Matrix estimation by Universal Singular Value Thresholding

Sourav Chatterjee

Courant Institute, NYU
Let us begin with an example:

- Suppose that we have an undirected random graph $G$ on $n$ vertices.
- **Model:** There is a real symmetric matrix $P = (p_{ij})$ such that
  \[
  \text{Prob}(\{i, j\} \text{ is an edge of } G) = p_{ij},
  \]
  and edges pop up independently of each other.
- A statistical question: Given a single realization of the random graph $G$, under what conditions can we accurately estimate all the $p_{ij}$’s?
- The question is motivated by the study of the structure of real-world networks.
Of course, in the absence of any structural assumption about the matrix $P$, it is impossible to estimate the $p_{ij}$’s. They may be completely arbitrary.

The strongest structural assumption that one can make is that the $p_{ij}$’s are all equal to a single value $p$. This is the Erdős–Rényi model of random graphs. In this case $p$ may be easily estimated by the estimator

$$\hat{p} = \frac{\# \text{ edges of } G}{\binom{n}{2}}.$$ 

Then $\mathbb{E}(\hat{p} - p)^2 \to 0$ as $n \to \infty$, i.e., $\hat{p}$ is a consistent estimator of $p$. 

Sourav Chatterjee

Matrix estimation by Universal Singular Value Thresholding
The stochastic block model

- The **stochastic block model** assumes a little less structure than ‘all $p_{ij}$’s equal’.
- The vertices are divided into $k$ blocks (unknown to the statistician). For any two blocks $A$ and $B$, $p_{ij}$ is the same for all $i \in A$ and $j \in B$.
- Originated in the study of social networks. Studied by many authors over the last thirty years.
- A side remark: By the famous regularity lemma of Szemerédi, all dense graphs ‘look like’ as if they originated from a stochastic blockmodel.
The question of estimating the $p_{ij}$'s in the stochastic block model is a difficult question because the block membership is unknown. 

Condon and Karp (2001) were the first to give a consistent estimator when the number of blocks $k$ is fixed, all blocks are of equal size, and $n \to \infty$. 

Quite recently, Bickel and Chen (2009) solved the problem when the block sizes are allowed to be unequal. 

The work of Bickel and Chen was extended to allow $k \to \infty$ slowly as $n \to \infty$ by various authors. 

One cannot expect to solve the problem if $k$ is allowed to be of the same size as $n$, i.e. the number of blocks is comparable to the number of vertices. 

What if $k$ grows like $o(n)$? We will see later that indeed, consistent estimation is possible. This will solve the estimation problem of the stochastic block model in its entirety.
Latent space models

Here, one assumes that to each vertex $i$ is attached a hidden or latent variable $\beta_i$, and that

$$p_{ij} = f(\beta_i, \beta_j)$$

for some fixed function $f$.

Various authors have attempted to estimate the $\beta_i$’s from a single realization of the graph, but in all cases, $f$ is assumed to be some known function.

For example, in a recent paper with Persi Diaconis and Allan Sly, we showed that all the $\beta_i$’s may be simultaneously estimated from a single realization of the graph if $f(x, y) = e^{x+y}/(1 + e^{x+y})$.

What if $f$ is unknown? We will see later that the problem is solvable even if the statistician has absolutely no knowledge about $f$, as long as $f$ has some amount of smoothness.
A third approach to imposing structure is through the assumption that $P$ has low rank.

This has been investigated widely in recent years, beginning with the works of Candès and Recht (2009), Candès and Tao (2010) and Candès and Plan (2010).

Usually, the authors assume that a large part of the data is missing. This imposes an additional difficulty in detecting the structure.

Suppose that only a random fraction $q$ of the edges are ‘visible’ to the statistician, and that the matrix $P$ is of rank $r$. What is a necessary and sufficient condition, in terms of $r$, $n$ and $q$, under which the problem of estimating $P$ is solvable?

