A European application of random matrix theory

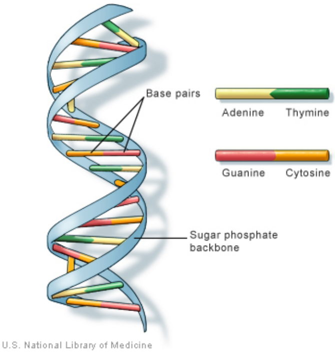
Brent Nelson

Michigan State University

Math 317H
December 5th, 2019

Genetics

[Novembre et al., Nature 2008]: Researchers analyzed genetic data from people with European ancestry: 1,387 people at 197,146 genetic loci.



$$\begin{array}{ccc} A & \mapsto & 1 \\ C & \mapsto & 2 \\ G & \mapsto & 3 \\ T & \mapsto & 4 \end{array}$$

They recorded these numbers as a 1387×197146 matrix X.

$$
\begin{array}{r c}
 & \begin{array}{cccccc} \textbf{locus } 1 & \textbf{locus } 2 & \cdots & \textbf{locus } j & \cdots & \textbf{locus } 197146 \end{array} \\
\begin{array}{r} \textbf{person } 1 \\ \textbf{person } 2 \\ \vdots \\ \textbf{person } \mathrm{i} \\ \vdots \\ \textbf{person } 1387 \end{array} &
\left(\begin{array}{cccccc}
x_{1,1} & x_{1,2} & \cdots & & \cdots & x_{1,197146} \\
x_{2,1} & x_{2,2} & \cdots & & & \vdots \\
\vdots & \vdots & \ddots & & & \\
 & & & x_{i,j} & & \\
\vdots & & & & \ddots & \vdots \\
x_{1387,1} & \cdots & & & \cdots & x_{1387,197146}
\end{array}\right) =: X
\end{array}
$$

They recorded these numbers as a 1387×197146 matrix X.

$$
\begin{array}{r c}
 & \begin{array}{cccccc} \textbf{locus } 1 & \textbf{locus } 2 & \cdots & \textbf{locus } j & \cdots & \textbf{locus } 197146 \end{array} \\
\begin{array}{r} \textbf{person } 1 \\ \textbf{person } 2 \\ \vdots \\ \textbf{person } \mathrm{i} \\ \vdots \\ \textbf{person } 1387 \end{array} &
\begin{pmatrix}
x_{1,1} & x_{1,2} & \cdots & & \cdots & x_{1,197146} \\
x_{2,1} & x_{2,2} & \cdots & & & \vdots \\
\vdots & \vdots & \ddots & & & \\
 & & & x_{i,j} & & \\
\vdots & & & & \ddots & \vdots \\
x_{1387,1} & \cdots & & & \cdots & x_{1387,197146}
\end{pmatrix} =: X
\end{array}
$$

They then analyzed this data by examining the singular values of X....

Probability Theory

- What does it mean for a coin to be *fair*?

- What does it mean for a coin to be *fair*?
- If you find a coin on the street, how can you determine if it is fair?

- What does it mean for a coin to be *fair*?
- If you find a coin on the street, how can you determine if it is fair?

Unfair Coins:

https://izbicki.me/blog/how-to-create-an-unfair-coin-and-prove-it-with-math.html

Probability Questions

- Flip a coin 100 times, how many times should it come up heads if the coin is fair?

Probability Questions

- Flip a coin 100 times, how many times should it come up heads if the coin is fair? (50)

Probability Questions

- Flip a coin 100 times, how many times should it come up heads if the coin is fair? (50)
- Roll a 6-sided die 60 times, how times should you roll a 1 if the die is fair?

Probability Questions

- Flip a coin 100 times, how many times should it come up heads if the coin is fair? (50)
- Roll a 6-sided die 60 times, how times should you roll a 1 if the die is fair? (10)

Probability Questions

- Flip a coin 100 times, how many times should it come up heads if the coin is fair? (50)
- Roll a 6-sided die 60 times, how times should you roll a 1 if the die is fair? (10)
- Play 600,000 hands of poker, how many times should you get a royal flush if the deck is fair?

