Tensor denoising and completion based on ordinal observations

Chanwoo Lee
chanwoo.lee@wisc.edu

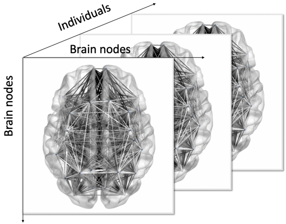
Miaoyan Wang
miaoyan.wang@wisc.edu

University of Wisconsin - Madison

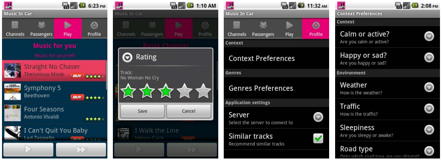
ICML 2020

Ordinal tensor data in applications

- Tensor in networks (Human Connectome Project (HCP)).
- Each entry $y_\omega \in \{\text{high}, \text{moderate}, \text{low}\}$.



- Tensor in recommendation system (Baltrunas et al., 2011).
- Each entry $y_\omega \in \{1, 2, 3, 4, 5\}$



(a) Tracks Proposed to Play (b) Rating a Track (c) Editing the User Profile (d) Configuring the Recommender

Challenges from ordinal tensor data

- Goal: learn a probabilistic tensor from multi-way ordinal observations.

Challenges from ordinal tensor data

- Goal: learn a probabilistic tensor from multi-way ordinal observations.
- Two key properties needed for a reasonable model.
  1. The model should be invariant under a reversal of categories

$$\textbf{like} \prec \textbf{neutral} \prec \textbf{dislike} \Longleftrightarrow \textbf{like} \succ \textbf{neutral} \succ \textbf{dislike}.$$

  2. The parameter interpretations should be consistent under merging or splitting of contiguous categories.

Challenges from ordinal tensor data

- Goal: learn a probabilistic tensor from multi-way ordinal observations.
- Two key properties needed for a reasonable model.
  1. The model should be invariant under a reversal of categories

$$\textbf{like} \prec \textbf{neutral} \prec \textbf{dislike} \Longleftrightarrow \textbf{like} \succ \textbf{neutral} \succ \textbf{dislike}.$$

  2. The parameter interpretations should be consistent under merging or splitting of contiguous categories.
- Two challenges for ordinal tensor model.
  1. The entries do not belong to exponential family distribution.
  2. The observation contains less information - neither the underlying signal nor the quantization operator is unknown.

Summary of our contribution

- We establish the recovery theory for signal tensors and quantization operators simultaneously from observed ordinal tensor data.
- Let $\mathcal{Y} \in \mathbb{R}^{d\times\cdots\times d}$ be an order-K, L-level ordinal tensor.

	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	unkonwn	✗	✓
Sample complexity for completion	d^K	Kd	Kd

- Preprint: https://arxiv.org/abs/2002.06524
- Software: https://cran.r-project.org/web/packages/tensorordinal/index.html

Probabilistic model: a cumulative link model

- $[L] = \{1, 2, \cdots, L\}$ denotes the ordinal level.
- Let $\mathcal{Y} = [\![y_\omega]\!] \in [L]^{d_1 \times \cdots \times d_K}$ be an ordinal tensor. The entries y_ω are independently distributed with cumulative probabilities:

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

ex) $f(x) = \frac{e^x}{1+e^x}$ is a logistic link.

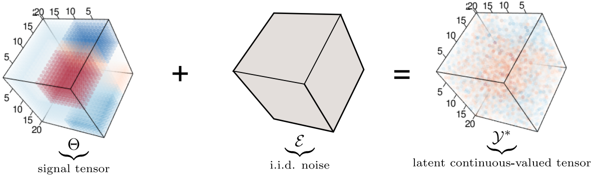
- The additive, cumulative model enjoys two key properties for ordinal tensor data.
- If f is a cumulative function,

$$\mathbb{P}(y_\omega = \ell) = f(b_\ell - \theta_\omega) - f(b_{\ell-1} - \theta_\omega) = \mathbb{P}(b_{\ell-1} < y^*_\omega \leq b_\ell),$$

where $\epsilon_\omega \overset{i.i.d}{\sim} f$ and $y^*_\omega = \theta_\omega + \epsilon_\omega$.

