Discriminating Tensor Spectral Clustering for High-Dimension-Low-Sample-Size Data

Yu Hu, Fei Qi, Yiu-Ming Cheung, *Fellow, IEEE*, and Hongmin Cai, *Senior Member, IEEE*

***Abstract*— Tensor spectral clustering (TSC) is a recently proposed approach to robustly group data into underlying clusters. Unlike the traditional spectral clustering (SC), which merely uses pairwise similarities of data in an affinity matrix, TSC aims at exploring their multiwise similarities in an affinity tensor to achieve better performance. However, the performance of TSC highly relies on the design of multiwise similarities, and it remains unclear especially for high-dimension-low-sample-size (HDLSS) data. To this end, this article has proposed a discriminating TSC (DTSC) for HDLSS data. Specifically, DTSC uses the proposed discriminating affinity tensor that encodes the pair-to-pair similarities, which are particularly constructed by the anchor-based distance. HDLSS asymptotic analysis shows that the proposed affinity tensor can explicitly differentiate samples from different clusters when the feature dimension is large. This theoretical property allows DTSC to improve the clustering performance on HDLSS data. Experimental results on synthetic and benchmark datasets demonstrate the effectiveness and robustness of the proposed method in comparison to several baseline methods.**

***Index Terms*— High-dimension-low-sample-size (HDLSS) data, similarity measurement, spectral clustering (SC), tensor, tensor SC (TSC).**

Nomenclature

$\mathcal{T}$	Affinity tensor.
$\mathcal{T}^\star$	Discriminating affinity tensor.
$\mathcal{D}/\mathcal{D}^\star$	Degree tensor of $\mathcal{T}/\mathcal{T}^\star$.
$\mathcal{L}/\mathcal{L}^\star$	Laplacian tensor of $\mathcal{T}/\mathcal{T}^\star$.
$\mathcal{A}_{:,r,:,s}$	rth frontal slice of the sth subtensor of $\mathcal{A}$.
$\mathcal{A}_{i,j,k,l}$	(i, j, k, l)th element of $\mathcal{A}$.
$\boldsymbol{A}_{p,q}$	(p, q)th element of the matrix $\boldsymbol{A}$.
$\boldsymbol{V}^{\star(i)}$	ith eigenmatrix of Laplacian tensor $\mathcal{L}^\star$.
$\boldsymbol{x}_i^{(t)}$	tth feature of the sample $\boldsymbol{x}_i$.
d_{ij}	Pairwise Euclidean distance of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.
$\phi(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k)$	Anchor-based distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.
$[m]$	Set of $\{1, 2, \dots, m\}$.
b	Number of eigenmatrices.
c	Number of clusters.
$\mathcal{F}_p/\boldsymbol{\mu}_p/\tau_p$	Population of pth cluster/location/overall scale.

I. Introduction

Clustering aims at grouping samples into their respective clusters in an unsupervised manner and has a variety of applications in machine learning and data mining [1], [2], [3], [4]. Over the past decades, spectral clustering (SC) [5] has been recognized as a representative technique in the related literature due to its empirical performance, simplicity, and theoretical foundations. Nevertheless, SC relies on an affinity matrix encoding pairwise similarities, which is noise-sensitive for clustering [6], [7], [8], [9]. Another issue is that the pairwise similarities measured by the Euclidean distance suffer from the concentration effect [10] when dealing with high-dimension-low-sample-size (HDLSS) data [11]. The concentration effect is that the pairwise similarities of two measured samples become indiscriminative when feature dimensions are extremely large [12]. Consequently, most of the clustering methods that rely on such pairwise similarities fail to achieve satisfactory performance. In the literature, a few recent works have attempted to address this issue. For example, Sarkar and Ghosh [11] have proposed an approach to tackle the concentration effect with a data-driven measure of dissimilarity by the anchor-based distance, named as the mean of absolute differences of pairwise distances (MADD). Using MADD, they adapt the classic clustering methods such as SC to HDLSS data. Nevertheless, MADD-based methods still rely on pairwise similarities, which are prone to noise contamination.

Recently, tensor SC (TSC) has been proposed to address the noise issue and the concentration effect simultaneously. TSC leverages an affinity tensor instead of an affinity matrix to characterize multiwise similarities [13], [14], [15], which are shown to be noise-robust and alleviate the concentration effect [16]. For instance, Ghoshdastidar and Dukkipati [15]

Manuscript received 14 March 2022; revised 19 June 2023 and 14 March 2024; accepted 20 June 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFE0112200; in part by the National Natural Science Foundation of China under Grant U21A20520 and Grant 62325204; in part by the Key-Area Research and Development Program of Guangzhou City under Grant 202206030009; in part by the NSFC/Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21; in part by the General Research Fund of RGC under Grant 12201321, Grant 12202622, and Grant 12201323; and in part by the RGC Senior Research Fellow Scheme under Grant SRFS2324-2S02. *(Corresponding author: Hongmin Cai.)*

Yu Hu, Fei Qi, and Hongmin Cai are with the Department of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: hmcai@scut.edu.cn).

Yiu-Ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong.

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2024.3422243, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2024.3422243

used the affinity tensor for characterizing multiwise similarities and applied a multilinear singular value decomposition (SVD) on the affinity tensor to obtain the spectral embedding for clustering. They demonstrate the superior performance of TSC over SC in several HDLSS datasets. Later, Ghoshdastidar and Dukkipati [15], [17] have developed a trace optimization on the affinity tensor and formulated a tensor sampling strategy [18] to save the computational cost. More recently, Peng et al. [16] proposed to construct the ratio-based pair-to-pair similarity encoded in a fourth-order affinity tensor and formulated a tensor decomposition on the affinity tensor to extract the high-order affinity matrix. The obtained high-order affinity matrix has achieved a promising clustering performance on a few HDLSS datasets. Nevertheless, the performance of TSC greatly depends on the construction of the multiwise similarities in the affinity tensor and has yet to be well-explored especially for HDLSS data. In Section III-A, we demonstrate that the concentration effect happens in the case of multiwise similarities proposed in [16], where, due to the use of Euclidean distance for constructing the multiwise similarities, the numerical values converge a constant when the feature dimension increases.

To this end, this article will propose a discriminating TSC (DTSC) method for HDLSS data. The DTSC consists of two major steps: 1) constructing a discriminating affinity tensor for HDLSS data and 2) performing a spectral analysis on the proposed affinity tensor to obtain the cluster labels. Specifically, as illustrated in Fig. 1, the first step constructs a fourth-order discriminating affinity tensor using the anchor-based distance to quantify the pair-to-pair similarities. Such a discriminating affinity tensor explicitly addresses the concentration effect via the asymptotic analysis. In other words, the discriminating affinity tensor can differentiate samples from distinct clusters especially when the feature dimension is large. In the second step, the high-order affinity matrix is extracted from the discriminating affinity tensor based on the tensor decomposition. Then, the cluster labels can be derived via the standard SC on the high-order affinity matrix. Extensive experiments on synthetic and benchmark datasets have been conducted to verify the effectiveness of our method and its robustness against noise compared with several recent baseline methods.

The main contributions of this article are as follows.

1) We demonstrate that the tradition affinity tensor suffers from the concentration effect, which adversely affects the clustering performance on HDLSS data. Then, we theoretically demonstrate that the reason for this is the Euclidean distance used for constructing the affinity tensor (see Section III-A).
2) The DTSC method is proposed to address the concentration effect brought by HDLSS data. The method constructs the discriminating affinity tensor with the anchor-based distance, and the asymptotic analysis of the discriminating affinity tensor provides a theoretical guarantee that allows us to improve the clustering performance of HDLSS data (see Section III-B).
3) Extensive experiments on synthetic and benchmark datasets, including DBWorld, COIL20, Lymphoma, and UCI gene, are conducted to demonstrate the competitive clustering performance of our method and its robustness against noise (see Section IV).

II. Related Works

A. High-Dimension-Low-Sample-Size Data Clustering

Clustering HDLSS data has recently attracted increasing research interests. HDLSS data refer to the one with the dimension of feature n being far greater than the size of samples m and are frequently encountered in computer vision, text mining, and bioinformatics [19], [20]. The major difficulty in clustering HDLSS data is that the Euclidean distance frequently used in plenty of methods suffers from the concentration effect [10] when the feature dimension is large and cannot differentiate samples from distinct clusters [11]. To address this issue, Ahn et al. [21] developed the maximum data piling (MDP) clustering that is inspired by the HDLSS asymptotics. Recently, Sarkar and Ghosh [11] proposed an approach to tackle HDLSS data clustering with a data-driven measure of dissimilarity by the anchor-based distance, named as MADD. The idea of MADD originates from the statistical asymptotic study in [12]. It showed that under some mild assumptions, the scaled Euclidean distance of two measured samples tends to be a constant as dimension n goes to infinity, which is referred to as geometric representation. MADD is proved to be discriminating for HDLSS data clustering under the statistical framework in [12]. Our analysis in Section III-A is partially inspired by the work from [11] and [12].

B. Tensor Spectral Clustering

TSC has recently been proposed to address HDLSS data, and it uses the affinity tensor to encode multiwise similarities for a better performance than using pairwise similarities [13], [14], [22], [23], [24], [25], [26], [27]. A series of works applied a combination of Euclidean distances among samples to construct the multiwise similarities encoded in the affinity tensor, and then used high-order SVD or tensor trace norm maximization to derive the tensor spectral embedding matrix for clustering. For instance, Ghoshdastidar and Dukkipati [15] proposed a multilinear SVD method to decompose the affinity tensor and showed that this decomposition amounted to clustering samples by maximizing the squared associativity of the partition. Ghoshdastidar and Dukkipati [17] applied a trace optimization on the affinity tensor and developed a tensor sampling strategy [18] to save the computational cost.

