Principal Component Analysis with Structured Factors

Yash Deshpande and Andrea Montanari

Stanford University

May 28, 2014

Principal Component Analysis

Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$

Find $\mathbf{U}, \mathbf{V}$ such that

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^{\mathsf{T}}, \quad \mathbf{U} \in \mathbb{R}^{n \times r}, \quad \mathbf{V} \in \mathbb{R}^{p \times r}$$

Dimensionality reduction: $r \ll n, p$

What happens if $\mathbf{U}, \mathbf{V}$ have special structure?

Principal Component Analysis

Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$

Find $\mathbf{U}, \mathbf{V}$ such that

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^{\mathsf{T}},$$
$$\mathbf{U} \in \mathbb{R}^{n \times r},$$
$$\mathbf{V} \in \mathbb{R}^{p \times r}$$

Dimensionality reduction: $r \ll n, p$

What happens if $\mathbf{U}, \mathbf{V}$ have special structure?

Principal Component Analysis

Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$

Find $\mathbf{U}, \mathbf{V}$ such that

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^{\mathsf{T}},$$
$$\mathbf{U} \in \mathbb{R}^{n \times r},$$
$$\mathbf{V} \in \mathbb{R}^{p \times r}$$

Dimensionality reduction: $r \ll n, p$

What happens if $\mathbf{U}, \mathbf{V}$ have special structure?

I will talk only about one type of structure. . .

Sparse Principal Component Analysis

Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$

Find $\mathbf{U}, \mathbf{V}$ such that

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^{\mathsf{T}},$$
$$\mathbf{U} \in \mathbb{R}^{n \times r},$$
$$\mathbf{V} \in \mathbb{R}^{p \times r} \qquad \text{sparse}$$

Dimensionality reduction: $r \ll n, p$

What happens if $\mathbf{U}, \mathbf{V}$ have special structure?

Equivalently: Superposition of sparse vectors

- Rows of $\mathbf{X}$: $\mathbf{x}_1, \mathbf{x}_2, \dots\dots \mathbf{x}_n \in \mathbb{R}^p$.
- Rows of $\mathbf{V}^{\mathsf{T}}$: $\mathbf{v}_1, \dots \mathbf{v}_r \in \mathbb{R}^p$.

$$\mathbf{x}_i \approx \sum_{\ell=1}^{r} u_{i\ell}\, \mathbf{v}_\ell \qquad \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r \ \text{sparse}$$

Example: Topic models

$$\mathbf{x}_i \approx \sum_{\ell=1}^{r} u_{\ell,i}\, \mathbf{v}_\ell$$

$\mathbf{x}_i$: Document i

$\mathbf{v}_\ell$: Topic ℓ

Document = Superposition of topics.

Topic = Sparse distribution over words

Example: Topic models

$$\mathbf{x}_i \approx \sum_{\ell=1}^{r} u_{\ell,i}\, \mathbf{v}_\ell$$

$\mathbf{x}_i$: Document i

$\mathbf{v}_\ell$: Topic ℓ

Document = Superposition of topics.
Topic = Sparse distribution over words

Example: Topic models

Table 1: Words associated with the top 5 sparse principal components in NYTimes

1st PC (6 words)	2nd PC (5 words)	3rd PC (5 words)	4th PC (4 words)	5th PC (4 words)
million	point	official	president	school
percent	play	government	campaign	program
business	team	united_states	bush	children
company	season	u_s	administration	student
market	game	attack		
companies				

Table 2: Words associated with the top 5 sparse principal components in PubMed

1st PC (5 words)	2nd PC (5 words)	3rd PC (5 words)	4th PC (4 words)	5th PC (4 words)
patient	effect	human	tumor	year
cell	level	expression	mice	infection
treatment	activity	receptor	cancer	age
protein	concentration	binding	maligant	children
disease	rat		carcinoma	child

[Zhang, El Ghaoui, 2011]

Other applications

- Dictionary learning
- Computer vision
- Dimensionality reduction
- ...

Outline

1. Model
2. State of the art
3. Algorithm and motivation
4. Analysis and simulations

[arXiv:1311.5179]

Model

Spiked covariance model

$$\mathbf{X} = \sum_{\ell=1}^{r} \sqrt{\beta_\ell}\, \mathbf{u}_\ell \mathbf{v}_\ell^{\mathsf{T}} + \mathbf{Z}$$

- $\mathbf{Z}_{ij} \sim_{i.i.d.} \mathsf{N}(0,1), \quad \mathbf{u}_\ell \sim \mathsf{N}(0, \mathrm{I}_{n\times n})$
- $p = \Theta(n)$.
- $\|\mathbf{v}_\ell\|_0 \le k$, $\min_{i\in \mathrm{supp}(\mathbf{v}_\ell)} |v_{\ell,i}| \ge v_{\min}/\sqrt{k}$
- r, β_ℓ bounded
- Separation $\beta_1 > \beta_2 > \cdots > \beta_r > 0$

Spiked covariance model

$$\mathbf{X} = \sum_{\ell=1}^{r} \sqrt{\beta_\ell}\, \mathbf{u}_\ell \mathbf{v}_\ell^{\mathsf{T}} + \mathbf{Z}$$

- $\mathbf{Z}_{ij} \sim_{i.i.d.} \mathsf{N}(0,1), \quad \mathbf{u}_\ell \sim \mathsf{N}(0, \mathrm{I}_{n\times n})$
- $p = \Theta(n)$.
- $\|\mathbf{v}_\ell\|_0 \le k$, $\min_{i\in \mathrm{supp}(\mathbf{v}_\ell)} |v_{\ell,i}| \ge v_{\min}/\sqrt{k}$
- r, β_ℓ bounded
- Separation $\beta_1 > \beta_2 > \cdots > \beta_r > 0$

