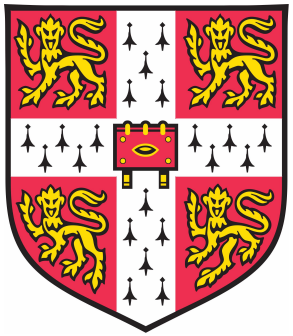
Trade-offs in Statistical Learning

Quentin Berthet - University of Cambridge

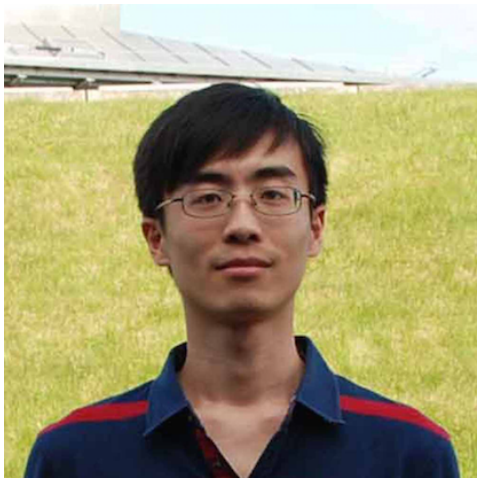


Computational and Statistical Trade-offs in Learning

Institut des Hautes Etudes Scientifiques - 2016

P. Rigollet (MIT)

R. Samworth (Cambridge)



T. Wang (Cambridge)

V. Chandrasekaran (Caltech)

J. Ellenberg (UW Madison)

"Big Data"

- Broad phenomenon, description of challenges in data sciences.
- Important aspect: Data collected without discernment
  - Most of the data not relevant to the problem at hand.
  - Data can be complex: heterogeneity, privacy, errors.
- Tradeoffs in inferential problems.

Libraries of Babel, past and present

Computational aspects of high-dimensional statistics

- Flood of data, high dimensional problems:
  - Higgs boson: 800 trillion events/year. Genome: 3,000 megabase pairs.
  - **Data:** $X_1, \ldots, X_n$ **High-dimensional parameter:** $X_i \sim \mathbf{P}_\theta$, $\theta \in \mathbf{R}^d$.
- Structure, sparsity $\rightarrow$ combinatorial problems in likelihood methods.
  - **Estimation** $\rightarrow$ **Optimization**

$$\hat{\theta} \in \operatorname*{argmax}_{\theta \in \mathcal{S}} \ell_{\mathbf{X},n}(\theta) \, .$$

  - **Hypothesis testing** $\rightarrow$ **Averages**

$$\Psi(\mathbf{X}, n) : \quad \frac{1}{|\mathcal{S}_1|} \sum_{\theta \in \mathcal{S}_1} L_{\mathbf{X},n}(\theta) > \frac{1}{|\mathcal{S}_0|} \sum_{\theta \in \mathcal{S}_0} L_{\mathbf{X},n}(\theta) \, .$$

- Objective: computationally efficient statistical methods.

Computational aspects of high-dimensional statistics

- Objective: computationally efficient statistical methods.

- Computational limits in statistics:
  - Worst-case hardness: 0th order information.
  - Algorithm for frequent instances of NP-hard problem:

    Clustering, Nonconvex regression, Alternating minimization.
  - Proxy functional and optimization problem.

    Lasso, Convex relaxations.

- Average-case hypotheses:

  Some task is hard to achieve consistently and efficiently.

Statistics & Computation

- Complex models in statistical problems.
- Structure assumptions (sparsity, clusters, etc.) $\rightarrow$ hard optimization problems.
- Can we design inference procedures with

Statistical performance

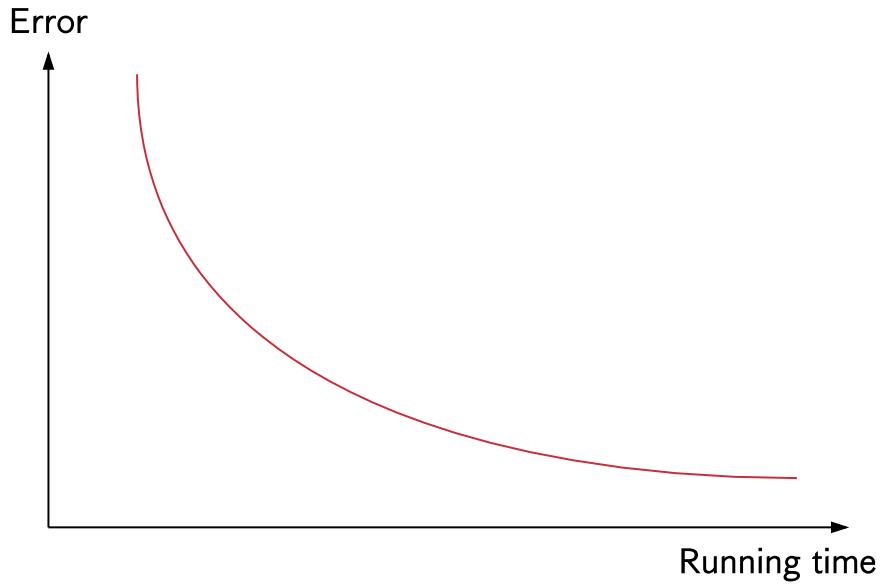
Algorithmic efficiency

Statistics & Computation

- Computational trade-offs
  - Computational lower bounds: gaps in performance for tractable methods.

    Blum et al. (03), B., Rigollet (12,13), Ma, Wu (13), Deshpande, Montanari (14), Wang et al. (16),...
  - Smooth trade-offs: analysis of a given method or framework.

    Chandrasekaran, Jordan (13), Feldman et al (13), Tropp, Bruer (15),...

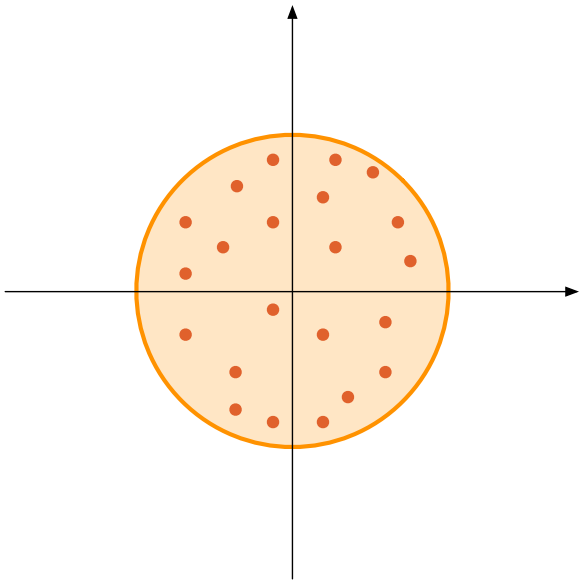


Sparse principal component detection

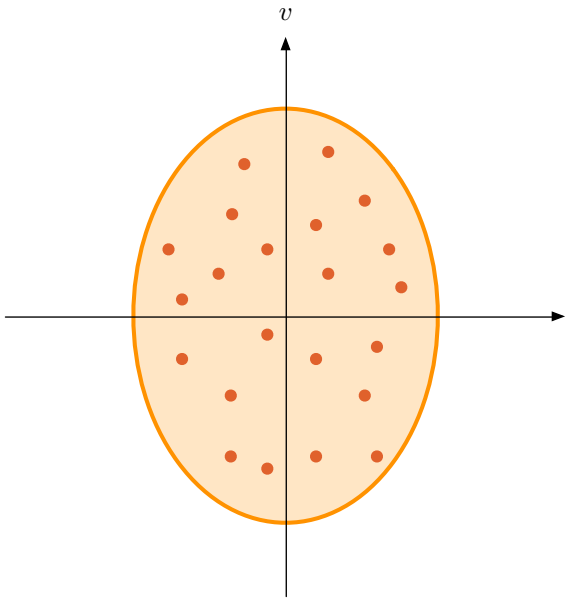
$X_1, \dots, X_n \in \mathbf{R}^d$ independent, centered Gaussian with unknown covariance.

$$\begin{cases} H_0 : I_d \\ H_1 : I_d + \theta v v^\top, \quad v \in \mathcal{B}_0(k) \end{cases}$$

v is a k-sparse unit vector. $\mathcal{B}_0(k) = \{v \in \mathbf{R}^d \, : \, |v|_2 = 1 \, , \, |v|_0 \leq k\}$.



Isotropy: $\mathcal{N}(0, I_d)$



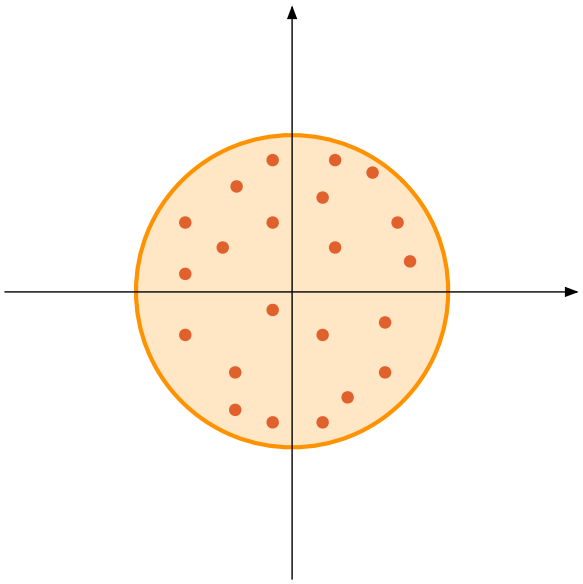
Sparse PC: $\mathcal{N}(0, I_d + \theta \, v v^\top)$

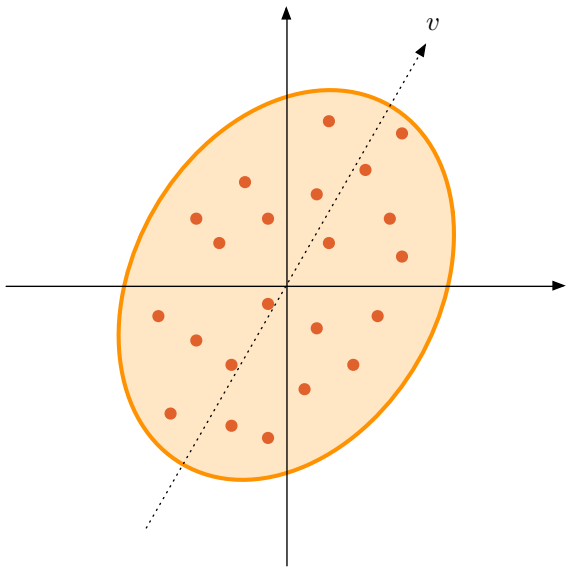
Importance of sparsity assumption

General problem of principal component detection: **Onatski et al. (13)**

$$\begin{cases} H_0 : I_d \\ H_1 : I_d + \theta v v^\top, \quad v \in \mathcal{S}^{d-1} \end{cases}$$

v is a unit vector. $\mathcal{S}^{d-1} = \{v \in \mathbf{R}^d \, : \, |v|_2 = 1\}$.





Isotropy: $\mathcal{N}(0, I_d)$

Principal Component: $\mathcal{N}(0, I_d + \theta\, vv^\top)$

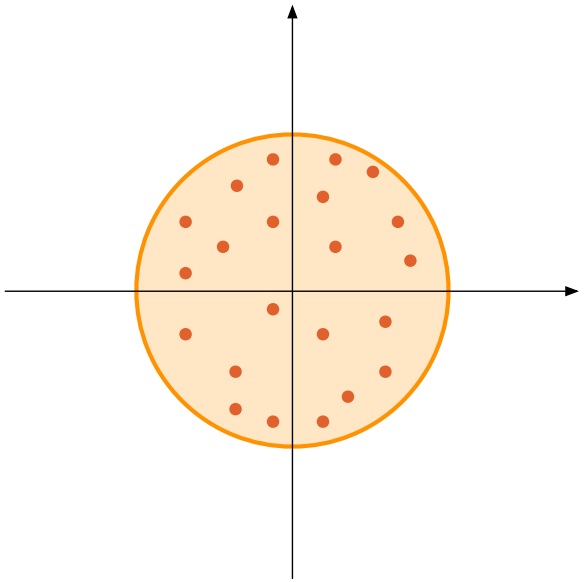
BBP transition of **Baik, Ben Arous, Péché (05)**. Strong signal: $\theta > \sqrt{\frac{d}{n}}$.

Sparse principal component detection

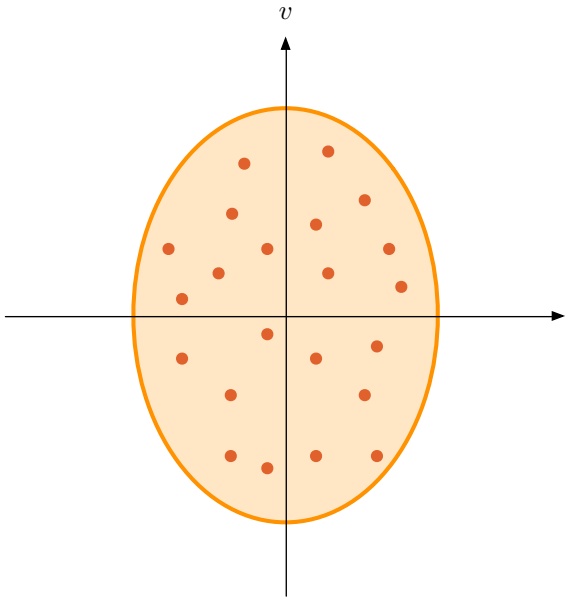
Testing problem between two hypotheses.

$$\begin{cases} H_0 : I_d \\ H_1 : I_d + \theta v v^\top, \quad v \in \mathcal{B}_0(k) \end{cases}$$

v is a k-sparse unit vector. $\mathcal{B}_0(k) = \{v \in \mathbf{R}^d \, : \, |v|_2 = 1 \, , \, |v|_0 \leq k\}$.



Isotropy: $\mathcal{N}(0, I_d)$



Sparse PC: $\mathcal{N}(0, I_d + \theta \, v v^\top)$

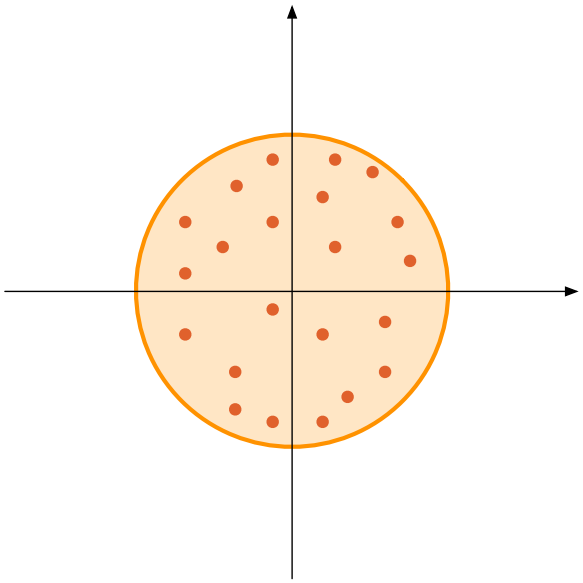
The **signal strength** is θ, quantifies the distance between the distributions.

