Tensor denoising and completion based on ordinal observations

Anonymous Authors[1]

Abstract

Higher-order tensors arise frequently in applications such as neuroimaging, recommendation system, social network analysis, and psychological studies. We consider the problem of low-rank tensor estimation from possibly incomplete, ordinal-valued observations. Two related problems are studied, one on tensor denoising and another on tensor completion. We propose a multi-linear cumulative link model, develop a rank-constrained M-estimator, and obtain theoretical accuracy guarantees. Our performance bound enjoys a faster convergence rate than previous results, and we show that the proposed estimator is minimax optimal under the class of low-rank models. An efficient completion method is further developed, which guarantees consistent recovery of an order-K $(d, \ldots, d)$-dimensional low-rank tensor using only $O(Kd)$ noisy, quantized observations. We demonstrate the outperformance of our approach over previous methods on the tasks of clustering and collaborative filtering.

1. Introduction (needs rewriting)

Higher-order tensors commonly arise in modern applications. In a context-aware recommendation system, for example, the survey results can be organized as a three-way tensor, where the (i, j, k)-th entry indicates the rating from user i to the item j at context k. While there are numerous methods for continuous-valued tensors, fewer methods are available for learning tensors based on qualitative observations. Qualitative observations usually take values in a limited set of categories which may be on an ordinal or on a purely nominal scale. The present work focuses on ordinal tensors; namely, the tensor entries are categorical variables with a natural ordering among the categories. (Prior work and our contribution. See Table 1 for comparison.)

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Preliminaries

Let $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote an order-K $(d_1, \ldots, d_K)$-dimensional tensor. We use y_ω to denote the tensor entry indexed by ω, where $\omega \in [d_1] \times \cdots \times [d_K]$. The Frobenius norm of $\mathcal{Y}$ is defined as $\|\mathcal{Y}\|_F = \sum_\omega y_\omega^2$ and the infinity norm of $\mathcal{Y}$ is defined as $\|\mathcal{Y}\|_\infty = \max_\omega |y_\omega|$. We use $\mathcal{Y}_{(k)}$ to denote the unfolded matrix of size d_k-by-$\prod_{i \neq k} d_k$, obtained by reshaping the tensor along the mode-k. The Tucker rank of $\mathcal{Y}$ is defined as a length-K vector $\boldsymbol{r} = (r_1, \ldots, r_K)$, where r_k is the rank of matrix $\mathcal{Y}_{(k)}$ for all $k \in [K]$. We say that an event A occurs "with very high probability" if $\mathbb{P}(A)$ tends to 1 faster than any polynomial of tensor dimension $d_{\min} = \min\{d_1, \ldots, d_K\}$ as $d_{\min} \to \infty$.

We use lower-case letters (e.g., a, b, c) for scalars/vectors, upper-case boldface letters (e.g., $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$) for matrices, and calligraphy letters (e.g., $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$) for tensors of order three or greater. For ease of notation, we allow the basic arithmetic operators (e.g., $\leq, +, -$) to be applied to pairs of vectors in an element-wise manner. We use the shorthand $[n]$ to denote the n-set $\{1, \ldots, n\}$ for $n \in N_+$.

3. Model formulation and motivation

3.1. Observation model

Let $\mathcal{Y}$ denote an order-K $(d_1, \ldots, d_K)$-dimensional tensor. Suppose the entries of $\mathcal{Y}$ are ordinal-valued, and the observation space consists of L ordered categories, denoted by $[L] := \{1, \ldots, L\}$. We propose a cumulative link model for the ordinal tensor $\mathcal{Y} = [\![y_\omega]\!] \in [L]^{d_1 \times \cdots \times d_K}$. Specifically, assume the entries y_ω are (conditionally) independently distributed with cumulative probabilities:

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

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

The formulation (1) imposes an additive model to the transformed probability of cumulative categories, rather than that of each single category. This modeling choice is to respect the ordering structure among the categories. For example, if

	Bhaskar [2016]	Ghadermarzy et al. [2018]	This paper
Higher-order tensors ($K \geq 3$)	✗	✓	✓
Multi-level categories ($L \geq 3$)	✓	✗	✓
Error rate for tensor denoising	d^{-1} for $K = 2$	$d^{-(K-1)/2}$	$d^{-(K-1)}$
Optimality guarantee under low-rank models	unknown	✗	✓
Sample complexity for tensor completion	d^K	Kd	Kd