Equivalently

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \sim_{i.i.d.} \mathsf{N}(0, \Sigma)$$

$$\Sigma = \sum_{\ell=1}^{r} \beta_\ell \mathbf{v}_\ell \mathbf{v}_\ell^{\mathsf{T}} + \mathrm{I}\,.$$

Equivalently

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \sim_{i.i.d.} \mathsf{N}(0, \Sigma)$$

$$\Sigma = \sum_{\ell=1}^{r} \beta_\ell \mathbf{v}_\ell \mathbf{v}_\ell^{\mathsf{T}} + \mathrm{I}\,.$$

For ease of exposition: $r = 1$

$$\mathbf{X} = \sqrt{\beta}\,\mathbf{u}\mathbf{v}^{\mathsf{T}} + \mathbf{Z}$$

A definition: Sample covariance

$$\widehat{\Sigma} \equiv \frac{1}{n} \mathbf{X}^{\mathsf{T}} \mathbf{X}$$
$$= \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}}$$

State of the art

Objective: Support recovery

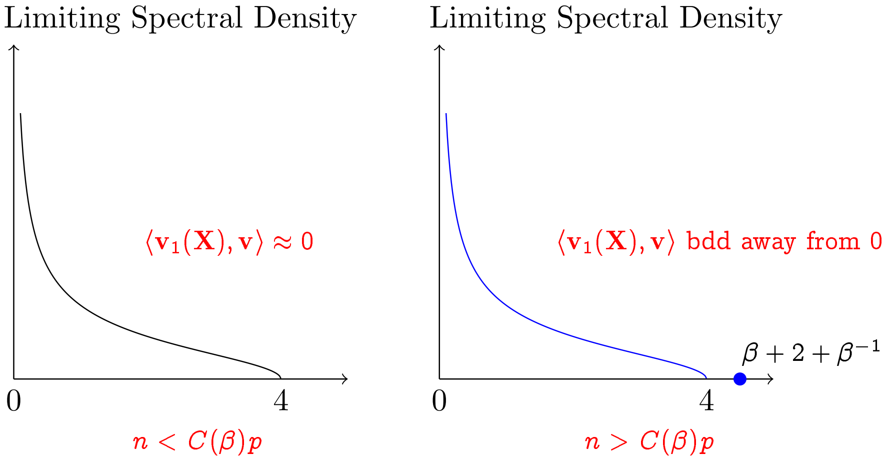
Want to reconstruct supp(v)

Simple PCA

Principal vector of $\mathbf{X}$:

$$\mathbf{v}_1(\mathbf{X})$$

Simple PCA: Spectral phase transition



Principal component is orthogonal to the signal unless $n > C(\beta)p$

[Baik, Ben Arous, Peche, 2005; Baik Silverstein, 2006; Paul, 2007]

Information theory lower bound

For $i \in \{1, \dots, n\}$

$$\mathbf{x}_i = \sqrt{\beta} u_i \mathbf{v} + \mathbf{z}_i$$

- Each sample yields $\Theta(1)$ bits
- Need $(k \log p)$ bits
- Doable if $n \geq C(\beta) k \log p$ (exhaustive search)

[Amini, Wainwright, 2009]

Information theory lower bound

For $i \in \{1, \dots, n\}$

$$\mathbf{x}_i = \sqrt{\beta} u_i \mathbf{v} + \mathbf{z}_i$$

- Each sample yields $\Theta(1)$ bits
- Need $(k \log p)$ bits
- Doable if $n \geq C(\beta) k \log p$ (exhaustive search)

[Amini, Wainwright, 2009]

Can we achieve this in polytime?

What about linear time?

Diagonal thresholding [Johnstone, Lu, 2004]

Idea

$$\Sigma_{ii} = 1 + \beta\, v_i^2$$

$$\widehat{\Sigma}_{ii} = 1 + \beta\, v_i^2 + \frac{1}{\sqrt{n}}\, W_i$$

$$W_i \approx \mathsf{N}(0,1)$$

Support estimate

$$\widehat{\mathsf{Q}} = \Big\{\; i \in [p] : \;\; \widehat{\Sigma}_{ii} \geq \lambda \;\Big\}\,.$$

Diagonal thresholding [Johnstone, Lu, 2004]

Idea

$$\Sigma_{ii} = 1 + \beta\, v_i^2$$

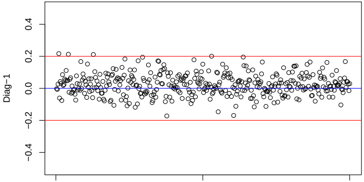
$$\widehat{\Sigma}_{ii} = 1 + \beta\, v_i^2 + \frac{1}{\sqrt{n}}\, W_i$$

$$W_i \approx \mathsf{N}(0,1)$$

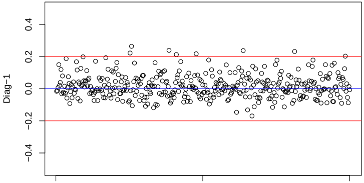
Support estimate

$$\widehat{\mathsf{Q}} = \left\{\; i \in [p] : \;\; \widehat{\Sigma}_{ii} \geq \lambda \;\right\}.$$

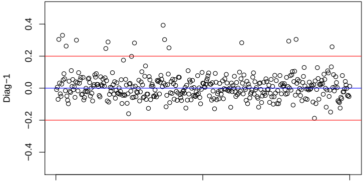
Example: $\beta = 1$, $p = 400$, $n = 300$



