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 mean squared error 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. Furthermore, the procedure developed serves as an efficient completion method which guarantees consistent recovery of an order-K $(d,\ldots,d)$-dimensional low-rank tensor using only $\tilde{\mathcal{O}}(Kd)$ noisy, quantized observations. We demonstrate the outperformance of our approach over previous methods on the tasks of clustering and collaborative filtering.

1. Introduction

Multidimensional arrays, a.k.a. tensors, arise in a variety of applications including recommendation systems (Baltrunas et al., 2011), social networks (Nickel et al., 2011), genomics (Hore et al., 2016), and neuroimaging (Zhou et al., 2013). There is a growing need to develop general methods for analyzing these noisy, high-dimensional datasets that can handle two main problems. The problem of tensor denoising – which aims to recover a signal tensor from its noisy entries – has gained increased attention in theory and applications (Xia & Zhou, 2019; Wang & Zeng, 2019). A related problem, tensor completion, examines the minimum number of entries needed for a consistent recovery (Ghadermarzy et al., 2018; 2019). Low-rankness is often imposed to the signal tensor, which efficiently reduces the intrinsic dimension in both problems.

A number of low-rank tensor estimation methods have been proposed (Kolda & Bader, 2009; Acar et al., 2010), revitalizing classical methods such as CANDECOMP/PARAFAC (CP) decomposition (Hitchcock, 1927) and Tucker decomposition (Tucker, 1966). These tensor methods treat the entries as continuous-valued. In many cases, however, we encounter datasets of which the entries are qualitative. For example, the Netflix problem records the ratings of users on movies over time. Each data is a rating on a nominal scale $\{$*very like, like, neutral, dislike, very dislike*$\}$. Another example is in the signal processing, where the digits are frequently rounded or truncated so that only integer values are available. The qualitative observations take values in a limited set of categories, making the learning problem harder compared to continuous observations.

Ordinal entries are categorical variables with an ordering among the categories; for example, *very like* $\prec$ *like* $\prec$ *neutral* $\prec \cdots$. The analyses of tensors with the ordinal entries are mainly complicated by two key properties needed for a reasonable model. First, the model should be invariant under a reversal of categories, say, from the Netflix example, *very like* $\succ$ *like* $\succ$ *neutral* $\succ \cdots$, but not under arbitrary label permutations. Second, the parameter interpretations should be consistent under merging or splitting of contiguous categories. The classical continuous tensor model (Kolda & Bader, 2009; Ghadermarzy et al., 2019) fails in the first aspect, whereas the binary tensor model (Ghadermarzy et al., 2018) lacks the second property. An appropriate model for ordinal tensors has yet to be studied.

Our contribution. This paper presents an efficient low-rank estimation method and theory for tensors with ordinal-valued entries. Our main contributions are summarized in Table 1. We propose a cumulative link model for higher-order tensors, develop a rank-constrained M-estimator, and obtain theoretical accuracy guarantees. The mean squared error bound is established, and we show that the obtained bound has minimax optimal rate in high dimensions under the low-rank model. Our estimator enjoys a faster convergence rate $\mathcal{O}(d^{-(K-1)/2})$ than $\mathcal{O}(d^{-K})$ in Ghadermarzy et al. (2018), which is a substantial improvement as the

[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.

	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 with previous work. For ease of presentation, we summarize the error rate and sample complexity assuming equal tensor dimension in all modes. K: tensor order; L: number of ordinal levels; d: dimension at each mode.

order K increases. Furthermore, our proposal serves as an efficient completion algorithm that guarantees consistent recovery of an order-K $(d, \ldots, d)$-dimensional low-rank tensor using only $\tilde{\mathcal{O}}(Kd)$ noisy, quantized observations.

Related work. Our work is related to, but clearly distinctive from, several lines of existing literature. Matrix completion from quantized samples was firstly introduced for binary observations (Cai & Zhou, 2013; Davenport et al., 2014; Bhaskar & Javanmard, 2015) and then extended to ordinal observations (Bhaskar, 2016). As we show in Section 4, applying existing matrix methods to an ordinal tensor results in a suboptimal estimator with a slower convergence rate. Therefore, a full exploitation of the tensor structure is necessary; this is the focus of the current paper.

