Heteroskedastic Sparse PCA in High Dimensions

Zhao Ren

Department of Statistics
University of Pittsburgh

Joint work with R. Kang and P. Zhang

Dec 18, 2023 CIRM

Outline

- **Introduction: PCA with Heteroskedastic Noise**
- **Algorithm**
  - Power Method and Main Ideas
  - Algorithm Part I: Power Method with Adaptive Thresholding
  - Algorithm Part II: Iterative Update of Diagonal Values
- **Optimality**
  - Theoretical Guarantee
  - A Minimax Lower Bound
- **Numerical Studies**
- **Extensions and Applications**

Introduction: PCA

Data:

$$X_1, \ldots, X_n \in \mathbb{R}^p \text{ with Covariance } \Sigma$$

Objective function:

$$\text{(One PC)} \quad \arg max_{u \in \mathbb{R}^p} \operatorname{Var}(u^\top X) \quad s.t. \ \|u\|_2 = 1$$

$$\text{(Multiple PCs)} \quad \arg max_{U \in \mathbb{R}^{p \times r}} \operatorname{Tr}(U^\top \Sigma U) \quad s.t. \ U^\top U = I_r$$

PCA procedures: Do eigen-decomposition on sample covariance $\widehat{\Sigma}$.

Introduction: High Dimensional PCA

Classical PCA does not work in high dimensions: $p/n \not\to 0$.

Paul (2007), D'Aspremont et al. (2007), Johnstone and Lu (2009)

- Inconsistency
- Interpretation

Introduction: High Dimensional PCA

Classical PCA does not work in high dimensions: $p/n \not\to 0$.

Paul (2007), D'Aspremont et al. (2007), Johnstone and Lu (2009)

- Inconsistency
- Interpretation

A common approach: Sparse PCA

- Zou et al. (2006), Witten et al. (2009): Lasso type of regularization
- D'Aspremont et al. (2007), Vu and Lei (2013): relaxed convex optimization such as fantope
- Yuan and Zhang (2013), Ma (2013): truncated/thresholded power method
- Cai et al. (2013), Cai et al. (2015), Gataric et al. (2020), ...

Spiked Covariance Model for PCA Analysis

Johnstone (2001) proposed Spiked Covariance Model, where Σ has a "spiked" eigen-structure: $\Sigma = U\Lambda U^\top$,

$$\Lambda = diag\{\lambda_1, \ldots, \lambda_r, 1, \ldots, 1\} \in \mathbb{R}^{p\times p} \tag{1}$$

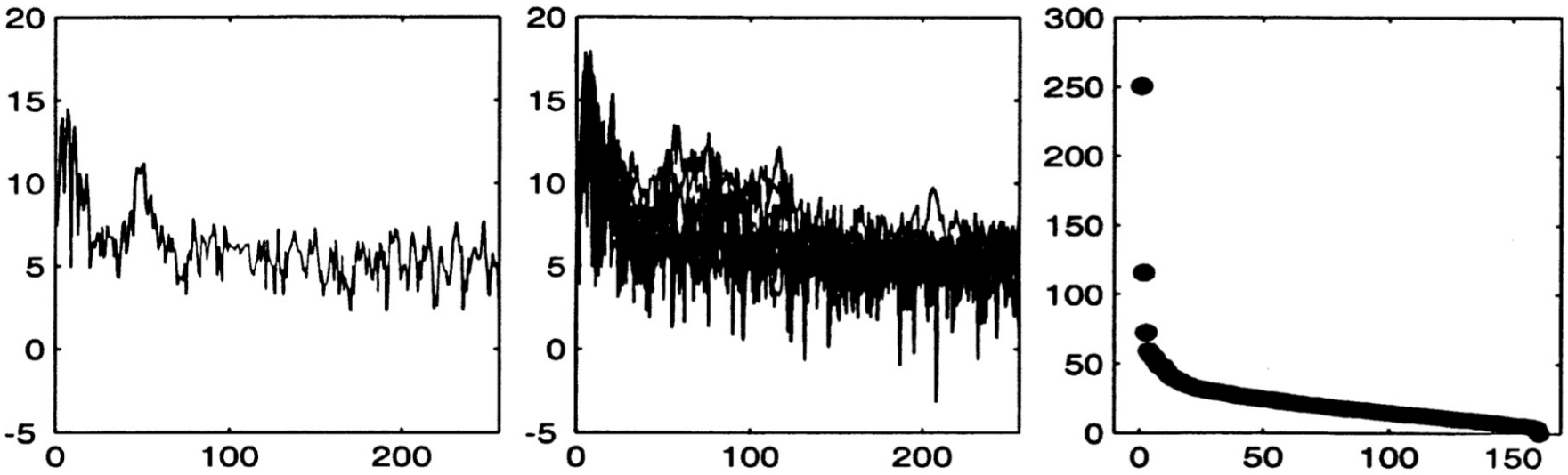


FIG. 1. (a) *a single instance of a periodogram from the phoneme dataset*; (b) *ten instances, to indicate variability*; (c) *screeplot of eigenvalues in phoneme example.*

Figure from Johnstone (2001). $n = 162$, $p = 256$ for this phoneme data set.

Spiked Covariance Model for PCA Analysis

- Model (1) can be represented as

$$\Sigma = \sum_{j=1}^{r} \lambda_j u_j u_j^\top + \sigma^2 I_p$$

Spiked Covariance Model for PCA Analysis

- Model (1) can be represented as

$$\Sigma = \sum_{j=1}^{r} \lambda_j u_j u_j^\top + \sigma^2 I_p$$

- It is equivalent to assume that

$$X_i = S_i + \varepsilon_i, \quad S_i \perp\!\!\!\perp \varepsilon_i$$

$$\mathrm{Cov}(S_i) = \sum_{j=1}^{r} \lambda_j u_j u_j^\top = U \Lambda U^\top, \quad \mathrm{Cov}(\varepsilon_i) = \sigma^2 I_p$$

where $U = (u_1, \ldots, u_r) \in \mathbb{R}^{p \times r}$, $\Lambda = diag\{\lambda_1, \ldots, \lambda_r\} \in \mathbb{R}^{r \times r}$.

