Generalized tensor response regression with multi-sided covariates

Anonymous Author 1
Unknown Institution 1

Anonymous Author 2
Unknown Institution 2

Anonymous Author 3
Unknown Institution 3

Abstract

We consider the problem of learning higher-order tensor with side information on a set of modes. Such data problems arise frequently arise in applications such as neuroimaging, network analysis, and ... We propose a new family of tensor response regression models that incorporate covariate information...

1 Introduction

Many contemporary scientific and engineering studies collect multi-way array data, a.k.a. tensor, accompanied by additional covariates. For example, in neuro-imaging analysis, researchers measure brain connections from a sample of individuals with the goal to identify the brain edges affected by age and gender. In social network analysis, how to explain the connection (e.g. community, transitive, etc.) by attributable of both nodes. ... (add two pictures; one for estimating network population; another for estimating link prediction)

In this article, we provide a general treatment to these seemingly different problems.

Comparison with other models
Our model is related to, but fundamentally different from, several lines of existing work.
Unsupervised tensor. supervised learning.
Tensor-predictor regression multilinear in coefficients.
Tensor-response regression multilinear in coefficient vs. multilinear in covariates.
Generalized linear model. In the high-dimensions both p and n increase while $p \leq d$. This is the case we consider. Classical GLM fixes p. Compared to Gaussian model, the log-likelihood is not strictly convex in the linear predictor. We allow various types of dependent variable.

Preliminary work. Under review by AISTATS 2020. Do not distribute.

2 Preliminaries

We begin by reviewing a few basic factors about tensors [?]. We use $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_K}$ to denote an order-K $(d_1,\ldots,d_K)$-dimensional tensor. The multilinear multiplication of a tensor $\mathcal{Y} \in \mathbb{R}^{d_1\times\cdots\times d_K}$ by matrices $\boldsymbol{M}_k = [\![m^{(k)}_{i_k,j_k}]\!] \in \mathbb{R}^{s_k\times d_k}$ is defined as

$$\mathcal{Y}\times_1\boldsymbol{M}_1\ldots\times_K\boldsymbol{M}_K = [\![\sum_{i_1,\ldots,i_K} y_{i_1,\ldots,i_K} m^{(1)}_{i_1,j_1}\ldots m^{(K)}_{i_K,j_K}]\!],$$

which results in an order-K tensor $(s_1,\ldots,s_K)$-dimensional tensor. For ease of notatio, we also write the above Tucker product $\mathcal{Y}\times\{\boldsymbol{M}_1,\ldots,\boldsymbol{M}_K\}$ for short. For any two tensors $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!]$, $\mathcal{Y}' = [\![y'_{i_1,\ldots,i_K}]\!]$ of identical order and dimensions, their inner product is defined as $\langle\mathcal{Y},\mathcal{Y}'\rangle = \sum_{i_1,\ldots,i_K} y_{i_1,\ldots,i_K} y'_{i_1,\ldots,i_K}$. The Frobenius norm of tensor $\mathcal{Y}$ is defined as $\|\mathcal{Y}\|_F = \langle\mathcal{Y},\mathcal{Y}\rangle^{1/2}$; it is the Euclidean norm of $\mathcal{Y}$ regarded as an $\prod_k d_k$-dimensional vector. We use lower-case letters $(a,b,c,\ldots)$ for scalars and vectors, upper-case boldface letters $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\ldots)$ for matrices, and calligraphy letter $(\mathcal{A},\mathcal{B},\mathcal{C},\ldots)$ for tensors of order 3 or greater.

3 Motivation and model

Let $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_K}$ denote an order-K data tensor of interest. In addition, suppose we observe covariates on a subset of modes. Let $\boldsymbol{X}_k \in \mathbb{R}^{d_k\times p_k}$ be the available covariates on the mode-k, where $p_k \leq d_k$. We propose the following multilinear structure in the mean of the tensor. Specifically,

$$\begin{aligned}&\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K) = f(\Theta),\ \text{where}\\ &\Theta = \mathcal{B}\times\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K\},\end{aligned} \tag{1}$$

where $f(\cdot)$ is a known link function, $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is called the linear predictor tensor, $\mathcal{B} \in \mathbb{R}^{p_1\times\cdots p_K}$ is the parameter tensor of interest, and $\times$ denotes the tensor Tucker product. The link function depends on the distribution family of the response. Some common choices are identity link for Gaussian tensor, logistic link for binary tensor, and log link for Poisson tensor. We give three examples of multi-covariates tensor regression model that arises in practice.

