Context-Aware Hypergraph Construction for Robust Spectral Clustering

Xi Li, Weiming Hu, Chunhua Shen, Anthony Dick, Zhongfei Zhang

Abstract—Spectral clustering is a powerful tool for unsupervised data analysis. In this paper, we propose a context-aware hypergraph similarity measure (CAHSM), which leads to robust spectral clustering in the case of noisy data. We construct three types of hypergraph—the pairwise hypergraph, the $k$-nearest-neighbor ($k$NN) hypergraph, and the high-order over-clustering hypergraph. The pairwise hypergraph captures the pairwise similarity of data points; the $k$NN hypergraph captures the neighborhood of each point; and the clustering hypergraph encodes high-order contexts within the dataset. By combining the affinity information from these three hypergraphs, the CAHSM algorithm is able to explore the intrinsic topological information of the dataset. Therefore, data clustering using CAHSM tends to be more robust. Considering the intra-cluster compactness and the inter-cluster separability of vertices, we further design a discriminative hypergraph partitioning criterion (DHPC). Using both CAHSM and DHPC, a robust spectral clustering algorithm is developed. Theoretical analysis and experimental evaluation demonstrate the effectiveness and robustness of the proposed algorithm.

Index Terms—Hypergraph construction, spectral clustering, graph partitioning, similarity measure.

1 INTRODUCTION

Spectral clustering is an effective means of clustering data with complex topological structure [8]–[13], [15], [16], [31], [34]–[36]. It plays an important role in unsupervised learning from data, and therefore has a wide range of applications, including circuit layout [1], [2], load balancing [3], image segmentation [4]–[7], [14], motion segmentation [25], video retrieval [26], etc. Typically, the affinity relationships between data samples are modeled by a graph, and therefore spectral clustering aims to optimize a graph partitioning criterion for data clustering based on local vertex similarities. However, there are still several unsolved issues for traditional spectral clustering methods: i) how to automatically discover the number of clusters; ii) how to correctly choose the scaling parameter for graph construction; iii) how to counteract the adverse effect of noise or outliers; and iv) how to incorporate different types of information to enhance the clustering performance.

In the literature, Zelnik-Manor and Perona [11] attempt to address issues i) and ii) by designing a local scaling mechanism, which adaptively calculates the affinity matrix and explores the intrinsic structural information on the energy eigenvalue spectrum of the normalized graph Laplacian to discover the number of clusters. However, this local scaling mechanism is susceptible to noise or outliers. Following [11], Li et al. [15] propose a noise robust spectral clustering (NRSC) algorithm to resolve issues i) and iii). The proposed NRSC algorithm can automatically estimate the number of clusters via computing the largest eigenvalue gap of the normalized graph Laplacian. In addition, the proposed NRSC algorithm maps the original data samples (vertices) into a new space, in which the clusters have a higher intra-cluster compactness and inter-cluster separability. If the noisy data samples are weakly interconnected with each other, this mapping relocates the noisy data samples around the origin of the new space. This usually results in a compact noise cluster. However, the real-world data samples (e.g., images and videos) within a cluster are often not densely interconnected due to problems with the visual feature description. Therefore, the mapping may lead to topological information loss for the clusters. As a result, the weakly interconnected samples in the ordinary clusters are also relocated around the origin of the new space. This may result in low separability between clusters.

More recently, hypergraph analysis [17], [18] has emerged as a popular tool for addressing issues iii) and iv). The fundamental idea of hypergraph analysis is to explore the underlying affinity relationships among vertices by constructing a hypergraph with a variety of hyperedges that capture affinity. This has been applied to many domains such as image matching [19], multi-label classification [20], video object segmentation [21], and image retrieval [22]. For example, Sun et al. [20] carry out hypergraph construction by sequentially introducing new vertices into existing hyperedges using clique expansion or star expansion. Huang et al. [22] propose a probabilistic hypergraph model that softly assigns a vertex to a hyperedge according to the similarity between the vertex and the centroid of the hyperedge.

Motivation and contribution In general, most existing spectral clustering algorithms only focus on the pairwise interactions between vertices. In other words, the pairwise similarity between two vertices is only based on the individual vertices themselves. If a vertex is corrupted, this pairwise similarity can change significantly. Consequently, their true affinity may not be stably represented. Thus, designing a robust similarity measure is one of the key problems in data clustering.

Here, we show that the high-order contextual information on vertices can help alleviate this problem. Contexts are groups of vertices that share some common properties. Once contexts have been computed, the vertex similarity measure depends on not only two individual vertices but also their corresponding contexts. The similarity measure that includes contextual information is much more stable because it takes into account local grouping and neighborhood information of each vertex. When a single vertex is corrupted, the high-order contextual similarity can still provide complementary information to counteract the impact of the corruption.

Motivated by this observation, we propose a robust spectral clustering algorithm based on a context-aware hypergraph similarity measure. We use three different types of hypergraphs: pairwise hypergraph, $k$-nearest-neighbor ($k$NN) hypergraph, and high-order over-clustering hypergraph. The pairwise hypergraph is capable of encoding pairwise affinity information on vertices. In contrast, the $k$NN hypergraph and the over-clustering hypergraph capture the underlying manifold structure on vertices by modeling their high-order neighborhood and contextual grouping properties, respectively. By combining these hypergraphs, we obtain the context-aware hypergraph similarity measure that characterizes the intrinsic connectivity relationships among vertices, resulting in the clustering robustness in the case of noise or outlier corruption.

The main contributions of this work are therefore three-fold.

• We introduce a high-order context into the spectral clustering process. The high-order context of a vertex is defined as a set of vertices with similar properties to the vertex. Each vertex in a context is influenced by other vertices in the same context.

