IEEE Transactions on Cybernetics

Localized Incomplete Multiple Kernel k-means with Matrix-induced Regularization

Journal:	*IEEE Transactions on Cybernetics*
Manuscript ID	CYB-E-2021-01-0012
Manuscript Type:	Regular Paper
Date Submitted by the Author:	13-Jan-2021
Complete List of Authors:	Xia, Jingyuan; Imperial College London, Department of Electric and Electronic Engineering Li, Miaomiao; National University of Defense Technology, School of Computer Xu, Huiying; Zhejiang Normal University, College of Mathematics and Computer Science Liao, Qing; Harbin Institute of Technology Shenzhen, Computer Sience Zhu, Xinzhong; Zhejiang Normal University, College of Mathematics, Physics and Information Engineering LIU, Xinwang; National University of Defense Technology, College of Computer
Keywords:	multiple kernel clustering, multi-view clustering, clustering ensemble

Localized Incomplete Multiple Kernel k-means with Matrix-induced Regularization

Jingyuan Xia, Miaomiao Li, Huiying Xu, Qing Liao, Xinzhong Zhu, Xinwang Liu, *Senior Member, IEEE*

***Abstract*—Localized incomplete multiple kernel k-means (LI-MKKM) is recently put forward to boost the clustering accuracy via optimally utilizing a quantity of pre-specified incomplete base kernel matrices. Despite achieving significant achievement in a variety of applications, we find out that LI-MKKM *does not sufficiently consider the diversity and the complementary of the base kernels*. This could make the imputation of incomplete kernels less effective, and vice versa degrades on the subsequent clustering. To tackle these problems, an improved LI-MKKM termed as LI-MKKM-MR is proposed by incorporating a matrix-induced regularization term to handle the correlation among base kernels. The incorporated regularization term is beneficial to decrease the probability of simultaneously selecting two similar kernels and increase the probability of selecting two kernels with moderate differences. After that, we establish a three-step iterative algorithm to solve the corresponding optimization objective and analyze its convergence. Moreover, we theoretically show that the local kernel alignment is a special case of its global one with normalizing each base kernel matrices. Based on the above observation, the generalization error bound of the proposed algorithm is derived to theoretically justify its effectiveness. Finally, extensive experiments on several public datasets have been conducted to evaluate the clustering performance of the LI-MKKM-MR. As indicated, the experimental results have demonstrated that our algorithm consistently outperforms the state-of-the-art ones, verifying the superior performance of the proposed algorithm.**

***Index Terms*—multiple view learning, multiple kernel clustering, incomplete kernel learning**

I. Introduction

Multiple kernel clustering (MKC) [1–8] sufficiently integrates a number of pre-calculated base kernel matrices to group samples into clusters, where similar samples are in the same cluster while dissimilar ones are partitioned into different ones. MKC has attracted much attention of the data mining researchers and has been widely studied in recent years [9–17]. The seminal work in [9] extends the multiple kernel learning from supervised learning to unsupervised learning, and proposes a margin-based MKC algorithm. It jointly optimizes the optimal kernel, the maximum margin hyperplane and the optimal clustering labels. The widely used kernel k-means method has been extended in [18] for clustering analysis, where an optimal kernel is learned from multiple data sources. Similarly, the work in [12] extends existing multiple kernel k-means algorithm (MKKM) by designing a localized MKKM one in order to well utilize the characteristics of each individual sample. To enhance the robustness of existing MKKM algorithms to noisy data, [13] proposes a robust MKKM algorithm by substituting the widely adopted squared error in existing k-means with an $\ell_{2,1}$-norm one, and simultaneously optimizes the best combination of kernels. To increase the diversity and decrease the redundancy of the selected base kernels, the recent work in [14] extends existing MKKM algorithms by designing a matrix-induced regularization term to sufficiently explore the correlation among pre-specified base kernels. More recently, an optimal neighborhood kernel clustering (ONKC) algorithm is proposed in [19], where the representability of the optimal kernel to learn is largely boosted and the negotiation between kernel learning and clustering is also reinforced. The aforementioned MKC algorithms have been applied into many cases and reached a superior performance [20–23].

As observed, these MKC algorithms share a common assumption: *all the pre-specified base kernels are complete*. Nevertheless, in some real world applications, some views of a sample are usually not collected due to various reasons [24, 25]. To address this issue, the work in the literature proposes to firstly impute the missing elements in base kernel matrices with imputation methods and then performs existing multiple kernel clustering on these imputed kernels. Several commonly used filling methods include zero-filling, mean value filling, k-nearest-neighbor filling, expectation-maximization (EM) filling [26], as well as several recently proposed to matrix imputation [27–30]. One disadvantage existing in the aforementioned "two-stage" algorithms is that the imputation is separated from the subsequent clustering. As a result, this may not be conducive to mutual negotiation between the imputation and clustering to reach the best performance. To overcome the above issue, the more recent literature [31–33] advocates to unify the learning procedure of imputation and clustering into a common framework, with the aim to learn an optimal imputation that best severe for the clustering tasks.

Although demonstrating superior clustering results in several practical applications, we find that these work *does not sufficiently consider the redundancy and diversity among pre-*

J. Xia is with Department of Electric and Electronic Engineering, Imperial College London, London, SW72AZ, UK.

M. Li and X. Liu are with College of Computer, National University of Defense Technology, Changsha, 410073, China. M. Li is also with college of electronic information and electrical engineering, Changsha University, Changsha, 410073, China.

H. Xu and X. Zhu are with the College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua 321004, China. X. Zhu is also with the Research Institute of Ningbo Cixing Co. Ltd, Ningbo 315336, China. (E-mail: {xyh, zxz}@zjnu.edu.cn).

Q. Liao is with the Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, China.

J. Xia, M. Li and H. Xu contribute equally to this work.

Corresponding authors: Xinzhong Zhu and Xinwang Liu (E-mail: xinwangliu@nudt.edu.cn).

Manuscript received January, 2021.

specified kernel matrices when performing incomplete multiple kernel clustering. This could lead to high redundancy and low diversity among the selected kernels [14], making the utilization ratio of these base kernel matrices insufficient and conversely decreasing the accuracy of clustering tasks. In our work, a localized incomplete multiple kernel k-means with matrix-induced regularization (LI-MKKM-MR) is propsoed to address the above-mentioned issue. By incorporating a matrix-induced regularization, LI-MKKM-MR is able to avoid selecting two similar kernel matrices simultaneously and increase the probability of selecting two kernel matrices with large diversity, making the base kernels better utilized for clustering. In addition, it inherits the advantage of LI-MKKM which only requires that the similarity of each sample to its top k-nearest neighbours be optimally aligned with the corresponding patch of the whole ideal similarity. This is helpful for LI-MKKM-MR to pay more attention on closer pairwise sample similarities that shall be put together, and prevents involving unreliable similarity evaluation for farther sample pairs. Furthermore, a three-step iterative optimization algorithm is designed to solve the corresponding optimization objective and its convergence has also been analyzed. After that, the generalization error bound of the clustering algorithm is derived, which theoretically guarantees its effectiveness. Comprehensive experiments on several public datasets have been conducted to evaluate the clustering performance of the proposed LI-MKKM-MR. As demonstrated, LI-MKKM-MR significantly and consistently outperforms existing two-step-based algorithms and the newly proposed algorithm [33]. Extensive experimental results have demonstrated the superiority of involving the matrix-induced regularization.

To summarize, this work makes the following major contributions.

- This is the first attempt to identify the kernel redundancy probelm in *incomplete multiple kernel clustering*. We then introduce a new algorithm to improve LI-MKKM by integrating a matrix-induced regularization to select low-redundant and high-diverse kernel matrices, and carefully establish three-step iterative algorithm to solve the corresponding optimization objective.
- We build the theoretical connection between global and local kernel alignment criteria, then we further derive the generalization error bound of the proposed LI-MKKM-MR, which theoretically justifies its effectiveness.
- Comprehensive experiments on ten public datasets have demonstrated that our LI-MKKM-MR achieves the state-of-the-art performance compared with existing advanced algorithms. This considerably verifies our identification of the aforementioned issue and the effectiveness of our solution.

Finally, we clarify the differences between LI-MKKM-MR and several recently proposed related work [14, 32]. The differences between LI-MKKM [32] and LI-MKKM-MR can be summarized from the following three aspects: i) LI-MKKM [32] *does not sufficiently consider the diversity and the complementarity of these incomplete base kernels*. This could make the imputation of incomplete kernels less effective, and incur the adverse effect on the subsequent clustering. Differently, LI-MKKM-MR is proposed by incorporating a matrix-induced regularization which is helpful to reduce the probability of simultaneously selecting two similar kernels and increase the probability of selecting two kernels with moderate differences, making the base kernels better utilized for clustering. ii) Compared with LI-MKKM [32], LI-MKKM-MR provides the generalization error analysis which measures the clustering performance of the learned clusters in training procedure on unseen samples. This theoretically justifies the effectiveness of the proposed LI-MKKM-MR. iii) As observed from the experimental results in Section V, LI-MKKM-MR significantly improves the clustering performance of LI-MKKM [32] in various benchmark datasets, which well validates our identification of the aforementioned issue in LI-MKKM and the effectiveness of our solution. We then summarize the differences between [14] and our work from the following aspects: In [14], a matrix-induced regularization is proposed to solve the kernel redundancy in multiple kernel clustering. However, it cannot effectively solve multiple kernel clustering with incomplete kernels. Differently, the proposed LI-MKKM-MR makes the first attempt to identify the kernel redundancy problem in *incomplete multiple kernel clustering*, proposes an effective solution and conducts comprehensive experiments to validate our identification of this issue and the superiority of our algorithm.

II. Related Work

In this part, we mainly introduce the methods of multiple kernel k-means (MKKM) clustering, MKKM with incomplete kernels (MKKM-IK) and its localized variant. Before introducing these algorithms, we present all notations which will be used in the following in Table I.

Table I: Notations summary

$\{\mathbf{x}_i\}_{i=1}^n$	n training samples
k	number of clusters
τ	ratio of the nearest neighbors
$\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_m]^\top$	kernel weights
$\kappa_p(\cdot,\cdot)$	the p-th kernel function
$\phi_p(\cdot)$	feature mapping corresponding to $\kappa_p(\cdot,\cdot)$
$\phi_{\boldsymbol{\gamma}}(\cdot)$	feature mapping corresponding to $\kappa_{\boldsymbol{\gamma}}(\cdot,\cdot)$
$\{\mathbf{K}_p\}_{p=1}^m$	m base kernel matrices
$\mathbf{e}_p$	observed sample indices of $\mathbf{K}_p$
$\mathbf{H}$	partition matrix
$\mathbf{K}_p^{(dd)}$	sub-matrix of $\mathbf{K}_p$ for observed samples
$\mathbf{U}^{(i)} \in \{0,1\}^{n\times \mathrm{round}(n*\tau)}$	neighborhood indication matrix of $\mathbf{x}_i$
$\mathbf{M}$	correlation matrix among m base kernels
$\hat{\mathbf{C}} = [\hat{\mathbf{C}}_1, \cdots, \hat{\mathbf{C}}_k]$	the learned k centroids

A. Multiple kernel k-means (MKKM)

Let $\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathcal{X}$ be n training samples, and $\phi_p(\cdot) : \mathbf{x} \in \mathcal{X} \mapsto \mathcal{H}_p$, $\mathbf{x}$ are mapped onto a reproducing kernel Hilbert space $\mathcal{H}_p$ $(1 \le p \le m)$ by the p-th feature. Each sample in multiple kernel clustering is represented by $\phi_{\boldsymbol{\gamma}}(\mathbf{x}) = [\gamma_1\phi_1^\top(\mathbf{x}), \cdots, \gamma_m\phi_m^\top(\mathbf{x})]^\top$, where $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_m]^\top$ represents the weights of the m pre-specified base kernel functions $\{\kappa_p(\cdot,\cdot)\}_{p=1}^m$. These kernel weights will be adaptively adjusted

during multiple kernel clustering. Under the aforementioned definition of $\phi_{\boldsymbol{\gamma}}(\mathbf{x})$, the corresponding kernel function can be expressed as follows.

$$\kappa_{\boldsymbol{\gamma}}(\mathbf{x}_i, \mathbf{x}_j) = \phi_{\boldsymbol{\gamma}}^{\top}(\mathbf{x}_i)\phi_{\boldsymbol{\gamma}}(\mathbf{x}_j) = \sum\nolimits_{p=1}^{m} \gamma_p^2 \kappa_p(\mathbf{x}_i, \mathbf{x}_j). \quad (1)$$

One can calculate a kernel matrix $\mathbf{K}_{\boldsymbol{\gamma}}$ on training samples $\{\mathbf{x}_i\}_{i=1}^n$ with the kernel function defined in Eq. (1). As a result, the objective of MKKM with $\mathbf{K}_{\boldsymbol{\gamma}}$ is formulated as

$$\begin{aligned} \min_{\mathbf{H},\boldsymbol{\gamma}} \ & \mathrm{Tr}\left(\mathbf{K}_{\boldsymbol{\gamma}}(\mathbf{I}_n - \mathbf{H}\mathbf{H}^{\top})\right) \\ s.t. \ & \mathbf{H}^{\top}\mathbf{H} = \mathbf{I}_k, \ \boldsymbol{\gamma}^{\top}\mathbf{1}_m = 1, \ \gamma_p \geq 0, \ \forall p, \end{aligned} \quad (2)$$

where $\mathbf{H} \in \mathbb{R}^{n\times k}$ is a soft version of the cluster assignment matrix, and $\mathbf{I}_k$ is a $k \times k$ identity matrix. Alternately updating $\mathbf{H}$ and $\boldsymbol{\gamma}$ can optimize Eq. (2).

Optimizing $\mathbf{H}$ with fixed $\boldsymbol{\gamma}$. With $\boldsymbol{\gamma}$ fixed, the optimization in Eq. (2) toward $\mathbf{H}$ is exactly the traditional kernel k-means presented in Eq. (3)

$$\max_{\mathbf{H}} \ \mathrm{Tr}\left(\mathbf{H}^{\top}\mathbf{K}_{\boldsymbol{\gamma}}\mathbf{H}\right) \ \ s.t. \ \mathbf{H} \in \mathbb{R}^{n\times k}, \mathbf{H}^{\top}\mathbf{H} = \mathbf{I}_k, \quad (3)$$

The optimal $\mathbf{H}$ in Eq. (3) consists of the k eigenvectors corresponding to the top-k eigenvalues of $\mathbf{K}_{\boldsymbol{\gamma}}$ [34].

Optimizing $\boldsymbol{\gamma}$ with fixed $\mathbf{H}$. With $\mathbf{H}$ fixed, the equivalent form of the optimization in Eq. (2) with regard to $\boldsymbol{\gamma}$ is as follows

$$\min_{\boldsymbol{\gamma}} \sum\nolimits_{p=1}^{m} \gamma_p^2 \mathrm{Tr}\left(\mathbf{K}_p(\mathbf{I}_n - \mathbf{H}\mathbf{H}^{\top})\right) \ s.t. \ \boldsymbol{\gamma}^{\top}\mathbf{1}_m = 1, \ \gamma_p \geq 0, \quad (4)$$

which has a closed-form solution.

