Supplementary file of 'Multi-typed Objects Multi-view Multi-instance Multi-label Learning'

Anonymous

I. Optimization

We alternatively consider three of them as constants and optimize the other one in the (1) about variable $\mathbf{G}, \mathbf{S}, \mathbf{W}^r$ and $\mathbf{W}^h$. Before elaborating on the updating rule, we introduce the Lagrangian multipliers $\{\boldsymbol{\lambda}_i\}_{i=1}^m$ for $\mathbf{G}_i \geq 0$, and reformulate the objective function of M4L-JMF as follows:

$$\begin{aligned}\min_{G\geq 0} \boldsymbol{\Omega}(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h,\boldsymbol{\lambda}) &= \sum_{i,j=1}^{m} \boldsymbol{W}_{ij}^r \|\boldsymbol{R}_{ij} - \boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\|_F^2 \\ &+ \|\boldsymbol{R}_{bm} - \boldsymbol{R}_{bi}\boldsymbol{G}_i\boldsymbol{S}_{im}\boldsymbol{G}_m^T\|_F^2 \\ &+ \sum_{p=1}^{m}\sum_{t=1}^{\tau} \boldsymbol{W}_{pt}^h tr(\boldsymbol{G}_p^T\boldsymbol{\Theta}_p^{(t)}\boldsymbol{G}_p) \\ &+ \alpha\|vec(\boldsymbol{W}^r)\|_F^2 + \beta\|vec(\boldsymbol{W}^h)\|_F^2 \\ &- \sum_{i=1}^{m} tr(\boldsymbol{\lambda}_i\boldsymbol{G}_i^T) s.t. \boldsymbol{W}^r \geq 0, \\ &\boldsymbol{W}^h \geq 0, \sum vec(\boldsymbol{W}_i^r) = 1, \\ &\sum vec(\boldsymbol{W}_i^h) = 1\end{aligned} \tag{1}$$

The iterative updating rules for $\boldsymbol{G}$ and $\boldsymbol{S}$ follow the idea in [1], [2]. We approximate $\boldsymbol{G}_i\boldsymbol{S}_{im}\boldsymbol{G}_m^T$ to $\boldsymbol{R}_{im}$, and reformulate the objective function of M4L-JMF as follows:

$$\begin{aligned}\min_{G\geq 0} \boldsymbol{\Omega}(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h,\boldsymbol{\lambda}) &= \boldsymbol{W}_{ij}^r tr((\boldsymbol{R}_{ij} - \boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T)^T(\boldsymbol{R}_{ij} \\ &- \boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T)) + \|\boldsymbol{R}_{bm} - \boldsymbol{R}_{bi}\boldsymbol{R}_{im}\|_F^2 \\ &+ \sum_{p=1}^{m}\sum_{t=1}^{\tau} \boldsymbol{W}_{pt}^h tr(\boldsymbol{G}_p^T\boldsymbol{\Theta}_p^{(t)}\boldsymbol{G}_p) \\ &+ \alpha\|vec(\boldsymbol{W}^r)\|_F^2 + \beta\|vec(\boldsymbol{W}^h)\|_F^2 \\ &- \sum_{i=1}^{m} tr(\boldsymbol{\lambda}_i\boldsymbol{G}_i^T) s.t. \boldsymbol{W}^r \geq 0, \\ &\boldsymbol{W}^h \geq 0, \sum vec(\boldsymbol{W}_i^r) = 1, \\ &\sum vec(\boldsymbol{W}_i^h) = 1\end{aligned} \tag{2}$$

Suppose $\boldsymbol{G}$, $\boldsymbol{W}^r$ and $\boldsymbol{W}^h$ are known, to obtain the optimal $\boldsymbol{S}_{ij}$ (if $\boldsymbol{R}_{ij} \in \mathbb{R}$), we can take the partial derivative of $\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h,\boldsymbol{\lambda})$ with respect to $\boldsymbol{S}_{ij}$:

$$\frac{\partial \Omega}{\partial \boldsymbol{S}_{ij}} = \boldsymbol{W}_{ij}^r(-2\boldsymbol{G}_i^T\boldsymbol{R}_{ij}\boldsymbol{G}_j + 2\boldsymbol{G}_i^T\boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j) \tag{3}$$

Let $\frac{\partial \Omega}{\partial \boldsymbol{S}_{ij}} = 0$, we can get:

$$\boldsymbol{S}_{ij} = (\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{R}_{ij}\boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{G})^{-1} \tag{4}$$

Same way, suppose $\boldsymbol{S}, \boldsymbol{W}^r$ and $\boldsymbol{W}^h$ are known, the partial derivative of $\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h,\boldsymbol{\lambda})$ with respect to $\boldsymbol{G}_i$is:

$$\begin{aligned}\frac{\partial \Omega}{\partial \boldsymbol{G}_i} &= \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} (-2\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T + 2\boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)\boldsymbol{W}_{ij}^r \\ &+ \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} (-2\boldsymbol{R}_{ji}\boldsymbol{G}_j\boldsymbol{S}_{ji} + 2\boldsymbol{G}_i\boldsymbol{S}_{ji}^T\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ji}^T)\boldsymbol{W}_{ji}^r \\ &+ \sum_{t=1}^{\tau} 2\boldsymbol{W}_{pt}^h\boldsymbol{\Theta}_i^{(t)}\boldsymbol{G}_i - \boldsymbol{\lambda}_i\end{aligned} \tag{5}$$

Multipliers $\boldsymbol{\lambda}_i$ can be obtained from (5) by letting $\frac{\partial \Omega}{\partial \boldsymbol{G}_i} = 0$, The KKT (Karush-Kuhn-Tucker) complementary condition [3] for nonnegativity of $\boldsymbol{G}_i$is:

$$\begin{aligned}0 &= \lambda_i \circ \boldsymbol{G}_i \\ &= [\sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} (-2\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T + 2\boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)\boldsymbol{W}_{ij}^r \\ &+ \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} (-2\boldsymbol{R}_{ji}\boldsymbol{G}_j\boldsymbol{S}_{ji} + 2\boldsymbol{G}_i\boldsymbol{S}_{ji}^T\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ji}^T)\boldsymbol{W}_{ji}^r \\ &+ \sum_{t=1}^{\tau} 2\boldsymbol{W}_{pt}^h\boldsymbol{\Theta}_i^{(t)}\boldsymbol{G}_i)]\circ\boldsymbol{G}_i\end{aligned} \tag{6}$$

where $\circ$ denotes the Hadamard product. Equation (6) is a fixed point equation and the solution must satisfy it at convergence. We can let

$$\begin{aligned}\boldsymbol{\Theta}_i^{(t)} &= [\boldsymbol{\Theta}_i^{(t)}]^+ - [\boldsymbol{\Theta}_i^{(t)}]^+ \\ \boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T &= (\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^+ - (\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^- \\ \boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T &= (\boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^+ - (\boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^- \\ \boldsymbol{R}_{ji}\boldsymbol{G}_j\boldsymbol{S}_{ji}^T &= (\boldsymbol{R}_{ji}\boldsymbol{G}_j\boldsymbol{S}_{ji}^T)^+ - (\boldsymbol{R}_{ji}\boldsymbol{G}_j\boldsymbol{S}_{ji}^T)^- \\ \boldsymbol{G}_i\boldsymbol{S}_{ji}^T\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ji} &= (\boldsymbol{G}_i\boldsymbol{S}_{ji}^T\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ji})^+ - (\boldsymbol{G}_i\boldsymbol{S}_{ji}^T\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ji})^-\end{aligned} \tag{7}$$