• The problem of building the high-order context is converted to that of hypergraph construction, which encodes the local affinity information using different types of hypergraphs. To this end, we design three types of hypergraphs to encode the pairwise, neighboring, and local grouping information on vertices. Based
on these hyperedges, we further propose a context-aware hypergraph similarity measure (CAHSM) to capture the intrinsic topological information on vertices. In essence, CAHSM is a generalization of traditional similarity measures, and aims to utilize the hypergraph context to explore the underlying affinity relationships between vertices.

- We propose a discriminative hypergraph partitioning criterion (DHPC) to characterize the intra-cluster compactness and the inter-cluster separability of vertices. By maximizing the DHPC, we effectively capture the discriminative information on vertices. The optimization of DHPC can be relaxed into a trace-ratio maximization problem. Using both CAHSM and DHPC, we develop a pairwise+kNN+over-clustering hypergraph spectral clustering algorithm (referred to as PKO+HSC) for data clustering.

## 2 CONTEXT-AWARE HYPERGRAPH CONSTRUCTION

In what follows, we first discuss how to construct a robust hypergraph similarity measure using three different types of hypergraphs, and then describe a discriminative hypergraph partitioning criterion for data clustering.

### 2.1 Context-aware hypergraph similarity measure

In order to effectively explore the high-order affinity relationships among vertices, we propose a hypergraph construction mechanism based on three types of hypergraphs, which are the pairwise hypergraph, the k-nearest-neighbor (kNN) hypergraph, and the over-clustering hypergraph. The pairwise hypergraph reflects the pairwise relationships between vertices; the kNN hypergraph characterizes the neighboring information on vertices, and the over-clustering hypergraph captures the local grouping relationships among vertices. By combining these three types of hypergraphs, the proposed hypergraph construction mechanism is capable of exploring the underlying high-order affinity relationships among vertices.

1) **Pairwise hypergraph.** For easy exposition, let $Z = \{z_i\}_{i=1}^N$ denote a sample set. Based on $Z = \{z_i\}_{i=1}^N$, we create a pairwise graph $G_p$ with $N$ vertices. Mathematically, the graph $G_p$ can be denoted as $G_p = (V, E_p, W_p)$, where $V = \{z_i\}_{i=1}^N$ is the vertex set corresponding to $\{z_i\}_{i=1}^N$, $E_p \subseteq V \times V$ is the edge set containing all possible pairwise edges, and $W_p$ is the edge-weight function returning the affinity value between two vertices. In practice, the graph $G_p$ is formulated as a weighted similarity matrix $A = (a_{ij})_{N \times N}$:

$$a_{ij} = \begin{cases} W_p(v_i, v_j) & \text{if } (v_i, v_j) \in E_p, \\ 0 & \text{otherwise}, \end{cases}$$

where $W_p(v_i, v_j) = \mathcal{G}(z_i, z_j)$ is a kernel function used for measuring the similarity between $z_i$ and $z_j$. Note that the above procedure of graph creation is independent of the choice of kernel functions.

In other words, it is easy to incorporate various kernel functions into the above graph creation process.

According to the hypergraph theory, the pairwise graph $G_p$ is merely a special hypergraph whose hyperedge cardinality equals 2. Therefore, we reformulate $G_p$ using the hypergraph terminologies. As a generalization of a traditional pairwise graph, a hypergraph is composed of many hyperedges, and each hyperedge corresponds to a set of vertices which have some common properties. Mathematically, these hyperedges are generally associated with a hypergraph incidence matrix $H_p = (h_p(v_i, e_p^k))_{|V| \times |E_p|}$:

$$h_p(v_i, e_p^k) = \begin{cases} 1 & \text{if } v_i \in e_p^k, \\ 0 & \text{otherwise}, \end{cases}$$

where $e_p^k = (v_m, v_n)$ is the $k$-th hyperedge of $E_p$. In order to measure the degree of the within-hyperedge vertices belonging to the same cluster, we introduce the notion of the pairwise hyperedge weight, which is defined as the pairwise similarity of the vertices in each hyperedge. So we can define the pairwise hypergraph similarity as:

$$u_{ij} = \sum_{e_p^k \in E_p} \eta_p(e_p^k) h_p(v_i, e_p^k) h_p(v_j, e_p^k) = a_{ij},$$

where $\eta_p(e_p^k)$ is the corresponding hyperedge weight of $e_p^k = (v_m, v_n)$ such that $\eta_p(e_p^k) = a_{mn}$. As a result, we have a weighted hypergraph similarity matrix $U = \{u_{ij}\}_{N \times N}$ that characterizes the pairwise affinity relationships between vertices. For simplicity, the resulting pairwise hypergraph similarity matrix is represented as its corresponding matrix form:

$$U = H_p \Sigma_p H_p^T = A,$$

where $\Sigma_p$ is a diagonal matrix whose diagonal elements are denoted as $\{\eta_p(e_p^k)\}_{e_p^k \in E_p}$.

2) **k-nearest-neighbor (kNN) hypergraph.** Based on the neighboring information on vertices, we further define a kNN hypergraph $G_n$, as shown in Fig. 1. For each vertex $v_i$, we search its corresponding $k$ nearest neighbors $\{v_j | v_j \in N_{kN}(v_i)\}$ (as shown in Fig. 2(a)), and then use these nearest neighbors to form a kNN hyperedge $e_p^k$. By concatenating all the kNN hyperedges, a kNN hyperedge set is generated as $E_n = \{e_p^k\}_{p=1}^P$, as illustrated in Fig. 3. To characterize the vertex-to-hyperedge membership, we define an indicator function as:

$$\mathbb{I}(v_i, e_p^k) = \begin{cases} 1 & \text{if } v_i \in e_p^k, \\ 0 & \text{otherwise}, \end{cases}$$
Based on this indicator function, we design a hypergraph model for softly assigning a vertex to each hyperedge:

$$h_n(v_i, e^o_j) = \frac{a_{k_j} \sqrt{T(v_i, e^o_j)}}{\sqrt{\sum_{k=1}^N \delta_k d_{i_k}(v_i, e^o_k) a_{i_k}^o}},$$

where $a_k$ is the hyperedge weight associated with the $k$-th hyperedge $e^o_k$ such that $\delta_k = |\{k \in E_n \mid |\{i \in V \mid e^o_k \in E_n \} = 1\}|$, and $a_{k_j}$ represents the vertex-to-hyperedge similarity between $v_i$ and the $k$-th $kNN$ hyperedge $e^o_k$. Specifically, $e^o_k$ is composed of a centroid vertex $v_i$ and its corresponding $k$ nearest vertices. Based on Eq. (1), the vertex-to-hyperedge similarity $a_{k_j}$ is computed as the pairwise similarity between $v_i$ and $v_j$. As a result, we obtain a $kNN$ hypergraph incidence matrix $H_n = (h_n(v_i, e^o_j))|_{V \times |E_n|}$ for capturing the vertex-to-hyperedge relationships. Based on $H_n$, a $kNN$ hypergraph similarity $b_{ij}$ between $v_i$ and $v_j$ is derived as:

$$b_{ij} = \sum_{e^o_k \in E_n} \delta_k h_n(v_i, e^o_k) h_n(v_j, e^o_k)$$

$$= \frac{\sqrt{\sum_{k=1}^N \delta_k d_{i_k}(v_i, e^o_k) h_n(v_i, e^o_k) d_{i_k}(v_j, e^o_k) h_n(v_j, e^o_k)}}{\sqrt{\sum_{k=1}^N \delta_k d_{i_k}(v_i, e^o_k) h_n(v_i, e^o_k) d_{i_k}(v_j, e^o_k) h_n(v_j, e^o_k)}}$$

where $x_m = \left( a_{k_m} \sqrt{T(v_m, e^o_k) \delta_k} \right)^N$, $< \cdot, \cdot >$ is the inner product operator, and $|| \cdot ||$ is the 2-norm. Indeed, $x_m$ is a vertex-to-hyperedge feature vector that characterizes the correlation between $v_m$ and the $kNN$ hyperedges. For example, the $\ell$-th element of $x_m$ contains two terms: $a_{k_m}$ and $\sqrt{T(v_m, e^o_k) \delta_k}$. The first term $a_{k_m}$ is the pairwise similarity between $v_m$ and the vertex $v_i$ of the $\ell$-th $kNN$ hyperedge, and the second term $\sqrt{T(v_m, e^o_k) \delta_k}$ measures the cohesiveness of the $\ell$-th $kNN$ hyperedge by computing the average connection similarity between the vertex $v_i$ and the other vertices in the $\ell$-th $kNN$ hyperedge. Thus, the $kNN$ hypergraph similarity $b_{ij}$ can be interpreted as the cosine similarity between two vertex-to-hyperedge feature vectors $x_i$ and $x_j$. As a result, we obtain a weighted similarity matrix $B = (b_{ij})_{N \times N}$ associated with $G_n$. Essentially, $B = (b_{ij})_{N \times N}$ aims to explore the local neighboring relationships between vertices. For simplicity, the $kNN$ hypergraph similarity matrix $B$ is denoted as its corresponding matrix form: $B = H_n \Sigma_m H_n^T$ where $\Sigma_m$ is a diagonal matrix with the $\ell$-th diagonal element being $\delta_k$. Fig. 1 gives an illustration of the process of the $kNN$ hypergraph construction.

3) High-order over-clustering hypergraph. In practice, the vertices are often distributed in different cohesive vertex communities, and each community contains a set of mutually correlated vertices with some common properties. In order to effectively discover such cohesive vertex communities, we propose a high-order over-clustering hypergraph $G_o$ based on over-clustering (or over-segmentation) using different clustering methods. Specifically, an over-clustering mechanism is employed to generate a set of vertex groups, each of which corresponds to a cohesive vertex community (as shown in Fig. 2(a)). In this case, the vertices belonging to the same vertex community are mutually influenced, and work as the high-order contexts of the other vertices in the same vertex community, as illustrated in Fig. 4. Without loss of generality, we assume that there are $L$ vertex communities in total. For convenience, let $E_o = \{e^o_l\}_{l=1}^L$ denote these vertex communities, each of which corresponds to a hyperedge $e^o_l$. Based on these hyperedges, we define the high-order over-clustering hypergraph incidence matrix $H_o = (h_o(v_i, e^o_l))|_{V \times L}$ as:

$$h_o(v_i, e^o_l) = \frac{\sqrt{\mu_l(v_i, e^o_l)(1 + \frac{1}{|\mathcal{E}|} \sum_{m \in \mathcal{E}_l} a_{m_l})}}{\sqrt{\sum_{e^o_n \in E_o} \mu_l(v_i, e^o_n)(1 + \frac{1}{|\mathcal{E}|} \sum_{m \in \mathcal{E}_l} a_{m_l})}},$$

where $\mathcal{E}$ is the indicator function in Eq. (5), $\mathcal{E}_l$ is the corresponding vertex index set of the nearest neighbors of $v_i$ in the hyperedge $e^o_l$ (s.t. $|\mathcal{E}_l| = 3$ in the experiments), and $\mu_l$ is the associated hyperedge weight of $e^o_l$ such that:

$$\mu_l = \frac{1}{2} \left( 1 + \frac{1}{|\mathcal{E}|} \sum_{e^o_n \in E_o} \sum_{m \in \mathcal{E}_l} a_{m_l} |\mathcal{E}_l| \right).$$

