Statistical Inference for Noisy Matrix Completion

Dong XIA
Department of Mathematics
Hong Kong University of Science and Technology

Joint work with Ming Yuan (Columbia University)

Outline

- ***Motivation and Statistical Model***
- *Prior Works: (much) Estimation and (few) Inference*
- *Statistical Inference of Linear Forms*
- *Methodology: Double Sample-Splitting and Projection*
- *Theory: Data-driven Asymptotical Normality*
- *Numerical Experiments*

Examples of Matrix Completion

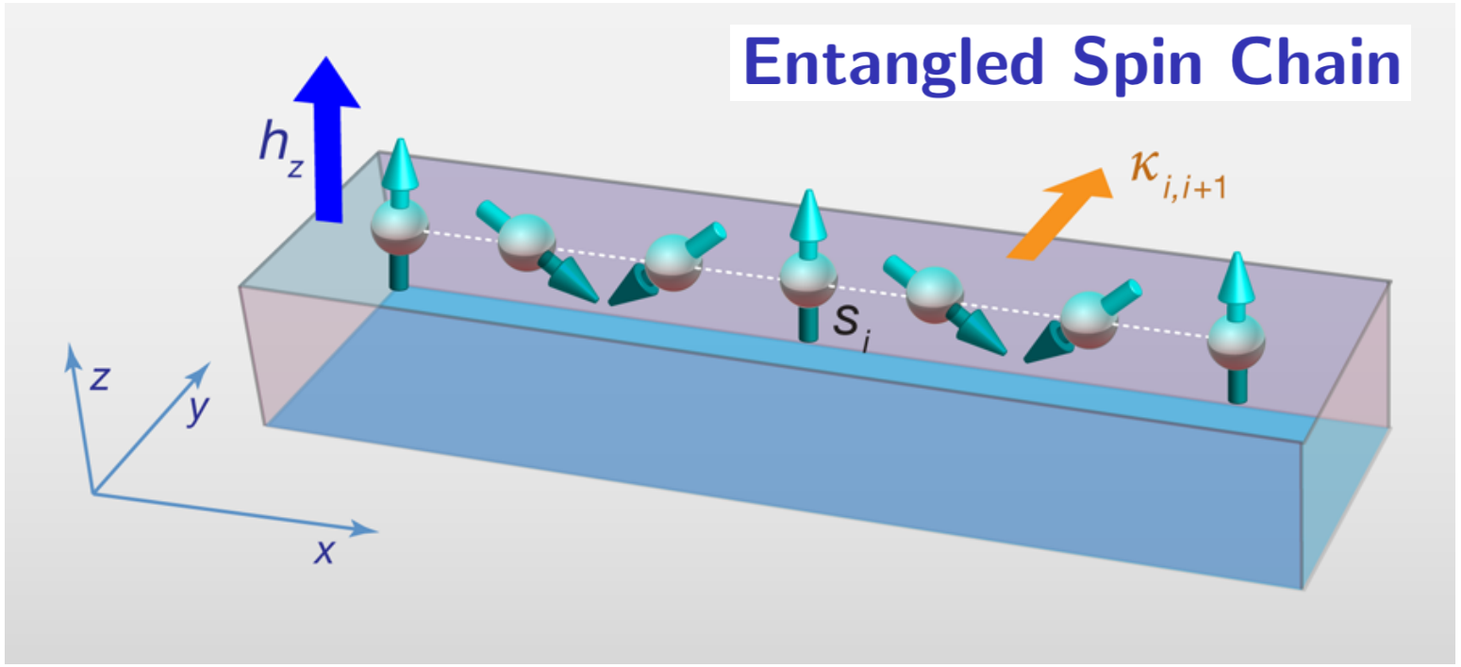
Recommender System

							.....
Lee	5	?	?	1	?	?	
Jin	?	?	4	?	?	?	
Yang	?	4	?	?	?	2	
Xu	?	?	3	?	5	?	
Zhang	3	4	?	?	5	?	
⋮							
⋮							

(Candes and Recht, 2009)

Examples of Matrix Completion

Quantum State Tomography



Density Matrix for b-Qubits Systems

A b-qubits system has density matrix $\rho \in \mathcal{S}_m$ with $m = 2^b$.

$$\rho \in \mathcal{S}_m := \left\{ S \in \mathbb{C}^{m \times m} : S = S^\dagger, S \succcurlyeq 0, \mathrm{tr}(S) = 1 \right\}.$$

Measurements

Each measurement is a Pauli matrix.

There are $m = 4^b$ Pauli matrices forming an orthonormal basis.

A generalization of matrix completion.

Goal

How to use a few Pauli measurements to learn the unknown density matrix ?

(Gross, 2011)

Statistical Model

An unknown low-rank matrix $M \in \mathbb{R}^{d_1 \times d_2}$ *with rank* $r \ll d_2 \leq d_1$.

A random pair (ω, Y) *satisfying*

$$Y = M(\omega) + \xi$$

(i) ω *is sampled uniformly from* $[d_1] \times [d_2]$

(ii) ξ *is centered sub-Gaussian with variance* σ_ξ^2 *and independent with* ω .

The data $\mathcal{D}_n$ *consists of* n *i.i.d. copies of* (ω, Y) .

$$\mathcal{D}_n = \{(\omega_i, Y_i) : i = 1, \cdots, n\}$$

The goal is to recover M *from* $\mathcal{D}_n$.

Incoherent Conditions

9	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Irregular settings
(impossible to reconstruct)

Singular Value Decomposition

$$M = U \Lambda V^\top$$

$$U^\top U = V^\top V = I_r \qquad \Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_r)$$

Incoherence Assumption

$$\max_{j \in [d_1]} \|e_j^\top U\| \le \mu_{\max} \sqrt{\frac{r}{d_1}} \quad \text{and} \quad \max_{j \in [d_2]} \|e_j^\top V\| \le \mu_{\max} \sqrt{\frac{r}{d_2}}$$

