Randomness in data analysis

Van H. Vu

Department of Mathematics
Yale University

S. O'rourke (Yale), T. Tao (UCLA), K. Wang (Rutgers)

Data and Noise

Data: A very large matrix A.

Data and Noise

Data: A very large matrix A.

Major tools to analyze data: Linear Algebra.

Data and Noise

Data: A very large matrix A.

Major tools to analyze data: Linear Algebra.

Example. Principal Component Analysis (Low rank approximation).
(1) Computing the first few singular vectors and values.

(2) Project onto the subspace spanned by the first few eigenvectors.

Random Noise is Inevitable.

Negative Impact

The data matrix A is perturbed by random noise. Thus, one works with $A + E$, where E is the noise matrix.

Negative Impact

The data matrix A is perturbed by random noise. Thus, one works with $A + E$, where E is the noise matrix.

$$\sum_{i=1}^{n} \sigma_i v_i v_i^{\ T} \rightarrow \sum_{i=1}^{k} \sigma_i v_i v_i^T \ \rightarrow \sum_{i=1}^{k} \sigma_i' v_i' v_i'^T.$$

How does the noise effect the accuracy of the analysis ?

Guarantee resilience and fault tolerance of software.

Estimate error rates, and increase the signal to noise ratio.

Positive impact

Once we understand (quantitatively) the effect of noise, we can make some use of it.

Artificial randomness.

Positive impact

Once we understand (quantitatively) the effect of noise, we can make some use of it.

Artificial randomness.

(1) Adding artificial randomness can speed up algorithms.

(2) "Small" noise does not influence the output significantly. (Spielman-Teng smoothed analysis.)

Positive impact

Once we understand (quantitatively) the effect of noise, we can make some use of it.

Artificial randomness.

(1) Adding artificial randomness can speed up algorithms.

(2) "Small" noise does not influence the output significantly. (Spielman-Teng smoothed analysis.)

Strong connections to random matrix theory and high dimensional geometry.

Computing the eigen/singular vectors

Problem. For a matrix A of size $n \times n$ with singular values $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, let $v_1, \ldots, v_n$ be the corresponding (unit) singular vectors. Compute $v_1, \ldots, v_k$, for some $k \leq n$.

Computing the eigen/singular vectors

Problem. For a matrix A of size $n \times n$ with singular values $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, let $v_1, \ldots, v_n$ be the corresponding (unit) singular vectors. Compute $v_1, \ldots, v_k$, for some $k \leq n$.

Typically n is large and k is relatively small (say 2 or 3). .

Goal: estimate the influence of noise on the vectors $v_1, \ldots, v_k$.

Let $v'_1, \ldots, v'_k$ be the first k singular vectors of $A + E$.

Question. When is v'_1 a good approximation of v_1 ?

sub-Question. Is it true that if the noise gets smaller, then v'_1 becomes a better approximation ?

A surprising answer

NO !! the singular vectors are *not* continuous. Let A be

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Singular vectors $(1, 0)$ and $(0, 1)$. Let E be

$$\begin{pmatrix} 0 & \epsilon \\ \epsilon & 0 \end{pmatrix}.$$

The perturbed matrix $A + E$ has the form

$$\begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}.$$

The singular vectors are $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$, *no matter how small* ϵ is.

The goal

Measure the distance between v and v' by $\sin \angle(v, v')$, where $\angle(v, v')$ is the angle between the vectors, taken in $[0, \pi/2]$.

Fix a small parameter $\epsilon > 0$, which represents a desired accuracy.

GOAL. Find a sufficient condition (for A) which guarantees that $\sin \angle(v_1, v_1') \leq \epsilon$.

CLASSICAL numerical linear algebra: The key parameter to look at is the gap (or separation)

$$\delta := \sigma_1 - \sigma_2,$$

between the first and second singular values of A.

Wedin and Davis-Kahan

Theorem (Wedin sin theorem)

$$\sin \angle(v_1, v_1') \leq \frac{\|E\|}{\delta}.$$

Corollary

For any small $\epsilon > 0$, *if* $\delta \geq \frac{\|E\|}{\epsilon}$, *then*

$$\sin \angle(v_1, v_1') \leq \epsilon.$$

Wedin and Davis-Kahan

Theorem (Wedin sin theorem)

$$\sin \angle(v_1, v_1') \leq \frac{\|E\|}{\delta}.$$

Corollary

For any small $\epsilon > 0$, *if* $\delta \geq \frac{\|E\|}{\epsilon}$, *then*

$$\sin \angle(v_1, v_1') \leq \epsilon.$$

A and $A + E$ are Hermitian: Davis-Kahan theorem.

