Supplementary Material: A probabilistic framework for multi-view feature learning with many-to-many associations via neural networks

A. Proof of Theorem 5.1

Since $g_* : [-M', M']^{2K_*} \to \mathbb{R}$ is a positive definite kernel on a compact set, it follows from Mercer's theorem that there exist positive eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$ and continuous eigenfunctions $\{\phi_k\}_{k=1}^{\infty}$ such that

$$g_*(\boldsymbol{y}_*, \boldsymbol{y}'_*) = \sum_{k=1}^{\infty} \lambda_k \phi_k(\boldsymbol{y}_*)\phi_k(\boldsymbol{y}'_*), \quad \boldsymbol{y}_*, \boldsymbol{y}'_* \in [-M', M']^{K_*},$$

where the convergence is absolute and uniform (Minh et al., 2006). The uniform convergence implies that for any $\varepsilon_1 > 0$ there exists $K_0 \in \mathbb{N}$ such that

$$\sup_{(\boldsymbol{y}_*, \boldsymbol{y}'_*) \in [-M', M']^{2K_*}} \left| g_*(\boldsymbol{y}_*, \boldsymbol{y}'_*) - \sum_{k=1}^{K} \lambda_k \phi_k(\boldsymbol{y}_*)\phi_k(\boldsymbol{y}'_*) \right| < \varepsilon_1, \quad K \geq K_0.$$

This means $g_*(\boldsymbol{y}_*, \boldsymbol{y}'_*) \approx \langle \boldsymbol{\Phi}_K(\boldsymbol{y}_*), \boldsymbol{\Phi}_K(\boldsymbol{y}'_*) \rangle$ for a feature map $\boldsymbol{\Phi}_K(\boldsymbol{y}_*) = (\sqrt{\lambda_k}\phi_k(\boldsymbol{y}_*))_{k=1}^{K}$.

We fix K and consider approximation of $h_k^{(d)}(\boldsymbol{x}) := \sqrt{\lambda_k}\phi_k(f_*^{(d)}(\boldsymbol{x}))$ below. Since $h_k^{(d)}$ are continuous functions on a compact set, there exists $C = C(K) > 0$ such that

$$\sup_{\boldsymbol{x} \in [-M, M]^{p_d}} |h_k^{(d)}(\boldsymbol{x})| < C, \quad k = 1, \ldots, K,\ d = 1, \ldots, D.$$

Let us write the neural networks as $f_{\boldsymbol{\psi}}^{(d)} = (f_1^{(d)}, \ldots, f_K^{(d)})$, where $f_k^{(d)} : \mathbb{R}^{p_d} \to \mathbb{R}$, $d = 1, \ldots, D$, $k = 1, \ldots, K$, are two-layer neural networks with T hidden units. Since $h_k^{(d)}$ are continuous functions, it follows from the universal approximation theorem (Cybenko, 1989; Telgarsky, 2017) that for any $\varepsilon_2 > 0$, there exists $T_0(K) \in \mathbb{N}$ such that

$$\sup_{\boldsymbol{x} \in [-M, M]^{p_d}} |h_k^{(d)}(\boldsymbol{x}) - f_k^{(d)}(\boldsymbol{x})| < \varepsilon_2, \quad k = 1, \ldots, K,\ d = 1, \ldots, D$$

