# 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

# Intro

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

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

# Intro

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

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

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

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 $\Sigma$

# 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 $\Sigma$

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 $p/n \to c$,

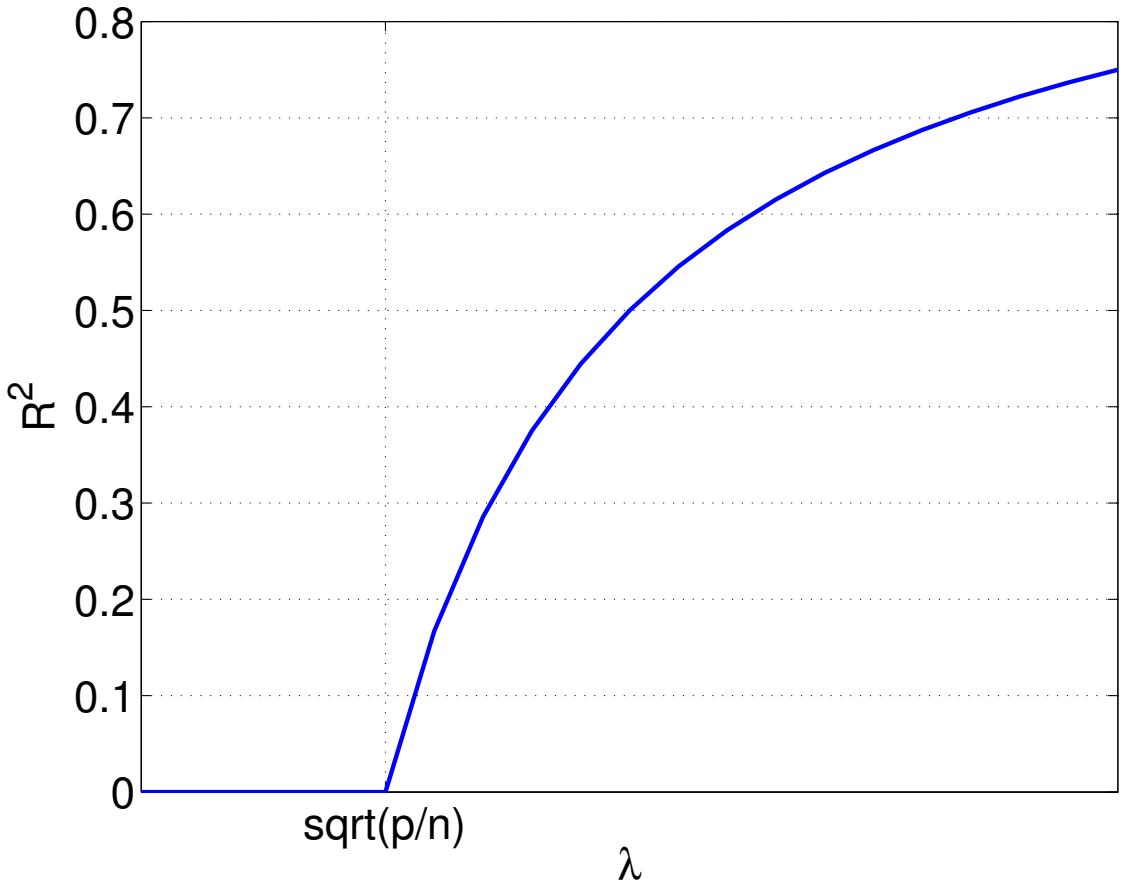
$$R_i^2 = |\langle \hat{\mathbf{v}}_i, \mathbf{v}_i \rangle|^2 \to \begin{cases} 0 & \lambda_i < \sigma^2 \sqrt{p/n} \\ \frac{\frac{\lambda_i^2}{c\sigma^2} - \sigma^2}{\frac{\lambda_i^2}{c\sigma^2} + \lambda_i} & \lambda_i > \sigma^2 \sqrt{p/n} \end{cases}$$

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

Key point:

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

# Breakdown of Classical PCA



# Sparsity and High-Dimensions

**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 $\mu$ - 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 $\mu$ - 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

# Sparsity and High-Dimensions

**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.

# Sparse Covariance Estimation

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

# 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

some theory, some algorithms, many open questions

# Prior Work: 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 $q \in [0, 1)$:

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

# Prior Work: 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

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

# Outlier/Heavy Tail breakdown of sample covariance

Key reason why thresholding works is following (deterministic) lemma

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

# Outlier/Heavy Tail breakdown of sample covariance

Key reason why thresholding works is following (deterministic) lemma

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

$$\max_{ij} |\hat{\Sigma}_{ij} - \Sigma_{ij}| < C\sqrt{\log p/n}$$

# Outlier/Heavy Tail breakdown of sample covariance

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

# 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.

# 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.

# 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 ?

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

# (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 $\mathbb{S}^{p-1}$ 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

# (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 $\mathbb{S}^{p-1}$ 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 $(S_p)_{ii} = 1$.

