

Size Matters: Cardinality-Constrained Clustering and Outlier Detection via Conic Optimization

Napat Rujeerapaiboon, Kilian Schindler, Daniel Kuhn

Risk Analytics and Optimization Chair, École Polytechnique Fédérale de Lausanne, Switzerland,
 napat.rujeerapaiboon@epfl.ch, kilian.schindler@epfl.ch, daniel.kuhn@epfl.ch

Wolfram Wiesemann

Imperial College Business School, Imperial College London, United Kingdom, ww@imperial.ac.uk

Plain vanilla K -means clustering is prone to produce unbalanced clusters and suffers from outlier sensitivity. To mitigate both shortcomings, we formulate a joint outlier detection and clustering problem, which assigns a prescribed number of datapoints to an auxiliary outlier cluster and performs cardinality-constrained K -means clustering on the residual dataset. We cast this problem as a mixed-integer linear program (MILP) that admits tractable semidefinite and linear programming relaxations. We propose deterministic rounding schemes that transform the relaxed solutions to feasible solutions for the MILP. We also prove that these solutions are optimal in the MILP if a cluster separation condition holds.

Key words: Semidefinite programming, K -means clustering, outlier detection, optimality guarantee

1. Introduction

Clustering aims to partition a set of datapoints into a set of clusters so that datapoints in the same cluster are more similar to another than to those in other clusters. Among the myriad of clustering approaches from the literature, K -means clustering stands out for his long history dating back to 1957 as well as its impressive performance in various application domains, ranging from market segmentation and recommender systems to image segmentation and feature learning (Jain 2010).

This paper studies the *cardinality-constrained K -means clustering problem*, which we define as

$$\begin{aligned} & \text{minimize} && \sum_{k=1}^K \sum_{i \in I_k} \left\| \boldsymbol{\xi}_i - \frac{1}{n_k} \left(\sum_{j \in I_k} \boldsymbol{\xi}_j \right) \right\|^2 \\ & \text{subject to} && (I_1, \dots, I_K) \in \mathfrak{P}(n_1, \dots, n_K). \end{aligned} \tag{1}$$

Problem (1) partitions N datapoints $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N \in \mathbb{R}^d$ into K clusters I_1, \dots, I_K of sizes n_1, \dots, n_K , $n_1 + \dots + n_K = N$, so as to minimize the sum of squared intra-cluster distances. Here,

$$\mathfrak{P}(n_1, \dots, n_K) = \left\{ (I_1, \dots, I_K) : |I_k| = n_k \ \forall k, \ \cup_{k=1}^K I_k = \{1, \dots, N\}, \ I_k \cap I_\ell = \emptyset \ \forall k \neq \ell \right\}$$

denotes the ordered partitions of the set $\{1, \dots, N\}$ into K sets of sizes n_1, \dots, n_K , respectively.

Our motivation for studying problem (1) is threefold. Firstly, it has been shown by Bennett et al. (2000) and Chen et al. (2006) that the algorithms commonly employed for the *unconstrained* K -means clustering problem frequently produce suboptimal solutions where some of the clusters contain very few or even no datapoints. In this context, cardinality constraints can act as a regularizer that avoids local minima of poor quality. Secondly, many application domains require the clusters I_1, \dots, I_K to be of comparable size. This is the case, among others, in distributed clustering (where different computer clusters should contain similar numbers of network nodes), market segmentation (where each customer segment will subsequently be addressed by a marketing campaign) and document clustering (where topic hierarchies should display a balanced view of the available documents), see Banerjee and Ghosh (2006) and Balcan et al. (2013). Finally, and perhaps most importantly, K -means clustering is highly sensitive to outliers. To illustrate this, consider the dataset in Figure 1, which accommodates three clusters as well as three individual outliers. The K -means clustering problem erroneously merges two of the three clusters in order to assign the three outliers to the third cluster (top left graph), whereas a clustering that disregards the three outliers would recover the true clusters and result in a significantly lower objective value (bottom left graph). The cardinality-constrained K -means clustering problem, where the cardinality of each cluster is set to be one third of all datapoints, shows a similar behavior on this dataset (graphs on the right). We will argue below, however, that the cardinality-constrained K -means clustering problem (1) offers an intuitive and mathematically rigorous framework to robustify K -means clustering against outliers.

