Tensor Regression and Its Application

Botao Hao

March 1, 2016

Outline

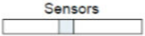
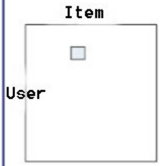
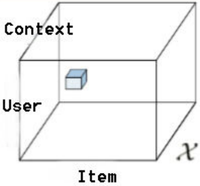
1. Introduction
   - Background
   - Overview
   - Tensorial Data Analysis
2. Tensor Predictor Regression
   - Introduction
   - CP-Decomposition
   - Convex Regularization
3. Future Work

Outline

1. Introduction
   - Background
   - Overview
   - Tensorial Data Analysis
2. Tensor Predictor Regression
   - Introduction
   - CP-Decomposition
   - Convex Regularization
3. Future Work

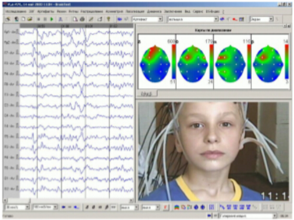
What is Tensor?

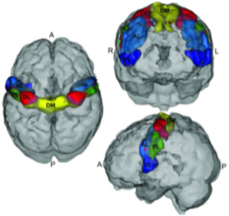
- Tensor is *Muti-Dimensional Array Data*, formally denoted as $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$.

Order	1st	2nd	3rd
Correspondence	Vector	Matrix	3D array
Example			

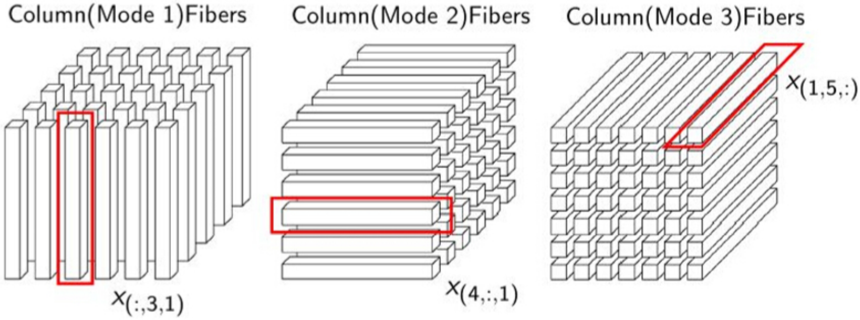
Some Tensorial Type Data

- Neuroscience
  - fMRI data:(time $\times$ x axis $\times$ y axis $\times$ z axis)
- Vision
  - image (video) data:
    (pixel $\times$ illumination $\times$ expression $\times$ viewpoints)

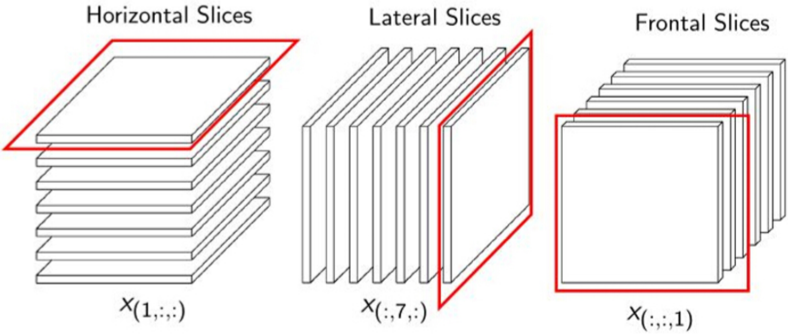




Fibers


Column(Mode 1)Fibers
Column(Mode 2)Fibers
Column(Mode 3)Fibers
X(:,3,1)
X(4,:,1)
X(1,5,:)

Slices



Outline

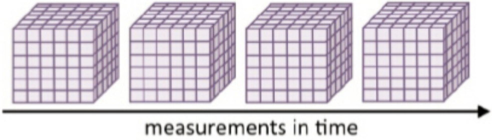
Recent Progress in Statistical Community

- Tensor decomposition.
  - *Learning latent variable model. (Anandkumar et al. JMLR 2014)*
  - *Bayesian tensor decomposition in contingency table (Zhou et al. JASA 2014), categorical classification (Yang and Dunson JASA 2014), log-linear model (Johndrow et al. AOS 2015)*
  - *High-dimensional tensor decomposition. (Sun et al. arXiv. 2015).*
- Tensor completion: Denoise the observational noise and complete the unobserved element.
- Tensor regression.
  - *Tensor predictor regression (Zhou et al. JASA 2013), tensor response regression (Li et al. arXiv 2015), high dimensional tensor regression (Yuan et al. arXiv 2015).*
  - *Bayesian tensor regression (Guhaniyogi et al. arXiv 2015).*

Outline

Brain Imaging Data Analysis

- Neuroimaging can explain the brain physiology
- Several types of neuroimaging: MRI fMRI
- Two popular approaches:
  - Voxel-based methods, multiple testing...
  - Regression-based method. Treat outcome as response.



