# Tensor Response Regression and Neuroimaging Analysis

Lexin Li

Division of Biostatistics
University of California, Berkeley

# Outline

- talk outline:
  - overview
  - motivating examples
  - tensor response regression: sparsity and low-rankness
  - tensor response regression: generalized sparsity and envelope approach
- collaborators:
  - William Jagust Lab @ UC Berkeley
  - Will Wei Sun @ U Miami; Xin Zhang @ FSU
- thanks:
  - NSF DMS-1310319, DMS-1613137
  - NIH 2R01AG034570-06A1 (PI: Jagust)

# Overview

- neuroimaging analysis is a super exciting area, because

# Overview

- neuroimaging analysis is a super exciting area, because
  - scientifically, a battery of important but challenging neurological disorders, e.g., Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), as well as normal aging

# Overview

- neuroimaging analysis is a super exciting area, because
  - scientifically, a battery of important but challenging neurological disorders, e.g., Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), as well as normal aging
  - statistically, an array of diverse statistical problems, constantly demanding new models, theory, algorithms

# Overview

- neuroimaging analysis is a super exciting area, because
  - scientifically, a battery of important but challenging neurological disorders, e.g., Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), as well as normal aging
  - statistically, an array of diverse statistical problems, constantly demanding new models, theory, algorithms
  - large public neuroimaging databases are becoming available

# Overview

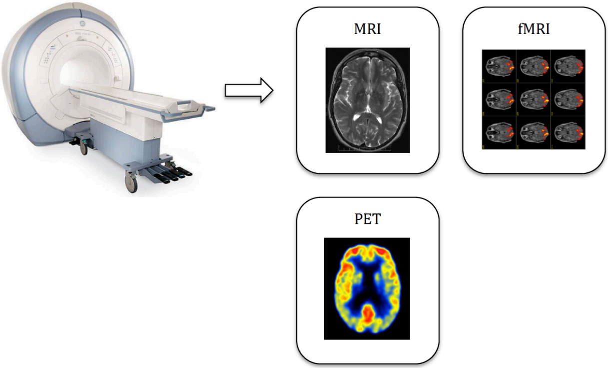
- neuroimaging analysis is a super exciting area, because
  - scientifically, a battery of important but challenging neurological disorders, e.g., Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), as well as normal aging
  - statistically, an array of diverse statistical problems, constantly demanding new models, theory, algorithms
  - large public neuroimaging databases are becoming available
  - not overly crowded, yet

# Overview

- neuroimaging analysis is a super exciting area, because
  - scientifically, a battery of important but challenging neurological disorders, e.g., Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), as well as normal aging
  - statistically, an array of diverse statistical problems, constantly demanding new models, theory, algorithms
  - large public neuroimaging databases are becoming available
  - not overly crowded, yet
  - even my in-laws got interested in what I am doing...

# Overview

- imaging modalities:
  - anatomical magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), positron emission tomography (PET), electroencephalography (EEG), . . .
  - a unifying form: **multidimensional array**, a.k.a. **tensor**



# Overview

- neuroimaging problems under investigation:
  - tensor regression
    - tensor predictor regression
    - **tensor response regression**
  - brain connectivity analysis
    - graphical model estimation (undirected, directed, Gaussian, non-Gaussian, static, dynamic)
    - graph inference
    - graph based regression (association) analysis
  - multimodal neuroimaging analysis
    - integrative classification
    - correlated region identification and inference
  - more topics
    - longitudinal imaging analysis
    - imaging genetics
    - imaging causal inference

# Overview

- neuroimaging problems under investigation:
  - tensor regression
    - tensor predictor regression
    - **tensor response regression**
  - brain connectivity analysis
    - graphical model estimation (undirected, directed, Gaussian, non-Gaussian, static, dynamic)
    - graph inference
    - graph based regression (association) analysis
  - multimodal neuroimaging analysis
    - integrative classification
    - correlated region identification and inference
  - more topics
    - longitudinal imaging analysis
    - imaging genetics
    - imaging causal inference
- **pick up another new topic here?**

# Motivation

- attention deficit hyperactivity disorder (ADHD) study:
  - one of the most commonly diagnosed child-onset neurodevelopmental disorders, with an estimated childhood prevalence of 5 – 10% worldwide
  - 776 subjects: 285 combined ADHD subjects and 491 normal controls
  - anatomical MRI images were acquired and **preprocessed**
  - MRI is in the form of **3D array**, $256 \times 198 \times 256$

