Approximate Inference by Semidefinite Relaxations

Andrea Montanari
[with Adel Javanmard, Federico Ricci-Tersenghi, Subhabrata Sen]

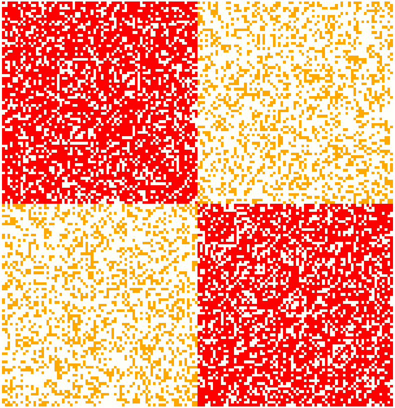
Stanford University

December 11, 2015

What is this talk about?

SDP for Matrix/Graph estimation

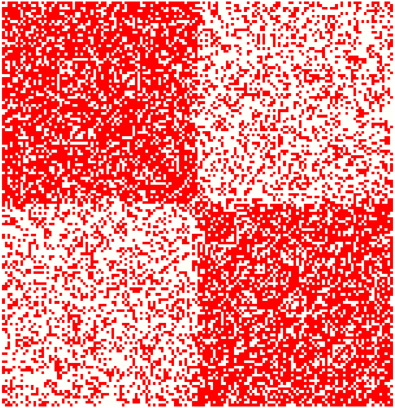
The hidden partition model



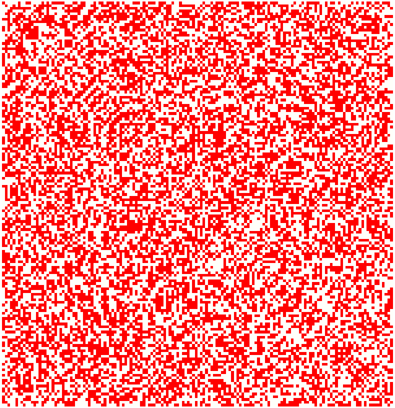
Vertices V, $|V| = n$, $V = V_+ \cup V_-$, $|V_+| = |V_-| = n/2$

$$\mathbb{P}\{(i,j) \in E\} = \begin{cases} p & \text{if } \{i,j\} \subseteq V_+ \text{ or } \{i,j\} \subseteq V_-, \\ q < p & \text{otherwise.} \end{cases}$$

Of course entries are not colored. . .



... and rows/columns are not ordered



Problem: Detect/estimate the partition

What is this talk about?

SDP for Matrix/Graph estimation

Exact phase transition(?)

Outline

Background

Statistical estimation

$$x_{0,i} = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-, \end{cases}$$

$$\mathbb{P}\{(i,j) \in E\} = \begin{cases} p & \text{if } x_{0,i} = x_{0,j}, \\ q < p & \text{otherwise.} \end{cases}$$

Estimator $\widehat{\mathbf{x}} \in \{+1,-1\}^n$

$$\text{Overlap}_n(\widehat{\mathbf{x}}) = \frac{1}{n}\mathbb{E}\big\{\big|\langle \widehat{\mathbf{x}}(G), \boldsymbol{x_0}\rangle\big|\big\}\,.$$

Statistical estimation ($p = a/n,\ q = b/n$)

$$x_{0,i} = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-, \end{cases}$$

$$\mathbb{P}\{(i,j) \in E\} = \begin{cases} a/n & \text{if } x_{0,i} = x_{0,j}, \\ b/n & \text{otherwise.} \end{cases}$$

Estimator $\widehat{\mathbf{x}} \in \{+1, -1\}^n$

$$\text{Overlap}_n(\widehat{\mathbf{x}}) = \frac{1}{n}\mathbb{E}\{\big|\langle \widehat{\mathbf{x}}(G), \boldsymbol{x_0}\rangle\big|\}\,.$$

Information theory threshold

Theorem (Mossel, Neeman, Sly, 2012)

