Tensor denoising and completion based on ordinal observations

Chanwoo Lee
University of Wisconsin – Madison
chanwoo.lee@wisc.edu

Miaoyan Wang
University of Wisconsin – Madison
miaoyan.wang@wisc.edu

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 $(d, \dots, d)$-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.

Keywords: Higher-order tensors, ordinal observation, tensor decomposition, tensor completion

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 that can handle two main problems for analyzing these noisy, high-dimensional datasets. The first problem is tensor denoising which aims to recover a signal tensor from its noisy entries (Hong et al., 2019; Wang and Zeng, 2019). The second problem is tensor completion which 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, thereby efficiently reducing the intrinsic dimension in both problems.

A number of low-rank tensor estimation methods have been proposed (Anandkumar et al., 2014; Wang and Song, 2017), 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 entry 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 and 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 for higher-order tensors. 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.

	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 when tensor rank $r = O(1)$. For ease of presentation, we summarize the error rate and sample complexity (neglecting log factors) when the tensor dimensions are equal in all modes. K: tensor order; L: number of ordinal levels; d: dimension at each mode.

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 and Zhou, 2013; Davenport et al., 2014; Bhaskar and 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 and 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 $d_{\min} = \min\{d_1, \ldots, d_K\} \to \infty$. For any two functions f, g depending on $(d_1, \ldots, d_K)$, we write $f = \mathcal{O}(g)$ to indicate that $f \leq Cg$, where $C > 0$ is a constant independent of tensor dimension. We write $f = \tilde{\mathcal{O}}(g)$ to indicate that $f \leq C(\log d_{\min})^\beta g$ for some $\beta > 0$.

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:

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

where $\boldsymbol{b} = (b_1, \ldots, b_{L-1})$ is a set of unknown scalars satisfying $b_1 < \cdots < b_{L-1}$, $\Theta = [\![\theta_\omega]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ 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:

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

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, 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 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 $f(\cdot) : \mathbb{R} \mapsto [0,1]$.

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

1. *$f(\theta)$ is strictly increasing and twice-differentiable in θ.*
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 and Javanmard, 2015). Specifically, we consider the parameter space:

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

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 and 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 and 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 and Zhou, 2013; Bhaskar and 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:

$$\begin{aligned} \hat{\Theta} &= \arg\max_{\Theta\in\mathcal{P}} \mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b}), \text{ where} \\ \mathcal{P} &= \Big\{ \Theta \in \mathbb{R}^{d_1\times\cdots\times d_K} : \operatorname{rank}(\Theta) \le \boldsymbol{r},\ \|\Theta\|_\infty \le \alpha \Big\}. \end{aligned} \tag{8}$$

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 Appendix A.

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

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

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 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):

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

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*

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

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 Section 8.1) 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 $\{\check{\Theta} \in \mathcal{P} : \mathcal{L}_{\mathcal{Y},\Omega}(\check{\Theta}) \ge \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\mathrm{true}})\}$. 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 4.1 (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,*

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

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

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

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 $\mathrm{MSE}(\tilde{\Theta}, \Theta^{\mathrm{true}}) \asymp d^{-(K-1)/2}$. 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),*

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

where $C = C(\alpha, L, f, \boldsymbol{b}) > 0$ and $c > 0$ are constants 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:

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

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,*

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

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 and 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 Section 8.3). However, the feasible set $\mathcal{P}$ is non-convex,

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. $\|\Theta^{(t+1)}\|_\infty \le \alpha$, 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. $\|\Theta^{(t+1)}\|_\infty \le \alpha$.

  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 $\Theta^{(t+1)}$: $\boldsymbol{b}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{b} \in \mathcal{B}} \mathcal{L}_{\mathcal{Y},\Omega}\big(\Theta^{(t+1)}, \boldsymbol{b}\big)$.

end for

return $(\hat{\Theta}, \hat{\boldsymbol{b}})$

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.

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 $\mathcal{L}_{\mathcal{Y},\Omega}(\hat{\Theta}) \ge \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}})$. 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.

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.

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

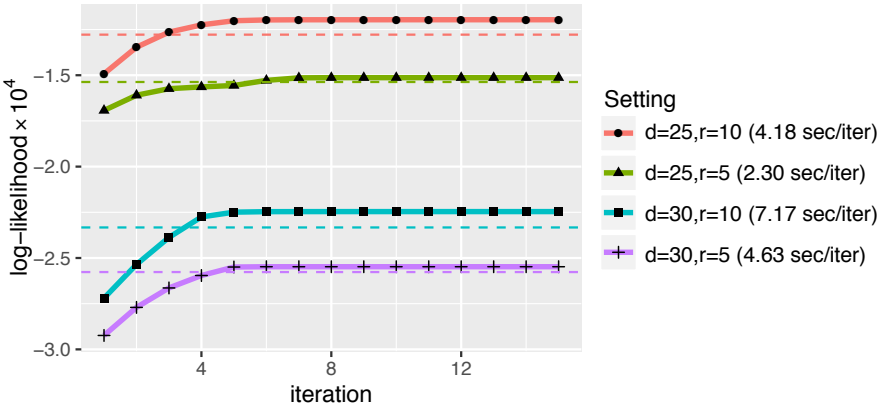


Figure 1: Trajectory of objective function with various d and r.

where $\hat{\Theta}(\boldsymbol{r}), \hat{\boldsymbol{b}}(\boldsymbol{r})$ are the estimates given the rank $\boldsymbol{r}$, and $p_e(\boldsymbol{r}) \stackrel{\text{def}}{=} \sum_k (d_k - r_k) r_k + \prod_k r_k$ 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 $\mathcal{N}(0, 1)$, 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 $n_{\text{sim}} = 30$ 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 and 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.

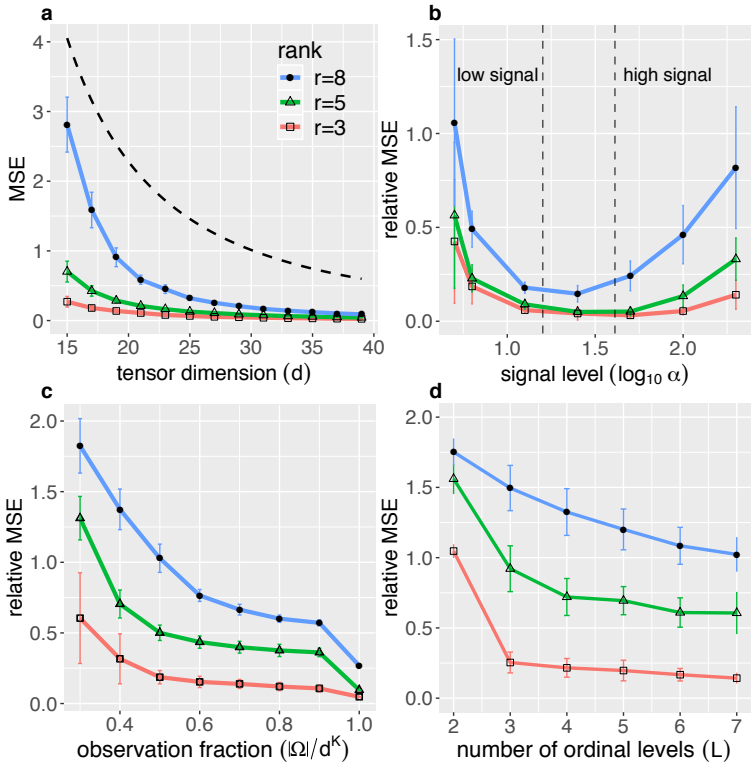


Figure 2: Empirical relationship between (relative) MSE versus (a) dimension d, (b) signal level α, (c) observation fraction ρ, and (d) number of ordinal levels L. In panels (b)-(d), we plot the relative MSE $= \|\hat{\Theta} - \Theta^{\text{true}}\|_F / \|\Theta^{\text{true}}\|_F$ for better visualization.

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, $\text{MAD}(\mathcal{Y}^{\text{mode}}, \hat{\mathcal{Y}}^{\text{mode}}) = d^{-K} \sum_\omega |y^{\text{mode}}_\omega - \hat{y}^{\text{mode}}_\omega|$, and misclassification rate, $\text{MCR}(\mathcal{Y}^{\text{mode}}, \hat{\mathcal{Y}}^{\text{mode}}) = d^{-K} \sum_\omega \mathbb{1}_{\{y^{\text{mode}}_\omega \neq \text{round}(\hat{y}^{\text{mode}}_\omega)\}}$, where round$(\cdot)$ denotes the nearest integer of the prediction (possibly continuous-valued returned by **Continuous-T**). Both metrics are widely used for evaluation of prediction accuracy. 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 labels, $y^{\text{median}}_\omega = \min\{\ell : \mathbb{P}(y_\omega = \ell) \geq 0.5\}$. Under the latent variable model (4) and Assumption 1, the median label is the quantized θ_ω without noise; i.e. $y^{\text{median}}_\omega = \sum_\ell \mathbb{1}_{\theta_\omega \in (b_{\ell-1}, b_\ell]}$. We utilize the same simulation setting as in the earlier experiment. Figure 4 shows that our method outperforms the others in both MCR and MAD. The improved accuracy comes from the incorporation of multilinear low-rank structure, multi-level observations, and the ordinal structure. Interestingly, for the three multilevel methods (**1bit-sign-T**, **Ordinal-M**, and **Ordinal-T**), the median estimator tends to yield smaller MAD than the mode estimator, $\text{MAD}(\mathcal{Y}^{\text{median}}, \hat{\mathcal{Y}}^{\text{median}}) \leq \text{MAD}(\mathcal{Y}^{\text{mode}}, \hat{\mathcal{Y}}^{\text{mode}})$ (Figures 3a-b vs. Figures 4a-b). On the other hand, the mode estimator tends to yield smaller MCR than the median estimator, $\text{MCR}(\mathcal{Y}^{\text{mode}}, \hat{\mathcal{Y}}^{\text{mode}}) \leq \text{MCR}(\mathcal{Y}^{\text{median}}, \hat{\mathcal{Y}}^{\text{median}})$ (Figures 3c-d vs. Figures 4c-d). This tendency is from the property that the median estimator $\hat{y}^{(\text{median})}_\omega$ minimizes $g_1(z) = \mathbb{E}_{y_\omega} |y_\omega - z|$, whereas the mode estimator $\hat{y}^{(\text{mode})}_\omega$ minimizes $g_2(z) = \mathbb{E}_{y_\omega} \mathbb{1}_{\{y_\omega = z\}}$. Here the expectation is over the categorical distribution of y_ω given parameters $\hat{\Theta}$ and $\hat{\boldsymbol{b}}$.

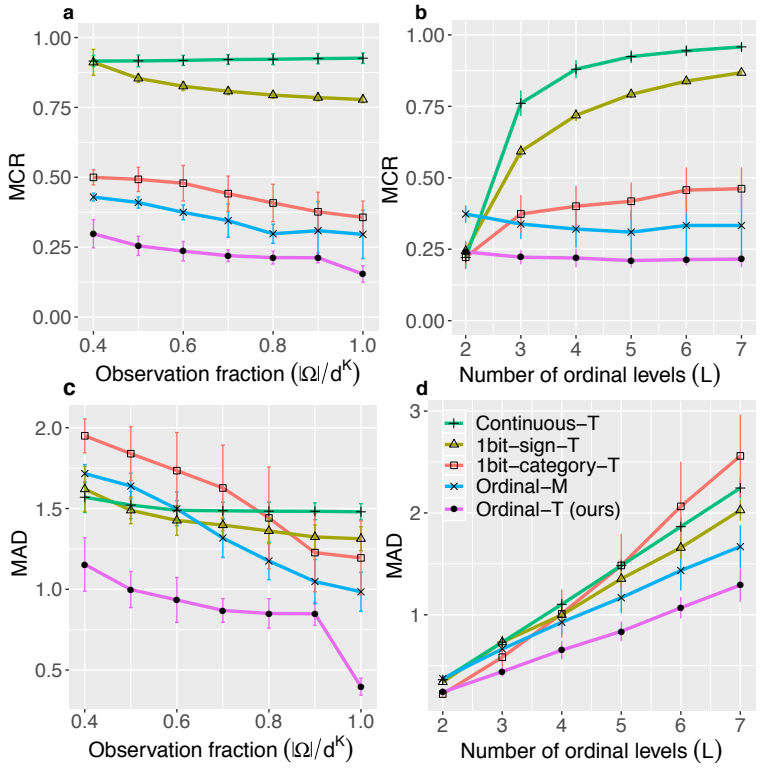


Figure 3: Performance comparison for predicting most likely labels. (a, c) Prediction error versus sample complexity $\rho = |\Omega|/d^K$ when $L = 5$. (b, d) Prediction error versus the number of ordinal levels L when $\rho = 0.8$.