$$\begin{aligned}\boldsymbol{G}_i^{(e)} &= \boldsymbol{W}_{ij}^r(\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^+ + \boldsymbol{W}_{ij}^r\boldsymbol{G}_i(\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^- \\ \boldsymbol{G}_i^{(d)} &= \boldsymbol{W}_{ij}^r(\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^- + \boldsymbol{W}_{ij}^r\boldsymbol{G}_i(\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^+ \\ \boldsymbol{G}_j^{(e)} &= \boldsymbol{W}_{ij}^r(\boldsymbol{R}_{ij}^T\boldsymbol{G}_i\boldsymbol{S}_{ij})^+ + \boldsymbol{W}_{ij}^r\boldsymbol{G}_j(\boldsymbol{S}_{ij}^T\boldsymbol{G}_j^T\boldsymbol{G}_i\boldsymbol{S}_{ij}^T)^- \\ \boldsymbol{G}_j^{(d)} &= \boldsymbol{W}_{ij}^r(\boldsymbol{R}_{ij}^T\boldsymbol{G}_i\boldsymbol{S}_{ij})^- + \boldsymbol{W}_{ij}^r\boldsymbol{G}_j(\boldsymbol{S}_{ij}^T\boldsymbol{G}_j^T\boldsymbol{G}_i\boldsymbol{S}_{ij}^T)^+\end{aligned} \tag{8}$$

$$\begin{aligned}\boldsymbol{G}_i^{(e)} + &= \boldsymbol{W}^h[\boldsymbol{\Theta}_i^{(t)}]^-\boldsymbol{G}_i \; for \; i = 1, 2, ..., m \\ \boldsymbol{G}_i^{(d)} + &= \boldsymbol{W}^h[\boldsymbol{\Theta}_i^{(t)}]^-\boldsymbol{G}_i \; for \; i = 1, 2, ..., m\end{aligned} \tag{9}$$

where the matrices with positive and negative symbols are defined as $\boldsymbol{A}^{+}=\frac{|\boldsymbol{A}|+\boldsymbol{A}}{2}$ and $\boldsymbol{A}^{-}=\frac{|\boldsymbol{A}|-\boldsymbol{A}}{2}$ respectively. We then update $\boldsymbol{G}$ as:

$$\boldsymbol{G}\longleftarrow \boldsymbol{G}\circ diag(\sqrt{\frac{\boldsymbol{G}_1^{(e)}}{\boldsymbol{G}_1^{(d)}}},\sqrt{\frac{\boldsymbol{G}_2^{(e)}}{\boldsymbol{G}_2^{(d)}}},...,\sqrt{\frac{\boldsymbol{G}_m^{(e)}}{\boldsymbol{G}_m^{(d)}}}) \tag{10}$$

where $\circ$ means the Hadamard product. The $\sqrt{\cdot}$ and $\div$ are entry-wise operations. After updating $\boldsymbol{S}$ and $\boldsymbol{G}$, we view them as known and take the partial derivative of $\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h,\boldsymbol{\lambda})$ with respect to $\boldsymbol{W}^r$and $\boldsymbol{W}^h$. Firstly, we introduce the optimization about $\boldsymbol{W}^r$. In this case, the second , the third and the fifth terms on the right of (2) are irrelevant to $\boldsymbol{W}^r$, and can be ignored. Then we can obtain.

$$\begin{aligned}\tilde{\boldsymbol{\Omega}}(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r)=\sum_{i,j=1}^{m}\boldsymbol{W}_{ij}^r\|\boldsymbol{R}_{ij}-\boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\|_F^2\\ +\alpha\|vec(\boldsymbol{W}^r)\|_F^2\ s.t.\boldsymbol{W}^r\geq 0,\sum vec(\boldsymbol{W}_i^r)=1\end{aligned} \tag{11}$$

Let $\boldsymbol{H}_{ij}=\|\boldsymbol{R}_{ij}-\boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\|_F^2$ be the reconstruction loss for $\boldsymbol{R}_{ij}$ then (11) can be updated as:

$$\begin{aligned}\tilde{\boldsymbol{\Omega}}(\boldsymbol{H},\boldsymbol{W}^r)=vec(\boldsymbol{W}^r)^Tvec(\boldsymbol{H})+\alpha vec(\boldsymbol{W}^r)^Tvec(\boldsymbol{W}^{\boldsymbol{r}})\\ s.t.\boldsymbol{W}^r\geq 0,\sum vec(\boldsymbol{W}^r)=1\end{aligned} \tag{12}$$

Equation (12) is a quadratic optimization problem with respect to $vec(\boldsymbol{W}^r)$. By introducing the Lagrangian multipliers ($\boldsymbol{\phi}\in\mathbb{R}^{m\times m}$ and γ) for the constraints of $\boldsymbol{W}^r$, (12) is formulated as:

$$\begin{aligned}\tilde{\boldsymbol{\Omega}}(\boldsymbol{H},\boldsymbol{W}^{\boldsymbol{r}},\boldsymbol{\phi},\boldsymbol{\gamma})=vec(\boldsymbol{W}^r)^Tvec(\boldsymbol{H})+\boldsymbol{\alpha}vec(\boldsymbol{W}^r)^Tvec(\boldsymbol{W}^{\boldsymbol{r}})\\ -\sum_{i,j=1}^{m}\boldsymbol{\phi}_{ij}\boldsymbol{W}_{ij}^r-\boldsymbol{\gamma}(\sum_{i,j=1}^{m}\boldsymbol{W}_{ij}^r-1)\end{aligned} \tag{13}$$

Based on the KKT conditions, the optimal $\boldsymbol{W}^{\boldsymbol{r}}$ should satisfy the following four conditions:
(i) Stationary condition: $\frac{\partial\tilde{\Omega}}{\partial\boldsymbol{W}^r}=\boldsymbol{H}+2\alpha\boldsymbol{W}^r-\boldsymbol{\phi}-\boldsymbol{\gamma}=\boldsymbol{0}$
(ii) Feasible condition: $\boldsymbol{W}^r\geq 0,\sum_{i,j=1}^{m}\boldsymbol{W}_{ij}^{\boldsymbol{r}}-1=0$
(iii) Dual feasibility: $\boldsymbol{\phi}_{ij}\geq 0,\forall\boldsymbol{R}_{ij}\in\mathbb{R}$
(iv) Complementary slackness: $\boldsymbol{\phi}_{ij}\boldsymbol{W}_{ij}^{\boldsymbol{r}}=0,\forall\boldsymbol{R}_{ij}\in\mathbb{R}$

From the stationary condition, $\boldsymbol{W}_{ij}$ can be computed as:

$$\boldsymbol{W}_{ij}^{\boldsymbol{r}}=\frac{\boldsymbol{\phi}_{ij}+\boldsymbol{\gamma}-\boldsymbol{H}_{ij}}{2\alpha} \tag{14}$$