Generalized Spiked Covariance Model

- Spiked covariance model with heteroskedastic noise (Bai and Yao (2012), Yao et al. (2015)):

$$X_i = S_i + \varepsilon_i, \quad S_i \perp\!\!\!\perp \varepsilon_i$$
$$\mathrm{Cov}(S_i) = U\Lambda U^\top, \quad \mathrm{Cov}(\varepsilon_i) = diag\{\sigma_1^2, \ldots, \sigma_p^2\}$$

Generalized Spiked Covariance Model

- Spiked covariance model with heteroskedastic noise (Bai and Yao (2012), Yao et al. (2015)):

$$X_i = S_i + \varepsilon_i, \quad S_i \perp\!\!\!\perp \varepsilon_i$$
$$\mathrm{Cov}(S_i) = U\Lambda U^\top, \quad \mathrm{Cov}(\varepsilon_i) = \mathit{diag}\{\sigma_1^2, \ldots, \sigma_p^2\}$$

- Examples include:
  - biological sequencing data: Marx (2013), Cao et al. (2020)
  - photon imaging data: Salmon et al. (2014)
  - network analysis: Sun et al. (2012)

Identifiability

- An example: $U = (1, 0, \ldots, 0)^\top$ or $U = (0, 1, \ldots, 0)^\top$?,

$$\Sigma = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 + \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 + \lambda_1 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}$$

- Incoherence:

  (Candès and Recht (2009), Candès and Tao (2010))

$$I(U) = (p/r) \max_{i \in [p]} \left\| e_i^\top U \right\|_2^2 \in [1, p/r]$$

Heteroskedastic PCA in high dimensions

- Heteroskedastic PCA in fixed/low dimensions ($n > p$) has been recently explored by Zhang et al. (2022) and Yan et al. (2023)

Heteroskedastic PCA in high dimensions

- Heteroskedastic PCA in fixed/low dimensions ($n > p$) has been recently explored by Zhang et al. (2022) and Yan et al. (2023)
- Our work focuses on the high dimensional setting ($p > n$), assuming a (row or joint) sparsity of the eigenvector U.

$$\|U\|_0 := |\{1 \le j \le p : U_{j\star} \neq 0\}| \le s$$

- Another type of heteroskedastic PCA: Hong et al. (2016, 2018, 2018)

Methodology

Model and Notations Set-up

Generalized Spiked Covariance Model:

$$X_i = S_i + \varepsilon_i, \quad S_i \perp\!\!\!\perp \varepsilon_i$$

$$\mathrm{Cov}(S_i) = U\Lambda U^\top, \quad \mathrm{Cov}(\varepsilon_i) = diag\{\sigma_1^2, \ldots, \sigma_p^2\}$$

$$U = (u_1, \ldots, u_r), \quad \|U\|_0 \le s$$

Model and Notations Set-up

Generalized Spiked Covariance Model:

$$X_i = S_i + \varepsilon_i, \quad S_i \perp\!\!\!\perp \varepsilon_i$$
$$\mathrm{Cov}(S_i) = U\Lambda U^\top, \quad \mathrm{Cov}(\varepsilon_i) = diag\{\sigma_1^2, \ldots, \sigma_p^2\}$$
$$U = (u_1, \ldots, u_r), \quad \|U\|_0 \le s$$

Goal: To estimate (the principal subspace spanned by) the leading eigenvectors $u_1, \ldots, u_r$.

Model and Notations Set-up

Generalized Spiked Covariance Model:

$$X_i = S_i + \varepsilon_i, \quad S_i \perp\!\!\!\perp \varepsilon_i$$
$$\mathrm{Cov}(S_i) = U\Lambda U^\top, \quad \mathrm{Cov}(\varepsilon_i) = diag\{\sigma_1^2, \ldots, \sigma_p^2\}$$
$$U = (u_1, \ldots, u_r), \quad \|U\|_0 \le s$$

Goal: To estimate (the principal subspace spanned by) the leading eigenvectors $u_1, \ldots, u_r$.

Loss function: $\|\sin\Theta(\widehat{U}, U)\| = \|\widehat{U}\widehat{U}^\top - UU^\top\|$

PCA: The Power Method

- Let $A = U\Lambda U^\top$, $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$, $U = (u_1, \ldots, u_p)$.

PCA: The Power Method

- Let $A = U\Lambda U^\top$, $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$, $U = (u_1, \ldots, u_p)$.
- To estimate u_1, the power method does

for $t = 1, 2, \ldots$

$$\bar{u}^{(t)} = Au^{(t-1)} \qquad (\textit{Multiplication})$$
$$u^{(t)} = \bar{u}^{(t)} / \left\|\bar{u}^{(t)}\right\|_2 \qquad (\textit{Normalization})$$

PCA: The Power Method

- Let $A = U\Lambda U^\top$, $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$, $U = (u_1, \ldots, u_p)$.
- To estimate u_1, the power method does

$$\textbf{for } t = 1, 2, \ldots$$

$$\bar{u}^{(t)} = Au^{(t-1)} \qquad (\textit{Multiplication})$$

$$u^{(t)} = \bar{u}^{(t)} / \left\|\bar{u}^{(t)}\right\|_2 \qquad (\textit{Normalization})$$

- Intuition: Assume $u^{(t-1)} = \sum_{j=1}^{p} a_j u_j$.
  Then $Au^{(t-1)} = \sum_{j=1}^{p} (\lambda_j a_j) u_j$.

The Power Method with Thresholding

What we learned from Ma (2013) for (homoscedastic) Sparse PCA?

Notation:

$$\Sigma = \mathrm{Cov}(X_i), \Sigma_0 = \mathrm{Cov}(S_i) = U\Lambda U^\top.$$

$$\Sigma = \Sigma_0 + diag\{\sigma^2, \ldots, \sigma^2\}.$$

Let $\widehat{\Sigma}$ be the sample covariance of $\{X_i\}_{i=1}^n$.

Sparse PCA: The Power Method with Thresholding

- Multiplication: $\bar{U}^{(t)} = \widehat{\Sigma}\widehat{U}^{(t-1)}$;
- Thresholding: $\widetilde{U}^{(t)} = H_\gamma(\bar{U}^{(t)})$;
- QR normalization: $\widehat{U}^{(t)}R^{(t)} = \widetilde{U}^{(t)}$.

