Handling Complex and Noisy Data: Robust Semiparametric Method

Fang Han and Han Liu
Outline

• Literature Review
• Motivation
• Robust Statistics:
  1. Marginal rank — Transelliptical (Elliptical copula)
  2. Multivariate rank — Elliptical
  3. Quantile statistics — Pair-elliptical
• Summary
Covariance matrix plays a key role in multivariate statistics:

- **Principal component analysis**: leading eigenvectors of $\Sigma$;
- **Graphical model estimation**: nonzero entries in $\Sigma^{-1}$;
- **Transition matrix estimation**: $\Sigma = \text{Cov}(X_t)$ and $\Sigma^1 = \text{Cov}(X_t, X_{t+1})$;
- **Linear discriminant analysis**: $(x - \mu_a)\Sigma^{-1}(\mu_1 - \mu_0)$.
- . . . . .
Robust Statistics

Robust alternatives to PCA and sparse PCA:

- **ECA**: Multivariate rank — Elliptical model
- **QCA**: Quantile statistics — Pair-elliptical model
- **TCA**: Marginal rank — Transelliptical (Elliptical copula) model
PCA: All About Covariance Estimation

Let $u_1(\Sigma)$ denote the leading eigenvector of $\Sigma$ and $u_1(\hat{\Sigma})$ denote the leading eigenvector of $\hat{\Sigma}$, an estimator $\Sigma$.

- **Estimation:** Davis-Kahan inequality (sin $\Theta$ theorem):
  \[ |\sin \angle(u_1(\hat{\Sigma}), u_1(\Sigma))| \leq \frac{1}{\lambda_1(\Sigma) - \lambda_2(\Sigma)} \|\hat{\Sigma} - \Sigma\|_2, \quad (\text{Davis and Kahan, 1970}). \]

- **Lower bound:**
  \[ \|\hat{\Sigma} - \Sigma\|_2 \gtrsim \Sigma_2 \sqrt{\frac{r^*(\Sigma)}{n}}, \quad (\text{Lounici, 2012}). \]

- **Upper bound:** ((sub)Gaussian condition and Sample covariance matrix $S$)
  \[ \|S - \Sigma\|_2 \lesssim \|\Sigma\|_2 \sqrt{\frac{r^*(\Sigma) \log d}{n}}, \quad (\text{Lounici, 2012, Bunea and Xiao, 2013}). \]

$(r^*(\Sigma) := \text{Tr}(\Sigma)/\lambda_1(\Sigma)$ is called the "efficient rank").
Why Sparse PCA?

• Empirically, motivated from the inclination to select a subset of features when \( d \) is too large.

• Theoretically, motivated from the possibly inconsistency in PCA (Johnstone, 2001).

• A global optimum:

\[
 u_{1,s}(M) := \arg \max_{\|v\|_0 \leq s, v \in \mathbb{S}^{d-1}} |v^T M v|.
\]

(Estimated based on a global search algorithm).
Sparse PCA: All about Covariance Estimation

When the dimension $d \gg n$ and $\lambda_1(\Sigma)$ is not large enough compared to others:

- **Estimation:** Davis-Kahan type inequality (Vu and Lei, 2012):
  \[
  \sin \angle(u_1, s(\hat{\Sigma}), u_1(\hat{\Sigma})) \leq \frac{2\sqrt{2}}{\lambda_1(\Sigma) - \lambda_2(\Sigma)} \cdot \max_{\|v\|_2 = 1, \|v\|_0 \leq 2s} \frac{|v^T(\hat{\Sigma} - \Sigma)v|}{\|\hat{\Sigma} - \Sigma\|_2, 2s}.
  \]

- **Lower bound:**
  \[
  \|\hat{\Sigma} - \Sigma\|_{2,s} \gtrsim \|\Sigma\|_2 \sqrt{\frac{s \log(d/s)}{n}}, \quad (Vu \text{ and Lei, 2012}),
  \]

- **Upper bound:** ((sub)Gaussian and Sample covariance matrix $S$)
  \[
  \|S - \Sigma\|_{2,s} \lesssim \|\Sigma\|_2 \sqrt{\frac{s \log(d/s)}{n}}. \quad (Vu \text{ and Lei, 2012, 2013})
  \]
What Matters in the Performance?