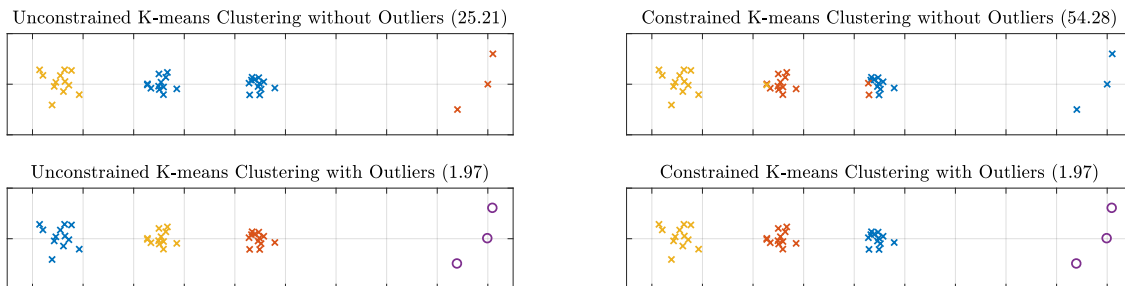


Figure 1 Sensitivity of the (un)constrained K -means clustering problem to outliers.

To our best knowledge, to date only two solution approaches have been proposed for problem (1). Bennett et al. (2000) combine a classical local search heuristic for the unconstrained K -means clustering problem due to Lloyd (1982) with the repeated solution of linear assignment problems to solve a variant of problem (1) that imposes lower bounds on the cluster sizes n_1, \dots, n_K . Banerjee and Ghosh (2006) solve the balanced version of problem (1), where $n_1 = \dots = n_K$, by sampling

a subset of the datapoints, performing a clustering on this subset, and subsequently populating the resulting clusters with the remaining datapoints while adhering to the cardinality constraints. Although the local search schemes of Bennett et al. (2000) and Banerjee and Ghosh (2006) tend to quickly produce solutions of high quality, they are not guaranteed to terminate in polynomial time, they do not provide bounds on the suboptimality of the identified solutions, and their performance may be sensitive to the choice of the initial solution. Moreover, neither of these local search schemes accommodates for outliers.

In recent years, several conic optimization schemes have been proposed to alleviate the shortcomings of these local search methods for the unconstrained K -means clustering problem (Peng and Wei 2007, Awasthi et al. 2015). Peng and Wei (2007) develop two semidefinite programming relaxations of the unconstrained K -means clustering problem. Their weaker relaxation admits optimal solutions that can be characterized by means of an eigenvalue decomposition. They further use this eigenvalue decomposition to set up a modified K -means clustering problem where the dimensionality of the datapoints is reduced to $K - 1$ (provided that their original dimensionality was larger than that). To obtain an upper bound, they solve this K -means clustering problem of reduced dimensionality, which can be done either exactly by enumerating Voronoi partitions, as described in Inaba et al. (1994), or by approximation methods such as those in Hasegawa et al. (1993). Using either approach, the runtime grows polynomially in the number of datapoints N but not in the number of desired clusters K . Hence, this method is primarily suitable for small K . Similar conic approximation schemes have been developed by Elhamifar et al. (2012) and Nellore and Ward (2015) in the context of unconstrained exemplar-based clustering.

Awasthi et al. (2015) and Iguchi et al. (2017) develop probabilistic recovery guarantees for the stronger semidefinite relaxation of Peng and Wei (2007) when the data is generated by a stochastic ball model (*i.e.*, datapoints are drawn randomly from rotation symmetric distributions supported on unit balls). More specifically, they use primal-dual arguments to establish conditions on the cluster separation under which the semidefinite relaxation of Peng and Wei (2007) recovers the underlying clusters with high probability as the number of data points N increases. The condition of Awasthi et al. (2015) requires less separation in low dimensions, while the condition of Iguchi et al. (2017) is less restrictive in high dimensions. In addition, Awasthi et al. (2015) consider a linear programming relaxation of the unconstrained K -means clustering problem, and they derive similar recovery guarantees for this relaxation as well.

