Attention boosted Individualized Regression

Presenter: **YANG Guang**
Department of Data Science
City University of Hong Kong

Joint with **CAO Yuan** and **FENG Long**
Department of Statistics and Actuarial Science
The University of Hong Kong

NeurIPS 2024

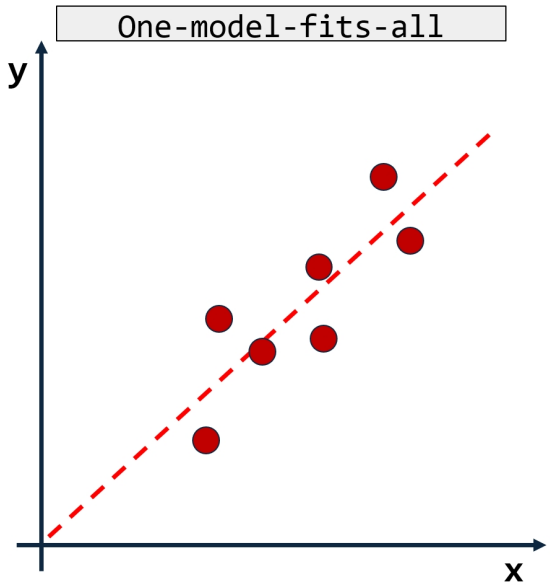
Contents

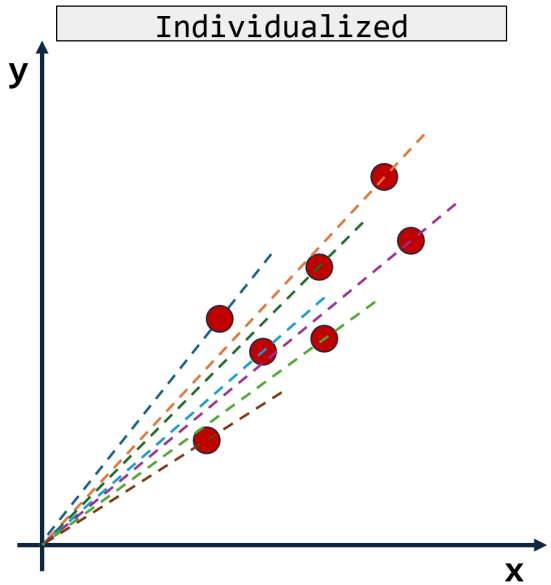
Contents

Introduction

Principle

- One-model-fits-all
- Individualized model
  - Varying-coefficient model (Time, spatially, etc.)
  - Sample-specific model (Coefficients similarity)
  - Self-attention mechanism (Contextual relationship)





Contents

Individualized model

Individualized linear regression

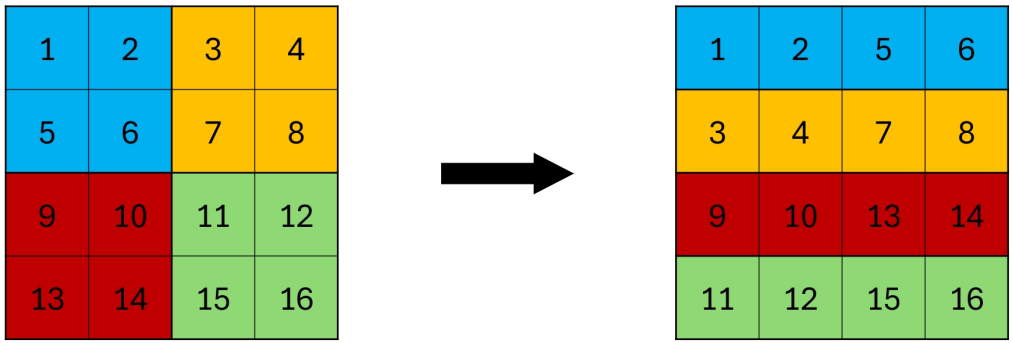
- Ordinary linear regression model for matrix-valued input

$$y_i = \langle \boldsymbol{X}_i, \boldsymbol{C} \rangle + \varepsilon_i, \ \ i = 1, \ldots, n.$$

- Individualized linear regression model for matrix-valued input

$$y_i = \langle \boldsymbol{X}_i, \boldsymbol{C}_i \rangle + \varepsilon_i, \ \ i = 1, \ldots, n.$$

To study internal relations among patches, we take reshaped images as input, namely $\boldsymbol{X}_i = \mathcal{R}_{(d_1,d_2)}(\boldsymbol{X}_i^{\text{ori}})$. Subsequently, $\boldsymbol{X}_i \boldsymbol{X}_i^\top$ reflects relations among patches within i-th sample, called **internal relations**.