Table 1. Comparison of various low-rank estimation methods based on categorical observations. For an order-K dimensional-$(d, \ldots, d)$ ordinal tensor with L-level observations, we report the error bound in the recovered signal (for tensor denoising) and required sample complexity (for tensor completion) as functions of tensor dimensions. (font in the table seems odd...) Add a row "algorithm stability"

we choose the inverse link $f^{-1}(x) = \log \frac{x}{1-x}$ to be the log odds, then the model (1) assumes linear spacing between the proportional odds:

$$\log \frac{\mathbb{P}(y_\omega \leq \ell)}{\mathbb{P}(y_\omega > \ell)} - \log \frac{\mathbb{P}(y_\omega \leq \ell - 1)}{\mathbb{P}(y_\omega > \ell - 1)} = b_\ell - b_{\ell-1}, \quad (2)$$

for all tensor entries y_ω. When there are only two categories in the observation space (e.g. binary tensors), the cumulative model (1) is equivalent to the usual multinomial link model. In general, however, when the number of categories $L \geq 3$, the proportional odds assumption (2) is more parsimonious, in that, the ordered categories can be envisaged as contiguous intervals on the continuous scale, where the points of division are exactly $b_1 < \cdots < b_{L-1}$. This interpretation will be made more explicit in the next section.

3.2. Latent-variable interpretation

We show that the ordinal tensor model (1) is equivalent to an L-level quantization model on $\mathcal{Y} = [\![y_\omega]\!]$:

$$y_\omega = \begin{cases} 1, & \text{if } y^*_\omega \in (-\infty, b_1], \\ 2, & \text{if } y^*_\omega \in (b_1, b_2], \\ \vdots & \vdots \\ L, & \text{if } y^*_\omega \in (b_{L-1}, \infty), \end{cases} \quad (3)$$

for all $\omega \in [d_1] \times \cdots \times [d_k]$. Here, $\mathcal{Y}^* = [\![y^*_{i_1,\ldots,i_K}]\!]$ is a latent continuous-valued tensor following an additive noise model:

$$\underbrace{\mathcal{Y}^*}_{\text{latent continuous-valued tensor}} = \underbrace{\Theta}_{\text{signal tensor}} + \underbrace{\mathcal{E}}_{\text{i.i.d. noise}}, \quad (4)$$

where $\mathcal{E} = [\![\varepsilon_\omega]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is a noise tensor with i.i.d. entries following distribution $\mathcal{F}(\varepsilon)$. From the viewpoint of (4), the parameter tensor Θ can be interpreted as the latent signal tensor prior to contamination and quantization.

The equivalence between the latent-variable model (3) and the cumulative link model (1) is established if the link f behaves like a cumulative distribution function of noise ε, i.e., $f(\theta) = \mathbb{P}(\varepsilon \geq -\theta)$. We describe several common choices of link f, or equivalently, the distribution of $\mathcal{E}$.

Example 1 (Logistic model)**.** The logistic model is characterized by (1) with $f(\theta) = (1 + e^{-\theta/\sigma})^{-1}$, where $\sigma > 0$ is the scale parameter. Equivalently, the noise ε_ω in (3) follows i.i.d. logistic distribution with scale parameter σ.

Example 2 (Probit model)**.** The probit model is characterized by (1) with $f(\theta) = \mathbb{P}(z \leq \theta/\sigma)$, where $z \sim N(0, 1)$. Equivalently, the noise ε_ω in (3) follows i.i.d. $N(0, \sigma^2)$.

Example 3 (Laplace model)**.** The Laplacian model is characterized by (1) with

$$f(\theta) = \begin{cases} \frac{1}{2} e^{\frac{\theta}{\sigma}}, & \text{if } \theta < 0, \\ 1 - \frac{1}{2} e^{-\frac{\theta}{\sigma}}, & \text{if } \theta \geq 0, \end{cases}$$

where $\sigma > 0$ is the scale parameter. Equivalently, the noise ε_ω in (3) follows i.i.d. Laplace distribution with scale parameter σ.

Other link functions are also possible, such as Cauchy function, inverse log-log, etc (McCullagh, 1980). All the models share the property that the ordered categories can be thought of as contiguous interval on some continuous scale. We should point out that, although the latent-variable interpretation is incisive, our estimation procedure does not require the existence of $\mathcal{Y}^*$. Therefore, our model (1) is general and still valid in the absence of quantization process. More generally, we make the following assumptions to the link function f.

Assumption 1. *The link function is assumed to satisfy:*

1. *$f(\theta)$ is strictly increasing, strictly log-concave, and twice-differentiable in $\theta \in \mathbb{R}/\{0\}$.*
2. *$f'(\theta)$ is unimodal, and symmetric with respect to $\theta = 0$.*

3.3. Problem 1: Tensor denoising

The first question we aim to address is tensor denoising:

(P1) Given the quantization process induced by f and the cut-off points $\boldsymbol{b}$, how accurate can we estimate the latent signal tensor Θ from the ordinal observation $\mathcal{Y}$?

