A Highly-Accurate Approach to Dynamic Networks Representation Learning based on Biased Nonnegative Tensor Ring Factorization Supplementary File

Qu Wang, Hao Wu, *Member, IEEE*, and Xin Luo, *Fellow, IEEE*

I. Introduction

This is the supplementary file for paper entitled *A Highly-Accurate Approach to Dynamic Networks Representation Learning based on Biased Nonnegative Tensor Ring Factorization*. Supplementary definitions, formulas, experimental results are put into this file.

II. Supplementary Definitions

A. Canonical Polyadic Decomposition

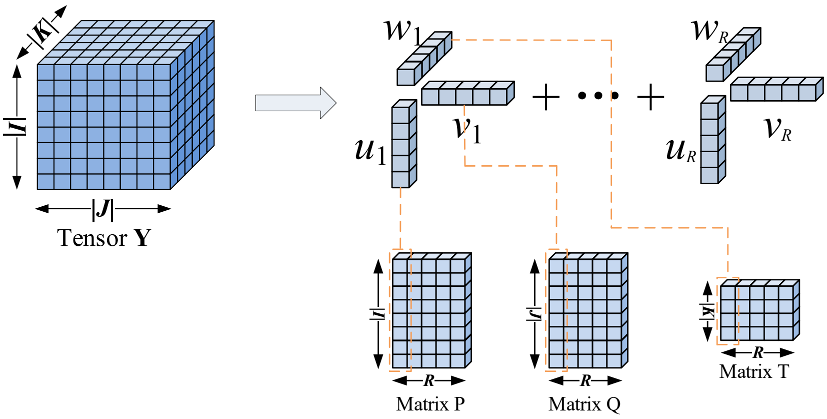


Fig. S1. CP decomposition.

Definition S1. (Canonical Polyadic decomposition) CP decomposition is a classic tensor decomposition method that decomposes a high-order tensor into the sum of multiple rank-one tensors. As shown in Fig. S1, given a third-order tensor, CP decomposition represents it as the sum of *R* rank-one tensors, where each rank-one tensor is obtained by the outer product of three vectors, one from each column of the corresponding factor matrix. So, the formula is:

$$\mathbf{Y} \approx \hat{\mathbf{Y}} = \sum_{r=1}^{R} u_r \circ v_r \circ w_r, \tag{S1}$$

where *R* is the rank of the decomposition, o represents the outer product, and u_r, v_r, w_r are the *r*-th column vectors of the factor matrices U, V, W respectively. Of course, using CP decomposition, we can get the calculation method of a single element in the low-rank approximation tensor $\hat{\mathbf{Y}}$:

$$\hat{y}_{ijk} = \sum_{r=1}^{R} u_{ir} v_{jr} w_{kr}. \tag{S2}$$

The core idea of CP decomposition is to capture the potential structure in the tensor through low-rank approximation, thereby achieving data dimensionality reduction and feature extraction. Its main advantage lies in its simple model structure and low computational complexity, but this also limits its representation capabilities.

B. Tucker Decomposition

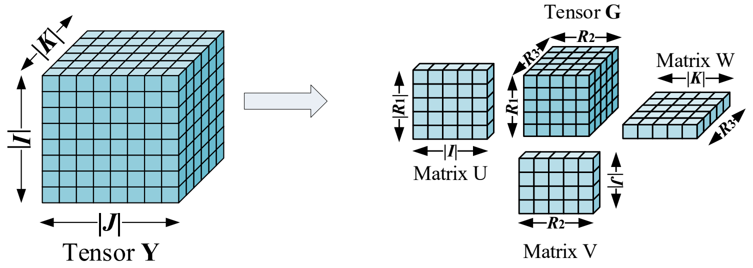


Fig. S2. Tucker decomposition.