Attention boosted individualized regression

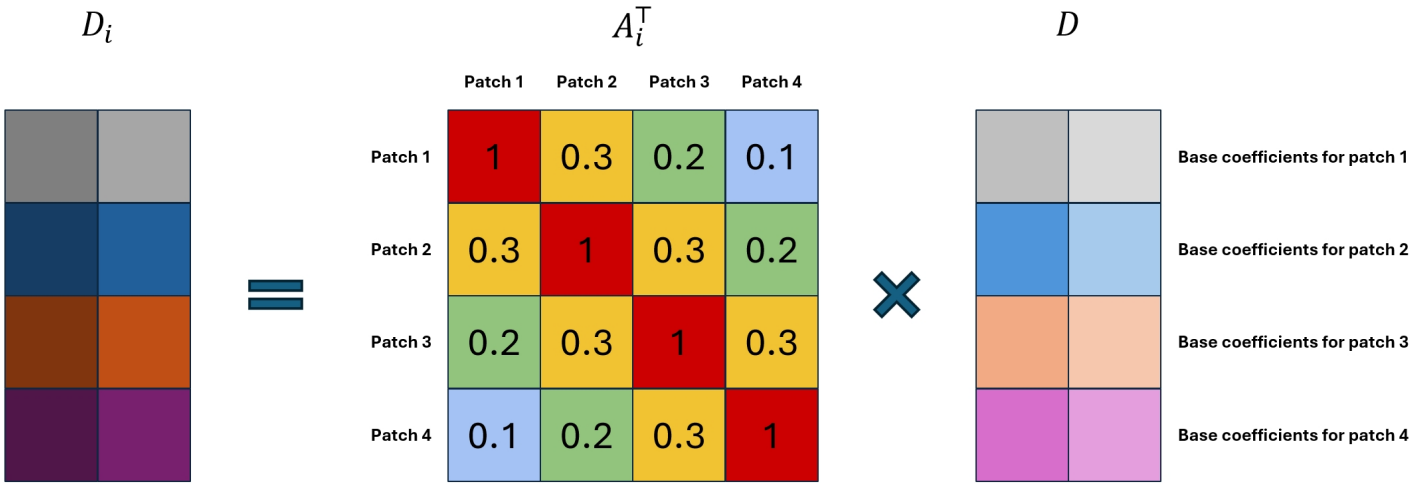
Decomposition of coefficients

I. Decomposition:

$$\underset{\text{Individualized coefficient}}{\boldsymbol{C}_i} = \underset{\text{homogeneous coefficient}}{\boldsymbol{C}} + \underset{\text{heterogeneous coefficient}}{\boldsymbol{D}_i}.$$

II. Aggregation: $\boldsymbol{D}_i = \boldsymbol{A}_i^\top \boldsymbol{D}$.

III. Internal relations: $\boldsymbol{A}_i = g(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{X}_i^\top)$.



Attention boosted individualized regression

Rotational correlation

For any two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, the rotational correlation is defined as

$$\max_{\boldsymbol{H}} \boldsymbol{u}^\top \boldsymbol{H} \boldsymbol{v},$$

where the matrix $\boldsymbol{H}$ is usually required to be orthogonal. This rotational correlation aims to find the maximized correlation between $\boldsymbol{u}$ and $\boldsymbol{v}$ with the best possible rotation.

Explanation

Note that the (j,k)-th element of the sample-specific factor can be written as

$$\{\boldsymbol{A}_i\}_{jk} = \{\boldsymbol{X}_i\}_{j\cdot} \boldsymbol{W} \{\boldsymbol{X}_i\}_{k\cdot}^\top,$$

where $\{\boldsymbol{X}_i\}_{j\cdot}$ and $\{\boldsymbol{X}_i\}_{k\cdot}$ are the j-th and k-th rows of $\boldsymbol{X}_i$, respectively.
To say, $\{\boldsymbol{A}_i\}_{jk}$ is related to the rotational correlation between $\{\boldsymbol{X}_i\}_{j\cdot}$ and $\{\boldsymbol{X}_i\}_{k\cdot}$.

Our goal is not to maximize the rotational correlation between $\{\boldsymbol{X}_i\}_{j\cdot}$ and $\{\boldsymbol{X}_i\}_{k\cdot}$, but to find the optimal rotation that achieves the best fitting for the responses.

Attention boosted individualized regression

Attention boosted individualized regression (AIR)

Now we obtain our individualized model in the following form:

$$y_i = \underbrace{\langle \boldsymbol{X}_i, \boldsymbol{C} \rangle}_{\text{homogeneous}} + \underbrace{\langle \boldsymbol{X}_i, g(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{X}_i^\top)^\top \boldsymbol{D} \rangle}_{\text{heterogeneous}} + \varepsilon_i.$$

Here, $\boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{p \times d}$, and $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ are the coefficient matrices to be learned.

Optimization

To learn $\boldsymbol{C}$, $\boldsymbol{D}$ and $\boldsymbol{W}$, we propose the following penalized minimization problem

