# Sparse PCA
# and
# Sparse Covariance Estimation

## Theory, Algorithms and Applications

Boaz Nadler

Department of Computer Science and Applied Mathematics
The Weizmann Institute of Science

Dec 2021

# Introduction

We observe $n$ i.i.d. realizations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $N(0, \Sigma_p)$.

$S_n$ - sample covariance matrix.

If $n = O(p)$ or $n \ll p$, $S_n$ may be poor approximation to $\Sigma$, and eigenvectors of $S_n$ may be poor approximation to eigenvectors of $\Sigma$.

# Different Sparsity Assumptions

To obtain a better estimator (of $\Sigma, \Sigma^{-1}, PCA(\Sigma)$ ), need some additional assumptions on structure of problem.

# Different Sparsity Assumptions

To obtain a better estimator (of $\Sigma, \Sigma^{-1}, PCA(\Sigma)$ ), need some additional assumptions on structure of problem.

Several Recent Suggestions:

- *Sparse-PCA:* few large eigenvalues with sparse eigenvectors.

# Different Sparsity Assumptions

To obtain a better estimator (of $\Sigma, \Sigma^{-1}, PCA(\Sigma)$ ), need some additional assumptions on structure of problem.

Several Recent Suggestions:

- *Sparse-PCA:* few large eigenvalues with sparse eigenvectors.
- Sparse Covariance matrix (most entries close to zero).

# Different Sparsity Assumptions

To obtain a better estimator (of $\Sigma, \Sigma^{-1}, PCA(\Sigma)$ ), need some additional assumptions on structure of problem.

Several Recent Suggestions:

- *Sparse-PCA:* few large eigenvalues with sparse eigenvectors.
- Sparse Covariance matrix (most entries close to zero).
- Sparse Inverse covariance (graphical models, conditional independence).

# Different Sparsity Assumptions

To obtain a better estimator (of $\Sigma, \Sigma^{-1}, PCA(\Sigma)$ ), need some additional assumptions on structure of problem.

Several Recent Suggestions:

- *Sparse-PCA:* few large eigenvalues with sparse eigenvectors.
- Sparse Covariance matrix (most entries close to zero).
- Sparse Inverse covariance (graphical models, conditional independence).
- Robust-PCA: Matrix $\Sigma =$ low rank $+$ sparse (outlier noise).

# Introduction

Questions:

Under different sparsity assumptions,

- How to construct better estimators ?

# Introduction

Questions:

Under different sparsity assumptions,

- How to construct better estimators ?

- What are fundamental limits for detection/estimation.

# Sparse PCA

Consider single spike model

$$\mathbf{x} = \sqrt{\lambda} s \mathbf{v} + \sigma \boldsymbol{\xi}$$

Assume eigenvector $\mathbf{v}$ is approximately sparse:

$$L_q(C) = \{\mathbf{v} \in \mathbb{R}^p \mid \|\mathbf{v}\|_2 = 1, \sum_j |v_j|^q \leq C^q\}$$

$q \in [0, 2)$,
smaller value for $q$ - a sparser vector $\mathbf{v}$.

# Sparse PCA

Consider single spike model

$$\mathbf{x} = \sqrt{\lambda} s \mathbf{v} + \sigma \boldsymbol{\xi}$$

Assume eigenvector $\mathbf{v}$ is approximately sparse:

$$L_q(C) = \{\mathbf{v} \in \mathbb{R}^p \mid \|\mathbf{v}\|_2 = 1, \sum_j |v_j|^q \leq C^q\}$$

$q \in [0, 2)$,

smaller value for $q$ - a sparser vector $\mathbf{v}$.

$q = 0$ then, $\mathbf{v}$ is sparse with at most $C$ non-zero entries.

