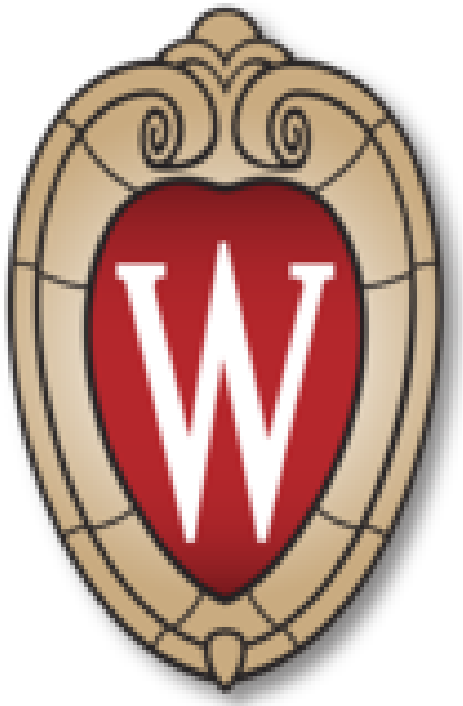
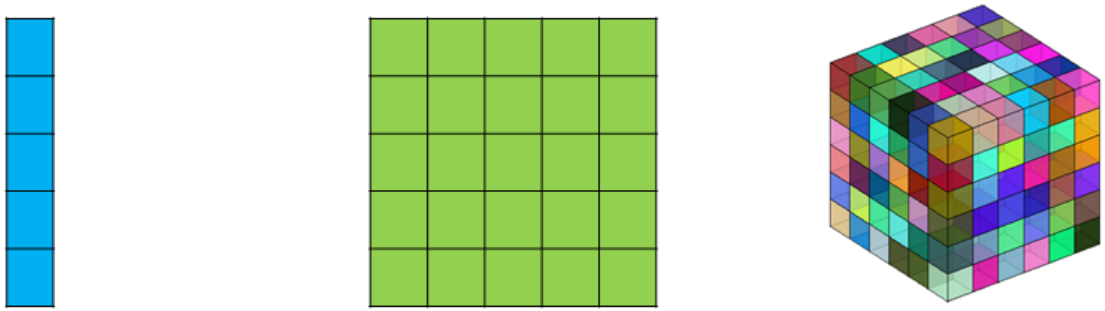
Efficient and Optimal Tensor Regression via Importance Sketching

Anru Zhang
Department of Statistics
University of Wisconsin-Madison



Introduction

- Tensors are arrays with multiple directions.



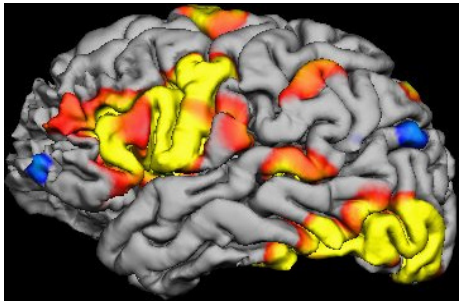
Order-1 tensor: vector Order-2 tensor: matrix Order-3 tensor

- Tensors of order three or higher are called high-order tensors.

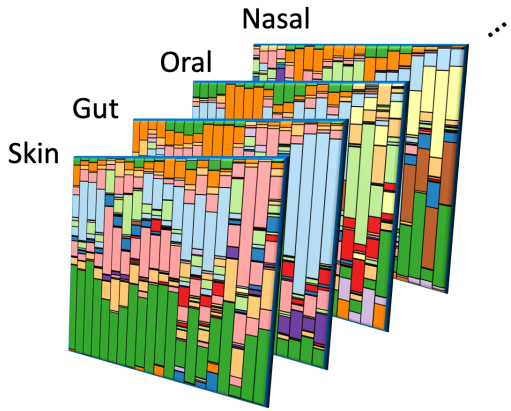
$$\mathcal{A} \in \mathbb{R}^{p_1 \times \cdots \times p_d}, \qquad \mathcal{A} = (A_{i_1 \cdots i_d}), \qquad 1 \le i_k \le p_k, \quad k = 1, \ldots, d.$$

More High-Order Data Are Emerging

- Brain imaging

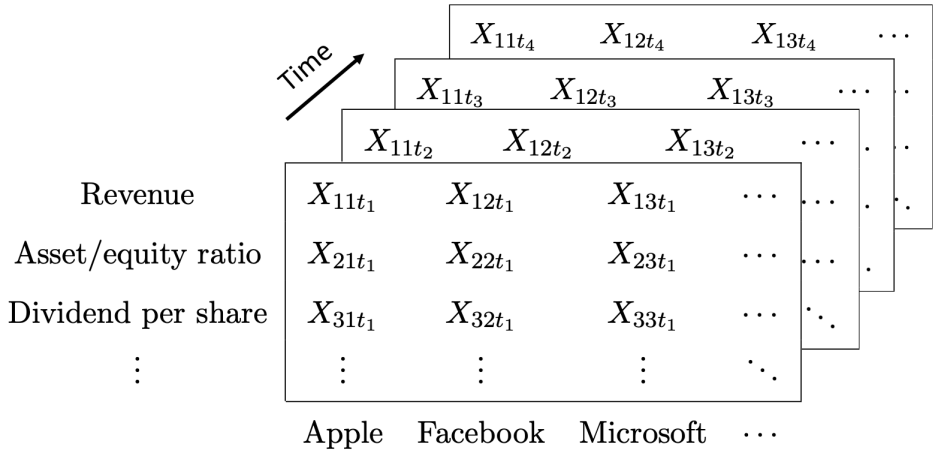


- Microbiome studies

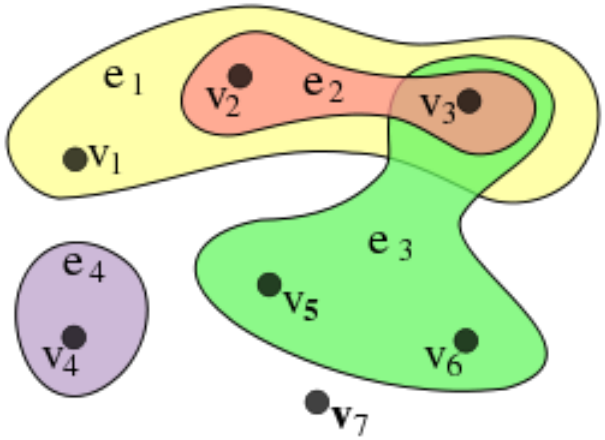


More High-Order Data Are Emerging

- Matrix-valued time series



- Hypergraphs



High Order Enables Solutions for Harder Problems

High-order Interaction Pursuits

- Model (Hao, **Z.**, Cheng, 2018)

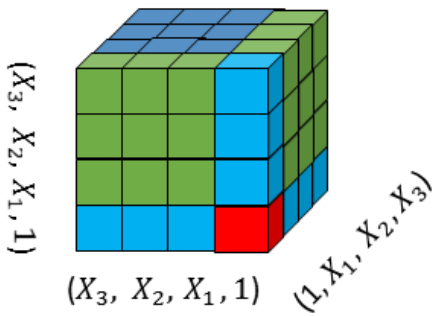
$$y_i = \beta_0 + \underbrace{\sum_i X_i\beta_i}_{\text{Main effect}} + \underbrace{\sum_{i,j} \gamma_{ij}X_iX_j}_{\text{Pairwise interaction}} + \underbrace{\sum_{i,j,k} \eta_{ijk}X_iX_jX_k}_{\text{Triple-wise}} + \varepsilon_i, \quad i = 1, \ldots, n.$$

- Rewrite as

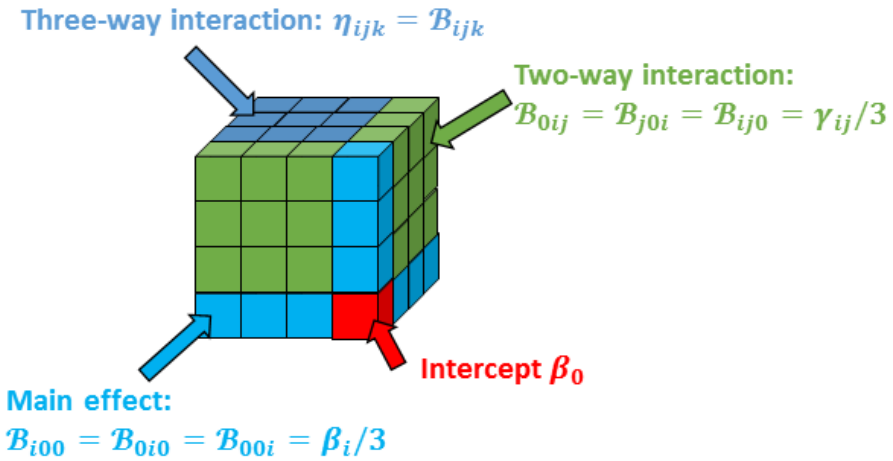
$$y_i = \langle \mathcal{B}, \mathcal{X}_i \rangle + \varepsilon_i.$$

Tensorized Covariates

$\mathcal{X} = (1, \boldsymbol{X}) \circ (1, \boldsymbol{X}) \circ (1, \boldsymbol{X})$



High-order Interaction Effects $\mathcal{B}$



High Order Enables Solutions for Harder Problems

Estimation of Mixture Models

- A mixture model incorporates subpopulations in an overall population.

- Examples:
  - Gaussian mixture model (Lindsay & Basak, 1993; Hsu & Kakade, 2013; Wu & Yang, 2019; Wu & Zhang, 2019)
  - Topic modeling (Arora et al, 2013)
  - Hidden Markov Process (Anandkumar, Hsu, & Kakade, 2012)
  - Independent component analysis (Miettinen, et al., 2015)
  - Additive index model (Balasubramanian, Fan & Yang, 2018)
  - Mixture regression model (De Veaux, 1989; Jordan & Jacobs, 1994)
  - ...

- Method of Moment (MoM):
  - First moment $\rightarrow$ vector;
  - Second moment $\rightarrow$ matrix;
  - **High-order moment $\rightarrow$ high-order tensors.**

High Order is ...

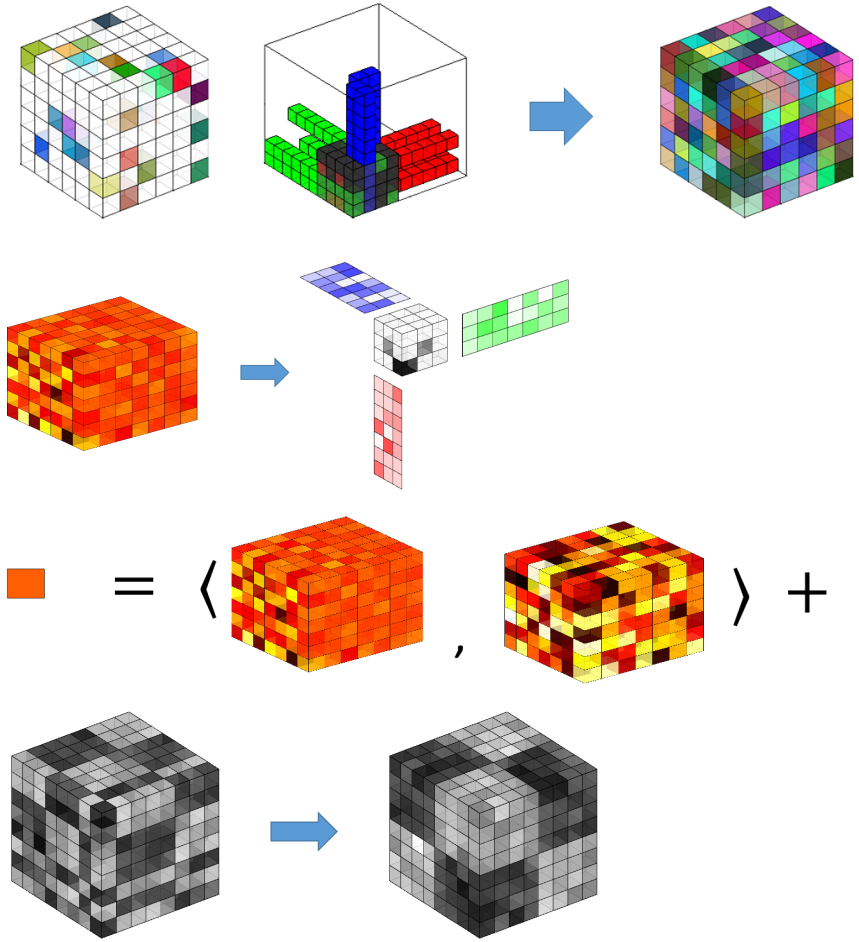
- **High order is more charming!**
- **High order is harder!**

  Tensor problems are **far more than** extension of matrices.
  - More structures
  - High-dimensionality
  - Computational difficulty
  - Many concepts not well defined or NP-hard

  Research of many basic high order problems is still in its infancy.