• Inverse covariance matrix/Graph estimation: element-wise supremum norm $\|\hat{\Sigma} - \Sigma\|_{\text{max}}$;

• Linear Regression/PCA: spectral norm $\|\hat{\Sigma} - \Sigma\|_2$;

• Lasso/Sparse PCA: restricted spectral norm $\|\hat{\Sigma} - \Sigma\|_{2,s}$. 
When Sample Covariance Matrix Works?

• Marginal subgaussian:

\[
\exp(tX) \leq \exp \left( -\frac{K^2t^2}{2} \right) \Rightarrow \| S - \Sigma \|_{\text{max}} \lesssim \sqrt{\frac{\log d}{n}};
\]

• Multivariate subgaussian:

\[
\exp(t \cdot b^TX) \leq \exp \left( -\frac{K^2t^2}{2} \right) \Rightarrow \| S - \Sigma \|_{2,s} \lesssim \sqrt{\frac{s \log(d/s)}{n}};
\]

Key idea: Moment constraint, so light-tailed distribution.
Today we usually need to handle the data with **high dimensions** (Genomics, Brain Imaging, Finance, etc.):

- “Heavy tail” is not a big issue in low dimensions: **CLT**.
- “Heavy tail” is a key problem in high dimensions: **Curse and Bless of Dimensionality** (Donoho, 2000).
452 stocks that were consistently in the S&P 500 index between January 1, 2003 though January 1, 2008. We are interested in the daily log-return value.

Figure 1. Illustration of the asymmetry issue of the log-return stock data.
Understanding the Empirical Mean

What will happen when subgaussian condition is violated? We characterize the heavy tails by the $L_p$ norm.

**Theorem (Upper bound).** Suppose that $X = (X_1, \ldots, X_d)^T \in \mathbb{R}$ with population mean $\mu$ satisfies that $\|X_j\|_{L_p} \leq K$ and $p = 2 + 2\gamma + \delta$. For $d = O(n^\gamma)$, letting $\bar{\mu}$ be the sample mean of $n$ observations $X_1, \ldots, X_n$ of $X$, we then have, with probability not smaller than $1 - 2d^{-2.5} - (\log d)^{p/2} n^{-\delta/2}$,

$$\|\bar{\mu} - \mu\|_\infty \leq 12K \cdot \sqrt{\frac{\log d}{n}}.$$ 

This means for heavy tailed distributions, we have optimal rate of convergence but suboptimal scaling.
The scalings with regard to \((d, n)\) in the last theorem cannot be improved in general.

**Theorem (Lower bound).** Consider any absolute constant \(C\) and \(p = 2 + 2\gamma\) with \(\gamma > 0\). There exists some distribution \(X = (X_1, \ldots, X_d)^T \in \mathbb{R}^d\) with \(\|X_j\|_{L_q} < \infty\) for all \(q \leq p\) and \(\|X_j\|_{L_q} \to \infty\) for all \(q > p\), such that, when \(d = n^{\gamma+\delta_0}\) with \(\delta_0 > 0\), we have with probability tending to 1,

\[
\|\bar{\mu} - \mu\|_\infty \geq \sqrt{\frac{C \log d}{n}}.
\]
Heavy Tailedness: First Attempt

The elliptical distribution is defined as:

\[ X \sim_d \mu + \xi A U, \]

where \( AA^T = \Sigma \) of rank \( q \), \( U \) uniformly distributed in \( \mathbb{S}^{q-1} \) independent of the nonnegative random variable \( \xi \geq 0 \). \( \xi \) is unspecified. In this case, we represent it by \( EC_d(\mu, \Sigma, \xi) \). \( \Sigma \) is called the scatter matrix.

Some examples: Gaussian, rank-deficient Gaussian, multivariate t, Cauchy, Logistic, and etc..

Note: Elliptical distributions can be very heavy tailed and even have infinite first or second moments.
Elliptical distribution is the scaled Gaussian: $X - \mu \overset{d}{=} \xi' \cdot N(0, \Sigma)$.

Reference: Liu, Han, and Zhang (2013).