Definition S2. (Tucker decomposition) Tucker decomposition is a high-order principal component analysis method that decomposes a high-order tensor into a core tensor and multiple factor matrices. As shown in Fig. S2, for a third-order tensor **Y**, the Tucker decomposition is as follows:

$$\mathbf{Y} \approx \hat{\mathbf{Y}} = \mathbf{G} \times_1 \mathrm{U} \times_2 \mathrm{V} \times_3 \mathrm{W}, \tag{S3}$$

where $\times_1$ represents the modal multiplication operation, which multiplies the factor matrix U with the first dimension of the core tensor. Of course, using Tucker decomposition, we can get the calculation method of a single element in the low-rank approximation tensor $\hat{\mathbf{Y}}$:

$$\hat{y}_{ijk} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} g_{r_1 r_2 r_3} u_{ir_1} v_{jr_2} w_{kr_3}, \tag{S4}$$

where $\{R_1, R_2, R_3\}$ represents the rank set of Tucker decomposition. Tucker decomposition reduces the dimensionality of data through core tensors and factor matrices. Due to the introduction of core tensors, Tucker decomposition can better capture the complex interactions between data compared to CP decomposition, and is particularly suitable for scenarios where data has a strong correlation structure. However, the computational complexity of Tucker decomposition grows exponentially with the dimensionality of the core tensor, so huge computing resources are required when processing high-order tensors.

III. Supplementary proof of convergence

A. Build Lagrangian Function

To establish the First, we let $\tilde{\mathbf{U}}$, $\tilde{\mathbf{V}}$, $\tilde{\mathbf{W}}$, $\tilde{\mathrm{D}}$, $\tilde{\mathrm{E}}$, $\tilde{\mathrm{F}}$ be Lagrangian multipliers due to the non-negativity constraints on latent feature tensors U, V, W and bias matrices D, E, F. Then, the Lagrangian function L for (14) is

$$\begin{aligned} L = \varepsilon\left(\mathbf{U},\mathbf{V},\mathbf{W},\mathrm{D},\mathrm{E},\mathrm{F}\right) - \sum_{r_3=1}^{R} \sum_{i=1}^{|I|} \sum_{r_1=1}^{R} \tilde{u}_{r_3 i r_1} u_{r_3 i r_1} - \sum_{r_1=1}^{R} \sum_{j=1}^{|J|} \sum_{r_2=1}^{R} \tilde{v}_{r_1 j r_2} v_{r_1 j r_2} \\ - \sum_{r_2=1}^{R} \sum_{k=1}^{|K|} \sum_{r_3=1}^{R} \tilde{w}_{r_2 k r_3} w_{r_2 k r_3} - \sum_{i=1}^{|I|} \sum_{r=1}^{R} \tilde{d}_{ir} d_{ir} - \sum_{j=1}^{|J|} \sum_{r=1}^{R} \tilde{e}_{jr} e_{jr} - \sum_{k=1}^{|K|} \sum_{r=1}^{R} \tilde{f}_{kr} f_{kr}. \end{aligned} \tag{S5}$$

Considering that U, V, W are highly similar, so are D, E, and F. So we only take partial derivatives of $u_{r_3 i r_1}$ and d_{ir} of L.

$$\begin{gathered} \begin{cases} \dfrac{\partial L}{\partial u_{r_3 i r_1}} = \sum\limits_{y_{ijk} \in \Lambda(i)} \left(\left(y_{ijk} - \hat{y}_{ijk} \right) \left(- \left(\sum\limits_{r_2=1}^{R} v_{r_1 j r_2} w_{r_2 k r_3} \right) \right) + \lambda_1 u_{r_3 i r_1} \right) - \tilde{u}_{r_3 i r_1} = 0 \\ \dfrac{\partial L}{\partial d_{ir}} = \sum\limits_{y_{ijk} \in \Lambda(i)} \left(\left(y_{ijk} - \hat{y}_{ijk} \right) \left(-e_{jr} f_{kr} \right) + \lambda_2 d_{ir} \right) - \tilde{d}_{ir} = 0 \end{cases} \\ \Downarrow \\ \begin{cases} \tilde{u}_{r_3 i r_1} = \sum\limits_{y_{ijk} \in \Lambda(i)} \left(\left(y_{ijk} - \hat{y}_{ijk} \right) \left(- \left(\sum\limits_{r_2=1}^{R} v_{r_1 j r_2} w_{r_2 k r_3} \right) \right) + \lambda_1 u_{r_3 i r_1} \right) \\ \tilde{d}_{ir} = \sum\limits_{y_{ijk} \in \Lambda(i)} \left(\left(y_{ijk} - \hat{y}_{ijk} \right) \left(-e_{jr} f_{kr} \right) + \lambda_2 d_{ir} \right) \end{cases} \end{gathered} \tag{S6}$$

