An Optimal Transport View for Subspace Clustering and Spectral Clustering

Yuguang Yan[1], Zhihao Xu[1], Canlin Yang[1], Jie Zhang[2], Ruichu Cai[1,3*], Michael Kwok-Po Ng[4]

[1]School of Computer Science, Guangdong University of Technology, Guangzhou, China
[2]Department of Mathematics, The University of Hong Kong, Hong Kong, China
[3]Peng Cheng Laboratory, Shenzhen, China
[4]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
ygyan@gdut.edu.cn, xzhsqzr987@outlook.com, {yangcl0608,cairuichu}@gmail.com,
zj199607@connect.hku.hk, michael-ng@hkbu.edu.hk

Abstract

Clustering is one of the most fundamental problems in machine learning and data mining, and many algorithms have been proposed in the past decades. Among them, subspace clustering and spectral clustering are the most famous approaches. In this paper, we provide an explanation for subspace clustering and spectral clustering from the perspective of optimal transport. Optimal transport studies how to move samples from one distribution to another distribution with minimal transport cost, and has shown a powerful ability to extract geometric information. By considering a self optimal transport model with only one group of samples, we observe that both subspace clustering and spectral clustering can be explained in the framework of optimal transport, and the optimal transport matrix bridges the spaces of features and spectral embeddings. Inspired by this connection, we propose a spectral optimal transport barycenter model, which learns spectral embeddings by solving a barycenter problem equipped with an optimal transport discrepancy and guidance of data. Based on our proposed model, we take advantage of optimal transport to exploit both feature and metric information involved in data for learning coupled spectral embeddings and affinity matrix in a unified model. We develop an alternating optimization algorithm to solve the resultant problems, and conduct experiments in different settings to evaluate the performance of our proposed methods.

Introduction

Clustering aims to partition data samples into different clusters so that similar samples are grouped together (Xu and Wunsch 2005). As one of the most fundamental problems in machine learning and data mining, clustering has been applied in many real-world applications, from image clustering (Yang et al. 2010) to text analysis (Liu et al. 2015), and many clustering methods have been proposed in the literature. Among them, subspace clustering and spectral clustering are important approaches being studied extensively (Nie et al. 2016; Bai and Liang 2020; Wang et al. 2022).

Over the past decades, spectral clustering and subspace clustering methods have been widely investigated due to their promising performance. Spectral clustering first constructs an affinity matrix to capture the similarities of samples, and then adopts the eigenvectors with the first k smallest eigenvalues of the Laplacian matrix as the new representations of data (Ng, Jordan, and Weiss 2001). After that, K-means is conducted on the new representations to partition samples into clusters. Usually, the affinity matrix is based on the Gaussian kernel or learned from data according to the pairwise distance (Bai and Liang 2020). For instance, to learn a better affinity matrix, subspace clustering assumes that samples approximately lie in linear subspaces. Based on this, a coefficient matrix is learned by considering self-expressiveness, *i.e.*, each sample can be represented as a linear combination of other samples in the same subspace, and the affinity matrix is constructed based on the coefficient matrix (Lu et al. 2018). Under this paradigm, different regularizers are applied to the learning model of the coefficient matrix (Lu et al. 2012; Elhamifar and Vidal 2013; Liu, Lin, and Yu 2010; Lu et al. 2018).

In this paper, we provide an explanation for subspace clustering and spectral clustering from the perspective of optimal transport, which studies how to move a set of samples from one distribution to another with the minimal transport cost (Peyré and Cuturi 2017). Optimal transport is first proposed by Monge in (Monge 1781) and then extended by Kantorovich in (Kantorovitch 1958). Optimal transport develops a powerful computational tool to capture geometric information involved in data (Benamou et al. 2015; An et al. 2022) and has been widely applied for estimating the discrepancy between two probability distributions (Courty, Flamary, and Tuia 2014; Courty et al. 2017b,a; Yan et al. 2018; Flamary et al. 2018). However, the connection between optimal transport and clustering has not been well investigated in the literature.

To fill this gap, we consider a self optimal transport model involving only one group of samples, and present the connection from optimal transport to subspace clustering and spectral clustering. We find that both subspace clustering and spectral clustering can be explained in the framework of optimal transport. In specific, the optimal transport matrix bridges the original feature space and the spectral embedding space, and spectral clustering can be interpreted as finding spectral embeddings in the Stiefel manifold with a consistent affinity matrix to the original features.

Based on this connection, we model spectral clustering as a special optimal transport barycenter problem, and pro-

*Corresponding author.

pose a model named Spectral Optimal Transport bArycenter (SOTA), which is a unified model to learn coupled spectral embeddings and the affinity matrix. On the strength of optimal transport, we propose to exploit both feature and metric information of data for spectral clustering, which is achieved by leveraging the Wasserstein and Gromov-Wasserstein discrepancies in the theory of optimal transport. In specific, we develop two spectral clustering algorithms, one considers a fixed marginal distribution for samples, and the other one considers an unfixed marginal distribution with an ability to recognize outliers during clustering. We adopt an alternating strategy to optimize the resultant problems, and analyze the convergence property of the algorithms. Experiments are conducted on simulation data and benchmark datasets to evaluate the performance of our proposed methods.

We summarize our major contributions as follows:

- We provide optimal transport explanations for subspace clustering and spectral clustering by considering a self optimal transport model.
- We model spectral clustering as an optimal transport barycenter problem equipped with an underlying optimal transport discrepancy.
- We develop spectral clustering algorithms based on our proposed model considering both feature and metric information involved in data.

From Optimal Transport to Clustering

In this section, we first present the notations used in the paper, and then show the connection between optimal transport and clustering.

Notations

In clustering, we are given a data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, where n is the number of samples, and d is the number of features. $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th samples and also the transpose of the i-th row of $\mathbf{X}$. Clustering aims to partition these samples into k clusters to minimize distances between samples within clusters.

Throughout the paper, $[n]$ denotes a set including the elements $\{1,\dots,n\}$. $\mathbf{1}_n$ denotes a vector in the space $\mathbb{R}^n$ with all the elements being 1. For a matrix $\mathbf{A}$, the (i,j)-th element of $\mathbf{A}$ is denoted as A_{ij}, and $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$. The trace of a square matrix $\mathbf{A}$ is defined as $\text{tr}(\mathbf{A}) = \sum_i A_{ii}$. Given two matrices $\mathbf{A}$ and $\mathbf{B}$ with the same size, the inner product of them is defined as

$$\langle \mathbf{A}, \mathbf{B}\rangle = \sum_i \sum_j A_{ij}B_{ij} = \text{tr}(\mathbf{A}^\top\mathbf{B}) = \text{tr}(\mathbf{A}\mathbf{B}^\top). \quad (1)$$

For a vector $\mathbf{v}$, $\text{diag}(\mathbf{v})$ represents a diagonal matrix with the diagonal elements being $\mathbf{v}$.