There is an estimator that achieves $\liminf_{n\to\infty} \mathrm{Overlap}_n(\widehat{\mathrm{x}}) \geq \varepsilon > 0$ *if and only if* $a + b > 2$ *and*

$$\frac{a-b}{\sqrt{2(a+b)}} > 1\,.$$

[Proves conjecture by Decelle, Krzakala, Moore, Zdeborova, 2011]

Information theory threshold

Theorem (Mossel, Neeman, Sly, 2012)

There is an estimator that achieves $\liminf_{n\to\infty} \text{Overlap}_n(\hat{\mathbf{x}}) \geq \varepsilon > 0$ *if and only if* $a + b > 2$ *and*

$$\frac{a-b}{\sqrt{2(a+b)}} > 1\,.$$

[Proves conjecture by Decelle, Krzakala, Moore, Zdeborova, 2011]

Computational threshold

- Dyer, Frieze 1989 $p = na > q = nb$ fixed.
- Condon, Karp 2001 $a - b \gg n^{1/2}$
- McSherry 2001 $a - b \gg \sqrt{b \log n}$
- Coja-Oghlan 2010 $a - b \gg \sqrt{b}$
- Massoulie 2013 and Mossel, Neeman, Sly, 2013

$$\frac{a-b}{\sqrt{2(a+b)}} > 1$$

Very ingenious spectral methods!

Computational threshold

- Dyer, Frieze 1989 $p = na > q = nb$ fixed.
- Condon, Karp 2001 $a - b \gg n^{1/2}$
- McSherry 2001 $a - b \gg \sqrt{b \log n}$
- Coja-Oghlan 2010 $a - b \gg \sqrt{b}$
- Massoulie 2013 and Mossel, Neeman, Sly, 2013

$$\frac{a-b}{\sqrt{2(a+b)}} > 1$$

Very ingenious spectral methods!

Computational threshold

- Dyer, Frieze 1989 $p = na > q = nb$ fixed.
- Condon, Karp 2001 $a - b \gg n^{1/2}$
- McSherry 2001 $a - b \gg \sqrt{b \log n}$
- Coja-Oghlan 2010 $a - b \gg \sqrt{b}$
- Massoulie 2013 and Mossel, Neeman, Sly, 2013

$$\frac{a - b}{\sqrt{2(a + b)}} > 1$$

Very ingenious spectral methods!

What if I am not ingenious?

Maximum Likelihood

Posterior probability

Candidate partiton $\boldsymbol{\sigma} \in \{+1, -1\}^n$

$$\mathbb{P}(\boldsymbol{x_0} = \boldsymbol{\sigma} | G) \approx \frac{1}{Z(G)} \prod_{(i,j)\in E} \{ a\, \mathbb{I}(\sigma_i = \sigma_j) + b\, \mathbb{I}(\sigma_i \neq \sigma_j) \} \, \mathbb{I}\Big(\sum_{i=1}^{n} \sigma_i = 0 \Big)$$

Pairwise binary graphical model

Adjacency matrix

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

$$\boldsymbol{A} = (A_{ij})_{1 \le i,j \le n}$$

Maximum likelihood

$$\sigma_i = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-. \end{cases}$$

$$\begin{aligned} \text{maximize} \quad & \sum_{i,j=1}^{n} A_{ij}\, \sigma_i \sigma_j \,, \\ \text{subject to} \quad & \sum_{i=1}^{n} \sigma_i = 0 \,, \\ & \sigma_i \in \{+1, -1\} \,. \end{aligned}$$

Maximum likelihood

$$\sigma_i = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-. \end{cases}$$

$$\begin{aligned} \text{maximize} \quad & \sum_{i,j=1}^{n} A_{ij}\, \sigma_i \sigma_j\,, \\ \text{subject to} \quad & \sum_{i=1}^{n} \sigma_i = 0\,, \\ & \sigma_i \in \{+1,-1\}\,. \end{aligned}$$

Maximum likelihood

$$\sigma_i = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-. \end{cases}$$

$$\begin{aligned} \text{maximize} \quad & \sum_{i,j=1}^{n} A_{ij}\, \sigma_i \sigma_j \,, \\ \text{subject to} \quad & \sum_{i=1}^{n} \sigma_i = 0\,, \\ & \sigma_i \in \{+1, -1\}\,. \end{aligned}$$

