Soft Tensor Regression

Georgia Papadogeorgou

with Zhengwu Zhang & David B. Dunson

CMStatistics, December 14, 2019

Motivational setting

- In many applications, data naturally have an array or tensor structure
  - For example, $R \times R \times p$ array containing features measuring the strength of connections between an individual's R brain regions
- Characterize the relationship between a tensor predictor and a scalar outcome within a regression framework
- Scalar ~ Tensor

Statistical approaches for tensor regression

Estimation requires some type of parameter regularization or dimensionality reduction

1. Estimating coefficients with entry-specific penalization
   (Cox and Savoy, 2003; Craddock et al., 2009)
   Does not account for the array structure of the predictor

Statistical approaches for tensor regression

Estimation requires some type of parameter regularization or dimensionality reduction

1. Estimating coefficients with entry-specific penalization
   (Cox and Savoy, 2003; Craddock et al., 2009)
   Does not account for the array structure of the predictor
2. Use low-dimensional summaries of the tensor predictor
   (Zhang et al., 2019; Zhai and Li, 2019)
   Unsupervised, performance depends on number and choice

Statistical approaches for tensor regression

Estimation requires some type of parameter regularization or dimensionality reduction

1. Estimating coefficients with entry-specific penalization
   (Cox and Savoy, 2003; Craddock et al., 2009)
   Does not account for the array structure of the predictor
2. Use low-dimensional summaries of the tensor predictor
   (Zhang et al., 2019; Zhai and Li, 2019)
   Unsupervised, performance depends on number and choice
3. Estimate a coefficient tensor assuming a low-rank structure
   (Zhou et al., 2013; Li et al., 2018; Guhaniyogi et al., 2017; Guha and Rodriguez, 2018; Wang et al., 2018)
   Attractive, can suffer if the true tensor is not low-rank

Challenge 1

- Estimation of high-dimensional tensor model
- Respecting the predictor's array structure

Challenge 2

- Low rank approximations can perform poorly

Our goal: Develop a tensor regression framework that

1. accommodates the **predictor's structure**
2. adaptively **expands away** from low-rank

Notation

- Y_i: continuous outcome of unit i
- $\boldsymbol{X}_i$: K-mode tensor of dimensions $p_1, p_2, \ldots, p_K$ with entries $[\boldsymbol{X}_i]_{j_1 j_2 \ldots j_K} = X_{i, j_1 j_2 \ldots j_K}$
- Assume model

$$Y_i = \mu + \langle \boldsymbol{X}_i, \boldsymbol{B} \rangle_F + \epsilon_i$$

where

$\boldsymbol{B}$ is K-mode coefficient tensor of dimensions $p_1, p_2, \ldots, p_K$

$\langle \boldsymbol{X}_i, \boldsymbol{B} \rangle_F = \sum_{j_1=1}^{p_1} \sum_{j_2=1}^{p_2} \cdots \sum_{j_K=1}^{p_K} X_{i, j_1 j_2 \ldots j_K} B_{j_1 j_2 \ldots j_K}$

PARAFAC decomposition

A tensor $\boldsymbol{B} \in \mathbb{R}^{p_1 \times p_2 \times \dots p_K}$ can be written as

$$\boldsymbol{B} = \sum_{d=1}^{D} \beta_1^{(d)} \otimes \beta_2^{(d)} \otimes \cdots \otimes \beta_K^{(d)}$$

for $\beta_k^{(d)} \in \mathbb{R}^{p_k}$. The minimum value of D is referred to as its rank.