More recently, Peng et al. [16] proposed integrating tensor similarity and pairwise similarity (IPS2) that constructs the ratio-based pair-to-pair similarities encoded in a fourth-order affinity tensor and used a tensor decomposition on the affinity tensor to extract the high-order affinity matrix. IPS2 has shown promising performance on several HDLSS datasets and is used as the baseline of our method to conduct the theoretical analysis and comparison. In what follows, IPS2 will be briefly introduced and the notations used in this manuscript are summarized in Nomenclature. Specifically, it measures the ratio between the intra- and interpair Euclidean distances of samples as the fourwise similarities and has demonstrated promising empirical performance on some benchmark HDLSS datasets. Formally, a fourth-order **affinity tensor**

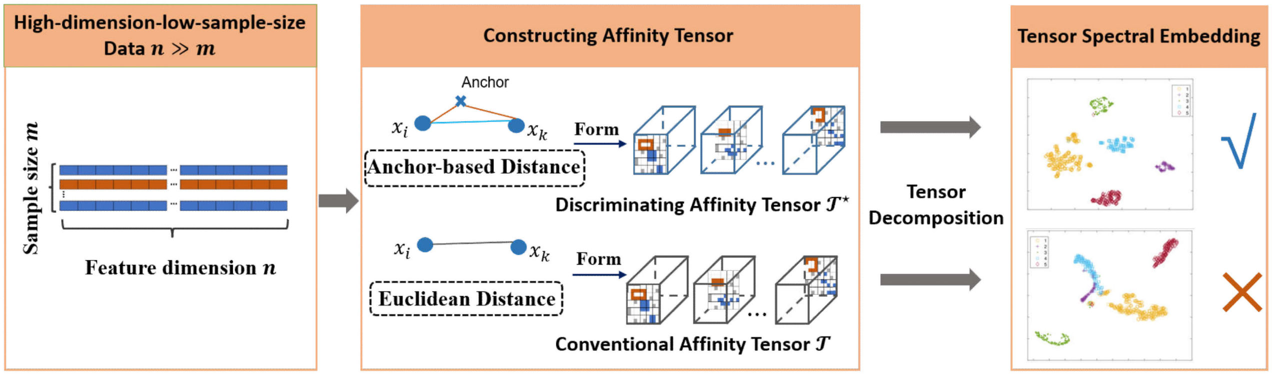

High-dimension-low-sample-size Data $n \gg m$
Sample size m
Feature dimension n
Constructing Affinity Tensor
Anchor
x_i
x_k
Form
Anchor-based Distance
Discriminating Affinity Tensor $\mathcal{T}^\star$
x_i
x_k
Form
Euclidean Distance
Conventional Affinity Tensor $\mathcal{T}$
Tensor Decomposition
Tensor Spectral Embedding
√
×

Fig. 1. Workflow of our DTSC and the difference from conventional methods. Similar to conventional methods (bottom), DTSC (top) has two major steps: constructing the affinity tensor that encodes the multiwise similarities and performing spectral analysis on the affinity tensor to obtain cluster labels. The difference is that conventional methods (bottom) apply the Euclidean distance, whereas DTSC adopts the anchor-based distance to construct the multiwise similarities. The former leads to the concentration effect when the feature dimension is exceptionally large, and the latter helps explicitly address the concentration effect and improves the clustering performance.

$\mathcal{T} \in \mathbb{R}^{m\times m\times m\times m}$ for m samples is defined and visually shown in Fig. 2, with its each entry being the following:

Definition 1 (Affinity Tensor):

$$\mathcal{T}_{i,j,k,l} = \exp\left(-\sigma \frac{d_{ij}+d_{kl}+d_{il}+d_{jk}}{d_{ik}+d_{jl}+\varepsilon}\right) \quad (1)$$

for $i,j,k,l \in [m]$, where d_{ij} denotes the pairwise Euclidean distance between samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

The numerator, $d_{ij}+d_{kl}+d_{il}+d_{jk}$, and the denominator, $d_{ik}+d_{jl}$, are the inter- and intrapair distances, respectively. In general, each element of the fourth-order affinity tensor, $\mathcal{T}_{i,j,k,l}$, characterizes the pair-to-pair similarities by the ratio between the interpair and intrapair distances. Intuitively, if any of the interpair distances, e.g., d_{ij}, d_{kl}, d_{il}, and d_{jk}, are large, and any of intrapair distances, d_{ik} or d_{jl}, are small, those two pairs come with small similarities. The parameter ε is a given small parameter to overcome the instability caused by a zero denominator and set to be 0.001. In contrast, σ is an empirical parameter and set to be 1 for simplicity in this article.

One can note that $\mathcal{T}$ naturally depicts a more comprehensive spatial structure among four samples than pairwise similarities used in SC [5] and enables a robust similarity estimation, alleviating noise contamination [16]. As articulated in [16], the noise robustness may stem from the fact that the computation of multiwise similarities serves as a natural filter against noise corruption, as it alleviates the effects of random variations across a larger sample set. Furthermore, it has been shown in [16] that a high-order affinity matrix can be learned from the Laplacian tensor of $\mathcal{T}$, defined as follows.

Definition 2 (Laplacian Tensor): Let $\mathcal{T}$ be a fourth-order m-dimension affinity tensor and $\mathcal{D}$ is its degree tensor. The tensor $\mathcal{L}$ is called the Laplacian tensor of $\mathcal{T}$ if

$$\mathcal{L}_{:,p,:,q} = \mathcal{I}_{:,p,:,q} - \mathcal{D}^{-\frac{1}{2}}_{:,p,:,q}\mathcal{T}_{:,p,:,q}\mathcal{D}^{-\frac{1}{2}}_{:,p,:,q} \quad \forall p,q \in [m]. \quad (2)$$

Here, the definitions of the degree tensor $\mathcal{D}$ and identity tensor $\mathcal{I}$ are moved to Section A in the Supplementary Material for ease of explanation. Accordingly, the high-order affinity matrix is given as follows.

Theorem 1 (High-Order Affinity Matrix): Let $\mathcal{L}$ be the Laplacian tensor, there exists a nonzero square matrix $\boldsymbol{V} \in \mathbb{R}^{m\times m}$, which is termed as high-order affinity matrix, and a scalar $\lambda \in \mathbb{R}$, satisfying

$$\mathcal{L}\cdot \boldsymbol{V} = \lambda \boldsymbol{V} \quad (3)$$

where

$$(\mathcal{L}\cdot\boldsymbol{V})_{i,j} = \sum_{k=1}^{m}\sum_{l=1}^{m}\mathcal{L}_{i,j,k,l}\boldsymbol{V}_{k,l} \quad \forall i,j,k,l \in [m]. \quad (4)$$

The numerical strategy for solving (3) is provided in Section E in the Supplementary Material. In practice, solving the eigenmatrix problem in (3) yields multiple eigenmatrices [16]. One can average the multiple eigenmatrices to form the high-order affinity matrix.

III. Method

This section details the proposed DTSC, which is specifically tailored for HDLSS data. DTSC follows a similar pipeline with IPS2 [16], but differs in the definition of the pair-to-pair similarities in the affinity tensor. Section III-A demonstrates the HDLSS asymptotic behavior of the affinity defined in IPS2 and illustrates that it suffers from the concentration effect and is not suitable for clustering HDLSS data. Section III-B shows our approach can address the concentration effect by applying the anchor-based distance to construct the discriminating affinity tensor. Section III-C demonstrates the sampling strategy to achieve a balance between computational cost and clustering performance.

A. Asymptotic Behavior of Affinity Tensor

Suppose we have m samples $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m] \in \mathbb{R}^{n\times m}$ from c clusters, corresponding to c populations $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_c$. The overarching goal of the affinity tensor for clustering is characterizing a discriminative pair-to-pair similarities. In other words, the affinity tensor aims to compare four samples concurrently and distinguish whether they are from the same population or not. For the sake of convenience, we consider a simplified case and divide four samples into two pairs. Given any two pairs of samples $(\boldsymbol{x}_i, \boldsymbol{x}_k)$ and $(\boldsymbol{x}_j, \boldsymbol{x}_l)$, $\forall i,j,k,l \in [m]$, the affinity tensor $\mathcal{T}_{i,j,k,l}$ should be dependent on whether those two pairs' samples come from the same population. For example, $\mathcal{T}_{i,j,k,l}$ should be much

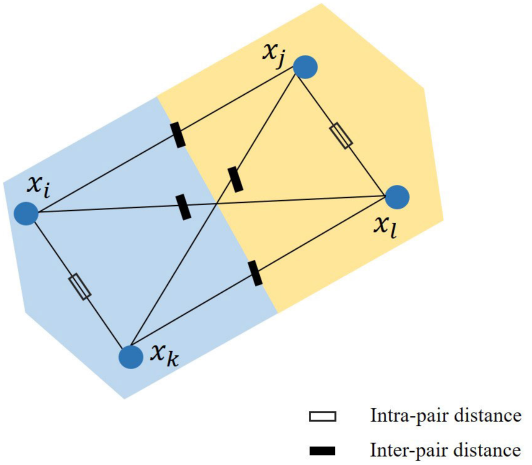


Fig. 2. Intuitive figure to explain the affinity tensor. In general, an element of the fourth-order affinity tensor characterizes the pair-to-pair similarities, which can be quantified by the ratio between the inter- and intrapair distances.

smaller when $(\boldsymbol{x}_i, \boldsymbol{x}_k)$ and $(\boldsymbol{x}_j, \boldsymbol{x}_l)$ come from two distinct populations, compared with that when they come from the same one.

