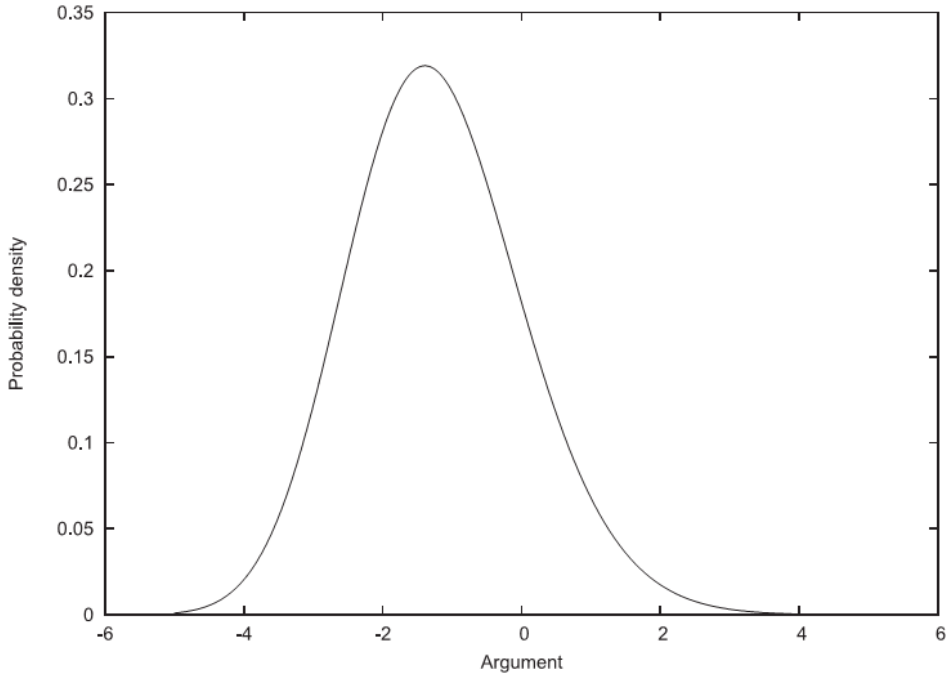
- Tracy–Widom Theory
  - X is a Wishart matrix. $\{\lambda_i\}_{1 \le i \le m}$ be the eigenvalues of X.

$$\mu(m,n) = \frac{\left(\sqrt{n-1} + \sqrt{m}\right)^2}{n}$$

$$\sigma(m,n) = \frac{\left(\sqrt{n-1} + \sqrt{m}\right)}{n} \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{m}}\right)^{1/3}$$

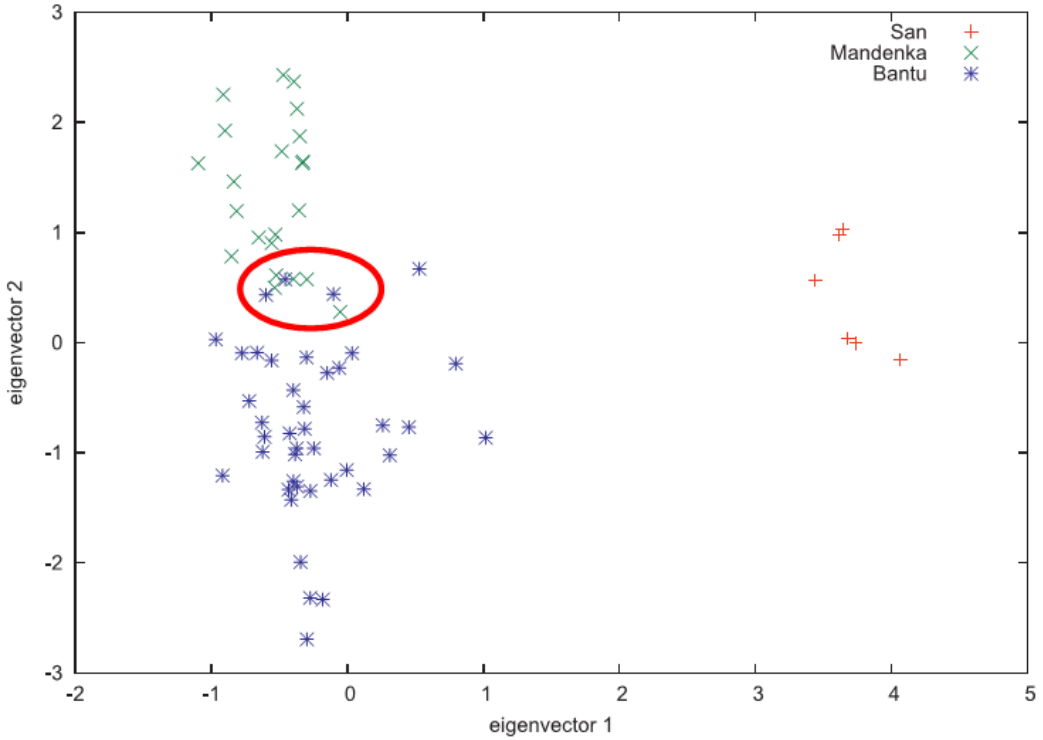
$$x = \frac{\lambda_1 - \mu(m,n)}{\sigma(m,n)}$$

