ICML Paper Draft

Chanwoo Lee

Jan 12, 2020

1 Introduction

Multidimensional array data, a.k.a. a tensor, appears in a huge variety of applications including recommendation systems (Kutty, Chen, and Nayak 2012; Adomavicius et al. 2008; W. W. Sun et al. 2015), social networks (J. Sun et al. 2009; Nickel, Tresp, and Kriegel 2011), genomics (Wang, Fischer, and Song 2017), neuroimaging (EEG, fMRI) (Miwakeichi et al. n.d.) and signalprocessing (Sidiropoulos, Bro, and Giannakis 2000; Cichocki et al. 2015). Instead of unfolding those data tensors into matrices where many analysis methods have been proposed, we preserve its inherent multi-modal structure. Studying tensor data while respecting the structure allows us to examine complex interactions among tensor entries. Thereby we can provide extra more interpretation that cannot be addressed by traditional matrix analysis. Also, It has been shown that the tensor preserving analysis improves performance (Zare et al. 2018; Wang and Li 2018). With those reasons, there is a growing need to develop dimension reduction method without losing multi modal structure. In the line of the attempts, a number of tensor decomposition methods have been proposed in many applications. CANDECOMP/PARAFAC (CP) decomposition was first introduced (Hitchcock 1928) and revitalized in psychometrics (Harshman 1970) and in linguistics (Smilde, Bro, and Geladi 2004). The Tucker decomposition was proposed in psychometrics (Tucker 1964; Tucker 1966).

Classical tensor completion with those decompositions has treated the entries of data as real-valued. In many cases, However, we encounter data of which the entries are not real-valued but discrete or quantized i.e. binary-valued or multiple-valued. For example, many survey data takes the integer values. To be specific, the data in the Netflix problem has the 3 modes of 'user','movie' and 'date of grade'. The entries of the data are the gradings from the users which take integer value from 1 to 5. Also, there are many cases that the data are quantized in real application. In signal processing, the data are frequently rounded or truncated so that only integer values are available. Also, in graph theory, adjacency matrix can be labeled as from 1 to 3 taking 3 when pairs of vertexes have strong connection and giving 1 when two vertexes have the weak connection according to a given threshold. If we add one more mode on adjacency matrix such as 'context' or 'individual', the data turns into tensor data with 3 integer values.

Therefore, performance improvement can be achieved when the observations are treated as discrete value not as continuous value. In matrix case, there has been many achievements to complete matrix for discrete cases: Models for the case of binary or 1-bit were introduced and studied (Davenport et al. 2012; Bhaskar and Javanmard 2015). Furthermore, Bhaskar (2016) suggested matrix completion method for general ordinal observations. In tensor case, however, only binary tensor has gotten an attention and achieved performance improvement using binary tensor decomposition methods (Hore et al. 2016; Wang and Li 2018; Hong, Kolda, and Duersch 2018; Hu et al. 2018). Accordingly, a general method for the data which has more than 2 ordered label is needed.

In this article, we develop a probabilistic model and the associated theory for ordinal tensor decomposition. For a ordinal tensor $\mathcal{Y} = [\![y_{i_1,\cdots,i_N}]\!] \in \{1,2,\cdots,K\}^{d_1\times\cdots\times d_N}$, whose entries are integer value in from 1 to K indexed by the $N-$tuplet $(i_1,\cdots,i_N)$, we proposed the following low-rank multinomial model:

$$\mathcal{Y}|\Theta \sim \text{Multinomial}\{\boldsymbol{f}(\Theta)\}, \quad \text{where } \boldsymbol{f} = (f_1,\cdots,f_K), \quad \text{rank}(\Theta) = \boldsymbol{r} = (r_1,\cdots,r_N). \tag{1}$$

In this model, we assume that the entries of $\mathcal{Y}$ are realizations of independent multinomial random variable with parameter function $\boldsymbol{f}:\mathbb{R}\to[0,1]^N$ such that $\sum_{l=1}^N f_l = 1$. The latent parameter tensor, $\Theta = [\![\theta_{i_1,\cdots,i_N}]\!]$ has the same dimension as $\mathcal{Y}$ and takes continuous value in $\mathbb{R}$. Also, we assume that Θ admits a low rank $\boldsymbol{r}$-Tucker decomposition.

