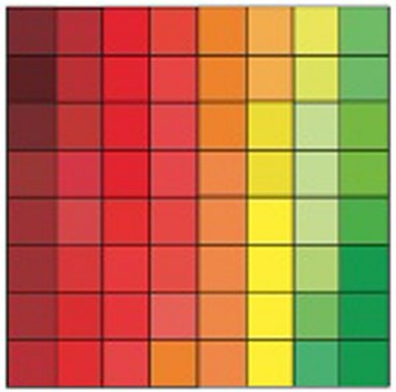
Efficient Nonconvex Regularized Tensor Completion with Structure-aware Proximal Iterations

Presenter: *Quanming Yao* (4Paradigm)

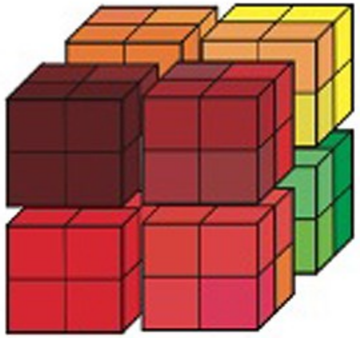
Joint work with *James T. Kwok* (HKUST) and *Bo Han* (RIKEN)

What is Tensor?

Vector



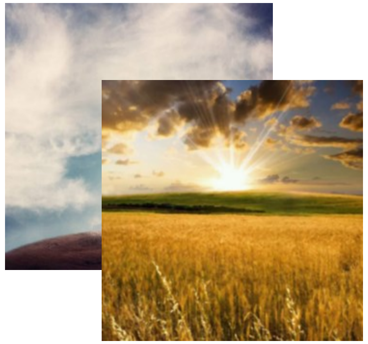
Matrix

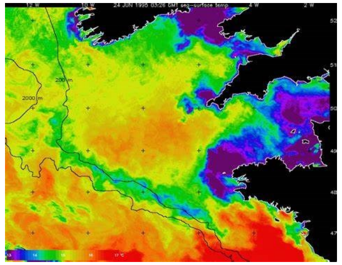


Tensor

Capture higher order interactions inside the data

Color images





Remote sensing data

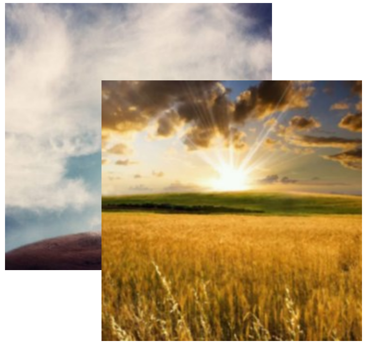
Spatiotemporal recommendation

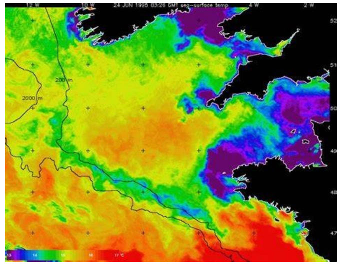
where to eat: (user, location, action)

Examples of 3-order tensors

Why needs Tensor Completion?

Color images





Remote sensing data

Spatiotemporal recommendation

where to eat: (user, location, action)

