Robust Sparse Covariance Estimation
by Thresholding Tyler’s M-estimator

Boaz Nadler

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

Based on joint work with

John Goes, Gilad Lerman (Minnesota)

Feb 2018
Talk Outline

1. Brief Intro: Covariance Matrices and PCA
2. Prior Work: Sparse covariance estimation, sub-Gaussian case
3. Sparse covariance estimation with heavy tailed data
Let $X \in \mathbb{R}^p$ be a $p$ dimensional random variable

Observe $x_1, \ldots, x_n$: $n$ i.i.d. realizations of $X$
Let $X \in \mathbb{R}^p$ be a $p$ dimensional random variable

Observe $x_1, \ldots, x_n$: $n$ i.i.d. realizations of $X$

In principle, $X$ fully characterized by its density $f(x)$
Let $X \in \mathbb{R}^p$ be a $p$ dimensional random variable

Observe $x_1, \ldots, x_n$: $n$ 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:**

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

**Covariance**

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

**Principal Components** leading eigenvalues/vectors ($\lambda_j, v_j$) of $\Sigma$

examples: dimension reduction, denoising, regression, classification etc.
Low order moments

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

**Mean:**

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

**Covariance**

$$\Sigma = \mathbb{E}[(x - \mu)(x - \mu)^T]$$
Luckily, many statistical tasks need only low order moments of \( X \).

**Mean:**

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

**Covariance**

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

**Principal Components**

leading eigenvalues/vectors \((\lambda_j, v_j)\) of \( \Sigma \)
Low order moments

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

**Mean:**

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

**Covariance**

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

**Principal Components**

leading eigenvalues/vectors $(\lambda_j, v_j)$ of $\Sigma$

examples: dimension reduction, denoising, regression, classification etc
Sample / Empirical Estimates

sample mean:

\[ \hat{\mu} = \bar{x} = \frac{1}{n} \sum_i x_i \]
sample mean:

\[ \hat{\mu} = \bar{x} = \frac{1}{n} \sum_i x_i \]

sample covariance matrix:

\[ \hat{\Sigma} = \frac{1}{n-1} \sum_i (x_i - \bar{x})(x_i - \bar{x})^T \]

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

\[ \hat{\Sigma} = \sum_i \ell_i \hat{v}_i \hat{v}_i^T \]
Sample / Empirical Estimates

sample mean:

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

sample covariance matrix:

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

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".

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: $\hat{\Sigma} \to \Sigma$ and for all $\lambda_i$ with multiplicity one, $\hat{v}_i \to v_i$.

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']
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*:

$$\hat{\Sigma} \to \Sigma \quad \text{and for all } \lambda_i \text{ with multiplicity one, } \hat{v}_i \to v_i$$
The good old days

Datasets had ”small \( p \) - large \( n \”).

Asymptotic analysis: dimension \( p \) fixed, sample size \( n \rightarrow \infty \), under mild conditions on \( X \), asymptotic consistency of \( \hat{\mu}, \hat{\Sigma} \) to their population counterparts.

Similarly, sample PCA is \textit{asymptotically consistent}:

\[
\hat{\Sigma} \rightarrow \Sigma \quad \text{and for all } \lambda_i \text{ with multiplicity one, } \hat{v}_i \rightarrow v_i
\]

\textit{However} in high dimensions, as \( p, n \rightarrow \infty \) with \( p/n \rightarrow c > 1 \),

\[
\|\hat{\mu} - \mu\| = O_p(p/n), \quad \|\hat{\Sigma} - \Sigma\| \geq \lambda_{\min}(\Sigma)
\]

\textit{sample PCA is inconsistent.}

[Johnstone & Lu, 09’]
Consider $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0) + \sigma^2 I_p$

Spiked Covariance Model with $k$ spikes
Inconsistency of Sample PCA

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

Spiked Covariance Model with \( k \) spikes

As \( p, n \to \infty \) with \( p/n \to c \),

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

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

Key point:

\[
R^2 = 1 - \frac{\sigma^2}{\lambda} \frac{p}{n} + \ldots
\]
Breakdown of Classical PCA

\[ \sqrt{\frac{p}{n}} \]

\[ \lambda \]

\[ R^2 \]

\[ \sqrt{\frac{p}{n}} \]

Boaz Nadler

Robust Sparse Covariance
Key Question:
Can one do better under sparsity assumptions?
**Key Question:**

Can one do better under sparsity assumptions?