Sparse PCA: The Power Method with Thresholding

- Multiplication: $\bar{U}^{(t)} = \widehat{\Sigma}\widehat{U}^{(t-1)}$;
- Thresholding: $\widetilde{U}^{(t)} = H_\gamma(\bar{U}^{(t)})$;
- QR normalization: $\widehat{U}^{(t)}R^{(t)} = \widetilde{U}^{(t)}$.
- The thresholding function: $H_\gamma(U) = (h_{\gamma_k}(U_{jk}))_{1\le j\le p, 1\le k\le r}$

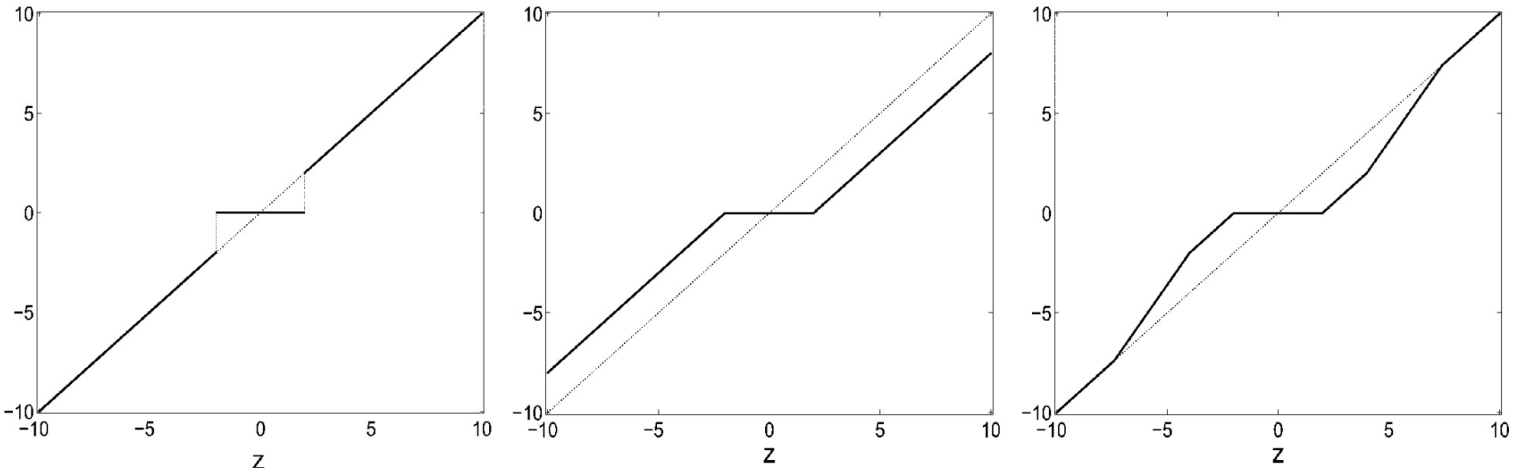


Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.

Figure from Fan and Li (2001).

Sparse PCA: The Power Method with Thresholding

- Multiplication: $\bar{U}^{(t)} = \widehat{\Sigma}\widehat{U}^{(t-1)}$;
- Thresholding: $\widetilde{U}^{(t)} = H_\gamma(\bar{U}^{(t)})$;
- QR normalization: $\widehat{U}^{(t)}R^{(t)} = \widetilde{U}^{(t)}$.
- The thresholding function: $H_\gamma(U) = (h_{\gamma_k}(U_{jk}))_{1\le j\le p, 1\le k\le r}$

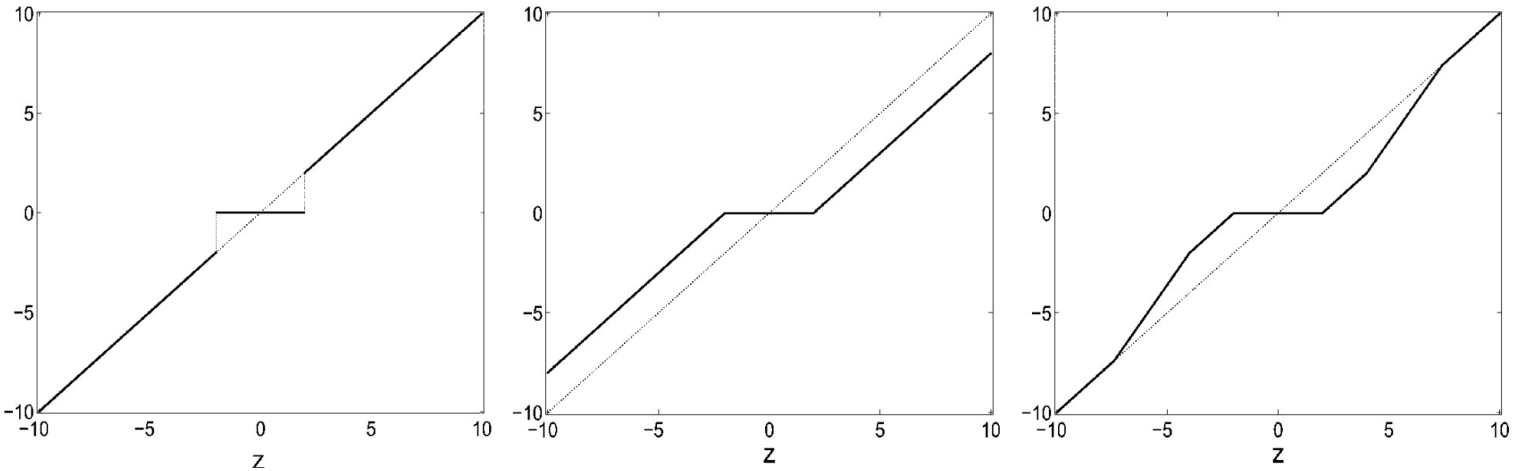


Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.

Figure from Fan and Li (2001).

- Thresholding level $\gamma = \gamma_0\sigma\left(\sigma + \sqrt{\lambda_k(\Lambda)}\right)\sqrt{\frac{\log p}{n}}$.

Challenges with Heteroskedastic Noises

Main Ideas

- Heteroskedastic noise brings bias. Notice that

$$\mathrm{Cov}(X_i) = U\Lambda U^\top + diag\{\sigma_1^2, \ldots, \sigma_p^2\}$$

Main Ideas

- Heteroskedastic noise brings bias. Notice that

$$\mathrm{Cov}(X_i) = U\Lambda U^\top + \mathit{diag}\{\sigma_1^2, \ldots, \sigma_p^2\}$$

Eigenvectors of $\Sigma = \mathrm{Cov}(X_i)$ are generally not the same as U, the eigenvectors of $\Sigma_0 = U\Lambda U^\top$.

Main Ideas

- Heteroskedastic noise brings bias. Notice that

$$\mathrm{Cov}(X_i) = U\Lambda U^\top + \mathit{diag}\{\sigma_1^2, \ldots, \sigma_p^2\}$$

  Eigenvectors of $\Sigma = \mathrm{Cov}(X_i)$ are generally not the same as U, the eigenvectors of $\Sigma_0 = U\Lambda U^\top$.

- Key fact:
  (Off-diagonal) $\Delta(\widehat{\Sigma}) \to \Delta(\Sigma_0)$ (no bias);
  (Diagonal) $D(\widehat{\Sigma}) \not\to D(\Sigma_0)$ (biased).

