Iterative Spectral Method for Alternative Clustering

Chieh Wu   **Stratis Ioannidis**   **Mario Sznaier**
Xiangyu Li   **David Kaeli**   **Jennifer G. Dy**
Electrical and Computer Engineering Dept., Northeastern University, Boston, MA

Abstract

Given a dataset and an existing clustering as input, alternative clustering aims to find an alternative partition. One of the state-of-the-art approaches is Kernel Dimension Alternative Clustering (KDAC). We propose a novel Iterative Spectral Method (ISM) that greatly improves the scalability of KDAC. Our algorithm is intuitive, relies on easily implementable spectral decompositions, and comes with theoretical guarantees. Its computation time improves upon existing implementations of KDAC by as much as 5 orders of magnitude.

1   Introduction

Clustering, i.e., the process of grouping similar objects in a dataset together, is a classic problem. It is extensively used for exploratory data analysis. Traditional clustering algorithms typically identify a single partitioning of a given dataset. However, data is often multi-faceted and can be both interpreted and clustered through multiple viewpoints (or, *views*). For example, the same face data can be clustered based on either identity or based on pose. In real applications, partitions generated by a clustering algorithm may not correspond to the view a user is interested in.

In this paper, we address the problem of finding an *alternative clustering*, given a dataset and an existing, pre-computed clustering. Ideally, one would like the alternative clustering to be *novel* (i.e., non-redundant) w.r.t. the existing clustering to reveal a new viewpoint to the user. Simultaneously, one would like the result to reveal partitions of high clustering *quality*. Several recent papers propose algorithms for alternative

clustering [1–6]. Among them, Kernel Dimension Alternative Clustering (KDAC) is a flexible approach, shown to have superior performance compared to several competitors [6]. KDAC is as powerful as spectral clustering in discovering arbitrarily-shaped clusters (including ones that are not linearly separable) that are non-redundant w.r.t. an existing clustering. As an additional advantage, KDAC can simultaneously learn the subspace in which the alternative clustering resides.

The flexibility of KDAC comes at a price: the KDAC formulation involves optimizing a non-convex cost function constrained over the space of orthogonal matrices (i.e, the Stiefel manifold). Niu et al. [6] proposed a Dimension Growth (DG) heuristic for solving this optimization problem, which is nevertheless highly computationally intensive. We elaborate on its complexity in Section 2; experimentally, DG is quite slow, with a convergence time of roughly 46 hours on an Intel Xeon CPU, for a 624 sample-sized face data (c.f. Section 4). This limits the applicability of KDAC in interactive exploratory data analysis settings, which often require results to be presented to a user within a few seconds. It also limits the scalability of KDAC to large data. Alternately, one can solve the KDAC optimization problem by gradient descent on a Stiefel manifold (SM) [7]. However, given the lack of convexity, both DG or SM are prone to get trapped to local minima. Multiple iterations with random initializations are required to ameliorate the effect of locality. This increases computation time, and also decreases in effectiveness as the dimensionality of the data increases: the increase in dimension rapidly expands the search space and the abundance of local minima. As such, with both DG and SM, the clustering quality is negatively affected by an increase in dimension.

Our Contributions. Motivated by the above issues, we make the following contributions:

- We propose an Iterative Spectral Method (ISM), a *novel algorithm* for solving the non-convex optimization constrained on a Stiefel manifold problem inherent in KDAC. Our algorithm has several highly de-

sirable properties. First, it *significantly outperforms* traditional methods such as DG and SM in terms of both computation time and quality of the produced alternative clustering. Second, the algorithm relies on an *intuitive use of iterative spectral decompositions*, making it both easy to understand as well as easy to implement, using off-the-shelf libraries.

- ISM has a natural initialization, constructed through a Taylor approximation of the problem's Lagrangian. Therefore, high quality results can be obtained without random restarts in search of a better initialization. We show that this initialization is a contribution in its own right, as its use improves performance of competitor algorithms.
- We provide *theoretical guarantees* on its fixed point. In particular, we establish conditions under which the fixed point of ISM satisfies both the 1st and 2nd order necessary conditions for local optimality.
- We extensively evaluate the performance of ISM in solving KDAC with synthetic and real data under various clustering quality and cost measures. Our results show an improvement in execution time by up to a factor of roughly 70 and 10^5, compared to SM and DG, respectively. At the same time, ISM outperforms SM and DG in clustering quality measures along with significantly lower computational cost.

Related Work. There exist two general modes of discovering alternative clusterings – simultaneously or iteratively. Simultaneous approaches find the multiple alternative clusterings at the same time [3, 8–13]. Iterative approaches find an alternative clustering given existing clustering [2]. Since this work focuses on the iterative paradigm, we elaborate on the related work along these lines. Alternative clustering methods differ in how they measure novelty and cluster quality. Gondek and Hofmann [1] find an alternative clustering by conditional information (CI) bottleneck. Bae and Bailey [14] perform agglomerative clustering with cannot-link constraints imposed on the data points that belong together in the existing clustering. Cui et al. [5] find an alternative clustering by projecting the data to a subspace orthogonal to the existing clustering. Qi and Davidson [15] search for novelty by minimizing the Kullback-Leiber (KL) divergence between the original data and the transformed data subject to the constraint that the sum-squared-error between samples in the projected space with the existing clusters is small. Dang and Bailey [16] find quality clusters by maximizing the mutual information (MI) between the alternative clusters and the data while simultaneously ensuring novelty by minimizing the MI between alternative and existing clusterings.