Entropic Optimal Transport

Optimal transport studies how to transport mass from a group of samples to another group with the minimal cost (Peyré and Cuturi 2017), where the minimal transport cost can be used to measure the distribution discrepancy between these two groups. In specific, given two groups of samples $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\tilde{\mathbf{x}}_j\}_{j=1}^m$, the corresponding empirical distributions are $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\mu}}$ with the simplex constraint, *i.e.*, $\boldsymbol{\mu} \in \Sigma_n$, $\tilde{\boldsymbol{\mu}} \in \Sigma_m$, where $\Sigma_n = \{\mathbf{v} \in \mathbb{R}^n \mid v_i \geq 0 \quad \forall i \in [n], \|\mathbf{v}\|_1 = 1\}$. A transport plan is denoted as a matrix $\mathbf{T}$, which is also a joint distribution of $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\mu}}$. The domain of the definition of $\mathbf{T}$ is given as follows:

$$\mathcal{T}(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}}) = \{\mathbf{T} \in (\mathbb{R}^+)^{n\times m} \mid \mathbf{T}\mathbf{1}_m = \boldsymbol{\mu}, \mathbf{T}^\top\mathbf{1}_n = \tilde{\boldsymbol{\mu}}\}, \quad (2)$$

where T_{ij} indicates how many masses are transported from $\mathbf{x}_i$ to $\tilde{\mathbf{x}}_j$. For the transport cost from $\mathbf{x}_i$ to $\tilde{\mathbf{x}}_j$, the squared Euclidean distance is commonly used in the existing works:

$$C_{ij}^X = \|\mathbf{x}_i - \tilde{\mathbf{x}}_j\|_2^2. \quad (3)$$

Based on this, the squared Wasserstein distance $\mathcal{W}_2^2$ (Courty et al. 2017b) can be defined as

$$\mathcal{W}_2^2(\mathbf{C}^X, \boldsymbol{\mu}, \tilde{\boldsymbol{\mu}}) = \min_{\mathbf{T}\in\mathcal{T}(\boldsymbol{\mu},\tilde{\boldsymbol{\mu}})} \langle \mathbf{C}^X, \mathbf{T}\rangle. \quad (4)$$

In order to obtain a smooth solution and speed up the optimization, (Cuturi 2013) introduces a negative entropic regularization on $\mathbf{T}$ as

$$\Omega(\mathbf{T}) = \sum_{i=1}^n \sum_{j=1}^m T_{ij}\log T_{ij} - T_{ij}, \quad (5)$$

which induces an entropic Wasserstein discrepancy

$$\mathcal{W}_\epsilon(\mathbf{C}^X, \boldsymbol{\mu}, \tilde{\boldsymbol{\mu}}) = \min_{\mathbf{T}\in\mathcal{T}(\boldsymbol{\mu},\tilde{\boldsymbol{\mu}})} \langle \mathbf{C}^X, \mathbf{T}\rangle + \epsilon\Omega(\mathbf{T}), \quad (6)$$

which is an approximation to the squared Wasserstein distance $\mathcal{W}_2^2$ (Peyré and Cuturi 2017).

Self Entropic Optimal Transport

In this part, we present a self entropic optimal transport model to move a group of samples to itself. This model is used in the following discussions regarding the connection between optimal transport and clustering. In specific, our self entropic optimal transport considers one group of samples, and aims to move samples to samples except themselves with the minimal transport cost, which is formalized as follows:

$$\begin{aligned}\mathcal{W}_\epsilon^s(\mathbf{C}^X, \boldsymbol{\mu}, \boldsymbol{\mu}) = \min_{\mathbf{T}} \ & \langle \mathbf{C}^X, \mathbf{T}\rangle + \epsilon\Omega(\mathbf{T}) \\ \text{s.t.} \quad & \mathbf{T} \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\mu}), \ T_{ii} = 0 \ \forall i \in [n]. \end{aligned} \quad (7)$$

It is worth mentioning that self transport here refers to transport a group to itself instead of transporting a sample to itself. We name $\mathcal{W}_\epsilon^s(\mathbf{C}^X, \boldsymbol{\mu}, \boldsymbol{\mu})$ as self entropic Wasserstein discrepancy. Actually, the optimal transport matrix $\mathbf{T}$ reflects the similarity relationships between samples. If two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are far away from each other, the transport cost C_{ij}^X between them will be large, resulting in a small transport mass T_{ij}. As a result, most of the mass will be transported between two samples with a small distance, *i.e.*, large similarity. In this sense, the total transport cost $\mathcal{W}_\epsilon^s(\mathbf{C}^X, \boldsymbol{\mu}, \boldsymbol{\mu})$ reflects the density degree of the samples.

Now we briefly discuss how to solve Problem (7) to obtain $\mathbf{T}$. First, without considering the constraints $T_{ii} = 0$,

based on the Sinkhorn algorithm in (Cuturi 2013), we can construct a matrix

$$\mathbf{K} = \exp(-\mathbf{C}^X/\epsilon), \tag{8}$$

and the optimal transport matrix $\mathbf{T}$ can be calculated as

$$\mathbf{T} = \text{diag}(\mathbf{u})\mathbf{K}\text{diag}(\mathbf{u}), \tag{9}$$

where $\mathbf{u}$ can be obtained by the Sinkhorn-Knopp algorithm. The optimization details can be found in (Cuturi 2013).

The constraints $T_{ii} = 0$ are non-trivial to address. In practice, we can punish the transport cost from a sample to itself by assigning a large value to the diagonal elements C^X_{ii}, By doing this, the diagonal elements T_{ii} will be close to 0.

It is obvious that the matrix in Eq. (8) is exactly the Gaussian kernel of the samples in $\mathbf{X}$, and Eq. (9) is a normalization operation to make the sum of the i-th row (or i-th column, since $\mathbf{T}$ is symmetric here) be μ_i. Therefore, $\mathbf{T}$ learned by the self entropic optimal transport model (7) is a normalized affinity matrix that can be used in spectral clustering.

A similar result has also been presented in (Landa, Coifman, and Kluger 2021), which constructs a kernel matrix similar to Eq. (8) except that the diagonal elements being 0, equivalent to our trick that assigns a large value to C^X_{ii}. Eq. (9) is similar to doubly stochastic normalization of the Gaussian kernel with zero main diagonal in (Landa, Coifman, and Kluger 2021).

Connection with Subspace Clustering

Now we show the connections between optimal transport with two versions of subspace clustering, respectively.

Subspace clustering aims to learn an affinity matrix $\mathbf{S}$ for data reconstruction, which is formalized as follows:

$$\begin{aligned} &\min_{\mathbf{S}} \ \ell(\mathbf{X}, \mathbf{SX}) + \gamma\mathcal{R}(\mathbf{S}) \\ &\text{s.t. } \mathbf{S} = \mathbf{S}^\top, \mathbf{S} \geq 0, \text{diag}(\mathbf{S}) = \mathbf{0}, \end{aligned} \tag{10}$$

where γ is the trade-off parameter, $\mathcal{R}(\mathbf{S})$ is a regularization term that can be set as a low-rank term, or the negative entropy in Eq. (5), as shown in (Bai and Liang 2020). Next, we consider two implementations for the loss function $\ell(\mathbf{X}, \mathbf{SX})$, which is used for data reconstruction.

