NO MORE THAN 6FT APART: ROBUST K-MEANS VIA RADIUS UPPER BOUNDS

Ahmed Imtiaz Humayun[†] *Randall Balestriero*[‡] *Anastasios Kyrillidis*[†] *Richard Baraniuk*[†]

[†]Rice University [‡]Meta AI Research

ABSTRACT

Centroid based clustering methods such as k-means, k-medoids and k-centers are heavily applied as a go-to tool in exploratory data analysis. In many cases, those methods are used to obtain *representative* centroids of the data manifold for visualization or summarization of a dataset. Real world datasets often contain inherent abnormalities, e.g., repeated samples and sampling bias, that manifest imbalanced clustering. We propose to remedy such scenario by introducing a maximal radius constraint r on the clusters formed by the centroids, i.e., samples from a same cluster should not be more than $2r$ apart in terms of ℓ_2 distance. We achieve this constraint by solving a semi-definite program, followed by a linear assignment problem with quadratic constraints. Through qualitative results, we show that our proposed method is robust towards dataset imbalances and sampling artifacts. To the best of our knowledge, ours is the first constrained k-means clustering method with hard radius constraints.[1]

Index Terms— robust k-means, radius constraint, constrained optimization, data imbalance, clustering

1. INTRODUCTION

K-clustering methods offer the benefit of producing summarized dataset representations via a set of learned centroids or centers. Such representations find many applications from denoising, anomaly detection, visual summarization, as initial parameters for downstream algorithms such as Gaussian Mixture Models, and as plastic features for life-long machine learning classifiers [1]. The fundamental assumptions governing the success of K-means lies in having clusters with roughly the same number of samples and intra-cluster data covariance that is isotropic with the form σI. Furthermore, σ should be roughly the same between clusters. Whenever the data does not align with those assumptions, K-means gets skewed toward producing an incorrect representation. For example, even in the simplest case of having a dataset made of a mixture of Gaussians but with varied number of samples per mixture, K-means centroids will naturally shift toward the mode with the greatest number of samples.

[1]Codes at https://bit.ly/kmeans-constrained

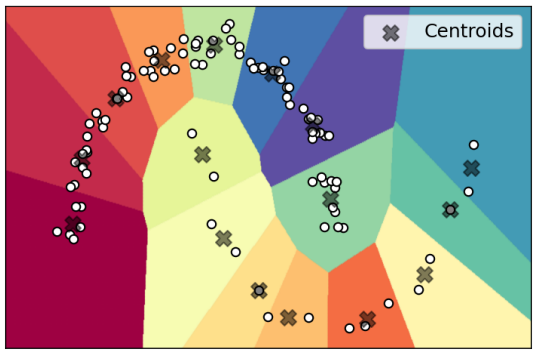

Fig. 1. Centroids generated by our proposed maximal radius constrained k-means method for $K = 16$ on imbalanced two-moons data $(85/15)$. Even though the concave moon is oversampled more than 5 times, our method produces equal number of centroids for both moons. See Fig 2 for comparisons.

The implication of those cases can be dramatic as any downstream task relying on those representations, would be negatively impacted causing, e.g., bias in facial recognition models [2], gender bias in word-level language models [3]. This has led to the birth of many K-means alternatives, each aiming at fixing a particular limitation, e.g., the presence of outliers among others [4, 5]. There also exists k-clustering methods focused on (fair) data summarization [6, 7, 8], imbalanced data clustering [9] and robustness to specific transformations of the data [10, 11, 12]. Some of these methods require specifications on the cardinality of the demographics [6, 13, 14], weak labels of imbalance [9], the data transformations to be robust against [12], or a priori knowledge of the data/outlier distributions [15].

In this paper we propose radius constrained clustering as a method to introduce robustness to K-means without requiring any domain specific knowledge. That is, the algorithm will produce regions/clusters for which the pairwise distance between samples within that region is upper bounded by a chosen constant. Our proposed method (Fig. 1) generates uniformly spaced centroids on the data manifold, while being robust towards sampling inconsistencies. This offers great advantages, e.g., when using K-means to obtain points from a manifold robust to the distribution of samples. We compare