KDAC [6] discovers an alternative clustering by max-

imizing for cluster quality based on the spectral clustering objective and at the same time maximizing for novelty based on a non-linear dependence measure, HSIC [17], on the projected subspace of the alternative clustering. KDAC's ability to detect arbitrarily-shaped clusters is due to its use of the Hilbert-Schmidt Independence Criterion (HSIC) [17] as a cluster quality measure. HSIC is motivated by the objective function of spectral clustering. Moreover, since HSIC models non-linear dependence, it is also utilized by KDAC to measure novelty. In contrast, e.g., the orthogonal subspace projection approach in [5] is limited, as it only captures linear dependencies. Other approaches, such as [1, 16], can take non-linear dependencies into account by utilizing information theoretic measures. However, doing so requires estimating joint probability distributions. The advantage of KDAC over such approaches is that it utilizes HSIC for measuring novelty and cluster quality, which can capture non-linear dependencies through kernels, without having to explicitly learn the joint probability distributions; empirically, it significantly outperforms aforementioned schemes in clustering quality [6].

2 Kernel Dimension Alternative Clustering (KDAC)

In alternative clustering, a dataset is provided along with existing clustering labels. Given this as input, we seek a *new* clustering that is (a) distinct from the existing clustering, and (b) has high quality with respect to a clustering quality measure. An example illustrating this is shown in Figure 1. This dataset comprises 400 points in $\mathcal{R}^4$. Projected to the first two dimensions, the dataset contains two clusters of intertwining parabolas shown as Clustering A. Projected to the last two dimensions, the dataset contains two Gaussian clusters shown as Clustering B. Points clustered together in one view can be in different clusters in the alternative view. In alternative clustering, given (a) the dataset, and (b) one of the two possible clusterings (e.g., Clustering B), we wish to discover the alternative clustering illustrated by the different view.
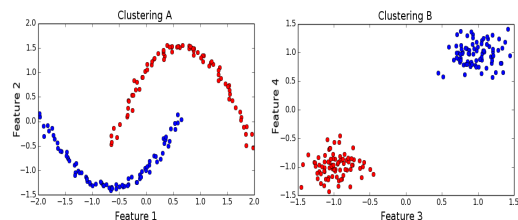

Figure 1: Four-dimensional moon dataset. Projection into the first two dimensions reveals different clusters than projection to the latter two dimensions.

Formally, let $X \in \mathcal{R}^{n \times d}$ be a dataset with n samples and d features, along with an existing clustering $Y \in \mathcal{R}^{n \times k}$, where k is the number of clusters. If x_i belongs to cluster j, then $Y_{i,j} = 1$; otherwise, $Y_{i,j} = 0$. We wish to discover an alternative clustering $U \in \mathcal{R}^{n \times k}$ on some lower dimensional subspace of dimension $q \ll d$. Let $W \in \mathcal{R}^{d \times q}$ be a projection matrix such that $XW \in \mathcal{R}^{n \times q}$.

We seek the optimal projection W and clustering U that maximizes the statistical dependence between XW with U, yielding a high clustering quality, while minimizing the dependence between XW and Y, ensuring the novelty of the new clustering. Denoting DM as a Dependence Measure function, and using λ as a weighing constant, this optimization can be written as:

$$\text{Maximize:} \quad \text{DM}(XW, U) - \lambda \, \text{DM}(XW, Y), \quad (1a)$$

$$\text{s.t}: \quad W^T W = I, U^T U = I. \quad (1b)$$

As in spectral clustering, the labels of the alternative clustering are retrieved by performing K-means on matrix U, treating its rows as samples. There are many potential choices for DM. The most well-known measures are correlation and mutual information (MI). While correlation performs well in many applications, it lacks the ability to measure non-linear relationships. Although there is clear relationship in Clustering A in Figure 1, correlation would mistakenly yield a value of nearly 0. As a dependence measure, MI is superior in that it also measures non-linear relationships. However, due to the probabilistic nature of its formulation, a joint distribution is required. Depending on the distribution, the computation of MI can be prohibitive.

For these reasons, the Hilbert Schmidt Independence Criterion (HSIC) [17] has been proposed for KDAC [6]. Like MI, it captures non-linear relationships. Unlike MI, HSIC does not require estimating a joint distribution, and it relaxes the need to discretize continuous variables. In addition, as shown by Niu et al. [6], HSIC is mathematically equivalent to spectral clustering, further implying that a high HSIC between the data and U yields high clustering quality. A visual comparison of HSIC and correlation can be found in Figure 5 of Appendix I in the supplement.

Using HSIC as a dependence measure, the objective of KDAC becomes

$$\text{Maximize:} \quad \text{HSIC}(XW, U) - \lambda \, \text{HSIC}(XW, Y), \quad (2a)$$

$$\text{subject to:} \quad W^T W = I, U^T U = I. \quad (2b)$$

where $\text{HSIC}(X, Y) \equiv \frac{1}{(n-1)^2} \text{Tr}(K_X H K_Y H)$. Here, the variables K_X and K_Y are Gram matrices, and the H matrix is a centering matrix where $H =$

Algorithm 1: KDAC Algorithm

Input : dataset X, original clustering Y
Output: alternative clustering Y
Initialize W_0 using W_{init} from (12)
Initialize U_0 from original clustering
Initialize D_0 from W and original clustering
while *(U not converged) or (W not converged)* **do**
    Update D
    Update W by solving Equation (4)
    Update U by solving Equation (3)
Clustering Result $\leftarrow$ Apply K-means to U

