Distant-supervision of heterogeneous multitask learning for social event forecasting with multilingual indicators (Supplementary Materials)

Proof of Theorem 1

We aim to solve the following optimization problem.

$$\arg\min_{Q_{s,t,l}} h(Q_{s,t,l}) \tag{1}$$

$$h(Q_{s,t,l}) = \|Z_{s,t} - \max_l Q_{s,t,l} + \Lambda_{2,s,t}/\rho\|_F^2 + \sum\nolimits_l^L \|Q_{s,t,l} - U_l^T \Theta_l X_{s,t,l}^T + \Lambda_{3,s,t}/\rho\|_F^2$$

This subsection presents the proof of Theorem 1.

Theorem 1. *The solution to the problem in Equation* (1) *is as follows:*

$$Q_{s,t,l} = \begin{cases} U_l^T \Theta_l X_{s,t,l}^T - \Lambda_{3,s,t,l}/\rho, & l \neq \arg\min_i Q_{s,t,i} \\ \sum_{i=1}^k (\tilde{Q}_{s,t,i} + U_l^T \Theta_l X_{s,t,l}^T + \Lambda_{2,s,t}/\rho)/(k+1), & l = \arg\min_i Q_{s,t,i} \end{cases}$$

where $\tilde{Q}_{s,t}$ is the decreasing ordered list whose elements are the set $\{U_l^T \Theta_l X_{s,t,l} - \Lambda_{3,s,t,l}/\rho\}_l^L$, and k is equal to the solution of the following problem:

$$k = \arg\min_j \ j, \ s.t. \ \sum\nolimits_{i=1}^j (\tilde{Q}_{s,t,i} + U_l^T \Theta_l X_{s,t,l}^T + \Lambda_{2,s,t}/\rho)/(j+1) > \tilde{Q}_{s,t,j-1} \tag{2}$$

Proof: There are two possible situations for the solution of $Q_{s,t,l}$: 1) Situation 1: $Q_{s,t,l} = \max_i Q_{s,t,i}$; and 2) Situation 2: $Q_{s,t,l} < \max_i Q_{s,t,i}$. It is easily seen that the solution for Situation 2 has closed-form: $U_l^T \Theta_l X_{s,t,l}^T - \Lambda_{3,s,t,l}/\rho$. In the following, we focus on proving the solution for Situation 1, and detail how to identify which situation each $Q_{s,t,l}$ should lie in. Assume $x = \max_l Q_{s,t,l}$ and define an index set $\mathcal{C} = \{l | l \in L, Q_{s,t,j} = \max_i Q_{s,t,i}\}$ which implies $Q_{s,t,l} (l \in \mathcal{C})$ belongs to Situation 1, while the complementary set $\bar{\mathcal{C}} = L - \mathcal{C}$. Therefore, by integrating the closed-form solutions of the variables in Situation 2 into the objective function, the problem Equation (1) is simplified into the following problem:

$$\min_x \|Z_{s,t} - x + \Lambda_{2,s,t}/\rho\|_F^2 + \sum\nolimits_{l \in \mathcal{C}}^L \|x - U_l^T \Theta_l X_{s,t,l}^T + \Lambda_{3,s,t}/\rho\|_F^2 \tag{3}$$

The optimal solution has the following closed-form:

$$x^* = (\sum\nolimits_{l \in \mathcal{C}} (U_l^T \Theta_l X_{s,t,l}^T - \Lambda_{3,s,t}/\rho) + U_l^T \Theta_l X_{s,t,l}^T + \Lambda_{2,s,t}/\rho)/(|\mathcal{C}| + 1) \tag{4}$$

Then remaining issue is to determine the set $\mathcal{C}$ such that $Q_{s,t,l} (l \in \mathcal{C})$ belongs to Situation 1. Rank the set $Q'_{s,t} = \{U_l^T \Theta_l X_{s,t,l}^T - \Lambda_{3,s,t,l}/\rho\}_l^L$ by a decreasing order and thus form an ordered list $\tilde{Q}_{s,t}$, where $\tilde{Q}_{s,t,i}$ is the ith-largest element in it. Therefore, the problem for determining $\mathcal{C}$ is equivalent to identify how many largest elements should be selected from $\tilde{Q}_{s,t}$, where k is the number of elements in $\mathcal{C}$. In other words, $\mathcal{C}$ is composed of the top k largest elements in $\tilde{Q}_{s,t}$. Assume $a_{s,t,k} = x^*$ because x^* is a function of k, then we need to prove the objective function $h(Q_{s,t,l}, |\mathcal{C}| = k)$ increases monotonously with k. In fact,

