Statistical learning methods based on tensor data

1 Research goals

datum

Rapid developments in modern technologies have made large-scale multidimensional data available in science and engineering. Tensors, or multidimensional arrays, provide a generalized data. Analyzing tensor data with increasing dimensionality and complexity requires the development of novel statistical methods. A naive approach for tensor data analysis is to transform tensors to matrices or vectors and apply classical methods. However, this transformation would destroy the structural information of the data tensors. Furthermore, the reshaping tensors to matrices ~~or vectors~~ results in high-dimensionality which leads to overfitting. Tensors are not simply vectors with more indices; rather they are mathematical objects possessing structural information. ~~I believe that~~ tensor data analysis preserving structural information will boost new scientific discoveries, ~~which can not be revealed from unstructured data.~~ To this end, I have established three main statistical learning methods based on tensor data in different contexts and developed theoretical guarantees.

Learning with non-Guassian tensors

2 Aim 1: Tensor denoising and completion based on ordinal observation

data tensor

We explore the problem of tensor denosing and completion based on ordinal measurements. Such tensor data arises in several applications such as recommendation system, social networks, and neuroimaging. One example is the Netflix problem, which records the ratings of users on movies over time. Each data entry is a rating on a nominal scale {*very like, like, neutral, dislike, very dislike*}. This data can be naturally described as a three-way tensor of users × movies × time and each entry indicates the ordinal values, ~~say,~~ $\{1, 2, 3, 4, 5\}$. The effective ordinal model satisfies two key properties. First, the model should be invariant under a reversal of categories, ~~say,~~ from the Netflix example, the labeling "*very like* $\succ$ *like* $\succ \cdots$" and "*very like* $\prec$ *like* $\prec \cdots$" should have the same result. Second, the parameter interpretation should be consistent under merging or splitting of contiguous categories. The classical continuous tensor model (Kolda and Bader, 2009; Ghadermarzy et al., 2019) and the binary model (Ghadermarzy et al., 2018) lack the properties. ~~We~~ propose an appropriate model for ordinal tensors which satisfies the properties and achieves theoretical guarantees. Based on the proposed model, we can estimate missing entries of the ordinal tensor (completion) and obtain good interpretation of the ordinal tensor data through examining the latent tensor (denoising).

a.k.a. completion problem

We —> I

Model formulation and estimation

Let $\mathcal{Y} = [\![y_\omega]\!] \in \{1, \ldots, L\}^{d_1 \times \cdots \times d_K}$ denote ordinal tensor observation where L is an ordinal level. We propose a cumulative link model that can explain the ordinal tensor observation

$$\mathbb{P}(y_\omega \leq \ell) = f(b_\ell - \theta_\omega), \text{ for all } \ell \in \{1, \ldots, L-1\}, \tag{1}$$

where $\boldsymbol{b} = (b_1, \ldots, b_{L-1})$ is a set of unknown scalars satisfying $b_1 < \cdots < b_{L-1}, \Theta = [\![\theta_\omega]\!] \in \mathbb{R}^{d_1 \times \cdots d_K}$ is a continuous-valued tensor, and $f(\cdot) \colon \mathbb{R} \mapsto [0, 1]$ is a known, strictly increasing function. We refer to Θ as the signal tensor, $\boldsymbol{b}$ the cut-off points, and f the link function. Furthermore, we assume a low-rank structure on Θ:

$$\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_k \boldsymbol{M}_K, \tag{2}$$

where where $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is a core tensor , $\boldsymbol{M}_k \in \mathbb{R}^{d_k \times r_k}$ orthonormal factor matrices, and $\times_k$ denotes the tensor-by-matrix multiplication. The low-rank structure is a commonly used dimension reduction tool in tensor data analysis. The proposed ordinal tensor model can be regarded as an entry-wise quantization of a noisy continuous-valued low-rank tensor. Figure 1 shows how observed ordinal tensor is generated from our model.