Our work is also connected to non-Gaussian tensor decomposition. Existing work focuses exclusively on univariate observations such as binary- or continuous-valued entries (Wang & Li, 2018; Hong et al., 2019; Ghadermarzy et al., 2018). As we mentioned earlier, the ordinal observations add considerable challenges to the model formulation. We address the problems from two perspectives. From statistical perspective, our proposed model generalizes the usual binary tensor model while preserving palindromic invariance (McCullagh, 1980) for ordinal observations. From algorithm perspective, our alternating optimization compares favorably to the approximate (non-convex) algorithm developed in the context of binary tensors (Ghadermarzy et al., 2018). We numerically compare the two approaches in Section 6.

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 $d_{\min} = \min\{d_1, \ldots, d_K\} \to \infty$.

We use lower-case letters $(a, b, c, \ldots)$ for scalars/vectors, upper-case boldface letters $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \ldots)$ for matrices, and calligraphy letters $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots)$ 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 tensors 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 data tensor. Suppose the entries of $\mathcal{Y}$ are ordinal-valued, and the observation space consists of L ordered levels, 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:

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

where $\boldsymbol{b} = (b_1, \ldots, b_{L-1})$ is a set of unknown scalars satisfying $b_1 < \cdots < b_{L-1}$, $\Theta = [\![\theta_\omega]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is a continuous-valued parameter tensor satisfying certain low-dimensional structure (to be specified later), 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. This modeling choice is to respect the ordering structure among the categories. For example, if we choose the inverse link $f^{-1}(x) = \log \frac{x}{1-x}$ to be the log odds, then the model (1) implies 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

The ordinal tensor model (1) with certain types of link f has the equivalent representation as an L-level quantization model on $\mathcal{Y} = [\![y_\omega]\!]$:

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

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

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

where $\mathcal{E} = [\![\varepsilon_\omega]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is a noise tensor with i.i.d. entries according to distribution $\mathbb{P}(\varepsilon)$. 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 is chosen to be the cumulative distribution function of noise ε, i.e., $f(\theta) = \mathbb{P}(\varepsilon \leq \theta)$. We describe two common choices of link f, or equivalently, the distribution of ε.

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. $N(0, \sigma^2)$.

Other link functions are also possible, such as Laplace, Cauchy, 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 refer to 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 about the link f.

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

1. *$f(\theta)$ is strictly increasing and twice-differentiable in $\theta \in \mathbb{R}/\{0\}$.*
2. *$f'(\theta)$ is strictly log-concave 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 accurately 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:

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

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

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

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

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 imposed 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 rank (Hitchcock, 1927) and train rank (Oseledets, 2011). Some notation of low-rankness may lead to mathematically ill-posed optimization; for example, the best low CP-rank tensor approximation may not exist (De Silva & Lim, 2008). We choose Tucker representation for well-posedness of optimization and easy 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 avoids the degeneracy in probability estimation with ordinal observations.

3.4. Problem 2: Tensor completion

Motivated by applications in collaborative filtering, we 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 observations $\{y_\omega\}_{\omega \in \Omega}$, how many sampled entries do we need to consistently recover Θ based on the model (1)?

The answer to (P2) depends on the choice of Ω. We consider 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 is drawn with replacement using distribution Π. This sampling model relaxes the uniform sampling in literature and is arguably a better fit in 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 requirement for completion. In classical matrix completion, the incoherence is often imposed on the singular vectors. This assumption is recently relaxed for "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

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

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

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

In practice, the cut-off points $\boldsymbol{b}$ are unknown and should be jointly estimated with Θ. For technical convenience, we assume in this section that the cut-off points $\boldsymbol{b}$ are known. The adaptation of unknown $\boldsymbol{b}$ is addressed in Section 5 and the Supplement.

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

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

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 including flatness and convexity of the latent noise distribution. Under the Assumption 1, all these quantities are strictly positive and independent of tensor dimension.

4.1. Estimation error for tensor denoising

For the tensor denoising problem, we assume that the full set of tensor entries are observed. We assess the estimation accuracy using the mean squared error (MSE):

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

The next theorem establishes the upper bound for the MSE of 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*

$$MSE(\hat{\Theta},\Theta^{true})\leq\min\left(4\alpha^2,\ \frac{c_2U_\alpha^2r_{\max}^{K-1}}{L_\alpha^2}\frac{\sum_k d_k}{\prod_k d_k}\right), \tag{9}$$

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

Theorem 4.1 establishes the statistical convergence for the estimator (8). In fact, the proof of this theorem (see the Supplement) shows that the same statistical rate holds, not only for the global optimizer (8), but also for any local optimizer $\check{\Theta}$ in the level set $\{\check{\Theta}\in\mathcal{P}:\mathcal{L}_{\mathcal{Y},\Omega}(\check{\Theta})\geq\mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}})\}$. This suggests that the local optimality itself is not necessarily a severe concern in our context, as long as the convergent objective is large enough. In Section 5, we perform empirical studies to assess the algorithmic stability.

