On Sparse Principal Components and Sparse Covariance Estimation in High Dimensions

Boaz Nadler

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

Based on joint works with
Iain Johnstone (Stanford), Debashis Paul (UC-Davis), Aharon Birnbaum

Robert Krauthgamer (Weizmann), Danny Vilenchik (Weizmann)

John Goes, Gilad Lerman (Minnesota)

Apr 2017

Talk Outline

1. Covariance Matrices and PCA
2. Sparse PCA, ℓ_q sparsity
3. Sparse PCA, ℓ_0 sparsity
4. Sparse covariance estimation with heavy tailed data

Intro

p dimensional random variable $X \in \mathbb{R}^p$

Observe $\mathbf{x}_1, \ldots, \mathbf{x}_n$: n i.i.d. realizations of X

Intro

p dimensional random variable $X \in \mathbb{R}^p$

Observe $\mathbf{x}_1, \ldots, \mathbf{x}_n$: n i.i.d. realizations of X

In principle, X fully characterized by its density $f(x)$

Intro

p dimensional random variable $X \in \mathbb{R}^p$

Observe $\mathbf{x}_1, \ldots, \mathbf{x}_n$: n i.i.d. realizations of X

In principle, X fully characterized by its density $f(x)$

but

Curse of Dimensionality:

accurate non-parametric estimate of f requires $n \propto \exp(p)$

Low order moments

Luckily, many statistical tasks need only low order moments of X.

Mean:

$$\mu = \mathbb{E}[\mathbf{x}]$$

Low order moments

Luckily, many statistical tasks need only low order moments of X.

Mean:

$$\mu = \mathbb{E}[\mathbf{x}]$$

Covariance

$$\Sigma = \mathbb{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T]$$

Low order moments

Luckily, many statistical tasks need only low order moments of X.

Mean:

$$\mu = \mathbb{E}[\mathbf{x}]$$

Covariance

$$\Sigma = \mathbb{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T]$$

Principal Components

leading eigenvalues/vectors $(\lambda_j, \mathbf{v}_j)$ of Σ

Low order moments

Luckily, many statistical tasks need only low order moments of X.

Mean:

$$\mu = \mathbb{E}[\mathbf{x}]$$

Covariance

$$\Sigma = \mathbb{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T]$$

Principal Components

leading eigenvalues/vectors $(\lambda_j, \mathbf{v}_j)$ of Σ

examples: dimension reduction, denoising, regression, classification etc

Sample / Empirical Estimates

sample mean:

$$\hat{\mu} = \bar{\mathbf{x}} = \frac{1}{n} \sum_i \mathbf{x}_i$$

Sample / Empirical Estimates

sample mean:

$$\hat{\mu} = \bar{\mathbf{x}} = \frac{1}{n} \sum_i \mathbf{x}_i$$

sample covariance matrix:

$$\hat{\Sigma} = \frac{1}{n-1} \sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Sample PCA: eigen-decomposition of $\hat{\Sigma}$

$$\hat{\Sigma} = \sum_i \ell_i \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T$$

Sample / Empirical Estimates

sample mean:

$$\hat{\mu} = \bar{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i$$

sample covariance matrix:

$$\hat{\Sigma} = \frac{1}{n-1}\sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Sample PCA: eigen-decomposition of $\hat{\Sigma}$

$$\hat{\Sigma} = \sum_i \ell_i \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T$$

Use $\hat{\mathbf{v}}_i$ as estimate of i-th principal component $\mathbf{v}_i$

The good old days

Datasets had ”small p - large n”.

The good old days

Datasets had "small p - large n".

Asymptotic analysis: dimension p fixed, sample size $n \to \infty$,

under mild conditions on X, asymptotic consistency of $\hat{\mu}, \hat{\Sigma}$ to their population counterparts.

Similarly, sample PCA is *asymptotically consistent*:

$$\hat{\Sigma} \to \Sigma \quad \text{and for all } \lambda_i \text{ with multiplicity one, } \hat{\mathbf{v}}_i \to \mathbf{v}_i$$

The good old days

Datasets had "small p - large n".

Asymptotic analysis: dimension p fixed, sample size $n \to \infty$,

under mild conditions on X, asymptotic consistency of $\hat{\mu}, \hat{\Sigma}$ to their population counterparts.

Similarly, sample PCA is *asymptotically consistent*:

$$\hat{\Sigma} \to \Sigma \quad \text{and for all } \lambda_i \text{ with multiplicity one, } \hat{\mathbf{v}}_i \to \mathbf{v}_i$$

However in high dimensions, as $p, n \to \infty$ with $p/n \to c > 1$,

$$\|\hat{\mu} - \mu\| = O_p(p/n), \; \|\hat{\Sigma} - \Sigma\| \geq \lambda_{\min}(\Sigma)$$

sample PCA is *inconsistent*.