The theory that I am going to present shows that $r \ll nq$ is necessary and sufficient.
Recall: We have an undirected random graph $G$ on $n$ vertices, and there is a real symmetric matrix $P = (p_{ij})$ such that

$$\text{Prob}\{\{i, j\} \text{ is an edge of } G\} = p_{ij},$$

and edges occur independently of each other.

Given a single realization of the random graph $G$, under what conditions can we accurately estimate all the $p_{ij}$’s?

Instead of the graph $G$, we can visualize our data as the adjacency matrix $X = (x_{ij})$ of $G$.

The problem may be generalized beyond graphs by considering any random symmetric matrix $X$ whose entries on and above the diagonal are independent and $\mathbb{E}(x_{ij}) = p_{ij}$. 
The estimation problem can be solved only if we assume that the matrix $P$ has some ‘structure’.

We have seen three kinds of structural assumption: the stochastic block models, the latent space models, and the low rank assumption. There are various other kinds of assumptions that people make.

Questions: Can all these structural assumptions arise as special cases of a single assumption? That is, can there be a ‘universal’ notion of structure? And if so, does there exist a ‘universal’ algorithm that solves the estimation problem whenever structure is present (and in particular, solves all of the previously stated problems)?

Answer: Yes.
Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $P$. Recall that elements of $P$ are in $[0, 1]$.

Define the randomness coefficient of $P$ as the number

$$R(P) := \frac{\sum_{i=1}^{n} |\lambda_i|}{n^{3/2}}.$$ 

Incidentally, $\sum |\lambda_i|$ is commonly known as the ‘nuclear norm’ or ‘trace norm’ of $P$ and denoted by $\|P\|_\ast$. 

Sourav Chatterjee
Matrix estimation by Universal Singular Value Thresholding
The randomness coefficient

- **Claim:** $0 \leq R(P) \leq 1$ for any $P$.
- **Proof:** Simple consequence of the Cauchy-Schwarz inequality:

$$n^{3/2} R(P) = \sum_{i=1}^{n} |\lambda_i| \leq \left( n \sum_{i=1}^{n} \lambda_i^2 \right)^{1/2}$$

$$= (n \text{Tr}(P^2))^{1/2} = (n \sum_{i,j=1}^{n} p_{ij}^2)^{1/2} \leq n^{3/2}.$$ 

- When $R(P)$ is close to zero, we will interpret it as saying that $P$ has some amount of structure.
- Suppose that $n$ is large. When is $R(P)$ not close to zero?
- The only construction of a large matrix $P$ with $R(P)$ away from zero that I could come up with is a matrix with independent random entries.
- For example, one can show that such a construction is not possible with $p_{ij} = f(i/n, j/n)$ for some a.e. continuous $f$. 

Sourav Chatterjee  
Matrix estimation by Universal Singular Value Thresholding
Examples of matrices with structure (i.e. low randomness)

- **Latent space models.**
  - Suppose that $\beta_1, \ldots, \beta_n$ are values in $[0, 1]$ and $f : [0, 1]^2 \to [0, 1]$ is a Lipschitz function with Lipschitz constant $L$.
  - Suppose that $p_{ij} = f(\beta_i, \beta_j)$.
  - Then $R(P) \leq C(L)n^{-1/3}$, where $C(L)$ depends only on $L$.

- **Stochastic block models.**
  - Suppose that $P$ is described by a stochastic block model with $k$ blocks, possibly of unequal sizes.
  - Then $R(P) \leq \sqrt{k/n}$.

- **Low rank matrices.**
  - Suppose that $P$ has rank $r$.
  - Then $R(P) \leq \sqrt{r/n}$.

- **Distance matrices.**
  - Suppose that $(K, d)$ is a compact metric space and $p_{ij} = d(x_i, x_j)$, where $x_1, \ldots, x_n$ are arbitrary points in $K$.
  - Then $R(P) \leq C(K, d, n)$, where $C(K, d, n)$ is a number depending only on $K$, $d$ and $n$ that tends to zero as $n \to \infty$. 
Examples, continued

- **Positive definite matrices.**
  - Suppose that $P$ is positive definite with all entries in $[-1, 1]$.
  - Then $R(P) \leq 1/\sqrt{n}$.