# Sparse Eigenvector Estimation

Loss Function: quality of estimate $\hat{\mathbf{v}}$ of $\mathbf{v}$

$$L(\hat{\mathbf{v}}, \mathbf{v}) = \min\{\|\hat{\mathbf{v}} - \mathbf{v}\|_2^2, \|\hat{\mathbf{v}} + \mathbf{v}\|_2^2\}$$

# Sparse Eigenvector Estimation

Loss Function: quality of estimate $\hat{\mathbf{v}}$ of $\mathbf{v}$

$$L(\hat{\mathbf{v}}, \mathbf{v}) = \min\{\|\hat{\mathbf{v}} - \mathbf{v}\|_2^2, \|\hat{\mathbf{v}} + \mathbf{v}\|_2^2\}$$

Remark:

$$L(\mathbf{a}, \mathbf{b}) = 2(1 - |\mathbf{a}^T \mathbf{b}|).$$

# Sparse Eigenvector Estimation

Loss Function: quality of estimate $\hat{\mathbf{v}}$ of $\mathbf{v}$

$$L(\hat{\mathbf{v}}, \mathbf{v}) = \min\{\|\hat{\mathbf{v}} - \mathbf{v}\|_2^2, \|\hat{\mathbf{v}} + \mathbf{v}\|_2^2\}$$

Remark:

$$L(\mathbf{a}, \mathbf{b}) = 2(1 - |\mathbf{a}^T\mathbf{b}|).$$

**Theorem** In the joint limit $p, n \to \infty$, if $\lambda > \sqrt{p/n}$, then

$$R^2 = |\hat{\mathbf{v}}_{\text{PCA}}^T \mathbf{v}|^2 \to \frac{\frac{n}{p}\lambda^2 - 1}{\frac{n}{p}\lambda^2 + \lambda}$$

if $\lambda \le \sqrt{p/n}$ then $R \to 0$.

# Diagonal Thresholding

Question: Assuming a sparse eigenvector, is it possible to achieve smaller errors ?

# Diagonal Thresholding

Question: Assuming a sparse eigenvector, is it possible to achieve smaller errors ?

Very simple method [Johnstone and Lu, J. Am. Stat. Assoc. 2009]:

- Compute diagonal of $S_n$.

- $I = \{i \,|\, (S_n)_{ii} > \sigma^2 t(\alpha)\}$

- Compute eigenvector of $(S_n)|_I$.

# Diagonal Thresholding

Questions:

- How should the threshold $t(\alpha)$ be chosen ?

- What is the resulting error ?

- Is this method rate optimal ? (what is optimal ?)

# Diagonal Thresholding

The threshold $t(\alpha)$: Assume $\mathbf{v}$ was truly sparse, with few $\ll p$ non-zero entries.

$$\Pr[S_{ii} > t(\alpha) \,|\, v_i = 0] \approx \alpha/p$$

or

$$\Pr[\max_i S_{ii} > t(\alpha) \,|\, v_i = 0] \approx \alpha$$

# Diagonal Thresholding

The threshold $t(\alpha)$: Assume $\mathbf{v}$ was truly sparse, with few $\ll p$ non-zero entries.

$$\Pr[S_{ii} > t(\alpha) \,|\, v_i = 0] \approx \alpha/p$$

or

$$\Pr[\max_i S_{ii} > t(\alpha) \,|\, v_i = 0] \approx \alpha$$

maxima of many i.i.d. random variables is a classical problem in **extreme value theory**.

# Diagonal Thresholding

**Theorem** Let $Z_i$ be $p$ i.i.d. $N(0,1)$ r.v.'s. Then, as $p \to \infty$

$$\max_i X_i \to \sqrt{2 \ln p}$$

# Diagonal Thresholding

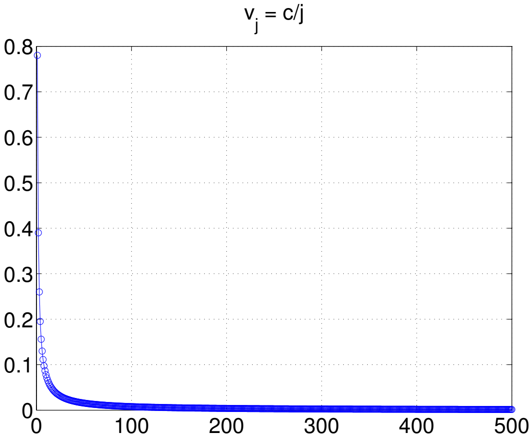
**Theorem** Let $Z_i$ be $p$ i.i.d. $N(0,1)$ r.v.'s. Then, as $p \to \infty$