$k = 100$



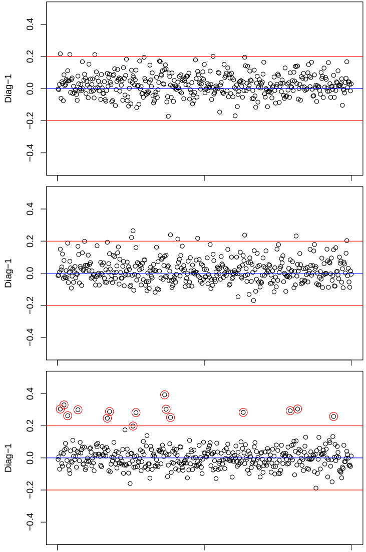
$k = 50$



$k = 15$

noise level $\approx \sqrt{(2 \log p)/n}$, signal $\approx \beta/k$

Example: $\beta = 1$, $p = 400$, $n = 300$



$k = 100$

$k = 50$

$k = 15$

noise level $\approx \sqrt{(2 \log p)/n}$, signal $\approx \beta/k$

Diagonal thresholding

$$\text{noise level} \approx \sqrt{\frac{2\log p}{n}}, \qquad \text{signal} \approx \frac{\beta}{k}$$

Works if

$$\frac{\beta}{k} \geq 10\sqrt{\frac{\log p}{n}}$$

$$k \leq C(\beta)\sqrt{\frac{n}{\log p}}$$

Diagonal thresholding

$$\text{noise level} \approx \sqrt{\frac{2\log p}{n}}, \qquad \text{signal} \approx \frac{\beta}{k}$$

Works if

$$\frac{\beta}{k} \geq 10\sqrt{\frac{\log p}{n}}$$

$$k \leq C(\beta)\sqrt{\frac{n}{\log p}}$$

Diagonal thresholding

$$\text{noise level} \approx \sqrt{\frac{2\log p}{n}}, \qquad \text{signal} \approx \frac{\beta}{k}$$

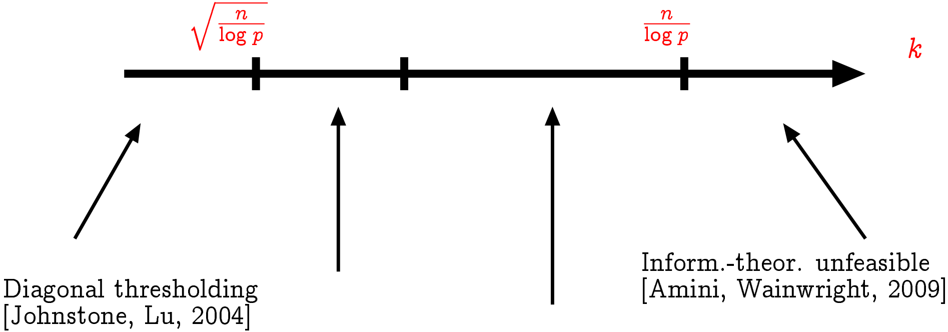
Works if

$$\frac{\beta}{k} \geq 10\sqrt{\frac{\log p}{n}}$$

$$k \leq C(\beta)\sqrt{\frac{n}{\log p}}$$

Executive summary

$r = 1$, $k = \|\mathbf{v}\|_0$ (smaller $k \Rightarrow$ easier)



Complaints about diagonal thresholding

- Sup-optimal sample size
- Sensitive to the i.i.d. noise assumption

Anything better?

SDP relaxation (d'Aspremont, El Ghaoui, Jordan, Lanckriet, 2004)

$$\begin{aligned} \text{maximize} \quad & \mathrm{Tr}(\widehat{\Sigma}\mathbf{W}), \\ \text{subjectto} \quad & \mathbf{W} \succeq 0, \\ & \mathrm{Tr}(\mathbf{W}) = 1, \\ & \sum_{i,j=1}^{p} |\mathbf{W}_{ij}| \le \xi. \end{aligned}$$

- Amini, Wainwright 2009: Conditionally positive results
- Krauthgamer, Nadler, Vilechnik, 2013: Fails for $k \gtrsim \sqrt{n}$

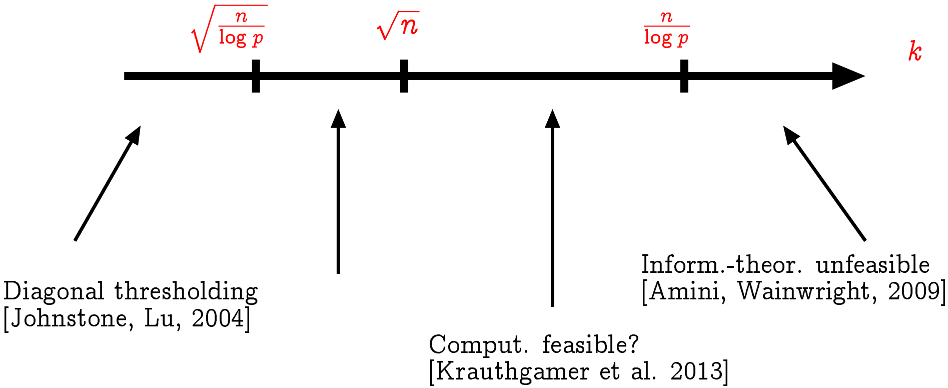
SDP relaxation (d'Aspremont, El Ghaoui, Jordan, Lanckriet, 2004)

$$
\begin{aligned}
\text{maximize} \quad & \mathsf{Tr}(\widehat{\Sigma}\mathbf{W}), \\
\text{subjectto} \quad & \mathbf{W} \succeq 0, \\
& \mathsf{Tr}(\mathbf{W}) = 1, \\
& \sum_{i,j=1}^{p} |\mathbf{W}_{ij}| \le \xi .
\end{aligned}
$$

- Amini, Wainwright 2009: Conditionally positive results
- Krauthgamer, Nadler, Vilechnik, 2013: Fails for $k \gtrsim \sqrt{n}$

Executive summary

$r = 1$, $k = \|\mathbf{v}\|_0$ (smaller $k \Rightarrow$ easier)



A computational barrier?