Probability Questions

- Flip a coin 100 times, how many times should it come up heads if the coin is fair? (50)
- Roll a 6-sided die 60 times, how times should you roll a 1 if the die is fair? (10)
- Play 600,000 hands of poker, how many times should you get a royal flush if the deck is fair? (< 1)

Probability Questions

- Flip a coin 100 times, how many times should it come up heads if the coin is fair? (50)
- Roll a 6-sided die 60 times, how times should you roll a 1 if the die is fair? (10)
- Play 600,000 hands of poker, how many times should you get a royal flush if the deck is fair? (< 1)
- You text someone 5 times, how many minutes should it take for them to reply if they aren't ghosting you?

Probability Questions

- Flip a coin 100 times, how many times should it come up heads if the coin is fair? (50)
- Roll a 6-sided die 60 times, how times should you roll a 1 if the die is fair? (10)
- Play 600,000 hands of poker, how many times should you get a royal flush if the deck is fair? (< 1)
- You text someone 5 times, how many minutes should it take for them to reply if they aren't ghosting you? (asking for a friend)

Probability Questions

- Flip a coin 100 times, how many times should it come up heads if the coin is fair? (50)
- Roll a 6-sided die 60 times, how times should you roll a 1 if the die is fair? (10)
- Play 600,000 hands of poker, how many times should you get a royal flush if the deck is fair? (< 1)
- You text someone 5 times, how many minutes should it take for them to reply if they aren't ghosting you? (asking for a friend)

Moral: In order to know if something unexpected has happened, you first need to know what the expected (i.e average) outcome is.

Probability Questions

- Flip a coin 100 times, how many times should it come up heads if the coin is fair? (50)
- Roll a 6-sided die 60 times, how times should you roll a 1 if the die is fair? (10)
- Play 600,000 hands of poker, how many times should you get a royal flush if the deck is fair? (< 1)
- You text someone 5 times, how many minutes should it take for them to reply if they aren't ghosting you? (asking for a friend)

Moral: In order to know if something unexpected has happened, you first need to know what the expected (i.e average) outcome is.

Probability theory: Provides the tools needed to compute the expected outcome.

Random Matrix Theory

Definition

A **random matrix** is an $n \times m$ matrix X with at least one *randomly* generated entry.

Definition

A **random matrix** is an $n \times m$ matrix X with at least one *randomly* generated entry.

Example

Flip a coin, define $x = 1$ if the coin comes up heads and $x = -1$ if the coin comes up tails. Roll a 6-sided die and let y be the result. Then

$$A = \begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix} \qquad B = \begin{pmatrix} 1 & y \\ y & 1 \end{pmatrix}$$

are random matrices.

Definition

A **random matrix** is an $n \times m$ matrix X with at least one *randomly* generated entry.

Example

Flip a coin, define $x = 1$ if the coin comes up heads and $x = -1$ if the coin comes up tails. Roll a 6-sided die and let y be the result. Then

$$A = \begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix} \qquad B = \begin{pmatrix} 1 & y \\ y & 1 \end{pmatrix}$$

are random matrices.

Since the matrix is (at least partially) random, the data associated to the matrix is potentially random as well: the entries of X, $\det(X)$, $\mathrm{Tr}(X)$, eigenvalues of X, eigenvectors of X, etc.

Least Squares?

Suppose you collect the following data set:

$$(0, -1), \qquad (2, 3), \qquad (q, 2).$$

However, all you remember about the last x-coordinate is that $3 \leq q \leq 5$.

Least Squares?

Suppose you collect the following data set:

$$(0, -1), \qquad (2, 3), \qquad (q, 2).$$

However, all you remember about the last x-coordinate is that $3 \leq q \leq 5$. You want to find the line $y = ax + b$ that best fits this data:

$$\begin{pmatrix} 0 & 1 \\ 2 & 1 \\ q & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \\ 2 \end{pmatrix}$$

Least Squares?