$$\max_i X_i \to \sqrt{2 \ln p}$$

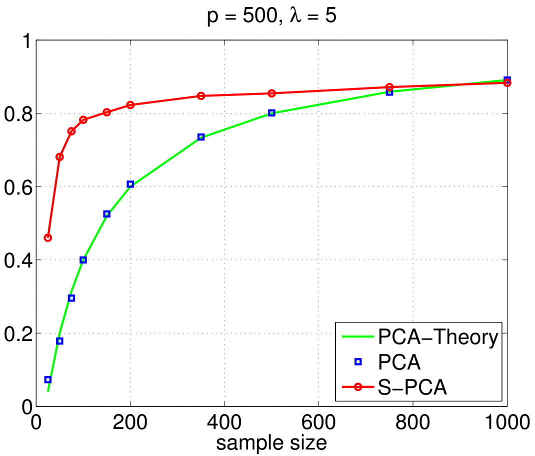
In our case, in the absence of a signal, all $S_{ii}$ are i.i.d., with distribution $\chi^2_n/n$.
For large $n$, $\chi^2_n/n \approx 1 + \sqrt{\frac{2}{n}} N(0,1)$.
Then threshold

$$t(\alpha) \approx 1 + \sqrt{\frac{2 \ln p}{n}}(1 + o(1)).$$

# Diagonal Thresholding



# Diagonal Thresholding



# Diagonal Thresholding

Which coordinate strengths can be detected ?

# Diagonal Thresholding

Which coordinate strengths can be detected ?

For signal coordinate, $S_{ii} = (1 + \lambda v_i^2)\chi_n^2/n$.

$$\mathbb{E}[S_{ii}] = 1 + \lambda v_i^2 > 1 + \sqrt{2 \ln p/n}$$

Thus

$$\lambda v_i^2 > C\sqrt{\ln p/n}$$

# Diagonal Thresholding

[Joint work with A. Birnbaum, D. Paul and I.M. Johnstone]

What is the worst error rate of Diagonal Thresholding algorithm ?

# Diagonal Thresholding

[Joint work with A. Birnbaum, D. Paul and I.M. Johnstone]

What is the worst error rate of Diagonal Thresholding algorithm ?

**Theorem:** Consider single signal, $\mathbf{v} \in L_q(C)$. Then,

$$\sup_{\mathbf{v} \in L_q(C)} \|\hat{\mathbf{v}}_{\text{DT}} - \mathbf{v}\|^2 \geq K(C^q - 1)\, n^{-\frac{1}{2}(1-\frac{q}{2})}$$

# Diagonal Thresholding

[Joint work with A. Birnbaum, D. Paul and I.M. Johnstone]

What is the worst error rate of Diagonal Thresholding algorithm ?

**Theorem:** Consider single signal, $\mathbf{v} \in L_q(C)$. Then,

$$\sup_{\mathbf{v} \in L_q(C)} \|\hat{\mathbf{v}}_{\text{DT}} - \mathbf{v}\|^2 \geq K(C^q - 1)\, n^{-\frac{1}{2}(1-\frac{q}{2})}$$

Question: Can one do better ?

# Oracle Rate

Consider an oracle that tells us all large coordinates of $\mathbf{v}$,

$$I_\delta = \{j \in \{1, 2, \ldots, p\} \mid |v_j| > \delta\}$$

Then, we could do PCA only on $S$ restricted to $I$.