Algorithm 2: ISM Algorithm

Input : U, D, X, Y
Output: W^*
Initialize W_0 to the previous value of W in the master loop of KDAC.
while W *not converged* **do**
    $W \leftarrow \text{eig}_{\min}(\Phi(W))$;

$I - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ with $\mathbf{1}$ the n-sized vector of all ones. The elements of K_X and K_Y are calculated by kernel functions $k_X(x_i, x_j)$ and $k_Y(y_i, y_j)$. The kernel functions for Y and U used in KDAC are $K_Y = YY^T$ and $K_U = UU^T$, and the kernel function for XW is the Gaussian $k_{XW}(x_i, x_j) = \exp(-\text{Tr}[(x_i - x_j)^T WW^T (x_i - x_j)]/(2\sigma^2))$. Due to the equivalence of HSIC and spectral clustering, the practice of normalizing the kernel K_{XW} is adopted from spectral clustering by Niu et al. [6]. That is, for K_{XW} the unnormalized Gram matrix, the normalized matrix is defined as $D^{-1/2} K_{XW} D^{-1/2}$ where $D = \text{diag}(\mathbf{1}_n^T K_{XW})$ is a diagonal matrix whose elements are the column-sums of K_{XW}.

KDAC Algorithm. The optimization problem (2) is non-convex. The KDAC algorithm solves (2) using alternate maximization between the variables U, W and D, updating each while holding the other two fixed. After convergence, motivated by spectral clustering, U is discretized via K-means to provide the alternative clustering. The algorithm proceeds in an iterative fashion, summarized in Algorithm 1. In each iteration, variables D, U, and W are updated as follows:

Updating D: While holding U and W constant, D is computed as $D = \text{diag}(\mathbf{1}_n^T K_{XW})$. Matrix D is subsequently treated as a scaling constant throughout the rest of the iteration.

Updating U: Holding W and D constant and solving for U, (2) reduces to :

$$\max_{U:U^T U=I} \text{Tr}(U^T \mathcal{Q} U), \quad (3)$$

where $\mathcal{Q} = HD^{-1/2} K_{XW} D^{-1/2} H$. This is precisely spectral clustering [18]: (3) can be solved by setting U's columns to the k most dominant eigenvectors of $\mathcal{Q}$, which can be done in $O(n^3)$ time.

Updating W: While holding U and D constant to

solve for W, (2) reduces to:

Minimize: $\quad F(W) = -\sum_{i,j} \gamma_{i,j} e^{-\frac{\text{Tr}[W^T A_{i,j} W]}{2\sigma^2}}$ (4a)

subject to: $\quad W^T W = I$ (4b)

where $\gamma_{i,j}$ are the elements of matrix $\gamma = D^{-1/2} H(UU^T - \lambda YY^T) H D^{-1/2}$, and $A_{i,j} = (x_i - x_j)(x_i - x_j)^T$ (see Appendix A in the supplement for the derivation). This objective, along with a Stiefel Manifold constraint, $W^T W = I$, pose a challenging optimization problem as neither is convex. Niu et al. [6] propose solving (4) through an algorithm termed Dimensional Growth (DG). This algorithm solves for W by computing individual columns of W separately through gradient descent (GD). Given a set of computed columns, the next column is computed by GD projected to a subspace orthogonal to the span of computed set. Since DG is based on GD, the computational complexity is dominated by computing the gradient of (4a). The latter is given by:

$$\nabla F(W) = \sum_i^n \sum_j^n \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\text{Tr}[W^T A_{i,j} W]}{2\sigma^2}} A_{i,j} W. \quad (5)$$

The complexity of DG is $O(t_{DG} n^2 d^2 q)$, where n, d are the dataset size and dimension, respectively, q is the dimension of the subspace of the alternative clustering, and t_{DG} is the number of iterations of gradient descent. The calculation of the gradient contributes the term $O(n^2 d^2 q)$. Although this computation is highly parallelizable, the algorithm still suffers from slow convergence rate. Therefore, t_{DG} often dominates the computation cost.

An alternative approach to optimize (4) is through classic methods for performing optimization on the Stiefel Manifold (SM) [7]. The computational complexity of this algorithm is dominated by the computation of the gradient and a matrix inversion with t_{SM} iterations. This yields a complexity of $O(t_{SM} n^2 d^2 + t_{SM} d^3)$ for SM. Finally, as gradient methods applied to a non-convex objective, both SM and DG require multiple executions from random initialization points to find improved local minima. This approach becomes less effective as the dimension d increases.

3 An Iterative Spectral Method

The computation of KDAC is dominated by the W updates in Algorithm 1. Instead of using DG or SM to solve the optimization problem for W in KDAC, we propose an Iterative Spectral Method (ISM). Our algorithm is motivated from the following observations.

The Lagrangian of (4) is:

$$\mathcal{L}(W, \Lambda) = -\sum_{i,j} \gamma_{i,j} \exp\left(-\frac{\text{Tr}(W^T A_{i,j} W)}{2\sigma^2}\right)$$
$$- \frac{1}{2} \text{Tr}(\Lambda(W^T W - I)) \quad (6)$$

Setting $\nabla_W \mathcal{L}(W, \Lambda) = 0$ gives us the equation:

$$\Phi(W)W = W\Lambda, \quad (7)$$

where

$$\Phi(W) = \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} \exp(-\frac{\text{Tr}[W^T A_{i,j} W]}{2\sigma^2}) A_{i,j}, \quad (8)$$

and Λ is a diagonal matrix. Recall that a feasible W, satisfying (4b), is orthonormal. (7) is an eigenequation; thus, a stationary point W of the Lagrangian (6) comprises of q eigenvectors of $\Phi(W)$ as columns. Motivated by this observation, ISM attempts to find such a W in the following iterative fashion. Let W_0 be an initial matrix. Given W_k at iteration k, the matrix W_{k+1} is computed as:

$$W_{k+1} = \text{eig}_{\min}(\Phi(W_k)), \quad k = 0, 1, 2, \ldots,$$

where the operator $\text{eig}_{\min}(A)$ returns a matrix whose columns are the q eigenvectors corresponding to the smallest eigenvalues of A.

