Reduced-Rank Tensor-on-Tensor Regression and Tensor-Variate ANOVA

Carlos Llosa Vite
Join work with Ranjan Maitra

Deparment of Statistics
Iowa State University

Outline

1. Motivation: Suicide Risk Assessment
2. Methodology
   - A Primer on Tensors
   - Tensor-on-Tensor Regression
   - Maximum Likelihood Estimation and Sampling Distributions
3. Performance and Data applications
   - Andean Camelids Simulation
   - Labeled Faces in the Wild
   - Assessing Suicide Risk

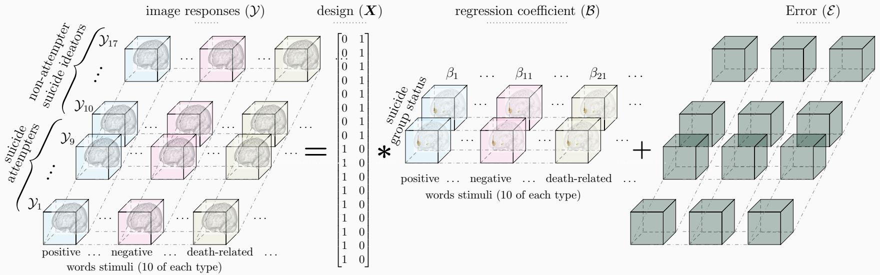
Outline

1. Motivation: Suicide Risk Assessment
2. Methodology
   - A Primer on Tensors
   - Tensor-on-Tensor Regression
   - Maximum Likelihood Estimation and Sampling Distributions
3. Performance and Data applications
   - Andean Camelids Simulation
   - Labeled Faces in the Wild
   - Assessing Suicide Risk

Suicide Risk Assessment

- 2nd leading cause of death for people under 35 in the US.
- Accounted for 2/3 of all US homicides in 2017.
- Assessment is challenging: 78% of patients who die of suicide deny ideation in their last communication with a professional.
- Just et al. (2017) provided fMRI data from 17 young suicide ideators, 9 have attempted suicide.
- Each subject was exposed to 10 positive, 10 negative, and 10 death-related word stimuli.
- Stimuli extracted from time series using the general linear model. βs used as the response in the tensor-linear model.

Suicide Risk Assessment



$$\mathcal{Y} = \boldsymbol{X}\mathcal{B} + \mathcal{E}$$

- The ith subject has $\mathcal{Y}_i \in \mathbb{R}^{3\times10\times43\times56\times20}$ as a tensor-response.
- This is MANOVA with vec($\mathcal{Y}_i$) as response.
- $\mathcal{B}$ has 96,320 unconstrained parameters.
- Error covariance has 1,043,724,242,400 unconstrained parameters.
- Goal: Identifying brain regions associated with a significant interaction between suicide group status and type of stimuli.

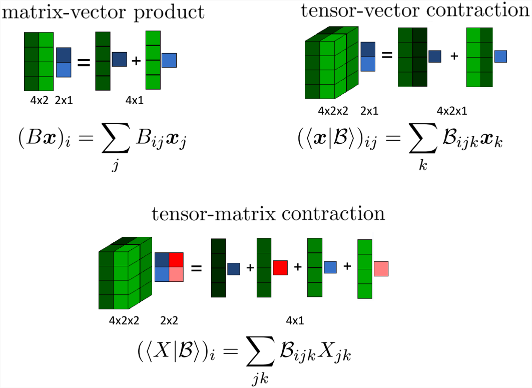
Outline

Tensor Partial Contraction

Tensor-on-Tensor Regression:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i,$$

$$\mathcal{X}_i \in \mathbb{R}^{h_1 \times \ldots \times h_l}, \quad \mathcal{Y}_i \in \mathbb{R}^{m_1 \times \ldots \times m_p}, \quad \mathcal{B} \in \mathbb{R}^{h_1 \times \ldots \times h_l \times m_1 \times \ldots \times m_p}.$$