Suppose you collect the following data set:

$$(0,-1), \qquad (2,3), \qquad (q,2).$$

However, all you remember about the last x-coordinate is that $3 \le q \le 5$. You want to find the line $y = ax + b$ that best fits this data:

$$\begin{pmatrix} 0 & 1 \\ 2 & 1 \\ q & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \\ 2 \end{pmatrix}$$

Can still find the least squares solution using this random matrix:

$$\begin{pmatrix} 4+q^2 & 2+q \\ 2+q & 3 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 6+2q \\ 4 \end{pmatrix}$$

Least Squares?

Suppose you collect the following data set:

$$(0,-1), \qquad (2,3), \qquad (q,2).$$

However, all you remember about the last x-coordinate is that $3 \leq q \leq 5$. You want to find the line $y = ax + b$ that best fits this data:

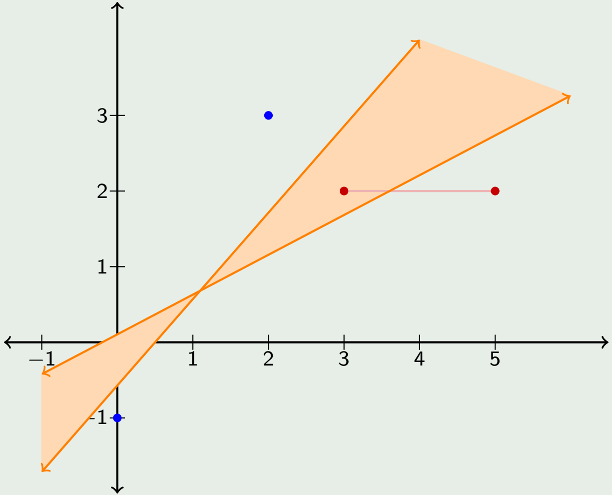
$$\begin{pmatrix} 0 & 1 \\ 2 & 1 \\ q & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \\ 2 \end{pmatrix}$$

Can still find the least squares solution using this random matrix:

$$\begin{pmatrix} 4+q^2 & 2+q \\ 2+q & 3 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 6+2q \\ 4 \end{pmatrix} \rightsquigarrow y = \frac{q+5}{q^2-2q+4}x + \frac{q^2-5q+2}{q^2-2q+4}$$

Least Squares? (continued)

$$y = \frac{q+5}{q^2-2q+4}x + \frac{q^2-5q+2}{q^2-2q+4}$$



Not McConnell

McConnell

Eugene Wigner

Sad McConnell

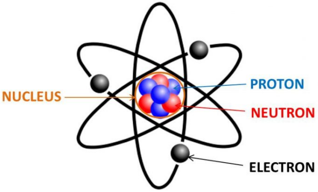
Wigner was motivated by chemistry and physics.

Wigner was motivated by chemistry and physics.

- **Goal:** given an atom, try to understand the possible energy levels of electrons

Wigner was motivated by chemistry and physics.

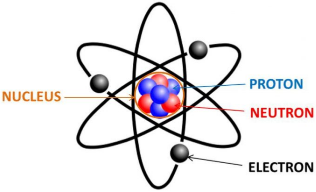
- **Goal:** given an atom, try to understand the possible energy levels of electrons
- **Hamiltonian** linear operator that models the dynamics of the subatomic particles (protons, electrons, neutrons), and its eigenvalues give you the possible energy levels



http://www.whoinventedfirst.com/who-discovered-the-atom/

Wigner was motivated by chemistry and physics.

- **Goal:** given an atom, try to understand the possible energy levels of electrons
- **Hamiltonian** linear operator that models the dynamics of the subatomic particles (protons, electrons, neutrons), and its eigenvalues give you the possible energy levels

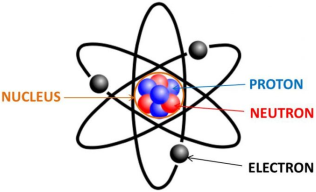


http://www.whoinventedfirst.com/who-discovered-the-atom/

- Can be computed explicitly for the hydrogen atom: 1 proton, 1 electron

Wigner was motivated by chemistry and physics.

