Understanding Doubly Stochastic Clustering

Tianjiao Ding [1] **Derek Lim** [2] **René Vidal** [1] **Benjamin D. Haeffele** [1]

Abstract

The problem of projecting a matrix onto the space of *doubly stochastic* matrices finds several applications in machine learning. For example, in spectral clustering, it has been shown that forming the normalized Laplacian matrix from a data affinity matrix has close connections to projecting it onto the set of doubly stochastic matrices. However, the analysis of why this projection improves clustering has been limited. In this paper we present theoretical conditions on the given affinity matrix under which its doubly stochastic projection is an ideal affinity matrix (i.e., it has no false connections between clusters, and is well-connected within each cluster). In particular, we show that a necessary and sufficient condition for a projected affinity matrix to be ideal reduces to a set of conditions on the input affinity that decompose along each cluster. Further, in the *subspace clustering* problem, where each cluster is defined by a linear subspace, we provide geometric conditions on the underlying subspaces which guarantee correct clustering via a continuous version of the problem. This allows us to explain theoretically the remarkable performance of a recently proposed doubly stochastic subspace clustering method.

1. Introduction

Spectral clustering is a core technique in machine learning, allowing one to cluster data with relatively general geometric arrangements based on pairwise measures of similarity (or affinity) between data points. The steps of spectral clustering are well-known: 1) Define an *affinity matrix* whose $(i,j)^{\text{th}}$ entry measures the similarity between points i and j; 2) *Normalize* the affinity matrix and compute a *Laplacian*; 3) Use eigenvectors of the Laplacian to define an embedding of the data; and 4) Cluster the embedding. However, despite the relative simplicity in defining these steps, there are several implementation details that can have a considerable impact on the overall clustering performance. For example, the definition of the affinity matrix appears to be the most important choice one needs to consider for good performance. However, a key detail that often receives relatively little attention but can have a significant impact on performance is how to normalize the affinity matrix. Indeed, the two most popular spectral clustering methods, Ratio Cut (Hagen & Kahng, 1992) and Normalized Cut (NCut) (Shi & Malik, 2000), use two different normalizations of the affinity, which emerge as continuous relaxations of two different clustering objectives. An alternative point of view is provided by Zass & Shashua (2006), who argue that these methods for normalizing the affinity matrix are closely related to projecting the affinity onto the space of doubly stochastic matrices, with the difference between methods being what distance[3] is minimized in the projection.

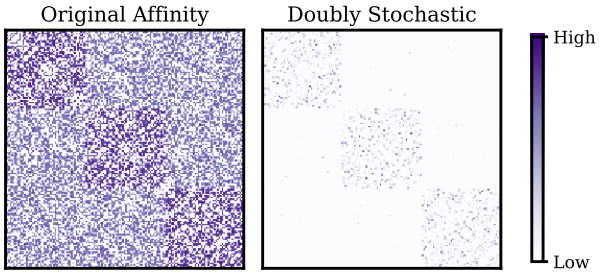


Figure 1. Left: An affinity from a weighted stochastic block model of 3 clusters, where the probability of a inter- and intra-cluster connection are both 0.5, and the weights of inter- and intra-cluster connections are from $\mathcal{N}(1,.1^2)$ and $\mathcal{N}(1.25,.1^2)$ respectively. Right: A projection of the affinity onto the set of doubly stochastic matrices under the ℓ_2 metric, which has few false connections and recovers the 3 clusters.

With this context, recall that an ideal affinity matrix satisfies two properties: 1) *Connectivity*, i.e., the non-zero entries (connections) between points within a cluster form a fully-connected graph and 2) *No false connections*, i.e., there are no connections for points between two different clusters. Prior work has studied conditions on the data under which

[1]Mathematical Institute for Data Science, Johns Hopkins University, USA [2]Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[3]For simplicity, we abuse the word distance to include other more general 'distance-like' measures which to do not necessarily satisfy all axioms of a distance (e.g., divergences).

the affinity matrix computed from the data satisfies some of these properties. For example, the problem of *subspace clustering* considers clustering data which is (approximately) supported on a union of low-dimensional linear subspaces, where each linear subspace defines a cluster. In this case, an affinity matrix is typically computed by expressing each data point as a linear combination of all other data points and enforcing sparse or low-rank properties on the matrix of coefficients, which is then used to define the affinity (Vidal et al., 2016). While several sufficient conditions on the data under which no false connections exist between points from different subspaces (Property 2) have been derived (Vidal et al., 2016), such conditions do not guarantee the connectivity of the affinity. This issue is discussed in (You et al., 2016), where the trade off between sparsity and connectivity is studied experimentally. Moreover, even when the conditions on the data that guarantee no false connections of the affinity are violated, suitable normalization can still lead to an ideal affinity. Indeed, a wide number of ad-hoc approaches have been proposed to normalize an affinity matrix beyond the standard normalization inherent to NCut or RatioCuts (Liu et al., 2013; Ji et al., 2014; Elhamifar & Vidal, 2013), and it has even been argued that much of the benefit claimed by many proposed clustering algorithms can actually be attributed largely to ad-hoc affinity normalization (Haeffele et al., 2020). Seeking a more principled approach to affinity normalization, Lim et al. (2020) follow the interpretation of Zass & Shashua (2006) and show empirically that normalizing the affinity matrix by projecting it to the space of doubly stochastic matrix under the ℓ_2 distance achieves state-of-the-art performance across a wide variety of common clustering datasets. This is illustrated in Figure 1, which shows that a very noisy affinity matrix with numerous false connections between the 3 underlying clusters becomes nearly ideal following projection to a doubly stochastic matrix under the ℓ_2 metric.

Contributions. In this work, we give a rigorous theoretical analysis of doubly stochastic projection of affinity matrices with the ℓ_2 metric (Problem (2)) to help explain its empirical success. First, we prove a necessary and sufficient condition on the input affinity for the projected affinity to have no false connections (Theorem 2.2) even when the original affinity matrix might contain a large number of false connections. While the optimality condition couples the primal and dual variables from different clusters together which could complicate the analysis, the condition we provide relates quantities from decoupled sub-problems, each of which concerns the doubly stochastic projection of the entries within only one cluster. Further, under some models of the input affinity, this allows us to characterize conditions under which an input affinity with false connections will have no false connections *and* be well-connected within each cluster following doubly stochastic projection (Corollaries 2.3 and 2.4). Then, we specialize to the subspace clustering data model, where the data is assumed to be generated from a union of linear subspaces, each defining a cluster. For this setting, we develop a continuous problem (21), i.e., in the limit when the number of data points becomes very large and uniformly distributed, followed by a continuous counterpart of the decoupling theorem (Theorem 3.2). This allows for an analysis of false connections (Theorem 3.5) and the connectivity of the projected affinities (Theorem 3.8), which depends solely on the subspace dimensions, percentage of points within each subspace, and the angles between the subspaces. In particular, we show that there will be no false connections in the normalized affinity matrix if the subspaces are sufficiently separated in angle, or have sufficiently low dimensions, or are well balanced in terms of mixture weights. Finally, we conduct a variety of experiments that illustrate our theoretical findings and demonstrate the utility of doubly stochastic projection in different settings.

2. Doubly Stochastic Clustering

2.1. Problem Formulation

To define the doubly stochastic clustering problem that we study, consider n data points drawn from k underlying clusters, where cluster l contains n_l points and $n = n_1 + \cdots + n_k$. Define the mixture weight of each cluster l as $\pi_l = \frac{n_l}{n}$.

Given an affinity matrix $\boldsymbol{K} \in \mathbb{R}^{n\times n}$, where $K_{ij} = K_{ji} \geq 0$ denotes the similarity between data points i and j, the method of doubly stochastic clustering consists of two steps. First, one projects a scaled version of the given $\boldsymbol{K}$ onto the set of doubly stochastic matrices[4]

$$\mathcal{A}_n := \{\boldsymbol{A} \in \mathbb{R}^{n\times n} : A_{ij} \geq 0, \forall i, j;\ \boldsymbol{A}\mathbf{1} = \boldsymbol{A}^\top \mathbf{1} = n\mathbf{1}\}$$

under some notion of distance $\mathfrak{d}$

$$\boldsymbol{A}^* = \operatorname*{arg\,min}_{\boldsymbol{A}\in\mathcal{A}_n} \mathfrak{d}(\boldsymbol{A}, \frac{1}{\eta}\boldsymbol{K}), \tag{1}$$

where $\mathbf{1}$ is the vector of all ones, and $\eta > 0$ is a scaling parameter. After that, one runs spectral clustering on the obtained $\boldsymbol{A}^*$ to produce the final clustering. One immediate merit of projecting onto $\mathcal{A}_n$ is that one does not need to choose between Ratio Cut and NCut, since both are equivalent[5] on a doubly stochastic affinity $\boldsymbol{A}^*$.

Choice of the Distance $\mathfrak{d}$. As briefly mentioned in §1, classical spectral clustering methods correspond to projecting the given affinity $\boldsymbol{K}$ onto the set of doubly stochastic matrices $\mathcal{A}_n$ under different distances: Ratio Cut is closely related to

[4]The usual definition of a doubly stochastic matrix requires the row and column sums to be 1. Here, without loss of generality, we define the sums to be n to simplify notation in our later analysis.

