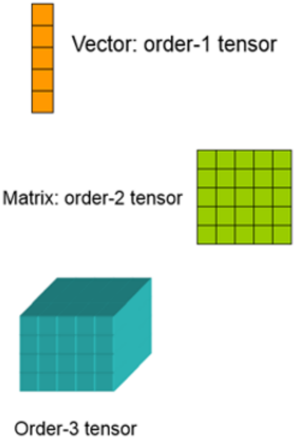
Sparse and Low-Rank Tensor Recovery via Cubic-Sketching

Botao Hao
Department of Statistics
Purdue University

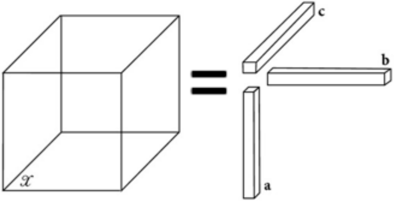
2017 International Conference on Data Science
Dec. 18, 2017

Joint work with Anru Zhang, and Guang Cheng

Tensor: Multi-dimensional Array



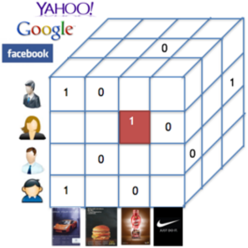
Rank-one Tensor



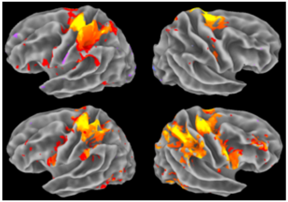
Tensor Data Example

Color image

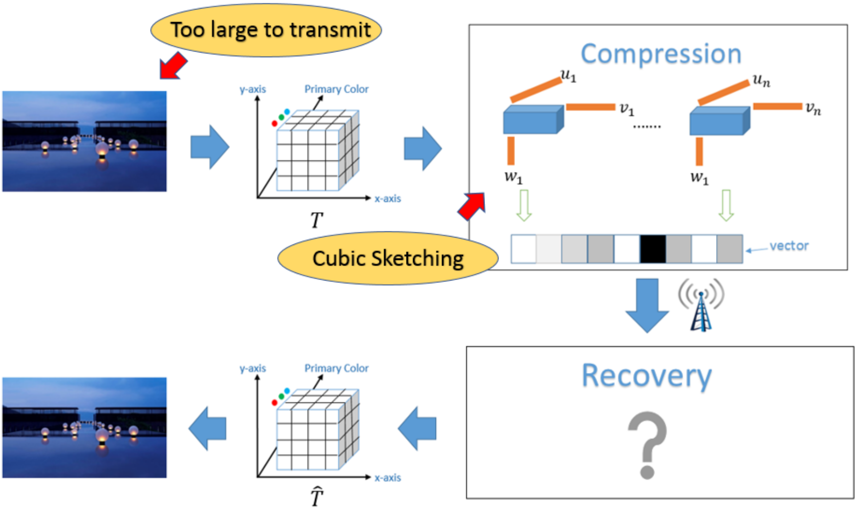
Advertisement

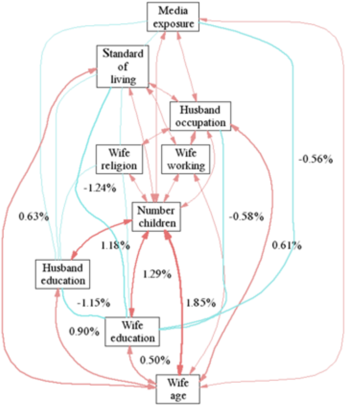


fMRI



Motivation: Compressed Image Transmission





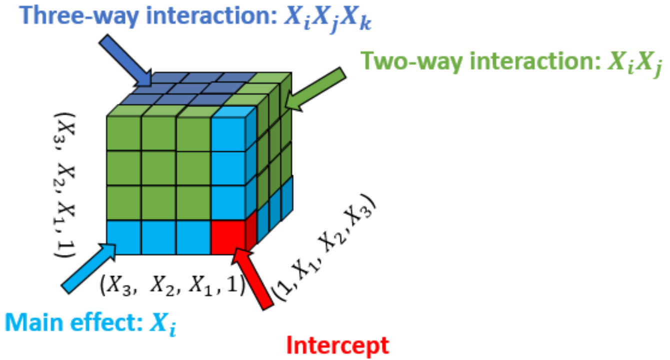
covariate

$$\mathbb{E}(y|\boldsymbol{X}) = \beta_0 + \sum_{i=1}^{p} X_i\beta_i + \sum_{i,j=1}^{p} \gamma_{ij}X_iX_j + \sum_{i,j,k=1}^{p} \eta_{ijk}X_iX_jX_k$$

Intercept | Main effect | Pairwise interaction | Triple-wise interaction

source: Contraceptive Method Choice dataset from UCI

Motivation: Interaction Effect Model



 $=(1, X_1, X_2, X_3) \circ (1, X_1, X_2, X_3) \circ (1, X_1, X_2, X_3) \in R^{p\times p\times p}$

Sparse and Low-Rank Tensor Recovery

Noisy Cubic Sketching Model

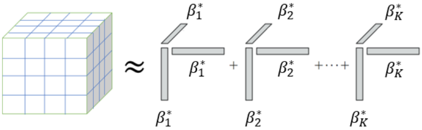
- Observe $\{y_i, \mathscr{X}_i\}$ from noisy cubic sketching model,

$$\underbrace{y_i}_{\text{scalar}} = \underbrace{\langle \mathscr{T}^*, \mathscr{X}_i \rangle}_{\text{tensor inner product}} + \underbrace{\epsilon_i}_{\text{noise}}, \quad i = 1, \ldots, n.$$

- Goal: Recover unknown third-order tensor parameter $\mathscr{T}^*$.

Key Assumptions on Tensor Parameter

