# Supplemental Materials: Multiplicative Sparse Tensor Factorization for Multi-View Multi-Task Learning

## 1 Proof of details for Section: Methodology

In Sec 4.3 of the main paper, we have the conclusion: $\mathcal{W}_t = \mathcal{G}_t \times_1 \mathbf{F}^{(1)} \times_2 \mathbf{F}^{(2)} \cdots \times_V \mathbf{F}^{(V)}$ in (7) of the main paper, $\mathbf{F}^{(v)} \in \mathbb{R}^{(d_v+1)\times r_v} = diag(\mathbf{a}^{(v)})\mathbf{C}^{(v)}$. Now we prove this here.

$$\begin{aligned}\mathcal{W}_t &= (\mathbf{a}^{(1)} \circ \cdots \circ \mathbf{a}^{(V)}) \odot (\mathcal{G}_t \times_1 \mathbf{C}^{(1)} \cdots \times_V \mathbf{C}^{(V)}) \\ &= \mathcal{G}_t \times_1 \mathbf{C}^{(1)} \cdots \times_V \mathbf{C}^{(V)} \times_1 diag(\mathbf{a}^{(1)}) \cdots \times_V diag(\mathbf{a}^{(V)}) \\ &= \mathcal{G}_t \times_1 (diag(\mathbf{a}^{(1)})\mathbf{C}^{(1)}) \cdots \times_V (diag(\mathbf{a}^{(V)})\mathbf{C}^{(V)}) \\ &= \mathcal{G}_t \times_1 \mathbf{F}^{(1)} \times_2 \mathbf{F}^{(2)} \cdots \times_V \mathbf{F}^{(V)}.\end{aligned}$$

The second equation holds because

$$(\mathbf{x}_1 \circ \cdots \circ \mathbf{x}_n) \odot \mathcal{Y} = \mathcal{Y} \times_1 diag(\mathbf{x}_1) \cdots \times_n diag(\mathbf{x}_n).$$

The third equation holds because

$$\mathcal{X} \times_n \mathbf{E} \times_n \mathbf{F} = \mathcal{X} \times_n (\mathbf{F}\mathbf{E}).$$

Then we can easily prove (7) in the main paper.

## 2 Proof of Theorem 1 in the main paper

We first give the general form objective function for vMSTF:

$$\min_{\mathcal{A}\geq 0, \mathcal{B}_t} \sum_{t=1}^{m} L(\mathcal{W}_t) + \lambda_1 \|\mathcal{A}\|_k^k + \lambda_2 \sum_{t=1}^{m} \|\mathcal{B}_t\|_p^p, \quad s.t.\ \mathcal{W}_t = \mathcal{A} \odot \mathcal{B}_t, \quad t = 1, 2, ..., m. \tag{1}$$

**Theorem 1** *Let $(\hat{\mathcal{A}}, \hat{\mathcal{B}}_t)$ be the optimal solution of problem (1) and $\hat{\mathcal{W}}$ be the optimal solution of the following optimization problem,*

$$\min_{\mathcal{W}} \sum_{t=1}^{m} L(\mathcal{W}_t) + \gamma \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} \sqrt{\|\mathbf{w}_{i_1 \ldots i_V :}\|_p^{p/q}}, \tag{2}$$

*where $\mathbf{w}_{i_1 \ldots i_V :} \in \mathbb{R}^m$ is the mode-$(V+1)$ fiber of $\mathcal{W} = [\mathcal{W}_1; ...; \mathcal{W}_m] \in \mathbb{R}^{(d_1+1)\times \cdots \times (d_V+1)\times m}$. When $\gamma = 2\sqrt{\lambda_1^{p/kq} \lambda_2^{2-p/kq}}$ and $q = (k+p)/2k$, we have $\hat{\mathcal{W}}_t = \hat{\mathcal{A}} \odot \hat{\mathcal{B}}_t$. In addition, task-shared $\hat{\mathcal{A}}$ is related to task-specific $\hat{\mathcal{B}}_t$ by:*

$$(\mathcal{A})_{i_1 \ldots i_V} = (\frac{\lambda_2}{\lambda_1})^{\frac{1}{k}} \sqrt[k]{\sum_{t=1}^{m} |(\mathcal{B}_t)_{i_1 \ldots i_V}|^p}, \quad \forall i_1, ..., i_V. \tag{3}$$

**proof** In order to prove (1) is equivalent to (2), we need to introduce a new objective function (4) and Lemmas 1 and 2. In (4), $\Sigma \in \mathbb{R}^{(d_1+1)\times \cdots \times (d_V+1)}$. Lemma 1 shows that the optimal solution of (2) is equivalent to the optimal solution of (4), when we choose proper values of $\gamma, \mu_1$ and $\mu_2$. Lemma 2 shows that the optimal solution of (1) is equivalent to the optimal solution of (4), under some circumstance. Then we can use (4) as a bridge to prove that (1) is equivalent to (2).

$$\min_{\mathcal{W}, \Sigma \geq 0} \sum_{t=1}^{m} L(\mathcal{W}_t) + \mu_1 \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} \sigma_{i_1 \ldots i_V} + \mu_2 \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} (\sigma_{i_1 \ldots i_V})^{-1} \|\mathbf{w}_{i_1 \ldots i_V :}\|_p^{p/q}. \tag{4}$$

**Lemma 1** *When $\gamma = 2\sqrt{\mu_1\mu_2}$, the optimal solution of (2) is equivalent to the optimal solution of (4).*

**proof** If we apply Cauchy-Schwarz inequality to problem (4), we can get

$$\begin{aligned} & \sum_{t=1}^{m} L(\mathcal{W}_t) + \mu_1 \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} \sigma_{i_1 \ldots i_V} + \mu_2 \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} (\sigma_{i_1 \ldots i_V})^{-1} \|\mathbf{w}_{i_1 \ldots i_V:}\|_p^{p/q} \\ \geq & \sum_{t=1}^{m} L(\mathcal{W}_t) + 2\sqrt{\mu_1\mu_2} \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} \sqrt{\|\mathbf{w}_{i_1 \ldots i_V:}\|_p^{p/q}}. \end{aligned} \tag{5}$$

The equality holds when $\sigma_{i_1 \ldots i_V} = \mu_1^{-1/2}\mu_2^{1/2}\sqrt{\|\mathbf{w}_{i_1 \ldots i_V:}\|_p^{p/q}}$. No matter what values $\|\mathbf{w}_{i_1 \ldots i_V:}\|_p^{p/q}$ takes, as long as $\sigma_{i_1 \ldots i_V} = \mu_1^{-1/2}\mu_2^{1/2}\sqrt{\|\mathbf{w}_{i_1 \ldots i_V:}\|_p^{p/q}}$, we can reach the conclusion that the equation in (5) is the lower bound of the objective in (4). As a result, when $\gamma = 2\sqrt{\mu_1\mu_2}$, (4) equals to (2). So the optimal solution of (2) is equivalent to the optimal solution of (4) when $\gamma = 2\sqrt{\mu_1\mu_2}$.

