Detection threshold in very sparse matrix completion

Charles Bordenave

Work in collaboration with Raj Rao Nadakuditi (Univ. Michigan) and Simon Coste (Inria Paris)

ArXiv:2005.06062

Matrix estimation

Matrix estimation

Let $P \in M_{n,m}(\mathbb{R})$ be a large rectangular matrix $n = \Theta(m)$.

We observe each entry independently with probability d/n. The other entries remain hidden.

$d =$ *average number of observed entries per row.*

We assume that the matrix P has a simple structure: notably small rank and some spectral incoherence.

Completion and estimation

Matrix completion: can we reconstruct exactly P thanks to this observation?

Possible in the regime $d \geqslant C \log n$.

Completion and estimation

Matrix completion: can we reconstruct exactly P thanks to this observation?

Possible in the regime $d \geqslant C \log n$.

Matrix estimation: we look for a matrix with a small mean square error

$$\mathrm{MSE}(\widehat{P}) = \sum_{i,j} |\widehat{P}_{ij} - P_{ij}|^2 = \|\widehat{P} - P\|_F^2.$$

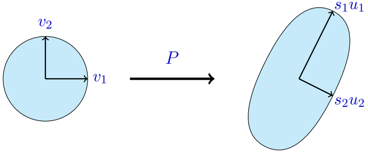
Best bounds for $d \ll \log n$: $\mathrm{MSE}(\widehat{P}) = O(mn/d)$.

Candès-Tao 09, Candès-Recht 10, Keshavan-Montanari-Oh 09 ...

Singular value decomposition of $P \in M_{n,m}(\mathbb{C})$:

$$P = UDV^* = \sum_{k=1}^{n} s_k u_k v_k^*,$$

where $D = \mathrm{diag}(s_1, \ldots, s_n) \in M_{n,m}(\mathbb{C})$ and $s_1 \geqslant \ldots \geqslant s_n \geqslant 0$ are the singular values de P.



$$\|P\|_F^2 = \sum_k s_k^2.$$

Matrix detection

⋆ Above which value of d can we reconstruct a consistent estimator of s_k?

Matrix detection

⋆ Above which value of d can we reconstruct a consistent estimator of s_k?

⋆ Fix $0 < \gamma < 1$. Find the smallest d such that there is with high probability an estimator $\hat{u}_k$ of u_k with

$$|\langle \hat{u}_k, u_k \rangle| \geqslant \gamma.$$

⋆ Above which value of d can we reconstruct a consistent estimator of s_k?

⋆ Fix $0 < \gamma < 1$. Find the smallest d such that there is with high probability an estimator $\hat{u}_k$ of u_k with

$$|\langle \hat{u}_k, u_k \rangle| \geqslant \gamma.$$

⋆ Fix $0 < \gamma < 1$. Find the smallest d such that there is with high probability an estimator $\hat{P}$ of P with

$$\mathrm{MSE}(\hat{P}) = \|P - \hat{P}\|_F^2 \geqslant \gamma \|P\|_F^2.$$

Applications

Numerous applications in global positioning, remote sensing, signal processing, computer vision, ... but the most famous is collaborative filtering.

Applications

Numerous applications in global positioning, remote sensing, signal processing, computer vision, ... but the most famous is collaborative filtering.

Guess what a user likes even before she knows it. The Netflix prize launched in 2006 consisted in minimizing the MSE (on a sample) with respect to the matrix:

$$P_{ij} = \text{ mark given by user } i \text{ on movie } j.$$

Estimation of a symmetric matrix

Estimation of a symmetric matrix

Let $P \in M_n(\mathbb{R})$ be a symmetric matrix.

Let $M = (M_{ij}) \in M_n(\mathbb{R})$ with $M_{ij} \in \{0, 1\}$ iid Bernoulli with parameter d/n.

Estimation of a symmetric matrix

Let $P \in M_n(\mathbb{R})$ be a symmetric matrix.

Let $M = (M_{ij}) \in M_n(\mathbb{R})$ with $M_{ij} \in \{0, 1\}$ iid Bernoulli with parameter d/n.

The non-symmetric observed matrix $A \in M_n(\mathbb{R})$ is

$$A_{ij} = \frac{n}{d} P_{ij} M_{ij}.$$

Estimation of a symmetric matrix

Let $P \in M_n(\mathbb{R})$ be a symmetric matrix.

Let $M = (M_{ij}) \in M_n(\mathbb{R})$ with $M_{ij} \in \{0, 1\}$ iid Bernoulli with parameter d/n.

