Average-case integrality gap for non-negative principal component analysis

Afonso Bandeira, Dmitriy Kunisky, Alexander Wein

MSML 2021

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

Classical computational problem, especially in statistics:

$$\begin{array}{ll} \text{maximize} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subject to} & \boldsymbol{x} \in \mathbb{S}^{n-1} \end{array}$$

Optimizer is the leading eigenvector or “principal component” of $\boldsymbol{W}$ (e.g., $\boldsymbol{W}$ a sample covariance matrix).

Principal Component Analysis (PCA)

Classical computational problem, especially in statistics:

$$\begin{array}{ll} \text{maximize} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subject to} & \boldsymbol{x} \in \mathbb{S}^{n-1} \end{array}$$

Optimizer is the leading eigenvector or "principal component" of $\boldsymbol{W}$ (e.g., $\boldsymbol{W}$ a sample covariance matrix).

Elaborated version, allowing prior information:

$$\begin{array}{ll} \text{maximize} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subject to} & \boldsymbol{x} \in \mathcal{X} \subset \mathbb{S}^{n-1} \end{array}$$

This paper: non-negative PCA, with $\mathcal{X} = \mathbb{R}^n_+$.

Spiked Matrix Model

Well-studied toy model of PCA:

- Under $\mathbb{Q}$, observe $\boldsymbol{W} \sim \mathsf{GOE}(n)$ (i.i.d. Gaussian entries, symmetrized).
- Under $\mathbb{P}$, observe $\boldsymbol{W} = \boldsymbol{W}^{(0)} + \lambda \boldsymbol{x}\boldsymbol{x}^\top$ for $\boldsymbol{W}^{(0)} \sim \mathsf{GOE}(n)$, $\boldsymbol{x}$ drawn from some signal distribution over $\mathcal{X} \subset \mathbb{S}^{n-1}$.

Spiked Matrix Model

Well-studied toy model of PCA:

- Under $\mathbb{Q}$, observe $\boldsymbol{W} \sim \mathsf{GOE}(n)$ (i.i.d. Gaussian entries, symmetrized).
- Under $\mathbb{P}$, observe $\boldsymbol{W} = \boldsymbol{W}^{(0)} + \lambda \boldsymbol{x}\boldsymbol{x}^\top$ for $\boldsymbol{W}^{(0)} \sim \mathsf{GOE}(n)$, $\boldsymbol{x}$ drawn from some signal distribution over $\mathcal{X} \subset \mathbb{S}^{n-1}$.

Interesting questions:

- Can we **distinguish** $\mathbb{Q}$ from $\mathbb{P}$?
- Can we **estimate** the signal $\boldsymbol{x}$ under $\mathbb{P}$?
- Can we **optimize** the likelihood $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}$ over $\boldsymbol{x} \in \mathcal{X}$?

Spiked Matrix Model

Well-studied toy model of PCA:

- Under $\mathbb{Q}$, observe $\boldsymbol{W} \sim \mathsf{GOE}(n)$ (i.i.d. Gaussian entries, symmetrized).
- Under $\mathbb{P}$, observe $\boldsymbol{W} = \boldsymbol{W}^{(0)} + \lambda \boldsymbol{x}\boldsymbol{x}^\top$ for $\boldsymbol{W}^{(0)} \sim \mathsf{GOE}(n)$, $\boldsymbol{x}$ drawn from some signal distribution over $\mathcal{X} \subset \mathbb{S}^{n-1}$.

Interesting questions:

- Can we **distinguish** $\mathbb{Q}$ from $\mathbb{P}$?
- Can we **estimate** the signal $\boldsymbol{x}$ under $\mathbb{P}$?
- Can we **optimize** the likelihood $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}$ over $\boldsymbol{x} \in \mathcal{X}$?

This paper: optimization for $\mathbb{Q} = \mathsf{GOE}(n)$, $\mathcal{X} = \mathbb{R}^n_+ \cap \mathbb{S}^{n-1}$.

