ICML

International Conference
On Machine Learning

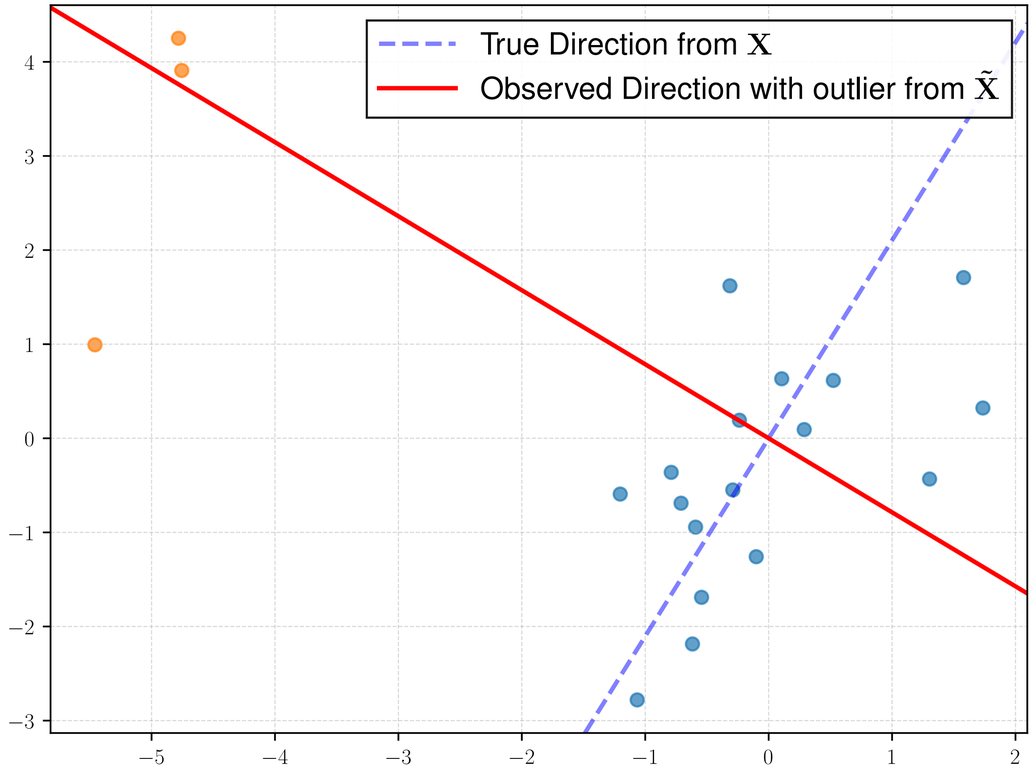
Mean-Shift PCA by Knockoff Mean

Removing noise is difficult, but adding noise is easy.

Mengda Li (CUHK-SZ), Zeng Li (SUSTech), Jianfeng Yao (CUHK-SZ)

Motivation: Why another “Robust PCA”?

- **K-Largest Principal Components** are too **sensitive** to noise!
- A small proportion of **mean-shift** outlier can shift the first principal component
- "Still" a **open problem in high dimensions**: no robust PCA algorithm can do it