Models of random noise

The entries of E are iid random variables with mean 0 and variance 1 (the value 1 is, of course, just matter of normalization).

Example. Bernoulli (± 1), Gaussian.

Models of random noise

The entries of E are iid random variables with mean 0 and variance 1 (the value 1 is, of course, just matter of normalization).

Example. Bernoulli (± 1), Gaussian.

We prefer Bernoulli over Gaussian:

Models of random noise

The entries of E are iid random variables with mean 0 and variance 1 (the value 1 is, of course, just matter of normalization).

Example. Bernoulli (± 1), Gaussian.

We prefer Bernoulli over Gaussian:

(1) In real life situations, noise is not always Gaussian (in fact rarely Gaussian).

(2) If one can analyze Bernoulli, one can analyze any distribution.

A classical bound

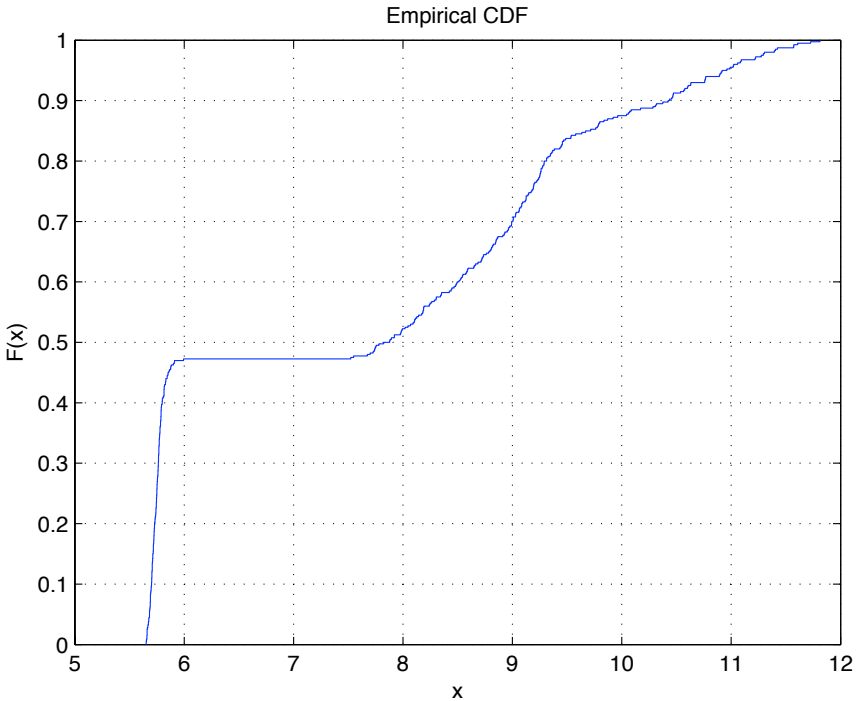
(Füredi-Kom'os) $\mathbf{E}\|\| \approx 2\sqrt{n}$, with high probability.

Corollary

A gap $\delta \geq \frac{\sqrt{n}}{\delta}$ guarantees (with high probability)

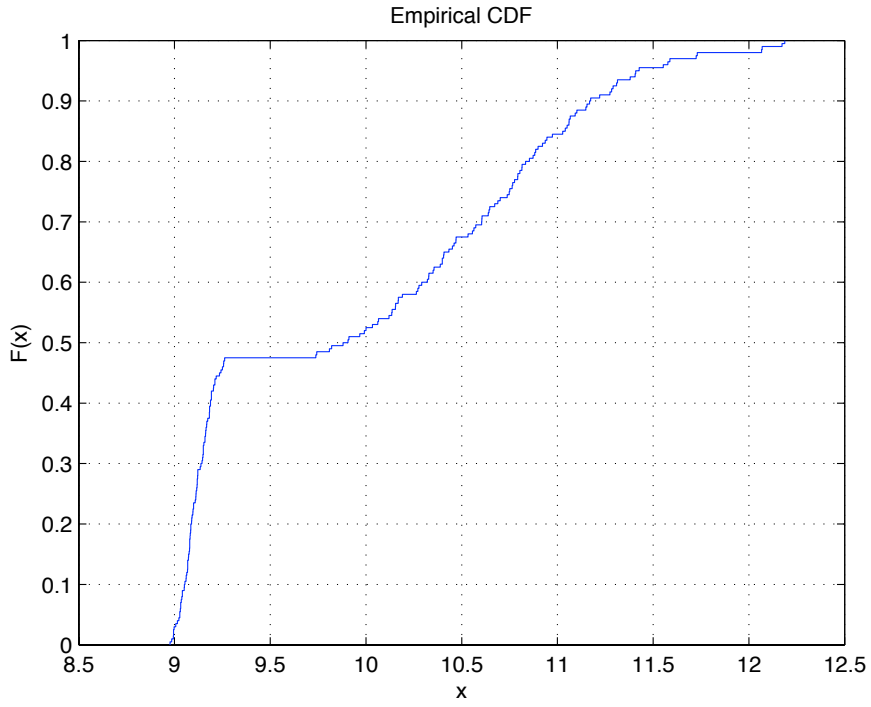
$$\sin \angle(v_1, v_1') \leq \epsilon.$$