The non-symmetric observed matrix $A \in M_n(\mathbb{R})$ is

$$A_{ij} = \frac{n}{d} P_{ij} M_{ij}.$$

The symmetric observed matrix $H \in M_n(\mathbb{C})$ is

$$H = \frac{A + A^*}{2}.$$

Estimation of a symmetric matrix

Let $P \in M_n(\mathbb{R})$ be a symmetric matrix.

Let $M = (M_{ij}) \in M_n(\mathbb{R})$ with $M_{ij} \in \{0, 1\}$ iid Bernoulli with parameter d/n.

The non-symmetric observed matrix $A \in M_n(\mathbb{R})$ is

$$A_{ij} = \frac{n}{d} P_{ij} M_{ij}.$$

The symmetric observed matrix $H \in M_n(\mathbb{C})$ is

$$H = \frac{A + A^*}{2}.$$

We have $\mathbb{E}A = \mathbb{E}H = P$, i.e. A and H are noisy unbiased observations of P.

Estimation of a symmetric matrix

We would like to compare the k-th largest eigenvalues of P and A or H and their corresponding eigenspaces

Estimation of a symmetric matrix

We would like to compare the k-th largest eigenvalues of P and A or H and their corresponding eigenspaces

It is smarter to consider the spectrum of A rather than the spectrum of H.

Estimation of a symmetric matrix

We would like to compare the k-th largest eigenvalues of P and A or H and their corresponding eigenspaces

It is smarter to consider the spectrum of A rather than the spectrum of H.

Benefits of asymmetry: in some situations, the spectrum of a matrix P is much less perturbed by a random asymmetric noise than by a random symmetric noise.

Detection threshold

We set

$$Q_{ij} = n|P_{ij}|^2 \qquad \text{and} \qquad \rho = \|Q\|.$$

The detection threshold is defined as

$$\theta = \max\left(\sqrt{\frac{\rho}{d}}, \frac{L}{d}\right),$$

with

$$L = n \max_{i,j} |P_{ij}|.$$

Incoherence

We order the real eigenvalues of P,

$$|\mu_1| \geqslant \cdots \geqslant |\mu_{r_0}| > \theta \geqslant |\mu_{r_0+1}| \geqslant \cdots \geqslant |\mu_n|.$$

In an ON basis of eigenvectors of P, for all $1 \leqslant k \leqslant r_0$,

$$\|\varphi_k\|_\infty = \max_i |\varphi_k(i)| \leqslant \frac{b}{\sqrt{n}}.$$

Stable numerical rank

The stable numerical rank is

$$r = \frac{\sum_k \mu_k^2}{\mu_1^2} \leqslant \operatorname{rank}(P).$$

In this talk, we assume that r, b, L, d, r_0 are $n^{o(1)}$. All results are quantitative but will be stated in an asymptotic way.

Estimation of eigenvalues

Recall:

$$A_{ij} = \frac{n}{d} P_{ij} M_{ij} \qquad \text{and} \qquad \theta = \max\left(\sqrt{\frac{\rho}{d}}, \frac{L}{d}\right).$$

Eigenvalues of P :

$$|\mu_1| \geqslant \cdots \geqslant |\mu_{r_0}| > \theta \geqslant |\mu_{r_0+1}| \geqslant \cdots \geqslant |\mu_n|.$$

Estimation of eigenvalues

Recall:

$$A_{ij} = \frac{n}{d} P_{ij} M_{ij} \qquad \text{and} \qquad \theta = \max\left(\sqrt{\frac{\rho}{d}}, \frac{L}{d}\right).$$

Eigenvalues of P :

$$|\mu_1| \geqslant \cdots \geqslant |\mu_{r_0}| > \theta \geqslant |\mu_{r_0+1}| \geqslant \cdots \geqslant |\mu_n|.$$

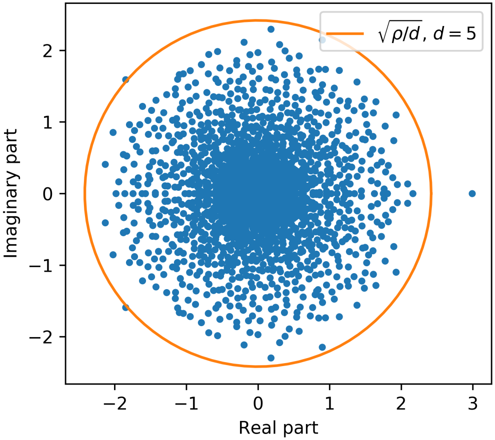
Theorem

With high probability, there exists an ordering of the eigenvalues $\lambda_1, \ldots, \lambda_n$ *of* A *such that*

$$\max_{1 \leqslant k \leqslant r_0} |\lambda_k - \mu_k| = o(1) \qquad \textit{and} \qquad \max_{r_0+1 \leqslant k \leqslant n} |\lambda_k| \leqslant (1 + o(1))\theta.$$

