TenIPS: Inverse Propensity Sampling for Tensor Completion

Chengrun Yang, Lijun Ding, Ziyang Wu, Madeleine Udell

Cornell University

Tensors

On an *order*-3 tensor $\mathcal{B}$, for each of the modes $n \in [3] := \{1,2,3\}$:

- size of the n-th mode: I_n
- mode-n fibers: fixing every index but the n-th. e.g., mode-1 fiber: $\mathcal{B}_{:jk}$
- mode-n unfolding: matrix $\mathcal{B}^{(n)}$, whose columns are mode-n fibers

tensor decomposition: CP, Tucker (this paper), tensor-train, . . .

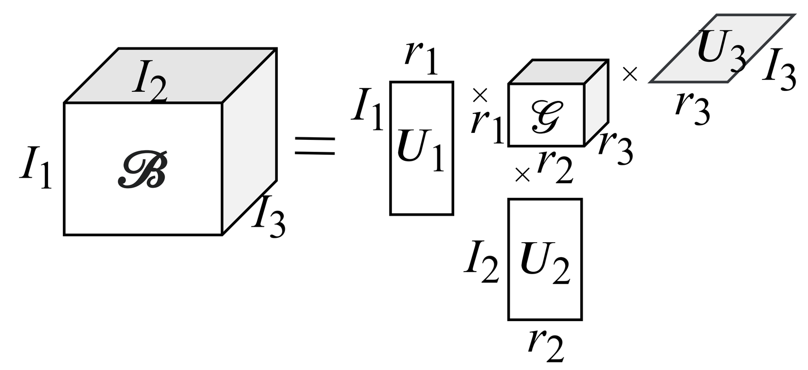


Figure 1: Tucker decomposition with multilinear rank (r_1, r_2, r_3): $\mathcal{B} = \mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3$.

tensor completion

Given a partially observed $\mathcal{B}_{\text{obs}} \in \mathbb{R}^{I_1\times\cdots\times I_N}$, we have

- observation pattern $\Omega \in \mathbb{R}^{I_1\times\cdots\times I_N}$: $\Omega_{i_1\ldots i_N} = 1$ if $\mathcal{B}_{i_1\ldots i_N}$ is observed, and 0 otherwise
- observation probability $\mathcal{P} \in \mathbb{R}^{I_1\times\cdots\times I_N}$: $\mathcal{P}_{i_1\ldots i_N} = \mathbb{P}(\Omega_{i_1\ldots i_N} = 1) = \mathbb{P}(\mathcal{B}_{i_1\ldots i_N}$ is observed)

missingness types	$\{\mathcal{P}_{i_1\ldots i_N}\}$
missing-completely-at-random (MCAR)	uniform
missing-not-at-random (MNAR)	non-uniform

1-bit matrix completion

Given a binary matrix $Y \in \{0,1\}^{m\times n}$, predict the parameter matrix $M \in \mathbb{R}^{m\times n}$

Assumptions:

- M is approximately low rank.
- There exists a link function σ: $\mathbb{R} \to [0,1]$, such that $\mathbb{P}(Y_{ij} = 1) = \sigma(M_{ij})$ for $(i,j) \in [m]\times[n]$.

Low rank surrogates for M: low nuclear norm, low max norm, . . .

Our problem formulation: MNAR tensor completion

Input: MNAR data tensor $\mathcal{B}_{\text{obs}} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$

Assumptions:

- true data tensor $\mathcal{B} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$ is **approximately low multilinear rank**
- **noiseless observation**: $(\mathcal{B}_{\text{obs}})_{i_1\ldots i_N} = \mathcal{B}_{i_1\ldots i_N}$ if $\mathcal{B}_{i_1\ldots i_N}$ is observed, and 0 otherwise
- unknown parameter tensor $\mathcal{A} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$ has the **same rank structure** as $\mathcal{B}$
- **1-bit observation**: With the observation propensity tensor $\mathcal{P} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$, $\mathbb{P}(\mathcal{B}_{i_1 i_2\cdots i_N}$ is observed$) = \mathcal{P}_{i_1 i_2\cdots i_N} = \sigma(\mathcal{A}_{i_1 i_2\cdots i_N})$, in which σ: $\mathbb{R} \to [0,1]$ is a non-decreasing link function.