Main Ideas

Two potential approaches:

- Diagonal deletion (Florescu and Perkins (2016)) : delete diagonal entries, i.e., replace $\widehat{\Sigma}$ by $\Delta(\widehat{\Sigma})$.
- Diagonal imputation (Zhang et al. (2022)) : use $D(\widehat{\Sigma}_0)$ with some estimator $\widehat{\Sigma}_0$ to approximate $D(\Sigma_0)$ instead, i.e., replace $\widehat{\Sigma}$ by $\Delta(\widehat{\Sigma}) + D(\widehat{\Sigma}_0)$.

Algorithm: HSPCA

- Input: Sample covariance matrix $\widehat{\Sigma}$, target subspace dimension r, thresholding parameters γ, initial orthonormal matrix $\widehat{U}^{(0)}$.

Algorithm: HSPCA

- Input: Sample covariance matrix $\widehat{\Sigma}$, target subspace dimension r, thresholding parameters γ, initial orthonormal matrix $\widehat{U}^{(0)}$.
- Initialization: $\widehat{\Sigma}_0^{(t)} = \Delta(\widehat{\Sigma})$ for $t = 0$.

Algorithm: HSPCA

- Input: Sample covariance matrix $\widehat{\Sigma}$, target subspace dimension r, thresholding parameters γ, initial orthonormal matrix $\widehat{U}^{(0)}$.
- Initialization: $\widehat{\Sigma}_0^{(t)} = \Delta(\widehat{\Sigma})$ for $t = 0$.

  For $t = 1, \ldots, T$, do the following:

Algorithm: HSPCA

- Input: Sample covariance matrix $\widehat{\Sigma}$, target subspace dimension r, thresholding parameters γ, initial orthonormal matrix $\widehat{U}^{(0)}$.
- Initialization: $\widehat{\Sigma}_0^{(t)} = \Delta(\widehat{\Sigma})$ for $t = 0$.

  For $t = 1, \ldots, T$, do the following:
- Part I: Power method: $\widehat{\Sigma}_0^{(t-1)} \to \widehat{U}^{(t)}$.
  - Multiplication: $\bar{U}^{(t)} = \widehat{\Sigma}_0^{(t-1)} \widehat{U}^{(t-1)}$;
  - Adaptive thresholding: $\widetilde{U}^{(t)} = H_\gamma(\bar{U}^{(t)})$;
  - QR normalization: $\widehat{U}^{(t)} R^{(t)} = \widetilde{U}^{(t)}$.

Algorithm: HSPCA

- Input: Sample covariance matrix $\widehat{\Sigma}$, target subspace dimension r, thresholding parameters γ, initial orthonormal matrix $\widehat{U}^{(0)}$.
- Initialization: $\widehat{\Sigma}_0^{(t)} = \Delta(\widehat{\Sigma})$ for $t = 0$.

  For $t = 1, \ldots, T$, do the following:
- Part I: Power method: $\widehat{\Sigma}_0^{(t-1)} \to \widehat{U}^{(t)}$.
  - Multiplication: $\bar{U}^{(t)} = \widehat{\Sigma}_0^{(t-1)} \widehat{U}^{(t-1)}$;
  - Adaptive thresholding: $\widetilde{U}^{(t)} = H_\gamma(\bar{U}^{(t)})$;
  - QR normalization: $\widehat{U}^{(t)} R^{(t)} = \widetilde{U}^{(t)}$.
- Part II: Update diagonal values: $\widehat{U}^{(t)} \to \widehat{\Sigma}_0^{(t)}$ (Recall $\Sigma_0 = U \Lambda U^\top$)
  - $\widehat{\Lambda}^{(t)} = \widehat{U}^{(t)\top} \widehat{\Sigma}_0^{(t-1)} \widehat{U}^{(t)}$;
  - $\widetilde{\Sigma}_0^{(t)} = \widehat{U}^{(t)} \widehat{\Lambda}^{(t)} \widehat{U}^{(t)\top}$;
  - $\widehat{\Sigma}_0^{(t)} = D(\widetilde{\Sigma}_0^{(t)}) + \Delta(\widehat{\Sigma})$.

HSPCA: The Power Method with Adaptive Thresholding

- Part I: Power method: $\widehat{\Sigma}_0^{(t-1)} \to \widehat{U}^{(t)}$.
  - Multiplication: $\bar{U}^{(t)} = \widehat{\Sigma}_0^{(t-1)} \widehat{U}^{(t-1)}$;
  - Adaptive thresholding: $\widetilde{U}^{(t)} = H_\gamma(\bar{U}^{(t)})$;
  - QR normalization: $\widehat{U}^{(t)} R^{(t)} = \widetilde{U}^{(t)}$.

- The thresholding function: $H_\gamma(U) = (h_{\gamma_{jk}}(U_{jk}))_{1 \le j \le p, 1 \le k \le r}$
- Adaptive Thresholding level $\gamma_{jk} = \gamma_0 \sigma_j \left(\sigma_{\max} + \sqrt{\lambda_k(\Lambda)} \right) \sqrt{\frac{\log p}{n}}$.

An Algorithm for Initializer: Off-Diagonal Sparse PCA

- Input: Sample covariance matrix $\widehat{\Sigma}$, target subspace dimension r, thresholding parameter γ_0.
- Step 1: Diagonal deletion with off-diagonal thresholding
  Let $\widehat{\Sigma}_0 = \left(\widehat{\sigma}_{ij}^{(0)}\right)$, where $\widehat{\sigma}_{ii}^{(0)} = 0$, for $i = 1, \ldots, p$,
  and $\widehat{\sigma}_{ij}^{(0)} = \widehat{\sigma}_{ij} \cdot I\left\{|\widehat{\sigma}_{ij}| \geqslant \gamma_0 (\Sigma_{ii}\Sigma_{jj})^{1/2} (\log p/n)^{1/2}\right\}$ for $i \neq j$.
- Step 2: SVD.
  Do SVD to $\widehat{\Sigma}_0$, get the matrix $U^{(0)}$ that contains the first r singular vectors.

Optimality and Theoretical Analysis

Theoretical Guarantee

Recall our generalized Spiked Covariance Model:

$$X_i = S_i + \varepsilon_i, \quad S_i \perp\!\!\!\perp \varepsilon_i$$
$$\mathrm{Cov}(S_i) = U\Lambda U^\top, \quad \mathrm{Cov}(\varepsilon_i) = diag\{\sigma_1^2, \ldots, \sigma_p^2\}$$
$$U = (u_1, \ldots, u_r), \quad \|u_i\|_0 \le s, \forall i = 1, \ldots, r$$

Let

$$\sigma^2_{\max} := \max_{1 \leqslant j \leqslant p} \sigma_j^2, \quad \sigma^2_{\mathrm{sum}} := \sum_{j \in S} \sigma_j^2$$

where $S := \{1 \leqslant j \leqslant p : U_{j\star} \neq 0\}$ is the support of the eigenvector matrix U.