# Motivation

- attention deficit hyperactivity disorder (ADHD) study:
  - one of the most commonly diagnosed child-onset neurodevelopmental disorders, with an estimated childhood prevalence of 5 – 10% worldwide
  - 776 subjects: 285 combined ADHD subjects and 491 normal controls
  - anatomical MRI images were acquired and **preprocessed**
  - MRI is in the form of **3D array**, $256 \times 198 \times 256$
- autism spectrum disorder (ASD) study:
  - an increasingly prevalent neurodevelopmental disorder; 1 in 68 american children according to CDC in 2015
  - 795 subjects: 362 ASD subjects and 433 normal controls
  - functional MRI images were acquired and **preprocessed** into 2 forms
  - fractional amplitude of low-frequency fluctuations (fALFF), which characterizes the intensity of spontaneous brain activities, and is in the form of **3D array**, $91 \times 109 \times 91$
  - partial correlation between brain regions of interest, which describes the conditional dependency and synchronization of brain systems, and is in the form of **2D symmetric matrix**, $116 \times 116$

# Motivation

- scientific question of interest:
  - understand the change of the tensor image or brain connectivity pattern as the predictors such as disease status varies, **after** adjusting for the demographical and other variables
  - identify brain regions exhibiting different patterns across subject groups — **"differentially expressed regions"**

# Motivation

- scientific question of interest:
  - understand the change of the tensor image or brain connectivity pattern as the predictors such as disease status varies, **after** adjusting for the demographical and other variables
  - identify brain regions exhibiting different patterns across subject groups — **"differentially expressed regions"**
- statistical formulation: **tensor response regression**
  - predictors: binary diagnostic status, age, gender, ...
  - response: **3D MRI, 3D fALFF, 2D symmetric connectivity matrix**
  - challenges: extremely high dimensionality and small sample size; complex data structure

# Motivation

- scientific question of interest:
  - understand the change of the tensor image or brain connectivity pattern as the predictors such as disease status varies, **after** adjusting for the demographical and other variables
  - identify brain regions exhibiting different patterns across subject groups — **"differentially expressed regions"**
- statistical formulation: **tensor response regression**
  - predictors: binary diagnostic status, age, gender, ...
  - response: **3D MRI, 3D fALFF, 2D symmetric connectivity matrix**
  - challenges: extremely high dimensionality and small sample size; complex data structure
  - solution I: **generalized sparsity and envelope approach**
  - solution II: **sparsity and low-rankness**

# Generalized sparsity and envelope

# Model

- model:

$$\boldsymbol{Y} = \boldsymbol{B} \times_{(D+1)} \boldsymbol{X} + \boldsymbol{\varepsilon}$$

  - $\boldsymbol{Y} \in \mathbb{R}^{r_1 \times \cdots \times r_D}$ = $D$th-order array-valued response; e.g., MRI scan
  - $\boldsymbol{X} \in \mathbb{R}^{p}$ = group indicator, plus additional covariates like age, gender
  - $\boldsymbol{B} \in \mathbb{R}^{r_1 \times \cdots \times r_D \times p}$ = $(D+1)$th-order **coefficient tensor** that captures the interrelation between $\boldsymbol{Y}$ and $\boldsymbol{X}$, and is our **parameter of interest**
  - $\times_{(m+1)}$ is the $(m+1)$-mode product of the tensor $\boldsymbol{B}$ and vector $\boldsymbol{X}$
  - $\boldsymbol{\varepsilon} \in \mathbb{R}^{r_1 \times \cdots r_D}$ = $m$th-order error tensor independent of $\boldsymbol{X}$
  - $\mathrm{vec}(\boldsymbol{\varepsilon}) \sim \mathrm{Normal}(0, \boldsymbol{\Sigma})$, where the covariance has a **separable Kronecker covariance structure** such that

$$\mathrm{cov}\{\mathrm{vec}(\boldsymbol{\varepsilon})\} = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_D \otimes \cdots \otimes \boldsymbol{\Sigma}_1$$

normality is not essential

# Key idea

- **assumption:** there exist subspaces $\mathcal{S}_d \subseteq \mathbb{R}^{r_d}$, $d = 1, \ldots, D$, st

$$\boldsymbol{Y} \times_d \boldsymbol{Q}_d | \boldsymbol{X} \sim \boldsymbol{Y} \times_d \boldsymbol{Q}_d, \quad \boldsymbol{Y} \times_d \boldsymbol{Q}_d \perp\!\!\!\perp \boldsymbol{Y} \times_d \boldsymbol{P}_d | \boldsymbol{X}$$

  - $\boldsymbol{P}_d \in \mathbb{R}^{r_d \times r_d}$ is the projection matrix onto $\mathcal{S}_d$, $\boldsymbol{Q}_d = \boldsymbol{I}_{r_d} - \boldsymbol{P}_d$ is the projection onto the complement space $\mathcal{S}_d^{\perp}$
  - $\times_d$ is the $d$-mode product

# Key idea

- **assumption:** there exist subspaces $\mathcal{S}_d \subseteq \mathbb{R}^{r_d}$, $d = 1, \ldots, D$, st

$$\boldsymbol{Y} \times_d \boldsymbol{Q}_d | \boldsymbol{X} \sim \boldsymbol{Y} \times_d \boldsymbol{Q}_d, \quad \boldsymbol{Y} \times_d \boldsymbol{Q}_d \perp\!\!\!\perp \boldsymbol{Y} \times_d \boldsymbol{P}_d | \boldsymbol{X}$$