Squared Loss Function For the loss function $\ell(\mathbf{X}, \mathbf{SX})$, one common choice is the self-expressive loss, which is defined as

$$\ell_{se}(\mathbf{X}, \mathbf{SX}) = \|\mathbf{X} - \mathbf{SX}\|_F^2. \tag{11}$$

We show that this loss function can be derived based on the Wasserstein barycenter defined in optimal transport.

According to the above subsection, $\mathbf{T}$ obtained by the self optimal tranport model in Problem (7) is a normalized affinity matrix. To apply $\mathbf{T}$ to reconstruct $\mathbf{X}$, we use $\mathbf{\Lambda} = \text{diag}(\boldsymbol{\mu})^{-1}$ to scale the transport matrix $\mathbf{T}$, which makes the sum of each row of $\mathbf{\Lambda T}$ be 1. As a result, Eq. (11) based on $\mathbf{T}$ can be rewritten as $\|\mathbf{X} - \mathbf{\Lambda TX}\|_F^2$.

Inspired by the discussion in (Courty et al. 2017b), for two distributions, based on the optimal transport matrix $\mathbf{T}$, we can move samples from one distribution (*i.e.*, source distribution) to the other distribution (*i.e.*, target distribution). These transported samples can be constructed based on Wasserstein barycenter and follow a similar distribution to the target distribution (Courty et al. 2017b). Based on the squared Euclidean distance defined in Eq. (3), the transported samples can be obtained by solving the following barycentric mapping

$$\hat{\mathbf{x}}_i = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \sum_j T_{ij}\|\mathbf{x} - \mathbf{x}_j\|_2^2, \tag{12}$$

and the closed-form solution is

$$\hat{\mathbf{X}} = \mathbf{\Lambda TX}. \tag{13}$$

Remind that the transport in Problem (7) is performed from the samples $\mathbf{X}$ to $\mathbf{X}$, *i.e.*, data of both source and target distributions are $\mathbf{X}$. Therefore, the transported samples $\hat{\mathbf{X}}$ follows the similar distribution of $\mathbf{X}$, which means that $\hat{\mathbf{X}}$ is expected to be close to $\mathbf{X}$, *i.e.*, $\|\mathbf{X} - \hat{\mathbf{X}}\|_F^2 = \|\mathbf{X} - \mathbf{\Lambda TX}\|_F^2$ is approximately minimized.

Linear Loss Function Another implementation of $\ell(\mathbf{X}, \mathbf{SX})$ is proposed in (Bai and Liang 2020), in which $\ell(\mathbf{X}, \mathbf{SX})$ is assumed to be a linear function of $\mathbf{S}$, *i.e.*,

$$\ell(\mathbf{X}, \mathbf{SX}) = \sum_{i=1}^n \sum_{j=1}^n A_{ij}S_{ij} + B_{ij}. \tag{14}$$

Based on the self-expressive condition $\mathbf{X} = \mathbf{SX}$ and the constraints $\mathbf{S1}_n = \mathbf{1}_n$, (Bai and Liang 2020) presents that the objective function of subspace clustering in Problem (10) with the loss function in Eq. (14) and an entropic term on $\mathbf{S}$ can be rewritten as

$$\sum_{i=1}^n \sum_{j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 S_{ij} + \gamma \sum_{i=1}^n \sum_{j=1}^n S_{ij}\log S_{ij}, \tag{15}$$

which is almost the objective function of self entropic optimal transport in Problem (7), except that $\mathbf{S}$ is a scaled $\mathbf{T}$. When $\boldsymbol{\mu}$ is a uniform distribution, we have $\mathbf{\Lambda T} = n\mathbf{T}$, and the objective functions in Eqs. (7) and (15) are equivalent.

In conclusion, the optimal transport matrix $\mathbf{T}$ in Problem (7) can be used as an affinity matrix, which can be used for spectral clustering.

Connection with Spectral Clustering

Given a symmetric affinity matrix $\mathbf{S}$ and its Laplacian matrix $\mathbf{L} = \text{diag}(\mathbf{S1}) - \mathbf{S}$, spectral clustering aims to learn a spectral embedding $\mathbf{H}$ by solving the following problem

$$\min_{\mathbf{H}\in\mathcal{M}_k^n} \text{tr}(\mathbf{H}^\top\mathbf{LH}), \tag{16}$$

where $\mathcal{M}_k^n$ is a Stiefel manifold defined as

$$\mathcal{M}_k^n = \{\mathbf{H} \in \mathbb{R}^{n\times k} \mid \mathbf{H}^\top\mathbf{H} = \mathbf{I}\}. \tag{17}$$

A common choice of $\mathbf{S}$ is the Gaussian kernel, *i.e.*,

$$S_{ij} = \exp\Big(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/\sigma\Big). \tag{18}$$

According to Eqs. (8) and (9), the Gaussian kernel used in spectral clustering is the solution to our self entropic optimal transport model in Problem (7).

Besides, given the affinity matrix $\mathbf{T}$ obtained by Problem (7) and its Laplacian matrix $\mathbf{L} = \text{diag}(\mathbf{T1}) - \mathbf{T}$, the objective function of spectral clustering can be rewritten as

$$\text{tr}(\mathbf{H}^\top \mathbf{L}\mathbf{H}) = \frac{1}{2}\sum_{ij} T_{ij}\|\mathbf{h}_i - \mathbf{h}_j\|_2^2, = \frac{1}{2}\langle \mathbf{C}^H, \mathbf{T}\rangle, \quad (19)$$

where $\mathbf{C}^H$ is the squared Euclidean distance matrix with $C_{ij}^H = \|\mathbf{h}_i - \mathbf{h}_j\|_2^2$, $\mathbf{h}_i$ is the spectral embedding of $\mathbf{x}_i$ and also the transpose of the i-th row in $\mathbf{H}$. From this model, spectral clustering is to find spectral embeddings $\mathbf{H} \in \mathcal{M}_k^n$ which share the affinity matrix $\mathbf{T}$ with $\mathbf{X}$. In other words, the similarity relationship is modeled as a transport matrix $\mathbf{T}$ learned from features $\mathbf{X}$, and then applied in the Stiefel manifold $\mathcal{M}_k^n$ to learn spectral embeddings $\mathbf{H}$. This observation inspires us to learn the affinity matrix and spectral embeddings simultaneously, as shown in (Wang et al. 2022), while in which the optimal transport explanation has not been established.