Assumptions

Condition (1)

(Incoherence condition) $I(U) = (p/r)\max_{i\in[p]} \left\|e_i^\top U\right\|_2^2 \leqslant c_I p/r$ with some universal constant $c_I < 1$.

Condition (2)

$\lambda_1(\Lambda)/\lambda_r(\Lambda) \leqslant C_\Lambda$ with some universal constant $C_\Lambda > 1$.

Condition (3)

(i) $\max\{\frac{\log p}{n}, \sqrt{\frac{\log p}{n}}\left(\frac{\sigma_{\text{sum}}\sigma_{\max}}{\lambda_r(\Lambda)} + \frac{\sigma_{\text{sum}}}{\sqrt{\lambda_r(\Lambda)}}\right)\} = o(1)$;

(ii) $\sqrt{\frac{\log p}{n}}\frac{\sigma_{\text{sum}}^2}{\lambda_r(\Lambda)} \leq c_s$ for some sufficiently small universal constant $c_s < 1$.

Main Theorem

Theorem (Initializer)

Assume Conditions 1-3 hold. Under Gaussian assumption, the output $\widehat{U}^{(0)}$ *of our algorithm satisfies*

$$\|\sin\Theta(\widehat{U}^{(0)}, U)\| \lesssim \sqrt{\frac{\log p}{n}}\left(1 + \frac{\sigma_{\text{sum}}\sigma_{\max}}{\lambda_r(\Lambda)} + \frac{\sigma_{\text{sum}}^2}{\lambda_r(\Lambda)}\right) + c_I C_\Lambda$$

with probability at least $1 - c_1 p^{-2}$*, where* c_1 *is a universal constant independent of* (n, p, s)*.*

Remark: Whenever $c_s + c_I C_\Lambda < 1$, and n is large, we have

$$\|\sin\Theta(\widehat{U}^{(0)}, U)\| < 1.$$

Main Theorem

Theorem

Assume Conditions 1-3 hold. Under Gaussian assumption, after $T \asymp \log n$ *iterations, the output* $\widehat{U}^{(T)}$ *of our algorithm satisfies*

$$\|\sin\Theta(\widehat{U}^{(T)}, U)\| \lesssim \sqrt{\frac{\log p}{n}}\left(\frac{\sigma_{\text{sum}}\sigma_{\max}}{\lambda_r(\Lambda)} + \frac{\sigma_{\text{sum}}}{\sqrt{\lambda_r(\Lambda)}}\right) \wedge 1$$

with probability at least $1 - c_1 p^{-2}$*, where* c_1 *is a universal constant independent of* (n, p, s)*.*

Remarks

- Consider the homoskedastic PCA where $\sigma_1^2 = \ldots = \sigma_p^2 := \sigma^2$. Then

$$\|\sin\Theta(\widehat{U}^{(T)}, U)\| \lesssim \sqrt{\frac{s\log p}{n}}\left(\frac{\sigma^2}{\lambda_r(\Lambda)} + \frac{\sigma}{\sqrt{\lambda_r(\Lambda)}}\right) \wedge 1$$

Remarks

- Consider the homoskedastic PCA where $\sigma_1^2 = \ldots = \sigma_p^2 := \sigma^2$. Then

$$\|\sin\Theta(\widehat{U}^{(T)}, U)\| \lesssim \sqrt{\frac{s\log p}{n}}\left(\frac{\sigma^2}{\lambda_r(\Lambda)} + \frac{\sigma}{\sqrt{\lambda_r(\Lambda)}}\right) \wedge 1$$

- Let $\widetilde{s} = \sigma_{\text{sum}}^2/\sigma_{\max}^2$. Then

$$\|\sin\Theta(\widehat{U}^{(T)}, U)\| \lesssim \sqrt{\frac{\widetilde{s}\log p}{n}}\left(\frac{\sigma_{\max}^2}{\lambda_r(\Lambda)} + \frac{\sigma_{\max}}{\sqrt{\lambda_r(\Lambda)}}\right) \wedge 1$$

Remarks

- Consider the homoskedastic PCA where $\sigma_1^2 = \ldots = \sigma_p^2 := \sigma^2$. Then

$$\|\sin\Theta(\widehat{U}^{(T)}, U)\| \lesssim \sqrt{\frac{s\log p}{n}}\left(\frac{\sigma^2}{\lambda_r(\Lambda)} + \frac{\sigma}{\sqrt{\lambda_r(\Lambda)}}\right) \wedge 1$$

- Let $\widetilde{s} = \sigma_{\text{sum}}^2/\sigma_{\text{max}}^2$. Then

$$\|\sin\Theta(\widehat{U}^{(T)}, U)\| \lesssim \sqrt{\frac{\widetilde{s}\log p}{n}}\left(\frac{\sigma_{\text{max}}^2}{\lambda_r(\Lambda)} + \frac{\sigma_{\text{max}}}{\sqrt{\lambda_r(\Lambda)}}\right) \wedge 1$$

- $\widetilde{s} = \sigma_{\text{sum}}^2/\sigma_{\text{max}}^2$ can be viewed as "effective dimension" for heteroskedastic sparse PCA. (noting that $\widetilde{s} \leq s$.)

Remarks

- Non-adaptive thresholding in Part 1 leads to suboptimal rate of convergence: $\widetilde{s}$ is replaced by s.

Remarks

- Non-adaptive thresholding in Part 1 leads to suboptimal rate of convergence: $\widetilde{s}$ is replaced by s.