[Johnstone & Lu, 09']

Inconsistency of Sample PCA

Consider $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = diag(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0) + \sigma^2 \mathbf{I}_p$

Spiked Covariance Model with k spikes

Inconsistency of Sample PCA

Consider $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = diag(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0) + \sigma^2 \mathbf{I}_p$

Spiked Covariance Model with k spikes

As $p, n \to \infty$ with $p/n \to c$,

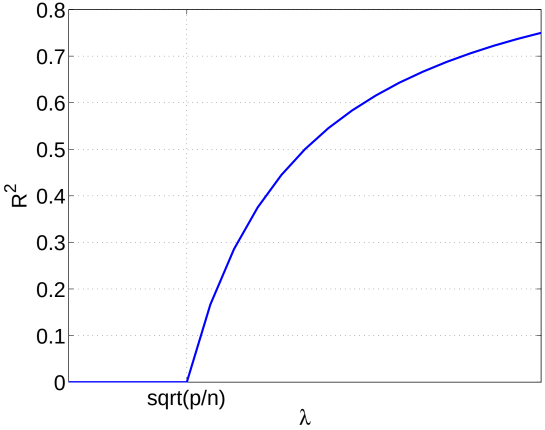
$$R_i^2 = |\langle \hat{\mathbf{v}}_i, \mathbf{v}_i \rangle|^2 \to \begin{cases} 0 & \lambda_i < \sigma^2 \sqrt{p/n} \\ \frac{\frac{\lambda_i^2}{c\sigma^2} - \sigma^2}{\frac{\lambda_i^2}{c\sigma^2} + \lambda_i} & \lambda_i > \sigma^2 \sqrt{p/n} \end{cases}$$

[statistical mechanics literature 90's]
[Paul 07', Nadler 08']

Key point:

$$R^2 = 1 - \frac{\sigma^2}{\lambda} \frac{p}{n} + \ldots$$

(Sparse)-PCA



Key Question:

Can one do better under sparsity assumptions ?

Sparsity and High-Dimensions

Key Question:

Can one do better under sparsity assumptions ?

[Donoho & Johnstone 94', others]

For estimation of μ - well studied sparse normal means problem

Sparsity and High-Dimensions

Key Question:

Can one do better under sparsity assumptions ?

[Donoho & Johnstone 94', others]

For estimation of μ - well studied sparse normal means problem

[Bickel & Levina, El-Karoui, Cai & Zhou, etc]

Sparse covariance estimation by thresholding, minimax lower bounds, works for sub-Gaussian r.v.'s

Sparsity and High-Dimensions

Key Question:

Can one do better under sparsity assumptions ?

[Donoho & Johnstone 94', others]

For estimation of μ - well studied sparse normal means problem

[Bickel & Levina, El-Karoui, Cai & Zhou, etc]

Sparse covariance estimation by thresholding, minimax lower bounds, works for sub-Gaussian r.v.'s

In this talk:

- Estimation of sparse PCA
- Sparse covariance estimation under heavy tails

The Sparse PCA problem

Given $\{\mathbf{x}_i\}_{i=1}^n$ iid with population covariance

$$\Sigma = diag(\lambda_1, \dots, \lambda_k, 0, \dots, 0) + \sigma^2 \mathbf{I}_p$$

Assume k leading eigenvectors $\mathbf{v}_i$ are *sparse*:

How well can we estimate them ?

Sparse PCA

Two settings:

a) Approximate sparsity: for $q \in (0, 2)$,

$$\|\mathbf{v}\|_2 = 1 \text{ and } \mathbf{v} \in \ell_q(C) = \{\mathbf{z} \in \mathbb{R}^p \mid \|z\|_q < C\}$$

b) Exact L_0 sparsity, $\|\mathbf{v}\|_0 = k$.

Sparse-PCA: Theoretical Questions

- Minimax Rates of Estimation ?

Sparse-PCA: Theoretical Questions

- Minimax Rates of Estimation ?
- (Computationally Efficient) Methods that achieve those ?

Sparse-PCA: Theoretical Questions

- Minimax Rates of Estimation ?
- (Computationally Efficient) Methods that achieve those ?
- What happens when $\mathbf{v} \in \ell_0$?

Sparse PCA

Many works (algorithms, analysis, optimizations) on sparse PCA.

Sparse PCA

Many works (algorithms, analysis, optimizations) on sparse PCA.

Diagonal Thresholding: [Johnstone & Lu, JASA 09]:

- Compute *only* diagonal entries of covariance matrix $\hat{\Sigma}_{ii}$.

- Variable selection by thresholding

$$I = \{i \,|\, \hat{\Sigma}_{ii} > t(\alpha, p, n)\}$$

- Compute $\hat{\Sigma}|_I$ and its leading eigenvectors via PCA.

Sparse PCA

Many works (algorithms, analysis, optimizations) on sparse PCA.

Diagonal Thresholding: [Johnstone & Lu, JASA 09]:

- Compute *only* diagonal entries of covariance matrix $\hat{\Sigma}_{ii}$.

- Variable selection by thresholding

$$I = \{i \,|\, \hat{\Sigma}_{ii} > t(\alpha, p, n)\}$$

- Compute $\hat{\Sigma}|_I$ and its leading eigenvectors via PCA.

Algorithm extremely fast $O(pn^2)$. Is it (rate) optimal ?

Minimax Rate for Sparse PCA

[with A. Birnbaum, I. Johnstone and D. Paul]
[Annals of Statistics, 2013]

Theorem: If $\mathbf{v}$ is *sparse*, then

$$\min_{\hat{\mathbf{v}}} \max_{\mathbf{v} \in \ell_q(C)} \mathbb{E}[\|\hat{\mathbf{v}} - \mathbf{v}\|^2] \geq C(\lambda, q) \left(\frac{\ln p}{n} \right)^{1-q/2}$$

Minimax Rate for Sparse PCA

[with A. Birnbaum, I. Johnstone and D. Paul]
[Annals of Statistics, 2013]