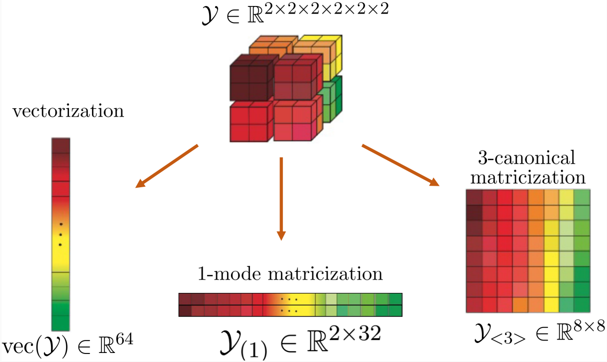


Tensor Reshapings

Tensor-on-Tensor Regression:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i,$$

$$\mathcal{X}_i \in \mathbb{R}^{h_1 \times \ldots \times h_l}, \quad \mathcal{Y}_i \in \mathbb{R}^{m_1 \times \ldots \times m_p}, \quad \mathcal{B} \in \mathbb{R}^{h_1 \times \ldots \times h_l \times m_1 \times \ldots \times m_p}.$$



The Tensor-Variate Normal (TVN) Distribution

Tensor-on-Tensor Regression:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{m_1, m_2 \ldots, m_p}(0, \sigma^2 \Sigma_1, \Sigma_2, \ldots, \Sigma_p),$$

$$\mathcal{E} \sim \mathcal{N}_{m_1, m_2, \ldots, m_p} (\mathcal{M}, \Sigma_1, \Sigma_2, \ldots, \Sigma_p)$$

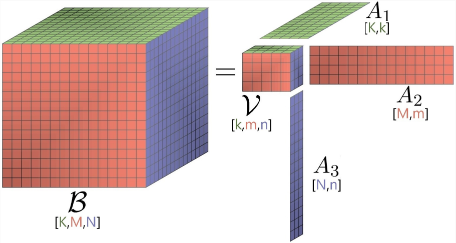
$$\iff \text{vec}(\mathcal{E}) \sim \mathcal{N}_{\prod_{k=1}^{p} m_k} \left(\text{vec}(\mathcal{M}), \bigotimes_{k=p}^{1} \Sigma_k \right)$$

- Introduced by Hoff (2011); Akdemir and Gupta (2011); Ohlson et al. (2013); Manceur and Dutilleul (2013).
- Unconstrained Σ is of size $(\prod_{k=1}^{p} m_k)(\prod_{k=1}^{p} m_k + 1)/2$.
- Constrained Σ is of size $\sum_{k=1}^{p} [m_k(m_k + 1)/2]$.
- $\Sigma_k(1, 1) = 1$ to deal with identifiability.

The Tucker Simplifying Format

Tensor-on-Tensor Regression:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{B} \in \mathbb{R}^{h_1 \times h_2 \times \ldots \times h_l \times m_1 \times m_2 \times \ldots \times m_p}$$



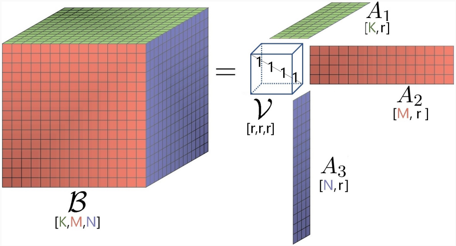
$$\mathcal{B} = [\![\mathcal{V}; A_1, A_2, A_3]\!]$$

- See Tucker (1966); Kolda and Bader (2009).
- Unconstrained $\mathcal{B} \in \mathbb{R}^{15 \times 15 \times 15}$ has 3,375 parameters.
- Constrained to a Tucker format of rank (3,4,5) leads to 240.

The Canonical CP Simplifying format

Tensor-on-Tensor Regression:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{B} \in \mathbb{R}^{h_1 \times h_2 \times \ldots \times h_l \times m_1 \times m_2 \times \ldots \times m_p}$$



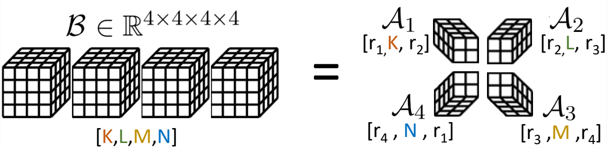
$$\mathcal{B} = [\![A_1, A_2, A_3]\!]$$

- See Hitchcock (1927); Kolda and Bader (2009).
- Unconstrained $\mathcal{B}$ has 3,375 parameters.
- Constrained to a CP format of rank 4 leads to only 180.