We organized this paper as follows. In section 2, we discuss detailed assumptions and descriptions about our probabilistic model in (1). Also, we suggest the estimation method for the latent parameters and related algorithm. In section 3, we provide the statistical properties of the upper, lower bounds and the phase-transition. We then provide the numerical experiments. Our model is applied to real-world data to check validity and performance in section 5. Finally, we wrap up the paper with a discussion.

2 Model

2.1 Low-rank cumulative model

For the N-mode ordinal tensor $\mathcal{Y} = [\![y_{i_1,\cdots,i_N}]\!] \in \{1,2,\cdots,K\}^{d_1\times\cdots\times d_N}$, we assume that its entries are realizations of independent multinomial random variables, such that

$$\mathbb{P}(y_{i_1,\cdots,i_N} = l) = f_l(\theta_{i_1,\cdots,i_N}), \quad (i_1,\cdots,i_N) \in [d_1]\times\cdots\times[d_N]. \tag{2}$$

In this model, a twice differentiable function $f_l:\mathbb{R}\to[0,1]$ with $l\in[K]$ is strictly increasing and satisfies $\sum_{l=1}^K f_l(\theta) = 1$ for a fixed θ. The tensor $\Theta = [\![\theta_{i_1,\cdots,i_N}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_N}$ is a hidden parameter which we are interested in. We assume the parameter tensor Θ is continuous value and admits a rank $\boldsymbol{r} = (r_1, \cdots, r_N)$ Tucker decomposition,

$$\Theta = \mathcal{C} \times_1 A_1 \cdots \times_N A_N, \tag{3}$$

where $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_N}$ is a core tensor and $A_n \in \mathbb{R}^{d_n \times r_n}$, for $n \in [N]$ is a factor matrix. We can get information about the influence of each mode by checking factor matrices.

The tensor model (2) has an equivalent representation as a latent model with K-level quantization. (Davenport et al. 2012; Lan, Studer, and Baraniuk 2014; Bhaskar and Javanmard 2015; Cai and Zhou 2013) where $\mathcal{Y} = [\![y_{i_1,\cdots,i_N}]\!]$ is a quantized value such that,

$$y_{i_1,\cdots,i_N} = \mathcal{Q}(\theta_{i_1,\cdots,i_N} + \epsilon_{i_1,\cdots,i_N}), \quad (i_1, \cdots, i_N) \in [d_1] \times \cdots [d_N], \tag{4}$$

where $\mathcal{E} = [\![\epsilon_{i_1,\cdots,i_N}]\!]$ is a noise tensor of *i.i.d.* from cumulative distribution function $\Phi(\cdot)$ and the function $\mathcal{Q} : \mathbb{R} \to [K]$ is a quantizer having the following rule.

$$\mathcal{Q}(x) = l, \text{ if } \omega_{l-1} < x \leq \omega_l, l \in [K], \tag{5}$$

where $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_{K-1})$ is a cutpoints vector such that $-\infty = \omega_0 < \omega_1 < \cdots < \omega_K = \infty$. That is, the entries of observed tensor $\mathcal{Y}$ fall in category l when the associated entries of the latent tensor $\Theta + \mathcal{E}$ fall in the lth interval of values. Then we have

$$\begin{aligned} f_l(\theta_{i_1,\cdots,i_N}) &= \mathbb{P}(y_{i_1,\cdots,i_N} = l) \\ &= \Phi(\omega_l - \theta_{i_1,\cdots,i_N}) - \Phi(\omega_{l-1} - \theta_{i_1,\cdots,i_N}). \end{aligned} \tag{6}$$

We can diversify our model by the choices of $\Phi(\cdot)$, or equivalently the distribution of $\mathcal{E}$ with given $\boldsymbol{\omega}$. The followings are 2 common choices of $\Phi(\cdot)$.