Example 1 (Spatio-temporal growth model)**.** Let $\mathcal{Y} = [\![y_{ijk}]\!] \in \mathbb{R}^{d\times m\times n}$ denote the pH measurements of d lakes at m levels of depth and for n time points. Suppose the sampled lakes belong to q types, with p lakes in each type. Let $\{\ell_j\}_{j\in[m]}$ denote the sampled depth levels and $\{t_k\}_{k\in[n]}$ the time points. Assume the expected pH trend in depth is a polynomial of order r and that the expected trend in time is a polynomial of order s. Then, a classical spatio-temporal growth model can be represented as

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) = \mathcal{B} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\},$$

where $\mathcal{B} \in \mathbb{R}^{p\times(r+1)\times(s+1)}$ is the coefficient tensor of interest, $\boldsymbol{X}_1 = \text{blockdiag}\{\boldsymbol{1}_p, \ldots, \boldsymbol{1}_p\} \in \{0,1\}^{d\times p}$ is the design matrix for lake types,

$$\boldsymbol{X}_2 = \begin{pmatrix} 1 & \ell_1 & \cdots & \ell_1^r \\ 1 & \ell_2 & \cdots & \ell_2^r \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \ell_m & \cdots & \ell_m^r \end{pmatrix}, \ \boldsymbol{X}_3 = \begin{pmatrix} 1 & t_1 & \cdots & t_1^s \\ 1 & t_2 & \cdots & t_2^s \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & \cdots & t_n^s \end{pmatrix}$$

are the design matrices for spatial and temporal effects, respectively.

Example 2 (Network population model)**.** Network response model is recently developed in the context of neuroimanig analysis. The goal is to study the relationship between the network-valued response with the individual covariates. Suppose we observe n i.i.d. observations $\{(\boldsymbol{Y}_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$, where $\boldsymbol{Y}_i \in \{0,1\}^{d\times n}$ is the brain connectivity network on the i-th individual and $\boldsymbol{x}_i \in \mathbb{R}^p$ is the subject covariate such as age, gender. The network-response model has the form

$$\text{logit}(\mathbb{E}(\boldsymbol{Y}_i|\boldsymbol{x}_i)) = \mathcal{B} \times_3 \boldsymbol{x}_i, \quad \text{for } i = 1, \ldots, n \qquad (2)$$

where $\mathcal{B} \in \mathbb{R}^{d\times d\times p}$ is the coefficient tensor of interest. In fact, the model (2) is a special case of our multilinear tensor-response model. To see this, let $\mathcal{Y} \in \{0,1\}^{d\times d\times n}$ denote the response tensor by stacking $\{\boldsymbol{Y}_i\}$ together along the 3$^{\text{rd}}$ mode and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{p\times n}$, then model (2) can be expressed as

$$\text{logit}(\mathbb{E}(\mathcal{Y}|\boldsymbol{X})) = \mathcal{B} \times_3 \boldsymbol{X} = \mathcal{B} \times \{\boldsymbol{I}_d, \boldsymbol{I}_d, \boldsymbol{X}\},$$

where $\boldsymbol{I}_d$ denotes the identity matrix of dimension d.