A similar conclusion is obtained for the prediction error, measured in Kullback-Leibler (KL) divergence, between the categorical distributions in the observation space.

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 $L\geq 2$. Then, with very high probability,

$$\text{KL}(\mathbb{P}_{\mathcal{Y}}||\hat{\mathbb{P}}_{\mathcal{Y}})\leq\frac{c_2U_\alpha^2r_{\max}^{K-1}}{L_\alpha^2}\frac{(4L-6)\dot{f}^2(0)}{A_\alpha}\frac{\sum_k d_k}{\prod_k d_k}, \tag{10}$$

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 with equal dimension in all modes, i.e., $d_1=\cdots=d_K=d$. In such a case, our bound (9) reduces to

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

Hence, our estimator achieves consistency with polynomial convergence rate. We compare the bound with existing literature. In the special case $L=2$, Ghadermarzy et al. (2018) proposed a max-norm constrained estimator $\tilde{\Theta}$ with $\text{MSE}(\tilde{\Theta},\Theta^{\text{true}})\asymp d^{-(K-1)/2}$. In contrast, our estimator converges at a rate of $d^{-(K-1)}$, which is substantially faster than theirs. This provides a positive answer to the open question posed in Ghadermarzy et al. (2018) whether the square root in the bound is removable. The improvement stems from utilizing the exact low-rankness of Θ, whereas the surrogate rank measure employed in Ghadermarzy et al. (2018) is scale-sensitive.

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

We next show the statistical optimality of our estimator $\hat{\Theta}$. The result is based on the information theory, and applies to all estimators in $\mathcal{P}$, including but not limited to $\hat{\Theta}$ in (8).

Theorem 4.2 (Minimax lower bound)**.** *Assume the same set-up as in Theorem 4.1, and $d_{\max} = \max_k d_k \geq 8$. 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),*

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

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

We see that the lower bound matches the upper bound in (9) on the polynomial order of tensor dimension. Therefore, our estimator (8) is order-optimal.

4.2. Sample complexity for tensor completion

We now consider the tensor completion problem, 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 error:

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

Note that the recovery error depends on the distribution Π. In particular, tensor entries with higher sampling probabilities have more influence on the recovery accuracy, compared to the ones with lower sampling probabilities.

Remark 1. If we assume each entry is sampled with strictly positive probability; i.e. there exits a constant $\mu > 0$ s.t.

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

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

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

The equality is attained under uniform sampling with $\mu = 1$.

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 $r_{\max} = \mathcal{O}(1)$. Then, with very high probability,*

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

Theorem 4.3 shows that our estimator achieves consistent recovery using as few as $\tilde{\mathcal{O}}(Kd)$ noisy, quantized observations from an order-K $(d, \ldots, d)$-dimensional tensor. Note that $\tilde{\mathcal{O}}(Kd)$ roughly matches the degree of freedom for an order-K tensor of fixed rank $\boldsymbol{r}$, suggesting the optimality of our sample requirement. This sample complexity substantially improves over earlier result $\mathcal{O}(d^{\lceil K/2 \rceil})$ based on square matricization (Mu et al., 2014), or $\mathcal{O}(d^{N/2})$ based on tensor nuclear-norm regularization (Yuan & Zhang, 2016). Existing methods that achieve $\tilde{\mathcal{O}}(Kd)$ sample complexity require either a deterministic cross sampling design (Zhang et al., 2019) or univariate measurements (Ghadermarzy et al., 2018). Our method extends the conclusions to multi-level measurements under a broader class of sampling schemes.