- **Goal:** given an atom, try to understand the possible energy levels of electrons
- **Hamiltonian** linear operator that models the dynamics of the subatomic particles (protons, electrons, neutrons), and its eigenvalues give you the possible energy levels

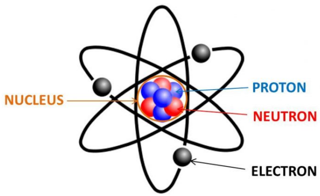


http://www.whoinventedfirst.com/who-discovered-the-atom/

- Can be computed explicitly for the hydrogen atom: 1 proton, 1 electron
- But for atoms with "heavy nuclei" (e.g uranium-238: 92 protons, 92 electrons, 146 neutrons), too complicated to solve explicitly

Wigner was motivated by chemistry and physics.

- **Goal:** given an atom, try to understand the possible energy levels of electrons
- **Hamiltonian** linear operator that models the dynamics of the subatomic particles (protons, electrons, neutrons), and its eigenvalues give you the possible energy levels



http://www.whoinventedfirst.com/who-discovered-the-atom/

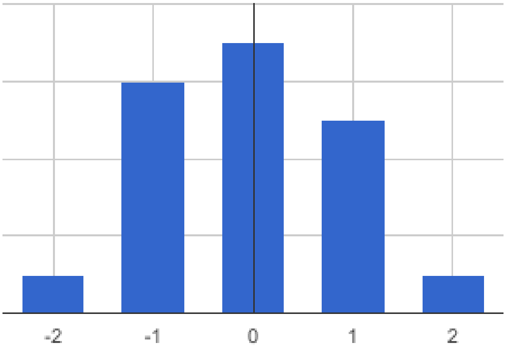
- Can be computed explicitly for the hydrogen atom: 1 proton, 1 electron
- But for atoms with "heavy nuclei" (e.g uranium-238: 92 protons, 92 electrons, 146 neutrons), too complicated to solve explicitly
- **Wigner's idea:** treat the Hamiltonian as a random matrix

- For each $n \in \mathbb{N}$, let A_n be an $n \times n$ matrix with all entries random and independent of one another.

- For each $n \in \mathbb{N}$, let A_n be an $n \times n$ matrix with all entries random and independent of one another.
- Define $X_n := \frac{1}{n}(A_n + A_n^T)$, which is an $n \times n$ symmetric random matrix.

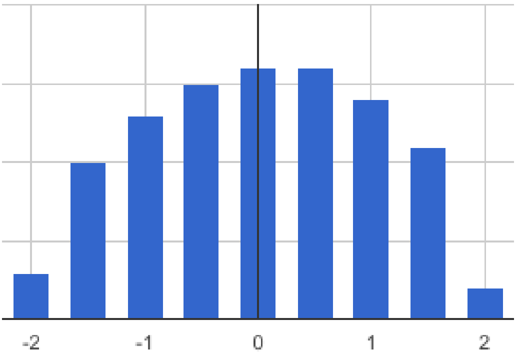
- For each $n \in \mathbb{N}$, let A_n be an $n \times n$ matrix with all entries random and independent of one another.
- Define $X_n := \frac{1}{n}(A_n + A_n^T)$, which is an $n \times n$ symmetric random matrix.
- Make a histogram of $\#$eigenvalues of X_n in each interval of length $\sim \frac{1}{n}$.

- For each $n \in \mathbb{N}$, let A_n be an $n \times n$ matrix with all entries random and independent of one another.
- Define $X_n := \frac{1}{n}(A_n + A_n^T)$, which is an $n \times n$ symmetric random matrix.
- Make a histogram of #eigenvalues of X_n in each interval of length $\sim \frac{1}{n}$.