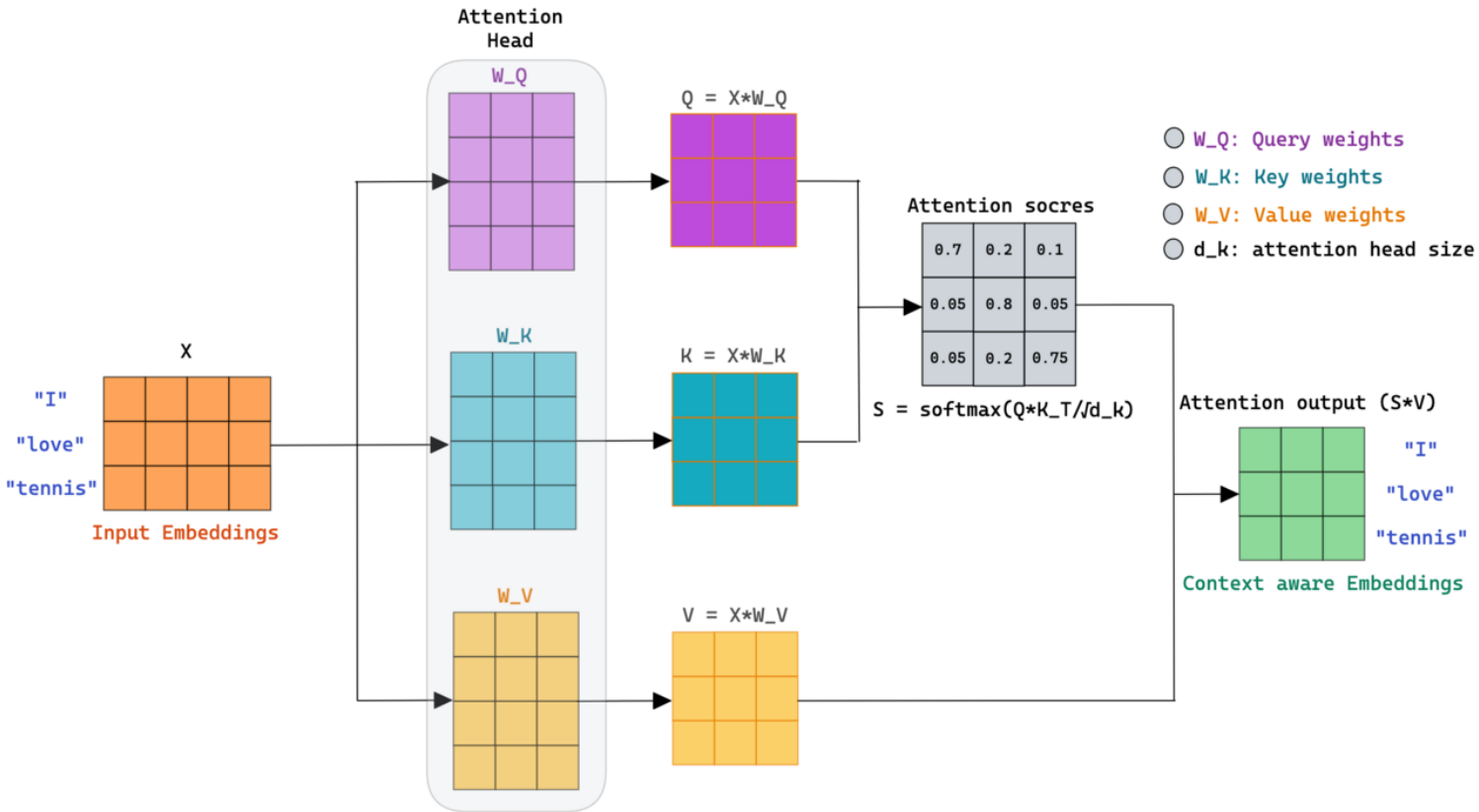
$$\begin{aligned} \min_{\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{W}} \quad & \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{X}_i, \boldsymbol{C}_i \rangle \right)^2 + \lambda_1 \|\boldsymbol{C}\|_F^2 + \lambda_2 \|\boldsymbol{D}\|_F^2, \\ \text{s.t.} \quad & \boldsymbol{C}_i = \boldsymbol{C} + g(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{X}_i^\top)^\top \boldsymbol{D}, \ \|\boldsymbol{W}\|_F = 1, \end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm and λ_1 and λ_2 are regularization parameters to balance the homogeneous and heterogeneous effects.

Self-attention mechanisms

Given the input $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, with $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q$, $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K$ and $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V$, the Scaled Dot-Product Attention mechanism computes

$$f(\boldsymbol{X}) = \text{softmax}\left(\boldsymbol{Q}\boldsymbol{K}^\top / \sqrt{d_k}\right)\boldsymbol{V}.$$



Self-attention mechanisms

General self-attention

Beyond softmax function, normalization in attention could also be accomplished using a general function $g(\cdot)$. We obtain the following generalized attention:

$$f(\boldsymbol{X}) = g\left(\boldsymbol{Q}\boldsymbol{K}^\top\right)\boldsymbol{V}.$$

Linear self-attention

Shen et al. (2021) considered a linear function as scaling normalization:

$$f(\boldsymbol{X}) = \frac{1}{n}\boldsymbol{Q}\boldsymbol{K}^\top\boldsymbol{V}.$$

Linear attention mechanisms are efficient because they bypass the need to compute $n \times n$ matrices by using associative multiplication, reducing complexity from $O(n^2)$ to $O(n)$. While on the other hand, experiments show that Linear attentions does not result in a significant compromise in performance.

Individualized regression and attention mechanism

Focus on the individualized model with only heterogeneous part

$$y_i = \langle \boldsymbol{X}_i, g(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{X}_i^\top)^\top \boldsymbol{D} \rangle + \varepsilon_i.$$

Proposition 1

Suppose the above model holds and $\boldsymbol{W}$ and $\boldsymbol{D}$ could be decomposed as below
(I) $\boldsymbol{W} = \boldsymbol{W}_Q \boldsymbol{W}_K^\top$ for two matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d \times d_k}$ with $d_k \leqslant d$,
(II) $\boldsymbol{D} = \boldsymbol{B} \boldsymbol{W}_V^\top$ for two matrices $\boldsymbol{B}, \boldsymbol{W}_V \in \mathbb{R}^{d \times d_v}$ with $d_v \leqslant d$.
where $d_k, d_v \leqslant d$. Then, for each sample $\boldsymbol{X}_i$, the following equation holds

$$\left\langle \boldsymbol{X}_i,\ g(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{X}_i^\top)^\top \boldsymbol{D} \right\rangle = \left\langle g(\boldsymbol{Q}_i \boldsymbol{K}_i^\top) \boldsymbol{V}_i,\ \boldsymbol{B} \right\rangle,$$

where

$$\boldsymbol{Q}_i = \boldsymbol{X}_i \boldsymbol{W}_Q, \quad \boldsymbol{K}_i = \boldsymbol{X}_i \boldsymbol{W}_K, \quad \boldsymbol{V}_i = \boldsymbol{X}_i \boldsymbol{W}_V. \tag{1}$$

Computation

Equation

$$\left\langle \boldsymbol{X}_i, \boldsymbol{X}_i \boldsymbol{W}^\top \boldsymbol{X}_i^\top \boldsymbol{D} \right\rangle = \left\langle \boldsymbol{X}_i^\top \boldsymbol{D} \boldsymbol{X}_i^\top \boldsymbol{X}_i, \boldsymbol{W} \right\rangle = \left\langle \boldsymbol{X}_i \boldsymbol{W} \boldsymbol{X}_i^\top \boldsymbol{X}_i, \boldsymbol{D} \right\rangle.$$