Now we provide an explanation for spectral clustering from the view of Wasserstein barycenter. According to (Cuturi and Doucet 2014), a Wasserstein barycenter of multiple distributions is a group of samples $\{\hat{\mathbf{x}}_i\}$ and its corresponding empirical probability distribution $\hat{\boldsymbol{\mu}}$ whose average Wasserstein distance to the distributions is minimized. Here, we consider a special barycenter problem considering self entropic Wasserstein discrepancy $\mathcal{W}_\epsilon^s$ in Eq. (7) in the spectral embedding space $\mathcal{M}_k^n$ with fixed distribution $\boldsymbol{\mu}$, which is modeled as follows:

$$\min_{\mathbf{H}\in\mathcal{M}_k^n} \mathcal{W}_\epsilon^s(\mathbf{C}^H, \boldsymbol{\mu}, \boldsymbol{\mu}). \quad (20)$$

This model aims to find $\mathbf{H} \in \mathcal{M}_k^n$ such that total transport cost is minimized by an optimal plan $\mathbf{T}$ found in $\mathcal{W}_\epsilon^s(\mathbf{C}^H, \boldsymbol{\mu}, \boldsymbol{\mu})$, in which the transport cost between two samples is measured by their squared Euclidean distance.

Different from the model in Eq. (19) where the affinity matrix $\mathbf{T}$ is learned from $\mathbf{X}$ and applied to learn $\mathbf{H}$, Problem (20) is lack of information since the features $\mathbf{X}$ are not involved. In order to obtain meaningful spectral embeddings for clustering, information from data $\mathbf{X}$ should be introduced. In the next section, we enrich this model by considering $\mathbf{X}$, and propose our spectral optimal transport barycenter model to learn $\mathbf{H}$ for spectral clustering.

Spectral Optimal Transport Barycenter

To learn an effective $\mathbf{H}$ for clustering with the data $\mathbf{X}$, we introduce $\mathbf{X}$ to construct the cost matrix $\mathbf{C}^{X,H}$ relying on both $\mathbf{X}$ and $\mathbf{H}$, which will be implemented in the next section, and propose the following model called spectral Wasserstein barycenter:

$$\min_{\mathbf{H}\in\mathcal{M}_k^n} \mathcal{W}_\epsilon^s(\mathbf{C}^{X,H}, \boldsymbol{\mu}, \boldsymbol{\mu}), \quad (21)$$

which finds $\mathbf{H} \in \mathcal{M}_k^n$ with guidance from $\mathbf{X}$ such that $\mathcal{W}_\epsilon^s(\mathbf{C}^{X,H}, \boldsymbol{\mu}, \boldsymbol{\mu})$ considering both original features $\mathbf{X}$ and spectral embeddings $\mathbf{H}$ is minimized. The solution to this model can be interpreted as a barycenter with the fixed probability distribution $\boldsymbol{\mu}$ and the underlying optimal transport discrepancy $\mathcal{W}_\epsilon^s$, and at the same time, satisfying the constraints that a part of representations is given as $\mathbf{X}$ in the original feature space $\mathbb{R}^{n\times d}$ while the other part is found in the spectral embedding space $\mathcal{M}_k^n$. The transport matrix $\mathbf{T}$ is the bridge between these two spaces and optimized in the inner optimal transport problem, as shown in Figure 1.

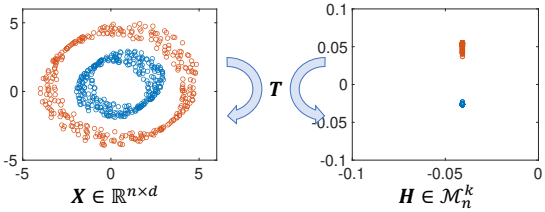


Figure 1: Illustration of the connection between optimal transport and spectral clustering. The self transport matrix $\mathbf{T}$ bridges the feature space and spectral embedding space. Different colors indicate different clusters.

Based on the above connection between optimal transport and spectral clustering, we propose a generalized learning problem from the theory of optimal transport for spectral clustering. Remind that the barycenter problem in (21) is based on the entropic Wasserstein discrepancy in Eq. (6), which approximates the squared Wasserstein distance in Eq. (4). By generalizing this underlying metric to a general optimal transport discrepancy, we propose the following optimal transport model for spectral clustering:

$$\min_{\mathbf{H}\in\mathcal{M}_k^n} \mathcal{OTD}^s(\mathbf{X}, \mathbf{H}, \boldsymbol{\mu}), \quad (22)$$

where $\mathcal{OTD}^s(\mathbf{X}, \mathbf{H}, \boldsymbol{\mu})$ is a self optimal transport discrepancy with the constraint that the involving transport matrix $\mathbf{T}$ has zero diagonal elements. $\mathcal{OTD}^s(\mathbf{X}, \mathbf{H}, \boldsymbol{\mu})$ involves the original features $\mathbf{X}$, the corresponding probability distribution $\boldsymbol{\mu}$ and the spectral embeddings $\mathbf{H}$. This generalized model allows us to implement it by adopting a discrepancy in the theory of optimal transport, such as the entropic Wasserstein discrepancy or the entropic Gromov-Wasserstein discrepancy (Peyré, Cuturi, and Solomon 2016). When adopting the discrepancy defined in Eq. (6) considering both $\mathbf{X}$ and $\mathbf{H}$, Problem (22) is implemented as Problem (21).

In addition, for a barycenter problem, the probability distribution $\boldsymbol{\mu}$ can also be learnable (Cuturi and Doucet 2014). $\boldsymbol{\mu}$ indicates the masses each sample has, and can be taken as a weight vector reflecting the contribution of each sample in a specific application (Yan et al. 2019). By optimizing both spectral embeddings $\mathbf{H}$ and the weights $\boldsymbol{\mu}$, we further propose the following model named Spectral Optimal Transport bArycenter (SOTA):

$$\min_{\mathbf{H}\in\mathcal{M}_k^n, \boldsymbol{\mu}\in\Sigma_n} \mathcal{OTD}^s(\mathbf{X}, \mathbf{H}, \boldsymbol{\mu}). \quad (23)$$

On the strength of optimal transport, in the following, we consider one implementation of this model by incorporating structure information in the metric space.

Implementation Based on Metric Information

Besides transport in the feature space, we also exploit structure information in the metric space to learn the affinity matrix $\mathbf{T}$ and spectral embeddings $\mathbf{H}$. This is achieved by further considering the Gromov-Wasserstein discrepancy, which considers metrics $\mathbf{M}$ and $\tilde{\mathbf{M}}$ (similarity or distance) constructed on two groups of samples $\mathbf{X}$ and $\tilde{\mathbf{X}}$ with probability distributions $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\mu}}$, respectively. And the transport is conducted from a pair $(\mathbf{x}_i, \mathbf{x}_{i'})$ to a pair $(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_{j'})$ inducing a 4-order transport cost tensor

$$\mathbf{L}(\mathbf{M}, \tilde{\mathbf{M}}) = \frac{1}{2}(M_{ii'} - \tilde{M}_{jj'})^2_{ii'jj'}, \tag{24}$$

and the transport cost from $\mathbf{x}_i$ to $\tilde{\mathbf{x}}_j$ is calculated by the following tensor-matrix multiplication (Peyré, Cuturi, and Solomon 2016)

$$\mathbf{L}(\mathbf{M}, \tilde{\mathbf{M}}) \otimes \mathbf{T} = \big(\sum_{i'j'} L_{ii'jj'} T_{i'j'}\big)_{ij}. \tag{25}$$