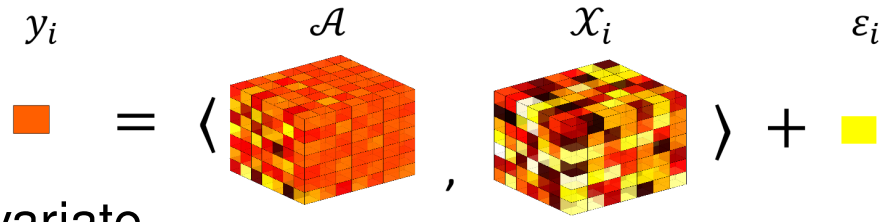
High Order Casts New Problems and Challenges

- Tensor Completion
- Tensor SVD
- Tensor Regression
- Biclustering/Triclustering
- ...



- In this talk, we focus on **tensor regression**.

$$y_i = \langle \mathcal{A}, \mathcal{X}_i \rangle + \varepsilon_i, \quad i = 1, \ldots, n.$$



  - $\mathcal{X}_i$: tensor covariate
  - y_i: response
  - ε_i: noise
  - $\mathcal{A}$: target tensor to be estimated
    low-rank / sparse / smooth ...

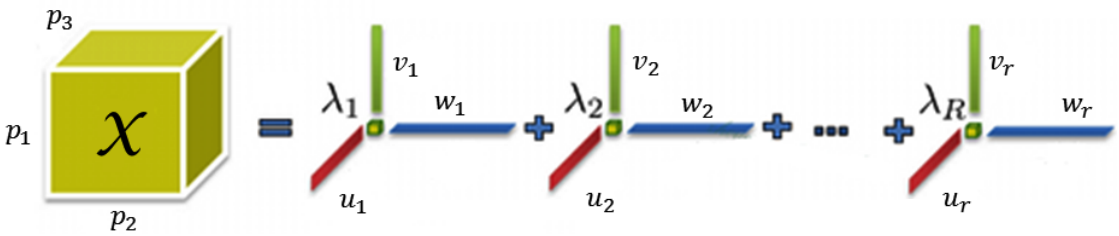
- **Goal: estimating $\mathcal{A}$ based on $(y_i, \mathcal{X}_i)$**
- Examples:
  - Degree of ADHD $\sim$ MRI Brain imaging data
  - Phenotypes $\sim$ Microbiome data from multiple body sites

Z., Luo, Y., Raskutti, G., and Yuan, M. (2019+). ISLET: fast and optimal low-rank tensor regression via importance sketchings. *SIAM Journal on Mathematics of Data Science*, major revision under review.

Tensor Rank Has No Uniform Definition

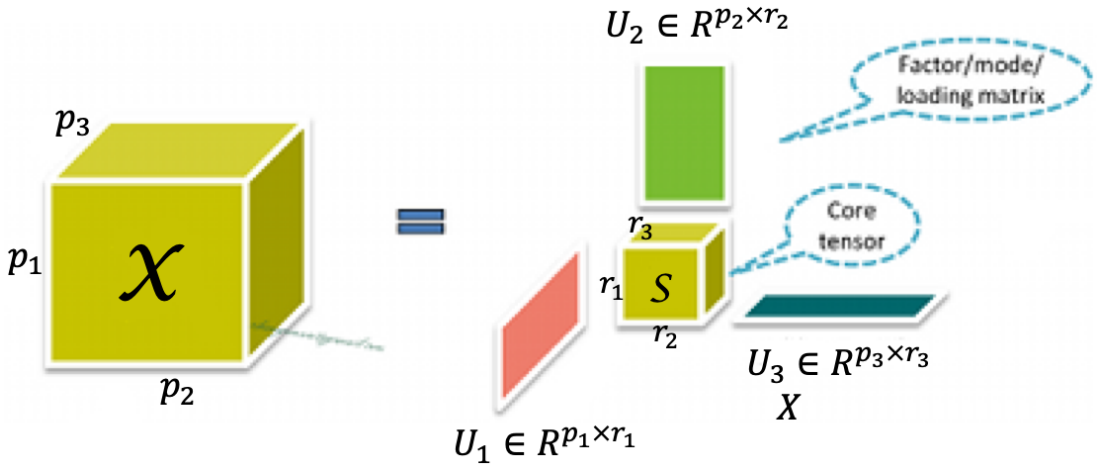
- Canonical polyadic (CP) rank:

$$r_{cp} = \min r \quad \text{s.t.}$$
$$\mathcal{X} = \sum_{i=1}^{r} \lambda_i \cdot u_i \circ v_i \circ w_i$$



- Tucker rank:

$$\mathcal{X} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3$$
$$\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, \boldsymbol{U}_k \in \mathbb{R}^{p_k \times r_k}$$



Smallest possible (r_1, r_2, r_3) are Tucker rank of $\mathcal{X}$.

- If $\mathcal{X}$ is CP rank-r, it is also Tucker rank-(r, r, r).

Picture Source: Guoxu Zhou's website. http://www.bsp.brain.riken.jp/ zhougx/tensor.html

Previous Ideas: Convex Regularization

Convex regularization is commonly used to estimate structured high-dimensional parameters, e.g., low-rank matrices.

- Rank minimization:

$$\hat{\boldsymbol{A}} = \underset{\boldsymbol{A}}{\operatorname{argmin}} \{\mathcal{L}(\boldsymbol{A}) + \lambda \cdot \operatorname{rank}(\boldsymbol{A})\}$$

  - $\mathcal{L}(\cdot)$ is the loss function;
  - Computational infeasible $\rightarrow$ $\operatorname{rank}(\boldsymbol{A})$ is non-convex.

- **Nuclear norm minimization** (Fazel, 2004):

$$\hat{\boldsymbol{A}} = \underset{\boldsymbol{A}}{\operatorname{argmin}} \{\mathcal{L}(\boldsymbol{A}) + \lambda \cdot \|\boldsymbol{X}\|_*\}$$

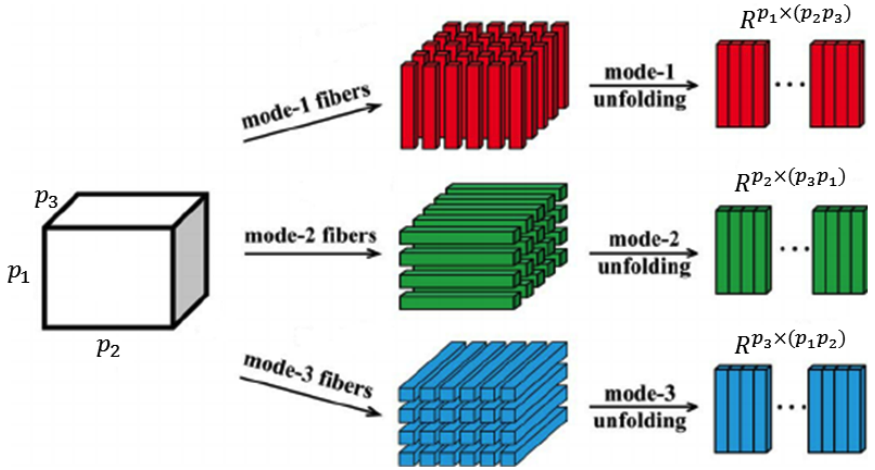
  - $\|\boldsymbol{X}\|_* = \sum_k \sigma_k(\boldsymbol{X})$ is the matrix nuclear norm;
  - Statistical and computational guarantees have been established.

Convex Regularization for Low-rank Tensor Estimation

Overlapped nuclear norm minimization (Liu et al, 2009; Tomioka & Suzuki, 2013; Mu, Huang, Wright, & Goldfarb, 2014)

$$\mathcal{A} = \operatorname*{argmin}_{\mathcal{A}} \left\{ \mathcal{L}(\mathcal{A}) + \lambda \sum_k \|\mathcal{M}_k(\mathcal{A})\|_* \right\}$$

$\mathcal{M}_k(\mathcal{A})$: matricizations of $\mathcal{A}$



- Advantages:
  - Computational feasible (implementations: SDP, ADMM)
  - Easier to analyze
- Disadvantages:
  - Sub-optimal statistical performance

Convex Regularization for Low-rank Tensor Estimation

Tensor nuclear norm minimization (Yuan & Zhang, 2014, 2016)

$$\hat{\mathcal{A}} = \underset{\mathcal{A}}{\operatorname{argmin}} \{\mathcal{L}(\mathcal{A}) + \lambda ||\mathcal{A}||_*\}$$

$\|\cdot\|_*$: tensor nuclear norm; the dual norm of tensor spectral norm

$$||\mathcal{A}||_* = \max_{||\mathcal{B}||_{sp} \le 1} \langle \mathcal{B}, \mathcal{A} \rangle, \quad ||\mathcal{B}||_{sp} = \sup_{u,v,w} \frac{\mathcal{B} \times_1 u \times_2 v \times_3 w}{||u||_2 ||v||_2 ||w||_2}.$$

- Advantage:
  - Provable optimal statistical performance (in some scenarios);
- Disadvantage:
  - NP-hard to compute (Hillar and Lim, 2013)!

Previous Ideas: Proximal Gradient Descent

Proximal gradient descent (PGD) for low-rank matrix estimation: (Toh and Yun, 2010; Chen and Wainwright, 2015)

$$\boldsymbol{B}^{(t+1)} = \boldsymbol{A}^{(t)} - \eta \nabla \mathcal{L}(\boldsymbol{A}^{(t)}) \quad \boldsymbol{A}^{(t+1)} = s_\lambda(\boldsymbol{B}^{(t+1)}), \quad t = 0, 1, \ldots$$

- $s_\lambda(\cdot)$ is the thresholding operator

$$s_\lambda(\boldsymbol{A}) = \sum_k (\sigma_k - \lambda)_+ u_k v_k^\top, \quad \boldsymbol{A} = \sum_k \sigma_k u_k v_k^\top \text{ is the SVD.}$$

Pitfalls of PGD in **low-rank tensor estimation**:

- Thresholding for tensor is not well-defined.
- The alternative, exact projection is NP-hard to evaluate for tensors.
  - Only approximation is available.
- Statistical optimality is not clear.
- Iteratively performing tensor projection is computational expensive.