However, such requirement is radically violated when the affinity tensor in (1) is applied for HDLSS data, due to the adoption of the pairwise Euclidean distance. The critical reason is that the applied Euclidean distance suffers from the **concentration effect** [10]. To demonstrate this, we apply the clustering method on a synthetic dataset satisfying Assumption 1. Each sample of the dataset is independently drawing from two Gaussian distributions $\mathcal{N}_1(\mathbf{0}^n, \sigma_1^2 \boldsymbol{\Sigma}^n)$ and $\mathcal{N}_2(\boldsymbol{\mu}^n, \sigma_2^2 \boldsymbol{\Sigma}^n)$. Herein, $\mathbf{0}^n = (0, \ldots, 0)^\top$, $\boldsymbol{\mu}^n = (1, -1, \ldots, (-1)^{n+1})^\top$, and $\boldsymbol{\Sigma}^n$ is a block-diagonal matrix, in which $\boldsymbol{\Sigma}^n_{i,i} = 1$, $\forall n$, $\boldsymbol{\Sigma}^n_{2i-1,2i} = \boldsymbol{\Sigma}^n_{2i,2i-1} = 0.98$ for $i = 1, 2, \ldots, \lfloor n/2 \rfloor$, and $\boldsymbol{\Sigma}^n_{i,j} = 0$ otherwise. For ease of analysis, we drop the parameters ϵ and σ for simplicity. Consequently, the concentration effect can be demonstrated as follows.

Lemma 1 (Concentration Effect in Gaussian Distribution Case [11]): Given $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k \in \mathbb{R}^n$ with $\boldsymbol{x}_i, \boldsymbol{x}_k \sim \mathcal{N}_2$ and $\boldsymbol{x}_j \sim \mathcal{N}_1$, as $n \to \infty$, the scaled Euclidean distance $n^{-1/2}||\boldsymbol{x}_i - \boldsymbol{x}_j||_2 \xrightarrow{P} (\sigma_1^2 + \sigma_2^2 + 1)^{1/2}$, whereas the scaled Euclidean distance $n^{-1/2}||\boldsymbol{x}_i - \boldsymbol{x}_k||_2 \xrightarrow{P} (2\sigma_2^2)^{1/2}$.

Accordingly, the affinity tensor $\boldsymbol{\mathcal{T}}$ in (1) follows.

Proposition 1: Suppose all the samples in $\boldsymbol{X}$ are independently drawn from populations following either $\mathcal{N}_1$ or $\mathcal{N}_2$. As $n \to \infty$, the affinity tensor $\boldsymbol{\mathcal{T}}$ in (1) tends to a constant tensor where

$$\forall i, j, k, l \in [m], \quad \boldsymbol{\mathcal{T}}_{i,j,k,l} \xrightarrow{P} \begin{cases} \exp\left(-\dfrac{2\sqrt{\sigma_1^2+\sigma_2^2+1}}{\sqrt{2\sigma_2^2}}\right), & \text{if } (\boldsymbol{x}_i, \boldsymbol{x}_k) \sim \mathcal{N}_2 \text{ and } \\ & (\boldsymbol{x}_j, \boldsymbol{x}_l) \sim \mathcal{N}_1 \\ \exp(-2) \approx 0.135, & \text{if } (\boldsymbol{x}_i, \boldsymbol{x}_k) \sim \mathcal{N}_2 \text{ and } \\ & (\boldsymbol{x}_j, \boldsymbol{x}_l) \sim \mathcal{N}_2. \end{cases} \tag{5}$$

Proof: The proof is provided in Section C in the Supplementary Material. □

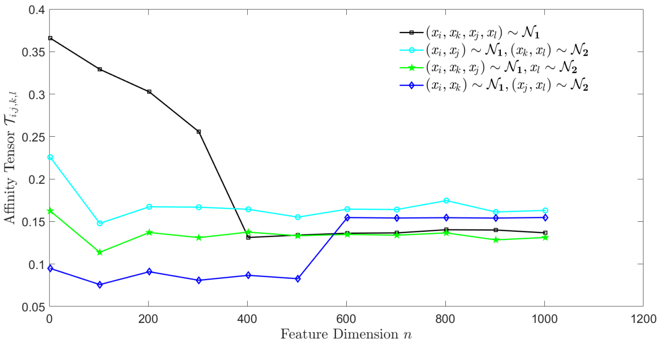


Fig. 3. Affinity tensor value $\boldsymbol{\mathcal{T}}_{i,j,k,l}$ on SD-1 data with dimension $n = 2$ to $n = 900$. The black, cyan, green, and blue lines denote the discriminating affinity tensor value on $(\boldsymbol{x}_i, \boldsymbol{x}_k) \sim \mathcal{N}_1$ and $(\boldsymbol{x}_j, \boldsymbol{x}_l) \sim \mathcal{N}_1$, $(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{N}_1$ and $(\boldsymbol{x}_k, \boldsymbol{x}_l) \sim \mathcal{N}_2$, $(\boldsymbol{x}_i, \boldsymbol{x}_k, \boldsymbol{x}_j) \sim \mathcal{N}_1$ and $\boldsymbol{x}_l \sim \mathcal{N}_2$, and $(\boldsymbol{x}_i, \boldsymbol{x}_k) \sim \mathcal{N}_1$ and $(\boldsymbol{x}_j, \boldsymbol{x}_l) \sim \mathcal{N}_2$, respectively. The figure shows that as n increases, the values of $\boldsymbol{\mathcal{T}}_{i,j,k,l}$ start to converge.

For example, let $\sigma_1^2 = 0.5$ and $\sigma_2^2 = 2$, and when two pairs samples are drawn from the same cluster, i.e., $(\boldsymbol{x}_i, \boldsymbol{x}_k) \sim \mathcal{N}_2$ and $(\boldsymbol{x}_j, \boldsymbol{x}_l) \sim \mathcal{N}_2$, one has $\boldsymbol{\mathcal{T}}_{i,j,k,l} \approx 0.135$. In comparison, when two pairs are drawn from distinct clusters, i.e., $(\boldsymbol{x}_i, \boldsymbol{x}_k) \sim \mathcal{N}_2$ and $(\boldsymbol{x}_j, \boldsymbol{x}_l) \sim \mathcal{N}_1$, one has $\boldsymbol{\mathcal{T}}_{i,j,k,l} \approx 0.154$. In this regard, $\boldsymbol{\mathcal{T}}_{i,j,k,l}$ obtains a smaller value when $(\boldsymbol{x}_i, \boldsymbol{x}_k)$ and $(\boldsymbol{x}_j, \boldsymbol{x}_l)$ are drawn from distinct clusters, compared with that when they are drawn from the same cluster. We visually demonstrate such concentration effect in Fig. 3. This violates the overarching goal of the fourth-order affinity tensor, which ought to be discriminating on given sample pairs.

In what follows, we generalize the above special case and apply the following assumptions and lemma.

Assumption 1: The collected samples are assumed to satisfy the following statements.

1) $\forall t \in [n], i \in [m]$, the fourth moments on each feature $\boldsymbol{x}_i^{(t)}$ are uniformly bounded.
2) For arbitrary two independent samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $\mathrm{Var}(\sum_{t=1}^n (\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_j^{(t)})^2) = \mathbf{o}(n^2)$, where $\mathbf{o}(n^2)$ represents an infinitesimal of n^2.
3) $\forall p, q \in [c]$, as $n \to \infty$, $\exists a_{pq}, \tau_p, \tau_q < \infty$, such that $n^{-1}||\boldsymbol{\mu}_p - \boldsymbol{\mu}_q||_2^2 \xrightarrow{P} a_{pq}$, $n^{-1}\mathrm{tr}(\boldsymbol{\Sigma}_p) \xrightarrow{P} \tau_p^2$, and $n^{-1}\mathrm{tr}(\boldsymbol{\Sigma}_q) \xrightarrow{P} \tau_q^2$, where $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$ are called the mean vector and dispersion matrix of the population $\mathcal{F}_p$, respectively, and the notation $\xrightarrow{P}$ means converge to in probability.

Generally, $\boldsymbol{\mu}_p$ and τ_p are termed as the **location** and **overall scale** of the population $\mathcal{F}_p$, respectively. We note that these three assumptions are commonly used in HDLSS literature [11], [12]. In practice, those assumptions are widely tested in the area of genome-wide association studies (GWASs) [28]. Based on the assumptions, the concentration effect of the Euclidean distance is formalized as follows.

Lemma 2 (Concentration Effect of Euclidean Distance [12]): Suppose all the samples in $\boldsymbol{X}$ satisfy Assumption 1. If $\boldsymbol{x}_i \sim \mathcal{F}_p$ and $\boldsymbol{x}_j \sim \mathcal{F}_q$ are two independent samples in $\boldsymbol{X}$, as $n \to \infty$, the following holds:

$$n^{-1/2} \sqrt{\sum_{t=1}^{n} \left(\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_j^{(t)}\right)^2} \xrightarrow{P} \left(\tau_p^2 + \tau_q^2 + a_{pq}\right)^{1/2}. \tag{6}$$

Building upon Lemma 2, the following proposition is clear.

Proposition 2: (Concentration Effect of Affinity Tensor): Suppose all the samples in $\boldsymbol{X}$ satisfy Assumption 1. For a general case, i.e., $i \neq j \neq k \neq l$, if $\boldsymbol{x}_i \sim \mathcal{F}_p$, $\boldsymbol{x}_j \sim \mathcal{F}_q$, $\boldsymbol{x}_k \sim \mathcal{F}_r$, and $\boldsymbol{x}_l \sim \mathcal{F}_s$ are four independent samples in $\boldsymbol{X}$, the affinity tensor $\mathcal{T}$ in (1) tends to a constant tensor as $n \to \infty$ where

$$\mathcal{T}_{i,j,k,l} \xrightarrow{P} \exp\left(-\frac{\begin{array}{c}\mathrm{const}(p,q)+\mathrm{const}(r,s)+\mathrm{const}(p,s)\\ +\,\mathrm{const}(q,r)\end{array}}{\mathrm{const}(p,r)+\mathrm{const}(q,s)}\right) \tag{7}$$

where $\mathrm{const}(\alpha,\beta) = (\tau_\alpha^2 + \tau_\beta^2 + a_{\alpha\beta})^{1/2}$, and $\alpha,\ \beta \in \{p,q,r,s\}$.