- Computational lower bound: Condition 3 (ii) $\sqrt{\frac{\log p}{n}} \frac{\sigma_{\text{sum}}^2}{\lambda_r(\Lambda)} \leq c_s$ is necessary. (Berthet and Rigollet (2013), Wang et al. (2016))

A Minimax Lower Bound for Heteroskedastic Sparse PCA

Consider the heteroskedastic sparse spiked covariance model:

$$\begin{aligned}&\mathcal{F}_{n,p}\left(\sigma_{\max}, \sigma_{\text{sum}}, \nu, s\right)\\ =&\Bigg\{\Sigma = U\Lambda U^\top + \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right) : U \in \mathbb{O}_{p,r}, \|U\|_0 \leqslant s, I(U) \leqslant c_I p/r,\\ &\lambda_1(\Lambda)/\lambda_r(\Lambda) \leqslant \kappa, \lambda_r(\Lambda) \geqslant \nu, \max_{1\leqslant j\leqslant p} \sigma_j^2 \leqslant \sigma_{\max}^2, \sum_{j\in S} \sigma_j^2 \leqslant \sigma_{\text{sum}}^2 \Bigg\}.\end{aligned}$$

Theorem

$$\inf_{\widehat{U}} \sup_{\Sigma\in\mathcal{F}_{n,p}} \mathbb{E}\|\sin\Theta(\widehat{U}, U)\| \gtrsim \sqrt{\frac{\log(\frac{p}{s})}{n}}\left(\frac{\sigma_{\text{sum}}\sigma_{\max}}{\nu} + \frac{\sigma_{\text{sum}}}{\sqrt{\nu}}\right) \wedge 1$$

Numeric Results

Simulation Scenarios

- Heteroskedastic noise variance $(\sigma_1^2, \ldots, \sigma_p^2)$:

$$v_1, \ldots, v_p \overset{i.i.d}{\sim} \mathrm{Unif}(0,1), \quad \sigma_i^2 = \frac{\beta \times p \times v_i^\alpha}{\sum_{k=1}^p v_k^\alpha}$$

- Single spike $(r = 1)$: $U = \frac{1}{\sqrt{s}} \left(1_s^\top, 0_{p-s}^\top\right)^\top$
- Three spikes $(r = 3)$: $U = \left[U_S^\top, 0_{p-s,p-s}\right]^\top$, where

$$U_S = \begin{bmatrix} \frac{1}{\sqrt{s/3}} 1_{s/3} & & \\ & \frac{1}{\sqrt{s/3}} 1_{s/3} & \\ & & \frac{1}{\sqrt{s/3}} 1_{s/3} \end{bmatrix}$$

- $\beta = 0.1, s = 0.1p, p = 2n, n = 256.$

Simulation Performance

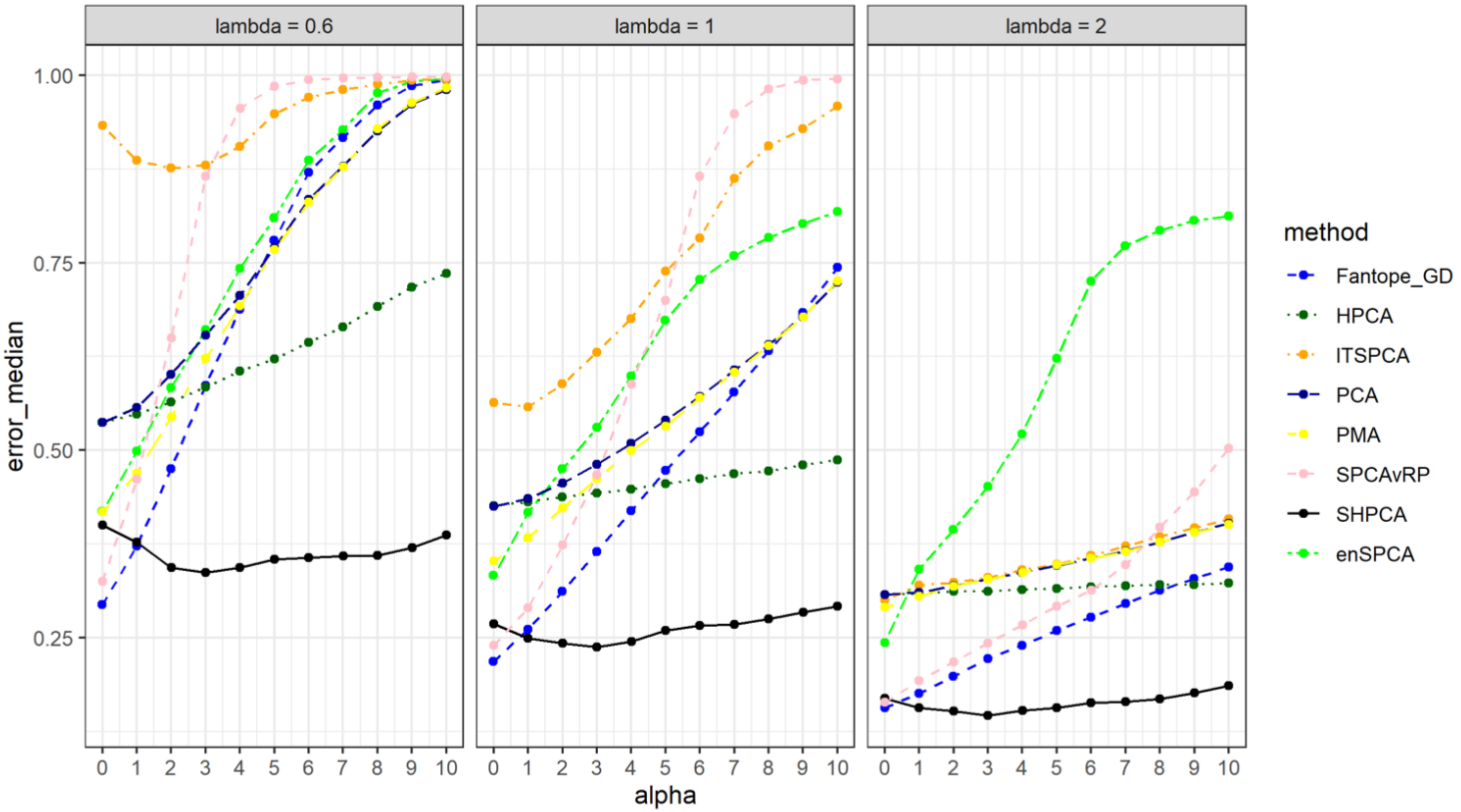


Figure: $r = 1$. L: $\lambda = 1$; M: $\lambda = 2$; R: $\lambda = 5$.