[Donoho & Johnstone 94’, others]

For estimation of $\mu$ - well studied sparse normal means problem
Key Question:

Can one do better under sparsity assumptions?

[Donoho & Johnstone 94’, others]

For estimation of $\mu$ - well studied sparse normal means problem

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

Models for sparse covariance matrices. Simple thresholding-based estimators, minimax lower bounds
Key Question:
Can one do better under sparsity assumptions?

[Donoho & Johnstone 94’, others]
For estimation of $\mu$ - well studied sparse normal means problem

[Bickel & Levina, El-Karoui, Cai & Zhou, etc]
Models for sparse covariance matrices. Simple thresholding-based estimators, minimax lower bounds

[Meinshausen & Buhlmann, Rothman et al, Cai and Liu, etc]
Sparse inverse covariance estimators.
Most prior works assumed random variable $X$ is sub-Gaussian but in various applications, data is heavy tailed.
Sparse Covariance Estimation

Most prior works assumed random variable $X$ is sub-Gaussian but in various applications, data is heavy tailed.

Key Challenge:
combine high dimension + sparsity + robustness
Sparse Covariance Estimation

Most prior works assumed random variable $X$ is sub-Gaussian

but in various applications, data is heavy tailed.

Key Challenge:
combine high dimension + sparsity + robustness

In this talk:
Sparse covariance estimation under heavy tails, specifically under an elliptical distribution
Most prior works assumed random variable $X$ is sub-Gaussian

but in various applications, data is heavy tailed.

Key Challenge:

combine high dimension + sparsity + robustness

**In this talk:**

Sparse covariance estimation under heavy tails, specifically under an elliptical distribution

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

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

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

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

Let $X$ 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

$$\|\tau_t(\hat{\Sigma}) - \Sigma\| = O_P \left( s_p (\log(p)/n)^{(1-q)/2} \right)$$
Key reason why thresholding works is following (deterministic) lemma

**Lemma:** Assume $B \in \mathcal{U}(q, s_p, M, s_{\text{max}})$. 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.

$$\|\tau_t(A) - B\| \leq C_2 s_p (\log p / n)^{(1-q)/2}$$
Key reason why thresholding works is following (deterministic) lemma

**Lemma:** Assume $B \in \mathcal{U}(q, s_p, M, s_{\text{max}})$. 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
Outlier/Heavy Tail breakdown of sample covariance

Key reason why thresholding works is following (deterministic) lemma

**Lemma:** Assume $B \in \mathcal{U}(q, s_p, M, s_{\text{max}})$. 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 w.h.p.

$$\max_{i,j} |\hat{\Sigma}_{ij} - \Sigma_{ij}| < C \sqrt{\log p/n}$$
Problem: For heavy-tailed data the sample covariance may be a poor entry-wise estimator of Σ
**Problem:** For heavy-tailed data the sample covariance may be a poor entry-wise estimator of $\Sigma$

Thresholding it will be a poor estimator of $\Sigma$ in spectral norm.
**Problem:** For heavy-tailed data the sample covariance may be a poor entry-wise estimator of $\Sigma$

Thresholding it will be a poor estimator of $\Sigma$ in spectral norm.