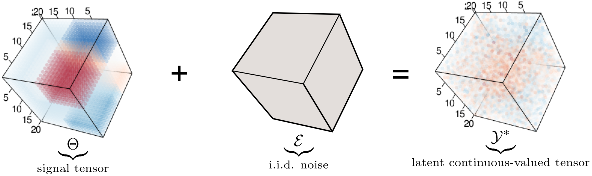
Latent variable interpretation

- We can interpret the ordinal tensor model (1) as an L-level quantization model.

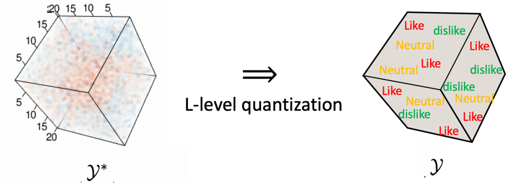


Latent variable interpretation

- We can interpret the ordinal tensor model (1) as an L-level quantization model.



- Given intervals from the cut-off points vector $\boldsymbol{b}$.



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

Probabilistic model: assumptions on Θ

- The parameter Θ admits the Tucker decomposition:

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

  where $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots r_K}$ is a core tensor, $\boldsymbol{M}_k \in \mathbb{R}^{d_k \times r_k}$ are factor matrices.
- Entries of Θ are uniformly bounded in magnitude by a constant $\alpha \in \mathbb{R}_+$.

Rank constrained M-estimation

- Let $\Omega \subset [d_1] \times \cdots \times [d_K]$ denote the set of observed indices. Ω could be the full set or incomplete set (for completion).
- The log-likelihood associated with the observations is

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

Rank constrained M-estimation

- Let $\Omega \subset [d_1] \times \cdots \times [d_K]$ denote the set of observed indices.
  Ω could be the full set or incomplete set (for completion).
- The log-likelihood associated with the observations is

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

- We propose a rank-constrained maximum likelihood estimation for Θ.

$$(\hat{\Theta}, \hat{\boldsymbol{b}}) = \arg\max_{\Theta \in \mathcal{P}, \boldsymbol{b} \in \mathcal{B}} \mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b}) \quad \text{where,}$$

$$\mathcal{P} = \{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \text{rank}(\mathcal{P}) \le \boldsymbol{r},\ \|\Theta\|_\infty \le \alpha,\ \underbrace{\langle \Theta, \mathcal{J} \rangle = 0}_{\text{identifiability condition}} \},$$

$$\mathcal{B} = \{\boldsymbol{b} \in \mathbb{R}^{L-1} : \|\boldsymbol{b}\|_\infty \le \beta,\ \min_\ell (b_\ell - b_{\ell-1}) \ge \Delta > 0\}.$$

Here, $\mathcal{J} = [\![1]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denotes a tensor of all ones.

Theoretical results: tensor denoising

- **Tensor denoising:**
  - (Q1) How accurately can we estimate the latent signal tensor Θ from the ordinal observation $\mathcal{Y}$?

Theoretical results: tensor denoising

- **Tensor denoising:**
  - (Q1) How accurately can we estimate the latent signal tensor Θ from the ordinal observation $\mathcal{Y}$?
  - (A1) Let us define $\text{MSE}(\hat{\Theta}, \Theta^{\text{true}}) = \frac{1}{\prod_k d_k} \|\hat{\Theta} - \Theta^{\text{true}}\|_F^2$.

Statistical convergence (L. and Wang, 2020)

With very high probability, our estimator $\hat{\Theta}$ satisfies

$$\text{MSE}(\hat{\Theta}, \Theta^{\text{true}}) \leq \min\left(4\alpha^2,\ c_1 r_{\max}^{K-1} \frac{\sum_k d_k}{\prod_k d_k}\right),$$

where $c_1 = c(f, K) > 0$ is a constant.

Theoretical results: tensor denoising

- **Tensor denoising:**
  - (Q1') Is this bound optimal?

Theoretical results: tensor denoising

- **Tensor denoising:**
  - (Q1') Is this bound optimal?
  - (A1')

Minimax lower bound (L. and Wang, 2020)

Under some mild technical conditions,

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

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

- So our estimation bound is rate-optimal.

Theoretical results: tensor completion

- **Tensor completion:**
  - (Q2) How many sampled entries do we need to consistently recover Θ?