$$\begin{aligned} & h(Q_{s,t,l}, |\mathcal{C}| = k+1) - h(Q_{s,t,l}, |\mathcal{C}| = k) \\ &= \sum\nolimits_{i=1}^{k+1} (a_{s,t,k+1} - \tilde{Q}_{s,t,i})^2 + (a_{s,t,k+1} - Z_{s,t} - \Lambda_{2,s,t})^2 - \sum\nolimits_{i=1}^{k} (a_{s,t,k} - \tilde{Q}_{s,t,i})^2 - (a_{s,t,k} - Z_{s,t} - \Lambda_{2,s,t})^2 \\ &= (a_{s,t,k+1} - a_{s,t,k})((k+1)(a_{s,t,k+1} + a_{s,t,k}) - 2\sum\nolimits_{i=1}^{k} \tilde{Q}_{s,t,i} - 2Z_{s,t} - 2\Lambda_{2,s,t}) \\ &= (a_{s,t,k+1} - a_{s,t,k})(\tilde{Q}_{s,t,k+1} - a_{s,t,k+1}) \\ &= (a_{s,t,k+1} - a_{s,t,k})((k+2)a_{s,t,k+1} - (k+1)a_{s,t,k} - a_{s,t,k+1}) \\ &= (k+1)(a_{s,t,k+1} - a_{s,t,k})^2 \geqslant 0 \end{aligned}$$

Therefore, we need to find the smallest k that satisfies the Situation 1, which is equal to solving the following optimization problem:

$$k = \arg\min_j \ j, \ s.t. \ \sum\nolimits_{i=1}^j (\tilde{Q}_{s,t,i} + U_l^T \Theta_l X_{s,t,l}^T + \lambda_{2,s,t}/\rho)/(j+1) > \tilde{Q}_{s,t,j-1} \tag{5}$$

The proof is completed □

Proof of Lemma 1

Lemma 1. *$\gamma_{\min}$ is the lower bound of $\|\Theta_l\|_1$ such that $\forall m \in \{1,2,\cdots,d\}$: $\gamma_{\min} = d \cdot \arg\min_{\|\Theta_{l,m}\|_2=1} \|\Theta_{l,m}\|_1$, where $\Theta_{l,m}$ is any row of Θ_l.*

Proof: First we know that the lower bound of $\|\Theta_l\|_1$ is $\arg\min_{\Theta_l\Theta_I^T=I} \|\Theta_l\|_1$, which is equal to $\arg\min_{\|\Theta_{l,m}\|_2=1} \sum_{m=1}^{d} \|\Theta_{l,m}\|_1$. We also know that $\arg\min_{\|\Theta_{l,m}\|_2=1} \sum_{m=1}^{d} \|\Theta_{l,m}\|_1 = d \cdot \arg\min_{\|\Theta_{l,m}\|_2=1} \|\Theta_{l,m}\|_1$ ($\forall m \in \{1,2,\cdots,d\}$). Therefore, the proof is completed. $\square$

Proof of Theorem 2

The proof of Theorem 2 is shown in this section. In order to prove Theorem 2, we need to prove Theorem 6 which requires the proof of Theorem 5. Theorem 5 requires the proof of Theorems 3 and 4. Therefore, we first present Theorems 3 and 4.

Theorem 2. *Let $\epsilon > 0$ and let μ be probability measure on $\mathbb{R}$. With probability of at least $1-\epsilon$ in the draw of $M \sim \mu^{|S|\cdot|T|}$, we have:*

$$
\begin{aligned}
\mathbb{E}(\Theta^*_{(M)}, U^*_{(M)}) - \mathbb{E}(\Theta^*, U^*) &= \mathbb{E}_{M\sim\mu}\Big[\frac{1}{|S|\cdot|T|}\sum_{s,t}^{S,T} \mathcal{L}(\max_l F_l([U^*_{(M)}]_l^T[\Theta^*_{(M)}]_l X^T_{s,t,l}), Y_{s,\tau})\Big] \\
&\quad - \inf_{\Theta\in\mathcal{F}_2, U\in\mathcal{F}_1} \mathbb{E}_{M\sim\mu}\Big[\frac{1}{|S|\cdot|T|}\sum_{s,t}^{S,T} \mathcal{L}(\max_l F_l(U_l^T\Theta_l X^T_{s,t,l}), Y_{s,\tau})\Big] \\
&\leq 2C\alpha\sqrt{\frac{2\mathcal{C}_1(X)|L|(d+12)}{|S|\cdot|T|}} + 2C|L|\alpha\sqrt{\frac{8\mathcal{C}_\infty(X)\ln(2d)}{|S|\cdot|T|}} + 2\sqrt{\frac{2\ln 2/\epsilon}{|S|\cdot|T|}}
\end{aligned}
$$

Theorem 3. *Define $F_U = F_U(\sigma) = \sup_{\Theta_l\in\mathcal{F}_2}\sum_{s,t}^{S,T} \sigma_{s,t}\max_l \left\langle U_l^T\Theta_l, X_{s,t,l}\right\rangle$, we have:*

$$
\mathbb{E}_\sigma F_U = \mathbb{E}_\sigma \sup_{\Theta_l\in\mathcal{F}_2}\sum_{s,t,l}^{S,T,L} \sigma_{s,t}\left\langle U_l^T\Theta_l, X_{s,t,l}\right\rangle \leq \alpha\sqrt{d|L||S||T|\mathcal{C}_1(X)} \tag{6}
$$

where $\mathcal{C}_1(X) = d\cdot\|\hat{\Sigma}(X)\|_*$.

