Tensor denoising and completion based on ordinal observations

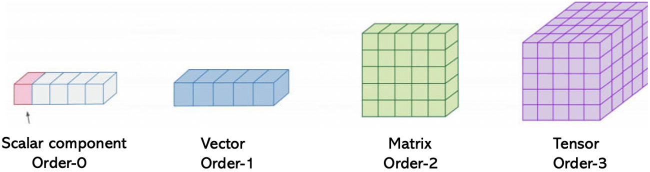
Chanwoo Lee
Joint work with Miaoyan Wang

Department of Statistics
University of Wisconsin - Madison

IFDS, March 9, 2020

Introduction: what is a tensor?

- Tensors are generalizations of vectors and matrices:



- We focus on tensors of order 3 or greater, also called higher-order tensors.
- Denote an order-$K(d_1, \cdots, d_K)$ dimensional tensor as $\mathcal{Y} = [\![y_\omega]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ where $\omega \in [d_1] \times \cdots \times [d_K]$.

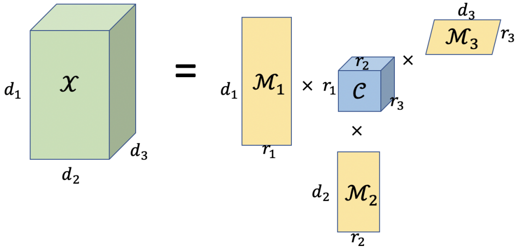
Introduction: Tucker decomposition

- Tucker decomposition
  - Generalization of matrix SVD to higher orders.
  - $\mathcal{X} = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \boldsymbol{M}_2 \times_3 \boldsymbol{M}_3$.
  - Tucker rank of an order-3 tensor is defined as

$$r(\mathcal{X}) = (r_1, r_2, r_3).$$

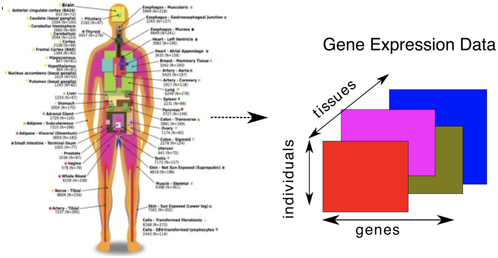
  - Degree of freedom (the number of parameters) is

$$\sum_k (d_k - r_k) r_k + \prod_k r_k \approx \mathcal{O}\left(\sum_k d_k\right) \text{ when } r_k = \mathcal{O}(1).$$

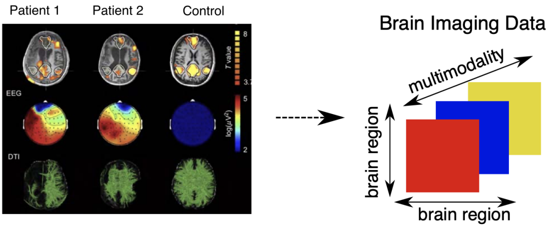


Introduction: tensor data in applications

- Tensor in genomics.

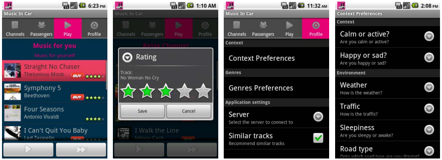


- Tensor in neuroimaging.



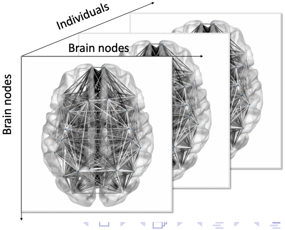
Introduction: **ordinal** tensor data in applications

- 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

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



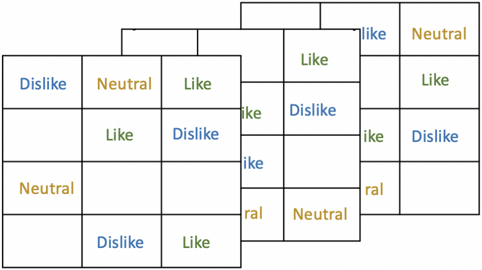
Tensor-based learning is an active but challenging field

- Tensor decomposition (Anandkumar et al, JMLR'14; Wang and Song, AISTATS'17; Han and Zhang, JASA'19).
- Tensor regression (Zhou et al JASA'13; Chen, Raskutti, and Yuan,JMLR'20; Xu, Hu and Wang'19).
- Tensor denoising (Wang and Li'18; Hong, Kolda, and Duersch, SIAM AR'19; Zeng and Wang, NeurIIPS'19).
- Tensor completion (Montanari and Sun, CPAM'16; Zhang AOS'19; Ghadermarzy, Plan and Yilmaz, I&A'19).