B. MKKM with Incomplete Kernels (MKKM-IK)

MKKM has recently been extended to handle incomplete multiple kernel clustering in [31, 33]. Previous algorithms first manage to impute the incomplete kernel matrices and then apply existing MKKM on the imputed kernel matrices. In contrast, they propose to unify the learning process of imputation and clustering into a common learning framework and establish an effective optimization algorithm to optimize each of them alternately. In MKKM-IK, the clustering procedure provides a guidance for the imputation of the incomplete base kernel matrices, and the clustering is further enhanced by the imputed kernels. Both procedures are alternated performed until achieving optimal results. The above idea can be achieved as follows

$$\begin{aligned} \min_{\mathbf{H},\,\boldsymbol{\gamma},\,\{\mathbf{K}_p\}_{p=1}^m} \ & \mathrm{Tr}\left(\mathbf{K}_{\boldsymbol{\gamma}}(\mathbf{I}_n - \mathbf{H}\mathbf{H}^{\top})\right) \\ s.t. \ & \mathbf{H} \in \mathbb{R}^{n\times k}, \mathbf{H}^{\top}\mathbf{H} = \mathbf{I}_k, \\ & \boldsymbol{\gamma}^{\top}\mathbf{1}_m = 1, \gamma_p \geq 0, \\ & \mathbf{K}_p(\mathbf{e}_p, \mathbf{e}_p) = \mathbf{K}_p^{(dd)}, \ \mathbf{K}_p \succeq 0, \ \forall p, \end{aligned} \quad (5)$$

where $\mathbf{e}_p\,(1 \leq p \leq m)$ denotes the sample indices, the p-th view is observed and $\mathbf{K}_p^{(dd)}$ denotes the kernel sub-matrix. Note that we impose the constraint $\mathbf{K}_p(\mathbf{e}_p, \mathbf{e}_p) = \mathbf{K}_p^{(dd)}$ to make the known entries of $\mathbf{K}_p$ kept unchanged during the learning course. The imputation of incomplete kernels can be regarded as a by-product of learning, because the ultimate goal of Eq. (5) is clustering. A tri-level optimization strategy is developed in [31] develops to solve Eq. (5) alternately.

Optimizing $\mathbf{H}$ with $\boldsymbol{\gamma}$ and $\{\mathbf{K}_p\}_{p=1}^m$ fixed. Given $\boldsymbol{\gamma}$ and $\{\mathbf{K}_p\}_{p=1}^m$, the optimization in Eq. (5) with respect to $\mathbf{H}$ is equivalent to a kernel k-means problem solved by Eq. (3);

Optimizing $\{\mathbf{K}_p\}_{p=1}^m$ with $\boldsymbol{\gamma}$ and $\mathbf{H}$ fixed. Given $\boldsymbol{\gamma}$ and $\mathbf{H}$, Eq. (5) towards each $\mathbf{K}_p$ is equivalently reformulated as follows,

$$\begin{aligned} \min_{\mathbf{K}_p} \ & \mathrm{Tr}\left(\mathbf{K}_p(\mathbf{I}_n - \mathbf{H}\mathbf{H}^{\top})\right) \\ s.t. \ & \mathbf{K}_p(\mathbf{e}_p, \mathbf{e}_p) = \mathbf{K}_p^{(dd)}, \ \mathbf{K}_p \succeq 0. \end{aligned} \quad (6)$$

It is proven in [31] that the optimal $\mathbf{K}_p$ in Eq. (6) has the closed-form solution as in Eq. (7), where $\mathbf{Z} = \mathbf{I}_n - \mathbf{H}\mathbf{H}^{\top}$ and taking the elements of $\mathbf{Z}$ corresponding to the observed and unobserved sample indices can construct $\mathbf{Z}^{(dm)}$. For more details, please refer to [31].

Optimizing $\boldsymbol{\gamma}$ with $\mathbf{H}$ and $\{\mathbf{K}_p\}_{p=1}^m$ fixed. Given $\mathbf{H}$ and $\{\mathbf{K}_p\}_{p=1}^m$, Eq. (5) with respect to $\boldsymbol{\gamma}$ reduces to a quadratic programming with linear constraints.

C. Localized Incomplete MKKM (LI-MKKM)

Although it is ingenious to unify clustering and imputation into one learning process, which is achieved by *globally* maximizing the alignment between the optimal kernel matrix $\mathbf{K}_{\boldsymbol{\gamma}}$ and the ideal matrix $\mathbf{H}\mathbf{H}^{\top}$, as presented in Eq. (2). This criterion does not take full advantage of the local distribution of data, and requires that all paired samples, whether closer or farther, should be consistent with the ideal similarity without distinction.

Instead of calculating the alignment between the optimal kernel and the idea matrix in a global manner as in Eq. (5), localized incomplete MKKM (LI-MKKM) [32] is proposed to utilize the local structure among data by only requiring the similarity of each sample to align with its nearest neighbours. Specifically, the objective function of LI-MKKM is as follows,

$$\begin{aligned} \min_{\boldsymbol{\gamma},\,\{\mathbf{K}_p\}_{p=1}^m,\mathbf{H}} \ & \sum\nolimits_{i=1}^{n} \mathrm{Tr}(\mathbf{K}_{\boldsymbol{\gamma}}(\mathbf{A}^{(i)} - \mathbf{A}^{(i)}\mathbf{H}\mathbf{H}^{\top}\mathbf{A}^{(i)})) \\ s.t. \ & \mathbf{H} \in \mathbb{R}^{n\times k}, \ \mathbf{H}^{\top}\mathbf{H} = \mathbf{I}_k, \ \boldsymbol{\gamma}^{\top}\mathbf{1}_m = 1, \ \gamma_p \geq 0, \\ & \mathbf{K}_p(\mathbf{e}_p, \mathbf{e}_p) = \mathbf{K}_p^{(dd)}, \ \mathbf{K}_p \succeq 0, \ \forall p, \end{aligned} \quad (8)$$

where $\mathbf{A}^{(i)} = \mathbf{U}^{(i)}\mathbf{U}^{(i)\top}$ with $\mathbf{U}^{(i)} \in \{0,1\}^{n\times \mathrm{round}(n*\tau)}\,(1 \leq i \leq n)$ denoting the neighborhood index matrix of the i-th sample. $\mathbf{U}_{jv}^{(i)} = 1$ reprensents that $\mathbf{x}_j$ is the v-th nearest neighbor of $\mathbf{x}_i$, where $1 \leq v \leq \mathrm{round}(n * \tau)$ and τ is the ratio of the nearest neighbors.

Similar to [31], the work in [32] develops a tri-step optimization algorithm to solve Eq. (8) and theoretically proves its convergence. Please refer to [32] for more details.

III. Localized Incomplete Multiple Kernel k-means with Matrix-induced Regularization

A. The Formulation

Although aligning the optimal kernel with the ideal similarity locally can improve the clustering performance, LI-MKKM dose not explicitly take the correlation among base kernels into account. This would prevent these incomplete base kernels from being well utilized. To overcome this problem,

$$\mathbf{K}_p = \begin{bmatrix} \mathbf{K}_p^{(dd)} & -\mathbf{K}_p^{(dd)}\mathbf{Z}^{(dm)}(\mathbf{Z}^{(mm)})^{-1} \\ -(\mathbf{Z}^{(mm)})^{-1}\mathbf{Z}^{(dm)\top}\mathbf{K}_p^{(dd)} & (\mathbf{Z}^{(mm)})^{-1}\mathbf{Z}^{(dm)\top}\mathbf{K}_p^{(dd)}\mathbf{Z}^{(dm)}(\mathbf{Z}^{(mm)})^{-1} \end{bmatrix} \tag{7}$$

we propose an improved algorithm based on LI-MKKM via introducing a matrix-induced regularization $\boldsymbol{\gamma}^\top\mathbf{M}\boldsymbol{\gamma}$ to decrease the redundancy and enhance the diversity of the selected base kernels, where M_{pq} measures the correlation between $\mathbf{K}_p$ and $\mathbf{K}_q$. By integrating this regularization into Eq. (8), the following objective is obtained:

$$\begin{aligned} \min_{\boldsymbol{\gamma},\{\mathbf{K}_p\}_{p=1}^m,\mathbf{H}} & \sum_{i=1}^{n} \mathrm{Tr}\left(\mathbf{K}_{\boldsymbol{\gamma}}(\mathbf{A}^{(i)} - \mathbf{A}^{(i)}\mathbf{H}\mathbf{H}^\top\mathbf{A}^{(i)})\right) + \frac{\lambda}{2}\boldsymbol{\gamma}^\top\mathbf{M}\boldsymbol{\gamma} \\ s.t.\ & \mathbf{H} \in \mathbb{R}^{n\times k},\ \mathbf{H}^\top\mathbf{H} = \mathbf{I}_k, \\ & \boldsymbol{\gamma}^\top\mathbf{1}_m = 1,\ \gamma_p \geq 0, \\ & \mathbf{K}_p(\mathbf{e}_p,\mathbf{e}_p) = \mathbf{K}_p^{(dd)},\ \mathbf{K}_p \succeq 0,\ \forall p, \end{aligned} \tag{9}$$

where λ is a hyper-parameter to balance the regularization on kernel weights and the loss of local kernel k-means.

In this work, we adopt $M_{pq} = \mathrm{Tr}(\mathbf{K}_p\mathbf{K}_q)$ to measure the correlation between $\mathbf{K}_p$ and $\mathbf{K}_q$. On one hand, the incorporation of $\boldsymbol{\gamma}^\top\mathbf{M}\boldsymbol{\gamma}$ is helpful for well utilizing the base kernels, which is utilized to boost the clustering performance. On the other hand, it makes the resultant optimization more challenging since the optimization on each $\mathbf{K}_p$ is a quadratic semi-defined programming, whose computational cost is intensive and this prevents it from being applied to practical applications. To reduce the computation overhead of Eq. (9), we propose to approximate M_{pq} by $\tilde{M}_{pq} = \mathrm{Tr}(\mathbf{K}_p^{(0)}\mathbf{K}_q^{(0)})$ and keep it unchanged during the learning course, where $\mathbf{K}_p^{(0)}$ is an initial imputation of $\mathbf{K}_p$. By substituting $\mathbf{M}$ with $\tilde{\mathbf{M}}$, the objective function of the proposed LI-MKKM-MR can be expressed as follows,

$$\begin{aligned} \min_{\boldsymbol{\gamma},\{\mathbf{K}_p\}_{p=1}^m,\mathbf{H}} & \sum_{i=1}^{n} \mathrm{Tr}(\mathbf{K}_{\boldsymbol{\gamma}}(\mathbf{A}^{(i)} - \mathbf{A}^{(i)}\mathbf{H}\mathbf{H}^\top\mathbf{A}^{(i)})) + \frac{\lambda}{2}\boldsymbol{\gamma}^\top\tilde{\mathbf{M}}\boldsymbol{\gamma} \\ s.t.\ & \mathbf{H} \in \mathbb{R}^{n\times k},\ \mathbf{H}^\top\mathbf{H} = \mathbf{I}_k, \\ & \boldsymbol{\gamma}^\top\mathbf{1}_m = 1,\ \gamma_p \geq 0, \\ & \mathbf{K}_p(\mathbf{e}_p,\mathbf{e}_p) = \mathbf{K}_p^{(dd)},\ \mathbf{K}_p \succeq 0,\ \forall p. \end{aligned} \tag{10}$$

It is reasonable to measure the correlation of pairwise kernels via observed similarity. Consequently, the approximation $\tilde{\mathbf{M}}$ can be regarded as a prior of $\mathbf{M}$. Also, although this approximation is simple, its advantages are three-folds. Firstly, it fulfills our requirement on the kernel coefficients to enhance the diversity and decrease the redundancy. Secondly, it simplifies the optimization on $\{\mathbf{K}_p\}_{p=1}^m$, making it admit a closed-form solution. This significantly increases the computational cost. Thirdly, the effectiveness of the proposed approximation can be demonstrated by experiments.

Although the matrix-induced regularization may be exploited in other related aspects such as multiple kernel clustering [14], this is the first work in literature to study the regularization on incomplete multiple kernel clustering and design a reasonable approximation for the convenience of computation. Moreover, this would trigger more research on incomplete multiple kernel clustering such as designing more informative $\mathbf{M}$, updating $\mathbf{M}$ with learned kernel weights and the imputation at each iteration, to name just a few. More importantly, our experimental study shows that the incorporation of matrix-induced regularization helps to utilize the incomplete kernels, leading to significantly improvement on clustering performance. We develop a tri-step optimization strategy to solve it alternately in the following parts.

B. Alternate Optimization of LI-MKKM-MR

Optimizing $\mathbf{H}$ with $\boldsymbol{\gamma}$ and $\{\mathbf{K}_p\}_{p=1}^m$ fixed. Given $\boldsymbol{\gamma}$ and $\{\mathbf{K}_p\}_{p=1}^m$, the optimization objective w.r.t $\mathbf{H}$ in Eq. (10) redefines to

$$\begin{aligned} \max_{\mathbf{H}}\ & \mathrm{Tr}\left(\mathbf{H}^\top\textstyle\sum_{i=1}^{n}(\mathbf{A}^{(i)}\mathbf{K}_{\boldsymbol{\gamma}}\mathbf{A}^{(i)})\mathbf{H}\right) \\ s.t.\ & \mathbf{H} \in \mathbb{R}^{n\times k},\ \mathbf{H}^\top\mathbf{H} = \mathbf{I}_k, \end{aligned} \tag{11}$$

which is transformed into a classical kernel k-means-based optimization objective, and can be conveniently tackled by the existing public toolkit.

Optimizing $\{\mathbf{K}_p\}_{p=1}^m$ with $\boldsymbol{\gamma}$ and $\mathbf{H}$ fixed. Given $\boldsymbol{\gamma}$ and $\mathbf{H}$, the optimization objective w.r.t $\{\mathbf{K}_p\}_{p=1}^m$ in Eq. (10) can be formulated as

$$\begin{aligned} \min_{\{\mathbf{K}_p\}_{p=1}^m}\ & \sum_{p=1}^{m}\gamma_p^2\mathrm{Tr}\left(\mathbf{K}_p\sum_{i=1}^{n}\mathrm{Tr}(\mathbf{A}^{(i)} - \mathbf{A}^{(i)}\mathbf{H}\mathbf{H}^\top\mathbf{A}^{(i)})\right) \\ s.t.\ & \mathbf{K}_p(\mathbf{e}_p,\mathbf{e}_p) = \mathbf{K}_p^{(dd)},\ \mathbf{K}_p \succeq 0,\ \forall p. \end{aligned} \tag{12}$$

It is difficult to solve the optimization problem in Eq. (12) since there are multiple kernel matrices to be optimized simultaneously. By cautiously analyze the optimization, we observe that: i) each kernel matrix $\mathbf{K}_p$ has its own separate constraint; and ii) the objective in Eq. (12) is a sum generated by calculating $\mathbf{K}_p$. As a result, Eq. (12) can be reformulated as m uncorrelated sub-objectives equivalently, as shown in Eq. (13),

$$\begin{aligned} \min_{\mathbf{K}_p}\ & \mathrm{Tr}(\mathbf{K}_p\mathbf{Q}) \\ s.t.\ & \mathbf{K}_p(\mathbf{e}_p,\mathbf{e}_p) = \mathbf{K}_p^{(dd)},\ \mathbf{K}_p \succeq 0, \end{aligned} \tag{13}$$

where $\mathbf{Q} = \sum_{i=1}^{n}(\mathbf{A}^{(i)} - \mathbf{A}^{(i)}\mathbf{H}\mathbf{H}^\top\mathbf{A}^{(i)})$.

It seems that directly solving the Eq. (13) is difficult because of the equality and PSD constraints imposed on $\mathbf{K}_p$. By following [32], we parameterize each $\mathbf{K}_p$ as

$$\mathbf{K}_p = \begin{bmatrix} \mathbf{K}_p^{(dd)} & \mathbf{K}_p^{(dd)}\mathbf{Z}_p \\ \mathbf{Z}_p^\top\mathbf{K}_p^{(dd)} & \mathbf{Z}_p^\top\mathbf{K}_p^{(dd)}\mathbf{Z}_p \end{bmatrix}, \tag{14}$$

where $\mathbf{Z}_p \in \mathbb{R}^{d\times m}$. d and m refer to the number of observed samples and unobserved ones, respectively. With Eq. (14), we assume that the observed ones represent the missing kernel entries. It is shown in [32] that $\mathbf{K}_p$ in Eq. (14) automatically satisfies both constraints after this parameterization.

Based on the parameterization in Eq. (14) , the constrained optimization in Eq. (13) is equivalent to

$$\min_{\mathbf{Z}_p} \operatorname{Tr}\left(\begin{bmatrix} \mathbf{K}_p^{(dd)} & \mathbf{K}_p^{(dd)}\mathbf{Z}_p \\ \mathbf{Z}_p^\top \mathbf{K}_p^{(dd)} & \mathbf{Z}_p^\top \mathbf{K}_p^{(dd)}\mathbf{Z}_p \end{bmatrix} \begin{bmatrix} \mathbf{Q}^{(dd)} & \mathbf{Q}^{(dm)} \\ \mathbf{Q}^{(dm)\top} & \mathbf{Q}^{(mm)} \end{bmatrix}\right), \tag{15}$$

where $\mathbf{Q}$ is decomposed into the following sub-matrices $\begin{bmatrix} \mathbf{Q}^{(dd)} & \mathbf{Q}^{(dm)} \\ \mathbf{Q}^{(dm)\top} & \mathbf{Q}^{(mm)} \end{bmatrix}$.