  - $\boldsymbol{P}_d \in \mathbb{R}^{r_d \times r_d}$ is the projection matrix onto $\mathcal{S}_d$, $\boldsymbol{Q}_d = \boldsymbol{I}_{r_d} - \boldsymbol{P}_d$ is the projection onto the complement space $\mathcal{S}_d^{\perp}$
  - $\times_d$ is the $d$-mode product
  - in plain English: **some parts of $\boldsymbol{Y}$ are irrelevant**

# Key idea

- **assumption:** there exist subspaces $\mathcal{S}_d \subseteq \mathbb{R}^{r_d}$, $d = 1, \ldots, D$, st

$$\boldsymbol{Y} \times_d \boldsymbol{Q}_d | \boldsymbol{X} \sim \boldsymbol{Y} \times_d \boldsymbol{Q}_d, \quad \boldsymbol{Y} \times_d \boldsymbol{Q}_d \perp\!\!\!\perp \boldsymbol{Y} \times_d \boldsymbol{P}_d | \boldsymbol{X}$$

  - $\boldsymbol{P}_d \in \mathbb{R}^{r_d \times r_d}$ is the projection matrix onto $\mathcal{S}_d$, $\boldsymbol{Q}_d = \boldsymbol{I}_{r_d} - \boldsymbol{P}_d$ is the projection onto the complement space $\mathcal{S}_d^{\perp}$
  - $\times_d$ is the $d$-mode product

  - in plain English: **some parts of $\boldsymbol{Y}$ are irrelevant**
  - $\boldsymbol{Y} \times_d \boldsymbol{Q}_d$ is the irrelevant information to the regression, while $\boldsymbol{Y} \times_d \boldsymbol{P}_d$ contains all the relevant information

# Key idea

- **assumption:** there exist subspaces $\mathcal{S}_d \subseteq \mathbb{R}^{r_d}$, $d = 1, \ldots, D$, st

$$\boldsymbol{Y} \times_d \boldsymbol{Q}_d | \boldsymbol{X} \sim \boldsymbol{Y} \times_d \boldsymbol{Q}_d, \quad \boldsymbol{Y} \times_d \boldsymbol{Q}_d \perp\!\!\!\perp \boldsymbol{Y} \times_d \boldsymbol{P}_d | \boldsymbol{X}$$

  - $\boldsymbol{P}_d \in \mathbb{R}^{r_d \times r_d}$ is the projection matrix onto $\mathcal{S}_d$, $\boldsymbol{Q}_d = \boldsymbol{I}_{r_d} - \boldsymbol{P}_d$ is the projection onto the complement space $\mathcal{S}_d^{\perp}$
  - $\times_d$ is the $d$-mode product

  - in plain English: **some parts of $\boldsymbol{Y}$ are irrelevant**
  - $\boldsymbol{Y} \times_d \boldsymbol{Q}_d$ is the irrelevant information to the regression, while $\boldsymbol{Y} \times_d \boldsymbol{P}_d$ contains all the relevant information
  - sound familiar?

# Key idea

- **assumption:** there exist subspaces $\mathcal{S}_d \subseteq \mathbb{R}^{r_d}$, $d = 1, \ldots, D$, st

$$\boldsymbol{Y} \times_d \boldsymbol{Q}_d | \boldsymbol{X} \sim \boldsymbol{Y} \times_d \boldsymbol{Q}_d, \quad \boldsymbol{Y} \times_d \boldsymbol{Q}_d \perp\!\!\!\perp \boldsymbol{Y} \times_d \boldsymbol{P}_d | \boldsymbol{X}$$

  - $\boldsymbol{P}_d \in \mathbb{R}^{r_d \times r_d}$ is the projection matrix onto $\mathcal{S}_d$, $\boldsymbol{Q}_d = \boldsymbol{I}_{r_d} - \boldsymbol{P}_d$ is the projection onto the complement space $\mathcal{S}_d^{\perp}$
  - $\times_d$ is the $d$-mode product

  - in plain English: **some parts of $\boldsymbol{Y}$ are irrelevant**
  - $\boldsymbol{Y} \times_d \boldsymbol{Q}_d$ is the irrelevant information to the regression, while $\boldsymbol{Y} \times_d \boldsymbol{P}_d$ contains all the relevant information
  - sound familiar?

