Tensor denoising and completion based on ordinal observations

Chanwoo Lee
chanwoo.lee@wisc.edu

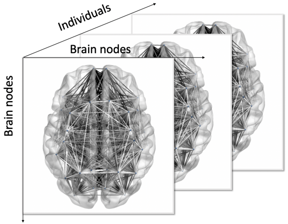
Miaoyan Wang
miaoyan.wang@wisc.edu

University of Wisconsin - Madison

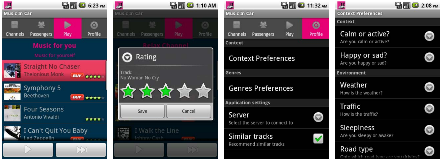
APS March Meeting

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

Summary of our contribution

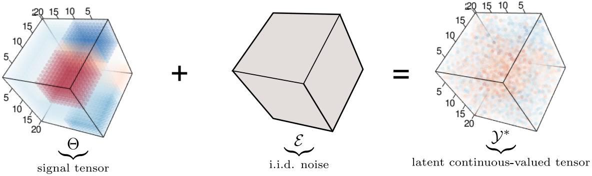
- We establish the recovery theory for signal tensors and quantization operators simultaneously from observed ordinal tensor data.
- Let $\mathcal{Y} \in \mathbb{R}^{d\times\cdots\times d}$ be an order-K, L-level ordinal tensor.

	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

- Paper: http://proceedings.mlr.press/v119/lee20i/lee20i.pdf
- Software: https://cran.r-project.org/web/packages/TensorComplete/index.html

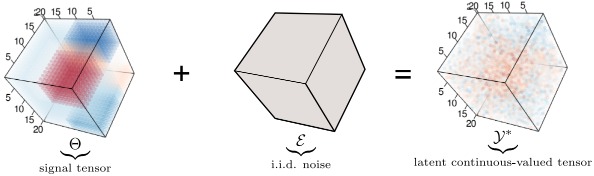
Probabilistic model: generating process

- Let $\mathcal{Y} = [\![y_\omega]\!] \in [L]^{d_1 \times \cdots \times d_K}$ where $[L] = \{1, 2, \cdots, L\}$.
- We assume the latent tensor $\mathcal{Y}^*$ before an L-level quantization.

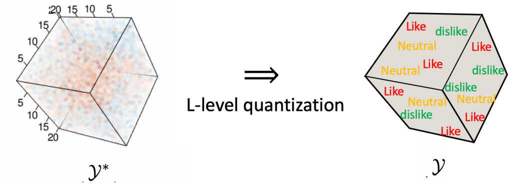


Probabilistic model: generating process

- Let $\mathcal{Y} = [\![y_\omega]\!] \in [L]^{d_1 \times \cdots \times d_K}$ where $[L] = \{1, 2, \cdots, L\}$.
- We assume the latent tensor $\mathcal{Y}^*$ before an L-level quantization.



- 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: a cumulative link model

- We summarize the generating process as

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

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

- More generally, we represent a cumulative link model:

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

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

Probabilistic model: a cumulative link model

- We summarize the generating process as

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

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

- More generally, we represent a cumulative link model:

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

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

- We assume the signal tensor Θ admits 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.

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.

$$(\hat{\Theta}, \hat{\boldsymbol{b}}) = \underset{\text{rank}(\Theta) \leq \boldsymbol{r}^*}{\arg\max} \ \mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b})$$

*There are more technical constraints on Θ and $\boldsymbol{b}$ in the main paper.

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.

- We show that our estimation bound is minimax 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 $\|\Theta - \hat{\Theta}\|^2_{F,\Pi} = \sum_{\omega \in [d_1]\times\cdots\times[d_K]} \pi_\omega (\Theta_\omega - \hat{\Theta}_\omega)^2$.

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,

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

- 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;

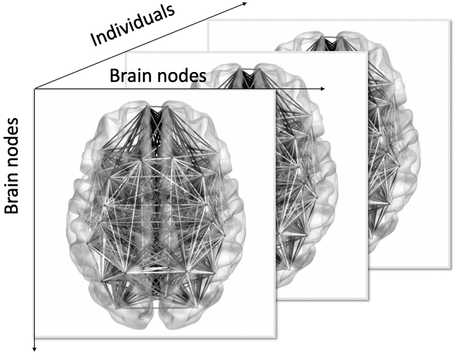
Repeat until converge;

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

end

Data application: Human Connectome Project (HCP)

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

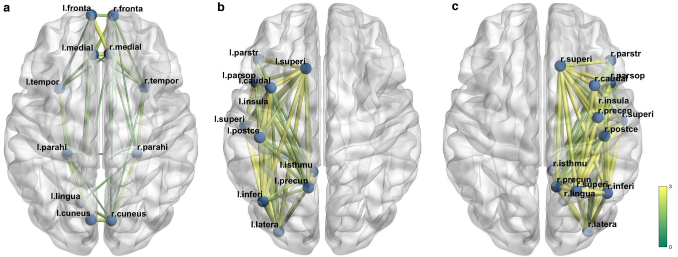


Data application: Human Connectome Project (HCP)

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

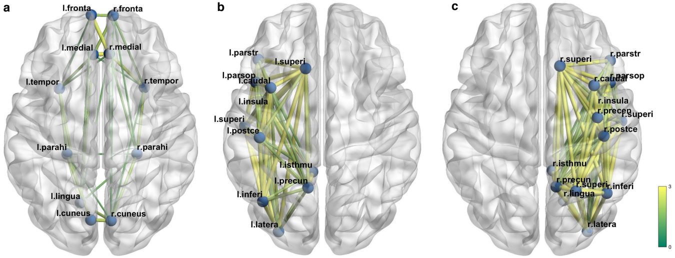
Data application: Human Connectome Project (HCP)

- 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)

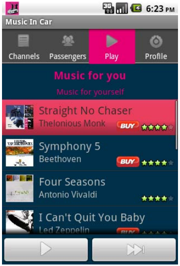
- The clustering based on the estimated $\hat{\Theta}$ identifies 11 (3+8) clusters among 68 brain nodes.
- The top three clusters capture the global separation among brain nodes.