Theorem: If $\mathbf{v}$ is *sparse*, then

$$\min_{\hat{\mathbf{v}}} \max_{\mathbf{v}\in\ell_q(C)} \mathbb{E}[\|\hat{\mathbf{v}}-\mathbf{v}\|^2] \geq C(\lambda,q)\left(\frac{\ln p}{n}\right)^{1-q/2}$$

Theorem: Diagonal thresholding is *not* rate optimal.

$$\max_{\mathbf{v}\in\ell_q(C)} \mathbb{E}[\|\hat{\mathbf{v}}_{DT}-\mathbf{v}\|^2] \geq C(\lambda,q)\left(\frac{1}{n}\right)^{\frac{1}{2}(1-q/2)}$$

[related work by Z. Ma and by Vu and Lei]

Some Intuition

[N. discussion in JASA 09']

Diagonal Thresholding:

$$\hat{\Sigma}_{ii}/\sigma^2 \geq 1 + C\sqrt{\ln p/n}$$

threshold set to avoid too many false detections.

Some Intuition

[N. discussion in JASA 09']

Diagonal Thresholding:

$$\hat{\Sigma}_{ii}/\sigma^2 \geq 1 + C\sqrt{\ln p/n}$$

threshold set to avoid too many false detections.

Detect only signal coordinates $v_i = O\left((\ln p/n)^{1/4}\right)$

Some Intuition

[N. discussion in JASA 09']

Diagonal Thresholding:

$$\hat{\Sigma}_{ii}/\sigma^2 \geq 1 + C\sqrt{\ln p/n}$$

threshold set to avoid too many false detections.

Detect only signal coordinates $v_i = O\left((\ln p/n)^{1/4}\right)$

For minimax: choose all coordinates $v_i \geq O\left((\ln p/n)^{1/2}\right)$

Need to look at *off-diagonal* entries of covariance matrix.

2-Step Sparse-PCA

Given $p \times p$ sample covariance matrix

1. Run Diagonal Thresholding

$$I = \{i \,|\, S_{ii} > t(\alpha, p, n)\}$$

2. Eigendecomposition of $S|_I$

$$S|_I = \sum_j \ell_j \mathbf{w}_j \mathbf{w}_j^T$$

3. Keep only m significant eigenvalues.
4. Find coordinates with high covariance to eigenvector

$$\tilde{I} = \{i \,|\, |\mathbf{e}_i^T S \mathbf{w}_j| > t'(\alpha, p, n)\}$$

5. Eigendecomposition of S on variable set $I \cup \tilde{I}$.

Sparse PCA

Theorem: For sufficiently strong signal, above computationally efficient 2-step estimator achieves the (lower bound on) minimax rate.

Z. Ma - different (iterative) estimator also achieves same rates.

Theorem: For sufficiently strong signal, above computationally efficient 2-step estimator achieves the (lower bound on) minimax rate.

Z. Ma - different (iterative) estimator also achieves same rates.

Comparison with Sparse Covariance Estimation: Under similar sparsity model, with $q \in (0, 1)$

[Bickel and Levina, Cai & at.]

$$\min\max \mathbb{E}[\|\hat{\Sigma} - \Sigma\|^2] \propto \left(\frac{\ln p}{n}\right)^{1-q}$$

Sparse PCA and Sparse Covariance Estimation are *different* problems

L_0 sparse PCA

What happens if $\mathbf{v} \in \ell_0$?

Typical problems with ℓ_0 norm are NP-Hard...

L_0 sparse PCA

[Amini and Wainwright, AoS 09]

Consider the 'hardest' case in $\ell_0(k)$, (single spike)

$$\mathbf{v} = \frac{1}{\sqrt{k}}(1, 0, \ldots, -1, 0, \ldots, 1, \ldots, 0)$$

Information limit: As $n, p \to \infty$ no recovery possible unless

$$n \geq Ck\ln(p)$$

For recovery by diagonal thresholding, as $n, p \to \infty$

$$n \geq Ck^2\ln(p)$$

Question: computationally efficient method that closes gap ?

SDP for sparse-PCA

[d'Aspremont et. al., Bach et. al.]

Semi-Definite formulation (relaxation) for Sparse PCA.

$$\max \, Trace(\hat{\Sigma} X)$$

subject to

a) $Trace(X) = 1$,

b) $X \in \mathcal{S}_+^p = \{X \in \mathbb{R}^{p \times p} : X = X^T, X \succeq 0\}$

c) Sparsity: $\|X\|_1 = \sum_{i,j} |X_{ij}| \leq k$.

SDP for sparse-PCA

[d'Aspremont et. al., Bach et. al.]

Semi-Definite formulation (relaxation) for Sparse PCA.

$$\max \mathit{Trace}(\hat{\Sigma} X)$$

subject to

a) $\mathit{Trace}(X) = 1$,

b) $X \in \mathcal{S}_+^p = \{X \in \mathbb{R}^{p \times p} : X = X^T, X \succeq 0\}$

c) Sparsity: $\|X\|_1 = \sum_{i,j} |X_{ij}| \leq k$.

Theorem:[Amini & Wainwright] *If* SDP has rank one solution, then SDP is statistically optimal, able to recover support with

$$n > C' k \ln p$$

Result seems to close gap between information and computation

Does SDP really solve L_0 Sparse PCA ?

[with D. Vilenchik and R. Krauthgamer, AoS 15']

Questions:

- Is SDP solution indeed rank one up to information limit ?
- If it is close to rank one (say $\lambda_1(X) = 0.99$), what is relation between leading eigenvector and true spike ?

Does SDP really solve L_0 Sparse PCA ?