**Lemma 2** *Let $(\hat{\mathcal{A}}, \hat{\mathcal{B}}_t)$ be the optimal solution of (1) and $(\hat{\mathcal{W}}, \hat{\Sigma})$ be the optimal solution of (4). They are equivalent to each other with $\hat{\mathcal{W}}_t = \hat{\mathcal{A}} \odot \hat{\mathcal{B}}_t$ and $\hat{\sigma}_{i_1 \ldots i_V} = ((\hat{\mathcal{A}})_{i_1 \ldots i_V})^k$, when $\lambda_1 = \mu_1, \lambda_2 = \mu_1^{\frac{kq-p}{2kq-p}} \mu_2^{\frac{kq}{2kq-p}}$ and $k = p/(2q-1)$.*

**proof**

- First, we show if $\hat{\mathcal{W}}$ and $\hat{\Sigma}$ optimize (4), then $(\hat{\mathcal{A}})_{i_1 \ldots i_V} = \sqrt[k]{\hat{\sigma}_{i_1 \ldots i_V}}$ and $\hat{\mathcal{B}}_t$, where $\hat{\mathcal{B}}_t$ satisfies the condition $\hat{\mathcal{W}}_t = \hat{\mathcal{A}} \odot \hat{\mathcal{B}}_t$, optimize (1) when $\lambda_1 = \mu_1, \lambda_2 = \mu_1^{\frac{kq-p}{2kq-p}} \mu_2^{\frac{kq}{2kq-p}}$ and $k = p/(2q-1)$.

  Substitute $\hat{W}$ and $\hat{\Sigma}$ with $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}$, we can get the optimal value of (4) is

$$\sum_{t=1}^{m} L(\mathcal{W}_t) + \mu_1 \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} ((\hat{\mathcal{A}})_{i_1 \ldots i_V})^k + \mu_2 \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} ((\hat{\mathcal{A}})_{i_1 \ldots i_V})^{\frac{p-kq}{q}} \|\hat{\mathbf{b}}_{i_1 \ldots i_V:}\|_p^{p/q}, \tag{6}$$

  where $\hat{\mathbf{b}}_{i_1 \ldots i_V:} \in \mathbb{R}^m$ is the mode-$(V+1)$ fiber of $\hat{\mathcal{B}}$ and $\hat{\mathcal{B}} = [\hat{\mathcal{B}}_1; \ldots; \hat{\mathcal{B}}_m] \in \mathbb{R}^{(d_1+1) \times \cdots \times (d_V+1) \times m}$. From Lemma 1, we know that $\hat{\sigma}_{i_1 \ldots i_V} = \mu_1^{-1/2}\mu_2^{1/2}\sqrt{\|\hat{\mathbf{w}}_{i_1 \ldots i_V:}\|_p^{p/q}}$ is the condition when (4) reaches the optimal value. Substitute $\hat{\mathcal{W}}$ and $\hat{\Sigma}$ with $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}$, the above condition becomes

$$(\hat{\mathcal{A}})_{i_1 \ldots i_V} = (\mu_1^{-1}\mu_2 \|\hat{\mathbf{b}}_{i_1 \ldots i_V:}\|_p^{p/q})^{\frac{q}{2kq-p}}. \tag{7}$$

  Then put (7) into (6) and substitute $\mu_1$ and $\mu_2$ with $\lambda_1$ and $\lambda_2$ using the conditions mentioned in Lemma 2, we can get the optimal value of (4) is

$$\sum_{t=1}^{m} L(\mathcal{W}_t) + \mu_1 \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} ((\hat{\mathcal{A}})_{i_1 \ldots i_V})^k + \lambda_2 \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} \|\hat{\mathbf{b}}_{i_1 \ldots i_V:}\|_p^p. \tag{8}$$

  We can find that (8) is equivalent to (1), which means if $\hat{\mathcal{W}}$ and $\hat{\Sigma}$ optimize problem (4), then $(\hat{\mathcal{A}})_{i_1 \ldots i_V} = \sqrt[k]{\hat{\sigma}_{i_1 \ldots i_V}}$ and $\hat{\mathcal{B}}_t$ optimize (1), when $\lambda_1 = \mu_1, \lambda_2 = \mu_1^{\frac{kq-p}{2kq-p}} \mu_2^{\frac{kq}{2kq-p}}$ and $k = p/(2q-1)$.

- Then we show if $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}_t$ optimize (1), then $\hat{\mathcal{W}}_t = \hat{\mathcal{A}} \odot \hat{\mathcal{B}}_t$ and $\hat{\sigma}_{i_1 \ldots i_V} = ((\hat{\mathcal{A}})_{i_1 \ldots i_V})^k$ optimize (4), when $\lambda_1 = \mu_1, \lambda_2 = \mu_1^{\frac{kq-p}{2kq-p}} \mu_2^{\frac{kq}{2kq-p}}$ and $k = p/(2q-1)$.

  By substituting $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}_t$ with $\hat{\mathcal{W}}$ and $\hat{\Sigma}$, we can get the optimal value of (1):

$$\sum_{t=1}^{m} L(\mathcal{W}_t) + \lambda_1 \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} \hat{\sigma}_{i_1 \ldots i_V} + \lambda_2 \sum_{i_1=1}^{d_1+1} \cdots \sum_{i_V=1}^{d_V+1} \|\hat{\mathbf{w}}_{i_1 \ldots i_V:}\|_p^p (\hat{\sigma}_{i_1 \ldots i_V})^{-p/k}. \tag{9}$$

Apply Cauchy-Schwarz inequality to (9), we find that $\hat{\sigma}_{i_1\dots i_V} = (\lambda_1^{-1}\lambda_2)^{\frac{k}{k+p}}(\|\hat{\mathbf{w}}_{i_1\dots i_V:}\|_p^p)^{\frac{k}{k+p}}$ is the condition when (9) reaches the optimal value. Then put this condition back into (9) and substitute $\lambda_1$ and $\lambda_2$ with $\mu_1$ and $\mu_2$ using the conditions mentioned in **Lemma 2**, we can get the final equation which is equivalent to (4). This means if $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}_t$ optimize (1), then $\hat{\mathcal{W}}_t = \hat{\mathcal{A}} \odot \hat{\mathcal{B}}_t$ and $\hat{\sigma}_{i_1\dots i_V} = ((\hat{\mathcal{A}})_{i_1\dots i_V})^k$ optimize (4), when $\lambda_1 = \mu_1, \lambda_2 = \mu_1^{\frac{kq-p}{2kq-p}}\mu_2^{\frac{kq}{2kq-p}}$ and $k = p/(2q-1)$.

In short, after introducing Lemmas 1 and 2, we successfully use (4) as a bridge to prove the equivalence between (1) and (2). In order to prove (3) in **Theorem 1**, we substitute $\mu_1$ and $\mu_2$ with $\lambda_1$ and $\lambda_2$ in (7) using the conditions mentioned in Lemma 2, and we have

$$(\hat{\mathcal{A}})_{i_1\dots i_V} = (\frac{\lambda_2}{\lambda_1})^{\frac{1}{k}}(\|\hat{\mathbf{b}}_{i_1\dots i_V:}\|_p^{\frac{p}{2kq-p}}). \tag{10}$$

Since we also have $q = (k+p)/2k$, we can obtain (3) by combining it with (10). In this way, **Theorem 1** is proved.