Algorithmic Approaches [Montanari, Richard 2015]

Algorithmic Approaches [Montanari, Richard 2015]

Scale to have $\lambda_{\mathsf{max}}(\boldsymbol{W}) \approx 2$. Then, using **approximate message-passing** algorithm have

$$\lambda^+(\boldsymbol{W}) := \left\{ \begin{array}{ll} \text{maximize} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{x}\| = 1 \\ & x_i \geq 0 \end{array} \right\} \approx \sqrt{2}$$

Algorithmic Approaches [Montanari, Richard 2015]

Scale to have $\lambda_{\max}(\boldsymbol{W}) \approx 2$. Then, using **approximate message-passing** algorithm have

$$\lambda^+(\boldsymbol{W}) := \left\{ \begin{array}{ll} \text{maximize} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{x}\| = 1 \\ & x_i \geq 0 \end{array} \right\} \approx \sqrt{2}$$

Alternative: **semidefinite program** with $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ relaxed to

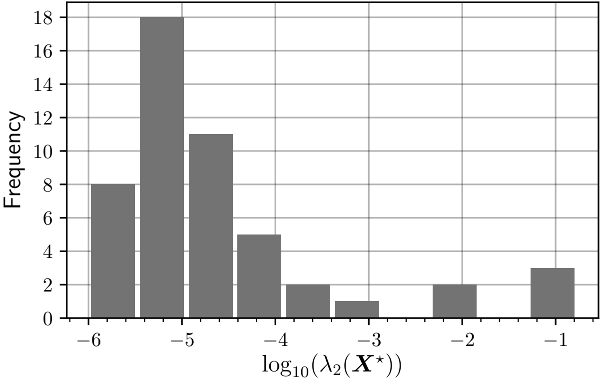
$$\mathsf{SDP}(\boldsymbol{W}) := \left\{ \begin{array}{ll} \text{maximize} & \langle \boldsymbol{X}, \boldsymbol{W} \rangle \\ \text{subject to} & \boldsymbol{X} \succeq \boldsymbol{0} \\ & \mathsf{Tr}(\boldsymbol{X}) = 1 \\ & X_{ij} \geq 0 \end{array} \right\} \geq \lambda^+(\boldsymbol{W})$$

Seems to recover under $\mathbb{P}$ $\rightsquigarrow$ expect $\mathsf{SDP}(\boldsymbol{W}) \approx \sqrt{2}$ under $\mathbb{Q}$.

Small SDP Experiments $(n = 150)$

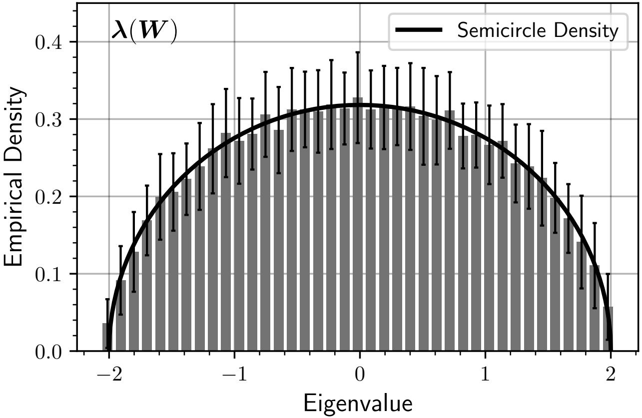
$$\mathsf{SDP}(\boldsymbol{W}) = \left\{ \begin{array}{ll} \text{maximize} & \langle \boldsymbol{X}, \boldsymbol{W} \rangle \\ \text{subject to} & \boldsymbol{X} \succeq \boldsymbol{0}, \mathsf{Tr}(\boldsymbol{X}) = 1, X_{ij} \geq 0 \end{array} \right\}$$

Optimizer $\boldsymbol{X}^\star$ has numerical rank ≈ 1:



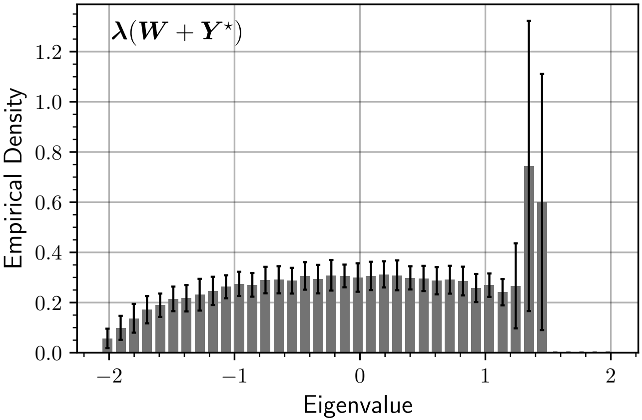
Small SDP Experiments $(n = 150)$

$$\mathsf{SDP}(\boldsymbol{W}) := \left\{ \begin{array}{ll} \text{minimize} & \lambda_{\max}(\boldsymbol{W} + \boldsymbol{Y}) \\ \text{subject to} & Y_{ij} \geq 0 \end{array} \right\}$$



Small SDP Experiments $(n = 150)$

$$\mathsf{SDP}(\boldsymbol{W}) := \left\{ \begin{array}{ll} \text{minimize} & \lambda_{\max}(\boldsymbol{W} + \boldsymbol{Y}) \\ \text{subject to} & Y_{ij} \geq 0 \end{array} \right\} \overset{?}{\approx} \sqrt{2}$$



An Unpleasant Surprise

Theorem 1: With high probability $\mathsf{SDP}(\boldsymbol{W}) \approx 2 \approx \lambda_{\max}(\boldsymbol{W})$.

An Unpleasant Surprise

Theorem 1: With high probability $\mathsf{SDP}(\boldsymbol{W}) \approx 2 \approx \lambda_{\max}(\boldsymbol{W})$.

Proof Sketch: Need $\boldsymbol{X}$ feasible with $\langle \boldsymbol{X}, \boldsymbol{W} \rangle \approx \lambda_{\max}(\boldsymbol{W})$.

An Unpleasant Surprise

Theorem 1: With high probability $\mathsf{SDP}(\boldsymbol{W}) \approx 2 \approx \lambda_{\mathsf{max}}(\boldsymbol{W})$.

Proof Sketch: Need $\boldsymbol{X}$ feasible with $\langle \boldsymbol{X}, \boldsymbol{W} \rangle \approx \lambda_{\mathsf{max}}(\boldsymbol{W})$.

Let $\boldsymbol{P}$ be the projector to top δn eigenspaces of $\boldsymbol{W}$.

$$\boldsymbol{X}^{(0)} := \frac{1}{\delta n} \boldsymbol{P}$$

almost feasible; only $X_{ij}^{(0)} < 0$ sometimes.

An Unpleasant Surprise

Theorem 1: With high probability $\mathsf{SDP}(\boldsymbol{W}) \approx 2 \approx \lambda_{\mathsf{max}}(\boldsymbol{W})$.

Proof Sketch: Need $\boldsymbol{X}$ feasible with $\langle \boldsymbol{X}, \boldsymbol{W} \rangle \approx \lambda_{\mathsf{max}}(\boldsymbol{W})$.

Let $\boldsymbol{P}$ be the projector to top δn eigenspaces of $\boldsymbol{W}$.

$$\boldsymbol{X}^{(0)} := \frac{1}{\delta n}\boldsymbol{P}$$

almost feasible; only $X_{ij}^{(0)} < 0$ sometimes.