5. Numerical Implementation

We describe the algorithm to seek the optimizer of (7). In practice, the cut-off points $\boldsymbol{b}$ are often unknown, so we choose to maximize $\mathcal{L}_{\mathcal{Y},\Omega}$ jointly over $(\Theta, \boldsymbol{b}) \in \mathcal{P} \times \mathcal{B}$. The objective $\mathcal{L}_{\mathcal{Y},\Omega}$ is concave in $(\Theta, \boldsymbol{b})$ whenever f' is log-concave (see the Supplement). However, the feasible set $\mathcal{P}$ is non-convex, which makes the optimization (7) a non-convex problem. We employ the alternating optimization approach by utilizing the Tucker representation of Θ. Specifically, based on (6) and (7), the objective function consists of $K + 2$ blocks of variables, one for the cut-off points $\boldsymbol{b}$, one for the core tensor $\mathcal{C}$, and K for the factor matrices $\boldsymbol{M}_k$'s. The optimization is a simple convex problem if any $K + 1$ out of the $K + 2$ blocks are fixed. We update one block at a time while holding others fixed, and alternate the optimization throughout the iteration. The convergence is guaranteed whenever $\mathcal{L}_{\mathcal{Y},\Omega}$ is bounded from above, since the alternating procedure monotonically increases the objective. The Algorithm 1 gives the full description.

Algorithm 1 Ordinal tensor decomposition

```
Input: Ordinal data 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: $(\hat{\Theta}, \hat{\boldsymbol{b}}) = \arg\max_{(\Theta, \boldsymbol{b}) \in \mathcal{P} \times \mathcal{B}} \mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b})$.
Random initialization of core tensor $\mathcal{C}^{(0)}$, factor matrices $\{\boldsymbol{M}_k^{(0)}\}$, and cut-off points $\boldsymbol{b}^{(0)}$.
for $t = 1, 2, \cdots,$ do
    for $k = 1, 2, \cdots, K$ do
        Update $\boldsymbol{M}_k^{(t+1)}$ 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. $\|\Theta^{(t+1)}\|_\infty \leq \alpha$, where $\Theta^{(t+1)}$ is the parameter tensor based on the current block estimates.
    end for
    Update $\mathcal{C}^{(t+1)}$ 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. $\|\Theta^{(t+1)}\|_\infty \leq \alpha$.
    Update $\Theta^{(t+1)}$ 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 $\boldsymbol{b}^{(t+1)}$ while fixing $\Theta^{(t+1)}$:
    $\boldsymbol{b}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{b} \in \mathcal{B}} \mathcal{L}_{\mathcal{Y},\Omega}\left(\Theta^{(t+1)}, \boldsymbol{b}\right)$.
end for
return $(\hat{\Theta}, \hat{\boldsymbol{b}})$
```

We comment on two implementation details before concluding this section. 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 $\mathcal{L}_{\mathcal{Y},\Omega}(\hat{\Theta}) \geq \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}})$. In practice, we find the convergence point $\hat{\Theta}$ upon random initialization 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) with $d_1 = d_2 = d_3 = d$ and $r_1 = r_2 = r_3 = r$. The dashed line is the objective value at the true parameter $\mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}})$. We find 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.

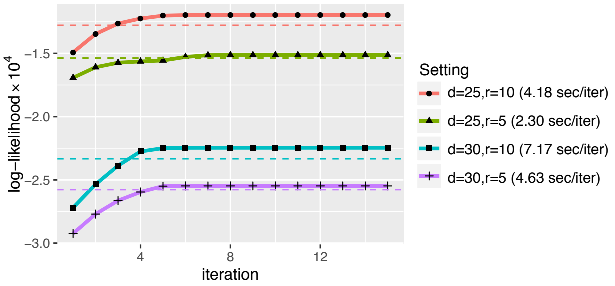

Figure 1. Trajectory of objective function with various d and r.

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.

$$\begin{aligned}\hat{\boldsymbol{r}} &= \arg\min_{\boldsymbol{r} \in \mathbb{N}_+^K} \text{BIC}(\boldsymbol{r}) \\ &= \arg\min_{\boldsymbol{r} \in \mathbb{N}_+^K} \{-2\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}(\boldsymbol{r}), \hat{\boldsymbol{b}}(\boldsymbol{r})) + p_e(\boldsymbol{r}) \log(\prod_k d_k)\},\end{aligned}$$

where $\hat{\Theta}(\boldsymbol{r}), \hat{\boldsymbol{b}}(\boldsymbol{r})$ are the estimates 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 $\hat{\boldsymbol{r}}$ that minimizes BIC through a grid search. The choice of BIC is intended to balance between the goodness-of-fit for the data and the degrees of freedom in the population model.