Algorithm Step 1: propensity recovery

Given a mask tensor Ω, get a predicted propensity tensor $\widehat{\mathcal{P}}$.

	algorithm	hyperparameters
Choice 1: ConvexPE	proximal-proximal-gradient	τ and γ
Choice 2: NonconvexPE	gradient descent	target rank and step size

ConvexPE: convex and provable

1. Get the **square set** and **square unfolding** [5] of $\Omega \in \mathbb{R}^{I_1\times\cdots\times I_N}$:
   - square set $S_\square := \arg\min_{S\subset[N]} \left|\prod_{n\in S} I_n - \prod_{n\in[N]\setminus S} I_n\right|$,
   - square unfolding $\Omega_\square := \texttt{reshape}(\pi_{S_\square}(\Omega)^{(1)}, \prod_{n\in S_\square} I_n, \prod_{n\in[N]\setminus S_\square} I_n)$, in which $\pi_S = (S_1, \ldots, S_{|S|}, S_1^C, \ldots, S_{N-|S|}^C)$ is a permutation map of the N modes
2. Compute parameter tensor $\mathcal{A}$ by **logistic loss minimization** (by proximal-proximal-gradient [6])
$$\widehat{\mathcal{A}}_\square = \operatorname*{argmin}_{\Gamma\in\mathcal{S}_{\tau,\gamma}} \sum_{i=1}^{I_\square}\sum_{j=1}^{I_{\square^C}} -(\Omega_\square)_{i,j}\log\sigma(\Gamma_{i,j}) - [1-(\Omega_\square)_{i,j}]\log[1-\sigma(\Gamma_{i,j})],$$
where $\mathcal{S}_{\tau,\gamma} = \left\{\Gamma \in \mathbb{R}^{I_\square\times I_{\square^C}} : \|\Gamma\|_\star \le \tau\sqrt{I_{[N]}}, \|\Gamma\|_{\max} \le \gamma\right\}$.
3. Estimate propensities: $\widehat{\mathcal{P}} = \sigma(\widehat{\mathcal{A}})$

NonconvexPE: nonconvex, gradient descent

1. Initialize core tensor and factor matrices $\mathcal{G}^{\mathcal{A}}, U_1^{\mathcal{A}}, \ldots, U_N^{\mathcal{A}} \leftarrow \bar{\mathcal{G}}^{\mathcal{A}}, \bar{U}_1^{\mathcal{A}}, \ldots, \bar{U}_N^{\mathcal{A}}$
2. Define objective
$$f(\mathcal{G}^{\mathcal{A}}, \{U_n^{\mathcal{A}}\}_{n\in[N]}) = \sum_{i_1\cdots i_N} -\Omega_{i_1\cdots i_N}\log\sigma(\widehat{\mathcal{A}}_{i_1\cdots i_N}) - (1-\Omega_{i_1\cdots i_N})\log[1-\sigma(\widehat{\mathcal{A}}_{i_1\cdots i_N})],$$
in which $\widehat{\mathcal{A}} = \mathcal{G}^{\mathcal{A}} \times_1 U_1^{\mathcal{A}} \times_2 \cdots \times_N U_N^{\mathcal{A}}$.
3. **Gradient descent updates**
4. Estimate propensities: $\widehat{\mathcal{P}} = \sigma(\mathcal{G}^{\mathcal{A}} \times_1 U_1^{\mathcal{A}} \times_2 \cdots \times_N U_N^{\mathcal{A}})$

Algorithm Step 2: tensor completion

TenIPS: Given $\widehat{\mathcal{P}}$ and MNAR observations $\mathcal{B}_{\text{obs}}$, get $\widehat{\mathcal{B}}$

1. Form an **entrywise inverse propensity estimator** for data tensor $\mathcal{B}$ as $\bar{\mathcal{X}}(\widehat{\mathcal{P}}) = \sum_{(i_1,i_2,\ldots,i_N)\in\Omega} \frac{1}{\widehat{\mathcal{P}}_{i_1\cdots i_N}}\mathcal{B}_{\text{obs}} \odot \mathcal{E}(i_1,\ldots,i_N)$, in which
   - $\Omega := \{(i_1,\ldots,i_N) | \mathcal{B}_{i_1\ldots i_N}$ is observed$\}$
   - $\mathcal{E}(i_1\ldots,i_N)$ is a binary tensor with the same shape as $\mathcal{B}$, with value 1 at the $(i_1,i_2,\ldots,i_N)$-th entry and 0 elsewhere.
2. Do **Tucker decomposition** on $\bar{\mathcal{X}}(\widehat{\mathcal{P}})$, get core tensor $\mathcal{W}(\widehat{\mathcal{P}})$ and factor matrices $\{Q_n(\widehat{\mathcal{P}})\}_{n\in[N]}$.
3. Estimate $\mathcal{B}$ by $\widehat{\mathcal{B}}(\widehat{\mathcal{P}}) = \mathcal{W}(\widehat{\mathcal{P}}) \times_1 Q_1(\widehat{\mathcal{P}}) \times_2 \cdots \times_N Q_N(\widehat{\mathcal{P}})$.