Initialization

Let $\boldsymbol{w} = \mathrm{vec}(\boldsymbol{W})$ and $\boldsymbol{d} = \mathrm{vec}(\boldsymbol{D})$ be the vectorization of $\boldsymbol{W}$ and $\boldsymbol{D}$. It holds that

$$\left\langle \boldsymbol{X}_i^\top \boldsymbol{D} \boldsymbol{X}_i^\top \boldsymbol{X}_i, \boldsymbol{W} \right\rangle = \left\langle \boldsymbol{Z}_i, \boldsymbol{w}\boldsymbol{d}^\top \right\rangle, \quad \text{where } \boldsymbol{Z}_i = \left(\boldsymbol{X}_i^\top \boldsymbol{X}_i \right) \otimes \boldsymbol{X}_i^\top$$

We start our algorithm by initializing $\boldsymbol{w}$ as the top left singular vector of $\sum_{i=1}^n y_i \boldsymbol{Z}_i$,

$$\widehat{\boldsymbol{w}}^{(0)} = \mathrm{SVD}_u \left(\sum_{i=1}^n y_i \boldsymbol{Z}_i \right).$$

Computation

Alternating minimization algorithm

Given $\widehat{\boldsymbol{W}}^{(t-1)}$, denote $\boldsymbol{U}_i^{(t-1)} = \boldsymbol{X}_i \widehat{\boldsymbol{W}}^{(t-1)} \boldsymbol{X}_i^\top \boldsymbol{X}_i$,

$$\left(\widehat{\boldsymbol{C}}^{(t)}, \widehat{\boldsymbol{D}}^{(t)}\right) = \underset{\boldsymbol{C},\boldsymbol{D}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left\langle \left[\boldsymbol{X}_i, \boldsymbol{U}_i^{(t-1)}\right], \left[\boldsymbol{C}, \ \boldsymbol{D}\right] \right\rangle\right)^2 + \lambda_1 \|\boldsymbol{C}\|_F^2 + \lambda_2 \|\boldsymbol{D}\|_F^2.$$

Given $\left(\widehat{\boldsymbol{C}}^{(t)}, \widehat{\boldsymbol{D}}^{(t)}\right)$,

$$\widehat{\boldsymbol{W}}^{(t)} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left\langle \boldsymbol{X}_i, \widehat{\boldsymbol{C}}^{(t)} \right\rangle - \left\langle \boldsymbol{X}_i^\top \widehat{\boldsymbol{D}}^{(t)} \boldsymbol{X}_i^\top \boldsymbol{X}_i, \boldsymbol{W} \right\rangle\right)^2,$$

$$\widehat{\boldsymbol{W}}^{(t)} = \widehat{\boldsymbol{W}}^{(t)} / \|\widehat{\boldsymbol{W}}^{(t)}\|_F.$$

Contents

Theoretical analysis

We focus on the heterogeneous part of the model. Let $\boldsymbol{w} = \text{vec}(\boldsymbol{W})$ and $\boldsymbol{d} = \text{vec}(\boldsymbol{D})$, the optimization problem could be written as

$$\min_{\boldsymbol{d},\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i - \left\langle \left(\boldsymbol{X}_i^\top \boldsymbol{X}_i \right) \otimes \boldsymbol{X}_i^\top, \boldsymbol{w}\boldsymbol{d}^\top \right\rangle \right\}^2 + \lambda_2 \|\boldsymbol{d}\|_2^2.$$

For the rearranged images $\boldsymbol{X}_i$ for $i = 1, \ldots, n$, we define

$$\boldsymbol{Z} = \left(\text{vec}\left\{ \left(\boldsymbol{X}_1^\top \boldsymbol{X}_1 \right) \otimes \boldsymbol{X}_1^\top \right\}, \ \ldots, \ \text{vec}\left\{ \left(\boldsymbol{X}_n^\top \boldsymbol{X}_n \right) \otimes \boldsymbol{X}_n^\top \right\} \right)^\top.$$

For the new feature matrix $\boldsymbol{Z}$, we suppose the following RIP condition.

Condition 1

(Restricted Isometry Property) For each integer $r = 1, 2, \ldots$, a matrix $\boldsymbol{P} \in \mathbb{R}^{n \times q_1 q_2}$ is said to satisfy the r-RIP condition with constant $\delta_r \in (0, 1)$, if for all $\boldsymbol{M} \in \mathbb{R}^{q_1 \times q_2}$ of rank at most r, it holds that

$$(1 - \delta_r)\|\boldsymbol{M}\|_F^2 \leqslant 1/n \|\boldsymbol{P}\text{vec}(\boldsymbol{M})\|_2^2 \leqslant (1 + \delta_r)\|\boldsymbol{M}\|_F^2.$$

Theorems

Theorem 1. Estimation (Informal)

Suppose the heterogeneous model holds and solved by alternating minimization algorithm. Denote μ_0 as the initial distance, $\kappa_1, \kappa_2 < 1$ as contraction quantities and τ_1 and τ_2 as noise-related terms. Then, after t iterations we have

$$\text{dist}\left(\widehat{\boldsymbol{W}}^{(t)}, \boldsymbol{W}\right) \leqslant (\kappa_1\kappa_2)^t\mu_0 + \frac{\kappa_1\tau_2 + \tau_1}{1-\kappa_1\kappa_2},$$
$$\text{dist}\left(\widehat{\boldsymbol{D}}^{(t)}, \boldsymbol{D}\right) \leqslant \kappa_1^{t-1}\kappa_2^t\mu_0 + \frac{\kappa_2\tau_1 + \tau_2}{1-\kappa_1\kappa_2}.$$

Theorem 2. Prediction (Informal)