Previous Ideas: Alternating Gradient Descent

Alternating gradient descent (AGD) for low-rank matrix estimation: (Jain, Netrapalli, & Sanghavi, 2013)

$$\text{Factorize } \boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^\top, \quad \boldsymbol{U} \in \mathbb{R}^{p_1 \times r}, \boldsymbol{V} \in \mathbb{R}^{p_2 \times r};$$

$$\begin{cases} \boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} - \eta \nabla \mathcal{L}_{\boldsymbol{U}}(\boldsymbol{U}^{(t)}(\boldsymbol{V}^{(t)})^\top), \\ \boldsymbol{V}^{(t+1)} = \boldsymbol{V}^{(t+1)} - \eta \nabla \mathcal{L}_{\boldsymbol{V}}(\boldsymbol{U}^{(t+1)}(\boldsymbol{V}^{(t)})^\top) \end{cases} \qquad t = 0, 1, \ldots$$

- AGD found great empirical and theoretical successes in various low-rank matrix estimation problems.

Pitfalls of AGD in **low-rank tensor estimation**

- Theoretical Analysis for AGD is involving.
  - Statistical optimality is not clear.
- Evaluation of full likelihood is prohibitively expensive in high-dimensions.

New Method: Importance Sketching

$$y_i = \langle \mathcal{A}, \boldsymbol{X}_i \rangle + \varepsilon_i, \quad i = 1, \ldots, n$$

- Direct generalization of matrix methods to tensor regression has not worked very well.
- We introduce a new framework for tensor estimation via

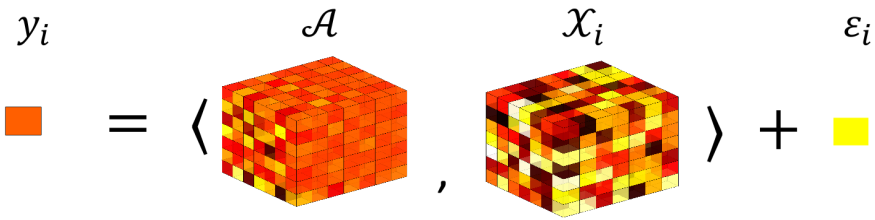
importance sketching

Idea: incorporates information from low-dimensional structure of $\mathcal{A}$ to perform dimension reduction on $\boldsymbol{X}$.

Model

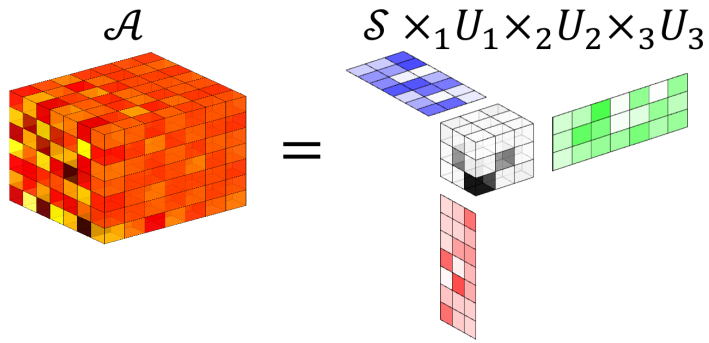
- For convenience, we focus on order-3 low-rank tensor regression,

$$y_i = \langle \mathcal{A}, \mathcal{X}_i \rangle + \varepsilon_i, \quad i = 1, \ldots, n.$$



Here, $\mathcal{A}$ is Tucker low-rank,

$$\mathcal{A} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3 \in \mathbb{R}^{p_1 \times p_2 \times p_3},$$
$$\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, \quad \boldsymbol{U}_k \in \mathbb{R}^{p_k \times r_k}, \quad k = 1, 2, 3.$$



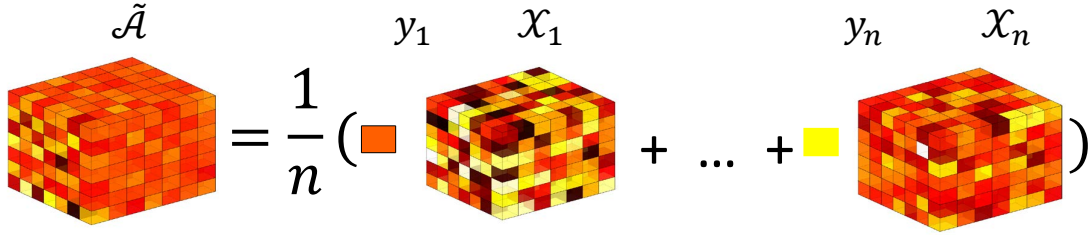
- Goal: estimate $\mathcal{A}$ based on $\{y_i, \mathcal{X}_i\}_{i=1}^n$.

Procedure of Importance Sketching Low-rank Estimation for Tensors (ISLET)

Step 1. Probing Importance Sketching Direction

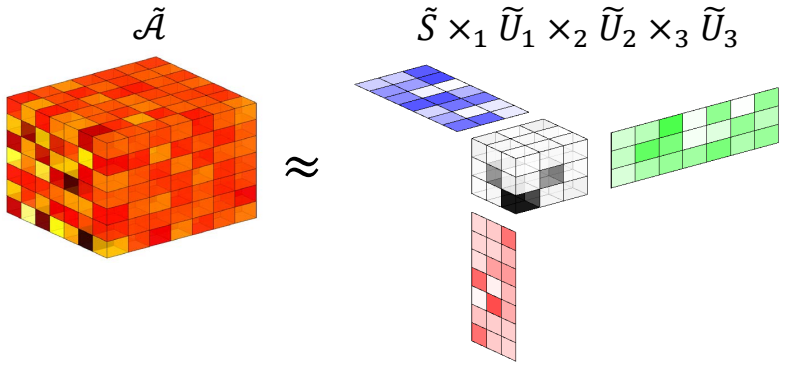
- **(Step 1.1)** Evaluate the sample covariance tensor

$$\widetilde{\mathcal{A}} = \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{\mathcal{X}}_i$$



- **(Step 1.2)** Apply high-order orthogonal iteration (HOOI) to obtain a low-rank factorization of $\widetilde{\mathcal{A}}$

$$\widetilde{\mathcal{A}} \approx \widetilde{\mathcal{S}} \times_1 \widetilde{\boldsymbol{U}}_1 \times_2 \widetilde{\boldsymbol{U}}_2 \times_3 \widetilde{\boldsymbol{U}}_3$$



- **(Step 1.3)** Perform QR orthogonalization $\widetilde{\boldsymbol{V}}_k = \mathrm{QR}\left(\mathcal{M}_k^{\top}\left(\widetilde{\mathcal{S}}\right)\right)$.
- **Outcome of Step 1:** $\{\widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{V}}_k\}_{k=1}^{3}$.

HOOI: tensor factorization method based on power iterations.

- An analog of SVD for high-order tensors.
- Introduced by De Lathauwer, De Moor, Vandewalle (2000).

On the Best Rank-1 and Rank-(R_1,R_2,. . .,R_N) Approximation of **Higher-Order** Tensors

L De Lathauwer, B De Moor, J Vandewalle - SIAM journal on Matrix Analysis …, 2000 - SIAM

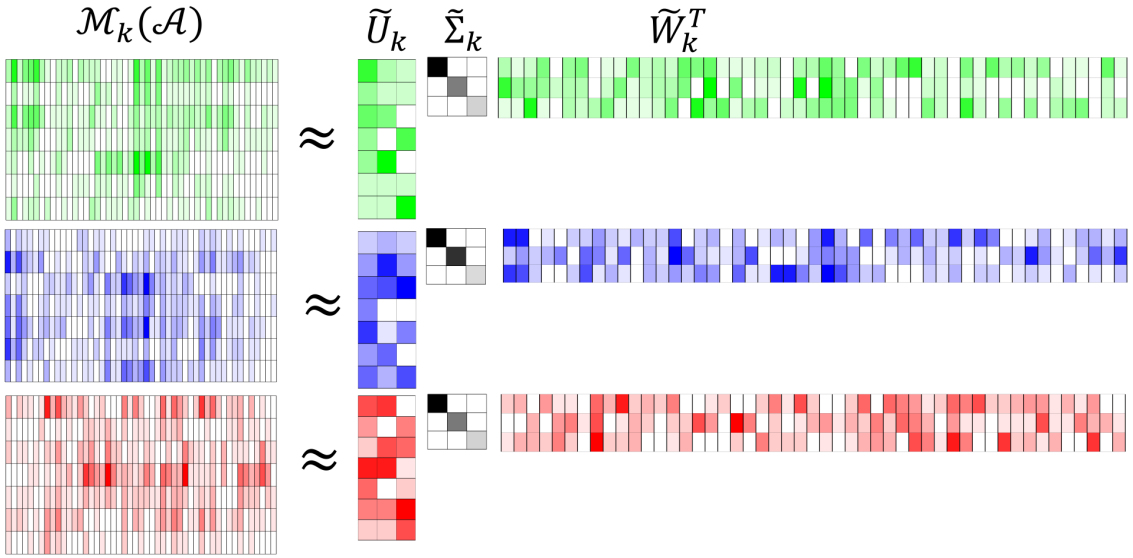
… (The generalization to **orders higher** than three … 2.1. Multiplication of a **higher-order** tensor by a matrix. **Higher-order** power and **orthogonal iterations** involve a multilinear equivalent of matrix-vector and matrix-matrix multiplications … ALGORITHM 3.2. **Higher-Order** Power Method …

☆ 99 Cited by 1282 Related articles All 19 versions

- **Z.** and Xia (*IEEE-TIT*, 2018)
  - proposed a statistical framework for HOOI,
  - studied the statistical and computational limits of this framework,
  - established the optimal statistical guarantees for HOOI.

Interpretations of Step 1

$$\mathcal{M}_k(\mathcal{A}) \approx \widetilde{\boldsymbol{U}}_k \widetilde{\Sigma}_k \widetilde{\boldsymbol{W}}_k^\top, \quad \widetilde{\boldsymbol{W}}_k = (\widetilde{\boldsymbol{U}}_{k+2} \otimes \widetilde{\boldsymbol{U}}_{k+1}) \widetilde{\boldsymbol{V}}_k, \quad k = 1, 2, 3,$$