Two more papers study the recovery guarantees of conic relaxations under a stochastic block model (*i.e.*, the dataset is characterized by a similarity matrix where the expected pairwise similarities of points in the same cluster are higher than those of points in different clusters). Ames (2014) considers the densest K -disjoint-clique problem whose aim is to split a given complete graph into

K subgraphs such as to maximize the sum of the average similarities of the resulting subgraphs. K -means clustering can be considered as a specific instance of this broader class of problems. By means of primal-dual arguments, the author derives conditions on the means in the stochastic block model such that his semidefinite relaxation recovers the underlying clusters with high probability as the cardinality of the smallest cluster increases. Vinayak and Hassibi (2016) develop a semidefinite relaxation and regularize it with the trace of the cluster assignment matrix. Using primal-dual arguments they show that, for specific ranges of the regularization parameter, their regularized semidefinite relaxation recovers the true clusters with high probability as the cardinality of the smallest cluster increases. The probabilistic recovery guarantees of Ames (2014) and Vinayak and Hassibi (2016) can also be extended to datasets containing outliers.

	Awasthi et al.	Iguchi et al.	Ames	Vinayak and Hassibi	This Paper
data generating model	stochastic ball	stochastic ball	stochastic block	stochastic block	none/arbitrary
type of relaxation	SDP + LP	SDP	SDP	SDP	SDP + LP
type of guarantee	stochastic	stochastic	stochastic	stochastic	deterministic
guarantee depends on N	yes	yes	yes	yes	no
guarantee depends on d	yes	yes	no	no	no
requires balancedness	yes	yes	no	no	yes
proof technique	primal-dual	primal-dual	primal-dual	primal-dual	valid cuts
access to cardinalities	no	no	no	no	yes
outlier detection	no	no	yes	yes	yes

Table 1 Comparison of Recovery Guarantees for K -means Clustering Relaxations.

In this paper, we propose the first conic optimization scheme for the cardinality-constrained K -means clustering problem (1). Our solution approach relies on an exact reformulation of problem (1) as an intractable mixed-integer linear program (MILP) to which we add a set of valid cuts before relaxing the resulting model to a tractable semidefinite program (SDP) or linear program (LP). The set of valid cuts is essential in strengthening these relaxations. Both relaxations provide lower bounds on the optimal value of problem (1), and they recover the optimal value of (1) whenever a cluster separation condition is met. Our relaxations also give rise to deterministic rounding schemes which produce feasible solutions that are provably optimal in (1) whenever the cluster separation condition holds. Table 1 compares our recovery guarantees to the results available in the literature. We emphasize that our guarantees are deterministic, that they apply to arbitrary data generating models, that they are dimension-independent, and that they hold for both our SDP and LP relaxations. Finally, our algorithms extend to instances of (1) that are contaminated by outliers and whose cluster cardinalities n_1, \dots, n_K are not known precisely. We summarize the paper’s contributions as follows.

1. We derive a novel MILP reformulation of problem (1) that only involves $\mathcal{O}(NK)$ binary variables, as opposed to the standard reformulation that contains $\Omega(N^2)$ binary variables.

2. We develop lower bounds which exploit the cardinality information in problem (1). Our bounds are tight whenever a cluster separation condition is met. Unlike similar results for other classes of clustering problems, our separation condition is deterministic, model-free and dimension-independent. Furthermore, our proof technique does not rely on the primal-dual argument of SDPs and LPs.
3. We propose deterministic rounding schemes that transform the relaxed solutions to feasible solutions for problem (1). The solutions are optimal in (1) if the separation condition holds. To our best knowledge, we propose the first tractable solution scheme for problem (1) with optimality guarantees.
4. We show that our lower bounds and rounding schemes extend to instances of problem (1) that are contaminated by outliers and whose cluster cardinalities are not known precisely.

The remainder of the paper is structured as follows. Section 2 analyzes the cardinality-constrained K -means clustering problem (1) and derives the MILP reformulation underlying our solution scheme. Sections 3 and 4 propose and analyze our conic rounding approaches for problem (1) in the absence and presence of outliers, respectively. Section 5 concludes with numerical experiments. Finally, a detailed description of closely related results (*i.e.*, the algorithm of Bennett et al. 2000 and the SDP relaxations of Peng and Wei 2007, Awasthi et al. 2015) is provided in the appendix.