The Tensor Ring (TR) Simplifying Format

Tensor-on-Tensor Regression:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{B} \in \mathbb{R}^{h_1 \times h_2 \times \ldots \times h_l \times m_1 \times m_2 \times \ldots \times m_p}$$



$$\mathcal{B} = \text{tr}\left(\mathcal{A}_1 \times^1 \mathcal{A}_2 \times^1 \mathcal{A}_3 \times^1 \mathcal{A}_4\right)$$

- See Affleck et al. (1987); Zhao et al. (2016).
- Referred to matrix product state (MPS) in many-body physics.
- Unconstrained $\mathcal{B}$ has 256 parameters.
- Constrained to a TR format of rank (2,2,2,2) leads to only 64.

Review of Tensor-on-Tensor Regression

Tensor-on-Tensor Regression:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{m_1, m_2 \ldots, m_p}(0, \sigma^2 \Sigma_1, \Sigma_2, \ldots, \Sigma_p),$$

- Identity matrices for Σs and CP-formatted $\mathcal{B}$ in Lock (2017).
- OP-formatted $\mathcal{B}$ in Hoff (2014).

$$\mathcal{B}_{OP}(i, j, k, l) = M_1(i, j) M_2(k, l)$$

- Tucker format in Li and Zhang (2017):
    - Vector-variate $\mathcal{X}_i$.
    - Specific structure of $\Sigma_1, \ldots, \Sigma_p$.
    - Tucker format on last p sides of $\mathcal{B}$ only.

Our Contributions to Tensor-on-Tensor Regression

Tensor-on-Tensor Regression:

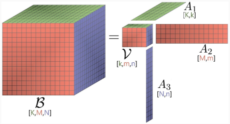
$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{m_1, m_2 \ldots, m_p}(0, \sigma^2 \Sigma_1, \Sigma_2, \ldots, \Sigma_p),$$

- We propose a unified framework for multiple formats:
    - We frame Hoff (2014) in the context of ToTR.
    - We extend Lock (2017) to allow for TVN error.
    - We extend Li and Zhang (2017) to full Tucker format, arbitrary Σs.
    - We propose the TR format.
- We don't make hard assumptions on $\Sigma_1, \ldots, \Sigma_p$.
- We study the distribution of $\widehat{\mathcal{B}}$ and perform inference.

Maximum Likelihood Estimation

ToTR with Tucker Format:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{m_1, m_2 \ldots, m_p}(0, \sigma^2 \Sigma_1, \Sigma_2, \ldots, \Sigma_p),$$
$$\mathcal{B} \;=\; [\![\mathcal{V}; L_1, L_2, \ldots, L_l, M_1, M_2, \ldots, M_p]\!],$$
$$M_k' \Sigma_k^{-1} M_k = I_{d_k}\;, \quad \Sigma_k(1,1) = 1.$$



- Loglikelihood where $\Sigma = \otimes_{k=p}^{1} \Sigma_k$:

$$\ell = -\frac{n}{2} \log |\sigma^2 \Sigma| - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[\text{vec}(\mathcal{Y}_i - \langle \mathcal{X}_i | \mathcal{B} \rangle)' \Sigma^{-1} \text{vec}(\mathcal{Y}_i - \langle \mathcal{X}_i | \mathcal{B} \rangle) \right]$$

- Block relaxation de Leeuw (1994): partitions the parameter and iteratively optimizes each while fixing the others.

Maximum Likelihood Estimation

ToTR with Tucker Format:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{m_1, m_2 \ldots, m_p}(0, \sigma^2 \Sigma_1, \Sigma_2, \ldots, \Sigma_p),$$
$$\mathcal{B} \;=\; [\![\mathcal{V}; L_1, L_2, \ldots, L_l, M_1, M_2, \ldots, M_p]\!].$$

Algorithm 1: ML Estimation of ToTR under Tucker Format

```
while convergence is not met do
    for k ∈ {1, 2, ..., l} do
        Estimate L_k given fixed values of L_{-k}, M, Σ
    end
    for k ∈ {1, 2, ..., p} do
        Estimate M_k given fixed values of L, M_{-k}, Σ
    end
    Estimate 𝒱 given fixed values of L, M, Σ for
     k ∈ {1, 2, ..., p} do
        Estimate (σ^2, Σ_k) given fixed values of L, M, Σ_{-k}
    end
end
```

Estimating L_1

ToTR with Tucker Format:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{m_1, m_2 \ldots, m_p}(0, \sigma^2 \Sigma_1, \Sigma_2, \ldots, \Sigma_p),$$
$$\mathcal{B} \;=\; [\![\mathcal{V}; L_1, L_2, \ldots, L_l, M_1, M_2, \ldots, M_p]\!],$$

- Express the model as multivariate multiple linear regression

$$\mathrm{vec}(\mathcal{Y}_i) = H_i \,\mathrm{vec}(L_1) + \boldsymbol{e}_i, \quad \boldsymbol{e}_i \overset{iid}{\sim} \mathcal{N}_m(0, \sigma^2 \Sigma),$$

  where the matrix H_i involves $(\mathcal{X}_i, \mathcal{V}, L_2, \ldots, L_l, M_1, \ldots, M_p)$.
- The MLE is the generalized least squares estimator.

