Fast and Provable Algorithms for Sparse PCA with Improved Sample Complexity

Jian-Feng Cai, Zhuozhi Xian, Jiaxi Ying

Department of Mathematics
Hong Kong University of Science and Technology

Forty-Second International Conference on Machine Learning

Spiked covariance model

- The spiked covariance model was introduced by (Johnstone, 2001).
- In the spiked covariance model, it observes n noisy samples

$$\boldsymbol{x}_i = \sqrt{\lambda} g_i \boldsymbol{v} + \boldsymbol{\xi}_i, \quad i = 1, \ldots, n, \tag{1}$$

  where

  - $\boldsymbol{v} \in \mathbb{R}^p$ is a k-sparse unit unknown vector,
  - $g_i \in \mathbb{R}$ are coefficients independently sampled from $\mathcal{N}(0,1)$*,
  - $\boldsymbol{\xi}_i \in \mathbb{R}^p$ are noisy vectors independently drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$†,
  - g_i and $\boldsymbol{\xi}_i$ are mutually independent,
  - $\lambda > 0$ is the signal strength.

- We focus on how many samples are sufficient to estimate $\boldsymbol{v}$ from n nosiy samples of (1) up to a constant error (in polynomial time).

*standard Gaussain distribution with mean 0 and variance 1

†multivariate Gaussain distribution with mean $\mathbf{0}$ and variance $\boldsymbol{I}$

Sparse PCA

- For n samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ drawn from (1), $\boldsymbol{x}_i$ are zero-mean and
  - empirical covariance matrix: $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$,
  - population covariance matrix: $\boldsymbol{\Sigma} = \mathbb{E}\big[\hat{\boldsymbol{\Sigma}}\big] = \lambda \boldsymbol{v}\boldsymbol{v}^T + \boldsymbol{I}$.
- To estimate $\boldsymbol{v}$, we consider the sparse PCA (SPCA) problem:

$$\max_{\boldsymbol{w}} \boldsymbol{w}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{w}, \quad \text{subject to } \|\boldsymbol{w}\|_2 = 1, \ \|\boldsymbol{w}\|_0 \leq k, \tag{2}$$

  where the solution is an estimator of $\boldsymbol{v}$.
- The SPCA problem (2) is non-convex and NP-hard.

Sample complexity

- Information-theoretic sample complexity is $n = \Omega(k \log p)^*$.
- Existing polynomial-time algorithms require at least $O(k^2)$ samples for successful recovery (Deshpande and Montanari, 2016), highlighting a significant gap in sample efficiency.
- Reductions from the planted-clique conjecture imply that, without further assumptions, no polynomial-time algorithm can attain the information-theoretic sample complexity[†].
- Question: *Can we design a polynomial-time algorithm to bridge the gap under some assumption of the model* (1)*?*

[*] Vu and Lei, 2013; Berthet and Rigollet, 2013

[†] Berthet and Rigollet, 2013; Krauthgamer et al., 2015; Wang et al., 2016; Gao et al., 2017; Brennan et al., 2018

Existing polynomial-time algorithms

- Diagonal thresholding (Johnstone and Lu, 2009): Find the top k elements of the diagonal of $\hat{\boldsymbol{\Sigma}}$ and compute the largest eigenvector of the corresponding $k \times k$ submatrix of $\hat{\boldsymbol{\Sigma}}$.
  - Sample complexity: $\Omega(k^2 \log p)$ (Amini and Wainwright, 2009)
  - Computational cost: $O(np + nk^2)$
- Covariance thresholding (Deshpande and Montanari, 2016): Soft-thresholding to $\hat{\boldsymbol{\Sigma}}$ and find the top k elements of the largest eigenvector of the thresholded $\hat{\boldsymbol{\Sigma}}$.
  - Sample complexity: $\Omega(k^2)$ (Deshpande and Montanari, 2016)
  - Computational cost: $O(np^2 + p^3)$
- Semi-definite programming relaxation (d'Aspremont et al., 2004): Relax the SPCA problem as a convex problem by using a new variable $\boldsymbol{W} = \boldsymbol{w}\boldsymbol{w}^T$ and modifying the ℓ_0-constriant.
  - Sample complexity: $\Omega(k^2 \log p)$ (Berthet and Rigollet, 2013)
  - Computational cost: $O(np^2 + p^4 \log p)$ (d'Aspremont et al., 2004)