Theorem (Berthet, Rigollet, 2013)

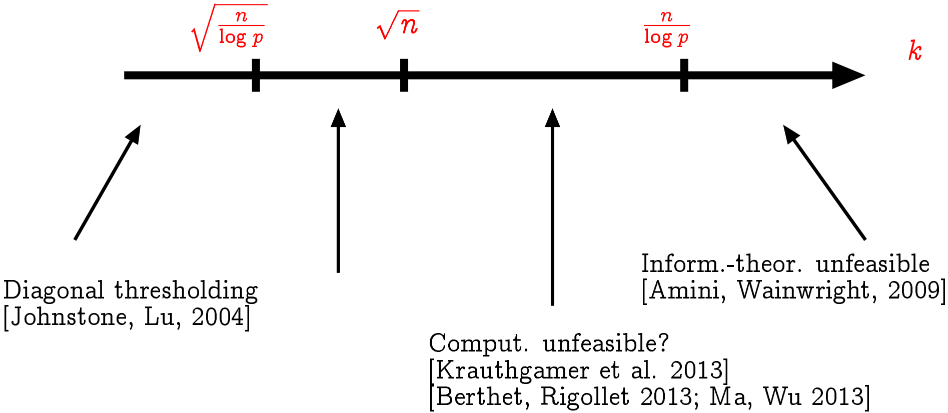
Assume that PLANTEDCLIQUE *cannot be solved in polynomial time for clique size* $n^{0.001} \le |\text{Clique}| \le n^{0.499}$.
Then[a] $\text{supp}(\mathbf{v})$ *cannot be found in polynomial time for* $k \le n^{0.499}$.

[a] Slightly different model

[See also Ma, Wu 2013]

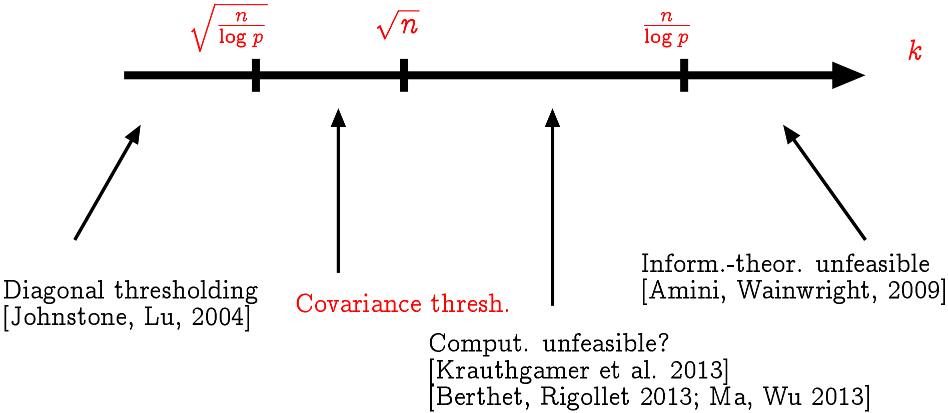
Executive summary

$r = 1$, $k = \|\mathbf{v}\|_0$ (smaller $k \Rightarrow$ easier)



Executive summary: This paper

$r = 1$, $k = \|\mathbf{v}\|_0$ (smaller $k \Rightarrow$ easier)



Algorithm and motivation

Sample covariance

Population covariance

$$\Sigma = \beta \mathbf{v}\mathbf{v}^{\mathsf{T}} + \mathrm{I}$$

Sample covariance

$$\widehat{\Sigma} \equiv \frac{1}{n}\mathbf{X}^{\mathsf{T}}\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}}$$

$$\widehat{\Sigma} = \beta\, \mathbf{v}\mathbf{v}^{\mathsf{T}} + \mathrm{I} + \text{noise}$$

Sample covariance

Population covariance

$$\Sigma = \beta \mathbf{v}\mathbf{v}^{\mathsf{T}} + \mathrm{I}$$

Sample covariance

$$\widehat{\Sigma} \equiv \frac{1}{n}\mathbf{X}^{\mathsf{T}}\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}}$$

$$\widehat{\Sigma} = \beta\, \mathbf{v}\mathbf{v}^{\mathsf{T}} + \mathrm{I} + \text{noise}$$

Covariance thresholding

- Bickel, Levina 2009
- Proposed for SPCA by Krauthgamer, Nadler, Vilechnik 2013

Covariance thresholding

$$\widehat{\Sigma} - \mathrm{I} = \beta\, \mathbf{v}\, \mathbf{v}^{\mathsf{T}} \qquad \text{Sparse, Norm} = \beta$$

$$+ \text{noise} \qquad \text{Dense, Norm} = c\sqrt{p/n}$$

Threshold entries at level $\lambda = \tau/\sqrt{n}$

$$\mathrm{ST}_\lambda(\widehat{\Sigma}) - c\mathrm{I} \approx \mathrm{ST}_\lambda(\beta\, \mathbf{v}\, \mathbf{v}^{\mathsf{T}}) \qquad \text{Norm} \approx \beta$$

$$+ \text{noise} \qquad \text{Norm} \approx \varepsilon(\tau)\sqrt{p/n}$$

$\mathrm{ST}_\lambda \equiv$ soft thresholding at level λ

Covariance thresholding

$$\widehat{\Sigma} - \mathrm{I} = \beta\, \mathbf{v}\, \mathbf{v}^{\mathsf{T}} + \text{noise}$$

Sparse, Norm $= \beta$

Dense, Norm $= c\sqrt{p/n}$

Threshold entries at level $\lambda = \tau/\sqrt{n}$

$$\mathsf{ST}_\lambda(\widehat{\Sigma}) - c\mathrm{I} \approx \mathsf{ST}_\lambda(\beta\, \mathbf{v}\, \mathbf{v}^{\mathsf{T}}) + \text{noise}$$