No existing method is able to analyze oridnal-valued tensors.

Motivating problems

- How can we fill the missing ordinal values from the available tensor data?
- How many ordinal samples do we need to complete the tensor?



- This talk is based on: L. and M. Wang. Tensor denoising and completion based on ordinal observations. arXiv:2002.06524, 2020.

Probabilistic model

- 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.
- Continuous tensor model lacks the first property.
- Binary tensor model lacks the second property.

Probabilistic model

Proposal: 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:

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

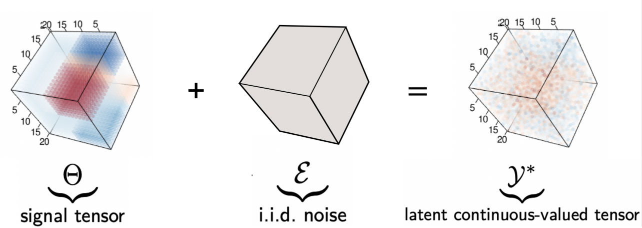
  ex) $f(x) = \frac{e^x}{1+e^x}$ is a logistic link.
- The additive, cumulative model enjoys two key properties for ordinal tensor data.
- If f is a cumulative function,

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

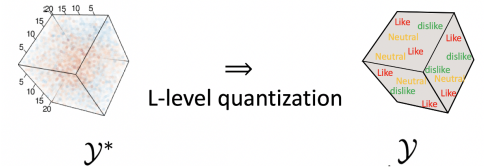
  where $\epsilon_\omega \overset{i.i.d}{\sim} f$ and $y^*_\omega = \theta_\omega + \epsilon_\omega$.

Latent variable interpretation

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



- 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 f

- The link function is assumed to satisfy:
  - $f(\theta)$ is strictly increasing and twice-differentiable in θ.
  - $f'(\theta)$ is strictly log-concave and symmetric with respect to $\theta = 0$.
- Many cumulative functions satisfy the above two assumptions.

Probabilistic model: assumptions on Θ

- The parameter Θ admits the Tucker decomposition:

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

  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

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

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

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

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

$$\mathcal{B} = \{\boldsymbol{b} \in \mathbb{R}^{L-1} \colon \|\boldsymbol{b}\|_\infty \leq \beta,\ \min_\ell (b_\ell - b_{\ell-1}) \geq \Delta\}.$$

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

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 1: Alternating optimization

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

Random initialization;

Repeat until converge;

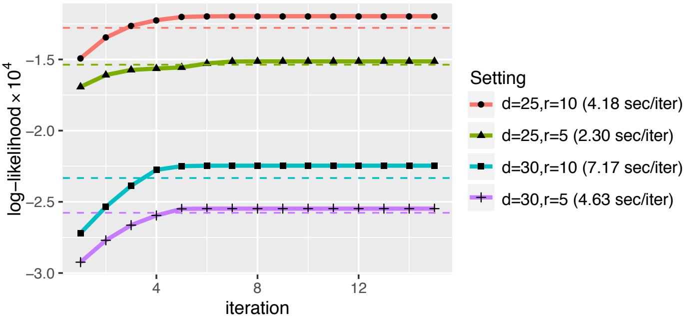
$$\mathcal{C}^{(n)} = \arg\max_{\mathcal{C}} \mathcal{L}_{\mathcal{Y},\Omega}(\mathcal{C}, \mathcal{M}_1^{(n-1)}, \cdots, \mathcal{M}_k^{(n-1)}, \boldsymbol{b}^{(n-1)}).$$
$$\mathcal{M}_1^{(n)} = \arg\max_{\mathcal{M}_1} \mathcal{L}_{\mathcal{Y},\Omega}(\mathcal{C}^{(n)}, \mathcal{M}_1, \cdots, \mathcal{M}_k^{(n-1)},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

end

- There is no guarantee on global optimality.

Algorithm

- However, our theoretical results hold as long as $\mathcal{L}_{\mathcal{Y},\Omega}(\hat{\Theta}) \geq \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}})$.
- The algorithm performs well in simulations and data applications.



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 also have general results for incomplete data, or unknown $\boldsymbol{b}$ cases.

unknown b case

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,

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

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,

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

- We allow both uniform and non-uniform sampling.
- 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.