Numerical experiments



400×400 matrix of rank 2, gap $= 8$;

Wedin's bound: gap $\geq \frac{2\times\sqrt{40}}{12/90} = 300$.



1000×1000 matrix of rank 2, gap $= 10$.

Wedin bound: gap $\geq \frac{2 \times \sqrt{1000}}{12/90} \approx 450$.

A new phenomenon: Real dimension matters

Low dimensional data and improved bounds

In a large variety of problems, the data is of small dimension, namely, $r := \mathit{rank}\ A \ll n$ (e.g: Compress sensing, Candes-Tao)

A new phenomenon: Real dimension matters

Low dimensional data and improved bounds

In a large variety of problems, the data is of small dimension, namely, $r := \mathit{rank}\ A \ll n$ (e.g: Compress sensing, Candes-Tao)

FINDING. the efficient gap depends on the *real dimension* r, rather than the dimension n of the matrix.

Theorem

A gap $\delta \geq C\frac{\sqrt{r \log n}}{\epsilon}$. *guarantees (with high probability)*

$$\sin \angle(v_1, v_1') \leq \epsilon. \tag{1}$$

Proof combines ideas from theory of random matrices, high dimensional geometry and concentration of measure.

Directions of research

Directions of research

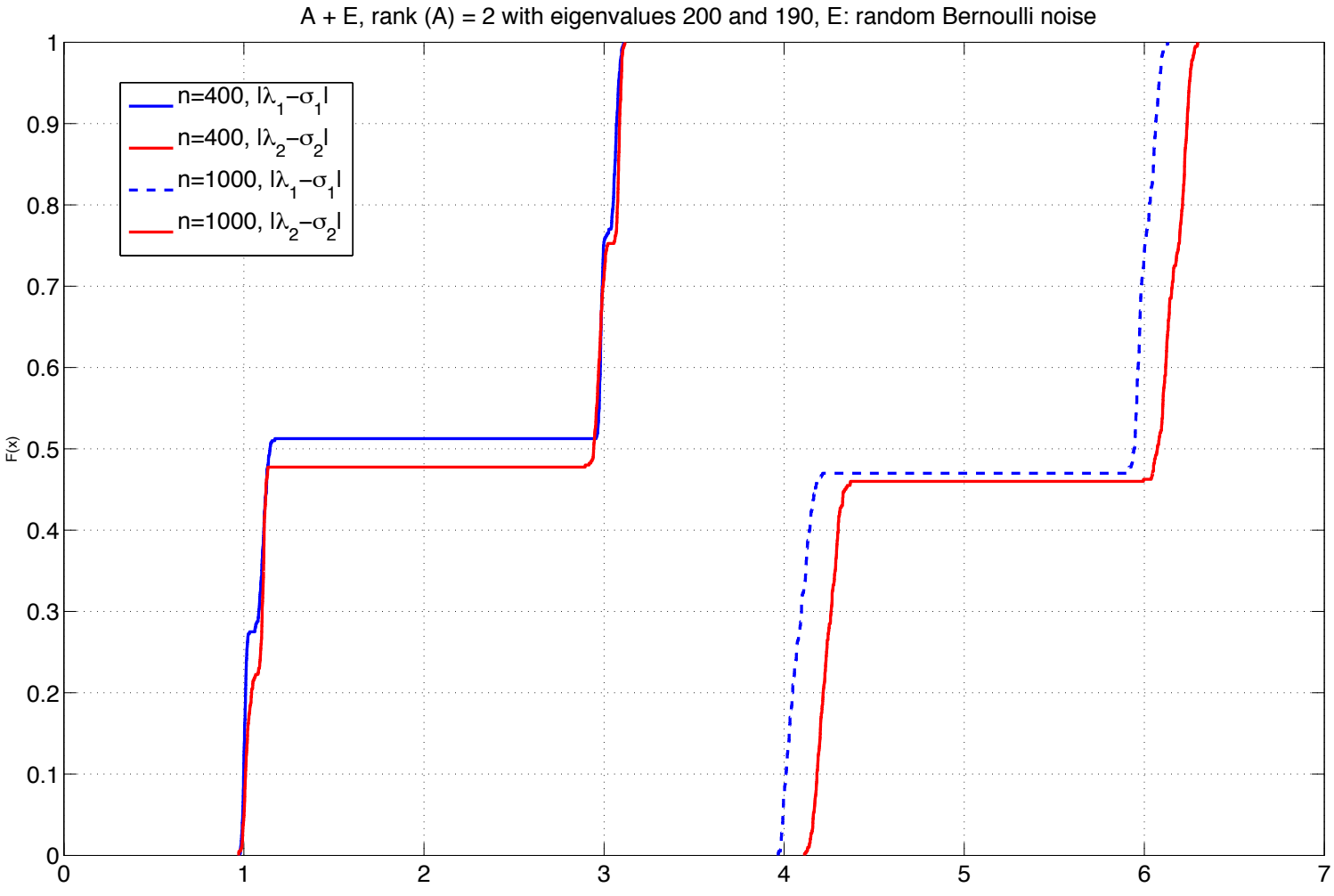
- Eigenvalues.
- Improve bounds.
- Angles between subspaces.
- Other models of random matrices (not iid entries).