# Oracle Rate

Consider an oracle that tells us all large coordinates of $\mathbf{v}$,

$$I_\delta = \{j \in \{1,2,\ldots,p\} \mid |v_j| > \delta\}$$

Then, we could do PCA only on $S$ restricted to $I$.

$$\|v_I - \hat{v}_I\|^2 \approx \frac{1}{\lambda_I}\frac{|I|-1}{n}$$

# Oracle Rate

Consider an oracle that tells us all large coordinates of $\mathbf{v}$,

$$I_\delta = \{j \in \{1, 2, \ldots, p\} \,|\, |v_j| > \delta\}$$

Then, we could do PCA only on $S$ restricted to $I$.

$$\|v_I - \hat{v}_I\|^2 \approx \frac{1}{\lambda_I} \frac{|I| - 1}{n}$$

**Overall Error:**

$$\|\mathbf{v} - \hat{v}_I\|^2 = \|v_I^\perp\|^2 + \|v_I - \hat{v}_I\|^2$$

# Oracle Rate

**Overall Error:**

$$\|\mathbf{v} - \hat{v}_I\|^2 = \underbrace{\|v_I^\perp\|^2}_{\text{squared bias}} + \underbrace{\|v_I - \hat{v}_I\|^2}_{\text{variance}}$$

# Oracle Rate

**Overall Error:**

$$\|\mathbf{v} - \hat{v}_I\|^2 = \underbrace{\|v_I^\perp\|^2}_{\text{squared bias}} + \underbrace{\|v_I - \hat{v}_I\|^2}_{\text{variance}}$$

or

$$\|\mathbf{v} - \hat{v}_I\|^2 = \underbrace{\|v_I^\perp\|^2}_{\text{approximation err}} + \underbrace{\|v_I - \hat{v}_I\|^2}_{\text{estimation err}}$$

# Oracle Rate

**Overall Error:**

$$\|\mathbf{v} - \hat{v}_I\|^2 = \underbrace{\|v_I^\perp\|^2}_{\text{squared bias}} + \underbrace{\|v_I - \hat{v}_I\|^2}_{\text{variance}}$$

or

$$\|\mathbf{v} - \hat{v}_I\|^2 = \underbrace{\|v_I^\perp\|^2}_{\text{approximation err}} + \underbrace{\|v_I - \hat{v}_I\|^2}_{\text{estimation err}}$$

**Bias - Variance Tradeoff**

# Oracle Rate

**Claim:** Optimal oracle threshold is to choose all coordinates up to

$$v_i^2 > \frac{C}{n}.$$

compare to $\sqrt{2\ln p/n}$

# Oracle Rate

**Claim:** Optimal oracle threshold is to choose all coordinates up to

$$v_i^2 > \frac{C}{n}.$$

compare to $\sqrt{2 \ln p / n}$

**Claim:** Let $\hat{\mathbf{v}}_{\text{oracle}}$ denote the PCA estimator using coordinates chosen by oracle with optimal threshold. Then,

$$\sup_{\mathbf{v} \in L_q(C)} \|\hat{\mathbf{v}}_{\text{oracle}} - \mathbf{v}\|^2 = O(n^{(1-q/2)})$$

# Oracle Rate

**Claim:** Optimal oracle threshold is to choose all coordinates up to

$$v_i^2 > \frac{C}{n}.$$

compare to $\sqrt{2 \ln p/n}$

**Claim:** Let $\hat{\mathbf{v}}_{\text{oracle}}$ denote the PCA estimator using coordinates chosen by oracle with optimal threshold. Then,

$$\sup_{\mathbf{v} \in L_q(C)} \|\hat{\mathbf{v}}_{\text{oracle}} - \mathbf{v}\|^2 = O(n^{(1-q/2)})$$

compare to DT rate of $n^{-\frac{1}{2}(1-q/2)}$

# Oracle Rate

**Claim:** Optimal oracle threshold is to choose all coordinates up to

$$v_i^2 > \frac{C}{n}.$$

compare to $\sqrt{2 \ln p / n}$

**Claim:** Let $\hat{\mathbf{v}}_{\text{oracle}}$ denote the PCA estimator using coordinates chosen by oracle with optimal threshold. Then,

$$\sup_{\mathbf{v} \in L_q(C)} \|\hat{\mathbf{v}}_{\text{oracle}} - \mathbf{v}\|^2 = O(n^{(1-q/2)})$$

compare to DT rate of $n^{-\frac{1}{2}(1-q/2)}$

Question: Can we close this gap ?

# Sparse PCA

Key Idea: 2 step procedure.

# Sparse PCA

Key Idea: 2 step procedure.