for $T \geq T_0(K)$. Therefore, for all $d, e \in \{1, 2, \ldots, D\}$, we have

$$\begin{aligned}
&\sup_{(\boldsymbol{x}, \boldsymbol{x}') \in [-M, M]^{p_d + p_e}} \left| g_*\left(f_*^{(d)}(\boldsymbol{x}), f_*^{(e)}(\boldsymbol{x}')\right) - \sum_{k=1}^{K} f_k^{(d)}(\boldsymbol{x}) f_k^{(e)}(\boldsymbol{x}') \right| \\
&\quad \leq \sup_{(\boldsymbol{x}, \boldsymbol{x}') \in [-M, M]^{p_d + p_e}} \left| g_*\left(f_*^{(d)}(\boldsymbol{x}), f_*^{(e)}(\boldsymbol{x}')\right) - \sum_{k=1}^{K} h_k^{(d)}(\boldsymbol{x}) h_k^{(e)}(\boldsymbol{x}') \right| \\
&\qquad + \sup_{(\boldsymbol{x}, \boldsymbol{x}') \in [-M, M]^{p_d + p_e}} \left| \sum_{k=1}^{K} h_k^{(d)}(\boldsymbol{x}) \left(h_k^{(e)}(\boldsymbol{x}') - f_k^{(e)}(\boldsymbol{x}')\right) \right| \\
&\qquad + \sup_{(\boldsymbol{x}, \boldsymbol{x}') \in [-M, M]^{p_d + p_e}} \left| \sum_{k=1}^{K} \left(h_k^{(d)}(\boldsymbol{x}) - f_k^{(d)}(\boldsymbol{x})\right) f_k^{(e)}(\boldsymbol{x}') \right| \\
&\quad \leq \sup_{\boldsymbol{y}_*, \boldsymbol{y}'_* \in [-M', M']^{K_*}} \left| g_*\left(\boldsymbol{y}_*, \boldsymbol{y}'_*\right) - \sum_{k=1}^{K} \lambda_k \phi_k(\boldsymbol{y}_*)\phi_k(\boldsymbol{y}'_*) \right| \\
&\qquad + \sum_{k=1}^{K} \sup_{\boldsymbol{x} \in [-M, M]^{p_d}} |h_k^{(d)}(\boldsymbol{x})| \sup_{\boldsymbol{x}' \in [-M, M]^{p_e}} \left| h_k^{(e)}(\boldsymbol{x}') - f_k^{(e)}(\boldsymbol{x}') \right| \\
&\qquad + \sum_{k=1}^{K} \sup_{\boldsymbol{x} \in [-M, M]^{p_d}} \left| h_k^{(d)}(\boldsymbol{x}) - f_k^{(d)}(\boldsymbol{x}) \right| \sup_{\boldsymbol{x}' \in [-M, M]^{p_e}} |f_k^{(e)}(\boldsymbol{x}')| \\
&\quad < \varepsilon_1 + KC\varepsilon_2 + K\varepsilon_2(C + \varepsilon_2).
\end{aligned}$$

By letting $\varepsilon_1 = \varepsilon/2, \varepsilon_2 = \min(C, \varepsilon/(6KC))$, the last formula becomes smaller than ε, thus proving

$$\sup_{(\boldsymbol{x},\boldsymbol{x}')\in[-M,M]^{p_d+p_e}} \left| g_* \left(f_*^{(d)}(\boldsymbol{x}), f_*^{(e)}(\boldsymbol{x}') \right) - \sum_{k=1}^{K} f_k^{(d)}(\boldsymbol{x}) f_k^{(e)}(\boldsymbol{x}') \right| < \varepsilon, \quad d, e = 1, \ldots, D.$$

□

B. Consistency of MLE in PMvGE

In this section, we provide technical details of the argument of Section 5.2.

For proving the consistency of MLE, we introduce the following generative model. Let d_i, $i = 1, \ldots, n$, be random variables independently distributed with the probability $\mathbb{P}(d_i = d) = \eta^{(d)} \in (0, 1)$ where $\sum_{d=1}^{D} \eta^{(d)} = 1$. Data vectors are also treated as random variables. The conditional distribution of $\boldsymbol{x}_i$ given d_i is

$$\boldsymbol{x}_i \mid d_i \overset{\text{indep.}}{\sim} q^{(d_i)}, \quad i = 1, \ldots, n,$$

where $q^{(d_i)}$ is a distribution on a compact support in $\mathbb{R}^{p_{d_i}}$. Let us denote eq. (3) as $\mu_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j, d_i, d_j | \boldsymbol{\alpha}, \boldsymbol{\psi})$ for indicating the dependency on $(\boldsymbol{x}_i, \boldsymbol{x}_j, d_i, d_j)$. The conditional distributions of link weights are already specified in (1) as

$$w_{ij} \mid \boldsymbol{x}_i, \boldsymbol{x}_j, d_i, d_j \overset{\text{indep.}}{\sim} \mathrm{Po}(\mu_{ij}^*), \quad i, j = 1, \ldots, n,$$

where $\mu_{ij}^* = \mu_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j, d_i, d_j | \boldsymbol{\alpha}_*, \boldsymbol{\psi}_*)$ with a true parameter $(\boldsymbol{\alpha}_*, \boldsymbol{\psi}_*)$. Due to the constraints $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\top$ and $w_{ij} = 0$ $((d_i, d_j) \notin \mathcal{D})$, the vector of free parameters in $\boldsymbol{\alpha}$ is $\boldsymbol{\alpha}_{\mathcal{D}} := \{\alpha^{(d,e)}\}_{(d,e)\in\mathcal{D}, d\le e} \in \mathbb{R}_{\ge 0}^{|\mathcal{D}|}$, and we write $\boldsymbol{\theta} := (\boldsymbol{\alpha}_{\mathcal{D}}, \boldsymbol{\psi})$. Let $\tilde{n} := |\mathcal{I}_n| = O(n^2)$ denote the number of terms in the sum of $\ell_n(\boldsymbol{\theta})$ in eq. (6). Then the expected value of $\tilde{n}^{-1}\ell_n(\boldsymbol{\theta})$ under the generative model with the true parameter $\boldsymbol{\theta}_*$ is expressed as

$$\ell(\boldsymbol{\theta}) := E_{\boldsymbol{x}_1,\boldsymbol{x}_2,d_1,d_2} \Big[\mu_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2, d_1, d_2 | \boldsymbol{\theta}_*) \log \mu_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2, d_1, d_2 | \boldsymbol{\theta}) - \mu_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2, d_1, d_2 | \boldsymbol{\theta}) \Big].$$