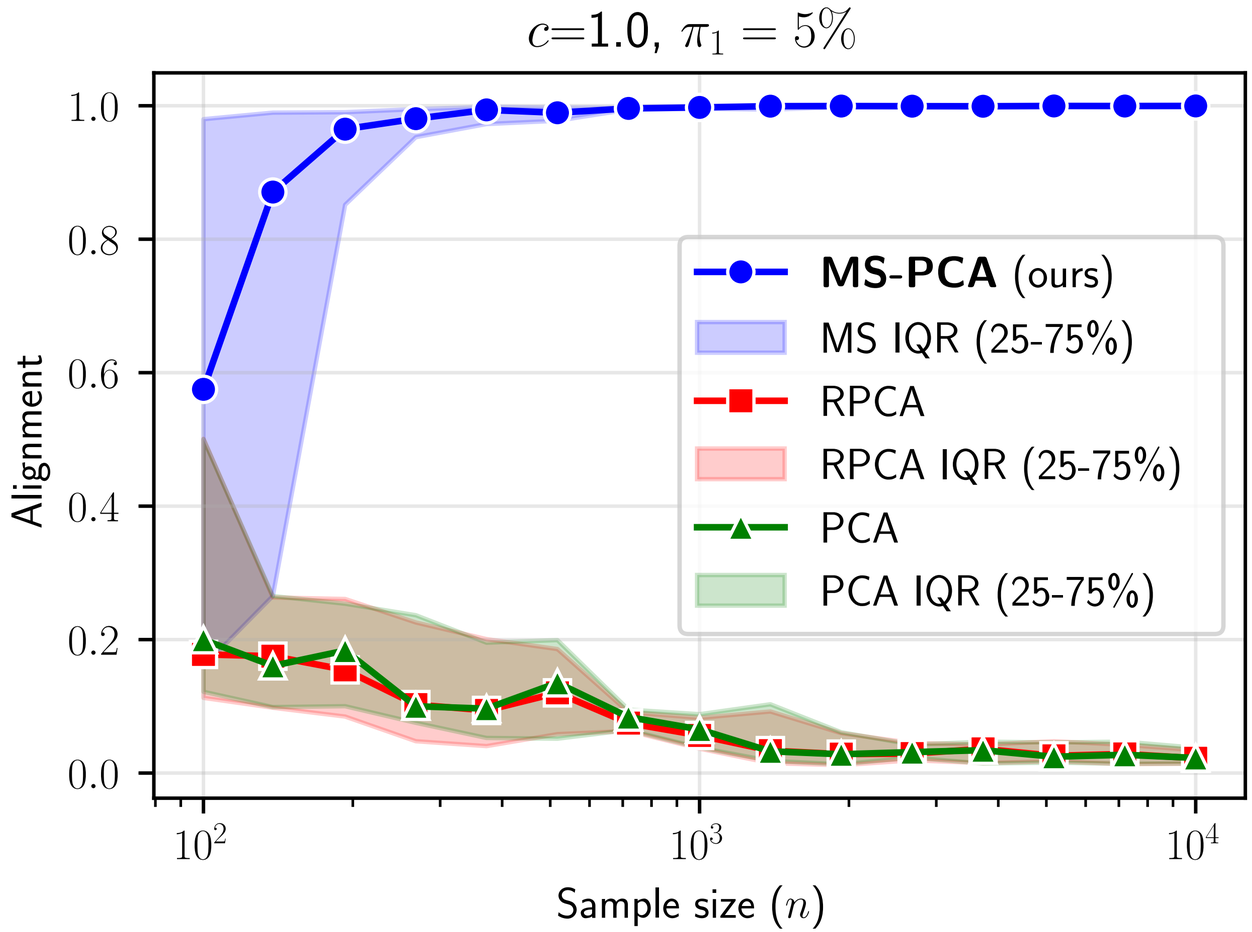


PCA on Gaussian Data with one Mean-Shift Cluster.

- Blue points : centered inlier component
- Orange points: outlier component with a mean shift
- Red line: the first principal component by PCA

Our contribution: an algorithm which can do it (in high dimensions)!

- A 2-fold PCA algorithm based on **Random Matrix Theory (RMT)**
- **Best Performance in Mean-Shift Mixture**
- **Lowest Time Complexity** among Robust PCA variants
- Slogan: **Adding Extra** mean-shift **Noise** to **Remove** mean-shift **Noise**



Failure of Robust PCA in high dimensions with only 5% noisy samples

- Blue : our algorithm based on RMT
- RPCA: SOTA Robust PCA algorithm (AAP, Cai et al., 2019)
- Green: cosine similarity of PC1 with true PC1

Performance VS Robust PCAs

Table 2. **Comparison with robust estimators and preprocessing baselines.** Alignment of the largest principal component under mean-shift Gaussian mixture contamination. Values are percentages, reported as mean ± standard deviation over 200 independent trials. Here $d = 900$, $n = 10^3$, and higher is better.

π_1	MS-PCA	RPCA-AAP	Tyler	Huber	ℓ_1-PCA	winsorized-PCA	center-PCA
5%	95.85 ± 12.53	8.40 ± 6.87	9.33 ± 7.46	9.72 ± 7.84	14.01 ± 10.73	9.26 ± 7.42	9.27 ± 7.37
10%	97.16 ± 6.96	8.75 ± 6.71	10.67 ± 8.00	10.95 ± 8.11	14.25 ± 10.46	10.63 ± 8.01	10.64 ± 8.10
15%	97.39 ± 7.03	7.47 ± 6.13	10.24 ± 8.12	10.51 ± 8.22	12.06 ± 9.33	10.22 ± 8.09	10.36 ± 8.07
20%	96.17 ± 11.82	7.74 ± 5.67	11.46 ± 8.63	11.53 ± 8.72	11.85 ± 8.81	11.47 ± 8.63	11.70 ± 8.70

Running Time VS Robust PCAs

Table 4. **Runtime scaling with dimension.** Runtime in milliseconds for estimating the largest PC with $c = d/n = 1$, reported as mean $\pm$ standard deviation.