Proof:

$$
\begin{aligned}
\mathbb{E}_\sigma F_U &= \mathbb{E}_\sigma \sup_{\Theta_l \in \mathcal{F}_2} \sum\nolimits_{s,t}^{S,T} \sigma_{s,t} \max_l \left\langle U_l^T \Theta_l, X_{s,t,l} \right\rangle && (7)\\
&= \mathbb{E}_\sigma \sup_{\Theta_l \in \mathcal{F}_2} \sum\nolimits_{s,t}^{S,T} \sigma_{s,t} \max_l \sum\nolimits_i^d \langle U_{l,i}\Theta_{l,i}, X_{s,t,l} \rangle && (8)\\
&\leq \mathbb{E}_\sigma \sup_{\Theta_l \in \mathcal{F}_2} \sum\nolimits_{s,t}^{S,T} \max_l \sum\nolimits_i^d \langle \Theta_{l,i}, \sigma_{s,t} U_{l,i} X_{s,t,l} \rangle && (9)\\
&\leq \mathbb{E}_\sigma \sup_{\Theta_l \in \mathcal{F}_2} \sum_{s,t}^{S,T} \max_l \sum_i^d \|\Theta_{l,i}\| \|\sigma_{s,t} X_{s,t,l} U_{l,i}\| \text{ (Cauchy-Schwarz inequality)} && (10)\\
&\leq \mathbb{E}_\sigma \sup_{\Theta_l \in \mathcal{F}_2} \sum_{s,t}^{S,T} \sum_l^L \sum_i^d \|\Theta_{l,i}\| \|\sigma_{s,t} X_{s,t,l} U_{l,i}\| && (11)\\
&\leq \mathbb{E}_\sigma \sup_{\Theta_l \in \mathcal{F}_2} \sum_{s,t}^{S,T} \left((\sum_{l,i}^{L,d} \|\Theta_{l,i}\|^2)^{\frac{1}{2}} (\sum_{l,i}^{L,d} \|\sigma_{s,t} X_{s,t,l} U_{l,i}\|^2)^{\frac{1}{2}} \right) \quad \text{(Cauchy-Schwarz inequality)} && (12)\\
&\leq \sum_{s,t}^{S,T} \left(\mathbb{E}_\sigma \sup_{\Theta_l \in \mathcal{F}_2} (\sum_{l,i}^{L,d} \|\Theta_{l,i}\|^2)^{\frac{1}{2}} (\sum_{l,i}^{L,d} \|\sigma_{s,t} X_{s,t,l} U_{l,i}\|^2)^{\frac{1}{2}} \right) && (13)\\
&= \sum_{s,t}^{S,T} \left(\sup_{\Theta_l \in \mathcal{F}_2} (\sum_{l,i}^{L,d} \|\Theta_{l,i}\|^2)^{\frac{1}{2}} \cdot \mathbb{E}_\sigma (\sum_{l,i}^{L,d} \|\sigma_{s,t} X_{s,t,l} U_{l,i}\|^2)^{\frac{1}{2}} \right) && (14)\\
&= \sqrt{|L| \cdot d} \cdot \sum_{s,t}^{S,T} \mathbb{E}_\sigma \left(\sum_{l,i}^{L,d} \left(\|\sigma_{s,t} U_{l,i} X_{s,t,l}\|^2 \right) \right)^{1/2} \quad (\Theta_l \Theta_l^T = I) && (15)\\
&\leq \sqrt{|L| \cdot d} \cdot \sum_{s,t}^{S,T} \left(\sum_{l,i}^{L,d} \left(\|U_{l,i}\|^2 \cdot \mathbb{E}_\sigma \|\sigma_{s,t} X_{s,t,l}\|^2 \right) \right)^{1/2} \quad \text{(Cauchy-Schwarz inequality)} && (16)\\
&\leq \sqrt{|L| \cdot d} \cdot \sum_{s,t}^{S,T} \left(\sum_{l,i}^{L,d} \|U_{l,i}\|^2 \cdot \mathbb{E}_\sigma \sum_{l,i}^{L,d} \|\sigma_{s,t} X_{s,t,l}\|^2 \right)^{1/2} && (17)\\
&\leq \cdot \sqrt{|L| \cdot d} \cdot \sum_{s,t}^{S,T} \left(\left(\sum_i^d \left(\sum_l^L \|U_{l,i}\|^2 \right)^{\frac{1}{2}} \right)^2 \mathbb{E}_\sigma \sum_{l,i}^{L,d} \|\sigma_{s,t} X_{s,t,l}\|^2 \right)^{1/2} && (18)\\
&\leq \alpha \cdot \sqrt{|L| \cdot d} \cdot \sum_{s,t}^{S,T} \left(\mathbb{E}_\sigma \sum_{l,i}^{L,d} \|\sigma_{s,t} X_{s,t,l}\|^2 \right)^{1/2} \quad (\sum_i^d \left(\sum_l^L \|U_{l,i}\|^2 \right)^{\frac{1}{2}} \leq \alpha) && (19)\\
&\leq \alpha \cdot d\sqrt{|L|} \cdot \sum_{s,t}^{S,T} \left(\sum_l^L \|X_{s,t,l}\|^2 \right)^{1/2} && (20)\\
&= \alpha \cdot d\sqrt{|L|} \cdot \left(|S||T| \sum_{s,t,l}^{S,T,L} \|X_{s,t,l}\|^2 \right)^{1/2} \qquad (\left(\sum_k^K x_k \right)^2 \leq K \sum_k^K x_k^2) && (21)\\
&= \alpha \sqrt{d|L||S||T|\mathcal{C}_1(X)} \qquad (\mathcal{C}_1(X) = \sum_{s,t,l}^{S,T,L} d\|X_{s,t,l}\|^2) && (22)
\end{aligned}
$$