**Theorem.** Let $X \sim EC_d(\mu, \Sigma, \xi)$ be an elliptical distribution with $\Sigma = AA^T$. It takes another stochastic representation:

$$X \overset{d}{=} \mu + \frac{\xi}{\|A^\dagger Z\|_2} \cdot Z,$$

where $Z \sim N_d(0, \Sigma)$, $\xi \geq 0$ is independent of $Y/\|A^\dagger Y\|_2$ and $A^\dagger$ is the Moore-Penrose pseudoinverse of $A$. 
A Journey Away from the Normal I

elliptical

Gaussian

multivariate t
Population Multivariate Kendall’s tau:

\[ K := \mathbb{E} \left( \frac{(X - \tilde{X})(X - \tilde{X})^T}{\|X - \tilde{X}\|_2^2} \right) = \mathbb{E} S(X)S(X)^T. \]

\[ S(X) := (X - \tilde{X})/\|X - \tilde{X}\|_2 \] is the self-normalized and centralized data.

**Remark 1:** \( K \) can be considered as the covariance matrix of the self-normalized data of the pair-wise differences \( \{x_i - x_{i'}\}_{i < i'} \).

**Remark 2:** Self-normalized samples are bounded and enjoy various “moment constraint free” property. Check Shao, 1997 for more details.
**Marden's Theorem**

**Theorem.** Let $X \sim EC_d(\mu, \Sigma, \xi)$ be a continuous distribution and $K$ be the population multivariate Kendall’s tau statistic. Then $K$ and $\Sigma$ share the same eigenspace and

$$
\lambda_j(K) = \mathbb{E} \left( \frac{\lambda_j(\Sigma)Y_j^2}{\lambda_1(\Sigma)Y_1^2 + \ldots + \lambda_d(\Sigma)Y_d^2} \right), \quad (Y_1, \ldots, Y_d)^T \sim N_d(0, I_d).
$$

Reference: Marden (1999); Croux et.al. (2002).
An Illustrative Proof

Prove “\( \lambda_j(\Sigma) > \lambda_k(\Sigma) \) implies \( \lambda_j(K) > \lambda_k(K) \):

\[
\frac{\lambda_k(K)}{\lambda_j(K)} = \frac{\mathbb{E}_j \frac{\lambda_k(\Sigma)Y_k^2}{\lambda_j(\Sigma)Y_j^2 + \lambda_k(\Sigma)Y_k^2 + E}}{\mathbb{E}_j \frac{\lambda_j(\Sigma)Y_j^2}{\lambda_j(\Sigma)Y_j^2 + \lambda_k(\Sigma)Y_k^2 + E}} < \frac{\mathbb{E}_j \frac{\lambda_k(\Sigma)Y_k^2}{\lambda_j(\Sigma)Y_j^2 + \lambda_k(\Sigma)Y_k^2 + E}}{\mathbb{E}_j \frac{\lambda_j(\Sigma)Y_j^2}{\lambda_j(\Sigma)Y_j^2 + \lambda_j(\Sigma)Y_j^2 + E}} = \frac{\mathbb{E}_j \frac{Y_k^2}{Y_j^2 + Y_k^2 + E / \lambda_k(\Sigma)}}{\mathbb{E}_j \frac{Y_k^2}{Y_j^2 + Y_k^2 + E / \lambda_j(\Sigma)}} < 1.
\]
Elliptical Component Analysis (ECA)

Let \(x_1, \ldots, x_n\) be \(n\) independent realizations of \(X\). We use the multivariate Kendall’s tau estimate:

\[
\hat{K} = \frac{2}{n(n-1)} \sum_{i < i'} \frac{(x_i - x_{i'})^T (x_i - x_{i'})}{\|x_i - x_{i'}\|^2_2}.
\]

Method: Plugging \(\hat{K}\) into any (sparse) principal component algorithm.

Remark: \(\hat{K}\) is a U statistic with kernel \(k_{KM}(\cdot)\) bounded: \(\|k_{KM}(\cdot)\|_2 \leq 1\).
Theoretical Results (non-sparse setting)

**Theorem.** Under the elliptical model, if \( n \gtrsim r^*(K) \log d \), we have

\[
\| \hat{K} - K \|_2 \lesssim \| K \|_2 \sqrt{\frac{r^*(K) \log d}{n}}.
\]

**Proof:** Matrix Bernstein inequality for U statistics (Tropp, 2011; Han and Liu, 2013).

**Implication:** \( | \sin \angle (u_1(\hat{K}), u_1(K)) | \lesssim \frac{\lambda_1(K)}{\lambda_1(K) - \lambda_2(K)} \cdot \sqrt{\frac{r^*(K) \log d}{n}} \).