One fMRI Observation from One Subject

Brain Imaging Data Analysis via Regression

- Tensor predictor regression
  - **Goal:** Understand the change of a clinical outcome as the tensor image varies. Disease diagnosis.
  - The covariates are the tensor.
- Tensor response regression
  - **Goal:** Study the change of the image as the predictors such as the disease status and age vary. Identify brain regions.
  - The responses are the tensor.

Limitation of Classical Regression

- Why tensor regression? Ultrahigh dimensionality!
- **Naive approach:** Tuning an image array into a vector
  - e.g. a MRI image: Third-order array with size $256 \times 256 \times 256$ requires $256^3 = 16,777,215$ regression parameters.
  - Why not screening first? Ignores spatial and temporal correlation.
- **Tensor Regression**: directly model each tensor observation in regression model.

Outline

Generalized Linear Model with Tensor Covariate

- In classical GLM, Y belongs to an exponential family with density:

$$p(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

- The tensor GLM relates a tensor-valued $\boldsymbol{X} \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}$ to the mean $\mu = \mathbb{E}(Y|\boldsymbol{X})$ via

$$g(u) = \alpha + \gamma^\top \boldsymbol{Z} + \langle \boldsymbol{B}, \boldsymbol{X} \rangle \tag{2.2}$$

where $\boldsymbol{B}$ is of same size, namely $\prod_{d=1}^{D} p_d$, as $\boldsymbol{X}$.

- $\langle \cdot, \cdot \rangle$ is the inner product.

Special Structures on $\boldsymbol{B}$

- Vector case in linear regression: we assume true parameter β^* is sparse, which means $\|\beta^*\|_0 \leq \lambda$.
- For matrix value, there are two kinds of structure assumptions:
  - Sparsity assumption: $\|A\|_0 \leq \lambda$
  - Low-rank assumption: $\text{rank}(A) \leq r$.
    - PCA or SVD
    - Matrix nuclear norm
- Tensor type data has more flexible structure assumptions.
  - Sparsity assumption: $\|\boldsymbol{B}\|_0 \leq \lambda$, element-wise, fiber-wise, slice-wise .
  - Low-rank assumption:
    - CP-decomposition
    - Tensor nuclear norm

Low-rank extension is non-trivial!

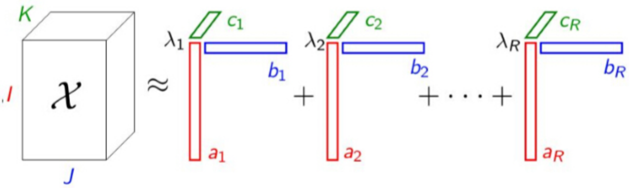
Outline

CANDECOMP/PARAFAC Decomposition

Definition

The rank of a tensor $\mathcal{X}$, denoted $\text{rank}(\mathcal{X})$, is defined as the smallest number of rank-one tensors that generate $\mathcal{X}$ as their sum.

Directly computing CP-rank is NP-hard!



$$\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \cdot a_r \circ b_r \circ c_r$$

Rank-R Generalized Linear Tensor Regression Model

- Suppose $\boldsymbol{B}$ admits a rank-R decomposition, then

$$g(\mu) = \alpha + \gamma^\top \boldsymbol{Z} + \langle \sum_{r=1}^{R} \beta_1^{(r)} \circ \beta_2^{(r)} \circ \dots \beta_D^{(r)}, \boldsymbol{X} \rangle \quad (2.3)$$

$$= \alpha + \gamma^\top \boldsymbol{Z} + \langle (\boldsymbol{B}_D \odot \dots \odot \boldsymbol{B}_1) I_R, \text{vec}\boldsymbol{X} \rangle \quad (2.4)$$

- $\boldsymbol{B}_d = [\beta_d^{(1)}, \dots, \beta_d^{(R)}] \in \mathbb{R}^{p_d \times R}$.
- The tensor rank R is set to be known.
- A key observation is that although $g(\mu)$ is not linear in $\boldsymbol{B}$ jointly, it is linear in $\boldsymbol{B}_d$ individually.
- $\odot$ is the Khatri-Rao product.