Method	$d = 1000$	$d = 2000$	$d = 3000$	$d = 4000$	$d = 5000$	$d = 10000$
MS-PCA	15.9 ± 0.0109	49.2 ± 0.0795	102 ± 0.22	216 ± 1.1	237 ± 0.516	832 ± 1.26
RPCA-AAP	86 ± 0.102	369 ± 0.945	1050 ± 3.01	2490 ± 4.39	2750 ± 6.78	10500 ± 21.6

Table 3. **Runtime for estimating leading PCs.** Runtime in milliseconds for $d = 900, n = 1000$, reported as mean $\pm$ standard deviation.

	MS-PCA	RPCA-AAP	Tyler	Huber	ℓ_1-PCA
1 PC	14.2 ± 0.0137	79.0 ± 0.176	172 ± 0.159	3860 ± 4.37	8230 ± 294
9 PCs	19.2 ± 0.131	99.6 ± 0.210	174 ± 0.351	3860 ± 8.5	N/A

Data Model: Spikes from Covariance and mean

The original data consist of n i.i.d. samples arranged column-wise:

$$\mathbf{X}_n = \left[\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}\right]_{d\times n}, \qquad \mathbf{x}_{(i)} \in \mathbb{R}^d.$$

The contaminated matrix $\widetilde{\mathbf{X}}_n$ is obtained by adding a structured mean-shift matrix $\mathbf{A}_n$:

$$\widetilde{\mathbf{X}}_n = \mathbf{X}_n + \mathbf{A}_n, \qquad \mathbf{A}_n = \sum_{i=1}^{k} \mathbf{m}_{(i)}\boldsymbol{\gamma}_{(i)}^\top. \qquad (1)$$

Notation 2. *We can decompose* $\mathbf{A}_n$ *as:* $\mathbf{A}_n = \sqrt{n}\mathbf{V}_n\boldsymbol{\Theta}\mathbf{W}_n^\top$ *where* $\mathbf{V}_n = \left((\mathbf{v}_{(i)})_{i=1}^k \right) \in \mathbb{R}^{d\times k}$, $\boldsymbol{\Theta} = \operatorname{diag}(\theta_i)_{i=1}^k$ *with* $\theta_i = \sqrt{\pi_i}\|\mathbf{m}_{(i)}\|$, *and* $\mathbf{W}_n = \frac{1}{\sqrt{n\pi_i}}\left((\boldsymbol{\gamma}_{(i)})_{i=1}^k \right) \in \mathbb{R}^{n\times k}$.

$$\mathbf{X}_n = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}_n,$$

where $\mathbf{Z}_n$ consists of i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ columns.

$$\boldsymbol{\Sigma} = \mathbf{I}_d + \mathbf{P}$$

The matrix $\mathbf{P} \in \mathbb{R}^{d\times d}$ is symmetric with finite rank r, its non-zero eigenvalues are denoted by $(\ell_i)_{i=1}^r$. Fur

- **A** introduces mean-spiked eigenvalues, parametrized by θ
- **P** introduces covariance-spiked eigenvalues, parametrized by ℓ