Theoretical results: summary

- 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

Simulations

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

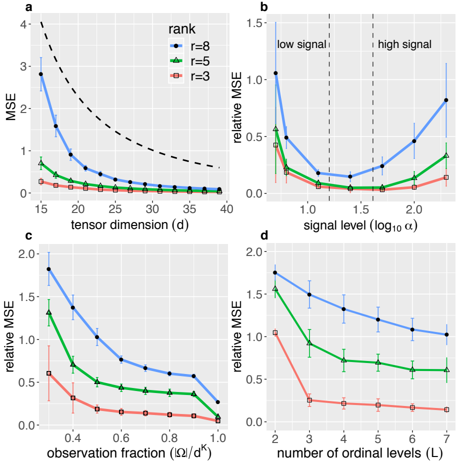
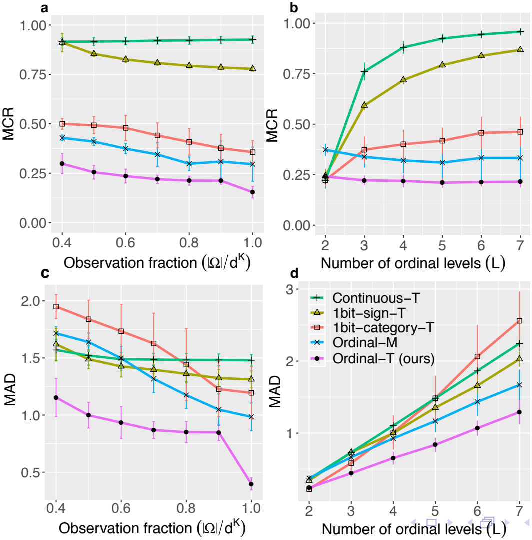


Figure: the relative MSE $= \|\hat{\Theta} - \Theta^{\text{true}}\|_F / \|\Theta^{\text{true}}\|_F$ for better visualization.

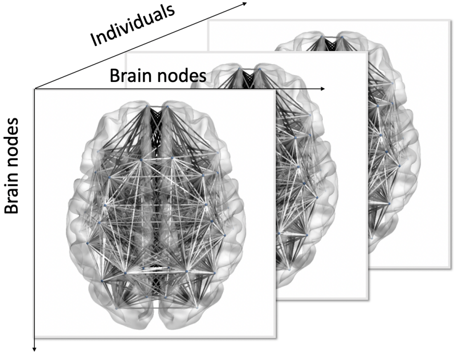
Simulations

- We compare our method to other 4 alternatives.
- Our method outperforms across a range of missingness and ordinal levels.



Data application: Human Connectome Project (HCP)

- An ordinal tensor consisting of structural connectivities among 68 brain nodes for 136 individuals (Van Essen et al., 2013).
- Each entry $y_\omega \in \{\text{high}, \text{moderate}, \text{low}\}$.

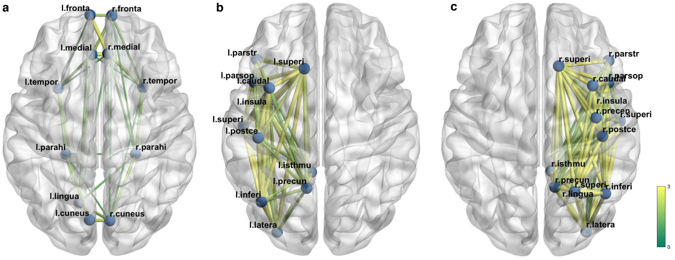


Data application: Human Connectome Project (HCP)

- The BIC suggests $\boldsymbol{r} = (23, 23, 8)$.
- The clustering based on the estimated $\hat{\Theta}$ identifies 11 (3+8) clusters among 68 brain nodes. clustering

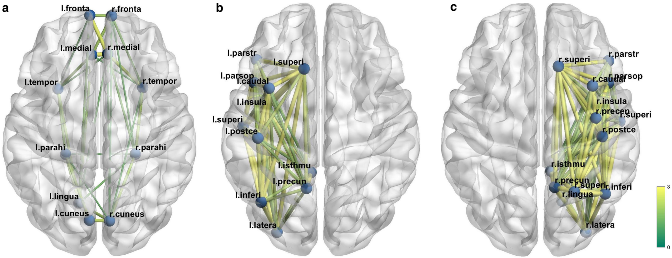
Data application: Human Connectome Project (HCP)

- The BIC suggests $\boldsymbol{r} = (23, 23, 8)$.
- The clustering based on the estimated $\hat{\Theta}$ identifies 11 (3+8) clusters among 68 brain nodes. clustering
- The top three clusters capture the global separation among brain nodes.



Data application: Human Connectome Project (HCP)

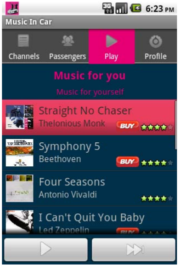
- The BIC suggests $\boldsymbol{r} = (23, 23, 8)$.
- The clustering based on the estimated $\hat{\Theta}$ identifies 11 (3+8) clusters among 68 brain nodes. clustering
- The top three clusters capture the global separation among brain nodes.