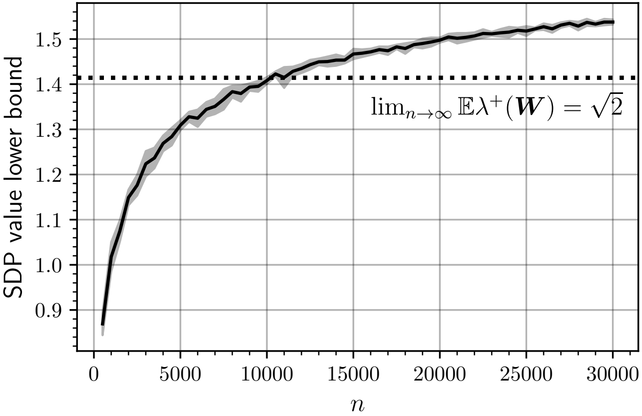
So, **nudge** towards feasible set:

$$\boldsymbol{X} := \alpha \frac{1}{n}\mathbf{1}\mathbf{1}^\top + (1-\alpha)\boldsymbol{X}^{(0)}.$$

Analysis of off-diagonal entries $\rightsquigarrow$ feasible for any $\alpha > 0$.

Larger Experiments ($\delta = 1/25$, α as small as possible)

Expect to need $n \gtrsim 10^4$ to see $\mathsf{SDP}(\boldsymbol{W}) > \sqrt{2}$. Much bigger than scale of quickly solving SDP!



General Certification Algorithms

What's going on—is the SDP just not fancy enough and there are better similar methods?

General Certification Algorithms

What's going on—is the SDP just not fancy enough and there are better similar methods?

Natural follow-up question: does there exist a better efficient **certification algorithm**? That is, $c(\boldsymbol{W})$ with

$$\lambda^+(\boldsymbol{W}) \le c(\boldsymbol{W})$$

such that

- The bound holds **for all** $\boldsymbol{W} \in \mathbb{R}^{n \times n}_{\mathsf{sym}}$.
- $c(\boldsymbol{W}) \le 2 - \epsilon$ with high probability when $\boldsymbol{W} \sim \mathsf{GOE}(n)$, for some $\epsilon > 0$.

Low-Degree Lower Bound

A two-part argument that such an algorithm (running in subexponential time) **cannot exist**.

Low-Degree Lower Bound

A two-part argument that such an algorithm (running in subexponential time) **cannot exist**.

Part 1: Reduction. If such existed, it could be used to distinguish two models of **bottom** eigenspaces of $\boldsymbol{W}$:

- $\boldsymbol{y}_1, \dots, \boldsymbol{y}_{(1-\delta)n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$.
- $\boldsymbol{y}_1, \dots, \boldsymbol{y}_{(1-\delta)n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n - \beta \boldsymbol{x}\boldsymbol{x}^\top)$ for $\boldsymbol{x}$ close to positive orthant.

Low-Degree Lower Bound

A two-part argument that such an algorithm (running in subexponential time) **cannot exist**.

Part 1: Reduction. If such existed, it could be used to distinguish two models of **bottom** eigenspaces of $\boldsymbol{W}$:

- $\boldsymbol{y}_1, \dots, \boldsymbol{y}_{(1-\delta)n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$.
- $\boldsymbol{y}_1, \dots, \boldsymbol{y}_{(1-\delta)n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n - \beta \boldsymbol{x}\boldsymbol{x}^\top)$ for $\boldsymbol{x}$ close to positive orthant.

Part 2: Hardness. Low-degree polynomials of the observations **cannot** distinguish these two models.

$\rightsquigarrow$ Assuming optimality of low-degree polynomial tests, no efficient certifier can beat $\mathsf{SDP}(\boldsymbol{W}) \approx \lambda_{\max}(\boldsymbol{W})$.

Takeaway Messages

Takeaway Messages

1. Small experiments can be misleading about the asymptotic behavior of semidefinite programs.

   There are strong finite-size effects up to current "laptop scale" with off-the-shelf solvers.

Takeaway Messages

1. Small experiments can be misleading about the asymptotic behavior of semidefinite programs.

   There are strong finite-size effects up to current “laptop scale” with off-the-shelf solvers.

2. Among certification algorithms (in particular convex relaxations), it is hard to beat the simple spectral bound $\lambda^+(\boldsymbol{W}) \le \lambda_{\mathsf{max}}(\boldsymbol{W})$.

   In this regard, non-negative PCA behaves like many other constrained PCA problems studied in previous literature.

Thank You!