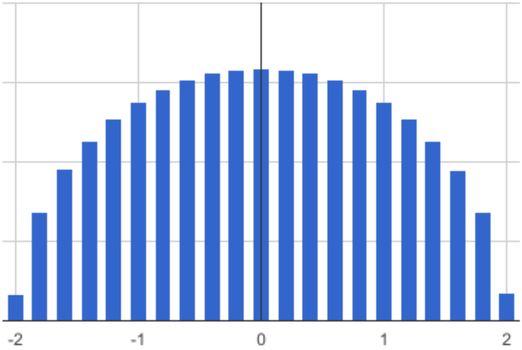
$n = 20$

- For each $n \in \mathbb{N}$, let A_n be an $n \times n$ matrix with all entries random and independent of one another.
- Define $X_n := \frac{1}{n}(A_n + A_n^T)$, which is an $n \times n$ symmetric random matrix.
- Make a histogram of #eigenvalues of X_n in each interval of length $\sim \frac{1}{n}$.



$n = 100$

- For each $n \in \mathbb{N}$, let A_n be an $n \times n$ matrix with all entries random and independent of one another.
- Define $X_n := \frac{1}{n}(A_n + A_n^T)$, which is an $n \times n$ symmetric random matrix.
- Make a histogram of #eigenvalues of X_n in each interval of length $\sim \frac{1}{n}$.

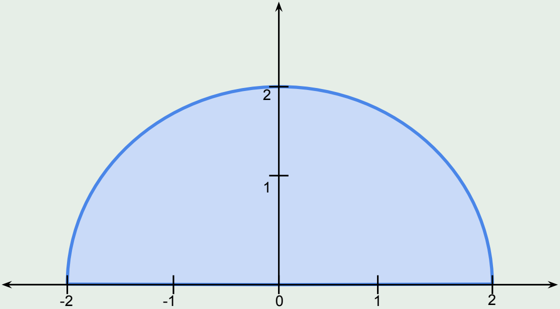


$n = 10,000$

Semicircle Distribution

The histograms get closer and closer to the **semicircle distribution**:

$$s(t) = \begin{cases} \frac{1}{2\pi}\sqrt{4-t^2} & \text{if } -2 \leq t \leq 2 \\ 0 & \text{otherwise} \end{cases}.$$



Theorem (Wigner's Semicircle Law)

Let X_n, $n \in \mathbb{N}$, be the sequence of symmetric random matrices as above. For any interval $[a, b] \subset \mathbb{R}$,

$$\lim_{n \to \infty} \frac{\#\{\text{eigenvalues of } X_n \text{ in the interval } [a, b]\}}{n} = \int_a^b s(t)\ dt.$$

Theorem (Wigner's Semicircle Law)

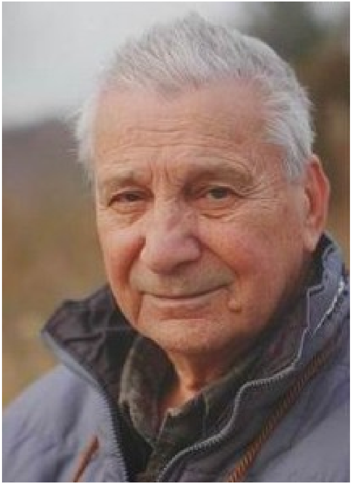
Let X_n, $n \in \mathbb{N}$, be the sequence of symmetric random matrices as above. For any interval $[a, b] \subset \mathbb{R}$,

$$\lim_{n\to\infty} \frac{\#\{\text{eigenvalues of } X_n \text{ in the interval } [a,b]\}}{n} = \int_a^b s(t)\ dt.$$

That is, there exists an $N \in \mathbb{N}$ so that for any $n \geq N$

$$\frac{\#\{\text{eigenvalues of } X_n \text{ in the interval } [a,b]\}}{n} \approx \frac{1}{2\pi} \int_a^b \sqrt{4 - t^2}\ dt.$$