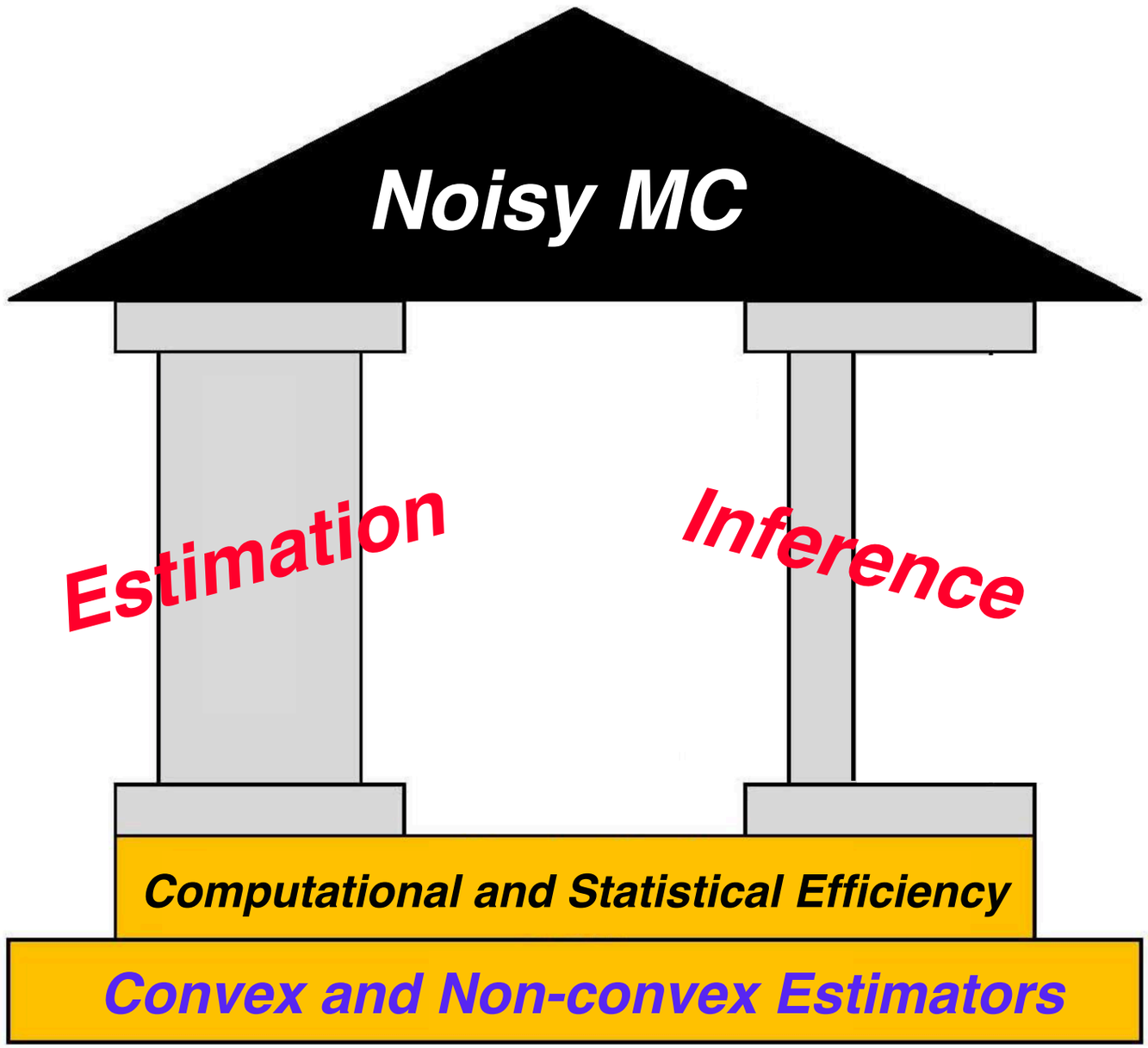
or $\sqrt{d_1 d_2} \|M\|_{\max} / \|M\|_{\mathrm{F}} \le \mu_{\max}$

Outline

- *Motivation and Statistical Model*
- ***Prior Works: (much) Estimation and (few) Inference***
- *Statistical Inference of Linear Forms*
- *Methodology: Double Sample-Splitting and Projection*
- *Theory: Data-driven Asymptotical Normality*
- *Numerical Experiments*

Two Pillars for High Dimensional Statistics:

Estimation and Inference



Rank Constrained or Penalized LSE

Rank-constrained Least Squares Estimator

$$\text{Minimize } \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - M(\omega_i)\right)^2$$
$$\text{Subject to: } \operatorname{rank}(M) \le r$$

Generally intractable.

Rank-penalized Least Squares Estimator

$$\text{Minimize } \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - M(\omega_i)\right)^2 + \operatorname{rank}(M)$$

Generally intractable.

Convex Relaxation

Minimize matrix nuclear norm

$$\text{Minimize } \|A\|_*$$

$$\text{Subject to } \left\| \sum_{i=1}^{n} \big(Y_i - A(\omega_i)\big) e_{\omega_i} \right\|_{\text{F}} \leq \delta_n$$

(Candes and Plan, 2010)

Theorem (Candes and Plan, 2010)

If $n \gg r d_1 \log d_1$ and incoherence conditions hold, then

$$\|\hat{M} - M\|_{\text{F}} = O_P\left(\delta_n \cdot \sqrt{\frac{d_1 d_2^2}{n}}\right)$$

Statistically sub-optimal in general

Convex Relaxation

Matrix nuclear norm penalized (modified) LSE

$$\hat{M}^{\mathrm{KLT}} := \arg \min_{\|M\|_{\max} \le a} \|M\|^2_{L_2(\Pi)} - \frac{2}{n} \sum_{i=1}^{n} Y_i M(\omega_i) + \varepsilon \|M\|_*$$

(Koltchinskii, Lounici and Tsybakov, 2011)

Theorem (Koltchinskii, Lounici and Tsybakov, 2011)

If $n \gg r d_1 \log d_1$, then without incoherent conditions

$$\|\hat{M}^{\mathrm{KLT}} - M\|_{\mathrm{F}} = O_P\left((\sigma_\xi \vee a) \sqrt{\frac{r d_1 d_2 (d_1 + d_2) \log d_1}{n}} \right)$$

Rate is minimax optimal
(but not proportional to noise)

Non-convex Approaches

Gradient Descent on Grassmannian

$$\text{Minimize } f(U,V) = \min_{G\in\mathbb{R}^{r\times r}} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - (UGV^\top)(\omega_i)\right)^2$$

$$\text{Subject to } U^\top U = V^\top V = I_r$$

(Keshavan, Montanari and Oh, 2010)

Theorem (Keshavan, Montanari and Oh, 2010)

If $n \gg rd_1 \log d_1$ and incoherent conditions hold and signal-to-noise ratio (SNR) satisfy $\lambda_r \gg \sigma_\xi \sqrt{d_1^2 d_2 \log(d_1)/n}$, then $f(U,V)$ behaves like a parabola around the truth. Initialized by spectral methods, gradient descent yields an estimator

$$\|\hat{M}^{\mathrm{KMO}} - M\|_{\mathrm{F}} = O_P\left(\sigma_\xi \sqrt{\frac{rd_1d_2(d_1+d_2)\log d_1}{n}}\right)$$

Rate is minimax optimal (and proportional to noise)

Non-convex Approaches

Gradient Descent for symmetric matrix factorization $(M = M^\top)$

$$\text{Minimize } f(X) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - (XX^\top)(\omega_i)\right)^2$$

$$X^{t+1} = X^t - \eta_t \nabla f(X^t)$$

(Ma, Wang, Chi and Chen, 2019)