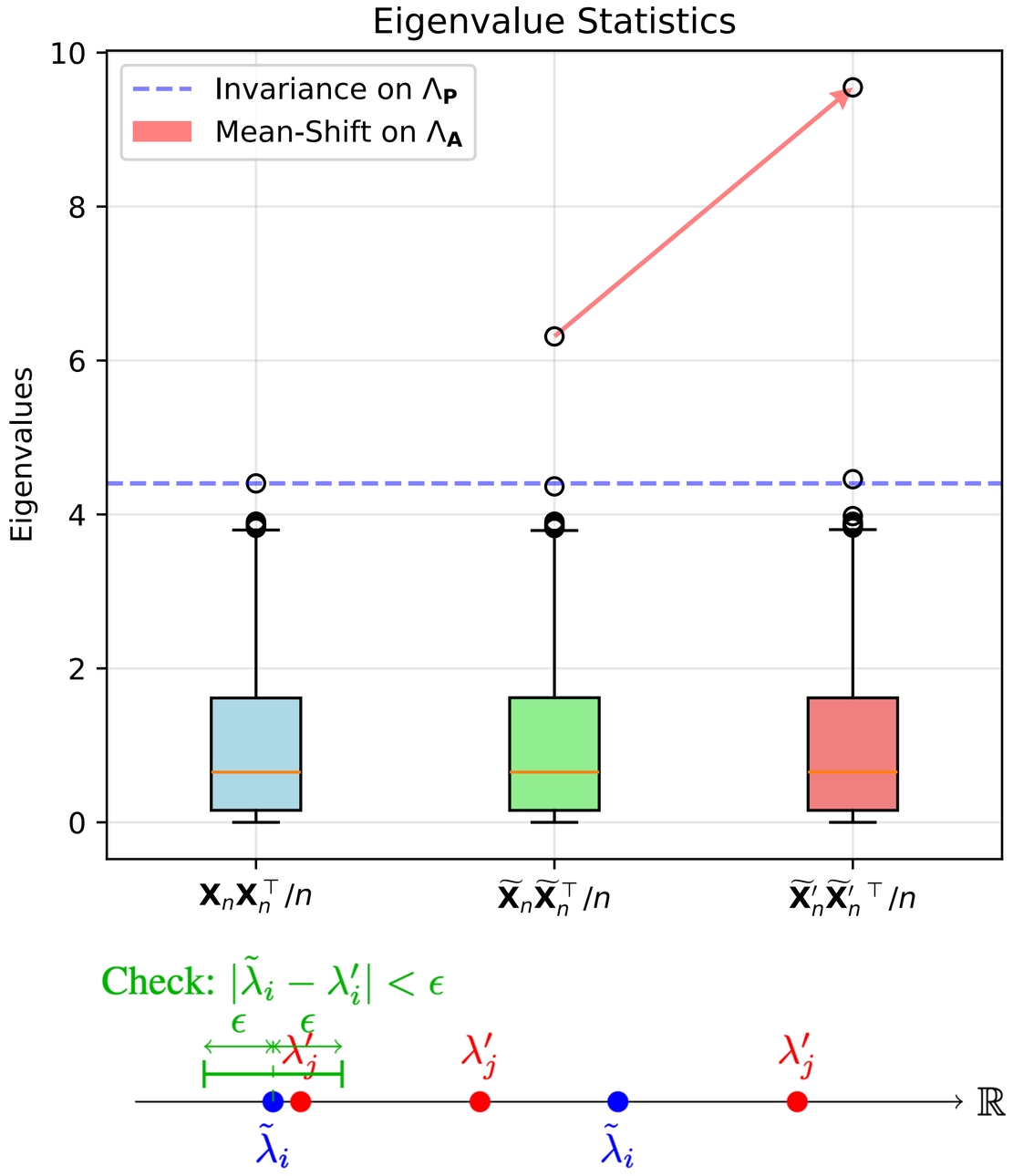
MS-PCA: find right spikes!

Algorithm 1 Mean-Shift PCA (MS-PCA)

1: **Input:** Contaminated data $\widetilde{\mathbf{X}}_n$, threshold constant C
2: **Initial PCA:** Compute the (largest spiked) eigenvalues $\{\tilde{\lambda}_i\}$ and eigenvectors $\{\tilde{\mathbf{u}}_i\}$ of $\widetilde{\mathbf{X}}_n\widetilde{\mathbf{X}}_n^\top/n$.
3: **Noise Injection:** Generate knock-off mean $\mathbf{m}'$ and its mixture weight $\boldsymbol{\gamma}'$. Form the mean-shift contamination matrix $\mathbf{A}'_n = \mathbf{m}'\boldsymbol{\gamma}'$ and the doublely perturbed data matrix $\widetilde{\mathbf{X}}'_n = \widetilde{\mathbf{X}}_n + \mathbf{A}'_n$.
4: **Second PCA:** Compute eigenvalues $\{\lambda'_i\}$ and eigenvectors $\{\mathbf{u}'_i\}$ of $\widetilde{\mathbf{X}}'_n\widetilde{\mathbf{X}}'^\top_n/n$.
5: **Invariance Check:** For each initially observed spiked eigenvalue $\tilde{\lambda}_i$, check if there exists a corresponding λ'_j such that

$$|\tilde{\lambda}_i - \lambda'_j| < \epsilon, \quad \text{where } \epsilon = Cn^{-1/2}. \tag{4}$$

Remove non-stable eigenspaces of $\widetilde{\mathbf{X}}_n\widetilde{\mathbf{X}}_n^\top/n$ and output the stable eigenspaces associated with λ_i satisfying equation 4.