Example 3 (Link model with node attributes)**.** Let $V = [n]$ be a set of vertices and explanatory variable $x_i \in \mathbb{R}^p$ associated to each $i \in V$. The network $G = (V, E)$ is described by the following matrix model. The edge connects the two vertices i and j independently of the others. The probability of connection is modeled as

$$\text{logit}(\mathbb{P}((i,j) \in E) = \boldsymbol{x}_i^T \boldsymbol{B} \boldsymbol{x}_j = \langle \boldsymbol{B}, \boldsymbol{x}_i^T \boldsymbol{x}_j \rangle.$$

Again, we show that this model is a special case of our tensor regression model. Let $\mathcal{Y} = [\![y_{ij}]\!]$ where $y_{ij} = \mathbb{1}_{(i,j)\in E}$. Define $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{p\times n}$. Then the above model can be expressed as

$$\text{logit}(\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X})) = \mathcal{B} \times \{\boldsymbol{X}, \boldsymbol{X}\}.$$

In the above three example and many other studies, researchers are interested in uncovering the variation in the data tensor that are explained by the covariates.

Without any structure on the coefficient tensor $\mathcal{B}$: *A naive approach is to regress the tensor entry, one at a time, on the covariates, and this model is repeatedly fitted for each tensor element.* Though this approach is scalable, it suffers from two drawbacks: (1) ignore the multilinear structure in the tensor (?) and (2) suffers from the multiplicity issue. To allow the structure among ..we further impose a multilinear low-rank structure on the coefficient tensor $\mathcal{B}$

$$\mathcal{P} = \{\mathcal{B} \in \mathbb{R}^{p_1\times\cdots\times p_K} : r_k(\mathcal{B}) \leq r_k \text{ for } k \in [K]\}, \quad (3)$$

where $r_k(\mathcal{B})$ is the Tucker rank of the tensor at mode k. We assume that r_k is known and $r_k \leq p_k$. Other low-rankness such as CP rank is also possible. We choose the Tucker decomposition due to the following observation.

The low-rank structure in (3) implies that the coefficient tensor can be expressed as $\mathcal{B} = \mathcal{C} \times_1 \boldsymbol{M}_1 \times \cdots \boldsymbol{M}_K$. Then, our tensor regression model (1) is equivalent to

$$f(\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)) = \mathcal{C} \times \{\boldsymbol{X}_1\boldsymbol{M}_1, \ldots, \boldsymbol{X}_k\boldsymbol{M}_k\}.$$

The goal is to find a joint dimension reduction of $\mathcal{Y}$ and $\boldsymbol{X}_K$ such that the unexplained variation in the mean tensor. The factorization is restricted to the space spanned by $\boldsymbol{X}_k$. Here $\boldsymbol{X}_1\boldsymbol{M}_1$ can be interpreted as the latent covariates that explains the variation in the response tensor. The core tensor $\mathcal{C}$ collects the interaction effects of latent covariates across the K modes.

(Question: the columns of $\boldsymbol{X}$ should be normalized??)

Question: any connection to tensor completion?? If $\boldsymbol{X}_K$ is a random design matrix?

4 Rank-constrained likelihood-based estimation

We develop a likelihood-based estimation procedure. The exponential family is a flexible framework for different data types. In a classical glm with a scalar response y and covariate $\boldsymbol{x}$, the density is

$$p(y|\boldsymbol{x}, \boldsymbol{\beta}) = c(y) \exp\left(\frac{y\theta - b(\theta)}{\phi}\right) \text{ with } \theta = \boldsymbol{\beta}^T\boldsymbol{x},$$

where $c(\cdot)$ and $b(\cdot)$ are known functions and θ is the linear predictor. Note that the canonical link function

f is chosen to be $f(\cdot) = b'(\cdot)$. Table 1 summarize the canonical link function for common distributions.

In our context, the log-likelihood of (1) is the (?) divergence between the conditional distribution of $\mathcal{Y}|\Theta$ and the exponential family

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) = \langle \mathcal{Y}, \Theta \rangle - \sum_{i_1,\ldots,i_K} b(\theta_{i_1,\ldots,i_K}),$$
$$\text{where } \Theta = \mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}$$

Assume that we have an additional information on an upper bound $a > 0$ such that $\|\Theta\|_\infty \leq \alpha$. (more comments? ill-condition in the bernoulli model?) We propose the following constrained maximum likelihood estimation for the tensor coefficient

$$\hat{\mathcal{B}} = \underset{\text{rank}(\mathcal{B}) \leq \boldsymbol{r}, \|\Theta(\mathcal{B})\|_\infty \leq \alpha}{\arg\max} \mathcal{L}_{\mathcal{Y}}(\mathcal{B}).$$