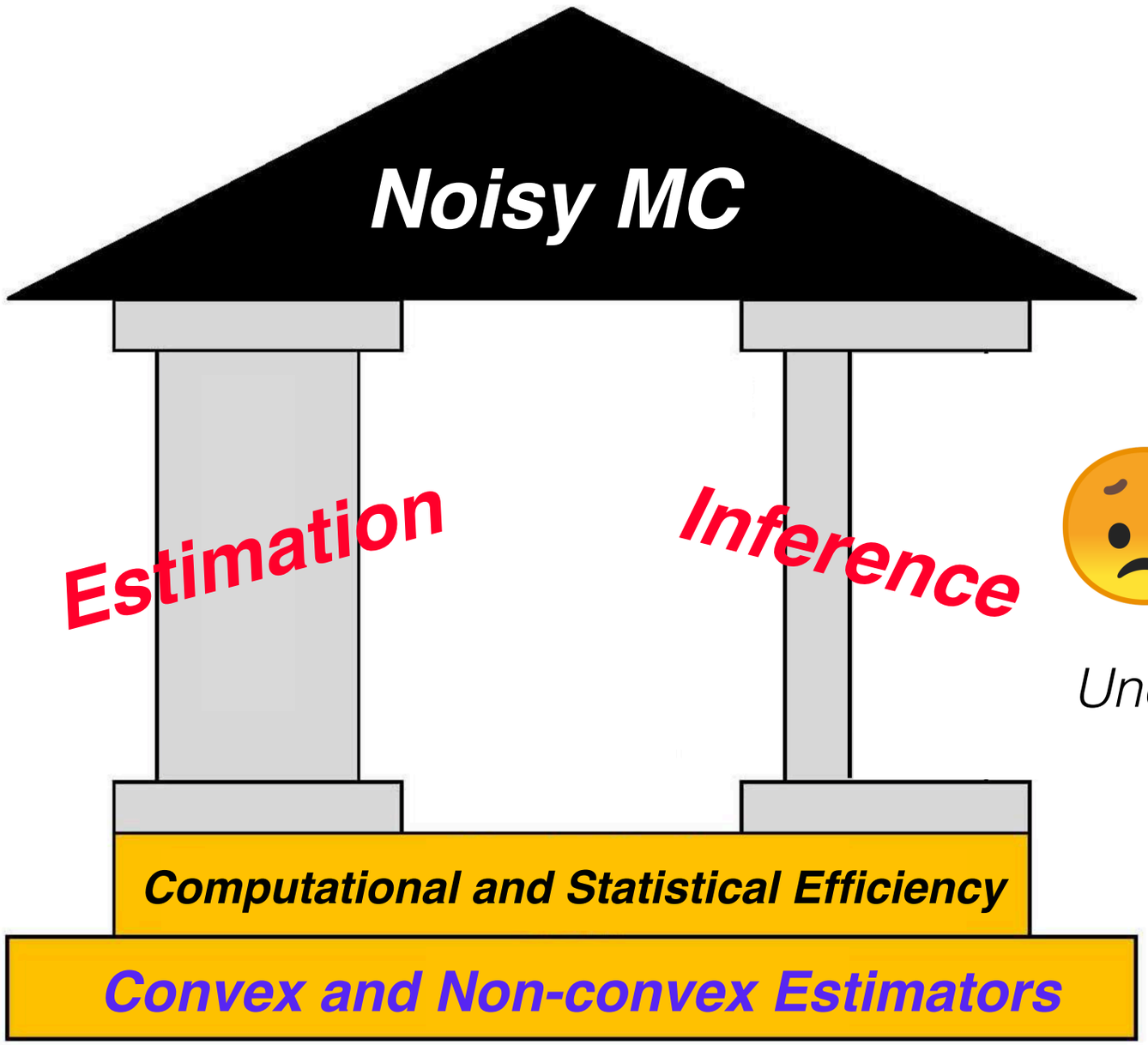
Theorem (Ma, Wang, Chi and Chen, 2017)

If $n \gg rd_1 \log d_1$ and incoherent conditions hold and signal-to-noise ratio (SNR) satisfy $\lambda_r \gg \sigma_\xi \sqrt{d_1^2 d_2 \log(d_1)/n}$, then gradient descent converges geometrically if initialized by spectral methods. The output estimator achieves rate

$$\|\hat{M}^{\mathrm{MWC}} - M\|_{\max} = O_P\left(\sigma_\xi \sqrt{\frac{rd_1 \log d_1}{n}}\right)$$

Rate is minimax optimal

Statistical Inferences



Uncertainty quantification

Statistical Inferences

De-biasing by sample splitting and confidence region w.r.t. Frobenius norm

$$\hat{R}_n = \frac{2}{n} \sum_{i \le n/2} \left(Y_i - \hat{M}^{\mathrm{KLT}}(\omega_i)\right)^2 - \sigma_\xi^2$$

$\hat{M}^{\mathrm{KLT}}$ *is computed from another half sample.*

(Carpentier, Klopp, Loffler and Nickl, 2019)

Theorem (Carpentier, Klopp, Loffler and Nickl, 2019)

If $n \gg r d_1 \log d_1$ and define

$$\mathcal{C}_n := \left\{ A \in \mathbb{R}^{d_1 \times d_2} : \frac{\|A - \hat{M}^{\mathrm{KLT}}\|_{\mathrm{F}}^2}{d_1 d_2} \le 2\left(\hat{R}_n + z\frac{d}{n} + \frac{\bar{z} + \xi_{\alpha,\sigma_\xi,a}}{\sqrt{n}} \right) \right\}$$

Then,

$$\mathbb{P}\left(M \in \mathcal{C}_n\right) \ge 1 - \frac{2\alpha}{3} - 2e^{-zd/(11a^2)}$$

not exact confidence interval

Statistical Inferences

De-biasing by linear regression technique and entry-wise confidence interval

$$(\Omega, \eta_n) = \arg\min_{\Omega,\eta}\{\eta : \|\mathcal{X}^*\mathcal{X}\Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \le \eta \ \forall 1 \le i \le p\}.$$

$$\tilde{M} := \hat{M} + \Omega\mathcal{X}^*(Y - \mathcal{X}(\hat{M})).$$

(Cai, Liang and Rakhlin, 2016)

Theorem (Cai, Liang and Rakhlin, 2016)

If $n \gg d_1 d_2$, then

$\tilde{M}_{ij} - M_{ij}$ follows approximately a centered normal distribution.

sample size requirement is too strong

Outline

- *Motivation and Statistical Model*
- *Prior Works: (much) Estimation and (few) Inference*
- ***Statistical Inference of Linear Forms***
- *Methodology: Double Sample-Splitting and Projection*
- *Theory: Data-driven Asymptotical Normality*
- *Numerical Experiments*

Statistical Inferences

Unsolved Fundamental Issues

Task 1 $H_0 : M_{ij} = 3 \quad \text{v.s.} \quad H_1 : M_{ij} > 3$

whether it worths to recommend item j to user i ?

Task 2 $H_0 : M_{ij} = M_{ik} \quad \text{v.s.} \quad H_1 : M_{ij} > M_{ik}$

should recommend item j or item k to user i ?

Statistical Inferences

Unsolved Fundamental Issues

Task 3

$$H_0 : M_{ij} = 2M_{ik} \quad \text{v.s.} \quad H_1 : M_{ij} > 2M_{ik}$$

If recommending item j costs 2 times of recommending item k, is it worth to recommend item j?

Task 4

$$H_0 : M_{i_1 j} + M_{i_2 j} + M_{i_3 j} = M_{i_1 k} + M_{i_2 k} + M_{i_3 k}$$

$$H_1 : M_{i_1 j} + M_{i_2 j} + M_{i_3 j} > M_{i_1 k} + M_{i_2 k} + M_{i_3 k}$$

should recommend item j or item k to a group of users $\{i_1, i_2, i_3\}$*?*

Why difficult ?

Statistical Bias Variance and bias trade-off on entries

$\mathbb{E}\widehat{M}_{ij}=M_{ij}$ but large $\mathrm{var}(\widehat{M}_{ij})$ ***V.S.*** $\mathbb{E}\widehat{M}_{ij}\neq M_{ij}$ but small $\mathrm{var}(\widehat{M}_{ij})$

Dependence across entries

$\widehat{M}_{i_1j_1}$ and $\widehat{M}_{i_2j_2}$ are dependent with non-linear relationship.

