Supplementary Materials for Non-negative low-rank approximations for multi-dimensional arrays on statistical manifold

Kazu Ghalamkari[1,2] **Mahito Sugiyama**[1,2]
[1]National Institute of Informatics
[2]The Graduate University for Advanced Studies, SOKENDAI
{gkazu,mahito}@nii.ac.jp

A Theoretical remarks

A.1 Connection between rank-1 approximation and fiber balancing

Given d tensors $\mathcal{C}^{\setminus k} \in \mathbb{R}_{\geq 0}^{I_1 \times \cdots \times I_{k-1} \times I_{k+1} \times \cdots \times I_d}$ for $k \in [d]$, the task of $\mathcal{C}$-fiber balancing is to rescale a tensor so that the sum of each fiber satisfies

$$\sum_{i_k=1}^{I_k} \mathcal{P}_{i_1,\dots,i_d} = \mathcal{C}^{\setminus k}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_d} \tag{A1}$$

for $i_k \in [I_k]$. Recall the definition of η-parameters, the above condition for $\mathcal{C}$-fiber balancing can be expressed as

$$\eta_{i_1,\dots,i_{k-1},1,i_{k+1},\dots,i_d} = \sum_{i'_1=i_1}^{I_1} \cdots \sum_{i'_{k-1}=i_{k-1}}^{I_{k-1}} \sum_{i'_{k+1}=i_{k+1}}^{I_{k+1}} \cdots \sum_{i'_d=i_d}^{I_d} \mathcal{C}^{\setminus k}_{i'_1,\dots,i'_{k-1},i'_{k+1},\dots,i'_d}. \tag{A2}$$

Let us define $\mathcal{C}$-fiber balancing space $\mathcal{Q}_{\mathcal{C}}$ as the set of $\mathcal{C}$-fiber balanced tensor, yielding $\mathcal{Q}_{\mathcal{C}} = \{p_\eta \mid \eta$ satisfies the condition (A2) $\}$. Note that fiber balancing is equivalent to slice balancing if $d = 2$, which in this case is called matrix balancing.

If we impose the $\mathcal{C}$-fiber balancing condition and rank-1 condition on a tensor $\mathcal{P}$ simultaneously, then any parameter that contains only one 1 in its index has been imposed both the rank-1 and $\mathcal{C}$-balancing conditions. We can prove the following theorem by examining when these two conditions stand together.

Theorem 3. *The intersection $\mathcal{Q}_{\mathcal{C}} \cap \mathcal{B}_1$ exists and is a singleton if and only if*

$$\mathcal{C}^{\setminus k}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_d} = \prod_{m \in [d] \setminus k} \left(\prod_{l \in [d] \setminus \{k,m\}} \sum_{i_l=1}^{I_l} \mathcal{C}^{\setminus k}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_d} \right).$$

B Additional information for numerical experiments

B.1 Implementation details

Because all methods were implemented in `Julia` 1.6 with `TensorToolbox` library (Periša and Arslan, 2019), runtime comparison is fair. We implemented lraSNTD referring to the original papers (Zhou et al., 2012). We used the `TensorLy` implementation (Kossaifi et al., 2019) for NTDs. Experiments were conducted on CentOS 6.10 with a single core of 2.2 GHz Intel Xeon CPU E7-8880

Table 1: Real dataset details for A1GM.

DataSet	# zeros	# negatives	URL
IndianPop	0	19	https://onl.bz/bcu3qCR
Autompg	0	0	https://www.kaggle.com/uciml/autompg-dataset
DailySunSpot	16994	3247	https://www.kaggle.com/abhinand05/daily-sun-spot-data-1818-to-2019
CaliforniaHousing	0	20640	https://www.kaggle.com/harrywang/housing?select=housing.csv
MTSLibrary	200932	0	https://www.kaggle.com/sharthz23/mts-library?select=interactions.csv
BigMartSaleForecas	526	0	https://www.kaggle.com/arashnic/big-mart-sale-forecast?select=train.csv
BoardGameGeekData	520624	21	https://www.kaggle.com/mandshaw/games-0918
CreditCardApproval	797	0	https://www.kaggle.com/redwuie/credit-card-approval?select=train.csv
HumanResourceAnaly	27510	0	https://www.kaggle.com/cezarschroeder/human-resource-analytics-dataset
concretemiss	1390	0	https://www.kaggle.com/datasets/izemdemirci/concrete-missing
heartdisease	1149	0	https://archive.ics.uci.edu/ml/datasets/Heart+Disease
lungcancer	107	0	https://archive.ics.uci.edu/ml/datasets/Lung+Cancer
PerthHousePrice	0	33656	https://www.kaggle.com/syuzai/perth-house-prices
SleepData	0	0	https://www.kaggle.com/mathurinache/sleep-dataset
HCVData	0	0	https://archive.ics.uci.edu/ml/datasets/HCV+data
arrhythmia	67256	14250	https://archive.ics.uci.edu/ml/datasets/Arrhythmia
Bostonhousing	812	0	https://www.kaggle.com/altavish/boston-housing-dataset
LifeExpectancyData	3385	0	https://www.kaggle.com/kumarajarshi/life-expectancy-who
HCCSurvivalDataSet	2416	0	https://archive.ics.uci.edu/ml/datasets/HCC+Survival
wiki4HE	1801	0	https://archive.ics.uci.edu/ml/datasets/wiki4HE

v4 and 3TB of memory. We used the default values of hyper-parameters of `tensorly` for NTD. We used default values of the hyper-parameters of `sklearn` (Pedregosa et al., 2011) for the NMF module in lraSNTD.

B.2 Dataset details

For experiments of LTR, we describe the details of each dataset below. 4DLFD is a $(9, 9, 512, 512, 3)$ tensor, which is produced by 4D Light Field Dataset described in (Honauer et al., 2016). Its license is Creative Commons Attribution-Non-Commercial-ShareAlike 4.0 International License. We used dino images and their depth and disparity map in training scenes and concatenated them to produce a tensor. AttFace is a $(92, 112, 400)$ tensor that is produced by the entire data in The Database of Faces (AT&T) (Samaria and Harter, 1994), which includes 400 grey-scale face photos. The size of each image is $(92, 112)$. AttFace is public on Kaggle, but the license is not specified. For experiments of A1GM, we provide the list of the source of the real datasets in Table 1.

B.3 Additional experimental results of LTR

The cost function of LTR is the KL divergence from the input tensor to the low-rank tensor. Our experimental results on real datasets in Figure 1 show that LTR also has competitive accuracy of approximation in terms of the KL divergence. We also provide the experimental results for real datasets in table format in Tables 2 and 3. In these tables, we evaluate the relative KL error as

$$D(\mathbf{X}, \tilde{\mathbf{X}}))/D(\mathbf{X}, \mathrm{LTR}(\mathbf{X})),$$

where $\mathrm{LTR}(\mathbf{X})$ and $\tilde{\mathbf{X}}$ are the low-rank reconstructed tensors by LTR and comparing methods, such as NTD_KL, NTD_LS, and lraSNTD. We also evaluate the relative LS error and relative running time. If these values are larger than 1, it means the proposed method LTR has better performance than the compared method.

B.4 Algorithm formats for implementation of LTR and A1GM

We provide the algorithms of proposed LTR and A1GM as algorithm formats in Algorithm 1 – 2.

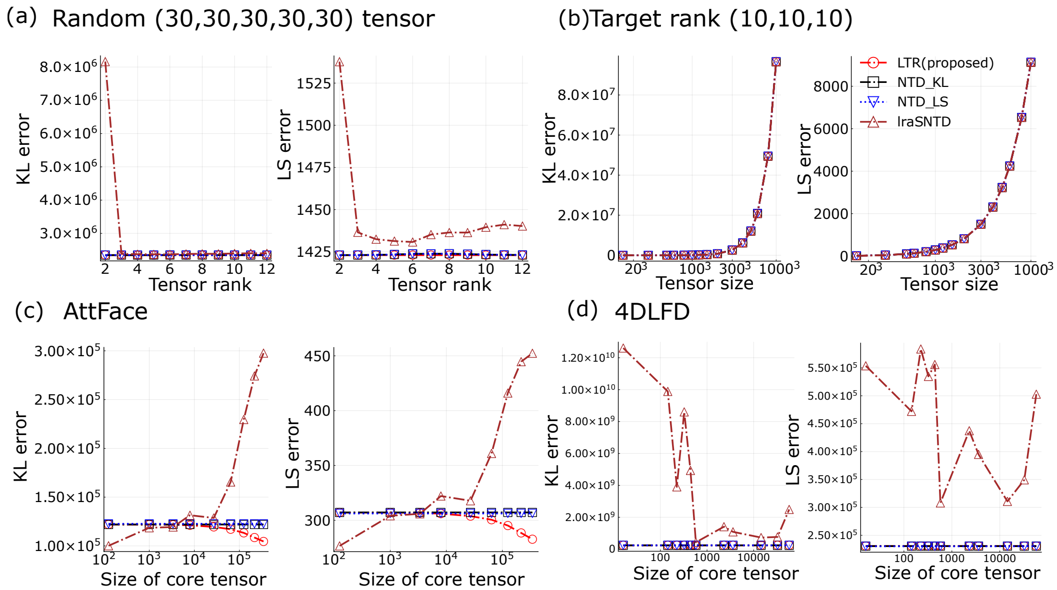

Figure 1: Experimental results for synthetic (a, b) and real-world (c, d) datasets. The left-hand panels are KL reconstruction error and the right-hand panels are LS reconstruction error. (a) The horizontal axis is r for target tensor rank (r, r, r, r, r). (b) The horizontal axis is n^3 for input (n, n, n) tensor.