4.1 Statistical properties

We assess the estimation accuracy using the deviation in the Frobenius norm. For the true coefficient tensor $\mathcal{B}_{\text{true}}$ and its estimator $\hat{\mathcal{B}}$, define

$$\text{Loss}(\mathcal{B}_{\text{true}}, \hat{\mathcal{B}}) = \frac{1}{\prod_k d_k} \|\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}\|_F^2$$

We focus on the high-dimensional region in which both $d_k \to \infty$ and $p_k \to \infty$ while $\frac{p_k}{d_k} \to \gamma_k \in [0, 1]$.

Assumption 1. *We make the following assumptions:*

1. *There exists two positive constants $c_1, c_2 > 0$ such that $c_1 \leq \sigma_{\min}(\boldsymbol{X}_k) \leq \sigma_{\max}(\boldsymbol{X}_k) < c_2$ for all $k \in [K]$.*

2. *There exist two positive constants L, M such that $L \leq Var(y_{i_1,\ldots,i_K}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) \leq M$ uniformly over the parameter space $\mathcal{P}$.*

2'. *Equivalently, $L \leq b''(x) \leq M$ for all $x \leq \alpha$, where $b(\cdot)$ is the known function in the exponential family distribution and α is the upper bound of the linear predictor.*

Theorem 4.1 (Convergence)**.** *Consider a generalized tensor regression model with multi-sided covariates. Let $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ be the tensor response and $\boldsymbol{X}_k \in \mathbb{R}^{p_k \times d_k}$ the covariates on mode-k. Suppose $\mathcal{Y}|\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}$ follows exponential family distribution generated from the low-rank tensor regression model* (1)*. Furthermore, the Assumption* (1) *holds. Then there exist two constants $C_1, C_2 > 0$ such that, with probability at least $1 - \exp(-C_1 \sum_k p_k)$,*

$$Loss(\hat{\mathcal{B}}, \mathcal{B}_{true}) \leq \frac{C_2 \sum_k p_k}{\prod_k d_k},$$

where $C_2 = C_2(\alpha, K, r_1, \ldots, r_K) > 0$ is a constant that does not depend on dimension $\{d_k\}$ and $\{p_k\}$.

To gain further insight on the bound we consider the special case when $d_1 = d_2 = \ldots = d_K = d$. 1. binary case; 2. large dimension region. $\mathcal{O}\left(\frac{p}{d^k}\right) \leq \mathcal{O}(d^{-(k-1)})$.

Corollary 1 (Spatio-temporal growth model)**.** Our method yields the convergence rate $\mathcal{O}\left(\frac{p+r+s}{dmn}\right)$. Note that $p \leq d$, $r \leq m$ and $s \leq m$, so consistent estimator.

Corollary 2 (Network population model)**.** Our method yields the convergence rate $\mathcal{O}\left(\frac{2d+p}{d^2 n}\right)$. Note that $p \leq m$, so this is a consistent estimator. In contrast, a naive repeated glm will give $\mathcal{O}(\frac{p}{n})$.

Corollary 3 (Link model with node attributes)**.** Our method yields the convergence rate is $\mathcal{O}\left(\frac{p}{d^2}\right)$. Note that $p \leq m$, so again a consistent estimator. In contrast, a naive repeated glm will give $\mathcal{O}(\frac{p}{n})$.

We provide the prediction error for $\mathbb{E}(\mathcal{Y})$ (for Bernoulli model only?)

Theorem 4.2.

4.1.1 Phase transition for logistic model

Third level headings are flush left, initial caps, bold, and in point size 10. Use one line space before the third level heading and one-half line space after the third level heading.

5 Numerical implementation

5.1 Alternating optimization

5.2 Tuning parameter selection

6 SUPPLEMENTARY MATERIAL

If you need to include additional appendices during submission, you can include them in the supplementary material file.