[5]That is, the two cuts yield the same Laplacian $\boldsymbol{I} - \frac{1}{n}\boldsymbol{A}^*$.

projection under the ℓ_1 metric, and NCut to projection under the KL divergence (Zass & Shashua, 2006). One interesting and perhaps natural alternative is to use the ℓ_2 norm as a distance. Remarkably, doubly stochastic projection under the ℓ_2 metric yields state-of-the-art performance in subspace clustering on a variety of realistic datasets (Lim et al., 2020). This is largely due to the empirical observation that such a projection strikes a balance between removing false connections (Property 2) and maintaining connectivity (Property 1), via a parameter η that controls the sparsity of the output affinity $\boldsymbol{A}^*$. This control of sparsity is, however, not present when projecting with the ℓ_1 norm[6] and also absent when projecting with the KL divergence[7].

Importance of Connectivity. Note that the absence of false connections is insufficient to guarantee correct clustering. This is because the connections within one cluster may not form a connected subgraph, resulting in the cluster being over-segmented. Further, the stability of spectral clustering (§1) against perturbations on the affinity is associated with whether each cluster is sufficiently well connected (see §7.1 of Von Luxburg, 2007). Hence, the stability of the final clustering benefits from the affinity being as connected as possible within each cluster.

Based on the discussion above, this paper focuses on doubly stochastic projection under the ℓ_2 metric, which we refer to as DS-D($\boldsymbol{K}, \eta$):

$$\boldsymbol{A}^* = \operatorname*{arg\,min}_{\boldsymbol{A} \in \mathcal{A}_n} \left\| \boldsymbol{A} - \frac{1}{\eta}\boldsymbol{K} \right\|_F^2. \tag{2}$$

2.2. Optimality Analysis

We first study the optimality conditions of DS-D($\boldsymbol{K}, \eta$). Since the optimization problem is strongly convex, a standard primal-dual analysis gives the following necessary and sufficient conditions for global optimality.

Proposition 2.1. *The optimality conditons to DS-D($\boldsymbol{K}, \eta$) are*

$$\boldsymbol{A}^* = \frac{1}{\eta}[\boldsymbol{K} - \boldsymbol{\alpha}^*\mathbf{1}^\top - \mathbf{1}\boldsymbol{\alpha}^{*\top}]_+, \tag{3}$$

$$\frac{1}{\eta}[\boldsymbol{K} - \boldsymbol{\alpha}^*\mathbf{1}^\top - \mathbf{1}\boldsymbol{\alpha}^{*\top}]_+^*\mathbf{1} = n\mathbf{1}, \tag{4}$$

where $\boldsymbol{A}^* \in \mathbb{R}^{n\times n}$ is the unique primal solution, $\boldsymbol{\alpha}^* \in \mathbb{R}^n$ is a dual solution which satisfies (4), and $[\cdot]_+ = \max(\cdot, 0)$ is applied entrywise.

Note that from the form of the primal solution for $\boldsymbol{A}^*$, the no false connections property is equivalent to saying that for all i, j coming from different clusters, $K_{ij} \leq \alpha_i^* + \alpha_j^*$. As such, we would like to find lower bounds on entries of $\boldsymbol{\alpha}^*$ to give sufficient conditions on $\boldsymbol{A}$ under which $\boldsymbol{A}^*$ enjoys the no-false-connection property. Nevertheless, it is non-trivial to directly bound $\boldsymbol{\alpha}^*$ in a meaningful way due to the coupling the entries of $\boldsymbol{\alpha}^*$ in (4). However, as we show in our next result, the no-false-connection property is satisfied if and only if the DS-D($\boldsymbol{K}, \eta$) problem can be decoupled into a sequence of doubly stochastic projection problems along the inter-cluster portions of the affinity matrix.

Without loss of generality, assume that the rows/columns of $\boldsymbol{K}$ are sorted according to their cluster membership, i.e.

$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{D}^{(1)} & * & \cdots & * \\ * & \boldsymbol{D}^{(2)} & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & \boldsymbol{D}^{(k)} \end{bmatrix}, \tag{5}$$

so that the l-th diagonal block of $\boldsymbol{K}$, $\boldsymbol{D}^{(l)} \in \mathbb{R}^{n_l \times n_l}$, contains the intra cluster affinities for cluster l. Further let $i \sim j$ and $i \not\sim j$ notate that points i and j are in the same or different clusters, respectively. With this notation we then have the following result.

Theorem 2.2. *The following statements are equivalent:*

1. *For each cluster l, there exist $\boldsymbol{\alpha}^{(l)} \in \mathbb{R}^{n_l}$ a dual solution of DS-D($\boldsymbol{D}^{(l)}, \frac{\eta}{\pi_l}$), such that $\alpha_i^\circ + \alpha_j^\circ \geq K_{ij}$ for all $i \not\sim j$, where $\boldsymbol{\alpha}^\circ := [\boldsymbol{\alpha}^{(1)\top}, \ldots, \boldsymbol{\alpha}^{(k)\top}]^\top \in \mathbb{R}^n$.*
2. *$\boldsymbol{A}^\circ := \operatorname{diag}(\frac{1}{\pi_1}\boldsymbol{A}^{(1)}, \ldots, \frac{1}{\pi_k}\boldsymbol{A}^{(k)})$ is the unique primal solution of DS-D($\boldsymbol{K}, \eta$), where for each cluster l, $\boldsymbol{A}^{(l)}$ is the unique primal solution of DS-D($\boldsymbol{D}^{(l)}, \frac{\eta}{\pi_l}$).*
3. *The unique primal solution of DS-D($\boldsymbol{K}, \eta$) has no false connections.*

A proof is given in the Appendix. The above theorem gives necessary and sufficient conditions for the projected doubly stochastic affinity to have no false connections. Notably, in words this theorem implies that if one solves a DS-D problem for each within-cluster block $\boldsymbol{D}^{(l)}$ of $\boldsymbol{K}$, then the solution of the DS-D for the entire $\boldsymbol{K}$ matrix will have no false connections if and only if the solution can be formed by concatenating all of the within-cluster solutions into a block diagonal matrix. From this result, we can give several immediate corollaries for simple properties of the affinity matrix that are sufficient to guarantee the no-false-connections property.

Corollary 2.3 (Constant Intra-cluster Connections)**.** *Suppose $\boldsymbol{K}$ is such that for each cluster l, all intra-cluster connections have values μ_l, i.e., $\boldsymbol{D}^{(l)} = \mu_l \mathbf{1}_{n_l}\mathbf{1}_{n_l}^\top$. Then, the unique primal optimal for DS-D($\boldsymbol{K}, \eta$) has no false connections and is fully-connected within each cluster, as*

[6] Since the output affinity $\boldsymbol{A}^*$ differs from the input $\boldsymbol{K}$ only by their diagonal entries as per Proposition 1 of (Zass & Shashua, 2006), the sparsity is almost unchanged.

[7] See Proposition 2 of (Zass & Shashua, 2006).

long as any connections in $\boldsymbol{K}$ between clusters $p \neq q$ have values at most $\frac{1}{2}(\mu_p + \mu_q - \frac{\eta}{\pi_p} - \frac{\eta}{\pi_q})$.

Proof. Suppose the upper bound on connections between clusters in $\boldsymbol{K}$ holds. We first show that statement 1. in Theorem 2.2 holds. Consider the problem DS-D($\boldsymbol{D}^{(l)}$, $\frac{\eta}{\pi_l}$). By Proposition 2.1, it can be seen that $\boldsymbol{\alpha}^{(l)} := \frac{1}{2}\left(\mu_l - \frac{\eta}{\pi_l}\right)\mathbf{1}_{n_l}$ and $\boldsymbol{A}^{(l)} := \frac{\pi_l}{\eta}[\boldsymbol{D}^{(l)} - \boldsymbol{\alpha}^{(l)}\mathbf{1}_{n_l}^\top - \mathbf{1}_{n_l}\boldsymbol{\alpha}^{(l)\top}]$ is respectively a dual optimal and the primal optimal for this problem. Indeed, a row sum of $\boldsymbol{A}^{(l)}$ takes the form

$$\frac{\pi_l}{\eta}\sum_{j=1}^{n_l}\left[\mu_l - \mu_l + \frac{\eta}{\pi_l}\right]_+ = \frac{\pi_l}{\eta}\sum_{j=1}^{n_l}\frac{\eta}{\pi_l} = n_l. \tag{6}$$