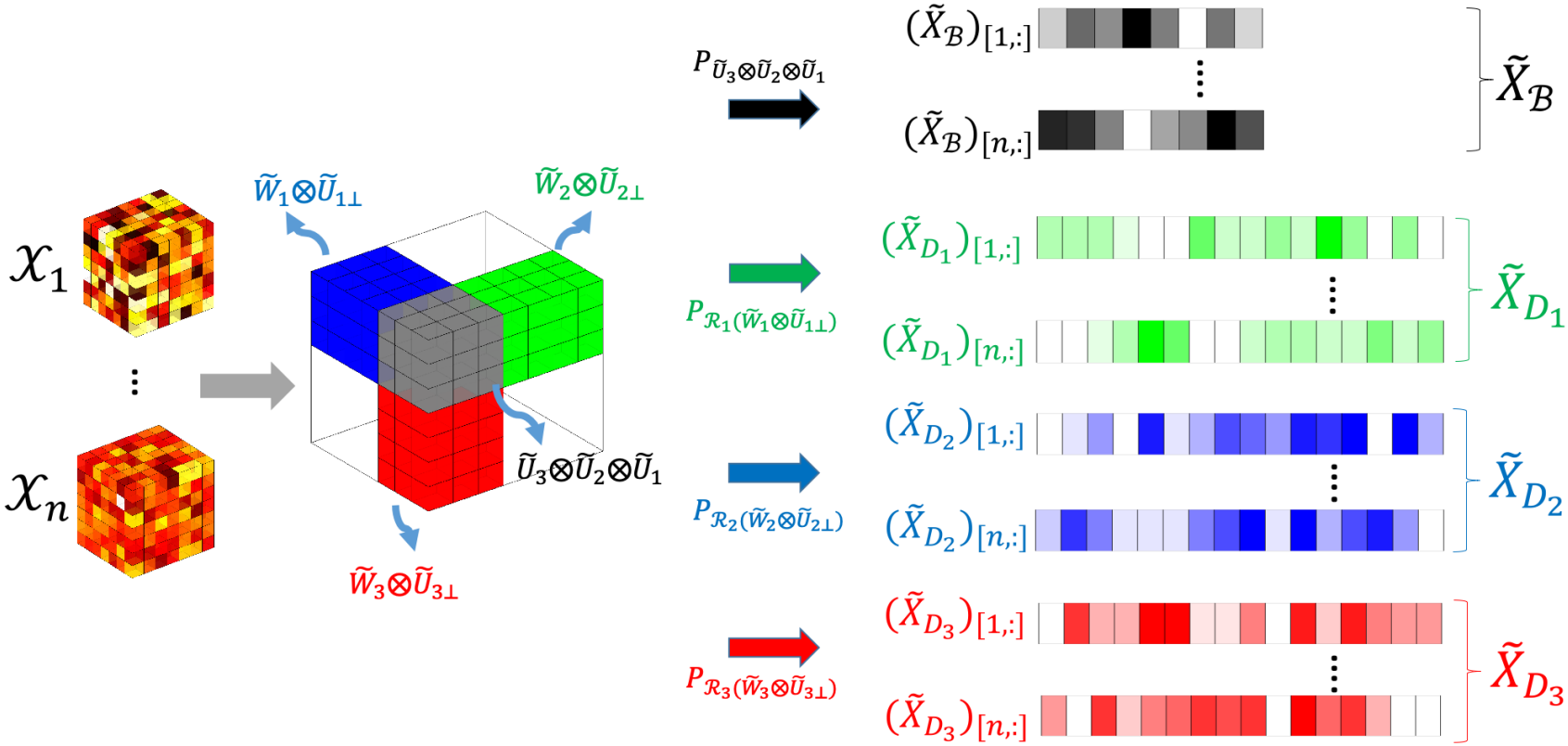
- $\{\widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{W}}_k\}$ are **importance sketching directions**.
  - $\{\widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{W}}_k\}$ are initial sample approximations of $\{\boldsymbol{U}_k, \boldsymbol{W}_k\}$, i.e., the left and right singular subspaces of $\mathcal{M}_k(\mathcal{A})$.
  - $\{\widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{W}}_k\}$ that best align with $\mathcal{A}$.

Step 2. Importance Sketching

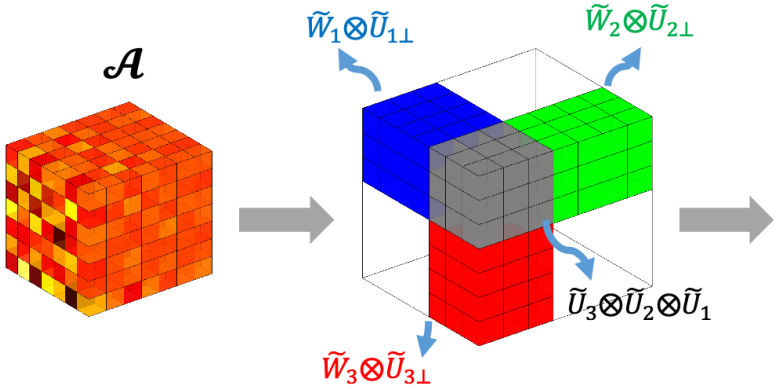
- Construct dimension-reduced covariates

$$\hat{\boldsymbol{X}}_{\mathcal{B}} \in \mathbb{R}^{n\times(r_1 r_2 r_3)}, \quad \left(\hat{\boldsymbol{X}}_{\mathcal{B}}\right)_{[i,:]} = \mathrm{vec}\left(\boldsymbol{\mathcal{X}}_i \times_1 \widetilde{\boldsymbol{U}}_1^\top \times_2 \widetilde{\boldsymbol{U}}_2^\top \times_3 \widetilde{\boldsymbol{U}}_3^\top\right)$$

$$\hat{\boldsymbol{X}}_{\boldsymbol{D}_k} \in \mathbb{R}^{n\times(p_k - r_k)r_k}, \quad \left(\hat{\boldsymbol{X}}_{\boldsymbol{D}_k}\right)_{[i,:]} = \mathrm{vec}\left(\widetilde{\boldsymbol{U}}_{k\perp}^\top \mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_i\right) \widetilde{\boldsymbol{W}}_k\right), \quad k = 1, 2, 3$$



Interpretation of Step 2



$$\widetilde{\gamma} = \begin{bmatrix} \mathrm{vec}(\widetilde{\mathcal{B}}) \\ \mathrm{vec}(\widetilde{D}_1) \\ \mathrm{vec}(\widetilde{D}_2) \\ \mathrm{vec}(\widetilde{D}_3) \end{bmatrix},$$

$$\widetilde{\mathcal{B}} := \mathcal{A} \times_1 \widetilde{U}_1^\top \times_2 \widetilde{U}_2^\top \times_3 \widetilde{U}_3^\top$$
$$\widetilde{D}_1 := \widetilde{U}_{1\perp}^\top \mathcal{M}_1(\mathcal{A}) \widetilde{W}_1$$
$$\widetilde{D}_2 := \widetilde{U}_{2\perp}^\top \mathcal{M}_2(\mathcal{A}) \widetilde{W}_2$$
$$\widetilde{D}_3 := \widetilde{U}_{3\perp}^\top \mathcal{M}_3(\mathcal{A}) \widetilde{W}_3$$

- Rewrite the regression model

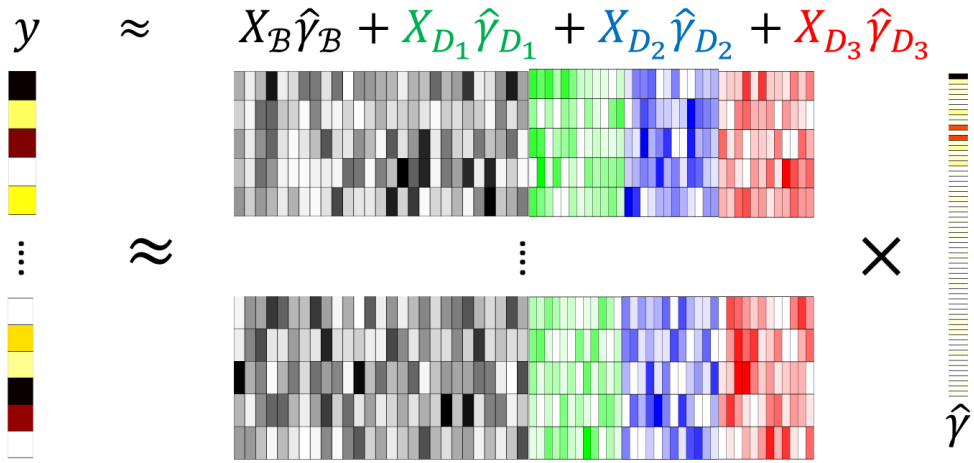
$$\begin{aligned} y_i =& \langle \mathcal{X}_i, \mathcal{A} \rangle + \varepsilon_i \\ =& (\widetilde{X}_{\mathcal{B}})_{[i,:]} \mathrm{vec}(\widetilde{\mathcal{B}}) + (\widetilde{X}_{D_1})_{[i,:]} \mathrm{vec}(\widetilde{D}_1) + (\widetilde{X}_{D_2})_{[i,:]} \mathrm{vec}(\widetilde{D}_2) \\ &+ (\widetilde{X}_{D_3})_{[i,:]} \mathrm{vec}(\widetilde{D}_3) + \mathrm{vec}(\mathcal{X}_i)^\top P_{\widetilde{U}_\perp} \mathrm{vec}(\mathcal{A}) + \varepsilon_i \\ =& \widetilde{X}_{[i,:]} \widetilde{\gamma} + \widetilde{\varepsilon}_i, \quad i = 1, \ldots, n. \end{aligned}$$

  - $\widetilde{X} = [\widetilde{X}_{\mathcal{B}}, \widetilde{X}_{D_1}, \widetilde{X}_{D_2}, \widetilde{X}_{D_3}]$ are sketching covariates;
  - $\widetilde{\varepsilon}_i = \mathrm{vec}(\mathcal{X}_i)^\top P_{\widetilde{U}_\perp} \mathrm{vec}(\mathcal{A}) + \varepsilon_i$ is the new noise;
  - $\widetilde{\gamma}$ is the sketch of $\mathcal{A}$.

Step 3. Dimension-Reduced Regression

- Perform dimension-reduced regression

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \|y - \widetilde{X}\gamma\|_2^2, \quad \widetilde{X} = \left[\widetilde{X}_{\mathcal{B}}\ \widetilde{X}_{D_1}\ \widetilde{X}_{D_2}\ \widetilde{X}_{D_3}\right].$$

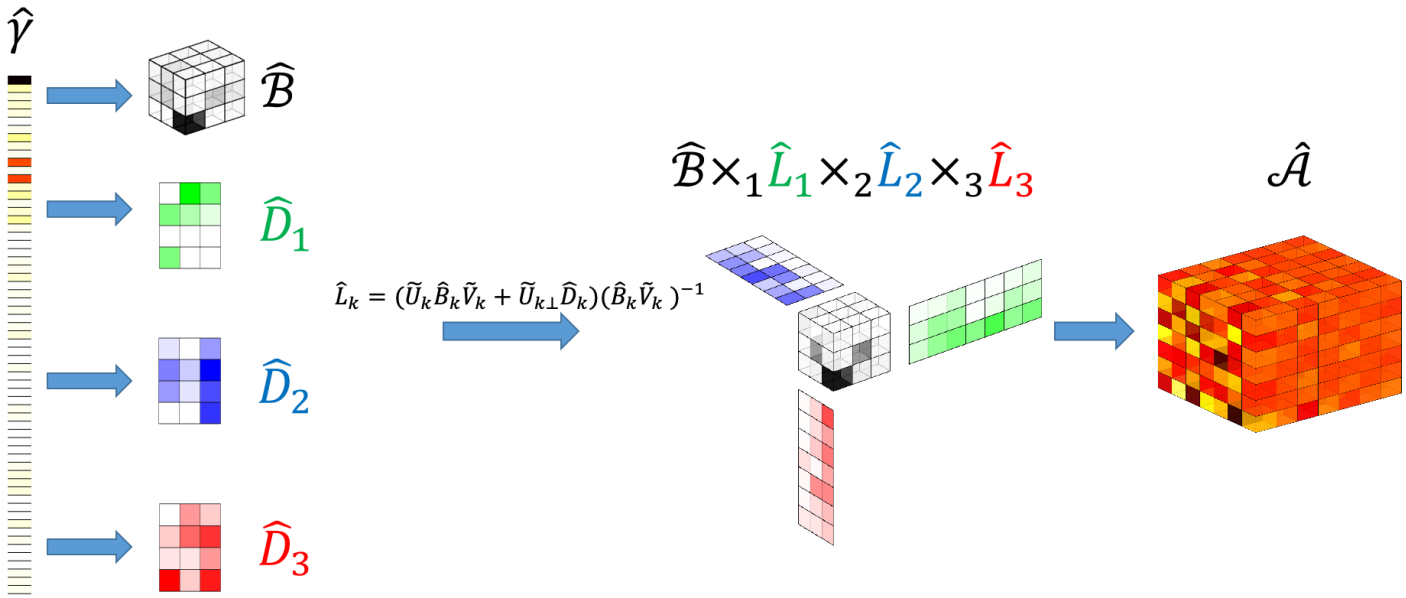


- Dimension of parameter is significantly reduced!