We can find that $\boldsymbol{W}_{ij}^r$ depends on the specification of $\boldsymbol{\phi}_{ij}$ and $\boldsymbol{\gamma}$, which can be analyzed in the following cases:

1) If $\boldsymbol{\gamma}>\boldsymbol{H}_{ij}$, then $\boldsymbol{W}_{ij}^r>0$;we get $\boldsymbol{\phi}_{ij}=0$ by complementary slackness, then $\boldsymbol{W}_{ij}^r=\boldsymbol{\gamma}-\boldsymbol{H}_{ij}\boldsymbol{2\alpha}$;
2) If $\boldsymbol{\gamma}\leq\boldsymbol{H}_{ij}$,we get $\boldsymbol{W}_{ij}^r=0$ because of $\boldsymbol{\phi}_{ij}\boldsymbol{W}_{ij}^r=0$ and $\boldsymbol{W}_{ij}^r\geq 0,\boldsymbol{W}_{ij}^r\geq 0$,it requires $\boldsymbol{\phi}_{ij}>0$;

So we can let $\boldsymbol{W}_{ij}^r$ as:

$$\boldsymbol{W}_{ij}^r=\begin{cases}\frac{\boldsymbol{\gamma}-\boldsymbol{H}_{ij}}{2\alpha},if\ \boldsymbol{\phi}-\boldsymbol{H}_{ij}>0\ and\ \boldsymbol{R}_{ij}\in\mathbb{R}\\ \\ 0,if\ \boldsymbol{\phi}-\boldsymbol{H}_{ij}\leq 0\ or\ \boldsymbol{R}_{ij}\notin\mathbb{R}\end{cases} \tag{15}$$

Algorithm 1 Seek h and compute $\boldsymbol{V}_{\boldsymbol{W}^r}$

Input:$\boldsymbol{V}_{\boldsymbol{H}},\alpha$
Output: $h,\boldsymbol{V}_{\boldsymbol{W}^r}$

Initialize $h=|R|,\boldsymbol{\gamma}=0$.
while $h>0$ **do**
  $\boldsymbol{\gamma}\leftarrow(2\boldsymbol{\alpha}+\sum_{h'=1}^{h}\boldsymbol{V}_{\boldsymbol{H}}(h^{'}))/h$.
  if $\boldsymbol{\gamma}-\boldsymbol{V}_{\boldsymbol{H}}(h)>0$ **then**
    break.
  else
    $h\leftarrow h-1$.
  end if
end while
$\boldsymbol{V}_{\boldsymbol{W}^r}(h^{'})\leftarrow(\boldsymbol{\gamma}-\boldsymbol{V}_{\boldsymbol{H}}(h))/2\boldsymbol{\alpha}$, for $h^{'}=1,\cdots,h$
$\boldsymbol{V}_{\boldsymbol{W}^r}(h^{'})\leftarrow 0$,for $h^{'}=h+1,\cdots,|R|$
return $h,\boldsymbol{V}_{\boldsymbol{W}^r}$.

Let $\boldsymbol{V}_H\in\mathbb{R}^{|R|}$ store the entries of vector $vec(\boldsymbol{H})$ in ascending order with entries corresponding to $\boldsymbol{R}_{ij}\notin\mathbb{R}$ removed. Accordingly, $\boldsymbol{V}_{W^r}\in\mathbb{R}^{|R|}$ stores the corresponding entries of vector $vec(\boldsymbol{W}^r)$ with entries corresponding to $\boldsymbol{R}_{ij}\notin\mathbb{R}$ removed. For a not too big predefined $\boldsymbol{\alpha}$, there exists $h\in\{1,2,\cdots,|R|\}$ with $\boldsymbol{V}_H(h+1)\geq\gamma$ and $\boldsymbol{V}_H(h)<\boldsymbol{\gamma}$, satisfying $\sum\boldsymbol{V}_H=\sum_{\boldsymbol{V}_H(h)<\boldsymbol{\gamma}}\frac{\boldsymbol{\gamma}-\boldsymbol{V}_H(h)}{2\boldsymbol{\alpha}}=1$, then $\boldsymbol{V}_H(h^{'})$ has the following explicit solution:

$$\boldsymbol{V}_{W^r}(h^{'})=\begin{cases}\frac{\boldsymbol{\gamma}-\boldsymbol{H}_{ij}}{2\boldsymbol{\alpha}},if\ h^{'}\leq h\\ \\ 0,if\ h^{'}>h\end{cases} \tag{16}$$

From $\sum_{h'=1}^{|\mathbb{R}|}\boldsymbol{V}_{\boldsymbol{W}^r}(h^{'})=\sum_{h'=1}^{h}\frac{\boldsymbol{\gamma}-\boldsymbol{V}_H(h)}{2\boldsymbol{\alpha}}=1$, we can get the value for $\boldsymbol{\gamma}$ as:

$$\boldsymbol{\gamma}=\frac{2\boldsymbol{\alpha}+\sum_{h'=1}^{h}\boldsymbol{V}_H(h^{'})}{h} \tag{17}$$

From the solution of $\boldsymbol{V}_{\boldsymbol{W}^r}(h^{'})$, we observe that, if $\boldsymbol{V}_H^r(h^{'})$ is smaller than $\boldsymbol{V}_{\boldsymbol{W}^r}(h^{''})$ and $\boldsymbol{\gamma}-\boldsymbol{V}_{\boldsymbol{H}}(h^{'})$ then a larger weight is assigned to the relational data source corresponding to $\boldsymbol{V}_{\boldsymbol{H}}(h^{'})$ than to $\boldsymbol{V}_{\boldsymbol{H}}(h^{''})$ because the data matrix of the former data source can be more well approximated than the latter.To find h that satisfies $\boldsymbol{\gamma}-\boldsymbol{V}_{\boldsymbol{H}}(h)>0$ and $\boldsymbol{\gamma}-\boldsymbol{V}_{\boldsymbol{H}}(h+1)\leq 0$, we decrease h from $|\mathbb{R}|$ to 1 step by step, and the procedure is summarized in **Algorithm 1**. It's the same way that calculate and update $\boldsymbol{W}^h$. we find that the first ,the second and the fourth terms on the right of (1) are irrelevant to $\boldsymbol{W}^h$, and can be ignored. Then we can obtain

$$\begin{aligned}\tilde{\boldsymbol{\Omega}}(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^h)=+\sum_{p=1}^{m}\sum_{t=1}^{\tau}\boldsymbol{W}_{pt}^h tr(\boldsymbol{G}_p^T\boldsymbol{\Theta}_p^{(t)}\boldsymbol{G}_p)\\ +\beta\|\boldsymbol{vec}(\boldsymbol{W}^h)\|_F^2-\sum_{i=1}^{m}tr(\boldsymbol{\lambda}_i\boldsymbol{G}_i^T)s.t.\boldsymbol{W}^h\geq 0,\\ \sum vec(\boldsymbol{W}_i^h)=1\end{aligned} \tag{18}$$