Sparse principal component detection

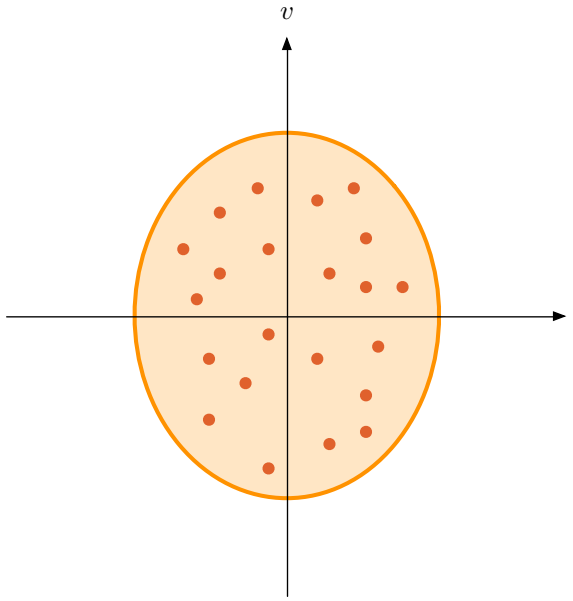
Testing problem between two hypotheses.

$$\begin{cases} H_0 : I_d \\ H_1 : I_d + \theta v v^\top, \quad v \in \mathcal{B}_0(k) \end{cases}$$

v is a k-sparse unit vector. $\mathcal{B}_0(k) = \{v \in \mathbf{R}^d \, : \, |v|_2 = 1 \, , \, |v|_0 \leq k\}$.



Isotropy: $\mathcal{N}(0, I_d)$



Sparse PC: $\mathcal{N}(0, I_d + \theta \, v v^\top)$, small θ

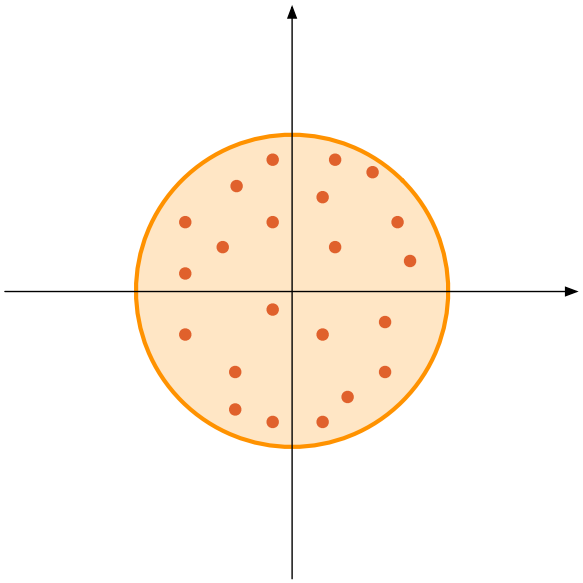
The **signal strength** is θ, quantifies the distance between the distributions.

Sparse principal component detection

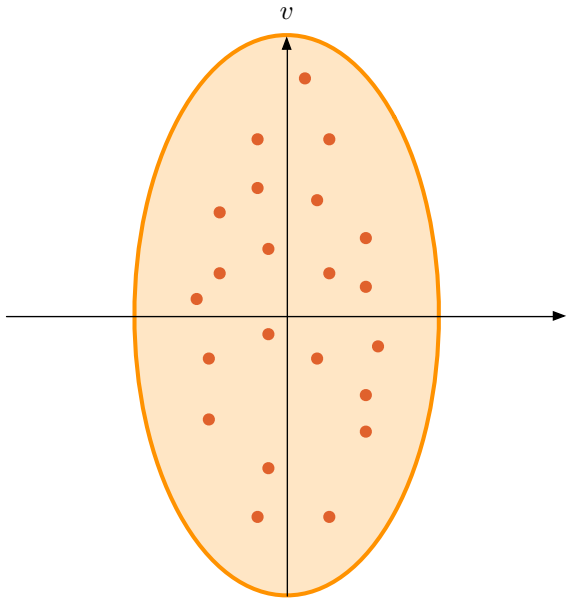
Testing problem between two hypotheses.

$$\begin{cases} H_0 : I_d \\ H_1 : I_d + \theta v v^\top, \quad v \in \mathcal{B}_0(k) \end{cases}$$

v is a k-sparse unit vector. $\mathcal{B}_0(k) = \{v \in \mathbf{R}^d \, : \, |v|_2 = 1 \, , \, |v|_0 \leq k\}$.



Isotropy: $\mathcal{N}(0, I_d)$



Sparse PC: $\mathcal{N}(0, I_d + \theta \, v v^\top)$, large θ

The **signal strength** is θ, quantifies the distance between the distributions.

Results - Detection

Statistics

- Maximizing $x^\top \hat{\Sigma} x$ over unit sparse vectors, $\lambda^k_{\max}(\hat{\Sigma})$. Test powerful in regime

$$\theta \gtrsim \sqrt{\frac{k \log(d)}{n}}.$$

- Lower bounds of the same order, optimal result.

Computations

- SDP relaxation of $\lambda^k_{\max}(\hat{\Sigma})$ by **d'Aspremont et al (07)**, requires

$$\theta \gtrsim \sqrt{\frac{k^2 \log(d)}{n}}.$$

- Can be better than $\lambda_{\max}(\hat{\Sigma})$, but still a suboptimal result.

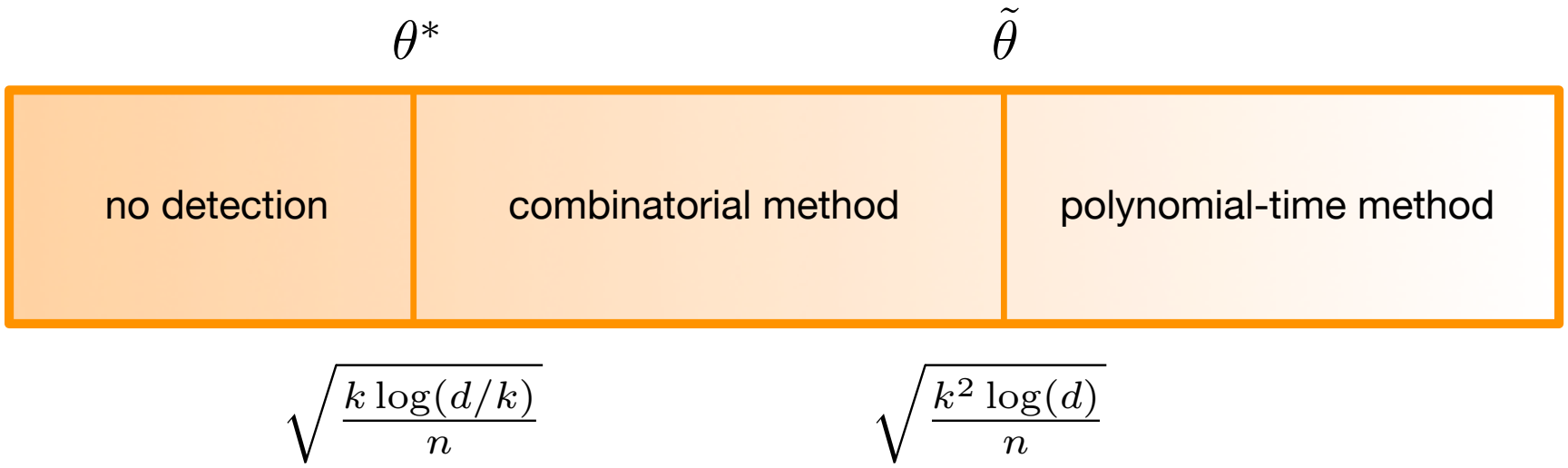
Overall picture - Testing

Computationally efficient tests seem to require

$$\theta \approx \sqrt{\frac{k^2 \log(d)}{n}}$$

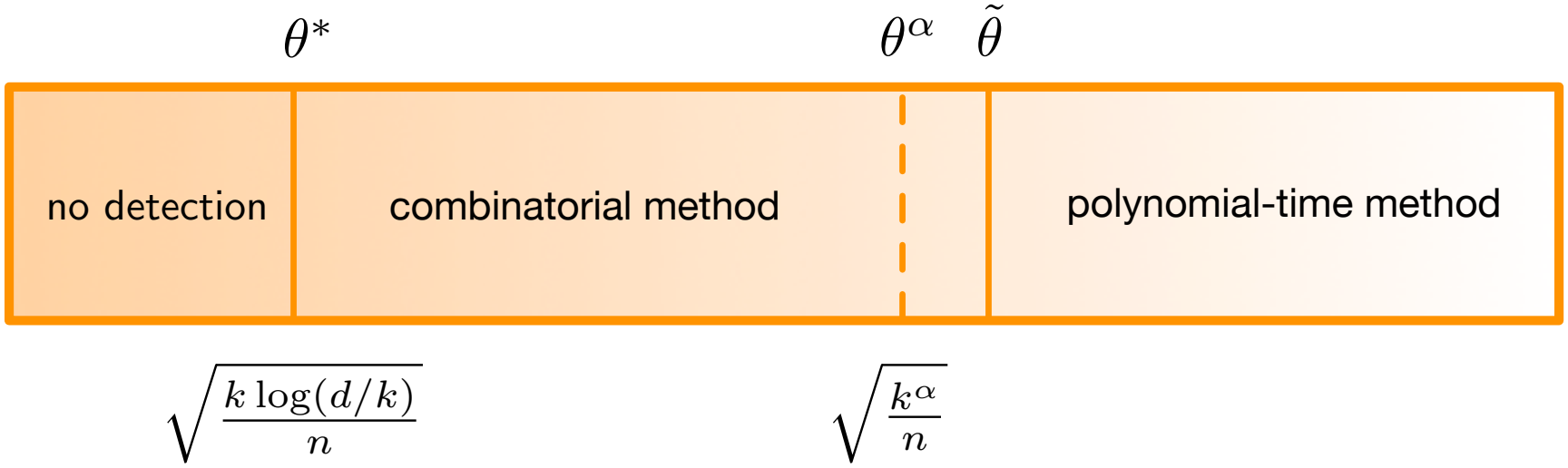
Rate observed for several explicit testing methods.

Tests: Diagonal method - **Johnstone (01)**, SDP - **d'Aspremont et al. (07)**, MDP - **Berthet and Rigollet (12)**, other heuristics.



Detection rates

So far, only upper bounds, suggestions



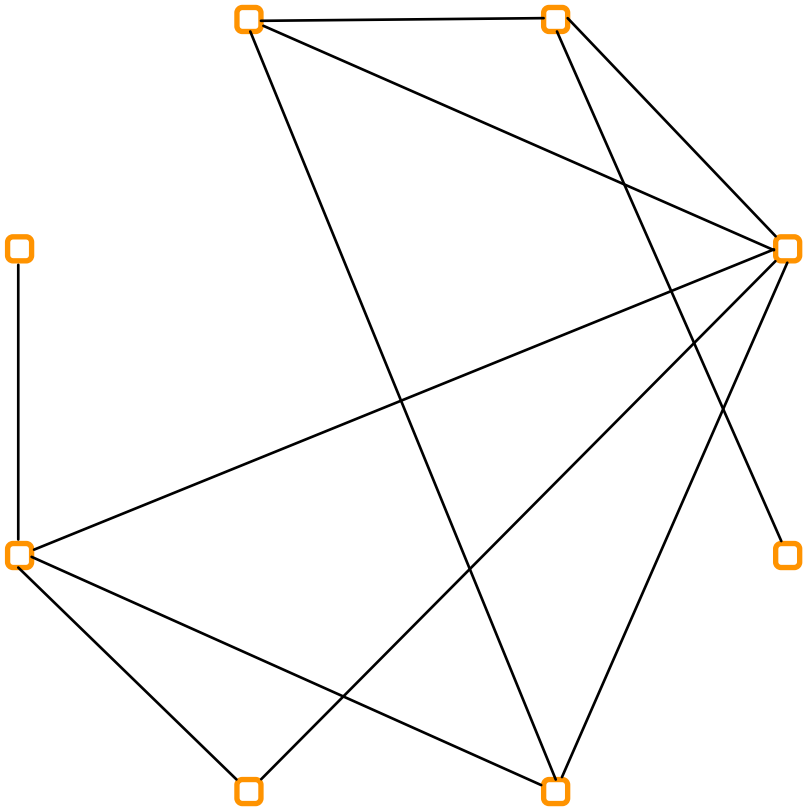
Situation could be very different, with an algorithm requiring only θ^α for $\alpha < 2$.

Need for **Complexity Theoretic Lower Bounds**

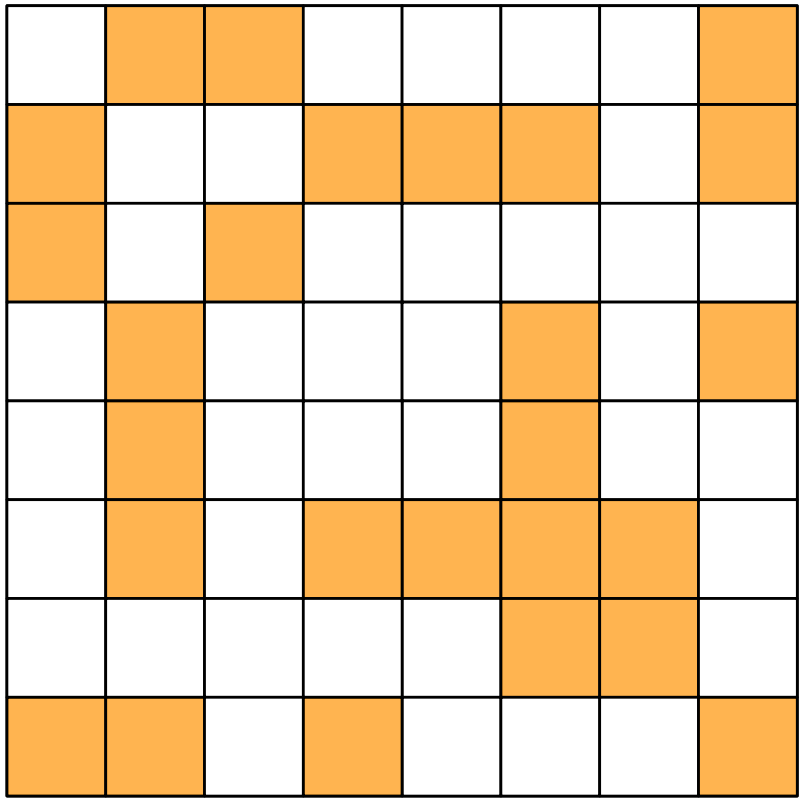
Planted clique problem

Erdős-Rényi graphs

$\mathcal{G}(m, 1/2)$: Each edge is randomly connected, with probability $1/2$, independently.