With similar conditions in Theorem 1, after t iterations we have

$$\|\widehat{\boldsymbol{Y}}^{(t)} - \boldsymbol{Y}\|_2 \leqslant 3\|\boldsymbol{D}\|_F\sqrt{1+\delta_2}\left\{(\nu_1\nu_2)^{t-1}\mu_0 + \frac{\tau_1+\tau_2}{1-\nu_1\nu_2}\right\}.$$

Contents

Simulation study

- Size of images $\boldsymbol{X}_i^{\text{ori}}$: 28×28.
- Sample size: 4000 for trianing and 1000 for testing.
- Noise: $\varepsilon_i \sim \mathcal{N}(0, 1)$.
- Coefficients: $\boldsymbol{C}$ and $\boldsymbol{D}$ are generated as two circles depicted in Figure 1.
- Internal relations: assumed among blocks of size 4×4 within each image, where two blocks at random locations are correlated.
- Consider two cases: $y_i = \langle \boldsymbol{X}_i, \boldsymbol{C} \rangle + \langle \boldsymbol{X}_i, \boldsymbol{D}_i \rangle + \varepsilon_i$, subject to $\boldsymbol{D}_i = \boldsymbol{A}_i^\top \boldsymbol{D}$.
  C1. With specific $\boldsymbol{W}$: $\boldsymbol{A}_i = \boldsymbol{X}_i \boldsymbol{W} \boldsymbol{X}_i^\top$ and $\boldsymbol{W} = 2 \cdot \boldsymbol{u}_1 \boldsymbol{v}_1^\top + 1 \cdot \boldsymbol{u}_2 \boldsymbol{v}_2^\top$.
  C2. Without specific $\boldsymbol{W}$: $\boldsymbol{A}_i = \text{coef}(\boldsymbol{X}_i)$.
- Comparison methods:
  - Regularized matrix regression (LRMR, Zhou and Li, 2014)
  - Tensor regression with lasso penalty (TRLasso, Zhou et al., 2013)
  - Deep kronecker network (DKN, Feng and Yang, 2023)
  - Vision transformer (ViT, Dosovitskiy et al., 2020)

Simulation study

Furthermore, we consider different degrees of individuation (DI) by the relative magnitude of the heterogeneous part and homogeneous part. Specifically, $\text{DI} = \sqrt{\sum_{i=1}^{n}\langle \boldsymbol{X}_i, \boldsymbol{D}_i\rangle^2 / \sum_{i=1}^{n}\langle \boldsymbol{X}_i, \boldsymbol{C}\rangle^2}$

Table: Prediction errors of different methods.

Methods	DI=0.5	DI=1.0	DI=2.0
	Case 1		
AIR	**4.422 (0.130)**	8.102 (0.325)	**10.599 (0.816)**
LRMR	6.616 (0.020)	13.101 (0.040)	26.239 (0.081)
TRLasso	8.215 (0.021)	14.465 (0.044)	27.007 (0.085)
DKN	4.886 (0.018)	**7.028 (0.032)**	11.741 (0.043)
ViT	18.429 (0.049)	18.351 (0.047)	24.098 (0.069)
	Case 2		
AIR	**3.590 (0.046)**	**6.632 (0.022)**	**13.002 (0.044)**
LRMR	6.766 (0.018)	13.408 (0.037)	26.864 (0.074)
TRLasso	8.337 (0.021)	14.739 (0.039)	27.484 (0.073)
DKN	8.269 (0.018)	14.964 (0.034)	28.686 (0.060)
ViT	24.492 (0.063)	29.939 (0.084)	44.036 (0.111)