$$\boldsymbol{A} \odot \boldsymbol{B} = [a_1 \otimes b_1, a_2 \otimes b_2, \dots, a_n \otimes b_n]$$

Estimation

- Maximum likelihood estimation. The log-likelihood function is

$$l(\alpha, \gamma, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = \sum_{i=1}^{n} \frac{y\theta - b(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y, \phi) \qquad (2.5)$$

- We rewrite the array inner product as

$$\begin{aligned} & \left\langle \sum_{r=1}^{R} \beta_1^{(r)} \circ \beta_2^{(r)} \circ \ldots \beta_D^{(r)}, \boldsymbol{X} \right\rangle \\ = \quad & \left\langle \boldsymbol{B}_d, \boldsymbol{X}_{(d)} (\boldsymbol{B}_D \odot \ldots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \ldots \odot \boldsymbol{B}_1) \right\rangle \end{aligned}$$

- Updating a block $\boldsymbol{B}_d$ is simply a classical GLM problem with Rp_d parameters.

Algorithm

Algorithm 1 Block relaxation algorithm for maximizing (5).

Initialize: $(\alpha^{(0)}, \gamma^{(0)}) = \text{argmax}_{\alpha,\gamma}\, \ell(\alpha, \gamma, \mathbf{0}, \ldots, \mathbf{0})$, $\boldsymbol{B}_d^{(0)} \in \mathbb{R}^{p_d \times R}$ a random matrix for $d = 1, \ldots, D$.

repeat

  for $d = 1, \ldots, D$ **do**

    $\boldsymbol{B}_d^{(t+1)} = \text{argmax}_{\boldsymbol{B}_d}\, \ell(\alpha^{(t)}, \gamma^{(t)}, \boldsymbol{B}_1^{(t+1)}, \ldots, \boldsymbol{B}_{d-1}^{(t+1)}, \boldsymbol{B}_d, \boldsymbol{B}_{d+1}^{(t)}, \ldots, \boldsymbol{B}_D^{(t)})$

  end for

  $(\alpha^{(t+1)}, \gamma^{(t+1)}) = \text{argmax}_{\alpha,\gamma}\, \ell(\alpha, \gamma, \boldsymbol{B}_1^{(t+1)}, \ldots, \boldsymbol{B}_D^{(t+1)})$

until $\ell(\boldsymbol{\theta}^{(t+1)}) - \ell(\boldsymbol{\theta}^{(t)}) < \epsilon$

Theory

Theorem (consistency)

Assume $\boldsymbol{B}_0$ is identifiable and the array covariates $\boldsymbol{X}_i$ are iid from a bounded distribution. The MLE is consistent, namely, $\widehat{\boldsymbol{B}}_n$ converges to $\boldsymbol{B}_0$ in probability .

Theorem (Asymptotic Normality)

For a fixed number of parameters p, an interior $\boldsymbol{B}_0 \in \mathcal{B}$ with singular information matrix $\boldsymbol{I}(\boldsymbol{B}_{01}, \ldots, \boldsymbol{B}_{0D})$,

$$\sqrt{n}[\mathit{vec}(\widehat{\boldsymbol{B}}_{n1}, \ldots, \widehat{\boldsymbol{B}}_{nD}) - \mathit{vec}(\boldsymbol{B}_{01}, \ldots, \boldsymbol{B}_{0D})] \tag{2.6}$$

converges in distribution to a normal with mean zero and covariance $I^{-1}(\boldsymbol{B}_{01}, \ldots, \boldsymbol{B}_{0D})$.

Outline

1. Introduction
   - Background
   - Overview
   - Tensorial Data Analysis
2. Tensor Predictor Regression
   - Introduction
   - CP-Decomposition
   - Convex Regularization
3. Future Work

General Tensor Regression Model

- Consider a general tensor regression problem where covariate tensors $\boldsymbol{X}^{(i)} \in \mathbb{R}^{d_1 \times \ldots \times d_M}$ and response tensors $\boldsymbol{Y}^{(i)} \in \mathbb{R}^{d_{M+1} \times \ldots \times d_N}$ are related through:

$$\boldsymbol{Y}^{(i)} = \langle \boldsymbol{X}^{(i)}, \boldsymbol{A} \rangle + \epsilon^{(i)}, \ i = 1, 2 \ldots, n \tag{2.7}$$

  where $\boldsymbol{A} \in \mathbb{R}^{d_1 \times \ldots \times d_N}$ is an unknown parameter of interest and $\epsilon^{(i)} s$ are $i.i.d$ noise tensors.
- Gaussian design: $X \sim \mathcal{N}(0, \Sigma)$. We vectorized tensor $\boldsymbol{X}$ by X. $c_l^2 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq c_u^2$.
- When $M = N$, it's a linear tensor regression model.
- When $M = 1$, it's a tensor response regression model.