Technically difficult

Lack of explicit representation formulas for $\widehat{M}_{ij}$

Our Contributions

A general framework for testing hypothesis of linear forms

$$H_0 : \langle M, T \rangle = a \quad \text{v.s.} \quad H_1 : \langle M, T \rangle > a$$

Task 1 $T = e_i e_j^\top, \quad a = 3$

Task 2 $T = e_i e_j^\top - e_i e_k^\top, \quad a = 0$

Task 3 $T = e_i e_j^\top - 2 e_i e_k^\top, \quad a = 0$

Task 4 $T = e_{i_1} e_j^\top + e_{i_2} e_j^\top + e_{i_3} e_j^\top - e_{i_1} e_k^\top - e_{i_2} e_k^\top - e_{i_3} e_k^\top$
$, a = 0$

Outline

- *Motivation and Statistical Model*
- *Prior Works: (much) Estimation and (few) Inference*
- *Statistical Inference of Linear Forms*
- ***Methodology: Double Sample-Splitting and Projection***
- *Theory: Data-driven Asymptotical Normality*
- *Numerical Experiments*

Estimating Procedure

Double-sample-splitting
(Chernozhukov et al. 2018)

$$n = 2n_0$$

$$\mathcal{D}_1 = \{(X_i, Y_i)\}_{i=1}^{n_0} \quad \text{and} \quad \mathcal{D}_2 = \{(X_i, Y_i)\}_{i=n_0+1}^{n}$$

Step 1 *(Initialization)*

Apply an initial estimating procedure on $\mathcal{D}_1$ and $\mathcal{D}_2$ separately to yield $\widehat{M}_1^{\text{init}}$ and $\widehat{M}_2^{\text{init}}$.

Step 2 *(Debiasing)*

Use the second data sample $\mathcal{D}_2$ to debias $\widehat{M}_1^{\text{init}}$

$$\widehat{M}_1^{\text{unbs}} = \widehat{M}_1^{\text{init}} + \frac{d_1 d_2}{n_0} \sum_{i=n_0+1}^{n} \big(Y_i - \widehat{M}_1^{\text{init}}(\omega_i)\big) e_{\omega_i}$$

Similarly, $\widehat{M}_2^{\text{unbs}} = \widehat{M}_2^{\text{init}} + \frac{d_1 d_2}{n_0} \sum_{i=1}^{n_0} \big(Y_i - \widehat{M}_2^{\text{init}}(\omega_i)\big) e_{\omega_i}$

Step 3 *(Projection)*

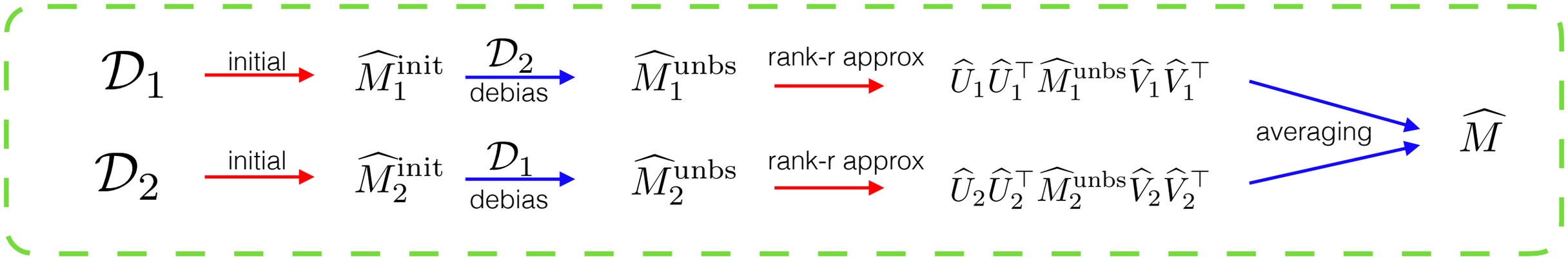
$$\widehat{M} = \frac{1}{2}\widehat{U}_1\widehat{U}_1^\top \widehat{M}_1^{\text{unbs}} \widehat{V}_1\widehat{V}_1^\top + \frac{1}{2}\widehat{U}_2\widehat{U}_2^\top \widehat{M}_2^{\text{unbs}} \widehat{V}_2\widehat{V}_2^\top$$

$\widehat{U}_k$ and $\widehat{V}_k$ are $\widehat{M}_k^{\text{unbs}}$'s top-r left and right singular vectors.

Step 4 *(Plug-in)*

Estimate $m_T = \langle M, T \rangle$ by $\hat{m}_T = \langle \widehat{M}, T \rangle$

Estimating Procedure



Double-sample-splitting avoids loss of statistical efficiency from sample splitting

$$\mathbb{E}\widehat{M}_1^{\text{unbs}} = \mathbb{E}\widehat{M}_2^{\text{unbs}} = M$$

$$\mathbb{E}\widehat{M} \neq M$$

Estimating Procedure

Remarks

$\mathrm{rank}(\widehat{M}) > r$?

Compute the best rank-r approximation

$\mathbb{E}\widehat{M}_1^{\mathrm{unbs}} = \mathbb{E}\widehat{M}_2^{\mathrm{unbs}} = M$

Unbiased but has large variance

$$\mathrm{sd}\Big((\widehat{M}_1^{\mathrm{unbs}})_{ij}\Big) \asymp \sigma_\xi \cdot \frac{d_1 d_2}{n}$$

$\mathbb{E}\widehat{M} \neq M$

Negligible bias with small variance

$$\mathrm{sd}\Big(\widehat{M}_{ij}\Big) \asymp \sigma_\xi \cdot \sqrt{\frac{r d_1}{n}}$$

$$\mathrm{bias}\Big(\widehat{M}_{ij}\Big) = o\Big(\sigma_\xi \cdot \sqrt{\frac{r d_1}{n}}\Big)$$

Outline

- *Motivation and Statistical Model*
- *Prior Works: (much) Estimation and (few) Inference*
- *Statistical Inference of Linear Forms*
- *Methodology: Double Sample-Splitting and Projection*
- ***Theory: Data-driven Asymptotical Normality***
- *Numerical Experiments*

Asymptotic Normality of $\mathrm{tr}(\widehat{M}^\top T)$

Assumption on initialization

There exists a sequence $\gamma_{n,d_1,d_2} \to 0$ as $n, d_1, d_2 \to \infty$ so that with probability at least $1 - d_1^{-2}$

$$\|\widehat{M}_1^{\mathrm{init}} - M\|_{\max} + \|\widehat{M}_2^{\mathrm{init}} - M\|_{\max} \le C\gamma_{n,d_1,d_2} \cdot \sigma_\xi$$

The dependence on initialization is fairly weak.

The rate γ_{n,d_1,d_2} ***needs not to be optimal.***

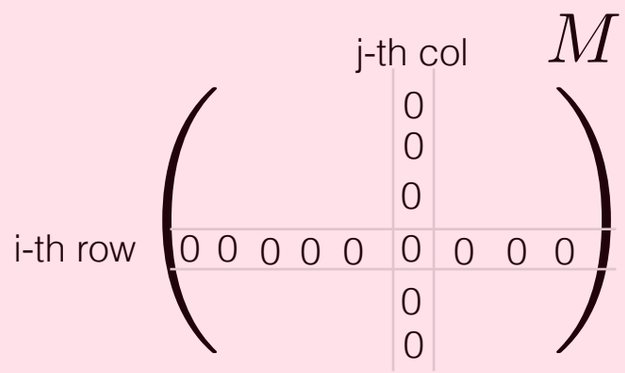
The resulting estimates $\mathrm{tr}(\widehat{M}^\top T)$ ***are asymptotically equivalent as long as*** $\gamma_{n,d_1,d_2} = o(1)$ ***.***

Asymptotic Normality of $\mathrm{tr}(\widehat{M}^\top T)$

Goal Establish distribution of $\mathrm{tr}(\widehat{M}^\top T) - \mathrm{tr}(M^\top T)$