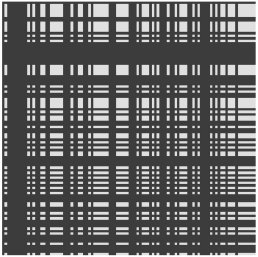
- The $(j_1 j_2 \dots j_K)$ entry of $\boldsymbol{B}$ is equal to

$$\boldsymbol{B}_{j_1 j_2 \dots j_K} = \sum_{d=1}^{D} \beta_{1 j_1}^{(d)} \beta_{2 j_2}^{(d)} \dots \beta_{K j_K}^{(d)}$$

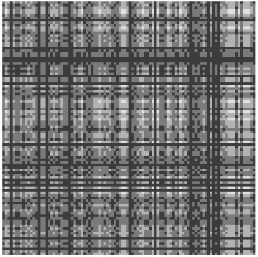
- Row j_k along mode k has **fixed** importance to all coefficient entries that include it

- Natural approximation of the coefficient tensor (Zhou et al., 2013; Guhaniyogi et al., 2017)

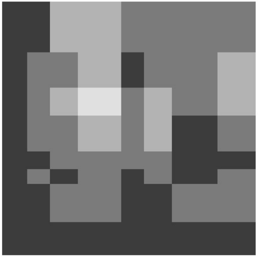
Block structure of the PARAFAC



Rank 1



Rank 3



Rank 3 ordered

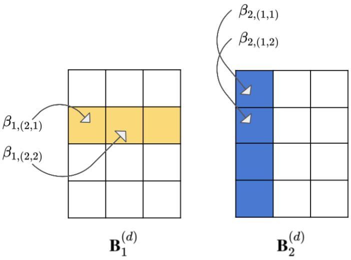
- We refer to it as the **hard** PARAFAC

Soft tensor regression

- Write $\boldsymbol{B} = \sum_{d=1}^{D} \boldsymbol{B}_1^{(d)} \circ \boldsymbol{B}_2^{(d)} \circ \ldots \circ \boldsymbol{B}_K^{(d)}$ with $\boldsymbol{B}_k^{(d)}$ equal dimension to $\boldsymbol{B}$
- Now $\boldsymbol{B}_{\utilde{j}} = \sum_{d=1}^{D} \beta_{1\utilde{j}}^{(d)} \beta_{2\utilde{j}}^{(d)} \ldots \beta_{K\utilde{j}}^{(d)}$, for $\utilde{j} = (j_1, j_2, \ldots, j_K)$

Soft tensor regression

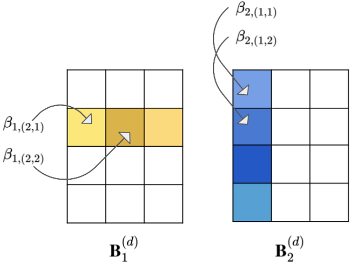
- Write $\boldsymbol{B} = \sum_{d=1}^{D} \boldsymbol{B}_1^{(d)} \circ \boldsymbol{B}_2^{(d)} \circ \ldots \circ \boldsymbol{B}_K^{(d)}$ with $\boldsymbol{B}_k^{(d)}$ equal dimension to $\boldsymbol{B}$
- Now $\boldsymbol{B}_{\underset{\sim}{j}} = \sum_{d=1}^{D} \beta_{1\underset{\sim}{j}}^{(d)} \beta_{2\underset{\sim}{j}}^{(d)} \ldots \beta_{K\underset{\sim}{j}}^{(d)}$, for $\underset{\sim}{j} = (j_1, j_2, \ldots, j_K)$
- Hard PARAFAC can be written like this by setting $\beta_{k,\underset{\sim}{j}}^{(d)} = \gamma_{k,j_k}^{(d)}$



Soft PARAFAC structure

$$\beta_{k,\underset{\sim}{j}}^{(d)} \sim N(\gamma_{k,j_k}^{(d)}, \sigma_k^2 \zeta^{(d)})$$

- Hard PARAFAC-centered: $\mathbb{E}[\boldsymbol{B}_{\underset{\sim}{j}} | \Gamma, S, Z] = \sum_{d=1}^{D} \gamma_{1j_1}^{(d)} \gamma_{2j_2}^{(d)} \dots \gamma_{Kj_K}^{(d)}$
- $\gamma_{k,j_k}^{(d)}$ represents overall importance of row j_k
- Allows variation within the mode$-k$ slices