Profiling $\mathcal{V}$ from the loglikelihood

ToTR with Tucker Format:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{m_1, m_2 \dots, m_p}(0, \sigma^2 \Sigma_1, \Sigma_2, \dots, \Sigma_p),$$
$$\mathcal{B} \;=\; [\![\mathcal{V}; L_1, L_2, \dots, L_l, M_1, M_2, \dots, M_p]\!],$$

- Rewrite the problem based on $M = \bigotimes_{k=p}^{1} M_k$ and $L = \bigotimes_{k=l}^{1} L_k$ as

$$\operatorname{vec} \mathcal{Y}_i = M \mathcal{V}'_{<l>} L' (\operatorname{vec} \mathcal{X}_i) + \boldsymbol{e}_i, \quad \boldsymbol{e}_i \overset{iid}{\sim} \mathcal{N}_m(0, \sigma^2 \Sigma).$$

- Recall $M' \Sigma^{-1} M$ is an identity matrix.

Block for M_1 after profiling $\mathcal{V}$

ToTR with Tucker Format:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{m_1, m_2 \dots, m_p}(0, \sigma^2 \Sigma_1, \Sigma_2, \dots, \Sigma_p),$$
$$\mathcal{B} \;=\; [\![\mathcal{V}; L_1, L_2, \dots, L_l, M_1, M_2, \dots, M_p]\!],\; M_k' \Sigma_k^{-1} M_k = I_{d_k}$$

- The profiled (on $\widehat{\mathcal{V}}$) loglikelihood in terms of M_1 is

$$\ell_p(M_1) = ||M_1' \Sigma_1^{-1} Q_1||_2^2$$

  for some Q_1 that involves $(\mathcal{X}, \Sigma_{-1}, M_{-1}, L)$.
- For $\Sigma_1^{-1/2} Q_1 = UDV'$, we have

$$\hat{M}_1 = \underset{M_1' \Sigma_1^{-1} M_1 = I_{d_1}}{\arg\max} \; \ell_p(M_1) = \Sigma_1^{1/2} U^*$$

  where U^* are the leading d_k columns of U.

Block for (Σ_1, σ^2)

ToTR with Tucker Format:

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{m_1, m_2 \ldots, m_p}(0, \sigma^2 \Sigma_1, \Sigma_2, \ldots, \Sigma_p),$$
$$\Sigma_k(1,1) = 1$$

- The loglikelihood in terms of Σ_1 is

$$\ell_1(\Sigma_1) = -\frac{nm_{-1}}{2} \log |\Sigma_1| - \frac{1}{2\sigma^2} \operatorname{tr}(\Sigma_1^{-1} S_1)$$

  for some matrix S_1 that involves $(\mathcal{B}, \Sigma_{-1})$, $m_{-1} = \prod_{k \neq 1} m_k$.
- Constrained optimization by Glanz and Carvalho (2018)

$$\hat{\Sigma}_1 = \underset{\Sigma_1(1,1)=1}{\arg\max} (\ell_1(\Sigma_1)) = ADJUST(nm_{-1}, \sigma^2, S_1).$$

- $\hat{\sigma}^2 = \dfrac{1}{nm} \operatorname{tr}(\widehat{\Sigma}_1^{-1} S_1)$ estimated alternatingly.

Asymptotic Sampling Distributions

Theorem 3.1:

Let $X = [(\text{vec}\,\mathcal{X}_1)\ldots(\text{vec}\,\mathcal{X}_n)]$, $\widehat{\mathcal{B}}_{TK} = [\![\widehat{\mathcal{V}}; \widehat{L}_1,\ldots,\widehat{L}_l, \widehat{M}_1,\ldots,\widehat{M}_p]\!]$, $\mathcal{B}_{TK} = [\![\mathcal{V}; L_1,\ldots,L_l, M_1,\ldots,M_p]\!]$. and $P_L = [L(L'L)^{-1}L']$. Then

$$\text{vec}(\hat{\mathcal{B}}_{TK}) \xrightarrow{d} \mathcal{N}\left(\text{vec}\left(\mathcal{B}_{TK}\right), \sigma^2 MM' \otimes (P_L(XX')^{-1}P_L)\right)$$

as $n \to \infty$.