[with D. Vilenchik and R. Krauthgamer, AoS 15']

Questions:

- Is SDP solution indeed rank one up to information limit ?
- If it is close to rank one (say $\lambda_1(X) = 0.99$), what is relation between leading eigenvector and true spike ?

[Berthet & Rigollet, 2013]

sparse-PCA $\sim$ hidden clique:

If $\exists$ polynomial algorithm to detect spike of sparsity $k \gg \sqrt{n}$ then can detect in polynomial time hidden clique of size $r \ll \sqrt{n}$ in random graph $G(n, 1/2)$.

hidden clique believed to be computationally hard problem

Does SDP really solve L_0 Sparse PCA ?

Challenge: No closed form expression for SDP solution.

Does SDP really solve L_0 Sparse PCA ?

Challenge: No closed form expression for SDP solution.

Theorem 1: Let $p, n, k \to \infty$ with $p/n \to c > 1$, $k = o(n)$ but $k \geq p/\sqrt{n}$, then if $\lambda < \sqrt{c}$

$$\frac{p}{n} \leq SDP(X_{opt}) \leq (1 + \sqrt{\frac{p}{n}})^2$$

Remark: If $p/n \gg 1$ lower and upper bounds are relatively close.

Does SDP really solve L_0 Sparse PCA ?

Challenge: No closed form expression for SDP solution.

Theorem 1: Let $p, n, k \to \infty$ with $p/n \to c > 1$, $k = o(n)$ but $k \geq p/\sqrt{n}$, then if $\lambda < \sqrt{c}$

$$\frac{p}{n} \leq SDP(X_{opt}) \leq (1 + \sqrt{\frac{p}{n}})^2$$

Remark: If $p/n \gg 1$ lower and upper bounds are relatively close.

Theorem 2: Let $p, n, k \to \infty$, with $p/n \to c > 10$, $\lambda < 1$ and $p/\sqrt{n} \leq k \leq Cp/(\ln p)^2$. If X is a rank-one feasible matrix, then

$$SDP(X) \leq \frac{3}{5}\frac{p}{n}.$$

Corollary: Exist (p, n, k) where SDP solution is *not* rank one.

Does SDP really solve L_0 Sparse PCA ?

Suppose X is almost rank one,
largest eigenvalue $\lambda_1 = \lambda_1(X)$, corresponding eigenvector $\mathbf{w}_1$.

Theorem 3 If signal strength < 1, as $p, n, k \to \infty$

$$|\langle \mathbf{w_1}, \mathbf{v} \rangle|^2 \leq \frac{O(1)}{\lambda_1}\sqrt{\frac{n}{p}}$$

Does SDP really solve L_0 Sparse PCA ?

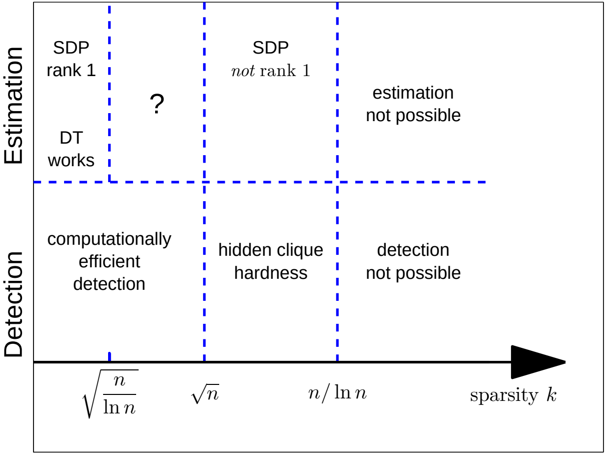
Suppose X is almost rank one,
largest eigenvalue $\lambda_1 = \lambda_1(X)$, corresponding eigenvector $\mathbf{w}_1$.

Theorem 3 If signal strength < 1, as $p, n, k \to \infty$

$$|\langle \mathbf{w_1}, \mathbf{v} \rangle|^2 \leq \frac{O(1)}{\lambda_1} \sqrt{\frac{n}{p}}$$

Corollary: if $p/n \gg 1$, largest eigenvector weakly related to sparse spike $\mathbf{v}$

L_0 sparsity

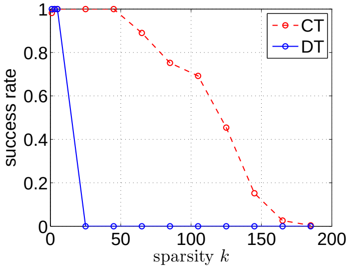

Estimation
Detection
SDP
rank 1
DT
works
?
SDP
not rank 1
estimation
not possible
computationally
efficient
detection
hidden clique
hardness
detection
not possible
$\sqrt{\frac{n}{\ln n}}$
$\sqrt{n}$
$n/\ln n$
sparsity k

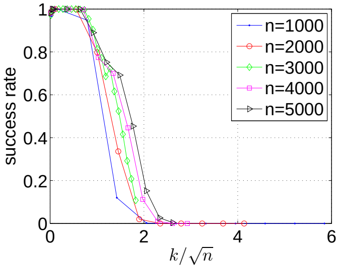
Covariance Thresholding for L_0 sparsity

Motivated by Bickel and Levina:

- compute sample covariance matrix $\hat{\Sigma}$
- threshold it at suitable threshold
- compute leading eigenvectors
- possibly threshold them.