Theoretical results: tensor completion

- **Tensor completion:**
  - (Q2) How many sampled entries do we need to consistently recover Θ?
  - (A2) Let us define $\|\Theta - \hat{\Theta}\|_{F,\Pi}^2 = \sum_{\omega \in [d_1]\times\cdots\times[d_K]} \pi_\omega (\Theta_\omega - \hat{\Theta}_\omega)^2$.

Sample complexity (L. and Wang, 2020)

Let $\{y_\omega\}_{\omega\in\Omega}$ be the ordinal observation, where Ω is chosen at random with replacement according to a probability distribution Π. Then, with very high probability,

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

  - The number of free parameters is roughly on the order of $\sum_k d_k$.
  - The sample complexity $|\Omega| \gg \mathcal{O}(\sum_k d_k)$ is almost optimal.

Algorithm

- The rank r is unknown

Algorithm

- The rank r is unknown $\Longrightarrow$ Bayesian information criterion (BIC).

Algorithm

- The rank r is unknown $\Longrightarrow$ Bayesian information criterion (BIC).
- Non-convex problem

Algorithm

- The rank r is unknown $\Longrightarrow$ Bayesian information criterion (BIC).
- Non-convex problem $\Longrightarrow$ Alternating optimization approach.

Algorithm

- The rank r is unknown $\Longrightarrow$ Bayesian information criterion (BIC).
- Non-convex problem $\Longrightarrow$ Alternating optimization approach.
  - Let $\mathcal{L}_{\mathcal{Y},\Omega}(\mathcal{C},\mathcal{M}_1,\cdots,\mathcal{M}_K,\boldsymbol{b}) = \mathcal{L}_{\mathcal{Y},\Omega}(\Theta,\boldsymbol{b})$.

Algorithm: Alternating optimization

Result: Estimated Θ, together with core tensor and factor matrices

Random initialization;

Repeat until converge;

$$
\begin{aligned}
\mathcal{C}^{(n)} &= \arg\max_{\mathcal{C}} \mathcal{L}_{\mathcal{Y},\Omega}(\mathcal{C},\mathcal{M}_1^{(n-1)},\cdots,\mathcal{M}_k^{(n-1)},\boldsymbol{b}^{(n-1)}).\\
\mathcal{M}_1^{(n)} &= \arg\max_{\mathcal{M}_1} \mathcal{L}_{\mathcal{Y},\Omega}(\mathcal{C}^{(n)},\mathcal{M}_1,\cdots,\mathcal{M}_k^{(n-1)},\boldsymbol{b}^{(n-1)}).\\
&\vdots\\
\mathcal{M}_K^{(n)} &= \arg\max_{\mathcal{M}_K} \mathcal{L}_{\mathcal{Y},\Omega}(\mathcal{C}^{(n)},\mathcal{M}_1^{(n)},\cdots,\mathcal{M}_k,\boldsymbol{b}^{(n-1)}).\\
\boldsymbol{b}^{(n)} &= \arg\max_{\boldsymbol{b}} \mathcal{L}_{\mathcal{Y},\Omega}(\mathcal{C}^{(n)},\mathcal{M}_1^{(n)},\cdots,\mathcal{M}_k^{(n)},\boldsymbol{b}).
\end{aligned}
$$

end

Simulations

- The decay in the error appears to behave on the order of d^{-2} when $K = 3$.
- A larger estimation error is observed when the signal is too small or large.
- There is a big improvement from $L = 2$ to $L \geq 3$.

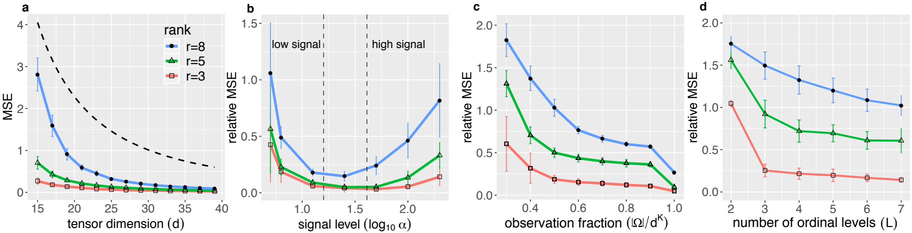
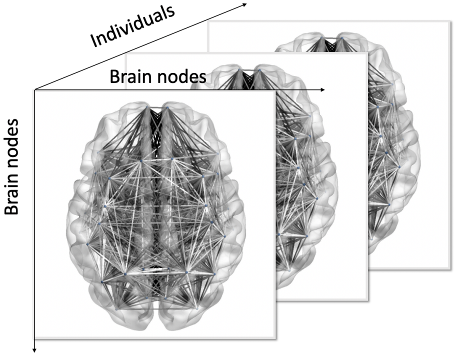


Figure: the relative MSE $= \|\hat{\Theta} - \Theta^{\text{true}}\|_F / \|\Theta^{\text{true}}\|_F$ for better visualization.