7 INSTRUCTIONS FOR CAMERA-READY PAPERS

For the camera-ready paper, if you are using LaTeX, please make sure that you follow these instructions. (If you are not using LaTeX, please make sure to achieve the same effect using your chosen typesetting package.)

1. Download `fancyhdr.sty` – the `aistats2020.sty` file will make use of it.

2. Begin your document with

```
\documentclass[twoside]{article}
\usepackage[accepted]{aistats2020}
```

Algorithm 1 Unsupervised binary tensor decomposition

Input: Y, the shape of core tensor r_1, r_2, r_3; the maximum iteration time N; the link function f; significant increment criterion ϵ

Output: The estimate of factor matrices $\hat{M}_1, \hat{M}_2, \hat{M}_3$ and core tensor $\hat{\mathcal{C}}$.

Use Tucker decomposition to get the initial $\mathcal{C}^{(0)}, M_1^{(0)}, M_2^{(0)}, M_3^{(0)}$; Calculate the initial log-likelihood value $l^{(0)}$; Set the iteration time $t = 0$

while The increment of log-likelihood $l^{(t)} - l^{(t-1)} \geq \epsilon$ and $t \leq N$ **or** $t = 0$ **do**

  (a) Upgrade the iteration index $t = t + 1$;

  (b) Upgrade $\mathcal{C}^{(t-1)} \leftarrow \mathcal{C}^{(t)}$ by solveing one *glm* model with $r_1 \times r_2 \times r_3$ coefficients and link function f. The responses and predictors are formed by Y and $M_1^{(t-1)}, M_2^{(t-1)}, M_3^{(t-1)}$.

  for i = 1,2,3 **do**

    Upgrade $M_i^{(t-1)} \leftarrow M_i^{(t)}$ by solving d_i *glm* models. The responses and predictors are formed by Y and $M_j^{(t)}, M_l^{(t-1)}, j < i, l > i$.

  (c) Calculate the log-likelihood value $l^{(t)}$ with $\mathcal{C}^{(t)}, M_1^{(t)}, M_2^{(t)}, M_3^{(t)}$

  (d) After each upgrading step, orthogonalize M_is through *tucker* decomposition,

The `twoside` option for the class article allows the package `fancyhdr.sty` to include headings for even and odd numbered pages. The option `accepted` for the package `aistats2020.sty` will write a copyright notice at the end of the first column of the first page. This option will also print headings for the paper. For the *even* pages, the title of the paper will be used as heading and for *odd* pages the author names will be used as heading. If the title of the paper is too long or the number of authors is too large, the style will print a warning message as heading. If this happens additional commands can be used to place as headings shorter versions of the title and the author names. This is explained in the next point.

3. If you get warning messages as described above, then immediately after `\begin{document}`, write

```
\runningtitle{Provide here an
alternative shorter version of the title
of your paper}
\runningauthor{Provide here the surnames
of the authors of your paper, all
separated by commas}
```

Note that the text that appears as argument in `\runningtitle` will be printed as a heading in the *even* pages. The text that appears as argument in `\runningauthor` will be printed as a heading in the *odd* pages. If even the author surnames do not fit, it is acceptable to give a subset of author names followed by "et al."

4. Use the file sample_paper.tex as an example.

5. The camera-ready versions of the accepted papers are 8 pages, plus any additional pages needed for references.

6. If you need to include additional appendices, you can include them in the supplementary material file.

7. Please, don't change the layout given by the above instructions and by the style file.

Acknowledgements

Use the unnumbered third level heading for the acknowledgements. All acknowledgements go at the end of the paper.

References

References follow the acknowledgements. Use an unnumbered third level heading for the references section. Any choice of citation style is acceptable as long as you are consistent. Please use the same font size for references as for the body of the paper—remember that references do not count against your page length total.

References

J. Alspector, B. Gupta, and R. B. Allen (1989). Performance of a stochastic learning microchip. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 1*, 748–760. San Mateo, Calif.: Morgan Kaufmann.

F. Rosenblatt (1962). *Principles of Neurodynamics.* Washington, D.C.: Spartan Books.

G. Tesauro (1989). Neurogammon wins computer Olympiad. *Neural Computation* **1**(3):321–323.