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) (Van Essen et al., 2013). 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).

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 $\mathcal{Y} \in [3]^{68\times68\times136}$ 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 $\{\hat{\boldsymbol{M}}_k\}$, we perform a clustering analysis via K-mean on the brain nodes (see Appendix B.1 for detailed procedure). The 68 brain nodes are grouped into eight clusters. We find that the clustering capture the spatial separation between brain regions (Supplementary Figure S2). In particular, the top two clusters represent the left and right hemispheres; the smaller clusters represent local regions driving by similar nodes (Supplementary Table S1). For example, the cluster III/IV consists of nodes in the supramarginal gyrus region in the left/right hemisphere. This region is known to be involved in visual word recognition and reading (Stoeckel et al., 2009). The spatial separation between clusters suggests the applicability of our clustering method even without knowledge of external annotations.

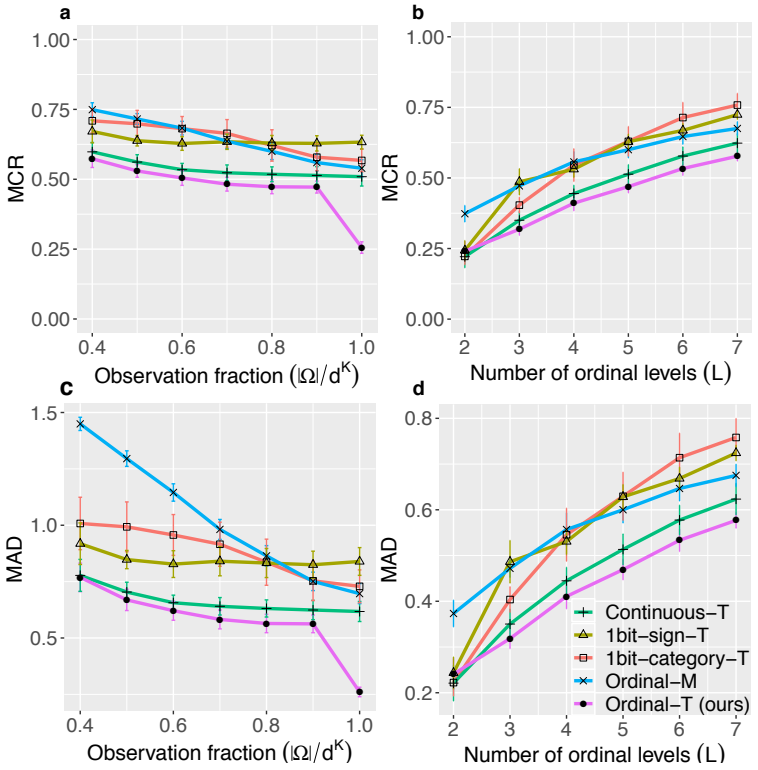


Figure 4: Performance comparison for predicting median labels. (a, c) Prediction error versus sample complexity $\rho = |\Omega|/d^K$ when $L = 5$. (b, d) Prediction error versus the number of ordinal levels L, when $\rho = 0.8$.

We also compare the goodness-of-fit of various tensor methods on the HPC data. We perform 5-fold cross-validation while preserving the same label proportions in train/test sets. Table 2 summarizes the prediction error averaged over 10 runs. Our method outperforms the others, especially in MAD.

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.

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

8 Proofs

Here, we provide proofs of the theoretical results presented in Sections 4.

8.1 Estimation error for tensor denoising

Proof of Theorem 4.1. We suppress the subscript Ω in the proof, because the tensor denoising assumes complete observation $\Omega = [d_1] \times \cdots \times [d_K]$. It follows from the expression of $\mathcal{L}_{\mathcal{Y}}(\Theta)$ that

$$\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial \theta_\omega} = \sum_{\ell \in [L]} \mathbb{1}_{\{y_\omega = \ell\}} \frac{\dot{g}_\ell(\theta_\omega)}{g_\ell(\theta_\omega)},$$
$$\frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial \theta_\omega^2} = \sum_{\ell \in [L]} \mathbb{1}_{\{y_\omega = \ell\}} \frac{\ddot{g}_\ell(\theta_\omega) g_\ell(\theta_\omega) - \dot{g}_\ell^2(\theta_\omega)}{g_\ell^2(\theta_\omega)} \text{ and } \quad \frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial \theta_\omega \theta_{\omega'}'} = 0 \text{ if } \omega \neq \omega', \tag{12}$$

for all $\omega \in [d_1] \times \cdots \times [d_K]$. Define $d_{\text{total}} = \prod_k d_k$. Let $\nabla_\Theta \mathcal{L}_{\mathcal{Y}} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote the tensor of gradient with respect to $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, and $\nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}$ the corresponding Hession matrix of size d_{total}-by-d_{total}. Here, $\text{Vec}(\cdot)$ denotes the operation that turns a tensor into a vector. By (12), $\nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}$ is a diagonal matrix. Recall that

$$U_\alpha = \max_{\ell \in [L], |\theta| \leq \alpha} \frac{|\dot{g}_\ell(\theta)|}{g_\ell(\theta)} > 0 \quad \text{and} \quad L_\alpha = \min_{\ell \in [L], |\theta| \leq \alpha} \frac{\dot{g}_\ell^2(\theta) - \ddot{g}_\ell(\theta) g_\ell(\theta)}{g_\ell^2(\theta)} > 0. \tag{13}$$

Therefore, the entries in $\nabla_\Theta \mathcal{L}_{\mathcal{Y}}$ are upper bounded in magnitude by $U_\alpha > 0$, and all diagonal entries in $\nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}$ are upper bounded by $-L_\alpha < 0$.

By the second-order Taylor's expansion of $\mathcal{L}_{\mathcal{Y}}(\Theta)$ around Θ^{true}, we obtain

$$\mathcal{L}_{\mathcal{Y}}(\Theta) = \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}) + \langle \text{Vec}(\nabla_\Theta \mathcal{L}_{\mathcal{Y}}),\ \text{Vec}(\Theta - \Theta^{\text{true}}) \rangle + \frac{1}{2} \text{Vec}(\Theta - \Theta^{\text{true}})^T \nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}(\check{\Theta}) \text{Vec}(\Theta - \Theta^{\text{true}}), \tag{14}$$

$\check{\Theta} = \gamma \Theta^{\text{true}} + (1 - \gamma)\Theta$ for some $\gamma \in [0, 1]$, and $\nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}(\check{\Theta})$ denotes the $\prod_k d_k$-by-$\prod_k d_k$ Hession matrix evaluated at $\check{\Theta}$.

We first bound the linear term in (14). Note that, by Lemma 4,

$$|\langle \text{Vec}(\nabla_\Theta \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}), \text{Vec}(\Theta - \Theta^{\text{true}}) \rangle| \leq \|\nabla_\Theta \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}})\|_\sigma \|\Theta - \Theta^{\text{true}}\|_*, \tag{15}$$

where $\|\cdot\|_\sigma$ denotes the tensor spectral norm and $\|\cdot\|_*$ denotes the tensor nuclear norm. Define

$$s_\omega = \frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial \theta_\omega}\Big|_{\Theta = \Theta^{\text{true}}} \quad \text{for all } \omega \in [d_1] \times \cdots \times [d_K].$$

Based on (12) and the definition of U_α, $\nabla_\Theta \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}) = [\![s_\omega]\!]$ is a random tensor whose entries are independently distributed satisfying

$$\mathbb{E}(s_\omega) = 0, \quad |s_\omega| \leq U_\alpha, \quad \text{for all } \omega \in [d_1] \times \cdots \times [d_K]. \tag{16}$$

By lemma 6, with probability at least $1 - \exp(-C_1 \sum_k d_k)$, we have

$$\|\nabla_\Theta \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}})\|_\sigma \leq C_2 U_\alpha \sqrt{\sum_k d_k}, \tag{17}$$

where C_1, C_2 are two positive constants that depend only on K. Furthermore, note that $\text{rank}(\Theta) \leq \boldsymbol{r}$, $\text{rank}(\Theta^{\text{true}}) \leq \boldsymbol{r}$, so $\text{rank}(\Theta - \Theta^{\text{true}}) \leq 2\boldsymbol{r}$. By lemma 3, $\|\Theta - \Theta^{\text{true}}\|_* \leq (2r_{\max})^{\frac{K-1}{2}} \|\Theta - \Theta^{\text{true}}\|_F$. Combining (15), (16) and (17), we have that, with probability at least $1 - \exp(-C_1 \sum_k d_k)$,

$$|\langle \text{Vec}(\nabla_\Theta \mathcal{L}_\mathcal{Y}(\Theta^{\text{true}})), \text{Vec}(\Theta - \Theta^{\text{true}}) \rangle| \leq C_2 U_\alpha \sqrt{r_{\max}^{K-1} \sum_k d_k} \|\Theta - \Theta^{\text{true}}\|_F. \tag{18}$$

We next bound the quadratic term in (14). Note that

$$\begin{aligned} \text{Vec}(\Theta - \Theta^{\text{true}})^T \nabla^2_\Theta \mathcal{L}_\mathcal{Y}(\check{\Theta}) \, \text{Vec}(\Theta - \Theta^{\text{true}}) &= \sum_\omega \left(\frac{\partial^2 \mathcal{L}_\mathcal{Y}}{\partial \theta_\omega^2} \Big|_{\Theta = \check{\Theta}} \right) (\theta_\omega - \theta_{\text{true},\omega})^2 \\ &\leq -L_\alpha \sum_\omega (\Theta_\omega - \Theta_{\text{true},\omega})^2 \\ &= -L_\alpha \|\Theta - \Theta^{\text{true}}\|_F^2, \end{aligned} \tag{19}$$

where the second line comes from the fact that $\|\check{\Theta}\|_\infty \leq \alpha$ and the definition of L_α.

Combining (14), (18) and (19), we have that, for all $\Theta \in \mathcal{P}$, with probability at least $1 - \exp(-C_1 \sum_k d_k)$,

$$\mathcal{L}_\mathcal{Y}(\Theta) \leq \mathcal{L}_\mathcal{Y}(\Theta^{\text{true}}) + C_2 U_\alpha \left(r_{\max}^{K-1} \sum_k d_k \right)^{1/2} \|\Theta - \Theta^{\text{true}}\|_F - \frac{L_\alpha}{2} \|\Theta - \Theta^{\text{true}}\|_F^2.$$

In particular, the above inequality also holds for $\hat{\Theta} \in \mathcal{P}$. Therefore,

$$\mathcal{L}_\mathcal{Y}(\hat{\Theta}) \leq \mathcal{L}_\mathcal{Y}(\Theta^{\text{true}}) + C_2 U_\alpha \left(r_{\max}^{K-1} \sum_k d_k \right)^{1/2} \|\hat{\Theta} - \Theta^{\text{true}}\|_F - \frac{L_\alpha}{2} \|\hat{\Theta} - \Theta^{\text{true}}\|_F^2.$$

Since $\hat{\Theta} = \arg\max_{\Theta \in \mathcal{P}} \mathcal{L}_\mathcal{Y}(\Theta)$, $\mathcal{L}_\mathcal{Y}(\hat{\Theta}) - \mathcal{L}_\mathcal{Y}(\Theta^{\text{true}}) \geq 0$, which gives

$$C_2 U_\alpha \left(r_{\max}^{K-1} \sum_k d_k \right)^{1/2} \|\hat{\Theta} - \Theta^{\text{true}}\|_F - \frac{L_\alpha}{2} \|\hat{\Theta} - \Theta^{\text{true}}\|_F^2 \geq 0.$$

Henceforth,

$$\frac{1}{\sqrt{\prod_k d_k}} \|\hat{\Theta} - \Theta^{\text{true}}\|_F \leq \frac{2 C_2 U_\alpha \sqrt{r_{\max}^{K-1} \sum_k d_k}}{L_\alpha \sqrt{\prod_k d_k}} = \frac{2 C_2 U_\alpha r_{\max}^{(K-1)/2}}{L_\alpha} \sqrt{\frac{\sum_k d_k}{\prod_k d_k}}.$$

This completes the proof. □

Proof of Corollary 4.1. The result follows immediately from Theorem 4.1 and Lemma 8. □

Proof of Theorem 4.2. Let $d_{\text{total}} = \prod_{k \in [K]} d_k$, and $\gamma \in [0, 1]$ be a constant to be specified later. Our strategy is to construct a finite set of tensors $\mathcal{X} = \{\Theta_i : i = 1, \ldots\} \subset \mathcal{P}$ satisfying the properties of (i)-(iv) in Lemma 9. By Lemma 9, such a subset of tensors exist. For any tensor $\Theta \in \mathcal{X}$, let $\mathbb{P}_\Theta$ denote the distribution of $\mathcal{Y}|\Theta$, where $\mathcal{Y}$ is the ordinal tensor. In particular, $\mathbb{P}_\mathbf{0}$ is the distribution of