To minimize Eq. (15), we take its derivative with respect to $\mathbf{Z}_p$ and let it vanish, leading to

$$\mathbf{Z}_p = -\mathbf{Q}^{(dm)}(\mathbf{Q}^{(mm)})^{-1}. \tag{16}$$

As a result, we obtain an analytical solution for the optimal $\mathbf{K}_p$ by substituting the $\mathbf{Z}_p$ in Eq. (16) into Eq. (14). As seen, Eq. (13) provides a guidance for the imputation of each base kernel by exploring the data structure in a local manner. Specifically, it locally estimates the alignment between the similarity of each sample and its τ-nearest neighbors with corresponding ideal matrix. This enables the proposed algorithm to better utilize the intra-cluster variations among samples. Therefore, the clustering performance could be improved, mainly attributing to an effective incomplete kernels imputation measure.

Optimizing $\boldsymbol{\gamma}$ with $\{\mathbf{K}_p\}_{p=1}^m$ and $\mathbf{H}$ fixed. Given $\{\mathbf{K}_p\}_{p=1}^m$ and $\mathbf{H}$, it is easy to present that Eq. (10) w.r.t. $\boldsymbol{\gamma}$ is as below,

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \ & \frac{1}{2}\boldsymbol{\gamma}^\top \left(2\mathbf{W} + \lambda\tilde{\mathbf{M}}\right)\boldsymbol{\gamma} \\ s.t. \ & \boldsymbol{\gamma}^\top \mathbf{1}_m = 1,\ \gamma_p \geq 0, \end{aligned} \tag{17}$$

where $\mathbf{W} = \operatorname{diag}([\operatorname{Tr}(\mathbf{K}_1\mathbf{Q}), \cdots, \operatorname{Tr}(\mathbf{K}_m\mathbf{Q})])$. The Theorem 1 in the following indicates that $\mathbf{W}$ is PSD.

Theorem 1: The Hessian matrix $2\mathbf{W} + \lambda\tilde{\mathbf{M}}$ in Eq. (17) is a symmetric PSD matrix.

Proof 1: By defining $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_k]$, we can find out that $\mathbf{H}\mathbf{H}^\top\mathbf{h}_c = \mathbf{h}_c (1 \leq c \leq k)$ since $\mathbf{H}^\top\mathbf{H} = \mathbf{I}_k$. This indicates that $\mathbf{H}\mathbf{H}^\top$ has k eigenvalue with 1. Besides, its rank does not exceed k. This means that its has $n-k$ eigenvalue with 0. $\mathbf{I}_n - \mathbf{H}\mathbf{H}^\top$ contains $n-k$ eigenvalue with 1 and k eigenvalue with 0. Consequently, $\mathbf{A}^{(i)}(\mathbf{I}_n - \mathbf{H}\mathbf{H}^\top)\mathbf{A}^{(i)}$ is PSD, which ensures that $\mathbf{Q} = \sum_{i=1}^n (\mathbf{A}^{(i)} - \mathbf{A}^{(i)}\mathbf{H}\mathbf{H}^\top\mathbf{A}^{(i)})$ is PSD. As a result, we have $w_p = \operatorname{Tr}(\mathbf{K}_p\mathbf{Q}) \geq 0, \forall p$, guaranteeing the positiveness of $\mathbf{W}$. Meanwhile, $\mathbf{W}$ is also a symmetric PSD matrix according to [35]. Consequently, $2\mathbf{W} + \lambda\tilde{\mathbf{M}}$ is a symmetric PSD matrix.

On the basis of Theorem 1, we can guarantee that the optimization in Eq. (17) w.r.t $\boldsymbol{\gamma}$ is a traditional quadratic programming (QP) with linear constraints. Therefore, it can be conveniently handled by existing optimization packages.

Algorithm 1 presents an outline of solving Eq. (10) by the proposed algorithm, where we adopt the zero-filling method to initially impute the missing elements of $\{\mathbf{K}_p^{(0)}\}_{p=1}^m$ and utilize $\text{obj}^{(t)}$ to represent the objective value at the t-th iteration. Besides, the neighbors of each sample remain unvaried during the optimization procedure in LI-MKKM-MR. In specific, we calculate the τ-nearest neighbors of each sample by $\mathbf{K}_{\boldsymbol{\gamma}^{(0)}}$. By this way, the optimization target of LI-MKKM-MR is guaranteed to be reduced in a monotonic manner when we update one variable and keep the others unchanged iteratively. Simultaneously, the objective is lower-bounded by zero.

Algorithm 1 The Proposed LI-MKKM-MR

Input: $\{\mathbf{K}_p^{dd)}\}_{p=1}^m$, $\{\mathbf{e}_p\}_{p=1}^m, k, \tau, \lambda$ and ϵ_0.
Output: $\mathbf{H}$, $\boldsymbol{\gamma}$ and $\{\mathbf{K}_p\}_{p=1}^m$.
Initialize $\boldsymbol{\gamma}^{(0)} = \mathbf{1}_m/m$, $\{\mathbf{K}_p^{(0)}\}_{p=1}^m$ and $t = 1$.
Generate $\mathbf{U}^{(i)}$ for i-th samples $(1 \leq i \leq n)$ by $\mathbf{K}_{\boldsymbol{\gamma}^{(0)}}$.
Calculate $\mathbf{A}^{(i)} = \mathbf{U}^{(i)}\mathbf{U}^{(i)^\top}$ for i-th samples $(1 \leq i \leq n)$.
repeat
  $\mathbf{K}_{\boldsymbol{\gamma}^{(t)}} = \sum_{p=1}^m \left(\gamma_p^{(t-1)}\right)^2 \mathbf{K}_p^{(t-1)}$.
  Update $\mathbf{H}^{(t)}$ by solving Eq. (11) with $\mathbf{K}_{\boldsymbol{\gamma}^{(t)}}$.
  Update $\{\mathbf{K}_p^{(t)}\}_{p=1}^m$ with $\mathbf{H}^{(t)}$ by Eq. (13).
  Update $\boldsymbol{\gamma}^{(t)}$ by solving Eq. (17) with $\mathbf{H}^{(t)}$ and $\{\mathbf{K}_p^{(t)}\}_{p=1}^m$.
  $t = t + 1$.
until $\left(\text{obj}^{(t-1)} - \text{obj}^{(t)}\right)/\text{obj}^{(t)} \leq \epsilon_0$

Hence, it is guaranteed that LI-MKKM-MR converges into a local optimal solution. Experimental results have demonstrated that our method usually converges quickly.

The end of this part analyzes the computational complexity of our method. In specific, the computational complexity of LI-MKKM-MR is $\mathcal{O}(n^3 + \sum_{p=1}^m n_p^3 + m^3)$ at each iteration, where n_p $(n_p \leq n)$ and m refer to the number of observed samples of $\mathbf{K}_p$ and base kernels. The complexity of LI-MKKM-MR can be compared to that of MKKM-IK [31] and LI-MKKM [32]. Moreover, each sample of $\mathbf{K}_p$ is independent so that they can be measured in a parallel manner. By this means, our LI-MKKM-MR can scale well regardless of the variation of the base kernels number.

IV. Theoretical Results

The generalization error of k-means clustering algorithm has been widely discussed in existing literature [36–38]. We first establish the theoretical connection between existing MKKM-IK [38] with LI-MKKM-MR, and further derive the generalization error bound of LI-MKKM-MR based on the theoretical results in [38]. The following Theorem 2 points out that the local kernel alignment adopted in our LI-MKKM-MR can be achieved by normalizing each base kernel matrix.

Theorem 2: The local kernel alignment criterion in Eq. (8) is equivalent to the widely adopted global kernel alignment by normalizing each base kernel matrix.

Proof 2: The objective function in Eq. (8) can be written as

$$\begin{aligned} & \sum_{i=1}^n \operatorname{Tr}\left(\mathbf{K}_{\boldsymbol{\gamma}}(\mathbf{A}^{(i)} - \mathbf{A}^{(i)}\mathbf{H}\mathbf{H}^\top\mathbf{A}^{(i)})\right) \\ &= \sum_{i=1}^n \langle \mathbf{A}^{(i)} \otimes \mathbf{K}_{\boldsymbol{\gamma}}, \mathbf{A}^{(i)} \otimes (\mathbf{I} - \mathbf{H}\mathbf{H}^\top)\rangle_{\mathrm{F}} \\ &= \sum_{i=1}^n \langle \mathbf{A}^{(i)} \otimes \mathbf{K}_{\boldsymbol{\gamma}}, \mathbf{I} - \mathbf{H}\mathbf{H}^\top\rangle_{\mathrm{F}} \\ &= \langle \left(\sum_{i=1}^n \mathbf{A}^{(i)}\right) \otimes \mathbf{K}_{\boldsymbol{\gamma}}, \mathbf{I} - \mathbf{H}\mathbf{H}^\top\rangle_{\mathrm{F}} \\ &= \sum_{p=1}^m \gamma_p^2 \langle \left(\sum_{i=1}^n \mathbf{A}^{(i)}\right) \otimes \mathbf{K}_p, \mathbf{I} - \mathbf{H}\mathbf{H}^\top\rangle_{\mathrm{F}} \\ &= \sum_{p=1}^m \gamma_p^2 \langle \tilde{\mathbf{K}}_p, \mathbf{I} - \mathbf{H}\mathbf{H}^\top\rangle_{\mathrm{F}} \\ &= \operatorname{Tr}\left(\tilde{\mathbf{K}}_{\boldsymbol{\gamma}}(\mathbf{I} - \mathbf{H}\mathbf{H}^\top)\right), \end{aligned} \tag{18}$$

where $\otimes$ denotes elementwise multiplication between two matrices, $\tilde{\mathbf{K}}_p = \left(\sum_{i=1}^n \mathbf{A}^{(i)}\right) \otimes \mathbf{K}_p$ can be treated as a normalized $\mathbf{K}_p$, and $\tilde{\mathbf{K}}_{\boldsymbol{\gamma}} = \sum_{p=1}^m \gamma_p^2 \tilde{\mathbf{K}}_p$. Consequently, by such normalization applied on each base kernel, we can clearly see that the local kernel alignment criterion in Eq. (8) is exactly the global kernel alignment in [38]. This completes the proof.

Let $t(\mathbf{x}^{(p)}) = 1$ if the p-th view of $\mathbf{x}$ is available, otherwise $\mathbf{x}^{(p)}$ should be optimized. It is worth pointing out that $t(\mathbf{x}^{(p)})$ is a random variable which depends on $\mathbf{x}$. Let $\hat{\mathbf{C}} = [\hat{\mathbf{C}}_1, \cdots, \hat{\mathbf{C}}_k]$ be the k centroids and $\hat{\boldsymbol{\gamma}}$ be the kernel weights learned by LI-MKKM-MR. k-means clustering should make the reconstruction error small

$$\mathbb{E}\left[\min_{\mathbf{y}\in\{\mathbf{e}_1,\cdots,\mathbf{e}_k\}} \left\|\phi_{\hat{\boldsymbol{\gamma}}}(\mathbf{x}) - \hat{\mathbf{C}}\mathbf{y}\right\|_{\mathcal{H}}^2\right], \tag{19}$$

where $\phi_{\hat{\boldsymbol{\gamma}}}(\mathbf{x}) = [\hat{\gamma}_1 t(\mathbf{x}^{(1)})\phi_1^\top(\mathbf{x}^{(1)}), \cdots, \hat{\gamma}_m t(\mathbf{x}^{(m)})\phi_m^\top(\mathbf{x}^{(m)})]^\top$, $\mathbf{e}_1, \cdots, \mathbf{e}_k$ form the orthogonal bases of $\mathbb{R}^k$.

We first define a function class:

$$\begin{aligned}\mathcal{F} = \Big\{ f : \mathbf{x} \mapsto \min_{\mathbf{y}\in\{\mathbf{e}_1,\ldots,\mathbf{e}_k\}} \|\phi_{\boldsymbol{\gamma}}(\mathbf{x}) - \mathbf{C}\mathbf{y}\|_{\mathcal{H}}^2 \,\Big|\, \boldsymbol{\gamma}^\top \mathbf{1}_m = 1,\ \gamma_p \geq 0,\\ \mathbf{C} \in \mathcal{H}^k,\ t(\mathbf{x}_i^{(p)}) t(\mathbf{x}_j^{(p)}) \tilde{\kappa}_p^\top(\mathbf{x}_i^{(p)}, \mathbf{x}_j^{(p)}) \leq b,\ \forall p, \forall \mathbf{x}_i \in \mathcal{X} \Big\},\end{aligned} \tag{20}$$

where $\mathcal{H}^k$ represents the multiple kernel Hilbert space and $\tilde{\kappa}(\cdot,\cdot)$ is a kernel function corresponding to $\tilde{\mathbf{K}}_p$.

Based on Theorem 2, we derive the generalization error bound of the proposed LI-MKKM-MR by following [38].

Theorem 3: For any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $f \in \mathcal{F}$:

$$\begin{aligned}\mathbb{E}[f(\mathbf{x})] \leq \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i) + \frac{4\sqrt{\pi m} b \mathcal{G}_{1n}(\boldsymbol{\gamma}, t)}{n} + \frac{4\sqrt{\pi m} b \mathcal{G}_{2n}(\boldsymbol{\gamma}, t)}{n}\\ + \frac{\sqrt{8\pi} b k^2}{\sqrt{n}} + 2b\sqrt{\frac{\log 1/\delta}{2n}},\end{aligned} \tag{21}$$

where

$$\mathcal{G}_{1n}(\boldsymbol{\gamma}, t) \triangleq \mathbb{E}_\gamma\left[\sup_{\boldsymbol{\gamma}, t} \sum_{i=1}^n \sum_{p,q=1}^m \gamma_{ipq} t(\mathbf{x}_i^{(p)}) t(\mathbf{x}_i^{(q)}) \gamma_p \gamma_q\right], \tag{22}$$

$$\mathcal{G}_{2n}(\boldsymbol{\gamma}, t) = \mathbb{E}_\gamma\left[\sup_{\boldsymbol{\gamma}, t} \sum_{i=1}^n \sum_{c=1}^k \sum_{p=1}^m \gamma_{icp} \gamma_p t(\mathbf{x}_i^{(p)})\right], \tag{23}$$

and $\gamma_{ipq}, \gamma_{icp}, i \in \{1,\ldots,n\}, p, q \in \{1,\ldots,m\}, c \in \{1,\ldots,k\}$ are i.i.d. Gaussian random variables with zero mean and unit standard deviation.

According to analyses in [38], our local kernel alignment criterion in Eq. (8), with normalized base kernel matrices, is an upper bound of $\frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)$. As a result, by minimizing $\mathrm{Tr}(\tilde{\mathbf{K}}_{\boldsymbol{\gamma}}(\mathbf{I}_n - \mathbf{H}\mathbf{H}^\top))$, one can get a small $\frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)$ for good generalization. This justifies the good generalization ability of the LI-MKKM-MR. The detailed proof has been presented in the supplementary material.

V. Experiments

A. Experimental Settings

In our experiments, we adopt ten widely used MKL benchmark data sets to verify the proposed algorithms, including Oxford Flower17 and Flower102[1], Caltech102[2], Digital[3] and Protein Fold Prediction[4], The information of them is shown in Table II. The kernel matrices of these datasets are precomputed, and can be directly obtained from the aforementioned link. Caltech102-5 refers to the number of samples belonging to each cluster is 5, and the same for the rest datasets. The publicly access codes for kernel k-means and MKKM can be found in the website [5].

