ISLET: Fast and Optimal Low-Rank Tensor Regression via Importance Sketching

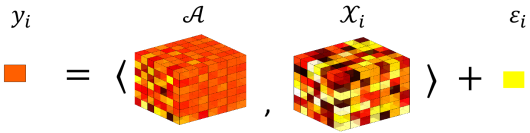
Anru R. Zhang Yuetian Luo Garvesh Raskutti Ming Yuan

SIAM J. Mathematics of Data Science (SIMODS), 2020
DOI: 10.1137/19M126476X

Limitations of Existing Tensor Regression

Tensor regression model:

$$y_i = \langle \mathcal{X}_i, \mathcal{B} \rangle + \varepsilon_i, \quad \mathcal{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_d}, \ \mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_d}.$$



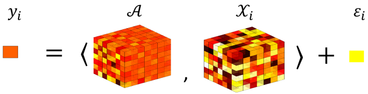
Existing regularization methods:

- **Convex surrogates (nuclear norms):** statistically accurate, but need repeated SVDs on large unfoldings $\Rightarrow$ extremely slow.
- **Nonconvex factorizations:** computationally cheaper, but sensitive to initialization and with weaker guarantees.

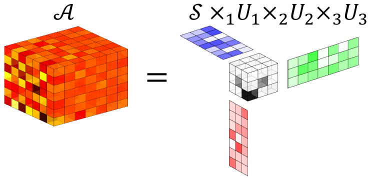
Model

- 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: $\mathcal{A} = \mathcal{S} \times_1 U_1 \times_2 U_2 \times_3 U_3$, where $\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, \quad U_k \in \mathbb{R}^{p_k \times r_k}, \; k = 1, 2, 3.$

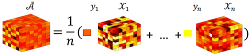


- **Goal:** estimate $\mathcal{A}$ based on $\{y_i, \mathcal{X}_i\}_{i=1}^n$.

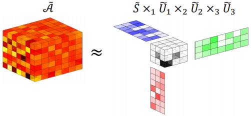
Step 1. Probing Importance Sketching Direction

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

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



- **(Step 1.2)** Apply high-order orthogonal iteration (HOOI) to obtain a low-rank factorization:



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

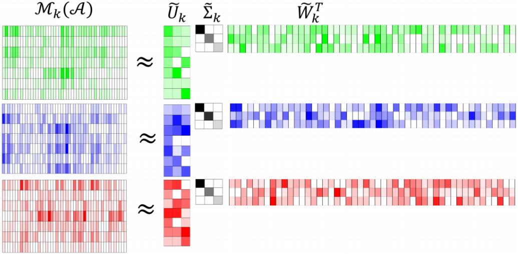
- **(Step 1.3)** Perform QR orthogonalization:

$$\widetilde{V}_k = \mathrm{QR}\left(\mathcal{M}_k^\top(\widetilde{\mathcal{S}})\right).$$

- **Outcome:** $\{\widetilde{U}_k, \widetilde{V}_k\}_{k=1}^3$.

Interpretations of Step 1

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



- $\{\widetilde{U}_k, \widetilde{W}_k\}$ are **importance sketching directions**.
- They are initial sample approximations of $\{U_k, W_k\}$, i.e. left/right singular subspaces of $\mathcal{M}_k(\mathcal{A})$.
- They best align with the true tensor $\mathcal{A}$.