Let $\boldsymbol{Q}_i^{(t)}=tr(\boldsymbol{G}_p^T\boldsymbol{\Theta}_p^{(t)}\boldsymbol{G}_p)$ be intra-association constraints for $\boldsymbol{G}_i$, then (18) can be update as:

$$\tilde{\mathbf{\Omega}}(\boldsymbol{Q}, \boldsymbol{W}^h) = vec(\boldsymbol{W}^h)^T vec(\boldsymbol{Q}) + \beta vec(\boldsymbol{W}^h) vec(\boldsymbol{W}^h)$$
$$s.t. \boldsymbol{W}^h \geq 0, \sum vec(\boldsymbol{W}_i^h) = 1 \tag{19}$$

Equation (19) is a quadratic optimization problem with respect to $vec(\boldsymbol{W}^h)$ By introducing the Lagrangian multipliers $\boldsymbol{\lambda}$ (and $\boldsymbol{\mu}$) for the constraints of $\boldsymbol{W}^h$, (19) is formulated as:

$$\begin{aligned}\tilde{\mathbf{\Omega}}(\boldsymbol{Q}, \boldsymbol{W}^h, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= vec(\boldsymbol{W}^h)^T vec(\boldsymbol{Q}) + \beta vec(\boldsymbol{W}^h) vec(\boldsymbol{W}^h) \\ &- \sum_{p=1}^{m}\sum_{t=1}^{\tau} \boldsymbol{\lambda}_i^{(t)} \boldsymbol{W}_{pt}^h - \boldsymbol{\mu}(\sum_{p=1}^{m}\sum_{t=1}^{\tau} \boldsymbol{W}_{pt}^h - 1) \\ &s.t. \boldsymbol{W}^h \geq 0, \sum vec(\boldsymbol{W}_i^h) = 1\end{aligned} \tag{20}$$

Base on the KKT conditions, the optional $\boldsymbol{W}_{pt}^h$ should meet the following four conditions:
(i) Stationary condition: $\frac{\partial \tilde{\Omega}}{\partial \boldsymbol{W}^h} = \boldsymbol{Q} + 2\beta \boldsymbol{W}^h - \boldsymbol{\lambda} - \boldsymbol{\mu} = \mathbf{0}$
(ii) Feasible condition:$\boldsymbol{W}^h \geq 0, \sum_{i=1}^{m}\sum_{t=1}^{\tau} \boldsymbol{W}_{pt}^h - 1 = 0$
(iii) Dual feasibility:$\boldsymbol{\lambda}_i^{(t)} \geq 0, \forall \boldsymbol{\Theta}_i^{(t)}, t \leq max_i t_i$
(iv) Complementary slackness:$\boldsymbol{\lambda}_i^{(t)} \boldsymbol{W}^h \geq 0, \forall \boldsymbol{\Theta}_i^{(t)}, t \leq max_i t_i$

From the stationary condition, $\boldsymbol{W}_{pt}^h$ can be computed as follows:

$$\boldsymbol{W}_{pt}^h = \frac{\boldsymbol{\lambda}_i^{(t)} + \boldsymbol{\mu} - \boldsymbol{Q}_i^{(t)}}{2\beta} \tag{21}$$

We can find that $\boldsymbol{W}_{pt}^h$ depends on the specification of $\boldsymbol{\lambda}_i^{(t)}$ and $\boldsymbol{\mu}$ and the specification of $\boldsymbol{\lambda}_i^{(t)}$ and $\boldsymbol{\mu}$ can be analyzed in the following cases:

1) If $\boldsymbol{\mu} > \boldsymbol{Q}_i^{(t)}$, then $\boldsymbol{W}_{pt}^h > 0$; we get $\boldsymbol{\lambda}_i^{(t)} = 0$ by complementary slackness,then $\boldsymbol{W}_{pt}^h = \frac{\boldsymbol{\mu} - \boldsymbol{Q}_i^{(t)}}{\mathbf{2\beta}}$;
2) If $\boldsymbol{\mu} \leq \boldsymbol{Q}_i^{(t)}$, we get $\boldsymbol{W}_{pt}^h = 0$ because of $\boldsymbol{\lambda}_i^{(t)} \boldsymbol{W}_{pt}^h = 0$ and $\boldsymbol{W}_{pt}^h \geq 0, \boldsymbol{W}_{pt}^h \geq 0$,it requires $\boldsymbol{\lambda}_i^{(t)} > 0$;

From the above analysis, we can set $\boldsymbol{W}_{pt}^h$ as:

$$\boldsymbol{W_{pt}^h} = \begin{cases} \frac{\boldsymbol{\mu} - \boldsymbol{Q}_i^{(t)}}{2\beta}, if\ \boldsymbol{\mu} > \boldsymbol{Q}_i^{(t)} > 0\ and\ t \leq max_i t_i \\ \\ 0, if\ \boldsymbol{Q}_i^{(t)} \leq 0\ and\ t > max_i t_i \end{cases} \tag{22}$$

where $\boldsymbol{Q}_i^{(t)}$ is the number of cannot-link constraints for i-th type of objects in t-th intra-association matrix. Let $V_{\boldsymbol{Q}}$ store the entries of vector $vec(\boldsymbol{Q})$ in ascending order with entries corresponding to $\boldsymbol{\Theta}_i^{(t)}, t > max_i t_i$ removed. Accordingly, $\boldsymbol{V}_{\boldsymbol{W}^h}$ stores the corresponding entries of $vec(\boldsymbol{W}^h)$ with entries corresponding to $\boldsymbol{\Theta}_i^{(t)}, t > max_i t_i$ removed. For a not too big predefined β, there exists $x \in \{1, 2, \cdots, |\boldsymbol{\Theta}|\}$, $\boldsymbol{\Theta}$ is the number of intra-association matrices for all the types, with $V_{\boldsymbol{Q}}(x) < \boldsymbol{\mu}$ and $V_{\boldsymbol{Q}}(x+1) \geq \boldsymbol{\mu}$, $\sum \boldsymbol{V}_{\boldsymbol{Q}} = \sum_{\boldsymbol{V}_{\boldsymbol{Q}}(x) < \boldsymbol{\mu}} \frac{\boldsymbol{\mu} - \boldsymbol{V}_{\boldsymbol{Q}}(x)}{2\beta} = 1$. Then $\boldsymbol{V}_{\boldsymbol{W}^h}(x')$ has the following explicit solution:

Algorithm 2 Seek x and compute $\boldsymbol{V}_{\boldsymbol{W}^h}$

Input:$\boldsymbol{V}_{\boldsymbol{Q}}, \beta$
Output: $x, \boldsymbol{V}_{\boldsymbol{W}^h}$

```
Initialize x = |Θ|, μ = 0.
while x > 0 do
    update μ using (24).
    if μ − V_Q(x) > 0 then
        break.
    else
        x ← x − 1.
    end if
end while
V_{W^h}(x') ← (μ − V_Q(x'))/2β, for x' = 1, ··· , x;
V(x')) ← 0 ,for x' = x + 1, ··· , |Θ|;
return  x and V_{W^h}.
```

$$\boldsymbol{V}_{\boldsymbol{W}^h}(x') = \begin{cases} \frac{\boldsymbol{\mu} - V_{\boldsymbol{Q}}(x')}{2\beta}, if\ x' \leq x \\ \\ 0, if\ x' > x \end{cases} \tag{23}$$