- When $\mathscr{T}^* \in \mathbb{R}^{p\times p\times p}$ is a symmetric tensor...
  1. CANDECOMP/PARAFAC(CP) low-rank:



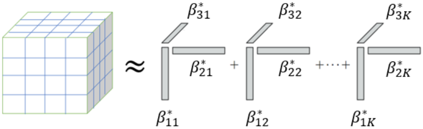
$$\mathscr{T}^* = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*, \text{ with } \|\boldsymbol{\beta}_k^*\|_2 = 1$$

  2. Sparse components: $\|\boldsymbol{\beta}_k^*\|_0 \leq s$ for $k \in [K]$.

- The cubic sketching tensor $\mathscr{X}_i$ for symmetric case is $\mathscr{X}_i = \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i$, where $\{\boldsymbol{x}_i\}_{i=1}^n$ are Gaussian random vectors.
- $\boldsymbol{\beta}_k^*$ and $\boldsymbol{\beta}_{k'}^*$ are not orthogonal. *Different from eigenvalue decomposition in matrix case.*

Key Assumptions on Tensor Parameter

- When $\mathscr{T}^* \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ is a non-symmetric tensor...
  1. CANDECOMP/PARAFAC(CP) low-rank:



$$\mathscr{T}^* = \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}^*_{1k} \circ \boldsymbol{\beta}^*_{2k} \circ \boldsymbol{\beta}^*_{3k}, \text{ with } \|\boldsymbol{\beta}^*_{1k}\|_2 = \|\boldsymbol{\beta}^*_{2k}\|_2 = \|\boldsymbol{\beta}^*_{3k}\|_2 = 1$$

  2. Sparse components: $\|\boldsymbol{\beta}^*_{1k}\|_0 \le s_1$, $\|\boldsymbol{\beta}^*_{2k}\|_0 \le s_2$, $\|\boldsymbol{\beta}^*_{3k}\|_0 \le s_3$ for $k \in [K]$.

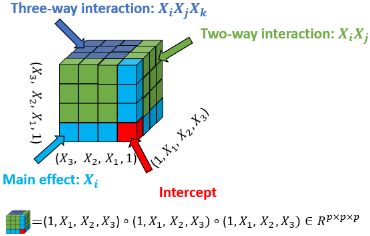
- The cubic sketching tensor $\mathscr{X}_i$ for non-symmetric case is $\mathscr{X}_i = \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i$, where $\{\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i\}_{i=1}^n$ are Gaussian random vectors.