Consider $T = e_i e_j^\top$ and $\mathrm{tr}(M^\top T) = M_{ij}$

Irregular case $e_i^\top U = e_j^\top V = \mathbf{0}$



All observed entries contributing information to M_{ij} are pure noise

Therefore, a natural requirement is $\|e_i^\top U\| + \|e_j^\top V\| \gg 0$

Asymptotic Normality of $\mathrm{tr}(\widehat{M}^\top T)$

General regularity assumption

There exists a constant $\alpha_T > 0$ such that

$$\|U^\top T\|_{\mathrm{F}} \geq \alpha_T \|T\|_{\mathrm{F}} \cdot \sqrt{\frac{r}{d_1}} \quad \textbf{or} \quad \|TV\|_{\mathrm{F}} \geq \alpha_T \|T\|_{\mathrm{F}} \cdot \sqrt{\frac{r}{d_2}}$$

The alignment parameter α_T is allowed to vanish as $n, d_1, d_2 \to \infty$.
In fact, the asymptotical theory requires

$$\alpha_T \geq \frac{\|T\|_{\ell_1}}{\|T\|_{\mathrm{F}}} \cdot \max\left\{\sqrt{\frac{r \log d_1}{d_2}}, \quad \frac{\sigma_\xi}{\lambda_r}\sqrt{\frac{r d_1^2 d_2 \log^2 d_1}{n}}\right\}$$

Asymptotic Normality of $\mathrm{tr}(\widehat{M}^\top T)$

Sample size requirements $\quad n \geq Crd_1 \log d_1$

Sub-Gaussian noise $\quad \mathbb{E}\xi = 0,\ \mathbb{E}\xi^2 = \sigma_\xi^2 \quad \text{and} \quad \mathbb{E}e^{s\xi} \leq e^{s^2\sigma_\xi^2},\ \forall s \in \mathbb{R}$

Signal-to-noise ratio (SNR)

$$\lambda_r \geq C\sigma_\xi \sqrt{\frac{rd_1^2 d_2 \log d_1}{n}}$$

The sample size and SNR conditions are almost optimal.

Asymptotic Normality of $\mathrm{tr}(\widehat{M}^\top T)$

Theorem (X and Yuan, 2019)

Under the aforesaid sample size and SNR conditions, and alignment condition, and initialization condition, then

$$\sup_x \left| \mathbb{P}\left(\frac{\mathrm{tr}(\widehat{M}^\top T) - \mathrm{tr}(M^\top T)}{\sigma_\xi (\|U^\top T\|_{\mathrm{F}}^2 + \|TV\|_{\mathrm{F}}^2)^{1/2} \cdot \sqrt{d_1 d_2 / n}} \le x \right) - \Phi(x) \right|$$

$$\lesssim \frac{\|T\|_{\ell_1}}{\alpha_T \|T\|_{\mathrm{F}}} \cdot \max\left\{ \sqrt{\frac{r \log d_1}{d_2}},\ \frac{\sigma_\xi}{\lambda_r} \sqrt{\frac{r d_1^2 d_2 \log^2 d_1}{n}} \right\}$$

$$+ \frac{\log d_1}{d_1^2} + \gamma_{n,d_1,d_2} \sqrt{\log d_1} + \sqrt{\frac{r d_1}{n}}$$

$\Phi(x)$ denotes the c.d.f. of standard normal distribution.

Asymptotic Normality of $\mathrm{tr}(\widehat{M}^\top T)$

$$\mathrm{tr}(\widehat{M}^\top T) - \mathrm{tr}(M^\top T)$$

$$\stackrel{\mathrm{d}}{=} \sigma_\xi(\|U^\top T\|_{\mathrm{F}}^2 + \|TV\|_{\mathrm{F}}^2)^{1/2}\sqrt{\frac{d_1 d_2}{n}} \cdot \left(Z + O_P(\gamma_{n,d_1,d_2}\sqrt{\log d_1})\right)$$

where $Z \sim \mathcal{N}(0,1)$

Dependence on Initialization $\gamma_{n,d_1,d_2}\sqrt{\log d_1} = o(1)$

For instance, if $\gamma_{n,d_1,d_2} = \sqrt{\frac{r d_1 \log d_1}{n}}$, then it suffices to require

$$n \gg r d_1 \log^2 d_1$$

Asymptotic Normality of $\mathrm{tr}(\widehat{M}^\top T)$

Example 1: $T = e_i e_j^\top$

Theorem (X and Yuan, 2019)

$$\frac{\widehat{M}_{ij} - M_{ij}}{\sigma_\xi (\|U^\top e_i\|^2 + \|V^\top e_j\|^2)^{1/2} \sqrt{d_1 d_2 / n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

If $n \gg r d_1 \log d_1$ and $\|U^\top e_i\| + \|V^\top e_j\| \geq \alpha_T \sqrt{\frac{r}{d_1}}$

$$\frac{1}{\alpha_T} \cdot \max\left\{ \sqrt{\frac{r \log d_1}{d_2}}, \frac{\sigma_\xi}{\lambda_r} \sqrt{\frac{r d_1^2 d_2 \log d_1}{n}} \right\} + \gamma_{n,d_1,d_2} \sqrt{\log d_1} \longrightarrow 0$$

It implies that

$$\mathbb{E}\|\widehat{M} - M\|_{\mathrm{F}}^2 = \big(1 + o(1)\big) \cdot \frac{\sigma_\xi^2 r d_1 d_2 (d_1 + d_2)}{n}$$

Asymptotic Normality of $\mathrm{tr}(\widehat{M}^\top T)$

Example 2: $T = e_i e_j^\top - e_i e_k^\top$

Theorem (X and Yuan, 2019)

$$\frac{(\widehat{M}_{ij} - \widehat{M}_{ik}) - (M_{ij} - M_{ik})}{(2\|U^\top e_i\|^2 + \|V^\top (e_j - e_k)\|^2)^{1/2} \cdot \sigma_\xi \sqrt{d_1 d_2 / n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

under similar sample size, SNR, alignment conditions

Asymptotic Normality of $\mathrm{tr}(\widehat{M}^\top T)$

Example 3: $T = e_i e_j^\top - 2e_i e_k^\top$

Theorem (X and Yuan, 2019)

$$\frac{(\widehat{M}_{ij} - 2\widehat{M}_{ik}) - (M_{ij} - 2M_{ik})}{(5\|U^\top e_i\|^2 + \|V^\top(e_j - 2e_k)\|^2)^{1/2} \cdot \sigma_\xi \sqrt{d_1 d_2 / n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

under similar sample size, SNR, alignment conditions

Asymptotic Normality of $\mathrm{tr}(\widehat{M}^\top T)$

Example 4: $\|T\|_{\ell_1}/\|T\|_{\mathrm{F}} \le s_0$

Theorem (X and Yuan, 2019)

$$\frac{\mathrm{tr}(\widehat{M}^\top T) - \mathrm{tr}(M^\top T)}{\sigma_\xi(\|U^\top T\|_{\mathrm{F}}^2 + \|TV\|_{\mathrm{F}}^2)^{1/2} \cdot \sqrt{d_1 d_2/n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

If $n \gg r d_1 \log d_1$ and $\|U^\top T\|_{\mathrm{F}} + \|TV\|_{\mathrm{F}} \ge \alpha_T \sqrt{\frac{r}{d_1}} \|T\|_{\mathrm{F}}$

$$\frac{\sqrt{s_0}}{\alpha_T} \cdot \max\left\{\sqrt{\frac{r\log d_1}{d_2}}, \frac{\sigma_\xi}{\lambda_r}\sqrt{\frac{r d_1^2 d_2 \log d_1}{n}}\right\} + \gamma_{n,d_1,d_2}\sqrt{\log d_1} \longrightarrow 0$$

In the case that T is well aligned with M in that $\alpha_T \asymp \sqrt{d_1/r}$ then our theory allows sparsity $s_0 = O(d_2)$