Here, $a_{m_l}$ determines the vertex-to-hyperedge similarity between $v_i$ and the $\ell$-th over-clustering hyperedge $e^o_l$. To ensure the robustness of similarity evaluation, we only take into account the affinity relationships between $v_i$ and its corresponding nearest neighbors (indexed by $\mathcal{E}_l$) in $e^o_l$. Therefore, $a_{m_l}$ is the similarity between $v_i$ and the $m$-th vertex of $\mathcal{E}_l$. With the definition of $H_o$, the high-order over-clustering hypergraph similarity $c_{ij}$ between $v_i$ and $v_j$ is formulated as:

$$c_{ij} = \sum_{e^o_n \in E_o} \mu_l h_o(v_i, e^o_l) h_o(v_j, e^o_l)$$

$$= \frac{\sum_{e^o_n \in E_o} \mu_l(v_i, e^o_l)(1 + \frac{1}{|\mathcal{E}|} \sum_{m \in \mathcal{E}_l} a_{m_l}) \mu_l(v_j, e^o_l)(1 + \frac{1}{|\mathcal{E}|} \sum_{m \in \mathcal{E}_l} a_{m_l})}{\sum_{e^o_n \in E_o} \mu_l(v_i, e^o_l)(1 + \frac{1}{|\mathcal{E}|} \sum_{m \in \mathcal{E}_l} a_{m_l}) \sum_{e^o_n \in E_o} \mu_l(v_j, e^o_l)(1 + \frac{1}{|\mathcal{E}|} \sum_{m \in \mathcal{E}_l} a_{m_l})}$$

where $\mathcal{E}$ is the $\mathcal{L}$-dimensional vector with the $\ell$-th element being $\sqrt{\mu_l(v_i, e^o_l)(1 + \frac{1}{|\mathcal{E}|} \sum_{m \in \mathcal{E}_l} a_{m_l})}$. Actually, $\mathcal{E}$ is a vertex-to-hyperedge feature vector, and its $\ell$-th element consists of two components: $\mu_l$ and $\sqrt{T(v_i, e^o_k) \delta_k}(1 + \frac{1}{|\mathcal{E}|} \sum_{m \in \mathcal{E}_l} a_{m_l})$. As defined in Eq. (9), the left term measures the average cross-link degree of the within-community-$\ell$ vertices with respect to the other vertices in the same community, while the right term reflects the average affinity relationships between $v_i$ and the vertices in the $\ell$-th community. Therefore, the high-order over-clustering hypergraph similarity $c_{ij}$ can be viewed as the cosine similarity between two vertex-to-hyperedge feature vectors $\mathcal{E}_i$ and $\mathcal{E}_j$. As a result, a weighted similarity matrix $C = (c_{ij})_{N \times N}$ is obtained to capture the local grouping information on vertices. For simplicity, the high-order over-clustering hypergraph similarity matrix $C$ can be expressed as its corresponding matrix form: $C = H_o \Sigma_o H_o^T$, where $\Sigma_o$ is a diagonal matrix with the $\ell$-th diagonal element being $\mu_l$. Fig. 5 gives an example of showing the different hypergraph similarity matrices on the ORL face dataset (referred to in Sec. 3).
factors such that
\[ \text{H} \]
where this measure, with its corresponding optimization procedure.

propose a discriminative hypergraph partitioning criterion based on manifold structure information among data samples at larger scales, order over-clustering hypergraph, we are capable of capturing the clustering.

\[ \text{K} \]
partitioning criterion, which aims to optimally partition the vertex

linearly combining the above three types of hypergraphs, a context-
aware hypergraph similarity matrix \( S = (s_{ij})_{N \times N} \) is obtained as follows:

\[
S = \mathcal{H} \left( \alpha \Sigma_p \ 0 \ 0 \ 0 \ 0 \ 0 \ (1 - \alpha - \beta) \Sigma_o \right) \mathcal{H}^T, \quad (11)
\]

where \( \mathcal{H} = (H_p, H_o, H_s) \) and \((\alpha, \beta)\) are the nonnegative weighting factors such that \((\alpha + \beta) \leq 1\). Encoding the local neighboring information, the kNN hypergraph plays a role in locally smoothing the clustering results (obtained by only using the pairwise hypergraph), as shown in the bottom-left part of Fig. 5. By constructing the high-order over-clustering hypergraph, we are capable of capturing the manifold structure information among data samples at larger scales, which leads to more accurate clustering results (shown in the bottom-right part of Fig. 5). Therefore, the final context-aware hypergraph similarity matrix keeps a balance among the three types of hypergraph information. In practice, it is easy to emphasize one particular type of hypergraph information by enlarging its associated weight. Fig. 2 (b) gives an example of illustrating the hypergraph clustering results based on the above hypergraph similarities.

2.2 Discriminative hypergraph partitioning for spectral clustering

Having defined a context aware vertex similarity measure, we now propose a discriminative hypergraph partitioning criterion based on this measure, with its corresponding optimization procedure.

1) Preliminaries of hypergraph partitioning. Hypergraph partitioning seeks an optimal hypergraph cut solution for effective data clustering. \( K \)-way normalized cut [12] is a well-known hypergraph partitioning criterion, which aims to optimally partition the vertex set \( \mathcal{V} \) into \( K \) disjoint subsets (i.e., \( \mathcal{V} = \bigcup_{k=1}^{K} \mathcal{V}_{k} \)) s.t. \( \forall \mathcal{V}_{m} \cap \mathcal{V}_{n} = \emptyset, \forall m \neq n \) by solving the following optimization problem:

\[
\text{max} \ f(X) = \frac{1}{K} \sum_{k=1}^{K} \frac{X_{i}^{T} S X_{i}}{D_{X_{i}}}, \quad (12)
\]