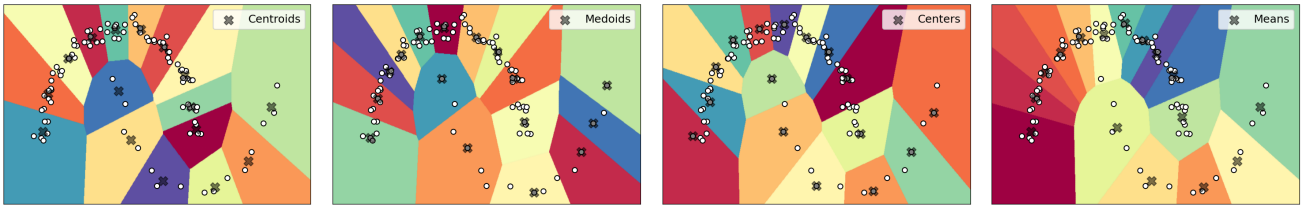

Fig. 2. From Left to Right, K-means, K-medoids, K-center and cardinality constrained K-means clustering [13] for the imbalanced two moons summarization task as in Fig.1. All four of the k-clustering algorithms are biased towards the concave moon with only $\{6, 6, 6, 5\}$ centroids being selected from the convex moon by each method respectively.

our proposed method with standard k-clustering methods and robust methods such as cardinality constrained clustering [13] and t-distribution K-means clustering [15]. Our contributions in this paper are summarized below:

- We present the first k-means algorithm with a hard radius constraint that is tractable. We use a convex relaxation of radius constrained k-means, and pose it as a mixed integer (MI) semi-definite program (SDP). We solve it via a linear SDP relaxation and subsequent rounding.
- We present empirical evidence on the efficacy of radius constrains on summarization of data, especially to be robust towards sampling biases.

The rest of the paper is organized as follows: in Section 2, we present the radius constraint K-means that we propose, starting from the definition of K-means and moving towards a Mixed Integer Semi-Definite Program (MISDP) formulation of our method. In Section 4 we discuss qualitative results comparing with different methods, and, in Section 5, we discuss future directions.

2. BACKGROUND: K-MEANS

We denote by $\Gamma = \{x_l\}_{l=1}^N$ the set of N data points in $\mathbb{R}^m$. K-means proposes a centroid based clustering, i.e., partition of Γ into k disjoint groups found by minimizing

$$\min_{\{\Gamma_k\}_{k=1}^K} \sum_{k=1}^K \sum_{l \in \Gamma_k} \|x_l - \gamma_k\|^2, \tag{1}$$

where $\Gamma_k \cap \Gamma_{k'} = \{\}, \forall k \neq k', \cup_{k=1}^K \Gamma_k = \Gamma$ and γ_k is the centroid of cluster k. Let, $\mathbf{1}_{\Gamma_k}$ be the indicator function of the k-th cluster, Eq. 1 becomes

$$\begin{aligned}\sum_{k=1}^K \sum_{l \in \Gamma_k} \|x_l - \gamma_k\|^2 &= \frac{1}{2} \sum_{k=1}^K \frac{1}{|\Gamma_k|} \sum_{l \in \Gamma_k, s \in \Gamma_k} \|x_l - x_s\|^2 \\ &= \frac{1}{2} \sum_{k=1}^K \frac{1}{|\Gamma_k|} \langle \mathbf{1}_{\Gamma_k} \mathbf{1}_{\Gamma_k}^T, \mathbf{D} \rangle \end{aligned} \tag{2}$$

where, $\langle \cdot, \cdot \rangle$ is the matrix inner product, and $\mathbf{D} \in \mathbb{R}^{N \times N}$ is the squared pairwise distance matrix with each element $d_{ij} = \|x_i - x_j\|^2$. The first equality in Eq. 2 comes from the equality relationship introduced in [16], relating the sum of pairwise distances with the sum of radial distance for any partition. The second equality is a simple matrix reformulation of the inner sum operation. Therefore, we can rewrite the k-means problem as,

$$\min_{\{\Gamma_k\}_{k=1}^K} \frac{1}{2} \sum_{k=1}^K \frac{1}{|\Gamma_k|} \langle \mathbf{1}_{\Gamma_k} \mathbf{1}_{\Gamma_k}^T, \mathbf{D} \rangle, \tag{3}$$

with $\cup_{k=1}^K \Gamma_k = \Gamma$ and $\Gamma_k \cap \Gamma_{k'} = \emptyset$ for $k \neq k'$, which is an NP hard problem [17]. Notice that in the above formulation, there are no explicit constraints on the number of samples per cluster, the intra-cluster radius or the weighting of different samples, e.g., to account for outliers. We propose to take one step into that direction by providing a hard constraint on the intra-cluster radius.