Example 1 (Logistic link/Logistic noise). *The logistic model is represented by* (2) *with* $f_l(\theta) = \Phi_{log}(\frac{\omega_l - \theta}{\sigma}) - \Phi_{log}(\frac{\omega_{l-1} - \theta}{\sigma})$ *where* $\Phi_{log}(x/\sigma) = (1 + e^{-x/\sigma})$. *Equivalently, the noise* $\epsilon_{i_1,\cdots,i_N}$ *in* (4) *follows i.i.d. logistic distribution with the scale parameter* σ.

Example 2 (Probit link/Gaussian noise). *The probit model is represented by* (2) *with* $f_l(\theta) = \Phi_{norm}(\frac{\omega_l - \theta}{\sigma}) - \Phi_{norm}(\frac{\omega_{l-1} - \theta}{\sigma})$ *where* Φ_{norm} *is the cumulative distribution function of the standard Gaussian. Equivalently, the noise* $\epsilon_{i_1,\cdots,i_N}$ *in* (4) *follows i.i.d.*$N(0, \sigma^2)$.

2.2 Rank-constrained likelihood-based estimation

Our goal is to estimate unknown parameter tensor Θ and a cutpoints vector $\boldsymbol{\omega}$ from observed tensor $\mathcal{Y}$ using a constrained likelihood approach. The log-likelihood function for (2) is

$$\mathcal{L}_{\mathcal{Y}}(\Theta, \boldsymbol{\omega}) = \sum_{i_1,\cdots,i_N} \left[\sum_{l\in[K]} \log(f_l(\theta_{i_1,\cdots,i_N}))\mathbb{1}_{\{y_{i_1,\cdots,i_N}=l\}} \right], \tag{7}$$

where the cutpoints vector $\boldsymbol{\omega}$ is implicitly contained in the function f_l. Considering the Tucker structure in (3), we have the following constrained optimization problem.

$$\begin{aligned} &\max_{(\Theta,\boldsymbol{\omega})\in\mathcal{D}\times\mathcal{W}} \mathcal{L}_{\mathcal{Y}}(\Theta, \boldsymbol{\omega}), \text{ where} \\ &\mathcal{D} = \{\Theta \in \mathbb{R}^{d_1\times\cdots\times d_N} : \text{rank}(\Theta) = \boldsymbol{r} \text{ and } \|\Theta\|_\infty \leq \alpha\} \\ &\mathcal{W} = \{\boldsymbol{\omega} \in \mathbb{R}^{K-1} : \omega_1 < \cdots < \omega_{K-1}\}, \end{aligned} \tag{8}$$

for a given rank $\boldsymbol{r} \in {\mathbb{N}_+}^N$ and a bound $\alpha \in \mathbb{R}_+$. The search space $\mathcal{D}$ has two constraints on unknown parameter Θ. The first constraint ensures that the unknown parameter Θ admits the Tucker decomposition with rank $\boldsymbol{r}$. The second constraint makes the entries of Θ bounded by a constant α. This bound condition is a technical assumption to help to recover Θ in the noiseless case. Similar conditions has been imposed in many literatures for the matrix case (Davenport et al. 2012; Bhaskar and Javanmard 2015; Cai and Zhou 2013; Bhaskar 2016). The search space $\mathcal{W}$ makes sure that the probability function f_l in (6) is positive.

2.3 Optimization

In this section, we describe the algorithm to seek the optimizer of (8). The objective function $\mathcal{L}_{\mathcal{Y}}(\Theta, \boldsymbol{\omega})$ is concave in $(\Theta, \boldsymbol{\omega})$ whenever $\Phi(x)$ is log-concave (McCullagh 1980; Burridge 1981). However, the feasible set $\mathcal{D}$ is not a convex set, which makes the optimization (8) a non-convex problem. One approach to handle this problem is utilizing the Tucker decomposition and converting optimization into a block-wise convex problem. Let us denote the objective function as

$$\mathcal{L}(\mathcal{C}, A_1, \cdots, A_N, \boldsymbol{\omega}) = \mathcal{L}_{\mathcal{Y}}(\mathcal{C} \times_1 A_1 \cdots \times_N A_N, \boldsymbol{\omega}). \tag{9}$$

From (9), we have $N+2$ blocks of variables in the objective function, one for the cutpoints vector $\boldsymbol{\omega}$, one for the core tensor $\mathcal{C}$ and N for the factor matrices A_n's. We can change the optimization problem to simple convex problem if any $N+1$ out of the $N+2$ blocks being fixed. Therefore, we can alternatively update one block at a time while other blocks being fixed. The algorithm 1 gives the full description.