Image inpainting

Missing super-pixel / bands

Predict unknown triplet

Tensor completion: predict missing entries in the tensor

On 2-order tensor: reduce to matrix completion

How? - Overlapped nuclear norm

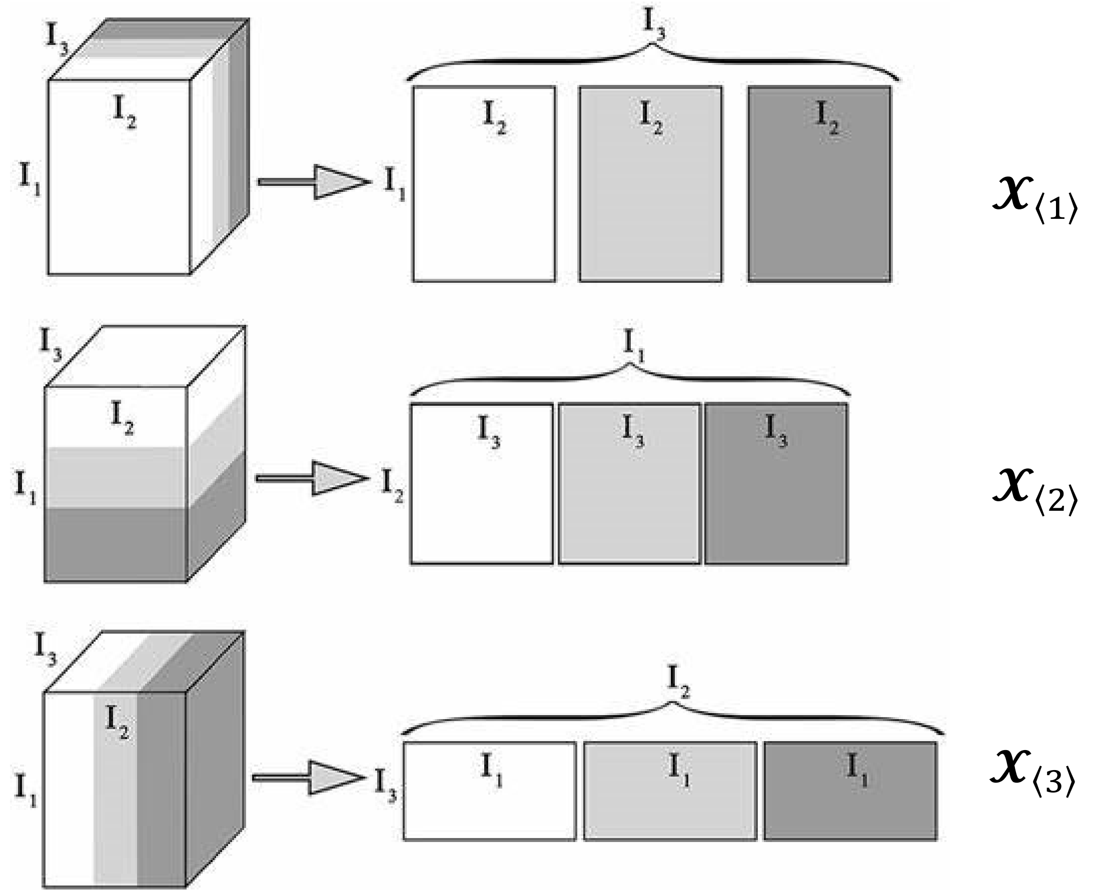
Nuclear norm $\|\mathbf{X}\|_*$ [Candes & Recht, 2009]

- Summation of all singular values of a matrix
- convex envelope of the matrix rank function

Tensor: overlapped nuclear norm [Tomioka et al., 2010]

Definition 1. *For a M-order tensor $\mathcal{X}$, the* overlapped nuclear norm *is* $\|\mathcal{X}\|_{overlap} = \sum_{m=1}^{M} \lambda_m \|\mathcal{X}_{\langle m \rangle}\|_*$, *where* $\{\lambda_m \geq 0\}$ *are hyperparameters.*

- $\mathcal{X}_{\langle m \rangle}$ unfold tensor along with mth mode
- encourage all unfolded matrix to be low-rank



unfolding operations

folding is the inverse of unfolding

Tensor Completion with Overlapped Nuclear Norm [Tomioka et al., 2010]

- Redundancy and correlations → low-rank approach is a power method in tensor completion
- Overlapped nuclear norm is a sound approach with statistical and convergence guarantee (compared with other tensor low-rank approaches [Tomioka et al., 2011; Liu et al., 2013; Guo et al., 2017])

$$\min_{\mathcal{X}} \underbrace{\frac{1}{2}\left\|P_{\Omega}\left(\mathcal{X}-\mathcal{O}\right)\right\|_F^2}_{\text{Squared loss on observed entries}} + \underbrace{\sum_{d=1}^{D} \lambda_m \left\|\mathcal{X}_{\langle m\rangle}\right\|_*}_{\text{Overlapped nuclear norm}}$$