Original dimension	Dimension-reduced regression
$p_1 p_2 p_3$	$m := r_1 r_2 r_3 + \sum_{k=1}^{3}(p_k - r_k)r_k$

- **Outcome of Step 3:** $\hat{\gamma}$.

Step 4. Assembling the Final Estimate



- Assemble via the Cross scheme (**Z.** *AoS*, 2018)

$$\hat{\mathcal{A}} = \hat{\mathcal{B}}\times_1\hat{L}_1\times_2\hat{L}_2\times_3\hat{L}_3,$$

$$\hat{L}_k = \left(\widetilde{U}_k\hat{B}_k\widetilde{V}_k + \widetilde{U}_{k\perp}\hat{D}_k\right)\left(\hat{B}_k\widetilde{V}_k\right)^{-1}, \quad \hat{B}_k = \mathcal{M}_k(\hat{\mathcal{B}}), \quad k = 1, 2, 3.$$

Algorithm Summary

1. Probing Importance Sketching Direction
   - Calculate $\widetilde{\mathcal{A}} = \frac{1}{n}\sum_{j=1}^{n} y_j \mathcal{X}_j$
   - $\left.\begin{array}{l} \text{HOOI: } \widetilde{\mathcal{A}} \approx \widetilde{\mathcal{S}} \times_1 \widetilde{\boldsymbol{U}}_1 \times_2 \widetilde{\boldsymbol{U}}_2 \times_3 \widetilde{\boldsymbol{U}}_3 \\ \widetilde{\boldsymbol{V}}_k = \mathrm{QR}[\mathcal{M}_k(\widetilde{\mathcal{S}})^\top], k = 1, 2, 3 \end{array}\right\} \Rightarrow$ Importance sketching direction $\{\widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{W}}_k\}$
2. Importance Sketching
   - $\mathcal{X} \overset{\widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{W}}_k}{\Longrightarrow} \widetilde{\mathcal{X}} = [\widetilde{\mathcal{X}}_{\mathcal{B}}\ \widetilde{\mathcal{X}}_{\boldsymbol{D}_1}\ \widetilde{\mathcal{X}}_{\boldsymbol{D}_2}\ \widetilde{\mathcal{X}}_{\boldsymbol{D}_3}]$
   - $\mathcal{A} \overset{\widetilde{\boldsymbol{U}}_k, \widetilde{\boldsymbol{W}}_k}{\Longrightarrow} \widetilde{\gamma} = [\mathrm{vec}(\widetilde{\mathcal{B}})^\top, \mathrm{vec}(\widetilde{\boldsymbol{D}_1})^\top, \mathrm{vec}(\widetilde{\boldsymbol{D}_2})^\top, \mathrm{vec}(\widetilde{\boldsymbol{D}_3})^\top]$
   - $y = \langle \mathcal{X}_i, \mathcal{A} \rangle + \varepsilon_i = \widetilde{\boldsymbol{X}}_{[i,:]}\widetilde{\gamma} + \widetilde{\varepsilon}_i$
3. Dimension Reduced Regression
   - Solve $\hat{\gamma} = \mathrm{argmin}_\gamma \|y - \widetilde{\mathcal{X}}\gamma\|_2^2$
4. Assembling the Final Estimate
   - $\hat{\gamma} \overset{\text{Cross Scheme}}{\Longrightarrow} \hat{\mathcal{A}}$

Comparison to Randomized Sketching

- **Randomized sketching** is often employed for dimension reduction
  - Vectors (Raskutti & Mahoney, *JMLR*, 2014 ; Pilanci & Wainwright, 2015 *IEEE T-IT*; Yang, Pilanci & Wainwright, *AoS*, 2017)
  - Matrices (Dasarathy, Shah, Bhaskar, & Nowak, *IEEE T-IT*, 2015; Tropp, Yurtsever, Udell, & Cevher, *SIAM J. Matrix Anal. Appl.,*, 2017)
  - Tensors (Wang, Tung, Smola, Anandkumar, *NeurIPS*, 2015; Li, Haupt, Woodruff, *NeurIPS*, 2017)
  - Survey Papers (Mahoney, 2011; Woodruff, 2014)
- Example: least square estimator: $\operatorname{argmin}_\beta \|y - \boldsymbol{X}\beta\|_2^2$,

  (Projections) $\operatorname{argmin}_\beta \|\boldsymbol{U}y - \boldsymbol{U}\boldsymbol{X}\beta\|_2^2$, $\boldsymbol{U}$: Gaussian ensemble;

  (Sub-sampling) $\operatorname{argmin}_\beta \|y_\Omega - \boldsymbol{X}_{[\Omega,:]}\beta\|_2^2$, Ω : random subset.
- **Often statistically suboptimal in estimation** (Raskutti & Mahoney, 2014; Dobriban and Liu, 2018)

Comparison to Randomized Sketching

- Instead of randomized sketching, ISLET is based on **importance sketching**
  - It is supervised by y;
  - It incorporates structural information in the parameter of interest.
- Therefore, ISLET achieves
  - much better statistical performance than randomized sketching,
  - while keep the advantage of low computational and storage costs.

Theoretical Analysis

Theoretical Analysis under General Design

- $\mathcal{X}_i$ has general design, no assumption on ε_i

Theorem

Assume $\theta = \max_k \left\{ \| \sin\Theta(\widetilde{\boldsymbol{U}}_k, \boldsymbol{U}_k) \|, \| \sin\Theta(\widetilde{\boldsymbol{W}}_k, \boldsymbol{W}_k) \| \right\} < 1/2$ *and* $\|\hat{\boldsymbol{D}}_k(\hat{\boldsymbol{B}}_k\widetilde{\boldsymbol{V}}_k)^{-1}\| \le \rho$. *Then,*

$$\left\| \hat{\mathcal{A}} - \mathcal{A} \right\|_{HS}^2 \le (1 + C(\theta + \rho)) \left\| (\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1} \widetilde{\boldsymbol{X}}^\top \widetilde{\varepsilon} \right\|_2^2 .$$

Here, $\widetilde{\boldsymbol{X}} = [\widetilde{\boldsymbol{X}}_{\mathcal{B}}\ \widetilde{\boldsymbol{X}}_{\boldsymbol{D}_1}\ \widetilde{\boldsymbol{X}}_{\boldsymbol{D}_2}\ \widetilde{\boldsymbol{X}}_{\boldsymbol{D}_3}]$ *is importance sketching covariates;* $\widetilde{\varepsilon} = (\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_n)^\top, \widetilde{\varepsilon}_i = vec(\mathcal{X}_i)^\top P_{\widetilde{U}_\perp} vec(\mathcal{A}) + \varepsilon_i$.

- θ, ρ: how well the sketching directions approximate the true ones;
- $\|(\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1} \widetilde{\boldsymbol{X}}^\top \widetilde{\varepsilon}\|_2^2$: error of dimension-reduced least squares.

Theoretical Analysis under Random Design

- Gaussian ensemble design: $\boldsymbol{\mathcal{X}}_i \overset{iid}{\sim} N(0,1)$, $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

Theorem

Let $\widetilde{\sigma}^2 = \|\boldsymbol{\mathcal{A}}\|_{HS}^2 + \sigma^2$, $\lambda_0 = \min_k \sigma_{r_k}(\boldsymbol{\mathcal{M}}_k(\boldsymbol{\mathcal{A}}))$, $p = \max\{p_1, p_2, p_3\}$, $r = \max\{r_1, r_2, r_3\}$. *Under regularity conditions and* $n = \Omega\left(p^{3/2}r + pr^2\right)$,

$$\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{HS}^2 \leq \frac{m}{n}\left(\sigma^2 + \frac{C\widetilde{\sigma}^2 p}{n}\right)\left(1 + C\sqrt{\frac{\log p}{m}} + C\sqrt{\frac{m\widetilde{\sigma}^2}{n\lambda_0^2}}\right)$$

with high probability. $m = r_1 r_2 r_3 + \sum_{k=1}^{3}(p_k - r_k)r_k$ *is degree of freedom of rank-*(r_1, r_2, r_3) *tensors in* $\mathbb{R}^{p_1 \times p_2 \times p_3}$.

- Sample complexity for provable consistent estimation:

$$n = \Omega\left(p^{3/2}r + pr^2\right).$$

(Outperforms the previous rate, e.g. $n = \Omega(p^2 r)$ (Chen, Raskutti, & Yuan))

Lower Bound

- $\boldsymbol{\mathcal{X}}_i \overset{iid}{\sim} N(0,1)$, $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$. Recall $m = r_1 r_2 r_3 + \sum_{k=1}^{3} r_k(p_k - r_k)$.

Theorem

Consider the following class of low-rank tensors,

$$\mathcal{A}_{\boldsymbol{p},\boldsymbol{r}} = \left\{\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times p_2 \times p_3} : rank(\boldsymbol{\mathcal{A}}) \le (r_1, r_2, r_3)\right\}.$$

If $n > m + 1$,

$$\inf_{\hat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}} \in \mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}} \mathbb{E}\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{HS}^2 \ge \frac{m}{n-m+1}\sigma^2.$$

If $n \le m + 1$,

$$\inf_{\hat{\boldsymbol{\mathcal{A}}}} \sup_{\boldsymbol{\mathcal{A}} \in \mathcal{A}_{\boldsymbol{p},\boldsymbol{r}}} \mathbb{E}\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\right\|_{HS}^2 = \infty.$$

- If $\frac{\sigma^2 + \|\boldsymbol{\mathcal{A}}\|_{HS}^2}{n}\left(\frac{p}{\sigma^2} + \frac{m}{\lambda_0^2}\right) = o(1)$, the upper bound of ISLET is sharp with matching constant to the lower bound.

A Long and Fulfilling Journey to Prove the Theorems

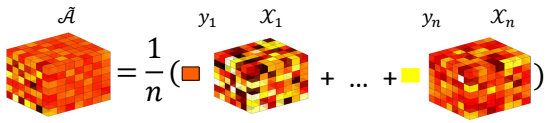
- Optimal one-sided perturbation bound (Cai and Z., *AoS* 2018)
- A sharp bound for HOOI tensor factorization (Z. and Xia, *IEEE-TIT* 2018)
- Sharp error bound under Cross scheme (Z., *AoS* 2018)
- High-order concentration inequality (Hao, Z., Cheng, *AoS* 2018, under revision)
- Careful handeling of tensor algebra
- ...