Covariance Thresholding for L_0 sparsity

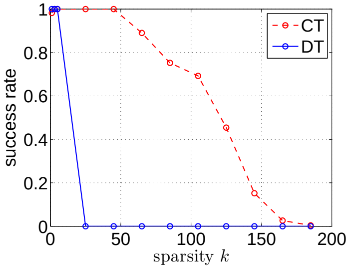


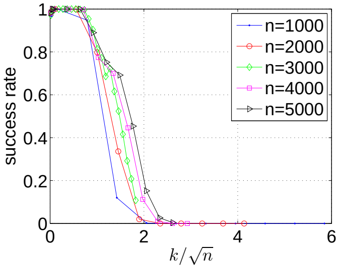


[Deshpande & Montanari]

Scaling is $k = O(\sqrt{n})$.

Covariance Thresholding for L_0 sparsity





[Deshpande & Montanari]

Scaling is $k = O(\sqrt{n})$.

Conjecture:

No computationally efficient method to recover L_0 spike for sparsity levels $k \gg \sqrt{n}$

Sparse Covariance Estimation

[Bickel and Levina, 08']

Let $\mathcal{U}(q, s_p, M, s_{\max})$ be the class of row/column s_p-sparse covariance matrices with sparsity parameter $q \in [0, 1)$:

$$\mathcal{U}(q, s_p, M, s_{\max}) := \left\{ S : \sigma_{ii} \leq M, \sum_{j=1}^{p} |\sigma_{ij}|^q \leq s_p, \|S\| \leq s_{\max} \right\}.$$

Sparse Covariance Estimation

[Bickel and Levina, 08']

Let $\mathcal{U}(q, s_p, M, s_{\max})$ be the class of row/column s_p-sparse covariance matrices with sparsity parameter $q \in [0, 1)$:

$$\mathcal{U}(q, s_p, M, s_{\max}) := \left\{ S : \sigma_{ii} \leq M, \sum_{j=1}^{p} |\sigma_{ij}|^q \leq s_p, \|S\| \leq s_{\max} \right\}.$$

X sub-Gaussian r.v. with mean zero, covariance $\Sigma \in \mathcal{U}$. Then, given n i.i.d. samples, thresholding $\hat{\Sigma}$ at $t = C\sqrt{\log p/n}$ gives

$$\|\tau_t(\hat{\Sigma}) - \Sigma\| = O_P(s_p(\log p/n)^{(1-q)/2}$$

Outlier/Heavy Tail breakdown of sample covariance

Key reason why thresholding works is following lemma

Lemma: Assume $B \in \mathcal{U}(q, s_p, M, s_{\max})$. Let A be close to B, s.t. $max_{i,j}|A_{ij} - B_{ij}| < C\sqrt{\log p/n}$. Then, for any $t = K\sqrt{\log p/n}$ with $K > C$, there is $C_2 = C_2(C, K, q)$ s.t.

$$\|\tau_t(A) - B\| \leq C_2 s_p (\log p/n)^{(1-q)/2}$$

Outlier/Heavy Tail breakdown of sample covariance

Key reason why thresholding works is following lemma

Lemma: Assume $B \in \mathcal{U}(q, s_p, M, s_{\max})$. Let A be close to B, s.t. $max_{i,j}|A_{ij} - B_{ij}| < C\sqrt{\log p/n}$. Then, for any $t = K\sqrt{\log p/n}$ with $K > C$, there is $C_2 = C_2(C, K, q)$ s.t.

$$\|\tau_t(A) - B\| \leq C_2 s_p (\log p/n)^{(1-q)/2}$$

bound on individual entries $\rightarrow$ global bound on spectral norm

Bickel & Levina: if X sub-Gaussian, then
$\max_{ij} |\hat{\Sigma}_{ij} - \Sigma_{ij}| < C\sqrt{\log p/n}$

Outlier/Heavy Tail breakdown of sample covariance

[with J. Goes and G. Lerman]

Problem: For heavy-tailed data the sample covariance may be a poor entry-wise estimator of Σ

Outlier/Heavy Tail breakdown of sample covariance

[with J. Goes and G. Lerman]

Problem: For heavy-tailed data the sample covariance may be a poor entry-wise estimator of Σ

Thresholding it will be a poor estimator of Σ in spectral norm.

Outlier/Heavy Tail breakdown of sample covariance

[with J. Goes and G. Lerman]

Problem: For heavy-tailed data the sample covariance may be a poor entry-wise estimator of Σ

Thresholding it will be a poor estimator of Σ in spectral norm.

Key Questions:

- Lower bounds - how well can one estimate a sparse covariance under heavy-tailed distributions.

Outlier/Heavy Tail breakdown of sample covariance

[with J. Goes and G. Lerman]

Problem: For heavy-tailed data the sample covariance may be a poor entry-wise estimator of Σ

Thresholding it will be a poor estimator of Σ in spectral norm.

Key Questions:

- Lower bounds - how well can one estimate a sparse covariance under heavy-tailed distributions.

- Computationally efficient rate optimal estimator ?

Outlier/Heavy Tail breakdown of sample covariance

[with J. Goes and G. Lerman]

Problem: For heavy-tailed data the sample covariance may be a poor entry-wise estimator of Σ

Thresholding it will be a poor estimator of Σ in spectral norm.

Key Questions:

- Lower bounds - how well can one estimate a sparse covariance under heavy-tailed distributions.

- Computationally efficient rate optimal estimator ?

Answer these questions for *elliptical* distributions

(Generalized) Elliptical Distribution

[Frahm 04']

Definition: X follows a generalized elliptical distribution with positive definite $p \times p$ shape matrix S_p if

$$X = US_p^{1/2}\xi$$

where $\xi \sim N(0, I_p)$ and $U \in \mathbb{R}$ is either stochastic or deterministic but $U \neq 0$.