Notation: We denote by $\mathbf{1}$ the vector of all ones and by $\|\cdot\|$ the Euclidean norm. For symmetric square matrices $\mathbf{A}, \mathbf{B} \in \mathbb{S}^N$, the relation $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite, while $\mathbf{A} \geq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is elementwise non-negative. Furthermore, we use $\text{diag}(\mathbf{A})$ to denote a vector in \mathbb{R}^N whose entries coincide with those of \mathbf{A} 's main diagonal. Conversely, for a vector $\mathbf{a} \in \mathbb{R}^N$, $\text{diag}(\mathbf{a})$ represents a diagonal matrix in \mathbb{S}^N with \mathbf{a} on its main diagonal. Finally, $\langle \mathbf{A}, \mathbf{B} \rangle = \text{Tr}(\mathbf{AB})$ denotes the trace inner product of \mathbf{A} and \mathbf{B} .

2. Problem Formulation and Analysis

We first prove that the clustering problem (1) is an instance of a *quadratic assignment problem* and transform (1) to an MILP with NK binary variables. Then, we discuss the complexity of (1) and show that an optimal clustering always corresponds to some Voronoi partition of \mathbb{R}^d . Throughout the paper we use $\mathbf{D} \in \mathbb{S}^N$ to denote the squared distance matrix with entries $d_{ij} = \|\xi_i - \xi_j\|^2$.

Our first result relies on the following auxiliary lemma, which we state without proof.

LEMMA 1. *For any vectors $\xi_1, \dots, \xi_n \in \mathbb{R}^d$, we have*

$$\sum_{i=1}^n \|\xi_i - (\sum_{j=1}^n \xi_j)/n\|^2 = \frac{1}{2n} \sum_{i,j=1}^n \|\xi_i - \xi_j\|^2.$$

Proof See Zha et al. (2002, p. 1060). □

PROPOSITION 1 (Quadratic Assignment Reformulation). *The clustering problem (1) can be cast as the quadratic assignment problem*

$$\underset{\sigma \in \mathfrak{S}^N}{\text{minimize}} \quad \frac{1}{2} \langle \mathbf{W}, \mathbf{P}_\sigma \mathbf{D} \mathbf{P}_\sigma^\top \rangle, \quad (2)$$

where $\mathbf{W} \in \mathbb{S}^N$ is a block diagonal matrix with blocks $\frac{1}{n_k} \mathbf{1}\mathbf{1}^\top \in \mathbb{S}^{n_k}$, $k = 1, \dots, K$, \mathfrak{S}^N is the set of permutations of $\{1, \dots, N\}$, and \mathbf{P}_σ is defined through $(\mathbf{P}_\sigma)_{ij} = 1$ if $\sigma(i) = j$; $= 0$ otherwise.

Proof We show that for any feasible solution of (1) there exists a feasible solution of (2) which attains the same objective value and vice versa. To this end, for any partition (I_1, \dots, I_K) feasible in (1), consider any permutation $\sigma \in \mathfrak{S}^N$ that satisfies $\sigma(\{1 + \sum_{i=1}^{k-1} n_i, \dots, \sum_{i=1}^k n_i\}) = I_k$ for all $k = 1, \dots, K$. This permutation is feasible in (2), and it achieves the same objective value as (I_1, \dots, I_K) in (1) because

$$\begin{aligned} \sum_{k=1}^K \sum_{i \in I_k} \|\xi_i - (\sum_{j \in I_k} \xi_j) / n_k\|^2 &= \frac{1}{2} \sum_{k=1}^K \frac{1}{n_k} \sum_{i, j \in I_k} d_{ij} = \frac{1}{2} \sum_{k=1}^K \frac{1}{n_k} \sum_{i, j \in \sigma^{-1}(I_k)} d_{\sigma(i)\sigma(j)} \\ &= \frac{1}{2} \langle \mathbf{W}, \mathbf{P}_\sigma \mathbf{D} \mathbf{P}_\sigma^\top \rangle, \end{aligned}$$

where the first equality is implied by Lemma 1, the second equality is a consequence of the definition of σ , and the third equality follows from the definition of \mathbf{W} .