Computation and Implementation of ISLET

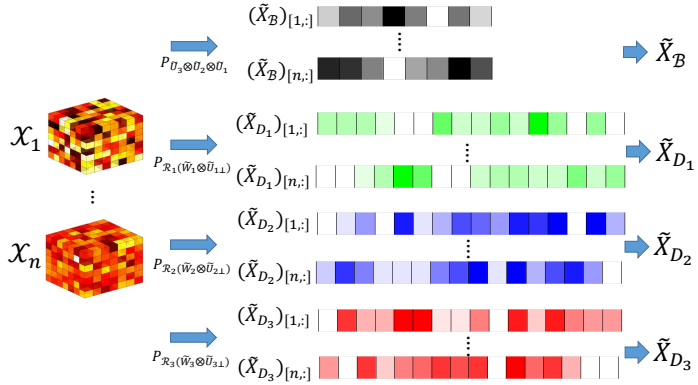
Computation and Implementation of ISLET

- ISLET accesses each sample only twice:

Covariance tensor (Step 1)



Importance sketching covariates (Step 2)

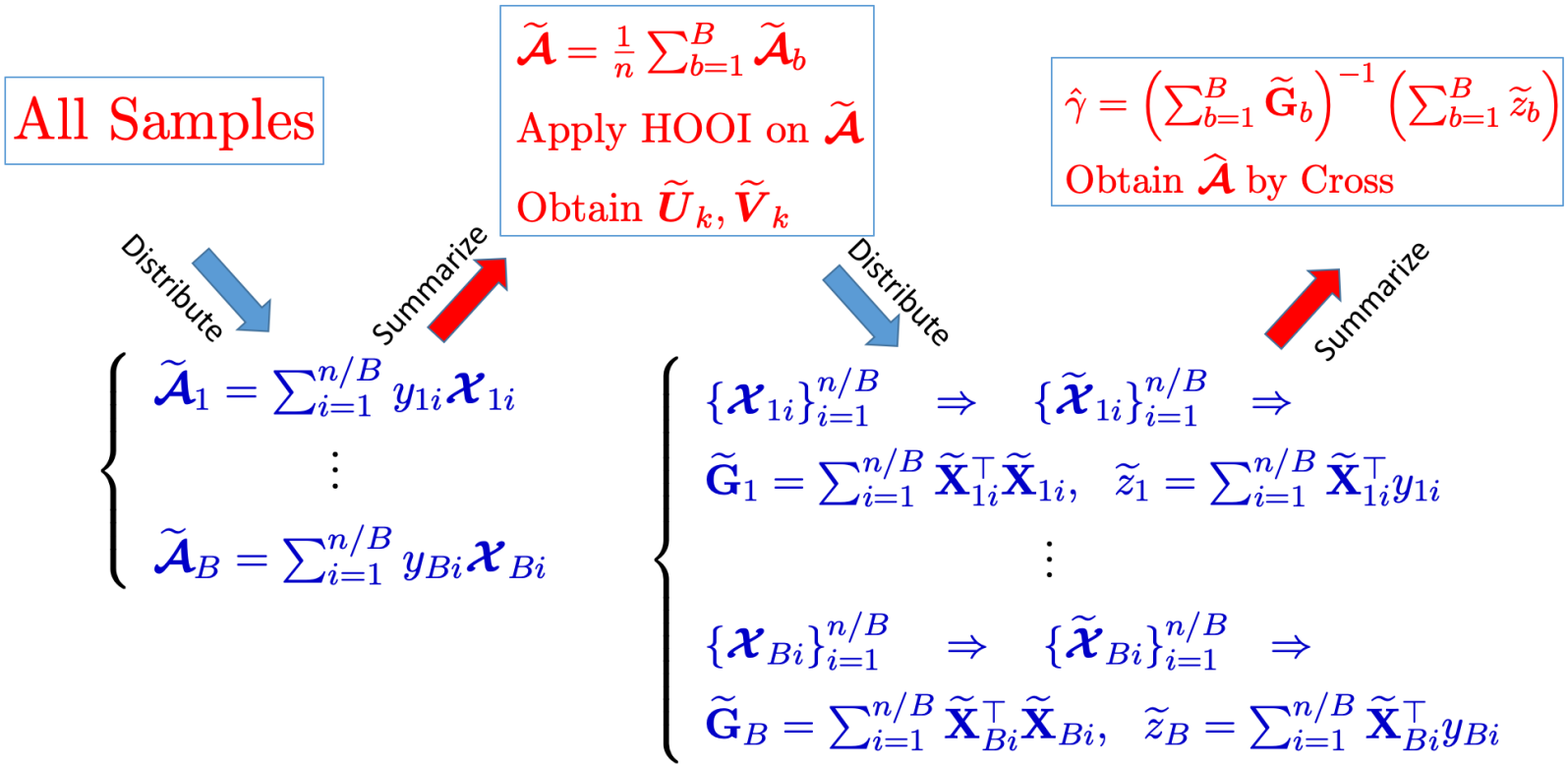


Great save in computational costs in large scale where it is difficult to store the whole dataset into RAM.

- ISLET requires $O(p^3 + n(pr + r^3))$ memory space.
  Computation complexity of ISLET is $O(np^3r + nr^6 + Tp^4)$.

Significantly better than previous methods.

ISLET allows parallel computing conveniently



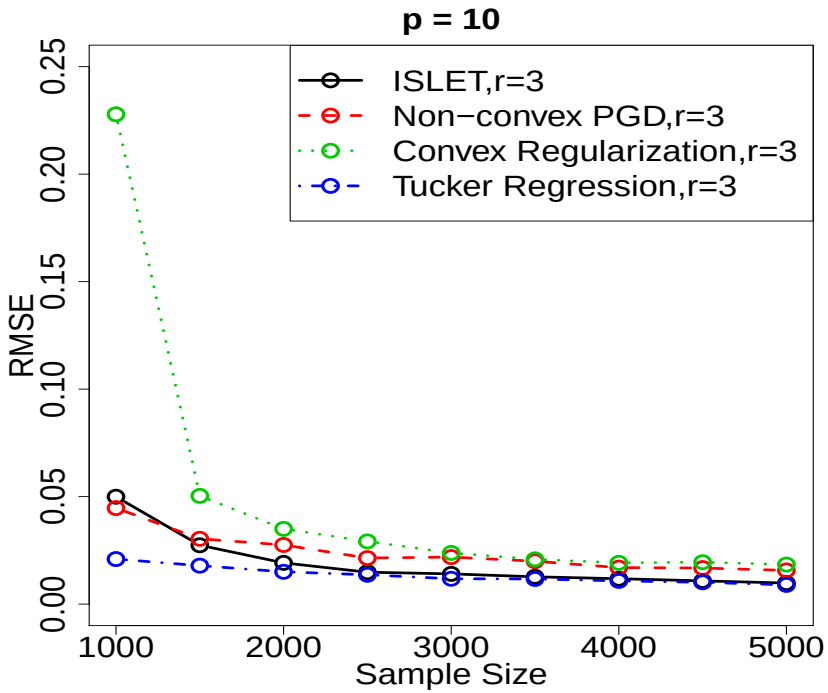
Both computation and storage are highly reduced.

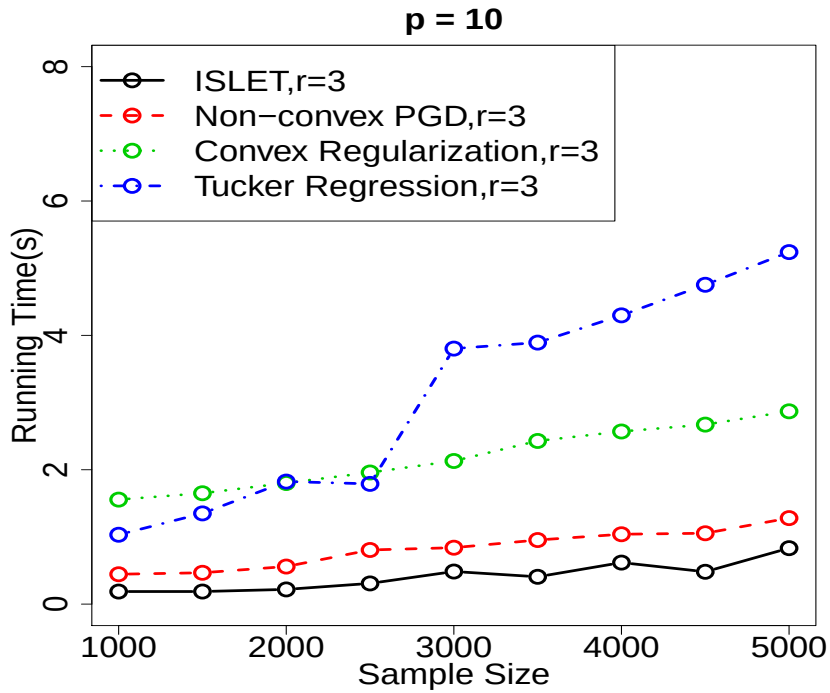
	Computation Complexity	Storage Space
non-parallel	$O\left(np^3r + nr^6 + Tp^4\right)$	$O\left(n(pr + r^3) + p^3\right)$
parallel	$O\left(\frac{np^3r+nr^6}{B} + Tp^4\right)$	$O\left(\frac{n(pr+r^3)}{B} + p^3\right)$

Simulation Studies

Simulation - Comparison with Previous Methods

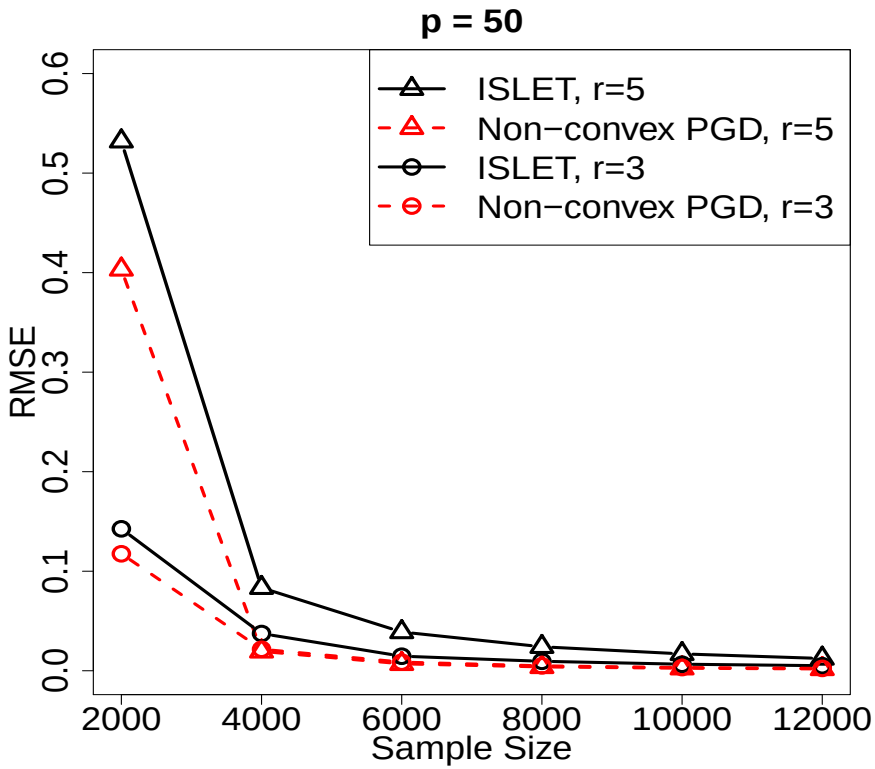
- We compare ISLET with
  1. Overlapped nuclear norm minimization (convex regularization) (Liu et al, 2013)
  2. Non-convex projected gradient descent (PGD) (Chen, Raskutti, Yuan, 2017)
  3. Tucker regression (AGD) (Zhou, Li, Zhu, 2013; Li, Xu, Zhou, Li, 2018)
- Let $p = 10, r = 3, n \in [1000, 5000]$

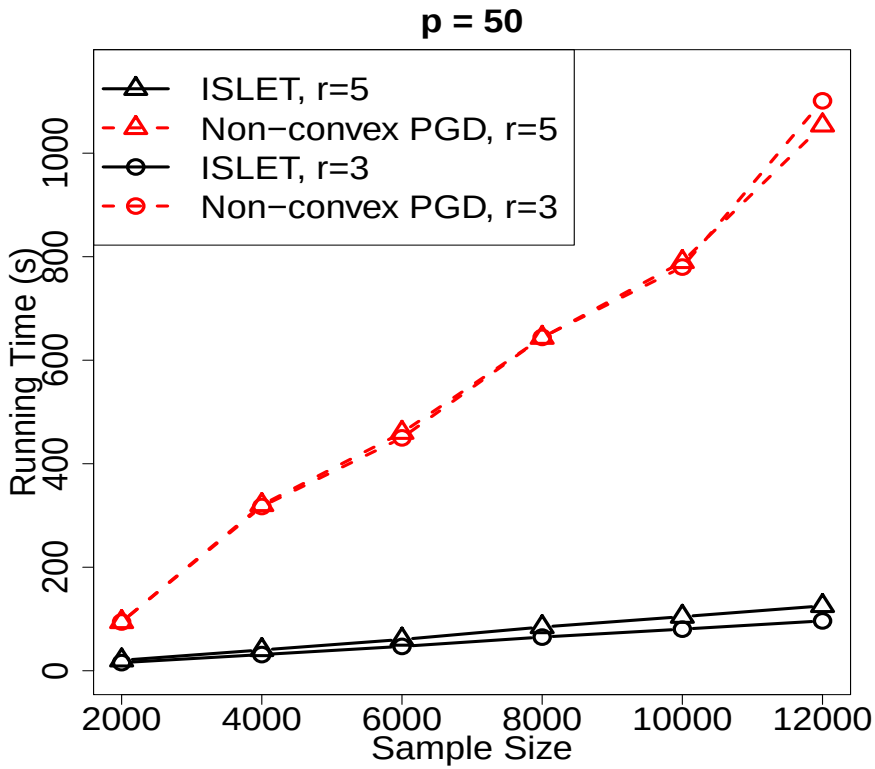




Simulation - Comparison with Previous Methods

- For $p = 50$, we compare ISLET with only non-convex PGD,
  - (The other methods are beyond our computation capability.)