# 3 Proof of Theorem 2 in the main paper

We present the general form objective function for MSTF:

$$\begin{aligned}&\min_{\mathbf{a}^{(v)}\geq 0, \mathbf{C}^{(v)}, \mathcal{G}_t} \sum_{t=1}^{m} L(\mathcal{W}_t) + \sum_{v=1}^{V}(\lambda_1\|\mathbf{a}^{(v)}\|_k^k + \lambda_2\|\mathbf{C}^{(v)}\|_p^p) + \lambda_3\sum_{t=1}^{m}\|\mathcal{G}_t\|_F^2,\\ &s.t. \mathcal{W}_t = \mathcal{G}_t \times_1 \mathbf{F}^{(1)} \cdots \times_V \mathbf{F}^{(V)}, \mathbf{F}^{(v)} = diag(\mathbf{a}^{(v)})\mathbf{C}^{(v)},\ \forall t, v.\end{aligned} \tag{11}$$

**Theorem 2** *Let $(\hat{\mathbf{a}}^{(v)}, \hat{\mathbf{C}}^{(v)}, \hat{\mathcal{G}}_t)$ be the optimal solution of problem (11) and $(\hat{\mathbf{F}}^{(v)}, \hat{\mathcal{G}}_t)$ be the optimal solution of the following problem,*

$$\min_{\mathbf{F}^{(v)}, \mathcal{G}_t} \sum_{t=1}^{m} L(\mathcal{W}_t) + \gamma\sum_{v=1}^{V}\sum_{i=1}^{d_v+1}\sqrt{\|(\mathbf{F}^{(v)})_{i:}\|_p^{p/q}} + \lambda_3\sum_{t=1}^{m}\|\mathcal{G}_t\|_F^2, \quad s.t.\ \mathcal{W}_t = \mathcal{G}_t \times_1 \mathbf{F}^{(1)} \times_2 \mathbf{F}^{(2)} \cdots \times_V \mathbf{F}^{(V)},\ \forall t, \tag{12}$$

*where $(\mathbf{F}^{(v)})_{i:} \in \mathbb{R}^{r_v}$ represents the i-th row of $\mathbf{F}^{(v)}$. When $\gamma = 2\sqrt{\lambda_1^{p/kq}\lambda_2^{2-p/kq}}$ and $q = (k+p)/2k$, we have $\mathbf{F}^{(v)} = diag(\mathbf{a}^{(v)})\mathbf{C}^{(v)}$. In addition, $\hat{\mathbf{a}}^{(v)}$ is related to $\hat{\mathbf{C}}^{(v)}$ according to:*

$$(\mathbf{a}^{(v)})_i = (\frac{\lambda_2}{\lambda_1})^{\frac{1}{k}}\sqrt[k]{\sum_{j=1}^{r_v}|(\mathbf{C}^{(v)})_{ij}|^p}, \quad \forall i. \tag{13}$$

**proof** In order to prove (11) is equivalent to (12), we need to introduce a new objective function (14) and Lemmas 3 and 4. In (14), $\sigma^{(v)} \in \mathbb{R}^{(d_v+1)}, \forall v$. Lemma 3 will show that the optimal solution of (12) is equivalent to the optimal solution of (14) when we choose proper values of $\gamma, \mu_1, \mu_2$. Lemma 4 will show that the optimal solution of (11) is equivalent to the optimal solution of (14) under some circumstance. Then we can use (14) as a bridge to prove that (11) is equivalent to (12).

$$\begin{aligned}&\min_{\mathbf{F}^{(v)}, \mathcal{G}_t, \Sigma^{(v)}\geq 0} \sum_{t=1}^{m} L(\mathcal{W}_t) + \sum_{v=1}^{V}[\mu_1\sum_{i=1}^{d_v+1}\sigma_i^{(v)} + \mu_2\sum_{i=1}^{d_v+1}(\sigma_i^{(v)})^{-1}\|(\mathbf{F}^{(v)})_{i:}\|_p^{p/q}] + \lambda_3\sum_{t=1}^{m}\|\mathcal{G}_t\|_F^2,\\ &s.t.\ \mathcal{W}_t = \mathcal{G}_t \times_1 \mathbf{F}^{(1)} \times_2 \mathbf{F}^{(2)} \cdots \times_V \mathbf{F}^{(V)},\ \forall t,\end{aligned} \tag{14}$$

**Lemma 3** *When $\gamma = 2\sqrt{\mu_1\mu_2}$, the optimal solution of (12) is equivalent to the optimal solution of (14).*

**proof** If we apply Cauchy-Schwarz inequality to (14), we can get

$$\sum_{t=1}^{m} L(\mathcal{W}_t) + \sum_{v=1}^{V}[\mu_1 \sum_{i=1}^{d_v+1} \sigma_i^{(v)} + \mu_2 \sum_{i=1}^{d_v+1} (\sigma_i^{(v)})^{-1} \|(\mathbf{F}^{(v)})_{i:}\|_p^{p/q}] + \lambda_3 \sum_{t=1}^{m} \|\mathcal{G}_t\|_F^2$$
$$\geq \sum_{t=1}^{m} L(\mathcal{W}_t) + \sum_{v=1}^{V} 2\sqrt{\mu_1\mu_2} \sum_{i=1}^{d_v+1} \sqrt{\|(\mathbf{F}^{(v)})_{i:}\|_p^{p/q}} + \lambda_3 \sum_{t=1}^{m} \|\mathcal{G}_t\|_F^2. \tag{15}$$

The equality holds when $\sigma_i^{(v)} = \mu_1^{-1/2}\mu_2^{1/2}\sqrt{\|(\mathbf{F}^{(v)})_{i:}\|_p^{p/q}}$. No matter what values $\|(\mathbf{F}^{(v)})_{i:}\|_p^{p/q}$ takes, as long as we let $\sigma_i^{(v)} = \mu_1^{-1/2}\mu_2^{1/2}\sqrt{\|(\mathbf{F}^{(v)})_{i:}\|_p^{p/q}}$, we can get (15) is the lower bound of the objective value of (14). As a result, when $\gamma = 2\sqrt{\mu_1\mu_2}$, (14) becomes the same as (12). So the optimal solution of (12) is equivalent to the optimal solution of (14) when $\gamma = 2\sqrt{\mu_1\mu_2}$.

**Lemma 4** *Let $(\hat{\mathbf{a}}^{(v)}, \hat{\mathbf{C}}^{(v)}, \hat{\mathcal{G}}_t)$ be the optimal solution of (11), and $(\hat{\mathbf{F}}^{(v)}, \hat{\mathcal{G}}_t, \hat{\sigma}^{(v)})$ be the optimal solution of (14). They are equivalent to each other where $\hat{\mathbf{F}}^{(v)} = diag(\hat{\mathbf{a}}^{(v)})\hat{\mathbf{C}}^{(v)}$ and $\hat{\sigma}_i^{(v)} = ((\mathbf{a}^{(v)})_i)^k$ when $\lambda_1 = \mu_1, \lambda_2 = \mu_1^{\frac{kq-p}{2kq-p}}\mu_2^{\frac{kq}{2kq-p}}$ and $k = p/(2q-1)$.*

**proof**

- First, we show if $\hat{\mathbf{F}}^{(v)}$, $\hat{\mathcal{G}}_t$ and $\hat{\sigma}^{(v)}$ optimize (14), then $(\hat{\mathbf{a}}^{(v)})_i = \sqrt[k]{\hat{\sigma}_i^{(v)}}$, $\hat{\mathbf{C}}^{(v)}$, where $\hat{\mathbf{C}}^{(v)}$ satisfies the condition $\hat{\mathbf{F}}^{(v)} = diag(\hat{\mathbf{a}}^{(v)})\hat{\mathbf{C}}^{(v)}$ and $\hat{\mathcal{G}}_t$ optimize (11) when $\lambda_1 = \mu_1, \lambda_2 = \mu_1^{\frac{kq-p}{2kq-p}}\mu_2^{\frac{kq}{2kq-p}}$ and $k = p/(2q-1)$.