s.t. \( X \in \{0, 1\}^{N \times K}, \quad X 1_{K} = 1_{N} \), where \( X \) is an \( N \times K \) partition matrix such that \( X^{T} X \) is a diagonal matrix, \( 1_{d} \) denotes a \( d \times 1 \) vector with each element being 1, \( D \) is an \( N \times N \) diagonal matrix with the \( m \)-th diagonal element being the sum of the elements belonging to the \( m \)-th row of \( S \) for \( 1 \leq m \leq N \), and \( X_{n} \) is the \( n \)-th column of \( X \) for \( 1 \leq n \leq K \). As pointed out in [12], the optimization problem (12) is typically relaxed to:

\[
\text{max} \ h(Z) = \frac{1}{K} \text{tr}(Z^{T} S Z), \quad (13)
\]

s.t. \( Z^{T} D Z = I_{K} \), where \( I_{K} \) is a \( K \times K \) identity matrix, \( \text{tr}(\cdot) \) denotes the trace of a matrix, and \( Z = X (X^{T} D X)^{-\frac{1}{2}} \). Eq. (13) is a trace maximization problem and can be solved by generalized eigenvalue decomposition. To simultaneously capture both intra-cluster compactness and the inter-cluster separability among the vertices in a unified clustering framework, we propose a discriminative hypergraph partitioning criterion which can be formulated as a trace-ratio optimization problem.

2) Discriminative hypergraph partitioning criterion (DHPC). The proposed DHPC considers both the inter-cluster separability and the intra-cluster compactness, and thus aims to solve the following optimization problem:

\[
\text{max} \ g(X) = \frac{1}{K} \sum_{k=1}^{K} \frac{X_{i}^{T} S X_{i}}{Q_{X_{i}}}, \quad (14)
\]

s.t. \( X \in \{0, 1\}^{N \times K}, \quad X 1_{K} = 1_{N} \), where \( Q = D - S \). In the proposed DHPC, the intra-cluster compactness and the inter-cluster separability are respectively captured by \( X_{i}^{T} S X_{i} \) and \( X_{i}^{T} Q X_{i} \), which are formulated as:

\[
X_{i}^{T} S X_{i} = \sum_{i \in \mathcal{V}_{n} \cap \mathcal{V}_{j}} s_{ij}, \quad X_{i}^{T} Q X_{i} = \sum_{i \in \mathcal{V}_{n} \cap \mathcal{V}_{j}} s_{ij}, \quad (15)
\]

where \( \mathcal{V}_{n} \) denotes the vertex set belonging to the \( n \)-th cluster. Furthermore, the value of \( X_{i}^{T} S X_{i} \) more compact the intra-cluster samples. The smaller the value of \( X_{i}^{T} Q X_{i} \), the more separable the inter-cluster samples. As a result, an optimal hypergraph partitioning solution is obtained by maximizing \( g(X) \) in Eq. (14). For simplicity, let \( P_{n} \) denote the vertex-to-cluster membership vector associated with the \( n \)-th cluster such that \( P_{n} = X_{n} (X_{n}^{T} X_{n})^{-\frac{1}{2}} \), and \( P \) denote the vertex-to-cluster membership matrix that is a concatenation of all the vertex-to-cluster membership vectors such that \( P = (P_{1}, P_{2}, \ldots, P_{K}) = X (X^{T} X)^{-\frac{1}{2}} \). It can be shown that \( P \) is an orthogonal matrix:

\[
P^{T} P = \frac{X (X^{T} X)^{-\frac{1}{2}}}{\text{tr}(X^{T} X)} [X (X^{T} X)^{-\frac{1}{2}}] = (X^{T} X)^{-\frac{1}{2}} = I_{K}, \quad (16)
\]

where \( X^{T} X \) is a diagonal matrix. According to the conclusion in [12], we obtain \( \hat{X} = \text{Diag}(\hat{\text{diag}}(\hat{\cdot}) (P^{T} P)^{-\frac{1}{2}}) \) that is the corresponding inverse transform of \( P = X (X^{T} X)^{-\frac{1}{2}} \). Here, \( \hat{\text{diag}}(\cdot) \) denotes a diagonal matrix formed from its vector argument, and \( \hat{\text{tr}}(\cdot) \) represents a column vector formed from the diagonal elements of its matrix argument. Consequently, the optimization problem in Eq. (14) can be rewritten as:

\[
\text{max} \ g(X) = \frac{1}{K} \sum_{n=1}^{K} \frac{P_{n}^{T} S P_{n}}{P_{n}^{T} Q P_{n}} = \frac{1}{K} \sum_{n=1}^{K} \frac{\text{tr}(P_{n}^{T} S P_{n})}{\text{tr}(P_{n}^{T} Q P_{n})}, \quad (17)
\]

s.t. \( P^{T} P = I_{K} \).
The trace-ratio optimization problem (19) has been investigated as a partitioning solution and thus does not satisfy the discrete-solution requirements for data clustering. As a result, an iterative refining process is chosen as:

$$\max_P f(P) = \frac{1}{K} \sum_{n=1}^{K} tr(P_n^T S P_n) = \frac{1}{K} \sum_{n=1}^{K} tr(P_n^T S P_n),$$

subject to

$$tr(P_n^T P_n) = I_K.$$

Due to $tr(P_n^T S P_n) = tr(P_n^T S P_n)$ and $tr(P_n^T P_n) = tr(P_n^T Q P_n)$, the above optimization problem (18) can be reformulated as:

$$\max_P f(P) = \frac{1}{K} \sum_{n=1}^{K} tr(P_n^T S P_n),$$

subject to

$$tr(P_n^T P_n) = I_K.$$

The trace-ratio optimization problem (19) has been investigated in [29, 30, 33, 38]. In order to obtain an effective solution to Eq. (19), we therefore utilize the Newton-Lanczos algorithm [33] for trace-ratio maximization. The Newton-Lanczos algorithm includes the following three iterative steps:

1. Compute the trace ratio $\rho = \frac{tr(P_n^T S P_n)}{tr(P_n^T P_n)}$.
2. Run the Lanczos algorithm [37] to compute the $K$ largest eigenvalues of $S - \rho Q$ as well as their associated eigenvectors (\(P_1, P_2, \ldots, P_K\)) of $P$.
3. Repeat the above two steps until convergence.