Estimation of the signal tensor Θ can give us interpretability and makes it possible to predict missing entries of the ordinal tensor. We propose a rank-constrained maximum likelihood estimator (MLE) for the signal

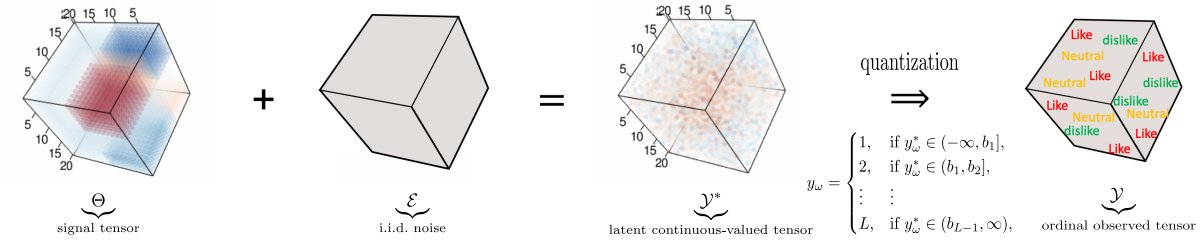

Figure 1: Schematic diagram for the cumulative link model

tensor Θ. The optimization is not jointly convex, but it is convex in each factor matrix individually with all other factor matrices fixed. This feature enables a block relaxation type minimization.

Theoretical results

We study the convergence property of the estimated signal tensor $\hat{\Theta}$. The classical asymptotic analysis does not apply straightforwardly because the number of parameters increases with the sample size. The following theorem shows that the convergence rate of our estimation.

Theorem 2.1 (Statistical convergence)**.** *Consider an ordinal tensor $\mathcal{Y} \in [L]^{d_1 \times \cdots \times d_K}$ generated from model (1), with the link function f and the true coefficient tensor $\Theta^{\text{true}} \in \mathcal{P}$. Define $r_{\max} = \max_k r_k$. Then, with very high probability, the MLE estimator satisfies*

$$\text{MSE}(\hat{\Theta}, \Theta^{\text{true}}) \le c_1 r_{\max}^{K-1} \frac{\sum_k d_k}{\prod_k d_k},$$

where $\text{MSE}(\hat{\Theta}, \Theta^{\text{true}}) \overset{def}{=} \frac{1}{\prod_k d_k} \|\hat{\Theta} - \Theta^{\text{true}}\|_F^2$ *and $c_1 > 0$ is a constant that depends only on K.*

Too technical.
1. idea that can be generalized. (take-away message)
2. last paragraph of each section:
pose open questions, not to draw conclusion

Theorem 2.1 establishes the statistical convergence for the estimator. In fact, the proof of this theorem shows that the same statistical rate holds, not only for the global optimizer, but also for any local optimizer as long as the convergent objective is large enough. In addition, we showed that the worst case MSE is always greater than the polynomial bound we have in Theorem 2.1. In this sense, our estimation is rate optimal.

3 Aim 2: Supervised tensor decomposition with interactive side information

Tensors are often collected with side information on multiple modes in modern scientific and engineering studies. A popular example can be found in neuroimaging (Zhou et al., 2013). The brain connectivity networks are collected from a sample of individuals, accompanied by individual characteristics such as age, gender, and diseases status (Figure 2a). Identifying the relationship between a high-dimensional tensor and side information is important yet challenging. One of the challenges comes from non Gaussian measurements. Many tensor datasets consist of non Gaussian entries. One example is the brain connectivity network dataset (Zhang et al., 2018) which is a collection of binary adjacency matrices. Many methods have been proposed but there is a gap between theory and practice. Classical tensor decomposition methods are based on minimizing the Frobenius norm of the reconstruction error resulting in suboptimal predictions for binary- or count-valued measurement. Many other supervised tensor methods have been proposed (Narita et al., 2012; Yu and Liu, 2016) to address tensor regression problems. These methods often assume Gaussian distribution on the tensor entries, or impose random design on feature matrices, both of which are less suitable for applications of our interests. We present a general model and associated methods for decomposing a data tensor whose entries are from exponential family with interactive side information. I believe that this model fills the gap between theory and practice and becomes powerful tools to boost scientific discoveries.

The first sentence of each subsection, section & last paragraph in each section.
==> I propose …. My proposal ….

The proposed project focuses on developing a general model for …. with multiple side information.
I will formulate the learning task as a structured tensor regression. Let ….