3. RADIUS CONSTRAINED K-MEANS: NO MORE THAN 6FT APART

Previously, [18] have provided formulations for soft radius constraints in online k-means clustering, where the constraint is introduced as an additional term in the optimization objective. We provide a formulation for hard radius constraints r, where r is fixed for every cluster. Since for any partition with a fixed radius, the maximal distance between two samples can be at most the diameter, we can write the maximal radius constraint as

$$\max\{\|x_l - x_s\|^2 | l, s \in \Gamma_k\} \leq 4r^2 \quad \text{for } k = 1, 2, 3..., K. \tag{4}$$

The k-means objective in Eq. 3 can therefore be rewritten with the maximal radius constraint as,

$$\min_{\{\Gamma_k\}_{k=1}^K} \frac{1}{2} \sum_{k=1}^K \frac{1}{|\Gamma_k|} \langle \mathbf{1}_{\Gamma_k} \mathbf{1}_{\Gamma_k}^T, \mathbf{D} \rangle \tag{5}$$

$$\text{s.t } \cup_{k=1}^K \Gamma_k = \Gamma, \ \Gamma_k \cap \Gamma_{k'} = \emptyset \text{ for } k \neq k' \tag{6}$$

$$d_{ij} \leq 4r^2 \ \ \forall i, j \in \Gamma_k \text{ for } k = 1, 2, 3..., K \tag{7}$$

Table 1. Comparison of partition radius, k-radius and number of k in the convex moon (Ω) for 100 random seeds. Note that our proposed constrained optimization model ($r = .189$) finds the optimal solution for given constraints.

Method	Max Partition Radius	Max k-Radius	k $\in \Omega$
K-means	.228($\pm$.03)	.258($\pm$.04)	5.77($\pm$.42)
K-medoids	.268($\pm$.05)	.379($\pm$.07)	5.56($\pm$.68)
K-center	.224($\pm$.02)	.289($\pm$.01)	7.52($\pm$.57)
card. K-means [13]	.222($\pm$.03)	.25($\pm$.03)	5.83($\pm$0.43)
tk-means [15]	.28($\pm$.06)	.309($\pm$.07)	4.92($\pm$0.63)
Ours	**0.181**	**.207**	**8**

. Note that by setting $r^2 = \infty$ one recovers the standard K-means form. Before going into the optimization method and empirical validations, we recall that our goal is to leverage the explicit constraint on $d_{i,j}$ to ensure that some regions can not cover samples that are too far apart in the space.

3.1. MILP formulation of Radius Constrained K-means

We start the Mixed Integer Linear Program (MILP) formulation of Eq. 5 by introducing NK binary variables $\pi_i^k \in \{0, 1\}$, where $\pi_i^k = 0$ if $x_i \notin \Gamma_k$ and $\pi_i^K = 1$ if $x_i \in \Gamma_k$. Therefore, the objective becomes

$$\min_{\forall \pi_i^k} \quad \frac{1}{2} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j=1}^{N} d_{ij} \pi_i^k \pi_j^k \tag{8}$$

$$\text{s.t} \quad \pi_i^k \in \{0, 1\}, \quad n_k \in \mathbb{Z}, 1 \leq n_k \leq N, \tag{9}$$

$$\sum_{i=1}^{N} \pi_i^k = n_k, \sum_{k=1}^{K} n_k = N, \ \sum_{k=1}^{K} \pi_i^k = 1, \tag{10}$$

$$d_{ij} \pi_i^k \pi_j^k \leq 4r^2, \forall i, j, \ \forall k, \tag{11}$$

where, n_k are integer variables between $[1, N]$ and $n_k = |\Gamma_k|$ at optimality. It can be easily verified that the constraints 10 and 11 are equivalent to constraints 6 and 7. The feasible set of the original k-means formulation in Eq. 1 is also a feasible set of the MILP formulation. Our formulation is closely related to the cardinality constrained k-means formulation in [14]. In our optimization model, we introduce a constraint on the squared pairwise distance inside each partition, while keeping its cardinality as an integer variable; whereas [14] allows specifying cardinality constraints for each partition. Note that our proposed model can also allow using different radius constraints r_k in Eq. 11 for different partitions without changing the model class. We avoid that for the sake of simplicity of our formulation. Another thing to note is that the partition radius upper bound r also upper bounds the k-radius, i.e., the maximal distance between any sample and its centroid, by $2r$.