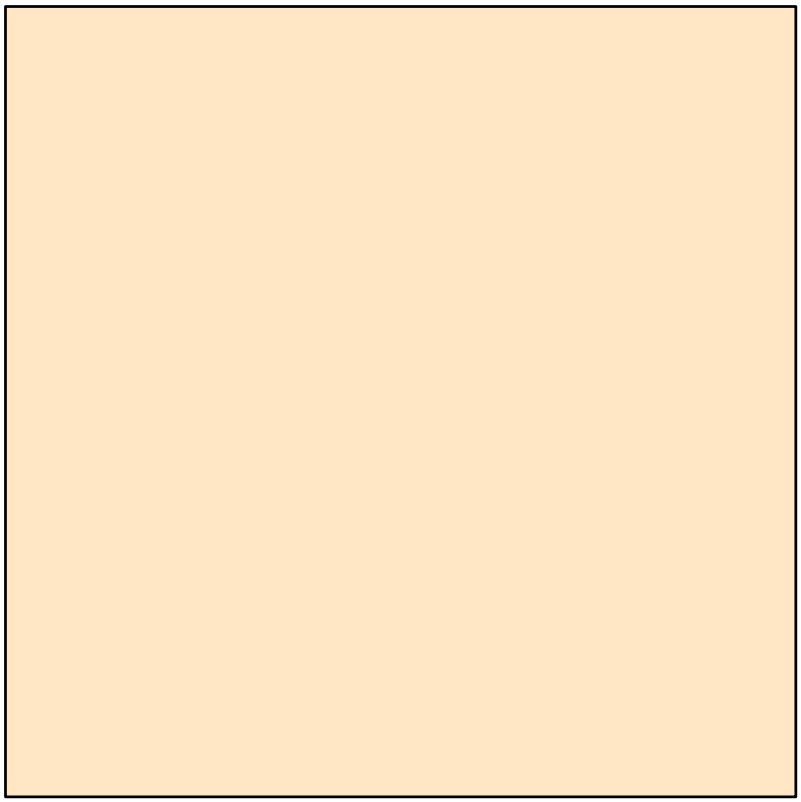
Graph



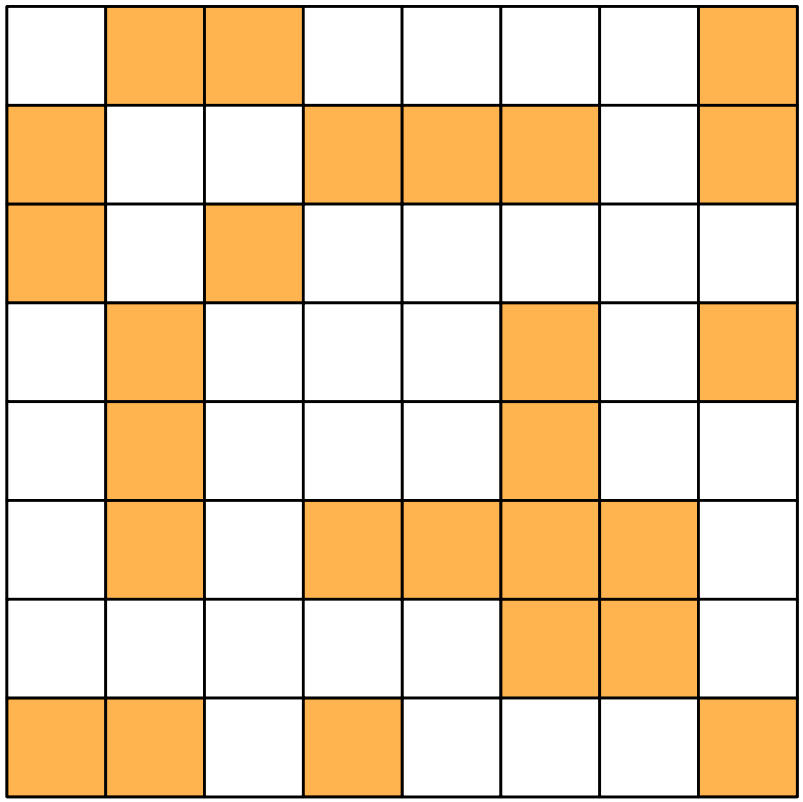
Adjacency matrix

Erdős-Rényi graphs

The expectation of the adjacency matrix is constant: pure noise setting.



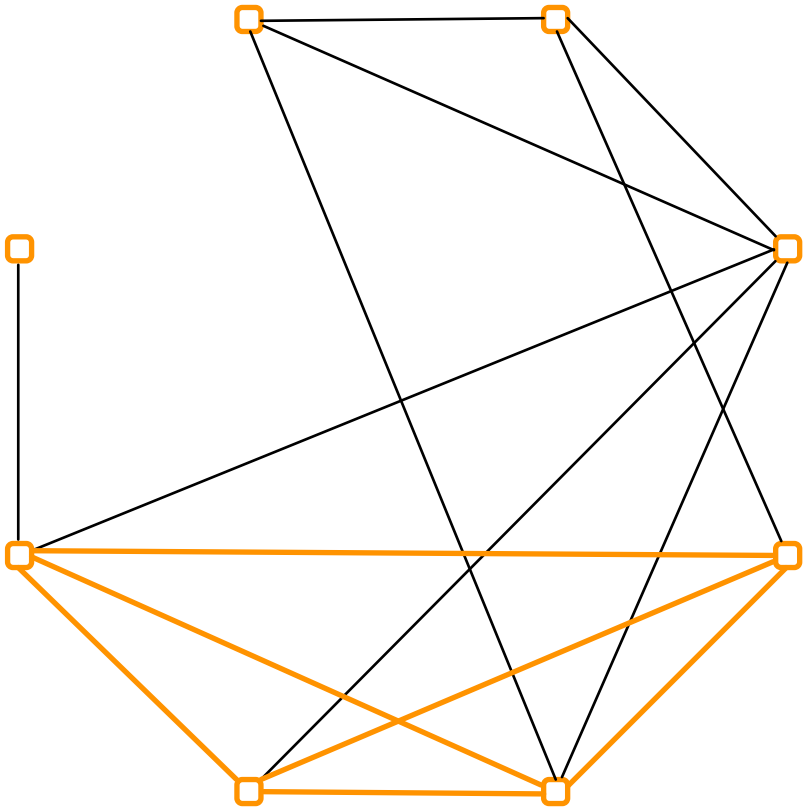
Expectation



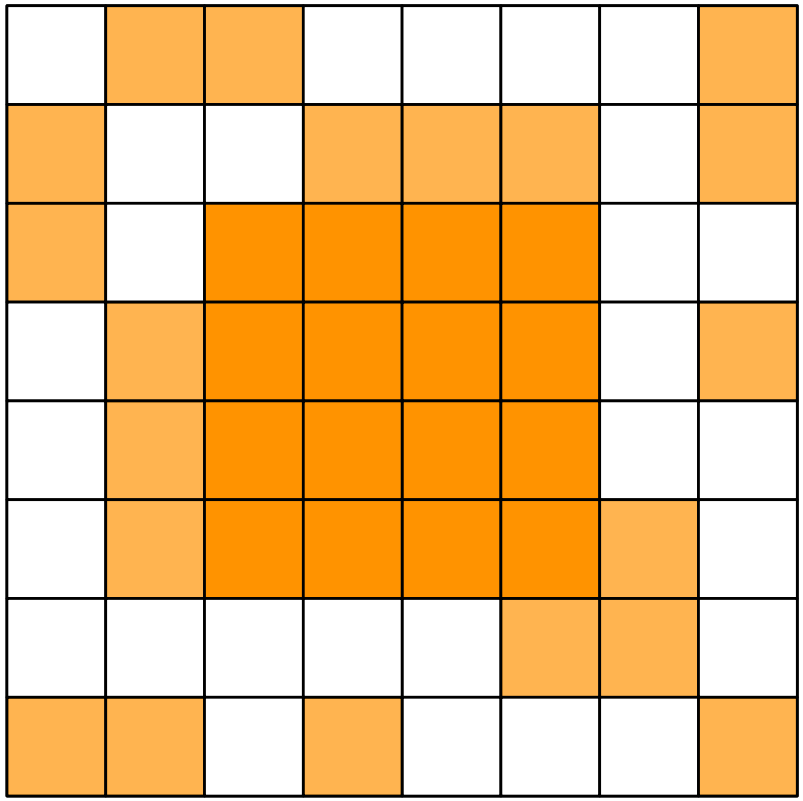
Random instance

Erdős-Rényi graphs

$\mathcal{G}(m, 1/2, \kappa)$: A clique of size κ is planted in a graph from $\mathcal{G}(m, 1/2)$.



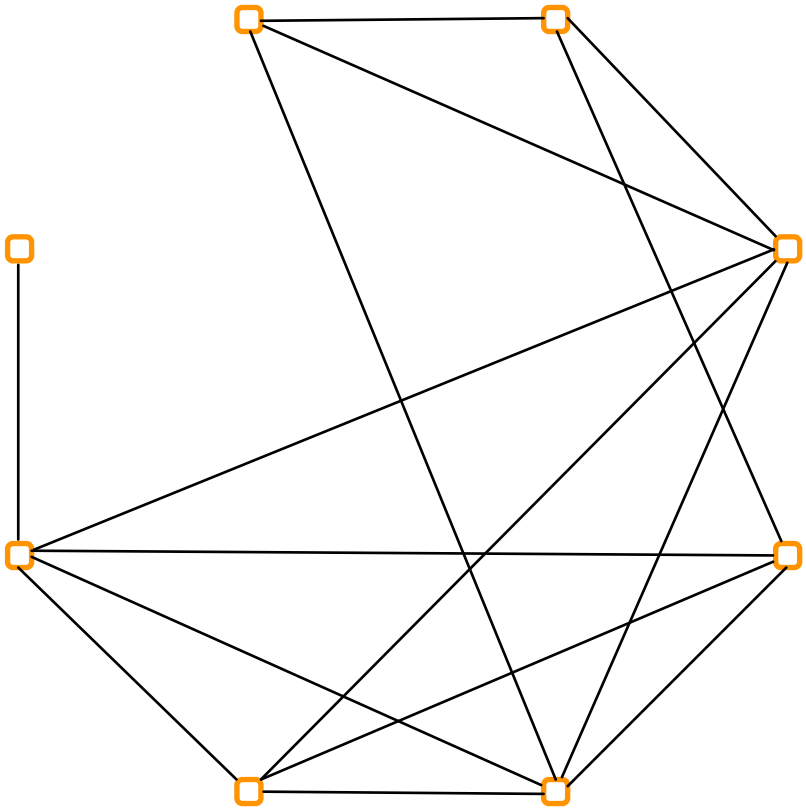
Graph



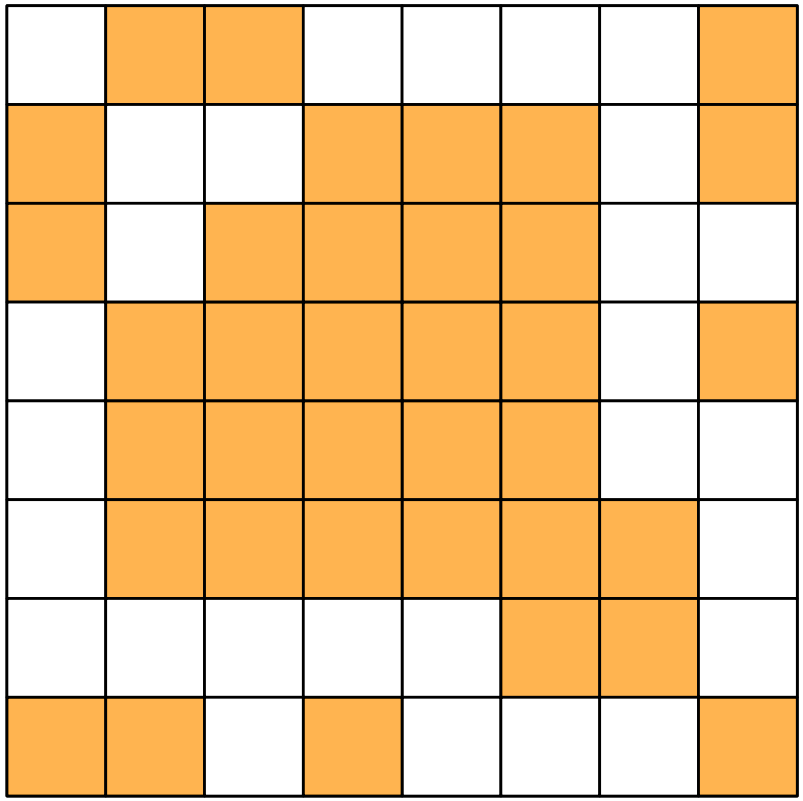
Adjacency matrix

Erdős-Rényi graphs

$\mathcal{G}(m, 1/2, \kappa)$: A clique of size κ is planted in a graph from $\mathcal{G}(m, 1/2)$.