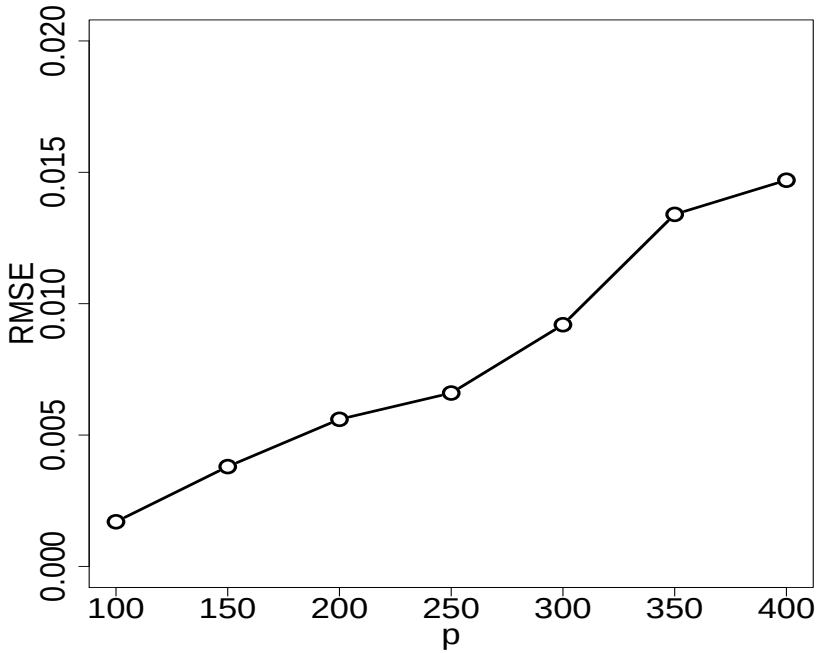


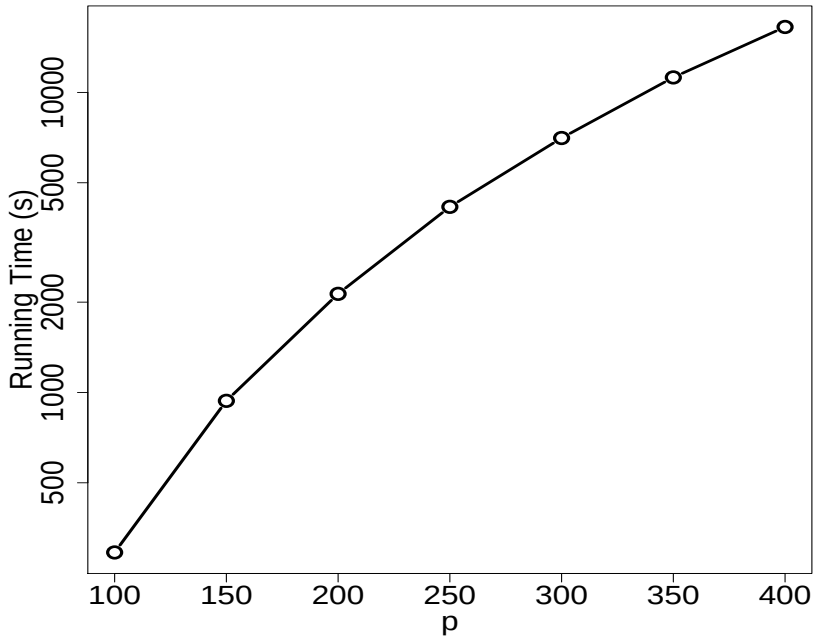


- ISLET achieves similar estimation error with much shorter runtime compared to baseline methods.

Simulation - Ultrahigh-dimensional Settings

- $r = 2, n = 30,000$, p grows to 400.
- Space cost for $\{\boldsymbol{\mathcal{X}}_i\}_{i=1}^n$ is $400^3 \times 30000 \times 4\text{bytes} = 7.68$ TBs
  - Far beyond the capacity of most PCs
  - Possible to perform ISLET since each sample was accessed only twice!
- Distribute on 40 cores and perform parallel computing





- ISLET performs stably in reasonable runtime!

Summary of Simulation Analysis

In summary, ISLET

- has similar or smaller estimation error compared with existing methods
- has much shorter run time
- is more scalable for ultrahigh-dimensional settings
- is better than the state-of-art matrix method for low-rank matrix regression (results are in appendix)

Sparse & Low-rank Tensor Regression and Sparse ISLET

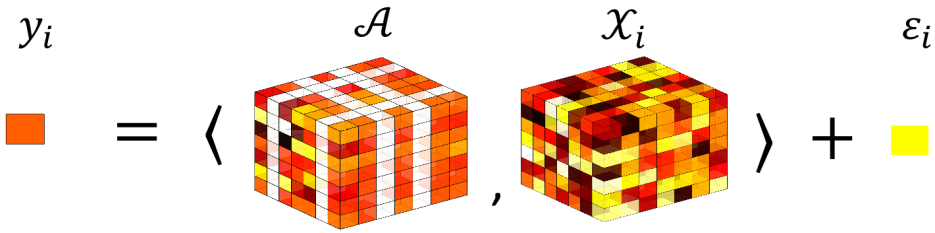
Sparse Settings

- Tensor data with sparsity structures:
  - Brain imaging analysis: only part of the regions are associated with the brain disorders.
  - Microbiome studies: only part of the bacterial taxa are related to clinical phenotypes.

- ISLET can be modified to utilize sparsity structures for better performance.

Sparse low-rank tensor regression model

$$y_i = \langle \mathcal{A}, \mathcal{X}_i \rangle + \varepsilon_i, \quad i = 1, \ldots, n.$$

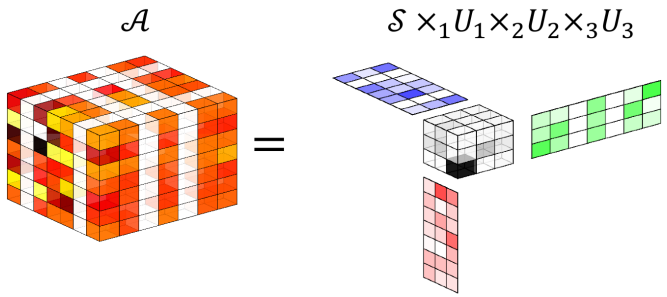


- $\mathcal{A}$ is Tucker low-rank

$$\mathcal{A} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3.$$

- Some of $\boldsymbol{U}_k$ satisfy row-wise sparsity,

$$\|\boldsymbol{U}_k\|_0 = \sum_i 1_{\{\boldsymbol{U}_{k,[i,:]} \neq 0\}} \leq s_k, \quad k \in J_s \subseteq [3]$$

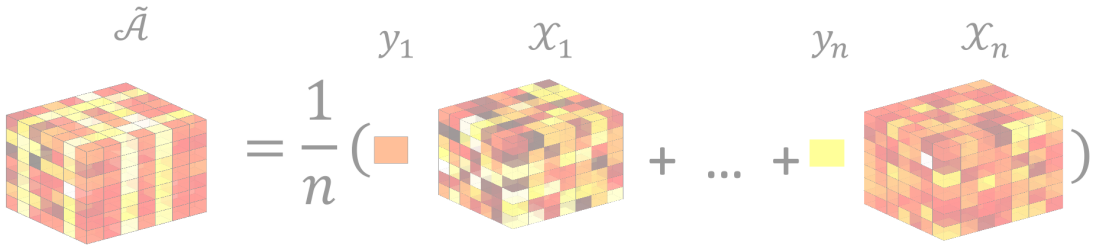


- Goal: estimate $\mathcal{A}$ based on $\{y_i, \boldsymbol{\mathcal{X}}_i\}_{i=1}^n$.

Step 1. Probing Importance Sketching Direction

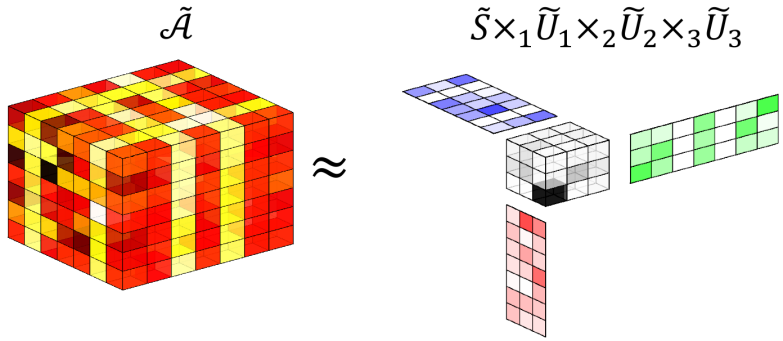
- **(Step 1.1)** Evaluate the sample covariance tensor.

$$\widetilde{\mathcal{A}} = \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{\mathcal{X}}_i$$



- **(Step 1.2)** Apply STAT-SVD (**Z.** and Han, *JASA* 2018) to obtain a factorization of $\widetilde{\mathcal{A}}$:

$$\widetilde{\mathcal{A}} \approx \widetilde{\mathcal{S}} \times_1 \widetilde{\boldsymbol{U}}_1 \times_2 \widetilde{\boldsymbol{U}}_2 \times_3 \widetilde{\boldsymbol{U}}_3$$

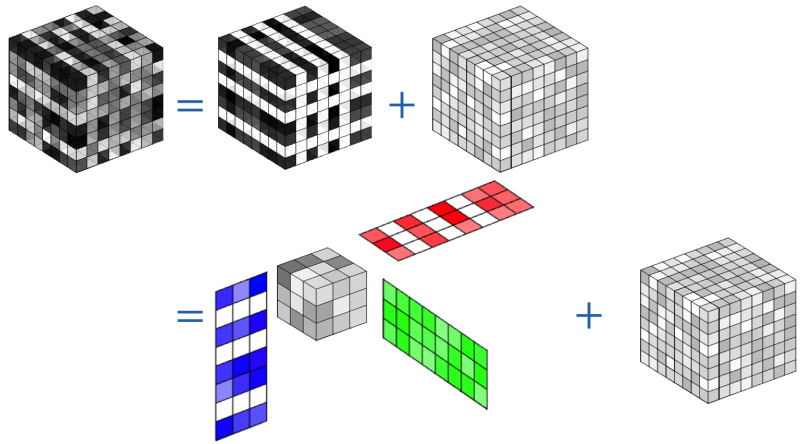


- **(Step 1.3)** Perform QR orthogonalization $\widetilde{V}_k = \mathrm{QR}\left(\mathcal{M}_k^\top\left(\widetilde{S}\right)\right)$.
- **Outcome of Step 1:** $\{\tilde{U}_k, \tilde{V}_k\}_{k=1}^3$.