If it were the case of i.i.d. observations of sample size n, we would have that $\tilde{n}^{-1}\ell_n(\boldsymbol{\theta})$ converges to $\ell(\boldsymbol{\theta})$ as $n \to \infty$ from the law of large numbers. In Theorem B.1, we actually prove the uniform convergence in probability, but we have to pay careful attention to the fact that n^2 observations of $(\boldsymbol{x}_i, \boldsymbol{x}_j, d_i, d_j)$ are not independent when indices overlap.

Theorem B.1 Let us assume that the parameter space of $\boldsymbol{\theta}$ is $\boldsymbol{\Theta} := [\delta, 1/\delta]^{|\mathcal{D}|} \times \boldsymbol{\Psi}$, where $\delta \in (0, 1)$ is a sufficiently small constant and $\boldsymbol{\Psi} \subset \mathbb{R}^q$ is a compact set. Assume also that the transformations $f_{\boldsymbol{\psi}}^{(d)}(\boldsymbol{x})$, $d = 1, \ldots, D$, are Lipschitz continuous with respect to $(\boldsymbol{\psi}, \boldsymbol{x})$. Then we have, as $n \to \infty$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})| \xrightarrow{p} 0. \tag{13}$$

Proof B.1 We refer to Corollary 2.2 in Newey (1991). This corollary shows $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\hat{Q}_n(\boldsymbol{\theta}) - \bar{Q}_n(\boldsymbol{\theta})| = o_p(1)$ under general setting of $\hat{Q}_n(\boldsymbol{\theta})$ and $\bar{Q}_n(\boldsymbol{\theta})$. Here we consider the case of $\hat{Q}_n(\boldsymbol{\theta}) = \tilde{n}^{-1}\ell_n(\boldsymbol{\theta})$ and $\bar{Q}_n(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta})$. For showing (13), the four conditions of the corollary are written as follows. (i) $\boldsymbol{\Theta}$ is compact, (ii) $\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}) \xrightarrow{p} 0$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, (iii) $\ell(\boldsymbol{\theta})$ is continuous, (iv) there exists $B_n = O_p(1)$ such that $|\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) - \tilde{n}^{-1}\ell_n(\boldsymbol{\theta}')| \le B_n \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$. The two conditions (i) and (iii) hold obviously, and thus we verify (ii) and (iv) below.

Before verifying (ii) and (iv), we first consider an array $\boldsymbol{Z} := (Z_{ij})$ of random variables $Z_{ij} \in \mathcal{Z}$, $(i, j) \in \mathcal{I}_n$, and a bounded and continuous function $h : \mathcal{Z} \to \mathbb{R}$. We assume that $\mathcal{Z} \subset \mathbb{R}$ is a compact set, and Z_{ij} is independent of Z_{kl} if

$k, l \in \mathcal{R}_n(i,j) := \{(k,l) \in \mathcal{I}_n \mid k,l \in \{1,\dots,n\} \setminus \{i,j\}\}$ for all $(i,j) \in \mathcal{I}_n$. Then we have

$$\begin{aligned} V_{\boldsymbol{Z}}\left[\frac{1}{\tilde{n}}\sum_{(i,j)\in\mathcal{I}_n} h(Z_{ij})\right] &= E_{\boldsymbol{Z}}\left[\left(\frac{1}{\tilde{n}}\sum_{(i,j)\in\mathcal{I}_n} h(Z_{ij})\right)^2\right] - E_{\boldsymbol{Z}}\left[\frac{1}{\tilde{n}}\sum_{(i,j)\in\mathcal{I}_n} h(Z_{ij})\right]^2 \\ &= \frac{1}{\tilde{n}^2}\left\{\sum_{(i,j)\in\mathcal{I}_n}\sum_{(k,l)\in\mathcal{I}_n} E_{\boldsymbol{Z}}\left[h(Z_{ij})h(Z_{kl})\right] - \left(\sum_{(i,j)\in\mathcal{I}_n} E_{\boldsymbol{Z}}\left[h(Z_{ij})\right]\right)^2\right\} \\ &= \frac{1}{\tilde{n}^2}\sum_{(i,j)\in\mathcal{I}_n}\sum_{(k,l)\in\mathcal{I}_n\setminus\mathcal{R}_n(i,j)} \left(E_{\boldsymbol{Z}}[h(Z_{ij})h(Z_{kl})] - E_{\boldsymbol{Z}}[h(Z_{ij})]E_{\boldsymbol{Z}}[h(Z_{kl})]\right). \end{aligned}$$