Clearly, the problem (P1) cannot be solved uniformly for all possible Θ. We focus on a class of "low-rank" and "flat"

signal tensors, which is a plausible assumption in practical applications (Zhou et al., 2013; Bhaskar & Javanmard, 2015). Specifically, we consider the parameter space:

$$\mathcal{P} = \left\{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \text{rank}(\Theta) \le \boldsymbol{r},\ \|\Theta\|_\infty \le \alpha\right\}. \quad (5)$$

where $\boldsymbol{r} = (r_1, \ldots, r_K)$ denotes the Tucker rank of Θ.

The candidate tensor of our interest satisfies two constraints. The first is that Θ is a low-rank tensor, with $r_k = O(1)$ for all $k \in [K]$. Equivalently, Θ admits the Tucker decomposition:

$$\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_1 \cdots \times_K \boldsymbol{M}_K, \quad (6)$$

where $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is a core tensor, $\boldsymbol{M}_k \in \mathbb{R}^{d_k \times r_k}$ are factor matrices with orthogonal columns, and $\times_k$ denotes the tensor-by-matrix multiplication (Kolda & Bader, 2009). The Tucker low-rankness is popularly used in tensor data analysis, and is shown to provide a reasonable tradeoff between model complexity and model flexibility. Note that, unlike matrices, there are various notations of tensor low-rankness, such as CP representation (Hitchcock, 1927) and train representation (Oseledets, 2011). Some notation of low-rankness may lead to mathematically ill-defined optimization; for example, the set of low-rank CP tensors is not necessarily compact (De Silva & Lim, 2008). We choose Tucker representation for parsimony and easier interpretation.

The second constraint is that the entries of Θ are uniformly bounded in magnitude by a constant $\alpha \in \mathbb{R}_+$. In view of (4), we refer to α as the signal level. The entry-wise bound assumption is a technical condition that ensures the well-posedness for the problems we considered. In the context of denoising with ordinal observations, the entry-wise bound helps avoid the degeneracy in probability estimation.

3.4. Problem 2: Tensor completion

Motivated by applications in collaborative filtering and recommendation system, we also also consider a more general setup when only a subset of tensor entries y_ω are observed. Let $\Omega \subset [d_1] \times \cdots \times [d_K]$ denote the set of observed indices. The second question we aim to address is stated as follows:

(P2) Given an incomplete set of ordinal observation $\{y_\omega\}_{\omega \in \Omega}$ based on the model (1), how many sampled entries do we need to consistently recover Θ?

The answer to (P2) depends on the choice of Ω. We propose a general model on Ω that allows both uniform and non-uniform sampling. Specifically, let $\Pi = \{\pi_{i_1,\ldots,i_K}\}$ denote a predefine probability distribution over the index set such that $\sum_{\omega \in [d_1] \times \cdots \times [d_K]} \pi_\omega = 1$. We assume that each index in Ω are drawn with replacement using distribution Π. This model relaxes the the commonly-used uniform sampling and is arguably a better fit in practical applications.

We consider the same parameter space (5) for the completion problem. In addition to the reasons mentioned in Section 3.3, the entrywise bound assumption also serves as the incoherence condition here. In classical matrix completion, the incoherence is often imposed on the singular vectors. This assumption was relaxed by considering "flat" matrices with bounded magnitude (Negahban et al., 2011; Cai & Zhou, 2013; Bhaskar & Javanmard, 2015). We adopt the same assumption for higher-order tensors.

4. Rank-constrained M-estimator

We present a general treatment to both problems mentioned above. With a little abuse of notation, we use Ω to denote either the full index set $\Omega = [d_1] \times \cdots \times [d_K]$ (for the tensor denoising) or a random subset induced from the sampling distribution Π (for the tensor completion). Define $b_0 = -\infty$, $b_L = \infty$, $f(-\infty) = 0$ and $f(\infty) = 1$. The log-likelihood associated with the observed entries is

$$\mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b}) = \sum_{\omega \in \Omega} \sum_{\ell \in [L]} \left\{\mathbb{1}_{\{y_\omega = \ell\}} \log\left[f(\theta_\omega + b_\ell) - f(\theta_\omega + b_{\ell-1})\right]\right\}. \quad (7)$$

We propose a rank-constrained maximum likelihood estimator (M-estimator) for Θ:

$$\hat{\Theta} = \arg\max_{\Theta \in \mathcal{P}} \mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b}), \text{ where}$$
$$\mathcal{P} = \left\{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \text{rank}(\Theta) \le \boldsymbol{r},\ \|\Theta\|_\infty \le \alpha\right\} \quad (8)$$

In practice, the cut-off points $\boldsymbol{b}$ are unknown and should be jointly estimated with Θ. For technical convenience, we assume that, in this section, the cut-off points $\boldsymbol{b}$ are known, or otherwise have been consistently estimated (??). The adaptation of unknown $\boldsymbol{b}$ will be addressed in Section 5. Similar treatment applies to the hyper-parameter $\boldsymbol{r}$.