Several well-known and widely used imputation methods, such as zero filling (ZF), mean filling (MF), k-nearest-neighbor filling (KNN), alignment-maximization filling (AF) are contained in [27]. After that, researchers take the imputed kernel matrices as the input of classical MKKM. The kind of two-stage methods are named MKKM+ZF, MKKM+MF, MKKM+KNN and MKKM+AF, respectively. Also, the newly proposed MKKM-IK [31], LI-MKKM [32], MVEC [39] and CG-IMVC [40] are also incorporated as strong baselines. The algorithms in [28, 29, 41] are not incorporated into our experimental comparison since that these algorithms only consider the missing of input features, rather than the rows or columns of base kernel matrices in our case.

In the experiment, ε is used to denote the percentage of incomplete samples. Intuitively, the clustering performance will become less accurate when the value of ε is increasing. In our simulation, we set ε as $[0.1 : 0.1 : 0.9]$ on all the 10 data sets. The performance metrics in this simulation include the clustering accuracy (ACC), normalized mutual information (NMI) and purity. For each method, we present the best result among 50 trials, where each trial started from a random initialization state. As a result, the effect of randomness caused by k-means could be alleviated. We generate "incomplete" patterns randomly for 10 times and report the statistical results. For all datasets, the quantity of clusters is given and set as the ground truth of classes. The generation of the missing vectors $\{\mathbf{s}_p\}_{p=1}^m$ follows the approach in [31]: (1) Randomly select $\mathrm{round}(\varepsilon * n)$ samples with the rounding function $\mathrm{round}(\cdot)$. (2) Generate a random vector $\mathbf{v} = (v_1, \cdots, v_k, \cdots, v_m)$, $v_k \in [0,1]$ and a scalar v_0, $v_0 \in [0,1]$ for each selected sample. (3) If $v_p \geq v_0$, it present the p-th view for this sample. (4) If there is no $v_p \geq v_0$, generate a new $\mathbf{v}$. Note that there is no requirement on complete view for each sample. In this instance, the index vector $\mathbf{s}_p$ is obtained to list the samples with the presentation on the p-th view.

B. Experimental Results

Experiments on Flower17 and Flower102. Three performance metrics, including the ACC, NMI and purity, of the testing algorithms with the variation of missing ratios in $[0.1, \cdots, 0.9]$ on the Flower17 and Flower102 datasets have been demonstrated in Figure 1. We have the following observations.

- The newly proposed MKKM-IK [33] (in green) has shown promising performance improvements

[1] http://www.robots.ox.ac.uk/~+vgg/data/flowers/
[2] http://files.is.tue.mpg.de/pgehler/projects/iccv09/
[3] http://ss.sysu.edu.cn/~+py/
[4] http://mkl.ucsd.edu/dataset/protein-fold-prediction/
[5] https://github.com/mehmetgonen/lmkkmeans/

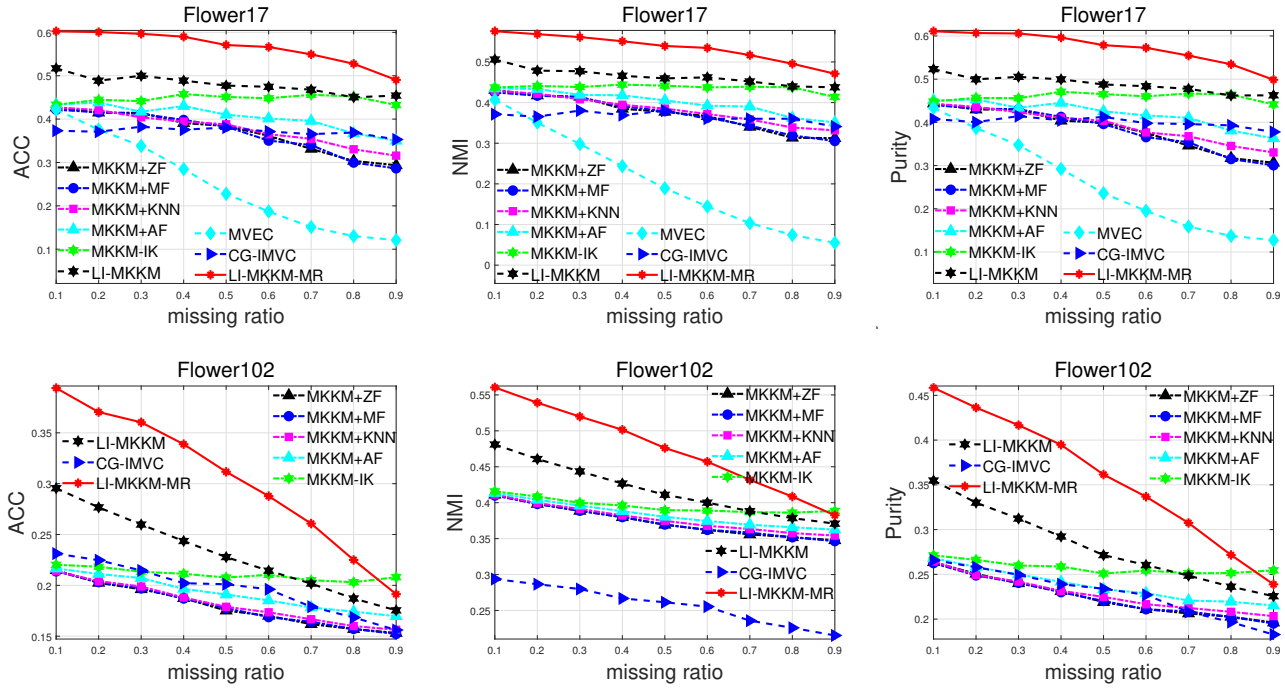

Figure 1: Clustering accuracy, NMI and purity comparison with the variation of missing ratios on Flower17 and Flower102 datasets.

Table III: Aggregated ACC, NMI and purity comparison (mean±std) of different kinds of clustering algorithms on Flower17 and Flower102 datasets.

Datasets	MKKM				MKKM-IK	LI-MKKM	MVEC	CG-IMVC	LI-MKKM-MR
	+ZF	+MF	+KNN	+AF [27]	[33]	[32]	[39]	[40]	Proposed
ACC									
Flower17	36.9 ± 0.8	36.8 ± 0.6	37.8 ± 0.6	40.5 ± 0.7	44.6 ± 0.6	48.0 ± 0.4	24.9 ± 0.4	37.1 ± 0.7	$\mathbf{56.6 \pm 0.3}$
Flower102	18.0 ± 0.2	18.0 ± 0.2	18.2 ± 0.1	19.2 ± 0.1	21.1 ± 0.2	23.1 ± 0.1	–	19.7 ± 0.3	$\mathbf{30.5 \pm 0.3}$
NMI									
Flower17	37.3 ± 0.4	37.3 ± 0.5	38.2 ± 0.5	40.1 ± 0.4	43.7 ± 0.3	46.4 ± 0.2	20.7 ± 0.4	36.5 ± 0.7	$\mathbf{53.5 \pm 0.2}$
Flower102	37.4 ± 0.1	37.4 ± 0.1	37.8 ± 0.1	38.4 ± 0.1	39.6 ± 0.1	41.8 ± 0.1	–	25.8 ± 0.3	$\mathbf{47.5 \pm 0.1}$
Purity									
Flower17	38.4 ± 0.6	38.3 ± 0.6	39.3 ± 0.6	42.0 ± 0.6	45.9 ± 0.5	48.9 ± 0.4	25.7 ± 0.4	40.1 ± 0.7	$\mathbf{57.3 \pm 0.2}$
Flower102	22.5 ± 0.1	22.4 ± 0.1	22.8 ± 0.1	23.7 ± 0.2	25.8 ± 0.2	28.1 ± 0.1	–	22.9 ± 0.3	$\mathbf{35.8 \pm 0.3}$

Table II: Datasets summary.

Dataset	#Samples	#Views	#Classes
Flower17	1360	7	17
Flower102	8189	4	102
Caltech102-5	510	48	102
Caltech102-10	1020	48	102
Caltech102-15	1530	48	102
Caltech102-20	2040	48	102
Caltech102-25	2550	48	102
Caltech102-30	3060	48	102
Digital	2000	3	10
ProteinFold	694	12	27

on the ACC, NMI and purity compared with the previous two-stage imputation methods. For example, the MKKM+AF outperforms MKKM-IK by $0.1\%, 0.6\%, 2.5\%, 2.8\%, 4.1\%, 4.7\%, 6.0\%, 8.5\%, 8.2\%$ in terms of clustering accuracy on Flower17, which clearly demonstrates the benefit of the joint optimization on imputation and clustering.

- Also, LI-MKKM outperforms MKKM-IK by $8.4\%, 4.4\%, 5.8\%, 3.1\%, 2.6\%, 2.6\%, 1.2\%, 0.2\%, 2.2\%$ on Flower17. This result clearly verifies that the utilizing data's local structure further boosts the clustering performance.
- Furthermore, our proposed LI-MKKM-MR (in red) significantly outperforms the LI-MKKM in all cases from Fig.1a to 1f in the aspect of clustering performance. For example, LI-MKKM-MR further outperforms LI-MKKM by $8.5\%, 11.2\%, 9.7\%, 10.1\%, 9.4\%, 9.2\%, 8.2\%, 7.7\%, 3.6\%$. This result indicates the effectiveness of incorporating the matrix-induced regularization.
- In addition, our newly proposed method demonstrates stronger advantage when compared to previous ones, especially under low missing ratios. It is noticeably that in Figure 1a, when the missing ratio is extremely

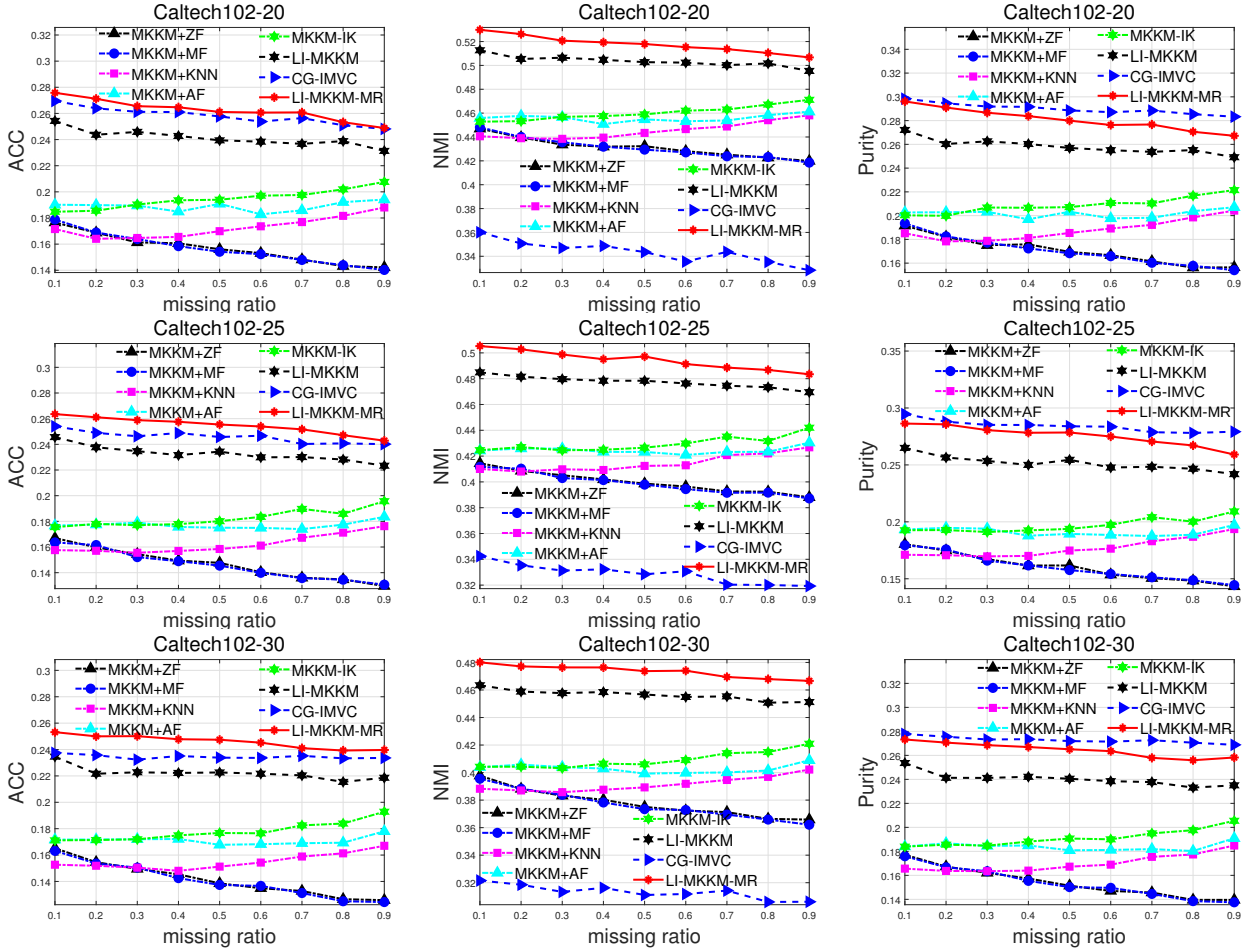

Figure 2: Clustering accuracy, NMI and purity comparison with the variation of missing ratios on Caltech102-20, Caltech102-25 and Caltech102-30.

Table IV: Total ACC, NMI and purity comparison (mean±std) of various clustering algorithms on Caltech102. On account of out of memory, the clustering results of MVEC [39] on Caltech102-15, Caltech102-20, Caltech102-25 and Caltech102-30 are not reported.

	MKKM				MKKM-IK	LI-MKKM	MVEC	CG-IMVC	LI-MKKM-MR
	+ZF	+MF	+KNN	+AF [27]	[33]	[32]	[39]	[40]	Proposed
ACC									
Cal102-5	26.1 ± 0.3	25.7 ± 0.3	27.3 ± 0.3	29.0 ± 0.3	28.9 ± 0.3	31.4 ± 0.3	26.8 ± 0.2	33.8 ± 0.2	$\mathbf{34.0 \pm 0.3}$
Cal102-10	19.7 ± 0.2	19.7 ± 0.2	21.5 ± 0.2	22.6 ± 0.2	22.7 ± 0.2	27.3 ± 0.2	22.4 ± 0.1	$\mathbf{28.9 \pm 0.2}$	$\mathbf{28.9 \pm 0.3}$
Cal102-15	17.1 ± 0.2	17.1 ± 0.2	18.9 ± 0.1	20.3 ± 0.2	20.8 ± 0.2	25.1 ± 0.2	–	$\mathbf{27.3 \pm 0.1}$	27.0 ± 0.4
Cal102-20	15.7 ± 0.1	15.7 ± 0.2	17.3 ± 0.2	18.9 ± 0.2	19.5 ± 0.1	24.1 ± 0.2	–	25.8 ± 0.2	$\mathbf{26.3 \pm 0.2}$
Cal102-25	14.7 ± 0.2	14.6 ± 0.1	16.2 ± 0.1	17.7 ± 0.2	18.3 ± 0.2	23.3 ± 0.2	–	24.6 ± 0.2	$\mathbf{25.5 \pm 0.2}$
Cal102-30	14.2 ± 0.1	14.1 ± 0.1	15.5 ± 0.2	17.1 ± 0.2	17.8 ± 0.2	22.2 ± 0.1	–	23.5 ± 0.1	$\mathbf{24.6 \pm 0.1}$
NMI									
Cal102-5	64.3 ± 0.2	63.9 ± 0.1	65.9 ± 0.2	66.6 ± 0.1	66.5 ± 0.2	67.1 ± 0.2	65.6 ± 0.1	52.9 ± 0.4	$\mathbf{68.6 \pm 0.2}$
Cal102-10	53.6 ± 0.1	53.7 ± 0.1	55.2 ± 0.1	55.7 ± 0.2	55.8 ± 0.1	58.7 ± 0.1	55.1 ± 0.1	40.4 ± 0.5	$\mathbf{59.2 \pm 0.3}$
Cal102-15	47.4 ± 0.1	47.4 ± 0.1	48.8 ± 0.1	49.7 ± 0.1	50.1 ± 0.1	53.6 ± 0.1	–	37.0 ± 0.3	$\mathbf{54.6 \pm 0.2}$
Cal102-20	43.1 ± 0.1	43.1 ± 0.2	44.5 ± 0.1	45.6 ± 0.2	46.0 ± 0.1	50.4 ± 0.1	–	34.4 ± 0.3	$\mathbf{51.8 \pm 0.1}$
Cal102-25	40.0 ± 0.1	39.9 ± 0.1	41.5 ± 0.1	42.5 ± 0.2	43.0 ± 0.2	47.7 ± 0.2	–	32.9 ± 0.3	$\mathbf{49.4 \pm 0.1}$
Cal102-30	37.8 ± 0.1	37.7 ± 0.1	39.2 ± 0.1	40.3 ± 0.1	40.9 ± 0.1	45.6 ± 0.1	–	31.3 ± 0.2	$\mathbf{47.4 \pm 0.1}$
Purity									
Cal102-5	26.7 ± 0.4	26.4 ± 0.3	27.9 ± 0.3	29.8 ± 0.3	29.6 ± 0.3	32.6 ± 0.3	27.3 ± 0.2	$\mathbf{35.9 \pm 0.2}$	35.5 ± 0.3
Cal102-10	21.0 ± 0.2	21.0 ± 0.2	22.9 ± 0.2	24.0 ± 0.3	24.2 ± 0.2	29.0 ± 0.2	23.3 ± 0.1	$\mathbf{31.7 \pm 0.2}$	30.8 ± 0.3
Cal102-15	18.5 ± 0.2	18.5 ± 0.2	20.4 ± 0.2	21.6 ± 0.2	22.2 ± 0.2	26.7 ± 0.2	–	$\mathbf{30.2 \pm 0.1}$	28.8 ± 0.3
Cal102-20	17.1 ± 0.1	17.0 ± 0.2	18.8 ± 0.2	20.2 ± 0.2	20.9 ± 0.1	25.8 ± 0.2	–	$\mathbf{29.0 \pm 0.2}$	28.1 ± 0.2
Cal102-25	16.0 ± 0.2	16.0 ± 0.2	17.7 ± 0.2	19.1 ± 0.2	19.7 ± 0.1	25.2 ± 0.2	–	$\mathbf{28.4 \pm 0.1}$	27.6 ± 0.2
Cal102-30	15.4 ± 0.1	15.4 ± 0.1	17.0 ± 0.1	18.4 ± 0.2	19.1 ± 0.2	24.0 ± 0.1	–	$\mathbf{27.3 \pm 0.1}$	26.5 ± 0.1

low (ε=0.1), LI-MKKM-MR improves the second best algorithm (LI-MKKM) by 8.5% in terms of clustering accuracy on Flower17.