- **Graphons.**
  - Suppose that $f : [0, 1]^2 \to [0, 1]$ is a measurable function.
  - Let $U_1, \ldots, U_n$ be i.i.d. Uniform$[0, 1]$ random variables.
  - Let $p_{ij} = f(U_i, U_j)$ and generate a random graph with these $p_{ij}$’s. Such graphs arise in the theory of graph limits recently developed by Lovász and coauthors.
  - In this case $R(P) \to 0$ as $n \to \infty$. The rate of convergence depends on $f$.

- **Monotone matrices.**
  - Suppose that there is a permutation $\pi$ of the vertices such that if $\pi(i) \leq \pi(i')$, then $p_{\pi(i)\pi(j)} \leq p_{\pi(i')\pi(j)}$ for all $j$.
  - Arises in certain statistical models, such as the Bradley–Terry model of pairwise comparison.
  - In this case, $R(P) \leq Cn^{-1/3}$, where $C$ is a universal constant.

- Basically, anything reasonable you can think of.
The USVT algorithm

- Suppose we have a random symmetric matrix $X = (x_{ij})$ of order $n$, all of whose entries are in $[0, 1]$ and are independent of each other on and above the diagonal. (Think of $X$ as the adjacency matrix of a random graph with independent edges.)

- Let $P = (p_{ij})$ where $p_{ij} = \mathbb{E}(x_{ij})$. In the random graph model, $p_{ij}$ is the probability that $\{i, j\}$ is an edge.

- Let $X = \sum_{i=1}^{n} \mu_i u_i u_i^T$ be the spectral decomposition of $X$.

- Define the estimate $\hat{P} = (\hat{p}_{ij})$ as

$$\hat{P} := \sum_{i : |\mu_i| \geq 1.01 \sqrt{n}} \mu_i u_i u_i^T.$$

- If $\hat{p}_{ij} > 1$ for some $i, j$, redefine $\hat{p}_{ij} = 1$. Similarly, if $\hat{p}_{ij} < 0$, redefine $\hat{p}_{ij} = 0$.

- This is a singular value thresholding algorithm. Since the threshold is universal, I call it Universal Singular Value Thresholding (USVT).
Remarks

- There exist other singular value thresholding algorithms in the literature, for example a recent one by Keshavan, Montanari and Oh (2010) or an old one by Achlioptas and McSherry (2001). But all previous algorithms use specific information about $P$.

- There is nothing special about the constant 1.01. Any constant strictly bigger than 1 is okay.
The main result

**Theorem (C., 2012)**

Let \( \hat{P} \) and \( P \) be as in the previous slide. Then

\[
\mathbb{E} \left( \frac{1}{n^2} \sum_{i,j=1}^{n} (\hat{p}_{ij} - p_{ij})^2 \right) \leq C \ R(P) + \frac{C}{n},
\]

where \( C \) is a universal constant and \( R(P) \) is the randomness coefficient of \( P \).
Theorem (C., 2012)

Fix $n$. Let $\tilde{P} = (\tilde{p}_{ij})$ be any estimator of $P$. Then for any $\delta \in [0, 1]$, there exists $P$ such that $R(P) \leq \delta$, and if this is the ‘true’ $P$, then

$$E\left(\frac{1}{n^2} \sum_{i,j=1}^{n} (\tilde{p}_{ij} - p_{ij})^2\right) \geq c \delta + \frac{c}{n},$$

where $c$ is a positive universal constant.
What if some entries are missing?