6. Experiments

In this section, we evaluate the empirical performance of our method. We investigate both the complete and the incomplete settings, and compare the recovery accuracy with other tensor-based methods. Unless otherwise stated, the ordinal data tensors are generated from model (1) using standard probit link f. We consider the setting with $K = 3$, $d_1 = d_2 = d_3 = d$, and $r_1 = r_2 = r_3 = r$. The parameter tensors are simulated based on (6), where the core tensor entries 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 $\ell \in [L]$, 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

The first experiment examines the performance under complete observations. We assess the empirical relationship between the MSE and various aspects of model complexity, such as 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). We find 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 the higher number of parameters to estimate, thus increasing the difficulty of the estimation. Figure 2b shows the estimation error versus the signal level under $d = 20$. Interestingly, a larger estimation error is observed when the signal is either too small or too large. The non-monotonic behavior may seem surprising, but this is an intrinsic feature in the estimation with ordinal data. In view of the latent-variable interpretation (see Section 3.2), estimation from ordinal observation can be interpreted as an inverse problem of quantization. Therefore, the estimation error diverges in the absence of noise $\mathcal{E}$, because it is impossible to distinguish two different signal tensors, e.g., $\Theta_1 = \boldsymbol{a}_1 \otimes \boldsymbol{a}_2 \otimes \boldsymbol{a}_3$ and $\Theta_2 = \text{sign}(\boldsymbol{a}_1) \otimes \text{sign}(\boldsymbol{a}_2) \otimes \text{sign}(\boldsymbol{a}_3)$, from the quantized observations. This phenomenon (Davenport et al., 2014; Sur & Candès, 2019) is clearly contrary to the classical continuous-valued tensor problem.

The second experiment investigates the incomplete observations. We consider L-level tensors with $d = 20$, $\alpha = 10$ and choose a subset of tensor entries via uniform sampling. Figure 2c shows the estimation error of $\hat{\Theta}$ versus the fraction of observation $\rho = |\Omega|/d^K$. As expected, the error reduces with increased ρ or decreased r. Figure 2d evaluates the impact of ordinal levels L to estimation accuracy, under the setting $\rho = 0.5$. An improved performance is observed as L grows, especially from binary observations ($L = 2$) to multi-level ordinal observations ($L \geq 3$). The result showcases the benefit of multi-level observations compared to binary observations.

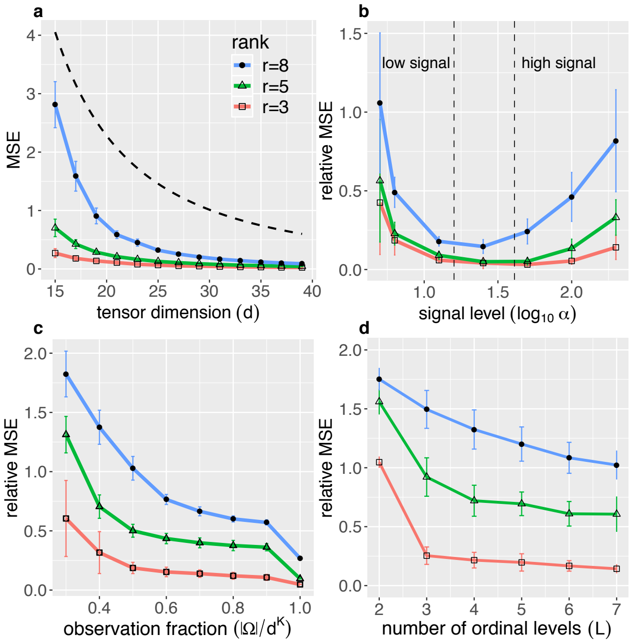

Figure 2. Empirical relationship between (relative) MSE versus (a) dimension d, (b) signal level α, (c) observation fraction ρ, and (d) number of ordinal levels L. In panels (b)-(d), we plot the relative $\text{MSE} = \|\hat{\Theta} - \Theta^{\text{true}}\|_F / \|\Theta^{\text{true}}\|_F$ for better visualization.

6.2. Comparison with alternative methods

Next, we compare our ordinal tensor method (**Ordinal-T**) with three popular low-rank methods:

- Continuous tensor decomposition (**Continuous-T**) (Acar et al., 2010) is a low-rank approximation method based on classical Tucker model.
- One-bit tensor completion (**1bit-T**) (Ghadermarzy et al., 2018) is a max-norm penalized tensor learning method based on partial binary observations.
- Ordinal matrix completion (**Ordinal-M**) (Bhaskar, 2016) is a rank-constrained matrix estimation method based on noisy, quantized observations.