Proof: The proof is provided in Section B in the Supplementary Material. □

B. Discriminating Affinity Tensor via Anchor-Based Distance

We note that the problem mainly stems from the adoption of the pairwise Euclidean distance when constructing the affinity tensor. To tackle this issue, we propose the **anchor-based distance** to form a new affinity tensor. Formally, given any two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with an anchor sample $\boldsymbol{x}_k$, the anchor-based distance between $\boldsymbol{x}_i, \boldsymbol{x}_j$ is defined by

$$\begin{aligned}\phi(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k) &= \left|d_{ik}-d_{jk}\right| \\ &= \left|\sqrt{\sum_{t=1}^{n}\left(x_i^{(t)}-x_k^{(t)}\right)^2}-\sqrt{\sum_{t=1}^{n}\left(x_j^{(t)}-x_k^{(t)}\right)^2}\right|.\end{aligned} \tag{8}$$

The main property is that for most of the anchors $\boldsymbol{x}_k$, the scaled anchor-based distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $n^{-1/2}\phi(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k)$, has a discriminating property when $n \to \infty$. Namely, taking the illustrative case from Section III-A as an example, we find that $n^{-1/2}\phi(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k) \xrightarrow{P} 0$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ come from either $\mathcal{N}_1$ or $\mathcal{N}_2$, when $n \to \infty$. In contrast, $n^{-1/2}\phi(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k) \xrightarrow{P} \lambda$, where $\lambda > 0$, if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are drawn from different populations, when $n \to \infty$. The general proof for the discriminating property of the anchor-based distance is provided in Suppl. Lemma 1 in Section D in the Supplementary Material.

The anchor-based distance in (8) is a semi-metric or pseudo-metric, since it satisfies the rule of symmetry and triangle inequality, but may violate the identity of indiscernible. That means it is possible to get $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ such that $\phi(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k) = 0$. However, it is worth noting that if $\boldsymbol{x}_i$, $\boldsymbol{x}_j$, and $\boldsymbol{x}_k$ come from absolutely continuous distributions, for $\boldsymbol{x}_i \neq \boldsymbol{x}_j$, $\phi(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k) > 0$ holds with probability 1 [11]. For practical purposes, it behaves like a metric.

Therefore, we use the anchor-based distance to define the affinity tensor $\mathcal{T}^\star = [\mathcal{T}^\star_{i,j,k,l}]$. Formally

$$\mathcal{T}^\star{}_{i,j,k,l} = \exp\left(-\sigma\frac{\zeta_{ijkl}}{\eta_{ijkl}+\varepsilon}\right) \tag{9}$$

where

$$\begin{aligned}\zeta_{ijkl} = {}&\phi^2(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k)+\phi^2(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_l)\\ &+\phi^2(\boldsymbol{x}_i,\boldsymbol{x}_l,\boldsymbol{x}_k)+\phi^2(\boldsymbol{x}_i,\boldsymbol{x}_l,\boldsymbol{x}_j)\\ &+\phi^2(\boldsymbol{x}_k,\boldsymbol{x}_j,\boldsymbol{x}_i)+\phi^2(\boldsymbol{x}_k,\boldsymbol{x}_j,\boldsymbol{x}_l)\\ &+\phi^2(\boldsymbol{x}_k,\boldsymbol{x}_l,\boldsymbol{x}_i)+\phi^2(\boldsymbol{x}_k,\boldsymbol{x}_l,\boldsymbol{x}_j)\end{aligned} \tag{10}$$

$$\begin{aligned}\eta_{ijkl} = \sqrt{n}\{&\phi(\boldsymbol{x}_i,\boldsymbol{x}_k,\boldsymbol{x}_j)+\phi(\boldsymbol{x}_i,\boldsymbol{x}_k,\boldsymbol{x}_j)\\ &+\phi(\boldsymbol{x}_j,\boldsymbol{x}_l,\boldsymbol{x}_k)+\phi(\boldsymbol{x}_j,\boldsymbol{x}_l,\boldsymbol{x}_i)\}.\end{aligned} \tag{11}$$

As detailed in Suppl. Lemma 2 in Section D in the Supplementary Material, the mathematical properties of the anchor-based distance ensure that its discriminating capability is invariant with respect to the choice of anchor. Since in most cases, the choice of the anchor makes no significant difference, when computing the distance of two samples, we use the rest two samples as the anchors as in (10) and (11). For example, when computing the anchor-based distance of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, we use $\boldsymbol{x}_k$ and $\boldsymbol{x}_l$ as the two anchors.

Regarding (9), we note that such an affinity tensor comes with a desirable property for tackling HDLSS data. For example in early synthetic dataset, when $(\boldsymbol{x}_i,\boldsymbol{x}_k) \sim \mathcal{N}_2$ and $(\boldsymbol{x}_j,\boldsymbol{x}_l) \sim \mathcal{N}_2$, we note $\mathcal{T}^\star_{i,j,k,l} \xrightarrow{P} 1$, as $n \to \infty$. In comparison, when $(\boldsymbol{x}_i,\boldsymbol{x}_k) \sim \mathcal{N}_2$ and $(\boldsymbol{x}_j,\boldsymbol{x}_l) \sim \mathcal{N}_1$, we note $\mathcal{T}^\star_{i,j,k,l} \xrightarrow{P} 0$, as $n \to \infty$. Thus, we term the proposed one in (9) as **discriminating affinity tensor**.

A general analysis of the proposed discriminating affinity tensor is given as follows.

Theorem 2: Suppose all the samples in $\boldsymbol{X}$ satisfy Assumption 1. For a general case, i.e., $i \neq j \neq k \neq l$, if $\boldsymbol{x}_i \sim \mathcal{F}_p$, $\boldsymbol{x}_j \sim \mathcal{F}_q$, $\boldsymbol{x}_k \sim \mathcal{F}_r$, and $\boldsymbol{x}_l \sim \mathcal{F}_s$ are four independent samples in $\boldsymbol{X}$, as $n \to \infty$, the affinity tensor $\mathcal{T}^\star$ in (9) has the following property:

$$\forall i,j,k,l \in [m]$$

$$\mathcal{T}^\star_{i,j,k,l} \xrightarrow{P} \begin{cases}0, & \text{if and only if } p=r, q=s, \text{ and } p \neq q\\ 1, & \text{if and only if } p=r=q=s\\ \lambda, & \text{where } 0<\lambda<1, \quad \text{otherwise.}\end{cases} \tag{12}$$

Proof: The proof is provided in Section D in the Supplementary Material. □

Compared with the affinity tensor defined in (1), the discriminating one defined in (9) enables separating samples from different clusters. The conventional affinity tensor fails to distinguish HDLSS samples from different clusters. The reason is that when $n \to \infty$, it converges to indiscriminating constants as proved in Proposition 2. In contrast, Theorem 2 shows that the redefined affinity tensor in (9) explicitly addresses such limitation. When $n \to \infty$, the pair-to-pair similarities converge to 1 when four samples belong to the same cluster, and 0 when two pairs come from distinct clusters. This property allows us to improve the clustering performance when dealing with HDLSS data. We perform an experiment in a synthetic dataset in Section IV-B1 to demonstrate that asymptotic behavior of the affinity tensor. One can observe from Fig. 4 that the affinity keeps discriminating especially when the feature dimension is large ($n > 400$).

With such discriminating affinity tensor, the associated high-order affinity matrix is expected to perform accurate clustering for HDLSS data. The associated Laplacian tensor is computed by [16]

$$\mathcal{L}^\star_{:,p,:,q} = \mathcal{I}_{:,p,:,q} - \mathcal{D}^{\star\,-\frac{1}{2}}_{:,p,:,q}\mathcal{T}^\star_{:,p,:,q}\mathcal{D}^{\star\,-\frac{1}{2}}_{:,p,:,q} \quad \forall p,q \in [m] \tag{13}$$

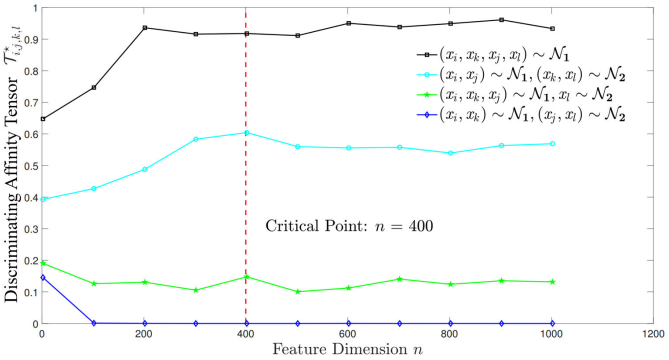


Fig. 4. Discriminating affinity tensor value $\mathcal{T}^{\star}_{i,j,k,l}$ on SD-1 data with dimension $n = 2$ to $n = 900$. The black, cyan, green, and blue lines denote the discriminating affinity tensor value on $(\boldsymbol{x}_i, \boldsymbol{x}_k) \sim \mathcal{N}_1$ and $(\boldsymbol{x}_j, \boldsymbol{x}_l) \sim \mathcal{N}_1$, $(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{N}_1$ and $(\boldsymbol{x}_k, \boldsymbol{x}_l) \sim \mathcal{N}_2$, $(\boldsymbol{x}_i, \boldsymbol{x}_k, \boldsymbol{x}_j) \sim \mathcal{N}_1$ and $\boldsymbol{x}_l \sim \mathcal{N}_2$, and $(\boldsymbol{x}_i, \boldsymbol{x}_k) \sim \mathcal{N}_1$ and $(\boldsymbol{x}_j, \boldsymbol{x}_l) \sim \mathcal{N}_2$, respectively. The figure shows that when $n > 400$, the values of $\mathcal{T}^{\star}_{i,j,k,l}$ start to approximate the theoretical results in Theorem 2.

where $\boldsymbol{\mathcal{D}}^{\star}$ is the associated degree tensor. Afterward, we solve the high-order affinity matrix via

$$\boldsymbol{\mathcal{L}}^{\star} \cdot \boldsymbol{V}^{\star} = \lambda \boldsymbol{V}^{\star}. \tag{14}$$

Upon having the eigenmatrices and then averaging them to have a high-order affinity matrix, the cluster labels can be obtained by performing a popular clustering method, such as standard SC, on it. We summarize the clustering procedures in Algorithm 1 and term the method as DTSC.