Simulation study

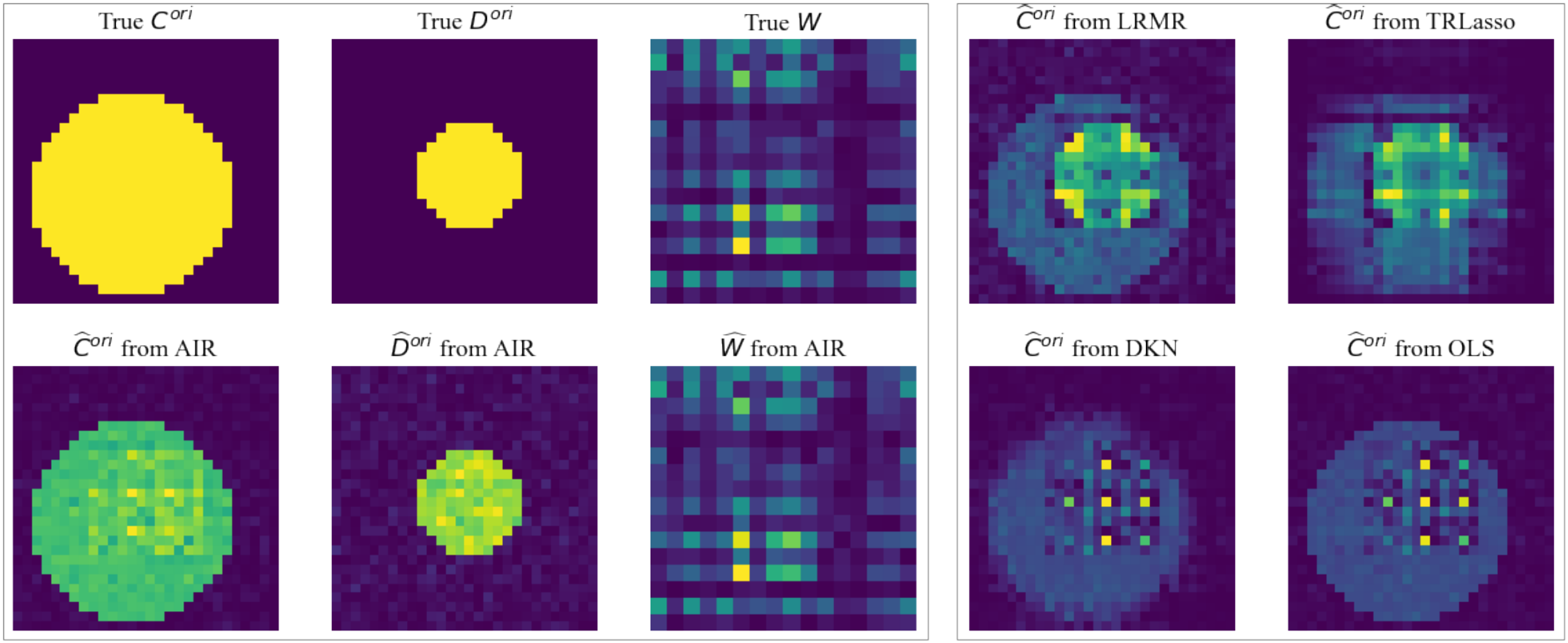


Figure: Case 1 simulation results with DI $= 1.0$. The first three columns show true parameters and estimations from AIR. The last two columns show estimations from other methods except ViT, as it has no explicit coefficient matrix. An additional OLS estimation is added for reference.

Simulation study

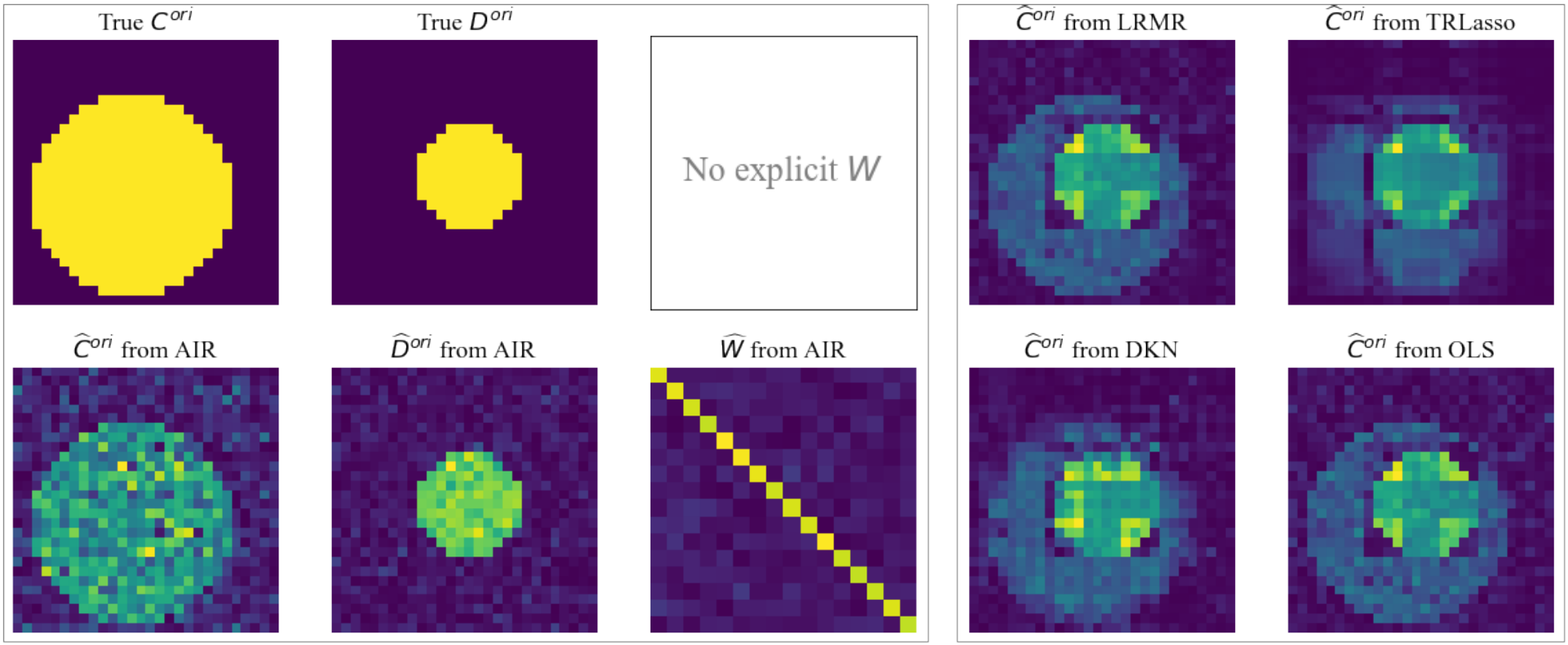


Figure: Case 2 simulation results with DI $= 1.0$. There does not exist an explicit true $\boldsymbol{W}$ while the internal relation matrix $\boldsymbol{A}_i$ is computed directly by patchwise Pearson correlation coefficients. Because such $\boldsymbol{A}_i$ is close to a diagnal matrix, it is rational that $\widehat{\boldsymbol{W}}$ from AIR is close to a diagnal matrix.