Inference about $\mathrm{tr}(M^\top T)$

Data-driven Estimates

$$\hat{\sigma}_\xi^2 = \frac{1}{2n_0} \sum_{i=n_0+1}^{n} \big(Y_i - \widehat{M}_1^{\mathrm{init}}(\omega_i)\big)^2 + \frac{1}{2n_0} \sum_{i=1}^{n_0} \big(Y_i - \widehat{M}_2^{\mathrm{init}}(\omega_i)\big)^2$$

$$\hat{s}_T^2 = \frac{1}{2}\big(\|\widehat{U}_1^\top T\|_{\mathrm{F}}^2 + \|\widehat{U}_2^\top T\|_{\mathrm{F}}^2 + \|T\widehat{V}_1\|_{\mathrm{F}}^2 + \|T\widehat{V}_2\|_{\mathrm{F}}^2\big)$$

Theorem (X and Yuan, 2019)

$$\frac{\mathrm{tr}(\widehat{M}^\top T) - \mathrm{tr}(M^\top T)}{\hat{\sigma}_\xi \hat{s}_T \cdot \sqrt{d_1 d_2 / n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

under similar sample size, SNR, alignment conditions

Confidence interval of $\mathrm{tr}(M^\top T)$

under optimal sample size condition

The $100(1-\theta)\%$ confidence interval of $\mathrm{tr}(M^\top T)$ can be defined by

$$\widehat{\mathrm{CI}}_{\theta,T} = \left[\mathrm{tr}(\widehat{M}^\top T) - z_{\theta/2}\cdot\hat{\sigma}_\xi\hat{s}_T\sqrt{\frac{d_1 d_2}{n}},\quad \mathrm{tr}(\widehat{M}^\top T) + z_{\theta/2}\cdot\hat{\sigma}_\xi\hat{s}_T\sqrt{\frac{d_1 d_2}{n}}\right]$$

Then,

$$\lim_{n,d_1,d_2\to\infty}\mathbb{P}\Big(\mathrm{tr}(M^\top T)\in\widehat{\mathrm{CI}}_{\theta,T}\Big) = 1-\theta$$

Hypothesis testing for $\mathrm{tr}(M^\top T)$

$$H_0 : \mathrm{tr}(M^\top T) = 0 \quad \text{v.s.} \quad H_1 : \mathrm{tr}(M^\top T) \neq 0$$

A two-sided test with significance level θ is

Reject H_0 if $|\hat{z}| > z_{\theta/2}$

where

$$\hat{z} = \frac{\mathrm{tr}(\widehat{M}^\top T)}{\hat{\sigma}_\xi \hat{s}_T \cdot \sqrt{d_1 d_2 / n}}$$

Hypothesis testing for $\mathrm{tr}(M^\top T)$

$$H_0 : M_{ij} = M_{ik} \quad \text{v.s.} \quad H_1 : M_{ij} > M_{ik}$$

A one-sided test with significance level θ is

Reject H_0 if $\hat{z} > z_\theta$

where

$$\hat{z} = \frac{\sqrt{2}(\widehat{M}_{ij} - \widehat{M}_{ik})}{\hat{\sigma}_\xi(\|\widehat{V}_1^\top(e_j - e_k)\|^2 + \|\widehat{V}_2^\top(e_j - e_k)\|^2 + 2\|\widehat{U}_1^\top e_i\|^2 + 2\|\widehat{U}_2^\top e_i\|^2)^{1/2}\sqrt{d_1 d_2/n}}$$

Initialization

Assumption on initialization

$$\|\widehat{M}_1^{\text{init}} - M\|_{\max} + \|\widehat{M}_2^{\text{init}} - M\|_{\max} \le C\gamma_{n,d_1,d_2} \cdot \sigma_\xi \qquad \gamma_{n,d_1,d_2}\sqrt{\log d_1} \to 0$$

Theorem (Ma, Wang, Chi and Chen, 2017)

$$\|\hat{M}^{\text{MWC}} - M\|_{\max} = O_P\left(\sigma_\xi \sqrt{\frac{r d_1 \log d_1}{n}}\right)$$

for symmetric case,
sampling without replacement

The only requirement $\gamma_{n,d_1,d_2}\sqrt{\log d_1} \to 0$

Gradient Descent on Grassmann

Partition $\mathcal{D}$ into $2m = 2C_1\lceil \log d_1 \rceil$ subsets $\mathfrak{D}_t = \{(X_j, Y_j)\}_{j=(t-1)N_0+1}^{tN_0}, \quad \forall\, t = 1, \cdots, 2m$

Algorithm 1 Rotation Calibrated Gradient descent on Grassmannians

Let $\widehat{U}^{(1)}$ and $\widehat{V}^{(1)}$ be the top-r left and right singular vectors of $d_1 d_2 N_0^{-1} \sum_{j\in\mathfrak{D}_1} Y_j X_j$.

Compute $\widehat{G}^{(1)} = \arg\min_{G\in\mathbb{R}^{r\times r}} L(\mathfrak{D}_2, (\widehat{U}^{(1)}, G, \widehat{V}^{(1)}))$ and its SVD $\widehat{G}^{(1)} = \widehat{L}_G^{(1)} \widehat{\Lambda}^{(1)} \widehat{R}_G^{(1)\mathsf{T}}$.

for $t = 1, 2, 3, \cdots, m-1$ **do**

Update by rotation calibrated gradient descent

$$\widehat{U}^{(t+0.5)} = \widehat{U}^{(t)} \widehat{L}_G^{(t)} - \eta \cdot \frac{d_1 d_2}{N_0} \sum_{j\in\mathfrak{D}_{2t+1}} \big(\langle \widehat{U}^{(t)} \widehat{G}^{(t)} \widehat{V}^{(t)}, X_j\rangle - Y_j\big) X_j \widehat{V}^{(t)} \widehat{R}_G^{(t)} (\widehat{\Lambda}^{(t)})^{-1}$$

$$\widehat{V}^{(t+0.5)} = \widehat{V}^{(t)} \widehat{R}_G^{(t)} - \eta \cdot \frac{d_1 d_2}{N_0} \sum_{j\in\mathfrak{D}_{2t+1}} \big(\langle \widehat{U}^{(t)} \widehat{G}^{(t)} \widehat{V}^{(t)}, X_j\rangle - Y_j\big) X_j^{\mathsf{T}} \widehat{U}^{(t)} \widehat{L}_G^{(t)} (\widehat{\Lambda}^{(t)})^{-1}$$

Compute the top-r left singular vectors

$$\widehat{U}^{(t+1)} = \mathrm{SVD}(\widehat{U}^{(t+0.5)}) \quad \text{and} \quad \widehat{V}^{(t+1)} = \mathrm{SVD}(\widehat{V}^{(t+0.5)})$$

Compute $\widehat{G}^{(t+1)}$ by

$$\widehat{G}^{(t+1)} = \underset{G\in\mathbb{R}^{r\times r}}{\arg\min}\, L\big(\mathfrak{D}_{2t+2}, (\widehat{U}^{(t+1)}, G, \widehat{V}^{(t+1)})\big) \text{ and its SVD } \widehat{G}^{(t+1)} = \widehat{L}_G^{(t+1)} \widehat{\Lambda}^{(t+1)} \widehat{R}_G^{(t+1)\mathsf{T}}$$

end for

Output: $(\widehat{U}^{(m)}, \widehat{G}^{(m)}, \widehat{V}^{(m)})$ and $\widehat{M}^{(m)} = \widehat{U}^{(m)} \widehat{G}^{(m)} (\widehat{V}^{(m)})^{\mathsf{T}}$.