The Marčenko–Pastur Law



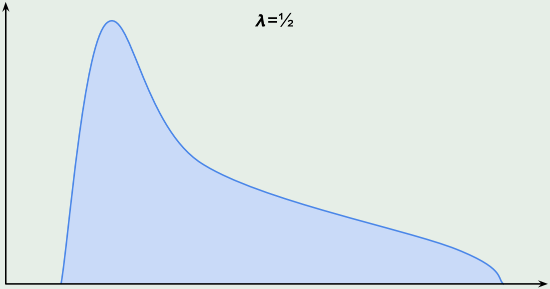
Vladimir Marčenko

Leonid Pastur

Marčenko–Pastur distribution

Fix $\lambda \in (0,1]$ and define $\lambda_\pm = (1 \pm \sqrt{\lambda})^2$. The **Marčenko–Pastur distribution** is

$$\nu(t) := \begin{cases} \frac{1}{2\pi\lambda} \frac{\sqrt{(t-\lambda_-)(\lambda_+-t)}}{t} & \text{if } \lambda_- \leq t \leq \lambda_+ \\ 0 & \text{otherwise} \end{cases}.$$



- Let $(p(n))_{n\in\mathbb{N}} \subset \mathbb{N}$ be a sequence satisfying

$$\lambda := \lim_{n\to\infty} \frac{p(n)}{n} \in (0,1].$$

- Let $(p(n))_{n\in\mathbb{N}} \subset \mathbb{N}$ be a sequence satisfying

$$\lambda := \lim_{n\to\infty} \frac{p(n)}{n} \in (0,1].$$

- For each $n \in \mathbb{N}$, let X_n be $p(n) \times n$ matrix all of whose entries are random and independent of one another.

- Let $(p(n))_{n\in\mathbb{N}} \subset \mathbb{N}$ be a sequence satisfying

$$\lambda := \lim_{n\to\infty} \frac{p(n)}{n} \in (0,1].$$

- For each $n \in \mathbb{N}$, let X_n be $p(n) \times n$ matrix all of whose entries are random and independent of one another.
- Define $Y_n = \frac{1}{n} X_n X_n^T$, a $p(n) \times p(n)$ symmetric random matrix

- Let $(p(n))_{n\in\mathbb{N}} \subset \mathbb{N}$ be a sequence satisfying

$$\lambda := \lim_{n\to\infty} \frac{p(n)}{n} \in (0,1].$$

- For each $n \in \mathbb{N}$, let X_n be $p(n) \times n$ matrix all of whose entries are random and independent of one another.
- Define $Y_n = \frac{1}{n} X_n X_n^T$, a $p(n) \times p(n)$ symmetric random matrix
- Eigenalues of Y_n are $\frac{1}{n}$ times the squares of the singular values of X_n.

- Let $(p(n))_{n\in\mathbb{N}} \subset \mathbb{N}$ be a sequence satisfying

$$\lambda := \lim_{n\to\infty} \frac{p(n)}{n} \in (0,1].$$

- For each $n \in \mathbb{N}$, let X_n be $p(n) \times n$ matrix all of whose entries are random and independent of one another.
- Define $Y_n = \frac{1}{n} X_n X_n^T$, a $p(n) \times p(n)$ symmetric random matrix
- Eigenalues of Y_n are $\frac{1}{n}$ times the squares of the singular values of X_n.

Theorem (The Marčenko–Pastur Law)

With Y_n, $n \in \mathbb{N}$, as above, for any interval $[a,b] \subset \mathbb{R}$,

$$\lim_{n\to\infty} \frac{\#\{\text{eigenvalues of } Y_n \text{ in the interval } [a,b]\}}{p(n)} = \int_a^b \nu(t)\, dt.$$

That is, there exists an $N \in \mathbb{N}$ so that for any $n \geq N$

$$\frac{\#\{\text{eigenvalues of } Y_n \text{ in the interval } [a,b]\}}{p(n)} \approx \frac{1}{2\pi\lambda} \int_a^b \frac{\sqrt{(t-\lambda_-)(\lambda_+ - t)}}{t}\, dt.$$