In practice, the initial solution $P$ is chosen as the $K$ principal eigenvectors (i.e., corresponding to the $K$ largest eigenvalues) of the matrix $Q^{-1} S$. If $Q$ is a singular matrix, $Q^{-1}$ is replaced with the matrix $(Q + \epsilon I_N)^{-1} S$, where $I_N$ is an $N \times N$ identity matrix and $\epsilon$ is a small positive constant ($\epsilon = 10^{-6}$ in the experiments).

After solving the trace-ratio optimization problem (19), we obtain a candidate solution $\bar{X}$ to Eq. (14) as follows:

$$\bar{X} = \text{Diag}(\text{diag}(\frac{1}{2} (P P^T))) P.$$  

However, the candidate solution $\bar{X}$ is a real-valued hypergraph partitioning solution, and thus does not satisfy the discrete-solution requirements for data clustering. As a result, an iterative refining procedure [12] may be used to find the optimal discrete hypergraph partitioning solution $X$ to Eq. (14) (more details can be found in Steps four to eight of the algorithm in [12]). After combining the constructed pairwise+KNN+over-clustering hypergraphs (referred to in Sec. 2.1), we have a DHPC-based spectral clustering algorithm called PKO+HSC (pairwise+kNN+over-clustering hypergraph spectral clustering), as listed in Algorithm 1.
datasets are 1024 (ORL), 1024 (YaleB), 256 (USPS), and 784 (MNIST), respectively. In addition, the corresponding image features for the Corel dataset are the 960-dimensional GIST descriptors (as in [23]) that are widely used in computer vision and pattern recognition. The corresponding features for the trajectory dataset are 18-dimensional discrete Fourier transform (DFT) coefficient features (as in [24]). The feature dimension for the Iris dataset is 4, as shown in the UCI repository. Moreover, the kernel function \( G(z_i, z_j) \) (as in Eq. (3)) is selected as follows: \( G(z_i, z_j) = \exp (-\|z_i - z_j\|^2 / 2\sigma^2) \) where \( \sigma \) is a scaling factor. In practice, \( \sigma \) is tuned from the set \( \{y / y = 0.2y + (\lambda - 1)0.2y \} \) where \( \lambda \) is a positive integer such that \( \lambda \in \{1, 2, \ldots, 15\} \) and \( y \) is the average of the distances from each sample \( z_i \) to the other samples. The parameter \( k \) in the \( kNN \) hypergraph is set to 3. The weighting factors \((\alpha, \beta)\) in Eq. (11) are chosen from the set \( \{0.2, 0.2, 0.2, 0.2, 0.4, 0.2, 0.6, 0.4, 0.2, 0.4, 0.6, 0.2, 0.6, 0.4, 0.2, 0.0, 0.0, 0.2\} \). The task of constructing the high-order-over-clustering hypergraph can be accomplished by using a set of existing clustering methods. In our case, we take advantage of classic spectral clustering [10] or multi-class spectral clustering [12] to generate a set of vertex communities. For each over-clustering method, the number of the vertex communities is chosen as \( 2K \) with \( K \) being the desired number of clusters (referred to in Algorithm 1). Since we focus on the issues of iii) and iv) referred to in Sec. I, \( K \) is directly set as the ground truth number of clusters for each dataset. The above experimental configurations remain the same for all the experiments below.

**Computational complexity analysis**

Given \( N \) data samples, our pairwise hypergraph construction requires \( N^2 \) kernel computation operations (referred to in Eq. (1)). Accordingly, the \( kNN \) hypergraph construction needs to calculate \( N^2 \) cosine similarities (defined in Eq. (7)) with respect to the \( N \) data samples. Similarly, the over-clustering hypergraph construction involves \( N^2 \) cosine similarity computation operations (mentioned in Eq. (10)). The main computational cost of graph partitioning lies in the eigenvalue decomposition of \( S - \rho Q \) while solving the optimization problem (19) using the Newton-Lanczos algorithm. According to [37], the time complexity of the Lanczos iterations in our graph partitioning is \( \mathcal{O}(N^2) \). Therefore, the overall time complexity of our method is \( \mathcal{O}(N^2) \), which is the same as standard spectral clustering methods. For example, the average running time of our method on the USPS dataset (s.t. \( N = 9298 \)) is 15.21 seconds. The spatial complexity of our method lies in the four aspects: 1) the \( N \times N \) pairwise hypergraph incidence matrix \( H_p \); 2) the \( N \times N \) kNN hypergraph incidence matrix \( H_y \); 3) the \( N \times N \) over-clustering hypergraph incidence matrix \( H_c \); and 4) the final \( N \times N \) similarity matrix \( S \). It therefore has \( \mathcal{O}(N^2) \) overall spatial complexity.

### 3.2 Competing algorithms

We compare the proposed PKO+HSC with several representative spectral clustering algorithms. These spectral clustering algorithms are recently proposed, and have significant impacts on the data clustering community. For descriptive convenience, they are respectively referred to as CSC (classic spectral clustering [10]), STSC (self-tuning spectral clustering [11]), and NRSC (noise-robust spectral clustering [15]).