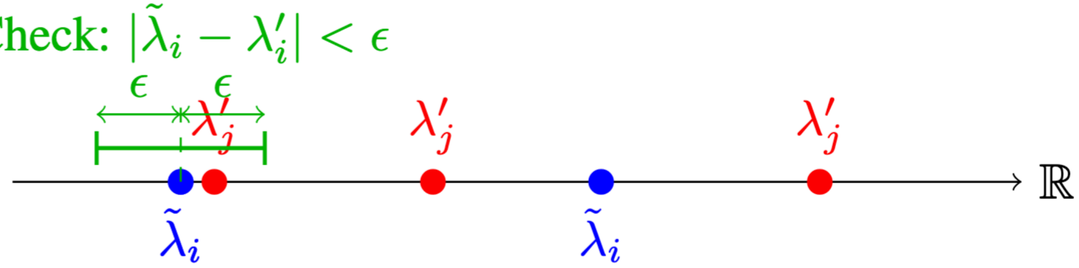
Only mean spikes MOVE after knockoff mean-shift!

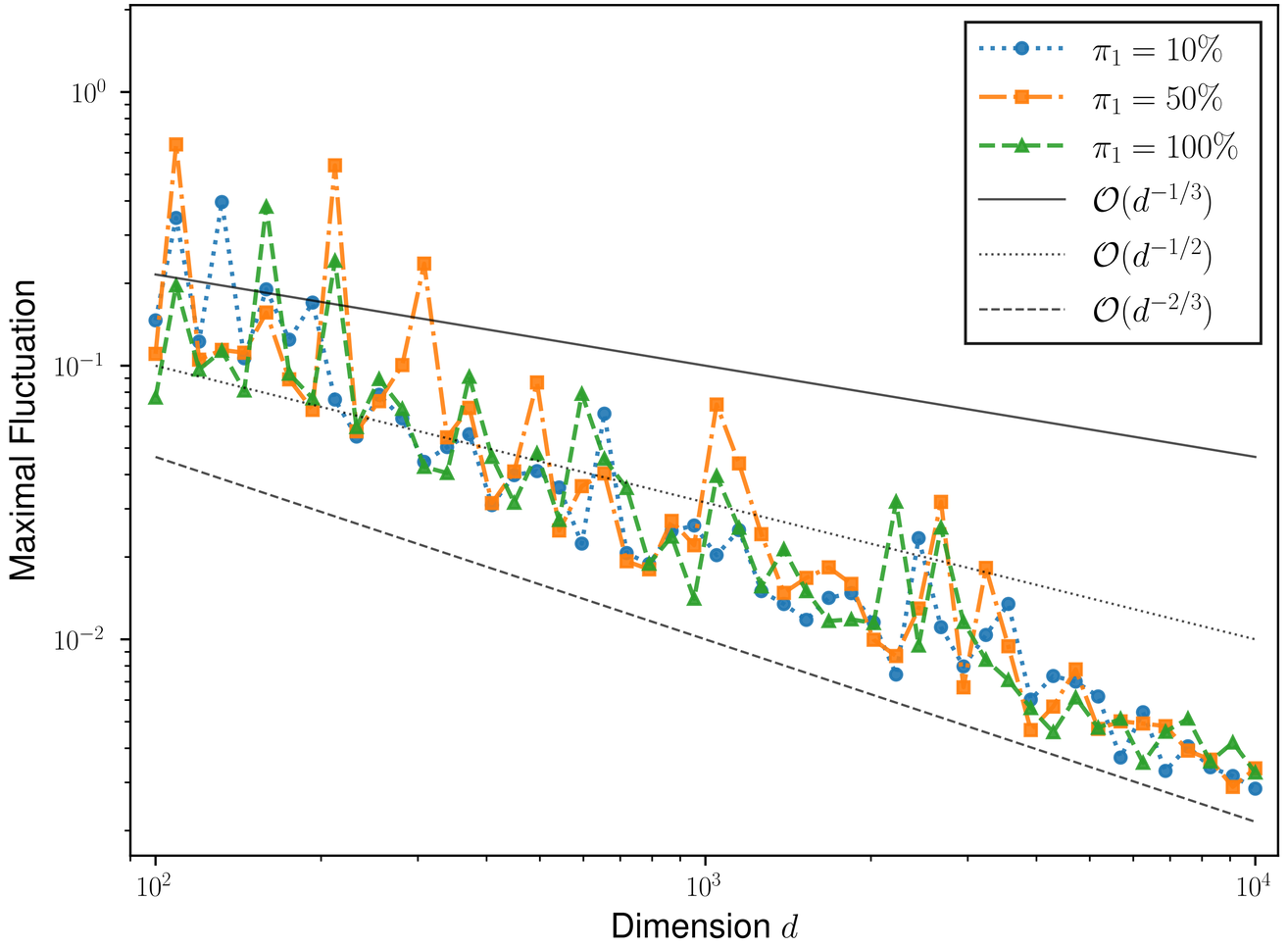
- Stable eigenvalues (covariance spikes)