Based on this, the entropic fused Gromov-Wasserstein discrepancy considering both Wasserstein and Gromov-Wasserstein is defined as (Titouan et al. 2019)

$$\begin{aligned}
&\mathcal{FGW}_\epsilon(\mathbf{C}, \mathbf{M}, \tilde{\mathbf{M}}, \boldsymbol{\mu}, \tilde{\boldsymbol{\mu}}) \\
&= \min_{\mathbf{T} \in \mathcal{T}(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}})} \alpha\langle \mathbf{C}, \mathbf{T}\rangle + (\mathbf{1} - \alpha)\langle \mathbf{L}(\mathbf{M}, \tilde{\mathbf{M}}) \otimes \mathbf{T}, \mathbf{T}\rangle \\
&\qquad + \epsilon\Omega(\mathbf{T}),
\end{aligned} \tag{26}$$

where $\mathbf{C}$ is the transport cost matrix from $\mathbf{X}$ to $\tilde{\mathbf{X}}$ based on the squared Euclidean distance. $\mathcal{FGW}_\epsilon$ leverages metric information for learning the transport matrix $\mathbf{T}$ in both the feature and the metric spaces. Based on this, we define the $\mathcal{FGW}^s_\epsilon$ on $\mathbf{X}$ and $\mathbf{H}$ under the situation of self optimal transport as

$$\begin{aligned}
&\mathcal{FGW}^s_\epsilon(\mathbf{C}^{X,H}, \mathbf{M}^{X,H}, \mathbf{M}^{X,H}, \boldsymbol{\mu}, \boldsymbol{\mu}) \\
&= \min_{\mathbf{T}} \quad \alpha\langle \mathbf{C}^{X,H}, \mathbf{T}\rangle \\
&\qquad + (\mathbf{1} - \alpha)\langle \mathbf{L}(\mathbf{M}^{X,H}, \mathbf{M}^{X,H}) \otimes \mathbf{T}, \mathbf{T}\rangle + \epsilon\Omega(\mathbf{T}) \\
&\text{s.t.} \quad \mathbf{T} \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\mu}), \;\; T_{ii} = 0 \;\; \forall i \in [n],
\end{aligned} \tag{27}$$

which can be adopted as the objective function of Problem (23). As a result, the model in Problem (23) can be implemented based on $\mathcal{FGW}^s_\epsilon$ as follows:

$$\min_{\mathbf{H} \in \mathcal{M}^n_k, \boldsymbol{\mu} \in \Sigma_n} \mathcal{FGW}^s_\epsilon(\mathbf{C}^{X,H}, \mathbf{M}^{X,H}, \mathbf{M}^{X,H}, \boldsymbol{\mu}, \boldsymbol{\mu}), \tag{28}$$

which we called Spectral Fused Gromov-Wasserstein model (SFGW).

We also consider a simplified situation where the probability distribution $\boldsymbol{\mu}$ is fixed, *e.g.*, a uniform distribution. Formally, this situation is modeled as

$$\min_{\mathbf{H} \in \mathcal{M}^n_k} \mathcal{FGW}^s_\epsilon(\mathbf{C}^{X,H}, \mathbf{M}^{X,H}, \mathbf{M}^{X,H}, \boldsymbol{\mu}, \boldsymbol{\mu}), \tag{29}$$

which we called SFGWH since it learns $\mathbf{H}$ with a fixed $\boldsymbol{\mu}$.

In the following, we provide the implementation details for Problems (29) and (28). We first develop an algorithm for SFGWH with a uniform distribution $\boldsymbol{\mu}$, which can be viewed as a spectral clustering problem learning coupled $\mathbf{T}$ and $\mathbf{H}$. Next, we develop an algorithm for SFGW with a learnable $\boldsymbol{\mu}$, which adaptively assigns weights for samples and can be used to recognize outliers during spectral clustering.

Algorithms

In this section, we first rewrite Problems (28) and (29) with more computational details, and then provide the optimization algorithms for them.

In specific, we calculate $\mathbf{C}^X$ and $\mathbf{C}^H$ based on the squared Euclidean distance defined in Eq. (3), and then construct $\mathbf{C}^{X,H}$ as $\mathbf{C}^{X,H} = \mathbf{C}^X + \lambda\mathbf{C}^H$, where λ control the effects of $\mathbf{C}^X$ and $\mathbf{C}^H$. We also assign a sufficiently large value to the diagonal elements of $\mathbf{C}^X$ to make the diagonal elements in the affinity matrix $\mathbf{T}$ be close to zero.

To leverage metric information in $\mathbf{X}$, we construct the metric matrix $\mathbf{M}^{X,H}$ based on a similarity or distance metric on $\mathbf{M}^X$. Here we adopt cosine similarity as follows:

$$M^X_{ii'} = (\mathbf{x}_i^\top \mathbf{x}_{i'}) \big/ \big(\|\mathbf{x}_i\| \|\mathbf{x}_{i'}\|\big). \tag{30}$$

We do not consider a similar $\mathbf{M}^H$ here since it brings an extra challenge for solving $\mathbf{H}$. Based on $\mathbf{C}^X$, $\mathbf{C}^H$ and $\mathbf{M}^X$, we rewrite the SFGW model in Problem (28) as

$$\begin{aligned}
\min_{\mathbf{H}, \boldsymbol{\mu}, \mathbf{T}} \;& \langle \alpha\mathbf{C}^{\mathbf{X}} + \alpha\lambda\mathbf{C}^{\mathbf{H}}, \mathbf{T}\rangle \\
&+ (\mathbf{1} - \alpha)\langle \mathbf{L}(\mathbf{M}^X, \mathbf{M}^X) \otimes \mathbf{T}, \mathbf{T}\rangle + \epsilon\Omega(\mathbf{T}) \\
\text{s.t.} \;& \mathbf{H} \in \mathcal{M}^n_k, \boldsymbol{\mu} \in \Sigma_n, \mathbf{T} \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\mu}).
\end{aligned} \tag{31}$$

According to Proposition 1 in (Peyré, Cuturi, and Solomon 2016), we have

$$\begin{aligned}
&\langle \mathbf{L}(\mathbf{M}^X, \mathbf{M}^X) \otimes \mathbf{T}, \mathbf{T}\rangle \\
&= \big\langle (\mathbf{M}^X \odot \mathbf{M}^X)\mathbf{T}\mathbf{1}_n\mathbf{1}_n^\top + \mathbf{1}_n\mathbf{1}_n^\top\mathbf{T}^\top(\mathbf{M}^X \odot \mathbf{M}^X)^\top \\
&\qquad - 2\mathbf{M}^X\mathbf{T}(\mathbf{M}^X)^\top, \mathbf{T}\big\rangle.
\end{aligned} \tag{32}$$