**Key Questions:**
- Lower bounds - how well can one estimate a sparse covariance under heavy-tailed distributions.
Problem: For heavy-tailed data the sample covariance may be a poor entry-wise estimator of $\Sigma$

Thresholding it will be a poor estimator of $\Sigma$ 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

**Problem:** For heavy-tailed data the sample covariance may be a poor entry-wise estimator of $\Sigma$

Thresholding it will be a poor estimator of $\Sigma$ 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

[Cambanis et al 81’, Frahm 04’]

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

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

where $\eta$ is uniformly distributed on unit sphere $S^{p-1}$ and $U \in \mathbb{R}$ is either stochastic or deterministic but $U \neq 0$. For generalized case $U$ may depend on $\eta$. 

(Generalized) Elliptical Distribution

[Campanis et al 81’, Frahm 04’]

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

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

where $\eta$ is uniformly distributed on unit sphere $S^{p-1}$ and $U \in \mathbb{R}$ is either stochastic or deterministic but $U \neq 0$. For generalized case $U$ may depend on $\eta$.

Common model in many applications involving heavy tails
(Generalized) Elliptical Distribution

[Campanis et al 81’, Frahm 04’]

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

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

where $\eta$ is uniformly distributed on unit sphere $\mathbb{S}^{p-1}$ and $U \in \mathbb{R}$ is either stochastic or deterministic but $U \neq 0$. For generalized case $U$ may depend on $\eta$.

Common model in many applications involving heavy tails

For unique scaling of shape matrix we assume $tr(S_p) = p$. Each variable has on average $(S_p)_{ii} = 1$. 
If distribution is not too heavy tailed, then population covariance of $X$ exists and $\Sigma = cS_p$. 
If distribution is not too heavy tailed, then population covariance of $X$ exists and $\Sigma = cS_p$.

**Question:** Given $n$ i.i.d. samples $x_1, \ldots, x_n$ from potentially heavy tailed elliptical distribution, accurately estimate its approximately sparse shape matrix $S_p$ in a computationally efficient way.
If distribution is not too heavy tailed, then population covariance of $X$ exists and $\Sigma = cS_p$.

**Question:** Given $n$ i.i.d. samples $x_1, \ldots, x_n$ 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

Solution to:

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

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

[Tyler, 87']

Intuition: iterative scaling by Mahalanobis distance
Tyler’s M-estimator

Solution to:

\[ \frac{p}{n} \sum_{i=1}^{n} \frac{x_i x_i^T}{x_i^T \Sigma^{-1} x_i} = \Sigma, \]

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

Solution can be obtained as limit of following iterations

\[ \hat{\Sigma}_{k+1} = \left( \sum_{i=1}^{n} \frac{x_i x_i^T}{x_i^T \hat{\Sigma}_{k}^{-1} x_i} \right) / Tr \left( \sum_{i=1}^{n} \frac{x_i x_i^T}{x_i^T \hat{\Sigma}_{k}^{-1} x_i} \right). \]
Tyler’s M-estimator

Solution to:

\[
\frac{p}{n} \sum_{i=1}^{n} \frac{x_i x_i^T}{x_i^T \Sigma^{-1} x_i} = \Sigma,
\]

normalized so that \( \text{Tr}(\Sigma) = 1 \).

Solution can be obtained as limit of following iterations

\[
\hat{\Sigma}_{k+1} = \sum_{i=1}^{n} \frac{x_i x_i^T}{x_i^T \hat{\Sigma}_{k}^{-1} x_i} / \text{Tr}\left( \sum_{i=1}^{n} \frac{x_i x_i^T}{x_i^T \hat{\Sigma}_{k}^{-1} x_i} \right).
\]

Intuition: iterative scaling by Mahalanobis distance
Tyler’s M-estimator is maximum likelihood estimate of shape matrix from elliptical distribution.
Tyler’s M-estimator is maximum likelihood estimate of shape matrix from elliptical distribution.

It is a robust estimate of $S_p$, consistent for $p$ fixed, $n \rightarrow \infty$. 
Tyler’s M-estimator is maximum likelihood estimate of shape matrix from elliptical distribution.