General Tensor Regression Model

- Here the definition of $\langle \boldsymbol{A}, \boldsymbol{B} \rangle$ is different from tensor inner product.
- For $\boldsymbol{A} \in \mathbb{R}^{d_1 \times \ldots \times d_M}$ and $\boldsymbol{B} \in \mathbb{R}^{d_1 \times \ldots \times d_N}$:

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \sum_{j_1=1}^{d_1} \ldots \sum_{j_M=1}^{d_M} \boldsymbol{A}_{j_1,\ldots,j_M} \boldsymbol{B}_{j_1,\ldots,j_M} \in \mathbb{R}$$

  is the usual inner product if $M = N$. And if $M < N$, then $\langle \boldsymbol{A}, \boldsymbol{B} \rangle \in \mathbb{R}^{d_{M+1} \times \ldots \times d_N}$ such that its $(j_{M+1}, \ldots, j_N)$ entry is given by

$$(\langle \boldsymbol{A}, \boldsymbol{B} \rangle)_{j_{M+1},\ldots,j_N} = \sum_{j_1=1}^{d_1} \ldots \sum_{j_M=1}^{d_M} \boldsymbol{A}_{j_1,\ldots,j_M} \boldsymbol{B}_{j_1\ldots,j_M,j_{M+1},\ldots,j_N}$$

  .
- It's a generalization for matrix and vector multiplication.

Convex Regularization Framework

- The regularized least-squares objective function:

$$\widehat{\boldsymbol{T}} \in \underset{\boldsymbol{A}\in\mathbb{R}^{d_1\times\ldots\times d_N}}{\operatorname{argmin}} \left\{ \frac{1}{2n}\sum_{i=1}^{n} \|\boldsymbol{Y}^{(i)} - \langle \boldsymbol{A}, \boldsymbol{X}^{(i)}\rangle\|_F^2 + \lambda\mathcal{R}(\boldsymbol{A}) \right\}$$

- Comparing to CP decomposition added on $\boldsymbol{A}$, we use convex regularizer to enforce low-dimensional structure.
- We require regularizer $\mathcal{R}(\boldsymbol{A})$ is weakly decomposable, extending the idea from vectors and matrices. (Negahban et al. 2012)

Sparsity Regularizers

- Entry-wise l_1 penalty:

$$\mathcal{R}(\boldsymbol{A}) := \sum_{j_1=1}^{d_1} \sum_{j_2=1}^{d_2} \sum_{j_3=1}^{d_3} |A_{j_1 j_2 j_3}|$$

- Fiber-wise group structure penalty

$$\mathcal{R}(\boldsymbol{A}) := \sum_{j_2=1}^{d_2} \sum_{j_3=1}^{d_3} \|A_{\cdot j_2 j_3}\|_2$$

- Slice-wise group structure penalty

$$\mathcal{R}(\boldsymbol{A}) := \sum_{j_3=1}^{d_3} \|A_{\cdot\cdot j_3}\|$$

Low-rankness Regularizers

- Tensor nuclear norm:

$$\mathcal{R}(\boldsymbol{A}) = \max_{\|\boldsymbol{B}\|_s \le 1} \langle \boldsymbol{A}, \boldsymbol{B} \rangle$$

where

$$\|\boldsymbol{B}\|_s = \max_{\|u\|_2, \|v\|_2, \|w\|_2 \le 1} \langle \boldsymbol{A}, u \otimes v \otimes w \rangle$$

is called tensor spectral norm.

- It is convex and weakly decomposable.

Finite Sample Risk Bound for Regularized Estimator

Theorem

Let $\widehat{\boldsymbol{T}}$ be the regularized estimator and $\mathcal{R}(\cdot)$ is decomposable. If

$$\lambda \geq \frac{2c_u(3+c_{\mathcal{R}})}{c_{\mathcal{R}}\sqrt{n}}\mathbb{E}[\mathcal{R}^*(G)], \tag{2.8}$$

then there exists a constant $c > 0$ such that with probability at least $1-\exp\{-c\mathbb{E}[\mathcal{R}^*(G)]\}$,

$$\max\left\{\|\widehat{\boldsymbol{T}}-\boldsymbol{T}\|_n^2, \|\widehat{\boldsymbol{T}}-\boldsymbol{T}\|_F^2\right\} \leq \frac{6(1+c_{\mathcal{R}})}{3+c_{\mathcal{R}}}\frac{9c_u^2}{c_l^2}s(\mathcal{A})\lambda^2 \tag{2.9}$$

when n is sufficiently large.