ISM is summarized in Alg. 2. Several important observations are in order. First, the algorithm ensures that W_k, for $k \geq 1$, is feasible, by construction: selected eigenvectors are orthonormal and satisfy (4b). Second, it is also easy to see that a fixed point of the algorithm will also be a stationary point of the Lagrangian (6) (see also Lemma 3). Though it is harder to prove, selecting eigenvectors corresponding to the *smallest* eigenvalues is key: we show that this is precisely the property that relates a fixed point of the algorithm to the local minimum conditions (see Thm. 1). Finally, ISM has several computational advantages. For t_{ISM} iterations, the calculation of $\Phi(W)$, and the ensuing eigendecomposition yields a complexity of $O(t_{ISM}(n^2 d^2 + d^3))$. Since $q \ll d$, various approximation methods [19] [20] [21] can be employed to find the few eigenvectors. For example, the Coordinate-wise Power Method [21], approximates the most dominant eigenvalue at $O(d)$ time, reducing ISM's complexity to $O(t_{ISM} n^2 d^2)$. This improvement is further confirmed experimentally (see Figure 3). Lastly, t_{ISM} is magnitudes smaller than both t_{DG} and t_{SM}. In general $t_{ISM} < 10$, while $t_{SM} > 50$ and $t_{DG} > 200$.

3.1 Convergence Guarantees

As mentioned above, the selection of the eigenvectors corresponding to the *smallest* eigenvalues of $\Phi(W_k)$ is crucial for the establishment of a stationary point. Namely, we establish the following theorem:

Theorem 1. *For large enough σ (satisfying Inequality* (10)*), a fixed point W^* of Algorithm 2 satisfies the necessary conditions of a local minimum of* (4) *if $\Phi(W^*)$ is full rank.*

Proof. The main body of the proof is organized into a series of lemmas proved in the supplement. Our first auxiliary lemma (from [22]), establishes conditions necessary for a stationary point of the Lagrangian to constitute local minimum.

Lemma 1. *[Nocedal,Wright, Theorem 12.5 [22]] (2nd Order Necessary Conditions) Consider the optimization problem:* $\min_{W:h(W)=0} f(W)$, *where $f : \mathbb{R}^{d \times q} \to \mathbb{R}$ and $h : \mathcal{R}^{d \times q} \to \mathcal{R}^{q \times q}$ are twice continuously differentiable. Let $\mathcal{L}$ be the Lagrangian of this optimization problem. Then, a local minimum must satisfy the following conditions:*

$$\nabla_W \mathcal{L}(W^*, \Lambda^*) = 0, \tag{9a}$$

$$\nabla_\Lambda \mathcal{L}(W^*, \Lambda^*) = 0, \tag{9b}$$

$$\text{Tr}(Z^T \nabla^2_{WW} \mathcal{L}(W^*, \Lambda^*) Z) \geq 0$$
$$\text{for all } Z \neq 0, \text{with } \nabla h(W^*)^T Z = 0. \tag{9c}$$

Armed with this result, we next characterize the properties of a fixed point of Algorithm 2:

Lemma 2. *Let W^* be a fixed point of Algorithm 2. Then it satisfies:* $\Phi(W^*)W^* = W^* \Lambda^*$, *where $\Lambda^* \in \mathcal{R}^{q \times q}$ is a diagonal matrix containing the q smallest eigenvalues of $\Phi(W^*)$ and $W^{*T}W^* = I$.*

The proof can be found in Appendix B. Our next result, whose proof is in Appendix C, states that a fixed point satisfies the 1st order conditions of Lemma 1.

Lemma 3. *If W^* is a fixed point and Λ^* is as defined in Lemma 2, then W^*, Λ^* satisfy the 1st order conditions (9a)(9b) of Lemma 1.*

Our last lemma, whose proof is in Appendix D, establishes that a fixed point satisfies the 2nd order conditions of Lemma 1, for large enough σ.

Lemma 4. *If W^* is a fixed point, Λ^* is as defined in Lemma 2, and $\Phi(W^*)$ is full rank, then given a large enough σ (satisfying Inequality* (10)*), W^* and Λ^* satisfy the 2nd order condition (9c) of Lemma 1.*

Thm. 1 therefore follows. $\square$

Thm. 1 is stated in terms of a large enough σ; we can characterize this constraint precisely. In the proof of

Lemma 4 we establish the following condition on σ:

$$\sigma^2 [\min_i (\bar{\Lambda}^*{}_i) - \max_j(\Lambda^*_j)] \geq$$

$$\sum_{i,j} \frac{|\gamma_{i,j}|}{\sigma^2} e^{-\frac{\text{Tr}((W^{*T} A_{i,j} W^*)}{2\sigma^2}} \text{Tr}(A_{i,j}^T A_{ij}). \tag{10}$$

Here, Λ^* is the set of q smallest eigenvalues of $\Phi(W)$, and $\bar{\Lambda}^*$ is the set of the remaining eigenvalues. The left-hand side (LHS) of the equation further motivates ISM's choice of eigenvectors corresponding to the q smallest eigenvalues. This selection guarantees that the LHS of the inequality is positive. Therefore, given a large enough σ, Inequality (10) and the 2nd order condition is satisfied.