  fluctuate with order $O(d^{-\frac{1}{2}})$

- After Knockoff mean-shift, mean spikes

  move >> $O(d^{-\frac{1}{2}})$





Maximal Fluctuation of stable Eigenvalues

Why we can do it? Spike independence! (Covariance vs Means)

The original data consist of n i.i.d. samples arranged column-wise:

$$\mathbf{X}_n = \left[\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}\right]_{d\times n}, \qquad \mathbf{x}_{(i)} \in \mathbb{R}^d.$$

The contaminated matrix $\widetilde{\mathbf{X}}_n$ is obtained by adding a structured mean-shift matrix $\mathbf{A}_n$:

$$\widetilde{\mathbf{X}}_n = \mathbf{X}_n + \mathbf{A}_n, \qquad \mathbf{A}_n = \sum_{i=1}^{k} \mathbf{m}_{(i)}\boldsymbol{\gamma}_{(i)}^\top. \qquad (1)$$

$$\mathbf{X}_n = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}_n,$$

where $\mathbf{Z}_n$ consists of i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ columns.

$$\boldsymbol{\Sigma} = \mathbf{I}_d + \mathbf{P}$$

The matrix $\mathbf{P} \in \mathbb{R}^{d\times d}$ is symmetric with finite rank r, its non-zero eigenvalues are denoted by $(\ell_i)_{i=1}^r$. Fur

Theorem 3.5 (Largest Eigenvalues). *Under Assumptions 3.1, 3.3 and 3.4, supposing all ℓ_i's are positive, let $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_{r+k}$ be the $r+k$ largest eigenvalues of the sample covariance matrix $\widetilde{\mathbf{X}}_n\widetilde{\mathbf{X}}_n^\top/n$, and let Λ be the set of the asymptotic largest spiked eigenvalues defined as:*

$$\Lambda_{\mathbf{P}} := \left\{ 1 + \ell_i + c\frac{1+\ell_i}{\ell_i} \;\middle|\; \ell_i > \sqrt{c},\ \forall i \in [r] \right\}$$

$$\Lambda_{\mathbf{A}} := \left\{ 1 + \theta_j^2 + c\frac{1+\theta_j^2}{\theta_j^2} \;\middle|\; \theta_j^2 > \sqrt{c}, \forall j \in [k] \right\}$$

$$\Lambda := \Lambda_{\mathbf{P}} \bigcup \Lambda_{\mathbf{A}}$$

Then, denoting λ_i^* as the i-th largest element in Λ,

$$\begin{aligned} \tilde{\lambda}_i &\xrightarrow[a.s.]{n\to\infty} \lambda_i^*, && \text{if } i \leq \operatorname{card}(\Lambda), \\ \tilde{\lambda}_i &\xrightarrow[a.s.]{n\to\infty} (1+\sqrt{c})^2, && \text{else} \end{aligned} \qquad (7)$$

- Eigenvalues **Order shifted** after mean-shift
- Covariance spike **value no change**

Why we can do it? Spike independence! (Covariance vs Means)

The original data consist of n i.i.d. samples arranged column-wise:

$$\mathbf{X}_n = \left[\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}\right]_{d\times n}, \qquad \mathbf{x}_{(i)} \in \mathbb{R}^d.$$