By considering $|\mathcal{I}_n \setminus \mathcal{R}_n(i,j)| = O(n)$, the last formula is $O(\tilde{n}^{-2}\cdot\tilde{n}\cdot n) = O(n^{-1})$. Therefore,

$$V_{\boldsymbol{Z}}\left[\frac{1}{\tilde{n}}\sum_{(i,j)\in\mathcal{I}_n} h(Z_{ij})\right] = O(n^{-1}). \tag{14}$$

Next we evaluate the variance of $\tilde{n}^{-1}\ell_n(\boldsymbol{\theta})$ to show (ii). Denoting $\boldsymbol{W} := (w_{ij}), \boldsymbol{X} := (\boldsymbol{x}_i), \boldsymbol{d} := (d_i)$,

$$\begin{aligned} V_{\boldsymbol{W},\boldsymbol{X},\boldsymbol{d}}[\tilde{n}^{-1}\ell_n(\boldsymbol{\theta})] &= E_{\boldsymbol{X},\boldsymbol{d}}[V_{\boldsymbol{W}}[\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) \mid \boldsymbol{X},\boldsymbol{d}]] + V_{\boldsymbol{X},\boldsymbol{d}}[E_{\boldsymbol{W}}[\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) \mid \boldsymbol{X},\boldsymbol{d}]] \\ &= E_{\boldsymbol{X},\boldsymbol{d}}\left[\frac{1}{\tilde{n}^2}\sum_{(i,j)\in\mathcal{I}_n}\mu_{ij}(\boldsymbol{\theta}_*)(\log\mu_{ij}(\boldsymbol{\theta}))^2\right] + V_{\boldsymbol{X},\boldsymbol{d}}\left[\frac{1}{\tilde{n}}\sum_{(i,j)\in\mathcal{I}_n}(\mu_{ij}(\boldsymbol{\theta}_*)\log\mu_{ij}(\boldsymbol{\theta}) - \mu_{ij}(\boldsymbol{\theta}))\right], \end{aligned}$$

for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The first term in the last formula is $O(\tilde{n}^{-1}\cdot\tilde{n}) = O(\tilde{n}^{-1}) = o(1)$, and the second term is $O(n^{-1}) = o(1)$ by applying eq. (14) with $Z_{ij} := (\boldsymbol{x}_i, \boldsymbol{x}_j, d_i, d_j)$, $h(Z_{ij}) = \mu_{ij}(\boldsymbol{\theta}_*)\log\mu_{ij}(\boldsymbol{\theta}) - \mu_{ij}(\boldsymbol{\theta})$. Therefore, $V_{\boldsymbol{W},\boldsymbol{X},\boldsymbol{d}}[\tilde{n}^{-1}\ell_n(\boldsymbol{\theta})] = o(1)$ and Chebyshev's inequality implies the pointwise convergence $\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) \xrightarrow{p} \ell(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where $\ell(\boldsymbol{\theta}) = E_{\boldsymbol{W},\boldsymbol{X},\boldsymbol{d}}[\tilde{n}^{-1}\ell_n(\boldsymbol{\theta})] = E_{\boldsymbol{x}_1,\boldsymbol{x}_2,d_1,d_2}[\mu_{12}(\boldsymbol{\theta}_*)\log\mu_{12}(\boldsymbol{\theta}) - \mu_{12}(\boldsymbol{\theta})]$. Thus, condition (ii) holds.

Finally, we work on condition (iv). Since $\mu_{ij}(\boldsymbol{\theta})$ is a composite function of C^1-functions on $\boldsymbol{\Theta}$, $\mu_{ij}(\boldsymbol{\theta})$ is Lipschitz continuous. The Lipschitz continuity of $\mu_{ij}(\boldsymbol{\theta})$ and $\mu_{ij}(\boldsymbol{\theta}) > 0$ ($\boldsymbol{\theta} \in \boldsymbol{\Theta}$) indicates the Lipschitz continuity of $\log\mu_{ij}(\boldsymbol{\theta})$. Therefore, there exist $M_1, M_2 > 0$ such that