We apply each of the above methods to L-level ordinal tensors $\mathcal{Y}$ generated from model (1). The **Continuous-T** is applied directly to $\mathcal{Y}$ by treating the L levels as continuous observations. The **Ordinal-M** is applied to the matrix $\mathcal{Y}_{(1)}$ obtained via 1-mode unfolding. The **1bit-T** is applied to $\mathcal{Y}$ in two ways. The first approach (**1bit-sign-T**) follows from Ghadermarzy et al. (2018) that transforms $\mathcal{Y}$ to a binary tensor, by taking the entrywise sign of the mean-adjusted tensor, $\mathcal{Y} - |\Omega|^{-1} \sum_\omega y_\omega$. The second approach (**1bit-category-T**) transforms the order-3 ordinal tensor $\mathcal{Y}$ to an order-4 binary tensor $\mathcal{Y}^\sharp = [\![y^\sharp_{ijkl}]\!]$ via dummy variable encoding; i.e., $y^\sharp_{ijk\ell} = \mathbb{1}_{\{y_{ijk} = \ell\}}$ for all $\ell \in [L-1]$. We evaluate the methods by their capabilities in predicting the most likely label for each entry, i.e., $y^{\text{mode}}_\omega = \arg\max_\ell \mathbb{P}(y_\omega = \ell)$. Two performance metrics are considered: mean absolute deviation, $\text{MAD} = d^{-K} \sum_\omega |y^{\text{mode}}_\omega - \hat{y}^{\text{mode}}_\omega|$, and misclassification rate, $\text{MCR} = d^{-K} \sum_\omega \mathbb{1}_{\{y^{\text{mode}}_\omega \neq \text{round}(\hat{y}^{\text{mode}}_\omega)\}}$, where $\text{round}(\cdot)$ denotes the nearest integer of the prediction (possibly continuous-valued returned by **Continuous-T**). Note that MAD penalizes the large deviation more heavily than MCR.

Figure 3 compares the prediction accuracy under the setting $\alpha = 10$, $d = 20$, and $r = 5$. The problem size we considered is comparable to Ghadermarzy et al. (2018). We find that our method outperforms the others in both MAD and MCR. In particular, methods built on multi-level observations (**Ordinal-T**, **Ordinal-M**, **1bit-category-T**) exhibit stable MCR over ρ and L, whereas the others two methods (**Continuous-T**, **1bit-sign-T**) generally fail except for $L = 2$ (Figures 3a-b). This observation highlights the necessity of modeling multi-level probabilities in classification task. Interestingly, although both **1bit-category-T** and our method **Ordinal-T** behave similarly for binary tensors ($L = 2$), the improvement of our method is substantial as L increases (Figures 3a and 3c). One possible reason is that our method incorporates the intrinsic ordering among the L levels via proportional odds assumption (2), whereas **1bit-category-T** ignores the ordinal structure and dependence among the induced binary entries. Figures 3c-d assess the prediction accuracy with sample size. We see a clear advantage of our method (**Ordinal-T**) over the matricization (**Ordinal-M**) in both complete and non-complete observations. When the observation fraction is small, e.g., $|\Omega|/d^K = 0.4$, the tensor-based completion shows $\sim 30\%$ reduction in error compared to the matricization.

We also compare the methods by their performance in predicting the median label, $y^{\text{median}}_\omega = \min\{\ell \colon \mathbb{P}(y_\omega = \ell) \geq 0.5\}$. Under the latent variable model (4) and Assump-