Algorithm 1 DTSC

```
Input:
    Dataset with $m$ samples $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$;
    Number of eigenmatrices $b$;
    Number of nearest neighbors $k$ for constructing $\boldsymbol{\mathcal{T}}^{\star}$.
Output:
    The clustering labels;

    Step 1: Constructing the discriminating affinity tensor $\boldsymbol{\mathcal{T}}^{\star}$ under the $k$ nearest neighbor sampling strategy, and computing its corresponding Laplacian tensor $\boldsymbol{\mathcal{L}}^{\star}$.
    Step 2: Solving (14) and obtaining eigen-matrices $\boldsymbol{V}^{\star(i)}$ for $i = 1, 2, \ldots, b$;
    Step 3: Averaging those eigenmatrices into a high-order affinity matrix: $\bar{\boldsymbol{V}}^{\star} = \frac{1}{b}\sum_{i=1}^{b} \boldsymbol{V}^{\star(i)}$;
    Step 4: Conducting spectral clustering on $\bar{\boldsymbol{V}}^{\star}$ to obtain the cluster labels.
```

C. Sampling Strategy and Complexity Analysis

The major computational cost in our method lies in constructing the discriminating affinity tensor $\boldsymbol{\mathcal{T}}^{\star}$. We propose to use a sampling strategy that selectively calculates $\boldsymbol{\mathcal{T}}^{\star}$ to save computational cost. Using Euclidean distance as the measurement, we only choose the top k nearest neighbors of $\boldsymbol{x}_i$ and generate $k(k-1)/2$ pairs to form the group $\boldsymbol{G}_i$, for $\forall i \in [m]$. Afterward, we calculate the pair-to-pair similarity using (9) among m groups, i.e., $\boldsymbol{G}_i$, for $\forall i \in [m]$. The remaining entries of $\boldsymbol{\mathcal{T}}^{\star}$ are set to 0. A similar scheme has been used by IPS2 [16] and achieved balanced performance. We will further examine the selection of the optimal parameter k in Section IV-D.

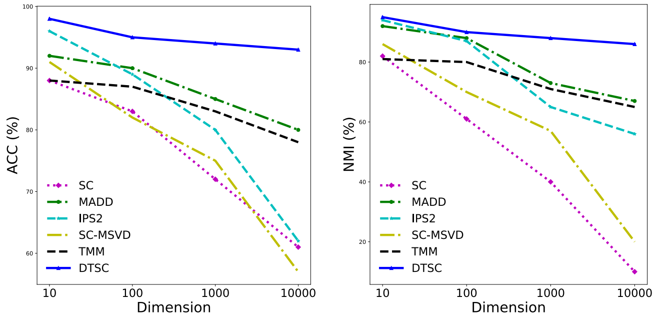


Fig. 5. ACC and NMI performance comparison of DTSC, IPS2, MADD, SC-MSVD, TMM, and SC on SD-1 with increasing dimensions and uniform noise. As can be seen, with the increasing dimension, IPS2, SC-MSVD and SC fail to obtain precise clustering, whereas DTSC keeps performing an accurate clustering.

Since we only use k-nearest-neighbor of each sample to build the affinity tensor, the computational cost is thereby $\mathcal{O}(m^2k^4)$. Furthermore, according to [16], we solve the sparse eigenvalue problem in (3) with computational time complexity bounded by $\mathcal{O}(m^2k^4)$. Finally, the computational cost of the standard SC is bounded by $\mathcal{O}(m^3)$. In all, the total cost is $\mathcal{O}(m^2k^4 + m^2k^4 + m^3) = \mathcal{O}(m^3 + m^2k^4)$.

IV. Experiments and Results

This section demonstrates the effectiveness of our method by comparing it with several state-of-the-art methods on synthetic and benchmark (HDLSS) datasets. All our experiments were performed on a desktop computer with a 3.70-GHz Intel[1] Core[2] i7-8700K CPU, 32.0 GB of RAM, and conducted by MATLAB R2016a (x64).

A. Experimental Settings

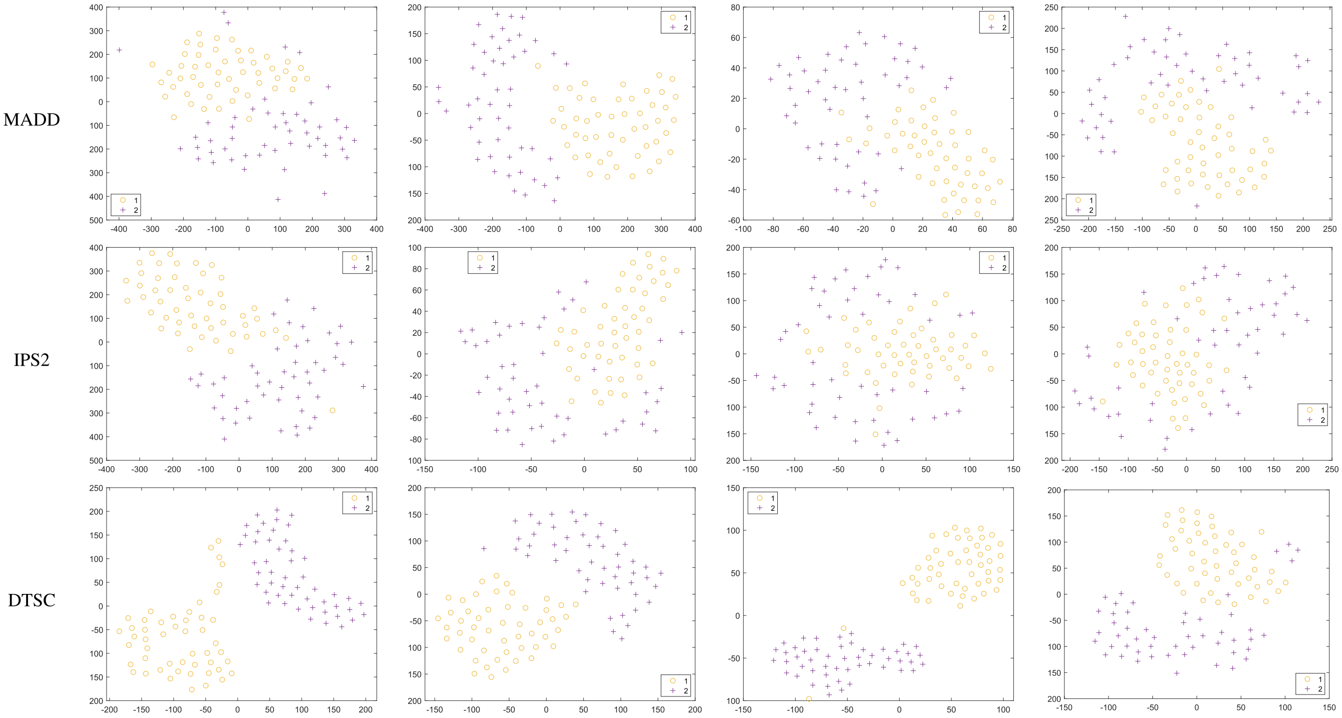
1) Performance Measurements: Throughout all the experiments, we use three widely accepted measurements to quantify clustering performance: accuracy (**ACC**), F-score (**F-SCORE**), and normalized mutual information (**NMI**). The detailed definitions are provided in Section F in the Supplementary Material.

2) Methods for Comparison: As baselines to our method, the competitors are: SC [5] (NIPS-2002), SC using multilinear SVD (SC-MSVD, AAAI-2015), uniform hypergraph partitioning: provable tensor methods and sampling techniques (TMM, JMLR-2017), MADD [11] (TPAMI-2019), IPS2 [16] (TPAMI-2020), robust matrix factorization with spectral embedding [29] (RMS, TNNLS-2020), and convex subspace clustering by adaptive block diagonal representation [30] (ABDR, TNNLS-2022).

3) Hyperparameter Setting and Computational Protocols: During the experiment, we empirically set the number of eigenmatrices b to be same with c, the number of clusters, and set the number of the nearest neighbor $k = 9$ for both DTSC and IPS2, as suggested in [16]. Regarding SC, SC-MSVD, TMM, and MADD, we adopted their default setting

[1] Registered trademark.

[2] Trademarked.



(a) (b) (c) (d)

Fig. 6. T-SNE visualization comparison on SD-1 with uniform noise and varying dimensions (a) $n = 10$, (b) $n = 100$, (c) $n = 1000$, and (d) $n = 10\,000$. The first, second, and third lines correspond MADD, IPS2, and DTSC, respectively. Here, "o" and "+" denote different classes.

from the original articles. In terms of a fair comparison, we ran each method 50 times for a fair comparison and calculated their mean values of the corresponding metrics. To further understand such behavior, we visualize their clustering results. To be specific, we used t-distributed stochastic neighbor embedding (T-SNE) to extract 2-D representations from the spectral embedding of the respective methods and then visually demonstrated their clustering behaviors.

B. Evaluation on Synthetic Dataset

We conducted three experiments on synthetic datasets to validate the performance of the proposed method. In the first example, we shall show that the proposed discriminating affinity tensor induced by the anchor-based distance preserves asymptotic behavior with the increase in the feature dimension n. The second and third examples are to test the robustness of the proposed method over noise contamination and concentration effect caused by the increase in feature dimension n.

