Towards a Regression using Tensors

Ming Hou

February 27, 2014
Outline

1 Background
   - Linear Regression
   - Tensorial Data Analysis

2 Tensor Basics
   - Definition
   - Tensor Operation
   - Tensor Decomposition

3 Generalized Linear Tensor Regression
   - Generalized Linear Tensor Regression Model
   - Attention Deficit Hyperactivity Disorder Data Analysis

4 Future Work Plan
   - Future Work Plan
Classical Linear Regression

- **Predict**
  e.g. speed, road conditions, weather ⇒ traffic accidents rates

- **Identify the key predictors**
  e.g. mental disease status ⇒ the regions of brain
Multi-Dimensional Array Data (Tensors)

- **Neuroscience**
  - EEG data: \((time \times frequency \times electrodes)\)
  - fMRI data: \((time \times x\ axis \times y\ axis \times z\ axis)\)

- **Vision**
  - image (video) data:
    \((pixel \times illumination \times expression \times viewpoints)\)

- **Chemistry**
  - fluorescence excitation-emission data:
    \((samples \times emission \times excitation)\)
Brain Imaging Data Analysis

- Mental health disorders are difficult to diagnose and treat
- Physiology of brain is not well understood
- Neuroimaging can explain the brain physiology
- Several types of neuroimaging EEG MRI fMRI
**Goal** is to find association between **brain images** and **clinical outcomes**.

Formulate as regression problem

- clinical outcome as response
- brain image (multi-dimensional array) as tensor predictor
Limitation of Classical Regression

**Naive approach**: turning an image array as vector predictor
- e.g. a fMRI image: 4D array with size $256 \times 256 \times 256 \times 100$
- yields a huge number of parameters (167 millions!)
- ignores spatial and temporal correlation

**New method**: treat each fMRI observation as one tensor predictor in regression model

One fMRI Observation from One Subject
What is Tensor?
What is Tensor? cont’d
What is Tensor? con’t

A tensor is formally denoted as $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$

- generalization of vector and matrix
- represented as multi-dimensional array

<table>
<thead>
<tr>
<th>Order</th>
<th>1st</th>
<th>2\textsuperscript{nd}</th>
<th>3\textsuperscript{rd}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Vector</td>
<td>Matrix</td>
<td>3D array</td>
</tr>
<tr>
<td></td>
<td>Sensors</td>
<td>Keywords</td>
<td></td>
</tr>
<tr>
<td>Example</td>
<td></td>
<td>Authors</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pots</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sources</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mathcal{X}$</td>
<td></td>
</tr>
</tbody>
</table>
Fibers

Column (Mode 1) Fibers

Column (Mode 2) Fibers

Column (Mode 3) Fibers

\( X(:,3,1) \)

\( X(:,4,1) \)

\( X(1,5,:) \)
Slices

Horizontal Slices

Lateral Slices

Frontal Slices

$X(1,:,:)$

$X(:,7,:)$

$X(:, :, 1)$
Matricization (Unfolding)

Convert a tensor to a matrix

\[ \chi \rightarrow \begin{bmatrix} \chi_1(l \times J) & \chi_2(l \times J) & \ldots & \chi_k(l \times J) \end{bmatrix} \]

Tube fibers are rearranged into the columns of a matrix
Matricization (Unfolding) Example

$X^{(1)} = \begin{pmatrix} 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{pmatrix}$

$X^{(2)} = \begin{pmatrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \end{pmatrix}$

$X^{(3)} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}$
The n-Mode Multiplication

Let $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, $\mathbf{B} \in \mathbb{R}^{M \times J}$, the 2-mode product of $\mathcal{X}$ with $\mathbf{B}$ is defined by

$$\mathcal{Y} = \mathcal{X} \times_2 \mathbf{B} \in \mathbb{R}^{I \times M \times K}$$

Elementwise

$$y_{imk} = \sum_j x_{ijk} b_{mj}$$

In matrix form

$$\mathbf{Y}_{(2)} = \mathbf{B} \mathbf{X}_{(2)}$$

Multiply each row (mode-2) fiber by $\mathbf{B}$
The n-Mode Multiplication Example

\[ \text{Type} \times \text{Time} \times \text{Clusters} \Rightarrow \text{Location} \]
Rank-1 Tensor

3-way outer product

\[ \mathcal{X} = a \circ b \circ c \]

Elementwise

\[ x_{ijk} = a_ib_jc_k \]


**CANDECOMP/PARAFAC Decomposition**

\[
\chi \approx \sum_{r=1}^{R} \lambda_r a_r \circ b_r \circ c_r
\]

\[
\chi \approx X \approx \sum_{r=1}^{R} \lambda_r a_r \circ b_r \circ c_r
\]

Ming Hou  
Towards a Regression using Tensors
Define **factor matrix** $A \in \mathbb{R}^{l \times R}$, $B \in \mathbb{R}^{J \times R}$ and $C \in \mathbb{R}^{K \times R}$

$$\chi \approx \sum_{r=1}^{R} \lambda_r a_r \circ b_r \circ c_r \equiv [\lambda; A, B, C]$$

$$\chi_{ijk} \approx \sum_{r=1}^{R} \lambda_r a_{ir} b_{jr} c_{kr}$$
Tucker decomposition