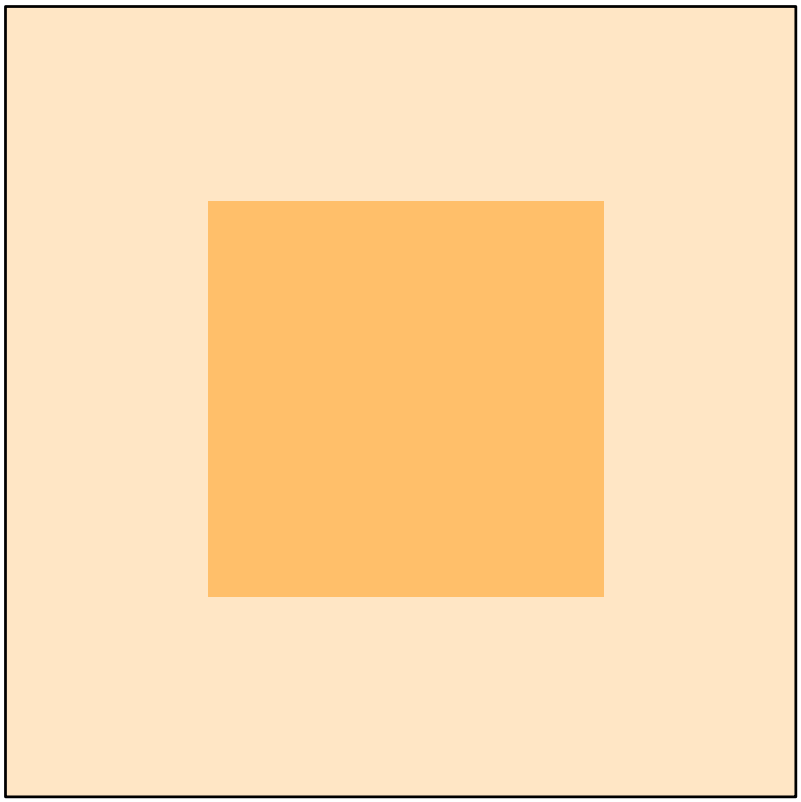
Graph



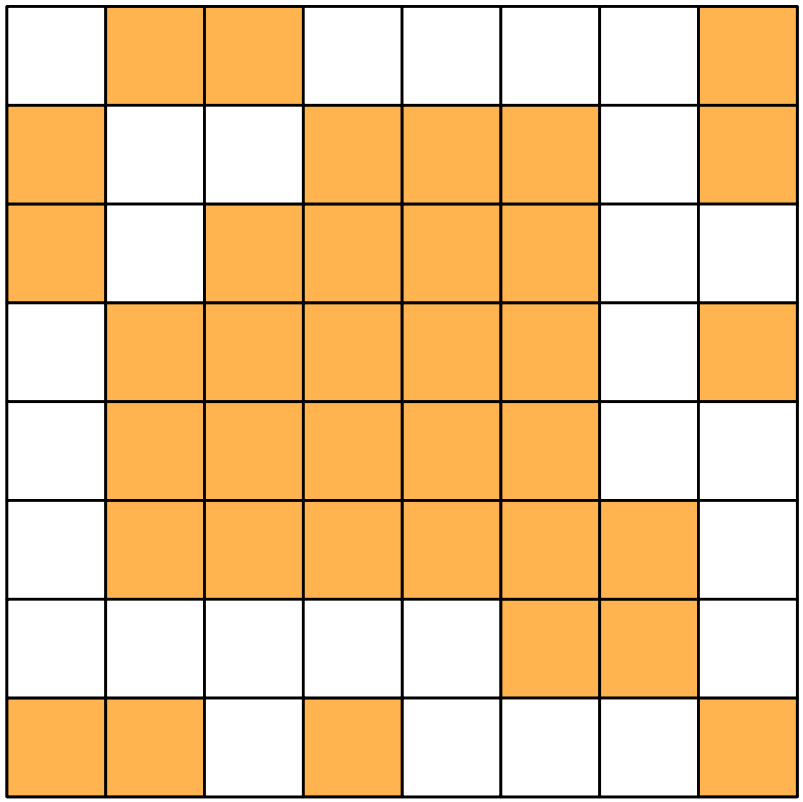
Adjacency matrix

Erdős-Rényi graphs

The expectation of the adjacency matrix has a sparse signal structure.



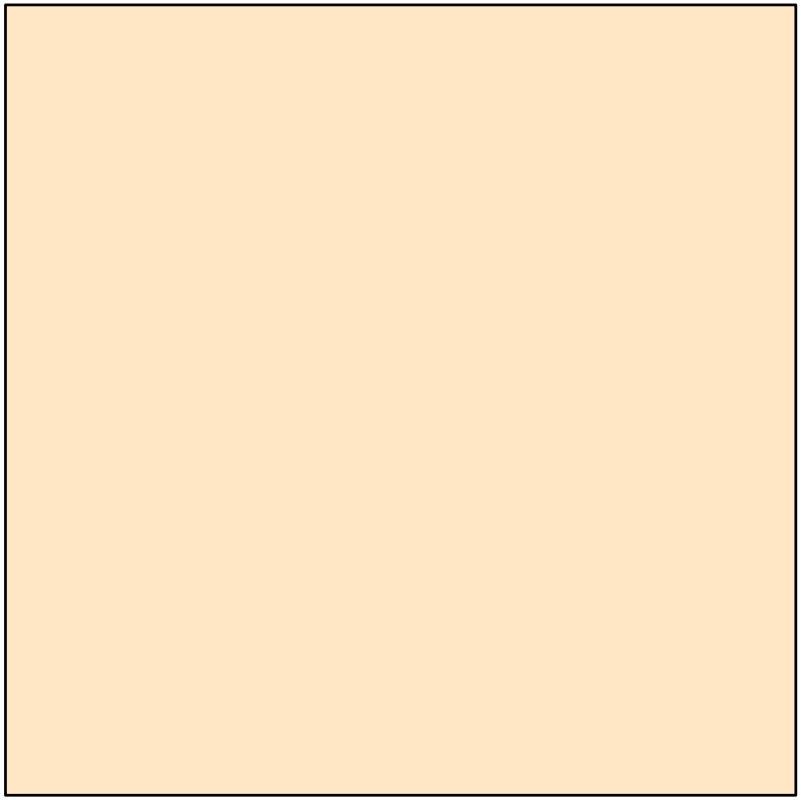
Expectation



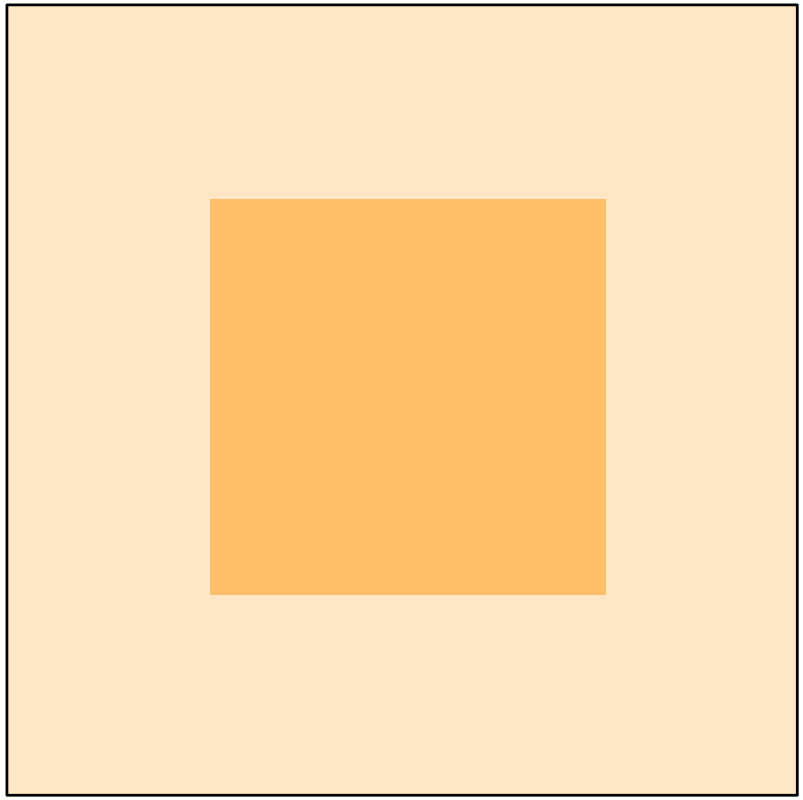
Random instance

Planted clique problem

Detection of a structured signal of sparsity κ in a random graph of size m.



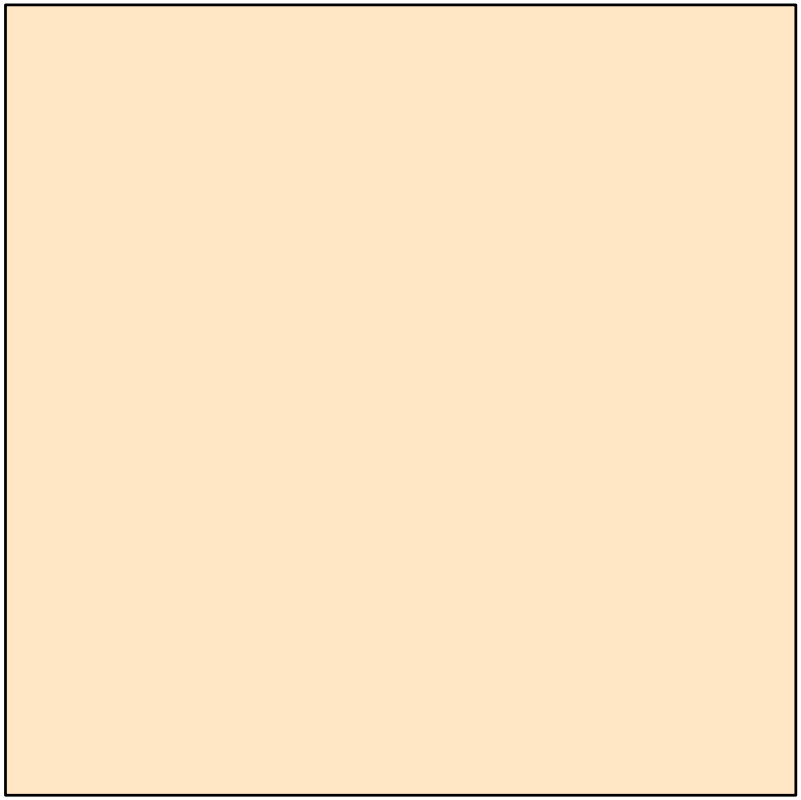
$$\mathcal{G}(m, 1/2) = \mathbf{P}_0^{(G)}$$



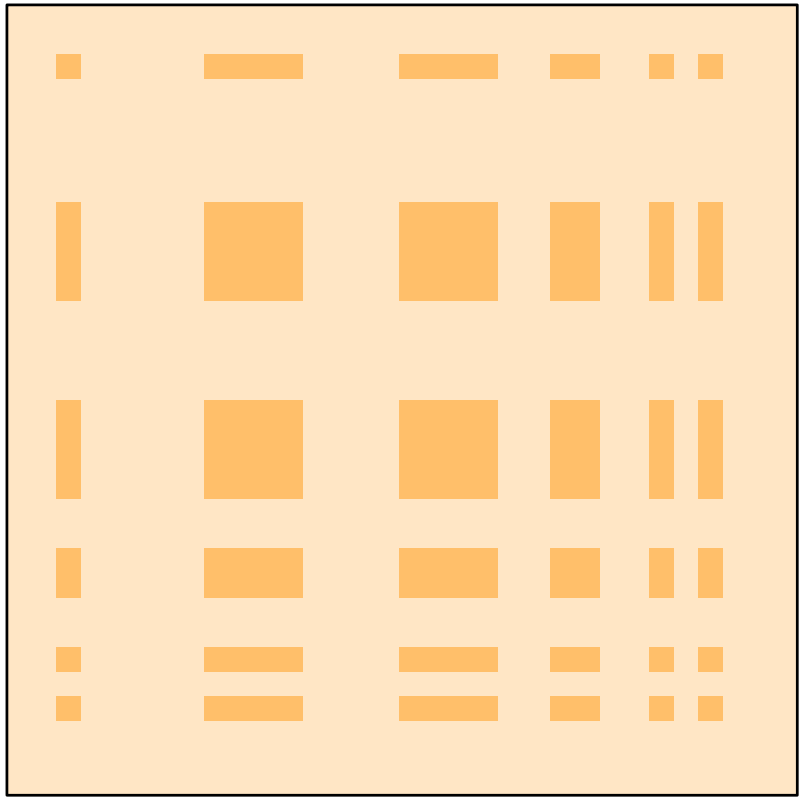
$$\mathcal{G}(m, 1/2, \kappa) = \mathbf{P}_1^{(G)}$$

Planted clique problem

Detection of a structured signal of sparsity κ in a random graph of size m.



$$\mathcal{G}(m, 1/2) = \mathbf{P}_0^{(G)}$$



$$\mathcal{G}(m, 1/2, \kappa) = \mathbf{P}_1^{(G)}$$

Planted clique problem

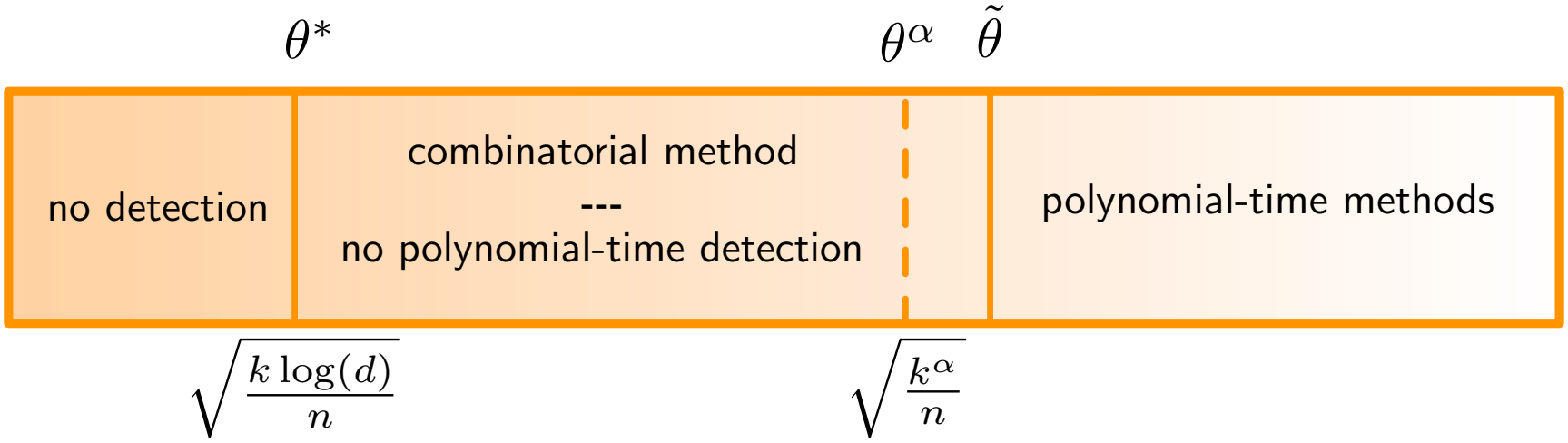
Distinguishing those two distributions is called the Planted Clique problem

$$\begin{cases} H_0 : G \sim \mathcal{G}(m, 1/2), & \omega(G) < 2\log_2(m) \\ H_1 : G \sim \mathcal{G}(m, 1/2, \kappa)\,. & \omega(G) \geq \kappa \end{cases}$$

- Detection for $\kappa > 2\log_2(m)$, **Erdős and Rényi (59)**.
- Polynomial time detection for $\kappa \geq C\sqrt{m}$, spectral method **Alon et al. (98)**.
- Hypothesis: detection impossible in polynomial time for $\kappa = O(m^c)$, $c < 1/2$.
  - **Ames and Vavasis (11), Dekel et al. (10), Feige et al. (00)**
  - **Jerrum (92), Feige and Krauthgamer (92), Feldman et al. (12)**
  - **Juels and Peinado (00), Alon et al. (07), Hazan et al. (11)**
- Primitive for **Average-case reduction**: link between problems.