Algorithm 1 Ordinal tensor decomposition

```
Input: Ordinal tensor $\mathcal{Y} \in \{1, \cdots, K\}^{d_1 \times \cdots \times d_N}$, rank $\boldsymbol{r} \in \mathbb{N}_+{}^{K-1}$, entrywise bound $\alpha \in \mathbb{R}_+$.
Output: Solution $(\hat{\Theta}, \hat{\boldsymbol{\omega}}) = \arg\max_{(\Theta, \boldsymbol{\omega}) \in \mathcal{D} \times \mathcal{W}} \mathcal{L}_{\mathcal{Y}}(\Theta, \boldsymbol{\omega})$.
Initialize random tensor $\mathcal{C}^{(0)} \in \mathbb{R}^{r_1 \times \cdots \times r_N}$, matrices $A^{(0)} = \{A_1^{(0)}, \cdots, A_N^{(0)}\}$ and $\boldsymbol{\omega}^{(0)} = (\omega_1^{(0)}, \cdots, \omega_{(K-1)}^{(0)})$ such that $\omega_1^{(0)} < \cdots < \omega_{K-1}^{(0)}$.
for t = 1,2,$\cdots$, do until convergence,
    Update $A_n$
    for n = 1,2,$\cdots$,N do
        $A_n^{(t+1)} \leftarrow \arg\max_{A_n \in \mathbb{R}^{d_n \times r_n}} \mathcal{L}\big(\mathcal{C}^{(t)}, A_1^{(t+1)}, \cdots, A_n, \cdots, A_N^{(t)}, \boldsymbol{\omega}^{(t)}\big)$
                such that $\|\mathcal{C}^{(t)} \times_1 A_1^{(t+1)} \cdots \times_n A_n \cdots \times_N A_N^{(t)}\|_\infty \leq \alpha$
    end for
    Update $\mathcal{C}$
    $\mathcal{C}^{(t+1)} \leftarrow \arg\max_{\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_N}} \mathcal{L}\big(\mathcal{C}, A_1^{(t+1)}, \cdots, A_N^{(t+1)}, \boldsymbol{\omega}^{(t)}\big)$
            such that $\|\mathcal{C} \times_1 A_1^{(t+1)} \cdots \times_N A_N^{(t+1)}\|_\infty \leq \alpha$
    $\Theta^{(t+1)} \leftarrow \mathcal{C}^{(t+1)} \times_1 A_1^{(t+1)} \cdots \times_N A_N^{(t+1)}$
    Update $\boldsymbol{\omega}$
    $\boldsymbol{\omega}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\omega} \in \mathbb{R}^{K-1}} \mathcal{L}_{\mathcal{Y}}\big(\Theta^{(t+1)}, \boldsymbol{\omega}\big)$ such that $\omega_1^{(0)} < \cdots < \omega_{K-1}^{(0)}$.
end for
return $\Theta, \boldsymbol{\omega}$
```

2.4 Rank selection

Algorithm 1 takes the rank $\boldsymbol{r}$ as an input variable. In practice, the rank $\boldsymbol{r}$ is hardly known. Estimating an appropriate rank $\boldsymbol{r}$ from a given tensor is an important issue. We suggest to use Bayesian Information Criterion(BIC) to choose the rank.

$$
\begin{aligned}
\hat{\boldsymbol{r}} &= \underset{\boldsymbol{r} \in \mathbb{N}_+^N}{\arg\min}\, BIC(\boldsymbol{r}) \\
&= \underset{\boldsymbol{r} \in \mathbb{N}_+^N}{\arg\min} \Big[-2\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}(\boldsymbol{r})) + \big(\prod_{n \in [N]} r_n + \sum_{n \in [N]} (d_n r_n - r_n^2) \big) \log(\prod_{n \in [N]} d_n) \Big],
\end{aligned} \tag{10}
$$

where $\hat{\Theta}(\boldsymbol{r})$ is a maximum likelihood estimater given the rank $\boldsymbol{r}$. The second part of the summation in (10) is the number of independent parameters in the model. We select the rank $\hat{\boldsymbol{r}}$ that minimizes BIC value through the grid search method.