However, two critical limitations

1. Inferior empirical performance
    - nuclear norm over penalize singular values
2. Expensive optimization
    - full tensor needs to be maintained due to folding / unfolding operations

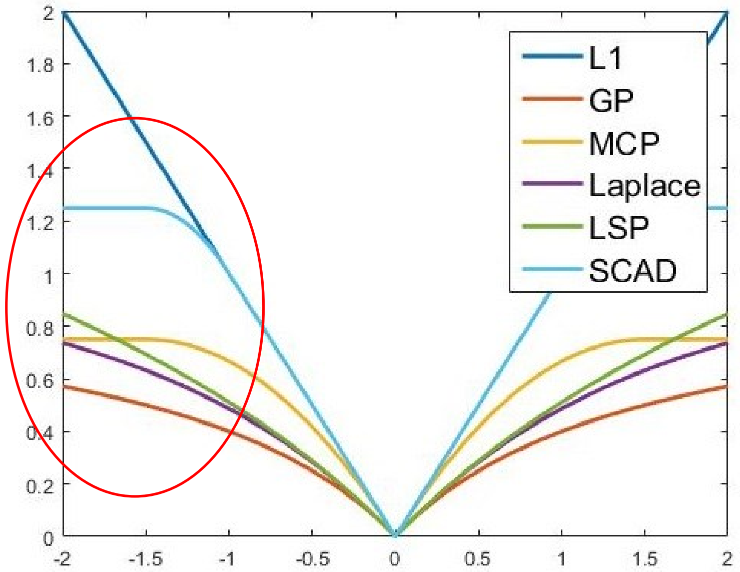
Proposed NORT: Nonconvex regularized tensor completion

Our contributions, propose NORT algorithm

1. Improve the performance of overlapped nuclear norm
   - Extend nonconvex regularization with overlapped nuclear norm
2. Speedup optimization with structure aware proximal iterations
   - Cheap iteration: propose a special way to do matrix multiplication without tensor folding/unfolding
   - Fast convergence: enhance proximal average with adaptive momentum

Improve Performance: nonconvex regularization

Objective: $$\min_{\mathcal{X}} F(\mathcal{X}) \equiv \frac{1}{2}\|P_{\Omega}(\mathcal{X}-\mathcal{O})\|_F^2 + \sum_{d=1}^{D}\frac{\lambda_d}{D}\phi(\mathcal{X}_{\langle d\rangle}).$$ where $$\phi(\mathbf{X}) = \overbrace{\sum_{i=1}^{n}\kappa(\sigma_i(\mathbf{X})),}^{\text{Nonconvex regularization}}$$



Less penalize large singular values, which are more informative

Common examples of $\kappa(\sigma_i(\mathbf{X}))$. Here, θ is a constant. For capped-ℓ_1, LSP and MCP, $\theta > 0$; for SCAD, $\theta > 2$; and for TNN, θ is a positive integer.

	$\kappa(\sigma_i(\mathbf{X}))$
nuclear norm	$\sigma_i(\mathbf{X})$
capped-ℓ_1	$\min(\sigma_i(\mathbf{X}), \theta)$
LSP	$\log(\sigma_i(\mathbf{X})/\theta + 1)$
TNN [27]	$\begin{cases} \sigma_i(\mathbf{X}) & \text{if } i > \theta \\ 0 & \text{otherwise} \end{cases}$
SCAD	$\begin{cases} \sigma_i(\mathbf{X}) & \text{if } \sigma_i(\mathbf{X}) \leq 1 \\ \frac{2\theta\sigma_i(\mathbf{X}) - \sigma_i(\mathbf{X})^2 - 1}{2(\theta-1)} & \text{if } 1 < \sigma_i(\mathbf{X}) \leq \theta \\ (\theta+1)^2/2 & \text{otherwise} \end{cases}$
MCP	$\begin{cases} \sigma_i(\mathbf{X}) - \alpha^2/2\theta & \text{if } \sigma_i(\mathbf{X}) \leq \theta \\ \theta^2/2 & \text{otherwise} \end{cases}$