STAT-SVD:

- *Sparse Tensor Alternating Thresholding-Singular Value Decomposition*;

- An efficient algorithm for sparse tensor factorization based on a novel double thresholding & projections.



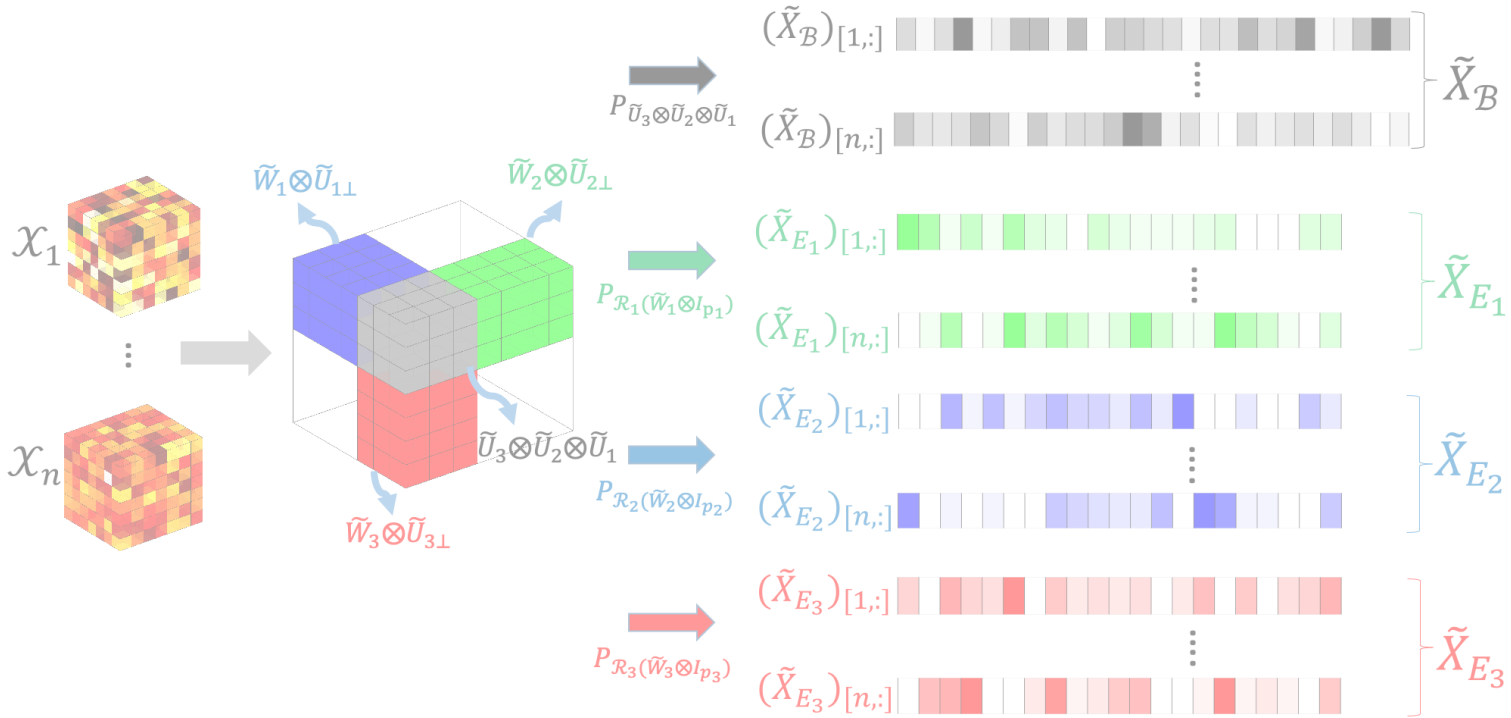
- **Z.** and Han (*JASA*, 2018) proposed the method and established the statistical optimality.

Step 2. Importance Sketching

Construct importance sketching covariates

$$\hat{\boldsymbol{X}}_{\mathcal{B}} \in \mathbb{R}^{n\times(r_1 r_2 r_3)}, \quad \left(\hat{\boldsymbol{X}}_{\mathcal{B}}\right)_{[i,:]} = \mathrm{vec}\left(\boldsymbol{\mathcal{X}}_i \times_1 \widetilde{\boldsymbol{U}}_1^\top \times_2 \widetilde{\boldsymbol{U}}_2^\top \times_3 \widetilde{\boldsymbol{U}}_3^\top\right)$$

$$\hat{\boldsymbol{X}}_{\boldsymbol{E}_k} \in \mathbb{R}^{n\times(p_k - r_k)r_k}, \quad \left(\hat{\boldsymbol{X}}_{\boldsymbol{E}_k}\right)_{[i,:]} = \mathrm{vec}\left(\mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_i \times_{k+1} \widetilde{\boldsymbol{U}}_{k+1}^\top \times_{k+2} \widetilde{\boldsymbol{U}}_{k+2}^\top\right)\right)$$



Step 3. Dimension-Reduced Regression

Perform dimension-reduced regression on importance sketching covariates

- Core: least square estimator

$$\hat{\mathcal{B}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, \quad \mathrm{vec}(\hat{\mathcal{B}}) = \operatorname*{argmin}_{\gamma} \|y - \widetilde{\boldsymbol{X}}_{\mathcal{B}}\gamma\|_2^2;$$

- Non-sparse loading(s): least square estimator

$$\text{for } k \notin J_s, \quad \mathrm{vec}(\hat{\mathcal{B}}) = \operatorname*{argmin}_{\gamma} \|y - \widetilde{\boldsymbol{X}}_{\boldsymbol{E}_k}\gamma\|_2^2;$$

- Sparse Loading(s): apply **group Lasso**

$$k \in J_s, \quad \mathrm{vec}(\hat{\boldsymbol{E}}_k) = \operatorname*{argmin}_{\gamma} \|y - \widetilde{\boldsymbol{X}}_{\boldsymbol{E}_k}\gamma\|_2^2 + \eta_k \sum_{j=1}^{p_k} \|\gamma_{G_j^k}\|_2;$$

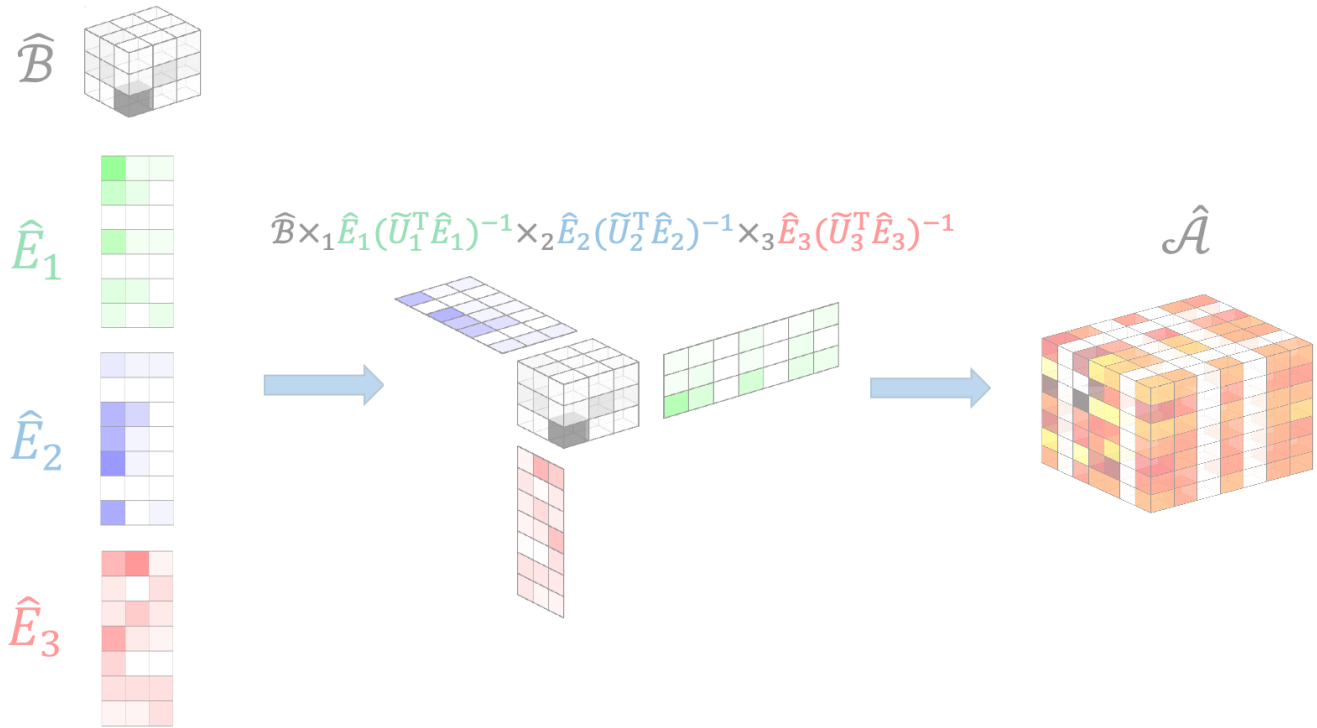
  - $\{G_j^k\}$ form a partition of indices that is induced by the sparsity of $\boldsymbol{U}_k$.

- **Outcome of Step 2**: $\hat{\mathcal{B}}, \hat{\boldsymbol{E}}_1, \hat{\boldsymbol{E}}_2, \hat{\boldsymbol{E}}_3$.

Step 4. Assembling the Final Estimate

Assemble via the Cross scheme (**Z.** 2018)

$$\hat{\mathcal{A}} = \hat{\mathcal{B}} \times_1 (\hat{E}_1 (\widetilde{U}_1^\top \hat{E}_1)^{-1}) \times_2 (\hat{E}_2 (\widetilde{U}_2^\top \hat{E}_2)^{-1}) \times_3 (\hat{E}_3 (\widetilde{U}_3^\top \hat{E}_3)^{-1}).$$



Remark

Sparse ISLET

- only accesses each sample **twice**;
- requires **significantly smaller computation and storage costs** than state-of-the-art methods;
- allows **parallel computing** conveniently;
- is among the first to achieve **provable minimax optimal statistical performance**!

Summary

- In this talk, we considered tensor regression and introduce the ISLET procedure.
  - Central idea: importance sketching
  - ISLET achieves minimax optimal performance
  - Computational efficient, allow parallel computing conveniently
  - Adapts to sparse tensor regression with minimax optimal statistical performance and computational advantages
- Wider applications:
  - Fast tensor/matrix completion
  - High-order interaction pursuits

References

- Zhang, A., Luo, Y., Raskutti, G., and Yuan, M. (2018). ISLET: fast and optimal low-rank tensor regression via importance sketchings. *SIAM Journal on Mathematics of Data Science*, major revision under review.
- Zhang, A. (2018). Cross: Efficient tensor completion. *Annals of Statistics*, to appear.
- Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64, 7311-7338.
- Cai, T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Annals of Statistics*, 43, 102-138.
- Zhang, A. and Han, R. (2018). Optimal denoising and singular value decomposition for sparse high-dimensional high-order data. *Journal of the American Statistical Association*, to appear.
- Hao, B., Zhang, A., and Cheng, G. (2018). Sparse and low-rank tensor estimation via cubic sketchings, under revision.