Data application: Human Connectome Project (HCP)

- An ordinal tensor consisting of structural connectivities among 68 brain nodes for 136 individuals.
- Each entry $y_\omega \in \{\text{high}, \text{moderate}, \text{low}\}$.

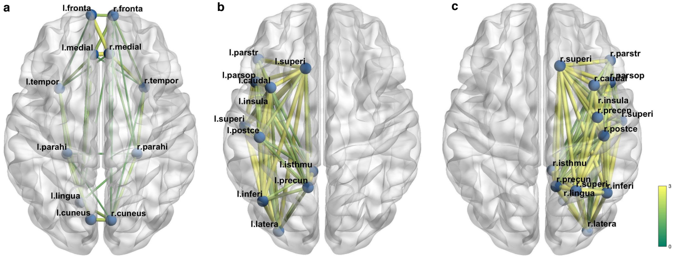


Data application: Human Connectome Project (HCP)

- The clustering based on the estimated $\hat{\Theta}$ identifies 11 (3+8) clusters among 68 brain nodes.

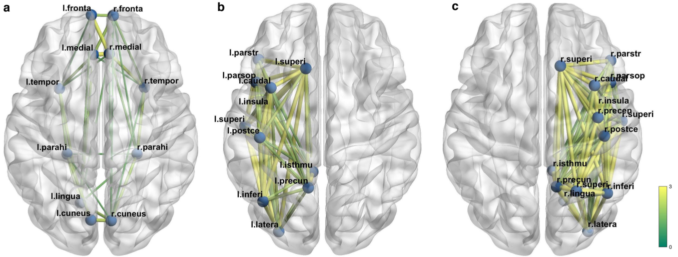
Data application: Human Connectome Project (HCP)

- The clustering based on the estimated $\hat{\Theta}$ identifies 11 (3+8) clusters among 68 brain nodes.
- The top three clusters capture the global separation among brain nodes.



Data application: Human Connectome Project (HCP)

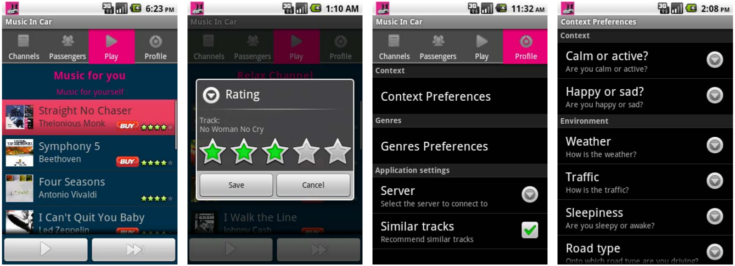
- The clustering based on the estimated $\hat{\Theta}$ identifies 11 (3+8) clusters among 68 brain nodes.
- The top three clusters capture the global separation among brain nodes.



- The small clusters represent local regions driving by similar nodes.

Data application: InCarMusic

- A tensor recording the ratings of 139 songs from 42 users on 26 contexts (Baltrunas et al., 2011).
- Each entry is a rating on a scale of 1 to 5 ($y_\omega \in \{1, 2, 3, 4, 5\}$).



(a) Tracks Proposed to Play (b) Rating a Track (c) Editing the User Profile (d) Configuring the Recommender

Data application: HCP, InCarMusic

- Our method achieves lower prediction error than others.

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

Table: Comparison of prediction error based on cross-validation (10 repetitions, 5 foldes). Standard errors are reported in parentheses. MAD: mean absolute error; MCR: misclassification error.

Summary

- We propose a cumulative probabilistic model for ordinal tensor observations.
- The model achieves optimal convergence rate and nearly optimal sample complexity.
- The model has good interpretation and prediction performance in HCP and InCarMusic application.
- Thank you!
- Preprint: https://arxiv.org/abs/2002.06524
- Software: https://cran.r-project.org/web/packages/tensorordinal/index.html