Gradient Descent on Grassmann

Theorem (X and Yuan, 2019)

$$\|\widehat{M}^m - M\|_{\max} = O_P\left(\sigma_\xi \sqrt{\frac{r^2 d_1 \log^2 d_1}{n}}\right)$$

Then, the initialization condition holds with

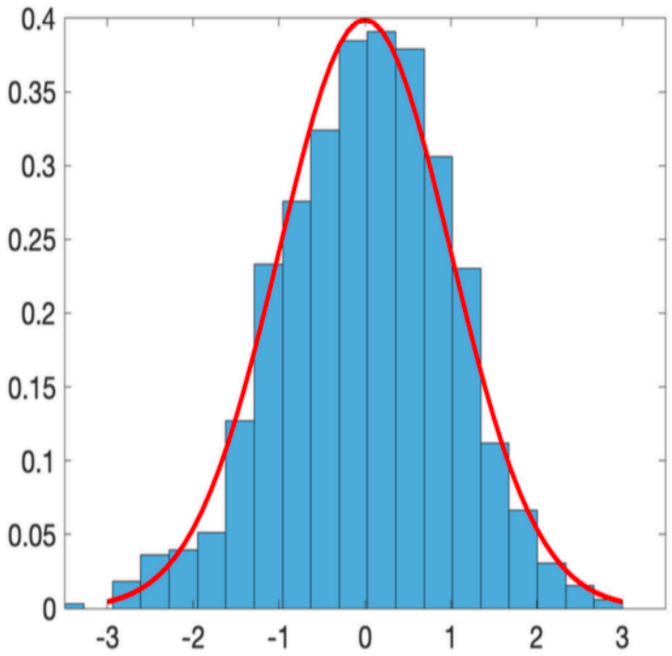
$$\gamma_n = \sqrt{\frac{r^2 d_1 \log^2 d_1}{n}}$$

Outline

- *Motivation and Statistical Model*
- *Prior Works: (much) Estimation and (few) Inference*
- *Statistical Inference of Linear Forms*
- *Methodology: Double Sample-Splitting and Projection*
- *Theory: Data-driven Asymptotical Normality*
- ***Numerical Experiments***

Numerical Experiments

Normal approximation of $\dfrac{\mathrm{tr}(\widehat{M}^\top T) - \mathrm{tr}(M^\top T)}{\hat{\sigma}_\xi \hat{s}_T \cdot \sqrt{d_1 d_2 / n}}$