Now, letting $\boldsymbol{\alpha}^\circ = [\boldsymbol{\alpha}^{(1)\top}, \ldots, \boldsymbol{\alpha}^{(k)\top}]^\top$ as in Theorem 2.2, note that if point i is in cluster p and point j is in cluster $q \neq p$, then $\alpha_i^\circ + \alpha_j^\circ = \frac{1}{2}\left(\mu_p + \mu_q - \frac{\eta}{\pi_p} - \frac{\eta}{\pi_q}\right)$. Since we assume that $\frac{1}{2}(\mu_p + \mu_q - \frac{\eta}{\pi_p} - \frac{\eta}{\pi_q}) \geq K_{ij}$, statement 1. in Theorem 2.2 holds. It follows from statement 2. that the primal optimal $\boldsymbol{A}^\circ$ for DS-D($\boldsymbol{K}, \eta$) is fully connected within cluster l, where each connection is of strength $\frac{1}{\pi_l}$; from statement 3. that $\boldsymbol{A}^\circ$ has no false connections. □

Corollary 2.4 (Constant Sum of Top Intra-cluster Connections)**.** *Suppose $\boldsymbol{K}$ is such that for each cluster l, there exists an integer σ_l, such that the σ_l-nearest-neighbour[8] graph of $\boldsymbol{D}^{(l)}$, denoted as $\boldsymbol{S}^{(l)}$, satisfy the following*

1. $\frac{1}{\sigma_l}\sum_{j\in S_1^{(l)}} D_{1j}^{(l)} = \cdots = \frac{1}{\sigma_l}\sum_{j\in S_{n_l}^{(l)}} D_{n_l j}^{(l)} := e_l$,
2. $\forall (i,j) \in S^{(l)}, D_{ij}^{(l)} \geq e_l - \frac{n\eta}{\sigma_l}$,
3. $\forall (i,j) \notin S^{(l)}, D_{ij}^{(l)} \leq e_l - \frac{n\eta}{\sigma_l}$.

Then, the unique primal optimal for DS-D($\boldsymbol{K}, \eta$) has no false connections and is connected within cluster l along the σ_l-nearest-neighbour graph, as long as any connections in $\boldsymbol{K}$ between clusters $p \neq q$ have values at most $\frac{1}{2}(e_p + e_q - \frac{n\eta}{\sigma_p} - \frac{n\eta}{\sigma_q})$.

Intuitively, the above corollaries suggest that as long as any two clusters have false connections smaller than the average of the largest intra-cluster connections, minus some gap that is proportional to the η parameter, the optimal doubly stochastic projection will have no false connections and is well connected within each cluster.

The next corollary shows that the doubly stochastic projection given by the solution of DS-D($\boldsymbol{K}, \eta$) is invariant to perturbations by low rank matrices of the form $\boldsymbol{v}\mathbf{1}^\top + \mathbf{1}\boldsymbol{v}^\top$ for any $\boldsymbol{v} \in \mathbb{R}^n$, such as an elementwise perturbation by a constant $c \in \mathbb{R}$. Thus, the doubly stochastic projection can remove certain additive corruptions on the input affinity $\boldsymbol{K}$.

Corollary 2.5. *For a vector $\mathbf{v} \in \mathbb{R}^n$, the primal optimal to DS-D($\boldsymbol{K}, \eta$) is the same as that of DS-D($\boldsymbol{K}+\mathbf{1}\mathbf{v}^\top+\mathbf{v}\mathbf{1}^\top, \eta$). In particular, adding a constant $c \in \mathbb{R}$ to each entry of $\boldsymbol{K}$ does not change the solution DS-D($\boldsymbol{K} + c\mathbf{1}\mathbf{1}^\top$).*

Proof. For any $\boldsymbol{v} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathcal{A}_n$, note that

$$\langle \mathbf{1}\boldsymbol{v}^\top + \boldsymbol{v}\mathbf{1}^\top, \boldsymbol{A}\rangle = \sum_i\sum_j v_i A_{ij} + \sum_i\sum_j v_j A_{ij} \tag{7}$$

$$= \sum_i v_i \sum_j A_{ij} + \sum_j v_j \sum_i A_{ij} \tag{8}$$

$$= \sum_i v_i n + \sum_j v_j n = 2n\sum_j v_j, \tag{9}$$

which is a constant independent of $\boldsymbol{A}$. Thus, the minimizer of the objective DS-D($\boldsymbol{K}$) is the same as the minimizer of the objective DS-D($\boldsymbol{K} + \mathbf{1}\boldsymbol{v}^\top + \boldsymbol{v}\mathbf{1}^\top$). The second part of the lemma follows from taking $\boldsymbol{v} = \frac{c}{2}\mathbf{1}$. □

3. Doubly Stochastic Subspace Clustering

Given our above analysis for the general DS-D problem, we now consider a specific data model to provide additional analysis. Specifically, we will analyze the subspace clustering model, where the data are assumed to lie on a union of (low-dimensional) linear subspaces and the goal is to cluster data points based on which linear subspace they lie in. This assumption is a reasonable model for many real-world data problems (Vidal, 2011), possibly after a preprocessing of the data such as the scattering transform (Bruna & Mallat, 2013). Past work has theoretically studied this data model in various settings (Soltanolkotabi & Candés, 2012; Soltanolkotabi et al., 2014; You & Vidal, 2015), and recent empirical work that uses doubly stochastic projection has achieved state-of-the-art empirical results for subspace clustering problems (Lim et al., 2020).

Specifically, consider k subspaces $\{\mathcal{S}_l\}_{l=1}^k$ of $\mathbb{R}^D$, each of dimension $\dim \mathcal{S}_l := d_l < D$. Suppose each $\mathcal{S}_l$ contains n_l points $\boldsymbol{X}^{(l)} = [\boldsymbol{x}_1^{(l)}, \ldots, \boldsymbol{x}_{n_l}^{(l)}]$ lying on the unit sphere $\mathbb{S}^{D-1}$. Let $\boldsymbol{\Phi} = \bigcup_{l=1}^k \mathcal{S}_l$ denote the union of the subspaces, and $\boldsymbol{X} = [\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k)}] \in \mathbb{R}^{D\times n}$ the collection of data. Given $\boldsymbol{X}$, one performs subspace clustering using doubly stochastic projection by first computing an affinity matrix $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ from $\boldsymbol{X}$ via *kernel functions*[9] or existing subspace clustering methods and the solving the DS-D($\boldsymbol{K}, \eta$) problem. Then, one can perform spectral clustering on the solution to DS-D($\boldsymbol{K}, \eta$) and obtain a final clustering.

[8] Namely, $\boldsymbol{S}^{(l)}$ contains indices of the largest σ_l entries of each row or column of $\boldsymbol{D}^{(l)}$ (recall $\boldsymbol{K}$ is symmetric).

[9] That is, for some positive definite function $\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, take $K_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for every i, j. Examples for κ include the Euclidean inner product and radial basis function kernel.

3.1. Analysis of the Continuous Problem

Since directly tackling the (discrete) DS-D problem with finite data points from a union of subspace model is non-trivial, we consider its continuous counterpart where the number of data points becomes infinitely large and uniformly distributed. First, we define a discrete measure associated with data $\boldsymbol{X}$

$$\mu_{\boldsymbol{X}}(\boldsymbol{z}) = \frac{1}{n}\sum_{j=1}^{n}\delta(\boldsymbol{z}-\boldsymbol{x}_j) \tag{10}$$

in which $\delta(\cdot)$ is the Dirac function on the sphere $\mathbb{S}^{D-1}$ such that for any $f:\mathbb{S}^{D-1}\to\mathbb{R}$ and any $\boldsymbol{z}_0\in\mathbb{S}^{D-1}$,

$$\int_{z\in\mathbb{S}^{D-1}} f(\boldsymbol{z})\delta(\boldsymbol{z}-\boldsymbol{z}_0)d\mu_{\mathbb{S}^{D-1}} = f(\boldsymbol{z}_0) \tag{11}$$

where $\mu_{\mathbb{S}^{D-1}}$ is the uniform measure on $\mathbb{S}^{D-1}$. With the above definitions, one can view the discrete objective (2) as

$$\left\|\frac{1}{\eta}\boldsymbol{K}-\boldsymbol{A}\right\|_F^2 = \sum_{i,j}\left(\frac{1}{\eta}K_{ij}-A_{ij}\right)^2 \tag{12}$$

$$\simeq \sum_{i,j}\left(\frac{1}{\eta}K(\boldsymbol{x}_i,\boldsymbol{x}_j)-A(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)^2 \tag{13}$$

$$= n^2 E_{\boldsymbol{y}\sim\mu_{\boldsymbol{X}}}E_{\boldsymbol{z}\sim\mu_{\boldsymbol{X}}}\left\{\left(\frac{1}{\eta}K(\boldsymbol{y},\boldsymbol{z})-A(\boldsymbol{y},\boldsymbol{z})\right)^2\right\}. \tag{14}$$

where we assume $K\in L^2(\mathbb{S}^{D-1}\times\mathbb{S}^{D-1})$ is a kernel function. Similarly, the constraints in (2) can be written as

$$\forall i,\ n = \sum_j A_{ij} \simeq \sum_j A(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{15}$$

$$= nE_{\boldsymbol{z}\sim\mu_{\boldsymbol{X}}}\{A(\boldsymbol{x}_i,\boldsymbol{z})\}, \tag{16}$$

$$\forall i,\ n \simeq nE_{\boldsymbol{z}\sim\mu_{\boldsymbol{X}}}\{A(\boldsymbol{z},\boldsymbol{x}_i)\}, \tag{17}$$

$$\forall i,j,\ A_{ij} = A(\boldsymbol{x}_i,\boldsymbol{x}_j)\geq 0. \tag{18}$$

Note further that the discrete measure can be separated as

$$\mu_{\boldsymbol{X}}(z) = \sum_{l=1}^{k}\frac{n_l}{n}\mu_{\boldsymbol{X}^{(l)}}(z) = \sum_{l=1}^{k}\pi_l\mu_{\boldsymbol{X}^{(l)}}(z). \tag{19}$$