  - **sparsity principle in variable selection**: a subset of **individual** predictors are irrelevant to the regression

# Key idea

- **assumption:** there exist subspaces $\mathcal{S}_d \subseteq \mathbb{R}^{r_d}$, $d = 1, \ldots, D$, st

$$\boldsymbol{Y} \times_d \boldsymbol{Q}_d | \boldsymbol{X} \sim \boldsymbol{Y} \times_d \boldsymbol{Q}_d, \quad \boldsymbol{Y} \times_d \boldsymbol{Q}_d \perp\!\!\!\perp \boldsymbol{Y} \times_d \boldsymbol{P}_d | \boldsymbol{X}$$

  - $\boldsymbol{P}_d \in \mathbb{R}^{r_d \times r_d}$ is the projection matrix onto $\mathcal{S}_d$, $\boldsymbol{Q}_d = \boldsymbol{I}_{r_d} - \boldsymbol{P}_d$ is the projection onto the complement space $\mathcal{S}_d^{\perp}$
  - $\times_d$ is the $d$-mode product

  - in plain English: **some parts of $\boldsymbol{Y}$ are irrelevant**
  - $\boldsymbol{Y} \times_d \boldsymbol{Q}_d$ is the irrelevant information to the regression, while $\boldsymbol{Y} \times_d \boldsymbol{P}_d$ contains all the relevant information
  - sound familiar?

  - **sparsity principle in variable selection**: a subset of **individual** predictors are irrelevant to the regression

  - **generalized sparsity principle**: shares the same spirit that only part of information is deemed useful for regressions and the rest irrelevant, but is also more flexible in that it permits **linear combination** of the variables to be irrelevant

# Tensor envelope

- why helpful?
  - **dimension reduction** on $\boldsymbol{Y}$: let $\boldsymbol{\Gamma}_d \in \mathbb{R}^{r_d \times u_d}$ be a basis for $\mathcal{S}_d$, and $\boldsymbol{\Gamma}_{0d} \in \mathbb{R}^{r_d \times (r_d - u_d)}$ the complement basis

$$\boldsymbol{Y} \in \mathbb{R}^{r_1 \times \cdots \times r_D} \Rightarrow [\![ \boldsymbol{Y}; \boldsymbol{\Gamma}_1^{\mathsf{T}}, \ldots, \boldsymbol{\Gamma}_D^{\mathsf{T}} ]\!] \in \mathbb{R}^{u_1 \times \cdots \times u_D}, \quad u_d \leq r_d$$

  - number of free parameters:
    - before: $p \prod_{d=1}^{D} r_d + \sum_{d=1}^{D} r_d(r_d + 1)/2$
    - after: $p \prod_{d=1}^{D} u_d + \sum_{d=1}^{D} \{u_d(r_d - u_d) + u_d(u_d + 1)/2 + (r_d - u_d)(r_d - u_d + 1)/2\}$
    - **difference:** $p\left\{\prod_{d=1}^{D} r_d - \prod_{d=1}^{D} u_d\right\}$
  - more **efficient** than OLS
- **tensor response envelope:**

$$\mathcal{T}_{\boldsymbol{\Sigma}}(\boldsymbol{B}) \equiv \mathcal{E}_{\boldsymbol{\Sigma}_D}\left(\boldsymbol{B}_{(D)}\right) \otimes \ldots \otimes \mathcal{E}_{\boldsymbol{\Sigma}_1}\left(\boldsymbol{B}_{(1)}\right)$$

# Tensor envelope

- why helpful?
  - **dimension reduction** on $\boldsymbol{Y}$: let $\boldsymbol{\Gamma}_d \in \mathbb{R}^{r_d \times u_d}$ be a basis for $\mathcal{S}_d$, and $\boldsymbol{\Gamma}_{0d} \in \mathbb{R}^{r_d \times (r_d - u_d)}$ the complement basis

$$\boldsymbol{Y} \in \mathbb{R}^{r_1 \times \cdots \times r_D} \Rightarrow [\![\boldsymbol{Y}; \boldsymbol{\Gamma}_1^{\mathsf{T}}, \ldots, \boldsymbol{\Gamma}_D^{\mathsf{T}}]\!] \in \mathbb{R}^{u_1 \times \cdots \times u_D}, \quad u_d \leq r_d$$

  - number of free parameters: — e.g., $r_1 = r_2 = r3 = 64, u_1 = u_2 = u_3 = 10, p = 3$
    - before: $p \prod_{d=1}^{D} r_d + \sum_{d=1}^{D} r_d(r_d+1)/2$ — $792,672$
    - after: $p \prod_{d=1}^{D} u_d + \sum_{d=1}^{D} \{u_d(r_d - u_d) + u_d(u_d+1)/2 + (r_d - u_d)(r_d - u_d + 1)/2\}$ — $9,240$
    - **difference:** $p\left\{\prod_{d=1}^{D} r_d - \prod_{d=1}^{D} u_d\right\}$ — save $783,432$ parameters
  - more **efficient** than OLS

- **tensor response envelope:**

$$\mathcal{T}_{\boldsymbol{\Sigma}}(\boldsymbol{B}) \equiv \mathcal{E}_{\boldsymbol{\Sigma}_D}\left(\boldsymbol{B}_{(D)}\right) \otimes \ldots \otimes \mathcal{E}_{\boldsymbol{\Sigma}_1}\left(\boldsymbol{B}_{(1)}\right)$$