$\mathcal{Y}$ induced by the zero parameter tensor $\mathbf{0}$, i.e., the distribution of $\mathcal{Y}$ conditional on the parameter tensor $\Theta = \mathbf{0}$. Based on the Remark for Lemma 8, we have

$$\mathrm{KL}(\mathbb{P}_\Theta || \mathbb{P}_{\mathbf{0}}) \leq C \|\Theta\|_F^2, \tag{20}$$

where $C = \frac{(4L-6)\dot{f}^2(0)}{A_\alpha} > 0$ is a constant independent of the tensor dimension and rank. Combining the inequality (20) with property (iii) of $\mathcal{X}$, we have

$$\mathrm{KL}(\mathbb{P}_\Theta || \mathbb{P}_{\mathbf{0}}) \leq \gamma^2 R_{\max} d_{\max}. \tag{21}$$

From (21) and the property (i), we deduce that the condition

$$\frac{1}{\mathrm{Card}(\mathcal{X}) - 1} \sum_{\Theta \in \mathcal{X}} \mathrm{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq \varepsilon \log_2 \{\mathrm{Card}(\mathcal{X}) - 1\} \tag{22}$$

holds for any $\varepsilon \geq 0$ when $\gamma \in [0, 1]$ is chosen to be sufficiently small depending on ε, e.g., $\gamma \leq \sqrt{\frac{\varepsilon \log 2}{8}}$. By applying Lemma 11 to (22), and in view of the property (iv), we obtain that

$$\inf_{\hat{\Theta}} \sup_{\Theta^{\mathrm{true}} \in \mathcal{X}} \mathbb{P}\left(\|\hat{\Theta} - \Theta^{\mathrm{true}}\|_F \geq \frac{\gamma}{8} \min\left\{ \alpha \sqrt{d_{\mathrm{total}}}, C^{-1/2} \sqrt{R_{\max} d_{\max}} \right\} \right) \geq \frac{1}{2} \left(1 - 2\varepsilon - \sqrt{\frac{16\varepsilon}{R_{\max} d_{\max} \log 2}} \right). \tag{23}$$

Note that $\mathrm{Loss}(\hat{\Theta}, \Theta^{\mathrm{true}}) = \|\hat{\Theta} - \Theta^{\mathrm{true}}\|_F^2 / d_{\mathrm{total}}$ and $\mathcal{X} \subset \mathcal{P}$. By taking $\varepsilon = 1/10$ and $\gamma = 1/11$, we conclude from (23) that

$$\inf_{\hat{\Theta}} \sup_{\Theta^{\mathrm{true}} \in \mathcal{P}} \mathbb{P}\left(\mathrm{Loss}(\hat{\Theta}, \Theta^{\mathrm{true}}) \geq c \min\left\{ \alpha^2, \frac{C^{-1} R_{\max} d_{\max}}{d_{\mathrm{total}}} \right\} \right) \geq \frac{1}{2} \left(\frac{4}{5} - \sqrt{\frac{1.6}{R_{\max} d_{\max} \log 2}} \right) \geq \frac{1}{8},$$

where $c = \frac{1}{88^2}$ and the last inequality comes from the condition for $d_{\max}$. This completes the proof. □

8.2 Sample complexity for tensor completion

Proof of Theorem 4.3. For notational convenience, we use $\|\Theta\|_{F,\Omega} = \sum_{\omega \in \Omega} \Theta_\omega^2$ to denote the sum of squared entries over the observed set Ω, for a tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$.

Following a similar argument as in the proof of Theorem 4.1, we have

$$\mathcal{L}_{\mathcal{Y},\Omega}(\Theta) = \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\mathrm{true}}) + \langle \mathrm{Vec}(\nabla_\Theta \mathcal{L}_{\mathcal{Y},\Omega}),\ \mathrm{Vec}(\Theta - \Theta^{\mathrm{true}}) \rangle + \frac{1}{2} \mathrm{Vec}(\Theta - \Theta^{\mathrm{true}})^T \nabla_{\check{\Theta}}^2 \mathcal{L}_{\mathcal{Y},\Omega}(\check{\Theta}) \mathrm{Vec}(\Theta - \Theta^{\mathrm{true}}), \tag{24}$$

where

1. $\nabla_\Theta \mathcal{L}_{\mathcal{Y},\Omega}$ is a $d_1 \times \cdots \times d_K$ tensor with $|\Omega|$ nonzero entries, and each entry is upper bounded by $U_\alpha > 0$.
2. $\nabla_\Theta^2 \mathcal{L}_{\mathcal{Y},\Omega}$ is a diagonal matrix of size d_{total}-by-d_{total} with $|\Omega|$ nonzero entries, and each entry is upper bounded by $-L_\alpha < 0$.

Similar to (15) and (19), we have

$$|\langle \mathrm{Vec}(\nabla_\Theta \mathcal{L}_{\mathcal{Y},\Omega}),\ \mathrm{Vec}(\Theta - \Theta^{\mathrm{true}}) \rangle| \leq C_2 U_\alpha \sqrt{r_{\max}^{K-1} \sum_k d_k} \|\Theta - \Theta^{\mathrm{true}}\|_{F,\Omega}$$

and

$$\mathrm{Vec}(\Theta - \Theta^{\mathrm{true}})^T \nabla^2_{\Theta} \mathcal{L}_{\mathcal{Y}}(\check{\Theta}) \,\mathrm{Vec}(\Theta - \Theta^{\mathrm{true}}) \le -L_\alpha \|\Theta - \Theta^{\mathrm{true}}\|^2_{F,\Omega}. \tag{25}$$

Combining (24)-(25) with the fact that $\mathcal{L}_{\mathcal{Y},\Omega}(\hat{\Theta}) \ge \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\mathrm{true}})$, we have

$$\|\hat{\Theta} - \Theta^{\mathrm{true}}\|_{F,\Omega} \le \frac{2C_2 U_\alpha r_{\max}^{(K-1)/2}}{L_\alpha} \sqrt{\sum_k d_k}. \tag{26}$$

Lastly, we invoke the result regarding the closeness of Θ to its sampled version Θ_Ω, under the entrywise bound condition. Note that $\|\hat{\Theta} - \Theta^{\mathrm{true}}\|_\infty \le 2\alpha$ and $\mathrm{rank}(\hat{\Theta} - \Theta^{\mathrm{true}}) \le 2\boldsymbol{r}$. By Lemma 2, $\|\hat{\Theta} - \Theta^{\mathrm{true}}\|_M \le 2^{(3K-1)/2}\alpha \left(\frac{\prod r_k}{r_{\max}}\right)^{3/2}$. Therefore, the condition in Lemma 12 holds with $\beta = 2^{(3K-1)/2}\alpha \left(\frac{\prod r_k}{r_{\max}}\right)^{3/2}$. Applying Lemma 12 to (26) gives

$$\begin{aligned}
\|\hat{\Theta} - \Theta^{\mathrm{true}}\|^2_{F,\Pi} &\le \frac{1}{m}\|\hat{\Theta} - \Theta^{\mathrm{true}}\|^2_{F,\Omega} + c\beta\sqrt{\frac{\sum_k d_k}{|\Omega|}} \\
&\le C_2 r_{\max}^{K-1}\frac{\sum_k d_k}{|\Omega|} + C_1 \alpha r_{\max}^{3(K-1)/2}\sqrt{\frac{\sum_k d_k}{|\Omega|}},
\end{aligned}$$

with probability at least $1 - \exp(-\frac{\sum_k d_k}{\sum_k \log d_k})$ over the sampled set Ω. Here $C_1, C_2 > 0$ are two constants independent of the tensor dimension and rank. Therefore,

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

provided that $r_{\max} = O(1)$. □

8.3 Convexity of the log-likelihood function

Theorem 8.1. *Define the function*

$$\mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b}) = \sum_{\omega\in\Omega}\sum_{\ell\in[L]} \left\{\mathbb{1}_{\{y_\omega = \ell\}} \log\left[f(b_\ell - \theta_\omega) - f(b_{\ell-1} - \theta_\omega)\right]\right\}, \tag{27}$$

where $f(\cdot)$ satisfies Assumption 1. Then, $\mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b})$ is concave in $(\Theta, \boldsymbol{b})$.

Proof. Define $d_{\mathrm{total}} = \prod_k d_k$. By abuse of notation, we use $(\Theta, \boldsymbol{b})$ to denote the length-$(d_{\mathrm{total}}+L-1)$-vector collecting all parameters together. Let us denote a bivariate function

$$\begin{aligned}
\lambda :&\mathbb{R}^2 \mapsto \mathbb{R} \\
(u, v) &\mapsto \lambda(u, v) = \log\left[f(u) - f(v)\right].
\end{aligned}$$

It suffices to show that $\lambda(u, v)$ is concave in (u, v) where $u > v$.

Suppose that the claim holds (which we will prove in the next paragraph). Based on (27), u, v are both linear functions of $(\Theta, \boldsymbol{b})$:

$$u = \boldsymbol{a}_1^T(\Theta, \boldsymbol{b}), \quad v = \boldsymbol{a}_2^T(\Theta, \boldsymbol{b}), \quad \text{for some } \boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^{d_{\mathrm{total}}+L-1}.$$

Then, $\lambda(u,v) = \lambda(\boldsymbol{a}_1^T(\Theta,\boldsymbol{b}),\ \boldsymbol{a}_2^T(\Theta,\boldsymbol{b}))$ is concave in $(\Theta,\boldsymbol{b})$ by the definition of concavity. Therefore, we can conclude that $\mathcal{L}_{\mathcal{Y},\Omega}(\Theta,\boldsymbol{b})$ is concave in $(\Theta,\boldsymbol{b})$ because $\mathcal{L}_{\mathcal{Y},\Omega}(\Theta,\boldsymbol{b})$ is the sum of $\lambda(u,v)$.

Now, we prove the concavity of $\lambda(u,v)$. Note that

$$\lambda(u,v) = \log\big[f(u) - f(v)\big] = \log\Big[\int \mathbb{1}_{[u,v]}(x) f'(x)dx\Big],$$

where $\mathbb{1}_{[u,v]}$ is an indicator function that equals 1 in the interval $[u,v]$, and 0 elsewhere. Furthermore, $\mathbb{1}_{[u,v]}(x)$ is log-concave in (u,v,x), and by Assumption 1, $f'(x)$ is log-concave in x. It follows that $\mathbb{1}_{[u,v]}(x)f'(x)$ is a log-concave in (u,v,x). By Lemma 1, we conclude that $\lambda(u,v)$ is concave in (u,v) where $u > v$. □

Lemma 1 (Corollary 3.5 in Brascamp and Lieb (2002)). *Let $F(x,y) : \mathbb{R}^{m+n} \to \mathbb{R}$ be an integrable function where $x \in \mathbb{R}^m, y \in \mathbb{R}^n$. Let*

$$G(x) = \int_{\mathbb{R}^n} F(x,y)dy.$$

If $F(x,y)$ is log concave in (x,y), then $G(x)$ is log concave in x.

8.4 Auxiliary lemmas

This section collects lemmas that are useful for the proofs of the main theorems.

Definition 1 (Atomic M-norm (Ghadermarzy et al., 2019)). Define $T_\pm = \{\mathcal{T} \in \{\pm 1\}^{d_1\times\cdots\times d_K} : \operatorname{rank}(\mathcal{T}) = 1\}$. The atomic M-norm of a tensor $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is defined as

$$\begin{aligned}\|\Theta\|_M &= \inf\{t > 0 : \Theta \in t\operatorname{conv}(T_\pm)\} \\ &= \inf\left\{\sum_{\mathcal{X}\in T_\pm} c_x : \Theta = \sum_{\mathcal{X}\in T_\pm} c_x\mathcal{X},\ c_x > 0\right\}.\end{aligned}$$

Definition 2 (Spectral norm (Lim, 2005)). The spectral norm of a tensor $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is defined as

$$\|\Theta\|_\sigma = \sup\left\{\langle\Theta, \boldsymbol{x}_1\otimes\cdots\otimes\boldsymbol{x}_K\rangle : \|\boldsymbol{x}_k\|_2 = 1,\ \boldsymbol{x}_k \in \mathbb{R}^{d_k},\ \text{for all } k\in[K]\right\}.$$

Definition 3 (Nuclear norm (Friedland and Lim, 2018)). The nuclear norm of a tensor $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is defined as

$$\|\Theta\|_* = \inf\left\{\sum_{i\in[r]}|\lambda_i| : \Theta = \sum_{i=1}^r \lambda_i \boldsymbol{x}_1^{(i)}\otimes\cdots\otimes\boldsymbol{x}_K^{(i)},\ \|\boldsymbol{x}_k^{(i)}\|_2 = 1,\ \boldsymbol{x}_k^{(i)} \in \mathbb{R}^{d_k},\ \text{for all} k\in[K],\ i\in[r]\right\},$$

where the infimum is taken over all $r \in \mathbb{N}$ and $\|\boldsymbol{x}_k^{(i)}\|_2 = 1$ for all $i \in [r]$ and $k \in [K]$.