Results - Detection

- Optimal signal level: information-theoretic barriers, combinatorial method.
- Link between sparse PCA and planted clique **B., Rigollet (12, 13)**



Take-home message: gap of $\sqrt{k}$ for computationally efficient methods.

Results - Estimation

- Optimal signal level: information-theoretic barriers, combinatorial method.

$$\mathbf{E}[\|\hat{v}\hat{v}^\top - vv^\top\|_F] \approx \frac{1}{\theta}\sqrt{\frac{k\log(d)}{n}}\,.$$

- Link between estimating v and clique recovery **Wang, B., Samworth (14)**

$$\mathbf{E}[\|\hat{v}\hat{v}^\top - vv^\top\|_F] \approx \frac{1}{\theta}\sqrt{\frac{k^2\log(d)}{n}}\,.$$

Take-home message: gap of $\sqrt{k}$ for computationally efficient methods.

Statistics & Computation

- Not a universal situation
  - Power of convex relaxation: Lasso, atomic norms
  - Algorithms for frequent instances:

    SGD for Non-convex regression, alternating minimization.

Statistical performance + Algorithmic efficiency

Statistics & Something else

- Other desired aspect/constraint of statistical procedure
  - Privacy, data security. **Agrawal, Skirant (00), Duchi et al. (13-15),...**
  - Robustness to errors. **Candes et al. (09), Loh, Wainwright (12),...**
  - Distributed data. **Zhang et al. (13), Bühlmann, Meinshausen (14),...**

Statistical performance + Some other thing

Statistics & Computation & Something else

- Study of multiple trade-offs

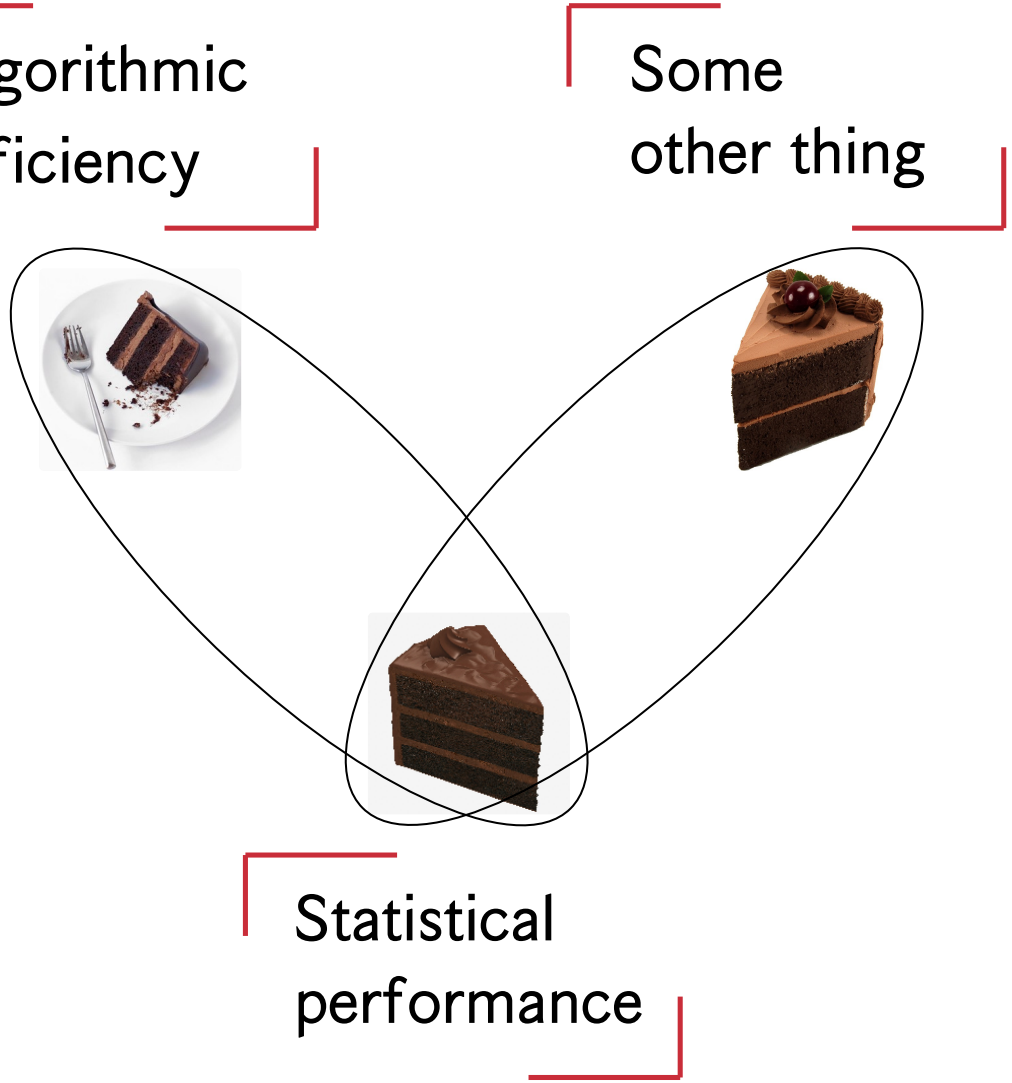
Algorithmic efficiency

Some other thing

Statistical performance

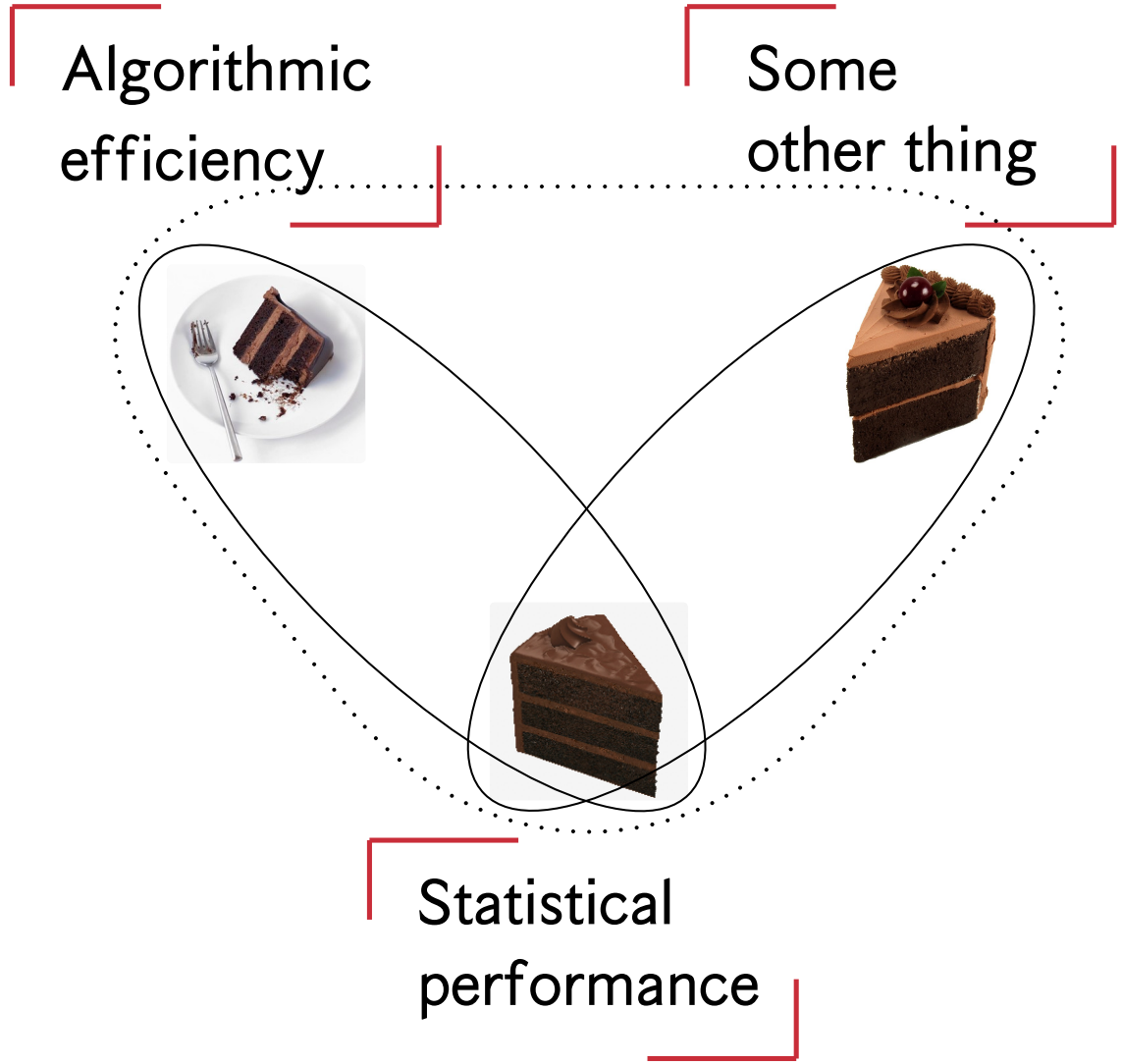
Statistics & Computation & Something else

- Study of multiple trade-offs



Statistics & Computation & Something else

- Study of multiple trade-offs



Distributed estimation - Sparse matrix signals

Distributed estimation

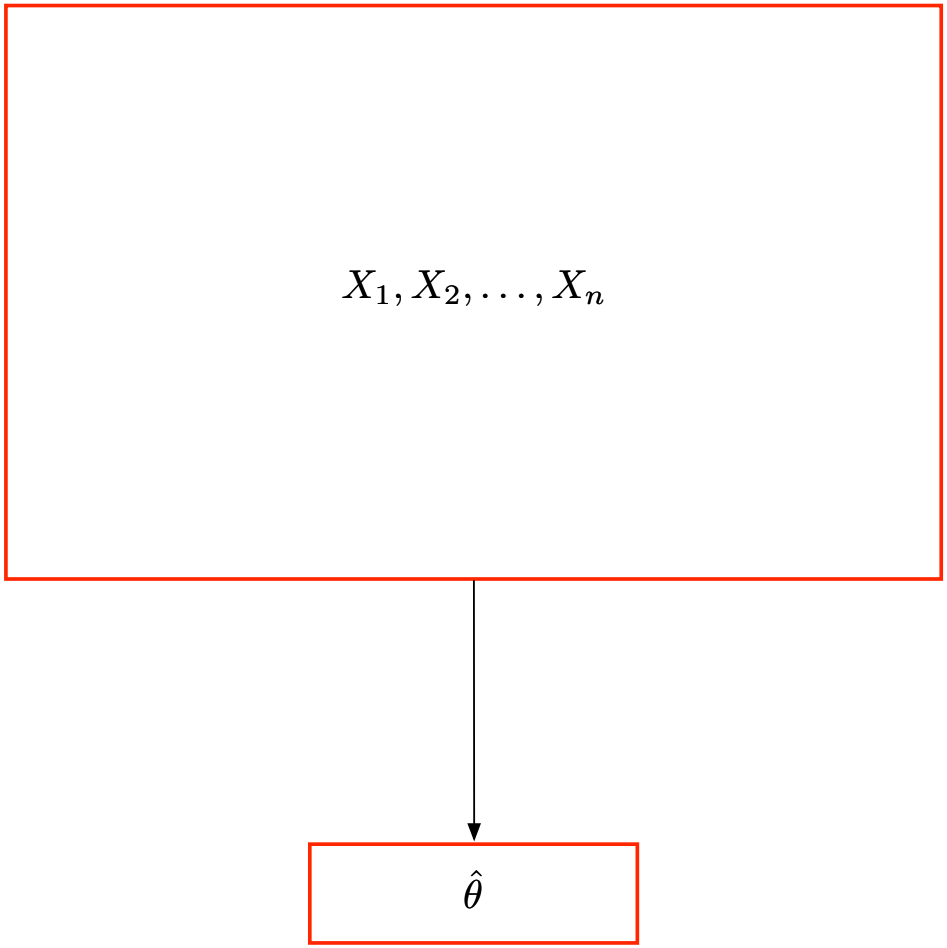
- Massive and/or sensitive datasets.
  - Too big to store, move around, do computations on.
  - Societal issues: Data cannot be shared freely.

Distributed estimation

- Massive and/or sensitive datasets.
  - Too big to store, move around, do computations on.
  - Societal issues: Data cannot be shared freely.

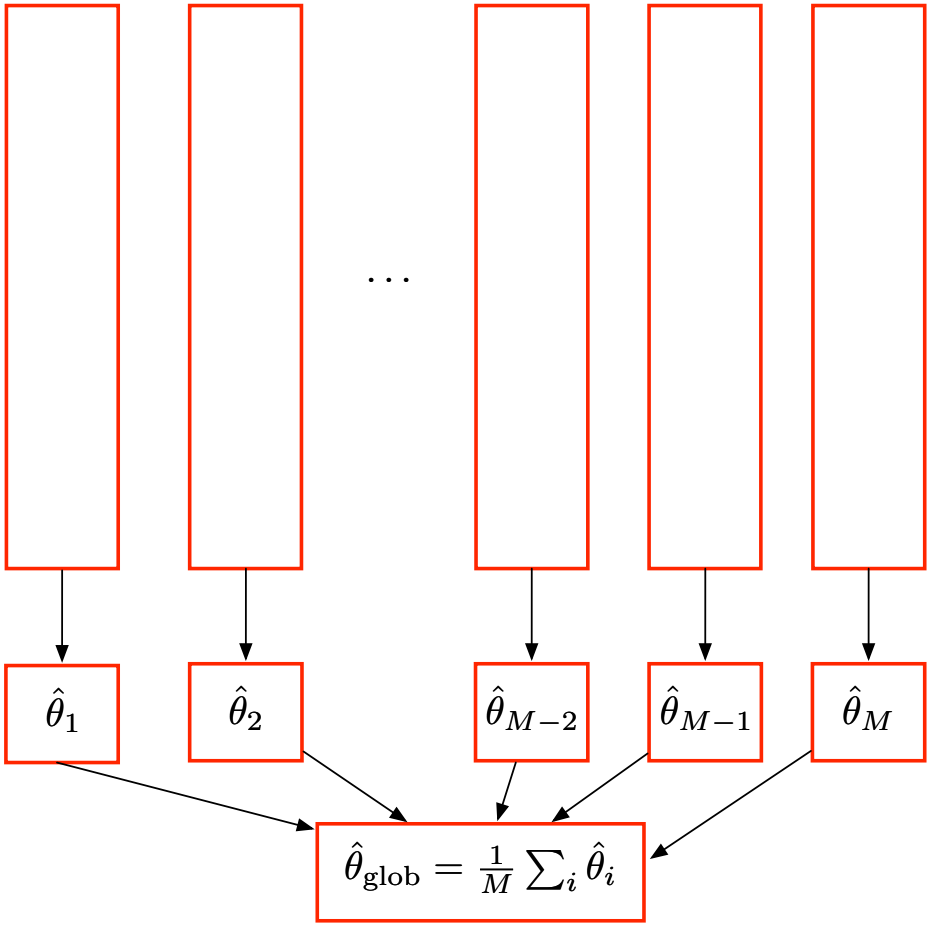
Parameter estimation

- Dataset $(X_1, \ldots, X_n) \sim \mathbf{P}_\theta$, estimator $\hat{\theta}$ function of the whole dataset
- Good properties for $\hat{\theta}$: deviation bounds, guarantees w.h.p., etc.