C Tensor operations

C.1 Kronecker product for vectors

Given d vectors $\boldsymbol{a}^{(1)} \in \mathbb{R}^{I_1}$, $\boldsymbol{a}^{(2)} \in \mathbb{R}^{I_2}$, $\dots$, and $\boldsymbol{a}^{(d)} \in \mathbb{R}^{I_d}$, the Kronecker product $\mathcal{P}$ of these d vectors, written as

$$\mathcal{P} = \boldsymbol{a}^{(1)} \otimes \boldsymbol{a}^{(2)} \otimes \cdots \otimes \boldsymbol{a}^{(d)},$$

is a tensor in $\mathbb{R}^{I_1 \times \cdots \times I_d}$, where each element of $\mathcal{P}$ is given as

$$\mathcal{P}_{i_1,\dots,i_d} = a^{(1)}_{i_1} a^{(2)}_{i_2} \dots a^{(d)}_{i_d}.$$

Table 2: Experimental results of LTR on AttFace dataset.

	Relative running time			Relative LS error			Relative KL error		
Rank	NTD_KL	NTD_LS	lraSNTD	NTD_KL	NTD_LS	lraSNTD	NTD_KL	NTD_LS	lraSNTD
(1,1,1)	4.7533	3.8874	0.5510	1.0000	0.9971	0.9971	1.0000	1.0056	1.0059
(3,3,3)	33.4752	26.3030	2.7253	0.9755	0.9743	0.9280	0.9440	0.9549	0.8685
(5,5,5)	38.2118	29.9251	4.0683	0.9777	0.9784	0.8757	0.9508	0.9633	0.7786
(10,10,10)	28.6684	22.7941	4.4291	0.9826	0.9834	0.9915	0.9621	0.9736	0.9795
(15,15,15)	34.6129	27.5794	6.2854	0.9865	0.9867	1.0090	0.9714	0.9819	1.0006
(20,20,20)	32.0121	25.9077	6.6242	0.9916	0.9912	1.0548	0.9817	0.9913	1.0890
(30,30,30)	43.9135	36.6030	13.2826	1.0007	0.9996	1.0710	0.9996	1.0082	1.1222
(40,40,40)	45.2404	38.8275	15.1266	1.0164	1.0149	1.2127	1.0296	1.0378	1.4427
(50,50,50)	62.7553	52.3623	25.9060	1.0385	1.0367	1.3627	1.0707	1.0786	1.8738
(60,60,60)	74.8940	60.5661	32.1785	1.0617	1.0595	1.5309	1.1150	1.1228	2.5023
(70,70,70)	41.5551	35.3710	21.7593	1.0888	1.0863	1.6287	1.1676	1.1755	2.9325
(80,80,80)	58.4794	47.3544	30.3658	1.1186	1.1160	1.5663	1.2278	1.2358	2.6274

Table 3: Experimental results of LTR on 4DLFD dataset.

	Relative running time			Relative LS error			Relative KL error		
Rank	NTD_KL	NTD_LS	lraSNTD	NTD_KL	NTD_LS	lraSNTD	NTD_KL	NTD_LS	lraSNTD
(1,1,1,1,1)	13.3018	13.3018	1.0019	1.0000	0.9985	0.9985	1.0000	1.0027	1.0027
(2,2,2,2,1)	22.9786	22.9786	2.8540	0.9870	0.9842	3.0129	0.9748	0.9766	14.3603
(3,3,4,4,1)	22.6299	22.6299	8.2270	0.9319	0.9210	1.1570	0.8765	0.8682	1.3131
(3,3,5,5,1)	22.9552	22.9552	4.9963	0.9303	0.9201	3.2399	0.8745	0.8672	28.4056
(3,3,6,6,1)	22.4911	22.4911	7.0107	0.9279	0.9183	2.6687	0.8704	0.8642	19.1505
(3,3,7,7,1)	23.7567	23.7567	6.2986	0.9294	0.9202	1.0066	0.8730	0.8674	1.0927
(3,3,8,8,1)	23.7998	23.7998	9.0809	0.9284	0.9198	1.3867	0.8710	0.8663	1.9389
(3,3,16,16,1)	23.6797	23.6797	7.1864	0.9256	0.9211	1.1852	0.8640	0.8659	1.5274
(3,3,20,20,1)	26.1782	26.1782	6.4697	0.9304	0.9265	1.4345	0.8719	0.8746	2.0486
(3,3,40,40,1)	26.4613	26.4613	4.6566	0.9426	0.9414	1.5224	0.8933	0.8996	2.2700
(3,3,60,60,1)	27.9254	27.9254	10.5136	0.9532	0.9528	1.3116	0.9128	0.9197	1.7012
(3,3,80,80,1)	26.4166	26.4166	8.1281	0.9619	0.9615	1.2247	0.9288	0.9354	1.8635

Algorithm 1: LTR

```
input  : Tensor $\mathcal{P}$, target Tucker-rank $\boldsymbol{r} = (r_1, \ldots, r_d)$
output : low-Tucker-rank tensor $\mathcal{T}$
LTR($\mathcal{P}$,$\mathbf{r}$)
    $(I_1, \ldots, I_d) \leftarrow$ the size of the input tensor $\mathcal{P}$
    foreach $k = 1, \ldots, d$ do
        Construct $\{c_1, \ldots, c_{r_k}\} \subseteq [I_k]$ by random sampling from $[I_k]$ without replacement,
          where we always assume that $c_1 = 1$ and $c_i < c_{i+1}$.
        foreach $l = 1, \ldots, r_k$ do
            if $c_l \neq c_{l+1} - 1$ then
                Replace the subtensor $\mathcal{P}_{c_l^{(k)}:c_{l+1}^{(k)}-1}$ of $\mathcal{P}$ by its rank-1 approximation as
                $\mathcal{P}_{c_l^{(k)}:c_{l+1}^{(k)}-1} \leftarrow \textsc{BestRank1}(\mathcal{P}_{c_l^{(k)}:c_{l+1}^{(k)}-1})$
    $\mathcal{T} \leftarrow \mathcal{P}$
    return $\mathcal{T}$
BestRank1($\mathcal{P}$)
    $(I_1, \ldots, I_d) \leftarrow$ the size of the input tensor $\mathcal{P}$
    foreach $k = 1, \ldots, d$ do
        foreach $i_k = 1, \ldots, I_k$ do
            $s_{i_k}^{(k)} \leftarrow \sum_{i_1=1}^{I_1} \cdots \sum_{i_{k-1}=1}^{I_{k-1}} \sum_{i_{k+1}=1}^{I_{k+1}} \cdots \sum_{i_d=1}^{I_d} \mathcal{P}_{i_1,\ldots i_{k-1},i_k,i_{k+1},\ldots,i_d}$
    $\lambda \leftarrow$ sum of all elements of $\mathcal{P}$
    $\overline{\mathcal{P}} \leftarrow \lambda^{1-d} \boldsymbol{s}^{(1)} \otimes \boldsymbol{s}^{(2)} \otimes \cdots \otimes \boldsymbol{s}^{(d)}$
    return $\overline{\mathcal{P}}$
```