Lemma 2 (M-norm and infinity norm (Ghadermarzy et al., 2019)). *Let $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ be an order-K, rank-$(r_1,\ldots,r_K)$ tensor. Then*

$$\|\Theta\|_\infty \le \|\Theta\|_M \le \left(\frac{\prod_k r_k}{r_{\max}}\right)^{\frac{3}{2}}\|\Theta\|_\infty.$$

Lemma 3 (Nuclear norm and F-norm). *Let $\mathcal{A} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ be an order-K tensor with Tucker rank$(\mathcal{A}) = (r_1, \ldots, r_K)$. Then*

$$\|\mathcal{A}\|_* \leq \sqrt{\frac{\prod_k r_k}{\max_k r_k}} \|\mathcal{A}\|_F,$$

where $\|\cdot\|_*$ denotes the nuclear norm of the tensor.

Proof. Without loss of generality, suppose $r_1 = \min_k r_k$. Let $\mathcal{A}_{(k)}$ denote the mode-k matricization of $\mathcal{A}$ for all $k \in [K]$. By Wang et al. (2017, Corollary 4.11), and the invariance relationship between a tensor and its Tucker core (Jiang et al., 2017, Section 6), we have

$$\|\mathcal{A}\|_* \leq \sqrt{\frac{\prod_{k \geq 2} r_k}{\max_{k \geq 2} r_k}} \|\mathcal{A}_{(1)}\|_*, \tag{28}$$

where $\mathcal{A}_{(1)}$ is a d_1-by-$\prod_{k \geq 2} d_k$ matrix with matrix rank r_1. Furthermore, the relationship between the matrix norms implies that $\|\mathcal{A}_{(1)}\|_* \leq \sqrt{r_1}\|\mathcal{A}_{(1)}\|_F = \sqrt{r_1}\|\mathcal{A}\|_F$. Combining this fact with the inequality (28) yields the final claim. □

Lemma 4. *Let $\mathcal{A}$, $\mathcal{B}$ be two order-K tensors of the same dimension. Then*

$$|\langle \mathcal{A}, \mathcal{B} \rangle| \leq \|\mathcal{A}\|_\sigma \|\mathcal{B}\|_*.$$

Proof. By Friedland and Lim (2018, Proposition 3.1), there exists a nuclear norm decomposition of $\mathcal{B}$, such that

$$\mathcal{B} = \sum_r \lambda_r \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)}, \quad \boldsymbol{a}_r^{(k)} \in \mathbf{S}^{d_k - 1}(\mathbb{R}), \quad \text{for all } k \in [K],$$

and $\|\mathcal{B}\|_* = \sum_r |\lambda_r|$. Henceforth we have

$$\begin{aligned} |\langle \mathcal{A}, \mathcal{B} \rangle| &= |\langle \mathcal{A}, \sum_r \lambda_r \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)} \rangle| \leq \sum_r |\lambda_r| |\langle \mathcal{A}, \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)} \rangle| \\ &\leq \sum_r |\lambda_r| \|\mathcal{A}\|_\sigma = \|\mathcal{A}\|_\sigma \|\mathcal{B}\|_*, \end{aligned}$$

which completes the proof. □

The following lemma provides the bound on the spectral norm of random tensors. The result was firstly presented in Nguyen et al. (2015), and we adopt the version from Tomioka and Suzuki (2014).

Lemma 5 (Spectral norm of random tensors (Tomioka and Suzuki, 2014)). *Suppose that $\mathcal{S} = [\![s_\omega]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is an order-K tensor whose entries are independent random variables that satisfy*

$$\mathbb{E}(s_\omega) = 0, \quad \text{and} \quad \mathbb{E}(e^{t s_\omega}) \leq e^{t^2 L^2 / 2}.$$

Then the spectral norm $\|\mathcal{S}\|_\sigma$ satisfies that,

$$\|\mathcal{S}\|_\sigma \leq \sqrt{8L^2 \log(12K) \sum_k d_k + \log(2/\delta)},$$

with probability at least $1 - \delta$.

Lemma 6. *Suppose that $\mathcal{S} = [\![s_\omega]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is an order-K tensor whose entries are independent random variables that satisfy*

$$\mathbb{E}(s_\omega) = 0, \quad \text{and} \quad |s_\omega| \leq U.$$

Then we have

$$\mathbb{P}\left(\|\mathcal{S}\|_\sigma \geq C_2 U \sqrt{\sum_k d_k} \right) \leq \exp\left(-C_1 \log K \sum_k d_k \right)$$

where $C_1 > 0$ is an absolute constant, and $C_2 > 0$ is a constant that depends only on K.

Proof. Note that the random variable $U^{-1} s_\omega$ is zero-mean and supported on $[-1, 1]$. Therefore, $U^{-1} s_\omega$ is sub-Gaussian with parameter $\frac{1-(-1)}{2} = 1$, i.e.

$$\mathbb{E}(U^{-1} s_\omega) = 0, \quad \text{and} \quad \mathbb{E}(e^{tU^{-1}s_\omega}) \leq e^{t^2/2}.$$

It follows from Lemma 5 that, with probability at least $1 - \delta$,

$$\|U^{-1}\mathcal{S}\|_\sigma \leq \sqrt{(c_0 \log K + c_1) \sum_k d_k + \log(2/\delta)},$$

where $c_0, c_1 > 0$ are two absolute constants. Taking $\delta = \exp(-C_1 \log K \sum_k d_k)$ yields the final claim, where $C_2 = c_0 \log K + c_1 + 1 > 0$ is another constant. □

Lemma 7. *Let X, Y be two discrete random variables taking values on L possible categories, with category probabilities $\{p_\ell\}_{\ell \in [L]}$ and $\{q_\ell\}_{\ell \in [L]}$, respectively. Suppose p_ℓ, $q_\ell > 0$ for all $i \in [L]$. Then, the Kullback-Leibler (KL) divergence satisfies that*

$$\mathrm{KL}(X||Y) \stackrel{def}{=} -\sum_{\ell \in [L]} \mathbb{P}_X(\ell) \log\left\{ \frac{\mathbb{P}_Y(\ell)}{\mathbb{P}_X(\ell)} \right\} \leq \sum_{\ell \in [L]} \frac{(p_\ell - q_\ell)^2}{q_\ell}.$$

Proof. Using the fact $\log x \leq x - 1$ for $x > 0$, we have that

$$\begin{aligned}
\mathrm{KL}(X||Y) &= \sum_{\ell \in [L]} p_\ell \log \frac{p_\ell}{q_\ell} \\
&\leq \sum_{\ell \in [L]} \frac{p_\ell}{q_\ell}(p_\ell - q_\ell) \\
&= \sum_{\ell \in [L]} \left(\frac{p_\ell}{q_\ell} - 1 \right)(p_\ell - q_\ell) + \sum_{\ell \in [L]} (p_\ell - q_\ell).
\end{aligned}$$

Note that $\sum_{\ell \in [L]} (p_\ell - q_\ell) = 0$. Therefore,

$$\mathrm{KL}(X||Y) \leq \sum_{\ell \in [L]} \left(\frac{p_\ell}{q_\ell} - 1 \right)(p_\ell - q_\ell) = \sum_{\ell \in [L]} \frac{(p_\ell - q_\ell)^2}{q_\ell}.$$

□

Lemma 8 (KL divergence and F-norm)**.** *Let $\mathcal{Y} \in [L]^{d_1 \times \cdots \times d_K}$ be an ordinal tensor generated from the model (1) with the link function f and parameter tensor Θ. Let $\mathbb{P}_\Theta$ denote the joint categorical distribution of $\mathcal{Y}|\Theta$ induced by the parameter tensor Θ, where $\|\Theta\|_\infty \leq \alpha$. Define*

$$A_\alpha = \min_{\ell \in [L], |\theta| \leq \alpha} \left[f(b_\ell - \theta) - f(b_{\ell-1} - \theta) \right]. \tag{29}$$

Then, for any two tensors Θ, Θ^* in the parameter spaces, we have

$$\mathrm{KL}(\mathbb{P}_\Theta || \mathbb{P}_{\Theta^*}) \leq \frac{2(2L-3)}{A_\alpha} \dot{f}^2(0) \|\Theta - \Theta^*\|_F^2.$$

Proof. Suppose that the distribution over the ordinal tensor $\mathcal{Y} = [\![y_\omega]\!]$ is induced by $\Theta = [\![\theta_\omega]\!]$. Then, based on the generative model (1),

$$\mathbb{P}(y_\omega = \ell | \theta_\omega) = f(b_\ell - \theta_\omega) - f(b_{\ell-1} - \theta_\omega),$$

for all $\ell \in [L]$ and $\omega \in [d_1] \times \cdots \times [d_K]$. For notational convenience, we suppress the subscribe in θ_ω and simply write θ (and respectively, θ^*). Based on Lemma 7 and Taylor expansion,

$$\begin{aligned}
\mathrm{KL}(\theta || \theta^*) &\leq \sum_{\ell \in [L]} \frac{\left[f(b_\ell - \theta) - f(b_{\ell-1} - \theta) - f(b_\ell - \theta^*) + f(b_{\ell-1} - \theta^*) \right]^2}{f(b_\ell - \theta^*) - f(b_{\ell-1} - \theta^*)} \\
&\leq \sum_{\ell=2}^{L-1} \frac{\left[\dot{f}(b_\ell - \eta_\ell) - \dot{f}(b_{\ell-1} - \eta_{\ell-1}) \right]^2}{f(b_\ell - \theta^*) - f(b_{\ell-1} - \theta^*)} (\theta - \theta^*)^2 + \frac{\dot{f}^2(b_1 - \eta_1)}{f(b_1 - \theta^*)} (\theta - \theta^*)^2 \\
&\quad + \frac{\dot{f}^2(b_{L-1} - \eta_{L-1})}{1 - f(b_{L-1} - \theta^*)} (\theta - \theta^*)^2,
\end{aligned}$$

where η_ℓ and $\eta_{\ell-1}$ fall between θ and θ^*. Therefore,

$$\mathrm{KL}(\theta || \theta^*) \leq \left(\frac{4(L-2)}{A_\alpha} + \frac{2}{A_\alpha} \right) \dot{f}^2(0) (\theta - \theta^*)^2 = \frac{2(2L-3)}{A_\alpha} \dot{f}^2(0) (\theta - \theta^*)^2, \tag{30}$$

where we have used Taylor expansion, the bound (29), and the fact that $\dot{f}(\cdot)$ peaks at zero for an unimodal and symmetric function. Now summing (30) over the index set $\omega \in [d_1] \times \cdots \times [d_K]$ gives

$$\mathrm{KL}(\mathbb{P}_\Theta || \mathbb{P}_{\Theta^*}) = \sum_{\omega \in [d_1] \times \cdots \times [d_K]} \mathrm{KL}(\theta_\omega || \theta^*_\omega) \leq \frac{2(2L-3)}{A_\alpha} \dot{f}^2(0) \|\Theta - \Theta^*\|_F^2.$$

□

Remark 2. In particular, let $\mathbb{P}_{\mathbf{0}}$ denote the distribution of $\mathcal{Y}|\mathbf{0}$ induced by the zero parameter tensor. Then we have

$$\mathrm{KL}(\mathbb{P}_\Theta || \mathbb{P}_{\mathbf{0}}) \leq \frac{2(2L-3)}{A_\alpha} \dot{f}^2(0) \|\Theta\|_F^2.$$

Lemma 9. *Assume the same setup as in Theorem 4.2. Without loss of generality, suppose $d_1 = \max_k d_k$. Define $R = \max_k r_k$ and $d_{total} = \prod_{k \in [K]} d_k$. For any constant $0 \leq \gamma \leq 1$, there exist a finite set of tensors $\mathcal{X} = \{\Theta_i : i = 1, \ldots\} \subset \mathcal{P}$ satisfying the following four properties:*

(i) Card$(\mathcal{X}) \geq 2^{Rd_1/8} + 1$, where Card denotes the cardinality;

(ii) $\mathcal{X}$ *contains the zero tensor* $\mathbf{0} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$;

(iii) $\|\Theta\|_\infty \le \gamma \min\left\{\alpha,\ C^{-1/2}\sqrt{\frac{Rd_1}{d_{total}}}\right\}$ *for any element* $\Theta \in \mathcal{X}$;