  By substituting $\hat{\mathbf{F}}^{(v)}$ and $\hat{\sigma}^{(v)}$ with $\hat{\mathbf{a}}^{(v)}$ and $\hat{\mathbf{C}}^{(v)}$, we can get the optimal value of (14) is

$$\sum_{t=1}^{m} L(\mathcal{W}_t) + \sum_{v=1}^{V}[\mu_1 \sum_{i=1}^{d_v+1} ((\hat{\mathbf{a}}^{(v)})_i)^k + \mu_2 \sum_{i=1}^{d_v+1} ((\hat{\mathbf{a}}^{(v)})_i)^{\frac{p-kq}{q}} \|(\hat{\mathbf{C}}^{(v)})_{i:}\|_p^{p/q}] + \lambda_3 \sum_{t=1}^{m} \|\hat{\mathcal{G}}_t\|_F^2, \tag{16}$$

  From **Lemma 3**, we know that $\hat{\sigma}_i^{(v)} = \mu_1^{-1/2}\mu_2^{1/2}\sqrt{\|(\hat{\mathbf{F}}^{(v)})_{i:}\|_p^{p/q}}$ is the condition when (14) reaches the optimal value. Substitute $\hat{\mathbf{F}}^{(v)}$ and $\hat{\sigma}^{(v)}$ with $\hat{\mathbf{a}}^{(v)}$ and $\hat{\mathbf{C}}^{(v)}$, the above condition becomes

$$(\hat{\mathbf{a}}^{(v)})_i = (\mu_1^{-1}\mu_2 \|(\hat{\mathbf{C}}^{(v)})_{i:}\|_p^{p/q})^{\frac{q}{2kq-p}}. \tag{17}$$

  Then put (17) into (16) and substitute $\mu_1$ and $\mu_2$ with $\lambda_1$ and $\lambda_2$ using the conditions mentioned in **Lemma 4**, we can get the optimal value of (14) is

$$\sum_{t=1}^{m} L(\mathcal{W}_t) + \sum_{v=1}^{V}[\lambda_1 \sum_{i=1}^{d_v+1} ((\hat{\mathbf{a}}^{(v)})_i)^k + \lambda_2 \sum_{i=1}^{d_v+1} \|(\hat{\mathbf{C}}^{(v)})_{i:}\|_p^{p}] + \lambda_3 \sum_{t=1}^{m} \|\hat{\mathcal{G}}_t\|_F^2. \tag{18}$$

  We can find that (18) is equivalent to (11), which means if $\hat{\mathbf{F}}^{(v)}$, $\hat{\mathcal{G}}_t$ and $\hat{\sigma}^{(v)}$ optimize problem (14), then $(\hat{\mathbf{a}}^{(v)})_i = \sqrt[k]{\hat{\sigma}_i^{(v)}}$, $\hat{\mathbf{C}}^{(v)}$, where $\hat{\mathbf{C}}^{(v)}$ satisfies the condition $\hat{\mathbf{F}}^{(v)} = diag(\hat{\mathbf{a}}^{(v)})\hat{\mathbf{C}}^{(v)}$ and $\hat{\mathcal{G}}_t$ optimize problem (11) when $\lambda_1 = \mu_1, \lambda_2 = \mu_1^{\frac{kq-p}{2kq-p}}\mu_2^{\frac{kq}{2kq-p}}$ and $k = p/(2q-1)$.

- Then we show if $\hat{\mathbf{a}}^{(v)}$, $\hat{\mathbf{C}}^{(v)}$ and $\hat{\mathcal{G}}_t$ optimize (11), then $\hat{\mathbf{F}}^{(v)} = diag(\hat{\mathbf{a}}^{(v)})\hat{\mathbf{C}}^{(v)}$ and $\hat{\sigma}_i^{(v)} = ((\mathbf{a}^{(v)})_i)^k$ optimize (14), when $\lambda_1 = \mu_1, \lambda_2 = \mu_1^{\frac{kq-p}{2kq-p}}\mu_2^{\frac{kq}{2kq-p}}$ and $k = p/(2q-1)$.

  Substitute $\hat{\mathbf{a}}^{(v)}$ and $\hat{\mathbf{C}}^{(v)}$ with $\hat{\mathbf{F}}^{(v)}$ and $\hat{\sigma}^{(v)}$, we can get the optimal value of (11) is

$$\sum_{t=1}^{m} L(\mathcal{W}_t) + \sum_{v=1}^{V}[\lambda_1 \sum_{i=1}^{d_v+1} \hat{\sigma}_i^{(v)} + \lambda_2 \sum_{i=1}^{d_v+1} \|(\hat{\mathbf{F}}^{(v)})_{i:}\|_p^{p} (\hat{\sigma}_i^{(v)})^{-p/k}] + \lambda_3 \sum_{t=1}^{m} \|\hat{\mathcal{G}}_t\|_F^2. \tag{19}$$

  Apply Cauchy-Schwarz inequality to (19), we can get $\hat{\sigma}_i^{(v)} = (\lambda_1^{-1}\lambda_2)^{\frac{k}{k+p}} (\|(\hat{\mathbf{F}}^{(v)})_{i:}\|_p^{p})^{\frac{k}{k+p}}$ is the condition when (19) reaches the optimal value. Then put this condition back into (19) and substitute $\lambda_1$ and $\lambda_2$ with

$\mu_1$ and $\mu_2$ using the conditions mentioned in **Lemma 4**, we can get the final equation which is equivalent to (14). This means if $\hat{\mathbf{a}}^{(v)}$, $\hat{\mathbf{C}}^{(v)}$ and $\hat{\mathcal{G}}_t$ optimize (11), then $\hat{\mathbf{F}}^{(v)} = diag(\hat{\mathbf{a}}^{(v)})\hat{\mathbf{C}}^{(v)}$ and $\hat{\sigma}_i^{(v)} = ((\mathbf{a}^{(v)})_i)^k$ optimize (14), when $\lambda_1 = \mu_1, \lambda_2 = \mu_1^{\frac{kq-p}{2kq-p}}\mu_2^{\frac{kq}{2kq-p}}$ and $k = p/(2q-1)$.

In short, after introducing Lemmas 3 and 4, we successfully use (14) as a bridge to prove that (11) is equivalent to (12). In order to prove (13) in **Theorem 2**, we substitute $\mu_1$ and $\mu_2$ with $\lambda_1$ and $\lambda_2$ in (17) using the conditions mentioned in Lemma 4, and we have

$$(\hat{\mathbf{a}}^{(v)})_i = (\frac{\lambda_2}{\lambda_1})^{\frac{1}{k}}(\|(\hat{\mathbf{C}}^{(v)})_{i:}\|_p^{\frac{p}{2kq-p}}). \tag{20}$$

Since we also have $q = (k+p)/2k$, then put it into (20), we can get (13). In this way, **Theorem 2** is proved.