  - Figure 1.The Tracy–Widom Density



# Population Structure and Eigenanalysis – software SMARTPCA

- Figure 4. Three African Populations




- In Table 1, the ANOVA p-value is obtained from the usual F-statistic, and we apply ANOVA to each of the first three eigenvectors

**Table 1.** Statistics from HGDP African Data

| Number | Eigenvalue | TW Statistic | TW *p*-Value | ANOVA *p*-Value |
|---|---|---|---|---|
| 1 | 2.07 | 46.2 | $<10^{-12}$ | $<10^{-12}$ |
| 2 | 1.40 | 6.717 | $3.08 \times 10^{-7}$ | $<10^{-12}$ |
| 3 | 1.31 | 0.380 | .108 | .74 |

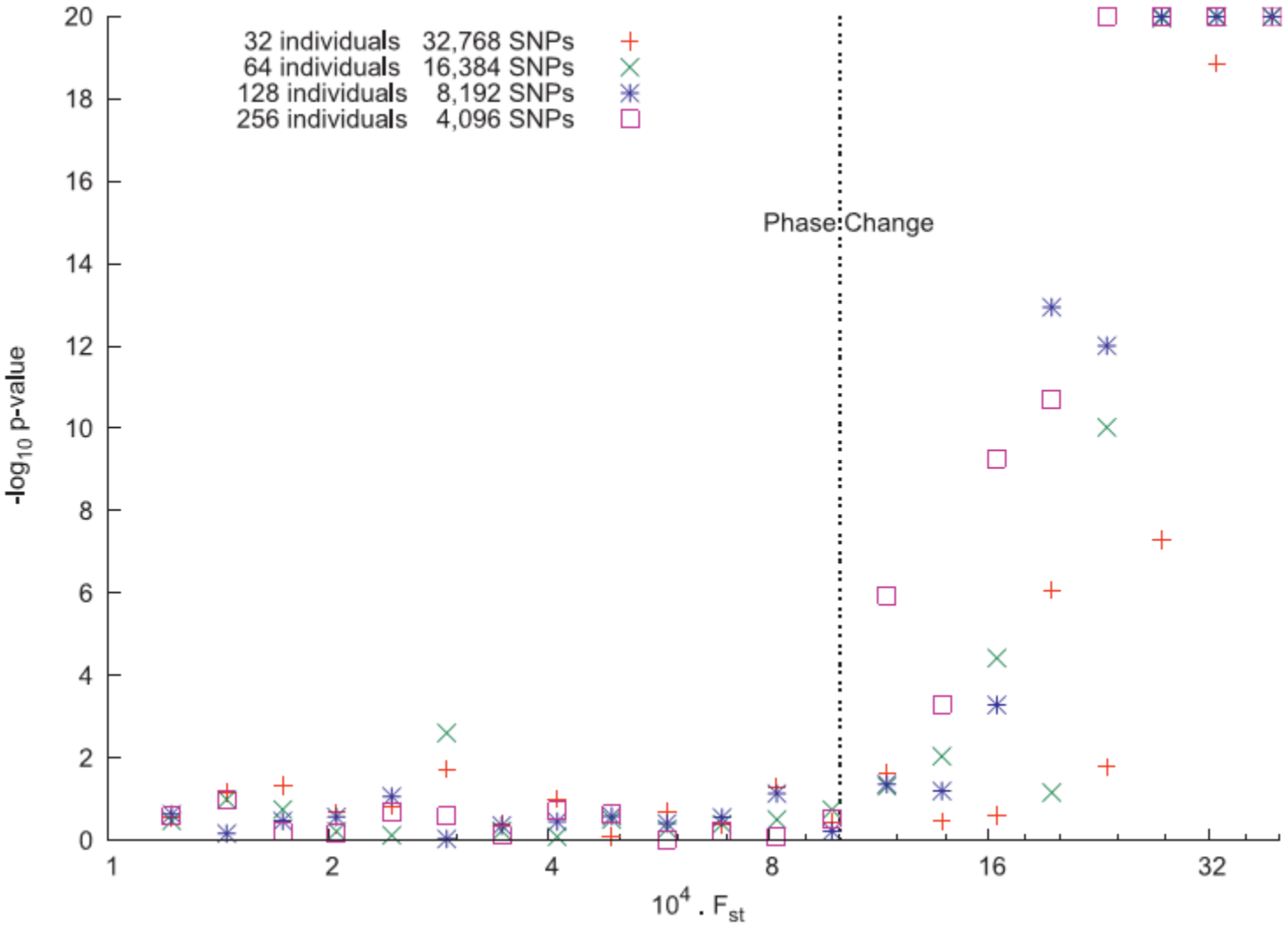
# Population Structure and Eigenanalysis – software SMARTPCA