(iv) $\|\Theta_i - \Theta_j\|_F \ge \frac{\gamma}{4}\min\left\{\alpha\sqrt{d_{total}},\ C^{-1/2}\sqrt{Rd_1}\right\}$ *for any two distinct elements* $\Theta_i,\ \Theta_j \in \mathcal{X}$,

Here $C = C(\alpha, L, f, \boldsymbol{b}) = \frac{(4L-6)f^2(0)}{A_\alpha} > 0$ *is a constant independent of the tensor dimension and rank.*

Proof. Given a constant $0 \le \gamma \le 1$, we define a set of matrices:

$$\mathcal{C} = \left\{\boldsymbol{M} = (m_{ij}) \in \mathbb{R}^{d_1 \times R} : a_{ij} \in \left\{0, \gamma \min\left\{\alpha, C^{-1/2}\sqrt{\frac{Rd_1}{d_{\text{total}}}}\right\}\right\},\ \forall (i,j) \in [d_1] \times [R]\right\}.$$

We then consider the associated set of block tensors:

$$\mathcal{B} = \mathcal{B}(\mathcal{C}) = \{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \Theta = \boldsymbol{A} \otimes \mathbf{1}_{d_3} \otimes \cdots \otimes \mathbf{1}_{d_K}, \\ \text{where } \boldsymbol{A} = (\boldsymbol{M}|\cdots|\boldsymbol{M}|\boldsymbol{O}) \in \mathbb{R}^{d_1 \times d_2},\ \boldsymbol{M} \in \mathcal{C}\},$$

where $\mathbf{1}_d$ denotes a length-d vector with all entries 1, $\boldsymbol{O}$ denotes the $d_1 \times (d_2 - R\lfloor d_2/R \rfloor)$ zero matrix, and $\lfloor d_2/R \rfloor$ is the integer part of d_2/R. In other words, the subtensor $\Theta(\boldsymbol{I}, \boldsymbol{I}, i_3, \ldots, i_K) \in \mathbb{R}^{d_1 \times d_2}$ are the same for all fixed $(i_3, \ldots, i_K) \in [d_3] \times \cdots \times [d_K]$, and furthermore, each subtensor $\Theta(\boldsymbol{I}, \boldsymbol{I}, i_3, \ldots, i_K)$ itself is filled by copying the matrix $\boldsymbol{M} \in \mathbb{R}^{d_1 \times R}$ as many times as would fit.

By construction, any element of $\mathcal{B}$, as well as the difference of any two elements of $\mathcal{B}$, has Tucker rank at most $\max_k r_k \le R$, and the entries of any tensor in $\mathcal{B}$ take values in $[0, \alpha]$. Thus, $\mathcal{B} \subset \mathcal{P}$. By Lemma 10, there exists a subset $\mathcal{X} \subset \mathcal{B}$ with cardinality $\text{Card}(\mathcal{X}) \ge 2^{Rd_1/8} + 1$ containing the zero $d_1 \times \cdots \times d_K$ tensor, such that, for any two distinct elements Θ_i and Θ_j in $\mathcal{X}$,

$$\|\Theta_i - \Theta_j\|_F^2 \ge \frac{Rd_1}{8}\gamma^2 \min\left\{\alpha^2, \frac{C^{-1}Rd_1}{d_{\text{total}}}\right\} \lfloor \frac{d_2}{R} \rfloor \prod_{k \ge 3} d_k \ge \frac{\gamma^2 \min\left\{\alpha^2 d_{\text{total}}, C^{-1}Rd_1\right\}}{16}.$$

In addition, each entry of $\Theta \in \mathcal{X}$ is bounded by $\gamma \min\left\{\alpha, C^{-1/2}\sqrt{\frac{Rd_1}{d_{\text{total}}}}\right\}$. Therefore the Properties (i) to (iv) are satisfied. □

Lemma 10 (Varshamov-Gilbert bound)**.** *Let* $\Omega = \{(w_1, \ldots, w_m) : w_i \in \{0, 1\}\}$. *Suppose* $m > 8$. *Then there exists a subset* $\{w^{(0)}, \ldots, w^{(M)}\}$ *of* Ω *such that* $w^{(0)} = (0, \ldots, 0)$ *and*

$$\|w^{(j)} - w^{(k)}\|_0 \ge \frac{m}{8}, \quad \text{for } 0 \le j < k \le M,$$

where $\|\cdot\|_0$ *denotes the Hamming distance, and* $M \ge 2^{m/8}$.

Definition 4 (Absolute continuity)**.** We use $\lambda \ll \mu$ to denote absolutely continuity of a measure λ with respect to another measure μ; that is, $\lambda(E) = 0$ whenever $\mu(E) = 0$.

Lemma 11 (Theorem 2.5 in Tsybakov (2008))**.** *Assume that a set* $\mathcal{X}$ *contains element* $\Theta_0, \Theta_1, \ldots, \Theta_M$ *($M \ge 2$) such that*

- $d(\Theta_j,\ \Theta_k) \ge 2s > 0$, $\forall 0 \le j \le k \le M$;

- $\mathbb{P}_j \ll \mathbb{P}_0$, $\forall j = 1, \ldots, M$, *and*

$$\frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_j || \mathbb{P}_0) \leq \alpha \log M$$

where $d\colon \mathcal{X} \times \mathcal{X} \mapsto [0, +\infty]$ *is a semi-distance function,* $0 < \alpha < 1/8$ *and* $P_j = P_{\Theta_j}$, $j = 0, 1 \ldots, M$.

Then

$$\inf_{\hat{\Theta}} \sup_{\Theta \in \mathcal{X}} \mathbb{P}_\Theta(d(\hat{\Theta}, \Theta) \geq s) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right) > 0.$$

Lemma 12 (Lemma 28 in Ghadermarzy et al. (2019))**.** *Define* $\mathbb{B}_M(\beta) = \{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \|\Theta\|_M \leq \beta\}$. *Let* $\Omega \subset [d_1] \times \cdots \times [d_K]$ *be a random set with* $m = |\Omega|$, *and assume that each entry in* Ω *is drawn with replacement from* $[d_1] \times \cdots \times [d_K]$ *using probability* Π. *Define*

$$\|\Theta\|_{F,\Pi}^2 = \frac{1}{m}\mathbb{E}_{\Omega \in \Pi}\|\Theta\|_{F,\Omega}^2.$$

Then, there exists a universal constant $c > 0$, *such that, with probability at least* $1 - \exp\left(-\frac{\sum_k d_k}{\sum_k \log d_k}\right)$ *over the sampled set* Ω,

$$\frac{1}{m}\|\Theta\|_{F,\Omega}^2 \geq \|\Theta\|_{F,\Pi}^2 - c\beta\sqrt{\frac{\sum_k d_k}{m}}$$

holds uniformly for all $\Theta \in \mathbb{B}_M(\beta)$.

9 Conclusion

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/tensor problem, they are often slow in practice (Ge and Ma, 2017; Chen et al., 2019). The interplay between computational efficiency and statistical accuracy in general tensor problems warrants future research.

Acknowledgments

This research is supported in part by NSF grant DMS-1915978 and the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

Appendix

A Extension of Theorem 4.1 to unknown cut-off points

When the cut-off points $\boldsymbol{b}$ is unknown, we estimate $(\hat{\Theta}, \hat{\boldsymbol{b}})$ by

$$(\hat{\Theta}, \hat{\boldsymbol{b}}) = \underset{(\Theta, \boldsymbol{b}) \in \mathcal{P} \times \mathcal{B}}{\arg\max}\, \mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b}), \tag{31}$$

where

$$\mathcal{P} = \{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \operatorname{rank}(\mathcal{P}) \le \boldsymbol{r},\ \|\Theta\|_\infty \le \alpha\}, \quad \mathcal{B} = \{\boldsymbol{b} \in \mathbb{R}^{L-1} : \|\boldsymbol{b}\|_\infty \le \beta,\ \min_\ell (b_\ell - b_{\ell-1}) \ge \Delta\}.$$

The estimation accuracy is assessed using the mean squared error (MSE):

$$\mathrm{MSE}\left((\hat{\Theta}, \hat{\boldsymbol{b}}), (\Theta, \boldsymbol{b})\right) = \frac{1}{\prod_k d_k + L - 1} \|(\hat{\Theta}, \hat{\boldsymbol{b}}) - (\Theta, \boldsymbol{b})\|_F^2$$

We introduce several quantities that will be used in our theory:

1. We make the convention that $b_0 = -\infty$, $b_L = \infty$, $f(-\infty) = 0$, $f(\infty) = 1$, and $\dot{f}(-\infty) = \ddot{f}(-\infty) = \dot{f}(\infty) = \ddot{f}(\infty) = 0$.
2. The difference function $g_\ell(\theta)$ is defined as $g_\ell(\theta) = f(b_\ell - \theta) - f(b_{\ell-1} - \theta)$ for all $\theta \in \mathbb{R}$ and $\ell \in [L]$.
3. Define $n_\ell = \sum_{\omega \in \Omega} \mathbb{1}_{\{y_\omega = \ell\}}$, i.e., the number of tensor entries taking value on $\ell \in [L]$.
4. With a little abuse of notation, we re-define the constants in (13) as

$$U_{\alpha,\beta,\Delta} = \max_{\substack{|\theta| \le \alpha, \\ \boldsymbol{b} \in \mathcal{B}}} \max_{\ell \in [L-1]} \frac{|\dot{g}_\ell(\theta)|}{g_\ell(\theta)}, \quad \text{and} \quad L_{\alpha,\beta,\Delta} = \min_{\substack{|\theta| \le \alpha, \\ \boldsymbol{b} \in \mathcal{B}}} \min_{\ell \in [L-1]} \frac{\dot{g}_\ell^2(\theta) - \ddot{g}_\ell(\theta) g_\ell(\theta)}{g_\ell^2(\theta)}. \tag{32}$$

5. We define three additional constants:

$$\begin{aligned}
C_{\alpha,\beta,\Delta} &= \max_{\substack{|\theta| \le \alpha, \\ \boldsymbol{b} \in \mathcal{B}}} \max_{\ell \in [L-1]} \max \left\{ \frac{\dot{f}(b_\ell - \theta)}{g_\ell(\theta)}, \frac{\dot{f}(b_\ell - \theta)}{g_{\ell+1}(\theta)} \right\}, \\
D_{\alpha,\beta,\Delta} &= \min_{\substack{|\theta| \le \alpha, \\ \boldsymbol{b} \in \mathcal{B}}} \min_{\ell \in [L-1]} \min \left\{ \frac{\partial}{\partial \theta} \left(\frac{\dot{f}(b_\ell - \theta)}{g_\ell(\theta)} \right), -\frac{\partial}{\partial \theta} \left(\frac{\dot{f}(b_\ell - \theta)}{g_{\ell+1}(\theta)} \right) \right\} \\
&= \min_{\substack{|\theta| \le \alpha, \\ \boldsymbol{b} \in \mathcal{B}}} \min_{\ell \in [L-1]} \min \left\{ -\frac{\ddot{f}(b_\ell - \theta) g_\ell(\theta) - \dot{f}(b_\ell - \theta) \dot{g}_\ell(\theta)}{g_\ell^2(\theta)}, \right. \\
&\qquad \left. \frac{\ddot{f}(b_\ell - \theta) g_{\ell+1}(\theta) - \dot{f}(b_\ell - \theta) \dot{g}_{\ell+1}(\theta)}{g_{\ell+1}^2(\theta)} \right\} \\
\tilde{D}_{\alpha,\beta,\Delta} &= \max_{\substack{|\theta| \le \alpha, \\ \boldsymbol{b} \in \mathcal{B}}} \max_{\ell \in [L-1]} \max \left\{ \frac{\partial}{\partial \theta} \left(\frac{\dot{f}(b_\ell - \theta)}{g_\ell(\theta)} \right), -\frac{\partial}{\partial \theta} \left(\frac{\dot{f}(b_\ell - \theta)}{g_{\ell+1}(\theta)} \right) \right\}.
\end{aligned} \tag{33}$$

We make the following assumptions about the link function.