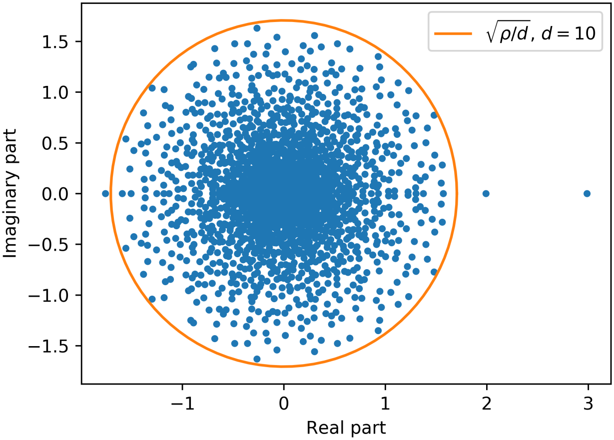
Simulation

For $n = 2000$ and $P = 3\varphi_1\varphi_1^* + 2\varphi_2\varphi_2^* + \varphi_3\varphi_3^*$ with φ_k uniform.



Simulation

For $n = 2000$ and $P = 3\varphi_1\varphi_1^* + 2\varphi_2\varphi_2^* + \varphi_3\varphi_3^*$ with φ_k uniform.



Estimation of eigenvectors

Recall:

$$A_{ij} = \frac{n}{d} P_{ij} M_{ij} \qquad \text{and} \qquad \theta = \max\left(\sqrt{\frac{\rho}{d}}, \frac{L}{d}\right).$$

We assume that the large eigenvalues of

$$P = \sum_k \mu_k \varphi_k \varphi_k^*$$

are well separated:

$$\left|1 - \frac{\mu_k}{\mu_l}\right| \geqslant \frac{\log d}{\log n} \qquad \text{for all} \quad 1 \leqslant k \neq l \leqslant r_0.$$

Estimation of eigenvectors

Recall:

$$A_{ij} = \frac{n}{d} P_{ij} M_{ij} \qquad \text{and} \qquad \theta = \max\left(\sqrt{\frac{\rho}{d}}, \frac{L}{d}\right).$$

We assume that the large eigenvalues of

$$P = \sum_k \mu_k \varphi_k \varphi_k^*$$

are well separated:

$$\left|1 - \frac{\mu_k}{\mu_l}\right| \geqslant \frac{\log d}{\log n} \qquad \text{for all } \; 1 \leqslant k \neq l \leqslant r_0.$$

Theorem

Let ψ_k be a unit eigenvector associated to k-th eigenvalue of A. There exists $\gamma_k > 0$ such that, with high probability, for $1 \leqslant k \leqslant r_0$,

$$|\langle \psi_k, \varphi_k \rangle| = \gamma_k + o(1).$$

Estimation of eigenvectors

The asymptotic scalar product $\gamma_k = |\langle \psi_k, \phi_k \rangle| + o(1)$ has an explicit formula:

$$\gamma_k = \frac{1}{\sqrt{\Gamma_{k,k}}}$$

with, for $1 \leqslant k, l \leqslant r_0$,

$$\Gamma_{k,l} = \sum_{i=1}^{n} w_{k,l}(i) \varphi_k(i) \varphi_l(i),$$

and

$$w_{k,l}(i) = \sum_j \left(I - \frac{Q}{\mu_k \mu_l d} \right)^{-1}_{i,j}.$$

Remark: $|\langle \psi_k, \psi_l \rangle| = |\Gamma_{k,l}| / \sqrt{\Gamma_{k,k} \Gamma_{l,l}} + o(1)$ is non-zero for $k \neq l$ if $\mathbf{1}$ is not an eigenvector of Q.