Furthermore, this equation provides a reasonable suggestion for the value of q. Since we wish to maximize the term $(\min_i(\bar{\Lambda}^*{}_i) - \max_j(\Lambda^*_j))$ to satisfy the inequality, the value q should be set where this gap is maximized. More formally, we will defined

$$\delta_{gap} = \min_i(\bar{\Lambda}^*{}_i) - \max_j(\Lambda^*_j). \tag{11}$$

as the eigengap.

3.2 Spectral Initialization via Taylor Approximation

ISM admits a natural initialization point, constructed via a Taylor approximation of the objective. As we show experimentally in Section 4, this initialization is a contribution in its own right: it improves both clustering quality and convergence time for ISM *as well as* competitor algorithms. To obtain a good initialization, observe that by using the 2nd order Taylor approximation of the objective function (4a) at $W = 0$, the Lagrangian can be approximated by

$$\tilde{\mathcal{L}}(W, \Lambda) \approx -\sum_{i,j} \gamma_{i,j} \left(1 - \frac{\text{Tr}(W^T A_{i,j} W)}{2\sigma^2} \right)$$
$$+ \frac{1}{2} \text{Tr}(\Lambda(I - W^T W)).$$

Setting $\nabla_W \tilde{\mathcal{L}}(W, \Lambda) = 0$ reduces the problem into a simple eigendecomposition, namely, the one defined by the system $\left[\sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} A_{i,j} \right] W = W\Lambda$. Hence, the 2nd order Taylor approximation of the original cost objective has a closed form global minimum that can be used as an initialization point, namely:

$$W_{\text{init}} = \text{eig}_{\min}(\sum_{i,j} \gamma_{i,j} A_{i,j}/\sigma^2). \tag{12}$$

We use this spectral initialization (SI) in the first master iteration of KDAC. In subsequent master iterations, W_0 (the starting point of ISM) is set to be the last value to which ISM converged to previously.
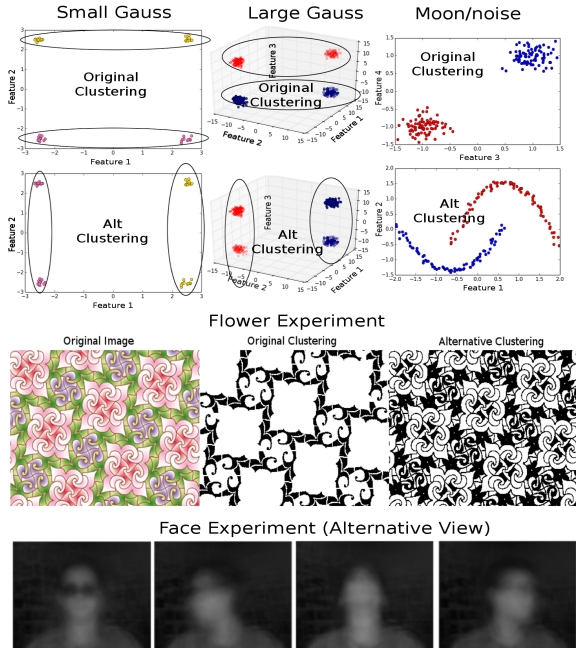
Figure 2: Figure includes all original and alternative displayable clusterings. The face data is originally clustered by the identity of the individual. Here the average image of the alternative clustering is displayed. It is observable that the original and alternative clusters are all visually obvious clusters and provide alternative views.

4 Experimental Results

We experimentally validate the performance of ISM in terms of both speed and clustering quality. The source code is publicly available at github[1]. Because [6] has already performed extensive comparisons of KDAC against other alternative clustering methods, this section will concentrate on comparing ISM to competing models for optimizing KDAC: Dimensional Growth (DG) [6] and gradient descent on the Stiefel Manifold (SM) [7]. SM is a generic approach, while DG is the approach originally proposed to solve KDAC [6].

In addition to introducing a new algorithm for optimizing KDAC, we also proposed an intelligent initialization scheme based on Taylor approximation in Section 3.2, we call Spectral Initialization (SI). We also investigate how SI effects the performance of the various algorithms compared to the standard random initialization (RI).

Datasets. We perform experiments on four synthetic and three real datasets. The synthetic data are displayed in Figure 2. Small Gaussian (SG) contains four Gaussian clusters with 40 samples and two features shown at the top left of the figure. When this data is projected down to feature 2, the original clustering is created. Rotating this projection onto feature