Assumption 2. *The link function $f\colon \mathbb{R} \mapsto [0,1]$ satisfies the following properties:*

1. *$f(\theta)$ is twice-differentiable and strictly increasing in θ.*
2. *$\dot{f}(\theta)$ is strictly log-concave and symmetric with respect to $\theta = 0$.*
3. *The function $\frac{\dot{f}(b_\ell - \theta)}{g_\ell(\theta)}$ is strictly increasing with respect to θ for all $\boldsymbol{b} \in \mathcal{B}$.*
4. *The function $\frac{\dot{f}(b_\ell - \theta)}{g_{\ell+1}(\theta)}$ is strictly decreasing with respect to θ for all $\boldsymbol{b} \in \mathcal{B}$.*

Remark 3. The condition $\Delta = \min_\ell(b_\ell - b_{\ell-1}) > 0$ on the feasible set $\mathcal{B}$ guarantees the strict positiveness of $g_\ell(\theta) = f(b_\ell - \theta) - f(b_{\ell-1} - \theta)$. Therefore, the denominators in the above quantities $U_{\alpha,\beta,\Delta}, L_{\alpha,\beta,\Delta}, C_{\alpha,\beta,\Delta}, D_{\alpha,\beta,\Delta}, \tilde{D}_{\alpha,\beta,\Delta}$ are well-defined. Furthermore, by Theorem 8.1, $g_\ell(\theta)$ is strictly log-concave in θ for all $\boldsymbol{b} \in \mathcal{B}$. Based on Assumption 2 and closeness of the feasible set, we have $U_{\alpha,\beta,\Delta} > 0$, $L_{\alpha,\beta,\Delta} > 0$, $C_{\alpha,\beta,\Delta} > 0$, $D_{\alpha,\beta,\Delta} > 0$, $\tilde{D}_{\alpha,\beta,\Delta} > 0$.

Remark 4. In particular, for logistic link $f(x) = \frac{1}{1+e^{-x}}$, we have

$$C_{\alpha,\beta,\Delta} = \max_{\substack{|\theta| \le \alpha, \\ \boldsymbol{b}\in\mathcal{B}}} \max_{\ell\in[L-1]} \max \left\{ \frac{1}{e^{b_\ell - b_{\ell-1}} - 1} \frac{1 + e^{-(b_{\ell-1}-\theta)}}{1 + e^{-(b_\ell - \theta)}}, \ \frac{1}{1 - e^{-(b_{\ell+1} - b_\ell)}} \frac{1 + e^{-(b_{\ell+1}-\theta)}}{1 + e^{-(b_\ell - \theta)}} \right\} > 0,$$

$$D_{\alpha,\beta,\Delta} = \min_{\substack{|\theta| \le \alpha, \\ \boldsymbol{b}\in\mathcal{B}}} \min_{\ell\in[L-1]} \frac{e^{-(b_\ell - \theta)}}{(1 + e^{-(b_\ell - \theta)})^2} > 0.$$

Theorem A.1 (Statistical convergence with unknown $\boldsymbol{b}$)**.** *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 parameters $(\Theta^{true}, \boldsymbol{b}^{true}) \in \mathcal{P} \times \mathcal{B}$. Suppose the link function f satisfies Assumption 2. Define $r_{\max} = \max_k r_k$. Then with very high probability, the estimator in (31) satisfies*

$$\mathrm{MSE}\left((\hat{\Theta}, \hat{\boldsymbol{b}}), (\Theta^{true}, \boldsymbol{b}^{true})\right) \le \frac{c_1 U_{\alpha,\beta,\Delta}^2}{L_{\alpha,\beta,\Delta}^2} r_{\max}^{(K-1)} \frac{\sum_k d_k}{\prod_k d_k} + \frac{C_2}{D_{\alpha,\beta,\Delta}^2} \frac{(\sum_k d_k)^2}{\min_\ell (n_\ell + n_{\ell+1})^2}.$$

Briefly,

$$\begin{aligned}\mathrm{MSE}\left((\hat{\Theta}, \hat{\boldsymbol{b}}), (\Theta^{true}, \boldsymbol{b}^{true})\right) &= \mathcal{O}\left(\frac{\sum_k d_k}{\prod_k d_k}\right) + \mathcal{O}\left(\frac{(\sum_k d_k)^2}{\min_\ell (n_\ell + n_{\ell+1})^2}\right) \\ &= \mathcal{O}\left(\frac{(\sum_k d_k)^2}{\min\left\{\sum_k d_k \prod_k d_k, \min_\ell (n_\ell + n_{\ell+1})^2\right\}}\right).\end{aligned}$$

where $c_1 > 0$ is a constant independent of the tensor dimension and rank, and $U_{\alpha,\beta,\Delta}$, $L_{\alpha,\beta,\Delta}$, $D_{\alpha,\beta,\Delta} > 0$ are constants defined in (32) and (33) and $C_2 > 0$ is a constant that depends on $C_{\alpha,\beta,\Delta}, \tilde{D}_{\alpha,\beta,\Delta}, L$.

Remark 5. The total MSE has two components where $\mathcal{O}\left(\frac{\sum_k d_k}{\prod_k d_k}\right)$ is from the error in $\hat{\Theta}$ and $\mathcal{O}\left(\frac{(\sum_k d_k)^2}{\min_\ell (n_\ell + n_{\ell+1})^2}\right)$ is from the error in $\hat{\boldsymbol{b}}$. When $\min_\ell(n_\ell + n_{\ell+1}) \gg \sum_k d_k \prod_k d_k$, the error in $\hat{\Theta}$ dominates the total MSE. When $\min_\ell(n_\ell + n_{\ell+1}) \ll \sum_k d_k \prod_k d_k$, the error in $\hat{\boldsymbol{b}}$ dominates the total MSE.

Remark 6. The constant bounds with α and β can be obtained trivially from the definition of the feasible sets. For notational convenience, we ignore this constant bound.

Proof of Theorem A.1. Similar to the proof of Theorem 4.1, we suppress Ω in the subscript. Based on the definition of $(\hat{\Theta}, \hat{\boldsymbol{b}})$, we have the following inequalities:

$$\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \hat{\boldsymbol{b}}) \geq \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}, \hat{\boldsymbol{b}}) \quad \text{and} \quad \mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \hat{\boldsymbol{b}}) \geq \mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \boldsymbol{b}^{\text{true}}). \tag{34}$$

Following the same argument as in Theorem 4.1 and the first inequality in (34), we obtain that

$$\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2 \leq \frac{c_1^2 U_{\alpha,\beta,\Delta}^2}{L_{\alpha,\beta,\Delta}^2} r_{\max}^{(K-1)} \sum_k d_k,$$

where $U_{\alpha,\beta,\Delta}, L_{\alpha,\beta,\Delta} > 0$ are two universal constants independent of the tensor dimension and rank. Next we bound $\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F^2$ given $\hat{\Theta}$. It follows from the expression of $\mathcal{L}_{\mathcal{Y}}(\Theta, \boldsymbol{b})$ that

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell} &= \sum_{\omega \in \Omega} \left[\mathbb{1}_{\{y_\omega = \ell\}} \frac{\dot{f}(b_\ell - \theta_\omega)}{g_\ell(\theta_\omega)} - \mathbb{1}_{\{y_\omega = \ell+1\}} \frac{\dot{f}(b_\ell - \theta_\omega)}{g_{\ell+1}(\theta_\omega)} \right], \\
\frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell^2} &= \sum_{\omega \in \Omega} \left[\mathbb{1}_{\{y_\omega = \ell\}} \frac{\ddot{f}(b_\ell - \theta_\omega) g_\ell(\theta_\omega) - \dot{f}^2(b_\ell - \theta_\omega)}{g_\ell^2(\theta_\omega)} - \mathbb{1}_{\{y_\omega = \ell+1\}} \frac{\ddot{f}(b_\ell - \theta_\omega) g_{\ell+1}(\theta_\omega) + \dot{f}^2(b_\ell - \theta_\omega)}{g_{\ell+1}^2(\theta_\omega)} \right], \\
&\quad \text{for all } \ell \in [L-1], \\
\frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell \partial b_{\ell+1}} &= \sum_{\omega \in \Omega} \mathbb{1}_{\{y_\omega = \ell+1\}} \frac{\dot{f}(b_\ell - \theta_\omega) \dot{f}(b_{\ell+1} - \theta_\omega)}{g_{\ell+1}^2(\theta_\omega)} \quad \text{and} \quad \frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell \partial b_{\ell'}} = 0 \text{ if } |\ell - \ell'| > 1.
\end{aligned}$$

Therefore, all entries in $\nabla_{\boldsymbol{b}} \mathcal{L}_{\mathcal{Y}}$ are upper bounded by $\{C_{\alpha,\beta,\Delta} \max_\ell (n_\ell + n_{\ell+1})\} > 0$, and $\nabla_{\boldsymbol{b}}^2 \mathcal{L}_{\mathcal{Y}}$ is a tridiagonal matrix.

We consider the profile log-likelihood $\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \boldsymbol{b})$ as a function of $\boldsymbol{b} \in \mathcal{B}$. For notational convenience, we drop $\hat{\Theta}$ from $\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \boldsymbol{b})$ and simply write $\mathcal{L}_{\mathcal{Y}}(\boldsymbol{b})$. By the second-order Taylor's expansion of $\mathcal{L}_{\mathcal{Y}}(\boldsymbol{b})$ around $\boldsymbol{b}^{\text{true}}$, we obtain

$$\mathcal{L}_{\mathcal{Y}}(\hat{\boldsymbol{b}}) = \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) + (\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}})^T \nabla_{\boldsymbol{b}} \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) + \frac{1}{2} (\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}})^T \nabla_{\boldsymbol{b}}^2 \mathcal{L}_{\mathcal{Y}}(\check{\boldsymbol{b}}) (\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}}), \tag{35}$$

where $\check{\boldsymbol{b}} = \gamma \boldsymbol{b}^{\text{true}} + (1-\gamma)\hat{\boldsymbol{b}}$ for some $\gamma \in [0,1]$, and $\nabla_{\boldsymbol{b}}^2 \mathcal{L}_{\mathcal{Y}}(\check{\boldsymbol{b}})$ denotes the $(L-1)$-by-$(L-1)$ Hession matrix evaluated at $\check{\boldsymbol{b}}$.

The linear term in (35) can be bounded by Cauchy-Schwartz inequality,

$$(\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}})^T \nabla_{\boldsymbol{b}} \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) \leq \|\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}}\|_F \|\nabla_{\boldsymbol{b}} \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}})\|_F \leq \|\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}}\|_F C \sqrt{L-1} \sum_k d_k \sqrt{\prod_k d_k}, \tag{36}$$

where the last inequality is followed from Lemma 14 stated as,

$$\left| \frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell} \Big|_{(\hat{\Theta}, \boldsymbol{b}^{\text{true}})} \right| \leq C \sum_k d_k \sqrt{\prod_k d_k}, \quad \text{for all } \ell \in [L-1] \text{ with high probability.}$$

We next bound the quadratic term in (35). Note that

$$(\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}})^T \nabla_{\boldsymbol{b}}^2 \mathcal{L}_{\mathcal{Y}}(\check{\boldsymbol{b}}) (\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}}) \tag{37}$$

$$
\begin{aligned}
&= \sum_{\ell\in[L-1]} \left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell^2}\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}\right)(\hat{b}_\ell - b_\ell^{\text{true}})^2 + 2\sum_{\ell\in[L-1]/\{1\}} \left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell \partial b_{\ell-1}}\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}\right)(\hat{b}_\ell - b_\ell^{\text{true}})(\hat{b}_{\ell-1} - b_{\ell-1}^{\text{true}}) \\
&\le \sum_{\ell\in[L-1]} \left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell^2}\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}\right)(\hat{b}_\ell - b_\ell^{\text{true}})^2 + \sum_{\ell\in[L-1]/\{1\}} \left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell \partial b_{\ell-1}}\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}\right)\left[(\hat{b}_\ell - b_\ell^{\text{true}})^2 + (\hat{b}_{\ell-1} - b_{\ell-1}^{\text{true}})^2\right] \\
&= \left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_1^2} + \frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_1 \partial b_2}\right)\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}(\hat{b}_1 - b_1^{\text{true}})^2 + \left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_{L-1}^2} + \frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_{L-2} \partial b_{L-1}}\right)\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}(\hat{b}_{L-1} - b_{L-1}^{\text{true}})^2 \\
&\quad + \sum_{\ell\in[L-2]/\{1\}} \left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell^2} + \frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell \partial b_{\ell-1}} + \frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_{\ell+1} \partial b_\ell}\right)\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}(\hat{b}_\ell - b_\ell^{\text{true}})^2 \\
&\le -D'_{\alpha,\beta,\Delta} \sum_{\ell\in[L-1]} (\hat{b}_\ell - b_{\text{true},\ell})^2 \\
&= -D'_{\alpha,\beta,\Delta} \|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F^2,
\end{aligned}
$$

where

$$
\begin{aligned}
D'_{\alpha,\beta,\Delta} &= \min_{\substack{|\theta|\le\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \min_{\ell\in[L-1]} -\left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell^2} + \frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell \partial b_{\ell-1}} + \frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_{\ell+1} \partial b_\ell}\right) \\
&= \min_{\substack{|\theta|\le\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \min_{\ell\in[L-1]} \left\{\sum_{\omega\in\Omega} -\mathbb{1}_{\{y_\omega=\ell\}} \left(\frac{\ddot{f}(b_\ell - \theta_\omega) g_\ell(\theta_\omega) - \dot{f}(b_\ell - \theta_\omega)\dot{g}_\ell(\theta_\omega)}{g_\ell^2(\theta_\omega)}\right)\right. \\
&\qquad\qquad \left. + \sum_{\omega\in\Omega} \mathbb{1}_{\{y_\omega=\ell+1\}} \left(\frac{\ddot{f}(b_\ell - \theta_\omega) g_{\ell+1}(\theta_\omega) - \dot{f}(b_\ell - \theta_\omega)\dot{g}_{\ell+1}(\theta_\omega)}{g_{\ell+1}^2(\theta_\omega)}\right)\right\} \\
&= \min_{\substack{|\theta|\le\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \min_{\ell\in[L-1]} \left\{\underbrace{-\sum_{\omega\in\Omega} \mathbb{1}_{\{y_\omega=\ell\}} \frac{\partial}{\partial\theta_\omega}\left(\frac{\dot{f}(b_\ell - \theta_\omega)}{g_\ell(\theta_\omega)}\right)}_{>0} + \underbrace{\sum_{\omega\in\Omega} \mathbb{1}_{\{y_\omega=\ell+1\}} \frac{\partial}{\partial\theta_\omega}\left(\frac{\dot{f}(b_\ell - \theta_\omega)}{g_{\ell+1}(\theta_\omega)}\right)}_{>0}\right\} \\
&\ge D_{\alpha,\beta,\Delta} \min_{\ell\in[L-1]} (n_\ell + n_{\ell+1})\,.
\end{aligned}
$$