We define a few key quantities that will be used in our theory. Let $g_\ell = f(\theta + b_\ell) - f(\theta + b_{\ell-1})$ for all $\ell \in [L]$, and

$$A_\alpha = \min_{\ell \in [L], |\theta| \le \alpha} g_\ell(\theta), \quad U_\alpha = \max_{\ell \in [L], |\theta| \le \alpha} \frac{\dot{g}_\ell(\theta)}{g_\ell(\theta)},$$
$$L_\alpha = \min_{\ell \in [L], |\theta| \le \alpha} \left[\frac{\dot{g}_\ell^2(\theta)}{g_\ell^2(\theta)} - \frac{\ddot{g}_\ell(\theta)}{g_\ell(\theta)}\right],$$

where $\dot{g}(\theta) = dg(\theta)/d\theta$, and α is the entrywise bound of Θ. In view of equation (4), these quantities characterize the geometry (such as flatness and convexity) of the latent noise distribution. Under the Assumption 1, all these quantities are strictly positive and independent of tensor dimensions.

4.1. Estimation error for tensor denoising

For the tensor denosing problem, we assume the full set of tensor entries are observed. In such a case, we assess the

estimation accuracy using the mean squared error (MSE):

$$\mathrm{MSE}(\hat{\Theta}, \Theta^{\text{true}}) = \frac{1}{\prod_k d_k}\|\Theta - \Theta^{\text{true}}\|_F^2.$$

The next theorem establishes the upper bound for the proposed $\hat{\Theta}$ in (8).

Theorem 4.1 (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 $\Theta^{true} \in \mathcal{P}$. Define $r_{\max} = \max_k r_k$. Then, with very high probability, the estimator in (8) satisfies*

$$\mathit{MSE}(\hat{\Theta}, \Theta^{\mathit{true}}) \le \min\left(4\alpha^2,\ \frac{c_2 U_\alpha^2 r_{\max}^{K-1}}{L_\alpha^2}\frac{\sum_k d_k}{\prod_k d_k}\right), \quad (9)$$

where $c_1, c_2 > 0$ are two constants that depend only on K.

Theorem 4.1 provides the statistical converge for the estimator (8). Its proof is given in the Supplements. In fact, the proof of this theorem shows that the same statistical rate holds not only for the global optimizer of (8), but also for any local optimizer $\check{\Theta}$ in the level set $\{\check{\Theta} \in \mathcal{P} : \mathcal{L}_{\mathcal{Y},\Omega}(\check{\Theta}) \ge \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}})\}$. This suggests that the local optimality itself is not necessarily a severe concern in our context, as long as the objective value at the estimator is large enough. In Section 6.1, we will perform empirical studies to assess the algorithmic stability and its impact to statistical efficiency.

A similar conclusion can be obtained for the prediction error, measured in Kullback-Leibler (KL) divergence, between the two categorical distributions.

Corollary 1 (Prediction error)**.** Assume the same set-up as in Theorem 4.1. Let $\mathbb{P}_{\mathcal{Y}}$ and $\hat{\mathbb{P}}_{\mathcal{Y}}$ denote the distributions generating the L-level ordinal tensor $\mathcal{Y}$, given the true parameter Θ and its estimator $\hat{\Theta}$, respectively. Assume $L \ge 2$. Then, with very high probability,

$$\mathrm{KL}(\mathbb{P}_{\mathcal{Y}}||\hat{\mathbb{P}}_{\mathcal{Y}}) \le \frac{c_2 U_\alpha^2 r_{\max}^{K-1}}{L_\alpha^2}\frac{(4L-6)\dot{f}^2(0)}{A_\alpha}\frac{\sum_k d_k}{\prod_k d_k}, \quad (10)$$

where $c_1, c_2 > 0$ are the same constants as in Theorem 4.1.

To gain insight into these bounds, we consider a special setting where the dimensions are the same in all modes, i.e., $d_1 = \cdots = d_K = d$. In such a case, our bound (9) reduces to

$$\mathrm{MSE}(\hat{\Theta}, \Theta^{\text{true}}) \asymp d^{-(K-1)}, \quad \text{as } d \to \infty.$$

Hence, the estimator achieves consistency with polynomial convergence rate. Comparing to max norm based estimator $\tilde{\Theta}$ proposed by Ghadermarzy et al. (2018), $\mathrm{MSE}(\tilde{\Theta}, \Theta^{\text{true}}) \asymp d^{-(K-1)/2}$, we see that our bound has a faster convergence rate. We believe the improvement comes from utilizing the exact low-rank structure of Θ, whereas an scale-sensitive surrogate rank measure was employed in Ghadermarzy et al. (2018).

Our bound also generalizes the previous results on ordinal matrices. The convergence rate for rank-constrained matrix estimation was $O(1/\sqrt{d})$ (Bhaskar, 2016), which fits into our special case when $K = 2$. Furthermore, our results (9) and (10) reveal that the convergence becomes favorably as the order of tensor data increases. Intuitively, for tensor data analysis, the sample size is the number of tensor entries, $\prod_k d_k$, and the number of free parameters is roughly on the order of $\sum_k d_k$, assuming $r_{\max} = O(1)$. A higher tensor order implies higher effective sample size per parameter, and thus exhibits a faster convergence rate in high dimensions.