In the continuous case, one replaces the discrete measure μ_X by its continuous counterpart

$$\mu'(z) := \sum_{l=1}^{k}\pi_l\mu_{\mathbb{S}^{D-1}\cap\mathcal{S}_l}(z), \tag{20}$$

where $\mu_{\mathbb{S}^{D-1}\cap\mathcal{S}_l}$ is the uniform measure on $\mathbb{S}^{D-1}\cap\mathcal{S}_l$. This leads to the continuous problem DS-C($\boldsymbol{K},\eta$)

$$\min_A E_{\boldsymbol{y}\sim\mu'}E_{\boldsymbol{z}\sim\mu'}\left\{\left(\frac{1}{\eta}K(\boldsymbol{y},\boldsymbol{z})-A(\boldsymbol{y},\boldsymbol{z})\right)^2\right\} \tag{21}$$

$$s.t.\ \ E_{\boldsymbol{z}\sim\mu'}\{A(\boldsymbol{y},\boldsymbol{z})\}=1,\ \boldsymbol{y}\in\mathbb{S}^{D-1}\cap\Phi\ a.e. \tag{22}$$

$$E_{\boldsymbol{y}\sim\mu'}\{A(\boldsymbol{y},\boldsymbol{z})\}=1,\ \boldsymbol{z}\in\mathbb{S}^{D-1}\cap\Phi\ a.e. \tag{23}$$

$$A(\boldsymbol{y},\boldsymbol{z})\geq 0,\ \boldsymbol{y},\boldsymbol{z}\in\mathbb{S}^{D-1}\cap\Phi\ a.e. \tag{24}$$

According to an analysis of the quadratically regularized optimal transport problems (Lorenz et al., 2021), $A^*\in L^2(\mathbb{S}^{D-1}\times\mathbb{S}^{D-1})$ is a solution if and only if there exists a dual function $\alpha^*\in L^2(\mathbb{S}^{D-1})$ such that

$$A^*(\boldsymbol{y},\boldsymbol{z}) = \frac{1}{\eta}[K(\boldsymbol{y},\boldsymbol{z})-\alpha^*(\boldsymbol{y})-\alpha^*(\boldsymbol{z})]_+, \tag{25}$$

$$E_{\boldsymbol{y}\sim\mu'}\{[K(\boldsymbol{y},\boldsymbol{z})-\alpha^*(\boldsymbol{y})-\alpha^*(\boldsymbol{z})]_+\}=\eta, \tag{26}$$

$$E_{\boldsymbol{z}\sim\mu'}\{[K(\boldsymbol{y},\boldsymbol{z})-\alpha^*(\boldsymbol{y})-\alpha^*(\boldsymbol{z})]_+\}=\eta, \tag{27}$$

where (25)-(27) are understood pointwise almost everywhere.

Remark 3.1. (21) is strongly convex, hence it has a unique primal solution[10]. However, infinitely-many dual solutions may exist due to $[\cdot]_+$ zeroing out negative inputs.

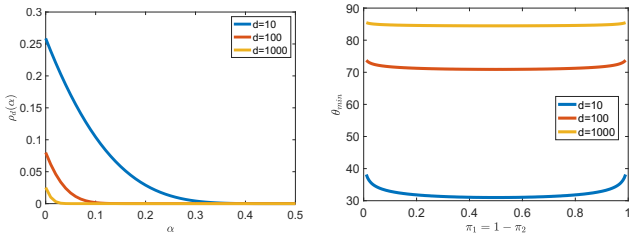
Similar to the case for the discrete problem (§2), the primal solution A^* satisfying the no false connections property (which we will equivalently refer to by saying the solution is *subspace preserving*) is equivalent to saying that for all $\boldsymbol{y},\boldsymbol{z}$ from different subspaces, $K(\boldsymbol{y},\boldsymbol{z})\leq\alpha^*(\boldsymbol{y})+\alpha^*(\boldsymbol{z})$ almost surely. Again, we can separate $\alpha^*(\boldsymbol{y})$ for $\boldsymbol{y}$ from different subspaces to show equivalent conditions for when the subspace preserving property will be satisfied.

Theorem 3.2. *For each $\mathcal{S}_l$, let $\alpha^{(l)}:\mathcal{S}_l\to\mathbb{R}$ be a dual optimal and $A^{(l)}$ the unique primal optimal of DS-C($\boldsymbol{K},\frac{\eta}{\pi_l}$) with measure $\mu_{\mathbb{S}^{D-1}\cap\mathcal{S}_l}$. Define $\alpha^\circ(\boldsymbol{x}):\Phi\to\mathbb{R}$ such that for $\boldsymbol{x}\in\mathcal{S}_l$, $\alpha^\circ(\boldsymbol{x})=\alpha^{(l)}(\boldsymbol{x})$.[11] Define $A^\circ:\Phi\times\Phi\to\mathbb{R}$, $A^\circ(\boldsymbol{y},\boldsymbol{z})=\begin{cases}\frac{1}{\pi_l}A^{(l)}(\boldsymbol{y},\boldsymbol{z}) & \boldsymbol{y},\boldsymbol{z}\in\mathcal{S}_l\\ 0 & o.w.\end{cases}$. The following statements are equivalent:*

1. *$\alpha^\circ(y)+\alpha^\circ(z)\geq K(y,z)$ for all $y\not\sim z$.*
2. *A° is the unique primal optimal for DS-C($\boldsymbol{K},\eta$).*
3. *The unique primal optimal for DS-C($\boldsymbol{K},\eta$) is subspace preserving.*

[10] By uniqueness, we mean that any optimal solutions can only differ from each other on a set of measure zero with respect to μ'.

[11] For $\boldsymbol{x}$ lying in multiple subspaces, $\alpha^\circ(\boldsymbol{x})$ can be defined arbitrary, since such $\boldsymbol{x}$ lie in a set of measure zero with respect to $\mu_{\mathbb{S}^{D-1}\cap\mathcal{S}_l}$, provided that the subspaces are independent, disjoint or intersecting.



(a) (b)

Figure 2. (a) Plot of $\rho_d(\alpha)$ with respect to α and dimension d. (b) Given two subspaces of mixture weights π_1 and $1-\pi_1$, each of dimension d, the subspaces must have first principal angle of at least θ_{min} to guarantee subspace preserving property.

3.2. Example: Inner Product Kernel

To demonstrate the effect of doubly stochastic projection for clustering subspaces, we consider the simplest kernel for K which is inner product kernel $K(\boldsymbol{y}, \boldsymbol{z}) = |\langle \boldsymbol{y}, \boldsymbol{z} \rangle|$. We first define the following quantity, which turns out useful in the analysis later on.

Definition 3.3 (Average height of spherical cap)**.** For any dimension d, define $\rho_d(\alpha) = \int_{\boldsymbol{z} \in \mathbb{S}^{d-1}} [|z_d| - 2\alpha]_+ d\mu_{\mathbb{S}^{d-1}}$.

Remark 3.4. $\rho_d(\alpha)$ is decreasing in α with d fixed and decreasing in d with α fixed. Figure 2a shows $\rho_d(\alpha)$ for $\alpha \in [0, 0.5]$ and $d \in \{10, 100, 1000\}$. More notes on $\rho_d(\alpha)$ are provided in the Appendix.

Theorem 3.5 (Subspace preserving property)**.** *Let K be the inner product kernel. If for any two different subspaces $\mathcal{S}_p, \mathcal{S}_q$, we have*

$$\max_{\substack{\boldsymbol{y} \in \mathbb{S}^{D-1} \cap \mathcal{S}_p \\ \boldsymbol{z} \in \mathbb{S}^{D-1} \cap \mathcal{S}_q}} |\langle \boldsymbol{y}, \boldsymbol{z} \rangle| \leq \rho_{d_p}^{-1}(\frac{\eta}{\pi_p}) + \rho_{d_q}^{-1}(\frac{\eta}{\pi_q}), \tag{28}$$

then the subspace preserving property holds for A^*.

Remark 3.6. Note that with η fixed, $\rho_d^{-1}(\frac{\eta}{\pi})$ is larger when the dimension d is smaller or when the mixture weight π is larger. Moreover, with a fixed subspace arrangement Φ and mixture weights $\{\pi_l\}_{l=1}^k$, subspace preserving property is guaranteed with sufficiently small η, since $\lim_{x \to 0} \rho_d^{-1}(x) = 0.5$.