Combining inequalities (35), (36) and (37) yields

$$
\mathcal{L}_{\mathcal{Y}}(\hat{\boldsymbol{b}}) \le \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) + C\sqrt{L-1}\sum_k d_k \sqrt{\prod_k d_k)}\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F - \frac{D_{\alpha,\beta,\Delta}}{2}\min_\ell(n_\ell + n_{\ell+1})\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F^2.
$$

Since $\hat{\boldsymbol{b}}$ satisfies $\mathcal{L}_{\mathcal{Y}}(\hat{\boldsymbol{b}}) - \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) \ge 0$, we have that

$$
C\sqrt{L-1}\sum_k d_k \sqrt{\prod_k d_k)}\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F - \frac{D_{\alpha,\beta,\Delta}}{2}\min_\ell(n_\ell + n_{\ell+1})\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F^2 \ge 0
$$

Finally,

$$
\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F^2 \le \frac{C_2}{D_{\alpha,\beta,\Delta}^2} \frac{(\sum_k d_k)^2 \prod_k d_k}{\min_\ell(n_\ell + n_{\ell+1})^2},
$$

where $C_2(C_{\alpha,\beta,\Delta}, \tilde{D}_{\alpha,\beta,\Delta}, L)$ is a positive constant. The following completes the proof.

$$
\text{MSE}\left((\hat{\Theta}, \hat{\boldsymbol{b}}), (\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})\right) \le \frac{1}{\prod_k d_k}\|(\hat{\Theta}, \hat{\boldsymbol{b}}) - (\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})\|_F^2
$$

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

□

Remark 7 (MSE for $\hat{\boldsymbol{b}}$)**.** In the proof, we have shown that

$$\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F \leq \frac{\sqrt{C_2}}{D_{\alpha,\beta,\Delta}} \frac{(\sum_k d_k)\sqrt{\prod_k d_k}}{\min_\ell(n_\ell + n_{\ell+1})}.$$

This bound is sharper than the trivial bound $\|\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}}\|_F \leq 2\beta\sqrt{L-1}$ which is from $|b_\ell^{\text{true}} - \hat{b}_\ell| \leq 2\beta$. In particular, $\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F \to 0$ as $\min_\ell(n_\ell + n_{\ell+1}) \asymp \tilde{\mathcal{O}}(\sum_k d_k\sqrt{\prod_k d_k}) \to \infty$.

Lemma 13 (CLT for i.d. Bernoulli)**.** *Suppose $\{X_n\}$ is a series of independent Bernoulli trials with possibly different success probabilities p_n. Let $s_n = \sum_{m=1}^n p_m(1-p_m)$ and $Y_n = X_n - p_n$. Then,*

$$\frac{1}{s_n}\sum_{m=1}^n Y_n \xrightarrow{\mathcal{D}} N(0,1), \qquad \text{as } n \to \infty,$$

whenever $s_n = \sum_{m=1}^n p_m(1-p_m) \to \infty$.

Proof. Let us look for a Lyapunov condition for $\delta = 1$. First, we have

$$\mathbb{E}[|Y_n|^3] = p_n(1-p_n)^3 + (1-p_n)p_n^3 \leq p_n(1-p_n)[(1-p_n)^2 + p_n^2] \leq p_n(1-p_n).$$

Hence, summation of the above inequality shows,

$$\sum_{m=1}^n \mathbb{E}[|Y_n|^3] \leq \sum_{m=1}^n p_m(1-p_m) = s_n^2.$$

Thus, the Lyapunov condition is satisfied whenever,

$$\frac{1}{s_n^3}\sum_{m=1}^n \mathbb{E}[|Y_n|^3] \leq \frac{s_n^2}{s_n^3} \to 0.$$

or simply $s_n = \sum_{m=1}^n p_m(1-p_m) \to \infty$. Central limit theorem on Y_n completes the proof.

□

Lemma 14 (Bound on Gradients)**.** *Consider the same set-up as in Theorem A.1. Then, with very high probability,*

$$\left|\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell}\right|_{(\hat{\Theta}, \boldsymbol{b}^{true})} \leq \left(C_{\alpha,\beta,\Delta} + \tilde{D}_{\alpha,\beta,\Delta}\|\Theta^{true} - \hat{\Theta}\|_F^2\right)\sqrt{\prod_k d_k}, \qquad \text{for all } \ell \in [L-1],$$

In particular, there exists a constant $d_0 \in \mathbb{N}_+$ such that for all $d_k \geq d_0$ for $k \in [K]$,

$$\left|\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell}\right|_{(\hat{\Theta}, \boldsymbol{b}^{true})} \leq C\sum_k d_k\sqrt{\prod_k d_k}, \qquad \text{for all } \ell \in [L-1],$$

where $C(C_{\alpha,\beta,\Delta}, \tilde{D}_{\alpha,\beta,\Delta}) > 0$ is a constant.

Proof. We only prove the case for $\ell = 1$. Other cases can be proved similarly. Note that

$$\left.\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\right|_{(\hat{\Theta},\boldsymbol{b}^{\text{true}})} = \underbrace{\left.\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\right|_{(\hat{\Theta},\boldsymbol{b}^{\text{true}})} - \mathbb{E}_{\mathcal{Y}}\left[\left.\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\right|_{(\hat{\Theta},\boldsymbol{b}^{\text{true}})}\right]}_{:=A} + \underbrace{\mathbb{E}_{\mathcal{Y}}\left[\left.\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\right|_{(\hat{\Theta},\boldsymbol{b}^{\text{true}})}\right] - \mathbb{E}_{\mathcal{Y}}\left[\left.\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\right|_{(\hat{\Theta^{\text{true}}},\boldsymbol{b}^{\text{true}})}\right]}_{:=B}. \tag{38}$$

We have used the fact that the score function has mean zero, $\mathbb{E}_{\mathcal{Y}}\left[\left.\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\right|_{(\hat{\Theta^{\text{true}}},\boldsymbol{b}^{\text{true}})}\right] = 0$. Here all expectations are taken with respect to $\mathcal{Y} \sim \mathbb{P}(\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})$.
We now bound the two deviation terms in (38) separately. The term A in (38) is the stochastic deviation of log-likelihood to its expectation:

$$\begin{aligned} A &= \sum_{\omega\in\Omega}\left\{[\mathbb{1}_{y_\omega=1} - g_1(\theta_\omega^{\text{true}})]\frac{\dot{f}(b_1-\hat{\theta}_\omega)}{g_1(\hat{\theta}_\omega)} - [\mathbb{1}_{y_\omega=2} - g_2(\theta_\omega^{\text{true}})]\frac{\dot{f}(b_1-\hat{\theta}_\omega)}{g_2(\hat{\theta}_\omega)}\right\} \\ &\le C_{\alpha,\beta,\Delta}\sum_{\omega\in\Omega}\underbrace{[\mathbb{1}_{y_\omega=1} + \mathbb{1}_{y_\omega=2} - g_1(\theta_\omega^{\text{true}}) - g_2(\theta_\omega^{\text{true}})]}_{:=W_\omega}. \end{aligned}$$

Note that $\{W_\omega\}$ are centered independent Bernoulli random variables with success probabilities $g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}})$. By Lemma 13, we have

$$\sum_{\omega\in\Omega} W_\omega \xrightarrow{\mathcal{D}} N\left(0, \sum_{\omega\in\Omega}\left(g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}})\right)\left(1 - g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}})\right)\right).$$

Hence, with the fact that $\sum_{\omega\in\Omega}\left(g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}})\right)\left(1 - g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}})\right) \le \frac{1}{4}\prod_k d_k$,

$$|A| \le C_{\alpha,\beta,\Delta}\left|\sum_{\omega\in\Omega} W_\omega\right| \le C_{\alpha,\beta,\Delta}\sqrt{\prod_k d_k}, \quad \text{with high probability.} \tag{39}$$

The second term B in (38) is the bias induced by the inaccuracy of $\hat{\Theta}$:

$$\begin{aligned} |B| &= \left|\sum_{\omega\in\Omega} g_1(\theta_\omega^{\text{true}})\left(\frac{\dot{f}(b_1-\hat{\theta}_\omega)}{g_1(\hat{\theta}_\omega)} - \frac{\dot{f}(b_1-\theta_\omega^{\text{true}})}{g_1(\theta_\omega^{\text{true}})}\right) - \sum_{\omega\in\Omega} g_2(\theta_\omega^{\text{true}})\left(\frac{\dot{f}(b_1-\hat{\theta}_\omega)}{g_2(\hat{\theta}_\omega)} - \frac{\dot{f}(b_1-\theta_\omega^{\text{true}})}{g_2(\theta_\omega^{\text{true}})}\right)\right| \\ &= \left|\sum_{\omega\in\Omega} g_1(\theta_\omega^{\text{true}})(\theta_\omega^{\text{true}} - \hat{\theta}_\omega)\left\{\frac{\partial}{\partial}\left(\frac{\dot{f}(b_1-\theta)}{g_1(\theta)}\right)\Big|_{\rho_\omega\hat{\theta}_\omega+(1-\rho_\omega)\theta_\omega^{\text{true}}}\right\}\right. \\ &\qquad \left. - \sum_{\omega\in\Omega} g_2(\theta_\omega^{\text{true}})(\theta_\omega^{\text{true}} - \hat{\theta}_\omega)\left\{\frac{\partial}{\partial}\left(\frac{\dot{f}(b_1-\theta)}{g_2(\theta)}\right)\Big|_{\rho'_\omega\hat{\theta}_\omega+(1-\rho'_\omega)\theta_\omega^{\text{true}}}\right\}\right| \\ &\le \tilde{D}_{\alpha,\beta,\Delta}\sum_{\omega\in\Omega}|g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}})||\theta_\omega^{\text{true}} - \hat{\theta}_\omega|. \end{aligned}$$

By Cauchy-Schwartz inequality with the fact that $g_1(\theta) + g_2(\theta) < 1$,

$$|B| \le \tilde{D}_{\alpha,\beta,\Delta}\sqrt{\prod_k d_k}\|\Theta^{\text{true}} - \hat{\Theta}\|_F. \tag{40}$$

Plugging (39) and (40) back to (38) yields that

$$\left.\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\right|_{(\hat{\Theta},\boldsymbol{b}^{\text{true}})} \leq \left(C_{\alpha,\beta,\Delta} + \tilde{D}_{\alpha,\beta,\Delta}\|\Theta^{\text{true}} - \hat{\Theta}\|_F\right)\sqrt{\prod_k d_k}.$$

holds with very high probability. The second inequality in the lemma comes from the fact that $\|\Theta^{\text{true}} - \hat{\Theta}\|_F \leq \mathcal{O}(\sum_k d_k)$. □

B Additional results of HCP analysis

B.1 Clustering based on Tucker representation

We perform clustering analyses based on the Tucker representation of the estimated tensor parameter $\hat{\Theta}$. The procedure is motivated from the higher-order extension of Principal Component Analysis (PCA) or Singular Value Decomposition (SVD). Recall that, in the matrix case, we usually perform clustering on an $m \times n$ (normalized) matrix X based on the following procedure. First, we factorize X into

$$X = U\Sigma V^T,$$

where Σ is a diagonal matrix and U, V are factor matrices with orthogonal columns. Second, we take each column of V as a principal axis and each row in $U\Sigma$ as principal component. A subsequent multivariate clustering method (such as K-means) is then applied to the m rows of $U\Sigma$.