- Proof outline:
  - $\widehat{\mathcal{V}}$ is a linear transformation of Gaussian data.
  - $(\widehat{L}_1, \ldots, \widehat{L}_l, \widehat{M}_1, \ldots \widehat{M}_p)$ are consistent.
  - $\text{vec}(\hat{\mathcal{B}}_{TK})$ is assembled using Slutsky's theorem.
- Similar results for CP and TR formats on $\mathcal{B}$.

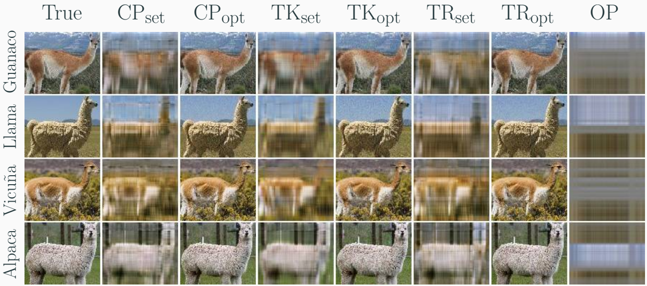
Outline

Andean Camelids Simulation

Matrix-on-Matrix Regression: Two Factor ANOVA

$$Y_{ijk} = \langle X_{ij} | \mathcal{B} \rangle + E_{ijk}, \quad E_{ijk} \overset{iid}{\sim} \mathcal{N}_{87,106}(0, \sigma^2 \Sigma_1, \Sigma_2)$$

- X_{ij} has 1 in position (i, j) and zeroes everywhere else.
- $\langle X_{ij} | \mathcal{B} \rangle$ is the jth RGB color matrix of the ith andean camelid.
- $k = 1, 2, \ldots, 50$ are repetitions.



BIC	1.692×10^7	1.586×10^7	1.695×10^7	1.612×10^7	1.703×10^7	1.590×10^7	1.772×10^7
% reduction	97.3%	67.1%	97.4%	41.5%	97.5%	57.9%	99.4%

Labeled Faces in the Wild

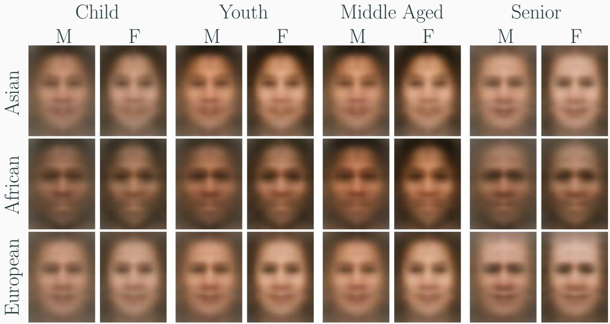
- We selected 605 facial images from more than 13,000.
- Three ethnic origins: African, European and Asian.
- Four cohorts: child, youth, middle-aged and senior.
- Two genders: male and female.
- Attributes taken from Afifi and Abdelhamed (2019).

Labeled Faces in the Wild

Tensor-on-Tensor Regression: Three Factor ANOVA

$$\mathcal{Y}_{ijkl} = \langle \mathcal{X}_{ijk} | \mathcal{B} \rangle + \mathcal{E}_{ijkl}, \quad \mathcal{E}_{ijkl} \overset{iid}{\sim} \mathcal{N}_{151,111,3}(0, \sigma^2 \Sigma_1, \Sigma_2, \Sigma_3)$$