Proposed thresholding algorithm

- Diagonal thresholding (Johnstone and Lu, 2009)
  - Statistical gap: $g_{\mathrm{d}} := \min_{j\in\mathcal{S}} \left|[\mathbb{E}[\hat{\boldsymbol{\Sigma}}]]_{jj}\right| - \max_{j\in\mathcal{S}^c} \left|[\mathbb{E}[\hat{\boldsymbol{\Sigma}}]]_{jj}\right| = \lambda \cdot \min_{j\in\mathcal{S}} \boldsymbol{v}_j^2$.*
  - A larger g_{d} permits the smaller required number of samples.

- For a larger statistical gap, we propose a novel thresholding algorithm‡:

1. Compute $\{\hat{\boldsymbol{\Sigma}}_{j,j}\}_{j=1}^{n}$ and set $j_0 = \arg\max_{1\le j\le n} \hat{\boldsymbol{\Sigma}}_{j,j}$;
2. Compute $\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_{j_0}$ and set $\hat{\mathcal{S}}$ as the indices of the top k elements of $\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_{j_0}$ in absolute value;†
3. Compute $[\hat{\boldsymbol{\Sigma}}]_{\hat{\mathcal{S}}}$, set $[\boldsymbol{v}^0]_{\hat{\mathcal{S}}}$ as the unit leading eigenvector of $[\hat{\boldsymbol{\Sigma}}]_{\hat{\mathcal{S}}}$ and set $[\boldsymbol{v}^0]_{\hat{\mathcal{S}}^c} = 0$;
4. Output $\boldsymbol{v}^0$ as the estimator of $\boldsymbol{v}$.

  - Statistical gap: $g := \min_{j\in\mathcal{S}} \left|\mathbb{E}[\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_{j_0}]_j\right| - \max_{j\in\mathcal{S}^c} \left|\mathbb{E}[\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_{j_0}]_j\right| \ge \lambda|\boldsymbol{v}_{j_0}| \cdot \min_{j\in\mathcal{S}} |\boldsymbol{v}_j|$.
  - $g \ge g_{\mathrm{d}}$
  - Computational cost: $O(np + nk^2)$

* $\mathcal{S}$: the support of $\boldsymbol{v}$

† $\boldsymbol{e}_j$: the j-th standard basis

‡ Inspired by (Wu and Rebeschini, 2021; Cai et al., 2023)

Proposed two-stage algorithm

- To enhance the estimation performance, we propose a two-stage algorithm:
  - Initialization stage: proposed thresholding algorithm
  - Refinement stage: truncated power method (Yuan and Zhang, 2013)

Proposed two-stage algorithm for enhancing estimation performance

Input: Samples $\{\boldsymbol{x}_i\}_{i=1}^n$, the sparsity k, parameter k'.

// Initialization stage:

Compute an initial estimate $\boldsymbol{v}^0$ by proposed thresholidng algorithm;

// Refinement stage:

for $t = 1, 2, \ldots$ **do**

$\widetilde{\boldsymbol{v}}^t = T_{k'}(\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^T\boldsymbol{v}^{t-1})\boldsymbol{x}_i)$;

$\boldsymbol{v}^t = \widetilde{\boldsymbol{v}}^t / \|\widetilde{\boldsymbol{v}}^t\|_2$;

end

Output: $\boldsymbol{v}^t$

- Computational cost: $O(np + nk^2)$ for first stage and $O(np)$ for each iteration in second stage

Theoretical results

- Noisy samples: $\boldsymbol{x}_i = \sqrt{\lambda} g_i \boldsymbol{v} + \boldsymbol{\xi}_i,\ i = 1, \ldots, n$ from (1).
- Error: $\mathrm{dist}(\boldsymbol{v}, \hat{\boldsymbol{v}}) := \min\left\{\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2, \|\boldsymbol{v} + \hat{\boldsymbol{v}}\|_2\right\}$.