1) Discriminating Affinity Tensor Converges Asymptotically to Discriminative Constants: We generated a synthetic dataset called SD-1, which consists of 100 samples drawing from two Gaussian distributions $\mathcal{N}_1(\mathbf{0}^d, \sigma_1^2\mathbf{\Sigma}^d)$ and $\mathcal{N}_2(\boldsymbol{\mu}^d, \sigma_2^2\mathbf{\Sigma}^d)$, with 50 samples for each. Herein, $\mathbf{0}^d = (0, \ldots, 0)^\top$, $\boldsymbol{\mu}^d = (1, -1, \ldots, (-1)^{d+1})^\top$, and $\mathbf{\Sigma}^d$ is a block-diagonal matrix, in which $\mathbf{\Sigma}_{i,i}^d = 1$, $\forall d$, $\mathbf{\Sigma}_{2i-1,2i}^d = \mathbf{\Sigma}_{2i,2i-1}^d = 0.98$ for $i = 1, 2, \ldots, \lfloor d/2 \rfloor$, and $\mathbf{\Sigma}_{i,j}^d = 0$ otherwise. We then computed the discriminating affinity tensor defined in (9). The similarity for the i, j, k, lth samples can be divided into four cases: 1) $(\boldsymbol{x}_i, \boldsymbol{x}_k, \boldsymbol{x}_j, \boldsymbol{x}_l) \sim \mathcal{N}_1$; 2) $(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{N}_1$ and $(\boldsymbol{x}_k, \boldsymbol{x}_l) \sim \mathcal{N}_2$; 3) $(\boldsymbol{x}_i, \boldsymbol{x}_k, \boldsymbol{x}_j) \sim \mathcal{N}_1$ and $\boldsymbol{x}_l \sim \mathcal{N}_2$; and 4) $(\boldsymbol{x}_i, \boldsymbol{x}_k) \sim \mathcal{N}_1$ and $(\boldsymbol{x}_j, \boldsymbol{x}_l) \sim \mathcal{N}_2$. For each case, we computed the similarity value versus the increase in the feature dimension n. The similarity $\mathcal{T}^\star_{i,j,k,l}$ is plotted in Fig. 4. It shows that when $n > 400$, the values of $\mathcal{T}^\star_{i,j,k,l}$ start to approximate the theoretical results in Theorem 2.

2) DTSC Is Robust Against Concentration Effect: To demonstrate the clustering performance on HDLSS data, we tested the clustering behaviors of each method on SD-1, by expanding the feature dimension with n being 10, 100, 1000, and 10 000, and we further added uniform noise to SD-1 with the lower and upper bounds being 0 and 0.2, consistently.

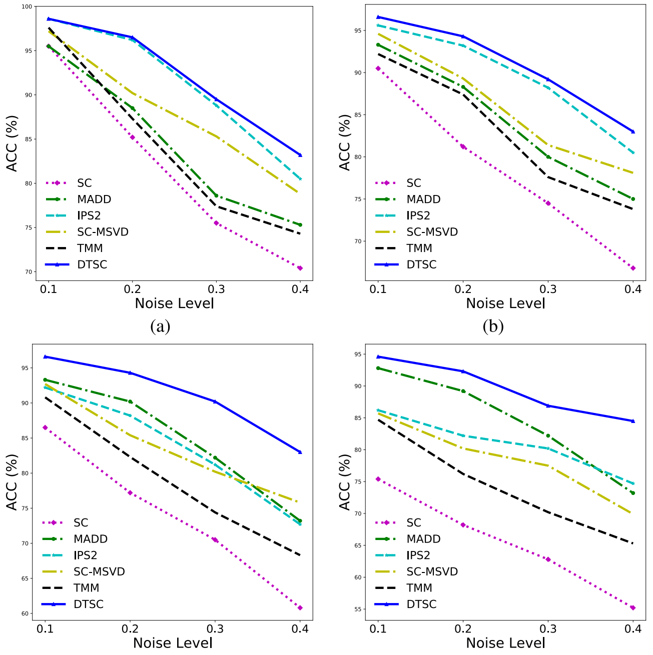
The results of ACC and NMI are shown in Fig. 5. In brief, DTSC comes at the top in ACC and NMI, showing its better ability to tackle HDLSS data with underlying noise. We further draw the following observations. First, comparing with IPS2 and DTSC, we can see that they both achieve satisfactory results when $n = 10$ with noise. However, the ACC and NMI of IPS2 gradually decrease when the dimension increases, whereas DTSC keeps its ability to clustering. To be specific, the ACC differences between DTSC and IPS2 are 2%, 6%, 14%, and 31%, when dimension equals 10, 100, 1000, and 10 000, respectively. Then, we performed T-SNE on the corresponding spectral embedding to obtain the low-dimensional representations and visual results. As can be seen in Fig. 6, the visual results are consistent with the ACC and NMI performance. In particular, IPS2 can learn a separable low-dimensional representation when $n = 10$, whereas it fails to maintain such separability when n approaches 1000. In contrast, DTSC keeps its ability of learning discriminating low-dimensional representations from $n = 10$ to $10\,000$. Second, DTSC, MADD, and TMM see their performance decrease from $n = 10$ to $10\,000$, with the decrease in ACC being 5%, 12%, and 10%, respectively. Third, all the pairwise distance-based methods, including SC, SC-MSVD, and IPS2, receive a striking performance drop when feature dimensions increase, with their declines in ACC being 27%, 34%, and 34%, respectively.

The foregoing results can be explained in the following. First, the drops of ACC and NMI performance are mainly due to the so-called concentration effect of the pairwise Euclidean distance. IPS2 adopts the Euclidean distance to characterize the multiwise similarities of samples, and SC-MSVD uses it to construct hypergraphs for clustering, achieving satisfactory performance when the dimension is 10. However, as the dimension increases, the concentration effect of Euclidean distance misleads IPS2 and SC-MSVD. The multiwise similarities captured by them start to lose discriminability, as analyzed in Section III-A. It is no wonder that IPS2, even with multiwise similarities, fails to clustering HDLSS data accurately. Second, DTSC, on the other hand, adopts the anchor-based distance to delineate the multiwise similarities, obtaining improved clustering across different dimensions. The anchor-based distance used in the discriminating affinity tensor allows our method to handle HDLSS data clustering. Third, despite the ability to mitigate the concentration effect, DTSC, MADD, and TMM receive a performance drop due to underlying noise. However, DTSC performs more stably than others as it characterizes a more comprehensive spatial structure and enables a stable similarity estimation for defying noise contamination. More noise contamination studies will be seen in what follows.

3) DTSC Is Robust Against Noise Contamination: To comprehensively evaluate the clustering behavior of the tested methods, the synthetic data, termed SD-2, were constructed. The SD-2 consists of three classes of samples, with each comprising 20 samples, independently drawing from multivariate Gaussian distribution with an equal standard deviation of 0.5 and different mean values of 0.1, 0.5, and 1. The feature dimensions vary from 500 to 10 000. SD-2 was further contaminated by uniform noise with varying levels. The lower bound of noise is set to 0 consistently, while the upper bound varies from 0.1 to 0.4.

In Fig. 7, we reported the ACC of the methods as mentioned earlier on SD-2 with varying noise levels and increasing dimensions. From Fig. 7, we underscore the following observations.

1) The curves corresponding to the proposed DTSC, denoted by the blue solid line with triangular, are located at the top in all the figures, indicating its best overall performance. These results demonstrate that DTSC is capable of dealing with HDLSS data and noise contamination simultaneously.
2) The DTSC is more noise-robust to other methods. Although all the methods face a performance decrease with the increasing noise levels, DTSC obtains a less performance decrease. For example, with noise level from 0.1 to 0.4, the ACC decrease in DTSC is 15.4%, 13.6%, 13.6%, and 10.1% for $n = 500, 1000, 5000, 10\,000$, whereas the ACC decrease in MADD is 20.2%, 18.3%, 20.1%, and 19.6%, respectively. Furthermore, TMM and SC-MSVD received a striking performance decrease. For $n = 500, 1000, 5000, 10\,000$, the ACC decrease in TMM is 23.3%, 18.4%, 22.5%, and 19.4%, respectively, while the one of SC-MSVD is 18.4%, 16.5%, 16.9%, and 25.8%, respectively. This implies that DTSC is relatively more stable against noise contamination.



(a)

(b)

(c)

(d)

Fig. 7. ACC performance comparison of DTSC, IPS2, MADD, SC-MSVD, TMM, and SC on SD-2 with varying noise levels ranging from 0.1 to 0.4 and dimensions (a) $n = 500$, (b) $n = 1000$, (c) $n = 5000$, and (d) $n = 10\,000$.

3) The DTSC achieves superior clustering performance than IPS2 under all levels of noise. Notably, as the dimension increases, the performance improvement of DTSC becomes more striking. For example, the ACC gap between DTSC and IPS2 is 2.7% when $n = 500$ and the noise level $= 0.4$, whereas the gap rises to 9.8% when $n = 10\,000$ with the same noise level.

We attribute the superiority of the proposed DTSC to the anchor-based distance underlying the discriminating multiwise similarities. In detail, first, DTSC leverages multiwise similarities for clustering, which is less susceptible to underlying noise. In contrast, MADD leverages the pairwise similarities, which are sensitive to noise even with a small one. In addition, since SC-MSVD depends on the pairwise distance to construct the multiwise similarities, it faces a significant performance drop when either feature dimension or noise increases. Second, DTSC adopts the anchor-based distance to build up multiwise similarities, while IPS2 still pins on the pairwise Euclidean distance. IPS2 faces a performance decrease when the dimensions increase due to the similar concentration effect. In contrast, DTSC keeps its discriminability for better clustering behavior, as validated by the experiment.