Remark 3.7. As seen in Figure 2b, to guarantee the subspace preserving property, the minimum angle between two subspaces must be larger, i.e., subspaces should be more separated, when the subspaces are more imbalanced.

Theorem 3.8 (Connectivity)**.** *Let K be the inner product kernel and suppose A^* satisfy subspace preserving property. For any subspace $\mathcal{S}_l$, any two points $\boldsymbol{y}, \boldsymbol{z} \in \mathcal{S}_l \cap \mathbb{S}^{D-1}$ except for a set of measure zero have a non-zero connection in A^* as long as*

$$|\langle \boldsymbol{y}, \boldsymbol{z} \rangle| > 2\rho_{d_l}^{-1}(\frac{\eta}{\pi_l}). \tag{29}$$

Note that from the above two results, we are guaranteed that the doubly stochastic projection will achieve the desired properties of being subspace preserving and being fully connected for an appropriate choice of parameter η.

4. Experiments

We now verify our theoretical analysis with a variety of numerical experiments.

Metrics. Since our theorems predict the no-false-connection and subspace preserving properties, we report the feature detection error[12] (FDE) defined as

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j: j \not\sim i} |\boldsymbol{A}_{ij}| / \|\boldsymbol{A}_i\|_1 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j: j \not\sim i} |\boldsymbol{A}_{ij}|, \tag{30}$$

and the percent of false connections (PFC). Both metrics take value 0 when $\boldsymbol{A}$ has no false connections. Likewise to asses the connectivity, we report the number of non-zeros (NNZ) of $\boldsymbol{A}$, defined as the average number of entries per row of $\boldsymbol{A}$ that are larger than 10^{-8}. To further evaluate the quality of the final clustering, we run spectral clustering on $\boldsymbol{A}$, and report the clustering accuracy (ACC) and normalized mutual information (NMI).

4.1. Weighted Stochastic Block Model

First, we consider the problem of clustering an affinity sampled from a Weighted Stochastic Block Model (WSBM) with random edge weights (Aicher et al., 2015). A sample affinity $\boldsymbol{K}$ is taken by first including each intra-cluster edge with a probability p and each inter-cluster edge with a probability q, then drawing edge weights for these chosen edges. Intra-cluster edge weights are drawn from a normal distribution $\mathcal{N}(1.25, .1^2)$ and inter-cluster edge weights are drawn from another normal distribution $\mathcal{N}(1, .1^2)$. In our experiments, we take p close to q, so a successful algorithm cannot just use the difference in sparsity between blocks, and must take into account the edge weight. We use a WSBM with 5 blocks and 50 points per block.

We run the DS-D($\boldsymbol{K}, \eta$) studied by this paper on the $\boldsymbol{K}$ with varying $\eta \in \{0.002, 0.004, 0.01\}$ to obtain a doubly stocahstic $\boldsymbol{A}$. Table 1 reports mean and median feature detection error, percent false connections, number of non-zeros, clustering accuracy and normalized mutual information of $\boldsymbol{A}$ over 10 $\boldsymbol{K}$ samples from weighted stochastic block models WSBM(.5, .4) and WSBM(.5, .5). Remarkably, even when the original affinity has a significant amount of false connections and achieves low clustering accuracy, the affinity normalized by DS-D has much fewer false connections

[12] This metric is commonly used in subspace clustering, known as feature detection error (Soltanolkotabi & Candés, 2012) or subspace preserving error (You et al., 2016; Lim et al., 2020).

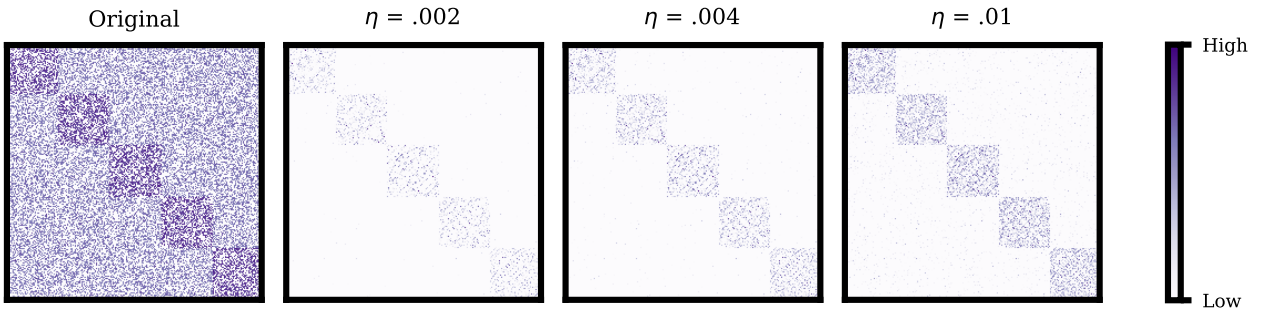


Figure 3. An affinity $\boldsymbol{K}$ sampled from the weighted stochastic block model WSBM(.5, .4) of 5 clusters and its doubly stochastic projections DS-D($\boldsymbol{K}, \eta$) with varying η.

Table 1. Feature detection error, number of non-zeros, clustering accuracy and normalized mutual information for the original affinity $\boldsymbol{K}$ and the output affinity from doubly stochastic clustering DS-D($\boldsymbol{K}, \eta$), with $\boldsymbol{K}$ generated by the weighted stochastic block model WSBM(p, q) of 5 clusters, where p, q are the probability of intra-cluster and inter-cluster edges respectively. Mean and standard deviations are taken over 10 trials.

Dataset	WSBM(.5, .4)					WSBM(.5, .5)				
Metrics	FDE	PFC	NNZ	ACC	NMI	FDE	PFC	NNZ	ACC	NMI
Original affinity	.72±.00	77±.2	104.5±.2	.79±.05	.53±.07	.77±.00	80±.2	124.5±.2	.33±.02	.07±.02
DS-D ($\eta = .002$)	.02±.00	**4.5**±.6	12.0±.2	**1.0**±.00	**1.0**±.00	.03±.00	**5.9**±.6	12.0±.3	**1.0**±.00	**1.0**±.00
DS-D ($\eta = .004$)	.04±.01	11±.9	18.5±.4	**1.0**±.00	**1.0**±.00	.06±.00	14±.8	18.7±.3	**1.0**±.00	**1.0**±.00
DS-D ($\eta = .01$)	.15±.01	35±1	34.1±.6	**1.0**±.00	**1.0**±.00	.17±.00	38±.9	35.9±.7	**1.0**±.00	**1.0**±.00

and perfect clustering accuracy. For example, the original affinity sampled from WSBM(.5, .4) has mean feature detection error and clustering accuracy of 0.72 and 0.79, while the one produced by DS-D($\eta = .004$) has 0.04 and 1.0 respectively. Further, the above conclusion holds even in the challenging case of WSBM(.5,.5), where the probability of an intra-cluster edge p is the same as that of an inter-cluster edge q. Last but not the least, with a smaller η, the affinity given by DS-D is sparser as expected, e.g., the mean number of non-zeros is 12 with $\eta = 0.002$ and 34 with $\eta = 0.01$.

4.2. Subspace Clustering

Here we conduct experiments to demonstrate the effect of doubly stochastic projection when the clusters are defined by linear subspaces. We first fix $\boldsymbol{K}$ to be the inner product kernel, and verify conditions on the subspace preserving property and connectivity studied[13] in §3.2 under various subspace angle θ_{min}, dimension d, and problem parameter η. Next, we further consider the kernel matrices from Least Squares Regression (Lu et al., 2012), and show the effect of doubly stochastic projection improving clustering.

Inner Product Kernel. We generate two subspaces of dimension d in $\mathbb{R}^{D=20}$. To control the angles between subspaces, we choose the basis of subspaces as

$$\boldsymbol{U}^{(1)} = \begin{bmatrix} \boldsymbol{I}_d \\ \boldsymbol{0}_{D-d,d} \end{bmatrix} \in \mathbb{R}^{D\times d},$$

$$\boldsymbol{U}^{(2)} = \begin{bmatrix} \cos(\theta_1) & 0 & \dots & 0 \\ 0 & \cos(\theta_2) & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \cos(\theta_d) \\ \sin(\theta_1) & 0 & \dots & 0 \\ 0 & \sin(\theta_2) & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \sin(\theta_d) \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \in \mathbb{R}^{D\times d}, \tag{31}$$

where $\theta_1 = \theta_{min}$ is the smallest angle between the subspaces, and $\cos(\theta_i)$ decreases linearly from $\cos(\theta_{min})$ to $\cos(\theta_d) = 0.5\cos(\theta_{min})$. Note that $\theta_{min} = 90°$ is a simple case where the subspaces are orthogonal, whereas $\theta_{min} = 0°$ is difficult since the subspaces have non-trivial intersections. After that, from each subspace we sample $n = 50d$ points of unit norm uniformly at random, which gives $100d$ points in total, and the input $\boldsymbol{K}$ is taken to the inner product kernel from those points.