Parameter estimation

- Dataset $(X_1, \ldots, X_n)$ separated into M blocks with n/M samples
- Estimators $\hat{\theta}_i$ functions of the block, $\hat{\theta}_{\text{glob}}$ average.



Distributed estimation

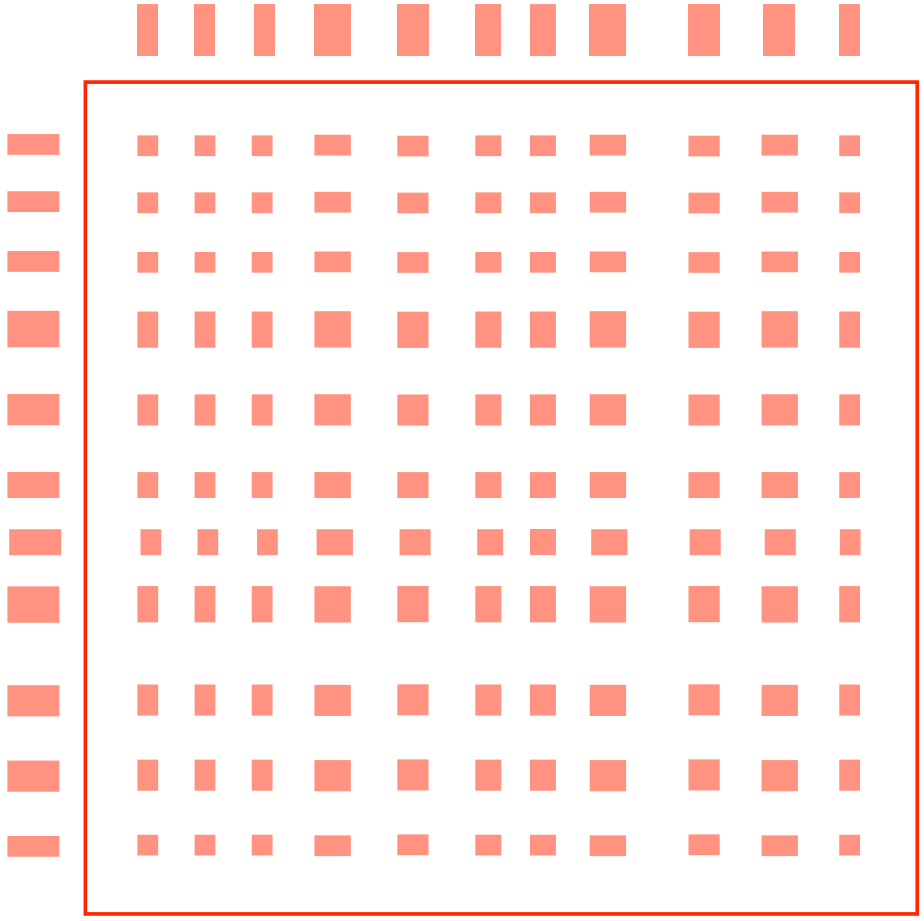
- **Practical advantages**
  - **Running time**: from $T(n)$ to $T_{\text{dist}}(n) = T(n/M)$ and $T_{\text{queue}} = MT(n/M)$

    E.g.: $M = n^{1/4}$, $T(n) = n^5$, to $T_{\text{dist}}(n) = n^{3.75}$ and $T_{\text{queue}} = n^4$
  - **Security**: Data is not communicated, just local estimator $\hat{\theta}_j$.
- **Properties**
  - **Loss of signal:** M estimators, each with less samples. How does it scale?
  - **Statistical results:** Large class of estimators "parallelize well". MLE, convex minimization, etc.
  - Can this be extended to any estimator for θ?

Simple problem

$$Y = A + Z$$

$A = aa^\top$ for sparse a, noise Z with i.i.d. coefficients, matrix of size n.



Simple problem

Noisy observation of interaction/comparison between n profiles

$$Y = \underbrace{A}_{\text{signal}} + \underbrace{Z}_{\text{ind. noise}}$$

- **Signal**: Symmetric low rank matrix $A = aa^\top$, with

$$a_i = \begin{cases} 0 & \text{w.p. } 1 - \frac{k}{n}, \\ \alpha_i \in [1/2, 1] & \text{w.p. } \frac{k}{n}. \end{cases}$$

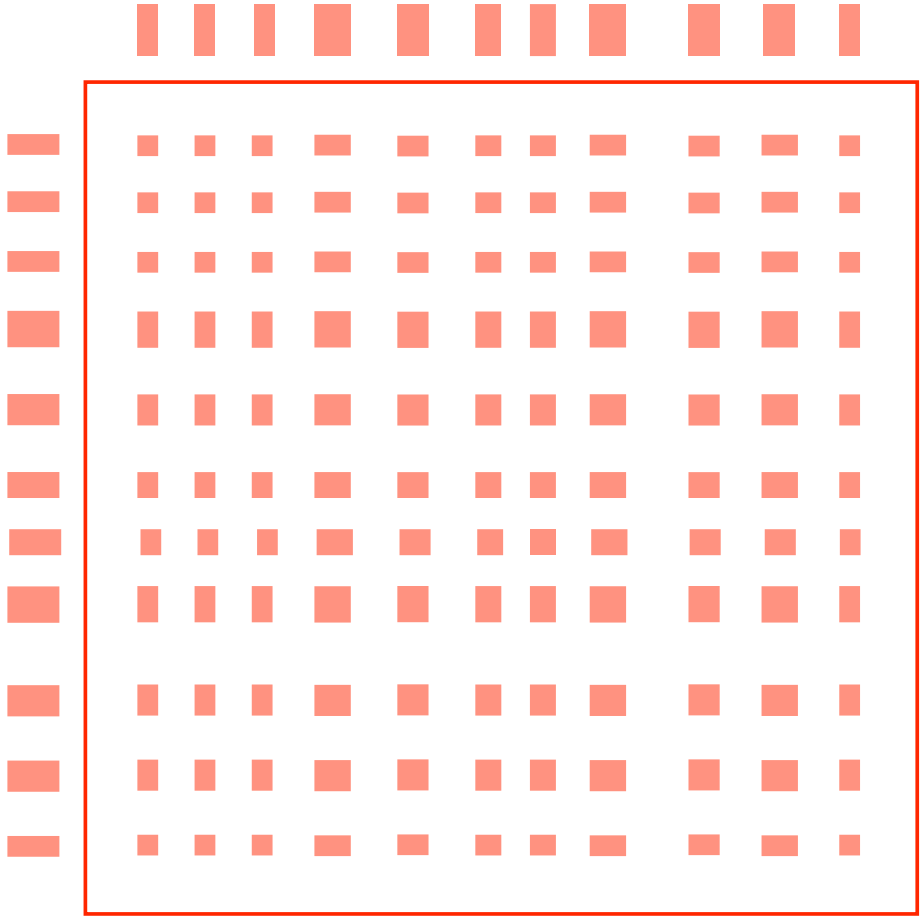
- **Noise**: Z indep. coeff., centred with variance 1 and sub-Gaussian tails.

In this problem a is sparse, with $k = C\sqrt{n}$.

Simple problem

$$Y = A + Z$$

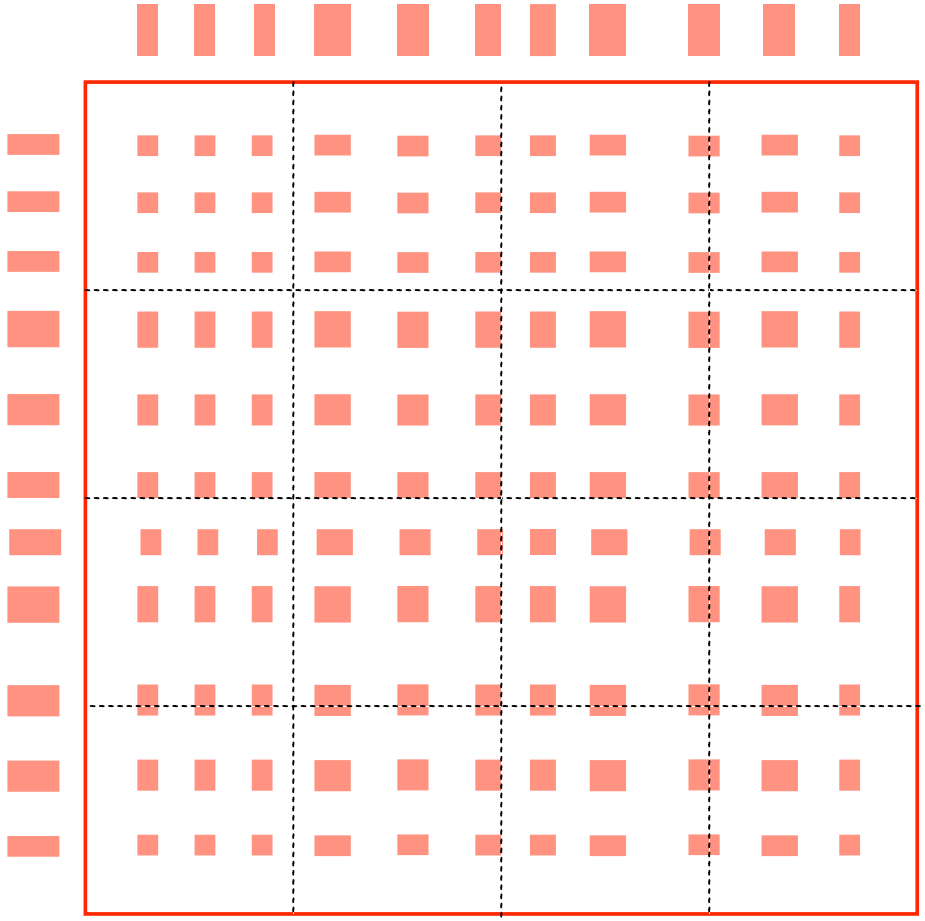
Estimate A or mean of the α_i: intensity of the similarities/interactions



Simple problem: distributed version

$$Y_S = A_S + Z_S$$

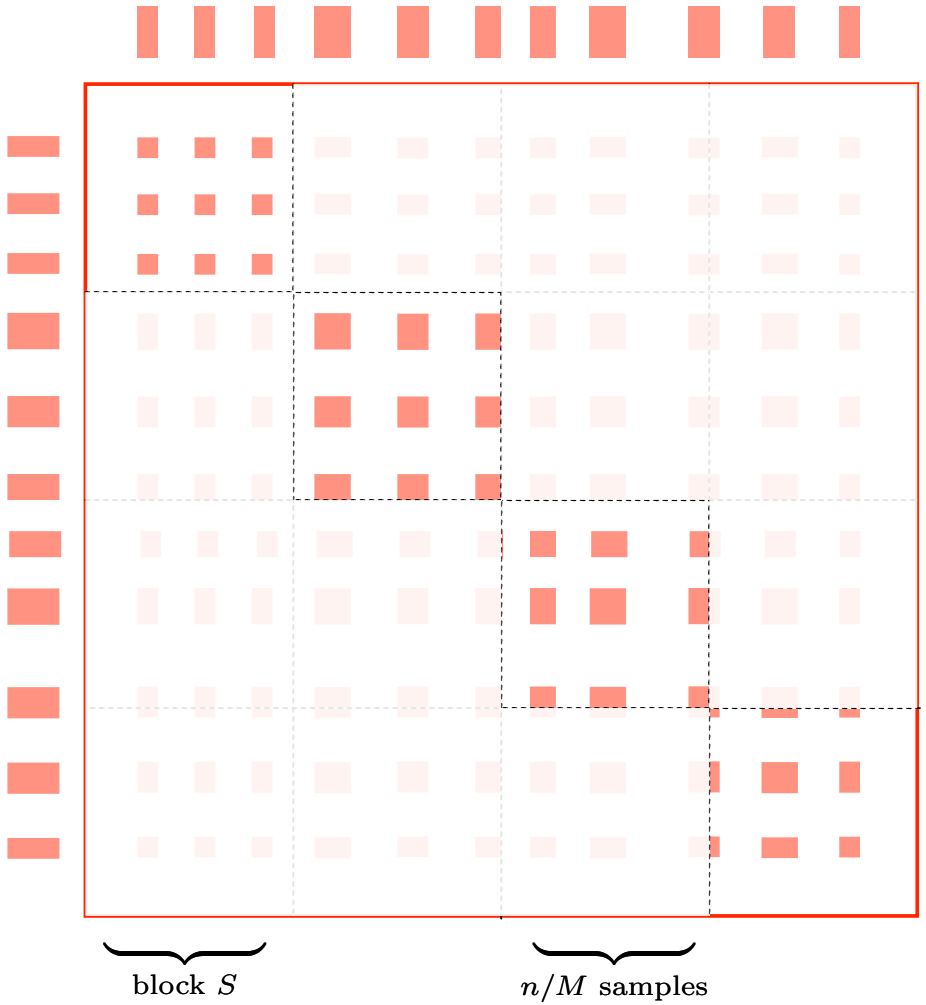
Estimate A or mean of the α_i: intensity of the simil./interactions, $M = n^{\varepsilon}$ **blocks**



Simple problem: distributed version

$$Y_S = A_S + Z_S\,,$$

Estimate A or mean of the α_i: intensity of the simil./interactions, $M = n^{\varepsilon}$ **blocks**

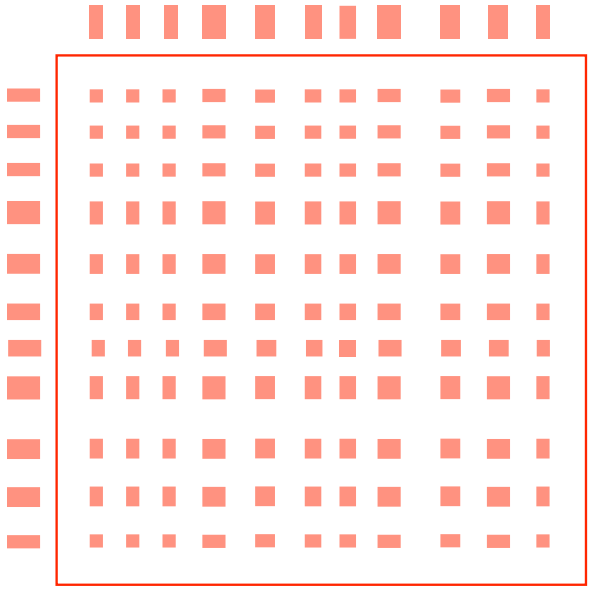


Issues in perspective

Heuristic for α, mean of the α_i: sum of the coefficients

$$\begin{aligned} \mathbf{1}^\top Y \mathbf{1} &= (\mathbf{1}^\top a)^2 + \mathbf{1}^\top Z \mathbf{1} \\ &\approx k^2\alpha^2 \pm \sqrt{n^2} \\ &\approx \alpha^2 C^2 n \pm n \end{aligned}$$