The contaminated matrix $\widetilde{\mathbf{X}}_n$ is obtained by adding a structured mean-shift matrix $\mathbf{A}_n$:

$$\widetilde{\mathbf{X}}_n = \mathbf{X}_n + \mathbf{A}_n, \qquad \mathbf{A}_n = \sum_{i=1}^{k} \mathbf{m}_{(i)}\boldsymbol{\gamma}_{(i)}^{\top}. \qquad (1)$$

$$\mathbf{X}_n = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}_n,$$

where $\mathbf{Z}_n$ consists of i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ columns.

$$\boldsymbol{\Sigma} = \mathbf{I}_d + \mathbf{P}$$

The matrix $\mathbf{P} \in \mathbb{R}^{d\times d}$ is symmetric with finite rank r, its non-zero eigenvalues are denoted by $(\ell_i)_{i=1}^{r}$. Fur

Noise Injection: Generate knock-off mean $\mathbf{m}'$ and its mixture weight $\boldsymbol{\gamma}'$. Form the mean-shift contamination matrix $\mathbf{A}'_n = \mathbf{m}'\boldsymbol{\gamma}'$ and the doublely perturbed data matrix $\widetilde{\mathbf{X}}'_n = \widetilde{\mathbf{X}}_n + \mathbf{A}'_n$.

Corollary 3.7. *Under the framework of Theorem 3.5, let Λ', $\Lambda'_{\mathbf{A}}$ and $\Lambda'_{\mathbf{P}}$ be the set of the asymptotic largest spiked eigenvalues of the sample covariance matrix $\widetilde{\mathbf{X}}'_n\widetilde{\mathbf{X}}'^{\top}_n/n$ after additional mean-shift contamination in Algorithm 1. Then, for sufficiently strong perturbation $\mathbf{A}'_n$, we have:*

$$\Lambda'_{\mathbf{P}} = \Lambda_{\mathbf{P}}, \qquad \Lambda'_{\mathbf{A}} \neq \Lambda_{\mathbf{A}}$$

We change again the mean spikes

by **adding knockoff Mean-Shift!**

Why we can do it? Spike independence! (Covariance vs Means)

The original data consist of n i.i.d. samples arranged column-wise:

$$\mathbf{X}_n = \left[\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}\right]_{d\times n}, \qquad \mathbf{x}_{(i)} \in \mathbb{R}^d.$$

The contaminated matrix $\widetilde{\mathbf{X}}_n$ is obtained by adding a structured mean-shift matrix $\mathbf{A}_n$:

$$\widetilde{\mathbf{X}}_n = \mathbf{X}_n + \mathbf{A}_n, \qquad \mathbf{A}_n = \sum_{i=1}^{k} \mathbf{m}_{(i)}\boldsymbol{\gamma}_{(i)}^\top. \qquad (1)$$

$$\mathbf{X}_n = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}_n,$$

where $\mathbf{Z}_n$ consists of i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ columns.

$$\boldsymbol{\Sigma} = \mathbf{I}_d + \mathbf{P}$$

The matrix $\mathbf{P} \in \mathbb{R}^{d\times d}$ is symmetric with finite rank r, its non-zero eigenvalues are denoted by $(\ell_i)_{i=1}^r$. Fur

Theorem 3.11 (Eigenspace Invariance)**.** *Under Assumptions 3.1 and 3.10, let* $\mathbf{u}$ *be an eigenvector of the uncontaminated sample covariance matrix* $\mathbf{X}_n\mathbf{X}_n^\top/n$ *associated with eigenvalue* λ*, i.e.,* $\frac{1}{n}\mathbf{X}_n\mathbf{X}_n^\top\mathbf{u} = \lambda\mathbf{u}$*, then*

$$\frac{1}{n}\widetilde{\mathbf{X}}_n\widetilde{\mathbf{X}}_n^\top\mathbf{u} = \lambda\mathbf{u} + \mathbf{r}_n, \quad \textit{with} \quad \|\mathbf{r}_n\|_2 = \mathcal{O}_p(n^{-1/2})$$

This asymptotic equivalence is denoted by: $\frac{1}{n}\widetilde{\mathbf{X}}_n\widetilde{\mathbf{X}}_n^\top\mathbf{u} \sim \lambda\mathbf{u}$*. Roughly speaking* $\mathbf{u}_\lambda \sim \tilde{\mathbf{u}}_\lambda$ *for* $\lambda \in Spec(\mathbf{X}_n\mathbf{X}_n^\top/n)$.

Eigenvectors are asymptotically unchanged

so just **find the stable eigenvalues!**