Reduced Symmetric Tensor Recovery Model

- For symmetric tensor recovery model

$$y_i = \langle \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^* \circ \boldsymbol{\beta}_k^*, \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i \rangle + \epsilon_i = \sum_{k=1}^{K} \eta_k^* \underbrace{(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k^*)^3}_{\text{non-linear}} + \epsilon_i$$

- Connect with *interaction effect model*.



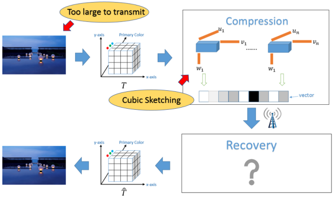
- New Goal: Recover $\{\eta_k^*, \boldsymbol{\beta}_k^*\}_{k=1}^K$

Reduced Non-symmetric Tensor Recovery Model

- For non-symmetric tensor recovery model

$$
\begin{aligned}
y_i &= \langle \sum_{k=1}^{K} \eta_k^* \boldsymbol{\beta}_{1k}^* \circ \boldsymbol{\beta}_{2k}^* \circ \boldsymbol{\beta}_{3k}^*, \boldsymbol{u}_i \circ \boldsymbol{v}_i \circ \boldsymbol{w}_i \rangle + \epsilon_i \\
&= \sum_{k=1}^{K} \eta_k^* \underbrace{(\boldsymbol{u}_i^\top \boldsymbol{\beta}_{1k}^*)(\boldsymbol{v}_i^\top \boldsymbol{\beta}_{2k}^*)(\boldsymbol{w}_i^\top \boldsymbol{\beta}_{3k}^*)}_{\text{non-linear}} + \epsilon_i
\end{aligned}
$$

- Connect with *compressed image transmission model*.



- New Goal: Recover $\{\eta_k^*, \boldsymbol{\beta}_{1k}^*, \boldsymbol{\beta}_{2k}^*, \boldsymbol{\beta}_{3k}^*\}_{k=1}^{K}$.

Empirical Risk Minimization

- Consider Empirical Risk Minimization

$$\widehat{\mathscr{T}} = \underset{\{\eta_k, \boldsymbol{\beta}_k\}}{\mathrm{argmin}} \underbrace{\sum_{i=1}^{n} (y_i - \sum_{k=1}^{K} \eta_k (\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)^3)^2}_{\mathcal{L}_1(\eta_k, \boldsymbol{\beta}_k)}$$

$$\widehat{\mathscr{T}} = \underset{\{\eta_k, \boldsymbol{\beta}_{ik}\}}{\mathrm{argmin}} \underbrace{\sum_{i=1}^{n} (y_i - \sum_{k=1}^{K} \eta_k (\boldsymbol{u}_i^\top \boldsymbol{\beta}_{1k})(\boldsymbol{v}_i^\top \boldsymbol{\beta}_{2k})(\boldsymbol{w}_i^\top \boldsymbol{\beta}_{3k}))^2}_{\mathcal{L}_2(\eta_k, \boldsymbol{\beta}_{ik})}$$

- Difficulties: *Non-convex optimization!* Non-convexity from cube structure or tri-convexity.

Our Contributions

1. Efficient two-stage implementation to non-convex optimization problem.
2. Non-asymptotic analysis. Provide optimal estimation rate.

Two-stage Implementation

Main Algorithm (Symmetric Recovery)

Input: observation pairs, the tensor rank, thresholding parameter

Initial Step

- Step one: *Construct empirical moment-based tensor*
- Step two: *Sparse tensor decompostion*

Output: Initial estimator

Gradient Update

- Step three: *Thresholded gradient decent*

Output: Final estimator

Initial Step: Construct unbiased estimator

- Construct an unbiased empirical moment based tensor $\mathcal{T}_s(y_i, \mathscr{X}_i) \in \mathbb{R}^{p\times p\times p}$ as following