It is a robust estimate of $S_p$, consistent for $p$ fixed, $n \to \infty$. Good potential candidate to threshold but not defined when $p > n$!
Regularized Tyler’s M-estimator

Solution to fixed point equation

\[ \hat{\Sigma}(\alpha) = \frac{1}{1 + \alpha} \frac{p}{n} \sum_i x_i x_i^T \frac{1}{\hat{\Sigma}(\alpha)^{-1} x_i} + \frac{\alpha}{1 + \alpha} I \]

where \( \alpha > 0 \) is regularization parameter.
Regularized Tyler’s M-estimator

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

Solution to fixed point equation

\[ \hat{\Sigma}(\alpha) = \frac{1}{1 + \alpha} \frac{p}{n} \sum_i x_i x_i^T \frac{x_i^T \hat{\Sigma}(\alpha)^{-1} x_i}{1 + \alpha} + \frac{\alpha}{1 + \alpha} I \]

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

\[ \hat{\Sigma}_{k+1}(\alpha) = \frac{1}{1 + \alpha} \frac{p}{n} \sum_i x_i x_i^T \frac{x_i^T \hat{\Sigma}_k(\alpha)^{-1} x_i}{1 + \alpha} + \frac{\alpha}{1 + \alpha} I. \]
Our Results

Let $\tau_t(A)$ be entrywise threshold operation on $A$ at level $t$.

$$\tau_t(A)_{ij} = a_{ij} \cdot 1(|a_{ij}| > t)$$

Consider following thresholding estimator for shape matrix:

**Case I:** $p < n$, threshold Tyler’s M-estimator

$$\hat{S}_p = \tau_t \left( p\hat{\Sigma}_{TME} \right)$$
Our Results

Let $\tau_t(A)$ be entrywise threshold operation on $A$ at level $t$.

$$
\tau_t(A)_{ij} = a_{ij} \cdot 1(|a_{ij}| > t)
$$

Consider following thresholding estimator for shape matrix:

**Case I:** $p < n$, threshold Tyler’s M-estimator

$$
\hat{S}_p = \tau_t \left( p\hat{\Sigma}_{\text{TME}} \right)
$$

**Case II:** Any values of $p, n$, threshold regularized TME,

$$
\hat{S}_p = \tau_t \left( p\frac{\hat{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha} I}{Tr(\hat{\Sigma}(\alpha) - \frac{\alpha}{1+\alpha} I)} \right).
$$
For both theorems, assume $S_p$ is approximately sparse.

**Theorem 1:** Let $n, p \to \infty$ with $p/n \to \gamma \in (0, 1)$. Then for any threshold $t_n = M' \sqrt{\log(p)/n}$ with large enough $M'$,

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

**Theorem 2:** Let $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. Assume $\lambda_{\min}(S_p) > s_{\min}$. Then for any $\alpha > \max(0, p/n - 1)$, for any threshold $t_n = M' \sqrt{\log p/n}$ with large enough $M'$,

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

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

For both theorems, assume $S_p$ is approximately sparse.

**Theorem 1:** Let $n, p \to \infty$ with $p/n \to \gamma \in (0, 1)$. Then for any threshold $t_n = M' \sqrt{\log(p)/n}$ with large enough $M'$,

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

**Theorem 2:** Let $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. Assume $\lambda_{\min}(S_p) > s_{\min}$. Then for any $\alpha > \max(0, p/n - 1)$, for any threshold $t_n = M' \sqrt{\log p/n}$ with large enough $M'$,

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

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

→ 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 $x_i \sim N(0, S_p)$.
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
   \[ \Sigma(\alpha) = \frac{p}{n} \frac{1}{1 + \alpha} \sum_{i} w_i \mathbf{x}_i \mathbf{x}_i^T + \frac{\alpha}{1 + \alpha} \mathbf{I} \]