Rank one projector

If $P = \varphi\varphi^*$, we find

$$\theta = \sqrt{\frac{n \sum_i |\varphi(i)|^4}{d}}$$

$$\gamma = \sqrt{1 - \frac{n \sum_i |\varphi(i)|^4}{d}}.$$

Estimation of eigenvectors

It is also possible to compute the scalar between the left ψ'_k and right ψ_k unit eigenvectors of the k-th eigenvalue of A:

$$\langle \psi'_k, \psi_k \rangle = \gamma_k^2 + o(1) = \frac{1}{\Gamma_{k,k}} + o(1).$$

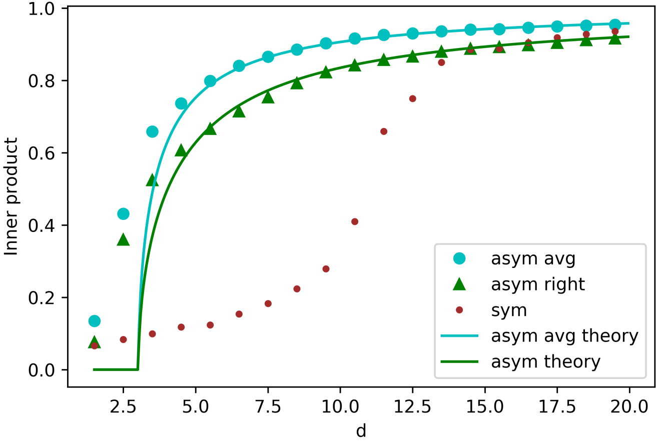
We get an estimator

$$\hat{\varphi}_k = \frac{\psi_k + \psi'_k}{\|\psi_k + \psi'_k\|_2}$$

such that

$$|\langle \varphi_k, \hat{\varphi}_k \rangle| = \sqrt{\frac{2\gamma_k^2}{1+\gamma_k^2}} + o(1).$$

For $n = 6000$ and $P = \varphi\varphi^*$ with φ uniform on the sphere.



Improved estimation with non-backtracking matrices

Put some symmetry back

We can improve the factor d in $2d$ in the detection threshold:

$$\theta = \max\left(\sqrt{\frac{\rho}{d}}, \frac{L}{d}\right).$$

We have not taken into account the information

$$P_{ij} = P_{ji}.$$

There is in fact an average of $2d$ observed entries per row.

Non-backtracking matrix

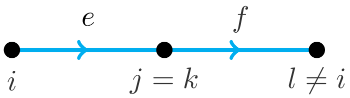
The set of symmetric observed entries is

$$E = \{(i,j) : (i,j) \text{ or } (j,i) \text{ is observed}\}.$$

We have $|E| \sim 2dn$.

We consider the non-symmetric matrix $B \in M_E(\mathbb{C})$ defined for all $(i,j),(k,l)$ in E by

$$B_{(i,j),(k,l)} = \frac{nP_{kl}}{2d}\mathbf{1}(j = k, l \neq i).$$



Non-backtracking matrix

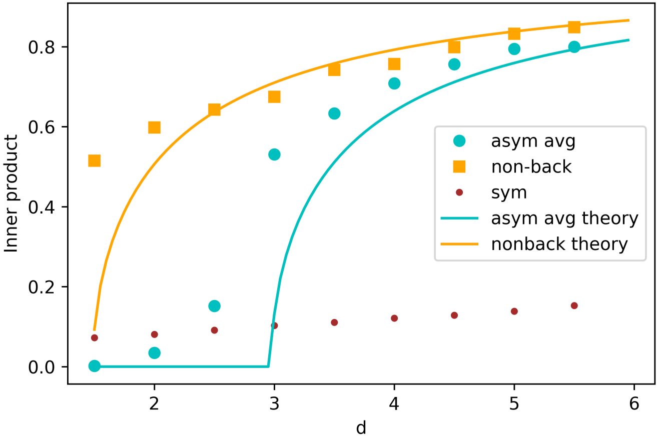
A vector $\varphi \in \mathbb{C}^n$ is lifted in $\mathbb{C}^E$ as

$$\varphi^+(i,j) = \varphi(j).$$

Theorem

The preceding results on A are true for the matrix B (with minor extra changes) with d replaced by $2d$.

For $n = 5000$ and $P = \varphi\varphi^*$ with φ uniform on the sphere.



The detection threshold

$$\theta = \max\left(\sqrt{\frac{\rho}{d}}, \frac{L}{d}\right).$$

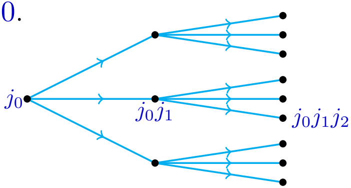
Lift of a matrix

Fix $j_0 \in [n] = \{1, \ldots, n\}$. Let V be the set of finite integer sequences in $[n]$, $(j_0, j_1, \ldots, j_k)$ starting with j_0.