Speedup optimization: Structure-aware proximal iterations

Proximal average algorithm [Bauschke et al., 2008; Yu, 2013]

$$\mathcal{X}_t = \frac{1}{K}\sum_{i=1}^{K}\mathcal{Y}_t^i,$$

$$\mathcal{Z}_t = \mathcal{X}_t - \frac{1}{\tau}\nabla f(\mathcal{X}_t),$$

$$\mathcal{Y}_{t+1}^i = \text{prox}_{\frac{\lambda_i}{\tau}g_i}(\mathcal{Z}_t), \quad i = 1, \dots, K.$$

maintain low-rank factorization →

$$\mathcal{X}_t = \frac{1}{D}\sum_{i=1}^{D}\left(\mathbf{U}_t^i(\mathbf{V}_t^i)^\top\right)^{\langle i\rangle}$$

sparse plus low-rank structure →

$$\mathcal{Z}_t = \frac{1}{D}\sum_{i=1}^{D}(\mathbf{U}_t^i(\mathbf{V}_t^i)^\top)^{\langle i\rangle} - \frac{1}{\tau}P_\Omega\left(\mathcal{X}_t - \mathfrak{O}\right)$$

proximal step with nonconvex regularization

$$\mathcal{Y}_{t+1}^i = \left[\text{prox}_{\frac{\lambda_i}{\tau}\phi}([\mathcal{Z}_t]_{\langle i\rangle})\right]^{\langle i\rangle}$$

matrix multiplications →

utilize sparse plus low-rank structure to efficient compute proximal step (Proposition 3.2)

$$(\mathcal{Z}_t)_{\langle i\rangle}\mathbf{b} = \frac{1}{D}\mathbf{U}_t^i[(\mathbf{V}_t^i)^\top\mathbf{b}] + \frac{1}{D}\sum_{j\neq i}[(\mathbf{U}_t^j(\mathbf{V}_t^j)^\top)^{\langle j\rangle}]_{\langle i\rangle}\mathbf{b} - \frac{1}{\tau}[P_\Omega\left(\mathcal{X}_t - \mathfrak{O}\right)]_{\langle i\rangle}\mathbf{b},$$

Needs folding/unfolding: full tensor computation

No folding/unfolding: fast and need less memory

NORT: the propose algorithm

Table 1. Comparison of the proposed NORT (Algorithm 1) and direct implementations of the PA algorithm.

	per-iteration time complexity	space	convergence
direct	$O(I_\times \sum_{i=1}^{D} I_i)$	$O(I_\times)$	slow
NORT	$O(\sum_{i=1}^{D} \sum_{j \neq i} (\frac{1}{I_i} + \frac{1}{I_j}) k_t^i k_{t+1}^i I_\times + \|\Omega\|_1 (k_t^i + k_{t+1}^i))$	$O(\sum_{i=1}^{D} \sum_{j \neq i} (\frac{1}{I_i} + \frac{1}{I_j}) k_t^i I_\times + \|\Omega\|_1)$	fast

Algorithm 1 NOnconvex Regularized Tensor (NORT).