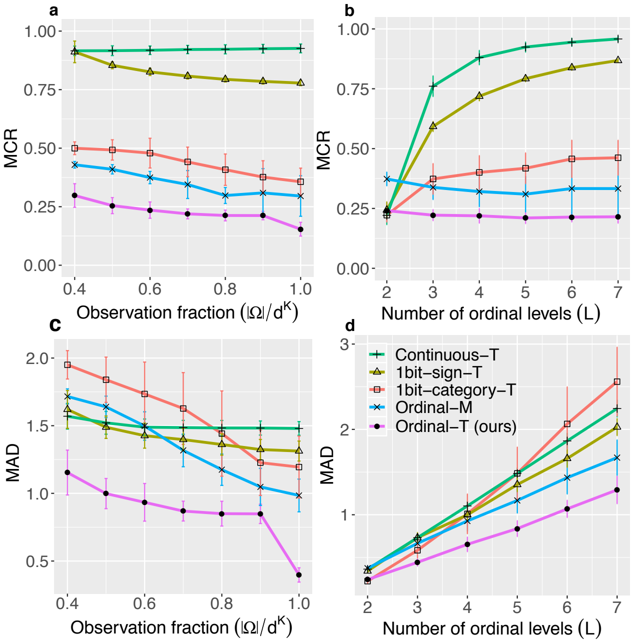

Figure 3. Performance comparison in MCR (a, b) and MAD (c, d). (b, d) Prediction errors versus the number of ordinal levels L when $\rho = 0.8$. (a, c) Prediction errors versus sample complexity $\rho = |\Omega|/d^K$ when $L = 5$.

tion 1, the median label is the quantized Θ without noise; i.e. $y_\omega^{\text{median}} = \sum_\ell \mathbb{1}_{\theta_\omega \in (b_{\ell-1}, b_\ell]}$. Similar results demonstrate the outperformance of our method (see the Supplement).

7. Data Applications

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

7.1. Human Connectome Project (HCP)

Each entry in the HCP dataset takes value on a nominal scale, $\{$*high, moderate, low*$\}$, indicating the strength level of fiber connection. We convert the dataset to a 3-level ordinal tensor $\mathcal{Y} \in [3]^{68\times 68\times 136}$ and apply the ordinal tensor method with a logistic link function. The BIC suggests $\boldsymbol{r} = (23, 23, 8)$ with $\mathcal{L}_{\mathcal{Y},\Omega}(\hat{\Theta}, \hat{\boldsymbol{b}}) = -216,646$. Based on the estimated Tucker factors $\{\hat{M}_k\}$, we perform a clustering analysis via K-mean on the brain nodes (see detailed procedure in the Supplement). We find that the 68 brain nodes are grouped into eight clusters, and the clustering capture the spatial separation of brain nodes very well (see the Supplement). In particular, the top two clusters represent the left and right hemispheres; the smaller clusters represent local regions driving by similar nodes. For example, the cluster IV consists of brain nodes *postcentral*, *posteriorcingulate*, *pericalcarine* and *temporalpole*, all of which are located at *SupraM*. The identified similarity among brain nodes suggests the applicability of our method.

We compare the goodness-of-fit of various tensor methods on the HPC data. Table 2 summarizes the prediction error via 5-fold stratified cross-validation averaged over 10 runs. Our method outperforms the others, especially in MAD.

Human Connectome Project (HCP) dataset		
Method	MAD	MCR
Ordinal-T (ours)	0.1607 (0.0005)	0.1606 (0.0005)
Continuous-T	0.2530 (0.0002)	0.1599 (0.0002)
1bit-sign-T	0.3566 (0.0010)	0.1563 (0.0010)
InCarMusic dataset		
Method	MAD	MCR
Ordinal-T (ours)	1.37 (0.039)	0.59 (0.009)
Continuous-T	2.39 (0.152)	0.94 (0.027)
1bit-sign-T	1.39 (0.003)	0.81 (0.005)

Table 2. Comparison of prediction error in the HPC and InCarMusic analyses. Standard errors are reported in parentheses.

7.2. InCarMusic recommendation system

We apply ordinal tensor completion to a recommendation system *InCarMusic*. *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 tensor completion to impute the unobserved entries in the $42 \times 139 \times 26$ ordinal tensor and thereby we can offer context-specific music recommendation to users. The data tensor consists of 2,884 observed entries. Table 2 shows the averaged prediction error via 5-fold cross validation. The high missing rate makes the accurate classification challenging. Nevertheless, our method achieves the best performance among the three.

8. Conclusions

We have developed a low-rank tensor estimation method based on possibly incomplete, ordinal-valued observations. A sharp error bound is established, and we demonstrate the outperformance of our approach compared to other methods. The work unlocks several directions of future research. One interesting question would be the inference problem, i.e.. to assess the uncertainty of the obtained estimates and the imputation. Other directions include the trade-off between (non)convex optimization and statistical/computational efficiency. While convex relaxations are popular approach for matrix-based problem, they are often slow in practice (Chen et al., 2019). The interplay between computational efficiency and statistical accuracy in general tensor problems warrants future research.