Model formulation and estimation

Let $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote the tensor observation and $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$ the side information encoded as multiple feature matrices. We propose a supervised tensor decomposition

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}, \ldots, \boldsymbol{X}_K) = f(\mathcal{C} \times_1 \boldsymbol{X}_1\boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{X}_k\boldsymbol{M}_K), \tag{3}$$

where $f(\cdot)$ is the known link function depending on the data type, $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is the core tensor, and $\boldsymbol{M}_i \in \mathbb{R}^{p_i \times r_i}$ are factor matrices with orthonormal columns. Figure 2b provides a schematic illustration of the model. The features $\boldsymbol{X}_k$ affect the distribution of tensor entries in $\mathcal{Y}$ through the form $\boldsymbol{X}_k\boldsymbol{M}_k$, which are r_k linear combinations of features on mode k. $\boldsymbol{X}_k\boldsymbol{M}_k$ can be viewed as sufficient features. The core tensor $\mathcal{C}$ collects the interaction effects between sufficient features across K modes.

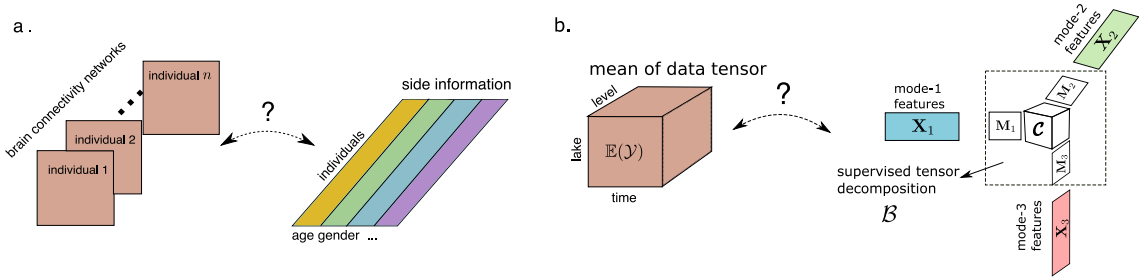

Figure 2: (a) Brain connectivity networks with individual side-information. (b) supervised tensor decomposition for an order-3 tensor with multiple side information.

Estimation of $\boldsymbol{M}_k$ and the corresponding $\mathcal{C}$ allows us to reveal the relationship between side information $\boldsymbol{X}_k$ and observed tensor $\mathcal{Y}$. We develop a likelihood-based procedure to estimate $\mathcal{C}$ and $\boldsymbol{M}_k$ in (3) (M-estimator). An alternating optimization is utilized to solve the rank constrained likelihood estimation.

Theoretical results

We provide the accuracy guarantee for the proposed M-estimator. We are interested in the high-dimensional region in which both d_k and p_k diverge; i.e. $d_k \to \infty$ and $p_k \to \infty$, while $p_k/d_k \to \gamma_k \in [0, 1)$. This setting is common in modern application when the tensor data and features are large-scale. The classical MLE theory cannot be applied directly. We leverage the recent development in random tensor theory and high-dimensional statistics to establish the error bound of the estimation.

Theorem 3.1. *Consider a data tensor generated from model (3). Let $(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_1, \ldots, \hat{\boldsymbol{M}}_K)$ be the M-estimator and $\hat{\mathcal{B}} = \hat{\mathcal{C}}_1 \times_1 \hat{\boldsymbol{M}}_1 \times_2 \cdots \times_K \hat{\boldsymbol{M}}_K$. Under mild assumptions on feature matrices $\boldsymbol{X}_k$ and the link function $f(\cdot)$, with high probability, the M-estimator satisfies*

$$\|\mathcal{B}_{true} - \hat{\mathcal{B}}\|_F^2 \leq \frac{c_2 \prod_k r_k \sum_k p_k}{\max_k r_k \prod_k d_k},$$

where $c_2 > 0$ is a constant independent of dimensions $\{d_k\}$ and $\{p_k\}$.

Consider a special case when tensor dimensions are equal at each of the modes, i.e., $d_k = d, p_k = \gamma d$, $\gamma \in [0, 1)$ for all $k \in [K]$. The result implies that the convergence rate of estimation is $\mathcal{O}\left(\frac{p}{d^K}\right)$. Therefore, our estimation is consistent as the tensor dimensions grows, and the convergence rate becomes favorable as the order of the tensor increases.