# 4 Algorithm for vMSTF for Section: Optimization

The objective function for vMSTF is shown in (1), which is convex w.r.t $\mathcal{A}$ and $\mathcal{B}_t$, respectively, and so we can use an alternating algorithm to solve it. Least squared loss is used here and the algorithm for other losses can be easily extended. For clarity, we reformulate the loss in a matrix form:

$$\sum_{t=1}^{m} \frac{1}{n_t}\|\mathbf{y}_t - \mathbf{X}_t\mathbf{w}_t\|_2^2. \tag{21}$$

Here $\mathbf{X}_t$ with the $i$-th row being $(vec(\mathcal{X}_{t,i}))^T$, and $\mathbf{w}_t = vec(\mathcal{W}_t) = diag(\mathbf{a}) \cdot \mathbf{b}_t$, where $\mathbf{a} = vec(\mathcal{A})$ and $\mathbf{b}_t = vec(\mathcal{B}_t)$. The algorithm repeats following two steps until convergence.

1. Update $\mathcal{B}_t$ with fixed $\mathcal{A}$. It can be separately optimized for each single task and the subproblem is:

$$\min_{\mathbf{b}_t} \frac{1}{n_t}\|\mathbf{y}_t - \mathbf{X}_t \cdot diag(\mathbf{a}) \cdot \mathbf{b}_t\|_2^2 + \lambda_2\|\mathbf{b}_t\|_p^p, \tag{22}$$

   When $p = 1$, (22) is lasso which can be solved by accelerated proximal gradient descent [Nesterov, 2013]. When $p = 2$, it is ridge regression which is solved by a closed-form solution, i.e.,

$$\mathbf{b}_t = (\frac{1}{n_t}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} + 2\lambda_2\mathbf{I})^{-1} \cdot (\frac{1}{n_t}\widetilde{\mathbf{X}}\mathbf{y}_t),$$

   where $\widetilde{\mathbf{X}} = \mathbf{X}_t \cdot diag(\mathbf{a})$.

2. Update $\mathcal{A}$ with fixed $\mathcal{B}_t$. According to Table 2 in the main paper, $\mathcal{A}$ can be solved by a closed-form solution.

For vMSTF, the time complexity updating $\mathcal{B}$ is $\mathcal{O}(mn(\prod_v(d_v+1)))$, where $n = \sum_t n_t$, and the time complexity of calculating $\mathcal{A}$ is $\mathcal{O}(m(\prod_v(d_v+1)))$. Therefore, the total time complexity of each iteration is $\mathcal{O}(mn(\prod_v(d_v+1)))$, which is linear in the number of samples and tasks.

# 5 Proof of details for Section: Optimization

In 6.1 of the main paper, when we update $\mathbf{C}^{(v)}$ with fixed $\mathbf{a}^{(v)}$ and $\mathcal{G}_t$ in MSTF, we have

$$\min_{\mathbf{C}^{(v)}} \sum_{t=1}^{m} \frac{1}{n_t}\|\mathbf{y}_t - \mathbf{X}_t\mathbf{w}_t\|_2^2 + \lambda_2\|vec(\mathbf{C}^{(v)})\|_p^p, \tag{23}$$

where $\mathbf{X}_t$ denotes the data matrix of the $t$-th task with the $i$-th row being $vec((\mathbf{X}_{t,i})_{(v)})^T$, and $\mathbf{w}_t = vec((\mathbf{W}_t)_{(v)})$. Let $(\mathbf{X}_{t,i})_{(v)}$, $(\mathbf{W}_t)_{(v)}$ and $(\mathbf{G}_t)_{(v)}$ be the mode-$v$ matricization of tensor $\mathcal{X}_{t,i}$, $\mathcal{W}_t$ and $\mathcal{G}_t$, respectively. Since $(\mathbf{W}_t)_{(v)} = \mathbf{F}^{(v)}(\mathbf{G}_t)_{(v)}(\mathbf{F}^{(V)} \cdots \otimes \mathbf{F}^{(v+1)} \otimes \mathbf{F}^{(v-1)} \cdots \otimes \mathbf{F}^{(1)})^T = \mathbf{F}^{(v)}(\mathbf{G}_t)_{(v)}(\mathbf{F}^{(-v)})^T$, where $\mathbf{F}^{(-v)} = \mathbf{F}^{(V)} \cdots \otimes \mathbf{F}^{(v+1)} \otimes \mathbf{F}^{(v-1)} \cdots \otimes \mathbf{F}^{(1)}$, then we have

$$\begin{aligned}
\mathbf{w}_t &= vec(\mathbf{F}^{(v)}(\mathbf{G}_t)_{(v)}(\mathbf{F}^{(-v)})^T) \\
&= vec(diag(\mathbf{a}^{(v)})\mathbf{C}^{(v)}(\mathbf{G}_t)_{(v)}(\mathbf{F}^{(-v)})^T) \\
&= [(\mathbf{F}_t^{(-v)}(\mathbf{G}_{(v)})^T) \otimes diag(\mathbf{a}^{(v)})] \cdot vec(\mathbf{C}^{(v)})
\end{aligned}$$

The third equation holds because

$$vec(\mathbf{ZXY}) = (\mathbf{Y}^T \otimes \mathbf{Z}) \cdot vec(\mathbf{X}).$$

Then we get the subproblem w.r.t. $\mathbf{C}^{(v)}$:

$$\min_{\mathbf{C}^{(v)}} \sum_{t=1}^{m} \frac{1}{n_t} \|\mathbf{y}_t - \mathbf{X}_t[(\mathbf{F}^{(-v)}(\mathbf{G}_t)_{(v)}{}^T) \otimes diag(\mathbf{a}_v)]vec(\mathbf{C}^{(v)})\|_2^2 + \lambda_2 \|vec(\mathbf{C}^{(v)})\|_p^p. \tag{24}$$

When we update $\mathcal{G}_t$ with fixed $\mathbf{a}^{(v)}$ and $\mathbf{C}^{(v)}$, let $\mathbf{X}_t$ with the $i$-th row being $vec(\mathcal{X}_{t,i})^T$, then each task can be independently optimized and the subproblem is:

$$\min_{\mathcal{G}_t} \frac{1}{n_t} \|\mathbf{y}_t - \mathbf{X}_t \mathbf{w}_t\|_2^2 + \lambda_3 \|vec(\mathcal{G}_t)\|_2^2, \tag{25}$$

where $\mathbf{w}_t = vec(\mathcal{W}_t) = (\mathbf{F}^{(V)} \otimes \cdots \otimes \mathbf{F}^{(1)})vec(\mathcal{G}_t)$, then we get the subproblem w.r.t. $\mathcal{G}_t$:

$$\min_{\mathcal{G}_t} \frac{1}{n_t} \|\mathbf{y}_t - \mathbf{X}_t(\mathbf{F}^{(V)} \otimes \cdots \otimes \mathbf{F}^{(1)})vec(\mathcal{G}_t)\|_2^2 + \lambda_3 \|vec(\mathcal{G}_t)\|_2^2, \tag{26}$$

# 6 Proof of Proposition 1 in the main paper

Here we give the proposition containing both algorithms:

**Proposition 1** *The proposed alternating algorithms do not increase the objective values of (1) and (11) at each iteration, indicating that*

$$J_1(\mathcal{A}^{(i+1)}, (\mathcal{B}_t)^{(i+1)}) \leq J_1(\mathcal{A}^{(i)}, (\mathcal{B}_t)^{(i)}), \tag{27}$$

$$J_2(\mathbf{a}^{(v)(i+1)}, \mathbf{C}^{(v)(i+1)}, \mathcal{G}_t^{(i+1)}) \leq J_2(\mathbf{a}^{(v)(i)}, \mathbf{C}^{(v)(i)}, \mathcal{G}_t^{(i)}), \tag{28}$$

*in the $(i+1)$-th iteration, with $J_1$ and $J_2$ denoting the objective values of (1) and (11), respectively.*

**proof**

- We first prove (27). When updating $\mathcal{B}_t$ with fixed $\mathcal{A}$, no matter $p = 1$ or $p = 2$, optimizing (22) can get the global optimal solution. So we can get the conclusion:

  $$J_1(\mathcal{A}^{(i)}, (\mathcal{B}_t)^{(i+1)}) \leq J_1(\mathcal{A}^{(i)}, (\mathcal{B}_t)^{(i)}).$$

  When updating $\mathcal{A}$ with fixed $\mathcal{B}_t$, we use (3) as the closed-form solution. In the proof of Theorem 1, we have mentioned that (3) is calculated from (7), which is the condition when (4) reaches the optimal value. Moreover, we know that equations (1), (2) and (4) are mutually equivalent. So the condition (3) can also reach the optimal value of (1), i.e.,

  $$J_1(\mathcal{A}^{(i+1)}, (\mathcal{B}_t)^{(i+1)}) \leq J_1(\mathcal{A}^{(i)}, (\mathcal{B}_t)^{(i+1)}).$$

  In short, we prove:

  $$J_1(\mathcal{A}^{(i+1)}, (\mathcal{B}_t)^{(i+1)}) \leq J_1(\mathcal{A}^{(i)}, (\mathcal{B}_t)^{(i+1)}) \leq J_1(\mathcal{A}^{(i)}, (\mathcal{B}_t)^{(i)}).$$

- We then prove (28) in a similar ways with (27). When updating $\mathbf{C}^{(v)}$ with fixed $\mathbf{a}^{(v)}$ and $\mathcal{G}_t$, we have the subproblem:

  $$\min_{\mathbf{C}^{(v)}} \sum_{t=1}^{m} \frac{1}{n_t} \|\mathbf{y}_t - \mathbf{X}_t[(\mathbf{F}^{(-v)}(\mathbf{G}_t)_{(v)}{}^T) \otimes diag(\mathbf{a}_v)]vec(\mathbf{C}^{(v)})\|_2^2 + \lambda_2 \|vec(\mathbf{C}^{(v)})\|_p^p. \tag{29}$$

  where $\mathbf{F}^{(-v)} = \mathbf{F}^{(V)} \cdots \otimes \mathbf{F}^{(v+1)} \otimes \mathbf{F}^{(v-1)} \cdots \otimes \mathbf{F}^{(1)}$. It is a convex optimization problem and we can get:

  $$J_2(\mathbf{a}^{(v)(i)}, \mathbf{C}^{(v)(i+1)}, \mathcal{G}_t^{(i)}) \leq J_2(\mathbf{a}^{(v)(i)}, \mathbf{C}^{(v)(i)}, \mathcal{G}_t^{(i)}).$$

  When updating $\mathbf{a}^{(v)}$ with fixed $\mathbf{C}^{(v)}$ and $\mathcal{G}_t$, we use (13) as the closed-form solution. In the proof of Theorem 2, we have mentioned that (13) is calculated from (17), which is the condition when problem

(14) reaches the optimal value. Moreover, we know that equations (11), (12) and (14) are mutually equivalent. So the condition (13) can also reach the optimal value of (11), i.e.,

$$J_2(\mathbf{a}^{(v)^{(i+1)}}, \mathbf{C}^{(v)^{(i+1)}}, \mathcal{G}_t^{(i)}) \leq J_2(\mathbf{a}^{(v)^{(i)}}, \mathbf{C}^{(v)^{(i+1)}}, \mathcal{G}_t^{(i)}).$$

When updating $\mathcal{G}_t$ with fixed $\mathbf{a}^{(v)}$ and $\mathbf{C}^{(v)}$, we have the subproblem:

$$\min_{\mathcal{G}_t} \frac{1}{n_t}\|\mathbf{y}_t - \mathbf{X}_t(\mathbf{F}^{(V)} \otimes \cdots \otimes \mathbf{F}^{(1)})vec(\mathcal{G}_t)\|_2^2 + \lambda_3\|vec(\mathcal{G}_t)\|_2^2, \tag{30}$$

which is also a convex optimization problem, and we can get:

$$J_2(\mathbf{a}^{(v)^{(i+1)}}, \mathbf{C}^{(v)^{(i+1)}}, \mathcal{G}_t^{(i+1)}) \leq J_2(\mathbf{a}^{(v)^{(i+1)}}, \mathbf{C}^{(v)^{(i+1)}}, \mathcal{G}_t^{(i)}).$$

In short, we reach the conclusion:

$$\begin{aligned} J_2(\mathbf{a}^{(v)^{(i+1)}}, \mathbf{C}^{(v)^{(i+1)}}, \mathcal{G}_t^{(i+1)}) &\leq J_2(\mathbf{a}^{(v)^{(i+1)}}, \mathbf{C}^{(v)^{(i+1)}}, \mathcal{G}_t^{(i)}) \\ &\leq J_2(\mathbf{a}^{(v)^{(i)}}, \mathbf{C}^{(v)^{(i+1)}}, \mathcal{G}_t^{(i)}) \leq J_2(\mathbf{a}^{(v)^{(i)}}, \mathbf{C}^{(v)^{(i)}}, \mathcal{G}_t^{(i)}). \end{aligned}$$

# 7 Implementation of vMSTF and MSTF for Section: Optimization

We provide the pseudocode of vMSTF and MSTF in Algorithms 1 and 2, respectively. We set the maximum number of iterations to be 1000 and terminate the algorithm once the relative change of its objective value is less than $10^{-4}$.

**Algorithm 1** Optimization Algorithm of vMSTF

**Input**: $\{\{\mathbf{X}_t^{(v)}\}_{v=1}^V\}_{t=1}^m, \{\mathbf{y}_t\}_{t=1}^m, \lambda_1, \lambda_2$
**Output**: $\mathcal{W}_t = \mathcal{A} \odot \mathcal{B}_t$.

initialize $\mathcal{A}$ and $\mathcal{B}_t$
**repeat**
  **for** $t = 1 : m$ **do**
    Update $\mathcal{B}_t$ by solving the problem: $\min_{\mathbf{b}_t} \frac{1}{n_t}\|\mathbf{y}_t - \mathbf{X}_t \cdot diag(\mathbf{a}) \cdot \mathbf{b}_t\|_2^2 + \lambda_2\|\mathbf{b}_t\|_p^p$.
  **end for**
  Update $\mathcal{A}$ by computing $(\mathcal{A})_{i_1 \ldots i_V} = (\frac{\lambda_2}{\lambda_1})^{\frac{1}{k}} \sqrt[k]{\sum_{t=1}^m |(\mathcal{B}_t)_{i_1 \ldots i_V}|^p}, \quad \forall i_1, ..., i_V$.
**until** convergence