# Estimation

- estimation:
  - maximum likelihood estimation: iterative optimization algorithm
  - approximation: one-step optimization algorithm

---

**for** $s = 0, \ldots, u_d - 1$ **do**
set $\boldsymbol{G}_d^s = \boldsymbol{0}$ if $s = 0$ and $\boldsymbol{G}_d^s = (\boldsymbol{g}_{d1}, \ldots, \boldsymbol{g}_{ds})$ otherwise
construct $\boldsymbol{G}_{0d}^s$ as an orthogonal basis complement to $\boldsymbol{G}_d^s$ in $\mathbb{R}^{\boldsymbol{r_d}}$
solve the objective function over $\boldsymbol{w} \in \mathbb{R}^{r-\boldsymbol{s}}$ subject to $\boldsymbol{w}^\mathsf{T}\boldsymbol{w} = 1$:

$$\begin{aligned} \boldsymbol{w}_{d+1} &= \arg\min_{\boldsymbol{w}} \log\left\{\boldsymbol{w}^\mathsf{T}\left((\boldsymbol{G}_{0d}^s)^\mathsf{T}\boldsymbol{\Sigma}_d^{(0)}\boldsymbol{G}_{0d}^s\right)\boldsymbol{w}\right\} + \\ & \qquad \log\left\{\boldsymbol{w}^\mathsf{T}\left((\boldsymbol{G}_{0d}^s)^\mathsf{T}\boldsymbol{N}_d^{(0)}\boldsymbol{G}_{0d}^s\right)^{-1}\boldsymbol{w}\right\} \end{aligned}$$

set $\boldsymbol{g}_{d+1} = \boldsymbol{G}_{0d}^s \boldsymbol{w}_{d+1} \in \mathbb{R}^{\boldsymbol{r_d}}$ and normalize to unit length
**end for**

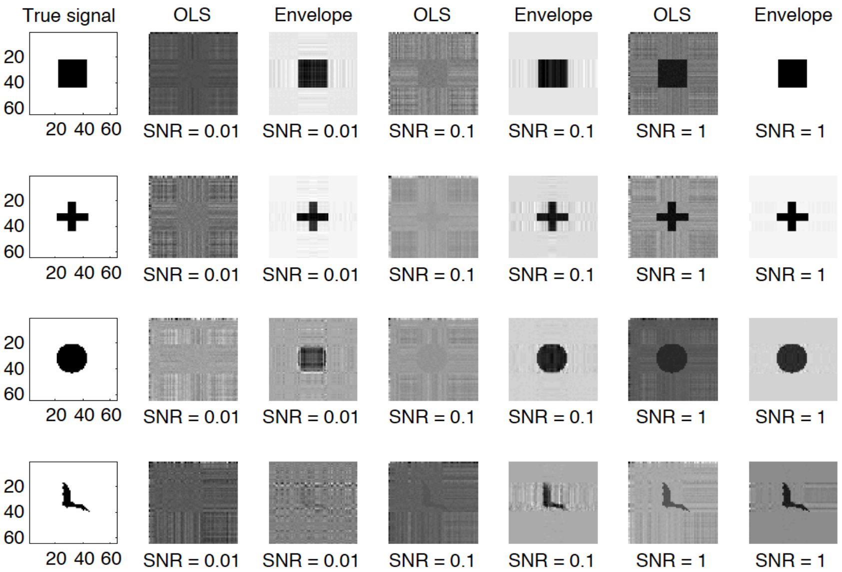
---

- envelope dimension estimation: a variant of BIC

# Theory

- asymptotics:
  assuming $\mathrm{vec}(\boldsymbol{\varepsilon}_i)$, $i = 1, \ldots, n$, are i.i.d. with finite fourth moments
  - **consistency:** $\widehat{\boldsymbol{B}}^{it}_{\mathrm{ENV}}$ and $\widehat{\boldsymbol{B}}^{os}_{\mathrm{ENV}}$ both converge at rate-$\sqrt{n}$ to the true tensor coefficient $\boldsymbol{B}_{\mathrm{TRUE}}$
  - **asymptotic normality:** $\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{B}}^{it}_{\mathrm{ENV}} - \boldsymbol{B}_{\mathrm{TRUE}}) \to N(0, \boldsymbol{U}_{\mathrm{ENV}})$
  - **efficiency:** $\widehat{\boldsymbol{B}}_{\mathrm{OLS}}$ satisfies that $\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{B}}_{\mathrm{OLS}} - \boldsymbol{B}_{\mathrm{TRUE}}) \to N(0, \boldsymbol{U}_{\mathrm{OLS}})$, and $\boldsymbol{U}_{\mathrm{ENV}} \le \boldsymbol{U}_{\mathrm{OLS}}$