Show tight concentration of weights to uniform vector

\[ \Pr(\max_{i} |nw_i - r| > \epsilon) < Cp^2 \exp(-cpe^{-2}) \]

where $r$ is solution of some complicated equation.
**Case I, TME:** we build upon recent result of Zhang, Cheng and Singer 16’, that $r = 1$. Namely, for TME, $w_i \approx 1/n$. So with Gaussian data, $\hat{\Sigma}_{TME}$ is close to sample covariance matrix $\hat{S} = \frac{1}{n} \sum x_i x_i^T$ in operator norm,

$$\|p\hat{\Sigma}_{TME} - \hat{S}\| = O_P(\sqrt{\log(p)/n}).$$

From here previous proof follows.
Case II, Regularized TME:
Here $r = r(\alpha)$. This means that

$$\Sigma(\alpha) - \frac{\alpha}{1+\alpha} I \text{ is close to } \frac{p}{n} \frac{r}{1+\alpha} \hat{S}$$

Need to show that $r \in [r_{\text{min}}, r_{\text{max}}]$ so things don't blow up.
Computational Complexity

Can one compute regularized TME in polynomial time?
Computational Complexity

Can one compute regularized TME in polynomial time?

Define $C(X) = \left\| \frac{1}{n} \sum_{i=1}^{n} \left( \sqrt{p} \frac{x_i}{\|x_i\|} \right) \left( \sqrt{p} \frac{x_i}{\|x_i\|} \right)^T \right\|$

Each iteration $O\left( \min(n, p) \right)^3$ operations due to matrix inversion.

For accuracy $\epsilon$ need only $O\left( \log(1/\epsilon) \right)$ iterations.

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

Boaz Nadler
Robust Sparse Covariance
Can one compute regularized TME in polynomial time?

Define \( C(X) = \| \frac{1}{n} \sum_{i=1}^{n} (\sqrt{p}\frac{x_i}{\|x_i\|})(\sqrt{p}\frac{x_i}{\|x_i\|})^{T} \| \)

\( C(X) \) is data dependent quantity that can be computed for any given dataset ahead of computing the regularized TME.

For elliptical data, \( C(X) \approx (1 + \sqrt{p/n})^2 \).
Can one compute regularized TME in polynomial time?

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

$C(X)$ is data dependent quantity that can be computed for any given dataset ahead of computing the regularized TME. For elliptical data, $C(X) \approx (1 + \sqrt{p/n})^2$.

**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 $O(\min(n, p)^3)$ operations due to matrix inversion. For accuracy $\epsilon$ need only $O(\log(1/\epsilon))$ iterations.
Can one compute regularized TME in polynomial time?

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

\( C(X) \) is a data dependent quantity that can be computed for any given dataset ahead of computing the regularized TME. For elliptical data, \( C(X) \approx (1 + \sqrt{p/n})^2 \).

**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 \( O(\min(n, p)^3) \) operations due to matrix inversion. For accuracy \( \epsilon \) 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

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

Three choices:
- Gaussian data
- *Laplace*, heavy tailed but all moments exist
- *Cauchy*, no moments exist
Simulation Results

Took approximately sparse matrix

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

Three choices:
- Gaussian data
- Laplace, heavy tailed but all moments exist
- Cauchy, no moments exist

$$p/n = \gamma = 1/2, 1 \text{ or } 2$$
Estimators

Compare 4 estimators:

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

Compare 4 estimators:

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

- $u_i = 1$
- $u_i \sim \text{Laplace}(0,1)$
- $u_i \sim \text{Cauchy}(0,1)$

$\gamma = 0.5$

$\gamma = 1$

$\gamma = 2$

$\gamma = 4$
Suppose $(1 - \epsilon)n$ of the data follow an elliptical distribution with a sparse shape matrix $S_p$.

remaining $\epsilon n$ samples follow a different elliptical distribution with shape $U \frac{pD}{\text{tr}(D)} U^T$, where $U$ is unitary matrix, randomly distributed with Haar measure.
Suppose \((1 - \epsilon)n\) of the data follow an elliptical distribution with a sparse shape matrix \(S_p\).

remaining \(\epsilon n\) samples follow a different elliptical distribution with shape \(U \frac{pD}{\text{tr}(D)} U^T\), where \(U\) is unitary matrix, randomly distributed with Haar measure.