- An Estimate for the Data Size Needed for Significance: a form of the conjecture, which we call the BBP conjecture.
- $l_1$ be the lead eigenvalue of the theoretical covariance matrix, with the remainder of the eigenvalues 1. Set $\gamma^2 = \frac{n}{m}$
- $L_1$ be the largest eigenvalue of the sample covariance
- the behavior of $L_1$ is qualitatively different depending on whether $l_1$ is greater or less than $1+ 1/\gamma$
- A phase-change phenomenon, as the BBP threshold. $1 + 1/\gamma = \frac{\sqrt{m} + \sqrt{n}}{\sqrt{n}}$
- Conclude: For two equal size subpopulation below which there will be essentially no evidence of population structure. Above the threshold, the evidence accumulates very rapidly, as we increase the divergence or the data size. Above the threshold for fixed data size mn, the evidence is stronger as we increase m, populations, there is a threshold value of $F_{ST}$,
- Most large genetic datasets with human data will show some detectable population structure

*Nick Patterson, et al(2006), PLoS Genetics, Volume 2 , Issue 12, e190*

# Population Structure and Eigenanalysis – software SMARTPCA

- Figure 6. The BBP Phase Change




*Nick Patterson, et al(2006), PLoS Genetics, Volume 2 , Issue 12, e190*

# Population Structure and Eigenanalysis – software SMARTPCA

- There remain issues of SMARTPCA:
  1. Recent admixture generates large-scale LD which may cause difficulties in a dense dataset as the allele distributions are not independent. STRUCTURE 2.0 allows careful modeling.
  2. More ancient admixture, especially if the admixed population is genetically now homogeneous, may lead to a causal eigenvalue not very different from the values generated by the sampling noise.
  3. Methods require that divergence is small, and that allele frequencies are divergent primarily because of drift.
  4. If ‘‘admixture LD’’ is present, so that in admixed individuals long segments of the genome originate from one founder population, simple PCA methods will not be as powerful as programs such as STRUCTURE 2.0. LD will seriously distort the eigenvector/eigenvalue structure, making results difficult to interpret.
- Solution for “admixture LD” problem.
  1. Form matrix M

$$M(i,j) = \frac{C(i,j) - \mu(j)}{\sqrt{p(j)(1-p(j))}}$$

  2. For each column j, set

$$\mathbf{a} = a_s^{[j]}(1 \le s \le k)$$

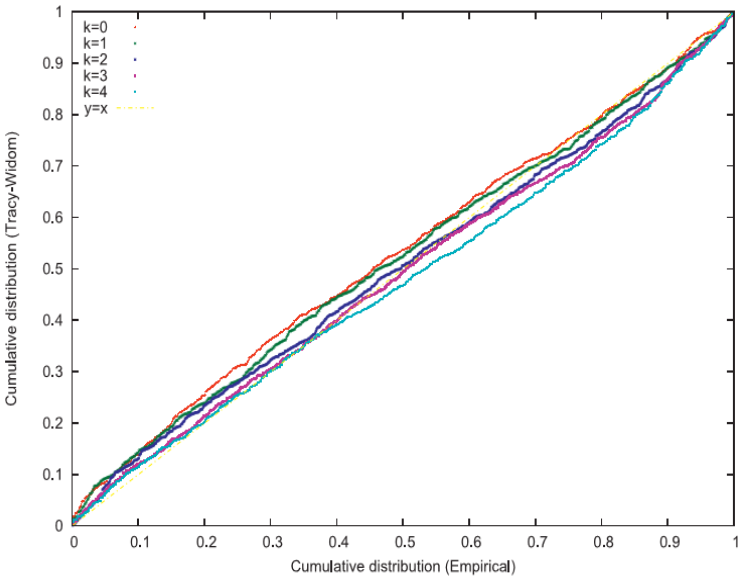
$$R(i,j) = M(i,j) - \sum_{s=1}^{k} a_s^{[j]} M(i,j-s)(1 \le i \le m)$$