Algorithm 2: A1GM

```
input  : A binary weight matrix $\mathbf{\Phi} \in \{0,1\}^{I \times J}$, a matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{I \times J}$
output : Dominant factors $\boldsymbol{c} \in \mathbb{R}_{\geq 0}^{I}$ and $\boldsymbol{d} \in \mathbb{R}_{\geq 0}^{J}$
A1GM($\mathbf{\Phi}$, $\mathbf{X}$)
    $S^{(1)} \leftarrow \varnothing$; $S^{(2)} \leftarrow \varnothing$
    for $(i,j) \in [I] \times [J]$ do
        if $\mathbf{\Phi}_{ij} = 0$ then
            $S^{(1)} \leftarrow S^{(1)} \cup \{i\}$; $S^{(2)} \leftarrow S^{(2)} \cup \{j\}$
    $B^{(1)} \leftarrow \{I - \mid S^{(1)} \mid +1, I - \mid S^{(1)} \mid +2, \ldots, I\}$
    $B^{(2)} \leftarrow \{J - \mid S^{(2)} \mid +1, J - \mid S^{(2)} \mid +2, \ldots, J\}$
    $\mathrm{perm1} \leftarrow (1, \ldots, I)$
    for $k \in \{1, 2, \ldots, \mid S^{(1)} \cap B^{(1)c} \mid\}$ do
        $i \leftarrow k$th smallest element of $S^{(1)} \cap B^{(1)c}$
        $j \leftarrow k$th smallest element of $S^{(1)c} \cap B^{(1)}$
        swap(perm1[$i$],perm1[$j$])
    $\mathrm{perm2} \leftarrow (1, \ldots, J)$
    for $k \in \{1, 2, \ldots, \mid S^{(2)} \cap B^{(2)c} \mid\}$ do
        $i \leftarrow k$th smallest element of $S^{(2)} \cap B^{(2)c}$
        $j \leftarrow k$th smallest element of $S^{(2)c} \cap B^{(2)}$
        swap(perm2[$i$],perm2[$j$])
    $\mathbf{X} \leftarrow \mathbf{X}[\mathrm{perm1}, \mathrm{perm2}]$
    $\boldsymbol{w}, \boldsymbol{h}, \boldsymbol{a}, \boldsymbol{b} \leftarrow$ the best rank-1 NMMF of $\mathbf{X}$
    $\boldsymbol{c} \leftarrow$ concate $\boldsymbol{w}$ and $\boldsymbol{a}$; $\boldsymbol{d} \leftarrow$ concate $\boldsymbol{h}$ and $\boldsymbol{b}$
    $\boldsymbol{c} \leftarrow \boldsymbol{c}[\mathrm{perm1}]$; $\boldsymbol{d} \leftarrow \boldsymbol{d}[\mathrm{perm2}]$
    return $\boldsymbol{c}, \boldsymbol{d}$
```

D Proofs

Proposition 1 (simultaneous rank-1 θ-condition)**.** *A tuple* $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{U})$ *is simultaneously rank-1 decomposable if and only if all of its many-body θ-parameters are* 0*.*
Proof. To make the following discussion clear, we define

$$\theta^{\mathbf{Y}}_{nj} = \theta_{nj}, \;\; \theta^{\mathbf{X}}_{ij} = \theta_{N+i,j}, \;\; \theta^{\mathbf{Z}}_{im} = \theta_{N+i,J+m}, \;\; \theta^{\mathbf{U}}_{lm} = \theta_{N+I+l,J+m}.$$

for all $i \in [I], j \in [J], n \in [N], m \in [M]$ and $k \in [K]$. First, we show a tuple $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{U})$ is simultaneously rank-1 decomposable $\Rightarrow$ all of its many-body natural parameters are 0. If a tuple $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{U})$ is simultaneously rank-1 decomposable, we can immediately confirm that $\mathrm{Rank}\,(\mathbf{X}) = \mathrm{Rank}\,(\mathbf{Y}) = \mathrm{Rank}\,(\mathbf{Z}) = \mathrm{Rank}\,(\mathbf{U}) = 1$. Then, from the Proposition 3 for $d = 2$,

$$\begin{aligned}
\theta^{\mathbf{Y}}_{nj} &= 0 \quad \text{if } n \neq 1 \text{ and } j \neq 1,\\
\theta^{\mathbf{X}}_{ij} &= 0 \quad \text{if } i \neq 1 \text{ and } j \neq 1,\\
\theta^{\mathbf{Z}}_{im} &= 0 \quad \text{if } i \neq 1 \text{ and } m \neq 1,\\
\theta^{\mathbf{U}}_{lm} &= 0 \quad \text{if } l \neq 1 \text{ and } m \neq 1.
\end{aligned}$$

Then, from the definition of the model in Equation (5), it holds that

$$\begin{aligned}
\mathbf{Y}_{nj} &= e^{\theta^{\mathbf{Y}}_{11}} \exp\left(\sum_{n'=2}^{n} \theta^{\mathbf{Y}}_{n'1}\right) \exp\left(\sum_{j'=2}^{j} \theta^{\mathbf{Y}}_{1j'}\right)\\
\mathbf{X}_{ij} &= e^{\theta^{\mathbf{X}}_{11}+\theta^{\mathbf{Y}}_{11}} \exp\left(\sum_{n'=2}^{N} \theta^{\mathbf{Y}}_{n'1} + \sum_{i'=2}^{i} \theta^{\mathbf{X}}_{i'1}\right) \exp\left(\sum_{j'=2}^{j} \theta^{\mathbf{Y}}_{1j'} + \sum_{j'=2}^{j} \theta^{\mathbf{X}}_{1j'}\right)\\
\mathbf{Z}_{im} &= e^{\theta^{\mathbf{X}}_{11}+\theta^{\mathbf{Y}}_{11}+\theta^{\mathbf{Z}}_{11}} \exp\left(\sum_{n'=2}^{N} \theta^{\mathbf{Y}}_{n'1} + \sum_{i'=2}^{i} \theta^{\mathbf{X}}_{i'1} + \sum_{i'=2}^{i} \theta^{\mathbf{Z}}_{i'1}\right)\\
&\quad \times \exp\left(\sum_{j'=2}^{J} \theta^{\mathbf{Y}}_{1j'} + \sum_{j'=2}^{J} \theta^{\mathbf{X}}_{1j'} + \sum_{m'=2}^{m} \theta^{\mathbf{Z}}_{1m'}\right)\\
\mathbf{U}_{lm} &= e^{\theta^{\mathbf{X}}_{11}+\theta^{\mathbf{Y}}_{11}+\theta^{\mathbf{Z}}_{11}+\theta^{\mathbf{U}}_{11}} \exp\left(\sum_{n'=2}^{N} \theta^{\mathbf{Y}}_{n'1} + \sum_{i'=2}^{I} \theta^{\mathbf{X}}_{i'1} + \sum_{i'=2}^{I} \theta^{\mathbf{Z}}_{i'1} + \sum_{l'=2}^{l} \theta^{\mathbf{U}}_{l'1}\right)\\
&\quad \times \exp\left(\sum_{j'=2}^{J} \theta^{\mathbf{Y}}_{1j'} + \sum_{j'=2}^{J} \theta^{\mathbf{X}}_{1j'} + \sum_{m'=2}^{m} \theta^{\mathbf{Z}}_{1m'} + \sum_{m'=2}^{m} \theta^{\mathbf{U}}_{1m'}\right).
\end{aligned}$$

If it does not hold that

$$\begin{aligned}
\theta^{\mathbf{X}}_{1j} &= 0 \quad \text{if } j \neq 1,\\
\theta^{\mathbf{Z}}_{i1} &= 0 \quad \text{if } i \neq 1,\\
\theta^{\mathbf{U}}_{1m} &= 0 \quad \text{if } m \neq 1,
\end{aligned}$$

matrices $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{U})$ have never shared factors. Then, the above condition holds if the matrices $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{U})$ is simultaneously rank-1 decomposable.

Next, we show that all of many-body natural parameters are $0 \Rightarrow$ the corresponding tuple $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{U})$ is simultaneously rank-1 decomposable. We put all many-body θ-parameters as 0 in Equation (5) and obtain

$$
\begin{aligned}
\mathbf{Y}_{nj} &= e^{\theta_{11}^{\mathbf{Y}}} \exp\left(\sum_{n'=2}^{n} \theta_{n'1}^{\mathbf{Y}}\right) \exp\left(\sum_{j'=2}^{j} \theta_{1j'}^{\mathbf{Y}}\right), \\
\mathbf{X}_{ij} &= e^{\theta_{11}^{\mathbf{X}}+\theta_{11}^{\mathbf{Y}}} \exp\left(\sum_{n'=2}^{N} \theta_{n'1}^{\mathbf{Y}} + \sum_{i'=2}^{i} \theta_{i'1}^{\mathbf{X}}\right) \exp\left(\sum_{j'=2}^{j} \theta_{1j'}^{\mathbf{Y}}\right), \\
\mathbf{Z}_{im} &= e^{\theta_{11}^{\mathbf{X}}+\theta_{11}^{\mathbf{Y}}+\theta_{11}^{\mathbf{Z}}} \exp\left(\sum_{n'=2}^{N} \theta_{n'1}^{\mathbf{Y}} + \sum_{i'=2}^{i} \theta_{i'1}^{\mathbf{X}}\right) \exp\left(\sum_{j'=2}^{J} \theta_{1j'}^{\mathbf{Y}} + \sum_{m'=2}^{m} \theta_{1m'}^{\mathbf{Z}}\right), \\
\mathbf{U}_{lm} &= e^{\theta_{11}^{\mathbf{X}}+\theta_{11}^{\mathbf{Y}}+\theta_{11}^{\mathbf{Z}}+\theta_{11}^{\mathbf{U}}} \exp\left(\sum_{n'=2}^{N} \theta_{n'1}^{\mathbf{Y}} + \sum_{i'=2}^{I} \theta_{i'1}^{\mathbf{X}} + \sum_{l'=2}^{l} \theta_{l'1}^{\mathbf{U}}\right) \exp\left(\sum_{j'=2}^{J} \theta_{1j'}^{\mathbf{Y}} + \sum_{m'=2}^{m} \theta_{1m'}^{\mathbf{Z}}\right).
\end{aligned}
$$