The proof is completed. □

Theorem 4. *If U satisfies $\|U\|_{2,1} \leq \alpha, \alpha > 0$, then for any $u \geq 0$*

$$\Pr\{F_U \geq \mathbb{E}[F_U] + u\} \leq \exp\left(\frac{-u^2}{\alpha^2 8|S||T|\mathcal{C}_\infty(X)}\right) \tag{23}$$

Proof: For any configuration σ of Rademacher variables, let

$$\Theta(\sigma) = \arg\max_{\Theta\in\mathcal{F}_2} F_U(\sigma) = \arg\max_{\Theta\in\mathcal{F}_2} \sum_{s,t}^{S,T} \sigma_{s,t} \max_l \left\langle U_l^T\Theta_l, X_{s,t,l}\right\rangle \tag{24}$$

For any $\hat{s} \in S, \hat{t} \in T$, and any $\sigma' \in \{-1, 1\}$ to replace $\sigma_{s,t}$ we have:

$$F_U(\sigma) - F_U(\sigma_{\hat{s},\hat{t}} \leftarrow \sigma') \tag{25}$$

$$= \sup_{\Theta\in\mathcal{F}_2} \sum_{s,t}^{S,T} \sigma_{s,t} \max_l \left\langle U_l^T\Theta_l, X_{s,t,l}\right\rangle - \sup_{\Theta\in\mathcal{F}_2} \sum_{s,t} \sigma'_{s,t} \max_l \left\langle U_l^T\Theta_l, X_{s,t,l}\right\rangle \tag{26}$$

$$= \sum_{s,t}^{S,T} \sigma_{s,t} \max_l \sigma_{s,t} \left\langle U_l^T\Theta_l(\sigma), X_{s,t,l}\right\rangle - \sup_{\Theta\in\mathcal{F}_2} \sum_{s,t}^{S,T} \sigma'_{s,t} \max_l \left\langle U_l^T\Theta_l, X_{s,t,l}\right\rangle \quad \text{(By definition)} \tag{27}$$

$$\leq \sum_{s,t}^{S,T} \sigma_{s,t} \max_l \sigma_{s,t} \left\langle U_l^T\Theta_l(\sigma), X_{s,t,l}\right\rangle - \sum_{s,t} \sigma'_{s,t} \max_l \left\langle U_l^T\Theta_l(\sigma), X_{s,t,l}\right\rangle \tag{28}$$

$$= \sigma_{\hat{s},\hat{t}} \max_l \left\langle U_l^T\Theta_l(\sigma), X_{\hat{s},\hat{t},l}\right\rangle - \sigma'_{\hat{s},\hat{t}} \max_l \left\langle U_l^T\Theta_l(\sigma), X_{\hat{s},\hat{t},l}\right\rangle \tag{29}$$

$$\leq 2|\max_l \left\langle U_l^T\Theta_l(\sigma), X_{\hat{s},\hat{t},l}\right\rangle| \tag{30}$$

Define $X_{\cdot,\cdot,l} = \{X_{s,t,l}\}_{s,t}^{S,T} \in \mathbb{R}^{(|S|\cdot|T|)\times(|V_l|+1)}$. Therefore, we have:

$$H(\sigma, U) = \sum_{\hat{s},\hat{t}}^{S,T} \left(F_U(\sigma) - \sup_{\sigma'\in\{-1,1\}} F_U(\sigma_{(\hat{s},\hat{t})\to\sigma'})\right)^2 \tag{31}$$

$$\leq 4 \sum_{\hat{s},\hat{t}}^{S,T} \max_l \left\langle U_l^T\Theta_l(\sigma), X_{\hat{s},\hat{t},l}\right\rangle^2 \tag{32}$$

$$\leq 4 \sum_{\hat{s},\hat{t}}^{S,T} \sum_l^L \left\langle U_l^T\Theta_l(\sigma), X_{\hat{s},\hat{t},l}\right\rangle^2 \tag{33}$$

$$= 4|S||T| \cdot \frac{1}{|S||T|} \sum_{\hat{s},\hat{t},l}^{S,T,L} \left\langle U_l^T\Theta_l(\sigma), X_{\hat{s},\hat{t},l}\right\rangle^2 \tag{34}$$

$$= 4|S||T| \sum_l^L (U_l^T\Theta_l(\sigma))^T \hat{\Sigma}(X_{\cdot,\cdot,l})(U_l^T\Theta_l(\sigma)) \tag{35}$$

$$\leq 4|S||T| \sum_l^L \lambda_{\max}(\hat{\Sigma}(X_{\cdot,\cdot,l}))\|U_l^T\Theta_l(\sigma)\|^2 \quad (\lambda_{\max}(x) \text{ is the largest eigen-value of } x) \tag{36}$$