**Step 1:** Compute diagonal thresholding as initial estimator for eigenvector.

# Sparse PCA

Key Idea: 2 step procedure.

**Step 1:** Compute diagonal thresholding as initial estimator for eigenvector.

**Step 2:** Find additional coordinates that are highly correlated with this eigenvector.

# Sparse PCA

Key Idea: 2 step procedure.

**Step 1:** Compute diagonal thresholding as initial estimator for eigenvector.

**Step 2:** Find additional coordinates that are highly correlated with this eigenvector.

**Theorem:** Under suitable conditions,

$$\mathbb{E}\left[\|\mathbf{v}-\hat{\mathbf{v}}\|^2\right] \leq C\left(\frac{\log p}{n}\right)^{1-q/2}$$

# Sparse Covariance Estimation

[Bickel and Levina 2008]

Class of approximately sparse matrices:

$$\mathcal{U}(q, c_0(p), M) = \left\{ \Sigma \geq 0 \,|\, \sigma_{ii} < M, \max_i \sum_j |\sigma_{ij}|^q \leq c_0(p) \right\}$$

# Sparse Covariance Estimation

[Bickel and Levina 2008]

Class of approximately sparse matrices:

$$\mathcal{U}(q, c_0(p), M) = \left\{ \Sigma \geq 0 \,|\, \sigma_{ii} < M, \max_i \sum_j |\sigma_{ij}|^q \leq c_0(p) \right\}$$

Each row is approximately sparse

# Sparse Covariance Estimation

[Bickel and Levina 2008]

Class of approximately sparse matrices:

$$\mathcal{U}(q, c_0(p), M) = \left\{ \Sigma \geq 0 \,|\, \sigma_{ii} < M, \max_i \sum_j |\sigma_{ij}|^q \leq c_0(p) \right\}$$

Each row is approximately sparse

$q \in [0, 1]$. If $q = 0$ matrix has many zeros provided $c_0(p) \ll p$.

# Sparse Covariance Estimation

[Bickel and Levina 2008]

Class of approximately sparse matrices:

$$\mathcal{U}(q, c_0(p), M) = \left\{ \Sigma \geq 0 \,|\, \sigma_{ii} < M, \max_i \sum_j |\sigma_{ij}|^q \leq c_0(p) \right\}$$

Each row is approximately sparse

$q \in [0,1]$. If $q = 0$ matrix has many zeros provided $c_0(p) \ll p$.

Question: Can we get accurate estimate of covariance matrix $\in \mathcal{U}$

# Sparse covariance estimation by thresholding

Very simple procedure:

# Sparse covariance estimation by thresholding

Very simple procedure:

Given $\mathbf{x}_1, \ldots, \mathbf{x}_n$ vectors in $\mathbb{R}^p$ with population covariance $\Sigma$ approximately sparse.

# Sparse covariance estimation by thresholding

Very simple procedure:

Given $\mathbf{x}_1, \ldots, \mathbf{x}_n$ vectors in $\mathbb{R}^p$ with population covariance $\Sigma$ approximately sparse.

**Step 1:** Compute sample covariance matrix.

$$S_n = \frac{1}{n} \sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

# Sparse covariance estimation by thresholding

Very simple procedure:

Given $\mathbf{x}_1, \ldots, \mathbf{x}_n$ vectors in $\mathbb{R}^p$ with population covariance $\Sigma$ approximately sparse.

**Step 1:** Compute sample covariance matrix.

$$S_n = \frac{1}{n} \sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

**Step 2:** Threshold it. Set to zero small entries.

$$\hat{\Sigma} = S_\lambda(S_n)$$

where

$$S_\lambda(t) = \begin{cases} 0 & |t| < \lambda \\ t & |t| \geq \lambda \end{cases}$$

# Sparse Covariance by Thresholding

**Theorem:** $\mathbf{x}_i$ i.i.d. from sub-Gaussian distribution with covariance $\Sigma$. Choose $\lambda = M'\sqrt{\log p/n}$ with $M'$ sufficiently large.
Then, for a wide variety of thresholding functions (hard/soft/SCAD/...) uniformly over $\mathcal{U}(q, c_0(p), M)$,

$$\|S_\lambda(S_n) - \Sigma\|_2 = O_P\left(c_0(p) \cdot \left(\frac{\log p}{n}\right)^{1-q/2}\right)$$