$$\begin{aligned} |\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) - \tilde{n}^{-1}\ell_n(\boldsymbol{\theta}')| &\le \left|\frac{1}{\tilde{n}}\sum_{(i,j)\in\mathcal{I}_n} w_{ij}(\log\mu_{ij}(\boldsymbol{\theta}) - \log\mu_{ij}(\boldsymbol{\theta}')) - \frac{1}{\tilde{n}}\sum_{(i,j)\in\mathcal{I}_n}(\mu_{ij}(\boldsymbol{\theta}) - \mu_{ij}(\boldsymbol{\theta}'))\right| \\ &\le \frac{1}{\tilde{n}}\sum_{(i,j)\in\mathcal{I}_n} w_{ij}|\log\mu_{ij}(\boldsymbol{\theta}) - \log\mu_{ij}(\boldsymbol{\theta}')| + \frac{1}{\tilde{n}}\sum_{(i,j)\in\mathcal{I}_n}|\mu_{ij}(\boldsymbol{\theta}) - \mu_{ij}(\boldsymbol{\theta}')| \\ &\le M_1\left(\tilde{n}^{-1}\sum_{(i,j)\in\mathcal{I}_n} w_{ij}\right)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 + M_2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2. \end{aligned}$$

Denoting by $B_n := M_1\cdot\tilde{n}^{-1}\sum_{(i,j)\in\mathcal{I}_n} w_{ij} + M_2$, we have

$$|\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) - \tilde{n}^{-1}\ell_n(\boldsymbol{\theta}')| \le B_n\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.$$

Since $\tilde{n}^{-1}\sum_{(i,j)\in\mathcal{I}_n} w_{ij} = O_p(1)$, the law of large numbers indicates $B_n = O_p(1)$. Thus, condition (iv) holds. □

Noticing that $\boldsymbol{\theta}_*$ is a maximizer of $\ell(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_n$ is a maximizer of $\ell_n(\boldsymbol{\theta})$, we would have the desired result $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_*$ by combining Theorem B.1 and continuity of $\ell(\boldsymbol{\theta})$. However it does not hold unfortunately. Instead, we define the set of parameter values equivalent to $\boldsymbol{\theta}_*$ as $\boldsymbol{\Theta}_* := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \mid \ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_*)\}$. Every $\boldsymbol{\theta} \in \boldsymbol{\Theta}_*$ gives the correct probability of link weights, because $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_*)$ holds if and only if $\mu_{12}(\boldsymbol{\theta}) = \mu_{12}(\boldsymbol{\theta}_*)$ almost surely w.r.t. $(\boldsymbol{x}_1, \boldsymbol{x}_2, d_1, d_2)$. With this setting, the theorem below states that $\hat{\boldsymbol{\Theta}}_n$ converges to $\boldsymbol{\Theta}_*$ in probability. This indicates that, $\hat{\boldsymbol{\theta}}_n$ will represent the true probability model for sufficiently large n.

Theorem B.2 Let $d_{\mathrm{H}}(\cdot,\cdot)$ denote the Hausdorff distance defined as max-min L^2-distance between two sets. We assume the same conditions as in Theorem B.1. Then we have, as $n \to \infty$,

$$d_H(\hat{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_*) \xrightarrow{p} 0. \tag{15}$$

Proof B.2 We refer to the case (1) of Theorem 3.1 in Chernozhukov et al. (2007) with $\hat{c} = 1$ under the condition C.1 This theorem shows, for general setting of $\hat{\boldsymbol{\Theta}}_I, \boldsymbol{\Theta}_I$, that

$$d_H(\hat{\boldsymbol{\Theta}}_I, \boldsymbol{\Theta}_I) \xrightarrow{p} 0. \tag{16}$$

Here $\hat{\boldsymbol{\Theta}}_I := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \mid \hat{Q}_n(\boldsymbol{\theta}) \le 1/a_n\}$ and $\boldsymbol{\Theta}_I := \arg\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q(\boldsymbol{\theta})$, where $\hat{Q}_n(\boldsymbol{\theta}), Q(\boldsymbol{\theta})$ are general functions satisfying $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_I} Q_n(\boldsymbol{\theta}) = o_p(1/a_n)$, and $a_n \to \infty$.

For proving (15), we consider the case of $\hat{Q}_n(\boldsymbol{\theta}) = -\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{n}^{-1}\ell_n(\boldsymbol{\theta})$, $Q(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta})$. The condition C.1 for (16) is re-written as follows. (i) $\boldsymbol{\Theta}$ is a (non-empty) compact set, (ii) $\tilde{n}^{-1}\ell_n(\boldsymbol{\theta})$ and $\ell(\boldsymbol{\theta})$ are continuous, (iii) $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})| \xrightarrow{p} 0$, and (iv) $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_*}(-\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{n}^{-1}\ell_n(\boldsymbol{\theta})) \xrightarrow{p} 0$. The conditions (i), (ii) are obvious. (iii) is shown in Theorem B.1. (iv) is verified by

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_*}\left(-\tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{n}^{-1}\ell_n(\boldsymbol{\theta})\right) = -\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_I} \tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) \xrightarrow{p} -\ell(\boldsymbol{\theta}_*) + \ell(\boldsymbol{\theta}_*) = 0,$$

where $\boldsymbol{\theta}_*$ is an element of $\boldsymbol{\Theta}_I$. Thus, (16) holds.