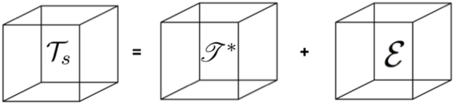
$$\mathcal{T}_s := \underbrace{\frac{1}{6}\Big[\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i \circ \boldsymbol{x}_i \circ \boldsymbol{x}_i - \mathcal{U}\Big]}_{\text{only depends on observations.}}$$

where the bias term
$\mathcal{U} = \sum_{j=1}^{p} \Big(\boldsymbol{m}_1 \circ \boldsymbol{e}_j \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{m}_1 \circ \boldsymbol{e}_j + \boldsymbol{e}_j \circ \boldsymbol{e}_j \circ \boldsymbol{m}_1\Big)$, and
$\boldsymbol{m}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i$. Here $\{\boldsymbol{e}_j\}_{j=1}^{p}$ are the basis vectors in $\mathbb{R}^p$.

Initial Step: Construct unbiased estimator

- Intuition: $\mathbb{E}[\mathcal{T}_s] = \mathscr{T}^*$.

Tensor Denosing Model: $\mathcal{T}_s = \mathscr{T}^* + \mathcal{E}$



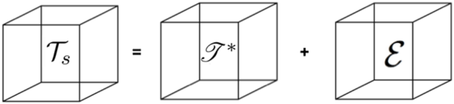
  - Observation $\mathcal{T}_s$.
  - Noise $\mathcal{E} = \mathcal{T}_s - \mathbb{E}(\mathcal{T}_s)$: approximation error.

- Decompose $\mathcal{T}_s$ to obtain $\{\eta_k^{(0)}, \beta_k^{(0)}\}$ through sparse tensor decomposition. See next slide for details.
- *Far from the optimal estimation, but good enough as a warm start.*

Initial Step: Construct unbiased estimator

- Intuition: $\mathbb{E}[\mathcal{T}_s] = \mathscr{T}^*$.

Tensor Denosing Model: $\mathcal{T}_s = \mathscr{T}^* + \mathcal{E}$



  - Observation $\mathcal{T}_s$.
  - Noise $\mathcal{E} = \mathcal{T}_s - \mathbb{E}(\mathcal{T}_s)$: approximation error.

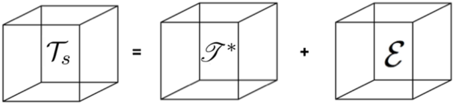
- Decompose $\mathcal{T}_s$ to obtain $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}$ through sparse tensor decomposition. See next slide for details.

- *Far from the optimal estimation, but good enough as a warm start.*

Initial Step: Construct unbiased estimator

- Intuition: $\mathbb{E}[\mathcal{T}_s] = \mathscr{T}^*$.

Tensor Denosing Model: $\mathcal{T}_s = \mathscr{T}^* + \mathcal{E}$



  - Observation $\mathcal{T}_s$.
  - Noise $\mathcal{E} = \mathcal{T}_s - \mathbb{E}(\mathcal{T}_s)$: approximation error.

- Decompose $\mathcal{T}_s$ to obtain $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}$ through sparse tensor decomposition. See next slide for details.
- *Far from the optimal estimation, but good enough as a warm start.*

Initial Step: Sparse Tensor Decomposition

⇒ Generate L staring points $\{\boldsymbol{\beta}_l^{\text{start}}\}_{l=1}^L$.

⇒ *For each starting point, compute a non-sparse component of moment-based $\mathcal{T}_s$ via symmetric tensor power update:*

$$\widetilde{\boldsymbol{\beta}}_l^{(t+1)} = \frac{\mathcal{T}_s \times_2 \boldsymbol{\beta}_l^{(t)} \times_3 \boldsymbol{\beta}_l^{(t)1}}{\|\mathcal{T}_s \times_2 \boldsymbol{\beta}_l^{(t)} \times_3 \boldsymbol{\beta}_l^{(t)}\|_2},$$

⇒ *Get a sparse solution $\boldsymbol{\beta}_l^{(t+1)}$ via thresholding or truncation.*

⇒ Cluster L sets of single component $\{\boldsymbol{\beta}_l^{(T)}, \boldsymbol{\beta}_l^{(T)}, \boldsymbol{\beta}_l^{(T)}\}_{l=1}^L$ into K clusters to obtain a rank-K decomposition $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}, \boldsymbol{\beta}_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$.