$$= 4|S||T| \sum_l^L \|\hat{\Sigma}(X_{\cdot,\cdot,l})\|_\infty \|U_l^T\Theta_l(\sigma)\|^2 \tag{37}$$

$$\leq 4\alpha^2|S||T| \sum_l^L \|\hat{\Sigma}(X_{\cdot,\cdot,l})\|_\infty \tag{38}$$

$$= 4\alpha^2|S||T|\mathcal{C}_\infty(X) \tag{39}$$

Denote $B^2 = \sup H(\sigma, U) = 4\alpha^2|S||T|\mathcal{C}_\infty(X)$, and apply Theorem 6.9 in the supplementary material of (Zhou et al. 2013), we have:

$$\Pr\{F_U \geq \mathbb{E}[F_U] + u\} \leq \exp\frac{-u^2}{2B^2} = \exp\frac{-u^2}{8\alpha^2|S||T|\mathcal{C}_\infty(X)} \tag{40}$$

□

Lemma 2.

$$\sum_l^L \|U_l^T\Theta_l\|^2 \tag{41}$$

$$\leq \sum_l^L \|\sum_i^d U_{l,i}\Theta_{l,i}\|^2 \tag{42}$$

$$\leq \sum_{l,i}^{L,d} \|U_{l,i}\Theta_{l,i}\|^2 = \sum_{l,i}^{L,d} \|U_{l,i}\|^2\|\Theta_{l,i}\|^2 \tag{43}$$

$$=|L|d\sum_i^d\sum_l^L \|U_{l,i}\|^2 \quad (\Theta_l\Theta_l^T = I) \tag{44}$$

$$\leq |L|d\left(\sum_i^d\left(\sum_l^L \|U_{l,i}\|^2\right)^{1/2}\right)^2 \tag{45}$$

$$\leq \alpha^2 d\cdot|L| \quad (\|U\|_{2,1}\leq\alpha) \tag{46}$$

Theorem 5. *we have:*

$$\mathbb{E}_\sigma \sup_{\Theta\in\mathcal{F}_2, U\in\mathcal{F}_1} \sum\nolimits_{s,t}^{S,T} \sigma_{s,t}\mathcal{L}(\max_l(U_l^T\Theta_l X_{s,t,l}^T), Y_{s,\tau}) \leq C\alpha\sqrt{|L|(d+12)|S||T|\mathcal{C}_1(X)} \tag{47}$$

$$+ C\alpha L\sqrt{8|S||T|\mathcal{C}_\infty(X)\ln(2d)} \tag{48}$$

Proof: Because of the Lipschitz property of the loss function $\mathcal{L}$, we have:

$$\mathbb{E}_\sigma \sup_{\Theta\in\mathcal{F}_2, U\in\mathcal{F}_1} \sum\nolimits_{s,t}^{S,T} \sigma_{s,t}\mathcal{L}(\max_l(U_l^T\Theta_l X_{s,t,l}^T), Y_{s,\tau})$$

$$\leq C\mathbb{E}_\sigma\cdot\sup_{\Theta,U}\sum\nolimits_{s,t}^{S,T}\sigma_{s,t}\max_l(U_l^T\Theta_l X_{s,t,l}^T) \tag{49}$$

$$= C\mathbb{E}_\sigma \max_{U\in\mathcal{F}_1} F_U \tag{50}$$

$$= C\mathbb{E}_\sigma \max_{U\in ext(F_1)} F_U \quad (F_U \text{ is linear in } U\text{; Linear function attains maxima at extreme points}) \tag{51}$$

$$\mathbb{E}_\sigma \max_{U\in\text{ext}(\mathcal{F}_1)^T} F_U = \int_0^\infty \Pr\left\{\max_{u\in\text{ext}(\mathcal{F}_1)^T} F_U > u\right\}\mathrm{d}u \tag{52}$$

$$\leq \alpha\sqrt{d|L||S||T|\mathcal{C}_1(X)} + \delta + \sum_{u\in\text{ext}(\mathcal{F}_1)}\int_{\alpha\sqrt{dL|S||T|\mathcal{C}_1(X)}+\delta}^\infty \Pr\{F_U > u\}\mathrm{d}u \tag{53}$$

$$\leq \alpha\sqrt{d|L||S||T|\mathcal{C}_1(X)} + \delta + \sum_{u\in\text{ext}(\mathcal{F}_1)}\int_\delta^\infty \Pr\{F_U > \mathbb{E}F_U + u\}\mathrm{d}u \text{ (Theorem 3)} \tag{54}$$

$$\leq \alpha\sqrt{d|L||S||T|\mathcal{C}_1(X)} + \delta + \sum_{u\in\text{ext}(\mathcal{F}_1)}\int_\delta^\infty \exp\left(\frac{-u^2}{8\alpha^2|S||T|\mathcal{C}_\infty(X)}\right)\mathrm{d}u \text{ (Theorem 4)} \tag{55}$$