In Table III, the aggregated ACC, NMI, purity, and the standard deviation are reported, where we show the highest performance one in bold. Similarly, the results also illustrates that MKKM+ZF, MKKM+MF, MKKM+KNN, MKKM+AF and MKKM-IK are outperformed by the proposed algorithm. Specifically, the second best one (LI-MKKM) is exceeded by the proposed LI-MKKM-MR by 7%.

Experiments on Caltech102 Dataset. Figure 2 presents ACC, NMI and purity of all the testing algorithms over variational missing ratios on Caltech102 datasets. We find out that the recently proposed MKKM-IK [33] (in green) achieves a comparable clustering performance with a representative two-stage imputation method MKKM+AF, while the proposed LI-MKKM outperforms MKKM-IK with significant improvements on all the performance criterions, details can be found in Fig.2a to 2i. More precisely, LI-MKKM obtains $6.4\%, 5.0\%, 5.1\%, 4.7\%, 4.6\%, 4.5\%, 3.8\%, 3.2\%, 2.6\%$ higher clustering accuracy than MKKM-IK when the missing ratios vary from 0.1 to 0.9 on Caltech102-30. This also illustrates that the well utilization of the local structure of data assures performance improvement. Furthermore, by taking into account the correlation among base kernels, LI-MKKM-MR further improves the clustering performance over the baseline LI-MKKM.

The aggregated ACC, NMI and purity, and the standard deviation on Caltech 102 datasets are reported in Table IV. Similarly, in comparison to the MKKM+ZF, MKKM+MF, MKKM+KNN, MKKM+AF and MKKM-IK, our method still achieves much better clustering performance. For instance, the proposed LI-MKKM-MR obtains $2.1\%, 2.1\%, 2.8\%, 2.4\%, 2.7\%, 2.4\%$ higher clustering accuracy than LI-MKKM. In addition, LI-MKKM-MR achieves comparable clustering performance with the newly proposed CG-IMVC [40] in terms of ACC and purity on Caltech102. However, LI-MKKM-MR significantly outperforms CG-IMVC in terms of NMI. The results on Caltech102-5, Caltech102-10 and Caltech102-15 are provided in the supplemental material due to space limit, whose results demonstrate the same conclusion as well.

Experiments on UCI-Digital Dataset. In this simulation, we apply all the testing methods on UCI-Digital dataset, which is widely utilized in in multiple kernel clustering as a benchmark. For each kind of missing ratio, we generate "incomplete patterns" for 10 times and report their averaged results.

The ACC, NMI and purity of all the testing methods over variational missing ratios are presented in Figure 3. It is clear that the latest proposed MKKM-IK provides unsatisfactory results on UCI-Digital, which is even worse than MKKM+KNN. However, LI-MKKM significantly outperforms the second best one (MKKM+KNN) by $22.2\%, 21.9\%, 20.6\%, 19.5\%, 17.9\%, 17.9\%, 20.4\%, 23.8\%$ and 23.2% on accuracy. In addition, the proposed LI-MKKM-MR further consistently improves the clustering performance of LI-MKKM. The aggregated clustering results in Table V also denote the same performance.

Experiments on Protein Fold Prediction Dataset. In this experiment, the protein fold dataset is applied to evaluate the testing methods, and we report all results in Figure 4 and Table VI. Also, we can find that our LI-MKKM-MR also achieves much better results than the rest algorithms on ACC, NMI and purity on the dataset.

In short, we think our algorithm has three advantages:

- *The joint optimization based on imputation and clustering*. First of all, the process of imputation is guided by the clustering results, which makes the imputation more direct to the final goal. Second, refining the clustering results can benefits from this meaningful imputation. These two learning processes work well together, thus leads to the clustering performance improvement. In contrast, MKKM+MF, MKKM+KNNMKKM+ZF, and MKKM+AF algorithms do not fully make use of the connection between the imputation and clustering procedures. This may produce imputation which does not well serve the subsequent clustering as originally expected, affecting the clustering performance;
- *Considerably utilizing data's local structure*. Our local kernel alignment criterion is flexible and it makes the pre-specified kernels aligned for better clustering performance;
- *Well considering the correlation of incomplete base kernels*. The incorporated matrix-induced regularization reduces the high redundancy and enforces low diversity among the selected kernels, making the pre-specified kernels be well utilized.

These factors have led to significant improvements in cluster performance.

C. Parameter Sensitivity of LI-MKKM-MR

In this part, we analyze that relationship between the clustering performance and matrix-induced regularization. Referring to the Eq.(10), LI-MKKM-MR induces the ratio of the nearest neighbors τ and regularization parameter λ. In the following, we conduct another experiment to show the variation of performance among different τ and λ on Flower17 dataset.

Figure 5a and 5b shows the ACC and NMI of our algorithm by varying τ in a huge range $[0.02 : 0.02 : 0.2]$ with $\lambda = 2^{-6}$. From these figures, we can find that: i) The ACC fluctuates with the monotonically increasing of τ. ii) The start points of the ACC curves are typically higher than the end points, which induces that when the matrix-induced regularization term is dominated at ending points while the local kernel alignment maximization is dominated at starting points. These observations verify the successful joint preservation of local structure of data and matrix-induced regularization term in our algorithm. Similarly, 5c and 5d presents the ACC and NMI of our algorithms by tuning λ from 2^{-9} to 2 with $\tau = 0.1$. In this scenario, our algorithm also shows stable performance over variational λ.

As aforementioned, we conclude that comparing to only preserving global kernel alignment, such as MKKM-IK in

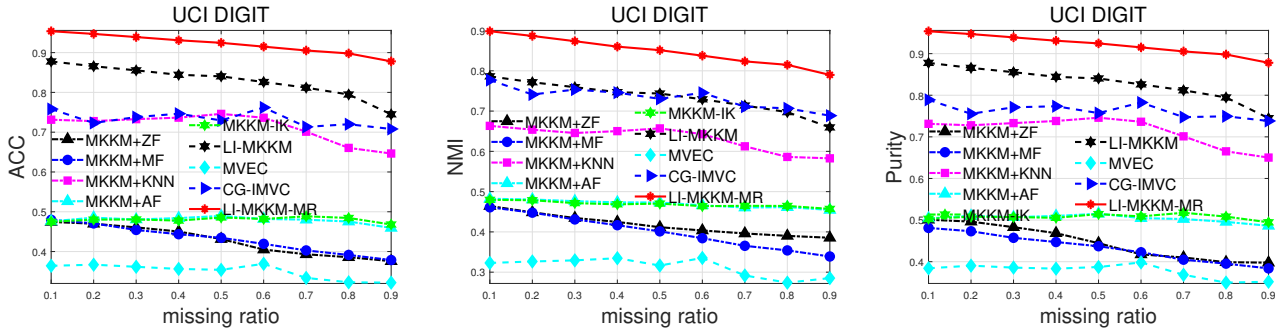

Figure 3: Clustering accuracy, NMI and purity comparison with the variation of missing ratios on UCI-digital dataset.

Table V: Total ACC, NMI and purity comparison (mean±std) of various clustering algorithms on UCI-Digital.

MKKM				MKKM-IK	LI-MKKM	MVEC	CG-IMVC	LI-MKKM-MR
+ZF	+MF	+KNN	+AF [27]	[33]	[32]	[39]	[40]	Proposed
ACC								
42.7 ± 0.4	43.1 ± 0.3	71.3 ± 1.0	47.9 ± 0.5	48.0 ± 0.4	82.9 ± 0.3	35.0 ± 0.8	73.3 ± 1.1	$\mathbf{92.1 \pm 0.3}$
NMI								
41.8 ± 0.2	40.0 ± 0.2	63.3 ± 0.5	47.0 ± 0.2	46.9 ± 0.2	73.4 ± 0.3	31.3 ± 1.1	73.3 ± 0.9	$\mathbf{84.8 \pm 0.4}$
Purity								
44.6 ± 0.5	43.4 ± 0.3	71.4 ± 0.7	50.4 ± 0.3	50.8 ± 0.4	82.9 ± 0.3	37.8 ± 0.8	76.3 ± 1.0	$\mathbf{92.1 \pm 0.3}$

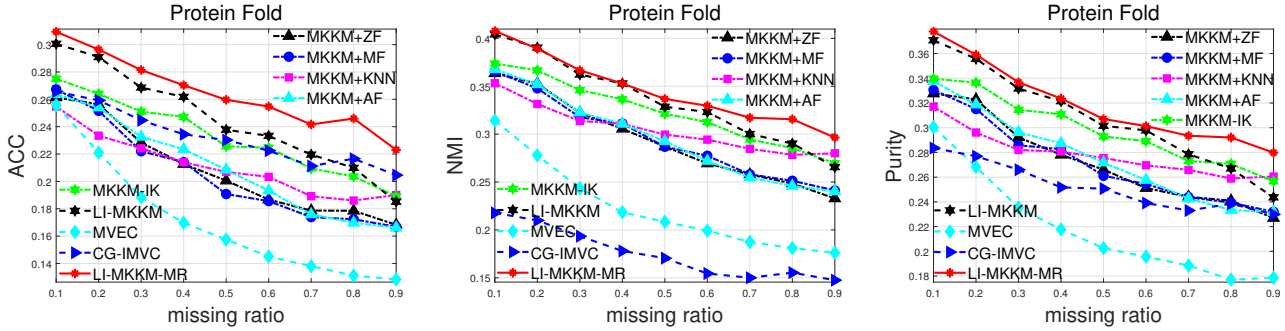

Figure 4: Clustering accuracy, NMI and purity comparison with the variation of missing ratios on protein Fold Prediction dataset.

Table VI: Total ACC, NMI and purity comparison (mean±std) of various clustering algorithms on Protein Fold Prediction dataset.

MKKM				MKKM-IK	LI-MKKM	MVEC	CG-IMVC	LI-MKKM-MR
+ZF	+MF	+KNN	+AF [27]	[33]	[32]	[39]	[40]	Proposed
ACC								
20.8 ± 0.2	20.5 ± 0.3	21.1 ± 0.5	21.0 ± 0.2	23.2 ± 0.6	24.5 ± 0.5	17.1 ± 0.2	23.2 ± 0.3	$\mathbf{26.5 \pm 0.2}$
NMI								
29.3 ± 0.4	29.5 ± 0.5	30.5 ± 0.4	29.5 ± 0.3	32.3 ± 0.6	33.5 ± 0.3	22.3 ± 0.2	17.5 ± 0.6	$\mathbf{34.6 \pm 0.2}$
Purity								
27.2 ± 0.4	27.2 ± 0.4	27.9 ± 0.5	27.5 ± 0.4	29.8 ± 0.7	30.8 ± 0.4	21.8 ± 0.2	25.2 ± 0.5	$\mathbf{31.9 \pm 0.3}$

[33], our proposed algorithms is more essential to the clustering performance by preserving the local structure of data. Meanwhile, the clustering performance could be further improved by incorporating the correlation among base kernels. By appropriately integrating these two factors, it is possible to obtain the best clustering performance. Practically, there exists a trade-off between the preservation of the local geometric structure and the correlation of base kernels to ensure the best clustering.

D. Convergence of LI-MKKM-MR

According to [42], the convergence of our proposed algorithm is guaranteed. We present one simulation trail of the proposed LI-MKKM-MR on Flower 17 dataset as an example in 6. It is clearly shown that the objective value of the proposed

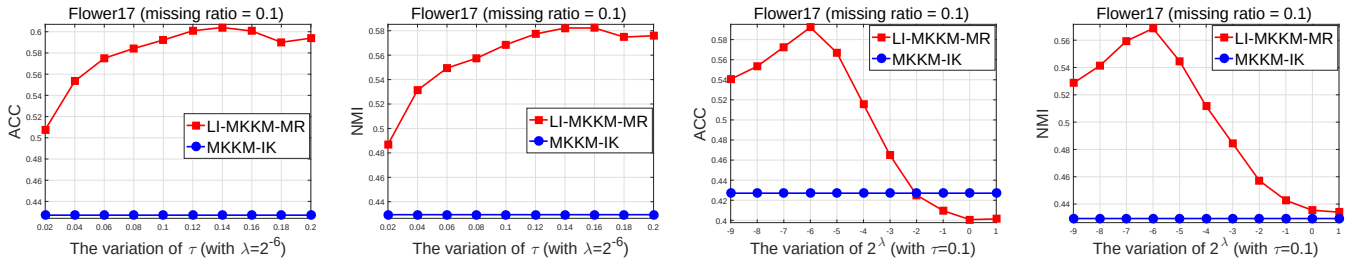

Figure 5: The sensitivity of the proposed LI-MKKM-MR with the variation of λ and τ.

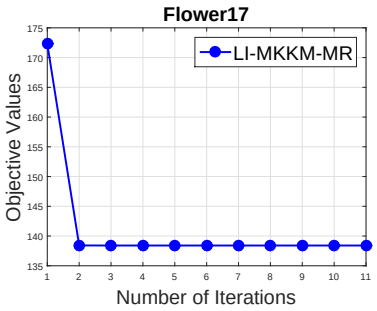

Figure 6: Proposed algorithm convergence illustration.

algorithm is monotonically decreased and converges in a few iteration.

VI. Conclusion

Though the newly proposed LI-MKKM is able to tackle the task of multiple kernel clustering with incomplete kernels, it takes the correlation among base kernels into account insufficiently. We propose to calculate the kernel alignment to address this issue together with a matrix-induced regularization in a local manner. The proposed algorithm efficiently solves the resultant optimization problem, and extensive experiments on benchmarks have demonstrated that LI-MKKM-MR consistently outperforms state-of-the-art baseline algorithms. In the future, instead of keeping the nearest neighbors of each sample unchanged, updating them automatically during the learning course will be further investigated for clustering performance improvement. Moreover, we will design efficient and effective algorithms to solve the optimization problem directly without approximating $\mathbf{M}$ in Eq. (9).