Next, we consider two sets $\hat{\boldsymbol{\Theta}}_n = \arg\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{n}^{-1}\ell_n(\boldsymbol{\theta}) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \mid Q_n(\boldsymbol{\theta}) = 0\}$ and $\hat{\boldsymbol{\Theta}}_I$. Since these sets satisfy $Q_n(\boldsymbol{\theta}) = 0$ $(\boldsymbol{\theta} \in \hat{\boldsymbol{\Theta}}_n), Q_n(\boldsymbol{\theta}') \le 1/a_n$ $(\boldsymbol{\theta}' \in \hat{\boldsymbol{\Theta}}_I)$ and $1/a_n \to 0$ as $n \to \infty$, we have $d_H(\hat{\boldsymbol{\Theta}}_n, \hat{\boldsymbol{\Theta}}_I) \xrightarrow{p} 0$. It follows from this convergence and (16) that, by noticing $\boldsymbol{\Theta}_* = \boldsymbol{\Theta}_I$,

$$d_H(\hat{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_*) = d_H(\hat{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_I) \le d_H(\hat{\boldsymbol{\Theta}}_n, \hat{\boldsymbol{\Theta}}_I) + d_H(\hat{\boldsymbol{\Theta}}_I, \boldsymbol{\Theta}_I) \xrightarrow{p} 0,$$

thus (15) holds. □

C. CDMCA is approximated by PMvGE with linear transformations

We argue an approximate relation between CDMCA and PMvGE, which is briefly explained in Section 3.6. In the below, we will derive the solution $\hat{\boldsymbol{\psi}}_{\text{CDMCA}}$ of a slightly modified version of CDMCA, and an approximate solution $\hat{\boldsymbol{\psi}}_{\text{Apr.PMvGE}}$ of PMvGE with linear transformations. We then show that these two solutions are equivalent up to a scaling in each axis of the shared space.

C.1. Solution of a modified CDMCA

The original CDMCA imposes the quadratic constraint (8) for maximizing the objective function (7). Here we replace w_{ij} in (8) with δ_{ij} so that the constraint becomes

$$\sum_{i=1}^{n} \boldsymbol{\psi}^{(d_i)\top} \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\psi}^{(d_i)} = \boldsymbol{I}.$$

This modification changes the scaling in the solution, but the computation below is essentially the same as that in Shimodaira (2016). Let us define the augmented data vector, called "simple coding" (Shimodaira, 2016), $\tilde{\boldsymbol{x}}_i := (\mathbf{0}_{p_1}, \ldots, \mathbf{0}_{p_{d_i-1}}, \boldsymbol{x}_i, \mathbf{0}_{p_{d_i+1}}, \ldots, \mathbf{0}_{p_D}) \in \mathbb{R}^p$ where $p := p_1 + p_2 + \cdots + p_D$. Now, data matrix is $\boldsymbol{X} := (\tilde{\boldsymbol{x}}_1^\top, \tilde{\boldsymbol{x}}_2^\top, \ldots, \tilde{\boldsymbol{x}}_n^\top)^\top \in \mathbb{R}^{n\times p}$, and the parameter matrix is $\boldsymbol{\psi} := (\boldsymbol{\psi}^{(1)\top}, \boldsymbol{\psi}^{(2)\top}, \ldots, \boldsymbol{\psi}^{(D)\top})^\top \in \mathbb{R}^{p\times K}$. With this augmented representation, the D-view embedding is now interpreted as a 1-view embedding. CDMCA maximizes the objective function

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\langle \boldsymbol{\psi}^{(d_i)\top}\boldsymbol{x}_i, \boldsymbol{\psi}^{(d_j)\top}\boldsymbol{x}_j\rangle = \mathrm{tr}\left(\boldsymbol{\psi}^\top \boldsymbol{H}\boldsymbol{\psi}\right) \quad (\boldsymbol{H} := \boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}),$$

with respect to $\boldsymbol{\psi}$ under constraint $\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{\psi} = \boldsymbol{I}$ where $\boldsymbol{G} := \boldsymbol{X}^\top\boldsymbol{X}$. Let $\boldsymbol{U}_K$ be the matrix composed of the top-K eigenvectors of $\boldsymbol{G}^{-1/2}\boldsymbol{H}\boldsymbol{G}^{-1/2}$. Then the solution of the modified CDMCA is

$$\hat{\boldsymbol{\psi}}_{\text{CDMCA}} := \boldsymbol{G}^{-1/2}\boldsymbol{U}_K.$$