Then, we can define the following shared factors on the tuple $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{U})$,

$$
\begin{cases}
\boldsymbol{a}_n = \exp\left(\sum_{n'=2}^{n} \theta_{n'1}^{\mathbf{Y}}\right), \\
\boldsymbol{h}_j = e^{\theta_{11}^{\mathbf{Y}}} \exp\left(\sum_{j'=2}^{j} \theta_{1j'}^{\mathbf{Y}}\right), \\
\boldsymbol{w}_i = e^{\theta_{11}^{\mathbf{X}}} \exp\left(\sum_{n'=2}^{N} \theta_{n'1}^{\mathbf{Y}} + \sum_{i'=2}^{i} \theta_{i'1}^{\mathbf{X}}\right) \\
\boldsymbol{b}_m = e^{\theta_{11}^{\mathbf{Y}}+\theta_{11}^{\mathbf{Z}}} \exp\left(\sum_{j'=2}^{J} \theta_{1j'}^{\mathbf{Y}} + \sum_{m'=2}^{m} \theta_{1m'}^{\mathbf{Z}}\right), \\
\boldsymbol{c}_l = e^{\theta_{11}^{\mathbf{X}}+\theta_{11}^{\mathbf{U}}} \exp\left(\sum_{n'=2}^{N} \theta_{n'1}^{\mathbf{Y}} + \sum_{i'=2}^{I} \theta_{i'1}^{\mathbf{X}} + \sum_{l'=2}^{l} \theta_{l'1}^{\mathbf{U}}\right),
\end{cases}
$$

which satisify $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{U}) = (\boldsymbol{w} \otimes \boldsymbol{h}, \boldsymbol{a} \otimes \boldsymbol{h}, \boldsymbol{w} \otimes \boldsymbol{b}, \boldsymbol{c} \otimes \boldsymbol{b})$. Thus, the proposition was proved. □

Proposition 2 (Homogeneity of rank-1 missing NMF)**.** *Let* $\mathrm{NMF}_1(\mathbf{\Phi}, \mathbf{X})$ *be the best rank-*1 *matrix* $\boldsymbol{w} \otimes \boldsymbol{h}$*, which minimizes the cost function in Equation* (7)*. For any permutation matrices* $\mathbf{G}$ *and* $\mathbf{H}$*, it holds that*

$$
\mathrm{NMF}_1\left(\mathbf{G\Phi H}, \mathbf{GKH}\right) = \mathbf{G}^{\mathrm{T}} \mathrm{NMF}_1\left(\mathbf{\Phi}, \mathbf{K}\right) \mathbf{H}^{\mathrm{T}}.
$$

Proof. We assume $\mathbf{K} \in \mathbb{R}_{\geq 0}^{n\times m}$ and $\mathbf{\Phi} \in \{0,1\}^{n\times m}$. Let $\mathbf{G} \in \{0,1\}^{n\times n}$ and $\mathbf{H} \in \{0,1\}^{m\times m}$ be the permutation matrices corresponding to the mappings $\mathcal{G} : i \mapsto g(i)$ and $\mathcal{H} : j \mapsto h(j)$, respectively. This means that, for a given matrix $\mathbf{X}$ and its row and column permutation $\mathbf{X}' = \mathbf{GXH}$, we have $\mathbf{X}'_{g(i),h(j)} = \mathbf{X}_{i,j}$. We can also apply the permutation matrices $\mathbf{G}$ and $\mathbf{H}$ to vectors $\boldsymbol{w} \in \mathbb{R}_{\geq 0}^{n}$ and $\boldsymbol{h} \in \mathbb{R}_{\geq 0}^{m}$. For $\boldsymbol{w}' = \mathbf{G}\boldsymbol{w}$ and $\boldsymbol{h}' = \mathbf{H}\boldsymbol{h}$, it holds that $w'_{g(i)} = w_i$ and $h'_{h(j)} = h_j$. We define $\boldsymbol{w}^*$ and $\boldsymbol{h}^*$ as

$$
\begin{aligned}
\boldsymbol{w}^*, \boldsymbol{h}^* &\equiv \operatorname*{argmin}_{\boldsymbol{w},\boldsymbol{h}} D_{\mathbf{G\Phi H}}\left(\mathbf{GXH}, \boldsymbol{w} \otimes \boldsymbol{h}\right) \\
&= \operatorname*{argmin}_{\boldsymbol{w},\boldsymbol{h}} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\mathbf{\Phi}_{g(i)h(j)} a_i b_j - \mathbf{\Phi}_{g(i)h(j)} \mathbf{X}_{g(i)h(j)} \log a_i b_j\right).
\end{aligned}
$$

We replace $\boldsymbol{w}$ with $\mathbf{G}\boldsymbol{w}$ and $\boldsymbol{h}$ with $\mathbf{H}\boldsymbol{h}$ and we get

$$
\begin{aligned}
\mathbf{G}\boldsymbol{w}^*, \mathbf{H}\boldsymbol{h}^* &= \operatorname*{argmin}_{\boldsymbol{w},\boldsymbol{h}} \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\mathbf{\Phi}_{g(i)h(j)} a_{g(i)} b_{h(j)} - \mathbf{\Phi}_{g(i)h(j)} \mathbf{X}_{g(i)h(j)} \log a_{g(i)} b_{h(j)}\right) \\
&= \operatorname*{argmin}_{\boldsymbol{w},\boldsymbol{h}} \sum_{g(i)=1}^{n}\sum_{h(j)=1}^{m} \left(\mathbf{\Phi}_{g(i)h(j)} a_{g(i)} b_{h(j)} - \mathbf{\Phi}_{g(i)h(j)} \mathbf{X}_{g(i)h(j)} \log a_{g(i)} b_{h(j)}\right) \\
&= \operatorname*{argmin}_{\boldsymbol{w},\boldsymbol{h}} \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\mathbf{\Phi}_{ij} a_i b_j - \mathbf{\Phi}_{ij} \mathbf{X}_{ij} \log a_i b_j\right) \\
&= D_{\mathbf{\Phi}}\left(\mathbf{X}, \boldsymbol{w} \otimes \boldsymbol{h}\right).
\end{aligned}
$$

Thus, it holds that $\boldsymbol{w}^* \otimes \boldsymbol{h}^* = \mathrm{NMF}_1\left(\mathbf{G\Phi H}, \mathbf{GXH}\right)$ and $\mathbf{G}\left(\boldsymbol{w}^* \otimes \boldsymbol{h}^*\right)\mathbf{H} = \mathrm{NMF}_1\left(\mathbf{\Phi}, \mathbf{X}\right)$. Therefore, we have

$$
\mathrm{NMF}_1\left(\mathbf{G\Phi H}, \mathbf{GTH}\right) = \mathbf{G}^{\mathrm{T}} \mathrm{NMF}_1\left(\mathbf{\Phi}, \mathbf{K}\right) \mathbf{H}^{\mathrm{T}}.
$$

We use the fact that permutation matrix is always orthogonal; that is, $\mathbf{G}^{-1} = \mathbf{G}^{\mathrm{T}}$ and $\mathbf{H}^{-1} = \mathbf{H}^{\mathrm{T}}$. □

Proposition 3 (rank-1 condition on θ-form)**.** *For any positive tensor $\overline{\mathcal{P}}$,* $\operatorname{rank}(\overline{\mathcal{P}}) = 1$ *if and only if all of its many-body θ-parameters are* 0*.*

Proof. First, we show that $\operatorname{rank}(\overline{\mathcal{P}}) = 1 \Rightarrow$ all many-body θ-parameters are 0. From the assumption of $\operatorname{rank}(\overline{\mathcal{P}}) = 1$, the m-th row of the mode-k expansion of $\overline{\mathcal{P}}$ has to be a constant multiple of the $(m-1)$-th row for all $m \in [2, I_k]$ and $k \in [d]$. That is,

$$
\begin{aligned}
\frac{\overline{\mathbf{P}}^{(k)}_{m,j}}{\overline{\mathbf{P}}^{(k)}_{m-1,j}} &= \frac{\overline{\mathcal{P}}_{i_1,\dots,i_{k-1},m,i_{k+1},\dots,i_d}}{\overline{\mathcal{P}}_{i_1,\dots,i_{k-1},m-1,i_{k+1},\dots,i_d}} \\
&= \exp\left(\sum_{i'_1=1}^{i_1} \cdots \sum_{i'_{k-1}=1}^{i_{k-1}} \sum_{i'_{k+1}=1}^{i_{k+1}} \cdots \sum_{i'_d=1}^{i_d} \theta_{i'_1,\dots,i'_{k-1},m,i'_{k+1},\dots,i'_d}\right)
\end{aligned}
$$

can depend on only m. If a many-body parameter $\theta_{i'_1,\dots,m,\dots i'_d}$ is not 0, the left side of the above equation depends on indices other than m. For example, if a many-body parameter $\theta_{2,1,\dots,1,m,1,\dots 1}$ is not 0, the right side of the equation depends on the value of i_1, which implies a contradiction with the assumption that the m-th row is a constant multiple of the $(m-1)$-th row. Therefore, all many-body θ-parameters of rank-1 tensor are 0.