We next establish the statistical optimality of our estimator $\hat{\Theta}$. The result is based on the information theory, and therefore applies to all estimators in $\mathcal{P}$, including but not limited to $\hat{\Theta}$ in (8). We show that this lower bound matches the upper bound in (9) as a function of tensor dimensions.

Theorem 4.2 (Minimax lower bound)**.** *Assume the same set-up as in Theorem 4.1, and $d_{\max} = \max_k d_k \ge 8$. Let $\inf_{\hat{\Theta}}$ denote the infimum over all estimators $\hat{\Theta} \in \mathcal{P}$ based on the ordinal tensor observation $\mathcal{Y} \in [L]^{d_1\times\cdots\times d_K}$. Then, under the model (1),*

$$\inf_{\hat{\Theta}}\sup_{\Theta^{\mathit{true}}\in\mathcal{P}} \mathbb{P}\Big\{\mathit{MSE}(\hat{\Theta}, \Theta^{\mathit{true}}) \ge \frac{1}{256}\min\left(\alpha^2,\ \frac{C r_{\max} d_{\max}}{\prod_k d_k}\right)\Big\} \ge \frac{1}{8},$$

where $C = C(\alpha, L, f, \boldsymbol{b}) > 0$ is a constant independent of tensor dimensions and the rank.

4.2. Sample complexity for tensor completion

Now we consider the problem of tensor completion, when only a subset of entries Ω are observed. We consider a general sampling procedure induced by Π. The recovery accuracy is assessed by the weighted squared deviation:

$$\|\Theta - \hat{\Theta}\|_{F,\Pi}^2 \overset{\text{def}}{=} \frac{1}{|\Omega|}\mathbb{E}_{\Omega\sim\Pi}\|\Theta - \hat{\Theta}\|_F^2 = \sum_{\omega\in[d_1]\times\cdots\times[d_K]} \pi_\omega(\Theta_\omega - \hat{\Theta}_\omega)^2. \quad (11)$$

Note that the recovery error depends on the distribution Π. In particular, the entries with higher sampling probabilities are equipped with better accuracy guarantee, compared to the ones with lower sampling probabilities.

Remark 1. If we assume each tensor entry is sampled with some positive probability, i.e., there exits a constant $\mu > 0$ such that

$$\pi_\omega \ge \frac{1}{\mu\prod_k d_k}, \quad \text{for all} \quad \omega \in [d_1]\times\cdots\times[d_K],$$

then the error measure in (11) provides an upper bound for

Algorithm 1 Ordinal tensor decomposition (allowing missing data)

Input: Ordinal tensor $\mathcal{Y} \in [L]^{d_1 \times \cdots \times d_K}$,
Rank $\boldsymbol{r} \in {\mathbb{N}_+}^K$,
Entry-wise bound $\alpha \in \mathbb{R}_+$.
Output: $(\hat{\Theta}, \hat{\boldsymbol{b}}) = \arg\max_{(\Theta, \boldsymbol{b}) \in \mathcal{D} \times \mathcal{B}} \mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b})$.
Initialize Core tensor $\mathcal{C}^{(0)}$,
Factor matrices $\{\boldsymbol{M}_1^{(0)}, \cdots, \boldsymbol{M}_K^{(0)}\}$,
Cut-off points $\boldsymbol{b}^{(0)}$.
for $t = 1, 2, \cdots$, **do**
  for $k = 1, 2, \cdots, K$ **do**
    Update $\boldsymbol{M}_k$ while fixing other blocks:
    $\boldsymbol{M}_k^{(t+1)} \leftarrow \arg\max_{\boldsymbol{M}_k \in \mathbb{R}^{d_k \times r_K}} \mathcal{L}_{\mathcal{Y},\Omega}(\boldsymbol{M}_k)$,
    s.t. $\|\Theta^{(t+1)}\|_\infty \le \alpha$, where $\Theta^{(t+1)}$ is the parameter tensor based on the current block estimates.
  end for
  Update $\mathcal{C}$ while fixing other blocks:
  $\mathcal{C}^{(t+1)} \leftarrow \arg\max_{\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_K}} \mathcal{L}_{\mathcal{Y},\Omega}(\mathcal{C})$, s.t. $\|\Theta^{(t+1)}\|_\infty \le \alpha$
  Update Θ based on the current block estimates:
  $\Theta^{(t+1)} \leftarrow \mathcal{C}^{(t+1)} \times_1 \boldsymbol{M}_1^{(t+1)} \cdots \times_K \boldsymbol{M}_K^{(t+1)}$
  Update $\boldsymbol{b}$ while fixing $\Theta^{(t+1)}$:
  $\boldsymbol{b}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{b} \in \mathbb{R}^{L-1}} \mathcal{L}_{\mathcal{Y},\Omega}\big(\Theta^{(t+1)}, \boldsymbol{b}\big)$
end for
return $\Theta, \boldsymbol{b}$

MSE:

$$\|\Theta - \hat{\Theta}\|_{F,\Pi}^2 \ge \frac{\|\Theta - \hat{\Theta}\|_F^2}{\mu \prod_k d_k} = \frac{1}{\mu}\text{MSE}(\hat{\Theta}, \Theta^{\text{true}}).$$

The equality is attained in the case of uniform sampling for which $\mu = 1$.