$$\leq \alpha\sqrt{d|L||S||T|\mathcal{C}_1(X)} + \delta + (2d)^{|L|}\int_\delta^\infty \exp\left(\frac{-u^2}{8\alpha^2 x|S||T|\mathcal{C}_\infty(X)}\right)\mathrm{d}u \text{ (Theorem 4) } (\text{card}(\text{ext}(\mathcal{F}_1)) = (2d)^{|L|}) \tag{56}$$

$$\leq \alpha\sqrt{d|L||S||T|\mathcal{C}_1(X)} + \delta + \frac{4\alpha^2|S||T|\mathcal{C}_\infty(X)(2d)^{|L|}}{\delta}\exp\left(\frac{-\delta^2}{8\alpha^2|S||T|\mathcal{C}_\infty(X)}\right) \tag{57}$$

$$\text{(Gaussian variable estimate) } (\text{card}(\text{ext}(\mathcal{F}_1)) = (2d)^{|L|}) \tag{58}$$

$$\tag{59}$$

Let $\delta = \sqrt{8|S||T|\mathcal{C}_\infty(X)\ln(e(2d)^T)}$, following the Proposition 12 in (Maurer, Pontil, and Romera-Paredes 2013), we have:

$$\mathbb{E}_\sigma \max_{e\in\text{ext}(\mathcal{F}_2)^T} F_U \leq \alpha\sqrt{2|L|(d+12)|S||T|\mathcal{C}_1(X)} + \alpha|L|\sqrt{8|S||T|\mathcal{C}_\infty(X)\ln(2d)} \tag{60}$$

which together with Equation (49) gives the result. □

Theorem 6. *Let $\epsilon > 0$, fix d and let μ be probability measures on $\mathbb{R}$. WIth probability of a least $1-\epsilon$ in the draw of $M \sim \mu$, we have $\forall \Theta \in \mathcal{F}_1$ and $\forall U \in \mathcal{F}_2$ that*

$$\begin{aligned}\mathbb{E}(\Theta, U) - \hat{\mathbb{E}}(\Theta, U|M) =& \mathbb{E}_{(X,Y)\sim\mu} \sum_{s,t}^{S,T} [\mathcal{L}(\max_l \left\langle U_l^T \Theta_l, X_{s,t,l}^T \right\rangle, Y_{s,\tau})] - \frac{1}{|S||T|} \sum_{s,t}^{S,T} \mathcal{L}(\max_l \left\langle U_l^T \Theta_l, X_{s,t,l}^T \right\rangle, Y_{s,\tau}) \\ \leq & 2C\alpha\sqrt{\frac{2(d+12)|L|\mathcal{C}_1(X)}{|S||T|}} + 2C|L|\alpha\sqrt{\frac{8\mathcal{C}_\infty(X)\ln(2d)}{|S||T|}} + \sqrt{\frac{9\ln(2/\epsilon)}{2|S||T|}} \end{aligned} \tag{61}$$

Proof:

$$\mathbb{E}(\Theta, U) = \mathbb{E}_{(X,Y)\sim\mu} \sum_{s,t}^{S,T} [\mathcal{L}(\max_l \left\langle U_l^T \Theta_l, X_{s,t,l}^T \right\rangle, Y_{s,\tau})] \tag{62}$$

$$\leq \frac{1}{|S||T|} \sum_{s,t}^{S,T} \mathcal{L}(\max_l \left\langle U_l^T \Theta_l, X_{s,t,l}^T \right\rangle, Y_{s,\tau}) + \hat{\mathcal{R}} + \sqrt{\frac{9\ln(2/\epsilon)}{2|S||T|}} \ \ (Theorem\ 6.12\ in\ (Zhouetal.2013)) \tag{63}$$

$$= \hat{\mathbb{E}}(\Theta, U|M) + \mathbb{E}_\sigma \sup_{\Theta\in\mathcal{F}_2, U\in\mathcal{F}_1} \frac{2}{|S||T|} \sum_{s,t}^{S,T} \sigma_{s,t} \mathcal{L}(\max_l \left\langle U_l^T \Theta_l, X_{s,t,l}^T \right\rangle, Y_{s,\tau}) + \sqrt{\frac{9\ln(2/\epsilon)}{2|S||T|}} \tag{64}$$

$$\leq \hat{\mathbb{E}}(\Theta, U|M) + 2C\alpha\sqrt{\frac{2(d+12)|L|\mathcal{C}_1(X)}{|S||T|}} + 2C|L|\alpha\sqrt{\frac{8\mathcal{C}_\infty(X)\ln(2d)}{|S||T|}} + \sqrt{\frac{9\ln(2/\epsilon)}{2|S||T|}} \ (Theorem\ 5) \tag{65}$$

which completes the proof. □

By Definitions 1 and 2, we have that

$$\hat{\mathbb{E}}(\Theta^*, U^*|M) - \hat{\mathbb{E}}(\Theta^*_{(M)}, U^*_{(M)}) > 0 \tag{66}$$