Next, we show that $\operatorname{rank}(\overline{\mathcal{P}}) = 1$ if all many-body θ-parameters are 0. If all many-body θ-parameters are 0, we have

$$
\overline{\mathcal{P}}_{i_1,\dots,i_d} = \exp\left(\theta_{1,1,\dots,1}\right) \prod_{k=1}^{d} \exp\left(\sum_{i'_k=2}^{i_k} \theta^{(k)}_{i'_k}\right).
$$

Then we can represent the tensor $\overline{\mathcal{P}}$ as the Kronecker products of d vectors $\boldsymbol{s}^{(1)} \in \mathbb{R}^{I_1}, \boldsymbol{s}^{(2)} \in \mathbb{R}^{I_2}, \dots,$ and $\boldsymbol{s}^{(d)} \in \mathbb{R}^{I_d}$, whose elements are described as

$$
s^{(k)}_{i_k} = \exp\left(\frac{\theta_{1,\dots,1}}{d}\right) \exp\left(\sum_{i'_k=2}^{i_k} \theta^{(k)}_{i'_k}\right)
$$

for each $k \in [d]$ and $i_k \in [I_k]$. Thus, $\operatorname{rank}(\overline{\mathcal{P}}) = 1$ followed by the definition of the tensor rank. □

Proposition 4 (rank-1 condition as η-form)**.** *For any positive dth-order tensor $\overline{\mathcal{P}} \in \mathbb{R}^{I_1 \times \cdots \times I_d}_{>0}$,* $\operatorname{rank}(\overline{\mathcal{P}}) = 1$ *if and only if all of its many-body η parameters are factorizable as*

$$
\overline{\eta}_{i_1,\dots,i_d} = \prod_{k=1}^{d} \overline{\eta}^{(k)}_{i_k}.
$$

Proof. First, we show that all many-body η-parameters are factorizable if $\mathrm{rank}(\overline{\mathcal{P}}) = 1$. Since we can decompose a rank-1 tensor as a product of normalized independent distributions $\boldsymbol{s}^{(k)} \in \mathbb{R}^{I_k}$ as shown in Section 3.3.5, we can decompose many-body η parameters of $\overline{\mathcal{P}}$ as follows:

$$
\begin{aligned}
\overline{\eta}_{i_1,\dots,i_d} &\overset{(14)}{=} \sum_{i'_1=i_1}^{I_1} \cdots \sum_{i'_d=i_d}^{I_d} \overline{\mathcal{P}}_{i'_1,\dots,i'_d} \\
&\overset{(19)}{=} \sum_{i'_1=i_1}^{I_1} \cdots \sum_{i'_d=i_d}^{I_d} \left(s^{(1)}_{i'_1} s^{(2)}_{i'_2} \dots s^{(d)}_{i'_d} \right) \\
&= \left(\sum_{i'_1=i_1}^{I_1} s^{(1)}_{i'_1} \right) \left(\sum_{i'_2=i_2}^{I_2} s^{(2)}_{i'_2} \right) \cdots \left(\sum_{i'_d=i_d}^{I_d} s^{(d)}_{i'_d} \right) \\
&= \prod_{k=1}^{d} \left(\sum_{i'_k=i_k}^{I_k} s^{(k)}_{i'_k} \right) \\
&= \prod_{k=1}^{d} \left(\sum_{i'_k=i_k}^{I_k} s^{(k)}_{i'_k} \sum_{i'_1=1}^{I_1} s^{(1)}_{i'_1} \sum_{i'_2=1}^{I_2} s^{(2)}_{i'_2} \cdots \sum_{i'_d=1}^{I_d} s^{(d)}_{i'_d} \right) \\
&= \prod_{k=1}^{d} \left(\overline{\eta}^{(k)}_{i_k} \sum_{i'_k=1}^{I_k} s^{(k)}_{i'_k} \right) = \prod_{k=1}^{d} \overline{\eta}^{(k)}_{i_k},
\end{aligned}
$$

where we use the normalization condition

$$
\sum_{i'_k=1}^{I_k} s^{(k)}_{i'_k} = 1
$$

for each $k \in [d]$.

Next, we show the opposite direction. If all many-body η-parameters are factorizable, it follows that

$$
\begin{aligned}
\overline{\mathcal{P}}_{i_1,\dots,i_d} &\overset{(15)}{=} \sum_{(i'_1,\dots,i'_d)\in\Omega_d} \left(\mu^{i'_1,\dots,i'_d}_{i_1,\dots,i_d} \prod_{k=1}^{d} \overline{\eta}^{(k)}_{i'_k} \right) \\
&= \sum_{(i'_1\dots i'_d)\in\Omega_d} \left(\prod_{k=1}^{d} \mu^{i'_k}_{i_k} \overline{\eta}^{(k)}_{i'_k} \right) \\
&= \prod_{k=1}^{d} \left(\overline{\eta}^{(k)}_{i_k} - \overline{\eta}^{(k)}_{i_k+1} \right) \equiv \prod_{k=1}^{d} s^{(k)}_{j_k}.
\end{aligned}
$$

This formula means that the tensor $\mathcal{P}$ can be represented as a Kronecker product of vectors $\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(d)}$. Thus, $\mathrm{rank}(\overline{\mathcal{P}}) = 1$ holds by the definition of the tensor rank. □

Proposition 5 (m-projection onto factorizable subspace)**.** *For any positive tensor $\mathcal{P} \in \mathbb{R}^{I_1\times\cdots\times I_d}_{>0}$, its m-projection onto the rank-1 space is given as*

$$
\overline{\mathcal{P}}_{i_1,\dots,i_d} = \prod_{k=1}^{d} \left(\sum_{i'_1=1}^{I_1} \cdots \sum_{i'_{k-1}=1}^{I_{k-1}} \sum_{i'_{k+1}=1}^{I_{k+1}} \cdots \sum_{i'_d=1}^{I_d} \mathcal{P}_{i'_1,\dots,i'_{k-1},i_k,i'_{k+1},\dots,i_d} \right). \tag{A3}
$$

Proof.

$$
\begin{aligned}
\overline{\mathcal{P}}_{i_1,\dots,i_d} &\overset{(15)}{=} \sum_{(i'_1\dots i'_d)\in\Omega_d} \mu_{i_1\dots i_d}^{i'_1,\dots,i'_d}\overline{\eta}_{i'_1,\dots,i'_d} \\
&\overset{(16)}{=} \sum_{(i'_1,\dots,i'_d)\in\Omega_d} \left(\mu_{i_1,\dots,i_d}^{i'_1,\dots,i'_d}\prod_{k=1}^{d}\overline{\eta}_{i'_k}^{(k)}\right) \\
&\overset{(17)}{=} \sum_{(i'_1\dots i'_d)\in\Omega_d} \left(\mu_{i_1,\dots,i_d}^{i'_1,\dots,i'_d}\prod_{k=1}^{d}\eta_{i'_k}^{(k)}\right) \\
&= \sum_{(i'_1,\dots,i'_d)\in\Omega_d} \left(\prod_{k=1}^{d}\mu_{i_k}^{i'_k}\eta_{i'_k}^{(k)}\right) \\
&= \prod_{k=1}^{d}\left(\eta_{i_k}^{(k)}-\eta_{i_k+1}^{(k)}\right) \\
&\overset{(14)}{=} \prod_{k=1}^{d}\left(\sum_{i'_1=1}^{I_1}\cdots\sum_{i'_{k-1}=1}^{I_{k-1}}\sum_{i'_{k+1}=1}^{I_{k+1}}\cdots\sum_{i'_d=1}^{I_d}\mathcal{P}_{i'_1,\dots,i'_{k-1},i_k,i'_{k+1},\dots,i_d}\right).
\end{aligned}
$$

□

Proposition 6 (Bingo and Tucker rank)**.** *If there are b_k bingos on mode-k, it holds that*

$$\mathrm{Rank}(\mathbf{P}^{(k)}) \le I_k - b_k.$$

Proof. If there is a bingo on mode-k, the m-th row of the mode-k expansion of $\mathcal{P}$ is a constant multiple of the $(m-1)$-th row, where m is a number determined by the bingo position. We can confirm that

$$
\begin{aligned}
\frac{\mathbf{P}^{(k)}_{m,j}}{\mathbf{P}^{(k)}_{m-1,j}} &= \frac{\mathcal{P}_{i_1,\dots,i_{k-1},m,i_{k+1},\dots,i_d}}{\mathcal{P}_{i_1,\dots,i_{k-1},m-1,i_{k+1},\dots,i_d}} \\
&= \frac{\exp\left(\sum_{i'_1=1}^{i_1}\cdots\sum_{i'_{k-1}=1}^{i_{k-1}}\sum_{i'_k=1}^{m}\sum_{i'_{k+1}=1}^{i_{k+1}}\cdots\sum_{i'_d=1}^{i_d}\theta_{i'_1,\dots,i'_d}\right)}{\exp\left(\sum_{i'_1=1}^{i_1}\cdots\sum_{i'_{k-1}=1}^{i_{k-1}}\sum_{i'_k=1}^{m-1}\sum_{i'_{k+1}=1}^{i_{k+1}}\cdots\sum_{i'_d=1}^{i_d}\theta_{i'_1,\dots,i'_d}\right)} \\
&= \exp\left(\sum_{i'_1=1}^{i_1}\cdots\sum_{i'_{k-1}=1}^{i_{k-1}}\sum_{i'_{k+1}=1}^{i_{k+1}}\cdots\sum_{i'_d=1}^{i_d}\theta_{i'_1,\dots,i'_{k-1},m,i'_{k+1},\dots,i'_d}\right) \\
&= \exp\left(\theta_{1,\dots,1,m,1,\dots,1}\right)
\end{aligned}
$$

is just a constant that does not depend on j. When a row is a constant multiple of another row, the rank of the matrix is reduced by a maximum of one, which means $\mathrm{Rank}(\mathbf{P}^{(k)}) \le I_k - 1$. In the same way, if there are b_k bingos, then b_k rows are constant multiple of the other rows, which means $\mathrm{Rank}(\mathbf{P}^{(k)}) \le I_k - b_k$. □