Theorem 4.3. *Assume the same set-up as in Theorem 4.1. Suppose that we observe a subset of tensor entries $\{y_\omega\}_{\omega \in \Omega}$, where Ω is chosen at random with replacement according to a probability distribution Π. Let $\hat{\Theta}$ be the solution to (8), and assume $r_{\max} = O(1)$. Then, with very high probability over the sampled set Ω (not over $\mathcal{Y}$?),*

$$\|\Theta - \hat{\Theta}\|_{F,\Pi}^2 \to 0, \quad \text{as} \quad \frac{|\Omega|}{\sum_k d_k} \to \infty.$$

Theorem 4.3 shows that our estimator achieves consistency when sample size is greater than $O(\sum_k d_k)$. add remarks comparing with prior work, and with the degree of freedom.

5. Numerical Implementation

In this section, we describe the algorithm to seek the optimizer of (7). The objective function $\mathcal{L}_{\mathcal{Y},\Omega}(\Theta)$ is concave in Θ whenever $\mathcal{P}$ is log-concave (McCullagh, 1980). However, the feasible set $\mathcal{P}$ is not a convex set, which makes the optimization (7) a non-convex problem. One approach to handle this problem is utilizing the Tucker decomposition and converting optimization into a block-wise convex problem. From (7) and (6), we have $K + 2$ blocks of variables in the objective function, one for the cut-points vector $\boldsymbol{b}$, one for the core tensor $\mathcal{C}$, and K for the factor matrices $\boldsymbol{M}_k$'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 update one block at a time while other blocks being fixed, and alternate the optimization via iteration. The Algorithm 1 gives the full description.

(highlight that $\boldsymbol{b}$ and Θ are jointly estimated...)

Before concluding this section, we comment on two implementation details. First, the problem (8) is non-convex, so Algorithm 1 usually has no theoretical guarantee on global optimality. Nevertheless, as shown in Section 4.1, the desired rate holds not only for the global optimizer, but also for the local optimizer with $\mathcal{L}_{\mathcal{Y},\Omega}(\hat{\Theta}) \ge \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}})$. We find that, in practice, the convergence point $\hat{\Theta}$ is often satisfactory, in that the corresponding objective $\mathcal{L}_{\mathcal{Y},\Omega}(\hat{\Theta})$ is close to and actually slightly larger than the objective evaluated at the true parameter $\mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}})$. Figure 5 shows the trajectory of the objective function that is output in the default setting of Algorithm 1, with the input tensor generated from probit model (1) using $d_1 = d_2 = d_3 = d$ and $r_1 = r_2 = r_3 = r$. The dashed line is the objective value at the true parameter $\mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}})$. It is observed that the algorithm generally converges quickly to a desirable value in reasonable number of steps. The actual running time per iteration is shown in the plot legend. In the case of complete observation, each sub-routine in the algorithm scales linear with the sample size $\mathcal{O}(\prod_k d_k)$, and this complexity matches the classical tensor methods (Kolda & Sun, 2008; Anandkumar et al., 2014).

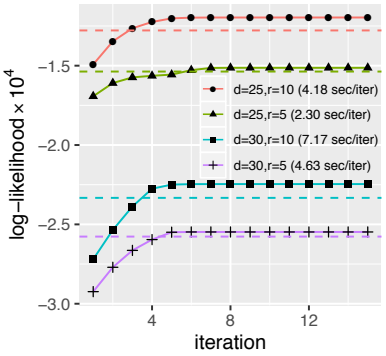

Figure 1. Trajectory of the objective function with varing d and r.

Second, the algorithm takes the rank $\boldsymbol{r}$ as an input. In practice, the rank $\boldsymbol{r}$ is hardly known and needs to be estimated from the data. We suggest to use Bayesian information

criterion (BIC) and choose the rank that minimizes BIC; i.e.

$$\hat{\boldsymbol{r}} = \arg\min_{\boldsymbol{r}\in\mathbb{N}_+^K} \mathrm{BIC}(\boldsymbol{r})$$
$$= \arg\min_{\boldsymbol{r}\in\mathbb{N}_+^K} \left[-2\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}(\boldsymbol{r}), \hat{\boldsymbol{b}}) + p_e(\boldsymbol{r}) \log(\prod_k d_k) \right],$$

where $\hat{\Theta}(\boldsymbol{r})$ is the estimator given the rank $\boldsymbol{r}$, and $p_e(\boldsymbol{r}) \overset{\text{def}}{=} \sum_k (d_k - r_k) r_k + \prod_k r_k$ is the effective number of parameters in the model. We select the rank $\hat{\boldsymbol{r}}$ that minimizes BIC through a grid search. Our choice of BIC aims to balance between the goodness-of-fit for the data and the degree of freedom in the population model.