Theorem (Proposed thresholding algorithm)

For any $\gamma \in (0, 1]$, there exists universal constants $C_1, C_2 > 0$ such that if $\lambda \geq C_1 \|\boldsymbol{v}\|_\infty^{-1}$ and $n \geq C_2 \gamma^{-2} k \log p$, with probability exceeding $1 - 5p^{-1}$, the output $\boldsymbol{v}^0$ satisfies $\mathrm{dist}(\boldsymbol{v}, \boldsymbol{v}^0) \leq \gamma$.

Theorem (Proposed two-stage algorithm)

There exist universal constants $C_3, C_4, C_5 > 0$ such that if $\lambda \geq C_3 \|\boldsymbol{v}\|_\infty^{-1}$ and $n \geq C_4 k \log p$, with probability exceeding $1 - 5p^{-1}$, the output $\boldsymbol{v}^t$ with parameter $k' = C_5 k$ and an initial estimate $\boldsymbol{v}^0$ generated by proposed thresholding algorithm satisfies

$$\mathrm{dist}(\boldsymbol{v}^t, \boldsymbol{v}) \leq \underbrace{d^t \cdot \mathrm{dist}(\boldsymbol{v}, \boldsymbol{v}^0)}_{\text{Optimization error}} + \underbrace{d' \sqrt{k \log p / n}}_{\text{Statistical error}}, \tag{3}$$

where $0 < d < 1$ and $d' > 0$ are constants.

Experiment results

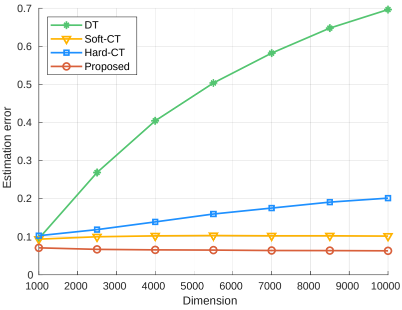

Figure 1: Comparisons of estimation error

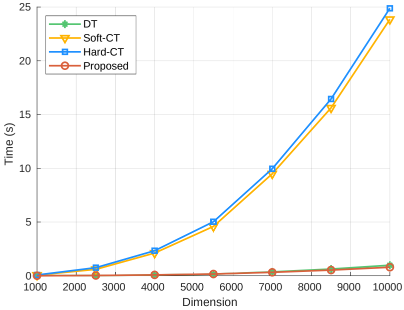

Figure 2: Comparisons of computational time

- Our proposed algorithm demonstrates both estimation accuracy and computational efficiency.

[*] DT: diagonal thresholding (Johnstone and Lu, 2009)
[†] Soft-CT: covariance thresholding in (Deshpande and Montanari, 2016)
[‡] Hard-CT: covariance thresholding in (Krauthgamer et al., 2015)

References I

Arash A. Amini and Martin J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5B):2877–2921, 2009.

Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780 – 1815, 2013.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Conference On Learning Theory*, pages 48–166. PMLR, 2018.

Jian-Feng Cai, Jingyang Li, and Juntao You. Provable sample-efficient sparse phase retrieval initialized by truncated power method. *Inverse Problems*, 39(7):075008, 2023.

Alexandre d'Aspremont, Laurent Ghaoui, Michael Jordan, and Gert Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *Advances in neural information processing systems*, 17, 2004.

Yash Deshpande and Andrea Montanari. Sparse PCA via covariance thresholding. *Journal of Machine Learning Research*, 17(141): 1–41, 2016.

Chao Gao, Zongming Ma, and Harrison H. Zhou. Sparse CCA: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2): 295–327, 2001.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

Robert Krauthgamer, Boaz Nadler, and Dan Vilenchik. Do semidefinite relaxations solve sparse PCA up to the information limit? *The Annals of Statistics*, 43(3):1300–1322, 2015.

Vincent Q. Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6), 2013.

Tengyao Wang, Quentin Berthet, and Richard J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896–1930, 2016.

Fan Wu and Patrick Rebeschini. Hadamard wirtinger flow for sparse phase retrieval. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 982–990, 2021.

Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(4), 2013.