Then, considering the KKT conditions of (17), i.e., $u_{r_3 i r_1}\tilde{u}_{r_3 i r_1}=0,\forall u_{r_3 i r_1},\tilde{u}_{r_3 i r_1}$ and $d_{ir}\tilde{d}_{ir}=0,\forall d_{ir},\tilde{d}_{ir}$, we achieve that:

$$\begin{gathered}\begin{cases} u_{r_3 i r_1}\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(y_{ijk}-\hat{y}_{ijk}\right)\left(-\left(\sum\limits_{r_2=1}^{R}v_{r_1 j r_2}w_{r_2 k r_3}\right)\right)+\lambda_1 u_{r_3 i r_1}\right)=0,\\ d_{ir}\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(y_{ijk}-\hat{y}_{ijk}\right)\left(-e_{jr}f_{kr}\right)+\lambda_2 d_{ir}\right)=0.\end{cases}\\ \Downarrow\\ \begin{cases} u_{r_3 i r_1}\sum\limits_{y_{ijk}\in\Lambda(i)}\left(y_{ijk}\sum\limits_{r_2=1}^{R}v_{r_1 j r_2}w_{r_2 k r_3}\right)=u_{r_3 i r_1}\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\hat{y}_{ijk}\sum\limits_{r_2=1}^{R}v_{r_1 j r_2}w_{r_2 k r_3}+\lambda_1 u_{r_3 i r_1}\right),\\ d_{ir}\sum\limits_{y_{ijk}\in\Lambda(i)}\left(y_{ijk}e_{jr}f_{kr}\right)=d_{ir}\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\hat{y}_{ijk}e_{jr}f_{kr}+\lambda_2 d_{ir}\right).\end{cases}\end{gathered}\tag{S7}$$

From (S5) to (S7), it becomes evident that the learning scheme based on SLF-NMU is closely related to the KKT condition of its learning target.

B. Proof of Proposition 1

Using (37), we have *G*(*x, x) = Fu*$_{r3ir1}$. Next, our goal is to demonstrate $G(x, u^{(t)}_{r3ir1}) \geq Fu_{r3ir1}(x)$. To achieve this, we initially derive the quadratic approximation to $Fu_{r3ir1}(x)$ at $u^{(t)}_{r3ir1}$.

$$F_{u_{r_3 i r_1}}(x)=F_{u_{r_3 i r_1}}\left(u^{(t)}_{r_3 i r_1}\right)+F'_{u_{r_3 i r_1}}\left(u^{(t)}_{r_3 i r_1}\right)\left(x-u^{(t)}_{r_3 i r_1}\right)+\frac{1}{2}F''_{u_{r_3 i r_1}}\left(u^{(t)}_{r_3 i r_1}\right)\left(x-u^{(t)}_{r_3 i r_1}\right)^2.\tag{S8}$$

Combining (36) to (S8), it can be observed that $G(x,\ ^{u(t)}_{r3ir1})$ serves as an auxiliary function for Fu_{r3ir1} if the following inequality holds:

$$\frac{\left(\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)\hat{y}_{ijk}\right)+\left(\sum\limits_{y_{ijk}\in\Lambda(i)}(\lambda_1)\right)u^{(t)}_{r_3 i r_1}\right)}{u^{(t)}_{r_3 i r_1}}\geq\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)^2\right)+\left(\sum\limits_{y_{ijk}\in\Lambda(i)}(\lambda_1)\right).\tag{S9}$$

Because of **Y**’s nonnegativity, we have $y_{ijk}\geq 0$, and factors in **U**, **V**, **W**, D, E, F are greater than 0 with SLF-NMU. Therefore, (S9) can be expressed as:

$$\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)\hat{y}_{ijk}\right)\geq u^{(t)}_{r_3 i r_1}\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)^2\right).\tag{S10}$$

Next, we transform the left-hand term of (S10) as follows:

$$\begin{aligned}\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)\hat{y}_{ijk}\right)&=u^{(t)}_{r_3 i r_1}\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)^2\right)+\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)\left(\sum\limits_{r_1=1}^{R_1}\sum\limits_{r_2=1}^{R_2}\sum\limits_{r_3=1}^{R_3}u_{r_3 i r_1}v_{r_1 j r_2}w_{r_2 k r_3}-u^{(t)}_{r_3 i r_1}\right)\right)\\&\geq u^{(t)}_{r_3 i r_1}\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)^2\right).\end{aligned}\tag{S11}$$

Note that (S9) holds with (S11), thereby establishing $G(x,\ ^{u(t)}_{r3ir1})$ as an auxiliary function for Fu_{r3ir1}.