We build an infinite matrix $\mathcal{P} = (\mathcal{P}_{uv})_{u,v \in V}$ by setting $u = (j_0, \ldots, j_k) \in V$ and $j \in [n]$

$$\mathcal{P}_{u,(u,j)} = P_{j_k,j}.$$

Otherwise $\mathcal{P}_{uv} = 0$.



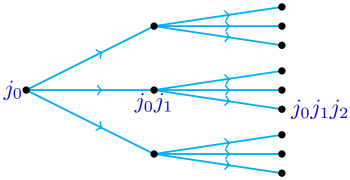
This defines a non-symmetric bounded operator on $\ell^2(V)$ build on an infinite n-ary tree.

If $P\varphi = \mu\varphi$ then Φ defined on V as

$$\Phi(j_0 \cdots j_k) = \varphi(j_k)$$

satisfies

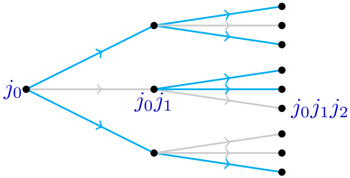
$$\mathcal{P}\Phi = \mu\Phi.$$



The function Φ is not in $\ell^2(V)$.

Percolation on the lift

We keep each edge with probability d/n.



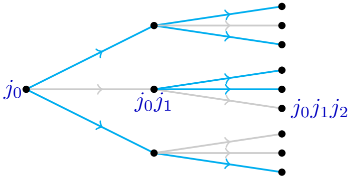
We denote by $\mathcal{P}_{\text{perc}}$ the corresponding operator and set

$$\mathcal{A} = \frac{n}{d}\mathcal{P}_{\text{perc}}.$$

The operator $\mathcal{A}$ is a local approximation of the matrix $A = (n/d)P \odot M$.

Percolation on the lift

$$\mathcal{A} = \frac{n}{d}\mathcal{P}_{\text{perc}}.$$



Since $\mathcal{P}\Phi = \mu\Phi$, for all $v \in V$, the process in $t \in \mathbb{N}$,

$$\Psi_t(v) = \mu^{-t}(\mathcal{A}^t\Phi)(v)$$

is a discrete martingale for the filtration of the successive generations in the tree.

The bracket of the martingale can be computed and we find:

$$\mathbb{E}|\Psi_{t+1}(v) - \Psi_t(v)|^2 = \frac{Q^t(\varphi^2)(v)}{(\mu^2 d)^t}.$$

Recall: $Q_{ij} = n|P_{ij}|^2$ and $\rho = \|Q\|$.

Hence $\Psi_t(v) = \mu^{-t}(\mathcal{A}^t \Phi)(v)$ converges a.s. and in L^2 toward $\Psi(v)$ if

$$|\mu| > \sqrt{\frac{\rho}{d}}.$$

This is called the Kesten-Stigum threshold.

Percolation on the lift

If $|\mu| > \sqrt{\rho/d}$, then $\Psi_t(v) = \mu^{-t}(\mathcal{A}^t\Phi)(v)$ converges a.s. and in L^2 toward $\Psi(v)$.

Since

$$\mathcal{A}\Psi_t = \mu\Psi_{t+1},$$

we can define a.s. a random eigenwave Ψ on V which satisfies

$$\mathcal{A}\Psi = \mu\Psi.$$

Back to finite dimension

This analysis and concentration inequalities allow to show that if $t \gg 1$ but not too large,

$$\|A^{t+1}\varphi_k - \mu_k A^t \varphi_k\|_2 = o(\|A^t \varphi_k\|_2).$$

Similarly $\varphi_k^* A^t$ is an approximate left eigenvector.

We decompose A^t in

$$A^t = \sum_{k=1}^{r_0} \mu_k^t u_k v_k^* + R_t,$$

with $u_k = A^t \varphi_k / \mu_k^t$ and $v_k = (A^t)^* \varphi_k / \mu_k^t$. We have

$$\langle u_k, v_l \rangle = \delta_{kl} + o(1).$$

Proof strategy

For $t = c \ln n / \log d$ well chosen,

$$A^t = \sum_{k=1}^{r_0} \mu_k^t u_k v_k^* + R_t.$$

- ⋆ Compute the inner products between these r_0^2 vectors;
- ⋆ Show that the Gram matrix is well-conditioned;
- ⋆ Show that $\|R_t\| \leqslant (\log n)^c \theta^t$;
- ⋆ Use an ad-hoc spectral perturbation theorem of a non-symmetric matrix of Bauer-Fike type.