Lagrangian

$$\begin{aligned} \text{maximize} \quad & \sum_{i,j=1}^{n} A_{ij}\,\sigma_i\sigma_j - \gamma\Big(\sum_{i=1}^{n}\sigma_i\Big)^2 . \\ \text{subject to} \quad & \sigma_i \in \{+1,-1\}\,. \end{aligned}$$

A good choice:

$$\gamma = \frac{a+b}{2n} \equiv \frac{d}{n}$$

Lagrangian

$$\begin{aligned} \text{maximize} \quad & \sum_{i,j=1}^{n} A_{ij}\,\sigma_i\sigma_j - \gamma\Big(\sum_{i=1}^{n}\sigma_i\Big)^2 . \\ \text{subject to} \quad & \sigma_i \in \{+1,-1\}\,. \end{aligned}$$

A good choice:

$$\gamma = \frac{a+b}{2n} \equiv \frac{d}{n}$$

Centered adjacency matrix

$$A^{\text{cen}}_{ij} = \begin{cases} 1 - (d/n) & \text{if } (i,j) \in E, \\ -(d/n) & \text{otherwise.} \end{cases}$$

$$\boldsymbol{A}^{\text{cen}} = \boldsymbol{A} - \frac{d}{n} \mathbf{1}\,\mathbf{1}^{\mathsf{T}}$$

Lagrangian

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{A}^{\text{cen}}, \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n \,. \end{aligned}$$

- NP-hard
- SDP($\boldsymbol{A}^{\text{cen}}$) is a very natural convex relaxation

Lagrangian

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{A}^{\text{cen}}, \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n \,. \end{aligned}$$

- NP-hard
- SDP($\boldsymbol{A}^{\text{cen}}$) is a very natural convex relaxation

Relaxation

$$
\begin{aligned}
\text{maximize} \quad & \langle \boldsymbol{A}^{\text{cen}}, \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} \rangle\,, \\
\text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n\,.
\end{aligned}
$$

SDP($\boldsymbol{A}^{\text{cen}}$):

$$
\begin{aligned}
\text{maximize} \quad & \langle \boldsymbol{A}^{\text{cen}}, \boldsymbol{X} \rangle\,, \\
\text{subject to} \quad & \boldsymbol{X} \in \mathbb{R}^{n \times n},\ \boldsymbol{X} \succeq 0\,, \\
& X_{ii} = 1\,.
\end{aligned}
$$

Estimator

- Compute principal eigenvector $v_1(\boldsymbol{X})$
- Threshold it $\hat{x}^{\rm SDP}(G) = \text{sign}(v_1(\boldsymbol{X}))$
- Randomized variation for proofs

This is really off-the-shelf

How well does it work?

Estimator

- Compute principal eigenvector $v_1(\boldsymbol{X})$
- Threshold it $\hat{\boldsymbol{x}}^{\rm SDP}(G) = {\rm sign}(v_1(\boldsymbol{X}))$
- Randomized variation for proofs

This is really **off-the-shelf**

How well does it work?

Near-optimality of SDP

Before we pass to SDP

- What's the problem with sparse graphs?
- What's the problem vanilla PCA?

Why PCA?

Ground truth

$$x_{0,i} = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-. \end{cases}$$

Data = RankOne + Wigner

$$\frac{1}{\sqrt{d}} A^{\text{cen}} = \frac{\lambda}{n} x_0 x_0^{\mathsf{T}} + W \,, \qquad \lambda \equiv \frac{a-b}{\sqrt{2(a+b)}}$$

$$E\{W_{ij}\} = 0 \,, \qquad \mathbb{E}\{W_{ij}^2\} \in \Big\{ \frac{a}{dn} \,, \frac{b}{dn} \Big\} \approx \frac{1}{n} \,.$$

Why PCA?