With $\hat{\alpha}^2 = \frac{1}{C^2 n}\mathbf{1}^\top Y \mathbf{1}$, estimate of order $\alpha \pm 1/C^2$: detect the presence of signal

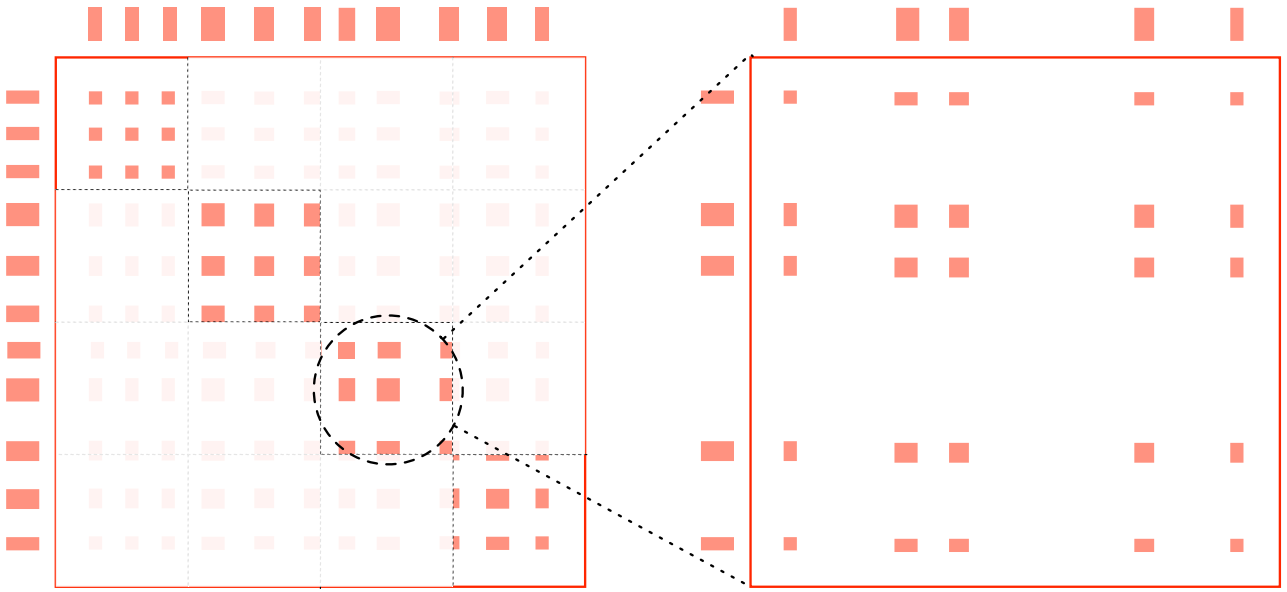


Issues in perspective

Sum of coefficients in each block, with $k_\varepsilon \approx n^{1-\varepsilon}\frac{k}{n} \approx n^{1/2-\varepsilon} \approx n_\varepsilon^{c_\varepsilon}$

$$\begin{aligned}\sum \mathbf{1}^\top Y_S \mathbf{1} &= \sum (\mathbf{1}^\top a_S)^2 + \mathbf{1}^\top Z_S \mathbf{1} \\ &\approx M k_\varepsilon^2 \alpha^2 \pm \sqrt{M n_\varepsilon^2} \\ &\approx \alpha^2 n^{1-\varepsilon} \pm n^{1-\varepsilon/2}\end{aligned}$$

The deviations are too large to detect the presence of signal and to estimate α.



Estimation methods: signal A

Benchmark: All of the data is available

First phase: sparse spectral seed

- Signal $A = aa^\top = \mu vv^\top$, where $|v|_2 = 1$ and $\mu = |a|_2^2 \approx k \approx M\sqrt{n}$
- Estimator $\hat{v}_k = \underset{\substack{|x|_2=1 \\ k \text{ sparse}}}{\operatorname{argmax}} x^\top Y x$, deviation bounds with high probability

$$\|\hat{v}_k \hat{v}_k^\top - vv^\top\| \leq \frac{c_1}{\mu}\|Y - A\|_{\text{op},k} \leq \frac{c_1}{k}\sqrt{k\log(n)} \leq \frac{c_1}{\sqrt{C}}\sqrt{\frac{\log(n)}{n^{1/2}}}.$$

Second phase: refinement

- Candidate set $\hat{W} = \{i : |\hat{v}_{k,i}| > \frac{1}{\sqrt{2k}}\}$, good intersection with the support of a.
- Recovery of the whole support by indexes with highest coefficients in $\hat{W}$.

Estimation methods: signal A

Distributed version: block estimators, and aggregation

First phase: sparse spectral seed

- Signal $A_S = a_S a_S^\top = \mu v_S v_S^\top$, where $|v_S|_2 = 1$ and $\mu = |a_S|_2^2 \approx k_\varepsilon \approx n_\varepsilon^{c_\varepsilon}$
- Estimator $\hat{v}_k = \underset{\substack{|x|_2=1 \\ k_\varepsilon \text{ - sparse}}}{\operatorname{argmax}} x^\top Y_S x$, deviation bounds with high probability

$$\|\hat{v}_{S,k}\hat{v}_{S,k}^\top - v_S v_S^\top\| \leq \frac{c_1}{\mu}\|Y_S - A_S\|_{\text{op},k_\varepsilon} \leq \frac{c_1}{k_\varepsilon}\sqrt{k_\varepsilon \log(n_\varepsilon)} \leq \frac{c_1}{\sqrt{C}}\sqrt{\frac{\log(n_\varepsilon)}{n_\varepsilon^{c_\varepsilon}}}.$$

Second phase: refinement

- Candidate set $\hat{W} = \{i : |\hat{v}_{k,i}| > \frac{1}{\sqrt{2k_\varepsilon}}\}$, good intersection with support of a_S.
- Recovery of the whole support by indexes with highest coefficients in $\hat{W}$.

Estimation methods: signal A

Algorithmic aspect: Spectral approach

Benchmark: All of the data is available

First phase: spectral seed

- Signal $A = aa^\top = \mu vv^\top$, where $|v|_2 = 1$ and $\mu = |a|_2^2 \approx k \approx M\sqrt{n}$
- Estimator $\hat{v}_k = \underset{|x|_2=1}{\operatorname{argmax}}\, x^\top Y x$, deviation bounds with high probability

$$\|\hat{v}_k \hat{v}_k^\top - vv^\top\| \leq \frac{c_1}{\mu}\|Y - A\|_{\text{op}} \leq \frac{c_1}{k}\sqrt{n} \leq \frac{c_1}{\sqrt{C}}\,.$$

Second phase: refinement

- Candidate set $\hat{W} = \{i : |\hat{v}_{k,i}| > \frac{1}{\sqrt{2k}}\}$, good intersection with the support of a.
- Recovery of the whole support by indexes with highest coefficients in $\hat{W}$.

Estimation methods: signal A

Algorithmic aspect: Spectral approach

Distributed version: block estimators, and aggregation

First phase: spectral seed

- Signal $A_S = a_S a_S^\top = \mu v_S v_S^\top$, where $|v_S|_2 = 1$ and $\mu = |a_S|_2^2 \approx k_\varepsilon \approx n_\varepsilon^{c_\varepsilon}$
- Estimator $\hat{v}_k = \underset{|x|_2=1}{\operatorname{argmax}} x^\top Y_S x$, deviation bounds with high probability

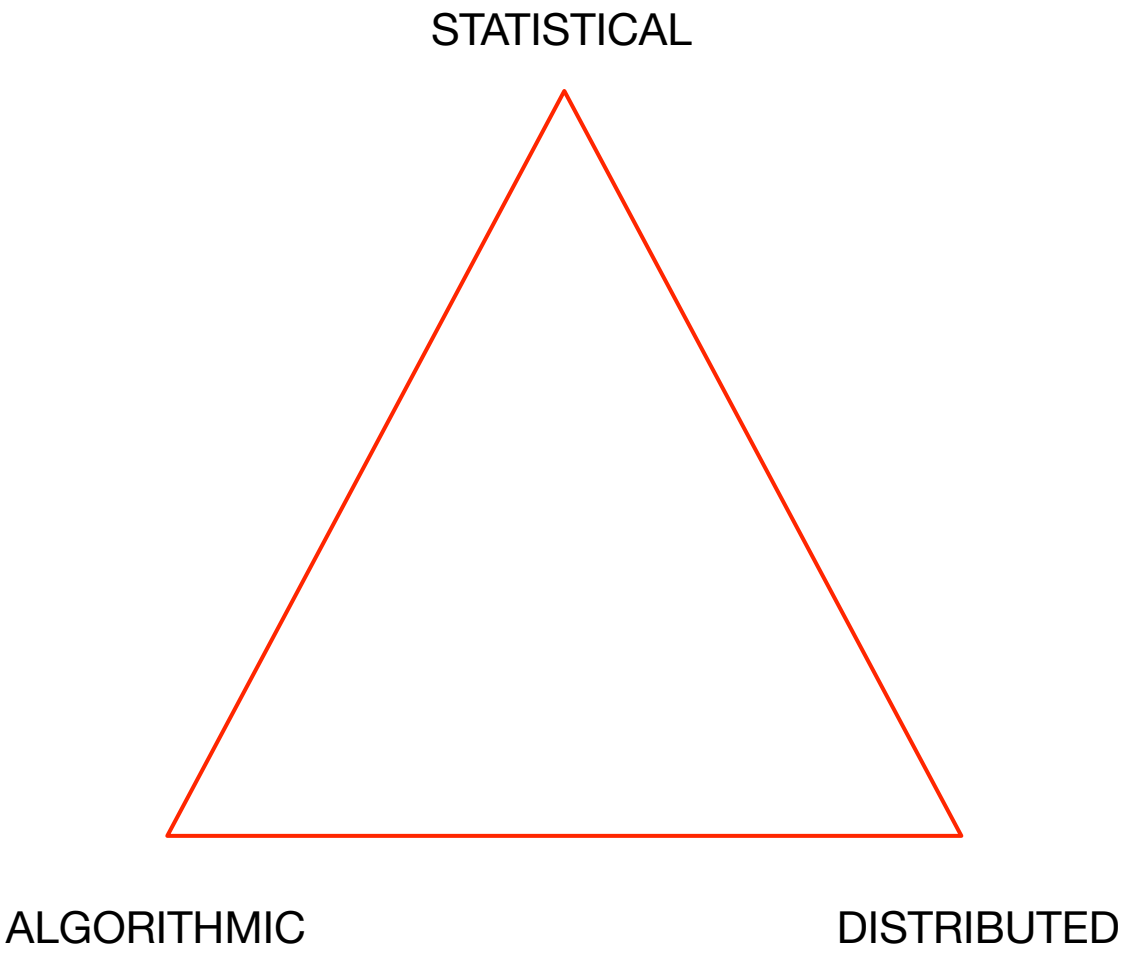
$$\|\hat{v}_{S,k}\hat{v}_{S,k}^\top - v_S v_S^\top\| \leq \frac{c_1}{\mu}\|Y_S - A_S\|_{\mathsf{op}} \leq \frac{c_1}{k_\varepsilon}\sqrt{n_\varepsilon} \approx n_\varepsilon^{1/2-c_\varepsilon} \to +\infty$$

Second phase: refinement

- ...

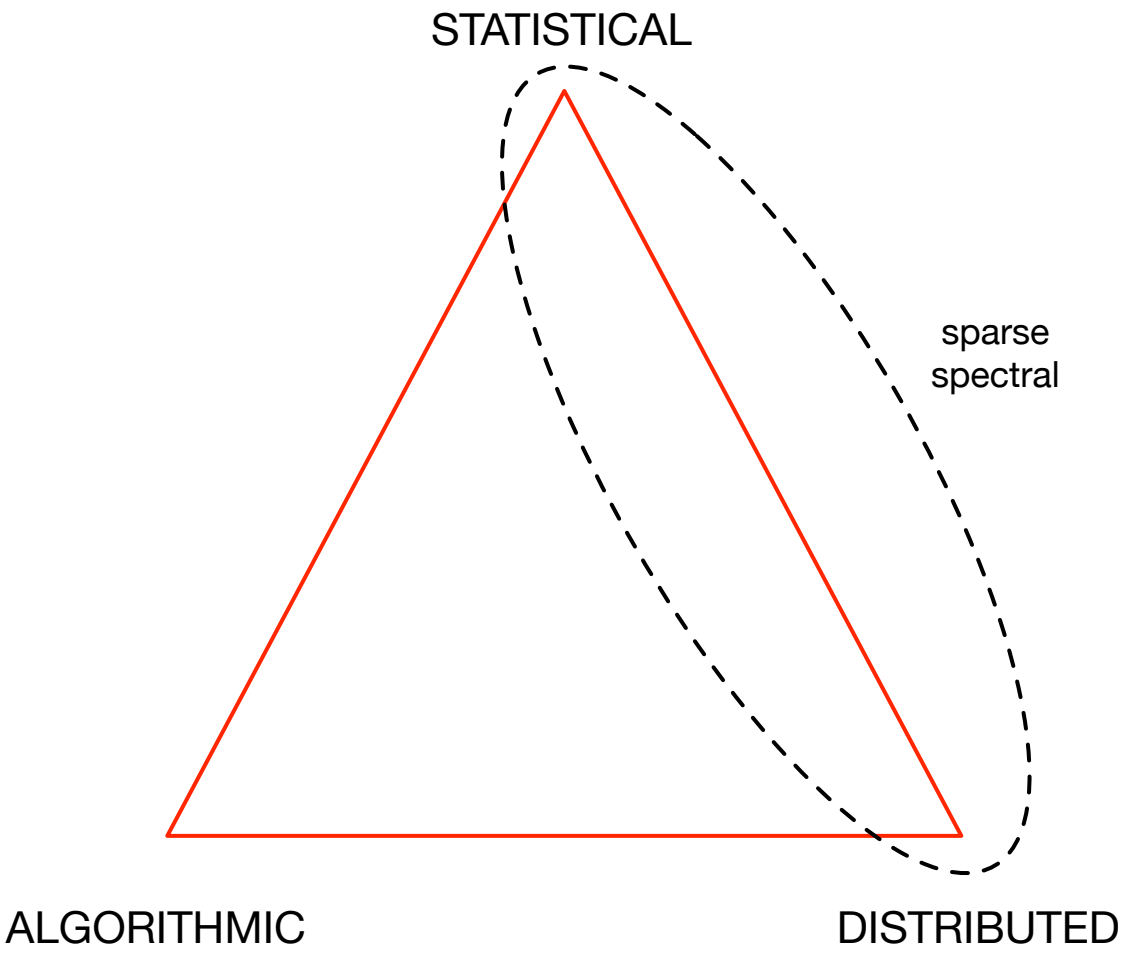
Overall picture

Objective: statistical procedure for A satisfying various properties.