  3. Choose a to minimize $\sum_i R^2(i,j)$

  4. Calculate X = RR’ instead of MM’

# Population Structure and Eigenanalysis – software SMARTPCA

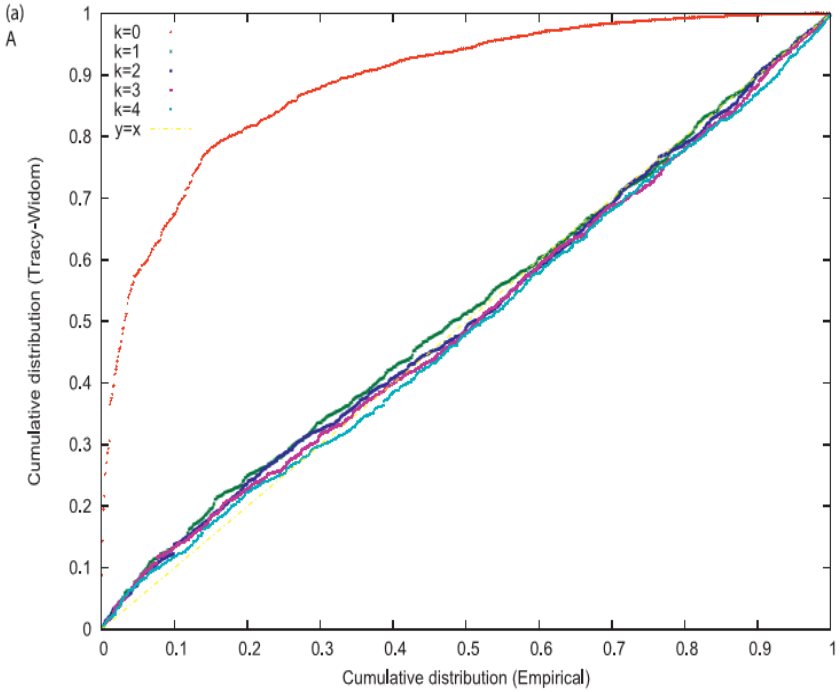
- Absence of LD the suggested correction does not seriously distort the Tracy–Widom statistic. Show P–P plots, uncorrected, and with five levels (k = 1,...,5) of correction.




- Problem: Markers are in complete LD, a large eigenvector of our Wishart matrix X will tend to correlate with the genotype pattern in the block. This will distort the eigenvector structure and also the distribution of eigenvalues.

Result: the uncorrected statistic is distributed quite differently than the Tracy–Widom distribution.

PCA strategy seems to work well



# Reference


1. J. Wishart, “The generalized product moment distribution in samples from a normal multivariate population,” Biometrika, vol. 20 A, pp. 32–52, 1928.
2. E. Wigner, “On the distribution of roots of certain symmetric matrices,” The Annals of Mathematics, vol. 67, pp. 325–327, 1958.
3. M. L. Mehta and M. Gaudin, “On the density of the eigenvalues of a random matrix,” Nuclear Physics, vol. 18, pp. 420–427, 1960.
4. V. A. Marcenk ˇ o and L. A. Pastur, “Distributions of eigenvalues for some sets of random matrices,” Math USSR-Sbornik, vol. 1, pp. 457–483, 1967
5. Laloux, L., Cizeau, P., Bouchaud, J. and Potters, M., Noise dressing of financial correlation matrices. Phys. Rev. Lett., 1999, 83, 1467–1470.
6. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T. and Stanley, H.E., Random matrix approach to cross correlations in financial data. Phys. Rev. E, 2002, 65, 066126-1–066126-18.
7. Rivoire O. Elements of coevolution in biological sequences. Phys Rev Lett. 2013 Apr; 110(17):178102.