Ground truth

$$x_{0,i} = \begin{cases} +1 & \text{if } i \in V_+, \\ -1 & \text{if } i \in V_-. \end{cases}$$

Data = RankOne + Wigner

$$\frac{1}{\sqrt{d}} \boldsymbol{A}^{\text{cen}} = \frac{\lambda}{n} \boldsymbol{x_0} \boldsymbol{x_0}^{\mathsf{T}} + \boldsymbol{W}, \qquad \lambda \equiv \frac{a-b}{\sqrt{2(a+b)}}$$

$$E\{W_{ij}\} = 0, \qquad \mathbb{E}\{W_{ij}^2\} \in \left\{ \frac{a}{dn}, \frac{b}{dn} \right\} \approx \frac{1}{n}.$$

The *right* parametrization

$$d = \frac{a+b}{2}, \qquad \lambda = \frac{a-b}{\sqrt{2(a+b)}}$$

Naive PCA

$$\widehat{\mathbf{x}}^{\mathrm{PCA}}(\boldsymbol{A}^{\mathrm{cen}}) = \sqrt{n}\, v_1(\boldsymbol{A}^{\mathrm{cen}})\,.$$

Does it work?

$$\frac{1}{\sqrt{d}} \boldsymbol{A}^{\text{cen}} = \frac{\lambda}{n} \boldsymbol{x_0} \boldsymbol{x_0}^{\mathsf{T}} + \boldsymbol{W}$$

Naive idea:

$$\|\boldsymbol{W}\|_2 \leq \text{const.}, \quad \left\|\frac{\lambda}{n} \boldsymbol{x_0} \boldsymbol{x_0}^{\mathsf{T}}\right\|_2 = \lambda \quad \Rightarrow \text{Works for } \lambda = O(1)$$

Wrong!

Does it work?

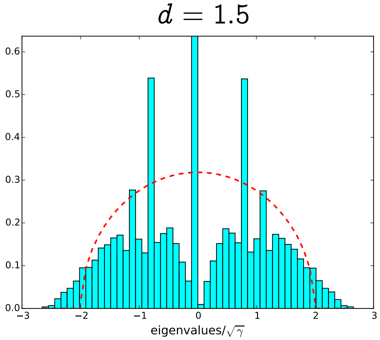
$$\frac{1}{\sqrt{d}}\boldsymbol{A}^{\text{cen}} = \frac{\lambda}{n}\boldsymbol{x_0}\boldsymbol{x_0}^{\mathsf{T}} + \boldsymbol{W}$$

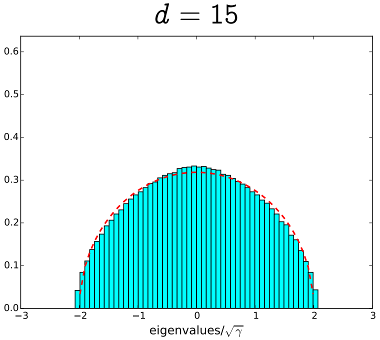
Naive idea:

$$\|\boldsymbol{W}\|_2 \leq \text{const.}, \quad \left\|\frac{\lambda}{n}\boldsymbol{x_0}\boldsymbol{x_0}^{\mathsf{T}}\right\|_2 = \lambda \quad \Rightarrow \text{Works for } \lambda = O(1)$$

Wrong!

Spectral relaxation bad in the sparse regime!



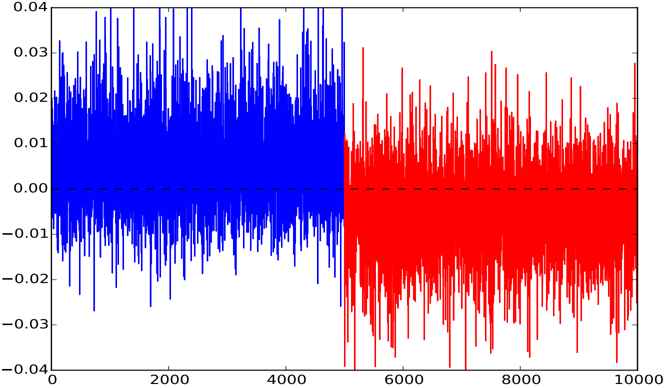


Theorem (Krivelevich, Sudakov 2003+Vu 2005)