3.2. Convex relaxation of the MILP formulation

We start the convex formulation by replacing the binary variables π_i^k with binary vector $\mathbf{b^k} = \{b_i^k\}_{i=1}^N$, $b_i^k \in \{-1, 1\}$ where b_i^k is 1 if $x_i \in \Gamma_k$ and -1 otherwise. This implies $b_i^k = 2\pi_i^k - 1$. The MILP objective function can be written in terms of $\mathbf{b^k}$ as:

$$\frac{1}{2} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j=1}^{N} d_{ij} \pi_i^k \pi_j^k \tag{12}$$

$$= \frac{1}{2} \langle \mathbf{D}, \sum_{k=1}^{K} \frac{1}{4n_k} (\mathbf{M^k} + \mathbf{1}\mathbf{1}^T + \mathbf{b^k}\mathbf{1}^T + \mathbf{1}(\mathbf{b^k})^T) \rangle \tag{13}$$

where $\mathbf{M^k} \in \mathbb{R}^{N \times N}$. For equality in the feasible set, we need $\mathbf{M^k} = \mathbf{b^k}(\mathbf{b^k})^T$ which yields each element in the right term of the inner product as $\frac{1}{4n_k} \sum_{k=1}^{K} b_i^k b_j^k + 1 + b_i^k + b_j^k$. The SDP relaxation of the problem therefore is,

$$\min_{\forall \mathbf{b^k}, \mathbf{M^k}} \quad \frac{1}{8} \langle \mathbf{D}, \sum_{k=1}^{K} \frac{1}{n_k} (\mathbf{M^k} + \mathbf{1}\mathbf{1}^T + \mathbf{b^k}\mathbf{1}^T + \mathbf{1}(\mathbf{b^k})^T) \rangle \tag{14}$$

$$\text{s.t} \quad -1 \leq b_i^k \leq 1, 1 \leq n_k \leq N, \mathbf{M^k} \succeq \mathbf{b^k}(\mathbf{b^k})^T,$$

$$\text{diag}(\mathbf{M^k}) = \mathbf{1}, \mathbf{1}^T \mathbf{b^k} = 2n_k - N, \ \sum_{k=1}^{K} n_k = N,$$

$$\sum_{k=1}^{K} \mathbf{b^k} = (2 - k)\mathbf{1}, d_{ij}(m_{ij}^k + 1 + b_i^k + b_j^k) \leq 16r^2,$$

$$\text{for} \quad i, j = 1, 2, ..., N \text{ and } k = 1, 2, ..., K$$

where, the semi-definite constraint $\mathbf{M} \succeq \mathbf{b}(\mathbf{b})^T$ can be converted into a linear matrix inequality using Schur's complement [19].

The objective in the current formulation is a linear fractional function which can be turned into a linear objective using Charnes-Cooper transformation [20]. Specifically, since the denominator in Eq. 12 is strictly positive as $n_k \geq 1$, the objective can be expressed as a perspective function [21]. Without the continuous relaxation of $\mathbf{b^k}$, the problem can be formally stated as a Mixed Integer Semi-definite program (MISDP) [22].

3.3. Rounding Algorithm

We define our rounding algorithm as a two-step linear assignment problem with quadratic constraints (Alg. 1). The first step in the algorithm is to find binary partition variables for each sample, we define it as $\mathbf{\Pi} \in \{0, 1\}^{N \times K}$ where each element π_i^k is 1 if x_i is assigned to cluster k. We solve two linear assignment objectives, one in which we maximize the inner product sum of the SDP solution and π_i^k with quadratic constraints adhering to the maximal radius constraint (Eq. 15). In the second step, we minimize the intra-cluster distance for the assignment variables (Eq. 16).

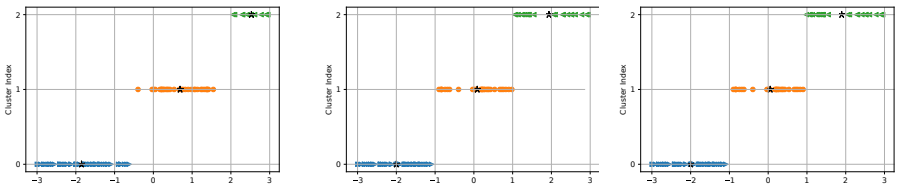

Fig. 3. From Left to Right, K-means, K-means with radius constraint$=1$ and K-means with maximal cardinality constraint$=55$.