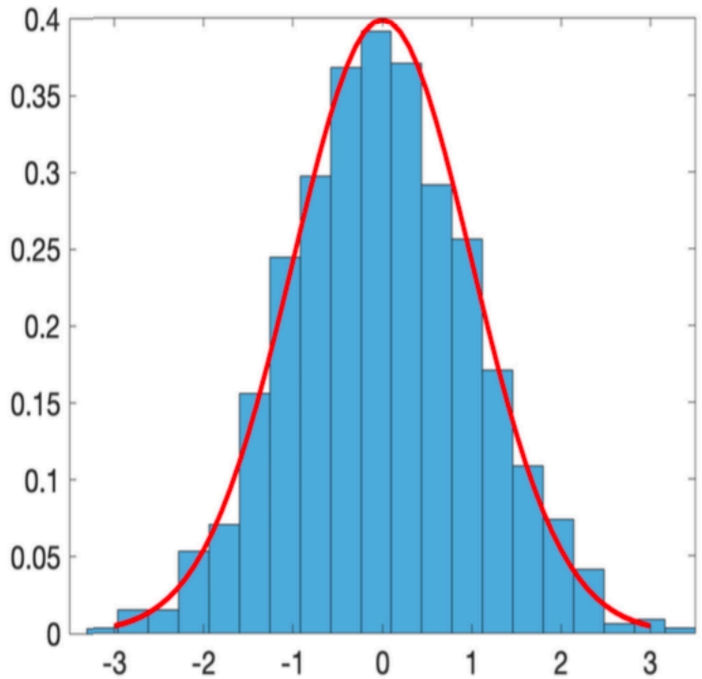


Gaussian noise

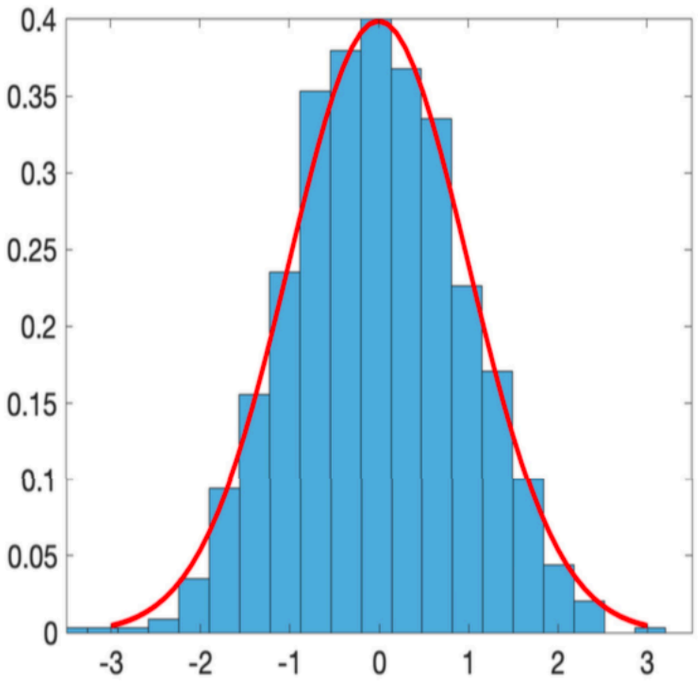
$d_1 = d_2 = d = 2000,$

$r = 3$

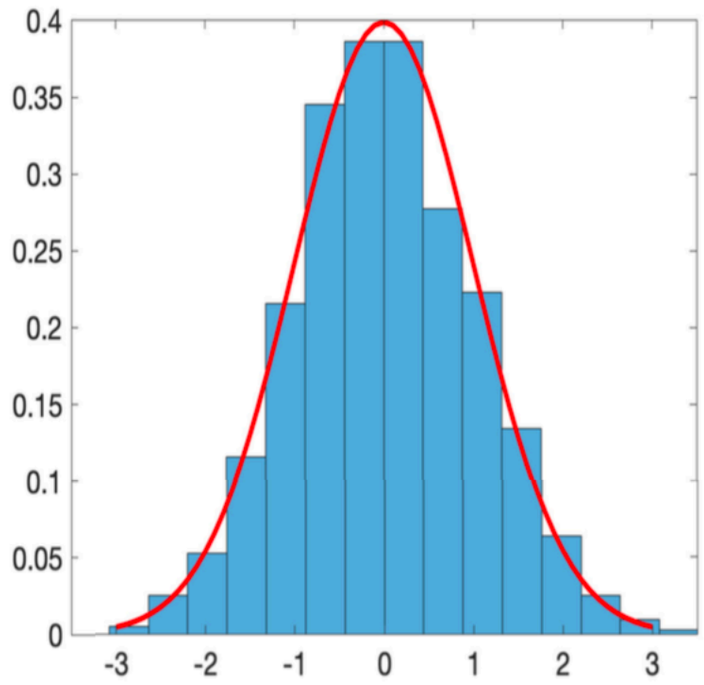
$n = 4r^2 d$



(a) $T = e_1 e_1^\top$



(b) $T = e_1 e_1^\top - e_1 e_2^\top$

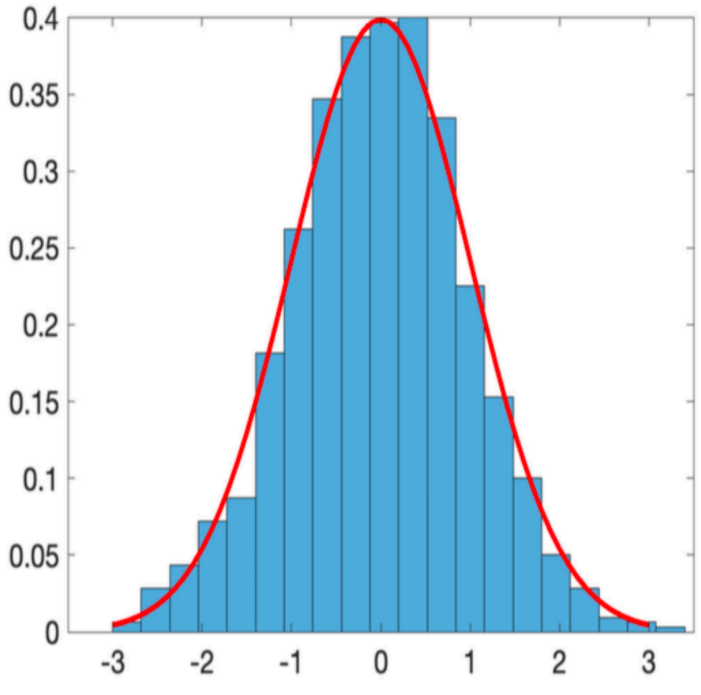


(c) $T = e_1 e_1^\top - e_1 e_2^\top + e_2 e_1^\top$

(d) $T = e_1 e_1^\top - e_1 e_2^\top + 2e_2 e_1^\top + 3e_2 e_2^\top$

Numerical Experiments

Normal approximation of $\dfrac{\mathrm{tr}(\widehat{M}^\top T) - \mathrm{tr}(M^\top T)}{\hat{\sigma}_\xi \hat{s}_T \cdot \sqrt{d_1 d_2 / n}}$

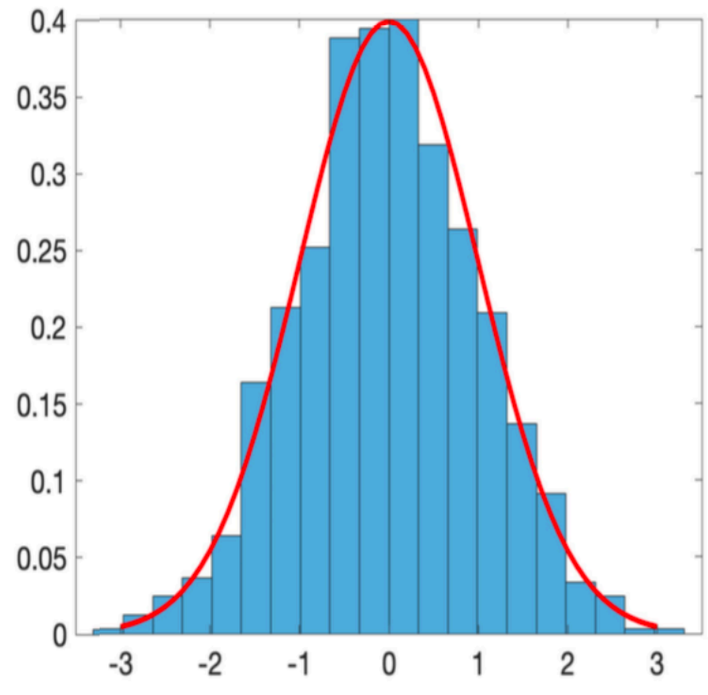


Uniform noise

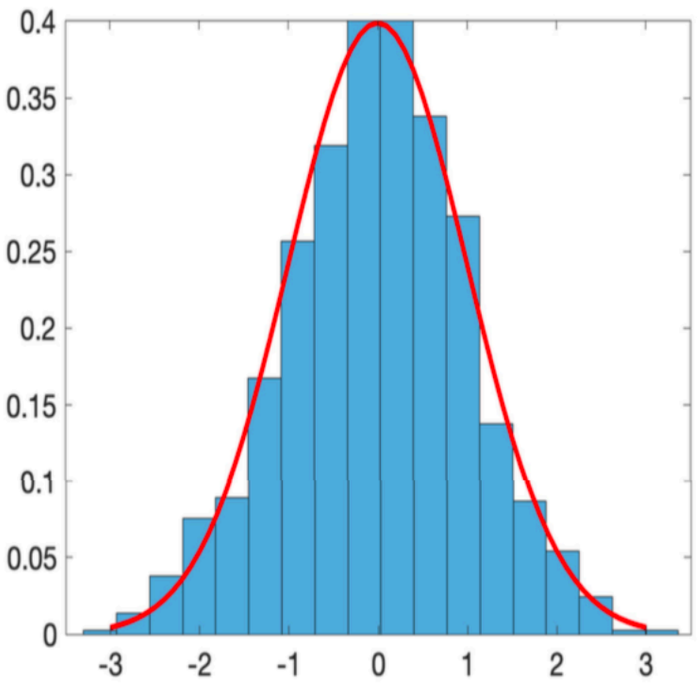
$d_1 = d_2 = d = 2000,$

$r = 3$

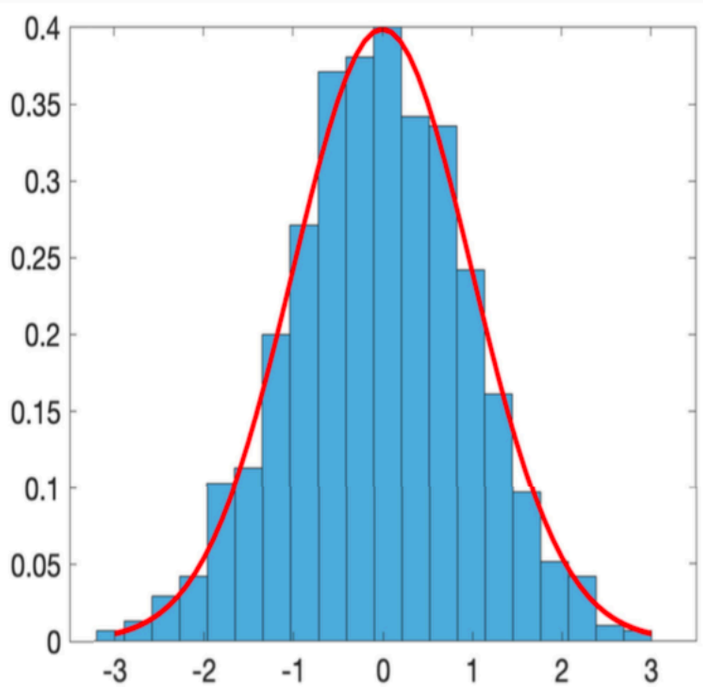
$n = 4r^2 d$



(a) $T = e_1 e_1^\top$



(b) $T = e_1 e_1^\top - e_1 e_2^\top$



(c) $T = e_1 e_1^\top - e_1 e_2^\top + e_2 e_1^\top$

(d) $T = e_1 e_1^\top - e_1 e_2^\top + 2e_2 e_1^\top + 3e_2 e_2^\top$

Conclusions

Any entry-wise consistent estimator can be used as initial estimate.

Double-sample-splitting + Spectral projection yields asymptotically normal estimators of linear forms.

Sample size requirement only depends on matrix rank and dimension.

Signal-to-noise ratio condition depends on
(i) sparsity of linear forms
(ii) Alignments of linear forms w.r.t. underlying singular spaces


Thank
You