Different from matrix SVD due to non-orthogonality.

[1] For $\mathcal{T}_s \in \mathbb{R}^{p\times p\times p}$ and $\boldsymbol{x} \in \mathbb{R}^p$, define $\mathcal{T}_s \times_2 \boldsymbol{x} \times_3 \boldsymbol{x} := \sum_{j,l} \boldsymbol{x}_j \boldsymbol{x}_l [\mathcal{T}]_{:,j,l}$

Gradient Update: Thresholded Gradient Decent

⇒ Input initial estimator $\{\eta_k^{(0)}, \boldsymbol{\beta}_k^{(0)}\}_{k=1}^K$.

⇒ *In each iteration step, update* $\{\boldsymbol{\beta}_k\}_{k=1}^K$ *as*

$$\widetilde{\boldsymbol{\beta}}_k^{(t+1)} = \boldsymbol{\beta}_k^{(t)} - \frac{\mu_t}{\phi} \nabla_{\boldsymbol{\beta}_k} \mathcal{L}_1(\eta_k^{(0)}, \boldsymbol{\beta}_k^{(t)})$$

where $\phi = \frac{1}{n}\sum_{i=1}^n y_i^2$, μ_t *is the step size.*

⇒ *Sparsify current update by thresholding* $\boldsymbol{\beta}_k^{(t+1)} = \varphi_\rho(\widetilde{\boldsymbol{\beta}}_k^{(t+1)})$.

⇒ Normalize final update $\boldsymbol{\beta}_k^{(T)} = \frac{\boldsymbol{\beta}_k^{(T)}}{\|\boldsymbol{\beta}_k^{(T)}\|_2}$ and update the weight $\widehat{\eta}_k = \eta_k^{(0)} \times \|\boldsymbol{\beta}_k^{(T)}\|_2^3$.

[1] Alternating update for non-symmetric tensor recovery.

Non-asymptotic Analysis

Non-asymptotic Upper Bound

Theorem

Suppose some regularity conditions for the true tensor parameter hold. Assume $n \geq C_0 s^{3/2} \log p$ *for some large constant* C_0*. Denote* $Z_k^{(t)} = \sum_{k=1}^{K} \| \sqrt[3]{\eta_k} \boldsymbol{\beta}_k^{(t)} - \sqrt[3]{\eta_k^*} \boldsymbol{\beta}_k^* \|_2^2$ *For any* $t = 0, 1, 2, \ldots$, *the factor-wise estimator satisfies*

$$Z_k^{(t+1)} \leq \underbrace{\kappa^t Z_k^{(t)}}_{\text{computational error}} + \underbrace{\frac{C_1 \eta_{\min}^{*-\frac{4}{3}}}{16} \frac{\sigma^2 s \log p}{n}}_{\text{statistical error}},$$

with high probability, where κ *is the contraction parameter between 0 and 1,* $\eta_{\min}^* = \min_k \{\eta_k^*\}$, σ *is the noise level and* C_0, C_1 *are some absolute constants.*

Remarks

- Interesting characterization for computational error and statistical error;
- Geometric convergence rate to the truth in the noiseless case and minimax optimal statistical rate shown later;
- The error bound is dominated by computation error in the first several iterations and then is dominated by statistical error. Useful guideline for choosing stopping rule.
- We conjecture that $n \gtrsim s^{3/2} \log p$ is the minimum requirement of sample complexity in most tensor problems. This has an essential difference with matrix case, where the optimal sample complexity is $\mathcal{O}(s \log p)$.