Conversely, for any $\sigma \in \mathfrak{S}^N$ feasible in (2), consider any partition (I_1, \dots, I_K) satisfying $I_k = \sigma(\{1 + \sum_{i=1}^{k-1} n_i, \dots, \sum_{i=1}^k n_i\})$ for all $k = 1, \dots, K$. This partition is feasible in (1), and a similar reasoning as before shows that the partition achieves the same objective value as σ in (2). \square

Generic quadratic assignment problems with N facilities and N locations can be reformulated as MILPs with $\Omega(N^2)$ binary variables via the Kaufmann and Broeckx linearization; see *e.g.*, Burkard et al. (1998). In Theorem 1 below we will show, however, that the intra-cluster permutation symmetry of the samples enables us to reduce the number of binary variables to $NK \ll \Omega(N^2)$. We also emphasize that existing MILP formulations of the cardinality-constrained clustering problem (1) involve $\Omega(N^2)$ binary variables; see Mulvey and Beck (1984).

THEOREM 1 (MILP Reformulation). *The clustering problem (1) is equivalent to the MILP*

$$\begin{aligned} \underset{}{\text{minimize}} \quad & \frac{1}{2} \sum_{k=1}^K \frac{1}{n_k} \sum_{i, j=1}^N d_{ij} \eta_{ij}^k \\ \text{subject to} \quad & \pi_i^k \in \{0, 1\}, \eta_{ij}^k \in \mathbb{R}_+ \quad i, j = 1, \dots, N, \quad k = 1, \dots, K \\ & \sum_{i=1}^N \pi_i^k = n_k \quad k = 1, \dots, K \\ & \sum_{k=1}^K \pi_i^k = 1 \quad i = 1, \dots, N \\ & \eta_{ij}^k \geq \pi_i^k + \pi_j^k - 1 \quad i, j = 1, \dots, N, \quad k = 1, \dots, K. \end{aligned} \quad (\mathcal{P})$$

The binary variable π_i^k in the MILP \mathcal{P} satisfies $\pi_i^k = 1$ if $i \in I_k$; $= 0$ otherwise. At optimality, $\eta_{ij}^k = \max\{\pi_i^k + \pi_j^k - 1, 0\}$ is equal to 1 iff $i, j \in I_k$ (i.e., $\pi_i^k = \pi_j^k = 1$) and 0 otherwise.

Proof of Theorem 1 At optimality, the decision variables η_{ij}^k in problem \mathcal{P} take the values $\eta_{ij}^k = \max\{\pi_i^k + \pi_j^k - 1, 0\}$. Accordingly, problem \mathcal{P} can equivalently be stated as

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j=1}^N d_{ij} \max\{\pi_i^k + \pi_j^k - 1, 0\} \\ & \text{subject to} && \pi_i^k \in \{0, 1\} && i = 1, \dots, N, \quad k = 1, \dots, K \\ & && \sum_{i=1}^N \pi_i^k = n_k && k = 1, \dots, K \\ & && \sum_{k=1}^K \pi_i^k = 1 && i = 1, \dots, N. \end{aligned} \tag{P'}$$

In the following, we show that any feasible solution of (1) gives rise to a feasible solution of \mathcal{P}' with the same objective value and vice versa. To this end, consider first a partition (I_1, \dots, I_K) that is feasible in (1). Choosing $\pi_i^k = 1$ if $i \in I_k$ and $\pi_i^k = 0$ otherwise, $k = 1, \dots, K$, is feasible in \mathcal{P}' and attains the same objective value as (I_1, \dots, I_K) in (1) since

$$\sum_{k=1}^K \sum_{i \in I_k} \|\xi_i - (\sum_{j \in I_k} \xi_j)/n_k\|^2 = \frac{1}{2} \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j \in I_k} d_{ij} = \frac{1}{2} \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j=1}^N d_{ij} \max\{\pi_i^k + \pi_j^k - 1, 0\}.$$

Here, the first equality is implied by Lemma 1, and the second equality follows from the construction of π_i^k . By the same argument, every π_i^k feasible in \mathcal{P}' gives rise to a partition (I_1, \dots, I_K) , $I_k = \{i : \pi_i^k = 1\}$ for $k = 1, \dots, K$, that is feasible in \mathcal{P}' and that attains the same objective value. \square

PROPOSITION 2. *K-means clustering with cardinality constraints is NP-hard even for $K = 2$. Hence, unless $P = NP$, there is no polynomial time algorithm for solving problem (1).*