C. Proof of Theorem 1

According to (31), (35), and (37), we have the following inference:

$$\begin{aligned}&u^{(t+1)}_{r_3 i r_1}=\arg\min\nolimits_x G\left(x,u^{(t)}_{r_3 i r_1}\right)\\&\Rightarrow F'_{u_{r_3 i r_1}}\left(u^{(t)}_{r_3 i r_1}\right)+\frac{\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)\hat{y}_{ijk}+\lambda_1 u^{(t)}_{r_3 i r_1}\right)}{u^{(t)}_{r_3 i r_1}}\left(x-u^{(t)}_{r_3 i r_1}\right)=0\\&\Rightarrow u^{(t+1)}_{r_3 i r_1}\leftarrow u^{(t)}_{r_3 i r_1}\frac{\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)y_{ijk}\right)}{\sum\limits_{y_{ijk}\in\Lambda(i)}\left(\left(\sum\limits_{r_2=1}^{R_2}v_{r_1 j r_2}w_{r_2 k r_3}\right)\hat{y}_{ijk}+\lambda_1 u^{(t)}_{r_3 i r_1}\right)}.\end{aligned}\tag{S12}$$

Based on (S12), it is evident that F remains nonincreasing with respect to (17). Naturally, (S12) holds ∀j∈J, i∈I, k∈K. Therefore, ***Theorem* 1** stands.

D. Proof of Theorem 2

The sequence converges with (17) based on (40). Let $\mathbf{U}^{(*)}$ denote a stationary point of $\mathbf{U}$, that is,

$$0 \le u^{(*)}_{r_3ir_1} = \lim_{t\to\infty} u^{(t)}_{r_3ir_1} < +\infty, \forall i \in I, r_1, r_3 \in \{1,2,...,R\}. \tag{S13}$$

Therefore, if $\mathbf{U}^{(*)}$ is an equilibrium points, the following KKT conditions on $\mathbf{U}$ in (14) should be satisfied:

$$\begin{aligned}&(a)\ \left.\frac{\partial L}{\partial u_{r_3ir_1}}\right|_{u_{r_3ir_1}=u^{(*)}_{r_3ir_1}} = \sum_{y_{ijk}\in\Lambda(i)}\left(\lambda_1 u^{(*)}_{r_3ir_1} - \left(y_{ijk}-\hat{y}_{ijk}\right)\left(\sum_{r_2=1}^{R_2} v_{r_1jr_2} w_{r_2kr_3}\right)\right) - \phi^{(*)}_{r_3ir_1} = 0,\\ &(b)\ \phi^{(*)}_{r_3ir_1}\cdot u^{(*)}_{r_3ir_1} = 0, \qquad (c)\ u^{(*)}_{r_3ir_1} \ge 0, \qquad (d)\ \phi^{(*)}_{r_3ir_1} \ge 0.\end{aligned} \tag{S14}$$

Note that condition (a) must satisfy (16). So, we have

$$\phi^{(*)}_{r_3ir_1} = \sum_{y_{ijk}\in\Lambda(i)}\left(\lambda_1 u^{(*)}_{r_3ir_1} - \left(y_{ijk}-\hat{y}_{ijk}\right)\left(\sum_{r_2=1}^{R_2} v_{r_1jr_2} w_{r_2kr_3}\right)\right). \tag{S15}$$

So, the focus is directed towards Conditions (c) and (d). We initiate the process by formulating $\theta^{(t)}_{r3ir1}$ as

$$\theta^{(t)}_{r_3ir_1} = \frac{\sum_{y_{ijk}\in\Lambda(i)}\left(\left(\sum_{r_2=1}^{R_2} v_{r_1jr_2} w_{r_2kr_3}\right) y_{ijk}\right)}{\sum_{y_{ijk}\in\Lambda(i)}\left(\left(\sum_{r_2=1}^{R_2} v_{r_1jr_2} w_{r_2kr_3}\right)\hat{y}_{ijk}\right) + \left|\Lambda(i)\right|\lambda u^{(t)}_{r_3ir_1}}. \tag{S16}$$