Acknowledgements

This work was supported by the Natural Science Foundation of China (project no. 61922088, 61773392 and 61976196), and Outstanding Talents of "Ten Thousand Talents Plan" in Zhejiang Province (project no. 2018R51001).

References

[1] K. Zhan, X. Chang, J. Guan, L. Chen, Z. Ma, and Y. Yang, "Adaptive structure discovery for multimedia analysis using multiple features," *IEEE Trans. Cybernetics*, vol. 49, no. 5, pp. 1826–1834, 2019.

[2] K. Zhan, F. Nie, J. Wang, and Y. Yang, "Multiview consensus graph clustering," *IEEE Trans. Image Processing*, vol. 28, no. 3, pp. 1261–1270, 2019.

[3] D. Huang, J. Lai, and C. Wang, "Robust ensemble clustering using probability trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1312–1326, 2016.

[4] C. Wang, J. Lai, and P. S. Yu, "Multi-view clustering based on belief propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 1007–1021, 2016.

[5] M. Yin, J. Gao, S. Xie, and Y. Guo, "Multiview subspace clustering via tensorial t-product representation," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 30, no. 3, pp. 851–864, 2019.

[6] Z. Ren, S. X. Yang, Q. Sun, and T. Wang, "Consensus affinity graph learning for multiple kernel clustering," *IEEE Transactions on Cybernetics*, pp. 1–12, 2020.

[7] K. Zhan, C. Niu, C. Chen, F. Nie, C. Zhang, and Y. Yang, "Graph structure fusion for multiview clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1984–1993, 2018.

[8] W. Liang, S. Zhou, J. Xiong, X. Liu, S. Wang, E. Zhu, Z. Cai, and X. Xu, "Multi-view spectral clustering with high-order optimal neighborhood laplacian matrix," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[9] B. Zhao, J. T. Kwok, and C. Zhang, "Multiple kernel clustering," in *SDM*, 2009, pp. 638–649.

[10] Z. Ren and Q. Sun, "Simultaneous global and local graph structure preserving for multiple kernel clustering," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2020.

[11] S. Li, Y. Jiang, and Z. Zhou, "Partial multi-view clustering," in *AAAI*, 2014, pp. 1968–1974.

[12] M. Gönen and A. A. Margolin, "Localized data fusion for kernel k-means clustering with application to cancer biology," in *NIPS*, 2014, pp. 1305–1313.

[13] L. Du, P. Zhou, L. Shi, H. Wang, M. Fan, W. Wang, and Y.-D. Shen, "Robust multiple kernel k-means clustering using ℓ_{21}-norm." in *IJCAI*, 2015, pp. 3476–3482.

[14] X. Liu, Y. Dou, J. Yin, L. Wang, and E. Zhu, "Multiple kernel k-means clustering with matrix-induced regularization," in *AAAI*, 2016, pp. 1888–1894.

[15] D. Huang, C. Wang, and J. Lai, "Locally weighted ensemble clustering," *IEEE Trans. Cybernetics*, vol. 48, no. 5, pp. 1460–1473, 2018.

[16] M. Li, X. Liu, L. Wang, Y. Dou, J. Yin, and E. Zhu,

"Multiple kernel clustering with local kernel alignment maximization," in *IJCAI*, 2016, pp. 1704–1710.

[17] S. Wang, X. Liu, E. Zhu, C. Tang, J. Liu, J. Hu, J. Xia, and J. Yin, "Multi-view clustering via late fusion alignment maximization." in *IJCAI*, 2019, pp. 3778–3784.

[18] S. Yu, L.-C. Tranchevent, X. Liu, W. Glänzel, J. A. K. Suykens, B. D. Moor, and Y. Moreau, "Optimized data fusion for kernel k-means clustering," *IEEE TPAMI*, vol. 34, no. 5, pp. 1031–1039, 2012.

[19] X. Liu, S. Zhou, Y. Wang, M. Li, Y. Dou, E. Zhu, and J. Yin, "Optimal neighborhood kernel clustering with multiple kernels," in *AAAI*, 2017, pp. 2266–2272.

[20] M. Yin, J. Gao, S. Xie, and Y. Guo, "Multiview subspace clustering via tensorial t-product representation," *IEEE TNNLS*, no. 99, pp. 1–14, 2018.

[21] Q. Wang, Z. Qin, F. Nie, and X. Li, "Spectral embedded adaptive neighbors clustering," *IEEE TNNLS*, no. 99, pp. 1–7, 2018.

[22] M. S. Chen, L. Huang, C. D. Wang, D. Huang, and P. S. Yu, "Multiview subspace clustering with grouping effect," *IEEE Transactions on Cybernetics*, pp. 1–14, 2020.

[23] D. Huang, C. Wang, and J. Lai, "Locally weighted ensemble clustering," *IEEE Transactions on Cybernetics*, vol. 48, no. 5, pp. 1460–1473, 2018.

[24] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye, "Multi-source learning with block-wise missing data for alzheimer's disease prediction," in *ACM SIGKDD*, 2013, pp. 185–193.

[25] R. Kumar, T. Chen, M. Hardt, D. Beymer, K. Brannon, and T. F. Syeda-Mahmood, "Multiple kernel completion and its application to cardiac disease discrimination," in *ISBI*, 2013, pp. 764–767.

[26] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an EM approach," in *NIPS*, 1993, pp. 120–127.

[27] A. Trivedi, P. Rai, H. Daumé III, and S. L. DuVall, "Multiview clustering with incomplete views," in *NIPS 2010: Machine Learning for Social Computing Workshop ,Whistler, Canada*, 2010.

[28] C. Xu, D. Tao, and C. Xu, "Multi-view learning with incomplete views," *IEEE Trans. Image Processing*, vol. 24, no. 12, pp. 5812–5825, 2015.

[29] W. Shao, L. He, and P. S. Yu, "Multiple incomplete views clustering via weighted nonnegative matrix factorization with $\ell_{2,1}$ regularization," in *ECML PKDD*, 2015, pp. 318–334.

[30] S. Bhadra, S. Kaski, and J. Rousu, "Multi-view kernel completion," in *arXiv:1602.02518*, 2016.

[31] X. Liu, M. Li, L. Wang, Y. Dou, J. Yin, and E. Zhu, "Multiple kernel k-means with incomplete kernels," in *AAAI*, 2017, pp. 2259–2265.

[32] X. Zhu, X. Liu, M. Li, E. Zhu, L. Liu, Z. Cai, J. Yin, and W. Gao, "Localized incomplete multiple kernel k-means," in *IJCAI*, 2018, pp. 3271–3277.

[33] X. Liu, X. Zhu, M. Li, L. Wang, E. Zhu, T. Liu, M. Kloft, D. Shen, J. Yin, and W. Gao, "Multiple kernel k-means with incomplete kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1191–1204, 2019.

[34] S. Jegelka, A. Gretton, B. Schölkopf, B. K. Sriperumbudur, and U. von Luxburg, "Generalized clustering via kernel embeddings," in *KI 2009: Advances in Artificial Intelligence, 32nd Annual German Conference on AI*, 2009, pp. 144–152.

[35] C. Cortes, M. Mohri, and A. Rostamizadeh, "Algorithms for learning kernels based on centered alignment," *JMLR*, vol. 13, pp. 795–828, 2012.

[36] A. Maurer and M. Pontil, "k-dimensional coding schemes in Hilbert spaces," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5839–5846, 2010.

[37] T. Liu, D. Tao, and D. Xu, "Dimensionality-dependent generalization bounds for k-dimensional coding schemes," *Neural computation*, vol. 28, no. 10, pp. 2213–2249, 2016.

[38] X. Liu, X. Zhu, M. Li, L. Wang, E. Zhu, T. Liu, M. Kloft, D. Shen, J. Yin, and W. Gao, "Multiple kernel k-means with incomplete kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[39] Z. Tao, H. Liu, S. Li, Z. Ding, and Y. Fu, "From ensemble clustering to multi-view clustering," in *IJCAI*, 2017, pp. 2843–2849.

[40] W. Zhou, H. Wang, and Y. Yang, "Consensus graph learning for incomplete multi-view clustering," in *Advances in Knowledge Discovery and Data Mining*, 2019, pp. 529–540.

[41] H. Zhao, H. Liu, and Y. Fu, "Incomplete multimodal visual data grouping," in *IJCAI*, 2016, pp. 2392–2398.

[42] J. C. Bezdek and R. J. Hathaway, "Convergence of alternating optimization," *Neural, Parallel Sci. Comput.*, vol. 11, no. 4, pp. 351–368, 2003.

Appendix of "Localized Incomplete Multiple Kernel k-means with Matrix-induced Regularization"

Jingyuan Xia, Miaomiao Li, Huiying Xu, Qing Liao, Xinzhong Zhu, Xinwang Liu

I. Summary of the Appendix

The following sections are arranged as follows. In Section II and III, we provide the generalization analysis of the proposed LI-MKKM-MR and give the detailed proof. After that, we report some additional experimental results in Section IV, including all clustering results on Caltech102-5, Caltech102-10 and Caltech102-15.

II. Theoretical Results

The generalization error of k-means clustering algorithm measures the clustering performance of the learned clusters in training procedure on unseen samples [1–3]. In this section, we first build the theoretical connection between existing MKKM-IK [3] with LI-MKKM-MR, and then derive the generalization error bound of the proposed LI-MKKM-MR based on the theoretical results in [3]. The following Theorem 1 states that the local kernel alignment adopted in our LI-MKKM-MR can be fulfilled by normalizing each base kernel matrix.

Theorem 1: The local kernel alignment criterion in Eq. (8) (in the manuscript) is equivalent to the widely adopted global kernel alignment by normalizing each base kernel matrix.

Proof 1: The objective function in Eq. (8) (in the manuscript) can be written as

$$\begin{aligned}
&\sum_{i=1}^{n} \operatorname{Tr}\left(\mathbf{K}_{\boldsymbol{\gamma}}(\mathbf{B}^{(i)}-\mathbf{B}^{(i)}\mathbf{H}\mathbf{H}^{\top}\mathbf{B}^{(i)})\right)\\
&=\sum_{i=1}^{n}\langle \mathbf{B}^{(i)}\otimes\mathbf{K}_{\boldsymbol{\gamma}}, \mathbf{B}^{(i)}\otimes(\mathbf{I}-\mathbf{H}\mathbf{H}^{\top})\rangle_{\mathrm{F}}\\
&=\sum_{i=1}^{n}\langle \mathbf{B}^{(i)}\otimes\mathbf{K}_{\boldsymbol{\gamma}}, \mathbf{I}-\mathbf{H}\mathbf{H}^{\top}\rangle_{\mathrm{F}}\\
&=\langle\Big(\sum_{i=1}^{n}\mathbf{B}^{(i)}\Big)\otimes\mathbf{K}_{\boldsymbol{\gamma}}, \mathbf{I}-\mathbf{H}\mathbf{H}^{\top}\rangle_{\mathrm{F}}\\
&=\sum_{p=1}^{m}\gamma_p^2\langle\Big(\sum_{i=1}^{n}\mathbf{B}^{(i)}\Big)\otimes\mathbf{K}_{p}, \mathbf{I}-\mathbf{H}\mathbf{H}^{\top}\rangle_{\mathrm{F}}\\
&=\sum_{p=1}^{m}\gamma_p^2\langle\tilde{\mathbf{K}}_{p}, \mathbf{I}-\mathbf{H}\mathbf{H}^{\top}\rangle_{\mathrm{F}}\\
&=\operatorname{Tr}\left(\tilde{\mathbf{K}}_{\boldsymbol{\gamma}}(\mathbf{I}-\mathbf{H}\mathbf{H}^{\top})\right),
\end{aligned} \tag{1}$$

where $\otimes$ denotes elementwise multiplication between two matrices, $\tilde{\mathbf{K}}_p=\left(\sum_{i=1}^{n}\mathbf{B}^{(i)}\right)\otimes\mathbf{K}_p$ can be treated as a normalized $\mathbf{K}_p$, and $\tilde{\mathbf{K}}_{\boldsymbol{\gamma}}=\sum_{p=1}^{m}\gamma_p^2\tilde{\mathbf{K}}_p$. Consequently, by such normalization applied on each base kernel, we can clearly see that the local kernel alignment criterion in Eq. (8) (in the manuscript) is exactly the global kernel alignment in [3]. This completes the proof.

Let $t(\mathbf{x}^{(p)})=1$ if the p-th view of $\mathbf{x}$ is available, otherwise $\mathbf{x}^{(p)}$ needs to be optimized. Note that $t(\mathbf{x}^{(p)})$ is a random variable which depends on $\mathbf{x}$. Let $\hat{\mathbf{C}}=[\hat{\mathbf{C}}_1,\cdots,\hat{\mathbf{C}}_k]$ be the k centroids and $\hat{\boldsymbol{\beta}}$ be the kernel weights learned by LI-MKKM-MR. k-means clustering should make the following reconstruction error small

$$\mathbb{E}\left[\min_{\mathbf{y}\in\{\mathbf{e}_1,\cdots,\mathbf{e}_k\}}\left\|\phi_{\hat{\boldsymbol{\gamma}}}(\mathbf{x})-\hat{\mathbf{C}}\mathbf{y}\right\|_{\mathcal{H}}^2\right], \tag{2}$$

where $\phi_{\hat{\boldsymbol{\gamma}}}(\mathbf{x})=[\hat{\gamma}_1 t(\mathbf{x}^{(1)})\phi_1^{\top}(\mathbf{x}^{(1)}),\cdots,\hat{\gamma}_m t(\mathbf{x}^{(m)})\phi_m^{\top}(\mathbf{x}^{(m)})]^{\top}$,
$\mathbf{e}_1,\cdots,\mathbf{e}_k$ form the orthogonal bases of $\mathbb{R}^k$.

J. Xia is with Department of Electric and Electronic Engineering, Imperial College London, London, SW72AZ, UK.
M. Li and X. Liu are with College of Computer, National University of Defense Technology, Changsha, 410073, China. M. Li is also with college of electronic information and electrical engineering, Changsha University, Changsha, 410073, China.
H. Xu and X. Zhu are with the College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua 321004, China. X. Zhu is also with the Research Institute of Ningbo Cixing Co. Ltd, Ningbo 315336, China. (E-mail: {zxz, xyh}@zjnu.edu.cn).
Q. Liao is with the Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, China.
J. Xia, M. Li and H. Xu contribute equally to this work.
Corresponding authors: Xinzhong Zhu and Xinwang Liu (E-mail: xinwangliu@nudt.edu.cn).
Manuscript received January, 2021.

We first define a function class:

$$\begin{aligned}\mathcal{F} = \Big\{ f : \mathbf{x} \mapsto \min_{\mathbf{y}\in\{\mathbf{e}_1,\ldots,\mathbf{e}_k\}} \|\phi_{\boldsymbol{\gamma}}(\mathbf{x}) - \mathbf{C}\mathbf{y}\|_{\mathcal{H}}^2 \Big| \boldsymbol{\gamma}^\top \mathbf{1}_m = 1,\ \gamma_p \geq 0, \\ \mathbf{C} \in \mathcal{H}^k,\ t(\mathbf{x}_i^{(p)}) t(\mathbf{x}_j^{(p)}) \tilde{\kappa}_p^\top (\mathbf{x}_i^{(p)}, \mathbf{x}_j^{(p)}) \leq b,\ \forall p, \forall \mathbf{x}_i \in \mathcal{X} \Big\},\end{aligned} \tag{3}$$

where $\mathcal{H}^k$ represents the multiple kernel Hilbert space and $\tilde{\kappa}(\cdot,\cdot)$ is a kernel function corresponding to $\tilde{\mathbf{K}}_p$.

Based on Theorem 1, we derive the generalization error bound of the proposed LI-MKKM-MR by following [3].