From $\sum_{x'=1}^{|\Theta|} \boldsymbol{V}_{\boldsymbol{W}^h}(x') = \sum_{x'=1}^{x} \frac{\boldsymbol{\mu} - \boldsymbol{V}_{\boldsymbol{Q}}(x')}{2\beta} = 1$, we can get the value for $\boldsymbol{\mu}$ as:

$$\boldsymbol{\mu} = \frac{2\beta + \sum_{x'=1}^{x} \boldsymbol{V}_{\boldsymbol{Q}}(x')}{x} \tag{24}$$

We compute $\boldsymbol{V}_{W^h}$ via finding an appropriate q that satisfies $\boldsymbol{\mu} - \boldsymbol{V}_{\boldsymbol{Q}}(x) > 0$ and $\boldsymbol{\mu} - V_{\boldsymbol{Q}}(x+1) \leq 0$. We decrease x from $|\boldsymbol{\Theta}|$ to 1 step by step, and list the procedure in Algorithm 2.

From (24), we can see that if $V_{\boldsymbol{Q}}(x')$ is smaller than $V_{\boldsymbol{Q}}(x'')$ ($x \in \{1, 2, \cdots, |\boldsymbol{\Theta}|\}$) and $\boldsymbol{\mu} - V_{\boldsymbol{Q}}(x') \leq 0$ then a larger weight is assigned to the relational data source corresponding to $V_{\boldsymbol{Q}}(x')$ than $V_{\boldsymbol{Q}}(x'')$,because the data matrix of the former data source can be more well approximated than the latter. From (23), it can be also observed that if $x > h, \boldsymbol{V}_{\boldsymbol{W}^h} = 0$ which means that the intra-relational matrices are automatically removed from the optimization process. That is because these data sources have larger reconstruction losses. The overall process of M4L-JMF is summarized in **Algorithm 3**.

II. Convergence proof

The updating rule of $\boldsymbol{G}$, $\boldsymbol{S}$, $\boldsymbol{W^r}$and $\boldsymbol{W}^h$ will be converged to the global optimum. This proof follows the concept of auxiliary functions that are often used in the proof of convergence of approximate matrix factorization algorithms [1], [4], [5]. This kind of proof focus on an appropriate function F($\boldsymbol{G}$,$\boldsymbol{G'}$,$\boldsymbol{W}^r$,$\boldsymbol{W}^h$), which is an auxiliary function of the objective function satisfying Ω($\boldsymbol{G}$,$\boldsymbol{S}$,$\boldsymbol{W}^r$,$\boldsymbol{W}^h$) with:

$$\begin{aligned} F(\boldsymbol{G}, \boldsymbol{G'}, \boldsymbol{W}^r, \boldsymbol{W}^h) &= \Omega(\boldsymbol{G'}, \boldsymbol{S}, \boldsymbol{W}^r, \boldsymbol{W}^h), \\ F(\boldsymbol{G}, \boldsymbol{G'}, \boldsymbol{W}^r, \boldsymbol{W}^h) &\geq \Omega(\boldsymbol{G}, \boldsymbol{S}, \boldsymbol{W}^r, \boldsymbol{W}^h). \end{aligned} \tag{25}$$

Algorithm 3 M4L-JMF pseudo-code

Input: $|\mathcal{R}|$ relational matrices $\boldsymbol{R}_{ij}$; Intra-association data matrices $\boldsymbol{\Theta}^{(t)}$ for m object types; ranks d.
Output: $\boldsymbol{G}, \boldsymbol{S}, \boldsymbol{W}^r, \boldsymbol{W}^h$

```
Initialize V_{W^r} ← 1/|R|; V_{W^h} ← 1/|Θ|.
for i = 1:m do
    G_i = SVD(R_ij, d)
end for
R_cell = {G_1, G_2, ...G_m}
S_cell = {G_1, G_2, ...G_m}
while not satisfy the stop condition do
    for i = 1:m do
        Update S_cell{i} via (4).
        Update R_cell{i} via (8)-(10).
    end for
    Seek h and compute V_{W^r} using Algorithm1
    Seek x and compute V_{W^h} using Algorithm2
end while
return G, S, W^r, W^h.
```

If such an auxiliary function F can be found and if $\boldsymbol{G}$ is updated in $(n+1)$-th iteration as:

$$\boldsymbol{G}^{n+1} = arg\min_{\boldsymbol{G}} F(\boldsymbol{G}, \boldsymbol{G^n}, \boldsymbol{W}^r, \boldsymbol{W}^h) \tag{26}$$

and then we can get the following inequality:

$$\begin{aligned}\Omega(\boldsymbol{G}^{n+1}, \boldsymbol{S}, \boldsymbol{W}^r) &\leq F(\boldsymbol{G}^{n+1}, \boldsymbol{G}^n, \boldsymbol{W}^r, \boldsymbol{W}^h)\\ &\leq F(\boldsymbol{G}^n, \boldsymbol{G}^n, \boldsymbol{W}^r, \boldsymbol{W}^h)\\ &= \Omega(\boldsymbol{G}^n, \boldsymbol{S}, \boldsymbol{W}^r, \boldsymbol{W}^h)\end{aligned} \tag{27}$$

That is, $\Omega(\boldsymbol{G}, \boldsymbol{S}, \boldsymbol{W}^r, \boldsymbol{W}^h)$ would be nonincreasing with such an auxiliary function F. The steps we update $\boldsymbol{S}$ in (10) are proved in [1] as an proper auxiliary function, which is also specified by Wang *et al.* [6](Appendix II). constructed an auxiliary function as $F_{Wang}(\boldsymbol{A}, \boldsymbol{A}'; \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ and showed it satisfies the conditions of auxiliary functions with the form $\Omega(\boldsymbol{A}; \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}) = tr(-2\boldsymbol{A}^T\boldsymbol{B} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T) + tr(\boldsymbol{A}^T\boldsymbol{C}\boldsymbol{A}))$, where $\boldsymbol{C}$ and $\boldsymbol{D}$ are symmetric and $\boldsymbol{A}$ is nonnegative. Given that, we treat our objective function of M4L-JMF in (1) of the main text as a special case of $\Omega(\boldsymbol{A}; \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$.