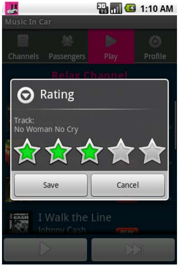
- The small clusters represent local regions driving by similar nodes.

Data application: InCarMusic

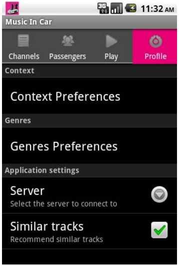
- A tensor recording the ratings of 139 songs from 42 users on 26 contexts (Baltrunas et al., 2011).
- Each entry is a rating on a scale of 1 to 5 ($y_\omega \in \{1, 2, 3, 4, 5\}$).



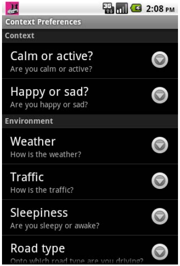
(a) Tracks Proposed to Play



(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
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.

Example

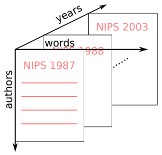
- True model:

$$\mathbb{P}(y_\omega = \ell) = \text{logistic}(b_\ell - \theta_\omega) - \text{logistic}(b_{\ell-1} - \theta_\omega) \text{ for } \ell = 1, 2,$$

where $\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \boldsymbol{M}_2 \times_3 \boldsymbol{M}_3$, and $(b_1, b_2) = (-0.2, 0.2)$.

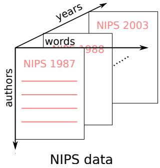
```
library(rTensor)
# Generate the signal tensor
alpha = 1
A_1 = matrix(runif(50*2,min=-1,max=1),nrow = 50)
A_2 = matrix(runif(50*2,min=-1,max=1),nrow = 50)
A_3 = matrix(runif(50*2,min=-1,max=1),nrow = 50)
C = as.tensor(array(runif(2^3,min=-1,max=1),dim = c(2,2,2)))
theta = ttm(ttm(ttm(C,A_1,1),A_2,2),A_3,3)@data
theta = alpha*theta/max(abs(theta))
omega = c(-0.2,0.2)
library(TensorComplete)
# Observed tensor
ttnsr <- realization(theta,omega)@data
# Estimation of parameters
ordinal_est = fit_ordinal(ttnsr,c(2,2,2),omega = TRUE,alpha = 1)
# Estimation error
mean((ordinal_est$theta-theta)^2)
```

Other data applications: NIPS



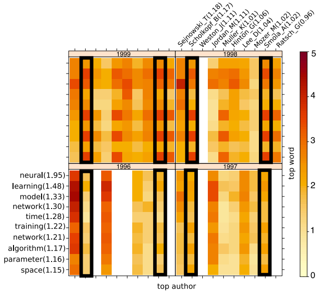
- The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003.
- Data tensor $\mathcal{Y} \in \mathbb{R}^{100\times 200\times 17}$.

Other data applications: NIPS



- The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003.
- Data tensor $\mathcal{Y} \in \mathbb{R}^{100\times 200\times 17}$.

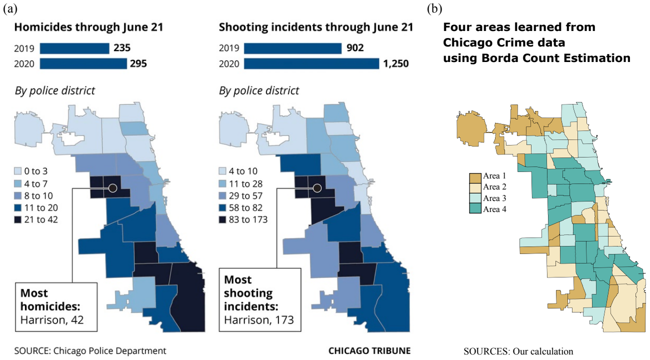
- We examine the estimated signal tensor $\hat{\Theta}$.
- Most frequent words are consistent with the active topics
- Strong heterogeneity among word occurrences across authors and years.
- Similar word patterns (B. Schölkopf and A. Smola).



Reference: C. Lee and M. Wang. Beyond the Signs: Nonparametric tensor completion via sign series. Advances in Neural Information Processing Systems 35 (NeurIPS), 2021.

Other data applications: Chicago crime data

- The Chicago dataset consists of crime counts from 24 hours, 77 community ares, and 32 crime types.
- Data tensor $\mathcal{Y} \in \mathbb{R}^{24\times 77\times 32}$.
- The estimated 4 clusters for crime areas are consistent with actual locations.



Reference: C. Lee and M. Wang. Smooth tensor estimation with unknown permutations. NeurIPS 2021 Workshop on Quantum Tensor Networks in Machine Learning. 2021.

Summary

We have developed efficient statistical methods for analyzing specially-structured tensors.

References:

- **C. Lee** and M. Wang. Tensor denoising and completion based on ordinal observations. Proceedings of International Conference on Machine Learning (ICML), 2020.
- **C. Lee** and M. Wang. Beyond the Signs: Nonparametric tensor completion via sign series. Advances in Neural Information Processing Systems 35 (NeurIPS), 2021.
- **C. Lee** and M. Wang. Smooth tensor estimation with unknown permutations. arXiv:2111.04681, 2021.
- **C. Lee** and M. Wang. Package 'TensorComplete'. R package, 2021.