**Reference:** Han and Liu, 2013.
Theoretical Results (sparse setting)

Some definitions and preliminary results:

- The subgaussian and sub-exponential norms:
  \[
  \|X\|_{\psi_2} := \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|X|^p)^{1/p} \quad \|X\|_{\psi_2} := \sup_{p \geq 1} p^{-1} (\mathbb{E}|X|^p)^{1/p}.
  \]

- \(X\) is subgaussian (sub-exponential) iff \(\|X\|_{\psi_2} < \infty\) (\(\|X\|_{\psi_1} < \infty\)).

- \(X\) is subgaussian iff \(X^2\) is sub-exponential:
  \[
  \|X\|_{\psi_2}^2 \leq \|X\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2.
  \]

- \(X\) is sub-exponential means it has Bernstein-type tail behavior.

- \(X\) gives us mean concentration, \(X^2\) gives us covariance concentration.
Theoretical Results (sparse setting)

**Theorem.** Let \(x_1, \ldots, x_n\) be \(n\) observations of \(X \sim EC_d(\mu, \Sigma, \xi)\) with \(\text{rank}(\Sigma) = q\) and \(\|u_1(\Sigma)\|_0 \leq s\). When \(n \gtrsim s \log(d/s)\), we have

\[
\|\hat{K} - K\|_{2,s} \lesssim \left( \frac{\lambda_1(\Sigma)}{q\lambda_q(\Sigma)} + \lambda_1(K) \right) \cdot \sqrt{\frac{s \log(d/s)}{n}}.
\]

**Proof:** Show that \(\sup_{v \in \mathbb{S}^{d-1}} \|v^T S(X)\|_{\psi_2} \leq \sqrt{\frac{\lambda_1(\Sigma)}{\lambda_q(\Sigma)} \cdot \frac{2}{q}}\) and the rest follows from the standard arguments in sparse PCA (using Hoeffding’s decoupling trick to handle the correlatedness between the kernels).

**Remark:** We would like to have \(\frac{\lambda_1(\Sigma)}{q\lambda_q(\Sigma)} \lesssim \lambda_1(K)\) for (sparse) PCA since \(|\sin \angle(u_1(\hat{K}), u_1(K))| \leq \frac{1}{\lambda_1(K) - \lambda_2(K)} \cdot \|\hat{K} - K\|_2\).
**Theoretical Results (sparse setting)**

**Theorem.** Let $x_1, \ldots, x_n$ be $n$ observations of $X \sim EC_d(\mu, \Sigma, \xi)$ with rank$(\Sigma) = q$ and $\|u_1(\Sigma)\|_0 \leq s$. Then when $n \gtrsim \log d$, we have

$$\|\hat{K} - K\|_{\max} \lesssim \left( \frac{\lambda_1(\Sigma)}{q\lambda_q(\Sigma)} + \|K\|_{\max} \right) \cdot \sqrt{\frac{\log d}{n}}.$$

**Proof:** Notice that $S(X)$ is multivariate subgaussian, implying that $S(X)$ is marginal subgaussian, too. The rest follows by using a concentration inequality for U statistics of subgaussian entries.
Theorem: When $\|\Sigma\|_F \log d = \text{Tr}(\Sigma) \cdot o(1)$ (e.g., condition number controlled),

$$
\lambda_j(K) = \mathbb{E} \left( \frac{\lambda_j(\Sigma)Y_j^2}{\lambda_1(\Sigma)Y_1^2 + \ldots + \lambda_d(\Sigma)Y_d^2} \right) \lesssim \frac{\lambda_j(\Sigma)}{\text{Tr}(\Sigma)}.
$$

Proof: Use the Hanson-Wright type inequality to control the quadratic term of the Gaussian in the denominator term.
Theorem. Under mild conditions (e.g., condition number of the matrix is controlled), we have

$$
\| \hat{K} - K \|_{2,s} \lesssim \| K \|_2 \sqrt{\frac{s \log(d/s)}{n}}, \quad \left( \text{in particular, } \| \hat{K} - K \|_2 \lesssim \| K \|_2 \sqrt{\frac{d}{n}} \right)
$$

and

$$
\| \hat{K} - K \|_{\text{max}} \lesssim \| K \|_{\text{max}} \sqrt{\frac{\log d}{n}}.
$$

Implication: 

$$
| \sin \angle (u_{1,s}(\hat{K}), u_{1,s}(K)) | \lesssim \frac{\lambda_1(K)}{\lambda_1(K) - \lambda_2(K)} \cdot \sqrt{\frac{s \log(d/s)}{n}}.
$$