For simplicity, we define

$$\begin{aligned}
\tilde{\mathbf{G}}^X =& (\mathbf{M}^X \odot \mathbf{M}^X)\mathbf{T}\mathbf{1}_n\mathbf{1}_n^\top + \mathbf{1}_n\mathbf{1}_n^\top\mathbf{T}^\top(\mathbf{M}^X \odot \mathbf{M}^X)^\top \\
&- 2\mathbf{M}^X\mathbf{T}(\mathbf{M}^X)^\top,
\end{aligned} \tag{33}$$

and achieve the following optimization problem

$$\begin{aligned}
\min_{\mathbf{H}, \boldsymbol{\mu}, \mathbf{T}} \;& \langle \lambda_X\mathbf{C}^X + \lambda_H\mathbf{C}^H + (\mathbf{1} - \lambda_X)\tilde{\mathbf{G}}^X, \mathbf{T}\rangle + \epsilon\Omega(\mathbf{T}) \\
\text{s.t.} \;& \mathbf{H} \in \mathcal{M}^n_k, \boldsymbol{\mu} \in \Sigma_n, \mathbf{T} \in \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\mu}), \mathbf{T} = \mathbf{T}^\top,
\end{aligned} \tag{34}$$

where $\lambda_X = \alpha$, $\lambda_H = \alpha\lambda$, and $\epsilon = 1$ in our experiments. The symmetric constraint is consider here since it is important for affinity learning. The symmetric constraint is easy to be satisfied in our optimization algorithm without an additional operation $\mathbf{T} := (\mathbf{T} + \mathbf{T}^\top)/2$, which is commonly used in subspace clustering.

Problem with Fixed $\boldsymbol{\mu}$

For the SFGWH model with a fixed $\boldsymbol{\mu}$, based on the condition $\mathbf{T}\mathbf{1}_n = \mathbf{T}^\top\mathbf{1}_n = \boldsymbol{\mu}$, we can simplify $\langle \tilde{\mathbf{G}}^X, \mathbf{T}\rangle$ as

$$\begin{aligned}
&\langle \tilde{\mathbf{G}}^X, \mathbf{T}\rangle \\
&= \big\langle (\mathbf{M}^X \odot \mathbf{M}^X)\mathbf{T}\mathbf{1}_n\mathbf{1}_n^\top + \mathbf{1}_n\mathbf{1}_n^\top\mathbf{T}^\top(\mathbf{M}^X \odot \mathbf{M}^X)^\top \\
&\qquad - 2\mathbf{M}^X\mathbf{T}(\mathbf{M}^X)^\top, \mathbf{T}\big\rangle \\
&= 2\mathrm{tr}\big((\mathbf{M}^X \odot \mathbf{M}^X)\boldsymbol{\mu}\boldsymbol{\mu}^\top\big) - 2\big\langle \mathbf{M}^X\mathbf{T}(\mathbf{M}^X)^\top, \mathbf{T}\big\rangle.
\end{aligned}$$

Only the second term is related to $\mathbf{T}$ while the first term is constant. Therefore, we can define

$$\mathbf{G}^X = -2\mathbf{M}^X\mathbf{T}(\mathbf{M}^X)^\top, \tag{35}$$

and simplify the model in (34) as follows:

$$\begin{aligned}\min_{\mathbf{H},\mathbf{T}}\ & \langle \lambda_X\mathbf{C}^X + \lambda_H\mathbf{C}^H + (\mathbf{1}-\lambda_X)\mathbf{G}^X, \mathbf{T}\rangle + \epsilon\Omega(\mathbf{T}) \\ \text{s.t.}\ & \mathbf{H}\in\mathcal{M}_k^n, \mathbf{T}\in\mathcal{T}(\boldsymbol{\mu},\boldsymbol{\mu}), \mathbf{T}=\mathbf{T}^\top.\end{aligned} \tag{36}$$

Problem with Unfixed $\boldsymbol{\mu}$

For the unfixed $\boldsymbol{\mu}$, we optimize $\mathbf{T}$ without considering the constraints $\mathbf{T}\mathbf{1} = \mathbf{T}^\top\mathbf{1} = \boldsymbol{\mu}$, instead of explicitly optimizing $\boldsymbol{\mu}$. To this end, we define the following domain for $\mathbf{T}$ with unfixed $\boldsymbol{\mu}$

$$\mathcal{T}(\Sigma_n,\Sigma_n) = \{\mathbf{T}\in(\mathbb{R}^+)^{n\times n} \mid \sum_{i=1}^n\sum_{j=1}^n T_{ij} = 1\}, \tag{37}$$

which removes the equality constraints regarding the marginal distribution $\boldsymbol{\mu}$, and only the total mass constraint $\sum_{ij}T_{ij}=1$ is taken into consideration.

Here we consider a modification regarding the entropic regularization term $\Omega(\mathbf{T})$. In Problem (36), the negative entropy regularization is on the joint distribution $\mathbf{T}$, which is proposed in (Cuturi 2013) to induce a smooth solution and speed up the optimization. For an unfixed probability distribution $\boldsymbol{\mu}$, to encourage more samples to be involved in the transport, we minimize the following negative entropy regularization on the marginal distributions $\mathbf{T}\mathbf{1}$ and $\mathbf{T}^\top\mathbf{1}$ (which are also $\boldsymbol{\mu}$) instead of the joint distribution $\mathbf{T}$

$$\begin{aligned}\tilde{\Omega}(\mathbf{T}) &= \Omega(\mathbf{T}\mathbf{1}) + \Omega(\mathbf{T}^\top\mathbf{1}) \\ &= \sum_{i=1}^n T_{i\cdot}(\log T_{i\cdot}-1) + \sum_{j=1}^n T_{\cdot j}(\log T_{\cdot j}-1),\end{aligned} \tag{38}$$

where $T_{i\cdot}$ and $T_{\cdot j}$ are defined as

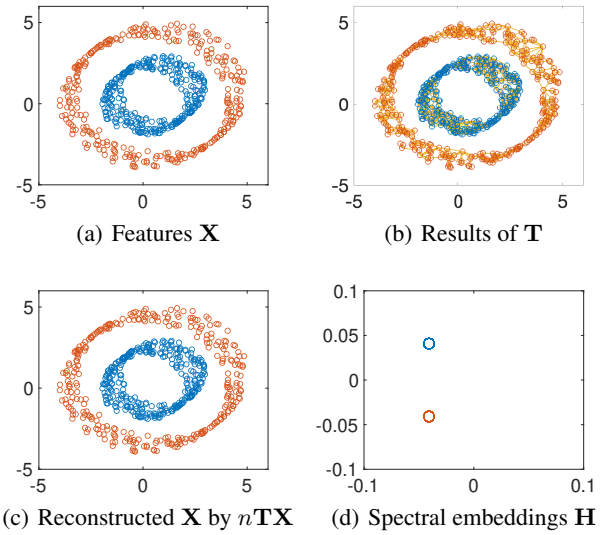
$$T_{i\cdot} = \sum_{j=1}^n T_{ij}, \qquad T_{\cdot j} = \sum_{i=1}^n T_{ij}. \tag{39}$$

As a result, we optimize the following SFGW model

$$\begin{aligned}\min_{\mathbf{H},\mathbf{T}}\ & \langle \lambda_X\mathbf{C}^X + \lambda_H\mathbf{C}^H + (1-\lambda_X)\mathbf{G}^X, \mathbf{T}\rangle + \epsilon\tilde{\Omega}(\mathbf{T}) \\ \text{s.t.}\ & \mathbf{H}\in\mathcal{M}_k^n, \mathbf{T}\in\mathcal{T}(\Sigma_n,\Sigma_n), \mathbf{T}=\mathbf{T}^\top.\end{aligned} \tag{40}$$

This model relaxed the constraints of $\mathbf{T}$ regarding the marginal distribution $\boldsymbol{\mu}$, and can adaptively assign weights for samples, thus is robust to outliers since the outliers are far away from other samples and will be assigned with little mass due to large transport costs.