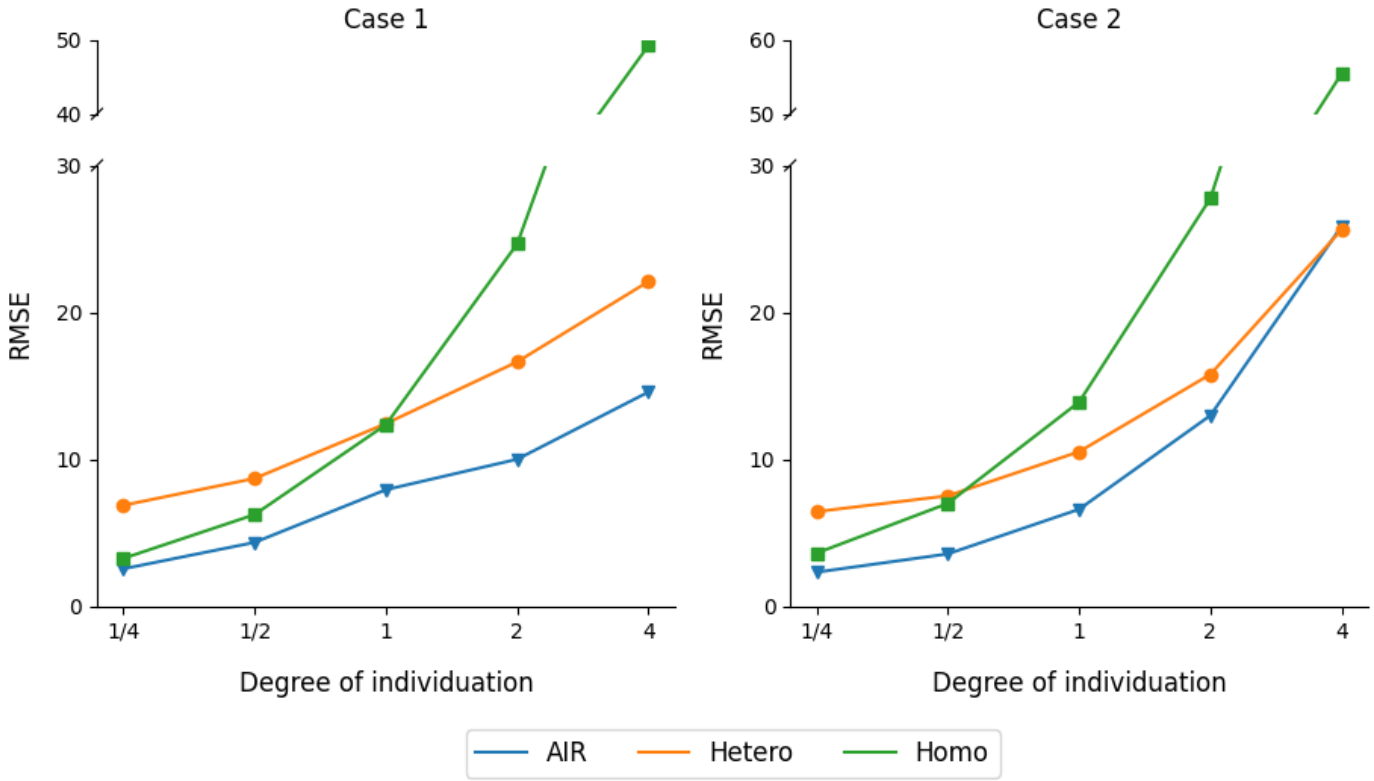
Ablation study

We compare

(1) AIR: $y_i = \langle \boldsymbol{X}_i, \boldsymbol{C} \rangle + \langle \boldsymbol{X}_i, \boldsymbol{D}_i \rangle + \varepsilon_i$, subject to $\boldsymbol{D}_i = \boldsymbol{X}_i \boldsymbol{W}^\top \boldsymbol{X}_i^\top \boldsymbol{D}$.

(2) Hetero: $y_i = \langle \boldsymbol{X}_i, \boldsymbol{D}_i \rangle + \varepsilon_i$, subject to $\boldsymbol{D}_i = \boldsymbol{X}_i \boldsymbol{W}^\top \boldsymbol{X}_i^\top \boldsymbol{D}$.

(3) Homo: $y_i = \langle \boldsymbol{X}_i, \boldsymbol{C} \rangle + \varepsilon_i$.

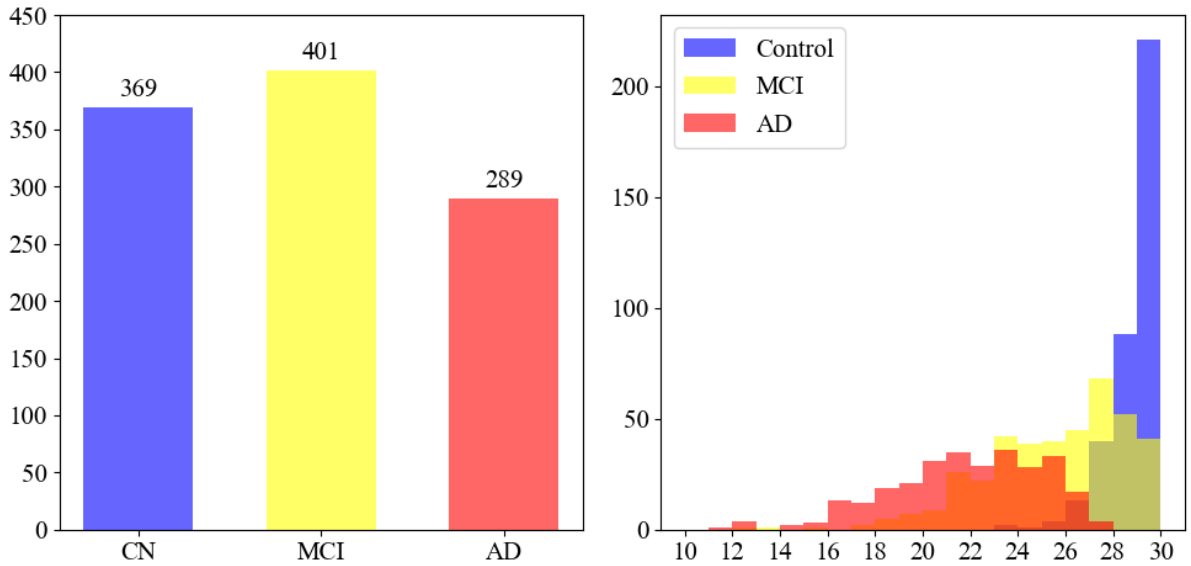


ADNI study

Alzheimer's Disease Neuroimaging Initiative (ADNI)

We collected a total of 1059 subjects from ADNI 1 and GO/2 phases with brain MRI scans, preprocessed to be of size $48 \times 60 \times 48$.