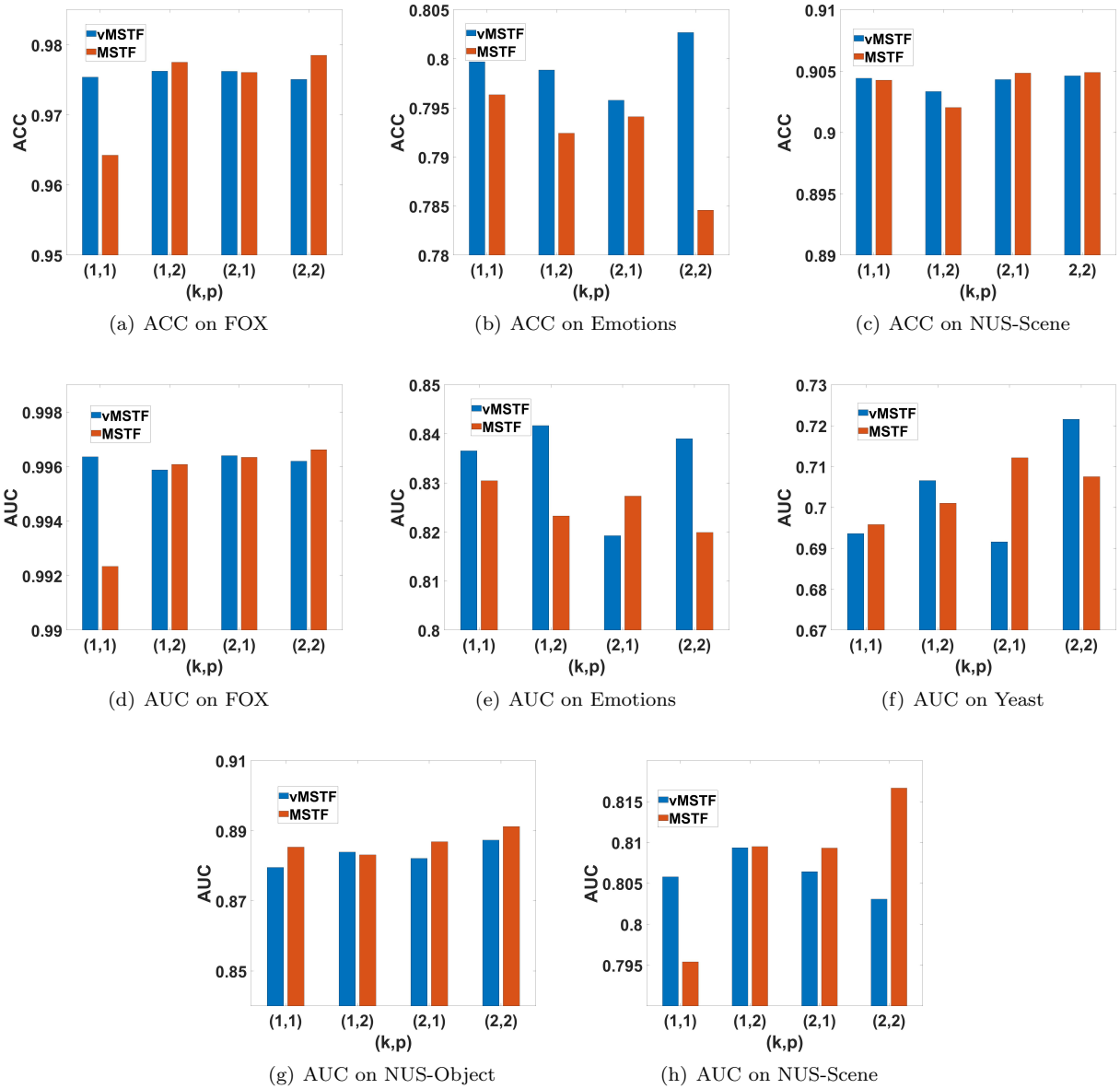
**Algorithm 2** Optimization Algorithm of MSTF

**Input**: $\{\{\mathbf{X}_t^{(v)}\}_{v=1}^V\}_{t=1}^m, \{\mathbf{y}_t\}_{t=1}^m, \lambda_1, \lambda_2, \lambda_3, k, p$
**Output**: $\mathcal{W}_t = \mathcal{G}_t \times_1 (diag(\mathbf{a}^{(1)})\mathbf{C}^{(1)}) \cdots \times_V (diag(\mathbf{a}^{(V)})\mathbf{C}^{(V)})$.

initialize $\mathcal{G}_t$, $\mathbf{a}^{(v)}$ and $\mathbf{C}^{(v)}$
**repeat**
  **for** $v = 1 : V$ **do**
    Update $\mathbf{C}^{(v)}$ by solving the problem:

$$\min_{\mathbf{C}^{(v)}} \sum_{t=1}^m \frac{1}{n_t}\|\mathbf{y}_t - \mathbf{X}_t[(\mathbf{F}^{(-v)}(\mathbf{G}_t)_{(v)}{}^T) \otimes diag(\mathbf{a}_v)]vec(\mathbf{C}^{(v)})\|_2^2 + \lambda_2\|vec(\mathbf{C}^{(v)})\|_p^p.$$

    Update $\mathbf{a}^{(v)}$ by computing $(\mathbf{a}^{(v)})_i = (\frac{\lambda_2}{\lambda_1})^{\frac{1}{k}} \sqrt[k]{\sum_{j=1}^{r_v} |(\mathbf{C}^{(v)})_{ij}|^p}, \quad \forall i$.
  **end for**
  **for** $t = 1 : m$ **do**
    Update $\mathcal{G}_t$ by solving the problem: $\min_{\mathcal{G}_t} \frac{1}{n_t}\|\mathbf{y}_t - \mathbf{X}_t(\mathbf{F}^{(V)} \otimes \cdots \otimes \mathbf{F}^{(1)})vec(\mathcal{G}_t)\|_2^2 + \lambda_3\|vec(\mathcal{G}_t)\|_2^2$,
  **end for**
**until** convergence




(a) ACC on FOX (b) ACC on Emotions (c) ACC on NUS-Scene

(d) AUC on FOX (e) AUC on Emotions (f) AUC on Yeast

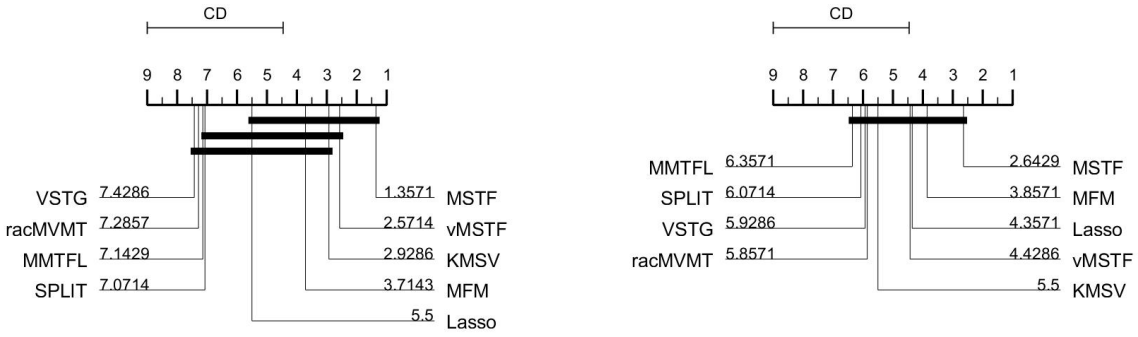
(g) AUC on NUS-Object (h) AUC on NUS-Scene

Figure 1: Comparison of different variants ($k, p \in \{1, 2\}$) of vMSTF and MSTF in ACC and AUC on real-world datasets.