# 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 $Tr(\Sigma)=1$.

[Kent and Tyler, 88']

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 $Tr(\Sigma)=1$.

[Kent and Tyler, 88']

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

# Tyler's M-estimator

Tyler's M-estimator is maximum likelihood etimate of shape matrix from elliptical distribution.

# Tyler's M-estimator

Tyler's M-estimator is maximum likelihood etimate of shape matrix from elliptical distribution.

It is a robust estimate of $S_p$, consistent for $p$ fixed, $n \to \infty$.

# Tyler's M-estimator

Tyler's M-estimator is maximum likelihood etimate 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

[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.

# 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

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

# Our Results

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

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

Consider following thresholding estimator for shape matrix:

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

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

# Our Results

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

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

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,

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

# 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 $M'$,

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

**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 $M'$,

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

# 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 $M'$,

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

**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]

$\to$ 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

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

Show tight concentration of weights to uniform vector

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

where $r$ is solution of some complicated equation.

# Proof Outline

**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, $\hat{\Sigma}_{\text{TME}}$ is close to sample covariance matrix $\hat{S} = \frac{1}{n}\sum \mathbf{x}_i \mathbf{x}_i^T$ in operator norm,

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

From here previous proof follows.

# Our Results

**Case II, Regularized TME:**

Here $r = r(\alpha)$.

This means that

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

Need to show that $r \in [r_{\min}, r_{\max}]$ 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 $C(X) = \| \frac{1}{n} \sum_{i=1}^{n} (\sqrt{p}\mathbf{x}_i / \|\mathbf{x}_i\|)(\sqrt{p}\mathbf{x}_i / \|\mathbf{x}_i\|)^T \|$

# Computational Complexity

Can one compute regularized TME in polynomial time ?

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

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

# Computational Complexity

Can one compute regularized TME in polynomial time ?

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

$C(X)$ 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.

# Computational Complexity

Can one compute regularized TME in polynomial time ?

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

$C(X)$ 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.

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:

- 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

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

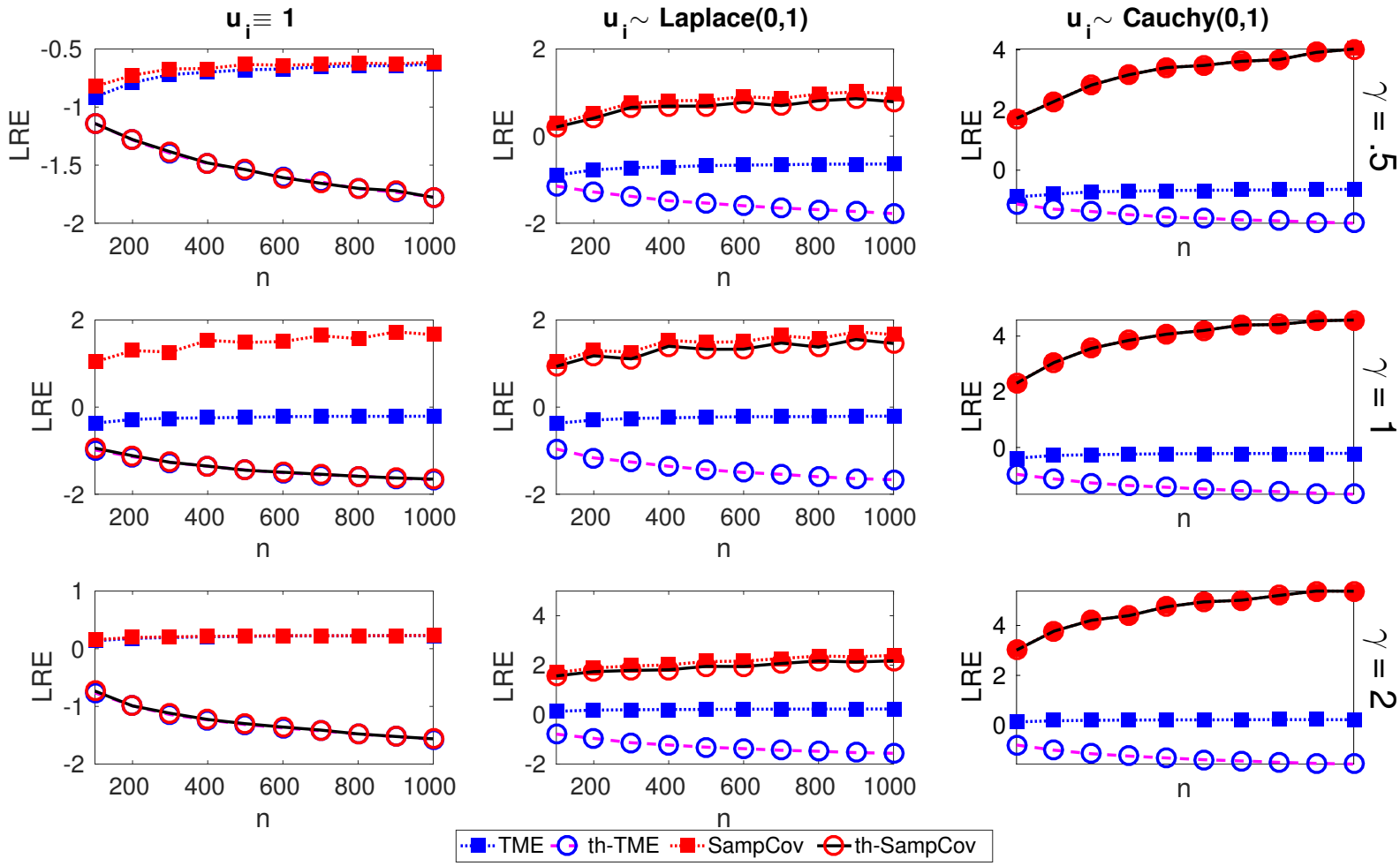
Compare 4 estimators:

- 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



# $\epsilon$-contamination

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 $U\frac{pD}{tr(D)}U^T$, where $U$ is unitary matrix, randomly distributed with Haar measure.

# $\epsilon$-contamination

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

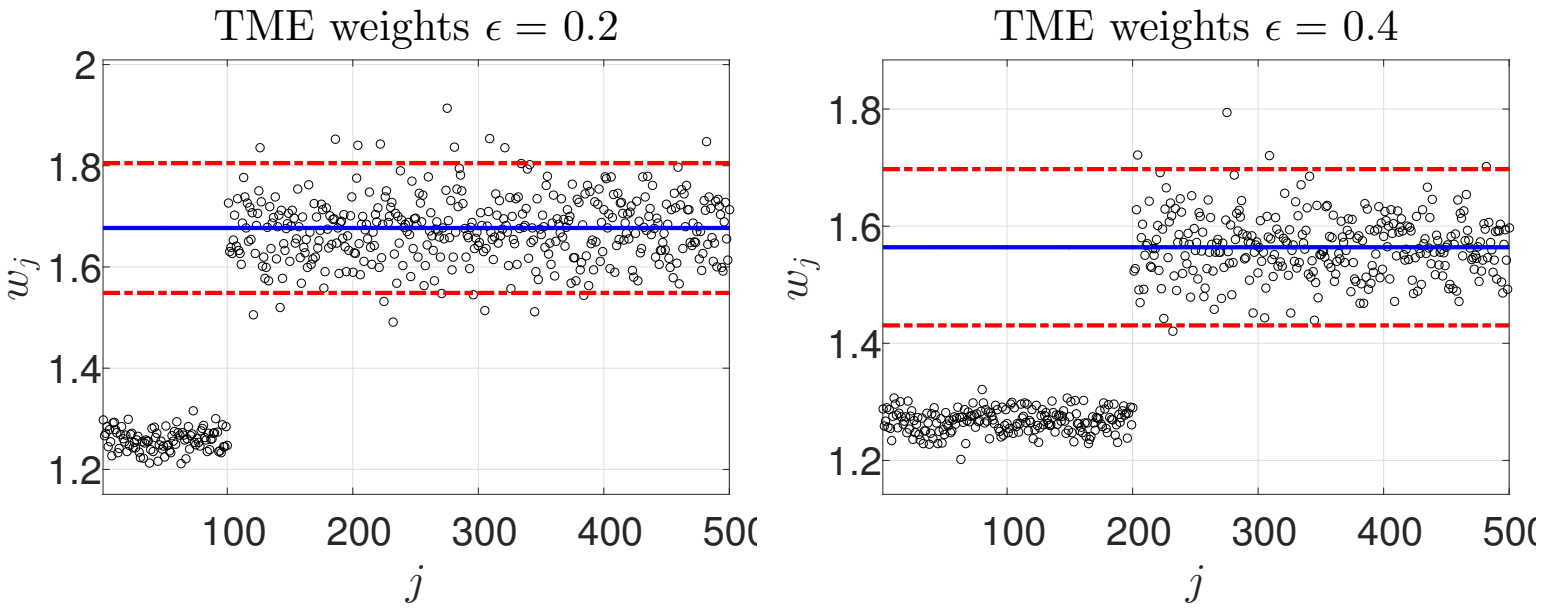
remaining $\epsilon n$ samples follow a different elliptical distribution with shape $U\frac{pD}{tr(D)}U^T$, where $U$ is unitary matrix, randomly distributed with Haar measure.