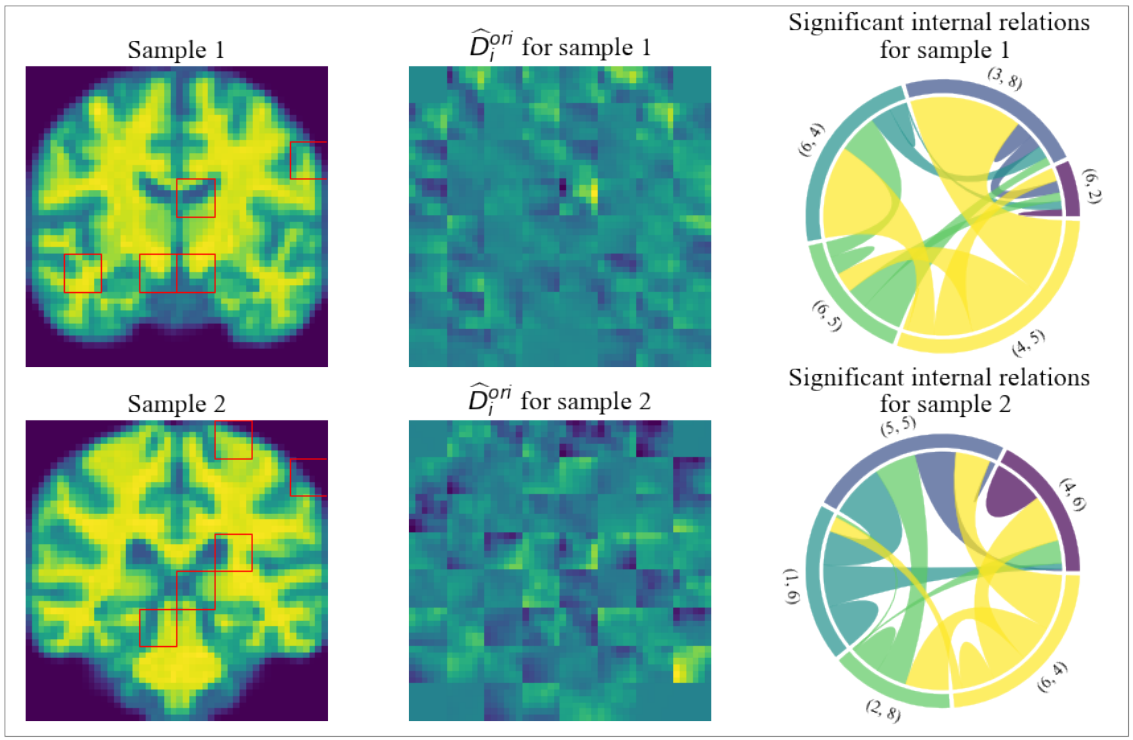
- $\boldsymbol{X}_i$: For each subject, we extracted 10 middle coronal slices, resulting in images of size 48×48.
- y_i: the Mini-Mental State Examination (MMSE) score.

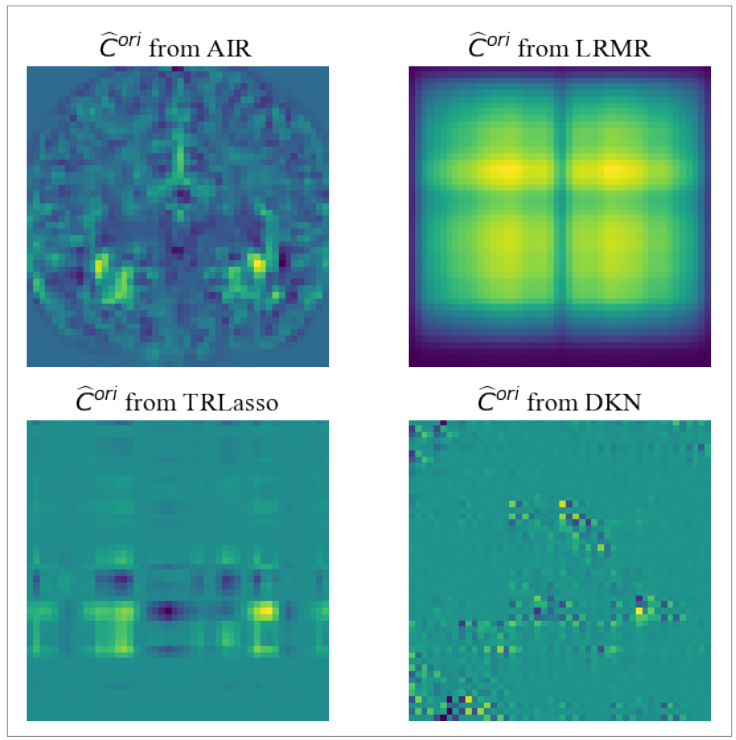


ADNI study

Table: Prediction errors of different methods.

AIR	LRMR	TRLasso	DKN	ViT
3.145 (0.019)	3.715 (0.008)	3.292 (0.023)	3.261 (0.017)	3.282 (0.025)





Thank you!

This paper

Yang, G., Cao, Y. and Feng, L., Attention boosted Individualized Regression. To appear in *NeurIPS 2024*.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng, L. and Yang, G. (2023). Deep Kronecker network. *Biometrika*, 111(2):707–714.

Shen, Z., Zhang, M., Zhao, H., Yi, S., and Li, H. (2021). Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):463–483.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552.