Proposition 7. *For any positive dth-order rank-1 tensor $\overline{\mathcal{P}} \in \mathbb{R}_{>0}^{I_1\times\cdots\times I_d}$, its one-body θ- and η-parameters satisfy the following equations*

$$\overline{\theta}_j^{(k)} = \log\left(\frac{\overline{\eta}_j^{(k)}-\overline{\eta}_{j+1}^{(k)}}{\overline{\eta}_{j-1}^{(k)}-\overline{\eta}_j^{(k)}}\right), \qquad \overline{\eta}_j^{(k)} = \frac{\sum_{i_k=j}^{I_k}\exp\sum_{i'_k=2}^{i_k}\overline{\theta}_{i'_k}^{(k)}}{1+\sum_{i_k=2}^{I_k}\exp\left(\sum_{i'_k=2}^{i_k}\overline{\theta}_{i'_k}^{(k)}\right)},$$

where we assume $\overline{\eta}_0^{(k)} = \overline{\eta}_{I_k+1}^{(k)} = 0$.

Proof. As shown in Theorem 2 in (Sugiyama et al., 2017), the relation between θ and η is obtained by the differentiation of Helmholtz's free energy $\psi(\theta)$, which is defined as the sign inverse normalization factor. For the rank-1 tensor $\overline{\mathcal{P}}$, Helmholtz's free energy $\psi(\theta)$ is given as

$$\psi(\overline{\theta}) = \log \prod_{k=1}^{d} \left(1 + \sum_{i_k=2}^{I_k} \exp \left(\sum_{i'_k=2}^{i_k} \overline{\theta}^{(k)}_{i'_k} \right) \right).$$

We obtain the expectation parameters η by the differentiation of Helmholtz's free energy $\psi(\theta)$ by θ, given as

$$\overline{\eta}^{(k)}_j = \frac{\partial}{\partial \overline{\theta}^{(k)}_j} \psi(\overline{\theta}) = \frac{\sum_{i_k=j}^{I_k} \exp \sum_{i'_k=2}^{i_k} \overline{\theta}^{(k)}_{i'_k}}{1 + \sum_{i_k=2}^{I_k} \exp \left(\sum_{i'_k=2}^{i_k} \overline{\theta}^{(k)}_{i'_k} \right)}.$$

By solving the above equation inverse, we obtain

$$\overline{\theta}^{(k)}_j = \log \left(\frac{\overline{\eta}^{(k)}_j - \overline{\eta}^{(k)}_{j+1}}{\overline{\eta}^{(k)}_{j-1} - \overline{\eta}^{(k)}_j} \right).$$

$\square$

Proposition 8. *Let θ denote natural parameters of given tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I_1 \times \cdots \times I_d}$ and $\overline{\theta}$ denote natural parameters of $\overline{\mathcal{P}}$, which is the best rank-1 approximation that minimizes KL divergence from $\mathcal{P}$. If a one-body natural parameter $\theta^{(j)}_{i_j} = 0$, its values after the best rank-1 approximation $\overline{\theta}^{(j)}_{i_j}$ remain 0.*

Proof. When $\theta^{(j)}_{i_j} = 0$, it holds that

$$\frac{\mathcal{P}_{1,\dots,1,i_j,1,\dots,1}}{\mathcal{P}_{1,\dots,1,i_j-1,1,\dots,1}} = \exp\left(\theta^{(j)}_{i_j}\right) = 1.$$

By using the closed formula of the best rank-1 approximation (11), we obtain

$$\begin{aligned}
\overline{\mathcal{P}}_{1,\dots,1,i_j,1,\dots,1} &= \prod_{k=1}^{d} \left(\sum_{i'_1=1}^{I_1} \cdots \sum_{i'_{k-1}=1}^{I_{k-1}} \sum_{i'_{k+1}=1}^{I_{k+1}} \cdots \sum_{i'_d=1}^{I_d} \mathcal{P}_{i'_1,\dots,i'_{k-1},1,i'_{k+1},\dots,i_j,\dots,i'_d} \right) \\
&= \prod_{k=1}^{d} \left(\sum_{i'_1=1}^{I_1} \cdots \sum_{i'_{k-1}=1}^{I_{k-1}} \sum_{i'_{k+1}=1}^{I_{k+1}} \cdots \sum_{i'_d=1}^{I_d} \mathcal{P}_{i'_1,\dots,i'_{k-1},1,i'_{k+1},\dots,i_j-1,\dots,i'_d} \right) \\
&= \overline{\mathcal{P}}_{1,\dots,1,i_j-1,1,\dots,1}.
\end{aligned}$$

It follows that

$$\frac{\overline{\mathcal{P}}_{1,\dots,1,i_j,1,\dots,1}}{\overline{\mathcal{P}}_{1,\dots,1,i_j-1,1,\dots,1}} = \exp\left(\overline{\theta}^{(j)}_{i_j}\right) = 1.$$

Finally, we obtain $\overline{\theta}^{(j)}_{i_j} = 0$. $\square$

Theorem 1 (the best rank-1 NMMF)**.** *For any given four positive matrices $\mathbf{X} \in \mathbb{R}_{>0}^{I \times J}$, $\mathbf{Y} \in \mathbb{R}_{>0}^{N \times J}$, $\mathbf{Z} \in \mathbb{R}_{>0}^{I \times M}$, and $\mathbf{U} \in \mathbb{R}_{>0}^{L \times M}$ and three parameters $\alpha, \beta, \gamma \geq 0$, five non-negative vectors $\boldsymbol{w} \in \mathbb{R}_{\geq 0}^{I}$, $\boldsymbol{h} \in \mathbb{R}_{\geq 0}^{J}$, $\boldsymbol{a} \in \mathbb{R}_{\geq 0}^{N}$, $\boldsymbol{b} \in \mathbb{R}_{\geq 0}^{M}$, and $\boldsymbol{c} \in \mathbb{R}_{\geq 0}^{L}$ that minimize the cost function in Equation* (4) *is* *given as*

$$w_i = \frac{\sqrt{S(\mathbf{X})}}{S(\mathbf{X}) + \beta S(\mathbf{Z})} \left(\sum_{j=1}^{J} \mathbf{X}_{ij} + \beta \sum_{m=1}^{M} \mathbf{Z}_{im} \right)$$

$$h_j = \frac{\sqrt{S(\mathbf{X})}}{S(\mathbf{X}) + \alpha S(\mathbf{Y})} \left(\sum_{i=1}^{I} \mathbf{X}_{ij} + \alpha \sum_{n=1}^{N} \mathbf{Y}_{nj} \right)$$

$$a_h = \frac{1}{\sqrt{S(\mathbf{X})}} \left(\sum_{j=1}^{J} \mathbf{Y}_{nj} \right)$$

$$b_m = \frac{S(\mathbf{Z})}{\beta S(\mathbf{Z}) + \gamma S(\mathbf{U})} \frac{1}{\sqrt{S(\mathbf{X})}} \left(\beta \sum_{i=1}^{I} \mathbf{Z}_{im} + \gamma \sum_{l=1}^{L} \mathbf{U}_{lm} \right)$$

$$c_l = \frac{\sqrt{S(\mathbf{X})}}{S(\mathbf{Z})} \left(\sum_{m=1}^{M} \mathbf{U}_{lm} \right)$$

Proof. First, we show this theorem with $\alpha = \beta = \gamma = 1$, followed by generalizing the result for any non-negative α, β and γ. Hereinafter, we use overline for quantities on the simultaneous rank-1 subspace; for example, $(\overline{\mathbf{X}}, \overline{\mathbf{Y}}, \overline{\mathbf{Z}}, \overline{\mathbf{U}})$ as rank-1 matrices sharing factors, $\overline{\eta}$ as expectation parameters for distributions on simultaneous rank-1 subspace. We decompose input matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and $\mathbf{U}$ as $\overline{\mathbf{X}} = \boldsymbol{w} \otimes \boldsymbol{h}$, $\overline{\mathbf{Y}} = \boldsymbol{a} \otimes \boldsymbol{h}$, $\overline{\mathbf{Z}} = \boldsymbol{w} \otimes \boldsymbol{b}$, and $\overline{\mathbf{U}} = \boldsymbol{c} \otimes \boldsymbol{b}$, respectively, so that they minimize the cost function

$$D(\mathbf{X}, \boldsymbol{w} \otimes \boldsymbol{h}) + \alpha D(\mathbf{Y}, \boldsymbol{a} \otimes \boldsymbol{h}) + \beta D(\mathbf{Z}, \boldsymbol{w} \otimes \boldsymbol{b}) + \gamma D(\mathbf{U}, \boldsymbol{c} \otimes \boldsymbol{b}).$$