Norm $\approx \beta$

Norm $\approx \varepsilon(\tau)\sqrt{p/n}$

$\mathsf{ST}_\lambda \equiv$ soft thresholding at level λ

Covariance thresholding

$$\widehat{\Sigma} - \mathrm{I} = \beta\, \mathbf{v}\, \mathbf{v}^{\mathsf{T}} + \text{noise}$$

Sparse, Norm $= \beta$

Dense, Norm $= c\sqrt{p/n}$

Threshold entries at level $\lambda = \tau/\sqrt{n}$

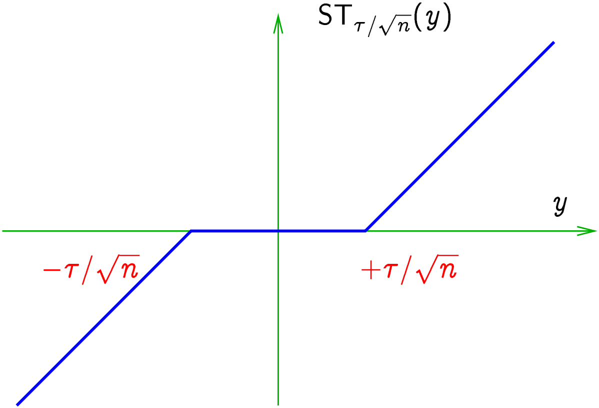
$$\mathsf{ST}_\lambda(\widehat{\Sigma}) - c\mathrm{I} \approx \mathsf{ST}_\lambda(\beta\, \mathbf{v}\, \mathbf{v}^{\mathsf{T}}) + \text{noise}$$

Norm $\approx \beta$

Norm $\approx \varepsilon(\tau)\sqrt{p/n}$

$\mathsf{ST}_\lambda \equiv$ soft thresholding at level λ

ST



Covariance thresholding

1: **Input:** Data $(\mathbf{x}_i)_{1\le i\le 2n}$, parameter $\tau \in \mathbb{R}_{\ge 0}$;

2: Compute $\widehat{\Sigma}$;

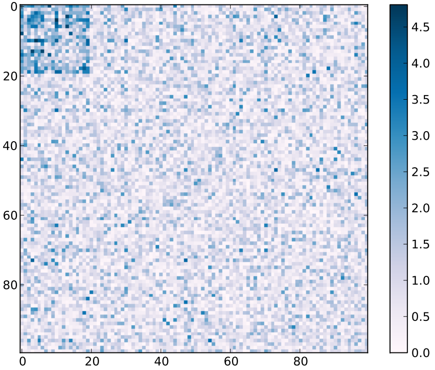
3: Set $\mathsf{ST}_{\tau/\sqrt{n}}(\widehat{\Sigma})_{ii} = 0$ and (for $i \neq j$):

$$\mathsf{ST}_{\tau/\sqrt{n}}(\widehat{\Sigma})_{ij} = \begin{cases} \widehat{\Sigma}_{ij} - \frac{\tau}{\sqrt{n}} & \text{if } \widehat{\Sigma}_{ij} \ge \tau/\sqrt{n}, \\ 0 & \text{if } -\tau/\sqrt{n} < \widehat{\Sigma}_{ij} < \tau/\sqrt{n}, \\ \widehat{\Sigma}_{ij} + \frac{\tau}{\sqrt{n}} & \text{if } \widehat{\Sigma}_{ij} \le -\tau/\sqrt{n}, \end{cases}$$

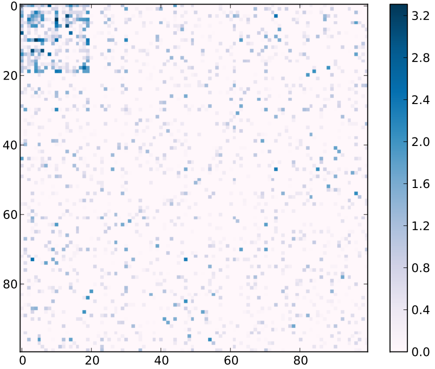
4: $\mathbf{v}_* =$ Principal eigenvector of $\mathsf{ST}_{\tau/\sqrt{n}}(\widehat{\Sigma})$;

5: 'Clean' $\mathbf{v}_*$ to estimate support $\widehat{\mathsf{Q}}$.

$\widehat{\Sigma}$



$\mathrm{ST}_{1.5/\sqrt{n}}(\widehat{\Sigma})$



Analysis and simulations

A theorem

Theorem (Deshpande, Montanari, 2013)

For any $\alpha, \beta, \varepsilon > 0$, there exists $C = C(\alpha, \beta, \varepsilon) > 0$ such that the following happens for signal to noise ratio β, and $p/n = \alpha$.

If $k \le C\sqrt{n}$, then, with high probability,

- $\|\mathbf{v}^* - \mathbf{v}\|_2 \le \varepsilon$
- $\widehat{\mathrm{Q}} = \mathrm{supp}(\mathrm{v})$

A theorem

Theorem (Deshpande, Montanari, 2013)

For any $\alpha, \beta, \varepsilon > 0$, there exists $C = C(\alpha, \beta, \varepsilon) > 0$ such that the following happens for signal to noise ratio β, and $p/n = \alpha$.

If $k \le C\sqrt{n}$, then, with high probability,

- $\|\mathbf{v}^* - \mathbf{v}\|_2 \le \varepsilon$
- $\widehat{\mathrm{Q}} = \mathrm{supp}(\mathrm{v})$

A theorem

Theorem (Deshpande, Montanari, 2013)

For any $\alpha, \beta, \varepsilon > 0$, there exists $C = C(\alpha, \beta, \varepsilon) > 0$ such that the following happens for signal to noise ratio β, and $p/n = \alpha$.