Remarks

- When $t \geq T$ for some enough T, the final estimator is bounded by
$$\left\| \mathscr{T}^{(T)} - \mathscr{T}^* \right\|_F^2 \leq \frac{C\sigma^2 K s \log p}{n},$$
with high probability.
- *Minimax optimal rate!*

Class of Sparse and Low-rank tensor

- Sparse CP decomposition

$$\mathscr{T} = \sum_{k=1}^{K} \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k \circ \boldsymbol{\beta}_k, \|\boldsymbol{\beta}_k\|_0 \leq s \text{ for } k \in [K]$$

- Incoherence condition(nearly orthogonal): The true tensor components are incoherent such that

$$\max_{k_i \neq k_j \in [K]} |\langle \boldsymbol{\beta}^*_{k_i}, \boldsymbol{\beta}^*_{k_j} \rangle| \leq \frac{C}{\sqrt{s}}.$$

Minimax Lower Bound

Theorem

Consider the class of tensor satisfy sparse CP-decomposition and incoherence condition. Suppose we sample via cubic measurements with i.i.d. standard normal sketches with i.i.d. $N(0, \sigma^2)$ noise, then we have the following lower bound result for recovery loss for this class of low-rank tensors,

$$\inf_{\widehat{\mathscr{T}}} \sup_{\mathscr{T} \in \mathcal{F}} \mathbb{E} \left\| \widehat{\mathscr{T}} - \mathscr{T} \right\|_F^2 \geq c\sigma^2 \frac{Ks \log(ep/s)}{n}.$$

Optimal Estimation Rate

Theorem

Consider the class of tensor $\mathcal{F}_{p,K,s}$ satisfy sparse CP-decomposition and incoherence condition. Suppose we observe n samples $\{y_i, \mathscr{X}_i\}_{i=1}^n$ from symmetric tensor cubic sketching model, where $n \geq C s^{3/2} \log p$ for some large constant C. Then the estimator $\widehat{\mathscr{T}}$ achieves

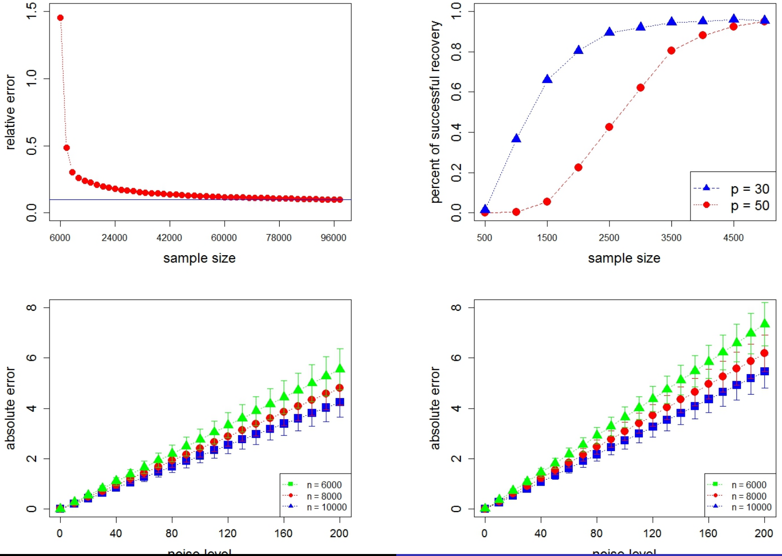
$$\inf_{\widetilde{\mathscr{T}}} \sup_{\mathscr{T} \in \mathcal{F}_{p,K,s}} \mathbb{E} \left\| \widetilde{\mathscr{T}} - \mathscr{T} \right\|_F^2 \asymp \underbrace{\sigma^2 \frac{K s \log(p/s)}{n}}_{R^*},$$

when $\log p \asymp \log p/s$. Here σ is the noise level.

Remarks

- Our analysis is non-asymptotic and our estimator is rate-optimal.
- In general, we have a trade-off $\rightarrow R^*$ is the outcome of *statistical error* and *optimization error* trade-off.
- Similar argument holds for non-symmetric case. *Different technical tools are used.*
- To overcome the obstacle from high-order Gaussian random variable, we develop novel high-order concentration inequality by the combination of *truncation argument* and ψ_α-*norm*.

symmetric tensor, $p = 50, K = 3, s = 0.3$, replication = 200.



Botao Hao
hao22@purdue.edu
Department of Statistics
Purdue University