To simplify, we define

$$\eta^{\mathbf{Y}}_{nj} = \eta_{nj}, \;\; \eta^{\mathbf{X}}_{ij} = \eta_{N+i,j}, \;\; \eta^{\mathbf{Z}}_{im} = \eta_{N+i,J+m}, \;\; \eta^{\mathbf{U}}_{lm} = \eta_{N+I+l,J+m},$$

for all $i \in [I], j \in [J], n \in [N], m \in [m]$ and $l \in [L]$. According to the expectation conservation law in this m-projection, it holds that

$$\eta^{\mathbf{Y}}_{n1} = \overline{\eta}^{\mathbf{Y}}_{n1}, \quad \eta^{\mathbf{Y}}_{1j} = \overline{\eta}^{\mathbf{Y}}_{1j}, \quad \eta^{\mathbf{X}}_{i1} = \overline{\eta}^{\mathbf{X}}_{i1}, \quad \eta^{\mathbf{Z}}_{1m} = \overline{\eta}^{\mathbf{Z}}_{1m}, \quad \eta^{\mathbf{U}}_{l1} = \overline{\eta}^{\mathbf{U}}_{l1}$$

where $(\eta^{\mathbf{X}}, \eta^{\mathbf{Y}}, \eta^{\mathbf{Z}}, \eta^{\mathbf{U}})$ is the expectation parameter of input, and $(\overline{\eta}^{\mathbf{X}}, \overline{\eta}^{\mathbf{Y}}, \overline{\eta}^{\mathbf{Z}}, \overline{\eta}^{\mathbf{U}})$ is the expectation parameter after the m-projection. By the definition of expectation parameters and the conservation law, we obtain

$$\begin{cases} \eta^{\mathbf{Y}}_{n1} - \eta^{\mathbf{Y}}_{n+1,1} = a_n S(\boldsymbol{h}), \\ \eta^{\mathbf{X}}_{i1} - \eta^{\mathbf{X}}_{i+1,1} = w_i \left(S(\boldsymbol{h}) + S(\boldsymbol{b}) \right), \\ \eta^{\mathbf{U}}_{l1} - \eta^{\mathbf{U}}_{l+1,1} = d_l S(\boldsymbol{b}), \\ \eta^{\mathbf{Y}}_{1j} - \eta^{\mathbf{Y}}_{1,j+1} = \left(S(\boldsymbol{a}) + S(\boldsymbol{w}) \right) h_j, \\ \eta^{\mathbf{Z}}_{1m} - \eta^{\mathbf{Z}}_{1,m+1} = \left(S(\boldsymbol{w}) + S(\boldsymbol{d}) \right) b_m. \end{cases}$$

We multiply these equations together and simplify them, resulting in

$$\overline{\mathbf{X}}_{ij} = w_i h_j = \frac{\left(\eta^{\mathbf{X}}_{i1} - \eta^{\mathbf{X}}_{i+1,1} \right) \left(\eta^{\mathbf{Y}}_{1j} - \eta^{\mathbf{Y}}_{1,j+1} \right)}{\left(S(\boldsymbol{w}) + S(\boldsymbol{a}) \right) \left(S(\boldsymbol{h}) + S(\boldsymbol{b}) \right)},$$

$$\overline{\mathbf{Y}}_{nj} = a_n h_j = \frac{\left(\eta^{\mathbf{Y}}_{n1} - \eta^{\mathbf{Y}}_{n+1,1} \right) \left(\eta^{\mathbf{Y}}_{1j} - \eta^{\mathbf{Y}}_{1,j+1} \right)}{\left(S(\boldsymbol{w}) + S(\boldsymbol{a}) \right) S(\boldsymbol{h})},$$

$$\overline{\mathbf{Z}}_{im} = w_i b_m = \frac{\left(\eta^{\mathbf{X}}_{i1} - \eta^{\mathbf{X}}_{i+1,1} \right) \left(\eta^{\mathbf{Z}}_{1m} - \eta^{\mathbf{Z}}_{1,m+1} \right)}{\left(S(\boldsymbol{h}) + S(\boldsymbol{b}) \right) \left(S(\boldsymbol{w}) + S(\boldsymbol{d}) \right)},$$

$$\overline{\mathbf{U}}_{lm} = d_l b_m = \frac{\left(\eta^{\mathbf{U}}_{l1} - \eta^{\mathbf{U}}_{l+1,1} \right) \left(\eta^{\mathbf{Z}}_{1m} - \eta^{\mathbf{Z}}_{1,m+1} \right)}{S(\boldsymbol{b}) \left(S(\boldsymbol{w}) + S(\boldsymbol{d}) \right)}.$$

Since the sum of each matrix $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and $\mathbf{U}$ are represented by conserved quantities, $S(\mathbf{X}) = \eta_{11}^{\mathbf{X}} - \eta_{11}^{\mathbf{Z}}$, $S(\mathbf{Y}) = \eta_{11}^{\mathbf{Y}} - \eta_{11}^{\mathbf{X}}$, $S(\mathbf{Z}) = \eta_{11}^{\mathbf{Z}} - \eta_{11}^{\mathbf{U}}$, and $S(\mathbf{U}) = \eta_{11}^{\mathbf{U}}$,the sum of each matrix also do not change in this projection; that is,

$$
\begin{aligned}
S(\mathbf{X}) = S(\overline{\mathbf{X}}) = S(\boldsymbol{w} \otimes \boldsymbol{h}) = S(\boldsymbol{w})S(\boldsymbol{h}),\\
S(\mathbf{Y}) = S(\overline{\mathbf{Y}}) = S(\boldsymbol{a} \otimes \boldsymbol{h}) = S(\boldsymbol{a})S(\boldsymbol{h}),\\
S(\mathbf{Z}) = S(\overline{\mathbf{Z}}) = S(\boldsymbol{w} \otimes \boldsymbol{b}) = S(\boldsymbol{w})S(\boldsymbol{b}),\\
S(\mathbf{U}) = S(\overline{\mathbf{U}}) = S(\boldsymbol{c} \otimes \boldsymbol{b}) = S(\boldsymbol{c})S(\boldsymbol{b}),
\end{aligned}
$$

and we obtain

$$
\begin{cases}
\overline{\mathbf{X}}_{ij} = \dfrac{S(\mathbf{X})\left(\sum_{j'}^{J} \mathbf{X}_{ij'} + \sum_{m}^{M} \mathbf{Z}_{im}\right)\left(\sum_{i'}^{I} \mathbf{X}_{i'j} + \sum_{n}^{N} \mathbf{Y}_{nj}\right)}{\big(S(\mathbf{X}) + S(\mathbf{Y})\big)\big(S(\mathbf{X}) + S(\mathbf{Z})\big)},\\
\overline{\mathbf{Y}}_{nj} = \dfrac{\left(\sum_{i}^{I} \mathbf{X}_{ij} + \sum_{n}^{N} \mathbf{Y}_{nj}\right)\left(\sum_{j'}^{J} \mathbf{Y}_{ij'}\right)}{S(\mathbf{X}) + S(\mathbf{Y})},\\
\overline{\mathbf{Z}}_{im} = \dfrac{S(\mathbf{Z})\left(\sum_{j}^{J} \mathbf{X}_{ij} + \sum_{m'}^{M} \mathbf{Z}_{im'}\right)\left(\sum_{i'}^{I} \mathbf{Z}_{i'm} + \sum_{l}^{L} \mathbf{U}_{lm}\right)}{\big(S(\mathbf{Z}) + S(\mathbf{U})\big)\big(S(\mathbf{X}) + S(\mathbf{Z})\big)}.\\
\overline{\mathbf{U}}_{lm} = \dfrac{\left(\sum_{i}^{I} \mathbf{Z}_{im} + \sum_{l'}^{L} \mathbf{U}_{l'j}\right)\left(\sum_{j'}^{J} \mathbf{U}_{ij'}\right)}{S(\mathbf{Z}) + S(\mathbf{U})}.
\end{cases}
$$

We used relations

$$
\begin{aligned}
\frac{1}{S(\mathbf{X}) + S(\mathbf{Y}) + S(\mathbf{Z}) + S(\boldsymbol{a})S(\boldsymbol{b})} &= \frac{S(\mathbf{X})}{\big(S(\mathbf{X}) + S(\mathbf{Y})\big)\big(S(\mathbf{X}) + S(\mathbf{Z})\big)},\\
\frac{1}{S(\mathbf{X}) + S(\mathbf{U}) + S(\mathbf{Z}) + S(\boldsymbol{h})S(\boldsymbol{c})} &= \frac{S(\mathbf{Z})}{\big(S(\mathbf{Z}) + S(\mathbf{U})\big)\big(S(\mathbf{Z}) + S(\mathbf{X})\big)}.
\end{aligned}
$$