If $k \le C\sqrt{n}$, then, with high probability,

- $\|\mathbf{v}^* - \mathbf{v}\|_2 \le \varepsilon$
- $\widehat{\mathrm{Q}} = \mathrm{supp}(\mathrm{v})$

Crucial lemma: Kernel random matrices

Lemma (Deshpande, Montanari, 2013)

Assume $\mathbf{Z} = (Z_{ij})_{i\le n, j\le p}$ *with* $Z_{ij} \sim_{i.i.d.} \mathsf{N}(0, 1/n)$, $p/n \to \alpha$. *Then, with high probability*

$$\left\| \mathsf{ST}_{\tau/\sqrt{n}}\Big(\mathbf{Z}\mathbf{Z}^{\mathsf{T}} - \mathrm{diag}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}})\Big) \right\|_2 \le C(\alpha)\,\tau^{-0.49}\,.$$

$\le C(\alpha, \tau)$ is easy

Crucial lemma: Kernel random matrices

Lemma (Deshpande, Montanari, 2013)

Assume $\mathbf{Z} = (Z_{ij})_{i \le n, j \le p}$ *with* $Z_{ij} \sim_{i.i.d.} \mathsf{N}(0, 1/n)$, $p/n \to \alpha$. *Then, with high probability*

$$\left\| \mathsf{ST}_{\tau/\sqrt{n}}\left(\mathbf{Z}\mathbf{Z}^{\mathsf{T}} - \mathrm{diag}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}})\right) \right\|_2 \le C(\alpha)\, \tau^{-0.49}\,.$$

$\le C(\alpha, \tau)$ is easy

Proof

- $\mathsf{S}(\,\cdot\,) \equiv \mathsf{ST}_{\tau/\sqrt{n}}(\,\cdot\,)$
- $\mathbf{N} = \mathsf{S}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}} - \mathrm{diag}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}}))$
- $T_\varepsilon \subseteq S_{p-1} \subseteq \mathbb{R}^p$ an ε-net, $|T_\varepsilon| \le (10/\varepsilon)^p$.

$$\mathbb{P}\Big\{\sup_{\mathbf{y}\in S_{p-1}} \langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle \ge \Delta\Big\} \le |T_\varepsilon| \sup_{\mathbf{y}\in T_\varepsilon} \mathbb{P}\Big\{\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle \ge (1-2\varepsilon)\Delta\Big\}$$

Sufficient to prove that

$$\sup_{\mathbf{y}\in T_\varepsilon} \mathbb{P}\Big\{\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle \ge C\tau^{-0.49}\Big\} \le 2\,e^{-cn}$$

In simple random matrix ensembles:

$\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle$ is Lipschitz in $\mathbf{Z}$ $\Rightarrow$ Gaussian isoperimetry

Proof

- $\mathsf{S}(\cdot) \equiv \mathsf{ST}_{\tau/\sqrt{n}}(\cdot)$
- $\mathbf{N} = \mathsf{S}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}} - \mathrm{diag}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}}))$
- $T_\varepsilon \subseteq S_{p-1} \subseteq \mathbb{R}^p$ an ε-net, $|T_\varepsilon| \le (10/\varepsilon)^p$.

$$\mathbb{P}\Big\{ \sup_{\mathbf{y}\in S_{p-1}} \langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle \ge \Delta \Big\} \le |T_\varepsilon| \sup_{\mathbf{y}\in T_\varepsilon} \mathbb{P}\Big\{ \langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle \ge (1-2\varepsilon)\Delta \Big\}$$

Sufficient to prove that

$$\sup_{\mathbf{y}\in T_\varepsilon} \mathbb{P}\Big\{ \langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle \ge C\tau^{-0.49} \Big\} \le 2\, e^{-cn}$$

In simple random matrix ensembles:

$\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle$ is Lipschitz in $\mathbf{Z}$ $\Rightarrow$ Gaussian isoperimetry

Proof

- $\mathsf{S}(\,\cdot\,) \equiv \mathsf{ST}_{\tau/\sqrt{n}}(\,\cdot\,)$
- $\mathbf{N} = \mathsf{S}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}} - \mathrm{diag}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}}))$
- $T_\varepsilon \subseteq S_{p-1} \subseteq \mathbb{R}^p$ an ε-net, $|T_\varepsilon| \le (10/\varepsilon)^p$.

$$\mathbb{P}\Big\{\sup_{\mathbf{y}\in S_{p-1}} \langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle \ge \Delta\Big\} \le |T_\varepsilon| \sup_{\mathbf{y}\in T_\varepsilon} \mathbb{P}\Big\{\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle \ge (1-2\varepsilon)\Delta\Big\}$$

Sufficient to prove that

$$\sup_{\mathbf{y}\in T_\varepsilon} \mathbb{P}\Big\{\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle \ge C\tau^{-0.49}\Big\} \le 2\, e^{-cn}$$

In simple random matrix ensembles:
$\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle$ is Lipschitz in $\mathbf{Z}$ $\Rightarrow$ Gaussian isoperimetry

Let's try the same approach

Columns of $\mathbf{Z}$: $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_p \sim \mathsf{N}(0, \mathrm{I}_{n\times n})$

$$\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = \sum_{i\neq j} y_i \mathsf{S}\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big) y_j$$

$$\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = 2\frac{y_i}{n}\sum_{j\in[p]\setminus i} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big)\mathbf{g}_j y_j$$

Problems:

- Need $\|\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle\|_2 \le C\tau^{-0.49}$
- $\mathbf{g}_j$ unbounded
- y_i can be $\sqrt{p}$ times its typical value
- If we use $|\mathsf{S}'(\,\cdot\,)| \le 1$, we loose the dependence in τ

Looks hopeless!