Theoretical guarantees

- **Upper bound for propensity recovery error [1, 3]**
  Assume that $\mathcal{P} = \sigma(\mathcal{A})$. Given a set $S \subset [N]$, together with the following assumptions:
  A1. $\mathcal{A}_S$ has bounded nuclear norm: there exists a constant $\theta > 0$ such that $\|\mathcal{A}_S\|_\star \le \theta\sqrt{I_{[N]}}$.
  A2. Entries of $\mathcal{A}$ have bounded absolute value: there exists a constant $\alpha > 0$ such that $\|\mathcal{A}\|_{\max} \le \alpha$.
  Suppose we run ConvexPE with thresholds satisfying $\tau \ge \theta$ and $\gamma \ge \alpha$ to obtain an estimate $\widehat{\mathcal{P}}$ of $\mathcal{P}$. With $L_\gamma := \sup_{x\in[-\gamma,\gamma]} \frac{|\sigma'(x)|}{\sigma(x)(1-\sigma(x))}$, there exists a universal constant $C > 0$ such that if $I_S + I_{S^C} \ge C$, with probability at least $1 - \frac{C}{I_S + I_{S^C}}$, the propensity estimation error $\frac{1}{I_{[N]}}\|\widehat{\mathcal{P}} - \mathcal{P}\|_F^2 \le 4eL_\gamma\tau\left(\frac{1}{\sqrt{I_S}} + \frac{1}{\sqrt{I_{S^C}}}\right)$.
- **Optimality of the square unfolding for propensity recovery**: Instate the same conditions as the previous lemma on propensity recovery error, and further assume that there exists a constant $c > 0$ such that $r_n^{\text{true}} \le cI_n$ for every $n \in [N]$. Then $S = S_\square$ gives the tightest upper bound on the propensity estimation error $\|\widehat{\mathcal{P}} - \mathcal{P}\|_F$ among all unfolding sets $S \subset [N]$.
- **Tensor completion error on cubical tensors** (same size in every mode):
  Consider an order-N cubical tensor $\mathcal{B}$ with size $I_1 = \cdots = I_N = I$ and multilinear rank $r_1^{\text{true}} = \cdots = r_N^{\text{true}} = r < I$, and two order-$N$ cubical tensors $\mathcal{P}$ and $\mathcal{A}$ with the same shape as $\mathcal{B}$. Each entry of $\mathcal{B}$ is observed with probability from the corresponding entry of $\mathcal{P}$. Assume $I \ge rN\log I$, and there exist constants $\psi, \alpha \in (0,\infty)$ such that $\|\mathcal{A}\|_{\max} \le \alpha$, $\|\mathcal{B}\|_{\max} = \psi$. Further assume that for each $n \in [N]$, the condition number $\sigma_1(\mathcal{B}^{(n)})/\sigma_r(\mathcal{B}^{(n)}) \le \kappa$ is a constant independent of tensor sizes and dimensions. Then under the conditions of the lemma on convex propensity recovery error, with probability at least $1 - I^{-1}$, the fixed multilinear rank $(r,r,\ldots,r)$ approximation $\widehat{\mathcal{B}}(\widehat{\mathcal{P}})$ computed from ConvexPE and TenIPS with thresholds $\tau \ge \theta$ and $\gamma \ge \alpha$ satisfies
$$\frac{\|\widehat{\mathcal{B}}(\widehat{\mathcal{P}}) - \mathcal{B}\|_F}{\|\mathcal{B}\|_F} \le CN\sqrt{\frac{r\log I}{I}},$$
  in which C depends on κ.