# Simulation


True signal
OLS
Envelope
OLS
Envelope
OLS
Envelope
20
40
60
20 40 60
SNR = 0.01
SNR = 0.01
SNR = 0.1
SNR = 0.1
SNR = 1
SNR = 1
20
40
60
20 40 60
SNR = 0.01
SNR = 0.01
SNR = 0.1
SNR = 0.1
SNR = 1
SNR = 1
20
40
60
20 40 60
SNR = 0.01
SNR = 0.01
SNR = 0.1
SNR = 0.1
SNR = 1
SNR = 1
20
40
60
20 40 60
SNR = 0.01
SNR = 0.01
SNR = 0.1
SNR = 0.1
SNR = 1
SNR = 1

# ADHD analysis

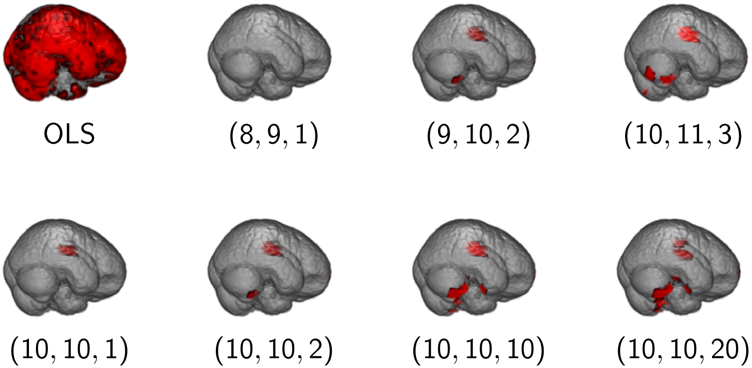



Figure: The $p$-value map, thresholded at 0.05, using the OLS and envelope method with varying working dimensions. BIC selected $(9, 10, 2)$.

- findings: superior temporal gyrus, and pyramid and uvula in cerebellum

# Sparsity and low-rankness

# Model

- model:

$$\boldsymbol{Y} = \boldsymbol{B} \times_{(D+1)} \boldsymbol{X} + \boldsymbol{\varepsilon}$$

  - $\boldsymbol{Y} \in \mathbb{R}^{r_1 \times \cdots \times r_D} = D$th-order array-valued response; can naturally handle **both a general tensor and a symmetric tensor**
  - $\boldsymbol{X} \in \mathbb{R}^{p} =$ group indicator, plus additional covariates like age, gender
  - $\boldsymbol{B} \in \mathbb{R}^{r_1 \times \cdots \times r_D \times p} = (D+1)$th-order **coefficient tensor** that captures the interrelation between $\boldsymbol{Y}$ and $\boldsymbol{X}$, and is our **parameter of interest**
  - $\times_{(m+1)}$ is the $(m+1)$-mode product of the tensor $\boldsymbol{B}$ and vector $\boldsymbol{X}$
  - $\boldsymbol{\varepsilon} \in \mathbb{R}^{r_1 \times \cdots r_D} = m$th-order error tensor independent of $\boldsymbol{X}$ (no Kronecker product structure imposed)

# Key idea

- low-rank structure:

$$\boldsymbol{B} = \sum_{k=1}^{K} w_k \boldsymbol{\beta}_{k,1} \circ \cdots \circ \boldsymbol{\beta}_{k,D} \circ \boldsymbol{\beta}_{k,D+1}$$

where $w_k \in \mathbb{R}, \boldsymbol{\beta}_{k,d} \in \mathbb{R}^d, \|\boldsymbol{\beta}_{k,d}\|_2 = 1$, and $\boldsymbol{\beta}_{k,D+1} \in \mathbb{R}^p$ encodes the predictor effect

# Key idea

- low-rank structure:

$$\boldsymbol{B} = \sum_{k=1}^{K} w_k \boldsymbol{\beta}_{k,1} \circ \cdots \circ \boldsymbol{\beta}_{k,D} \circ \boldsymbol{\beta}_{k,D+1}$$

where $w_k \in \mathbb{R}, \boldsymbol{\beta}_{k,d} \in \mathbb{R}^d, \|\boldsymbol{\beta}_{k,d}\|_2 = 1$, and $\boldsymbol{\beta}_{k,D+1} \in \mathbb{R}^p$ encodes the predictor effect

- for $D = 2$, $K = 1$, $\boldsymbol{B} = [\![\boldsymbol{B}_1, \boldsymbol{B}_2]\!], \boldsymbol{B}_1 = \boldsymbol{\beta}_1, \boldsymbol{B}_2 = \boldsymbol{\beta}_2$,

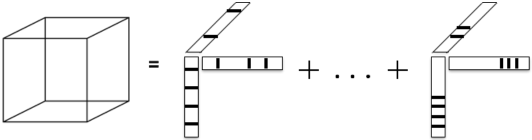
$$\boldsymbol{B} = w_1 \, \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2$$