Proof: We have 

$$
\frac{\lambda_1(\Sigma)}{q \lambda_q(\Sigma)} \sim \| K \|_{\text{max}} = \frac{\lambda_1(\Sigma)}{\text{Tr}(\Sigma)} \iff q \lambda_q(\Sigma) \asymp \text{Tr}(\Sigma).
$$
**Theorem.** When \( x_1, \ldots, x_n \) are \( n \) independent observations of an elliptically distributed random vector \( X \in \mathbb{R}^d \) (no moment constraint!), under mild conditions,

\[
\sqrt{\frac{\log d}{n}} \lesssim d\|\hat{K} - K\|_{\text{max}} \lesssim \sqrt{\frac{\log d}{n}}, \quad \text{(optimal and parametric rate)}
\]

\[
\sqrt{\frac{d}{n}} \lesssim d\|\hat{K} - K\|_{\text{max}} \lesssim \sqrt{\frac{d}{n}}, \quad \text{(optimal and parametric rate)}
\]

and

\[
\sqrt{\frac{s \log(d/s)}{n}} \lesssim d\|\hat{K} - K\|_{2,s} \lesssim \sqrt{\frac{s \log(d/s)}{n}}. \quad \text{(optimal and parametric rate)}
\]
Two Constraints of ECA

- Method-wise, ECA cannot estimate $\Sigma$.
- Model-wise, elliptical distribution cannot handle asymmetric data.
Generalized median absolute deviation (gMAD):

robust scale estimator to the standard deviation and covariance:

(marginal standard deviation) \( \sigma^M(X) = Q(|X - \text{median}(X)|, r) \),

(pairwise covariance) \( \sigma^M(X, Y) = \frac{1}{4}\{(\sigma^M(X + Y))^2 - (\sigma^M(X - Y))^2\} \).

Let \( M \) and \( \hat{M} \) be the population and empirical versions of the gMAD estimator.
First Solution: Pair-Elliptical Family

Let **pair-elliptical** be the family of distributions such that they are pairwise elliptically distributed.

**Theorem.** Elliptical Distribution is a strict sub-family of the pair-elliptical family.

**Theorem.** $M$ is a scatter matrix of $X$ under the pair-elliptical family.
A Journey Away from the Normal II

multivariate-t

elliptical

Gaussian

Pair-elliptical
Quantile Component Analysis (QCA)

Let $x_1, \ldots, x_n$ be $n$ independent realizations of $X$. We use the gMAD estimator $\widehat{M}$ to estimate a scatter matrix of $X$.

Method: Plugging $\widehat{M}$ into any (sparse) principal component algorithm.

Remark: QCA is a quantile-based statistics, and hence belongs to the general L-estimator family (van der Vaart (1998)).
Theory Summary

**Theorem.** When $X$ belongs to the pair-elliptical family and the marginal densities of $\{X_j\}$, around a small enough given region, are lower bounded by an absolute constant (no moment constraint!), we have

$$\sqrt{\frac{\log d}{n}} \lesssim \|\hat{M} - c\Sigma\|_{\text{max}} \lesssim \sqrt{\frac{\log d}{n}}, \quad \text{(optimal and parametric rate)}$$

and

$$? \lesssim \|\hat{M} - c\Sigma\|_{2,s} \lesssim s\sqrt{\frac{\log d}{n}}. \quad \text{(sub-parametric rate)}$$

**Proof:** Applying the Hoeffding’s inequality to the quantile statistics.

**Implication:** $|\sin \angle(u_{1,s}(\hat{M}), u_{1,s}(\Sigma))| \lesssim \frac{\lambda_1(\Sigma)}{\lambda_1(\Sigma) - \lambda_2(\Sigma)} \cdot s\sqrt{\frac{\log d}{n}}.$
**ECA v.s. QCA**

- **pro:** QCA can estimate the scatter/covariance matrix:
  \[
  \|\hat{M} - c\Sigma\|_{\text{max}} = O_P(\sqrt{\log d/n}); \text{ and } \|\hat{s} \cdot \hat{M} - \Sigma\|_{\text{max}} = O_P(\sqrt{\log d/n}).
  \]

- **con:** ECA achieves the optimal rate in estimating the leading eigenvector of \(\Sigma\), while for QCA it is sub-optimal (within the elliptical family).
**Second Solution: Transelliptical**