---

[1] https://github.com/neu-spiral/ISM

SG	NMI ↑	CQ ↑	Novelty ↓	Cost ↓	Time ↓
ISM SI	**1**	**2**	**0**	-1.2	**0.02**
ISM RI	0.4±0.49	1.6±0.478	**0.0±0.0**	-1.48±0.381	0.04±0.01
SM SI	**1**	1.99	**0**	**-1.791**	0.404
SM RI	**1±0.0**	1.79±0.398	**0.0±0.0**	-1.59±0.398	1.47±1.521
DG SI	0	1.629	1	-1.316	1.732
DG RI	0.93±0.196	0.99±0.0	0.03±0.104	-0.99±0.0	1.51±0.102
LG					
ISM SI	**1**	**1.9**	**0**	509	0.239
ISM RI	0.8±0.4	**1.9±0.011**	**0.0±0.0**	605±193	**0.18±0.038**
SM SI	**1**	1.9	**0**	509	0.413
SM RI	0.4±0.49	1.9±0.031	0.6±0.49	1095±479	11.21±11.01
DG SI	**1**	**1.9**	**0**	509.975	1595.174
DG RI	0.1±0.32	0.61±0.17	0.9±0.32	**101.5±30.8**	1688±551
Moon					
ISM SI	**1**	**2**	**0**	**149**	**0.575**
ISM RI	0.4±0.4	**2.0±0.0**	0.5±0.5	269.98±43	3.07±1.4
SM SI	**1**	1.998	**0**	**149**	2.653
SM RI	0.27±0.335	**2.0±0.0**	0.5±0.5	277±18	258±374
DG SI	**1**	1.998	**0**	149.697	359.188
DG RI	0.09±0.3	1.27±0.6	0.8±0.4	161.65±102	3212±1368
MoonN					
ISM SI	**1**	**2**	**0**	**15.3**	**0.3**
ISM RI	0.91±0.2	**2.0±0**	0.04±0.1	15.59±0.8	0.55±0.2
SM SI	**1**	**2**	**0**	**15.3**	0.452
SM RI	0.22±0.4	**2.0±0**	0.72±0.4	17.7±1	102±48.5
DG SI	**1**	**2**	**0**	15.4	3836.3
DG RI	0±0.0	1.6±0.04	1.0±0.0	16.2±0.06	3588±304
Flower					
ISM SI	-	**0.41**	**0**	20	**0.04**
ISM RI	-	**0.41±0.0**	**0.0±0.0**	19.51±0.0	0.1±0.01
SM SI	-	**0.41**	**0**	20	0.153
SM RI	-	0.3±0.2	0.01±0.012	**0.7±0.5**	0.8±0.6
DG SI	-	**0.41**	**0**	20	27.251
DG RI	-	0.35±0.02	0.2±0.4	20.6±1	37.0±5
Faces					
ISM SI	**0.57**	**0.61**	0.004	59.6	**1.5**
ISM RI	0.56±0.002	**0.61±0.0**	**0.0±0.0**	59.6±0.118	1.51±0.146
SM SI	0.562	0.608	0.004	59.559	7.96
SM RI	**0.57±0.002**	**0.61±0**	0.004±0	59.7±0.1	109±35.2
DG SI	0.564	0.6	0.004	59.8	100591.071
DG RI	0.458±0.045	0.565±0.018	0.024±0.018	**54.28±3**	166070±4342
WebKb					
ISM SI	**0.37**	0.286	**0**	-0.273	231
ISM RI	0.29±0.07	0.54±0.2	0.01±0.002	-0.52±0.2	137.45±15.7
SM SI	0.048	**3.296**	0.034	**-3.159**	**9.945**
SM RI	0.33±0.06	0.12±0.005	0.008±0.004	-0.11±0.004	13511±13342
DG SI	0.048	1.066	0.034	-1.019	199887.134
DG RI	0.23±0	1.06±0	0.1±0.005	0.616±0.04	727694±41068

Table 1: The Normalized Mutual Information, Clustering Quality, and clustering novelty are abbreviated in the first 3 columns as NMI, CQ, and Novelty. The cost of the objective and run time is displayed in the last two columns. For each optimization technique, spectral initialization (SI) and random initialization (RI) are separately tested. With RI, 10 random initial points have been tested with their mean and std displayed.
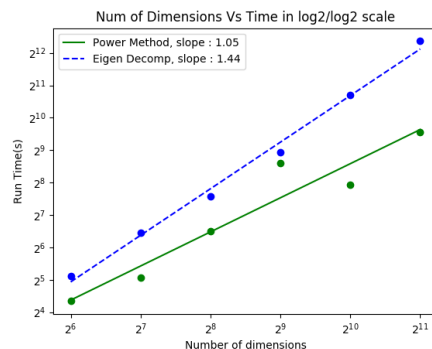

Figure 3: Growth in dimension vs time in log/log scale. A slope of 1 is linear growth.

1, the alternative clustering emerges. Large Gaussian (LG) contains four Gaussian clusters with 1000 samples and four dimensions is shown in the top center

location. The Gaussian clusters are rotated by 45 degrees to reside in 3D, and the fourth dimension is generated from a uniform noise distribution. This synthetic data is designed to test the response of KDAC within a noisy environment. We generate two additional synthetic datasets shown in top right: Moon and Moon+Noise (MoonN). Both datasets have the first two dimensions as two parabolas and the second two dimensions as Gaussian clusters. The MoonN dataset further includes three noisy dimensions generated from a uniform distribution with 1000 samples. Since the Gaussian clusters have a compact structure and the parabolas have a non-linear structure, these datasets demonstrate KDAC's ability to handle mixed clustering structures in a uniformly noisy environment. Due to the novelty of alternative clustering, there has been very few public repository data that have at least two alternative labels. The two traditional benchmark datasets are CMU's WebKB dataset [23] and face images from the UCI KDD repository [24]. CMU's WebKB dataset consists of 1041 html pages from 4 universities. One labelling is based on universities, and an alternative labelling based on topic (course, faculty, project, student). After preprocessing by removing rare and stop words, we are left with 500 words. The face dataset consists of 640 images from 20 people in four different poses. Each image has 32x30 pixels. Images are vectorized and PCA is then used to further condense the dataset by keeping 85% of the variance, resulting in a dataset of 624 samples and 20 features. We set the existing clustering based on identity, and seek an alternative clustering based on pose. The last real dataset is the Flower image by Alain Nicolas [25], a 350x256 pixel image. The RGB values of each pixel is taken as a single sample, with repeated samples removed. This results in a dataset of 256 samples and 3 features. Although this dataset does not have labels, the quality of the alternative clustering can be visually observed. With the exception of the high dimensional WebKB data, the remaining datasets are visualized in Figure 1 (b).