Therefore, we manipulate the terms to obtain:

$$\begin{aligned}\mathbb{E}(\Theta^*_{(M)}, U^*_{(M)}) =& \mathbb{E}(\Theta^*_{(M)}, U^*_{(M)}) - \mathbb{E}(\Theta^*, U^*) + \mathbb{E}(\Theta^*, U^*) && (67) \\ \leq & \hat{\mathbb{E}}(\Theta^*, U^*|M) - \hat{\mathbb{E}}(\Theta^*_{(M)}, U^*_{(M)}|M) - \mathbb{E}(\Theta^*, U^*) + \mathbb{E}(\Theta^*, U^*) && (68)\end{aligned}$$

Therefore we have:

$$\mathbb{E}(\Theta^*_{(M)}, U^*_{(M)}) - \mathbb{E}(\Theta^*, U^*) \leq \sup_{\Theta, U} |\mathbb{E}(\Theta, U) - \hat{\mathbb{E}}(\Theta, U|M)| + \hat{\mathbb{E}}(\Theta^*, U^*|M) - \mathbb{E}(\Theta^*, U^*) \tag{69}$$

The last two terms can be upper bounded using Hoeffding inequality . With probability of at least $1-\epsilon$, we have that:

$$\begin{aligned}\mathbb{E}(\Theta^*_{(M)}, U^*_{(M)}) - \mathbb{E}(\Theta^*, U^*) \leq & \sup_{\Theta, U} |\mathbb{E}(\Theta, U) - \hat{\mathbb{E}}(\Theta, U|M)| + \sqrt{\frac{\log(2/\epsilon)}{2|S||T|}} && (70) \\ \leq 2C\alpha\sqrt{\frac{2\mathcal{C}_1(X)|L|(d+12)}{|S|\cdot|T|}} + & 2C|L|\alpha\sqrt{\frac{8\mathcal{C}_\infty(X)\ln(2d)}{|S|\cdot|T|}} + \sqrt{\frac{8\ln 2/\epsilon}{|S|\cdot|T|}} \quad \text{(Theorem 6)} && (71)\end{aligned}$$

This completes the proof of Theorem 2.

Update Θ and U

Jointly optimizing Θ and U amounts to the following non-convex subproblem:

$$\begin{aligned}\min_{\Theta\geq 0, U} \lambda_1 \|U\|_{2,1} + \sum\nolimits_l^L \left\langle \Lambda_{1,l}, \Theta_l\Theta_l^T - I \right\rangle + \frac{\rho}{2} \sum\nolimits_l^L \|\Theta_l\Theta_l^T - I\|_F^2 + \\ \lambda_2 \sum\nolimits_l^L \|\Theta_l\|_1 + \frac{\rho}{2|S|\cdot|T|} \sum\nolimits_{s,t,l}^{S,T,L} \|U_l^T \Theta_l X_{s,t,l}^T - (Q_{s,t,l} - \Lambda_{3,s,t,l}/\rho)\|_F^2\end{aligned} \tag{72}$$

which contains a biconvex nonsmooth objective function of Θ and U as well as a quadratic equality constraint over Θ. To solve it, traditional methods like block coordinate descent (BCD) (Tseng and Yun 2009) may be easily trapped in a local minimizer in practice due to non-convexity and non-smoothness. To address this problem, we applied non-monotone strategy based on spectral projected gradient (SPG) method (Zhou et al. 2013). It is shown in (Lu and Zhang 2012) that under some suitable assumption the non-monotone SPG method has a linear convergence rate. The detailed algorithm procedures are shown in Algorithm 1, where Lines 3-4 are the calculation of the gradients with respect to Θ_l and U_l for the smooth part of the subproblem. Line 5 stores the historical max function value for the non-monotone SPG method. Then Lines 6-15 are the procedures of the non-monotone update of the step size η. Specifically, Lines 8-9 computes the proximal operators for the non-smooth parts of the Subproblem (72), Line 10 is the stop criterion and Line 13 is the update of new step size. Finally, Lines 16-20 are the calculations for the new Θ, U, and the residual ε. The details of the calcuations in Algorithm 1 is shown as follows.

The smooth part of function $L(\Theta, U)$ is:

$$\tilde{g}(\Theta, U) = \sum_{l}^{L} \left\langle \Lambda_{1,l}, \Theta_l \Theta_l^T - I \right\rangle + \frac{\rho}{2} \sum_{l}^{L} \|\Theta_l \Theta_l^T - I\|_F^2 + \tag{73}$$

$$\frac{\rho}{2S \cdot T} \sum_{s,t,l}^{S,T,L} \|U_l^T \Theta_l X_{s,t,l}^T - (Q_{s,t,l} - \Lambda_{3,s,t,l}/\rho)\|_F^2 \tag{74}$$

The gradient of $\tilde{g}$ with respect to Θ_l is given by:

$$\nabla_{\Theta_l} \tilde{g}(\Theta, U) = (\Lambda_{1l} + \Lambda_{1l}^T)\Theta_l + 2\rho\Theta_l(\Theta_l^T \Theta_l - I_\Theta) - \frac{\rho}{n} U_l \sum_{s,t}^{S,T} (Q_{s,t,l} - \frac{\Lambda_{3,s,t,l}}{\rho}) X_{s,t,l} \tag{75}$$