With high probability,

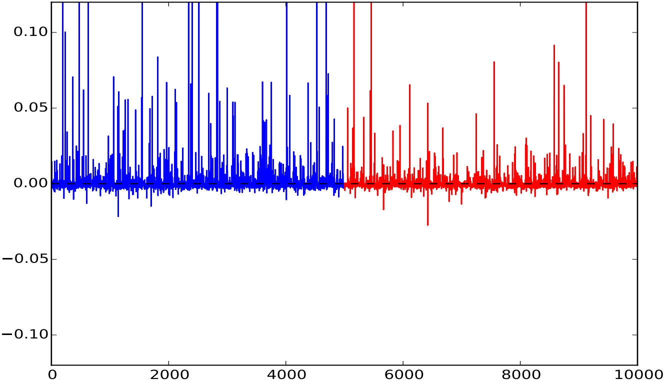
$$\lambda_{\max}(A^{cen}/\sqrt{d}) = \begin{cases} 2\,(1+o(1)) & \textit{if } d \gg (\log n)^4, \\ C\sqrt{\log n/(\log\log n)}(1+o(1)) & \textit{if } d = O(1). \end{cases}$$

Example: $d = 20$, $\lambda = 1.2$, $n = 10^4$



$v_1(A^{\text{cen}})$

Example: $d = 3$, $\lambda = 1.2$, $n = 10^4$



$v_1(A^{\rm cen})$

Why should SDP work better?

$$
\begin{aligned}
\text{maximize} \quad & \langle \boldsymbol{A}^{\text{cen}}, \boldsymbol{X} \rangle \,, \\
\text{subject to} \quad & \boldsymbol{X} \in \mathbb{R}^{n \times n},\ \boldsymbol{X} \succeq 0 \,, \\
& X_{ii} = 1 \,.
\end{aligned}
$$

Recall the ultimate limit

$G(n, d, \lambda)$ graph distribution with parameters

$$d = \frac{a+b}{2} > 1\,, \qquad \lambda = \frac{a-b}{\sqrt{2(a+b)}}$$

Theorem (Mossel, Neeman, Sly, 2012)

If $\lambda < 1$, *then*

$$\limsup_{n\to\infty} \big\|G(n, d, 0) - G(n, d, \lambda)\big\|_{\rm TV} < 1\,.$$

If $\lambda > 1$, *then*

$$\lim_{n\to\infty} \big\|G(n, d, 0) - G(n, d, \lambda)\big\|_{\rm TV} = 1\,.$$

SDP has nearly optimal threshold

Theorem (Montanari, Sen 2015)

Assume $G \sim \mathsf{G}(n, d, \lambda)$.
If $\lambda \leq 1$, *then, with high probability,*

$$\frac{1}{n\sqrt{d}}\mathsf{SDP}(\boldsymbol{A}_G^{cen}) = 2 + o_d(1)\,.$$

If $\lambda > 1$, *then there exists* $\Delta(\lambda) > 0$ *such that, with high probability,*

$$\frac{1}{n\sqrt{d}}\mathsf{SDP}(\boldsymbol{A}_G^{cen}) = 2 + \Delta(\lambda) + o_d(1)\,.$$

Consequence

Corollary (Montanari, Sen 2015)

Assume $\lambda \geq 1 + \varepsilon$*. Then there exists* $d_0(\varepsilon)$ *and* $\delta(\varepsilon) > 0$ *such that the randomized SDP-based estimator achieves, for* $d \geq d_0(\varepsilon)$*,*

$$\lim \inf_{n \to \infty} \mathsf{E}\{\text{Overlap}_n(\hat{x}^{SDP})\} \geq \delta(\varepsilon)\,.$$

Earlier/related work

Optimal spectral tests

- Massoulie 2013
- Mossel, Neeman, Sly, 2013
- Bordenave, Lelarge, Massoulie, 2015

SDP, $d = \Theta(\log n)$

- Abbe, Bandeira, Hall 2014
- Hajek, Wu, Xu 2015

SDP, detection

- Guédon, Vershynin, 2015 (requires $\lambda \geq 10^4$, very different proof)

How does SDP work 'in practice'?