initialize $\mathcal{X}_0 = \mathcal{X}_1 = 0$, $\tau > \rho + DL$ and $\gamma_1, p \in (0,1)$;
for $t = 1, \dots, T$ **do**
  $\mathcal{X}_{t+1} = \frac{1}{D} \sum_{i=1}^{D} (\mathbf{U}_{t+1}^i (\mathbf{V}_{t+1}^i)^\top)^{\langle i \rangle}$;
  $\bar{\mathcal{X}}_t = \mathcal{X}_t + \gamma_t (\mathcal{X}_t - \mathcal{X}_{t-1})$;
  if $F_\tau(\bar{\mathcal{X}}_t) \leq F_\tau(\mathcal{X}_t)$ **then**
    $\mathcal{V}_t = \bar{\mathcal{X}}_t$, $\gamma_{t+1} = \min(\frac{\gamma_t}{p}, 1)$;
  else
    $\mathcal{V}_t = \mathcal{X}_t$, $\gamma_{t+1} = p\gamma_t$;
  end if
  $\mathcal{Z}_t = \mathcal{V}_t - \frac{1}{\tau} P_\Omega (\mathcal{V}_t - \mathcal{O})$;
  // compute $P_\Omega (\mathcal{V}_t - \mathcal{O})$ using sparse tensor format;
  for $i = 1, \dots, D$ **do**
    $\mathbf{X}_{t+1}^i = \text{prox}_{\frac{\lambda_i}{\tau}\phi}((\mathcal{Z}_t)_{\langle i \rangle})$; // keep as $\mathbf{U}_t^i (\mathbf{V}_t^i)^\top$;
  end for
end for
output $\mathcal{X}_{T+1}$.

adaptive momentum

Theorem 3.5. *The sequence $\{\mathcal{X}_t\}$ generated from Algorithm 1 has at least one limit point, and all limits points are critical points of $F_\tau(\mathcal{X})$.*

Theorem 3.7. *Let $r_t = F_\tau(\mathcal{X}_t) - F_\tau^{\min}$. If F_τ has the uniformized KL property, for a sufficiently large t_0, we have*

1. *If $\beta = 1$, r_t reduces to zero in finite steps;*
2. *If $\beta \in [\frac{1}{2}, 1)$, $r_t \leq (\frac{d_1 C^2}{1 + d_1 C^2})^{t - t_0} r_{t_0}$ where $d_1 = \frac{2(\tau+\rho)^2}{\eta}$;*
3. *If $\beta \in (0, \frac{1}{2})$, $r_t \leq (\frac{C}{(t - t_0) d_2 (1 - 2\beta)})^{1/(1-2\beta)}$ where $d_2 = \min\{\frac{1}{2 d_1 C}, \frac{C}{1-2\beta}(2^{\frac{2\beta-1}{2\beta-2}} - 1) r_{t_0}\}$.*

- tensor size: $I_1 \times I_2 \times I_3$
- the speedup can be more the 100x on large tensors