- for $D = 2$, $K = 2$, $\boldsymbol{B} = [\![\boldsymbol{B}_1, \boldsymbol{B}_2]\!], \boldsymbol{B}_1 = [\boldsymbol{\beta}_1^{(1)}, \boldsymbol{\beta}_1^{(2)}], \boldsymbol{B}_2 = [\boldsymbol{\beta}_2^{(1)}, \boldsymbol{\beta}_2^{(2)}]$,

$$\boldsymbol{B} = w_1 \, \boldsymbol{\beta}_1^{(1)} \circ \boldsymbol{\beta}_2^{(1)} + w_2 \, \boldsymbol{\beta}_1^{(2)} \circ \boldsymbol{\beta}_2^{(2)}$$

# Key idea

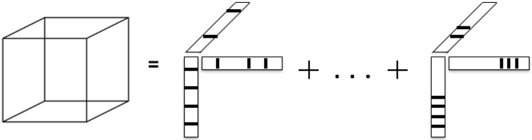
- **low-rank structure**:



- number of free parameters:
  - before: $p \prod_{d=1}^{D} r_d$
  - after: $K(p + \sum_{d=1}^{D} r_d)$
  - **difference:** $p \prod_{d=1}^{D} r_d - K(p + \sum_{d=1}^{D} r_d)$

# Key idea

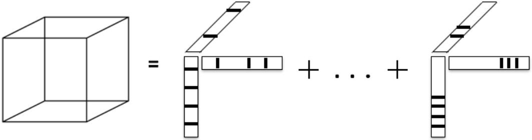
- **low-rank structure**:




- number of free parameters: — e.g., $r_1 = r_2 = r3 = 64, K = 3, p = 3$
  - before: $p \prod_{d=1}^{D} r_d$ — $786,432$
  - after: $K(p + \sum_{d=1}^{D} r_d)$ — $585$
  - **difference:** $p \prod_{d=1}^{D} r_d - K(p + \sum_{d=1}^{D} r_d)$ — save $785,847$ parameters

# Key idea

- low-rank structure:



  - number of free parameters: — e.g., $r_1 = r_2 = r3 = 64, K = 3, p = 3$
    - before: $p\prod_{d=1}^{D} r_d$ — $786,432$
    - after: $K(p + \sum_{d=1}^{D} r_d)$ — $585$
    - **difference:** $p\prod_{d=1}^{D} r_d - K(p + \sum_{d=1}^{D} r_d)$ — save $785,847$ parameters

- entry-wise sparsity:

$$\|\boldsymbol{\beta}_{k,d}\|_0 \leq s_d, \quad 1 \leq d \leq D$$

  - facilitate the interpretation
  - no sparsity constraint on $\boldsymbol{\beta}_{k,D+1}$

# Estimation

- objective function:

$$\min_{w_k,\beta_{k,1},\ldots,\beta_{k,D+1}} \frac{1}{n}\sum_{i=1}^{n}\left\| \boldsymbol{Y}_i - \sum_{k=1}^{K} w_k(\boldsymbol{\beta}_{k,D+1}^{\mathsf{T}}\boldsymbol{X}_i)\boldsymbol{\beta}_{k,1}\circ\cdots\circ\boldsymbol{\beta}_{k,D}\right\|_F^2,$$

$$\text{subject to } \|\boldsymbol{\beta}_{k,d}\|_2 = 1, \|\boldsymbol{\beta}_{k,d}\|_0 \le s_d$$

- alternating updating algorithm: thanks to the bi-convexity
  - update $\{w_k, \boldsymbol{\beta}_{k,1}, \ldots, \boldsymbol{\beta}_{k,D}\}$: solved by a hard-thresholding sparse tensor decomposition method
  - update $\boldsymbol{\beta}_{k,D+1}$: closed form solution
- symmetry can be obtained by setting $\boldsymbol{\beta}_{k,1} = \ldots \boldsymbol{\beta}_{k,D} = \boldsymbol{\beta}_k$
- rank estimation: a variant of BIC

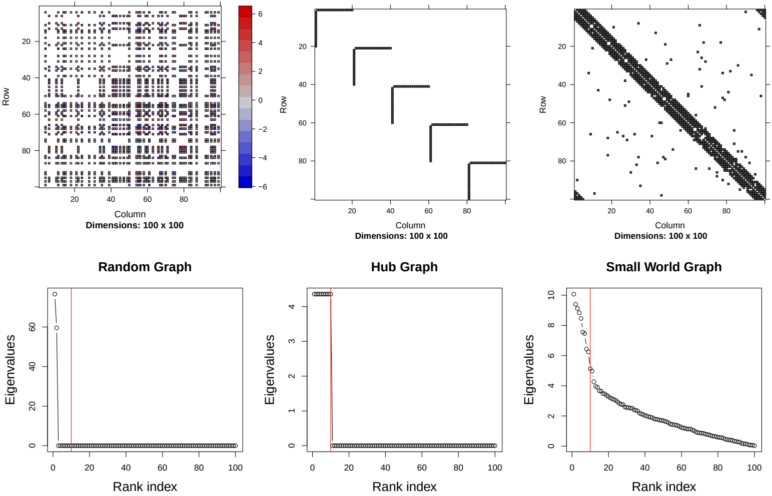
# Theory

- non-asymptotic error bound:

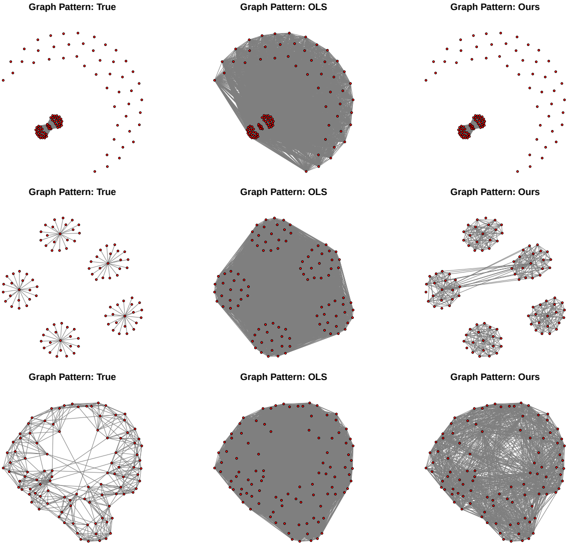
$$D\left(\widehat{\boldsymbol{\Theta}}^{(t)}, \boldsymbol{\Theta}^*\right) \leq \underbrace{\kappa^t \epsilon}_{\text{computational error}} + \underbrace{\frac{1}{1-\kappa} \max\left\{ C \cdot \eta\left( \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i, s\right), \frac{\tilde{C}}{\sqrt{n}} \right\}}_{\text{statistical error}},$$

  - for the actual minimizer obtained from our optimization algorithm, instead of a global minimizer that is not guaranteed to obtain
  - interplay between the computational efficiency and the statistical rate of convergence, i.e., the computational error decays geometrically with the iteration number $t$, whereas the statistical error remains the same when $t$ grows
  - choose the maximal number of iterations $T$, such that the computational error is dominated by the statistical error
  - the result holds for any distribution of the error tensor; further results when $\boldsymbol{\varepsilon}_i$ is a Gaussian tensor, or a symmetric matrix

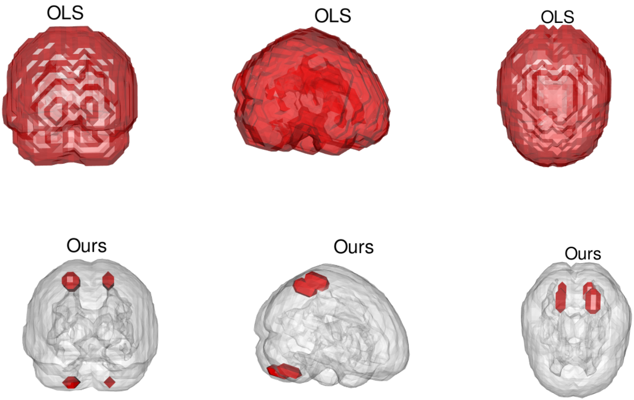
# Simulation


Row
Column
Dimensions: 100 x 100
Row
Column
Dimensions: 100 x 100
Row
Column
Dimensions: 100 x 100
Random Graph
Hub Graph
Small World Graph
Eigenvalues
Rank index
Eigenvalues
Rank index
Eigenvalues
Rank index

# Simulation

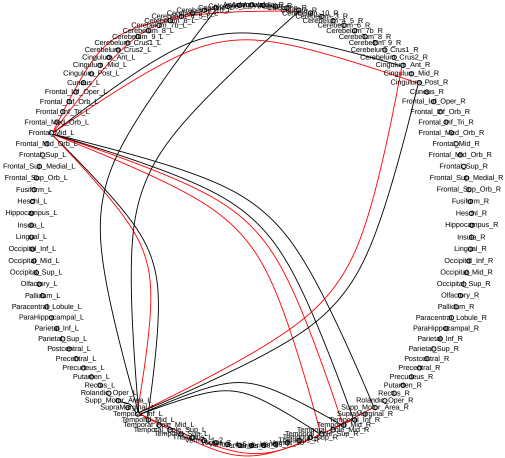

Graph Pattern: True
Graph Pattern: OLS
Graph Pattern: Ours
Graph Pattern: True
Graph Pattern: OLS
Graph Pattern: Ours
Graph Pattern: True
Graph Pattern: OLS
Graph Pattern: Ours

# ASD analysis




- findings: cerebellum, superior parietal lobule, precuneus

# ASD analysis




- findings: left middle frontal gyrus, temporal lobe

# Thank You!