Two models for diagonal \(D\):

1) \(d_{ii} \sim U[1, 5]\) so outliers diffuse

2) \(d_{11} = p, d_{22} = p/2\), and for \(i > 2\) all \(d_{ii} = 1\). Outliers approximately on 2-d random subspace. Here \(s_{\text{max}}\) of outliers is \(O(p)\) so does not satisfy our assumptions on bounded \(s_{\text{max}}\)
Simulation Results, Model 1

TME weights $\epsilon = 0.2$

TME weights $\epsilon = 0.4$
Consider \( 1 - \epsilon \) proportion of samples, \textit{inliers} from elliptical distribution with shape matrix \( S_{in} \).

\( \epsilon \) proportion, \textit{outliers} from elliptical distribution with shape matrix \( S_{out} \).
Consider $1 - \epsilon$ proportion of samples, \textit{inliers} from elliptical distribution with shape matrix $S_{in}$

$\epsilon$ proportion, \textit{outliers} from elliptical distribution with shape matrix $S_{out}$.

\textbf{Conjecture:} As $p, n \to \infty$, under suitable assumptions, the weights in TME concentrate around two values, $w_{in}$ and $w_{out}$. 
Proposed procedure:

Given (regularized)-TME weights, \( w_j \), compute non-parametric density estimate \( \hat{f}(w) \).

Choose \( w_{in} = \text{arg max} \hat{f}(w) \).

Retain all samples with weights in interval \([w_L, w_R]\) around \( w_{in} \) such that

\[
\hat{f}(w) > 0.7 \hat{f}(w_{in})
\]
Simulation Results, Model 1

$n = 500 \quad p = 500, \quad \alpha = 4$

- LRE
- Outliers Removed

- th-RegTME
- Outliers Removed
Simulation Results, Model 2

TME weights $\epsilon = 0.2$

TME weights $\epsilon = 0.4$
Simulation Results, Model 2

$n = 500 \ p = 500, \ \alpha = 4$

LRE

th-RegTME
Outliers Removed

0 0.1 0.2 0.3 0.4
outlier ratio
-1 0 1 2 3
LRE

0 0.1 0.2 0.3 0.4
outlier ratio
Open Questions

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

Chen, Gao, Ren [17'] proved minimax optimality for estimator based on Tukey’s depth function. But extremely computationally challenging (NP-hard?)
Open Questions

- Estimate optimal threshold in data-driven manner
- What if \( p = n^\beta \) for \( \beta > 1 \) ?
- \( \epsilon \)-contamination model ?

Chen, Gao, Ren [17'] proved minimax optimality for estimator based on Tukey’s depth function. But extremely computationally challenging (NP-hard ?)

Is there computationally efficient / practical robust estimator ?
Various contemporary applications involve ‘large $p – small n$’ data.
Summary

- Various contemporary applications involve ‘large \( p \) – small \( n \)’ data.
- Sparse covariance estimation for heavy tailed elliptical data
- Various contemporary applications involve ‘large $p$ – small $n$’ data.

- Sparse covariance estimation for heavy tailed elliptical data

- Is there computationally efficient method to handle arbitrary outliers?
- Various contemporary applications involve ‘large $p$ – small $n$’ data.

- Sparse covariance estimation for heavy tailed elliptical data

- Is there computationally efficient method to handle arbitrary outliers?

- Computationally efficient sparse PCA with heavy tailed data / outliers?

www.weizmann.ac.il/math/nadler
- Various contemporary applications involve ‘large $p – small n$’ data.

- Sparse covariance estimation for heavy tailed elliptical data

- Is there computationally efficient method to handle arbitrary outliers?

- Computationally efficient sparse PCA with heavy tailed data / outliers?

[www.weizmann.ac.il/math/nadler]

THE END / THANK YOU!