(Generalized) Elliptical Distribution

[Frahm 04']

Definition: X follows a generalized elliptical distribution with positive definite $p \times p$ shape matrix S_p if

$$X = US_p^{1/2}\xi$$

where $\xi \sim N(0, I_p)$ and $U \in \mathbb{R}$ is either stochastic or deterministic but $U \neq 0$.

Common model in multiple applications involving heavy tails.

(Generalized) Elliptical Distribution

[Frahm 04']

Definition: X follows a generalized elliptical distribution with positive definite $p \times p$ shape matrix S_p if

$$X = U S_p^{1/2} \xi$$

where $\xi \sim N(0, I_p)$ and $U \in \mathbb{R}$ is either stochastic or deterministic but $U \neq 0$.

Common model in multiple applications involving heavy tails.

For unique scaling of shape matrix we assume $tr(S_p) = p$.

Elliptical Distribution, sparse shape matrix

If distribution is not too heavy tailed, then population covariance of X exists and $\Sigma = cS_p$.

Elliptical Distribution, sparse shape matrix

If distribution is not too heavy tailed, then population covariance of X exists and $\Sigma = cS_p$.

Question: Given n i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from potentially heavy tailed elliptical distribution, accurately estimate its approximately sparse shape matrix S_p in a computationally efficient way.

Elliptical Distribution, sparse shape matrix

If distribution is not too heavy tailed, then population covariance of X exists and $\Sigma = cS_p$.

Question: Given n i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from potentially heavy tailed elliptical distribution, accurately estimate its approximately sparse shape matrix S_p in a computationally efficient way.

Key to solution: as in Bickel and Levina, need to construct some matrix $\hat{S}_p$ such that $\max_{ij} |\hat{S}_p - S_p| < C\sqrt{\log p/n}$

Tyler's M-estimator

[Tyler, 87']

Solution to:

$$\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i}=\Sigma,$$

normalized so that $Tr(\Sigma)=1$.

Tyler's M-estimator

[Tyler, 87']

Solution to:

$$\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i} = \Sigma,$$

normalized so that $Tr(\Sigma) = 1$.

Solution can be obtained as limit of following iterations

$$\hat{\Sigma}_{k+1} = \sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}_k^{-1}\mathbf{x}_i} \Big/ Tr\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}_k^{-1}\mathbf{x}_i}\right).$$

Tyler's M-estimator

[Tyler, 87']

Solution to:

$$\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i}=\Sigma,$$

normalized so that $Tr(\Sigma)=1$.

Solution can be obtained as limit of following iterations

$$\hat{\Sigma}_{k+1}=\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}_k^{-1}\mathbf{x}_i}\Bigg/ Tr\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}_k^{-1}\mathbf{x}_i}\right).$$

Intuition: iterative scaling by Mahalanobis distance

Tyler's M-estimator

[Tyler, 87']

Solution to:

$$\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i}=\Sigma,$$

normalized so that $Tr(\Sigma)=1$.

Solution can be obtained as limit of following iterations

$$\hat{\Sigma}_{k+1}=\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}_k^{-1}\mathbf{x}_i}\Big/ Tr\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}_k^{-1}\mathbf{x}_i}\right).$$

Intuition: iterative scaling by Mahalanobis distance

Robust estimate of S_p, consistent for p fixed, $n\to\infty$.

Tyler's M-estimator

[Tyler, 87']

Solution to:

$$\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i}=\Sigma,$$

normalized so that $Tr(\Sigma)=1$.

Solution can be obtained as limit of following iterations

$$\hat{\Sigma}_{k+1}=\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}_k^{-1}\mathbf{x}_i}\Big/ Tr\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}_k^{-1}\mathbf{x}_i}\right).$$

Intuition: iterative scaling by Mahalanobis distance

Robust estimate of S_p, consistent for p fixed, $n\to\infty$.

Good candidate to threshold but not defined when $p>n$!

Regularized Tyler's M-estimator

[Abramovich & Spencer 07', Wiesel 12', etc.]

Solution to fixed point equation

$$\hat{\Sigma}(\alpha) = \frac{1}{1+\alpha}\frac{p}{n}\sum_i \frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}(\alpha)^{-1}\mathbf{x}_i} + \frac{\alpha}{1+\alpha}\boldsymbol{I}$$

where $\alpha > 0$ is regularization parameter.

[Sun, Babu & Palomar 14']

If $\alpha > \max(0, p/n - 1)$ then regularized-TME exists and is limit of following iterations

$$\hat{\boldsymbol{\Sigma}}_{k+1}(\alpha) = \frac{1}{1+\alpha}\frac{p}{n}\sum_i \frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}_k(\alpha)^{-1}\mathbf{x}_i} + \frac{\alpha}{1+\alpha}\boldsymbol{I}.$$

Regularized Tyler's M-estimator

[Abramovich & Spencer 07', Wiesel 12', etc.]

Solution to fixed point equation

$$\hat{\Sigma}(\alpha) = \frac{1}{1+\alpha}\frac{p}{n}\sum_i \frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}(\alpha)^{-1}\mathbf{x}_i} + \frac{\alpha}{1+\alpha}\boldsymbol{I}$$

where $\alpha > 0$ is regularization parameter.