6. Experiments

In this section, we evaluate the empirical performance of our estimation method. We consider both the complete and the incomplete settings, and compare the recovery accuracy of our estimator with other tensor-based methods. Unless otherwise stated, we generated L-level ordinal data tensors from model (1) using standard probit link f. We consider order-3 parameter tensors with $d_1 = d_2 = d_3 = d$ and $r_1 = r_2 = r_3 = r$. The parameter tensor is simulated based on (6), where the entries in the core tensor $\mathcal{C}$ are i.i.d. drawn from $N(0, 1)$, and the factors $\boldsymbol{M}_k$ are uniformly sampled (with respect to Haar measure) from matrices with orthonormal columns. We set the cut-off points $b_\ell = f^{-1}(\ell/L)$ for $\ell = 0, \ldots, L$, such that $f(b_\ell)$ are evenly spaced from 0 to 1. In each simulation study, we report the summary statistics across n_{sim} = 30 replications.

6.1. Finite-sample performance

We first consider the case of complete observation, and assess the empirical relationship between the MSE and model complexity. In particular, we consider various settings of dimension d, rank r and signal level $\alpha = \|\Theta\|_\infty$. Figure 2a plots the estimation error versus the tensor dimension d for three different ranks $r \in \{3, 5, 8\}$. The decay in the error appears to behave on the order of d^{-2}, which is consistent with our theoretical results (9). It is seen that a higher rank leads to a larger error, as reflected by the upward shift of the curve as r increases. Indeed, a higher rank implies a higher intrinsic dimension, thus increasing the difficulty of the estimation.

Figure 2b shows the estimation error versus the signal level. A larger estimation error is observed when the signal is either too small or too large. This non-monotonic phenomenon may seem surprising, but we believe this is an intrinsic feature with ordinal-valued date problem. In light of the latent-variable interpretation (see Section 3.2), estimation from ordinal observation can be interpreted as an inverse problem of quantization. The estimation error will diverge when the noise $\mathcal{E}$ is negligible, because there is no way to distinguish two different signal tensors, say, $\Theta_1 = \boldsymbol{a}_1 \otimes \boldsymbol{a}_2 \otimes \boldsymbol{a}_3$ and $\Theta_2 = \mathrm{sign}(\boldsymbol{a}_1) \otimes \mathrm{sign}(\boldsymbol{a}_2) \otimes \mathrm{sign}(\boldsymbol{a}_3)$, from the quantized observations. This behavior is clearly contrary to that of classical continuous-valued tensor problem, and the phenomenon has attracted great interest recently (Davenport et al., 2014; Sur & Candès, 2018).

We then investigate the case of incomplete observation. We sample tensor entries via uniform sampling and assess the recovery error with sample complexity. Figure 2c shows the completion error versus the sample size. The error reduces roughly at a rate of $\mathcal{O}(d^{-2})$(?) as the sample size increases. This suggests....

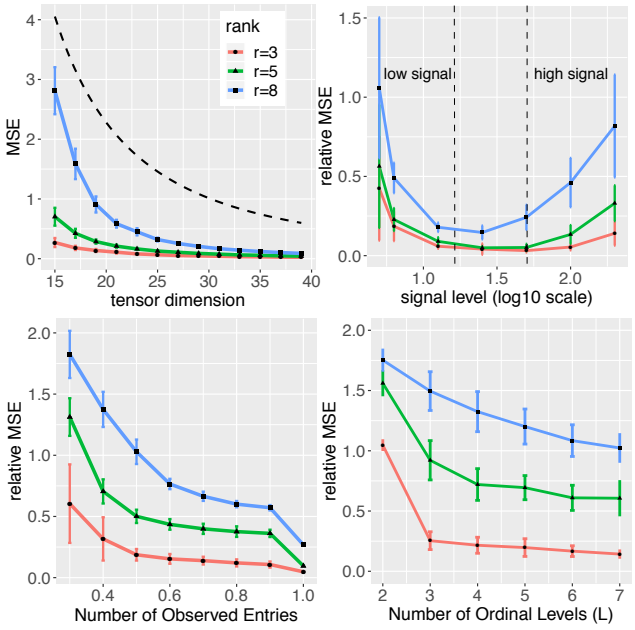

Figure 2. Finite sample performance

6.2. Comparison with alternative methods

Compare with

1. continuous tensor method (Filipović & Jukić, 2015);
2. 1-bit tensor method (Ghadermarzy et al., 2018);
3. ordinal matrix method (Bhaskar, 2016).