C. Evaluation on Benchmark HDLSS Dataset

Benchmark Data: In total, we included four publicly available benchmark datasets, all of which exhibit high dimensionality. Among them, DBWorld[3] is a text dataset,

[3] https://archive.ics.uci.edu/ml/datasets/DBWorld+e-mails

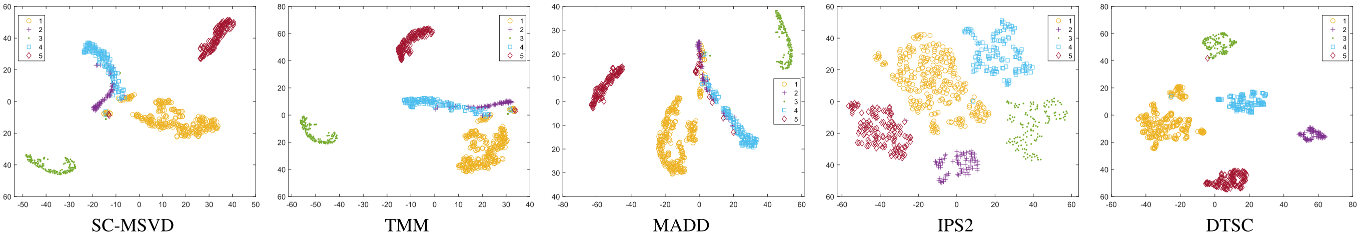


Fig. 8. T-SNE visualization of SC-MSVD, TMM, MADD, IPS2, and DTSC on UCI-gene, indicating the performance of different clustering methods. Here different shapes denote different classes. As can be seen, DTSC can learn a more separable representation with clear class boundaries.

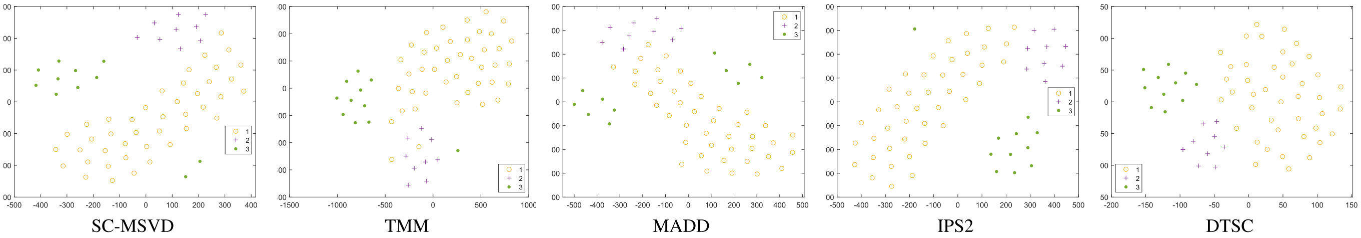


Fig. 9. T-SNE visualization of SC-MSVD, TMM, MADD, IPS2, and DTSC on Lymphoma, indicating the performance of different clustering methods. Here, different shapes denote different classes. As can be seen, DTSC can also learn a more separable representation with clear class boundaries.

TABLE I

STATISTICS OF DATASETS

Dataset	#Sample	#Feature	#Cluster
COIL20	1440	11078	20
UCI gene	801	20531	5
DBWorld	64	4702	2
Lymphoma	62	4702	3

TABLE II

PERFORMANCE COMPARISON ON DBWORLD AND COIL-20 (MEAN ± STANDARD DEVIATION, %). THE BEST RESULTS ARE **BOLDFACED** AND THE SECOND BEST ONES ARE UNDERLINED

Dataset	Algorithm	ACC	F-SCORE	NMI
DBWorld	SC (NIPS-2002)	65.9±0.1	43.3±0.1	36.3±0.1
	SC-MSVD (AAAI-2015)	77.6±0.7	72.8±0.8	42.9±0.6
	TMM (JMLR-2017)	78.1±1.2	73.2±1.1	43.5±1.2
	MADD (TPAMI-2019)	76.3±0.2	70.7±0.2	41.0±0.1
	IPS2 (TPAMI-2020)	80.1±2.2	77.2±2.0	44.5±1.8
	RMS (TNNLS-2020)	72.7±0.1	56.7 ±0.3	39.3 ±0.3
	ABDR (TNNLS-2022)	78.5±0.2	74.6 ±0.1	44.1 ±0.2
	DTSC ($k = 13$)	**89.0**±0.1	**80.0**±0.1	**50.1**±0.1
COIL-20	SC (NIPS-2002)	44.9±0.8	38.0±0.6	62.2±0.5
	SC-MSVD (AAAI-2015)	75.8±1.7	74.3±1.5	65.8±1.7
	TMM (JMLR-2017)	80.3±0.2	79.2±0.1	90.1±0.2
	MADD (TPAMI-2019)	76.8±1.2	76.3±1.4	89.8±1.2
	IPS2 (TPAMI-2020)	80.8±1.2	80.3±1.2	90.2±1.7
	RMS (TNNLS-2020)	75.4±0.5	73.8±0.7	87.4±0.6
	ABDR (TNNLS-2022)	81.6±1.1	80.9±1.0	91.1±1.0
	DTSC ($k = 8$)	**86.1**±0.3	**84.1**±0.3	**94.5**±0.5

COIL20[4] is image dataset, and the rest, i.e., Lymphoma [31] and UCI gene,[5] are bioinformatics datasets. We chose those datasets for two main reasons: 1) they are all from real-world scenarios, naturally exhibiting high-dimensionality and 2) they cover a wide range of applications, ideal for testing the performance of clustering methods. Statistics about this dataset are shown in Table I. The performance results are shown in Tables II and III. For ease of display, the best results are **boldfaced** and the second-best ones are underlined.

As can be seen in Tables II and III, DTSC consistently outperforms peer methods in terms of three metrics on all the datasets. In particular, our approach gains 8.9%, 5.3%, 7.7%, and 7.3% improvement over the second-best method in terms of ACC on DBWorld, COIL20, Lymphoma, and UCI gene, respectively. The striking results demonstrate the effectiveness of our proposed method, indicating its potential for dealing with a wide range of real-world clustering tasks with high dimensionality. Furthermore, we draw the following observations.

First, DTSC uniformly outperforms IPS2 and SC-MSVD on all the datasets. To be specific, DTSC is consistently superior to IPS2 in terms of all the three metrics on DBWorld, COIL20, Lymphoma, and UCI gene. Such consistent improvement arises from the difference between the anchor-based distance used in DTSC and the pairwise distance used in IPS2 and SC-MSVD. Because of the adoption of the pairwise distance, IPS2 and SC-MSVD suffer from the concentration effect for high-dimensional data, making it challenging to group samples accurately. In contrast, DTSC presents favorable discriminability for HDLSS data, leading to desirable clustering performance. In addition, from Figs. 8 and 9, we can note that DTSC learns a more discriminating cluster boundary than IPS2 and HSCs. That is to say, the low-dimensional embedding learned by DTSC scatters and gathers more distinctly, indicating a better clustering performance.

Second, DTSC exceeds MADD with a large margin on all the datasets across different metrics. For instance, the ACC gap is 17.8% and 22.4%, respectively, on Lymphoma and UCI gene. In addition, we can see from Figs. 8 and 9

[4] http://www1.cs.columbia.edu/CAVE/software/softlib/coil-20.php

[5] https://archive.ics.uci.edu

TABLE III

Performance Comparison on Lymphoma and UCI-Gene (Mean ± Standard Deviation, %). The Best Results Are **Boldfaced** and the Second Best Ones Are Underlined

Dataset	Algorithm	ACC	F-SCORE	NMI
Lymphoma	SC (NIPS-2002)	68.2±1.2	56.3±1.2	63.8±1.3
	SC-MSVD (AAAI-2015)	80.9±0.5	79.5±1.0	82.1±1.1
	TMM (JMLR-2017)	88.5±0.1	89.3±0.1	84.6±0.6
	MADD (TPAMI-2019)	80.6±0.3	71.7±0.2	71.3±0.3
	IPS2 (TPAMI-2020)	90.7±0.2	89.4±0.1	84.8±0.6
	RMS (TNNLS-2020)	77.8±0.4	71.4±0.3	73.6±0.2
	ABDR (TNNLS-2022)	88.9±0.4	87.3±0.1	81.9±0.2
	DTSC ($k = 11$)	**98.4**±0.1	**97.3**±0.1	**92.6**±0.1
UCI-gene	SC (NIPS-2002)	70.8±0.2	70.5±0.2	72.8±0.1
	SC-MSVD (AAAI-2015)	82.9±0.3	81.2±0.7	85.2±1.3
	TMM (JMLR-2017)	88.1±0.2	89.8±0.1	87.2±0.5
	MADD (TPAMI-2019)	77.4±0.2	75.0±0.1	74.2±0.2
	IPS2 (TPAMI-2020)	92.5±0.2	90.3±0.1	89.9±0.2
	RMS (TNNLS-2020)	79.8±0.3	76.2±0.2	74.9±0.1
	ABDR (TNNLS-2022)	86.7±0.4	84.5±0.3	82.8±0.3
	DTSC ($k = 9$)	**99.8**±0.1	**99.7**±0.1	**99.4**±0.1

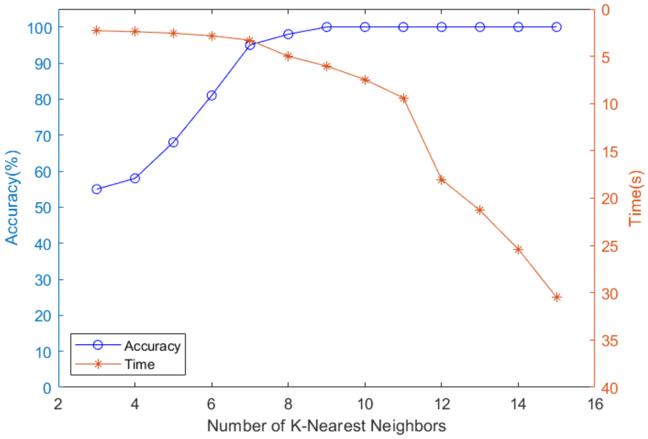


Fig. 10. Clustering ACC by the proposed method with a varying k nearest neighbors on SD-1 without noise with dimension $n = 1000$. This figure shows that as k increases both clustering ACC and running time increase, suggesting that the range [5, 15] can be appropriate to meet the balance between efficiency and effectiveness.

that visual results of MADD have vague boundaries among different classes of samples. MADD obtains intertwined samples across distinct classes. Combining the results from Fig. 7, we infer that the unsatisfactory performance of MADD stems from its pairwise similarity matrix. The pairwise similarity is susceptible to noise contamination, especially when facing real-world HDLSS data. Confronting HDLSS data with underlying noise, MADD can hardly preserve accurate clustering. At the same time, DTSC, which rests on affinity tensor, has shown its robustness against noise and thus achieves desirable clustering performance.