Thresholds

- $\lambda_c^{\mathrm{opt}}(d) \equiv$ Threshold for optimal test

- $\lambda_c^{\mathrm{SDP}}(d) \equiv$ Threshold for SDP-based test

What we know

- $\lambda_c^{\rm opt}(d) = 1$ [Mossel, Neeman, Sly, 2013]

- $\lambda_c^{\rm SDP}(d) = 1 + o_d(1)$ [Montanari, Sen, 2015]

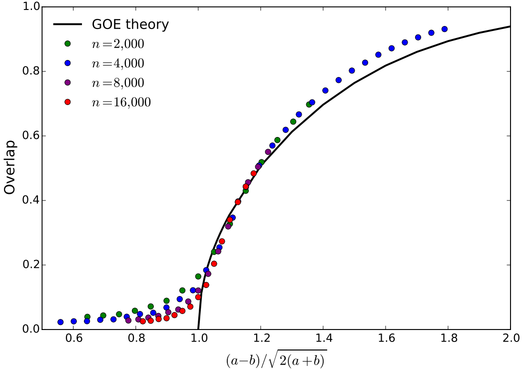
How big is the $o_d(1)$ gap?

What we know

- $\lambda_c^{\rm opt}(d) = 1$ [Mossel, Neeman, Sly, 2013]
- $\lambda_c^{\rm SDP}(d) = 1 + o_d(1)$ [Montanari, Sen, 2015]

How big is the $o_d(1)$ gap?

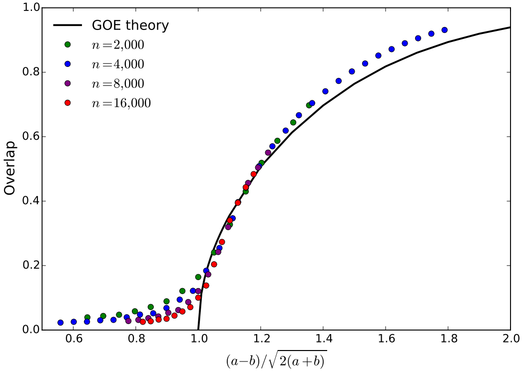
Simulations: $d = 5$, $N_{\rm sample} = 500$ (with Javanmard and Ricci)



SDP estimator $\hat{\boldsymbol{x}}^{\rm SDP} \in \{+1,-1\}^n$

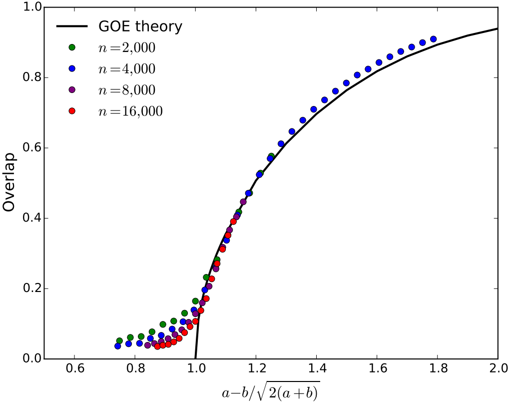
$$\text{Overlap}_n(\widehat{\mathrm{x}}) = \frac{1}{n}\mathbb{E}\big\{\big|\langle \hat{\boldsymbol{x}}^{\rm SDP}(G), \boldsymbol{x_0}\rangle\big|\big\}\,.$$

Simulations: $d = 5$, $N_{\text{sample}} = 500$



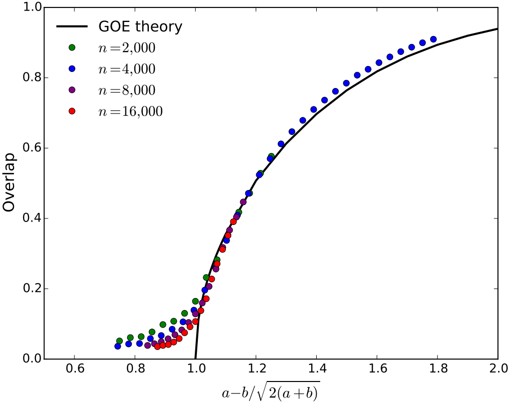
$$\lambda_c^{\text{SDP}}(d = 5) \approx 1\,.$$