C.2. Approximate solution of PMvGE with linear transformations

MLE of PMvGE maximizes $\ell_n(\boldsymbol{\alpha}, \boldsymbol{\psi})$ defined in (6). Here we modify it by adding an extra term as $\tilde{\ell}_n(\boldsymbol{\alpha}, \boldsymbol{\psi}) := \ell_n(\boldsymbol{\alpha}, \boldsymbol{\psi}) - \frac{1}{2}\sum_{i:(d_i,d_i)\in\mathcal{D}} \mu_{ii}(\boldsymbol{\alpha}, \boldsymbol{\psi})$. The difference approaches zero for large n, because $|\ell_n(\boldsymbol{\alpha}, \boldsymbol{\psi}) - \tilde{\ell}_n(\boldsymbol{\alpha}, \boldsymbol{\psi})|/|\ell_n(\boldsymbol{\alpha}, \boldsymbol{\psi})| = O(n^{-1})$. Since the parameter $\boldsymbol{\alpha}$ is not considered in CDMCA, we assume $\mathcal{D} := \{\text{all pairs of views}\}$ and $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}^{(de)}), \bar{\alpha}^{(de)} \equiv \alpha_0 > 0\ (\forall d, e)$. We further assume that the transformation of PMvGE is linear: $f^{(d)}_{\boldsymbol{\psi}}(\boldsymbol{x}) = \boldsymbol{\psi}^{(d)\top}\boldsymbol{x}\ (\forall d)$, and data vectors in each view are centered. With this setting, we will show that the maximizer of a quadratic approximation of $\tilde{\ell}_n$ is equivalent to CDMCA.

To rewrite the likelihood function as $\tilde{\ell}_n(\bar{\boldsymbol{\alpha}}, \boldsymbol{\psi}) = \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n S_{ij}(g_{ij}(\boldsymbol{\psi}))$, we define $g_{ij}(\boldsymbol{\psi}) := \langle \boldsymbol{\psi}^{(d_i)\top}\boldsymbol{x}_i, \boldsymbol{\psi}^{(d_j)\top}\boldsymbol{x}_j\rangle$ and $S_{ij}(g) := w_{ij}\log(\alpha_0\exp(g)) - \alpha_0\exp(g)$, $g \in \mathbb{R}$. Since $S_{ij}(g)$ is approximated quadratically around $g = 0$ by

$$S^Q_{ij}(g) = \alpha_0\left\{-\frac{1}{2}g^2 + \left(\frac{w_{ij}}{\alpha_0} - 1\right)g\right\} + S_{ij}(0),$$

$\tilde{\ell}_n(\bar{\boldsymbol{\alpha}}, \boldsymbol{\psi})$ is approximated quadratically by

$$\begin{aligned}\tilde{\ell}^Q_n(\boldsymbol{\psi}) &:= \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n S^Q_{ij}(g_{ij}(\boldsymbol{\psi})) \\ &= \alpha_0\left\{-\frac{1}{2}\mathrm{tr}\left((\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{\psi})^2\right) + \frac{1}{\alpha_0}\mathrm{tr}\left(\boldsymbol{\psi}^\top\boldsymbol{H}\boldsymbol{\psi}\right) - \underbrace{\mathrm{tr}\left(\boldsymbol{\psi}^\top\boldsymbol{X}^\top\mathbf{1}_n\mathbf{1}_n^\top\boldsymbol{X}\boldsymbol{\psi}\right)}_{=0\ (\because\{\boldsymbol{x}_i\}\text{ is centered.})}\right\} + \underbrace{\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n S_{ij}(0)}_{\text{Const.}}, \qquad (17)\end{aligned}$$

where $\boldsymbol{G} = \boldsymbol{X}^\top\boldsymbol{X}, \boldsymbol{H} = \boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}$. The function $\tilde{\ell}^Q_n(\boldsymbol{\psi})$ has rotational degrees of freedom: $\tilde{\ell}^Q_n(\boldsymbol{\psi}) = \tilde{\ell}^Q_n(\boldsymbol{\psi}\boldsymbol{O})$ for any orthogonal matrix $\boldsymbol{O} \in \mathbb{R}^{K\times K}$. Thus, we impose an additional constraint $\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{\psi} = \boldsymbol{\Gamma}_K := \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_K)$ for any $(\gamma_1, \ldots, \gamma_K) \in \mathbb{R}^K_{\geq 0}$. $\boldsymbol{\psi}$ satisfying this constraint is written as $\boldsymbol{\psi} = \boldsymbol{G}^{-1/2}\boldsymbol{V}_K\boldsymbol{\Gamma}_K^{1/2}$ where $\boldsymbol{V}_K \in \mathbb{R}^{P\times K}$ is a column-orthogonal matrix such that $\boldsymbol{V}_K^\top\boldsymbol{V}_K = \boldsymbol{I}$. By substituting $\boldsymbol{\psi} = \boldsymbol{G}^{-1/2}\boldsymbol{V}_K\boldsymbol{\Gamma}_K^{1/2}$ into eq. (17), we have