Proof In analogy to Theorem 1, one can show that the unconstrained K -means clustering problem can be formulated as a variant of problem \mathcal{P} that omits the first set of assignment constraints, which require that $\sum_{i=1}^N \pi_i^k = n_k$ for all $k = 1, \dots, K$, and replaces the (now unconstrained) cardinality n_k in the objective function by the size of I_k , which can be expressed as $\sum_{i=1}^N \pi_i^k$. If $K = 2$, we can thus solve the unconstrained K -means clustering problem by solving problem \mathcal{P} for all cluster cardinality combinations $(n_1, n_2) \in \{(1, N-1), (2, N-2), \dots, (\lfloor N/2 \rfloor, \lceil N/2 \rceil)\}$ and selecting the clustering with the lowest objective value. Thus, in this case, if problem \mathcal{P} was polynomial-time solvable, then so would be the unconstrained K -means clustering problem. This, however, would contradict Theorem 1 in Aloise et al. (2009), which shows that the unconstrained K -means clustering problem is NP-hard even for $K = 2$ clusters. \square

In K -means clustering *without* cardinality constraints, the convex hulls of the optimal clusters do not overlap, and thus each cluster fits within a separate cell of a Voronoi partition of \mathbb{R}^d ; see e.g., Hasegawa et al. (1993, Theorem 2.1). We demonstrate below that this property is preserved in the presence of cardinality constraints.

THEOREM 2 (Voronoi Partition). *For every optimal solution to problem (1), there exists a Voronoi partition of \mathbb{R}^d such that each cluster is contained in exactly one Voronoi cell.*

Proof We show that for every optimal clustering (I_1, \dots, I_K) of (1) and every $k, \ell \in \{1, \dots, K\}$, $k < \ell$, there exists a hyperplane separating the points in I_k from those in I_ℓ . This in turn implies the existence of the desired Voronoi partition. Denote the centers of the clusters I_k and I_ℓ by

$$\zeta_k = \frac{1}{n_k} \sum_{i \in I_k} \xi_i \quad \text{and} \quad \zeta_\ell = \frac{1}{n_\ell} \sum_{i \in I_\ell} \xi_i,$$

respectively, and let $\mathbf{h} = \zeta_k - \zeta_\ell$ be the vector that connects the two centers. The statement holds if $\mathbf{h}^\top(\xi_{i_k} - \xi_{i_\ell}) \geq 0$ for all $i_k \in I_k$ and $i_\ell \in I_\ell$ as \mathbf{h} itself determines a separating hyperplane for I_k and I_ℓ in that case. We thus assume that $\mathbf{h}^\top(\xi_{i_k} - \xi_{i_\ell}) < 0$ for some $i_k \in I_k$ and $i_\ell \in I_\ell$. However, this contradicts the optimality of the clustering (I_1, \dots, I_K) because

$$\begin{aligned} \mathbf{h}^\top(\xi_{i_k} - \xi_{i_\ell}) < 0 &\iff (\zeta_k - \zeta_\ell)^\top(\xi_{i_k} - \xi_{i_\ell}) < 0 \\ &\iff \xi_{i_k}^\top \zeta_k + \xi_{i_\ell}^\top \zeta_\ell < \xi_{i_k}^\top \zeta_\ell + \xi_{i_\ell}^\top \zeta_k \\ &\iff \|\xi_{i_\ell} - \zeta_k\|^2 + \|\xi_{i_k} - \zeta_\ell\|^2 < \|\xi_{i_k} - \zeta_k\|^2 + \|\xi_{i_\ell} - \zeta_\ell\|^2, \end{aligned}$$

where the last equivalence follows from multiplying both sides of the second inequality with 2 and then completing the squares by adding $\xi_{i_k}^\top \xi_{i_k} + \zeta_k^\top \zeta_k + \xi_{i_\ell}^\top \xi_{i_\ell} + \zeta_\ell^\top \zeta_\ell$ on both sides. Defining $\tilde{I}_k = I_k \cup \{i_\ell\} \setminus \{i_k\}$ and $\tilde{I}_\ell = I_\ell \cup \{i_k\} \setminus \{i_\ell\}$, the above would imply that