	Tensor decomposition method					
	Ordinal (logistic link)			Continuous-valued		
	RMSE	MAE	Error	RMSE	MAE	Error
Average	0.1503711	0.1502662	0.1502151	0.1603685	0.1600219	0.1598499

Table 1: Tensor completion for the HCP data. Two methods are compared: the proposed ordinal tensor decomposition and the classical continuous Tucker decomposition.

3 Real-world Data Applications

In this section, we apply our ordinal tensor decomposition method to two real-world datasets of ordinal tensors. In the first application, we use our model to analyze an ordinal tensor consisting of structural connectivity patterns among 68 brain regions for 136 individuals from Human Connectome Project (HCP) (Geddes 2016). In the second application, we perform tensor completion from the data with missing values. The data tensor records the ratings from scale 1 to 5 of 42 users to 139 songs on 26 contexts (Baltrunas et al. 2011).

3.1 Human Connectome Project (HCP)

The human connectome project (HCP) is a $68 \times 68 \times 136$ tensor where the first two modes have 68 indices representing brain regions and the last mode has 136 indices meaning individuals. All the individual images were preprocessed following a standard pipeline (Zhang et al. 2018), and the brain was parcellated to 68 regions of interest following the Desikan atlas (Desikan et al. 2006). The tensor entries consist of $\{1, 2, 3\}$, the strength of fiber connections between 68 brain regions for each of the 136 individuals. First, we apply our ordinal tensor decomposition method with a logistic link function to the HCP data and identify similar brain nodes based on latent parameters. The BIC result suggests $\boldsymbol{r} = (23, 23, 8)$ with $\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \hat{\boldsymbol{\omega}}) = -216645.8$.

Also, we compare our ordinal tensor decomposition method with a logistic link function with the classical continuous Tucker decomposition method. We use 5 folded cross validation method for the comparison. Specifically, we randomly split the tensor entries into 5 similar sized pieces and alternatively use each piece of entries as a test data using other 80% entries as a training data. The entries in the test data are encoded as missing and then predicted based on the tensor decomposition from the training data. We set the rank of the tensor as $\boldsymbol{r} = (23, 23, 8)$ which minimizes BIC value. Table 1 shows RMSE, MAE and error(false prediction rate), averaged over 5 test set results. We can check that our ordinal tensor decomposition model outperforms the classic continuous decomposition in all criteria.

3.2 Music data with missing values

InCarMusic is a mobile application that offers music recommendation to passengers of cars based on contexts (Baltrunas et al. 2011). Our goal is to get the tensor completion from the $42 \times 139 \times 26$ tensor using the ordinal tensor decomposition and thereby we can offer context-specific music recommendation to users. The tensor entries consist of $\{1, 2, 3, 4, 5\}$, the ratings of 42 users to 139 songs on 26 contexts and are encoded as -1 for missing values. The number of missing values is 148904 and the number of available values is 2884.

References

Adomavicius, Gediminas et al. (2008). "Context-aware recommender systems". In: *AI Magazine* 32, pp. 67–80.

Baltrunas, Linas et al. (2011). "InCarMusic: Context-Aware Music Recommendations in a Car". In: *EC-Web*.

Bhaskar, Sonia A. (2016). "Probabilistic Low-Rank Matrix Completion from Quantized Measurements". In: *J. Mach. Learn. Res.* 17, 60:1–60:34.

Bhaskar, Sonia A. and Adel Javanmard (2015). "1-bit matrix completion under exact low-rank constraint". In: *2015 49th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6.

Burridge, Jim (1981). "A Note on Maximum Likelihood Estimation for Regression Models using Grouped Data". In:

Cai, Tony and Wen-Xin Zhou (2013). "A max-norm constrained minimization approach to 1-bit matrix completion". In: *J. Mach. Learn. Res.* 14, pp. 3619–3647.