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 $s_{\max}$ of outliers is $O(p)$ so does not satisfy our assumptions on bounded $s_{\max}$

# Simulation Results, Model 1



# TME with Elliptical outliers

Consider $1-\epsilon$ proportion of samples, *inliers* from elliptical distribution with shape matrix $S_{in}$

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

# TME with Elliptical outliers

Consider $1-\epsilon$ proportion of samples, *inliers* from elliptical distribution with shape matrix $S_{in}$

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

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

# Outlier Detection via TME weights

Proposed procedure:

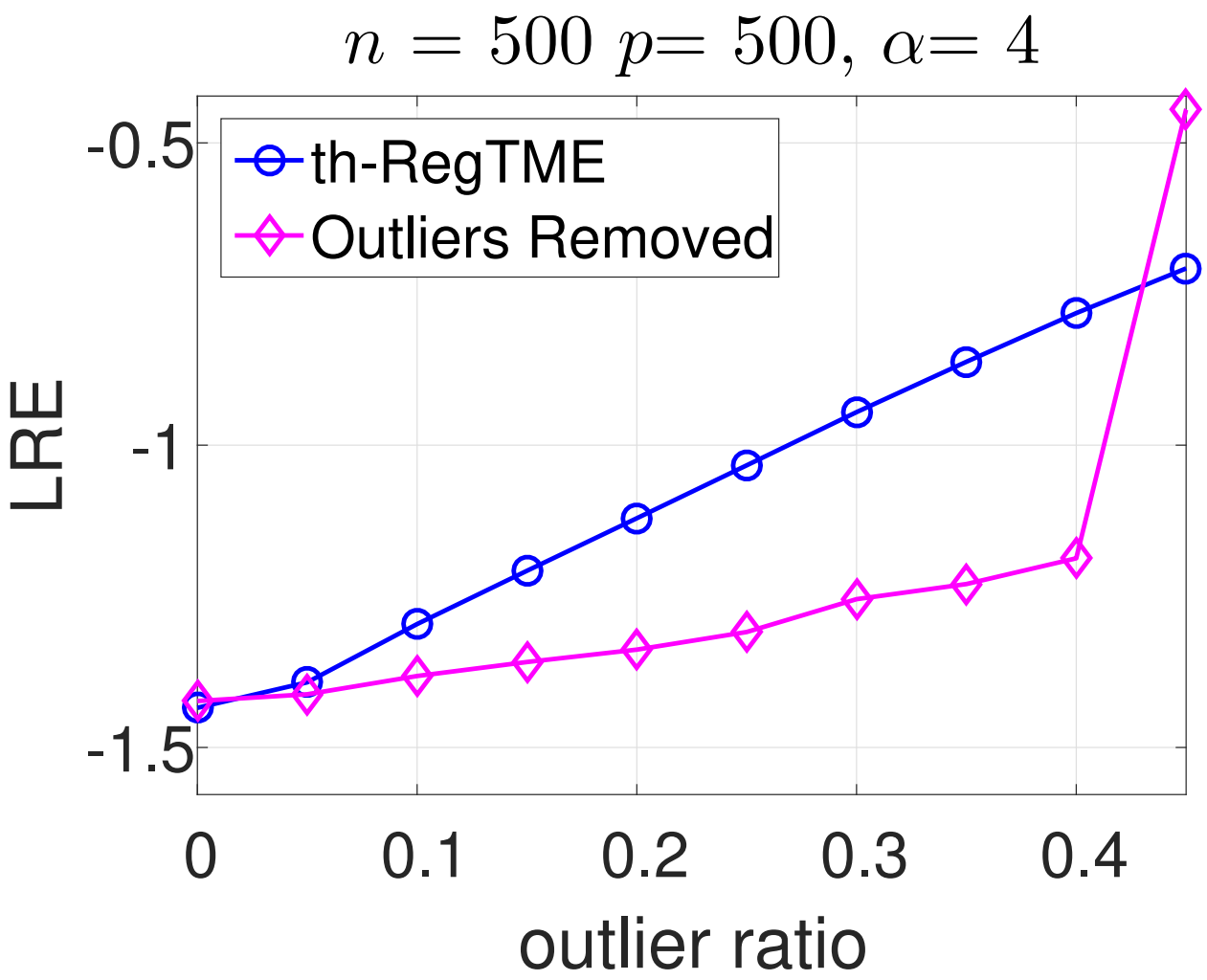
Given (regularized)-TME weights, $w_j$, compute non-parametric density estimate $\hat{f}(w)$.