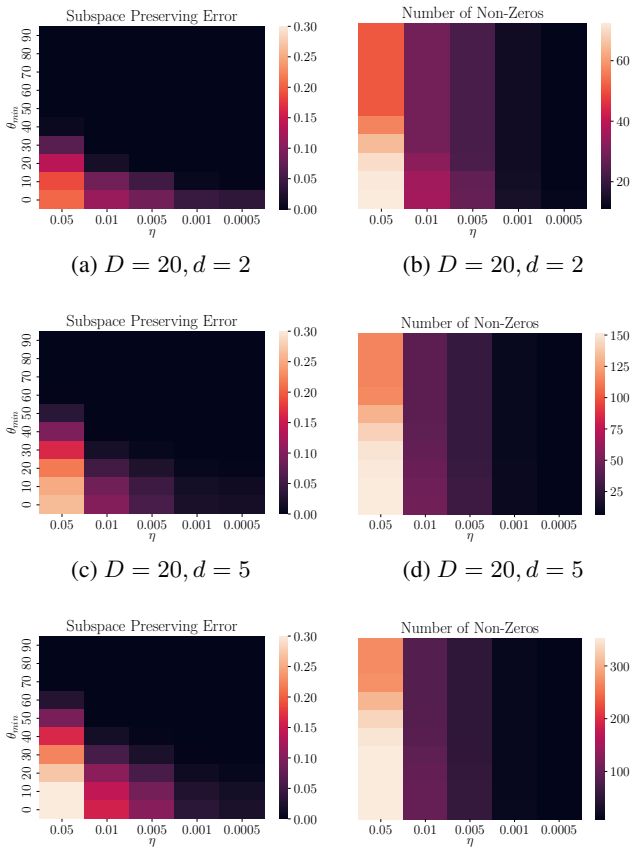
Figure 4 reports feature detection error and number of

[13] While our theorems for understanding subspace clustering are studied in the continuous limit, the experiments are conducted with finite data points.

non-zeros for $d \in \{2, 5, 10\}$, $\theta_{min} \in [0, 90]^\circ$ and $\eta \in \{5 \cdot 10^{-2}, 10^{-2}, 5 \cdot 10^{-3}, 10^{-3}\}$. Remarkably, even though the inner product kernel $\boldsymbol{K}$ is not subspace preserving when the subspaces are not orthogonal (i.e., $\theta_{min} \neq 90^\circ$), its doubly stochastic projection $\boldsymbol{A}^*$ is subspace preserving when the subspaces are far away enough from each other (i.e., θ_{min} is large) or when η is sufficiently small. For example, in Figures 4a, 4c and 4e, subspace preserving property holds for $\boldsymbol{A}^*$ when $\theta_{min} \geq 40^\circ$ and $\eta \leq 0.001$. Moreover, given η, the closest angle between two subspaces such that $\boldsymbol{A}^*$ does not have false connections is increasing in the subspace dimension d, e.g., using a rather loose $\eta = 0.05$, the smallest θ_{min} is below 40° for $d = 2$ (Figure 4a), while θ_{min} is above 60° for $d = 10$ (Figure 4e). Surprising as it may sound, the above phenomenon are expected from Theorem 3.5. Last but not the least, the doubly stochastic affinity A^* has fewer non-zero connections when η is smaller, which is expected from Theorem 3.8. As such, in practise one may want to tune η to balance between having fewer false connections and the connectivity. Nevertheless, there seems to be a range of suitable η, even for this arguably simplest inner product kernel. For instance, with $\eta \in [0.001, 0.01]$, $\boldsymbol{A}^*$ is subspace preserving whenever $\theta_{min} \geq 30^\circ$ and gets at least d-many connections inside the subspace.

Least Squares Regression Kernel. Beyond the inner product kernel, we also investigate the effect of doubly stochastic projection on other kernels used for subspace clustering, such as the LSR kernel (Lu et al., 2012). We generate k subspaces of dimension d in $\mathbb{R}^{D=20}$ uniformly at random, from which we further sample $n = 50d$ points of unit norm uniformly at random. First, we compute the LSR kernel $\boldsymbol{K}$ on the data, and record its performance (raw). The parameter γ in LSR is set to be 10. Then, we apply the doubly stochastic projection and report the metrics on $\boldsymbol{A}^*$.

Figure 5 reports feature detection error, number of non-zeros, clustering accuracy and normalized mutual information for $k \in \{2, 5\}$ and $d \in \{8, 14\}$. As expected, clustering seems to be simpler when one has fewer number of subspace ($k = 2$) or the subspaces are of lower dimension ($d = 8$). Interestingly, doubly stochastic projection seems to improve clustering over the LSR kernel. This is evidenced by the fact that doubly stochastic not only decreases feature detection error while still leaving a high connectivity, but also increases the final clustering quality as measured by clustering accuracy and normalized mutual information.



(a) $D = 20, d = 2$ (b) $D = 20, d = 2$

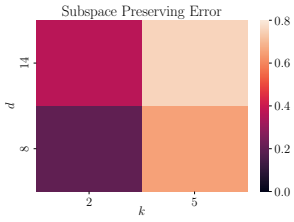
(c) $D = 20, d = 5$ (d) $D = 20, d = 5$

(e) $D = 20, d = 10$ (f) $D = 20, d = 10$

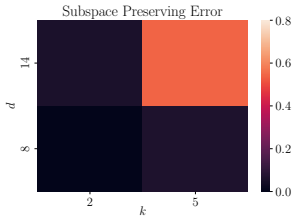
Figure 4. Feature detection error and number of non-zeros of the output affinity from doubly stochastic clustering DS-D($\boldsymbol{K}, \eta$), with the input affinity $\boldsymbol{K}$ being the inner product kernel of data coming from two subspaces of dimensions $d \in \{2, 5, 10\}$ in $\mathbb{R}^{D=20}$, each containing $n = 50d$ points of unit norm uniformly at random. For each d, metrics are shown for varying the smallest angle between subspaces $\theta_{min} \in [0, 90]^\circ$ and problem parameter $\eta \in \{5 \cdot 10^{-2}, 10^{-2}, 5 \cdot 10^{-3}, 10^{-3}\}$. This demonstrates the geometric conditions for subspace preserving property and connectivity in terms of θ_{min}, d, η, as predicted by Theorem 3.5 and Theorem 3.8 for the continuous problem (§3).

5. Conclusion

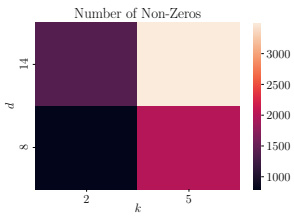
In this paper, we provide an analysis of projecting an affinity matrix onto the set of doubly stochastic matrices with the ℓ_2 distance metric, a technique commonly applied for spectral clustering yet whose theoretically properties have not been rigorously analyzed. In particular, we establish nec-



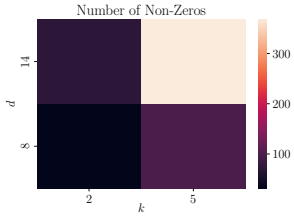
(a) Raw LSR Kernel



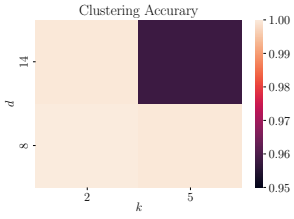
(b) LSR Kernel after DS



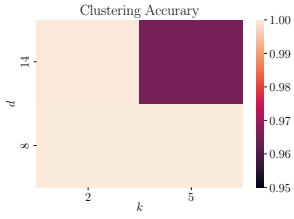
(c) Raw LSR Kernel



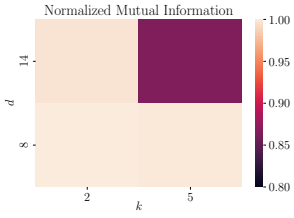
(d) LSR Kernel after DS



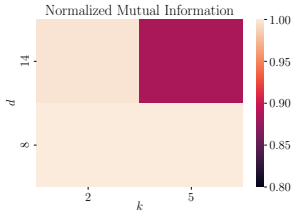
(e) Raw LSR Kernel



(f) LSR Kernel after DS



(g) Raw LSR Kernel



(h) LSR Kernel after DS

Figure 5. Feature detection error, number of non-zeros, clustering accuracy and normalized mutual information of the LSR kernel (Lu et al., 2012) before and after applying doubly stochastic projection, where the LSR kernel is applied on data from k subspaces of dimension d in $\mathbb{R}^{D=20}$, each containing $n = 50d$ points of unit norm. Both the subspaces and the points are sampled uniformly at random.

essary and sufficient conditions for the no-false-connection property and provide conditions on the input affinity matrix which will satisfy these conditions, along with conditions which guarantee the clusters will be connected. Moreover, in the case when the clusters are linear subspaces, we further provide analysis of the no-false-connection property and connectivity in terms of subspace dimensions, ratio of points within each subspace, and the angles between subspaces, via a continous extension of the doubly stochastic projection. Finally, via experiments under a variety of settings we compliment the theories and demonstrate the effect of doubly stochastic projection.

Acknowledgements

The authors are grateful to the anonymous reviewers for their comments on improving presentation and experiments. We thank Liangzu Peng for carefully proofreading a version of this manuscript. This work is supported by NSF grant 1704458 and Northrop Grumman Mission Systems Research in Applications for Learning Machines (REALM) initiative.