Note (S16) is constrained by nonnegative v_{r1jr2}, w_{r2kr3} and y_{ijk} as,

$$0 \le \theta^{(*)}_{r_3ir_1} = \lim_{t\to+\infty}\theta^{(t)}_{r_3ir_1} = \frac{\sum_{y_{ijk}\in\Lambda(i)}\left(\left(\sum_{r_2=1}^{R_2} v_{r_1jr_2} w_{r_2kr_3}\right) y_{ijk}\right)}{\sum_{y_{ijk}\in\Lambda(i)}\left(\left(\sum_{r_2=1}^{R_2} v_{r_1jr_2} w_{r_2kr_3}\right)\hat{y}_{ijk}\right) + \left|\Lambda(i)\right|\lambda_1 u^{(*)}_{r_3ir_1}}. \tag{S17}$$

Thus, the updating rule for u_{r3ir1} can be reformulated as,

$$u^{(t+1)}_{r_3ir_1} = u^{(t)}_{r_3ir_1}\theta^{(t)}_{r_3ir_1}. \tag{S18}$$

By combining (39) and (S18), we have the following inference:

$$\lim_{t\to+\infty}\left|u^{(t+1)}_{r_3ir_1} - u^{(t)}_{r_3ir_1}\right| = 0 \Rightarrow u^{(*)}_{r_3ir_1}\theta^{(*)}_{r_3ir_1} - u^{(*)}_{r_3ir_1} = 0. \tag{S19}$$

Considering the update rule from (17), $u^{(*)}_{r3ir1}$ is equal or above to zero with a nonnegative initial hypothesis. Thus, we have:

1) When $u^{(*)}_{r_3ir_1} > 0$. Utilizing (S16) and (S19), we have

$$\begin{aligned}&u^{(*)}_{r_3ir_1}\theta^{(*)}_{r_3ir_1} - u^{(*)}_{r_3ir_1} = 0,\\ &u^{(*)}_{r_3ir_1} > 0 \Rightarrow \theta^{(*)}_{r_3ir_1} = 1 \Rightarrow \left|\Lambda(i)\right|\lambda_1 u^{(*)}_{r_3ir_1} + \sum_{y_{ijk}\in\Lambda(i)}\hat{y}_{ijk}\left(\sum_{r_2=1}^{R_2} v_{r_1jr_2} w_{r_2kr_3}\right) - \sum_{y_{ijk}\in\Lambda(i)} y_{ijk}\left(\sum_{r_2=1}^{R_2} v_{r_1jr_2} w_{r_2kr_3}\right) = 0.\end{aligned} \tag{S20}$$

By merging (S15) and (S20), we deduce Condition (b) in (S14).

$$\begin{gathered}\phi^{(*)}_{r_3ir_1} = \left|\Lambda(i)\right|\lambda_1 u^{(*)}_{r_3ir_1} + \sum_{y_{ijk}\in\Lambda(i)}\hat{y}_{ijk}\left(\sum_{r_2=1}^{R_2} v_{r_1jr_2} w_{r_2kr_3}\right) - \sum_{y_{ijk}\in\Lambda(i)} y_{ijk}\left(\sum_{r_2=1}^{R_2} v_{r_1jr_2} w_{r_2kr_3}\right) = 0\\ \Downarrow\\ \phi^{(*)}_{r_3ir_1}\cdot u^{(*)}_{r_3ir_1} = 0.\end{gathered} \tag{S21}$$

Simultaneously, if $\Phi^{(*)}_{r3ir1}$ and $u^{(*)}_{r3ir1} > 0$, Conditions (c) and (d) are inherently satisfied. Then, (S14) is fulfilled when $u^{(*)}_{r3ir1} > 0$.

2) When $u_{r3ir1}^{(*)} = 0$, it's noteworthy that conditions (b) and (c) in (S14) are automatically satisfied. Therefore, our focus is on validating Condition (d). To do this, we redefine $u_{r3ir1}^{(*)}$ as:

$$u_{r_3 i r_1}^{(*)} = u_{r_3 i r_1}^{(0)} \lim_{t \to +\infty} \prod_{r=1}^{t} \theta_{r_3 i r_1}^{(r)}. \tag{S22}$$