Back to Genetics

Researchers recorded genetic data as the matrix

$$
\begin{array}{r c}
 & \begin{array}{cccccc} \textbf{locus } 1 & \textbf{locus } 2 & \cdots & \textbf{locus } j & \cdots & \textbf{locus } 197146 \end{array} \\
\begin{array}{r} \textbf{person } 1 \\ \textbf{person } 2 \\ \vdots \\ \textbf{person } \mathrm{i} \\ \vdots \\ \textbf{person } 1387 \end{array} &
\begin{pmatrix}
x_{1,1} & x_{1,2} & \cdots & & \cdots & x_{1,197146} \\
x_{2,1} & x_{2,2} & \cdots & & & \vdots \\
\vdots & \vdots & \ddots & & & \\
 & & & x_{i,j} & & \\
\vdots & & & & \ddots & \vdots \\
x_{1387,1} & \cdots & & & \cdots & x_{1387,197146}
\end{pmatrix}
\end{array} =: X
$$

Researchers recorded genetic data as the matrix

$$
\begin{array}{r c}
 & \begin{array}{ccccccc} \textbf{locus } 1 & \textbf{locus } 2 & \cdots & \textbf{locus } j & \cdots & \textbf{locus } 197146 \end{array} \\
\begin{array}{r} \textbf{person } 1 \\ \textbf{person } 2 \\ \vdots \\ \textbf{person } \mathrm{i} \\ \vdots \\ \textbf{person } 1387 \end{array} &
\begin{pmatrix}
x_{1,1} & x_{1,2} & \cdots & & \cdots & x_{1,197146} \\
x_{2,1} & x_{2,2} & \cdots & & & \vdots \\
\vdots & \vdots & \ddots & & & \\
 & & & x_{i,j} & & \\
\vdots & & & & \ddots & \vdots \\
x_{1387,1} & \cdots & & & \cdots & x_{1387,197146}
\end{pmatrix} =: X
\end{array}
$$

Set

$$
\begin{aligned} n &:= 197,146 \\ p(n) &:= 1,387 \end{aligned} \qquad \lambda := \frac{p(n)}{n} = \frac{1,387}{197,146} \approx 0.007035
$$

They computed eigenvalues of $Y := \frac{1}{n}XX^T$, which is a $p(n) \times p(n)$ symmetric matrix.

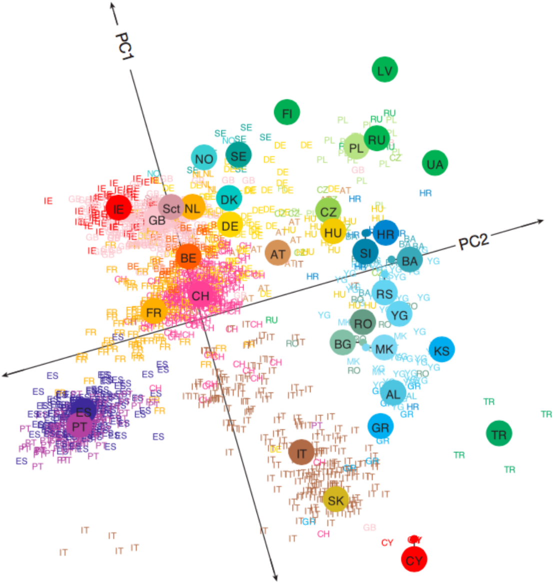
- Suppose—naively—that the entries of X are random. That is, that the nucleobases in each persons DNA were randomly assigned.