Fantope_GD: Qiu et. al. (2019); HPCA: Zhang et. al. (2022); ITSPCA: Ma (2013); PMA: Witten et. al. (2009); SPCAvRP: Gataric et. al. (2020); enSPCA: Zou et. al. (2006) SHPCA: Our method

Simulation Performance

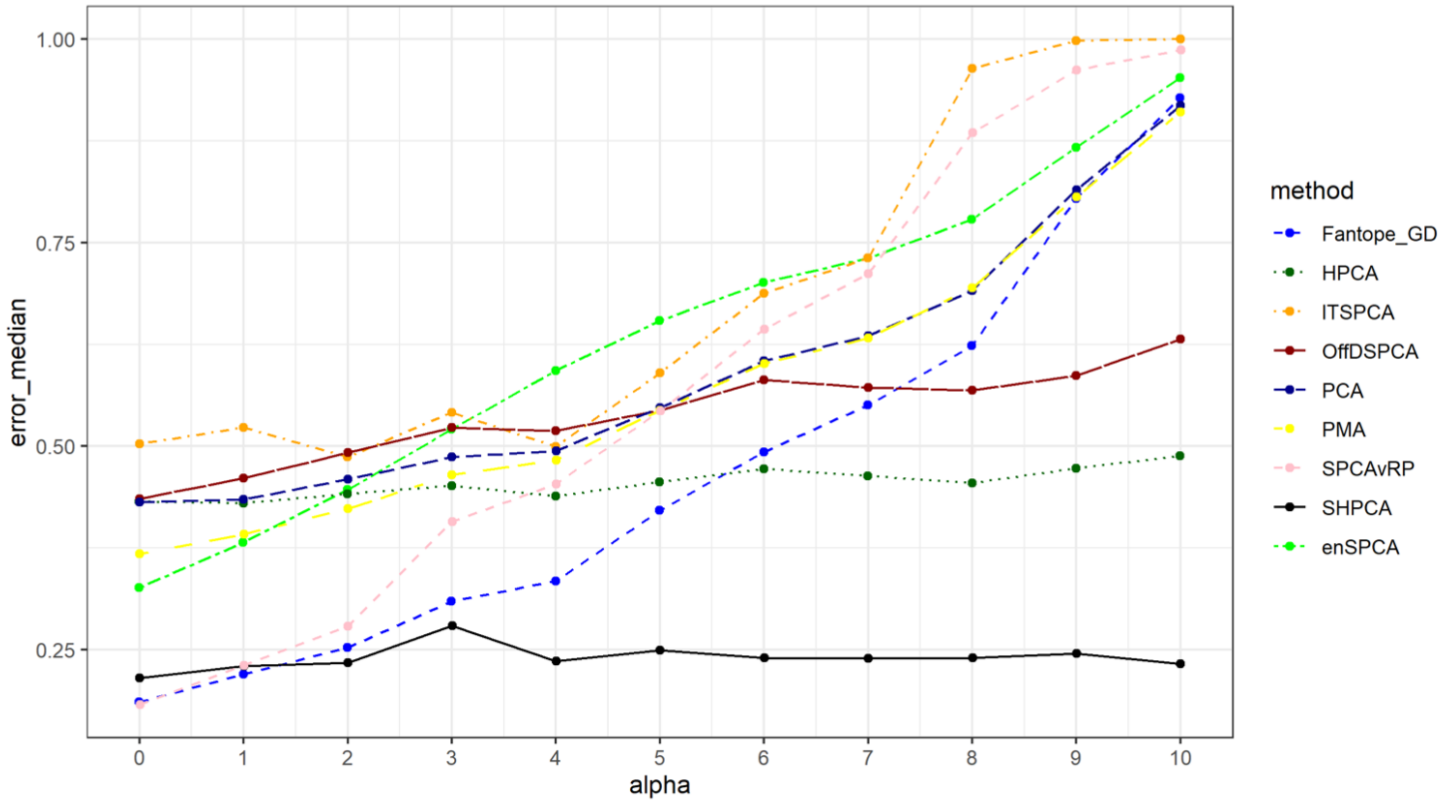


Figure: $r = 3$. $\Lambda = diag\{5, 2, 1\}$.

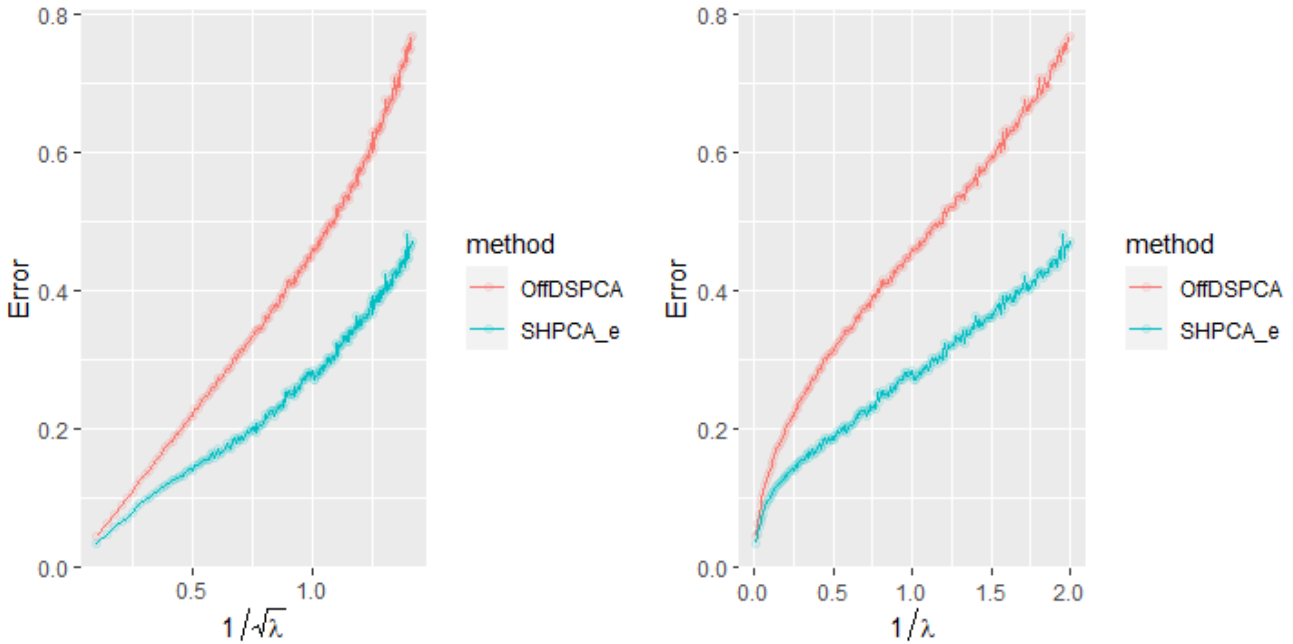
Fantope_GD: Qiu et. al. (2019); HPCA: Zhang et. al. (2022); ITSPCA: Ma (2013); PMA: Witten et. al. (2009); SPCAvRP: Gataric et. al. (2020); enSPCA: Zou et. al. (2006) OffDSPCA: Our Initializer; SHPCA: Our method

Rate of convergence

Recall our rate of convergence is

$$\sqrt{\frac{\log p}{n}}\left(\frac{\sigma_{\text{sum}}\sigma_{\max}}{\lambda_r(\Lambda)}+\frac{\sigma_{\text{sum}}}{\sqrt{\lambda_r(\Lambda)}}\right)$$