Drawing on (S22), we can make the following inferences:

$$\begin{aligned}
& u_{r_3 i r_1}^{(0)} > 0,\ u_{r_3 i r_1}^{(0)} \lim_{t \to +\infty} \prod_{r=1}^{t} \theta_{r_3 i r_1}^{(r)} = u_{r_3 i r_1}^{(*)} = 0 \\
& \Rightarrow \lim_{t \to +\infty} \prod_{r=1}^{t} \theta_{r_3 i r_1}^{(r)} = 0 \\
& \Rightarrow \lim_{t \to +\infty} \theta_{r_3 i r_1}^{(t)} = \theta_{r_3 i r_1}^{(*)} = \frac{\sum\limits_{y_{ijk} \in \Lambda(i)} \left(\left(\sum\limits_{r_2=1}^{R_2} v_{r_1 j r_2} w_{r_2 k r_3} \right) y_{ijk} \right)}{\sum\limits_{y_{ijk} \in \Lambda(i)} \left(\left(\sum\limits_{r_2=1}^{R_2} v_{r_1 j r_2} w_{r_2 k r_3} \right) \hat{y}_{ijk} \right) + \left(\sum\limits_{y_{ijk} \in \Lambda(i)} (\lambda_1) \right) u_{r_3 i r_1}^{(*)}} \le 1, \\
& \Rightarrow \phi_{r_3 i r_1}^{(*)} = u_{r_3 i r_1}^{(*)} \sum_{y_{ijk} \in \Lambda(i)} \lambda_1 + \sum_{y_{ijk} \in \Lambda(i)} \hat{y}_{ijk} \left(\sum_{r_2=1}^{R_2} v_{r_1 j r_2} w_{r_2 k r_3} \right) - \sum_{y_{ijk} \in \Lambda(i)} y_{ijk} \left(\sum_{r_2=1}^{R_2} v_{r_1 j r_2} w_{r_2 k r_3} \right) \ge 0.
\end{aligned} \tag{S23}$$

As a result, (S14) is fulfilled when $u_{r3ir1}^{(*)} > 0$. Similarly, we can demonstrate that the sequences $\{v_{r1jr2}^{(t)}\}$, $\{w_{r2kr3}^{(t)}\}$, $\{d_{ir}^{(t)}\}$, $\{e_{jr}^{(t)}\}$, $\{f_{kr}^{(t)}\}$ also converges to a stable equilibrium point of (14). To sum up, Theorem 2 holds.

