Boosting Spectral Clustering on Incomplete Data via Kernel Correction and Affinity Learning

Fangchen Yu[1], Runze Zhao[1], Zhan Shi[1], Yiwen Lu[1]
Jicong Fan[1,2], Yicheng Zeng[2], Jianfeng Mao[1,2], Wenye Li[1,2] (✉)

[1] The Chinese University of Hong Kong, Shenzhen (CUHK-SZ)
[2] Shenzhen Research Institute of Big Data (SRIBD)

1 · INTRODUCTION

Spectral Clustering on Complete Data:

Step 1. Construct A Similarity Matrix: $S = K$ or C

- Calculate a (Gaussian) kernel K from X
- Learn a self-expressive affinity matrix C from K

Step 2. Perform Normalized Cut Algorithm [1]

- Calculate the Laplacian matrix L
- Select the first k eigenvectors of L as W
- Conduct k-means clustering on W

Q: Spectral Clustering on Inomplete Data?

2 · AIM

When dealing with incomplete data, how to

1) Estimate a high-quality kernel;
2) Learn a high-quality self-expressive affinity,
benfiting the spectral clustering task.

3 · METHOD

I. Kernel Correction (KC) Algorithm

- A valid kernel should be positive semi-definite [2].
- **Step 1.** Estimate a naive kernel K^0 on incomplete data;
- **Step 2.** Find a nesrest PSD kernel as the new estimate.

$$\hat{K} = \underset{K \in R^{n \times n}}{\operatorname{argmin}} \|K - K^0\|_F^2 \; s.t. \; K \succcurlyeq 0, \; k_{ij} = k_{ji} \in [0,1]$$

- $\|K^* - \hat{K}\|_F \le \|K^* - K^0\|_F$ (K^* is the ground-truth) [3]

II. Affinity Learning Algorithms

- Kernel Self-expressive Affinity Learning with ℓ_p norm:
- **Proximal P-norm:** $\|C\|_p^p := \sum_{i,j=1}^n |c_{ij}|^p$

$$\min_{C \in R^{n \times n}} \|\phi(X) - \phi(X)C\|_F^2 + \lambda \|C\|_p^p \; s.t. \; C_{ij} \in [0,1]$$

- **Schatten P-norm:** $\|C\|_{S_p} := \left(\sum_{i=1}^n \sigma_i^p(C)\right)^{1/p}$

$$\min_{C \in R^{n \times n}} \|\phi(X) - \phi(X)C\|_F^2 + \lambda \|C\|_{S_p} \; s.t. \; C_{ij} \in [0,1]$$

- Adaptive Kernel Least-Squares Representation:

$$\min_{K \succcurlyeq 0, \, C} \|K - K^0\|_F^2 + \mathrm{Tr}(K - 2KC + C^\top KC) + \lambda \|C\|_F^2$$

4 · RESULTS

I. Performance on Gaussian Kernel Estimation

Relative error: **RE** $= \|\hat{K} - K^*\|_F / \|K^*\|_F$; **Recall**: measures the accuracy of top-10 neighbors.

Table 1. Gaussian kernel estimation on the Yale64 dataset with 80% random missing.

Metric	Naive	ZERO	MEAN	kNN	EM	SVT	GR	KFMC	DC	TRF	EE	KC
RE↓	0.113	0.382	0.195	0.381	0.195	0.380	0.376	0.335	0.180	0.112	0.097	**0.089**
Recall↑	0.721	0.063	0.275	0.063	0.275	0.066	0.070	0.183	0.571	0.722	0.751	**0.767**

Three ways to estimate $\hat{K}$:

- **Data Imputation:** $X^0 \to \hat{X} \to \hat{K}$
- **Distance Calibration:** $X^0 \to D^0 \to \hat{D} \to \hat{K}$
- **Kernel Correction (KC):** $X^0 \to K^0 \to \hat{K}$