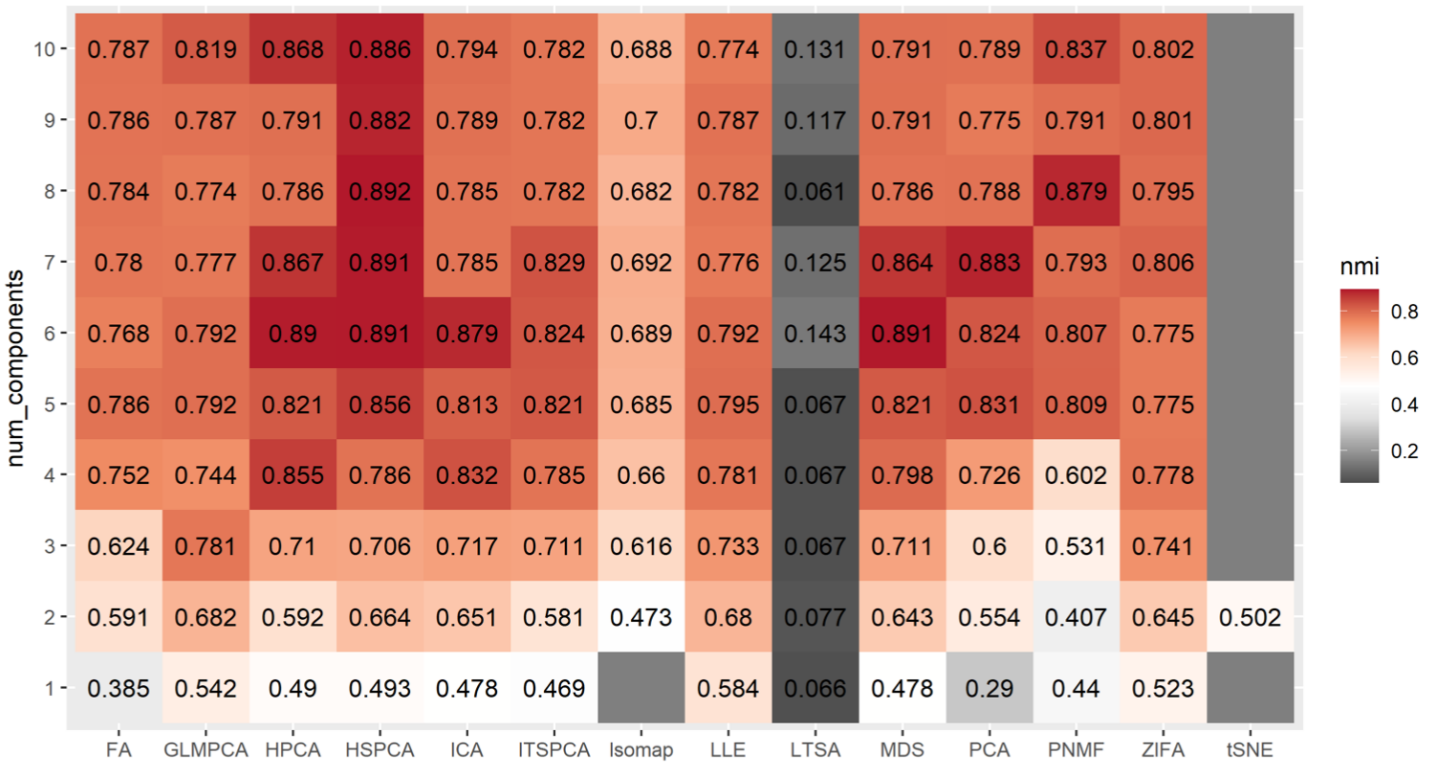
.



Application: Single-cell RNA-seq Data

Baron et al. (2016): transcript abundance across 1886 cells (into 13 clusters) and 14878 genes of mouse pancreatic islets;

- $p = 2000$ genes are picked using Seurat package; $n = 500$.
- Various dimension reduction methods together with k-means: t-SNE, Isomap, ICA, FA, Poisson Nonnegative Matrix Factorization, et al.
- Accuracy by normalized mutual information.



Extensions and Future works

Heteroskedastic Sparse SVD

- Model: $X = S + E$. $S \in \mathbb{R}^{p_1 \times p_2}$ is a low-rank matrix of interest with a sparse left singular vector matrix U; E is a noise matrix with independent entries.
- Examples include:
  - Possion PCA: Salmon et al. (2014)
  - Exponential family PCA: Liu et al. (2016)
  - Matrix denoising: Yang, Ma and Buja (2016)
  - ...
- Let $\Sigma = XX^\top$, $\Sigma_0 = SS^\top$, then

$$(\mathbb{E}\Sigma)_{ij} = \begin{cases} \Sigma_{0,ij}, & i \neq j \\ \Sigma_{0,ij} + \sum_{k=1}^{p_2} \mathrm{Var}\left(E_{ik}\right), & i = j \end{cases}$$

Summary

- Sparse Heteroskedastic PCA for Generalized Spiked Covariance Model.
- (i) Imputing diagonal values; (ii) Power method with adaptive thresholding.
- Lower bound: Fano's lemma.

Additional Remark on Adaptive Thresholding

Adaptive Threshold Levels for Algorithm 1: With the output of Algorithm 2, we can estimate the threshold levels defined in (2.4) by

$$\hat{\gamma}_{\nu j} = \gamma \hat{\sigma}_\nu \left(\hat{\sigma}_{\max} + \sqrt{\hat{\lambda}_j} \right) \sqrt{\frac{\log p}{n}}, \ 1 \leqslant \nu \leqslant p \text{ and } 1 \leqslant j \leqslant r, \tag{2.7}$$

where $\hat{\sigma}_i^2 = (\widehat{\Sigma}_{ii} - (\widehat{U}^{(0)} diag(\hat{\lambda}_1, \ldots, \hat{\lambda}_r)(\widehat{U}^{(0)})^\top)_{ii}) \vee 0$ and $\hat{\sigma}_{\max}^2 = \max_{1 \leqslant i \leqslant p} \hat{\sigma}_i^2$. The threshold levels here are adaptive to the signal and noise levels.

Assumption 5. *There exists a constant C_σ such that for $1 \leqslant i \leqslant p$, $\frac{\lambda_1}{\sigma_i^2} \leqslant C_\sigma$.*

Proposition 1. *Assume that Assumptions 1 - 5 hold with small enough c_s and c_I, there exist some constants C_1 and C_2 such that with probability at least $1 - cp^{-2}$:*

$$C_1 \gamma_{\nu j} \leqslant \hat{\gamma}_{\nu j} \leqslant C_2 \gamma_{\nu j}. \tag{3.5}$$