$$\begin{aligned} &\sum_{i \in \tilde{I}_k} \|\xi_i - \zeta_k\|^2 + \sum_{i \in \tilde{I}_\ell} \|\xi_i - \zeta_\ell\|^2 + \sum_{\substack{m=1, \dots, K \\ m \notin \{k, \ell\}}} \sum_{i \in I_m} \|\xi_i - \zeta_m\|^2 \\ &< \sum_{i \in I_k} \|\xi_i - \zeta_k\|^2 + \sum_{i \in I_\ell} \|\xi_i - \zeta_\ell\|^2 + \sum_{\substack{m=1, \dots, K \\ m \notin \{k, \ell\}}} \sum_{i \in I_m} \|\xi_i - \zeta_m\|^2, \end{aligned}$$

where ζ_m is defined analogously to ζ_k and ζ_ℓ . The left-hand side of the above inequality represents an upper bound on the sum of squared intra-cluster distances attained by the clustering $(I_1, \dots, \tilde{I}_k, \dots, \tilde{I}_\ell, \dots, I_K)$ since ζ_k and ζ_ℓ may not coincide with the minimizers $\frac{1}{n_k} \sum_{i \in \tilde{I}_k} \xi_i$ and $\frac{1}{n_\ell} \sum_{i \in \tilde{I}_\ell} \xi_i$, respectively. We thus conclude that the clustering $(I_1, \dots, \tilde{I}_k, \dots, \tilde{I}_\ell, \dots, I_K)$ attains a strictly lower objective value than (I_1, \dots, I_K) in problem (1), which is a contradiction. \square

3. Cardinality-Constrained Clustering without Outliers

We now relax the intractable MILP \mathcal{P} to tractable conic programs that yield efficiently computable lower and upper bounds on \mathcal{P} .

3.1. Convex Relaxations and Rounding Algorithm

We first eliminate the η_{ij}^k variables from \mathcal{P} by re-expressing the problem's objective function as

$$\frac{1}{2} \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j=1}^N d_{ij} \eta_{ij}^k = \frac{1}{2} \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j=1}^N d_{ij} \max\{\pi_i^k + \pi_j^k - 1, 0\} = \frac{1}{2} \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j=1}^N d_{ij} \pi_i^k \pi_j^k,$$

where the last equality holds because the variables π_i^k are binary. Next, we apply the variable transformation $x_i^k \leftarrow 2\pi_i^k - 1$, whereby \mathcal{P} simplifies to

$$\begin{aligned} & \text{minimize} && \frac{1}{8} \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j=1}^N d_{ij} (1 + x_i^k)(1 + x_j^k) \\ & \text{subject to} && x_i^k \in \{-1, +1\} && i = 1, \dots, N, \quad k = 1, \dots, K \\ & && \sum_{i=1}^N x_i^k = 2n_k - N && k = 1, \dots, K \\ & && \sum_{k=1}^K x_i^k = 2 - K && i = 1, \dots, N. \end{aligned} \tag{3}$$

Here, x_i^k takes the value $+1$ if the i -th datapoint is assigned to cluster k and -1 otherwise. Note that the constraints in (3) are indeed equivalent to the first two constraints in \mathcal{P} , respectively. In Theorem 3 below we will show that the reformulation (3) of the MILP \mathcal{P} admits the SDP relaxation

$$\begin{aligned} & \text{minimize} && \frac{1}{8} \left\langle \mathbf{D}, \sum_{k=1}^K \frac{1}{n_k} (\mathbf{M}^k + \mathbf{1}\mathbf{1}^\top + \mathbf{x}^k \mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top) \right\rangle \\ & \text{subject to} && (\mathbf{x}^k, \mathbf{M}^k) \in \mathcal{C}_{\text{SDP}}(n_k) \quad k = 1, \dots, K \\ & && \sum_{k=1}^K \mathbf{x}^k = (2 - K)\mathbf{1}, \end{aligned} \tag{\mathcal{R}_{\text{SDP}}}$$

where, for any $n \in \mathbb{N}$, the convex set $\mathcal{C}_{\text{SDP}}(n) \subset \mathbb{R}^N \times \mathbb{S}^N$ is defined as