Algorithm 1: Maximal radius constrained k-means

Input: $N; \mathbf{x} = \{x_i\}_{i=1}^N; k; r; \mathbf{D} \in \mathbb{R}^{N\times N}$

Output: binary assignments $\mathbf{\Pi} \in \{0,1\}^{N\times K}$

Step 1: Solve continuous relaxation of MISDP using interior point method $\rightarrow \mathbf{b^k}\ \forall k$;

Step 2: Let, $\mathbf{\Pi} \in \{0,1\}^{N\times K}$ binary partition assignment variables. Solve,

$$\max_{\mathbf{\Pi}} \quad \sum_{k=1}^{K}\sum_{i=1}^{N} \pi_i^k b_i^k \quad \text{s.t} \quad 1 \leq \sum_{i=1}^{N} \pi_i^k \leq N, \quad (15)$$
$$\text{and} \sum_{k=1}^{K} \pi_i^k = 1,\ d_{ij}\pi_i^k\pi_j^k \leq 4r^2 \quad \forall i,j,k$$

Step 3: Get partitions Γ_k and centroids $\gamma_k \quad \forall k$

Step 4: Solve for $\mathbf{\Pi} \in \{0,1\}^{N\times K}$

$$\min_{\mathbf{\Pi}} \sum_{k=1}^{K}\sum_{i=1}^{N} \pi_i^k \|x_i - \gamma_k\|^2 \text{ s.t } 1 \leq \sum_{i=1}^{N} \pi_i^k \leq N, \quad (16)$$
$$\text{and} \sum_{k=1}^{K} \pi_i^k = 1, d_{ij}\pi_i^k\pi_j^k \leq 4r^2, \forall i,j,k$$

4. EXPERIMENTS

2D experiments. Experimental results presented in Fig. 1 and Fig. 2 portray the efficacy of radius constrained k-means for imbalanced data summarization. We compare with cardinality constrained k-means [13] as an alternative constrained k-means method. We also compare with tk-means [15] which uses long tail assumptions for robustness. For comparison, we sweep the cardinality upper and lower bounds of [13] till infeasible to find the best balance. We see that both increasing the lower or decreasing the upper bounds from respectively 0 and $N = 100$ harms balanced centroid generation; increasing the lower bound makes it easier to achieve the lower bound for the convex moon, while decreasing the upper bound requires more centroids to cover the concave moon. For all experiments we choose N/K to be small since otherwise, k-means based clustering might return centroids off the data manifold, therefore yielding bad sketches/summaries. An added benefit radius constraints provide is a feasibility certificate- tighter radius bounds resulting in empty feasibility sets can be used to infer how to increase K to be able to cover all the samples. For the imbalanced two moons experiments, for a radius constraint of $r = .189$, we have seen that at least 16 centroids were required to be able to cover the whole manifold. For different methods, the partition radius and k-radius is presented in Table 1.

1D experiments. Let, we have $N = 102$ samples from three uniform distributions *U(-3,-1)*, *U(-1,1)* and *U(1,3)* with $51, 26, 25$ samples in each respectively. We draw comparisons between standard k-means, cardinality constrained k-means and radius constrained k-means in a $k = 3$ summarization task. Fig. 3 shows clustering performance for such a case; without any constraints, k-means will create an inconsistent partition, e.g., resulting in the mixing of different attributes represented by each random variable. This will yield centroids which are not proper summaries of the dataset. Whereas with radius constraint of 1 and cardinality constraint of 55 adhere to the correct partitioning. In such settings, cardinality constrained k-means require re-tuning the constraint when the dataset is resampled, whereas ours is robust.

Implementation. We use `MOSEK` to solve Step 1 in Alg. 1 and `Gurobi` to solve Steps 2 and 4. In our experiments, we did not require tuning of solver parameters.

5. CONCLUSION

We propose the first maximal radius constrained K-means as an MISDP optimization objective. Upon comparison with multiple k-clustering methods, we see that our method is more robust towards sampling bias/ data imbalance. The main limitation of our radius constrained k-means formulation is that both the order of variables and constraints are $\mathcal{O}(k.N^2)$ which is impractical for very large datasets. From preliminary experiments we see that replacing the SDP problem with a k-center problem in Step 1 of Alg. 1 has minimal effects on the centroid selection. This can be considered a future direction to improve computational complexity.