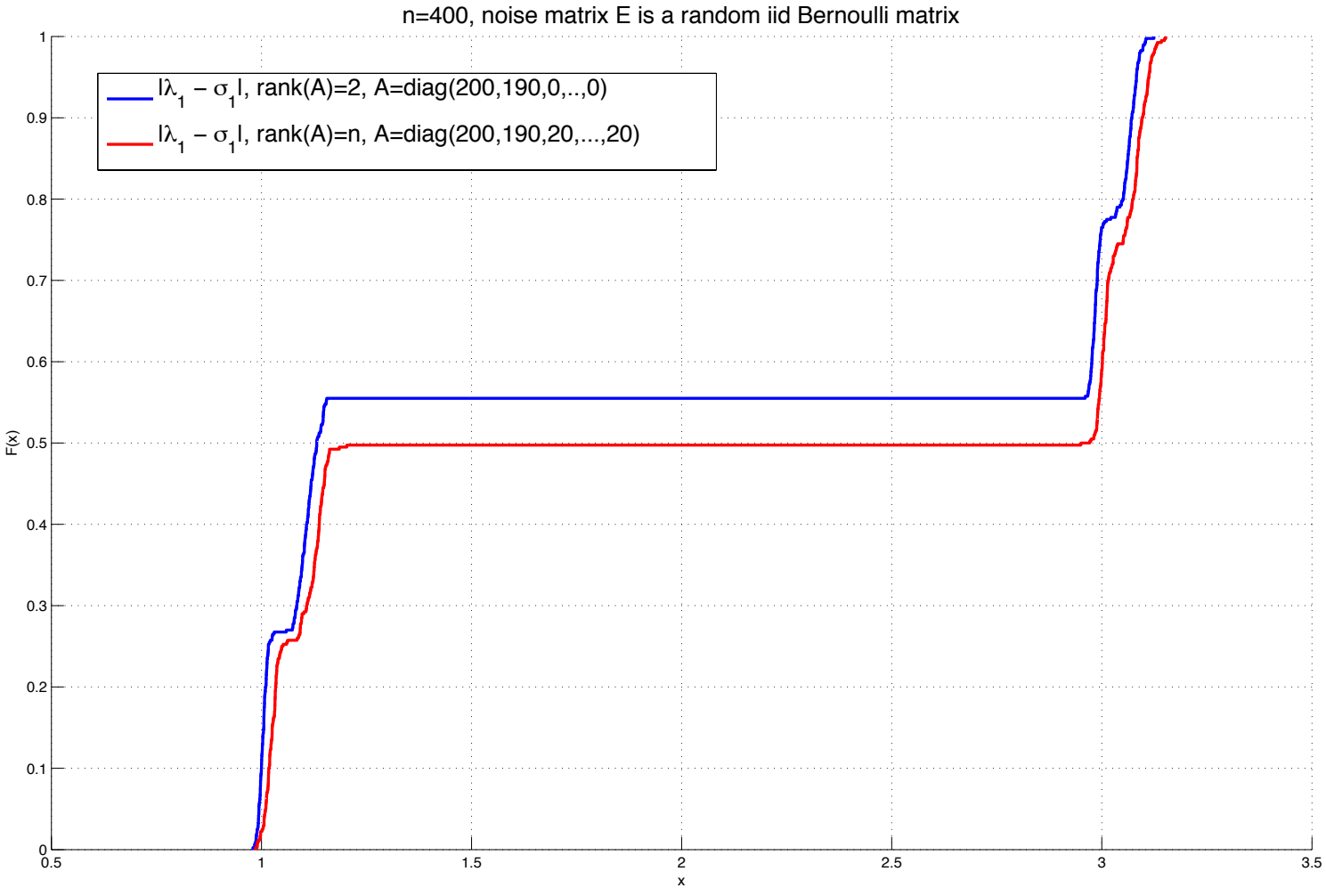
The eigenvalue problem

Problem. For a matrix A of size $n \times n$ with singular values $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. Compute $_1, \ldots, \sigma_k$, for some $k \leq n$.