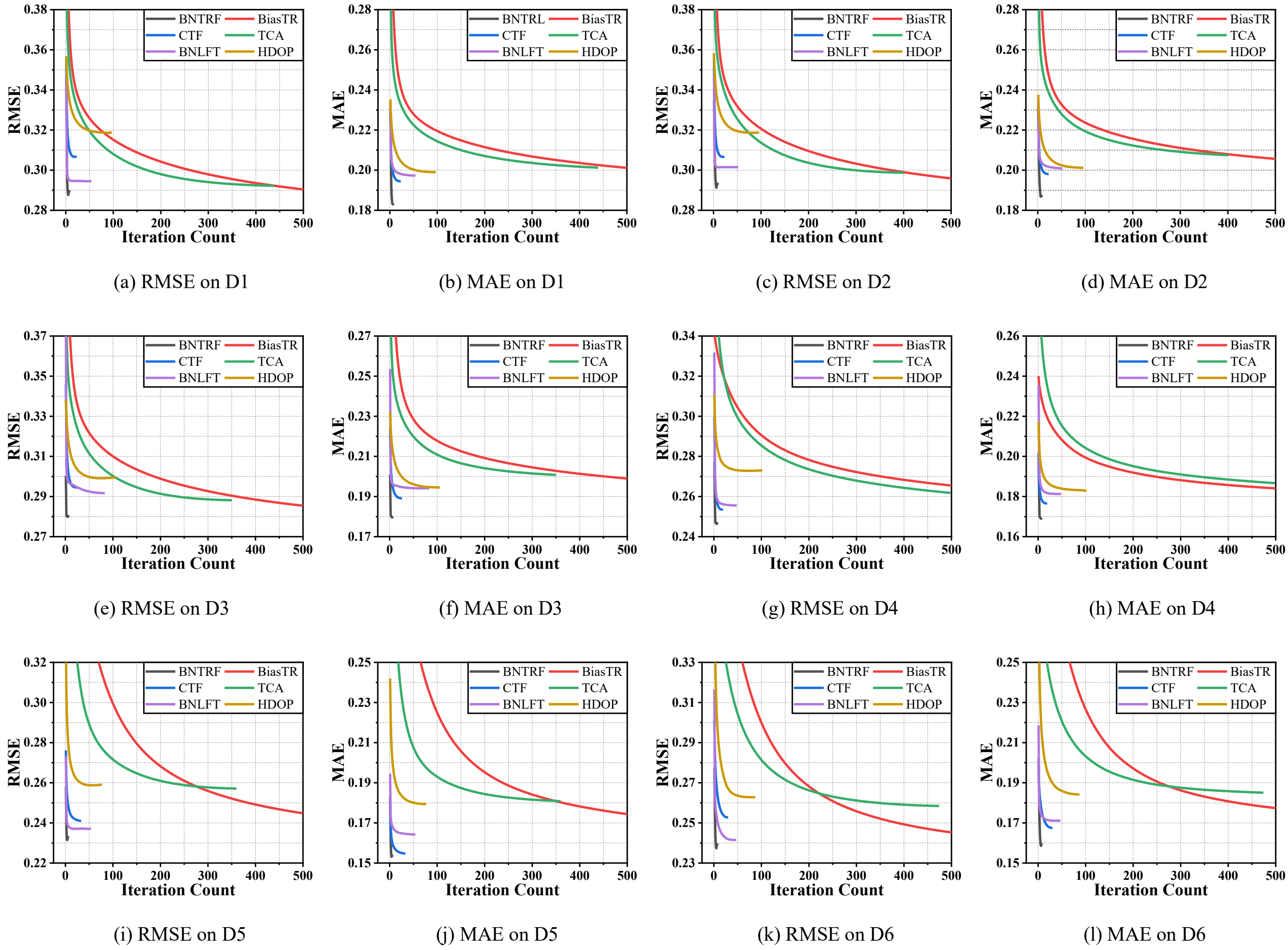
IV. Supplementary Tables and Figures

Here are some supplementary tables and figures. Table S1 records the hyper-parameters of all test models. Fig. S3 records the iterations of M1-M6 on D1-D6.

TABLE S1

Hyper-parameter Settings

Dataset	Hyper-parameter Settings			
D1	M1: Adaptive	M2: η_2=0.0001, λ_2=0.001	M3: λ_3=10^{-8}, γ=800	M4: η_4=0.002, λ_4=0.01
	M5: λ_a=0.4, λ_b=0.1	M6: η_6=0.004, λ_6=0.01	M7: η_7=0.1, λ_7=0.001	M8: η_8=0.02, λ_8=0.1
D2	M1: Adaptive	M2: η_2=0.0001, λ_2=0.001	M3: λ_3=10^{-8}, γ=800	M4: η_4=0.002, λ_4=0.01
	M5: λ_a=0.4, λ_b=0.1	M6: η_6=0.004, λ_6=0.01	M7: η_7=0.1, λ_7=0.001	M8: η_8=0.02, λ_8=0.1
D3	M1: Adaptive	M2: η_2=0.0001, λ_2=0.001	M3: λ_3=10^{-8}, γ=800	M4: η_4=0.0005, λ_4=0.01
	M5: λ_a=0.4, λ_b=0.1	M6: η_6=0.004, λ_6=0.01	M7: η_7=0.2, λ_7=0.001	M8: η_8=0.05, λ_8=0.1
D4	M1: Adaptive	M2: η_2=0.00002, λ_2=0.001	M3: λ_3=10^{-9}, γ=800	M4: η_4=0.0005, λ_4=0.01
	M5: λ_a=0.4, λ_b=0.1	M6: η_6=0.004, λ_6=0.01	M7: η_7=0.2, λ_7=0.001	M8: η_8=0.05, λ_8=0.1
D5	M1: Adaptive	M2: η_2=0.0001, λ_2=0.001	M3: λ_3=10^{-8}, γ=800	M4: η_4=0.002, λ_4=0.001
	M5: λ_a=0.4, λ_b=0.4	M6: η_6=0.004, λ_6=0.001	M7: η_7=0.2, λ_7=0.001	M8: η_8=0.1, λ_8=0.01
D6	M1: Adaptive	M2: η_2=0.0002, λ_2=0.001	M3: λ_3=10^{-8}, γ=800	M4: η_4=0.002, λ_4=0.001
	M5: λ_a=0.4, λ_b=0.4	M6: η_6=0.004, λ_6=0.001	M7: η_7=0.2, λ_7=0.001	M8: η_8=0.1, λ_8=0.01



(a) RMSE on D1

(b) MAE on D1

(c) RMSE on D2

(d) MAE on D2

(e) RMSE on D3

(f) MAE on D3

(g) RMSE on D4

(h) MAE on D4

(i) RMSE on D5

(j) MAE on D5

(k) RMSE on D6

(l) MAE on D6

Fig. S3. Training curves in RMSE and MAE of M1-6 on D1-6.