Let's try the same approach

Columns of $\mathbf{Z}$: $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_p \sim \mathsf{N}(0, \mathrm{I}_{n\times n})$

$$\langle \mathbf{y}, \mathbf{N}\mathbf{y} \rangle = \sum_{i \neq j} y_i \mathsf{S}\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j \rangle}{n}\Big) y_j$$

$$\nabla_{\mathbf{g}_i} \langle \mathbf{y}, \mathbf{N}\mathbf{y} \rangle = 2\frac{y_i}{n} \sum_{j \in [p] \setminus i} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j \rangle}{n}\Big) \mathbf{g}_j y_j$$

Problems:

- Need $\|\nabla_{\mathbf{g}_i} \langle \mathbf{y}, \mathbf{N}\mathbf{y} \rangle\|_2 \leq C\tau^{-0.49}$
- $\mathbf{g}_j$ unbounded
- y_i can be $\sqrt{p}$ times its typical value
- If we use $|\mathsf{S}'(\,\cdot\,)| \leq 1$, we loose the dependence in τ

Looks hopeless!

Let's try the same approach

Columns of $\mathbf{Z}$: $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_p \sim \mathsf{N}(0, \mathrm{I}_{n\times n})$

$$\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = \sum_{i\neq j} y_i \mathsf{S}\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big) y_j$$

$$\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = 2\frac{y_i}{n} \sum_{j\in[p]\setminus i} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big) \mathbf{g}_j y_j$$

Problems:

- Need $\|\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle\|_2 \le C\tau^{-0.49}$
- $\mathbf{g}_j$ unbounded
- y_i can be $\sqrt{p}$ times its typical value
- If we use $|\mathsf{S}'(\,\cdot\,)| \le 1$, we loose the dependence in τ

Looks hopeless!

Let's try the same approach

Columns of $\mathbf{Z}$: $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_p \sim \mathsf{N}(0, \mathrm{I}_{n\times n})$

$$\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = \sum_{i\neq j} y_i \mathsf{S}\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big) y_j$$

$$\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = 2\frac{y_i}{n}\sum_{j\in[p]\setminus i} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big)\mathbf{g}_j y_j$$

Problems:

- Need $\|\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle\|_2 \le C\tau^{-0.49}$
- $\mathbf{g}_j$ unbounded
- y_i can be $\sqrt{p}$ times its typical value
- If we use $|\mathsf{S}'(\,\cdot\,)| \le 1$, we loose the dependence in τ

Looks hopeless!

Let's try the same approach

Columns of $\mathbf{Z}$: $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_p \sim \mathsf{N}(0, \mathrm{I}_{n \times n})$

$$\langle \mathbf{y}, \mathbf{N}\mathbf{y} \rangle = \sum_{i \neq j} y_i \mathsf{S}\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j \rangle}{n}\Big) y_j$$

$$\nabla_{\mathbf{g}_i} \langle \mathbf{y}, \mathbf{N}\mathbf{y} \rangle = 2\frac{y_i}{n} \sum_{j \in [p] \setminus i} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j \rangle}{n}\Big) \mathbf{g}_j y_j$$

Problems:

- Need $\|\nabla_{\mathbf{g}_i} \langle \mathbf{y}, \mathbf{N}\mathbf{y} \rangle\|_2 \leq C\tau^{-0.49}$
- $\mathbf{g}_j$ unbounded
- y_i can be $\sqrt{p}$ times its typical value
- If we use $|\mathsf{S}'(\,\cdot\,)| \leq 1$, we loose the dependence in τ

Looks hopeless!

Let's try the same approach

Columns of $\mathbf{Z}$: $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_p \sim \mathsf{N}(0, \mathrm{I}_{n\times n})$

$$\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = \sum_{i\neq j} y_i \mathsf{S}\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big) y_j$$

$$\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = 2\frac{y_i}{n}\sum_{j\in[p]\setminus i} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big)\mathbf{g}_j y_j$$

Problems:

- Need $\|\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle\|_2 \leq C\tau^{-0.49}$
- $\mathbf{g}_j$ unbounded
- y_i can be $\sqrt{p}$ times its typical value
- If we use $|\mathsf{S}'(\,\cdot\,)| \leq 1$, we loose the dependence in τ

Looks hopeless!

Let's try the same approach

Columns of $\mathbf{Z}$: $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_p \sim \mathsf{N}(0, \mathrm{I}_{n\times n})$

$$\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = \sum_{i\neq j} y_i \mathsf{S}\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big) y_j$$

$$\nabla_{\mathbf{g}_i} \langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = 2\frac{y_i}{n} \sum_{j\in[p]\setminus i} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big) \mathbf{g}_j y_j$$

Problems:

- Need $\|\nabla_{\mathbf{g}_i} \langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle\|_2 \leq C\tau^{-0.49}$
- $\mathbf{g}_j$ unbounded
- y_i can be $\sqrt{p}$ times its typical value
- If we use $|\mathsf{S}'(\,\cdot\,)| \leq 1$, we loose the dependence in τ

Looks hopeless!

Let's try the same approach

Columns of $\mathbf{Z}$: $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_p \sim \mathsf{N}(0, \mathrm{I}_{n\times n})$

$$\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = \sum_{i\neq j} y_i \mathsf{S}\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big) y_j$$

$$\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = 2\frac{y_i}{n}\sum_{j\in[p]\setminus i} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big)\mathbf{g}_j y_j$$

Problems:

- Need $\|\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle\|_2 \le C\tau^{-0.49}$
- $\mathbf{g}_j$ unbounded
- y_i can be $\sqrt{p}$ times its typical value
- If we use $|\mathsf{S}'(\,\cdot\,)| \le 1$, we loose the dependence in τ

Looks hopeless!