Bayesian inference

$$
\begin{aligned}
\beta^{(d)}_{k,\underset{\sim}{j}} &\sim N(\gamma^{(d)}_{k,j_k}, \sigma^2_k \zeta^{(d)}) \\
\gamma^{(d)}_{k,j_k} &\sim N(0, \tau_\gamma \zeta^{(d)} w^{(d)}_{k,j_k}) \\
w^{(d)}_{k,j_k} &\sim Exp((\lambda^{(d)}_k)^2/2), \\
\lambda^{(d)}_k &\sim \Gamma(a_\lambda, b_\lambda) \\
\zeta &\sim \text{Dirichlet}(\alpha/D, \alpha/D, \ldots, \alpha/D) \\
\sigma^2_k &\sim \Gamma(a_\sigma, b_\sigma)
\end{aligned}
$$

τ_γ: Overall variance
$w^{(d)}_{k,j_k}$: Row-specific variance
$\zeta^{(d)}$: Component variance scaling
$\sigma^2_k\ \zeta^{(d)}$: PARAFAC softening

Underlying hard PARAFAC prior from Guhaniyogi et al. (2017)

Choosing the hyperparameters

- Prior coefficient variance V^*
- Percentage of prior variance due to softening AV^*

Proposition 1

For a matrix predictor, if

$$\frac{2b_\lambda^2}{(a_\lambda - 1)(a_\lambda - 2)} = \frac{b_\tau}{a_\tau}\sqrt{\frac{V^*(1 - AV^*)a_\tau}{C'(a_\tau + 1)}}$$

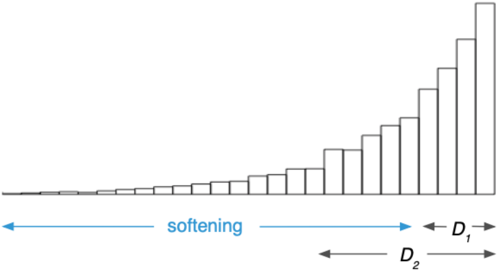
and

$$\frac{a_\sigma}{b_\sigma} = \sqrt{\frac{V^*(1 - AV^*)a_\tau}{C(a_\tau + 1)}}\left\{\sqrt{1 - \frac{a_\tau + 1}{a_\tau}\left\{1 - \left(1 - AV^*\right)^{-1}\right\}} - 1\right\}$$

then $\mathrm{Var}(\underset{\sim}{\boldsymbol{B}_j}) = V^*$, *and* $AV = AV^*$.

Dependence on the underlying PARAFAC rank

- Softer is more robust to the choice of D than the hard PARAFAC
- Hard PARAFAC with D_1 can capture D_1 largest eigenvalues
- Softening the D_1-PARAFAC can capture deviations arising from all eigenvalues



Full prior support and posterior consistency

For true coefficient tensor $\boldsymbol{B}^0$ for **any rank**:

Proposition 2

For $\epsilon > 0$, $\pi_{\boldsymbol{B}}\big(\mathcal{B}_\epsilon^\infty(\boldsymbol{B}^0)\big) > 0$ *where* $\mathcal{B}_\epsilon^\infty(\boldsymbol{B}^0) = \{\boldsymbol{B} : \max_{\underset{\sim}{j}} |\boldsymbol{B}^0_{\underset{\sim}{j}} - \boldsymbol{B}_{\underset{\sim}{j}}| < \epsilon\}$.

Proposition 3

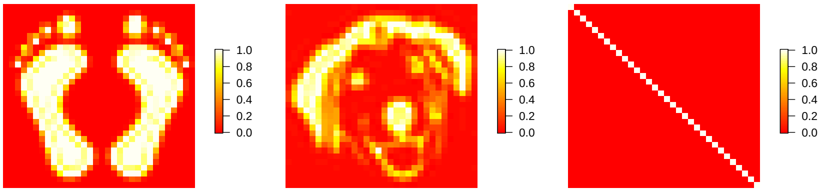
For any $\epsilon > 0$, *there exists* $\epsilon^* > 0$ *such that*

$$\Big\{\boldsymbol{B} : \max_{\underset{\sim}{j}} |\boldsymbol{B}^0_{\underset{\sim}{j}} - \boldsymbol{B}_{\underset{\sim}{j}}| < \epsilon^*\Big\} \subseteq \Big\{\boldsymbol{B} : KL(\boldsymbol{B}^0, \boldsymbol{B}) < \epsilon\Big\}$$