The gradient of $\tilde{g}$ with respect to U_l is given by:

$$\nabla_{U_l} \tilde{g}(\Theta, U) = \frac{\rho}{n} (U_l^T \Theta_l X_{s,t,l} X_{s,t,l}^T \Theta_l^T - \Theta_l X_{s,t,l} (Q_{s,t,l} - \frac{\Lambda_{3,l}}{\rho})^T)^T \tag{76}$$

We also need to solve the following proximal operators during the iterations:

$$\min_{x \geq 0} \frac{1}{2}\|x - \Theta_l\| + \beta\|x\|_1, \ \min_{x} \frac{1}{2}\|x - U\| + \beta\|U\|_{2,1} \tag{77}$$

with the following closed-form solutions:

$$\text{proj}_1(\Theta) = \max(\Theta - \beta, 0) \tag{78}$$

$$\text{proj}_{2,1}(U_{\cdot,i}) = \max(1 - \beta/\|U_{\cdot,i}\|_2) * U_{\cdot,i} \tag{79}$$

where $U_{\cdot,i} \in \mathbb{R}^{1 \times |L|}$, $i \in \{1, 2, \cdots, d\}$.

Algorithm 1 Update of Θ and U

Input: X,Λ, **and** ρ
Output: Θ and U

1: Initialize Θ, U, $n_g > 0$, and $0 < \gamma < 1$.
2: **repeat**
3: $\quad\nabla_{\Theta_l} \tilde{g}(\Theta, U) \leftarrow$ Equation (75), $\forall l \in L$
4: $\quad\nabla_{U_l} \tilde{g}(\Theta, U) \leftarrow$ Equation (76), $\forall l \in L$
5: $\quad g_{\max} \leftarrow$ max function value in latest n_g iterations.
6: $\quad$**repeat**
7: $\quad\quad\Theta_l' = \text{proj}_1(\Theta_l - \eta\nabla_{\Theta_l}\tilde{g}(\Theta, U))$ via Equation (78), $\forall, l \in L$
8: $\quad\quad U_{\cdot,i}' = \text{proj}_{2,1}(U_{\cdot,i} - \eta\nabla_{U_{\cdot,i}}\tilde{g}(\Theta, U))$ via Equation (79), $\forall i \in \{1, 2, \cdots, d\}$
9: $\quad\quad\delta = c\sum_l^L(\langle\Theta_l' - \Theta_l, \nabla_{\Theta_l}\tilde{g}(\Theta, U)\rangle) + \langle U_l' - U_l, \nabla_{U_l}\tilde{g}(\Theta, U)\rangle + c\lambda_1(\|U'\|_{2,1} - \|U\|_{2,1}) + c\lambda_2\sum_l^L(\|\Theta_l'\|_1 - \|\Theta_l\|_1)$
10: $\quad\quad$**if** $g(\Theta', U') \geq g_{\max} + \delta$ **then**
11: $\quad\quad\quad$break;
12: $\quad\quad$**else**
13: $\quad\quad\quad\eta \leftarrow \eta \cdot \gamma$
14: $\quad\quad$**end if**
15: $\quad$**until** forever
16: $\quad\Delta\Theta_l \leftarrow \Theta_l' - \Theta_l, \Delta U_l \leftarrow U_l' - U_l$
17: $\quad\Delta_g\Theta_l \leftarrow \nabla_{\Theta_l}\tilde{g}(\Theta', U') - \nabla_{\Theta_l}\tilde{g}(\Theta, U)$, $\Delta_g U_l \leftarrow \nabla_{U_l}\tilde{g}(\Theta', U') - \nabla_{U_l}\tilde{g}(\Theta, U)$
18: $\quad\eta \leftarrow \sum_l^L(((\langle\Delta\Theta_l, \Delta\Theta_l\rangle) + (\langle\Delta U_l, \Delta U_l\rangle))/\sum_l^L(((\langle\Delta\Theta_l, \Delta_g\Theta_l\rangle) + (\langle\Delta U_l, \Delta_g U_l\rangle))$
19: $\quad\varepsilon = \max(\max_d \|\text{proj}_{2,1}(U_{\cdot,d}' - \nabla_{U_{\cdot,d}}\tilde{g}(\Theta', U')) - U'\|_\infty, \|\max_l \text{proj}_{2,1}(\Theta' - \nabla_{\Theta_l}\tilde{g}(\Theta', U')) - \Theta_l'\|_\infty)$
20: $\quad\Theta \leftarrow \Theta'; U \leftarrow U'$
21: **until** ε <tolerance

References

[Lu and Zhang 2012] Lu, Z., and Zhang, Y. 2012. An augmented lagrangian approach for sparse principal component analysis. *Mathematical Programming* 1–45.

[Maurer, Pontil, and Romera-Paredes 2013] Maurer, A.; Pontil, M.; and Romera-Paredes, B. 2013. Sparse coding for multitask and transfer learning. In *ICML 2013*, 343–351.

[Tseng and Yun 2009] Tseng, P., and Yun, S. 2009. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming* 117(1):387–423.

[Zhou et al. 2013] Zhou, J.; Lu, Z.; Sun, J.; Yuan, L.; Wang, F.; and Ye, J. 2013. Feafiner: biomarker identification from medical data through feature generalization and selection. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1034–1042. ACM.