Ideas (1)

y_i can be $\sqrt{n}$ times its typical value

- Separate big entries of y (above $C/\sqrt{p}$)
- There are at most p/C big entries
- Control norm of all $(p/C) \times (p/C)$ submatrices

Ideas (1)

y_i can be $\sqrt{n}$ times its typical value

- Separate big entries of $\mathbf{y}$ (above $C/\sqrt{p}$)
- There are at most p/C big entries
- Control norm of all $(p/C) \times (p/C)$ submatrices

Ideas (2)

$$\nabla_{\mathbf{g}_i}\langle \mathbf{y}, \mathbf{N}\mathbf{y}\rangle = \frac{2y_i}{n}\sum_{j\in[p]\setminus i} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big)\mathbf{g}_j y_j$$

$$= \mathbf{Z}^{\mathsf{T}}\boldsymbol{\sigma}^i(\mathbf{y})$$

where

$$[\boldsymbol{\sigma}^i(\mathbf{y})]_j = \frac{2y_i}{n}\mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j\rangle}{n}\Big)y_j$$

Prove that, with overwhelming probability

- $\|\mathbf{Z}\|_2 \le$const. (known)
- $\|\boldsymbol{\sigma}^i(\mathbf{y})\|_2 \le a\tau^{-0.49}$ (work)

Ideas (2)

$$\nabla_{\mathbf{g}_i} \langle \mathbf{y}, \mathbf{N}\mathbf{y} \rangle = \frac{2y_i}{n} \sum_{j \in [p] \setminus i} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j \rangle}{n}\Big) \mathbf{g}_j y_j$$
$$= \mathbf{Z}^{\mathsf{T}} \boldsymbol{\sigma}^i(\mathbf{y})$$

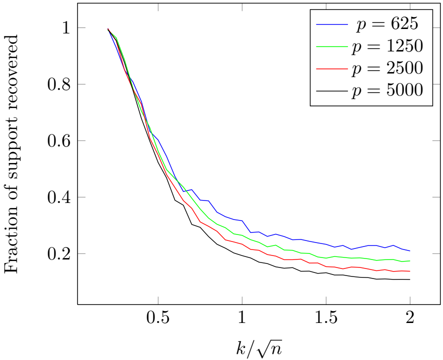
where

$$[\boldsymbol{\sigma}^i(\mathbf{y})]_j = \frac{2y_i}{n} \mathsf{S}'\Big(\frac{\langle \mathbf{g}_i, \mathbf{g}_j \rangle}{n}\Big) y_j$$

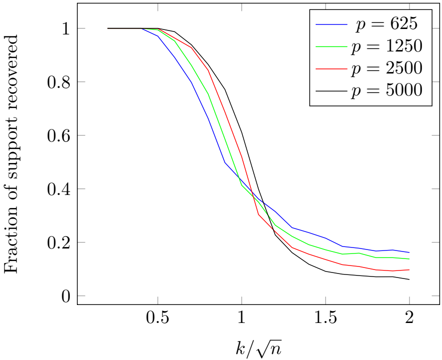
Prove that, with overwhelming probability

- $\|\mathbf{Z}\|_2 \leq$const. (known)
- $\|\boldsymbol{\sigma}^i(\mathbf{y})\|_2 \leq a\tau^{-0.49}$ (work)

Threshold behavior

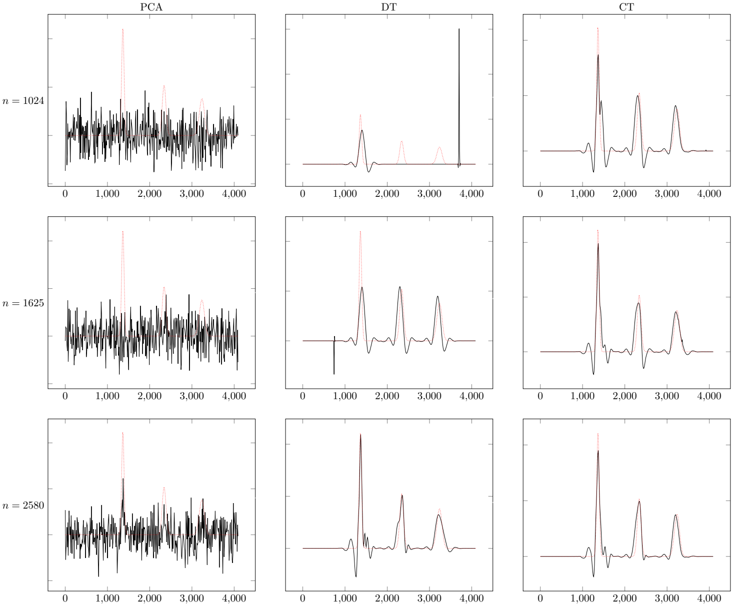


Diagonal thresholding



Covariance thresholding

Sparsity in wavelet domain



Conclusion/Open problems

- It would be nice to understand better kernel random matrices.
- $k = \Theta(\sqrt{n})$: Stronger lower bounds?
- Use sparsification to accelerate this
- Other algorithms for sparse PCA: ask me...

Thanks!

Conclusion/Open problems

- It would be nice to understand better kernel random matrices.
- $k = \Theta(\sqrt{n})$: Stronger lower bounds?
- Use sparsification to accelerate this
- Other algorithms for sparse PCA: ask me...

Thanks!

Conclusion/Open problems

- It would be nice to understand better kernel random matrices.
- $k = \Theta(\sqrt{n})$: Stronger lower bounds?
- Use sparsification to accelerate this
- Other algorithms for sparse PCA: ask me…

Thanks!

Conclusion/Open problems

- It would be nice to understand better kernel random matrices.
- $k = \Theta(\sqrt{n})$: Stronger lower bounds?
- Use sparsification to accelerate this
- Other algorithms for sparse PCA: ask me...

Thanks!

Conclusion/Open problems

- It would be nice to understand better kernel random matrices.
- $k = \Theta(\sqrt{n})$: Stronger lower bounds?
- Use sparsification to accelerate this
- Other algorithms for sparse PCA: ask me...

Thanks!

Conclusion/Open problems

- It would be nice to understand better kernel random matrices.
- $k = \Theta(\sqrt{n})$: Stronger lower bounds?
- Use sparsification to accelerate this
- Other algorithms for sparse PCA: ask me...

Thanks!