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

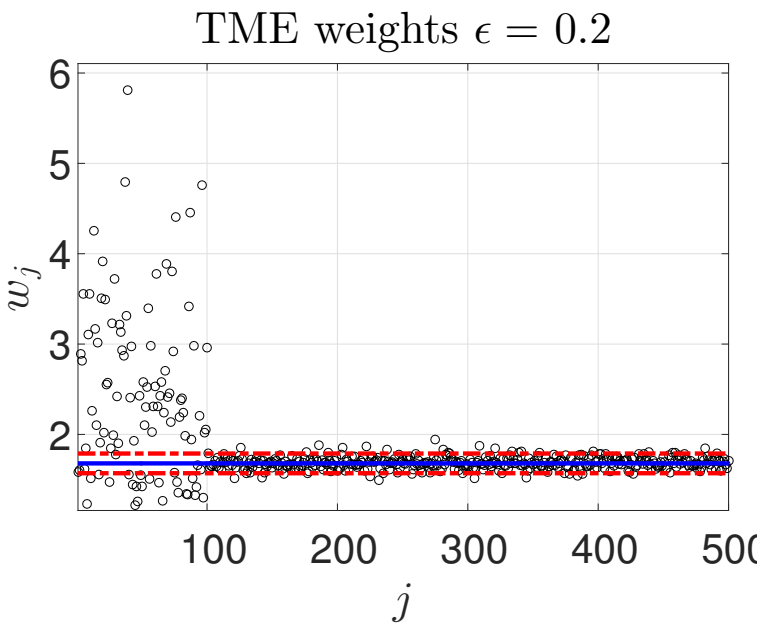
Retain all samples with weights in interval $[w_L, w_R]$ around $w_{in}$ such that