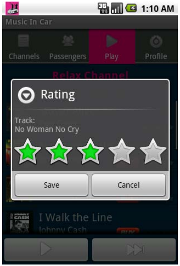
- The small clusters represent local regions driving by similar nodes.

Data application: InCarMusic

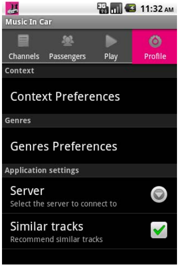
- An tensor recording the ratings from 42 users to 139 songs on 26 contexts (Baltrunas et al., 2011).
- Each entry is a rating on a scale of 1 to 5 ($y_\omega \in \{1, 2, 3, 4, 5\}$).



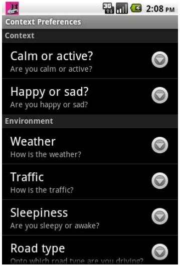
(a) Tracks Proposed to Play



(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)	0.59 (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.
- Future work:
  - Analysis of algorithmic error (global vs local).
  - Robustness of the model.
- Thank you!
- L. and M. Wang. Tensor denoising and completion based on ordinal observations. arXiv:2002.06524, 2020.

Unknown b case

- We make the following assumptions about the link function.

Assumption 1

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

1. *$f(z)$ is twice-differentiable and strictly increasing in z.*
2. *$\dot{f}(z)$ is strictly log-concave and symmetric with respect to $z = 0$.*

- We define the following constants that will be used in the theory:

$$C_{\alpha,\beta,\Delta} = \max_{|z|\leq\alpha+\beta} \max_{\substack{z'\leq z-\Delta \\ z''\geq z+\Delta}} \max\left\{\frac{\dot{f}(z)}{f(z)-f(z')}, \frac{\dot{f}(z)}{f(z'')-f(z)}\right\},$$

$$D_{\alpha,\beta,\Delta} = \max_{|z|\leq\alpha+\beta} \max_{\substack{z'\leq z-\Delta \\ z''\geq z+\Delta}} \max\left\{-\frac{\partial}{\partial z}\left(\frac{\dot{f}(z)}{f(z)-f(z')}\right), \frac{\partial}{\partial z}\left(\frac{\dot{f}(z)}{f(z'')-f(z)}\right)\right\}$$

$$A_{\alpha,\beta,\Delta} = \min_{|z|\leq\alpha+\beta} \min_{z'\leq z-\Delta} \left(f(z)-f(z')\right).$$

Unknown $\boldsymbol{b}$ case

- We have the following theorem corresponding to Theorem 1 in known $\boldsymbol{b}$ case.

Theorem 1 (Statistical convergence with unknown $\boldsymbol{b}$)

With very high probability,

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

and

$$\mathrm{MSE}\left(\hat{\boldsymbol{b}}, \boldsymbol{b}^{\mathrm{true}}\right) \leq \min\left(4\beta^2,\ c_1 r_{\max}^{K-1} \frac{L-1+\sum_k d_k}{(L-1)\prod_k d_K}\right),$$

where $c_1, C_{\alpha,\beta,\Delta}, D_{\alpha,\beta,\Delta}$ are positive constants independent of the tensor dimension, rank, and number of ordinal levels.

Clustering method

- In matrices case,
  1. Perform singular value decomposition,

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

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

Clustgering method

- In tensors case,
  1. Perform Tucker decompostion,
$$\hat{\Theta} = \hat{\mathcal{C}} \times_1 \hat{\boldsymbol{M}}_1 \times_2 \cdots \times_K \hat{\boldsymbol{M}}_K, \tag{2}$$
  2. The mode-k matricization of (2) gives
$$\hat{\Theta}_{(k)} = \hat{\boldsymbol{M}}_k \hat{\mathcal{C}}_{(k)} \left(\hat{\boldsymbol{M}}_K \otimes \cdots \otimes \hat{\boldsymbol{M}}_1 \right),$$
  3. Take each column in $\left(\hat{\boldsymbol{M}}_K \otimes \cdots \otimes \hat{\boldsymbol{M}}_1 \right)$ as principal axis and each row in $\hat{\boldsymbol{M}}_k \hat{\mathcal{C}}_{(k)}$ as principal component.
  4. A subsequent multivariate clustering method (such as K-means) is then applied to the d_k rows of the matrix $\hat{\boldsymbol{M}}_k \hat{\mathcal{C}}_{(k)}$.

References I

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.

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

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.