Weyl bound

$$|\sigma_i - \sigma_i'| \leq \|E\| \approx 2\sqrt{n}.$$


A + E, rank (A) = 2 with eigenvalues 200 and 190, E: random Bernoulli noise
n=400, |λ1−σ1|
n=400, |λ2−σ2|
n=1000, |λ1−σ1|
n=1000, |λ2−σ2|
F(x)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3
4
5
6
7



Weyl's bound $2\sqrt{400} = 40$.

A new method

Tao-V. Taylor's expansion approach (from random matrix theory, 2009):

$$F(A + tE) := F(A) + tF'(A) + \frac{t^2}{2!}F''(A) + \dots.$$

Set $t = 1, F := \sigma_1$ (or any parameter of interest).

Computing the derivatives

Set $M(t) = A + tE$.

$$M(t)v(t) := \lambda(t)v(t)$$

$$M'v + Mv' = \lambda'v + \lambda v'.$$

$$v^T M'v + v^T Mv' = \lambda' v^T v + \lambda v^T v'.$$

But as $vv^T = 1$, $vv'^T = 0$. So

$$v^T Ev = \lambda'.$$

E random, so Ev is *almost orthogonal to* v.

$|v^T Ev|$ is small; $|\lambda'|$ is small, so λ changes very little !!

Problem. The QR algorithm.

The QR algorithm (computing eigenvalues), dating to the early 1960s, is one of the jewels of numerical analysis. Its simplest form below can be seen as a stable procudure for computing QR factorization of the matrix power $M, M^2, M^3 \ldots$.

The Algorithm.

- Set $M^{(0)} := M$.
- For $k = 1, 2, \ldots$ compute $Q^{(k)} R^{(k)} = M^{(k-1)}$
- Set $M^{(k)} := R^{(k)} Q^{(k)}$.

The resulting matrices $M^{(k)}$ converges to the Schur form of M (upper -triangular of M is arbitrary and diagonal of M is symmetric). Our interest is in the *speed of convergence.*

The resulting matrices $M^{(k)}$ converges to the Schur form of M (upper -triangular of M is arbitrary and diagonal of M is symmetric). Our interest is in the *speed of convergence.*

Theorem

Let the pure QR algorithm be applied to a real symmetric matrix M with eigenvalues $|\lambda_1| > \cdots > |\lambda_n|$ and whose corresponding eigenvector matrix Q has all non-singular leading principal submatrix. Rate of convergence

$$\max_i \frac{|\lambda_i|}{|\lambda_i| - |\lambda_{i+1}|}.$$

KEY ISSUE. $\max_j |\lambda_{j+1}|/|\lambda_j|$ can be very close to one. If this is of order $1 + O(n^{-10})$, then the algorithm would take $\Omega(n^{10})$ steps.

In particular, the case when the matrix has eigenvalues with high multiplicities is rather troublesome.

KEY ISSUE. $\max_j |\lambda_{j+1}|/|\lambda_j|$ can be very close to one. If this is of order $1 + O(n^{-10})$, then the algorithm would take $\Omega(n^{10})$ steps.

In particular, the case when the matrix has eigenvalues with high multiplicities is rather troublesome.

IDEA. Adding artificial randomness to speed up !!

Run the algorithm on $M + \epsilon E$ where E is a random matrix.

The added randomness should create a gap between consecutive eigenvalues !!

The added randomness should create a gap between consecutive eigenvalues !!

But it also changes the eigenvalues slightly.

The added randomness should create a gap between consecutive eigenvalues !!

But it also changes the eigenvalues slightly.

It requires a delicate trade-off. (Work in process with Terence Tao (UCLA) and S. H. Teng (USC).)

The added randomness should create a gap between consecutive eigenvalues !!

But it also changes the eigenvalues slightly.

It requires a delicate trade-off. (Work in process with Terence Tao (UCLA) and S. H. Teng (USC).)

Application in control theory. The eigenvectors of a random graph do not have zero coordinate. (answered a question by M. Meshabi, Complex Network meeting 2010).