4 Aim 3: Nonparametric regression with matrix feature

We consider the problem of learning the relationship between a binary label response and high-dimensional matrix-valued predictors. Such data problems arise commonly in brain imaging studies, sensor network

localization, and personalized medicine. In Osteosarcoma treatment, for example, the degree of tumor necrosis is used to guide the choice of postoperative chemotherapy (Man et al., 2005). Patients with tumors, which reveal 90% necrosis ($Y = 1$), have a much better prognosis than those with 90% necrosis ($Y = 0$). The problem is that measuring Y involves an invasive biopsy. Suppose that $\boldsymbol{X}$ is a feature matrix collected from each individual, which saves information about gene expression levels on each tissue. The knowledge of the regression $\mathbb{P}(Y = 1|\boldsymbol{X})$ would allow for accurate treatment decision without biopsy. The classical logistic regression or linear discriminant analysis have been developed and popular for binary regression. However, these methods assume Gaussian distribution or specific parametric model, both of which are difficult to be justified in high dimensional setting such as matrix-valued feature space. We propose a flexible nonparametric framework for the binary regression that can consider the matrix structure of the predictors.

Framework

Suppose that we observe a sample of n data points, $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, identically and independently distributed according to a unknown distribution $\mathbb{P}(\boldsymbol{X}, Y)$ over $\mathbb{R}^{d_1 \times d_2} \times \{-1, 1\}$. We propose to estimate the regression function $\mathbb{P}(Y = 1|\boldsymbol{X}) \stackrel{\text{def}}{=} p(\boldsymbol{X})$ by two approximation steps for a given function class $\mathcal{F}$.

$$p(\boldsymbol{X}) \overset{\text{Step 1}}{\approx} \frac{1}{H} \sum_{h \in [H]} \mathbb{1}\left\{\boldsymbol{X} \colon p(\boldsymbol{X}) \leq \frac{h}{H}\right\}$$
$$\overset{\text{Step 2}}{\approx} \frac{1}{H} \sum_{h \in [H]} \mathbb{1}\left\{\boldsymbol{X} \colon \operatorname{sign}\left[\hat{f}_{\frac{h}{H}}(\boldsymbol{X})\right] = -1\right\},$$

where H is a smooth parameter and $\hat{f}_{\frac{h}{H}}$ is a weighted large margin classifier defined as

$$\hat{f}_{\frac{h}{H}} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1} \left[\left(1 - \frac{h}{H}\right) \sum_{y_i = 1} L(y_i f(\boldsymbol{X}_i)) + \frac{h}{H} \sum_{y_i = -1} L(y_i f(\boldsymbol{X}_i))\right] + \lambda J(f),$$

where $J(f)$ is a regularization term for model complexity and $L(z)$ is a margin loss that is a function of the functional margin $yf(\boldsymbol{X}_i)$. Step 1 discretizes the target function by level sets (Figure 3 visualizes the approximation). Step 2 estimates the level sets using a sequence of weighted classifiers. We can show that the error of Step 2 converges to 0 in probability when considered function class $\mathcal{F}$ is large enough. Through these two steps we estimate the regression function in the nonparametric way.

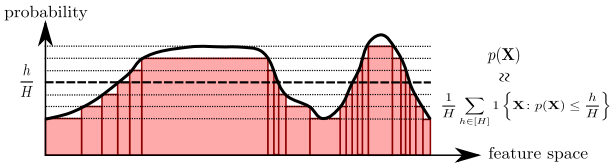

Figure 3: We can approximate the target function $p(\boldsymbol{X})$ by linear combinations of indicator functions of level sets.

One possibility is to choose ….

We choose the function class $\mathcal{F}$ that accounts for the structural information of feature matrices. One example is a linear function class with low rank structure on coefficient matrix.

$$\mathcal{F} = \{f \colon f(\cdot) = \langle \boldsymbol{U}\boldsymbol{V}^T, \cdot \rangle \text{ where } \boldsymbol{U} \in \mathbb{R}^{d_1 \times r}, \boldsymbol{V} \in \mathbb{R}^{d_2 \times r}\}.$$

In the linear case with $d = d_1 = d_2$, we guarantee that the estimated regression function converges to the true function in L_1 sense with the rate of $\mathcal{O}\left(\log(n/rd)/(n/rd)\right)$ under some assumptions on $p(\boldsymbol{X})$. We also develop nonlinear classifier for matrix predictors using a new family of matrix-input kernel.