Defined by factor matrix $A \in \mathbb{R}^{I \times R}$, $B \in \mathbb{R}^{J \times S}$ and $C \in \mathbb{R}^{K \times T}$, and core tensor $G \in \mathbb{R}^{R \times S \times T}$

\[ X \approx G \times_1 A \times_2 B \times_3 C \equiv [G; A, B, C] \]

\[ x_{i j k} = \sum_{r=1}^{R} \sum_{r=1}^{S} \sum_{r=1}^{T} g_{rst} a_{ir} b_{js} c_{kt} \]
The standard linear regression model $\mathbf{x} \in \mathbb{R}^p$, $y = \beta^T \mathbf{x} + \alpha + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ can be written

$$\mu = \beta^T \mathbf{x} + \alpha \quad y \sim \mathcal{N}(\mu, \sigma^2)$$

where $\mu = \mathbb{E}(Y|\mathbf{x})$

A generalized linear regression model (GLM) extends this to

$$g(\mu) = \beta^T \mathbf{x} + \alpha \quad y \sim \mathcal{E}\mathcal{F}(\mu, \phi)$$

- $\mathcal{E}\mathcal{F}(\mu, \phi)$ is any exponential family distribution (e.g. Normal, Poisson, Binomial)
- $g(\cdot)$ is any smooth monotonic link function
- $\beta^T \mathbf{x} + \alpha (= \eta)$ is the linear predictor
In classical **GLM** $Y$ belongs to an exponential family with **PMF**

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

The **GLM** relates $x \in \mathbb{R}^p$ to the mean $\mu = \mathbb{E}(Y|x)$ by

$$g(\mu) = \eta = \alpha + \beta^T x$$

The **GLM** for the matrix predictor $X$ given by

$$g(\mu) = \eta = \alpha + \gamma^T z + \beta_1^T X \beta_2$$
The GLM with the systematic part for tensor predictor given by

\[ g(\mu) = \eta = \alpha + \gamma^T z + \langle B, X \rangle \]

- D-dimensional tensor predictor \( X \in \mathbb{R}^{p_1 \times \cdots \times p_D} \)
- D-dimensional coefficient tensor \( B \in \mathbb{R}^{p_1 \times \cdots \times p_D} \)
- \( B \) has \( \prod_{d=1}^{D} p_d \) parameters, which is ultrahigh dimensional and far exceeds sample size
Generalized Linear CP Tensor Regression

- Univariate outcome $Y$ belongs to exponential family
- Tensor covariate $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$
- Assume coefficient tensor $\mathcal{B}$ has a rank-$R$ decomposition $[\mathcal{B}_1, \ldots, \mathcal{B}_D]$ where $\mathcal{B}_d \in \mathbb{R}^{p_d \times R}$

**Generalized linear CP tensor regression model** (Zhou et al. 2013) with the systematic part given by

$$g(\mu) = \eta = \alpha + \gamma^T z + \langle \sum_{r=1}^R \beta_1^{(r)} \circ \cdots \circ \beta_D^{(r)}, \mathcal{X} \rangle$$

$$= \alpha + \gamma^T z + \langle (\mathcal{B}_D \circ \cdots \circ \mathcal{B}_1) \mathbf{1}_R, \text{vec}(\mathcal{X}) \rangle$$
Generalized linear CP tensor regression model given by

\[ g(\mu) = \eta = \alpha + \gamma^T z + \langle (B_D \circ \cdots \circ B_1) 1_R, \text{vec}(X) \rangle \]

- substantial reduction in dimensionality to the scale of \( R \times \sum_{d=1}^{D} p_d \)

  e.g. For a 128-by-128-by-128 MRI image, the dimensionality reduce from 2,097,157 to 1,157 using rank-3 decomposition

- Zhou et al. (2013) showed that this low rank tensor model could provide a sound recovery of many low rank signals
Given $n$ iid data $\{(y_i, x_i, z_i), i = 1, ..., n\}$ the log-likelihood

$$\ell(\alpha, \gamma, B_1, ..., B_D) = \sum_{i=1}^{n} \frac{y_i \theta - b(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi)$$

find the parameters $(\alpha, \gamma, B_1, ..., B_D)$ that maximizes this function.
Generalized linear CP tensor regression model given by

\[ g(\mu) = \eta = \alpha + \gamma^T z + \langle (B_D \odot \cdots \odot B_1)1_R, \text{vec}(X) \rangle \]

A key observation is although \( g(\mu) \) is not linear in \((B_1, \ldots, B_D)\) jointly, it is linear in each \( B_d \) separately.

When updating \( B_d \in \mathbb{R}^{p_d \times R} \), the inner product part can be written as

\[ \langle B_d, X_{(d)}(B_D \odot \cdots \odot B_{d+1} \odot B_{d-1} \odot \cdots \odot B_1) \rangle \]

this yields the block relaxation algorithm, which converges to a stationary point.
Maximize a regularized log-likelihood function

$$\ell(\alpha, \gamma, B_1, \ldots, B_D) = \sum_{d=1}^{D} \sum_{r=1}^{R} \sum_{i=1}^{p_d} P_\lambda(|\beta_{di}^{(r)}|, \rho)$$

- scalar penalty function $P_\lambda(|\beta|, \rho)$
- **power family** $P_\lambda(|x|, \rho) = \rho|\beta|^\lambda$, $\lambda \in (0, 2]$
- in particular lasso ($\lambda = 1$)
ADHD-200 Data Results

[taken from (Zhou et al. 2013)]
Future Work Plan

- Extending the linear CP/Tucker tensor regression model to the linear $\mathcal{H}$-Tucker tensor regression model
  - like CP model, the number of parameters is free from exponential dependence on $D$
  - preserve the flexibility of Tucker model

- Comparing the performance of different tensor regression models (Tucker, $\mathcal{H}$-Tucker) when applying different regularization approaches (sparsity regularization, trace norm regularization)
Future Work Plan con’t

- Finding the appropriate model and algorithm to address the multi-block tensor regression problems

- Combing the kernel concept and partial least squares (PLS) techniques to deal with tensor (multi-block tensor) regression problem

- Applying tensor regression approaches listed above to the applications such as neuroimage data analysis, brain signal data analysis to test if the improved performance can be achieved