Firstly, we view $\Omega(\boldsymbol{G}^n, \boldsymbol{S}, \boldsymbol{W}^r, \boldsymbol{W}^h)$ in (4) as a function of $\boldsymbol{G}_1$ and construct the auxiliary function $F_{Wang}(\boldsymbol{A}, \boldsymbol{A}'; \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ as follows:

$$\begin{aligned}&\boldsymbol{A} = \boldsymbol{G}_1,\\ &\boldsymbol{B} = \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} \boldsymbol{W}_{1j}^r \boldsymbol{R}_{1j} \boldsymbol{G}_j \boldsymbol{S}_{1j}^T + \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} \boldsymbol{W}_{i1}^r \boldsymbol{R}_{i1} \boldsymbol{G}_i \boldsymbol{S}_{i1}^T\\ &\boldsymbol{C} = \sum_{t=1}^{\tau} \boldsymbol{W}_{pt}^h \boldsymbol{\Theta}_1^{(t)}\\ &\boldsymbol{D} = \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} \boldsymbol{W}_{1j}^r \boldsymbol{S}_{1j} \boldsymbol{G}_j^T \boldsymbol{G}_j \boldsymbol{S}_{1j}^T + \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} \boldsymbol{W}_{i1}^r \boldsymbol{S}_{i1} \boldsymbol{G}_i^T \boldsymbol{G}_i \boldsymbol{S}_{i1}^T\end{aligned} \tag{28}$$

Then we rewrite (4) as:

$$\begin{aligned}\Omega(\boldsymbol{A}; \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}) = tr(-2\boldsymbol{A}^T\boldsymbol{B}^+ + 2\boldsymbol{A}^T\boldsymbol{B}^- + \boldsymbol{A}\boldsymbol{D}^+\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{D}^-\boldsymbol{A}^T)\\ + tr(\boldsymbol{A}^T\boldsymbol{C}^+\boldsymbol{A} - \boldsymbol{A}^T\boldsymbol{C}^-\boldsymbol{A})\end{aligned} \tag{29}$$

by ignoring $tr(\boldsymbol{X}^T\boldsymbol{X})$, and based on the theorem 6 in literature [7], we have:

$$\begin{aligned}tr(\boldsymbol{A}^T\boldsymbol{C}^+\boldsymbol{A} &\leq \sum_{ij} \frac{(\boldsymbol{C}^+\boldsymbol{A}')_{ij}\boldsymbol{A}_{ij}^2}{\boldsymbol{A}'_{ij}}\\ tr(\boldsymbol{A}^T\boldsymbol{D}^+\boldsymbol{A}^T &\leq \sum_{ij} \frac{(\boldsymbol{A}'\boldsymbol{D}^+)_{ij}\boldsymbol{A}_{ij}^2}{\boldsymbol{A}'_{ij}}\end{aligned} \tag{30}$$

By the inequality

$$a \leq (a^2 + b^2)/2b,\ for\ \forall a, b > 0 \tag{31}$$

We have

$$tr(\boldsymbol{A}^T\boldsymbol{B}^-) = \sum_{ij} \boldsymbol{B}_{ij}^- \boldsymbol{A}_{ij} \leq \sum_{ij} \boldsymbol{B}_{ij}^- \frac{\boldsymbol{A}_{ij}^2 + \boldsymbol{A}_{ij}'^2}{2\boldsymbol{A}'_{ij}} \tag{32}$$

To obtain the lower bounds for the remaining terms, we use inequality that $z \geq 1 + logz$, which holds for any $z > 0$, then

$$\begin{aligned}tr(\boldsymbol{A}^T\boldsymbol{B}^+) &\geq \sum_{ij} \boldsymbol{B}_{ij}^+ \boldsymbol{A}'_{ij}(1 + log\frac{\boldsymbol{A}_{ij}}{\boldsymbol{A}'_{ij}})\\ tr(\boldsymbol{A}^T\boldsymbol{C}^-\boldsymbol{A}) &\geq \sum_{ijk} \boldsymbol{C}_{jk}^- \boldsymbol{A}'_{ji}\boldsymbol{A}'_{ki}(1 + log\frac{\boldsymbol{A}_{ji}\boldsymbol{A}_{ki}}{\boldsymbol{A}'_{ji}\boldsymbol{A}'_{ki}})\\ tr(\boldsymbol{A}^T\boldsymbol{D}^-\boldsymbol{A}^T) &\geq \sum_{ijk} \boldsymbol{D}_{jk}^- \boldsymbol{A}'_{ij}\boldsymbol{A}'_{ik}(1 + log\frac{\boldsymbol{A}_{ij}\boldsymbol{A}_{ik}}{\boldsymbol{A}'_{ij}\boldsymbol{A}'_{ik}})\end{aligned} \tag{33}$$

By summing all the bounds, we can get $P(\boldsymbol{A}, \boldsymbol{A}')$, which significantly satisfies (i) $P(\boldsymbol{A}, \boldsymbol{A}') \geq \Omega(\boldsymbol{A}')$ (ii) $P(\boldsymbol{A}', \boldsymbol{A}') = \Omega(\boldsymbol{A}')$. To find the minimum of $P(\boldsymbol{A}, \boldsymbol{A}')$, wa take

$$\begin{aligned}\frac{\partial p(\boldsymbol{A}, \boldsymbol{A}')}{\partial \boldsymbol{A}_{ij}} &= -2\boldsymbol{B}_{ij}^+ \frac{\boldsymbol{A}'_{ij}}{\boldsymbol{A}_{ij}} - 2\boldsymbol{B}_{ij}^- \frac{\boldsymbol{A}_{ij}}{\boldsymbol{A}'_{ij}} + 2\frac{(\boldsymbol{A}'\boldsymbol{D}^+)_{ij}\boldsymbol{A}'_{ij}}{\boldsymbol{A}'_{ij}}\\ &- \frac{2(\boldsymbol{A}'\boldsymbol{D}^-)_{ij}\boldsymbol{A}'_{ij}}{\boldsymbol{A}_{ij}} + \frac{2(\boldsymbol{C}^+\boldsymbol{A}')_{ij}\boldsymbol{A}_{ij}}{\boldsymbol{A}'_{ij}} + \frac{2(\boldsymbol{C}^-\boldsymbol{A}')_{ij}\boldsymbol{A}'_{ij}}{\boldsymbol{A}_{ij}}\end{aligned} \tag{34}$$

and the Hessian matrix for $P(\boldsymbol{A}, \boldsymbol{A}')$

$$\frac{\partial P(\boldsymbol{A}, \boldsymbol{A}')}{\partial \boldsymbol{A}_{ij}\partial \boldsymbol{A}_{kl}} = \varphi_{ik}\varphi_{jl}\boldsymbol{\Psi}_{ij} \tag{35}$$

is a diagonal matrix with positive diagonal elements

$$\begin{aligned}\boldsymbol{\Psi}_{ij} &= \frac{(2\boldsymbol{B}^+ + \boldsymbol{A}'\boldsymbol{D}^- + \boldsymbol{C}\boldsymbol{A}')_{ij}\boldsymbol{A}'_{ij}}{\boldsymbol{A}_{ij}^2}\\ &+ \frac{(2\boldsymbol{B}^- + \boldsymbol{A}'\boldsymbol{D}^+ + \boldsymbol{C}\boldsymbol{A}')_{ij}}{\boldsymbol{A}'_{ij}}\end{aligned} \tag{36}$$