$$\tilde{\ell}^Q_n(\boldsymbol{\psi}) = \alpha_0\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Gamma}_K^2\right) + \frac{1}{\alpha_0}\mathrm{tr}\left(\boldsymbol{\Gamma}_K\boldsymbol{S}_K\right)\right\} + \text{Const.} = \frac{\alpha_0}{2}\left\{\left\|\frac{1}{\alpha_0}\boldsymbol{S}_K\right\|_{\mathrm{F}}^2 - \left\|\boldsymbol{\Gamma}_K - \frac{1}{\alpha_0}\boldsymbol{S}_K\right\|_{\mathrm{F}}^2\right\} + \text{Const.}, \qquad (18)$$

where $\boldsymbol{S}_K = \boldsymbol{V}_K^\top\boldsymbol{G}^{-1/2}\boldsymbol{H}\boldsymbol{G}^{-1/2}\boldsymbol{V}_K$ and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. This objective function is maximized when $\boldsymbol{\Gamma}_K = \frac{1}{\alpha_0}\boldsymbol{S}_K$ and $\boldsymbol{V}_K = \boldsymbol{U}_K$, because $\min_{\boldsymbol{\Gamma}_K}\|\boldsymbol{\Gamma}_K - \frac{1}{\alpha_0}\boldsymbol{S}_K\|_{\mathrm{F}}^2 = 0$ is achieved by $\boldsymbol{\Gamma}_K = \frac{1}{\alpha_0}\boldsymbol{S}_K$, and $\max_{\boldsymbol{V}_K}\|\boldsymbol{S}_K\|_{\mathrm{F}}^2$ is achieved by $\boldsymbol{V}_K = \boldsymbol{U}_K$ where $\boldsymbol{U}_K$ is the matrix composed of the top-K eigenvectors of $\boldsymbol{G}^{-1/2}\boldsymbol{H}\boldsymbol{G}^{-1/2}$. Therefore, $\tilde{\ell}^Q_n(\boldsymbol{\psi})$ is maximized by

$$\hat{\boldsymbol{\psi}}_{\text{Apr.PMvGE}} = \boldsymbol{G}^{-1/2}\boldsymbol{U}_K\boldsymbol{\Gamma}_K^{1/2}.$$

By substituting $\boldsymbol{V}_K = \boldsymbol{U}_K$ into $\boldsymbol{\Gamma}_K = \frac{1}{\alpha_0}\boldsymbol{S}_K$, we verify that $\boldsymbol{\Gamma}_K$ is a diagonal matrix with $\gamma_k := \lambda_k/\alpha_0$ where λ_k is the k-th largest eigenvalue of $\boldsymbol{G}^{-1/2}\boldsymbol{H}\boldsymbol{G}^{-1/2}$ $(k = 1, 2, \ldots, K)$.

C.3. Equivalence of the two solutions up to a scaling

By comparing the two solutions, we have

$$\hat{\boldsymbol{\psi}}_{\text{Apr.PMvGE}} = \hat{\boldsymbol{\psi}}_{\text{CDMCA}}\boldsymbol{\Gamma}_K^{1/2}. \qquad (19)$$

This simply means that each axis in the shared space is scaled by the factor $\sqrt{\gamma_k}$, $k = 1, \ldots, K$. Let $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{y}}'$ be feature vectors in the shared space computed by the approximate PMvGE with linear transformations, and $\boldsymbol{y}$, $\boldsymbol{y}'$ be feature vectors in the shared space computed by the modified CDMCA. Then the inner product is weighted in PMvGE as

$$\langle\hat{\boldsymbol{y}}, \hat{\boldsymbol{y}}'\rangle = \sum_{k=1}^K \hat{y}_k\hat{y}'_k = \sum_{k=1}^K \gamma_k y_k y'_k.$$

References

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and Confidence Regions for Parameter Sets in Econometric Models. *Econometrica*, 75(5):1243–1284.

Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314.

Minh, H. Q., Niyogi, P., and Yao, Y. (2006). Mercer's Theorem, Feature Maps, and Smoothing. In *International Conference on Computational Learning Theory*, pages 154–168. Springer.

Newey, W. K. (1991). Uniform Convergence in Probability and Stochastic Equicontinuity. *Econometrica*, pages 1161–1167.

Shimodaira, H. (2016). Cross-validation of matching correlation analysis by resampling matching weights. *Neural Networks*, 75:126–140.

Telgarsky, M. (2017). Neural networks and rational functions. In Precup, D. and Teh, Y. W., editors, *Proceedings of the International Conference on Machine Learning (ICML)*.