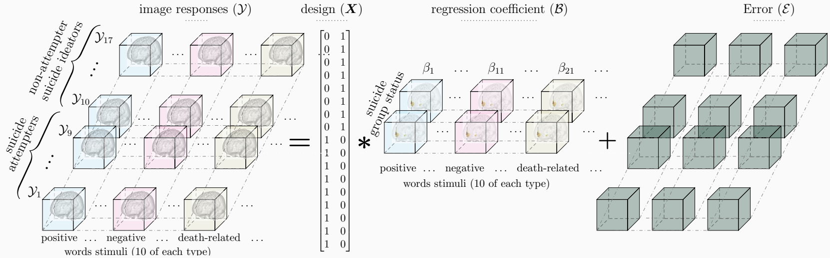
- $\mathcal{Y}_{ijkl}$ is the lth repetition of the color picture of someone with ith gender, jth ethnic origin and kth cohort.
- $\mathcal{X}_{ijk}$ has 1 in position (i, j, k) and zeroes everywhere else.
- $\mathcal{B} = \text{tr}(\mathcal{L}_1 \times^1 \mathcal{L}_2 \times^1 \mathcal{L}_3 \times^1 \mathcal{M}_1 \times^1 \mathcal{M}_2 \times^1 \mathcal{M}_3)$ has a TR format.



Assessing Suicide Risk

Tensor-on-Tensor Regression:

$$\mathcal{Y}_i = \langle \boldsymbol{x}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 \Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4, \Sigma_5),$$



- Tucker format on $\mathcal{B}$, with rank chosen using BIC.
- Dimension reduction in $\mathcal{B}$ of 97.3%.
- Σs chosen as AR(1), unconstrained and equicorrelation matrices.

Sampling Distribution of the Interaction Term

- Sampling distribution for $XX' = diag(9,8)$:

$$\widehat{\mathcal{B}} \xrightarrow{d} \mathcal{N}_{2,3,10,43,56,20}(\mathcal{B}, \sigma^2(XX')^{-1}, M_1M_1', M_2M_2', M_3M_3', M_4M_4', M_5M_5'),$$

- Interaction term: $\widehat{\mathcal{B}}_* = \widehat{\mathcal{B}} \times_1 \boldsymbol{c}_1 \times_2 C_2 \times_3 \boldsymbol{c}_3$, where

$$\boldsymbol{c}_1 = \begin{bmatrix} 1 & -1 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 1 & 0 & -1 \\ 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \text{ and } \boldsymbol{c}_3 = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}.$$

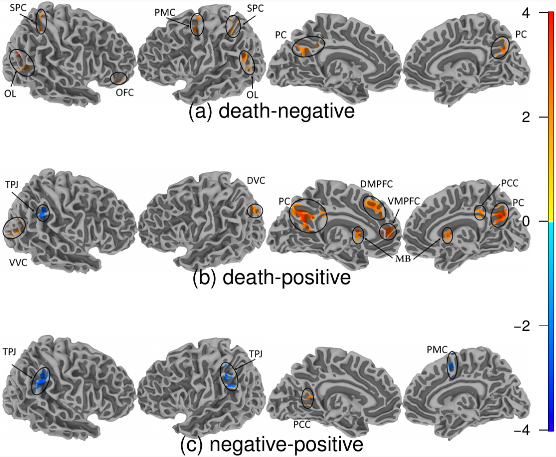
- Marginally standardize $\widehat{\mathcal{B}}_*$ to obtain $\widehat{\mathcal{Z}}$.
- Under the null hypothesis $H_o : \mathcal{B}_*(i,j,k,l) = 0$, we have that $\mathcal{Z}_*(i,j,k,l) \xrightarrow{d} N(0,1)$.

Interaction 3D maps

- PC associated with depression and rumination
- OFC associated with emotions' influence on decision-making

- PC is even more pronounced
- D/VMPFC, MB, PCC involved in processing emotional info

- Low TPJ and PMC, but not as much overall activation
- positive and negative words are more neurally similar



(PC) precuneus
(PMC) premotor cortex
(DVC) dorsal visual cortex
(OL) occipital lobe
(MB) mamillary bodies
(SPC) superior parietal cortex
(TPJ) temporal-parietal junction
(PCC) posterior cingulate cortex
(OFC) orbital frontal cortex
(VVC) ventral visual cortex
(DMPFC) dorsal medial frontal cortex
(VMPFC) ventral medial prefrontal cortex

Conclusions

- Extended MANOVA to tensor-variate structure.
- Different tensor formats compared using BIC.
- Provided MLE algorithms, their asymptotic properties, computational complexity, and evaluated them with simulation.
- Distinguished facial characteristics.
- Identified brain regions associated with suicide attempt, and negative, positive, or death-related stimuli

Further Work

- Time-series integrated Tensor-on-Tensor Regression for fMRI studies.
- Robust estimation for deviations from tensor-variate normality.
- Classification and prediction.
- Tensor-variate Mixed effects models and Gauge R& R