References

Aicher, C., Jacobs, A. Z., and Clauset, A. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, 2015.

Bruna, J. and Mallat, S. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, August 2013.

Elhamifar, E. and Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

Haeffele, B. D., You, C., and Vidal, R. A critique of Self-Expressive deep subspace clustering. In *International Conference on Learning Representations*, 2020.

Hagen, L. and Kahng, A. B. New spectral methods for ratio cut partitioning and clustering. *IEEE transactions on computer-aided design of integrated circuits and systems*, 11(9):1074–1085, 1992.

Ji, P., Salzmann, M., and Li, H. Efficient dense subspace clustering. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 461–468, March 2014.

Lim, D., Vidal, R., and Haeffele, B. D. Doubly Stochastic Subspace Clustering. 2020. URL http://arxiv.org/abs/2011.14859.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, January 2013.

Lorenz, D. A., Manns, P., and Meyer, C. Quadratically regularized optimal transport. *Applied Mathematics & Optimization*, 83(3):1919–1949, June 2021.

Lu, C.-Y., Min, H., Zhao, Z.-Q., Zhu, L., Huang, D.-S., and Yan, S. Robust and efficient subspace segmentation via least squares regression. In *European conference on computer vision*, pp. 347–360. Springer, 2012.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Soltanolkotabi, M. and Candés, E. J. A geometric analysis of subspace clustering with outliers. *Ann. Stat.*, 40(4): 2195–2238, 2012.

Soltanolkotabi, M., Elhamifar, E., and Candes, E. J. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

Tsakiris, M. C. and Vidal, R. Dual principal component pursuit. *Journal of Machine Learning Research*, 19:1–49, 2018.

Vidal, R. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

Vidal, R., Ma, Y., and Sastry, S. *Generalized Principal Component Analysis*. Springer Verlag, 2016.

Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

You, C. and Vidal, R. Geometric conditions for subspace-sparse recovery. In *International Conference on Machine Learning*, pp. 1585–1593. PMLR, 2015.

You, C., Robinson, D. P., and Vidal, R. Scalable sparse subspace clustering by orthogonal matching pursuit. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

Zass and Shashua. Doubly stochastic normalization for spectral clustering. *Advances in neural information processing systems*, 19, 2006.

A. Additional Proofs

Proof of Theorem 2.2.

Proof. (1 $\Rightarrow$ 2). Let $(\boldsymbol{A}^{(l)}, \boldsymbol{\alpha}^{(l)})$ be primal-dual optimal for DS-D($D^{(l)}, \frac{\eta}{\pi_l}$). By Proposition 2.1

$$\boldsymbol{A}^{(l)} = \frac{\pi_l}{\eta_2}[\boldsymbol{D}^{(l)} - \boldsymbol{\alpha}^{(l)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\alpha}^{(l)\top}]_+ \tag{32}$$

$$\frac{\pi_l}{\eta_2}[\boldsymbol{D}^{(l)} - \boldsymbol{\alpha}^{(l)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\alpha}^{(l)\top}]_+\mathbf{1} = n_l\mathbf{1}. \tag{33}$$

Assume now $\alpha_i^\circ + \beta_j^\circ \geq K_{ij}$ for any $i \not\sim j$. We verify if $(\boldsymbol{A}^\circ, \boldsymbol{\alpha}^\circ)$ are primal-dual optimal for P-DS($\boldsymbol{K}, \eta$). First, we check that $\frac{1}{\eta_2}[\boldsymbol{K} - \boldsymbol{\alpha}^\circ\mathbf{1}^\top - \mathbf{1}\boldsymbol{\alpha}^{\circ\top}]_+ = \boldsymbol{A}^\circ$. For any i, j coming from the same cluster l', $\frac{1}{\eta_2}[K_{ij} - \alpha_i^\circ - \alpha_j^\circ]_+ = \frac{1}{\pi_{l'}}\frac{\pi_{l'}}{\eta_2}[D^{(l')}_{i'j'} - \alpha^{(l')}_{i'} - \alpha^{(l')}_{j'}]_+ = \frac{1}{\pi_{l'}}A^{(l')}_{i'j'} = A^\circ_{ij}$, where $i' = i - n_1 - \cdots - n_{l'-1}$, $j' = j - n_1 - \cdots - n_{l'-1}$ are the counterpart of i, j for indexing $\boldsymbol{D}^{(l')}$. For i, j from different clusters, we have $\frac{1}{\eta_2}[K_{ij} - \alpha_i^\circ - \alpha_j^\circ]_+ = 0 = A^\circ_{ij}$. Now, we only need to verify that $\boldsymbol{A}^\circ\mathbf{1} = n\mathbf{1}$. This is immediate following the construction of A° and (33).

(2 $\Rightarrow$ 3). This is trivial.

(3 $\Rightarrow$ 1). Let A° be the primal optimal and α° be a dual optimal for DS-D($\boldsymbol{K}, \eta$). By Proposition 2.1,

$$\boldsymbol{A}^\circ = \frac{1}{\eta_2}[\boldsymbol{K} - \boldsymbol{\alpha}^\circ\mathbf{1}^\top - \mathbf{1}\boldsymbol{\alpha}^{\circ\top}]_+ \tag{34}$$

$$A^\circ\mathbf{1} = n\mathbf{1} \tag{35}$$

By assumption, A° has no false connections, i.e., $A^\circ_{ij} = [K_{ij} - \alpha_i^\circ - \alpha_j^\circ]_+ = 0$ for all $i \not\sim j$. That is, $\alpha_i^\circ + \alpha_j^\circ \geq K^\dagger_{ij}$. This fact, together with (35), yields $\frac{1}{\eta_2}[\boldsymbol{D}^{(l)} - \boldsymbol{\alpha}^{(l)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\alpha}^{(l)\top}]_+\mathbf{1} = n_l\mathbf{1}$ for all $l \in \{1, \ldots, k\}$. Therefore, $\boldsymbol{\alpha}^{(l)}$ is dual optimal for DS-D($D^{(l)}$). □

Lemma A.1. *Suppose K is rotational invariant, i.e., $K(\boldsymbol{y}, \boldsymbol{z}) = K(\boldsymbol{R}\boldsymbol{y}, \boldsymbol{R}\boldsymbol{z})$ for any orthogonal $\boldsymbol{R}$. For any dimension d and constant η_0, there exists a constant α such that $\int_{\boldsymbol{y}\in\mathbb{S}^{d-1}}[K(\boldsymbol{y}, \boldsymbol{z}) - 2\alpha]_+ d\mu_{\mathbb{S}^{d-1}} = \eta_0$.*

Proof. Taking $\boldsymbol{R}$ to be the householder matrix that sends $\boldsymbol{z}$ to $\boldsymbol{e}_1$ gives

$$\int_{\boldsymbol{y}\in\mathbb{S}^{d-1}}[K(\boldsymbol{y}, \boldsymbol{z}) - 2\gamma]_+ d\mu_{\mathbb{S}^{d-1}} = \int_{\boldsymbol{y}\in\mathbb{S}^{d-1}}[K(\boldsymbol{R}\boldsymbol{y}, \boldsymbol{R}\boldsymbol{z}) - 2\gamma]_+ d\mu_{\mathbb{S}^{d-1}} \quad \text{(K symmetric)} \tag{36}$$

$$= \int_{\boldsymbol{y}\in\mathbb{S}^{d-1}}[K(\boldsymbol{R}\boldsymbol{y}, \boldsymbol{e}_1) - 2\gamma]_+ d\mu_{\mathbb{S}^{d-1}} \quad \text{(Householder)} \tag{37}$$

$$= \int_{\boldsymbol{y}\in\mathbb{S}^{d-1}}[K(\boldsymbol{y}, \boldsymbol{e}_1) - 2\gamma]_+ d\mu_{\mathbb{S}^{d-1}} \quad \text{(Change of basis[14])} \tag{38}$$

This is to say there exist a $\gamma \in [0, 1]$, such that for any $z \in \mathbb{S}^{d-1}$, $\int_{y\in\mathbb{S}^{d-1}}[K(y, z) - 2\gamma]_+ d\mu_{\mathbb{S}^{d-1}} = \eta_2$, hence taking α to be a constant function will satisfy the optimality condition. □

Notes on Definition 3.3. For any dimension $d \in \mathbb{Z}_{\geq 2}$, let $z_1^2 + z_2^2 + ... + z_d^2 = 1$ be the coordinate representation of the unit sphere $\mathbb{S}^{d-1}$ of $\mathbb{R}^d$. Let $\alpha \in [0, 1]$. The goal is to compute $\rho_d(\alpha) = \int_{\mathbb{S}^{d-1}}[|z_d| - 2\alpha]_+ d\mu_{\mathbb{S}^{d-1}}$. Note that

$$\rho_d(\alpha) = \frac{\int_{\mathbb{S}^{d-1}}[|z_d| - 2\alpha]_+ dS}{\int_{\mathbb{S}^{d-1}} dS} = \frac{1}{A_{d-1}}\int_{\mathbb{S}^{d-1}}[|z_d| - 2\alpha]_+ dS \tag{39}$$

$$= \frac{2}{A_{d-1}}\int_{\mathbb{S}^{d-1}\,\cap\, z_d\geq 0}[z_d - 2\alpha]_+ dS \tag{40}$$

$$= \frac{2}{A_{d-1}}\int_{\mathbb{S}^{d-1}\,\cap\, z_d\geq 2\alpha}(z_d - 2\alpha)\, dS \tag{41}$$