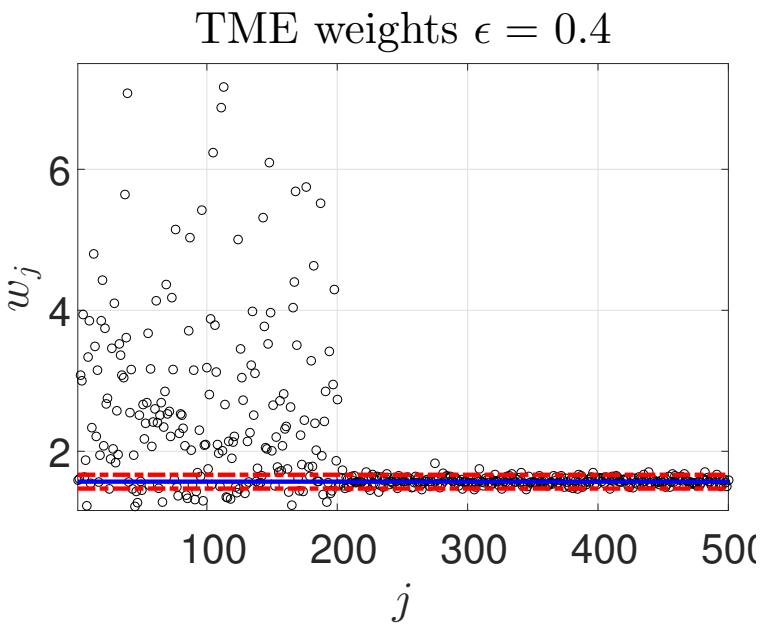
$$\hat{f}(w) > 0.7\hat{f}(w_{in})$$

# Simulation Results, Model 1

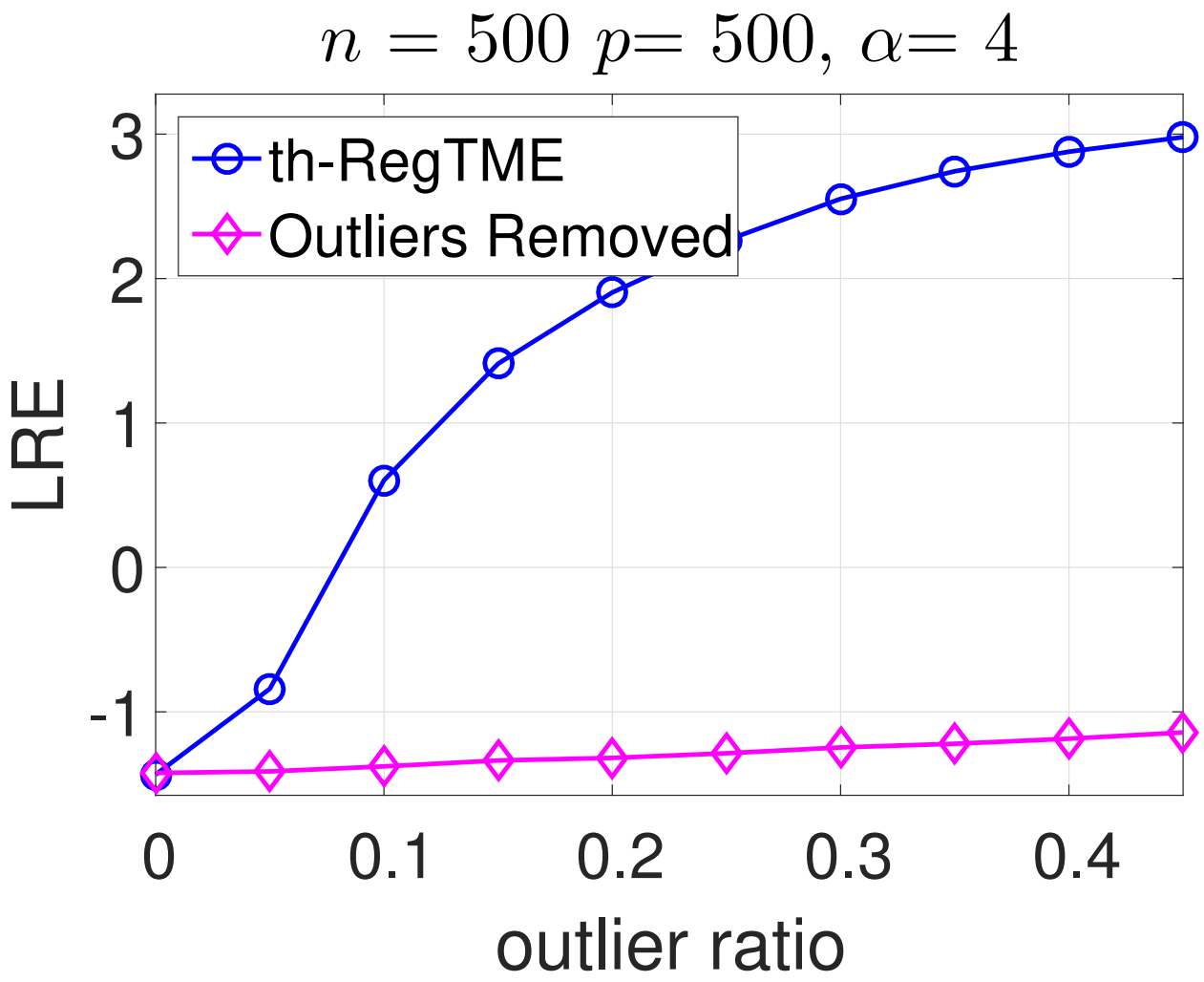


# Simulation Results, Model 2



# Simulation Results, Model 2



# 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 ?

# Summary

- 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

# Summary

- 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 ?

# Summary

- 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 ?

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

www.weizmann.ac.il/math/nadler

# Summary

- 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 ?

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

www.weizmann.ac.il/math/nadler

**THE END / THANK YOU !**