Numerics

ConvexPE to recover a size-8 cubical propensity tensor with approximately low rank:

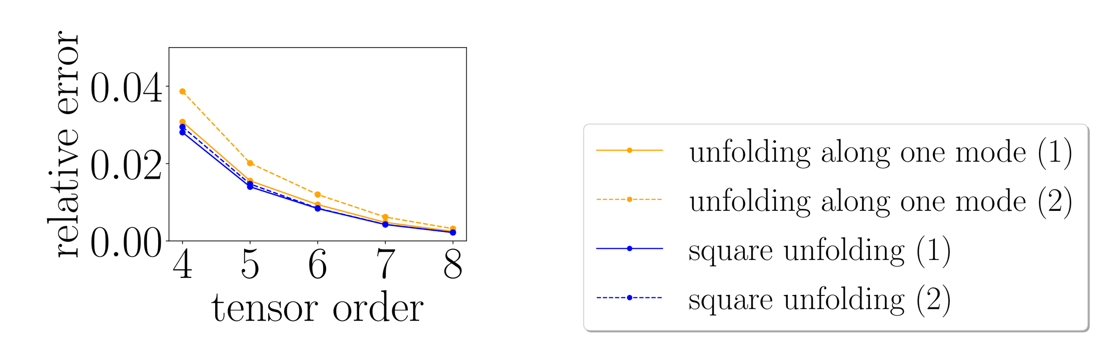


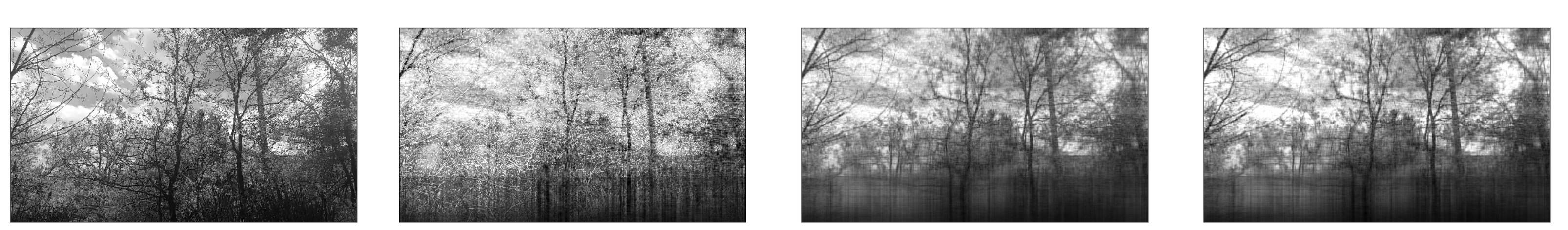
Figure 2: "(1)": setting $\tau = \theta$, $\gamma = \alpha$; "(2)": setting $\tau = 2\theta$, $\gamma = 2\alpha$

MNAR tensor completion on synthetic data:

Algorithm	time (s)	relative error $\|\widehat{\mathcal{B}}(\widehat{\mathcal{P}}) - \mathcal{B}\|_F / \|\mathcal{B}\|_F$		
		with $\mathcal{P}$	with $\widehat{\mathcal{P}}_1$	with $\widehat{\mathcal{P}}_2$
TenIPS	26	**0.110**	**0.110**	**0.109**
HOSVD_w [2]	35	0.129	0.116	0.110
SqUnfold	29	0.141	0.138	0.139
RectUnfold	**8**	0.259	0.256	0.256
LstSq	>600	-	-	-
SO-HOSVD [7]	>600	-	-	-

MNAR tensor completion on semi-synthetic data:

- real video tensor from [4]: $\mathcal{B} \in [0,255]^{2200\times1080\times1920}$
- synthetic parameter tensor $\mathcal{A} = (\mathcal{B} - 128)/64$



(a) original (b) TenIPS, assuming MCAR (c) TenIPS, assuming MNAR, with true $\mathcal{P}$ (d) TenIPS, assuming MNAR, with estimated $\widehat{\mathcal{P}}$

Thanks!

- Chengrun Yang: cy438@cornell.edu
- Madeleine Udell: udell@cornell.edu

Bibliography

[1] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[2] Longxiu Huang and Deanna Needell. Hosvd-based algorithm for weighted tensor completion. *arXiv preprint arXiv:2003.08537*, 2020.

[3] Wei Ma and George H Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. In *Advances in Neural Information Processing Systems*, pages 14871–14880, 2019.

[4] Osman Asif Malik and Stephen Becker. Low-rank tucker decomposition of large tensors using tensorsketch. In *Advances in Neural Information Processing Systems*, pages 10096–10106, 2018.

[5] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *International conference on machine learning*, pages 73–81, 2014.

[6] Ernest K Ryu and Wotao Yin. Proximal-proximal-gradient method. *arXiv preprint arXiv:1708.06908*, 2017.

[7] Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *arXiv preprint arXiv:1711.04934*, 2017.