[Sun, Babu & Palomar 14']

If $\alpha > \max(0, p/n - 1)$ then regularized-TME exists and is limit of following iterations

$$\hat{\boldsymbol{\Sigma}}_{k+1}(\alpha) = \frac{1}{1+\alpha}\frac{p}{n}\sum_i \frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\hat{\Sigma}_k(\alpha)^{-1}\mathbf{x}_i} + \frac{\alpha}{1+\alpha}\boldsymbol{I}.$$

Perhaps $\hat{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I}$ is good candidate to threshold as estimator of S_p.

Our Results

Consider following thresholding estimator for shape matrix:

$$\hat{S}_p = \tau_t\left(p\frac{\hat{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I}}{Tr(\hat{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I})}\right).$$

Our Results

Consider following thresholding estimator for shape matrix:

$$\hat{S}_p = \tau_t \left(p \frac{\hat{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I}}{Tr(\hat{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I})} \right).$$

Theorem: Let $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. Assume S_p is approximately sparse. Then for any $\alpha > \max(0, p/n - 1)$, for any threshold $t = M' \sqrt{\log p/n}$ with large enough M',

$$\left\| \tau_{t_n}\left(\hat{S}_p\right) - S_p \right\| = \mathcal{O}_P \left(s_p \left(\frac{\log p}{n} \right)^{(1-q)/2} \right).$$

Our Results

Consider following thresholding estimator for shape matrix:

$$\hat{S}_p = \tau_t \left(p \frac{\hat{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha} \boldsymbol{I}}{Tr(\hat{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha} \boldsymbol{I})} \right).$$

Theorem: Let $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. Assume S_p is approximately sparse. Then for any $\alpha > \max(0, p/n - 1)$, for any threshold $t = M' \sqrt{\log p/n}$ with large enough M',

$$\left\| \tau_{t_n} \left(\hat{S}_p \right) - S_p \right\| = \mathcal{O}_P \left(s_p \left(\frac{\log p}{n} \right)^{(1-q)/2} \right).$$

Remark: This is also minimax rate for sparse covariance estimation with sub-Gaussian data [Cai & Zhou]

$\rightarrow$ Our estimator is minimax rate optimal

Proof Outline

Quite involved. Relies on recent results from random matrix theory, concentration of quadratic forms, etc.

Key ideas:

1) regularized TME invariant to scaling, assume $\mathbf{x}_i \sim N(0, S_p)$.

Proof Outline

Quite involved. Relies on recent results from random matrix theory, concentration of quadratic forms, etc.

Key ideas:

1) regularized TME invariant to scaling, assume $\mathbf{x}_i \sim N(0, S_p)$.

2) Write

$$\Sigma(\alpha) = \frac{p}{n} \frac{1}{1+\alpha} \sum_i w_i \mathbf{x}_i \mathbf{x}_i^T + \frac{\alpha}{1+\alpha} \boldsymbol{I}$$

Show tight concentration of weights to uniform vector

$$\Pr(\max_i |n w_i - r| > \epsilon) < \ Cp^2 \exp(-cp\epsilon^2)$$

where r is solution of some complicated equation.

Proof Outline

Quite involved. Relies on recent results from random matrix theory, concentration of quadratic forms, etc.

Key ideas:

1) regularized TME invariant to scaling, assume $\mathbf{x}_i \sim N(0, S_p)$.

2) Write

$$\Sigma(\alpha) = \frac{p}{n}\frac{1}{1+\alpha}\sum_i w_i \mathbf{x}_i \mathbf{x}_i^T + \frac{\alpha}{1+\alpha}\boldsymbol{I}$$

Show tight concentration of weights to uniform vector

$$\Pr(\max_i |nw_i - r| > \epsilon) < \ Cp^2 \exp(-cp\epsilon^2)$$

where r is solution of some complicated equation.
This means that $\Sigma(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I}$ is close to S_p elementwise as needed for earlier proofs.

Computational Complexity

Can one compute regularized TME in polynomial time ?

Computational Complexity

Can one compute regularized TME in polynomial time ?

Define $C(X) = \|\frac{1}{n}\sum_{i=1}^{n}(\sqrt{p}\mathbf{x}_i/\|\mathbf{x}_i\|)(\sqrt{p}\mathbf{x}_i/\|\mathbf{x}_i\|)^T\|$

Computational Complexity

Can one compute regularized TME in polynomial time ?

Define $C(X) = \|\frac{1}{n}\sum_{i=1}^{n}(\sqrt{p}\mathbf{x}_i/\|\mathbf{x}_i\|)(\sqrt{p}\mathbf{x}_i/\|\mathbf{x}_i\|)^T\|$

Lemma if $1+\alpha > 5C(X)$ then regularized TME iterations converge *linearly*

$$\|\hat{\Sigma}_{k+1} - \Sigma(\alpha)\| < \frac{1}{2}\|\hat{\Sigma}_k - \Sigma(\alpha)\|$$

Each iteration $O(p^3)$ operations due to matrix inversion. For accuracy ϵ need only $O(\log(1/\epsilon))$ iterations.

Computational Complexity

Can one compute regularized TME in polynomial time ?

Define $C(X) = \|\frac{1}{n}\sum_{i=1}^{n}(\sqrt{p}\mathbf{x}_i/\|\mathbf{x}_i\|)(\sqrt{p}\mathbf{x}_i/\|\mathbf{x}_i\|)^T\|$

Lemma if $1+\alpha > 5C(X)$ then regularized TME iterations converge *linearly*

$$\|\hat{\Sigma}_{k+1} - \Sigma(\alpha)\| < \frac{1}{2}\|\hat{\Sigma}_k - \Sigma(\alpha)\|$$

Each iteration $O(p^3)$ operations due to matrix inversion. For accuracy ϵ need only $O(\log(1/\epsilon))$ iterations.