# Sparse Covariance by Thresholding

**Theorem:** $\mathbf{x}_i$ i.i.d. from sub-Gaussian distribution with covariance $\Sigma$. Choose $\lambda = M'\sqrt{\log p/n}$ with $M'$ sufficiently large.
Then, for a wide variety of thresholding functions (hard/soft/SCAD/...) uniformly over $\mathcal{U}(q, c_0(p), M)$,

$$\|S_\lambda(S_n) - \Sigma\|_2 = O_P\left(c_0(p) \cdot \left(\frac{\log p}{n}\right)^{1-q/2}\right)$$

Note different and slower rate from sparse eigenvector estimation $1-q$.

# Proof

**Bias-Variance Decomposition**

$$\|S_\lambda(S_n) - \Sigma\|_2 \leq \|S_\lambda(\Sigma) - \Sigma\|_2 + \|S_\lambda(S_n) - S_\lambda(\Sigma)\|_2$$

# Proof

**Bias-Variance Decomposition**

$$\|S_\lambda(S_n) - \Sigma\|_2 \leq \|S_\lambda(\Sigma) - \Sigma\|_2 + \|S_\lambda(S_n) - S_\lambda(\Sigma)\|_2$$

Bound Each term separately

# Bounding the Bias Term:

*Gershgorin Theorem:* Let $A$ be symmetric matrix, then

$$\|A\|_2 \leq \max_i \sum_j |A_{ij}|$$

# Bounding the Bias Term:

*Gershgorin Theorem:* Let $A$ be symmetric matrix, then

$$\|A\|_2 \leq \max_i \sum_j |A_{ij}|$$

Now for hard thresholding

$$\sum_j |S_\lambda(\sigma_{ij}) - \sigma_{ij}| = \sum_j |\sigma_{ij}| \mathbf{1}(|\sigma_{ij}| < \lambda)$$

# Bounding the Bias Term:

*Gershgorin Theorem:* Let $A$ be symmetric matrix, then

$$\|A\|_2 \leq \max_i \sum_j |A_{ij}|$$

Now for hard thresholding

$$\sum_j |S_\lambda(\sigma_{ij}) - \sigma_{ij}| = \sum_j |\sigma_{ij}| \mathbf{1}(|\sigma_{ij}| < \lambda)$$

Write sum,

$$\sum_j |\sigma_{ij}|^{1-q} |\sigma_{ij}|^q \mathbf{1}(|\sigma_{ij}| < \lambda) \leq \lambda^{1-q} c_0(p)$$

# Bounding the Variance Term

More complicated. Consider each term separately

$$|S_\lambda(S_n(i,j)) - S_\lambda(\Sigma_{ij})| \leq \begin{cases} S_n(i,j) & |S_n(i,j)| > \lambda, |\Sigma_{ij}| < \lambda \\ \Sigma_{ij} & |S_n(i,j)| < \lambda, |\Sigma_{ij}| > \lambda \\ S_n(i,j) - \Sigma_{ij} & |S_n(i,j)| > \lambda, |\Sigma_{ij}| > \lambda \\ 0 & \text{both terms smaller than } \lambda \end{cases}$$

Use sub-Gaussian assumption
$S_n(i,j)$ close to $\Sigma_{ij}$, deviation at most $C\sqrt{\log p/n}$.
**Key result:**

$$\|S_\lambda(S_n) - S_\lambda(\Sigma)\|_2 = O_P\left(c_0(p)\lambda^{-q}\sqrt{\frac{\log p}{n}} + c_0(p)\lambda^{1-q}\right)$$

Optimal $\lambda$ that minimizes overall error: $\lambda = M'\sqrt{\log p/n}$

# The Last Slide

No matter what, have a last slide that everyone will remember.

# The Last Slide

No matter what, have a last slide that everyone will remember.

# The Last Slide

No matter what, have a last slide that everyone will remember.


*Thank you*