Experiments: synthetic data

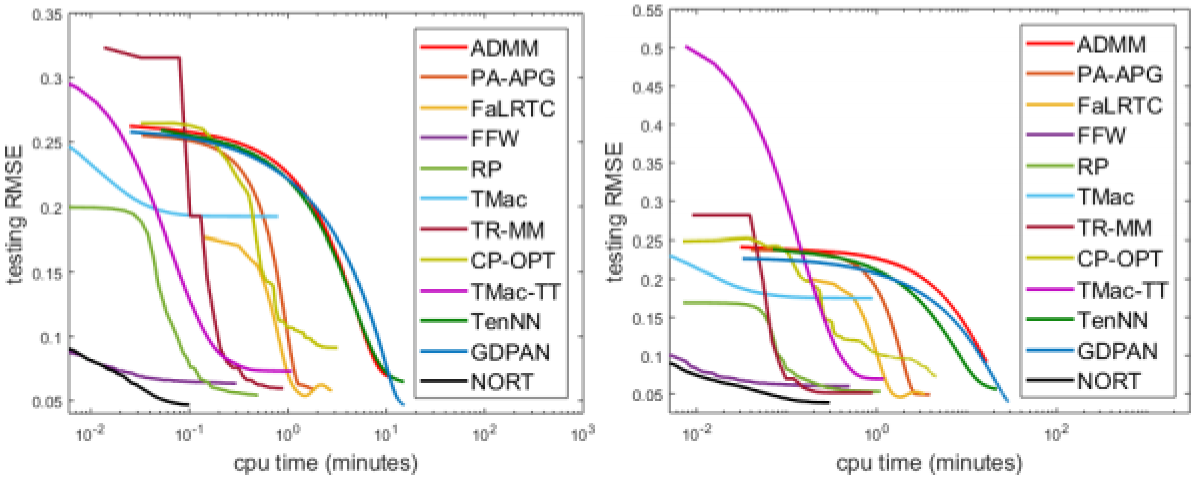
		small I_3: $\bar{c} = 100$, sparsity: 3.09%			large I_3: $\hat{c} = 40$, sparsity:2.70%		
		RMSE	space (MB)	time (sec)	RMSE	space (MB)	time (sec)
convex	PA-APG	0.0149±0.0011	302.4±0.1	2131.7±419.9	0.0098±0.0001	4804.5±598.2	6196.4±2033.4
(nonconvex) capped-ℓ_1	GDPAN	**0.0103±0.0001**	171.5±2.2	665.4±99.8	**0.0006±0.0001**	3243.3±489.6	3670.4±225.8
	sNORT	**0.0103±0.0001**	**14.0±0.8**	27.9±5.1	**0.0006±0.0001**	**44.6±0.3**	575.9±70.9
	NORT	**0.0103±0.0001**	14.9±0.9	**5.9±1.6**	**0.0006±0.0001**	66.3±0.6	89.4±13.4
(nonconvex) LSP	GDPAN	0.0104±0.0001	172.2±1.5	654.1±214.7	**0.0006±0.0001**	3009.3±376.2	3794.0±419.5
	sNORT	0.0104±0.0001	14.4±0.1	27.9±5.7	**0.0006±0.0001**	**44.6±0.2**	544.2±75.5
	NORT	0.0104±0.0001	15.1±0.1	**5.8±2.8**	**0.0006±0.0001**	62.1±0.5	**81.3±24.9**
(nonconvex) TNN	GDPAN	0.0104±0.0001	172.1±1.6	615.0±140.9	**0.0006±0.0001**	3009.2±412.2	3922.9±280.1
	sNORT	0.0104±0.0001	14.4±0.1	26.2±4.0	**0.0006±0.0001**	**44.7±0.2**	554.7±44.1
	NORT	**0.0103±0.0001**	15.1±0.1	**5.3±1.5**	**0.0006±0.0001**	63.1±0.6	**78.0±9.4**

- GDPAN is the direct proximal average algorithm
- Nonconvex regularization offers much lower testing RMSEs
- NORT is much faster, needs much less memory and achieves much lower testing RMSEs

Experiments: real data sets

Table 4. Testing RMSEs ($\times 10^{-1}$) on color images.

		rice	*tree*	*windows*
convex	ADMM	0.680±0.003	0.915±0.005	0.709±0.004
	PA-APG	0.583±0.016	0.488±0.007	0.585±0.002
	FaLRTC	0.576±0.004	0.494±0.011	0.567±0.005
	FFW	0.634±0.003	0.599±0.005	0.772±0.004
	TR-MM	0.596±0.005	0.515±0.011	0.634±0.002
	TenNN	0.647±0.004	0.562±0.004	0.586±0.003
factor-ization	RP	0.541±0.011	0.524±0.010	0.388±0.026
	TMac	1.923±0.005	1.750±0.006	1.313±0.005
	CP-OPT	0.912±0.086	0.733±0.060	0.964±0.102
	TMac-TT	0.729±0.022	0.697±0.147	1.045±0.107
noncvx	GDPAN	**0.467±0.002**	**0.388±0.012**	**0.296±0.007**
	NORT	**0.468±0.001**	**0.386±0.009**	**0.297±0.007**



(a) *rice*. (b) *tree*.

- NORT is fast and achieves lower testing RMSEs compared with other tensor low-rank approaches
- Same observations are on experiments with remote sensing data and multi-relational data (see our paper)

Thanks.

- Questions: yaoquanming@4paradigm.com
- Codes: available on my Github