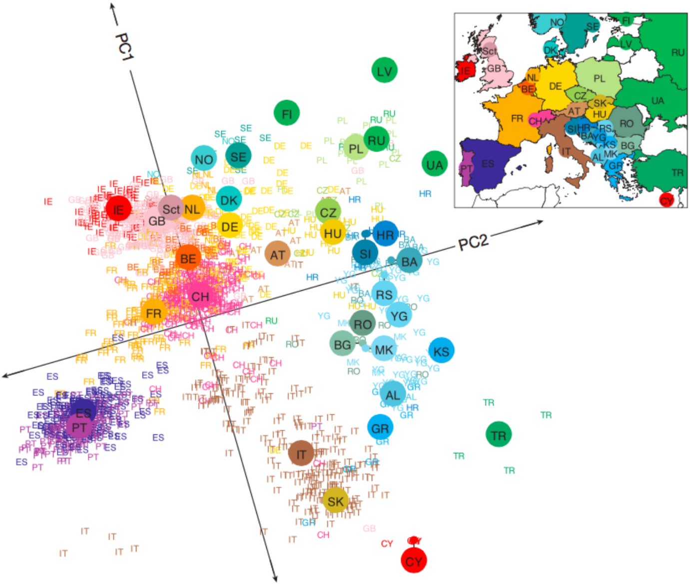
- Suppose—naively—that the entries of X are random. That is, that the nucleobases in each persons DNA were randomly assigned.
- Under our naive assumption, the Marčenko–Pastur law says that a histogram of the eigenvalues of Y should look like the graph of the Marčenko–Pastur distribution $\nu(t)$ with resolution λ. However, the actual data did not quite yield this: there were two outlying eigenvalues.

- Suppose—naively—that the entries of X are random. That is, that the nucleobases in each persons DNA were randomly assigned.
- Under our naive assumption, the Marčenko–Pastur law says that a histogram of the eigenvalues of Y should look like the graph of the Marčenko–Pastur distribution $\nu(t)$ with resolution λ. However, the actual data did not quite yield this: there were two outlying eigenvalues.
- Let $x, y \in \mathbb{R}^{p(n)}$ be their unit eigenvectors. Note that $p(n) = 1,387$, which was the number of people in the study.

- Suppose—naively—that the entries of X are random. That is, that the nucleobases in each persons DNA were randomly assigned.
- Under our naive assumption, the Marčenko–Pastur law says that a histogram of the eigenvalues of Y should look like the graph of the Marčenko–Pastur distribution $\nu(t)$ with resolution λ. However, the actual data did not quite yield this: there were two outlying eigenvalues.
- Let $x, y \in \mathbb{R}^{p(n)}$ be their unit eigenvectors. Note that $p(n) = 1,387$, which was the number of people in the study.
- Thus the eigenvectors x, y assign to each person a coordinate pair: person i is assigned the coordinate pair (x_i, y_i) where x_i and y_i are the ith entries of x and y, respectively.

- Suppose—naively—that the entries of X are random. That is, that the nucleobases in each persons DNA were randomly assigned.
- Under our naive assumption, the Marčenko–Pastur law says that a histogram of the eigenvalues of Y should look like the graph of the Marčenko–Pastur distribution $\nu(t)$ with resolution λ. However, the actual data did not quite yield this: there were two outlying eigenvalues.
- Let $x, y \in \mathbb{R}^{p(n)}$ be their unit eigenvectors. Note that $p(n) = 1,387$, which was the number of people in the study.
- Thus the eigenvectors x, y assign to each person a coordinate pair: person i is assigned the coordinate pair (x_i, y_i) where x_i and y_i are the ith entries of x and y, respectively.
- Something pretty incredible happens when you plot these coordinate pairs...


PC1
PC2
LV
FI
RU
PL
UA
NO
SE
DK
IE
Sct
NL
GB
DE
CZ
HU
HR
SI
BA
BE
AT
CH
RS
FR
YG
RO
BG
MK
KS
AL
ES
PT
GR
TR
IT
SK
CY


PC1
PC2
LV
FI
RU
PL
UA
NO
SE
DK
IE
Sct
GB
NL
DE
CZ
HU
HR
SI
BA
AT
BE
CH
FR
RS
YG
RO
BG
MK
KS
AL
GR
TR
ES
PT
IT
SK
CY