Simulations: $d = 10$, $N_{\text{sample}} = 500$



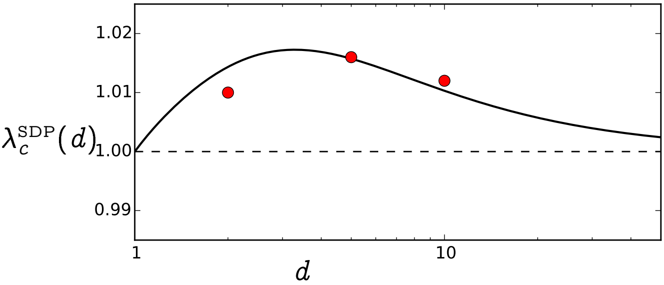
$$\lambda_c^{\text{SDP}}(d = 10) \approx 1\,.$$

Simulations: $d = 10$, $N_{\text{sample}} = 500$



Can we estimate $\lambda_c^{SDP}(d)$ from data?

$\lambda_c^{\text{SDP}}(d)$, $N_{\text{sample}} \geq 10^5$ (10 years CPU time)



- Dots: Numerical estimates
- Line: Non-rigorous analytical approximation (using statistical physics)
- **At most 2% sub-optimal!**

One last question

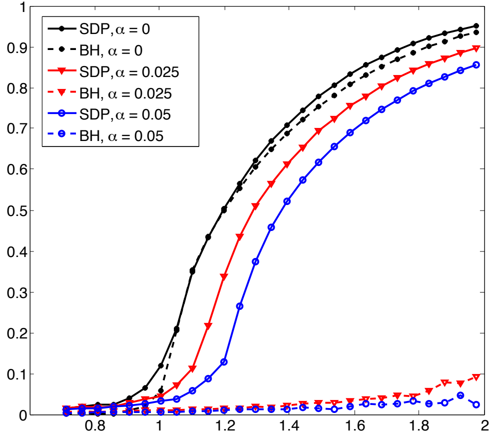
Is this approach robust to model miss-specifications?

An experiment

- Select $S \subseteq V$ uniformly at random. with $|S| = n\alpha$.

- For each $i \in S$, connect all of its neighbors.

An experiment



- Solid line:SDP
- Dashed line: Spectral
  (Non-backtracking walk [Krzakala, Moore, Mossel, Neeman, Sly, Zdeborova, Zhang, 2013])

Conclusion

Conclusion

- SDP $\gg$ PCA when data are heterogeneous
- Sharp information about eigenvalues of random matrices
- A lot of work on SDP with random data

  [Srebro, Fazel, Parrillo, Candés, Recht, Gross, myself, ...]
- **Little known about 'sharp SDP properties' and SDP vs PCA**

Thanks!

Conclusion

- SDP $\gg$ PCA when data are heterogeneous
- Sharp information about eigenvalues of random matrices
- A lot of work on SDP with random data

  [Srebro, Fazel, Parrillo, Candés, Recht, Gross, myself, ...]
- **Little known about 'sharp SDP properties' and SDP vs PCA**

Thanks!

Conclusion

- SDP $\gg$ PCA when data are heterogeneous
- Sharp information about eigenvalues of random matrices
- A lot of work on SDP with random data

  [Srebro, Fazel, Parrillo, Candés, Recht, Gross, myself, ...]
- **Little known about 'sharp SDP properties' and SDP vs PCA**

Thanks!

Conclusion

- SDP $\gg$ PCA when data are heterogeneous
- Sharp information about eigenvalues of random matrices
- A lot of work on SDP with random data

  [Srebro, Fazel, Parrillo, Candés, Recht, Gross, myself, ...]
- **Little known about 'sharp SDP properties' and SDP vs PCA**

Thanks!

Conclusion

- SDP $\gg$ PCA when data are heterogeneous
- Sharp information about eigenvalues of random matrices
- A lot of work on SDP with random data

  [Srebro, Fazel, Parrillo, Candés, Recht, Gross, myself, ...]
- **Little known about 'sharp SDP properties' and SDP vs PCA**

Thanks!