- $\mathcal{R}^*(\cdot)$ is the dual norm of $\mathcal{R}$.

- $\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_F^2$ measures the parameter estimation accuracy. $\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_n^2 := \frac{1}{n}\sum_{i=1}^n \|\langle \boldsymbol{X}^{(i)}, \widehat{\boldsymbol{T}} - \boldsymbol{T}\rangle\|_F^2$ measures the predictive accuracy.
- $\mathbb{E}[\mathcal{R}^*(G)]$ captures how large the $\mathcal{R}(\cdot)$ norm is relative to the $\|\cdot\|_F$. $s(\mathcal{A})$ captures the low dimension of the subspace $\mathcal{A}$.

Sparsity Regularizers

Lemma

Recall that vectorized l_1 regularizer: $\mathcal{R}(\boldsymbol{A}) := \sum\sum\sum |A_{j_1 j_2 j_3}|$. *Let*

$$\Theta_1(s) = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : \sum_{j_1=1}^{d_1} \sum_{j_2=1}^{d_2} \sum_{j_3=1}^{d_3} \mathbb{I}(A_{j_1 j_2 j_3} \neq 0) \leq s\},$$

The main theorem implies that

$$\sup_{T \in \Theta_1(s)} \|\widehat{T}_1 - T\|_{\mathrm{F}}^2 \lesssim \frac{s \log(d_1 d_2 d_3)}{n}$$

with high probability by taking $\lambda \asymp \sqrt{\frac{\log(d_1 d_2 d_3)}{n}}$.

Specific Statistical Problems

- Muti-Response regression with large p.
- Mutivariate Sparse Auto-regressive Models.
- In the specific statistical problems, they provide min-max lower bound and show that with proper choice of tuning parameter their estimator can achieve the min-max lower bound.

Future Work

- Efficient algorithm for convex regularization problem, especially for tensor nuclear norm.
- EM approach for heterogeneous tensor data (mixed tensor regression).

$$Y = \tau^T Z + \langle \boldsymbol{X}, L \cdot \boldsymbol{A} \rangle + \boldsymbol{W}$$

  L is latent variable with Rademacher distribution over $\{-1, 1\}$. We can analyze both the statistical error and optimization error for $\widehat{\boldsymbol{A}}$.

Reference I

- Anandkumar, A., Ge, R., Hsu, D., Kakade, S.M. and Telgarsky M. (2014). Tensor Decompositions for Learning Latent Variable Models. *Journal of Machine Learning Research.*
- Guhaniyogi, R., Qamar, S. and Dunson, D.B. (2015). Bayesian tensor regression. *arXiv.*
- Johndrow, J., Bhattacharya, A. and Dunson, D.B. (2015). Tensor decompositions and sparse log-linear models. *Annals of Statistics.*
- Raskutti, G., Yuan, M.(2015). Convex Regularization for High-Dimensional Tensor Regression. *arXiv.*

Reference II

- Yang, Y. and Dunson, D.B. (2015). Bayesian conditional tensor factorizations for high-dimensional classification. *Journal of the American Statistical Association.*
- Zhou, H., Li,L., and Zhu, H.(2013). Tensor regression with applications in neuroimaging data analysis. *Journal of American Statistical Association*.
- Zhou, J., Bhattacharya, A., Herring, A.H. and Dunson, D. (2015). Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*

Tensor Operator

- If $a_1 = (a_{11}, a_{12})$, $b_1 = (b_{11}, b_{12})$, then

$$a_1 \otimes b_1 = (a_{11}b_{11}, a_{11}b_{12}, a_{12}b_{11}, a_{12}b_{12})^\top$$

  Vectorized outer product.

- Given two matrices $A = [a_1 \dots a_n] \in \mathbb{R}^{m \times n}$ and $B = [b_1 \dots b_q] \in \mathbb{R}^{p \times q}$, the *Khatri-Rao* product is defined as the mp-by-n matrix

$$\boldsymbol{A} \odot \boldsymbol{B} = [a_1 \otimes b_1, a_2 \otimes b_2, \dots, a_n \otimes b_n]$$