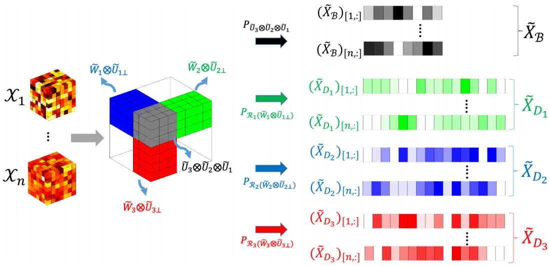
Step 2. Importance Sketching

- Construct **dimension-reduced covariates**:

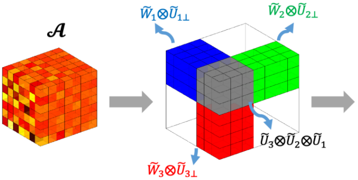
$$\widehat{X}_{\mathcal{B}}[i,:] = \mathrm{vec}\left(\mathcal{X}_i \times_1 \widetilde{U}_1^\top \times_2 \widetilde{U}_2^\top \times_3 \widetilde{U}_3^\top\right),$$

$$\widehat{X}_{D_k}[i,:] = \mathrm{vec}\left(\widetilde{U}_{k\perp}^\top \mathcal{M}_k(\mathcal{X}_i)\, \widetilde{W}_k\right), \quad k = 1, 2, 3.$$

- These sketches reduce both sample and feature dimensions.



Interpretation of Step 2



$$\widetilde{\gamma} = \begin{bmatrix} \mathrm{vec}(\widetilde{\mathcal{B}}) \\ \mathrm{vec}(\widetilde{D}_1) \\ \mathrm{vec}(\widetilde{D}_2) \\ \mathrm{vec}(\widetilde{D}_3) \end{bmatrix}, \qquad \begin{aligned} \widetilde{\mathcal{B}} &:= \mathcal{A} \times_1 \widetilde{U}_1^\top \times_2 \widetilde{U}_2^\top \times_3 \widetilde{U}_3^\top, \\ \widetilde{D}_1 &:= \widetilde{U}_{1\perp}^\top \mathcal{M}_1(\mathcal{A}) \widetilde{W}_1, \\ \widetilde{D}_2 &:= \widetilde{U}_{2\perp}^\top \mathcal{M}_2(\mathcal{A}) \widetilde{W}_2, \\ \widetilde{D}_3 &:= \widetilde{U}_{3\perp}^\top \mathcal{M}_3(\mathcal{A}) \widetilde{W}_3. \end{aligned}$$

- Rewrite the regression model:

$$y_i = \langle \mathcal{X}_i, \mathcal{A} \rangle + \varepsilon_i = \widetilde{X}[i,:]^\top \widetilde{\gamma} + \widetilde{\varepsilon}_i,$$

where

$$\widetilde{X} = [\widehat{X}_{\mathcal{B}}, \widehat{X}_{D_1}, \widehat{X}_{D_2}, \widehat{X}_{D_3}], \quad \widetilde{\gamma} = [\mathrm{vec}(\widetilde{\mathcal{B}}), \mathrm{vec}(\widetilde{D}_1), \mathrm{vec}(\widetilde{D}_2), \mathrm{vec}(\widetilde{D}_3)].$$

- $\widetilde{X}$ are sketching covariates.
- $\widetilde{\gamma}$ is the sketch of $\mathcal{A}$.

Step 3. Dimension-Reduced Regression

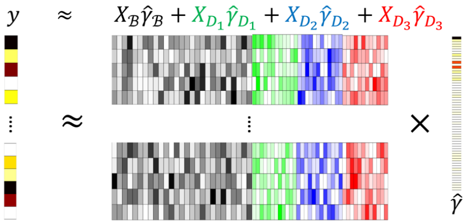
- Perform regression in reduced space:

$$\widehat{\gamma} = \arg\min_{\gamma} \|y - \widetilde{X}\gamma\|_2^2.$$

- Dimension of parameter reduces from $p_1 p_2 p_3$ to

$$m = r_1 r_2 r_3 + \sum_{k=1}^{3} (p_k - r_k) r_k.$$

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

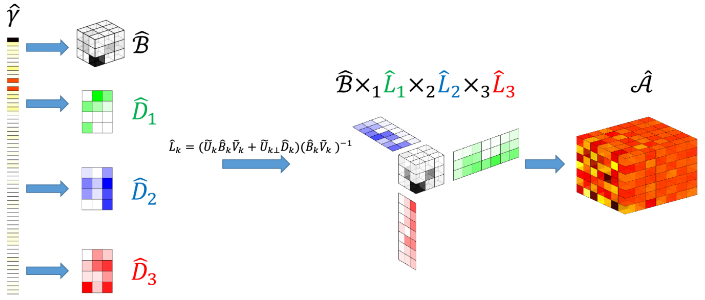


Step 4. Assembling the Final Estimate

- Reconstruct $\widehat{\mathcal{A}}$ via the **Cross scheme** (Z. AoS, 2018):