Using the general property of the KL divergence, $\mu D(\mathbf{P}, \mathbf{Q}) = D(\mu\mathbf{P}, \mu\mathbf{Q})$ for any matrices $\mathbf{P}, \mathbf{Q}$ and non-negative number μ, the above result with general α, β and γ is obtained by shifting matrices $\mathbf{Y} \to \alpha\mathbf{Y}$, $\mathbf{Z} \to \beta\mathbf{Z}$ and $\mathbf{U} \to \gamma\mathbf{U}$ in the both sides of the above equation as

$$
\begin{cases}
\overline{\mathbf{X}}_{ij} = \dfrac{S(\mathbf{X})\left(\sum_{j'}^{J} \mathbf{X}_{ij'} + \beta\sum_{m}^{M} \mathbf{Z}_{im}\right)\left(\sum_{i'}^{I} \mathbf{X}_{i'j} + \alpha\sum_{n}^{N} \mathbf{Y}_{nj}\right)}{\big(S(\mathbf{X}) + \alpha S(\mathbf{Y})\big)\big(S(\mathbf{X}) + \beta S(\mathbf{Z})\big)},\\
\overline{\mathbf{Y}}_{nj} = \dfrac{\left(\sum_{i}^{I} \mathbf{X}_{ij} + \alpha\sum_{n}^{N} \mathbf{Y}_{nj}\right)\left(\sum_{j'}^{J} \mathbf{Y}_{ij'}\right)}{S(\mathbf{X}) + \alpha S(\mathbf{Y})},\\
\overline{\mathbf{Z}}_{im} = \dfrac{S(\mathbf{Z})\left(\sum_{j}^{J} \mathbf{X}_{ij} + \beta\sum_{m'}^{M} \mathbf{Z}_{im'}\right)\left(\beta\sum_{i'}^{I} \mathbf{Z}_{i'm} + \gamma\sum_{l}^{L} \mathbf{U}_{lm}\right)}{\big(\beta S(\mathbf{Z}) + \gamma S(\mathbf{U})\big)\big(S(\mathbf{X}) + \beta S(\mathbf{Z})\big)}.\\
\overline{\mathbf{U}}_{lm} = \dfrac{\left(\beta\sum_{i}^{I} \mathbf{Z}_{im} + \gamma\sum_{l'}^{L} \mathbf{U}_{l'j}\right)\left(\sum_{j'}^{J} \alpha\mathbf{U}_{ij'}\right)}{\beta S(\mathbf{Z}) + \gamma S(\mathbf{U})}.
\end{cases}
$$

Thus, the theorem was proved. □

Theorem 2. *The intersection $\mathcal{Q}_{\boldsymbol{c}} \cap \mathcal{B}_1$ is a singleton.*

Proof. As we discussed in Section 2.2.1, we can identify a distribution using the mixture coordinate system (θ, η) that combines θ- and η-coordinates. Therefore, specifying $I_1 \times \cdots \times I_d$ parameters on the mixture coordinate (θ, η) uniquely identifies a tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I_1 \times \cdots \times I_d}$. There are two kinds of parameters: one-body parameters and many-body parameters. On the intersection $\mathcal{Q}_{\boldsymbol{c}} \cap \mathcal{B}_1$, the $\boldsymbol{c}$-balancing condition in Equation (26) determines all one-body η-parameters $\eta_{i_k}^{(k)}$ for all $i_k \in [I_k]$ and $k \in [d]$. On the other hand, the rank-1 condition in Proposition 3 determines all many-body θ-parameters $\theta_{i_1,\dots,i_d} = 0$. Now, the $\boldsymbol{c}$-balancing conditions and the rank-1 condition specify all $I_1 \times \cdots \times I_d$ parameters, so the mixture coordinate (θ, η) uniquely identifies the rank-1 $\boldsymbol{c}$-balanced tensor. □

Theorem 3. *The intersection $\mathcal{Q}_{\mathcal{C}} \cap \mathcal{B}_1$ exists and is a singleton if and only if*

$$\mathcal{C}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_d} = \prod_{m\in[d]\setminus k} \left(\prod_{l\in[d]\setminus\{k,m\}} \sum_{i_l=1}^{I_l} \mathcal{C}^{\setminus k}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_d} \right).$$

Proof. A tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I_1\times\cdots\times I_d}$ on the subspace $\mathcal{Q}_{\mathcal{C}}$ satisfies $\mathcal{C}$-balancing condition that can be expressed by η-parameters as

$$\eta_{i_1,\dots,i_{k-1},1,i_{k+1},\dots,i_d} = \sum_{i'_1=i_1}^{I_1} \cdots \sum_{i'_{k-1}=i_{k-1}}^{I_{k-1}} \sum_{i'_{k+1}=i_{k+1}}^{I_{k+1}} \cdots \sum_{i'_d=i_d}^{I_d} \mathcal{C}^{\setminus k}_{i'_1,\dots,i'_{k-1},i'_{k+1},\dots,i'_d}.$$

However, in the intersection $\mathcal{Q}_{\mathcal{C}} \cap \mathcal{B}_1$, these parameters $\eta_{i_1,\dots,i_{k-1},1,i_{k+1},\dots,i_d}$ have to satisfy the rank-1 condition. Recall that the necessary and sufficient conditions of a tensor to be rank-1 tensor in Equation (16); if and only if the condition $\mathcal{C}$ satisfies the following relation, the tensor $\mathcal{P}$ can be on $\mathcal{Q}_{\mathcal{C}} \cap \mathcal{B}_1$.

$$\begin{aligned}
&\mathcal{C}^{\setminus k}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_d} \\
&\overset{\text{(A1)}}{=} \sum_{i'_k=1}^{I_k} \mathcal{P}_{i_1,\dots,i_d} \\
&\overset{\text{(15)}}{=} \sum_{(i'_1,\dots,i'_{k-1},1,i'_{k+1},\dots,i'_d)\in\Omega_d} \mu^{i'_1,\dots,i'_{k-1},1,i'_{k+1},\dots,i'_d}_{i_1,\dots,i_{k-1},1,i_{k+1},\dots,i_d} \eta_{i'_1,\dots,i'_{k-1},1,i'_{k+1},\dots,i'_d} \\
&\overset{\text{(16)}}{=} \sum_{(i'_1,\dots,i'_{k-1},1,i'_{k+1},\dots,i'_d)\in\Omega_d} \mu^{i'_1,\dots,i'_{k-1},1,i'_{k+1},\dots,i'_d}_{i_1,\dots,i_{k-1},1,i_{k+1},\dots,i_d} \left(\prod_{m\in[d]\setminus k} \eta^{(m)}_{i'_m} \right) \\
&= \sum_{(i'_1,\dots,i'_{k-1},1,i'_{k+1},\dots,i'_d)\in\Omega_d} \left(\prod_{m\in[d]\setminus k} \mu^{i'_m}_{i_m} \eta^{(m)}_{i'_m} \right) \\
&= \prod_{m\in[d]\setminus k} \left(\eta^{(m)}_{i_m} - \eta^{(m)}_{i_{m+1}} \right) \\
&\overset{\text{(14)}}{=} \prod_{m\in[d]\setminus k} \left(\sum_{i'_1=1}^{I_1} \cdots \sum_{i'_{k-1}=1}^{I_{k-1}} \sum_{i'_{k+1}=1}^{I_{k+1}} \cdots \sum_{i'_d=1}^{I_d} \mathcal{P}_{i'_1,\dots,i'_{m-1},i_m,i'_{m+1},\dots,i_d} \right). \\
&\overset{\text{(A1)}}{=} \prod_{m\in[d]\setminus k} \left(\prod_{l\in[d]\setminus\{k,m\}} \sum_{i_l=1}^{I_l} \mathcal{C}^{\setminus k}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_d} \right)
\end{aligned}$$

Thus, the theorem was proved. □

References

Honauer, K., Johannsen, O., Kondermann, D., and Goldluecke, B. (2016). A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, pages 19–34. Springer.

Kossaifi, J., Panagakis, Y., Anandkumar, A., and Pantic, M. (2019). Tensorly: Tensor learning in Python. *Journal of Machine Learning Research*, 20(26):1–6.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Periša, L. and Arslan, A. (2019). lanaperisa/tensortoolbox.jl v1.0.2.

Samaria, F. S. and Harter, A. C. (1994). Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pages 138–142. IEEE.

Sugiyama, M., Nakahara, H., and Tsuda, K. (2017). Tensor balancing on statistical manifold. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3270–3279.

Zhou, G., Cichocki, A., and Xie, S. (2012). Fast nonnegative matrix/tensor factorization based on low-rank approximation. *IEEE Transactions on Signal Processing*, 60(6):2928–2940.