Regularized TME requires polynomial number of operations
practical: few seconds on standard PC for $p, n \approx 1000$.

Simulation Results

Took approximately sparse matrix

$$(S_p)_{ij} = (0.7^{|i-j|})$$

Three choices for U:

- $U = 1$, Gaussian data
- $U \sim$ *Laplace*, heavy tailed but all moments exist
- $U \sim$ *Cauchy*, no moments exist

Simulation Results

Took approximately sparse matrix

$$(S_p)_{ij} = (0.7^{|i-j})$$

Three choices for U:

- $U = 1$, Gaussian data
- $U \sim$ *Laplace*, heavy tailed but all moments exist
- $U \sim$ *Cauchy*, no moments exist

$p/n = \gamma = 1/2, 1$ or 2

Estimators

Compare 4 estimators:

- Scaled sample covariance $p\hat{\Sigma}/Tr(\hat{\Sigma})$
- Thresholding it
- Scaled Regularized TME $\Sigma(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I}$
- Thresholding regularized TME

Estimators

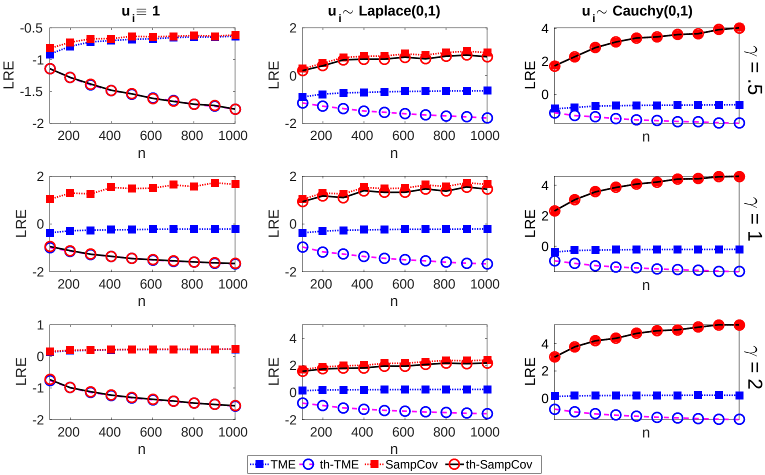
Compare 4 estimators:

- Scaled sample covariance $p\hat{\Sigma}/Tr(\hat{\Sigma})$
- Thresholding it
- Scaled Regularized TME $\Sigma(\alpha) - \frac{\alpha}{1+\alpha}\boldsymbol{I}$
- Thresholding regularized TME

Accuracy Measure: Log relative ratio

$$\text{LRE} = \log\left(\frac{\mathbb{E}[\|\hat{S}_p - S_p\|]}{\|S_p\|}\right).$$

Simulation Results



Open Questions

- Estimate optimal threshold in data-driven manner
- What if $p = n^{\beta}$ for $\beta > 1$?
- ϵ-contamination model ?

Chen,Gao, Ren [15'] proved minimax optimality for estimator based on Tukey's depth function. But NP-hard to compute.

Open Questions

- Estimate optimal threshold in data-driven manner

- What if $p = n^\beta$ for $\beta > 1$?

- ϵ-contamination model ?

Chen,Gao, Ren [15'] proved minimax optimality for estimator based on Tukey's depth function. But NP-hard to compute.

Is there computationally efficient / practical robust estimator ?

Summary

- Various contemporary applications involve 'large p – small n' data.

- Minimax Rates for Sparse - PCA with ℓ_q approximate sparsity.

- Computationally efficient algorithm achieves minimax rate in ℓ_q.

Summary

- Various contemporary applications involve 'large p – small n' data.

- Minimax Rates for Sparse - PCA with ℓ_q approximate sparsity.

- Computationally efficient algorithm achieves minimax rate in ℓ_q.

- Sparse covariance estimation for heavy tailed elliptical data

Summary

- Various contemporary applications involve 'large p – small n' data.

- Minimax Rates for Sparse - PCA with ℓ_q approximate sparsity.

- Computationally efficient algorithm achieves minimax rate in ℓ_q.

- Sparse covariance estimation for heavy tailed elliptical data

- Is there a computationally efficient method up to information limit for detection / estimation in ℓ_0 case ?

Summary

- Various contemporary applications involve 'large p – small n' data.

- Minimax Rates for Sparse - PCA with ℓ_q approximate sparsity.

- Computationally efficient algorithm achieves minimax rate in ℓ_q.

- Sparse covariance estimation for heavy tailed elliptical data

- Is there a computationally efficient method up to information limit for detection / estimation in ℓ_0 case ?

- Is there computiationally efficient method to handle arbitrary outliers ?

www.weizmann.ac.il/math/nadler

Summary

- Various contemporary applications involve 'large p – small n' data.

- Minimax Rates for Sparse - PCA with ℓ_q approximate sparsity.

- Computationally efficient algorithm achieves minimax rate in ℓ_q.

- Sparse covariance estimation for heavy tailed elliptical data

- Is there a computationally efficient method up to information limit for detection / estimation in ℓ_0 case ?

- Is there computiationally efficient method to handle arbitrary outliers ?

www.weizmann.ac.il/math/nadler

THE END / THANK YOU !