- Suppose that each element of $X$ is observed with probability $q$ and unobserved with probability $1 - q$, independent of each other.
- Let $\hat{q}$ be the proportion of observed entries.
- Put 0 in place of all the missing entries and call the resulting matrix $Y$.
- Let $Y = \sum_{i=1}^{n} \mu_i u_i u_i^T$ be the spectral decomposition of $Y$.
- Define
  \[ \hat{P} = \frac{1}{\hat{q}} \sum_{i : |\mu_i| \geq 1.01\sqrt{n\hat{q}}} \mu_i u_i u_i^T. \]
- As before, if $\hat{p}_{ij} > 1$, redefine $\hat{p}_{ij} = 1$ and if $\hat{p}_{ij} < 0$ redefine $\hat{p}_{ij} = 0$.
- This nice trick of replacing missing entries by zeros appeared for the first time in Keshavan, Montanari and Oh (2010).
Modified error bound and optimality

Theorem (C., 2012)

Suppose that $q \geq n^{-1+\epsilon}$ for some $\epsilon > 0$. Then

$$
\mathbb{E}\left( \frac{1}{n^2} \sum_{i,j=1}^{n} (\hat{p}_{ij} - p_{ij})^2 \right) \leq \frac{C R(P)}{\sqrt{q}} + \frac{C}{nq} + C(\epsilon)e^{-nq},
$$

where $C$ is a universal constant and $C(\epsilon)$ depends only on $\epsilon$.

Theorem (C., 2012)

If $\tilde{P}$ is any estimator, then for any $\delta \in [0, 1]$ there exists $P$ such that $R(P) \leq \delta$ and if this is the ‘true’ $P$, then

$$
\mathbb{E}\left( \frac{1}{n^2} \sum_{i,j=1}^{n} (\tilde{p}_{ij} - p_{ij})^2 \right) \geq \frac{c \delta}{\sqrt{q}} + \frac{c}{nq},
$$

where $c$ is a positive universal constant.
Non-symmetric and rectangular matrices

- Suppose that $P$ and $X$ are $m \times n$ matrices, with no symmetry assumption. Everything else as before.
- Let $X = \sum_{i=1}^{k} \mu_i u_i v_i^T$ be the singular value decomposition of $X$, where $k = \min\{m, n\}$ and $\mu_1, \ldots, \mu_k$ are the singular values of $X$.
- Then define

$$\hat{P} := \sum_{i: \mu_i \geq 1.01 \max\{\sqrt{m}, \sqrt{n}\}} \mu_i u_i v_i^T.$$ 

- The case of missing entries is dealt with exactly as before.
- The theorems remain just as they were, after modifying the definition of $R(P)$ as

$$R(P) = \frac{\sum_{i=1}^{k} \mu_i}{\sqrt{mnk}}.$$ 

Sourav Chatterjee
Matrix estimation by Universal Singular Value Thresholding
Let $n = 1000$. Let $\beta_1, \ldots, \beta_n$ and $\alpha$ be drawn independently and uniformly at random from $[0, 1]$.

Define

$$p_{ij} := \frac{1}{1 + e^{-\beta_i - \beta_j - \alpha \beta_i \beta_j}}.$$ 

This is a logistic model with interaction.

Generate a random graph on $n$ vertices by including the edge $\{i, j\}$ with probability $p_{ij}$, independently for all $i, j$.

Apply the USVT algorithm to this random graph to compute the estimates $\hat{p}_{ij}$. Note that the USVT algorithm knows nothing about the specific formula used to define $p_{ij}$, nor the values of $\beta_1, \ldots, \beta_n$.

To visually see how accurately $\hat{p}_{ij}$ estimates $p_{ij}$, take a random sample of 200 entries from the $1000 \times 1000$ matrix $P$ and plot them against the corresponding entries from $\hat{P}$.

The results are shown in the next slide.
Simulation result

**Figure**: Plot of $\hat{p}_{ij}$ versus $p_{ij}$ for a random sample of 200 entries.
USVT gives:

- A complete solution to the estimation problem in stochastic block models.
- A complete solution to the estimation problem in latent space models.
- A necessary and sufficient condition for estimability of low rank matrices with missing entries, and a simple and fast method for carrying out the estimation. (Note, however, that the methods of Candès and coauthors allow exact recovery under stronger assumptions, while USVT gives approximate recovery but under no additional assumptions.)
- A complete solution to the problem of distance matrix estimation.
- Many other applications, worked out in the manuscript on arXiv.
Proof sketch in the symmetric case with no missing entries

- Key ingredients: Random matrix theory + concentration of measure + matrix inequalities + lucky coincidence.