**Definition.** A random vector $X = (X_1, \ldots, X_d)^T$ is said to follow a transelliptical distribution if there exists a set of unspecified univariate increasing functions $\{f_j\}_{j=1}^d$ such that

$$(f_1(X_1), f_2(X_2), \ldots, f_d(X_d))^T \sim EC_d(0, \Sigma^0, \xi), \quad \text{where } \text{diag}(\Sigma^0) = I_d.$$ 

In this case, we represent $X \sim TE_d(\Sigma^0, \xi; f_1, \ldots, f_d)$. We call $\Sigma^0$ the latent generalized correlation matrix.

In particular, when $(f_1(X_1), f_2(X_2), \ldots, f_d(X_d))^T \sim N_d(0, \Sigma^0)$, we say $X$ follows a nonparanormal distribution.
Figure 2. Densities of Cauchy and two transelliptical distributions.
Transelliptical: Generating Scheme

\[ g_1(\cdot) \]

\[ g_2(\cdot) \]

Elliptical

Gaussian

Transelliptical
Transelliptical: A Three-Layer Hierarchical Structure

Layer 1: Observed transelliptical Variables

Layer 2: Latent Elliptical Variables

Layer 3: Latent Gaussian Variables

Latent Generalized Partial Correlation graph $G$
A Journey Away from the Normal III

- Transelliptical
- Elliptical
- Multivariate t
- Gaussian
- Nonparanormal
Population Kendall’s tau:

\[ \tau(X_j, X_k) = \text{Cov}(\text{sign}(X_j - \tilde{X}_j), \text{sign}(X_k - \tilde{X}_k)), \]

where \((\tilde{X}_j, \tilde{X}_k)^T\) is an independent copy of \((X_j, X_k)^T\).

Remark: Kendall’s tau can be considered as the covariance matrix of the sign of the pairwise difference: \(\{\text{sign}(x_i - x_{i'})\}_{i < i'}\).
**Theorem.** Given $X \sim T E_d(\Sigma^0, \xi; f_1, \ldots, f_d)$, for $j \neq k \in \{1, \ldots, d\}$, 

$$
\Sigma^0_{jk} = \sin \left( \frac{\pi}{2} \tau(X_j, X_k) \right).
$$

**Proof:** $\text{sign}(X - \bar{X}) = \text{sign}(Z - \bar{Z}) = \text{sign}(Y - \bar{Y})$, where $Y - \bar{Y}$ is a normal distribution with mean zero. The rest follows from the relation between Kendall’s tau and Pearson’s correlation (Kruskal, 1958).
Transelliptical Component Analysis (TCA)

Let $x_1, \ldots, x_n$ be $n$ independent realizations of $X$. Use the Kendall’s tau correlation coefficient estimate:

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \text{sign}(x_{ij} - x_{i'j}) (x_{ik} - x_{i'k}) .$$

Let $\hat{T} := [\hat{\tau}_{jk}]$ and $\hat{R} := [\sin(\frac{\pi}{2}\hat{\tau}_{jk})]$.

Method: Plugging $\hat{R}$ into any (sparse) principal component algorithm.

---

Remark 1: Kendall’s tau is a second order U statistic.

Remark 2: In using Kendall’s tau, we are conducting scale-invariant (sparse) PCA.
Theoretical Results (non-sparse setting)

**Theorem.** Let $x_1, \ldots, x_n$ be $n$ observations of $X \sim TE_d(\Sigma, \xi, f)$. Let $\hat{T}$ be the sample version marginal Kendall’s tau statistic. Provided $n$ satisfies $n \gtrsim d \log d$, we have

$$\| \hat{T} - T \|_2 \lesssim \| T \|_2 \sqrt{\frac{r^*(T) \log d}{n}}.$$

**Proof:** Matrix Bernstein inequality (Tropp, 2011).
Eigenvalues Relationship

Theorem. We have \( \frac{\pi}{2} \| \Sigma^0 \|_2 \leq \| T \|_2 \leq \| \Sigma^0 \|_2 \).

Proof. Using the Taylor expansion and the matrix spectral inequalities for Hadamard product.