Since the affinity matrix $\mathbf{T}$ and the spectral embedding $\mathbf{H}$ are coupled in the optimization problems, we adopt an iterative method to update $\mathbf{T}$ and $\mathbf{H}$ alternatively. Due to the page limitation, we present the optimization algorithms for Problems (36) and (40) in the appendix.



(a) Features $\mathbf{X}$ (b) Results of $\mathbf{T}$ (c) Reconstructed $\mathbf{X}$ by $n\mathbf{T}\mathbf{X}$ (d) Spectral embeddings $\mathbf{H}$

Figure 2: Results on simulation data, where different colors indicate different clusters. (a) shows two clusters and the samples, which are connected based on the optimal transport matrix $\mathbf{T}$ as shown in (b). (c) shows the reconstructed samples by $\mathbf{T}$, and (d) shows the spectral embeddings learned from $\mathbf{T}$.

Experiments

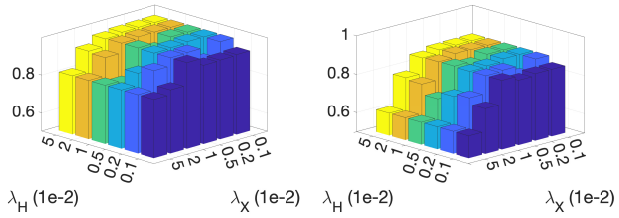
Simulation Study

We conduct experiments on simulation data with two clusters to evaluate the reconstruction property of the optimal transport matrix $\mathbf{T}$ discussed in Eq. (13). Figure 2(a) shows the original features $\mathbf{X}$, and Figure 2(b) shows the samples where the connected edges between samples are based on $\mathbf{T}$, which is obtained by solving Problem (7). Figure 2(c) shows the reconstructed data calculated by Eq. (13), which equals to $\hat{\mathbf{X}} = n\mathbf{T}\mathbf{X}$ here since $\boldsymbol{\mu}$ is a uniform distribution with the elements being $\frac{1}{n}$. Figure 2(d) shows the spectral embeddings learned from the affinity matrix $\mathbf{T}$. We observe that the optimal transport matrix $\mathbf{T}$ finds connections between samples in the same cluster. Although Problem (7) does not explicitly consider a self-expressive term, $n\mathbf{T}\mathbf{X}$ approximately reconstructs $\mathbf{X}$, which is consistent with our discussion in the connection between optimal transport and subspace clustering with a squared loss function.

Experimental Settings

Datasets. We conduct experiments on benchmark datasets from UCI machine learning repository (Asuncion and Newman 2007). The performance of SFGWH is evaluated on datasets without outliers, and the performance of SFGW is evaluated on datasets considering outliers. Following the setting used in (Liu et al. 2019; Zhang et al. 2021), we take the samples from the smallest clusters as outliers.

Compared Methods. For clustering without outliers, we compare the performance of SFGWH with K-means, spectral clustering (SC), and a state-of-the-art spectral clustering



(a) Accuracy (b) NMI

Figure 3: Results of parameter sensitivity on the iris data.

Dataset	K-means	SC	ERCAN	SFGWH
ecoli	56.9(6.0)	56.2(0.7)	69.1(0.0)	**73.7(1.1)**
glass	**53.1(2.0)**	45.4(0.7)	51.4(0.0)	52.0(1.1)
iris	66.7(0.0)	90.7(0.0)	95.9(0.0)	**96.7(0.0)**
landsat	35.8(3.3)	32.8(0.5)	51.0(0.0)	**68.3(1.2)**
seeds	**87.8(5.2)**	64.2(0.4)	82.8(0.0)	**87.8(0.9)**
zoo	65.7(7.3)	62.4(0.0)	**75.2(0.0)**	68.9(0.0)

Table 1: Accuracy results without considering outliers.

method Entropy Regularization for unsupervised Clustering with Adaptive Neighbors (ERCAN) (Wang et al. 2022). ERCAN simultaneously learns the spectral embeddings and the affinity matrix with an entropic regularization.

For clustering considering outliers, we compare the performance of SFGW with K-means, K-means– (Chawla and Gionis 2013), Clustering with Outlier Removal (COR) (Liu et al. 2019), and Fuzzy c-means (FCM) (Bezdek 2013). We follow (Liu et al. 2019) to set the number of clusters as the true number plus one for K-means, and the cluster with the smallest size is regarded as the outlier set. K-means– extends K-means to carry out clustering and outlier detection. COR employs Holoentropy to measure the compactness of the clusters for outlier recognition. FCM allows samples to belong to multiple clusters simultaneously with varying degrees of membership.

Evaluation Metrics. For the setting without outliers, we follow (Wang et al. 2022) to adopt two metrics *i.e.*, the accuracy and the normalized mutual information (NMI). For the settings considering outliers, we follow (Zhang et al. 2021) to adopt the outlier recall and NMI. For all the metrics, the higher the better. We repeatedly conduct experiments 20 times and report the average results.

Dataset	K-means	SC	ERCAN	SFGWH
ecoli	53.8(3.2)	53.4(0.1)	**68.0(0.0)**	58.7(5.4)
glass	36.0(2.6)	25.2(1.6)	38.3(0.0)	**39.0(1.4)**
iris	72.0(0.0)	79.6(0.0)	87.0(0.0)	**88.2(0.0)**
landsat	17.0(1.6)	16.2(0.1)	39.8(0.0)	**55.9(0.9)**
seeds	**68.2(6.5)**	45.3(0.7)	58.4(0.0)	64.6(0.0)
zoo	52.2(5.3)	66.7(0.0)	**76.9(0.0)**	66.8(0.0)

Table 2: NMI results without considering outliers.

Dataset	K-means	K-means–	COR	FCM	SFGW
ecoli	43.1	**59.9**	34.0	50.9	55.6
glass	10.6	**43.7**	32.9	20.9	41.0
landsat	40.3	44.2	45.2	43.9	**46.9**
seeds	1.4	29.1	37.7	42.9	**48.6**
zoo	8.1	11.0	11.9	19.9	**33.3**

Table 3: Recall results with the smallest clusters as outliers.