6. REFERENCES

[1] Wangli Hao, Junsong Fan, Zhaoxiang Zhang, and Guibo Zhu, "End-to-end lifelong learning: a framework to achieve plasticities of both the feature and classifier constructions," *Cognitive Computation*, vol. 10, no. 2, pp. 321–333, 2018.

[2] Joy Buolamwini and Timnit Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[3] Shikha Bordia and Samuel R Bowman, "Identifying and reducing gender bias in word-level language models," *arXiv preprint arXiv:1904.03035*, 2019.

[4] ZR Hesabi, Zahir Tari, A Goscinski, Adil Fahad, Ibrahim Khalil, and Carlos Queiroz, "Data summarization techniques for big data—a survey," in *Handbook on Data Centers*, pp. 1109–1152. Springer, 2015.

[5] Ashish Chiplunkar, Sagar Kale, and Sivaramakrishnan Natarajan Ramamoorthy, "How to solve fair k-center in massive data models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1877–1886.

[6] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern, "Fair k-center clustering for data summarization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3448–3457.

[7] Suman K Bera, Deeparnab Chakrabarty, Nicolas J Flores, and Maryam Negahbani, "Fair algorithms for clustering," *arXiv preprint arXiv:1901.02393*, 2019.

[8] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii, "Fair clustering through fairlets," *arXiv preprint arXiv:1802.05733*, 2018.

[9] Ch N Santhosh Kumar, K Nageswara Rao, A Govardhan, and K Sudheer Reddy, "Imbalanced k-means: An algorithm to cluster imbalanced-distributed data," *International Journal of Engineering and Technical Research*, vol. 2, no. 2, 2014.

[10] Luis Angel Garcia-Escudero and Alfonso Gordaliza, "Robustness properties of k means and trimmed k means," *Journal of the American Statistical Association*, vol. 94, no. 447, pp. 956–969, 1999.

[11] Katsuhiro Honda, Akira Notsu, and Hidetomo Ichihashi, "Fuzzy pca-guided robust k-means clustering," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 1, pp. 67–79, 2010.

[12] Olga Dorabiala, J Nathan Kutz, and Aleksandr Aravkin, "Robust trimmed k-means," *arXiv preprint arXiv:2108.07186*, 2021.

[13] Paul S Bradley, Kristin P Bennett, and Ayhan Demiriz, "Constrained k-means clustering," *Microsoft Research, Redmond*, vol. 20, no. 0, pp. 0, 2000.

[14] Napat Rujeerapaiboon, Kilian Schindler, Daniel Kuhn, and Wolfram Wiesemann, "Size matters: Cardinality-constrained clustering and outlier detection via conic optimization," *SIAM Journal on Optimization*, vol. 29, no. 2, pp. 1211–1239, 2019.

[15] Yiming Li, Yang Zhang, Qingtao Tang, Weipeng Huang, Yong Jiang, and Shu-Tao Xia, "tk-means: A robust and stable k-means variant," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3120–3124.

[16] Hongyuan Zha, Xiaofeng He, Chris Ding, Ming Gu, and Horst D Simon, "Spectral relaxation for k-means clustering," in *Advances in neural information processing systems*, 2001, pp. 1057–1064.

[17] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat, "Np-hardness of euclidean sum-of-squares clustering," *Machine learning*, vol. 75, no. 2, pp. 245–248, 2009.

[18] Shi Zhong and David DeMaris, "Diameter-constrained clustering: Algorithms and experiments for a layout coverage problem," in *7th Annual Austin CAS International Conference*. Citeseer, 2006.

[19] Fuzhen Zhang, *The Schur complement and its applications*, vol. 4, Springer Science & Business Media, 2006.

[20] Abraham Charnes and William W Cooper, "Programming with linear fractional functionals," *Naval Research logistics quarterly*, vol. 9, no. 3-4, pp. 181–186, 1962.

[21] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[22] Tristan Gally, Marc E Pfetsch, and Stefan Ulbrich, "A framework for solving mixed-integer semidefinite programs," *Optimization Methods and Software*, vol. 33, no. 3, pp. 594–632, 2018.