D. Hyperparameter Sensitivity and Running Time Analysis

We are also interested in how the number of nearest neighbors k of DTSC affects its clustering performance. We constructed synthetic data similar to SD-1, but with 300 rather 100 samples. The dimensions n were set to 1000. The results are shown in Fig. 10. Fig. 10 shows that TSC performs stably in terms of ACC for $k > 9$. In contrast, the running time increases rapidly as k goes from 11 to 15. Choosing a small value of k remarkably saves computational time. Yet, too small k impedes the performance. Therefore, we choose the range [5, 15] for tuning k to meet the balance between efficiency and effectiveness.

V. Conclusion

This article has proposed a TSC method to address the HDLSS data clustering problem. In particular, an anchor-based distance is introduced to form the discriminating affinity tensor that copes with the concentration effect raised by the HDLSS data. It is proved that under some mild conditions, the proposed method can differentiate samples from distinct clusters, addressing the concentration effect and improving clustering performance. Meanwhile, the discriminating affinity tensor is enabled to mitigate noise contamination by characterizing a comprehensive spatial structure of multiple samples and allows for a stable similarity estimation. Apart from theoretical analysis, extensive experiments have been conducted on synthetic and benchmark datasets to verify the promising performance of our method in comparison to the recent methods.

Overall, our approach demonstrates competitive clustering performance on HDLSS data, and there are a potential directions for improvement. Although our proposed method can effectively address the concentration effect in high-dimensional data clustering and outperforms several baseline methods on the benchmark dataset, it still confronts two limitations. One is that the proposed method needs to construct an affinity tensor, which requires high memory cost. The other is that this method is not able to perform feature learning and clustering jointly. To further address these two limitations, we consider incorporating a deep neural network to perform tensor-based deep learning. On one side, the deep neural network enables learning with stochastic optimization to reduce memory cost, and on the other side, it allows us to update deep feature learning with clustering.

References

[1] Y. Lu, Y.-M. Cheung, and Y. Yan Tang, “Adaptive chunk-based dynamic weighted majority for imbalanced data streams with concept drift,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 2764–2778, Aug. 2020.

[2] Y.-M. Cheung and Y. Zhang, “Fast and accurate hierarchical clustering based on growing multilayer topology training,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 876–890, Mar. 2019.

[3] H. Jia and Y.-M. Cheung, “Subspace clustering of categorical and numerical data with an unknown number of clusters,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3308–3325, Aug. 2018.

[4] Y. Hu, E. Guo, Z. Xie, X. Liu, and H. Cai, “Robust multi-view clustering through partition integration on Stiefel manifold,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 10, pp. 10397–10410, Mar. 2023.

[5] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in *Proc. Neural Inf. Process. Syst. Nat. Synthetic (NIPS)*, vol. 2, 2002, pp. 849–856.

[6] S. R. Bulò and M. Pelillo, “A game-theoretic approach to hypergraph clustering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1312–1327, Jun. 2013.

[7] S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. J. Kriegman, and S. J. Belongie, “Beyond pairwise clustering,” in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 838–845.

[8] Y. Hu and H. Cai, "Multi-view clustering through hypergraphs integration on Stiefel manifold," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[9] H. Peng, H. Wang, Y. Hu, W. Zhou, and H. Cai, "Multi-dimensional clustering through fusion of high-order similarities," *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108108.

[10] D. Francois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 7, pp. 873–886, Jul. 2007.

[11] S. Sarkar and A. K. Ghosh, "On perfect clustering of high dimension, low sample size data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2257–2272, Sep. 2020.

[12] P. Hall, J. S. Marron, and A. Neeman, "Geometric representation of high dimension, low sample size data," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 67, no. 3, pp. 427–444, Jun. 2005.

[13] V. M. Govindu, "A tensor decomposition for geometric grouping and segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 1150–1157.

[14] A. Shashua, R. Zass, and T. Hazan, "Multi-way clustering using super-symmetric non-negative tensor factorization," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 595–608.

[15] D. Ghoshdastidar and A. Dukkipati, "Spectral clustering using multilinear SVD: Analysis, approximations and applications," in *Proc. AAAI Conf. Artif. Intell.*, 2015, vol. 29, no. 1, pp. 2610–2616. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/2886521.2886684

[16] H. Peng, Y. Hu, J. Chen, H. Wang, Y. Li, and H. Cai, "Integrating tensor similarity to enhance clustering performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2582–2593, May 2022.

[17] D. Ghoshdastidar and A. Dukkipati, "A provable generalized tensor spectral method for uniform hypergraph partitioning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 400–409.

[18] D. Ghoshdastidar and A. Dukkipati, "Uniform hypergraph partitioning: Provable tensor methods and sampling techniques," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 1638–1678, 2017.

[19] T. Tian, J. Wan, Q. Song, and Z. Wei, "Clustering single-cell RNA-seq data with a model-based deep learning approach," *Nature Mach. Intell.*, vol. 1, no. 4, pp. 191–198, Apr. 2019.

[20] Y. Li, H. Peng, T. Dan, Y. Hu, G. Tao, and H. Cai, "Coarse-to-fine nasopharyngeal carcinoma segmentation in MRI via multi-stage rendering," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 623–628.

[21] J. Ahn, M. H. Lee, and Y. J. Yoon, "Clustering high dimension, low sample size data using the maximal data piling distance," *Statistica Sinica*, vol. 22, no. 2, pp. 443–464, Apr. 2012.

[22] S. Hu and L. Qi, "Algebraic connectivity of an even uniform hypergraph," *J. Combinat. Optim.*, vol. 24, no. 4, pp. 564–579, Nov. 2012.

[23] G. Li, L. Qi, and G. Yu, "The Z-eigenvalues of a symmetric tensor and its application to spectral hypergraph theory," *Numer. Linear Algebra Appl.*, vol. 20, no. 6, pp. 1001–1029, Dec. 2013.

[24] L. Qi, "H^+-eigenvalues of Laplacian and signless Laplacian tensors," *Commun. Math. Sci.*, vol. 12, no. 6, pp. 1045–1064, 2014.

[25] Y. Chen, L. Qi, and X. Zhang, "The Fiedler vector of a Laplacian tensor for hypergraph partitioning," *SIAM J. Sci. Comput.*, vol. 39, no. 6, pp. A2508–A2537, Jan. 2017.

[26] J. Chang, Y. Chen, L. Qi, and H. Yan, "Hypergraph clustering using a new Laplacian tensor with applications in image processing," *SIAM J. Imag. Sci.*, vol. 13, no. 3, pp. 1157–1178, Jan. 2020.

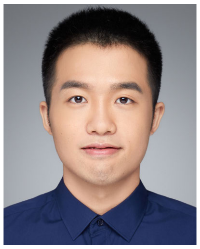
[27] H. Cai, Y. Hu, F. Qi, B. Hu, and Y.-M. Cheung, "Deep tensor spectral clustering network via ensemble of multiple affinity tensors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 5080–5091, Jul. 2024.

[28] M. Aoshima, D. Shen, H. Shen, K. Yata, Y.-H. Zhou, and J. Marron, "A survey of high dimension low sample size asymptotics," *Austral. New Zealand J. Statist.*, vol. 60, no. 1, pp. 4–19, 2018.

[29] M. Chen and X. Li, "Robust matrix factorization with spectral embedding," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5698–5707, Dec. 2021.

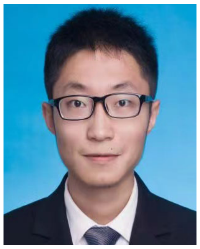
[30] Y. Lin and S. Chen, "Convex subspace clustering by adaptive block diagonal representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10065–10078, Dec. 2023.

[31] A. A. Alizadeh et al., "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, p. 503, 2000.



Yu Hu received the B.S. degree in electrical engineering and automation from the School of Information and Electric Engineering, China University of Mining and Technology, Xuzhou, China, in 2017, and the Ph.D. degree in computer science from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, in 2023.

He is currently a Research Associate with Guangdong Artificial Intelligence and Digital Economy Laboratory (Pazhou Laboratory), Guangzhou, China. His present research endeavors revolve around tensor-based machine learning, data mining and their applications in high-dimensional data clustering and multiview clustering.



Fei Qi received the B.S. and M.S. degrees from Xiamen University, Xiamen, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree in computer science and engineering with the South China University of Technology, Guangzhou, China.

His research interests include machine learning and image processing.

Yiu-Ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, in 2000.

He is currently a Chair Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning, pattern recognition, and visual computing.

Dr. Cheung is a fellow of the American Association for the Advancement of Science (AAAS), Institution of Engineering and Technology (IET), British Computer Society (BCS), and Asia-Pacific Artificial Intelligence Association (AAIA). He is the Editor-in-Chief of IEEE Transactions on Emerging Topics in Computational Intelligence. He also served as an Associate Editor for IEEE Transactions on Cybernetics, IEEE Transactions on Cognitive and Developmental Systems, IEEE Transactions on Neural Networks and Learning Systems from 2014 to 2020, *Pattern Recognition*, *Knowledge and Information Systems*, and *Neurocomputing*, just to name a few.

Hongmin Cai (Senior Member, IEEE) received the B.S. and M.S. degrees in mathematics from Harbin Institute of Technology, Harbin, China, in 2001 and 2003, respectively, and the Ph.D. degree in applied mathematics from The University of Hong Kong, Hong Kong, in 2007.

From 2005 to 2006, he was a Research Assistant with the Center of Bioinformatics, Harvard University, Cambridge, MA, USA, and the Section for Biomedical Image Analysis, University of Pennsylvania, Philadelphia, PA, USA. He is currently a Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His current research interests include bioinformatics, machine learning, and medical image analysis.