Theorem 2: For any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $f \in \mathcal{F}$:

$$\begin{aligned}\mathbb{E}[f(\mathbf{x})] \leq \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i) + \frac{4\sqrt{\pi} m b \mathcal{G}_{1n}(\boldsymbol{\gamma}, t)}{n} + \frac{4\sqrt{\pi m} b \mathcal{G}_{2n}(\boldsymbol{\gamma}, t)}{n} \\ + \frac{\sqrt{8\pi} b k^2}{\sqrt{n}} + 2b\sqrt{\frac{\log 1/\delta}{2n}},\end{aligned} \tag{4}$$

where

$$\mathcal{G}_{1n}(\boldsymbol{\gamma}, t) \triangleq \mathbb{E}_\beta \left[\sup_{\boldsymbol{\gamma}, t} \sum_{i=1}^n \sum_{p,q=1}^m \beta_{ipq} t(\mathbf{x}_i^{(p)}) t(\mathbf{x}_i^{(q)}) \gamma_p \gamma_q \right], \tag{5}$$

$$\mathcal{G}_{2n}(\boldsymbol{\gamma}, t) = \mathbb{E}_\beta \left[\sup_{\boldsymbol{\gamma}, t} \sum_{i=1}^n \sum_{c=1}^k \sum_{p=1}^m \beta_{icp} \gamma_p t(\mathbf{x}_i^{(p)}) \right], \tag{6}$$

and $\beta_{ipq}, \beta_{icp}, i \in \{1,\ldots,n\}, p, q \in \{1,\ldots,m\}, c \in \{1,\ldots,k\}$ are i.i.d. Gaussian random variables with zero mean and unit standard deviation.

According to the analysis in [3], our local kernel alignment criterion in Eq. (8) (in the manuscript), with normalized base kernel matrices, is an upper bound of $\frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)$. As a result, by minimizing $\mathrm{Tr}(\tilde{\mathbf{K}}_{\boldsymbol{\beta}}(\mathbf{I}_n - \mathbf{H}\mathbf{H}^\top))$, one can get a small $\frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)$ for good generalization. This justifies the good generalization ability of the proposed algorithm. We provide the detailed proof in the appendix due to space limit.

III. Proof of Theorem 2

In the following, we give the detailed proof of Theorem 2. For an i.i.d. given sample $\{\mathbf{x}_i\}_{i=1}^n$, multiple kernel k-means algorithm is to minimize an empirical reconstruction error, i.e.,

$$\min_{\mathbf{C}} \frac{1}{n} \sum\nolimits_{i=1}^n \min_{\mathbf{y}\in\{\mathbf{e}_1,\cdots,\mathbf{e}_k\}} \|\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}_i) - \mathbf{C}\mathbf{y}\|_{\mathcal{H}}^2, \tag{7}$$

where $\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}_i) = [\gamma_1 t(\mathbf{x}_i^{(1)}) \phi_1^\top(\mathbf{x}_i^{(1)}), \cdots, \gamma_m t(\mathbf{x}_i^{(m)}) \phi_m^\top(\mathbf{x}_i^{(m)})]^\top$, $\mathbf{e}_1, \cdots, \mathbf{e}_k$ form the orthogonal bases of $\mathbb{R}^k$.

Let

$$\hat{R}(\mathbf{C}, \boldsymbol{\gamma}, \{\mathbf{K}_p\}_{p=1}^m) = \frac{1}{n} \sum\nolimits_{i=1}^n \min_{\mathbf{y}\in\{\mathbf{e}_1,\cdots,\mathbf{e}_k\}} \|\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}_i) - \mathbf{C}\mathbf{y}\|_{\mathcal{H}}^2 \ . \tag{8}$$

Our proof idea is to upper bound

$$\sup_{\mathbf{C}, \boldsymbol{\gamma}, \{\mathbf{K}_p\}_{p=1}^m} \left(\mathbb{E}\left[\hat{R}(\mathbf{C}, \boldsymbol{\gamma}, \{\mathbf{K}_p\}_{p=1}^m) \right] - \hat{R}(\mathbf{C}, \boldsymbol{\gamma}, \{\mathbf{K}_p\}_{p=1}^m) \right), \tag{9}$$

and then upper bound the term $\hat{R}(\mathbf{C}, \boldsymbol{\gamma}, \{\mathbf{K}_p\}_{p=1}^m)$ by the proposed objective.

Note that in the proposed algorithms, the absent views are completed by exploiting the observed views. We assume that the kernel mappings for both the observed views and the completed views are upper bounded, i.e., every entry of $\mathbf{K}_p, p \in \{1,\ldots,m\}$, are no larger than b. Let us define a function class first:

$$\mathcal{F} = \Big\{ f : \mathbf{x} \mapsto \min_{\mathbf{y}\in\{e_1,\ldots,e_k\}} \|\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}) - \mathbf{C}\mathbf{y}\|_{\mathcal{H}}^2 \Big| \boldsymbol{\gamma}^\top \mathbf{1}_m = 1,\ \gamma_p \geq 0,\ \mathbf{C} \in \mathcal{H}^k,\ t(\mathbf{x}_i^{(p)}) t(\mathbf{x}_j^{(p)}) \kappa_p(\mathbf{x}_i^{(p)}, \mathbf{x}_j^{(p)}) \leq b,\ \forall p, \forall \mathbf{x}_i \in \mathcal{X} \Big\}, \tag{10}$$

where $\mathcal{H}^k$ stands for the multiple kernel Hilbert space.

Then, Eq. (20) becomes

$$\sup_{f\in\mathcal{F}} \left(\mathbb{E}\left[f(\mathbf{x}) \right] - \frac{1}{n} \sum\nolimits_{i=1}^n f(\mathbf{x}_i) \right). \tag{11}$$

Let $\boldsymbol{\gamma}$, the kernel matrices induced by absent views, and $\mathbf{C} = [\mathbf{C}_1, \ldots, \mathbf{C}_k]$ all be learned from predefined hypothesis classes. Note that

$$f(\mathbf{x}) = \min_{\mathbf{y}\in\{\mathbf{e}_1,\cdots,\mathbf{e}_k\}} \|\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}) - \mathbf{C}\mathbf{y}\|_{\mathcal{H}^k}^2 \ = \ \min \left\{ \|\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}) - \mathbf{C}_1\|_{\mathcal{H}^k}^2, \cdots, \|\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}) - \mathbf{C}_k\|_{\mathcal{H}^k}^2 \right\} \tag{12}$$

and for $v \in \{1, \cdots, k\}$

$$\begin{aligned}\|\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}) - \mathbf{C}_v\|_{\mathcal{H}^k}^2 &= \left\|\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}) - \frac{1}{|\mathbf{C}_v|}\sum\nolimits_{j\in\mathbf{C}_v}\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}_j)\right\|_{\mathcal{H}^k}^2 \leq 2\left(\phi_{\boldsymbol{\gamma},\mathbf{t}}^\top(\mathbf{x})\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}) + \frac{1}{|\mathbf{C}_v|^2}\sum\nolimits_{j_1,j_2\in\mathbf{C}_v}\phi_{\boldsymbol{\gamma},\mathbf{t}}^\top(\mathbf{x}_{j_1})\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}_{j_2})\right)\\ &= 2\left(\sum\nolimits_{p=1}^m \gamma_p^2 t^2(\mathbf{x}^{(p)})\phi_p^\top(\mathbf{x}^{(p)})\phi_p(\mathbf{x}^{(p)}) + \frac{1}{|\mathbf{C}_v|^2}\sum\nolimits_{j_1,j_2\in\mathbf{C}_v}\sum\nolimits_{p=1}^m \gamma_p^2 t(\mathbf{x}_{j_1}^{(p)})t(\mathbf{x}_{j_2}^{(p)})\phi_p^\top(\mathbf{x}_{j_1}^{(p)})\phi_p(\mathbf{x}_{j_2}^{(p)})\right)\\ &\leq 2\left(b\sum\nolimits_{p=1}^m \gamma_p^2 + \frac{b}{|\mathbf{C}_v|^2}\sum\nolimits_{j_1,j_2\in\mathbf{C}_v}\sum\nolimits_{p=1}^m \gamma_p^2\right) \leq 2\left(b\sum\nolimits_{p=1}^m \gamma_p + \frac{b}{|\mathbf{C}_v|^2}\sum\nolimits_{j_1,j_2\in\mathbf{C}_v}\sum\nolimits_{p=1}^m \gamma_p\right)\\ &= 4b.\end{aligned} \tag{13}$$

As a result, we have $f(\mathbf{x}) \leq 4b$.

By exploiting McDiarmid's concentration inequality, we have the following theorem [4].

Theorem 3: For any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $f \in \mathcal{F}$:

$$\mathbb{E}\left[f(\mathbf{x})\right] - \frac{1}{n}\sum\nolimits_{i=1}^n f(\mathbf{x}_i) \leq 2\mathfrak{R}_n(\mathcal{F}) + 4b\sqrt{\frac{\log 1/\delta}{2n}}, \tag{14}$$

where

$$\mathfrak{R}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum\nolimits_{i=1}^n \sigma_i f(\mathbf{x}_i)\right] \tag{15}$$

and $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables uniformly distributed from $\{-1, 1\}$.

Now, we are going to upper bound $\mathfrak{R}_n(\mathcal{F})$. Since there is a minimization function in f, it is not easy to directly upper $\mathfrak{R}_n(\mathcal{F})$. Similar to the proof method in [1], we upper bound it by introducing Gaussian complexities:

$$\mathfrak{G}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left[\sup\nolimits_{f\in\mathcal{F}}\sum\nolimits_{i=1}^n \beta_i f(\mathbf{x}_i)\right], \tag{16}$$

where $\beta_1, \cdots, \beta_n$ are i.i.d. Gaussian random variables with zero mean and unit standard deviation.

The following two lemmas [1] will be used in our proof.

Lemma 4:

$$\mathfrak{R}_n(\mathcal{F}) \leq \sqrt{\pi/2}\mathfrak{G}_n(\mathcal{F}). \tag{17}$$

Lemma 5: Let $G_f = \sum_{i=1}^n \beta_i G(\mathbf{x}_i, f)$ and $H_f = \sum_{i=1}^n \beta_i H(\mathbf{x}_i, f)$ be two zero mean, separable Gaussian processes. If for all $f_1, f_2 \in \mathcal{F}$,

$$\mathbb{E}\left[(G_{f_1} - G_{f_2})^2\right] \leq \mathbb{E}[(H_{f_1} - H_{f_2})^2]\ . \tag{18}$$

Then,

$$\mathbb{E}\left[\sup\nolimits_{f\in\mathcal{F}} G_f\right] \leq \mathbb{E}\left[\sup\nolimits_{f\in\mathcal{F}} H_f\right]. \tag{19}$$

In our case, let

$$G_{\boldsymbol{\gamma},t,\mathbf{C}} = \sum\nolimits_{i=1}^n \beta_i \min_{\mathbf{y}\in\{\mathbf{e}_1,\cdots,\mathbf{e}_k\}} \|\phi_{\boldsymbol{\gamma},\mathbf{t}}(\mathbf{x}_i) - \mathbf{C}\mathbf{y}\|_{\mathcal{H}^k}^2\ , \tag{20}$$

and

$$H_{\boldsymbol{\gamma},t,\mathbf{C}} = 2b\sqrt{m}\sum\nolimits_{i=1}^n\sum\nolimits_{p=1}^m \beta_{ip}\gamma_p^2 t^2(\mathbf{x}^p) + \sqrt{8}\sum\nolimits_{i=1}^n\sum\nolimits_{c=1}^k \beta_{ic}\phi_{\boldsymbol{\gamma},\mathbf{t}}^\top(\mathbf{x}_i)\mathbf{C}\mathbf{e}_c + 2\sum\nolimits_{i=1}^n\sum\nolimits_{c,l=1}^k \beta_{icl}\mathbf{e}_c^\top\mathbf{C}^\top\mathbf{C}\mathbf{e}_l. \tag{21}$$

we are going to prove that

$$\mathbb{E}_\gamma\left[(G_{\boldsymbol{\gamma}_1,t_1(\mathbf{x}),\mathbf{C}_1} - G_{\boldsymbol{\gamma}_2,t_2(\mathbf{x}),\mathbf{C}_2})^2\right] \leq \mathbb{E}_\gamma\left[(H_{\boldsymbol{\gamma}_1,t_1(\mathbf{x}),\mathbf{C}_1} - H_{\boldsymbol{\gamma}_2,t_2(\mathbf{x}),\mathbf{C}_2})^2\right]. \tag{22}$$

Specifically, for any $f_1, f_2 \in \mathcal{F}$, we have

$$
\begin{aligned}
&\left(\min_{\mathbf{y}} \left\|\phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}(\mathbf{x}) - \mathbf{C}_1\mathbf{y}\right\|_{\mathcal{H}^k}^2 - \min_{\mathbf{y}} \left\|\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}(\mathbf{x}) - \mathbf{C}_2\mathbf{y}\right\|_{\mathcal{H}^k}^2\right)^2 \\
&\le \left(\max_{\mathbf{y}} \left\{\left\|\phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}(\mathbf{x}) - \mathbf{C}_1\mathbf{y}\right\|_{\mathcal{H}^k}^2 - \left\|\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}(\mathbf{x}) - \mathbf{C}_2\mathbf{y}\right\|_{\mathcal{H}^k}^2\right\}\right)^2 \\
&= \left(\left(\left\|\phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}(\mathbf{x})\right\|_{\mathcal{H}^k}^2 - \left\|\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}(\mathbf{x})\right\|_{\mathcal{H}^k}^2\right) + \max_{\mathbf{y}} \left\{2\left(\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}^\top(\mathbf{x})\mathbf{C}_2 - \phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}^\top(\mathbf{x})\mathbf{C}_1\right)\mathbf{y} + \mathbf{y}^\top\left(\mathbf{C}_1^\top\mathbf{C}_1 - \mathbf{C}_2^\top\mathbf{C}_2\right)\mathbf{y}\right\}\right)^2 \\
&\le \left(\left(\left\|\phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}(\mathbf{x})\right\|_{\mathcal{H}^k}^2 - \left\|\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}(\mathbf{x})\right\|_{\mathcal{H}^k}^2\right) + \max_{\mathbf{y}} 2\left(\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}^\top(\mathbf{x})\mathbf{C}_2 - \phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}^\top(\mathbf{x})\mathbf{C}_1\right)\mathbf{y} + \max_{\mathbf{y}} \mathbf{y}^\top\left(\mathbf{C}_1^\top\mathbf{C}_1 - \mathbf{C}_2^\top\mathbf{C}_2\right)\mathbf{y}\right)^2 \\
&= \left(\left(\|\phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}(\mathbf{x})\|_{\mathcal{H}^k}^2 - \|\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}(\mathbf{x})\|_{\mathcal{H}^k}^2\right) + \max_{\mathbf{y}} 2\sum_{c=1}^k y_c\left(\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}^\top(\mathbf{x})\mathbf{C}_2 - \phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}^\top(\mathbf{x})\mathbf{C}_1\right)\mathbf{e}_c + \max_{\mathbf{y}} \sum_{c,l=1}^k y_c y_l \mathbf{e}_c^\top\left(\mathbf{C}_1^\top\mathbf{C}_1 - \mathbf{C}_2^\top\mathbf{C}_2\right)\mathbf{e}_l\right)^2 \\
&\le 4\left(\left\|\phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}(\mathbf{x})\right\|_{\mathcal{H}^k}^2 - \left\|\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}(\mathbf{x})\right\|_{\mathcal{H}^k}^2\right)^2 + 2\left(\max_{\mathbf{y}} 2\sum\nolimits_{c=1}^k y_c\left(\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}^\top(\mathbf{x})\mathbf{C}_2 - \phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}^\top(\mathbf{x})\mathbf{C}_1\right)\mathbf{e}_c\right)^2 \\
&\quad + 4\left(\max_{\mathbf{y}} \sum\nolimits_{c,l=1}^k y_c y_l \mathbf{e}_c^\top\left(\mathbf{C}_1^\top\mathbf{C}_1 - \mathbf{C}_2^\top\mathbf{C}_2\right)\mathbf{e}_l\right)^2 \\
&= 4\left(\sum_{p=1}^m \left(\gamma_{1p}^2 t_1^2(\mathbf{x}^{(p)}) - \gamma_{2p}^2 t_2^2(\mathbf{x}^{(p)})\right)\kappa_p(\mathbf{x}^{(p)},\mathbf{x}^{(p)})\right)^2 + 2\left(\max_{\mathbf{y}} 2\sum\nolimits_{c=1}^k y_c\left(\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}^\top(\mathbf{x})\mathbf{C}_2 - \phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}^\top(\mathbf{x})\mathbf{C}_1\right)\mathbf{e}_c\right)^2 \\
&\quad + 4\left(\max_{\mathbf{y}} \sum\nolimits_{c,l=1}^k y_c y_l \mathbf{e}_c^\top\left(\mathbf{C}_1^\top\mathbf{C}_1 - \mathbf{C}_2^\top\mathbf{C}_2\right)\mathbf{e}_l\right)^2 \\
&\le 4mb^2\sum_{p=1}^m \left(\gamma_{1p}^2 t_1^2(\mathbf{x}^{(p)}) - \gamma_{2p}^2 t_2^2(\mathbf{x}^{(p)})\right)^2 + 8\max_{\mathbf{y}}\left(\sum\nolimits_{c=1}^k y_c\left(\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}^\top(\mathbf{x})\mathbf{C}_2 - \phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}^\top(\mathbf{x})\mathbf{C}_1\right)\mathbf{e}_c\right)^2 \\
&\quad + 4\max_{\mathbf{y}}\left(\sum\nolimits_{c,l=1}^k y_c y_l \mathbf{e}_c^\top\left(\mathbf{C}_1^\top\mathbf{C}_1 - \mathbf{C}_2^\top\mathbf{C}_2\right)\mathbf{e}_l\right)^2 \\
&\le 4mb^2\sum_{p=1}^m \left(\gamma_{1p}^2 t_1^2(\mathbf{x}^{(p)}) - \gamma_{2p}^2 t_2^2(\mathbf{x}^{(p)})\right)^2 + 8\sum_{c=1}^k\left(\left(\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}^\top(\mathbf{x})\mathbf{C}_2 - \phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}^\top(\mathbf{x})\mathbf{C}_1\right)\mathbf{e}_c\right)^2 + 4\sum_{c,l=1}^k\left(\mathbf{e}_c^\top\left(\mathbf{C}_1^\top\mathbf{C}_1 - \mathbf{C}_2^\top\mathbf{C}_2\right)\mathbf{e}_l\right)^2
\end{aligned}
$$

where the last inequality holds because $(a+b+c)^2 \le 4a^2 + 2b^2 + 4c^2$, Cauchy-Schwarz inequality, and that $\sum_{c=1}^k y_c = 1$ and $\sum_{c,l=1}^k y_c y_l = 1$.