Reference: Wegkamp and Zhao (2013)
Theoretical Results (non-sparse setting)

**Theorem.** Let $x_1, \ldots, x_n$ be $n$ observations of $X \sim T E_d(\Sigma, \xi, f)$. Let $\hat{R} = \sin(\frac{\pi}{2} \hat{T})$ be the sample version transformed marginal Kendall’s tau statistic. Provided $n$ satisfies $n \gtrsim d \log d/n$, we have

$$\| \hat{R} - \Sigma^0 \|_2 \leq \| \Sigma^0 \|_2 \left( \sqrt{\frac{r^*(\Sigma^0) \log d}{n}} + \frac{r^*(\Sigma^0) \log d}{n} \right).$$

**Proof:** Taylor expansion for the entries in the matrix and the matrix spectral inequalities for Hadamard product.

**Implication:** $| \sin \angle(u_1(\hat{R}), u_1(\Sigma^0)) | \lesssim \frac{\lambda_1(\Sigma^0)}{\lambda_1(\Sigma^0) - \lambda_2(\Sigma^0)} \cdot \sqrt{\frac{r^*(\Sigma^0) \log d}{n}}$. 
Theorem. When $x_1, \ldots, x_n$ are $n$ independent observations of a transelliptically distributed random vector $X \in \mathbb{R}^d$ (no moment constraint!),

$$\sqrt{\frac{\log d}{n}} \lesssim \|\hat{R} - \Sigma^0\|_{\text{max}} \lesssim \sqrt{\frac{\log d}{n}}, \quad \text{(Optimal and parametric rate)}$$

$$\sqrt{\frac{d}{n}} \lesssim \|\hat{R} - \Sigma^0\|_2 \lesssim \sqrt{\frac{d \log d}{n}}, \quad \text{(near optimal and parametric rate)}$$

and

$$\sqrt{\frac{\log d}{n}} \lesssim \|\hat{R} - \Sigma^0\|_{2,s} \lesssim s \sqrt{\frac{\log d}{n}}. \quad \text{(sub-parametric rate)}$$
Theoretical Results cont'd

**Theorem.** When $x_1, \ldots, x_n$ are $n$ independent observations of a transelliptically distributed random vector $X \in \mathbb{R}^d$ with a sign subgaussian condition held, we have

$$\|\hat{R} - \Sigma^0\|_{2,s} \lesssim \sqrt{\frac{s \log(d/s)}{n}}. \quad \text{(Optimal and parametric rate)}$$

**Proof.** Remind that Kendall’s tau is working on the sign of the pairwise differences. So, as in the sparse PCA literature, showing that $\sup_{v \in \mathbb{S}^{d-1}} \|v^T \text{sign}(X)\|_{\psi_2} < \infty$ is enough.

**Implication:** $|\sin \angle (u_1,s(\hat{R}),u_1,s(\Sigma^0))| \lesssim \frac{\lambda_1(\Sigma^0)}{\lambda_1(\Sigma^0) - \lambda_2(\Sigma^0)} \cdot \sqrt{\frac{s \log(d/s)}{n}}.$
More About Sign Subgaussian Condition

• A simple definition: $X$ is said to have the sign subgaussian property if

$$\sup_{v \in S^{d-1}} \|v^T \text{sign}(X - \tilde{X})\|_{\psi_2} < \infty,$$

where $\tilde{X}$ is an independent copy of $X$.

• Examples: Elliptical distributions with block-compound-symmetry correlation matrices:

$$\Sigma = \begin{pmatrix}
\Sigma_1 & 0 & 0 & \ldots & 0 \\
0 & \Sigma_2 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & \Sigma_q
\end{pmatrix},$$

where $\Sigma_k \in \mathbb{R}^{d_k \times d_k}$ for $k = 1, \ldots, q$ is compound symmetry (i.e., has the same off-diagonal values).
Implications: Subgaussianity of Correlated Bernoulli

Definition. Using latent variable model:

- Latent: \( Z = (Z_1, \ldots, Z_d)^T \sim N_d(0, \Sigma) \), with \( \Sigma_{jj} = 1 \) \( (j = 1, \ldots, d) \);
- Observed: \( X = (X_1, \ldots, X_d)^T \), with \( X_j = \text{sign}(Z_j) \) \( (j = 1, \ldots, d) \).