In order to verify the effect of different hypergraph components, we compare the proposed PKO+HSC with PK+HSC (pairwise+\( kNN \) hypergraph spectral clustering) and PO+HSC (pairwise+over-clustering hypergraph spectral clustering). Actually, PK+HSC and PO+HSC are special cases of the proposed PKO+HSC with different configurations of \((\alpha, \beta)\). In order to evaluate the performance of different graph partitioning criteria, we perform a comparison experiment against PKO+NC (our context-aware hypergraph similarity measure together with the normalized cut criterion [12]). Furthermore, in order to demonstrate the effectiveness of our context-aware hypergraph similarity measure (defined in Eq. (11)), we make a quantitative comparison with another similarity measure called RWDSM (Random Walk Diffusion Similarity Measure [32]). We put both RWDSM and our similarity measure into the same discriminative hypergraph partitioning criterion (DHPC) for data clustering.

### 3.3 Evaluation criteria

For a quantitative comparison, we introduce two evaluation criteria—NMI (normalized mutual information) and clustering accuracy. Specifically, the NMI criterion is defined as:

\[
\text{NMI}(X, Y) = I(X,Y) / \sqrt{H(X)H(Y)}
\]

where \( X \) and \( Y \) are two random variables, \( H(X) \) and \( H(Y) \) are the corresponding entropies of \( X \) and \( Y \), and \( I(X,Y) \) is the mutual information on \( X \) and \( Y \). In principle, NMI\((X, Y)\) has the range of \([0, 1]\), and is equal to 1 when \( X = Y \). As far as data clustering is concerned, the NMI criterion is explicitly formulated as:

---

**Fig. 8:** Clustering performances of the eight clustering algorithms regarding different random noise levels on the first four datasets. The first row corresponds to their clustering performances in NMI; and the second row is associated with their clustering performances in accuracy. The \((\alpha, \beta)\) configurations for (ORL, YaleB)\(\rightarrow\)(0.6, 0.2) and (USPS, MNIST)\(\rightarrow\)(0.4, 0.4).
Iris

TABLE 3: Clustering performances of the eight clustering algorithms in NMI and accuracy on the seven datasets by averaging different random noise levels. Specifically, the upper part corresponds to NMI, and the lower part is associated with accuracy. The \((\alpha, \beta)\) configurations for weighting different types of hypergraphs on the seven datasets are as follows: (OIRL, YaleB, Traj.) \rightarrow (0.6, 0.2) and (USPS, MNIST, Iris, Corel) \rightarrow (0.4, 0.4).

3.4 Clustering results

In the experiments, we aim to evaluate the clustering performances of different clustering algorithms in the following three aspects: i) evaluating the clustering performances (i.e., NMI and clustering accuracy) on the original datasets; ii) quantitative comparisons on the datasets after noise-based feature perturbations; and iii) performance evaluations on the datasets with outlier corruptions. The purposes of the above-mentioned three aspects are to verify the clustering effectiveness and the clustering robustness.

For i), Tab. 1 reports the corresponding NMI and accuracies of the eight clustering algorithms. It is seen from Tab. 1 that the proposed PKO+HSC obtains higher NMI and accuracies than the other clustering algorithms. More specifically, the average NMI gains of PKO+HSC regarding the seven datasets are (5.53%, 3.65%, 2.26%, 8.07%, 20.78%, 11.16%, 19.67%) over those of (PK+HSC, PO+HSC, PKO+NC, CSC, NRSC, RWDSM, STSC), respectively; and the average accuracy gains are (2.83%, 2.42%, 2.80%, 9.02%, 25.55%, 10.72%, 21.58%), respectively. Furthermore, Tab. 2 reports the NMI and accuracies of the proposed PKO+HSC with different configurations of the weighting factors \((\alpha, \beta)\). From Tab. 2, we see that the clustering performances of the proposed PKO+HSC are not very sensitive to the configurations of the weighting factors.

For ii), noise-based feature perturbations are performed by using additive random noises. Figs. 8 and 9 show the NMI and accuracy curves with error bars in eleven different noise levels (i.e., \(L_0\) corresponds to i), and \(L_1 \rightarrow L_{10}\) are associated with ten ascending noise levels whose magnitudes are chosen from \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}). Clearly, the proposed PKO+HSC achieves the highest NMI and accuracies in most noise levels. Furthermore, Tab. 3 reports the average NMI and accuracies of the eight clustering algorithms regarding different noise levels. The average NMI and accuracy gains of the proposed PKO+HSC are (6.03%, 5.19%, 2.80%, 23.54%, 34.54%, 26.92%, 20.43%) and (7.01%, 6.71%, 3.14%, 24.79%, 35.19%, 28.81%, 24.32%) over those of (PK+HSC, PO+HSC, PKO+NC, CSC, NRSC, RWDSM, STSC), respectively.
Moreover, Fig. 12 shows the clustering performances of the proposed PKO+HSC with different configurations of the nearest-neighbor number $k$ for $k$NN hypergraph construction in NMI and accuracy on the four datasets without corruption. The $(\alpha, \beta)$ configurations for weighting different types of hypergraphs on the four datasets are as follows: (ORL, YaleB.)$\rightarrow$(0.6, 0.2) and (USPS, MNIST)$\rightarrow$(0.4,0.4).

respectively.

For iii), outlier corruptions are performed by randomly setting the feature elements to zeros. Fig. 10 displays the clustering performances of the eight clustering algorithms regarding four outlier corruption levels (i.e., $L_0$ corresponds to i), and $L_1 \rightarrow L_3$ are associated with three ascending corruption levels whose corruption ratios are chosen from \{0.2, 0.4, 0.6\}). From Fig. 10, we see that the proposed PKO+HSC consistently achieves higher NMI and accuracies than the other clustering algorithms. Moreover, Tab. 4 shows the average NMIs and accuracies of the eight clustering algorithms regarding different outlier corruption levels. The average NMI and accuracy gains of the proposed PKO+HSC are (7.87%, 2.96%, 2.99%, 11.02%, 30.55%, 14.83%, 30.58%) and (7.39%, 3.80%, 2.91%, 10.11%, 34.62%, 14.04%, 33.66%) over those of (PK+HSC, PO+HSC, PKO+NC, CSC, NRSC, RWDSM, STSC), respectively.