Thus, we have

$$
\begin{aligned}
&\mathbb{E}_\beta\left[\left(G_{\boldsymbol{\gamma}_1,t_1,\mathbf{C}_1} - G_{\boldsymbol{\gamma}_2,t_2,\mathbf{C}_2}\right)^2\right] \\
=\ &\mathbb{E}_\beta\left[\left(\sum\nolimits_{i=1}^n \beta_i\left[\min_{\mathbf{y}\in\{\mathbf{e}_1,\cdots,\mathbf{e}_k\}} \left\|\phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}(\mathbf{x}_i) - \mathbf{C}_1\mathbf{y}\right\|_{\mathcal{H}^k}^2 - \min_{\mathbf{y}\in\{\mathbf{e}_1,\cdots,\mathbf{e}_k\}} \left\|\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}(\mathbf{x}_i) - \mathbf{C}_2\mathbf{y}\right\|_{\mathcal{H}^k}^2\right]\right)^2\right] \\
=\ &\sum\nolimits_{i=1}^n\left(\min_{\mathbf{y}\in\{\mathbf{e}_1,\cdots,\mathbf{e}_k\}} \left\|\phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}(\mathbf{x}_i) - \mathbf{C}_1\mathbf{y}\right\|_{\mathcal{H}^k}^2 - \min_{\mathbf{y}\in\{\mathbf{e}_1,\cdots,\mathbf{e}_k\}} \left\|\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}(\mathbf{x}_i) - \mathbf{C}_2\mathbf{y}\right\|_{\mathcal{H}^k}^2\right)^2 \\
\le\ &4mb^2\sum\nolimits_{i=1}^n\sum\nolimits_{p=1}^m\left(\gamma_{1p}^2 t_1^2(\mathbf{x}^{(p)}) - \gamma_{2p}^2 t_2^2(\mathbf{x}^{(p)})\right)^2 + 8\sum\nolimits_{i=1}^n\sum\nolimits_{c=1}^k\left(\left(\phi_{\boldsymbol{\gamma}_2,\mathbf{t}_2}^\top(\mathbf{x})\mathbf{C}_2 - \phi_{\boldsymbol{\gamma}_1,\mathbf{t}_1}^\top(\mathbf{x})\mathbf{C}_1\right)\mathbf{e}_c\right)^2 \\
&\quad +4\sum\nolimits_{i=1}^n\sum\nolimits_{c,l=1}^k\left(\mathbf{e}_c^\top\left(\mathbf{C}_1^\top\mathbf{C}_1 - \mathbf{C}_2^\top\mathbf{C}_2\right)\mathbf{e}_l\right)^2 \\
=\ &\mathbb{E}_\beta\left[\left(H_{\boldsymbol{\gamma}_1,t_1,\mathbf{C}_1} - H_{\boldsymbol{\gamma}_2,t_2,\mathbf{C}_2}\right)^2\right].
\end{aligned} \tag{23}
$$

Moreover, since $t(\mathbf{x}^{(p)}) \in \{0,1\}$ and $\|\mathbf{C}\mathbf{e}_c\|_{\mathcal{H}}^2 \le b$ for all p and c, we have

$$
\begin{aligned}
&\mathbb{E}_\beta\left[2b\sqrt{m}\sup_{\boldsymbol{\gamma},t}\sum\nolimits_{i=1}^n\sum\nolimits_{p=1}^m \beta_{ip}\gamma_p^2 t^2(\mathbf{x}_i^{(p)})\right] \\
&= 2b\sqrt{m}\mathbb{E}_\beta\left[\sup_{\boldsymbol{\gamma},t}\sum_{i=1}^n\sum_{p=1}^m \beta_{ip}\left\langle\gamma_p t(\mathbf{x}_i^{(p)}), \gamma_p t(\mathbf{x}_i^{(p)})\right\rangle\right] \triangleq 2b\sqrt{m}\mathcal{G}_{1n}(\boldsymbol{\gamma},t).
\end{aligned} \tag{24}
$$

Note that when all the views are observed, using Hölder's inequality and Jensen's inequality, we have

$$\begin{aligned}\mathcal{G}_{1n}(\boldsymbol{\gamma},t) &= \mathbb{E}_\beta\left[\sup_{\boldsymbol{\gamma},t}\sum_{i=1}^{n}\sum_{p=1}^{m}\beta_{ip}\left\langle\gamma_p t(\mathbf{x}_i^{(p)}),\gamma_p t(\mathbf{x}_i^{(p)})\right\rangle\right]\\ &= \mathbb{E}_\beta\left[\sup_{\boldsymbol{\gamma}}\sum_{i=1}^{n}\sum_{p=1}^{m}\beta_{ip}\gamma_p^2\right]\\ &\leq \mathbb{E}_\beta\left[\sum_{p=1}^{m}\left|\sum\nolimits_{i=1}^{n}\beta_{ip}\right|\right]\leq m\sqrt{n}.\end{aligned}\tag{25}$$

Similarly, we have

$$\begin{aligned}&\mathbb{E}_\beta\left[\sqrt{8}\sup_{\boldsymbol{\gamma},t,\mathbf{C}}\sum\nolimits_{i=1}^{n}\sum\nolimits_{c=1}^{k}\beta_{ic}\phi_{\boldsymbol{\gamma},\mathbf{t}}^\top(\mathbf{x}_i)\mathbf{C}\mathbf{e}_c\right]\\ &=\sqrt{8}\mathbb{E}_\beta\left[\sup_{\boldsymbol{\gamma},t,\mathbf{C}}\sum\nolimits_{i=1}^{n}\sum\nolimits_{c=1}^{k}\beta_{ic}\left\langle[\gamma_1 t(\mathbf{x}_i^{(1)})\phi_1^\top(\mathbf{x}_i^{(1)}),\cdots,\gamma_m t(\mathbf{x}_i^{(m)})\phi_m^\top(\mathbf{x}_i^{(m)})]^\top,\mathbf{C}\mathbf{e}_c\right\rangle\right]\\ &\leq\sqrt{8}\mathbb{E}_\beta\left[\sup_{\boldsymbol{\gamma},t,\mathbf{C}}\sum\nolimits_{c=1}^{k}\left|\sum\nolimits_{i=1}^{n}\beta_{ic}\left\langle[\gamma_1 t(\mathbf{x}_i^{(1)})\phi_1^\top(\mathbf{x}_i^{(1)}),\cdots,\gamma_m t(\mathbf{x}_i^{(m)})\phi_m^\top(\mathbf{x}_i^{(m)})]^\top,\mathbf{C}\mathbf{e}_c\right\rangle\right|\right]\\ &=\sqrt{8}b\mathbb{E}_\beta\left[\sum\nolimits_{c=1}^{k}\left|\sum\nolimits_{i=1}^{n}\beta_{ic}\right|\right]\\ &\triangleq\sqrt{8}b\mathcal{G}_{2n}(\boldsymbol{\gamma},t),\end{aligned}\tag{26}$$

where the fourth line holds because β_{ic} are symmetric random variables. Also, when all the views are observed, we have $\mathcal{G}_{2n}(\boldsymbol{\gamma},t)\leq k\sqrt{n}$. At last, we have

$$\mathbb{E}_\beta\left[2\sup_{\mathbf{C}}\sum_{i=1}^{n}\sum_{c,l=1}^{k}\beta_{icl}\left\langle\mathbf{C}\mathbf{e}_c,\mathbf{C}\mathbf{e}_l\right\rangle\right]\leq\mathbb{E}_\beta\left[2b\sum_{c,l=1}^{k}\left|\sum\nolimits_{i=1}^{n}\beta_{icl}\right|\right]\leq 2bk^2\sqrt{n}.\tag{27}$$

Combining Lemmas 4 and 5, Eqs. (20) (21), and (23), we have

$$\begin{aligned}\mathfrak{R}_n(\mathcal{F}) &\leq\frac{1}{n}\sqrt{\pi/2}\mathbb{E}[\sup_{f\in\mathcal{F}}G_{\boldsymbol{\gamma},t,\mathbf{C}}]\leq\frac{1}{n}\sqrt{\pi/2}\mathbb{E}\left[\sup_{f\in F}H_{\boldsymbol{\gamma},t,\mathbf{C}}\right]\\ &\leq\frac{1}{n}\sqrt{\pi/2}\left(2b\sqrt{m}\mathcal{G}_{1n}(\boldsymbol{\gamma},t)+\sqrt{8}b\mathcal{G}_{2n}(\boldsymbol{\gamma},t)+2bk^2\sqrt{n}\right).\end{aligned}\tag{28}$$

Put the above inequality into Theorem 3, with probability at least $1-\delta$, the following holds for all $f\in\mathcal{F}$:

$$\mathbb{E}[f(\mathbf{x})]\leq\frac{1}{n}\sum_{i=1}^{n}f(\mathbf{x}_i)+\frac{4\sqrt{\pi}mb\mathcal{G}_{1n}(\boldsymbol{\gamma},t)}{n}+\frac{4\sqrt{\pi m}b\mathcal{G}_{2n}(\boldsymbol{\gamma},t)}{n}+\frac{\sqrt{8\pi}bk^2}{\sqrt{n}}+2b\sqrt{\frac{\log 1/\delta}{2n}}.\tag{29}$$

This completes the proof.

IV. The results on Caltech102-5, Caltech102-10 and Caltech102-15

References

[1] A. Maurer and M. Pontil, "k-dimensional coding schemes in Hilbert spaces," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5839–5846, 2010.

[2] T. Liu, D. Tao, and D. Xu, "Dimensionality-dependent generalization bounds for k-dimensional coding schemes," *Neural computation*, vol. 28, no. 10, pp. 2213–2249, 2016.

[3] X. Liu, X. Zhu, M. Li, L. Wang, E. Zhu, T. Liu, M. Kloft, D. Shen, J. Yin, and W. Gao, "Multiple kernel k-means with incomplete kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[4] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

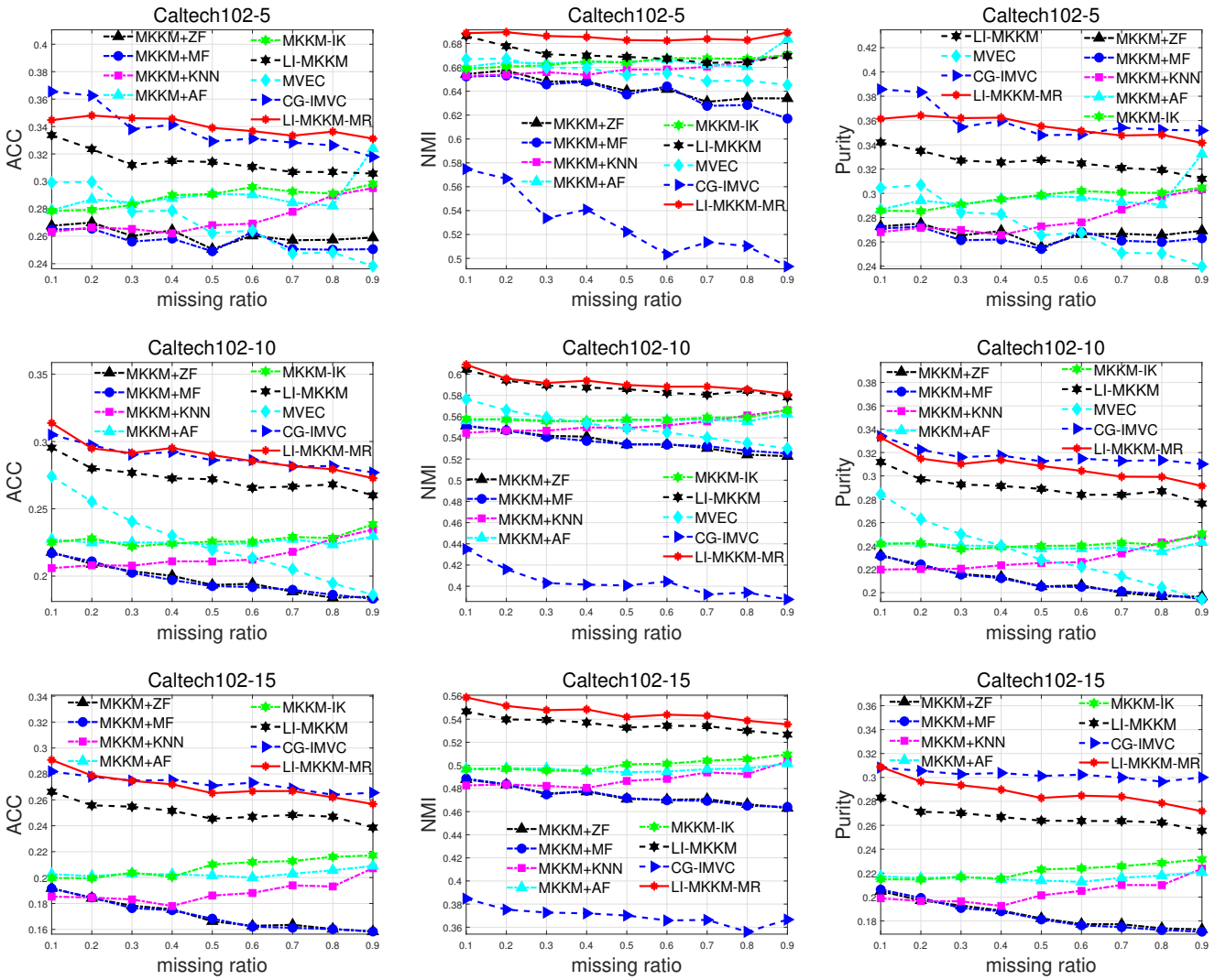

Figure 1: Clustering accuracy, NMI and purity comparison with the variation of missing ratios on Caltech102-5, Caltech102-10 and Caltech102-15. For each given missing ratio, the "incomplete patterns" are randomly generated for 10 times and their averaged results are reported.