Hyperparameter Tuning. ISM has 3 hyperparameters that require tuning, q, λ, and σ. The hyperparameter q with a potential range of $(0,d)$ is initially set to k, the number of clusters. A grid search was then conducted for $\lambda \in (0, 10]$ and $\sigma \in (0, 10]$ to find the highest CQ (λ, σ) pair that satisfies inequality (10) at an increment of 0.01. In the event that Inequality (10) cannot be satisfied, the highest CQ closest to satisfying (10) is used. At this point, since $\Phi(W^*)$ is computed, the q value that maximizes the eigengap as defined in Eq(11) can be used if (10) is not yet satisfied. Note that competing models do not have a natural way of selecting hyperparameters. Since all competing models optimize the same objective, we re-

port all results using the same hyperparameters. To ensure reproducible experiments, the hyperparameters utilized in each experiment is provided in Appendix J.

Evaluation Method. To exhaustively evaluate ISM on KDAC, both external and internal measures have been recorded in Table 1. More specifically, column 1 and 3 (NMI and Novelty) are external measures because they compare alternative cluster assignments against an externally known ground truth. To make the comparison, Normalized Mutual Information (NMI) as suggested by [26] was used. The NMI is a measure between 0 to 1 with 0 denoting no relationship and 1 as maximum relationship. Column 1 (NMI) is calculated by computing the NMI between the ground truth and the alternative cluster assignments. Column 3 (Novelty) is calculated by computing the NMI between the alternative cluster assignments against the original label. Ideally, we wish for the NMI of the alternative cluster assignments against the ground truth to be 1, and 0 against the original clustering. If we let U and L be two clustering assignments, NMI can be calculated with $NMI(L, U) = \frac{I(L,U)}{\sqrt{H(L)H(U)}}$, where $I(L, U)$ is the mutual information between L and U, and $H(L)$ and $H(U)$ are the entropies of L and U respectively. Lastly, since no ground truth exists for the Flower dataset, the NMI field is not applicable and indicated with a dash (-).

Columns 2, 4 and 5 (CQ, Cost, Time) are internal measures used to evaluate ISM. They are internal measures since they are computed solely from the data and the algorithm with no external knowledge applied during the comparison. The clustering quality (CQ) is computed with $HSIC(XW, U)$ with U as the clustering solution. This is equivalent to the spectral clustering objective with high values denoting high clustering quality. The cost quality records the objective function of KDAC in Eq. (2). The last column, Time, measures the execution time of the entire KDAC algorithm in seconds on an Intel Xeon E7 processor.

In Table 1, we report the performance of the various methods: our ISM, SM, and DG, paired with two initialization schemes: our spectral initialization (SI) and random initialization (RI). For random initialization, we repeat these 10 times and report the mean and standard deviation for each measure. The optimal direction of each measure is denoted by the ↑↓, with ↑ denoting a preference towards larger values and ↓ otherwise. For each field, the optimal result is printed in bold font.

Performance Comparison Results. Table 1 illustrates that ISM with SI outperforms its competitors in both internal and external quality measure for the first 3 columns. Since it is possible for the objective cost

to achieve a low cost with a trivial solution based on different (σ, λ) pairs, the objective should be low, but it is not always indicative of better clustering quality. This is demonstrated in the case of WebKb dataset, where SM+SI achieved a lower cost with a faster convergence. However, upon inspecting the clustering allocation in this case, we observed that it was a trivial classification, with almost all points in the same cluster. Since the demand for a faster KDAC was the original motivator, special attention should be paid to the Time column. The execution time of ISM significantly outperforms the original approach (DG) as well as the standard approach (SM). This speed improvement is especially true in the Faces dataset where ISM improves the speed by 5 folds.

The Effect of Spectral Initialization. By comparing SI against RI, we can isolate the effect of our proposed spectral initialization (SI). Comparing the rows, the spectral initialization improved both time and clustering quality for all methods. From this observation, we conclude that SI contributes to clustering quality by starting each algorithm at a desirable initialization. The convergence improvement came from placing each algorithm closer to the local minimum. Comparing all methods when using SI, we observed that ISM optimization still outperformed other algorithms in terms of time. From this, we conclude that the proposed SI has a greater impact on the clustering quality while the ISM optimization technique contribute towards faster convergence.

Scalability. Note that KDAC consists of optimizing U and W from Eq. (2). Since the computational bottleneck resided in the optimization of W, ISM was designed to speed up this portion. The scalability analysis, therefore, will concentrate on only the optimization of W. ISM has a computational complexity of $O(t_{ISM}(n^2d^2 + d^3)$ and the complexity can be divided into the calculation of the derivative and the eigendecomposition of $\Phi(W)$. Although the derivative contributes to $O(n^2d^2)$, it is a highly parallelizable operation that can be rendered trivial through the usage of GPUs. Therefore, ISM's true bottleneck is not the number of samples (n), but the number of dimensions due to the $O(d^3)$ operation of eigendecomposition. As the dimensionality of the data increase, the complexity grows at a cubic rate. Yet, while this growth of dimensionality may exacerbate algorithms such as SM, the negative influence on ISM is limited. This is because the $O(d^3)$ term from SM came from a matrix inversion while ISM uses a spectral method. Since only very few eigenvalues are required, there exists many approximation algorithms to speed up the eigendecomposition [19] [20] [21]. To demonstrate this point, Coordinate-wise Power Method (CPM) [21] was implemented and compared to the eigendecomposition operation from Numpy. In this experiment, noisy dimensions of uniform distributions are added to the synthetic dataset of Large Gauss (LG) such that the total dimension increase in multiples of 2. As the dimension of the data grows exponentially, the time of execution is recorded in Figure 3 in log scale. Given a log/log scale, a slope of 1 is linear while a slope of 2 is quadratic. From studying the slope of each operation, the eigendecomposition from Numpy grows at a rate between linear and quadratic. However, by utilizing CPM, the growth becomes roughly linear. Therefore, the refinement of utilizing a spectral optimization method is a key reason for speed improvement.