B-Lelarge-Massoulié 18.

Rectangular matrices

Linearization trick

If $P \in M_{m,n}(\mathbb{C})$, the matrix

$$\widetilde{P} = \begin{pmatrix} 0 & P \\ P^* & 0 \end{pmatrix}$$

is of size $(m+n) \times (m+n)$ and is Hermitian.

The singular value decomposition of $P = \sum_k s_k u_k v_k^*$ is equivalent to the diagonalization of $\widetilde{P}$:

$$\widetilde{P} = \sum_k s_k w_k^+ (w_k^+)^* - s_k w_k^- (w_k^-)^*,$$

with $w_k^\pm = (u_k, \pm v_k)'/\sqrt{2}$.

Recall

$$P = \sum_k s_k u_k v_k^*.$$

Consider $Z = (Z_{ij}) \in M_{m,n}(\mathbb{R})$ with iid $\{0,1\}$-Bernoulli entries with parameter $1/2$ and define

$$P_1 P_2^* \quad \text{with} \quad P_1 = P \odot Z\,,\; P_2 = P - P_1.$$

The k-th largest eigenvalue, say λ_k, of $P_1 P_2^*$ is a proxy for $s_k^2/4$.

The average of the left and right eigenvectors associated to λ_k is a proxy for the left singular vector u_k.

Matrix completion

Let $M = (M_{ij}) \in M_{m,n}(\mathbb{R})$ with iid $\{0,1\}$-Bernoulli entries with parameter d/n.

The observed matrix is

$$A = \frac{n}{d} P \odot M.$$

We perform the randomized asymmetric SVD on A.

At a higher computational cost, we may also consider the non-backtracking matrix associated to the linearized matrix $\widetilde{A}$.

In either case, if $n \asymp m$, we have explicit detection thresholds and formulas for the asymptotic inner products.

Recall

$$P = \sum_k s_k u_k v_k^*.$$

Once we have estimators $\hat{u}_k, \hat{v}_k$ of u_k and v_k, it is possible to design an estimator of P:

$$\hat{P} = \sum_{k=1}^{r_0} x_k \hat{u}_k \hat{v}_k^*$$

for some vector $x = (x_k) \in \mathbb{R}^{r_0}$ which asymptotically minimizes

$$\|\hat{P} - P\|_F$$

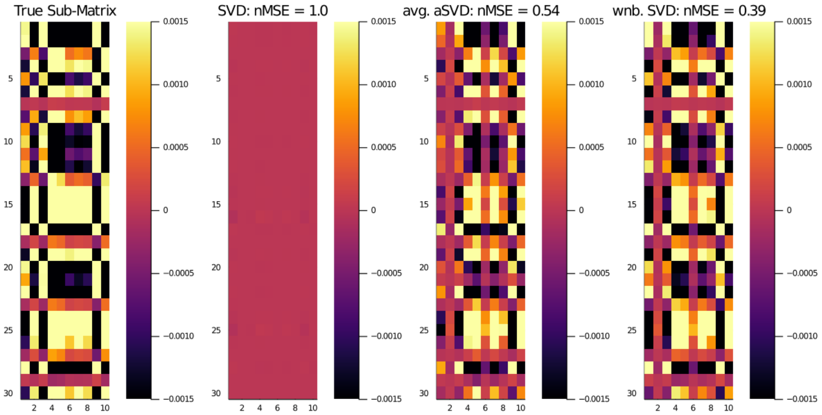
and compute an explicit asymptotic formula for

$$\text{MSE}(\hat{P}) = \|\hat{P} - P\|_F^2.$$

Nadakuditi 14.

Simulation

We take $d = 9.7$, $(m, n) = (2000, 3000)$ and $P = uv^*$ with u, v independent standard Gaussian vectors.



Concluding words

Conclusion

Spectral analysis methods on random non-symmetric matrices can be very efficient, *Chen-Cheng-Fan 18.*

Numerous possible extensions, for example include some extra noise, or models where the probability of observing an entry depends on the entry, *Stephan-Massoulié 20.*

There is nowadays a lot of activities on tensor completion.

Thank you for your attention!