7. Real data application

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

METHOD	RMSE	MAE	ERROR
ORIDNAL TENSOR DECOMPOSITION	0.1504	0.1503	0.1502
CONTI TENSOR DECOMPOSITION	0.1604	0.1600	0.1598
1BIT-COMPLETION (PROBIT)	0.3658	0.3633	0.3620
1BIT-COMPLETION (LOGISTIC)	0.4519	0.4219	0.4068

Table 2. Results of comparisons among 4 methods on the HCP data predicting the test data. Four methods are the ordinal tensor decomposition algorithm, the continuous tensor decomposition algorithm and the 1 bit tensor completion method with probit link function and logistic link function. Each method is evaluated by RMSE, MAE, error(false prediction rate).

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).

7.1. Human Connection Project (HCP)

The human connectome project (HCP) is a 68 68 136 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 $\mathcal{L}_{\mathcal{Y},\Omega}(\hat{\Theta}, \hat{\boldsymbol{b}}) = -216645.8$.

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.

7.2. InCarMusic recommendation system

InCarMusic is a mobile application that offers music recommendation to passengers of cars based on contexts (Baltrunas et al. 2011). Our goal is to perform the tensor completion to the $42 \times 139 \times 26$ ordinal tensor and thereby we can offer context-specific music recommendation to users. The tensor entries consist of ordinal observations on the scale 1 to 5, the ratings of 42 users to 139 songs on 26 contexts and are encoded as NA for missing values. The number of missing values is 148,904 and the number of available values is 2,884. We suggest three estimators $\hat{y}_\omega$ based on $(\hat{\theta}_\omega, \hat{\boldsymbol{b}})$:

- (Mean) $\hat{y}_\omega^{(\text{Mean})} = \sum_\ell \ell g_\ell(\hat{\theta}_\omega, \hat{\boldsymbol{b}})$;
- (Median) $\hat{y}_\omega^{(\text{Median})} = \min\{\ell \in [L] \colon f_\ell(\hat{\theta}_\omega, \hat{\boldsymbol{b}}) \geq 0.5\}$;
- (Mode) $\hat{y}_\omega^{(\text{Mode})} = \arg\max_\ell g_\ell(\hat{\theta}_\omega, \hat{\boldsymbol{b}})$

Note that, under the ordinal tensor model (1), the estimator $\hat{y}_\omega^{(\text{Mean})}$ (uniformally?) minimizes $\mathbb{E}(y_\omega - \hat{y}_\omega)^2$, and the estimator $\hat{y}_\omega^{(\text{Median})}$ minimizes $\mathbb{E}|y_\omega - \hat{y}_\omega|$ (?). In contrast, these three estimators degenerate to a single estimator under the continuous-valued tensor decomposition model. We then round $\hat{y}_\omega$ to the nearest integer and assess the accuracy using three metrics: Mean squared error (MSE), Mean absolute deviation (MAD), and Misclassification rate (MCR).

Results are averaged from 5 fold cross validation. add s.e. to the above table.

Critera	Our method			Continuous method
	$\hat{y}_\omega^{(\text{Mean})}$	$\hat{y}_\omega^{(\text{Median})}$	$\hat{y}_\omega^{(\text{Mode})}$	(Filipović & Jukić, 2015)
MSE	2.01	2.14	3.64	8.18
MAD	1.23	1.13	1.26	2.36
MCR	0.80	0.75	0.54	0.96

Table 3. Comparison of tensor completion performance on InCarMusic dataset.

We also attempted to run matrix ordinal mehod (Bhaskar, 2016) to the InCarMusic dataset. Unfortunately, the $\boldsymbol{b}$ cannot be estimated from their method.

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Bhaskar, S. A. Probabilistic low-rank matrix completion from quantized measurements. *The Journal of Machine Learning Research*, 17(1):2131–2164, 2016.

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

Cai, T. and Zhou, W.-X. A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

Davenport, M. A., Plan, Y., Van Den Berg, E., and Wootters, M. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

De Silva, V. and Lim, L.-H. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30 (3):1084–1127, 2008.

Filipović, M. and Jukić, A. Tucker factorization with missing data with application to low-rank tensor completion. *Multidimensional systems and signal processing*, 26(3): 677–692, 2015.

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

Hitchcock, F. L. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

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

Kolda, T. G. and Sun, J. Scalable tensor decompositions for multi-aspect data mining. In *2008 Eighth IEEE international conference on data mining*, pp. 363–372. IEEE, 2008.

McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

Negahban, S., Wainwright, M. J., et al. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Oseledets, I. V. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

Sur, P. and Candès, E. J. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*, 2018.

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