Dataset	K-means	K-means–	COR	FCM	SFGW
ecoli	58.7	60.0	57.1	44.6	**61.0**
glass	26.7	31.5	16.5	**35.1**	26.4
landsat	39.3	44.2	40.7	42.3	**45.0**
seeds	51.3	52.8	44.1	**74.8**	57.4
zoo	70.2	66.2	59.1	66.7	**77.4**

Table 4: NMI results with the smallest clusters as outliers.

Results and Discussions

We first conduct experiments on the iris data to evaluate the performance of SFGWH with different values of the parameters λ_X and λ_H, which are tuned in the set $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05\}$. Figure 3 shows the results of accuracy and NMI. We observe that the performance of SFGWH is relatively stable with respect to λ_H, and λ_X has a larger effect on the performance. Similar observations can be drawn from the other datasets.

Tables 1 and 2 report the results of experiments without outliers in terms of accuracy and NMI, respectively, where the standard derivations are given in the brackets. In general, ERCAN outperforms SC, which verifies the efficacy of affinity matrix learning in spectral clustering. SFGWH achieves promising performance on both accuracy and NMI on the datasets, which demonstrates the effect of the optimal transport matrix for spectral clustering. Tables 3 and 4 show the results of experiments considering the smallest clusters as outliers in terms of recall and NMI, respectively. SFGW achieves the best or highly competitive performance compared with the other methods, which demonstrates the effectiveness of our proposed method for clustering with outliers.

Conclusion

In this paper, we provide explanations for subspace clustering and spectral clustering from the perspective of optimal transport. Based on a self optimal transport model considering one distribution, we show that the optimal transport matrix bridges the spaces of features and spectral embeddings, and spectral clustering can be modeled as a barycenter problem with an underlying optimal transport discrepancy and guidance of features. We propose a model to learn coupled affinity matrix and spectral embeddings with the help of geometric information extracted by optimal transport, and develop algorithms to optimize the derived problems. The presented connection allows us to employ powerful tools in optimal transport for clustering in the future.

Acknowledgments

This research was supported in part by National Natural Science Foundation of China (62206061, 61876043, 61976052, 62206064), National Key R&D Program of China (2021ZD0111501), National Science Fund for Excellent Young Scholars (62122022), Guangzhou Basic and Applied Basic Research Foundation (2023A04J1700), and the major key project of PCL (PCL2021A12). The work of Michael K. Ng was supported in part by Hong Kong Research Grant Council GRF (17201020, 17300021), CRF (C1013-21GF), and Joint NSFC-RGC (N-HKU76921).

References

An, D.; Lei, N.; Xu, X.; and Gu, X. 2022. Efficient optimal transport algorithm by accelerated gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10119–10128.

Asuncion, A.; and Newman, D. 2007. UCI machine learning repository.

Bai, L.; and Liang, J. 2020. Sparse subspace clustering with entropy-norm. In *International conference on machine learning*, 561–568. PMLR.

Benamou, J.-D.; Carlier, G.; Cuturi, M.; Nenna, L.; and Peyré, G. 2015. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2): A1111–A1138.

Bezdek, J. C. 2013. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

Chawla, S.; and Gionis, A. 2013. k-means–: A unified approach to clustering and outlier detection. In *Proceedings of the 2013 SIAM international conference on data mining*, 189–197. SIAM.

Courty, N.; Flamary, R.; Habrard, A.; and Rakotomamonjy, A. 2017a. Joint distribution optimal transportation for domain adaptation. In *Annual Conference on Neural Information Processing Systems*, 3733–3742.

Courty, N.; Flamary, R.; and Tuia, D. 2014. Domain adaptation with regularized optimal transport. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 274–289.

Courty, N.; Flamary, R.; Tuia, D.; and Rakotomamonjy, A. 2017b. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9): 1853–1865.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Annual Conference on Neural Information Processing Systems*, 2292–2300.

Cuturi, M.; and Doucet, A. 2014. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, 685–693.

Elhamifar, E.; and Vidal, R. 2013. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11): 2765–2781.

Flamary, R.; Cuturi, M.; Courty, N.; and Rakotomamonjy, A. 2018. Wasserstein discriminant analysis. *Machine Learning*, 107(12): 1923–1945.

Kantorovitch, L. 1958. On the translocation of masses. *Management Science*, 5(1): 1–4.

Landa, B.; Coifman, R. R.; and Kluger, Y. 2021. Doubly stochastic normalization of the gaussian kernel is robust to heteroskedastic noise. *SIAM journal on mathematics of data science*, 3(1): 388–413.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 663–670.

Liu, H.; Li, J.; Wu, Y.; and Fu, Y. 2019. Clustering with outlier removal. *IEEE transactions on knowledge and data engineering*, 33(6): 2369–2379.

Liu, H.; Wu, J.; Tao, D.; Zhang, Y.; and Fu, Y. 2015. Dias: A disassemble-assemble framework for highly sparse text clustering. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, 766–774. SIAM.

Lu, C.; Feng, J.; Lin, Z.; Mei, T.; and Yan, S. 2018. Subspace clustering by block diagonal representation. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 487–501.

Lu, C.-Y.; Min, H.; Zhao, Z.-Q.; Zhu, L.; Huang, D.-S.; and Yan, S. 2012. Robust and efficient subspace segmentation via least squares regression. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*, 347–360. Springer.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Ng, A.; Jordan, M.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

Nie, F.; Wang, X.; Jordan, M.; and Huang, H. 2016. The constrained laplacian rank algorithm for graph-based clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Peyré, G.; and Cuturi, M. 2017. Computational Optimal Transport.

Peyré, G.; Cuturi, M.; and Solomon, J. 2016. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, 2664–2672.

Titouan, V.; Courty, N.; Tavenard, R.; and Flamary, R. 2019. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, 6275–6284. PMLR.

Wang, J.; Ma, Z.; Nie, F.; and Li, X. 2022. Entropy regularization for unsupervised clustering with adaptive neighbors. *Pattern Recognition*, 125: 108517.

Xu, R.; and Wunsch, D. 2005. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3): 645–678.

Yan, Y.; Li, W.; Wu, H.; Min, H.; Tan, M.; and Wu, Q. 2018. Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation. In *International Joint Conference on Artificial Intelligence*, 737–753.

Yan, Y.; Tan, M.; Xu, Y.; Cao, J.; Ng, M.; Min, H.; and Wu, Q. 2019. Oversampling for imbalanced data via optimal transport. In *AAAI Conference on Artificial Intelligence*, volume 33, 5605–5612.

Yang, Y.; Xu, D.; Nie, F.; Yan, S.; and Zhuang, Y. 2010. Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing*, 19(10): 2761–2773.

Zhang, Z.; Feng, Q.; Huang, J.; Guo, Y.; Xu, J.; and Wang, J. 2021. A local search algorithm for k-means with outliers. *Neurocomputing*, 450: 230–241.