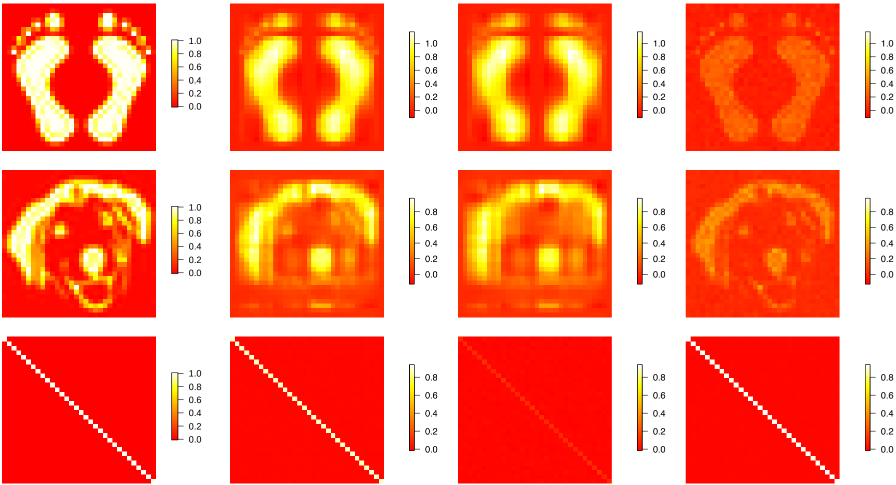
Proposition 3 → Weak consistency (Schwartz, 1965)

Simulations

- Matrix predictor of dimension 32×32
- Sample size: 400
- True coefficient tensors:



Simulation results



Truth Softer PARAFAC Lasso

Simulation conclusions

- Softer uses the low-rank structure of the PARAFAC when necessary, and diverge from it when needed
- We evaluated:
  1. MSE in coefficient estimatioon
  2. Frequentist coverage of 95% credible intervals
  3. Identification of important entries (sensitivity, specificity, FNR, FPR)
  4. Predictive MSE
- FPR much lower for Softer than hard PARAFAC
- Simulations with increasing rank of true coefficient tensor

Results from brain connectomics study

- We extended Softer to (semi-)symmetric tensors
- Extension to binary outcomes
- Employed tensor regression to analyze the relationship between
  - Features of structural brain connections, and
  - 15 human traits (personality, motor, etc)
- In the analysis
  - Methods had similar predictive performance
  - Up to 30% of the variance explained
  - Softer identified important structural connections for predicting three traits that agree with neuroscience literature

References

Soft Tensor Regression – arXiv:1910.09699

David D. Cox and Robert L. Savoy. Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2):261–270, 2003.

R. Cameron Craddock, Paul E. Holtzheimer, Xiaoping P. Hu, and Helen S. Mayberg. Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine*, 62(6):1619–1628, dec 2009.

Sharmistha Guha and Abel Rodriguez. Bayesian regression with undirected network predictors with an application to brain connectome data. Technical report, 2018.

Rajarshi Guhaniyogi, Shaan Qamar, and David B Dunson. Bayesian tensor regression. *Journal of Machine Learning Research*, 18:1–31, 2017.

Xiaoshan Li, Da Xu, Hua Zhou, and Lexin Li. Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10 (3):520–545, 2018.

Lorraine Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.

Lu Wang, Zhengwu Zhang, and David Dunson. Symmetric bilinear regression for signal subgraph estimation. *IEEE Transactions on Signal Processing*, PP(c):1, 2018.

Jian Zhai and Ke Li. Predicting brain age based on spatial and temporal features of human brain functional networks. *Frontiers in Human Neuroscience*, 13:Article 62, 2019.

Zhengwu Zhang, Genevera I. Allen, Hongtu Zhu, and David Dunson. Tensor network factorizations: Relationships between brain structural connectomes and traits. *NeuroImage*, 197:330–343, 2019.

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