Thus $P(\boldsymbol{A},\boldsymbol{A'})$ is a convex function of $\boldsymbol{A}$, then we can obtain the global minimum of $P(\boldsymbol{A},\boldsymbol{A'})$ by setting $\partial P(\boldsymbol{A},\boldsymbol{A'})/\partial \boldsymbol{A}_{ij} = 0$. Thus, we can update $\boldsymbol{G}$ via (8),(9) and (10) We repeat this process by constructing the rematining $q-1$ auxiliary function by separately considering $\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$ as a function of matrix factors $\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_q$. From the theory of auxiliary functions, it then follows that the objective function ($\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$) of M4L-JMF is nonincreasing under the update rules for each of $\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_q$. Letting $\Omega(\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_q\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h) = \Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$, and $\boldsymbol{M} = \Omega(\boldsymbol{G}_1^0,\boldsymbol{G}_2^0,\cdots,\boldsymbol{G}_q^0,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$ we can get:

$$\begin{aligned}\boldsymbol{M} &\geq \Omega(\boldsymbol{G}_1^1,\boldsymbol{G}_2^0,\cdots,\boldsymbol{G}_q^0,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)\\ &\geq \Omega(\boldsymbol{G}_1^1,\boldsymbol{G}_2^1,\cdots,\boldsymbol{G}_q^0,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)\\ &\geq \Omega(\boldsymbol{G}_1^1,\boldsymbol{G}_2^1,\boldsymbol{G}_3^1,\cdots,\boldsymbol{G}_q^0,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)\\ &\geq \cdots\\ &\geq \Omega(\boldsymbol{G}_1^1,\boldsymbol{G}_2^1,\boldsymbol{G}_3^1,\cdots,\boldsymbol{G}_q^1,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)\end{aligned} \tag{37}$$

Since $\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$ is certainly bounded from below by zero and its nonincreasing property, the convergence is proved.

A. Convergence proof of M4L-JMF

The updating rule of $\boldsymbol{G}$, $\boldsymbol{S}$, $\boldsymbol{W}^r$and $\boldsymbol{W}^h$ will be converged to the global optimum. This proof follows the concept of auxiliary functions that are often used in the proof of convergence of approximate matrix factorization algorithms [1], [4], [5]. This kind of proof focus on an appropriate function F($\boldsymbol{G}$,$\boldsymbol{G'}$,$\boldsymbol{W}^r$,$\boldsymbol{W}^h$), which is an auxiliary function of the objective function satisfying $\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$ with:

$$\begin{aligned}F(\boldsymbol{G'},\boldsymbol{G'},\boldsymbol{W}^r,\boldsymbol{W}^h) = \Omega(\boldsymbol{G'},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h),\\ F(\boldsymbol{G},\boldsymbol{G'},\boldsymbol{W}^r,\boldsymbol{W}^h) \geq \Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h).\end{aligned} \tag{38}$$

If such an auxiliary function F can be found and if $\boldsymbol{G}$ is updated in $(n+1)$-th iteration as:

$$\boldsymbol{G}^{n+1} = arg\min_{\boldsymbol{G}} F(\boldsymbol{G},\boldsymbol{G^n},\boldsymbol{W}^r,\boldsymbol{W}^h) \tag{39}$$

and then we can get the following inequality:

$$\begin{aligned}\Omega(\boldsymbol{G}^{n+1},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h) &\leq F(\boldsymbol{G}^{n+1},\boldsymbol{G}^n,\boldsymbol{W}^r,\boldsymbol{W}^h)\\ &\leq F(\boldsymbol{G}^n,\boldsymbol{G}^n,\boldsymbol{W}^r,\boldsymbol{W}^h)\\ &= \Omega(\boldsymbol{G}^n,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)\end{aligned} \tag{40}$$

That is, $\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$ would be nonincreasing with such an auxiliary function F. The steps we update $\boldsymbol{S}$ in (4) are proved in [1] as an proper auxiliary function, which is also specified by Wang *et al.* [6](Appendix II). constructed an auxiliary function as $F_{Wang}(\boldsymbol{A},\boldsymbol{A'};\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$ and showed it satisfies the conditions of auxiliary functions with the form $\Omega(\boldsymbol{A};\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}) = tr(-2\boldsymbol{A}^T\boldsymbol{B} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T) + tr(\boldsymbol{A}^T\boldsymbol{C}\boldsymbol{A}))$, where $\boldsymbol{C}$ and $\boldsymbol{D}$ are symmetric and $\boldsymbol{A}$ is nonnegative. Given that, we treat our objective function of M4L-JMF in (3) of the main text as a special case of $\Omega(\boldsymbol{A};\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$.

Firstly, we view $\Omega(\boldsymbol{G}^n,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$ in (4) as a function of $\boldsymbol{G}_1$ and construct the auxiliary function$F_{Wang}(\boldsymbol{A},\boldsymbol{A'};\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$ as follows:

$$\begin{aligned}&\boldsymbol{A} = \boldsymbol{G}_1,\\ &\boldsymbol{B} = \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} \boldsymbol{W}_{1j}^r\boldsymbol{R}_{1j}\boldsymbol{G}_j\boldsymbol{S}_{1j}^T + \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} \boldsymbol{W}_{i1}^r\boldsymbol{R}_{i1}\boldsymbol{G}_i\boldsymbol{S}_{i1}^T\\ &\boldsymbol{C} = \sum_{t=1}^{\tau} \boldsymbol{W}_{pt}^h\boldsymbol{\Theta}_1^{(t)}\\ &\boldsymbol{D} = \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} \boldsymbol{W}_{1j}^r\boldsymbol{S}_{1j}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{1j}^T + \sum_{j:\boldsymbol{R}_{ij}\in\mathbb{R}} \boldsymbol{W}_{i1}^r\boldsymbol{S}_{i1}\boldsymbol{G}_i^T\boldsymbol{G}_i\boldsymbol{S}_{i1}^T\end{aligned} \tag{41}$$

Then we rewrite (4) as:

$$\begin{aligned}\Omega(\boldsymbol{A};\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}) = tr(-2\boldsymbol{A}^T\boldsymbol{B}^+ + 2\boldsymbol{A}^T\boldsymbol{B}^- + \boldsymbol{A}\boldsymbol{D}^+\boldsymbol{A}^T\\ - \boldsymbol{A}\boldsymbol{D}^-\boldsymbol{A}^T) + tr(\boldsymbol{A}^T\boldsymbol{C}^+\boldsymbol{A} - \boldsymbol{A}^T\boldsymbol{C}^-\boldsymbol{A})\end{aligned} \tag{42}$$

where the matrices with positive and negative symbols are defined as $\boldsymbol{B}^+ = \frac{|\boldsymbol{B}|+\boldsymbol{B}}{2}$ and $\boldsymbol{B}^- = \frac{|\boldsymbol{B}|-\boldsymbol{B}}{2}$ respectively. By ignoring $tr(\boldsymbol{X}^T\boldsymbol{X})$, and based on the theorem 6 in literature [7], we have:

$$\begin{aligned}tr(\boldsymbol{A}^T\boldsymbol{C}^+\boldsymbol{A} \leq \sum_{ij}\frac{(\boldsymbol{C}^+\boldsymbol{A'})_{ij}\boldsymbol{A}_{ij}^2}{\boldsymbol{A}'_{ij}}\\ tr(\boldsymbol{A}^T\boldsymbol{D}^+\boldsymbol{A}^T \leq \sum_{ij}\frac{(\boldsymbol{A'}\boldsymbol{D}^+)_{ij}\boldsymbol{A}_{ij}^2}{\boldsymbol{A}'_{ij}}\end{aligned} \tag{43}$$