5 Conclusions

We have proposed an iterative spectral optimization technique for solving a non-convex optimization problem while being constrained on a Stiefel manifold. This new technique demonstrates speed improvement for any algorithm that could be formulated into Eq (4). The ISM algorithm is easy to implement with existing software libraries. Due to the usage of the spectral method, approximation techniques from existing research could be deployed to further speed up the eigendecomposition. Accompanied with the algorithm are the theoretical guarantees that satisfy the first and second order necessary conditions for a local minimum. Besides these guarantees, we also proposed a natural initialization point based on Taylor approximation of the original cost function. Experiments on synthetic and real data confirmed that our proposed spectral initialization improved the performance of all the optimization algorithms, ISM, SM and DG. Simultaneously, the experiments demonstrate that our proposed ISM algorithm had the best convergence rate compared to competing models with up to 5 folds of speed improvement. Although we focus this paper on an important application of ISM, alternative clustering, the optimization algorithm proposed and guarantees can be extended to other optimization problems involving Gaussian-kernel like objectives constrained over the Stiefel manifold, which is a common formulation for dimensionality reduction with a kernel-based objective. Understanding whether ISM can be applied to such objectives and be leveraged to solve broader classes of problems is a natural future direction for this work.

References

[1] David Gondek and Thomas Hofmann. Non-redundant data clustering. *Knowledge and Information Systems*, 12(1):1–24, 2007.

[2] Ying Cui, Xiaoli Z Fern, and Jennifer G Dy. Learning multiple nonredundant clusterings. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):15, 2010.

[3] Xuan Hong Dang and James Bailey. Generation of alternative clusterings using the cami approach. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 118–129. SIAM, 2010.

[4] Ian Davidson and Zijie Qi. Finding alternative clusterings using constraints. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 773–778. IEEE, 2008.

[5] Ying Cui, Xiaoli Z Fern, and Jennifer G Dy. Nonredundant multi-view clustering via orthogonalization. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 133–142. IEEE, 2007.

[6] Donglin Niu, Jennifer G Dy, and Michael I Jordan. Iterative discovery of multiple alternativeclustering views. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1340–1353, 2014.

[7] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

[8] Rich Caruana, Mohamed Elhawary, Nam Nguyen, and Casey Smith. Meta clustering. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 107–118. IEEE, 2006.

[9] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[10] Sajib Dasgupta and Vincent Ng. Mining clustering dimensions. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 263–270, 2010.

[11] Vikash K Mansinghka, Eric Jonas, Cap Petschulat, Beau Cronin, Patrick Shafto, and Joshua B Tenenbaum. Cross-categorization: A method for discovering multiple overlapping clusterings. In *Nonparametric Bayes Workshop at NIPS*, 2009.

[12] Donglin Niu, Jennifer Dy, and Zoubin Ghahramani. A nonparametric bayesian model for multiple clustering with overlapping feature views. In *Artificial Intelligence and Statistics*, pages 814–822, 2012.

[13] Leonard Poon, Nevin L Zhang, Tao Chen, and Yi Wang. Variable selection in model-based clustering: To do or to facilitate. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 887–894, 2010.

[14] Eric Bae and James Bailey. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 53–62. IEEE, 2006.

[15] ZiJie Qi and Ian Davidson. A principled and flexible framework for finding alternative clusterings. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 717–726. ACM, 2009.

[16] Xuan-Hong Dang and James Bailey. A hierarchical information theoretic technique for the discovery of non linear alternative clusterings. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 573–582. ACM, 2010.

[17] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

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

[19] Max Vladymyrov and Miguel Carreira-Perpinan. The variational nystrom method for large-scale spectral problems. In *International Conference on Machine Learning*, pages 211–220, 2016.

[20] Peter Richtárik. Generalized power method for sparse principal component analysis.

[21] Qi Lei, Kai Zhong, and Inderjit S Dhillon. Coordinate-wise power method. In *Advances in Neural Information Processing Systems*, pages 2064–2072, 2016.

[22] Stephen Wright and Jorge Nocedal. Numerical optimization. *Springer Science*, 35:67–68, 1999.

[23] CMU CMU. universities webkb data, 1997. 4.

[24] Stephen D Bay, Dennis Kibler, Michael J Pazzani, and Padhraic Smyth. The uci kdd archive of large data sets for data mining research and experimentation. *ACM SIGKDD Explorations Newsletter*, 2(2):81–85, 2000.

[25] Particles of tessellations. http://en.tessellations-nicolas.com/. Accessed: 2017-04-25.

[26] A Strehl and J Chosh. Knowledge reuse framework for combining multiple partitions. *Journal of Machine learning Research*, 33(3):583–617.