We apply a similar clustering procedure to the estimated parameter tensor $\hat{\Theta}$. Based on Tucker representation of $\hat{\Theta}$, we have

$$\hat{\Theta} = \hat{\mathcal{C}} \times_1 \hat{\boldsymbol{M}}_1 \times_2 \cdots \times_K \hat{\boldsymbol{M}}_K, \tag{41}$$

where $\hat{\mathcal{C}} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is the estimated core tensor, $\hat{\boldsymbol{M}}_k \in \mathbb{R}^{d_k \times r_k}$ are estimated factor matrices with orthogonal columns, and $\times_k$ denotes the tensor-by-matrix multiplication (Kolda and Bader, 2009). The mode-k matricization of (41) gives

$$\hat{\Theta}_{(k)} = \hat{\boldsymbol{M}}_k \hat{\mathcal{C}}_{(k)} \left(\hat{\boldsymbol{M}}_K \otimes \cdots \otimes \hat{\boldsymbol{M}}_1\right),$$

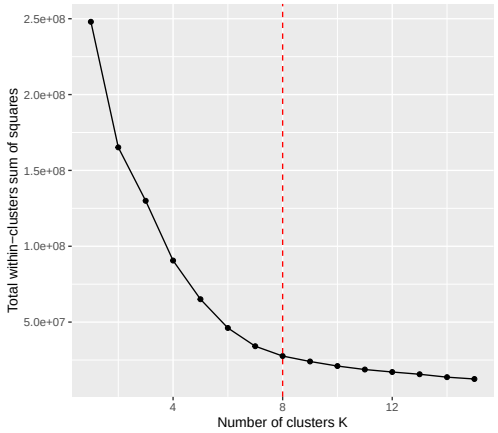
where $\hat{\Theta}_{(k)}, \hat{\mathcal{C}}_{(k)}$ denote the mode-k unfolding of $\hat{\Theta}$ and $\hat{\mathcal{C}}$, respectively. Then, the mode-k clustering can be performed as follows. First, we take columns in $\left(\hat{\boldsymbol{M}}_K \otimes \cdots \otimes \hat{\boldsymbol{M}}_1\right)$ as principal axes and rows in $\hat{\boldsymbol{M}}_k \hat{\mathcal{C}}_{(k)}$ as principal components. Then, we perform K-means clustering method to the d_k rows of the matrix $\hat{\boldsymbol{M}}_k \hat{\mathcal{C}}_{(k)}$.

B.2 Clustering results of HCP analysis

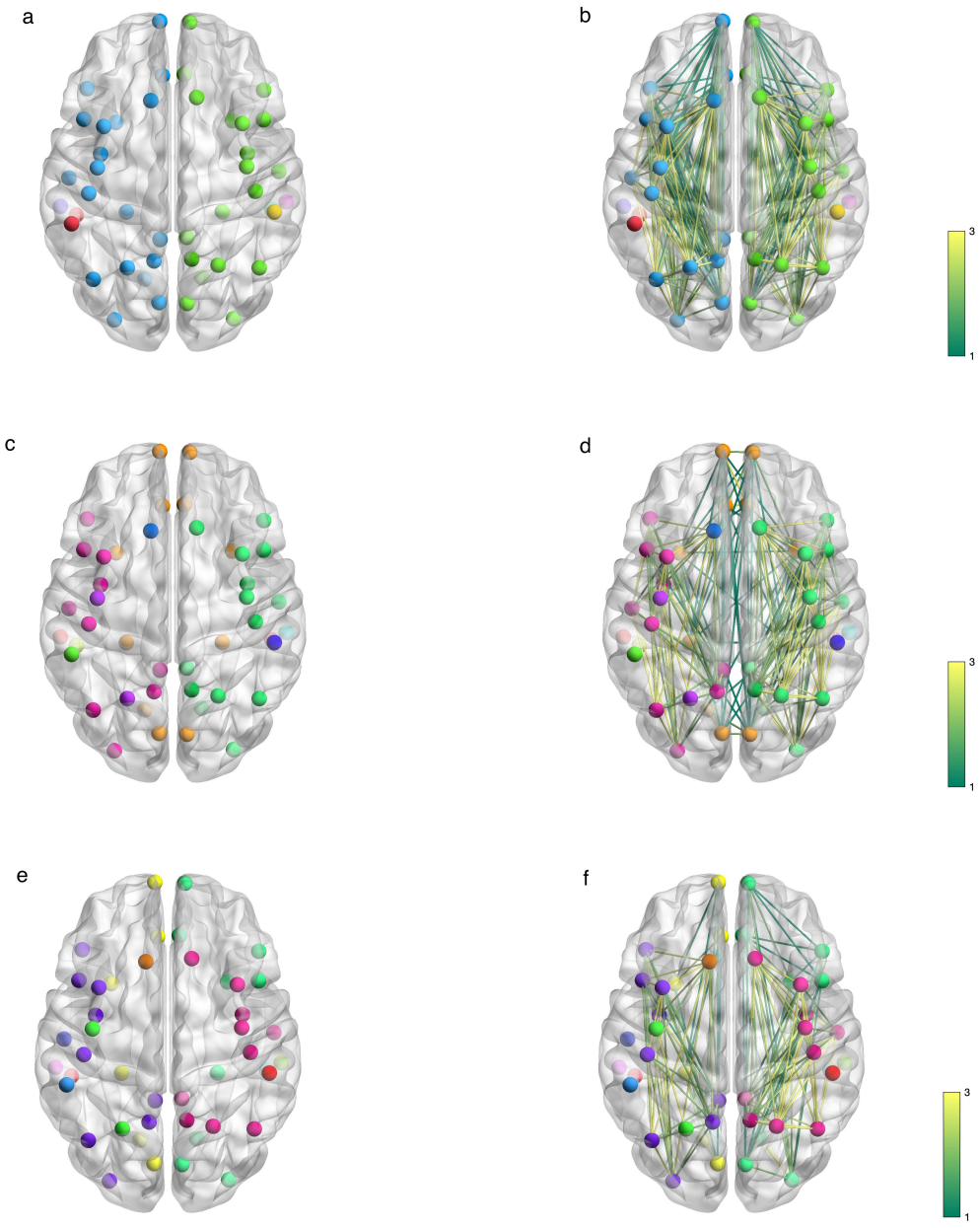
We perform a clustering analysis on the 68 brain nodes using the produce described above. Recall that our ordinal tensor method outputs the estimated parameter tensor $\hat{\Theta} \in \mathbb{R}^{68\times 68\times 136}$ with rank $(23, 23, 8)$. We apply K-means to the mode-1 principal component matrix of size 68×184 ($184 = 23\times 8$). The elbow method suggests eight clusters among the 68 nodes (see Figure S1).

Table S1 shows the eight clusters and their compositions. We find that most brain nodes fall into cluster I and the cluster II, which can be represented as the left and right hemispheres, respectively (Figure S2 a,b). Clusters III-VIII capture brain regions with same annotations. For example, clusters III/IV represents the supramarginal gyrus region on the left/right hemisphere, respectively. This region is known to be involved in visual word recognition and reading (Stoeckel et al., 2009).

We increase the number of clusters to 11 and 13. When the number of clusters is 11, there are three main clusters which most of brain nodes fall into (Figure S2 c,d). Two of them are represented as the left and right hemispheres like the previous cluster result. The other main cluster is newly captured and located in the bottom-middle part of the brain. Other small clusters are more or less the same as the small clusters in the result when the number of cluster is eight. When the number of clusters is 13, we find the four biggest clusters. We can characterize those clusters as top-left, bottom-left, top-right and bottom-right hemispheres (Figure S2 e,f). Other clusters remain almost the same as the small clusters in the previous results, which share same annotations in each cluster. The results demonstrate that our clustering successfully groups similar brain regions together without knowledge of external annotation.



Supplemental Figure S1: Elbow plot for determining the number of clusters in K-means.



Supplemental Figure S2: Cluster image of brain nodes when the number of clusters is 8 (a,b), 11 (c,d) and 13 (e,f). (a,c,e) Nodes from the same cluster are colored with the same color. (b,d,f) Predicted connectivity within clusters. Edges are colored based on predicted strength level averaged across individuals.

Cluster	I		
Brain nodes	l.superiorfrontal(2), l.superiortemporal(2), l.insula , l.frontalpole, l.caudalmiddlefrontal, l.parstriangularis, l.parsopercularis, l.precentral, l.temporalpole, l.postcentral, l.superiorparietal, l.inferiorparietal, l.lateraloccipital, l.medialorbitofrontal, l.superiorfrontal, l.precuneus, l.cuneus, l.parahippocampal, l.lingual, l.superiortemporal, l.isthmuscingulate, l.lateraloccipital		
Cluster	II		
Brain nodes	r.superiorfrontal(2), r.superiortemporal(2), r.insula, r.frontalpole, r.caudalmiddlefrontal, r.parstriangularis, r.parsopercularis, r.precentral, r.temporalpole, r.postcentral, r.superiorparietal, r.inferiorparietal, r.lateraloccipital, r.medialorbitofrontal, r.superiorfrontal, r.precuneus, r.cuneus, r.parahippocampal, r.lingual, r.superiortemporal, r.isthmuscingulate, r.lateraloccipital		
Cluster	III	IV	V
Brain nodes	l.supramarginal(4)	r.supramarginal(4)	l.inferiortemporal(3)
Cluster	VI	VII	VIII
Brain nodes	r.inferiortemporal(3)	l.middletemporal(3)	r.middletemporal(3)

Supplemental Table S1: Clustering result of brain nodes. The first alphabet in the node name indicates the left (L) or right (R) hemisphere. The number in the parentheses indicates the node count in each cluster.

References

Acar, E., Dunlavy, D. M., Kolda, T. G., and Mørup, M. (2010). Scalable tensor factorizations with missing data. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 701–712. SIAM.

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

Baltrunas, L., Kaminskas, M., Ludwig, B., Moling, O., Ricci, F., Aydin, A., Lüke, K.-H., and Schwaiger, R. (2011). Incarmusic: Context-aware music recommendations in a car. In *International Conference on Electronic Commerce and Web Technologies*, pages 89–100. Springer.

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

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

Brascamp, H. J. and Lieb, E. H. (2002). On extensions of the brunn-minkowski and prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. In *Inequalities*, pages 441–464. Springer.

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

Chen, H., Raskutti, G., and Yuan, M. (2019). Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–208.

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

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

Friedland, S. and Lim, L.-H. (2018). Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281.

Ge, R. and Ma, T. (2017). On the optimization landscape of tensor decompositions. In *Advances in Neural Information Processing Systems*, pages 3653–3663.

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

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

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

Hong, D., Kolda, T. G., and Duersch, J. A. (2019). Generalized canonical polyadic tensor decomposition. *SIAM Review. In press. arXiv:1808.07452*.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094.

Jiang, B., Yang, F., and Zhang, S. (2017). Tensor and its Tucker core: the invariance relationships. *Numerical Linear Algebra with Applications*, 24(3):e2086.

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

Lim, L.-H. (2005). Singular values and eigenvalues of tensors: a variational approach. In *1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005.*, pages 129–132. IEEE.

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

Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*, pages 73–81.

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

Nguyen, N. H., Drineas, P., and Tran, T. D. (2015). Tensor sparsification via a bound on the spectral norm of random tensors. *Information and Inference: A Journal of the IMA*, 4(3):195–229.

Nickel, M., Tresp, V., and Kriegel, H.-P. (2011). A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning*, volume 11, pages 809–816.

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

Stoeckel, C., Gough, P. M., Watkins, K. E., and Devlin, J. T. (2009). Supramarginal gyrus involvement in visual word recognition. *Cortex*, 45(9):1091–1096.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Tomioka, R. and Suzuki, T. (2014). Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79.

Wang, M., Duc, K. D., Fischer, J., and Song, Y. S. (2017). Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear Algebra and Its Applications*, 520:44–66.

Wang, M. and Li, L. (2018). Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *arXiv preprint arXiv:1811.05076*.

Wang, M. and Song, Y. (2017). Tensor decompositions via two-mode higher-order SVD (HOSVD). In *Artificial Intelligence and Statistics*, pages 614–622.

Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019). In press. arXiv:1906.03807*.

Yuan, M. and Zhang, C.-H. (2016). On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068.

Zhang, A. et al. (2019). Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47(2):936–964.

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