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

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



Algorithm: ISLET (Summary)

ISLET (Importance Sketching for Tensor Regression)

Input: samples $\{(\mathcal{X}_i, y_i)\}_{i=1}^n$, target ranks (r_1, r_2, r_3)
Output: estimate $\widehat{\mathcal{A}}$

Step 1: Probing directions

$$\widetilde{\mathcal{A}} \leftarrow \frac{1}{n}\sum_{i=1}^n y_i \mathcal{X}_i$$
$$(\widetilde{\mathcal{S}},\ \widetilde{U}_1,\ \widetilde{U}_2,\ \widetilde{U}_3) \leftarrow \mathrm{HOOI}(\widetilde{\mathcal{A}};\ r_1, r_2, r_3)$$
$$\widetilde{V}_k \leftarrow \mathrm{QR}\big(\mathcal{M}_k(\widetilde{\mathcal{S}})^\top\big) \text{ for } k = 1, 2, 3$$

Step 2: Importance sketching (build reduced covariates)

$$\widehat{X}_{\mathcal{B}}[i,:] \leftarrow \mathrm{vec}\big(\mathcal{X}_i \times_1 \widetilde{U}_1^\top \times_2 \widetilde{U}_2^\top \times_3 \widetilde{U}_3^\top\big)$$
$$\widetilde{W}_k \leftarrow (\widetilde{U}_{k+2} \otimes \widetilde{U}_{k+1})\, \widetilde{V}_k, \quad k = 1, 2, 3$$
$$\widehat{X}_{D_k}[i,:] \leftarrow \mathrm{vec}\big(\widetilde{U}_{k\perp}^\top\, \mathcal{M}_k(\mathcal{X}_i)\, \widetilde{W}_k\big), \quad k = 1, 2, 3$$
$$\widetilde{X} \leftarrow [\, \widehat{X}_{\mathcal{B}},\ \widehat{X}_{D_1},\ \widehat{X}_{D_2},\ \widehat{X}_{D_3}\,]$$

Step 3: Dimension-reduced regression

$$\widehat{\gamma} \leftarrow \arg\min_\gamma \|y - \widetilde{X}\gamma\|_2^2 \qquad \text{(use Group Lasso if sparse)}$$

Step 4: Assemble estimate

Reconstruct $\widehat{\mathcal{A}}$ from $\widehat{\gamma}$ and $(\widetilde{U}_k, \widetilde{V}_k)$ via the Cross scheme.

Oracle Inequalities (General Design)

Theorem (Oracle Inequality)

Consider order-3 tensor regression with low Tucker rank (r_1, r_2, r_3). Under mild conditions (angle error $\theta < \frac{1}{2}$, nonsingular sketches, GRIP in the sparse case), the ISLET estimator satisfies

$$\|\widehat{\mathcal{A}} - \mathcal{A}^\star\|_{\mathrm{HS}}^2 \ \leq \ C\Big(\frac{\sigma^2 m}{n} + \mathit{bias}(\theta)\Big),$$

where $m = r_1 r_2 r_3 + \sum_{k=1}^{3}(p_k - r_k) r_k$.

Takeaway. Error decomposes as variance $O(m/n)$ plus controlled sketching bias.