5 Future directions

There are several future directions that we would like to push forward. First, algorithmic convergence of suggested algorithms can be constructed from some conditions on the loss functions. Empirical error consists

of algorithmic error and statistical error. Therefore, through constructing algorithmic convergence, we expect to guarantee empirical performance of our models in actual applications based on already constructed statistical convergences.

Second, the nonparametric regression can be utilized to solve matrix completion and denoising problems. Let $\boldsymbol{P} = [\![p_{ij}]\!] \in [0,1]$ denote the probability matrix and $\boldsymbol{Y} = [\![y_{ij}]\!] \in \{0,1\}$ the binary observation from the probability matrix $\boldsymbol{P}$ with the rule $y_{ij} \overset{\text{ind}}{\sim} \text{Ber}(p_{ij})$ for $(i,j) \in [d_1] \times [d_2]$. From the proposed nonparametric approach, we can estimate $\boldsymbol{P}$ from the observation $\boldsymbol{Y}$ ~~(denoising).~~ Furthermore, we can impute missing entries of $\boldsymbol{Y}$ based on estimated $\hat{\boldsymbol{P}}$ ~~(completion). We generate the training set $\{(\boldsymbol{X}_{ij}, y_{ij}) : (i,j) \in [d_1] \times [d_2]\}$ where $\boldsymbol{X}_{ij}$ is an indicator matrix with 1 at $(i,j) - th$ position and 0's everywhere. We apply the proposed regression method based on the training set. Our simulation shows that the~~ proposed method successfully and robustly estimate the true probability (Figure 4). The promising result suggests that the new nonparametric completion method is worthwhile to study. substantially outperforms previous approaches.

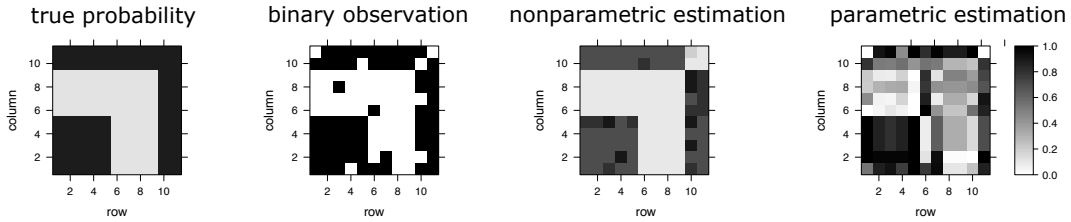

Figure 4: Probability matrix estimation result: our nonparametric approach outperforms parametric estimation based on logistic model.

Lastly, we can extend the nonparametric regression framework for matrix-valued predictors to tensor-valued one. Since tensors are not simple matrices with more indices, we need to develop a new method that can handle tensor predictors. One possible way is to impose Tucker low-rankness (2) on a tensor coefficient of a linear function. This new structure will allow us to examine structural information of tensor. ~~In addition,~~ this extension suggests a new nonparametric denoising and completion way addressed in Section 2 (Aim 1) based on the early mentioned connection.

References

Ghadermarzy, N., Plan, Y., and Yilmaz, O. (2018). Learning tensors from partial binary measurements. *IEEE Transactions on Signal Processing*, 67(1):29–40.

Ghadermarzy, N., Plan, Y., and Yilmaz, Ö. (2019). Near-optimal sample complexity for convex tensor completion. *Information and Inference: A Journal of the IMA*, 8(3):577–619.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Man, T.-K., Chintagumpala, M., Visvanathan, J., Shen, J., Perlaky, L., Hicks, J., Johnson, M., Davino, N., Murray, J., Helman, L., et al. (2005). Expression profiles of osteosarcoma that can predict response to chemotherapy. *Cancer research*, 65(18):8142–8150.

Narita, A., Hayashi, K., Tomioka, R., and Kashima, H. (2012). Tensor factorization using auxiliary information. *Data Mining and Knowledge Discovery*, 25(2):298–324.

Yu, R. and Liu, Y. (2016). Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning*, pages 373–381.

Zhang, J., Sun, W. W., and Li, L. (2018). Network response regression for modeling population of networks with covariates. *arXiv preprint arXiv:1810.03192*.

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