Besides, we report the clustering results of the proposed PKO+HSC with different configurations of the nearest-neighbor number $k$ for $k$NN hypergraph construction in Fig. 11. From Fig. 11, we see that the proposed PKO+HSC is not very sensitive to the settings of $k$. Moreover, Fig. 12 shows the clustering performances of the proposed PKO+HSC using different numbers of vertex communities for high-order over-clustering hypergraph construction in NMI and clustering accuracy on the two datasets. It is clearly seen from Fig. 12 that the proposed PKO+HSC is not very sensitive to the choice of the vertex community number.

Overall, the proposed PKO+HSC outperforms the other clustering algorithms. Considering three types of hypergraph information (i.e., pairwise $k$NN, and over-clustering), the proposed PKO+HSC is capable of effectively exploring the intrinsic topological information among vertices. By optimizing the discriminative hypergraph partitioning criterion (DHPC), the proposed PKO+HSC considers both intra-cluster compactness and inter-cluster separability, resulting in the overall clustering robustness.

4 Conclusion and Future Work

In this work, we have proposed a context-aware hypergraph similarity measure (CAHSM), which is based on three types of hypergraphs—pairwise hypergraph, $k$-nearest-neighbor ($k$NN) hypergraph, and high-order over-clustering hypergraph. These hypergraphs capture the pairwise, neighborhood, and local grouping information on vertices. By effectively combining these types of affinity information, CAHSM is capable of effectively exploring the intrinsic structural information on vertices, resulting in the robust clustering performance. In order to fully capture the intra-cluster compactness and the inter-cluster separability of vertices, we have also designed a discriminative hypergraph partitioning criterion (DHPC) that is solved by trace-ratio maximization. Based on both CAHSM and DHPC, a robust spectral clustering algorithm (referred to as PKO+HSC) is developed for data clustering. Experimental results on various datasets, with
TABLE 4: Clustering performances of the eight clustering algorithms by averaging different outlier corruption levels. Specifically, the upper part corresponds to NMI and the lower part is associated with accuracy. The ($\alpha, \beta$) configurations for weight different types of hypergraphs on the four datasets are as follows: (ORL, YaleB) $\rightarrow (0.6, 0.2)$ and (USPS, MNIST) $\rightarrow (0.4, 0.4)$, and without noisy perturbation and outlier corruption, demonstrate that the proposed PKO+HSC has higher clustering robustness and effectiveness than competing algorithms in most cases.

On the other hand, this work is likely to have two limitations: 1) it is incapable of adaptively combining the aforementioned three types of hypergraphs; and 2) the number of clusters in spectral clustering is required to be provided in advance. Therefore, our future work is to figure out an adaptive weighting mechanism for hypergraph combination and an effective scheme for automatically estimating the cluster number prior to spectral clustering.

ACKNOWLEDGEMENTS

This work is partially supported by ARC Discovery Project (DP1104764), NSFC (Grant No. 60935002), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), ZZ is supported in part by US NSF (IIS-0812114, CCF-1017828), National Basic Research Program of China (2012CB316400), and Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis.

REFERENCES


Weiming Hu received the Ph.D. degree from the department of computer science and engineering at the National University of Singapore in 1998. From April 1998 to March 2000, he was a postdoctoral research fellow with the Institute of Computer Science and Technology, Peking University. Now he is a professor in the Institute of Automation, Chinese Academy of Sciences. His research interests are in visual surveillance, and filtering of Internet objectionable information.

Chunhua Shen is a senior lecturer at School of Computer Science, The University of Adelaide. Before that, he was with the computer vision program at NICTA (National ICT Australia), Canberra Research Laboratory for about six years. His research interests are in the intersection of computer vision and statistical machine learning. Recent work has been on real-time object detection, large-scale image retrieval and classification, and scalable nonlinear optimization. He studied at Nanjing University, at Australian National University, and received his PhD degree from The University of Adelaide. From 2012 to 2016, he holds an Australian Research Council Future Fellowship.

Anthony Dick received the PhD degree in 2002 from the University of Cambridge, UK, where he worked on problems in 3D reconstruction of architecture from images. He is currently a Senior Lecturer at the University of Adelaide, Australia. His research interests include image based modeling, automated video surveillance, image search and the intersection of computer vision and graphics.

Xi Li is currently a senior researcher at the University of Adelaide, Australia. From 2009 to 2010, he worked as a postdoctoral researcher at CNRS Telecom ParisTech, France. In 2009, he got the doctoral degree from National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. His research interests include visual tracking, motion analysis, face recognition, web data mining, image and video retrieval.

Chunhua Shen is a senior lecturer at School of Computer Science, The University of Adelaide. Before that, he was with the computer vision program at NICTA (National ICT Australia), Canberra Research Laboratory for about six years. His research interests are in the intersection of computer vision and statistical machine learning. Recent work has been on real-time object detection, large-scale image retrieval and classification, and scalable nonlinear optimization. He studied at Nanjing University, at Australian National University, and received his PhD degree from The University of Adelaide. From 2012 to 2016, he holds an Australian Research Council Future Fellowship.
Zhongfei Zhang is a professor at the Computer Science Department, State University of New York (SUNY) at Binghamton, USA, and the director of the Multimedia Research Laboratory in the Department. He received a B.S. in Electronics Engineering (with Honors), an M.S. in Information Sciences, both from Zhejiang University, China, and a PhD in Computer Science from the University of Massachusetts at Amherst, USA. He was on the faculty of Computer Science and Engineering Department, and a research scientist at the Center of Excellence for Document Analysis and Recognition, both at SUNY Buffalo, before he joined the faculty of computer science at SUNY Binghamton.