[14] Let $x = Ry$, and note that $\mu_{\mathbb{S}^{d-1}}(dy) = \mu_{\mathbb{S}^{d-1}}(dR^{-1}x) = \mu_{\mathbb{S}^{d-1}}(dx)$, where the second equality is due to the fact that $\mu_{\mathbb{S}^{d-1}}$ is a uniform measure.

where A_{d-1} is the surface area of $\mathbb{S}^{d-1}$. For $z_d \geq 0$, we have the equation of the sphere as $z_d = \sqrt{1-(z_1^2+\cdots+z_{d-1}^2)} := g(z_1,\ldots,z_{d-1}^2)$, which yields the following identities

$$z_d \geq 2\alpha \Leftrightarrow \sqrt{z_1^2+\cdots+z_{d-1}^2} \leq \sqrt{1-4\alpha^2} \tag{42}$$

$$\sqrt{(\frac{\partial g}{\partial z_1})^2+\cdots+(\frac{\partial g}{\partial z_{d-1}})^2+1} = \frac{1}{\sqrt{1-(z_1^2+\cdots+z_{d-1}^2)}}. \tag{43}$$

With those, one can perform a change of measure to (41) and get

$$\rho_d(\alpha) = \frac{2}{A_{d-1}} \int_{\sqrt{z_1^2+\cdots+z_{d-1}^2} \leq \sqrt{1-4\alpha^2}} 1 - \frac{2\alpha}{\sqrt{1-(z_1^2+\cdots+z_{d-1}^2)}}\, dA \tag{44}$$

$$= \frac{2}{A_{d-1}} \left(V_{d-1}\sqrt{1-4\alpha^2}^{d-1} - 2\alpha S_{D-2} \int_0^{\sqrt{1-4\alpha^2}} \frac{r^{D-2}}{\sqrt{1-r^2}}\, dr \right), \tag{45}$$

where V_{d-1} is the $(d-1)$-dimensional volume of $\mathbb{S}^{d-2}$. Note that when $\alpha = 0$, (45) gives $\rho_d(\alpha) = \frac{2V_{d-1}}{A_{d-1}}$, which coincides with the c_d quantity as in (Tsakiris & Vidal, 2018). On the other hand, for a general $\alpha \in [0,1]$, we further have

$$\rho_d(\alpha) = \frac{2}{A_{d-1}} \left(V_{d-1}\sqrt{1-4\alpha^2}^{d-1} - 2\alpha A_{D-2} \int_0^{\sin^{-1}(\sqrt{1-4\alpha^2})} \sin\theta^{D-2}\, d\theta \right) \tag{46}$$

$$= \frac{2}{A_{d-1}} \left(V_{d-1}\sqrt{1-4\alpha^2}^{d-1} - \frac{2\alpha A_{D-2}}{d-1}\sqrt{1-4\alpha^2}^{D-1}\, {}_2F_1(\frac{1}{2},\frac{d-1}{2},\frac{D+1}{2},1-4\alpha^2) \right) \tag{47}$$

$$= \frac{2}{A_{d-1}}\sqrt{1-4\alpha^2}^{d-1} \left(V_{d-1} - \frac{2\alpha A_{D-2}}{d-1}\, {}_2F_1(\frac{1}{2},\frac{d-1}{2},\frac{D+1}{2},1-4\alpha^2) \right), \tag{48}$$

where ${}_2F_1$ is the hypergeometric function. Since the last term is hard to bound and interpret, we observe that (41) can be alternatively written as

$$\rho_d(\alpha) = \frac{2}{A_{d-1}} \int_{\mathbb{S}^{d-1} \cap\, z_d \geq 2\alpha} z_d\, dS - \frac{4\alpha}{A_{d-1}} \int_{\mathbb{S}^{d-1} \cap\, z_d \geq 2\alpha} dS \tag{49}$$

$$= \frac{2}{A_{d-1}} V_{d-1}\sqrt{1-4\alpha^2}^{d-1} - \frac{4\alpha}{A_{d-1}} \int_{\mathbb{S}^{d-1} \cap\, z_d \geq 2\alpha} dS, \tag{50}$$

where the second term is some scalar times the surface area of a spherical cap. With that, we have

$$\rho_d(\alpha) = \frac{2V_{d-1}}{A_{d-1}}\sqrt{1-4\alpha^2}^{d-1} - \frac{4\alpha}{A_{d-1}}\frac{1}{2}A_{d-1}I_{1-4\alpha^2}(\frac{d-1}{2},\frac{1}{2}) \tag{51}$$

$$= c_d(1-4\alpha^2)^{\frac{d-1}{2}} - 2\alpha I_{1-4\alpha^2}(\frac{d-1}{2},\frac{1}{2}), \tag{52}$$

where I is the regularized beta function.

Observe that as α increases from 0 to $\frac{1}{2}$, $\rho_d(\alpha)$ decreases from c_d to 0. Thus, it may be desirable to say $1 - \frac{\rho_d(\alpha)}{c_d} = O(\alpha^{\cdots})$. Indeed, we have

$$1 - \frac{\rho_d(\alpha)}{c_d} = 1 - (1-4\alpha^2)^{\frac{d-1}{2}} + \frac{2}{c_d}\alpha I_{1-4\alpha^2}(\frac{d-1}{2},\frac{1}{2}) \tag{53}$$

$$= 1 - \sum_{i=0}^{\frac{d-1}{2}} \binom{\frac{d-1}{2}}{i} (-1)^i (4\alpha^2)^i + \frac{2}{c_d}\alpha I_{1-4\alpha^2}(\frac{d-1}{2},\frac{1}{2}) \tag{54}$$

$$= -\sum_{i=1}^{\frac{d-1}{2}} \binom{\frac{d-1}{2}}{i} (-1)^i (4\alpha^2)^i + \frac{2}{c_d}\alpha I_{1-4\alpha^2}(\frac{d-1}{2},\frac{1}{2}) \tag{55}$$

Assuming d is odd, i.e., $\frac{d-1}{2}$ is integer, from the property of I we have

$$(55) = -\sum_{i=1}^{\frac{d-1}{2}} \binom{\frac{d-1}{2}}{i} (-1)^i (4\alpha^2)^i + \frac{2}{c_d}\alpha \left\{ 1 - \frac{2\alpha}{B(\frac{d-1}{2}, \frac{1}{2})} \sum_{i=0}^{\frac{d-1}{2}-1} (-1)^i \binom{\frac{d-1}{2}-1}{i} \frac{(4\alpha^2)^i}{i+\frac{1}{2}} \right\} \tag{56}$$

$$= \frac{2}{c_d}\alpha - \sum_{i=1}^{\frac{d-1}{2}} \binom{\frac{d-1}{2}}{i} (-1)^i (4\alpha^2)^i - \frac{4\alpha^2}{c_d B(\frac{d-1}{2}, \frac{1}{2})} \sum_{i=0}^{\frac{d-1}{2}-1} (-1)^i \binom{\frac{d-1}{2}-1}{i} \frac{(4\alpha^2)^i}{i+\frac{1}{2}} \tag{57}$$

$$= \frac{2}{c_d}\alpha - \sum_{i=1}^{\frac{d-1}{2}} \binom{\frac{d-1}{2}}{i} (-1)^i (4\alpha^2)^i - \frac{1}{c_d B(\frac{d-1}{2}, \frac{1}{2})} \sum_{i=0}^{\frac{d-1}{2}-1} (-1)^i \binom{\frac{d-1}{2}-1}{i} \frac{(4\alpha^2)^{i+1}}{i+\frac{1}{2}} \tag{58}$$

$$= \frac{2}{c_d}\alpha - \sum_{i=1}^{\frac{d-1}{2}} \binom{\frac{d-1}{2}}{i} (-1)^i (4\alpha^2)^i - \frac{1}{c_d B(\frac{d-1}{2}, \frac{1}{2})} \sum_{i=1}^{\frac{d-1}{2}} (-1)^{i-1} \binom{\frac{d-1}{2}-1}{i-1} \frac{(4\alpha^2)^i}{i-\frac{1}{2}} \tag{59}$$

$$= \frac{2}{c_d}\alpha - \sum_{i=1}^{\frac{d-1}{2}} (4\alpha^2)^i \left\{ \binom{\frac{d-1}{2}}{i} (-1)^i - \frac{1}{c_d B(\frac{d-1}{2}, \frac{1}{2})} (-1)^{i-1} \binom{\frac{d-1}{2}-1}{i-1} \frac{1}{i-\frac{1}{2}} \right\} \tag{60}$$