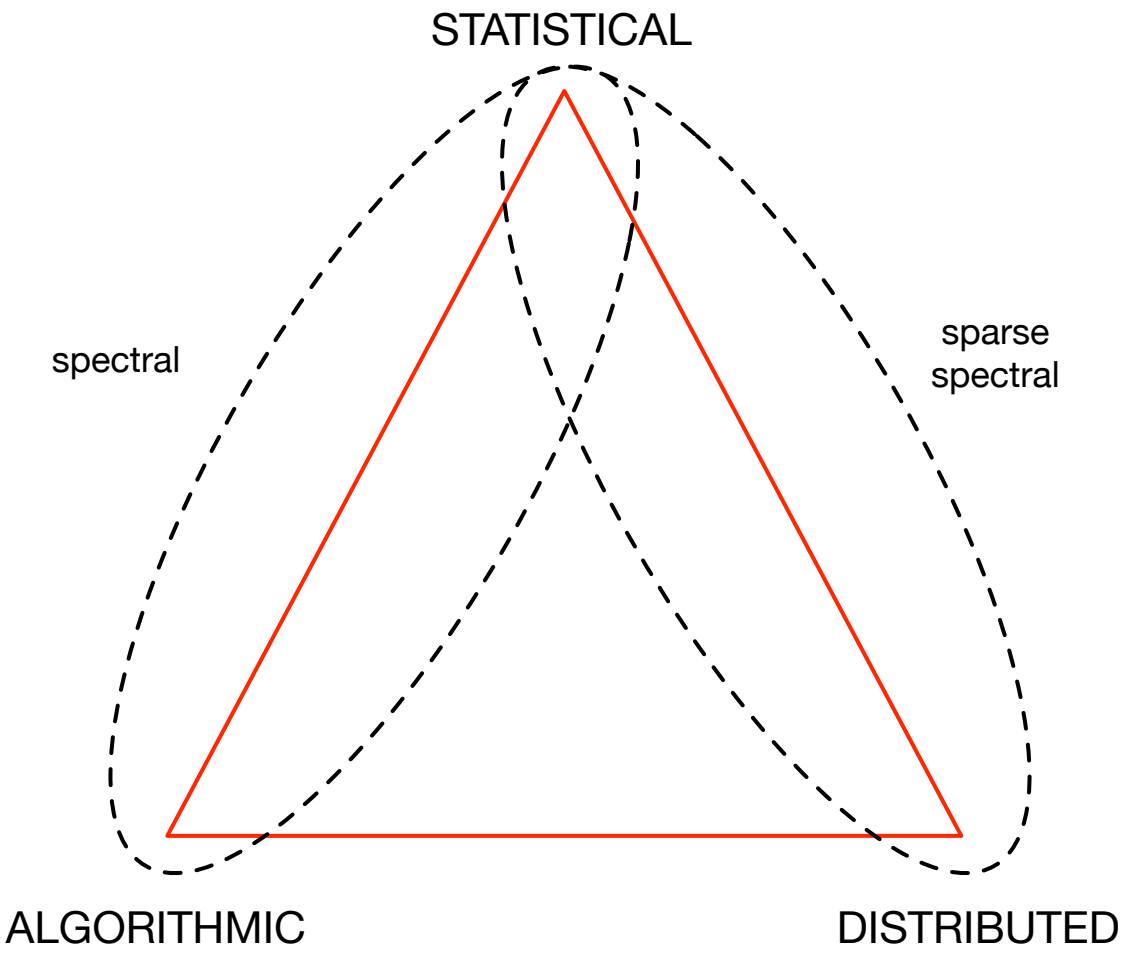
Overall picture

Objective: statistical procedure for A satisfying various properties.



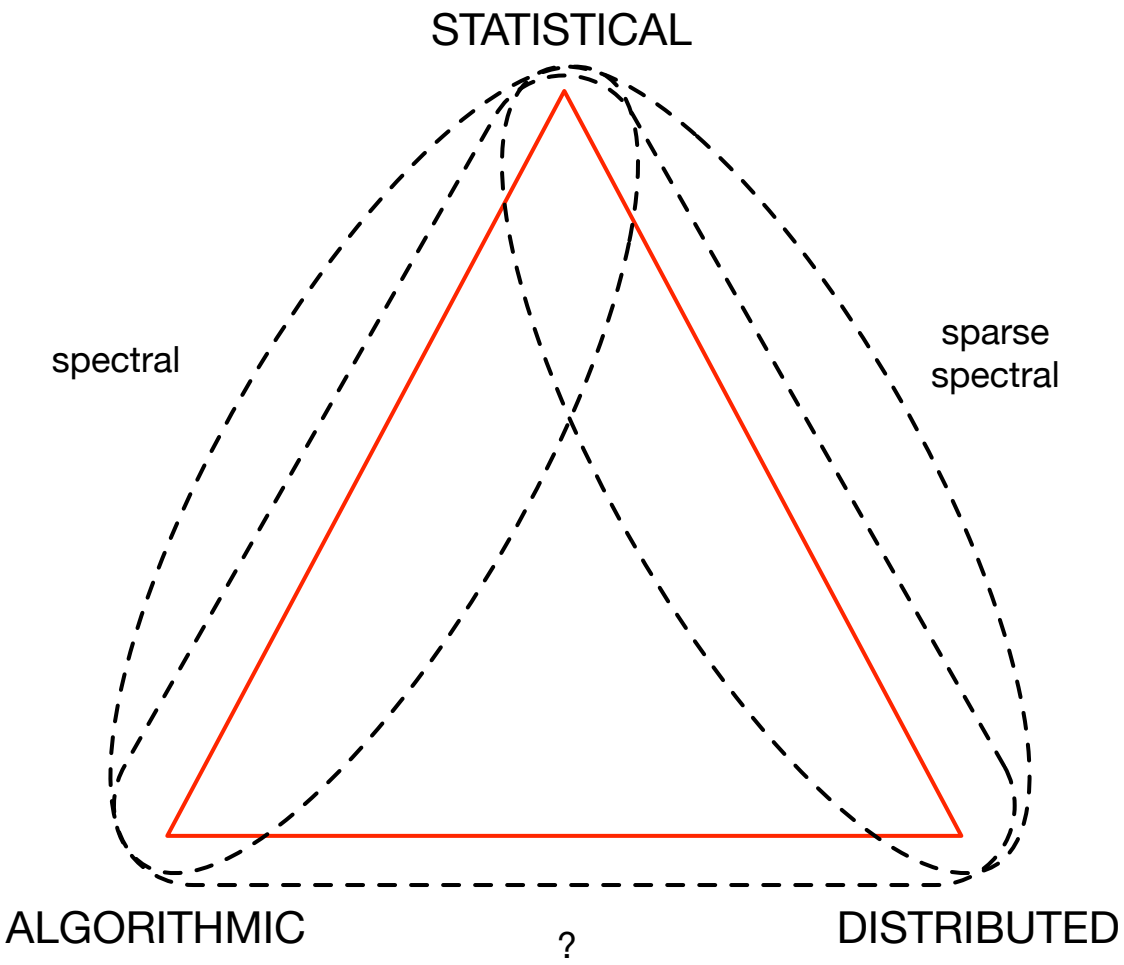
Overall picture

Objective: statistical procedure for A satisfying various properties.



Overall picture

Objective: statistical procedure for A satisfying various properties.

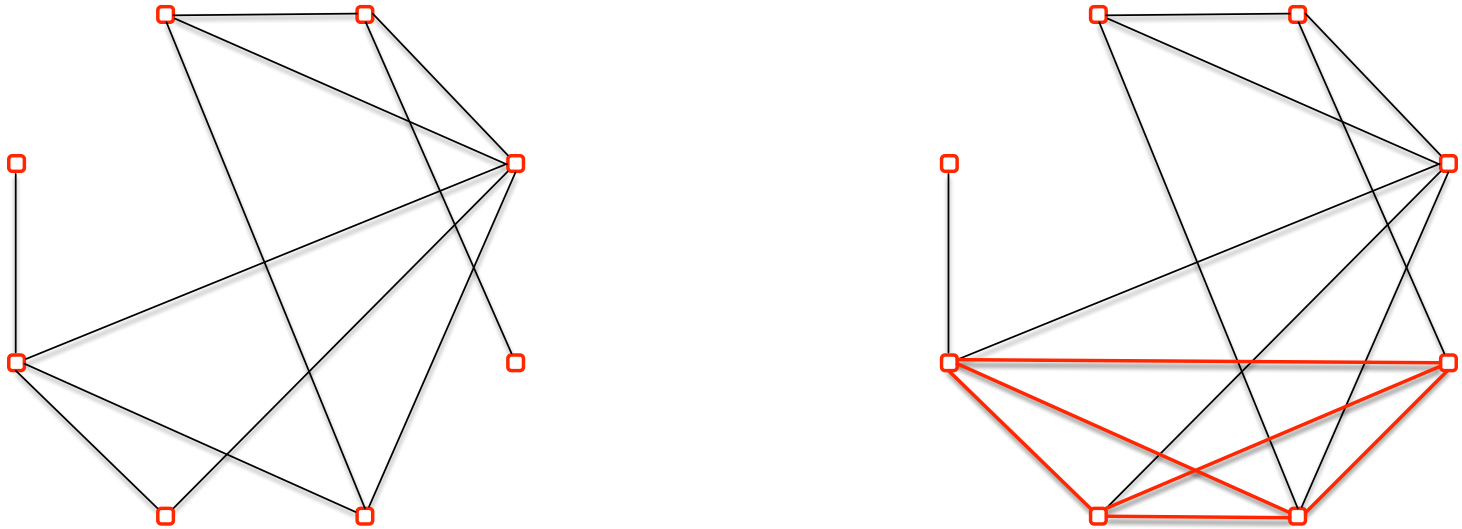


Computational limits

Impossibility result for three objectives simultaneously:

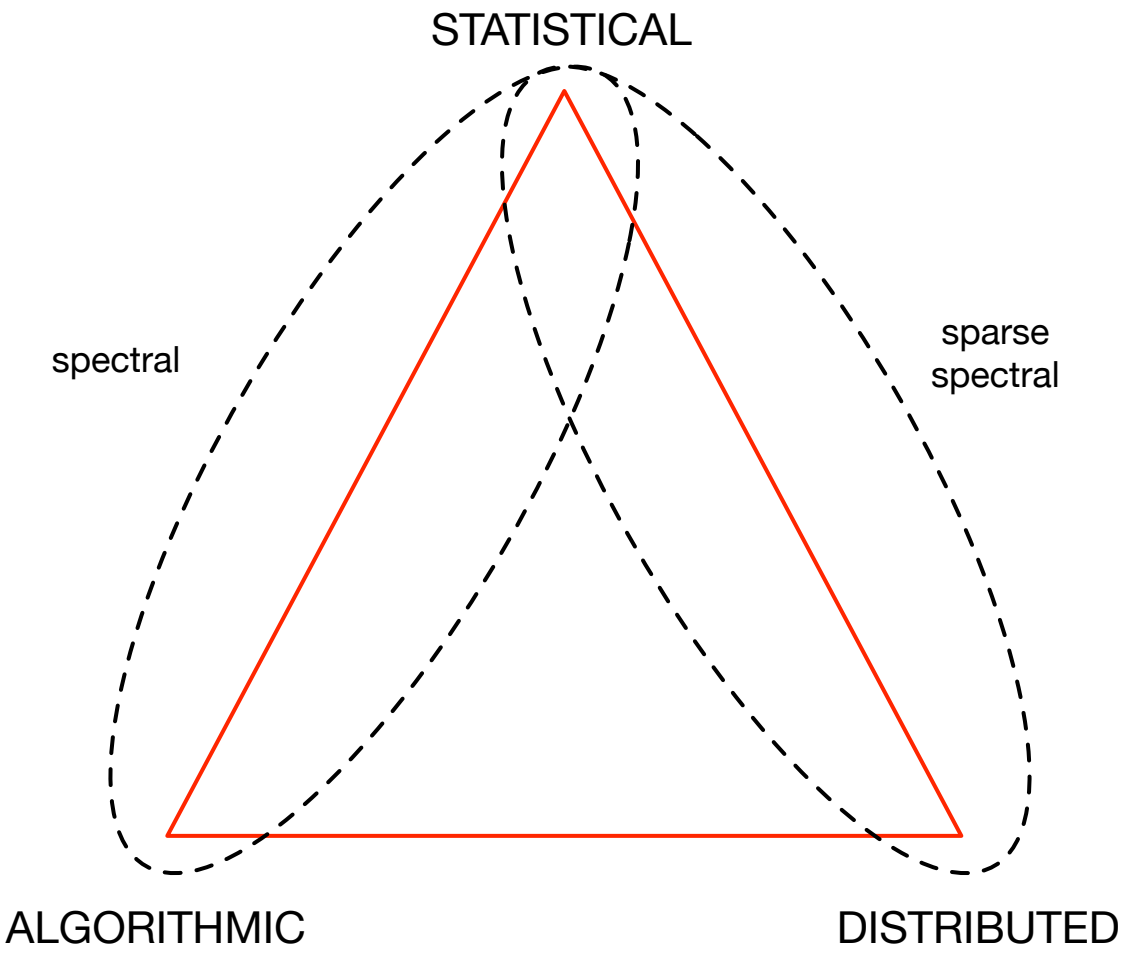
- Distributed estimation of A implies estimation of A_S
- Signal strength in each block too low for computationally efficient procedures:

Reduction from planted clique problem



Multiple trade-offs

Trade-offs between different objectives for a statistical procedure.



Multiple tradeoffs - other types

Robustness to errors

- Detection problem in $\{0,1\}^n$ related to satisfiability **B., Ellenberg (15)**
- Optimal statistical performance can be reached by NP-hard method.
- Two improvements can be made, with small statistical loss:
  - Using a computationally efficient testing method.
  - Allowing for a constant portion of errors.
- Two improvements cannot be made simultaneously:

  Reduction from Learning Parity with Noise.

Conclusion

- **Results**
  - Computational limits to doing estimation in a distributed manner.
  - **Redundancy**: Avoids some issues, for detection/intensity estimation

- **Questions**
  - How general is this situation? Systematically make datasets secure?
  - Finer analysis of these trade-offs? Complicated for lower bounds.

THANK YOU

Work supported by the Isaac Newton Trust.