Cichocki, Andrzej et al. (2015). "Tensor Decompositions for Signal Processing Applications: From two-way to multiway component analysis". In: *IEEE Signal Processing Magazine* 32, pp. 145–163.

Davenport, Mark A. et al. (2012). "1-Bit Matrix Completion". In: *ArXiv* abs/1209.3672.

Desikan, Rahul S. et al. (2006). "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest". In: *NeuroImage* 31, pp. 968–980.

Geddes, Linda (2016). "Human brain mapped in unprecedented detail". In:

Harshman, Richard A. (1970). "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-model factor analysis". In:

Hitchcock, Frank L (1928). "Multiple Invariants and Generalized Rank of a P-Way Matrix or Tensor". In: *Journal of Mathematics and Physics* 7.1-4, pp. 39–79. DOI: 10.1002/sapm19287139. eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1002/sapm19287139. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sapm19287139.

Hong, David, Tamara G. Kolda, and Jed A. Duersch (2018). “Generalized Canonical Polyadic Tensor Decomposition”. In: *ArXiv* abs/1808.07452.

Hore, Victoria et al. (2016). “Tensor decomposition for multi-tissue gene expression experiments”. In: *Nature Genetics*.

Hu, Qinghao et al. (2018). “Training Binary Weight Networks via Semi-Binary Decomposition”. In: *ECCV*.

Kutty, Sangeetha, Lin Chen, and Richi Nayak (2012). “A people-to-people recommendation system using tensor space models”. In: *SAC ’12*.

Lan, Andrew S., Christoph Studer, and Richard G. Baraniuk (2014). “Matrix recovery from quantized and corrupted measurements”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4973–4977.

McCullagh, Peter (1980). “Regression Models for Ordinal Data”. In:

Miwakeichi, Fumikazu et al. (n.d.). “Decomposing EEG data into space-time-frequency components using parallel factor analysis”. In: *Neuroimage* (), p. 2004.

Nickel, Maximilian, Volker Tresp, and Hans-Peter Kriegel (2011). “A Three-Way Model for Collective Learning on Multi-Relational Data”. In: *ICML*.

Sidiropoulos, Nikos D., Rasmus Bro, and Georgios B. Giannakis (2000). “Parallel factor analysis in sensor array processing”. In: *IEEE Trans. Signal Processing* 48, pp. 2377–2388.

Smilde, Age K., Rasmus Bro, and Paul Geladi (2004). “Multi-way Analysis with Applications in the Chemical Sciences”. In:

Sun, Jimeng et al. (2009). “MultiVis: Content-Based Social Network Exploration through Multi-way Visual Analysis”. In: *SDM*.

Sun, Will Wei et al. (2015). “Provable sparse tensor decomposition”. In:

Tucker, Ledyard (1964). “The extension of factor analysis to three-dimensional matrices”. In: *Contributions to mathematical psychology*. Ed. by H. Gulliksen and N. Frederiksen. New York: Holt, Rinehart and Winston, pp. 110–127.

— (1966). “Some mathematical notes on three-mode factor analysis”. In: *Psychometrika* 31.3, pp. 279–311. URL: https://EconPapers.repec.org/RePEc:spr:psycho:v:31:y:1966:i:3:p:279-311.

Wang, Miaoyan, Jonathan Fischer, and Yun S. Song (2017). “Three-way clustering of multi-tissue multi-individual gene expression data using constrained tensor decomposition”. In:

Wang, Miaoyan and Lexin Li (2018). “Learning from Binary Multiway Data: Probabilistic Tensor Decomposition and its Statistical Optimality”. In: *ArXiv* abs/1811.05076.

Zare, Ali et al. (2018). “Extension of PCA to Higher Order Data Structures: An Introduction to Tensors, Tensor Decompositions, and Tensor PCA”. In: *Proceedings of the IEEE* 106, pp. 1341–1358.

Zhang, Zhengwu et al. (2018). “Mapping population-based structural connectomes”. In: *NeuroImage* 172, pp. 130–145.