Optimal Risk under Gaussian Design

Theorem (Minimax Risk under Gaussian Design)

Suppose the observed variables is i.i.d. Gaussian and the noise is $\mathcal{N}(0, \sigma^2)$. *Then*

$$\mathbb{E}\|\widehat{\mathcal{A}} - \mathcal{A}^\star\|_{\mathrm{HS}}^2 = (1 + o(1))\frac{m\sigma^2}{n}, \qquad m = r_1 r_2 r_3 + \sum_{k=1}^{3}(p_k - r_k) r_k.$$

Takeaway. Achieves minimax-optimal rate with sharp constant; m equals the degrees of freedom of the Tucker rank class.

Simulation Study: Experimental Setup

Goal. Evaluate the performance of ISLET under synthetic low-rank tensor regression.

Data Generation.

- Covariates: $\mathcal{X}_j \in \mathbb{R}^{p \times p \times p}$ with i.i.d. $\mathcal{N}(0,1)$ entries.
- Coefficient tensor:

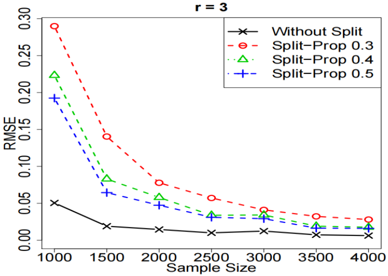
$$\mathcal{A} = [\![\mathcal{S}; E_1, E_2, E_3]\!]$$

  - *Nonsparse setting:* $\mathcal{S}$ and E_k Gaussian random. - *Sparse setting:* rows of E_k randomly zeroed out, sparsity level s_k.
- Responses: $y_j = \langle \mathcal{X}_j, \mathcal{A} \rangle + \varepsilon_j, \;\; \varepsilon_j \sim \mathcal{N}(0, \sigma^2)$.

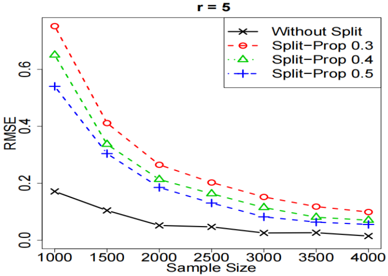
Evaluation.

- Normalized RMSE: $\|\widehat{\mathcal{A}} - \mathcal{A}\|_{\mathrm{HS}} / \|\mathcal{A}\|_{\mathrm{HS}}$.
- Results averaged over 100 repetitions.

Simulation Study: Results

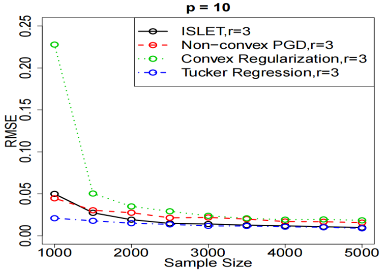


(a) $r = 3$

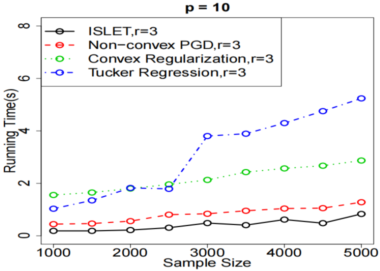


(b) $r = 5$

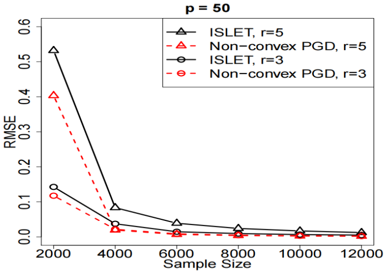
Simulation Study: Results



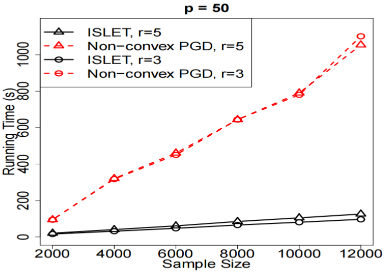
(a) RMSE



(b) Run Time



(c) RMSE



(d) Run Time