By the inequality

$$a \leq (a^2+b^2)/2b,\ for\ \forall a,b>0 \tag{44}$$

We have

$$tr(\boldsymbol{A}^T\boldsymbol{B}^-) = \sum_{ij}\boldsymbol{B}_{ij}^-\boldsymbol{A}_{ij} \leq \sum_{ij}\boldsymbol{B}_{ij}^-\frac{\boldsymbol{A}_{ij}^2+\boldsymbol{A'}_{ij}^2}{2\boldsymbol{A}'_{ij}} \tag{45}$$

To obtain the lower bounds for the remaining terms, we use inequality that $Z \geq 1+logz$, which holds for any $z>0$, then

$$\begin{aligned}tr(\boldsymbol{A}^T\boldsymbol{B}^+) \geq \sum_{ij}\boldsymbol{B}_{ij}^+\boldsymbol{A}'_{ij}(1+log\frac{\boldsymbol{A}_{ij}}{\boldsymbol{A}'_{ij}})\\ tr(\boldsymbol{A}^T\boldsymbol{C}^-\boldsymbol{A}) \geq \sum_{ijk}\boldsymbol{C}_{jk}^-\boldsymbol{A}'_{ji}\boldsymbol{A}'_{ki}(1+log\frac{\boldsymbol{A}_{ji}\boldsymbol{A}_{ki}}{\boldsymbol{A}'_{ji}\boldsymbol{A}'_{ki}})\\ tr(\boldsymbol{A}^T\boldsymbol{D}^-\boldsymbol{A}^T) \geq \sum_{ijk}\boldsymbol{D}_{jk}^-\boldsymbol{A}'_{ij}\boldsymbol{A}'_{ik}(1+log\frac{\boldsymbol{A}_{ij}\boldsymbol{A}_{ik}}{\boldsymbol{A}'_{ij}\boldsymbol{A}'_{ik}})\end{aligned} \tag{46}$$

By summing all the bounds, we can get $P(\boldsymbol{A},\boldsymbol{A'})$, which significantly satisfies (i) $P(\boldsymbol{A},\boldsymbol{A'}) \geq \Omega(\boldsymbol{A'})$ (ii) $P(\boldsymbol{A'},\boldsymbol{A'}) = \Omega(\boldsymbol{A'})$. To find the minimum of $P(\boldsymbol{A},\boldsymbol{A'})$, wa take

$$\frac{\partial p(\boldsymbol{A},\boldsymbol{A}')}{\partial \boldsymbol{A}_{ij}} = -2\boldsymbol{B}^{+}_{ij}\frac{\boldsymbol{A}'_{ij}}{\boldsymbol{A}_{ij}} - 2\boldsymbol{B}^{-}_{ij}\frac{\boldsymbol{A}_{ij}}{\boldsymbol{A}'_{ij}} + 2\frac{(\boldsymbol{A}'\boldsymbol{D}^{+})_{ij}\boldsymbol{A}'_{ij}}{\boldsymbol{A}'_{ij}}$$
$$- \frac{2(\boldsymbol{A}'\boldsymbol{D}^{-})_{ij}\boldsymbol{A}'_{ij}}{\boldsymbol{A}_{ij}} + \frac{2(\boldsymbol{C}^{+}\boldsymbol{A}')_{ij}\boldsymbol{A}_{ij}}{\boldsymbol{A}'_{ij}} + \frac{2(\boldsymbol{C}^{-}\boldsymbol{A}')_{ij}\boldsymbol{A}'_{ij}}{\boldsymbol{A}_{ij}} \tag{47}$$

and the Hessian matrix for $P(\boldsymbol{A},\boldsymbol{A}')$

$$\frac{\partial P(\boldsymbol{A},\boldsymbol{A}')}{\partial \boldsymbol{A}_{ij}\partial \boldsymbol{A}_{kl}} = \varphi_{ik}\varphi_{jl}\boldsymbol{\Psi}_{ij} \tag{48}$$

is a diagonal matrix with positive diagonal elements

$$\begin{aligned}\boldsymbol{\Psi}_{ij} = &\frac{(2\boldsymbol{B}^{+} + \boldsymbol{A}'\boldsymbol{D}^{-} + \boldsymbol{C}\boldsymbol{A}')_{ij}\boldsymbol{A}'_{ij}}{\boldsymbol{A}^2_{ij}} \\ &+ \frac{(2\boldsymbol{B}^{+} + \boldsymbol{A}'\boldsymbol{D}^{+} + \boldsymbol{C}\boldsymbol{A}')_{ij}}{\boldsymbol{A}^2_{ij}}\end{aligned} \tag{49}$$

Thus $P(\boldsymbol{A},\boldsymbol{A}')$ is a convex function of $\boldsymbol{A}$, then we can obtain the global minimum of $P(\boldsymbol{A},\boldsymbol{A}')$ by setting $\partial P(\boldsymbol{A},\boldsymbol{A}')/\partial \boldsymbol{A}_{ij} = 0$. Thus, we can update $\boldsymbol{G}$ in (8),(9) and (10). We repeat this process by constructing the rematining $q-1$ auxiliary function by separately considering $\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$ as a function of matrix factors $\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_q$. From the theory of auxiliary functions, it then follows that the objective function ($\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$) of M4L-JMF is nonincreasing under the update rules for each of $\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_q$. Letting $\Omega(\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_q\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h) = \Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$, and $\boldsymbol{M} = \Omega(\boldsymbol{G}^0_1,\boldsymbol{G}^0_2,\cdots,\boldsymbol{G}^0_q,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$ we can get:

$$\begin{aligned}\boldsymbol{M} &\geq \Omega(\boldsymbol{G}^1_1,\boldsymbol{G}^0_2,\cdots,\boldsymbol{G}^0_q,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h) \\ &\geq \Omega(\boldsymbol{G}^1_1,\boldsymbol{G}^1_2,\cdots,\boldsymbol{G}^0_q,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h) \\ &\geq \Omega(\boldsymbol{G}^1_1,\boldsymbol{G}^1_2,\boldsymbol{G}^1_3,\cdots,\boldsymbol{G}^0_q,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h) \\ &\geq \cdots \\ &\geq \Omega(\boldsymbol{G}^1_1,\boldsymbol{G}^1_2,\boldsymbol{G}^1_3,\cdots,\boldsymbol{G}^1_q,\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)\end{aligned} \tag{50}$$

Since $\Omega(\boldsymbol{G},\boldsymbol{S},\boldsymbol{W}^r,\boldsymbol{W}^h)$ is certainly bounded from below by zero and its nonincreasing property, the convergence is proved.

References

[1] M. Zitnik and B. Zupan, "Data fusion by matrix factorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 41–53, 2015.

[2] G. Yu, Y. Wang, J. Wang, G. Fu, M. Guo, and C. Domeniconi, "Weighted matrix factorization based data fusion for predicting lncrna-disease associations," in *BIBM*, 2018, pp. 572–577.

[3] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," pp. 556–562, 2000.

[5] Q. Gu and J. Zhou, "Co-clustering on manifolds," pp. 359–368, 2009.

[6] F. Wang, T. Li, and C. Zhang, "Semi-supervised clustering via matrix factorization." pp. 1–12, 2008.

[7] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," pp. 126–135, 2006.