- $P = (p_{ij})$ is a symmetric matrix of order $n$, and $X = (x_{ij})$ is a random matrix with independent entries on and above the diagonal, such that $x_{ij} \in [0, 1]$ and $\mathbb{E}(x_{ij}) = p_{ij}$ for all $i, j$.

- Let $X = \sum_{i=1}^{n} \mu_i u_i u_i^T$ be the spectral decomposition of $X$.

- The USVT estimate of $P$ is defined as

$$\hat{P} := \sum_{i : |\mu_i| \geq 1.01\sqrt{n}} \mu_i u_i u_i^T.$$

- For a symmetric matrix $A$ of order $n$ and eigenvalues $\theta_1, \ldots, \theta_n$,
  - the spectral norm of $A$ is defined as $\|A\| := \max_i |\theta_i|$, and
  - the Frobenius norm of $A$ is defined as

$$\|A\|_F := (\sum_{i,j} a_{ij}^2)^{1/2} = (\sum_i \theta_i^2)^{1/2}.$$  

- Clearly, $\|A\|_F \leq \sqrt{\text{rank}(A)} \|A\|$. 

Sourav Chatterjee  
Matrix estimation by Universal Singular Value Thresholding
From random matrix theory and concentration of measure it follows that

\[ \|X - P\| \leq 1.001\sqrt{n} \]

with probability tending to 1 as \( n \to \infty \). Call this event \( E \).

Let \( P = \sum_{i=1}^{n} \lambda_i v_i v_i^T \) be the spectral decomposition of \( P \).

Let

\[ P_1 := \sum_{i : |\lambda_i| \geq .009\sqrt{n}} \lambda_i v_i v_i^T. \]

Let \( S := \{ i : |\lambda_i| \geq .009\sqrt{n} \} \). Then

\[ \text{rank}(P_1) \leq |S| \leq \frac{\sum_{i=1}^{n} |\lambda_i|}{.009\sqrt{n}} \leq C n R(P), \]

where \( C \) is a universal constant.
Proof sketch continued

- Suppose that \( \lambda_i \)'s and \( \mu_i \)'s are arranged in decreasing order. Then from matrix inequalities it follows that

\[
\max_i |\lambda_i - \mu_i| \leq \|X - P\|.
\]

- Thus if the event \( E \) happens, then \( |\mu_i| \geq 1.01 \sqrt{n} \) implies that \( |\lambda_i| \geq 0.009 \sqrt{n} \).

- In particular, if \( E \) happens then the rank of \( \hat{P} \) is also bounded by \( CnR(P) \).

- Consequently, if \( E \) happens then

\[
\|\hat{P} - P_1\|_F \leq C \sqrt{nR(P)} \|\hat{P} - P_1\|
\]
\[
\leq C \sqrt{nR(P)} (\|\hat{P} - X\| + \|X - P\| + \|P - P_1\|)
\]
\[
\leq Cn \sqrt{R(P)}.
\]
Proof sketch continued

Moreover,

\[ \| P_1 - P \|_F = \left( \sum_{i: |\lambda_i| < 0.09\sqrt{n}} \lambda_i^2 \right)^{1/2} \]

\[ \leq \left( 0.09\sqrt{n} \sum_{i=1}^{n} |\lambda_i| \right)^{1/2} \]

\[ \leq Cn \sqrt{R(P)}. \]

The last two inequalities give the same bound in terms of \( R(P) \) (serendipity!). Combining, we see that if \( E \) happens, then

\[ \| \hat{P} - P \|_F \leq Cn \sqrt{R(P)}. \]

It is now easy to complete the proof because \( E \) happens with high probability.