Question. Is \( X \) a subgaussian distribution:

\[
\sup_{\|v\| \leq 1} \mathbb{E} \exp(t \cdot v^T X) \leq \exp(C\|\Sigma\|_2 t^2), \quad \text{for all } t \in \mathbb{R} ?
\]

Existing result. If \( \Sigma \) is compound symmetry, the above equation holds.
An Overview

transelliptical

elliptical*

pair-elliptical

nonparanormal

pair-normal

Gaussian
Table 1: Comparisons between ECA, QCA, and TCA

<table>
<thead>
<tr>
<th></th>
<th>ECA</th>
<th>QCA</th>
<th>TCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>model:</td>
<td>elliptical</td>
<td>Pair-elliptical family &amp; possibly non-i.d.d</td>
<td>transelliptical &amp; possibly non-i.d.d</td>
</tr>
<tr>
<td>estimation:</td>
<td>eigenvector of</td>
<td>covariance/ correlation matrix</td>
<td>latent correlation matrix only</td>
</tr>
<tr>
<td></td>
<td>covariance matrix</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>only</td>
<td></td>
<td></td>
</tr>
<tr>
<td>convergence</td>
<td>optimal</td>
<td>sub-parametric</td>
<td>optimal (nonparanormal), sub-optimal (general transelliptical)</td>
</tr>
<tr>
<td>rate (PCA):</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Do not forget:

They are robust!
Summary

- **Modeling flexibility**: The elliptical, pair-elliptical, and transelliptical modeling techniques, allowing data in complex structure;

- **Statistical efficiency**: A series of new optimal and sub-parametric rates of convergence under the double asymptotic framework;

- **Computational efficiency**: Procedure simple and computation efficiency;

- **Theoretical contribution**: First optimality results in high dimensional heavy-tailed distributions.
Multivariate rank (ECA) is the only optimal PCA method in the elliptical family. However, it is pretty *constrained to the PCA procedure*.

In contrast, Marginal rank and quantile-based estimators have broader interests:

*Covariance matrix estimation, inverse covariance matrix estimation (conditional independence graph estimation), regression, discriminant analysis, ...*
Discussion: Away from PCA

- Marginal rank (Kendall’s tau) is optimal in estimating the conditional independence graph with regard to the nonparanormal family.

- Quantile-based estimator (gMAD) attains the parametric rate in portfolio optimization: Asset allocation weight $\hat{w}^{opt}$

\[
\hat{w}^{opt} := \arg\min \sum_{w_i=1} w^T \Sigma w, \quad \text{s.t.} \quad \|w\|_1 \leq c \quad \text{(exposure constraint)}.
\]
## Discussion 2: Statistics and Computation Tradeoff

The illustration of the results in standard (sparse) PCA, TCA, and ECA for the leading eigenvector estimation. Sparse PCA is conducted via the Fantope projection proposed in **Vu, et.al. (2013)**.

<table>
<thead>
<tr>
<th></th>
<th>standard (sparse) PCA</th>
<th>TCA</th>
<th>ECA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>working model:</strong></td>
<td>subgaussian family</td>
<td>transelliptical family</td>
<td>elliptical family</td>
</tr>
<tr>
<td><strong>parameter of interest:</strong></td>
<td>eigenvectors of $\Sigma$</td>
<td>eigenvectors of $\Sigma^0$</td>
<td>eigenvectors of $\Sigma$</td>
</tr>
<tr>
<td><strong>input statistics:</strong></td>
<td>Pearson’s covariance matrix</td>
<td>Kendall’s tau</td>
<td>multivariate Kendall’s tau</td>
</tr>
<tr>
<td><strong>sparse setting (r.c.):</strong></td>
<td>$\sqrt{r^*(\Sigma) \log d/n}$</td>
<td>$\sqrt{r^*(\Sigma^0) \log d/n}$</td>
<td>$\sqrt{r^*(\Sigma) \log d/n}$</td>
</tr>
<tr>
<td><strong>n-s setting 1 (r.c.):</strong></td>
<td>$\sqrt{s \log(d/s)/n}$</td>
<td>$s\sqrt{\log d/n}$ (general),</td>
<td>$\sqrt{s \log(d/s)/n}$,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sqrt{s \log(d/s)/n}$ (sign subgaussian)</td>
<td></td>
</tr>
<tr>
<td><strong>n-s setting 2 (r,c):</strong></td>
<td>$\sqrt{s \log(d/s)/n}$</td>
<td>$s\sqrt{\log d/n}$</td>
<td>$s\sqrt{s \log(d/s)/n}$,</td>
</tr>
</tbody>
</table>

given $s^2 \log d/n \to 0$
Thanks!