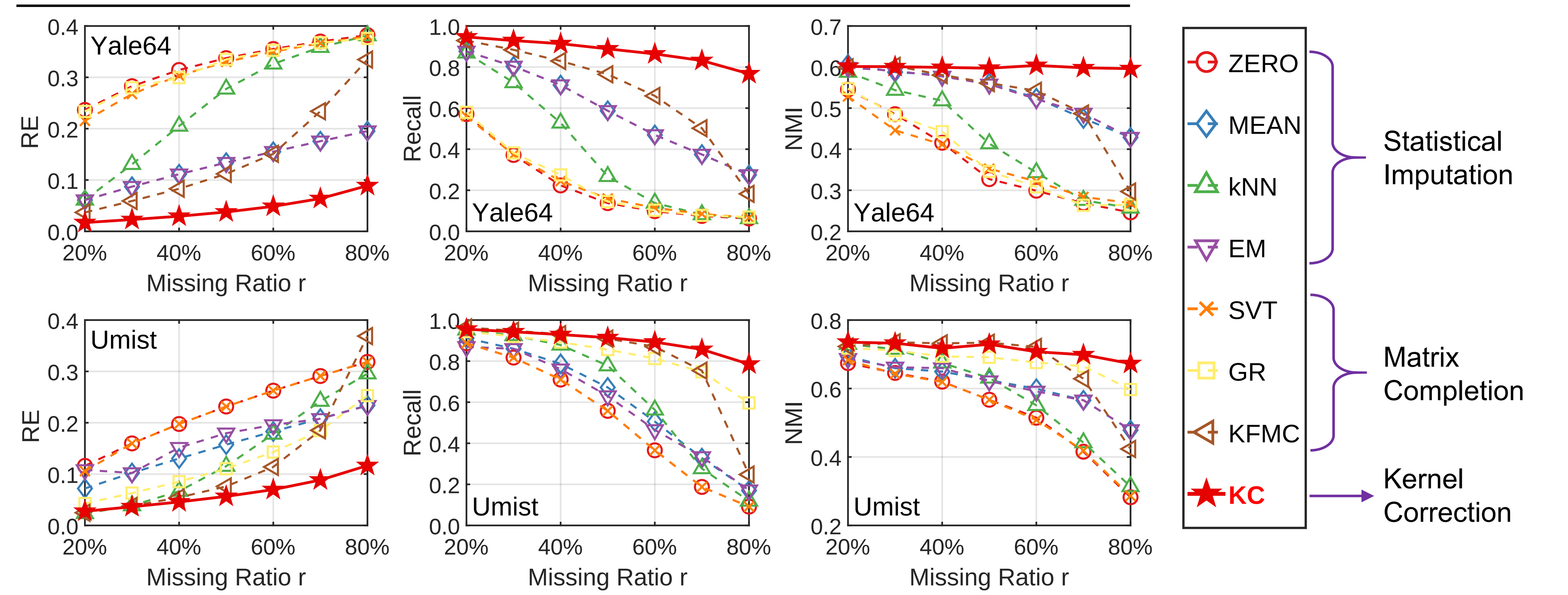
Smallest RE & Highest Recall

II. Performance on Standard Spectral Clustering

Table 2. NMI of standard spectral clustering on Yale64 and Umist with 80% missing.

Dataset	Naive	ZERO	MEAN	kNN	EM	SVT	GR	KFMC	DC	TRF	EE	KC
Yale64	0.588	0.246	0.429	0.257	0.428	0.269	0.264	0.297	0.551	0.587	0.593	**0.596**
Umist	0.669	0.282	0.478	0.314	0.479	0.286	0.597	0.423	0.488	0.667	0.669	**0.673**

- Missing completely at random
- an estimated Gaussian kernel
  → kNN graph
  → normalized cut algorithm
- Normalized Mutual Information



III. Performance on Self-expressive Affinity Learning

KSL-Sp and AKLSR employ corrected kernels to yield dependable affinity matrices.

Table 3. NMI of spectral clustering using learned affinity on Yale64 with 80% missing.

Method	Naive	ZERO	MEAN	kNN	EM	SVT	GR	KFMC	DC	TRF	EE	KC
KSSC	0.219	0.215	0.167	0.173	0.177	0.218	0.208	0.259	0.588	0.210	0.209	**0.616**
KLSR	0.606	0.311	0.604	0.320	0.609	0.321	0.327	0.318	0.597	0.603	0.604	**0.616**
KSL-Sp	0.370	0.315	0.581	0.303	0.579	0.305	0.304	0.295	0.555	0.364	0.599	**0.619**
AKLSR	0.452	0.327	0.606	0.338	0.605	0.308	0.338	0.312	0.570	0.464	0.575	**0.614**

- **KSSC [4]:** $\min_{C \in R^{n \times n}} \|\phi(X) - \phi(X)C\|_F^2 + \lambda \|C\|_1$
- **KLSR [5]:** $\min_{C \in R^{n \times n}} \|\phi(X) - \phi(X)C\|_F^2 + \lambda \|C\|_F^2$
- **KSL-Sp (Ours):** $\min_{C \in R^{n \times n}} \|\phi(X) - \phi(X)C\|_F^2 + \lambda \|C\|_{S_p}$
- **AKLSR (Ours):** $\min_{K \succcurlyeq 0, \, C} \|K - K^0\|_F^2 + \mathrm{Tr}(K - 2KC + C^\top KC) + \lambda \|C\|_F^2$

5 · CONCLUSIONS

Q: How to boost spectral clustering on incomplete data via effective kernel and affinity learning?

A: We propose an imputation-free framework:

- Start with a naive kernel instead of imputing;
- Learn a high-quality ***kernel*** *matrix* by the kernel correction method with a theoretical guarantee;
- Learn a high-quality ***affinity*** *matrix* by the kernel self-expressive affinity learning algorithms with an adaptive extension using joint optimization.

Experimental results show the superiority:

- **Experiments I & II:** show our kernel correction method significantly outperforms data imputation and distance calibration approaches, especially for a large missing ratio (i.e., 80%).
- **Experiment III:** shows the effectiveness of KSL-Sp and AKLSR with comparable performance.

6 · REFERENCES

[1] Andrew Ng, Michael Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm". Advances in Neural Information Processing Systems (2001).

[2] Rafic Nader et al. "On the positive semi-definite property of similarity matrices". Theoretical Computer Science 755 (2019), pp. 13–28.

[3] Fangchen Yu et al. "Online estimation of similarity matrices with incomplete data". Uncertainty in Artificial Intelligence. PMLR. (2023), pp. 2454–2464.

[4] Vishal M Patel and Ren´e Vidal. "Kernel sparse subspace clustering". 2014 IEEE International Conference on Image Processing. (2014), pp. 2849–2853.

[5] Jicong Fan, et al. "A simple approach to automated spectral clustering". Advances in Neural Information Processing Systems 35 (2022), pp. 9907–9921.

7 · MORE INFORMATION

Paper | Github | Homepage