# 8 Data preparation for Section: Experiments

We compare MSTF with three types of methods: MTL, MVMTL and baseline (Lasso [Tibshirani, 1996]). For MTL methods, we directly concatenate features from different views to make it suitable for the algorithm. We conduct experiments on seven real-world dual-heterogeneous datasets.

- **Fox**: It is an article classification dataset with 4 categories(*tasks*): health, science_technology, sports and travel, and each article has 2 *views*: text and image. The dataset we use contains 2711 and 747 features for the two views, respectively, and each task has 1523 samples. We use PCA to reduce the dimensions of the two views to 19 and 69, respectively.
- **Emotions**: It is a music emotion classification dataset, and the data has two *views*: rhythmic and timbre, which has 8 and 64 features, respectively. The dataset has 6 *tasks* and each has 593 samples.
- **Yeast**: It is a biololgical experiments dataset, which has 2 *views*: concatenation of the genetic expression (79 features) and the phylogenetic profile of a gene (24 features). It has 14 *tasks* and each task has 2417 samples.
- **NUS-Object**: It is an object image dataset extracted from the NUS-WIDE dataset [Chua et al., 2009]. Images labels, such as book, car and horse, can be treated as *tasks*. Each image contains 5 types (*views*)




(a) ACC (b) AUC

Figure 2: CD diagrams (0.05 significance level) of all the methods in ACC and AUC.

of low-level features: 64-D color histogram, 144-D color correlation, 73-D edge direction histogram, 128-D wavelet texture and 225-D block-wise color moments. We combine color histogram, olor correlation and edge direction histogram into one view, and wavelet texture and block-wise color moments into the other view in the experiment. We use PCA to reduce the dimensions of both two combined views to 39. In addition, for each task, we randomly select 2000 out of total 10370 samples in experiments.

- **NUS-Scene**: It is a scene image dataset extracted from the NUS-WIDE dataset [Chua et al., 2009]. Images labels, such as beach, building and road, can be treated as *tasks*. Each image contains 5 types (*views*) of low-level features: 64-D color histogram, 144-D color correlation, 73-D edge direction histogram, 128-D wavelet texture and 225-D block-wise color moments. We combine color histogram, color correlation and edge direction histogram into one view, and wavelet texture and block-wise color moments into the other view in the experiment. We use PCA to reduce the dimensions of both two combined views to 39. In addition, for each task, we randomly select 2000 out of total 16406 samples in experiments.

- **Mirflickr**: It is an image annotation dataset which has 6 *views*: RGB(4096 features), Lab(4096 features), Hsv(4096 features), Gist(512 features), DenseSift(1000 features) and DenseHue(100 features). The dataset has 38 *tasks* and each has 25000 features.

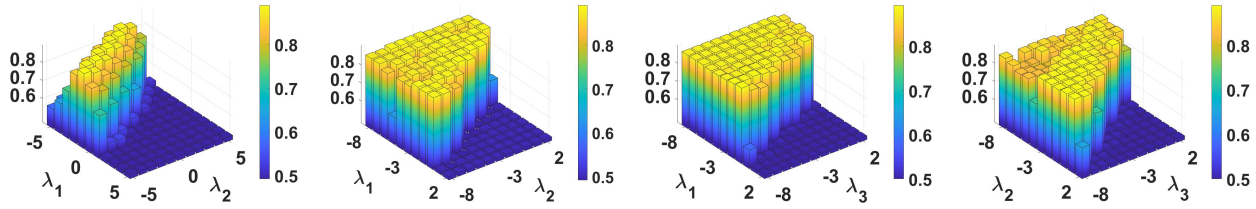
- **Corel5k**: It is an image annotation dataset which has 6 *views*: RGB(4096 features), Lab(4096 features), Hsv(4096 features), Gist(512 features), DenseSift(1000 features) and DenseHue(100 features). The dataset has 260 *tasks* and each has 4999 features.

# 9 Statistical test for Section: Experiments

We perform statistical test on the experimental results shown in Table 3 of the main paper. We use Nemenyi test [Demšar, 2006] to statistically evaluate the performance among vMSTF, MSTF and all the comparing methods. Figure. 2 shows the critical difference (CD) diagrams for ACC and AUC at 0.05 significance level. CD is used to show if two methods have significantly different performance. If two methods perform differently, their average ranks differ by at least a critical difference. From Figure. 2, we can find that for ACC, the proposed MSTF and vMSTF have highest ranks compared to other methods and they have superior performances than racMVMT, VSTG and MMTFL. For AUC, MSTF ranks the first and MFM ranks the second, which shows the necessity to capture full-order interactions in dualheterogeneous datasets.

# 10 Analysis on different variants of vMSTF and MSTF for Section: Experiments

In the main paper, we show the performance of different variants of vMSTF and MSTF in ACC on the Yeast and NUS-Object datasets. Here we show more results in Figure. 1. We can find that for different dataset, the




(a) vMSTF (b) MSTF (c) MSTF (d) MSTF

Figure 3: Sensitivity Analysis of vMSTF, and MSTF in AUC on the NUS-Object dataset. Values (shown in the logarithm scale) of $\lambda_1, \lambda_2$ in vMSTF are selected from $\{10^{-5}, 10^{-4}, ..., 10^5\}$, and $\lambda_1, \lambda_2, \lambda_3$ in MSTF are selected from $\{10^{-8}, 10^{-7}, ..., 10^2\}$.

degree of sparsity (the smaller values of $k$ and $p$, the stronger the sparsity) influences performance a lot. In practice, setting an appropriate degree of sparsity is preferred for both vMSTF and MSTF on different datasets..

## 11 Hyperparameter sensitivity analysis for Section: Experiments

In the main paper, we show the sensitivity analysis of MSTF in terms of $\lambda_1, \lambda_2, \lambda_3$ and $\alpha$ on the NUS-Object dataset. Here we show some other experimental results in Figure. 3. For vMSTF, $\lambda_1$ and $\lambda_2$ control the sparsity of task-shared component $\mathcal{A}$ and task-specific component $\mathcal{B}_t$, respectively. We select $\lambda_1, \lambda_2$ from $\{10^{-5}, 10^{-4}, ..., 10^5\}$. The result is shown in Figure 3(a). For MSTF, $\lambda_1$ and $\lambda_2$ control the sparsity of task-shared component $\mathcal{A}$ and task-specific component $\mathcal{B}_t$, respectively. $\lambda_3$ controls the strength of regularization on the core tensor $\mathcal{G}_t$ and the factor ratio $\alpha$ controls the number of latent factors. We select $\lambda_1, \lambda_2, \lambda_3$ from $\{10^{-8}, 10^{-7}, ..., 10^2\}$. In Figure. 3(b), the experiment is conducted by fixing $\lambda_3 = 10^{-3}$ and $\alpha = 0.2$. Similar setting is applied for other two experiments in Figure. 3(c) and Figure. 3(d).

## References

[Chua et al., 2009] Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y.-T. (2009). Nuswide: A real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9.

[Demšar, 2006] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30.

[Nesterov, 2013] Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.