$$\mathcal{C}_{\text{SDP}}(n) = \left\{ (\mathbf{x}, \mathbf{M}) \in \mathbb{R}^N \times \mathbb{S}^N : \begin{aligned} & \mathbf{1}^\top \mathbf{x} = 2n - N, \quad \mathbf{M}\mathbf{1} = (2n - N)\mathbf{x} \\ & \text{diag}(\mathbf{M}) = \mathbf{1}, \quad \mathbf{M} \succeq \mathbf{x}\mathbf{x}^\top \\ & \mathbf{M} + \mathbf{1}\mathbf{1}^\top + \mathbf{x}\mathbf{1}^\top + \mathbf{1}\mathbf{x}^\top \geq \mathbf{0} \\ & \mathbf{M} + \mathbf{1}\mathbf{1}^\top - \mathbf{x}\mathbf{1}^\top - \mathbf{1}\mathbf{x}^\top \geq \mathbf{0} \\ & \mathbf{M} - \mathbf{1}\mathbf{1}^\top + \mathbf{x}\mathbf{1}^\top - \mathbf{1}\mathbf{x}^\top \leq \mathbf{0} \\ & \mathbf{M} - \mathbf{1}\mathbf{1}^\top - \mathbf{x}\mathbf{1}^\top + \mathbf{1}\mathbf{x}^\top \leq \mathbf{0} \end{aligned} \right\}.$$

Note that $\mathcal{C}_{\text{SDP}}(n)$ is semidefinite-representable because Schur's complement allows us to express the constraint $\mathbf{M} \succeq \mathbf{x}\mathbf{x}^\top$ as a linear matrix inequality; see, *e.g.*, Boyd and Vandenberghe (2004).

We can further relax the above SDP to an LP, henceforth denoted by \mathcal{R}_{LP} , where the constraints $(\mathbf{x}^k, \mathbf{M}^k) \in \mathcal{C}_{\text{SDP}}(n_k)$ are replaced with $(\mathbf{x}^k, \mathbf{M}^k) \in \mathcal{C}_{\text{LP}}(n_k)$, and where, for any $n \in \mathbb{N}$, the polytope $\mathcal{C}_{\text{LP}}(n)$ is obtained by removing the non-linear constraint $\mathbf{M} \succeq \mathbf{x}\mathbf{x}^\top$ from $\mathcal{C}_{\text{SDP}}(n)$.

THEOREM 3 (SDP and LP Relaxations). *We have $\min \mathcal{R}_{\text{LP}} \leq \min \mathcal{R}_{\text{SDP}} \leq \min \mathcal{P}$.*

Proof The inequality $\min \mathcal{R}_{\text{LP}} \leq \min \mathcal{R}_{\text{SDP}}$ is trivially satisfied because $\mathcal{C}_{\text{SDP}}(n)$ is constructed as a subset of $\mathcal{C}_{\text{LP}}(n)$ for every $n \in \mathbb{N}$. To prove the inequality $\min \mathcal{R}_{\text{SDP}} \leq \min \mathcal{P}$, consider any set of binary vectors $\{\mathbf{x}^k\}_{k=1}^K$ feasible in (3) and define $\mathbf{M}^k = \mathbf{x}^k(\mathbf{x}^k)^\top$ for $k = 1, \dots, K$. By construction,

the objective value of $\{\mathbf{x}^k\}_{k=1}^K$ in (3) coincides with that of $\{(\mathbf{x}^k, \mathbf{M}^k)\}_{k=1}^K$ in \mathcal{R}_{SDP} . Moreover, the constraints in (3) imply that

$$\mathbf{M}^k \mathbf{1} = \mathbf{x}^k (\mathbf{x}^k)^\top \mathbf{1} = (2n_k - N) \mathbf{x}^k, \quad \text{diag}(\mathbf{M}^k) = \mathbf{1}, \quad \mathbf{M}^k \succeq \mathbf{x}^k (\mathbf{x}^k)^\top$$

and

$$\mathbf{M}^k + \mathbf{1}\mathbf{1}^\top + \mathbf{x}^k \mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top = +(\mathbf{1} + \mathbf{x}^k)(\mathbf{1} + \mathbf{x}^k)^\top \geq \mathbf{0}$$

$$\mathbf{M}^k + \mathbf{1}\mathbf{1}^\top - \mathbf{x}^k \mathbf{1}^\top - \mathbf{1}(\mathbf{x}^k)^\top = +(\mathbf{1} - \mathbf{x}^k)(\mathbf{1} - \mathbf{x}^k)^\top \geq \mathbf{0}$$

$$\mathbf{M}^k - \mathbf{1}\mathbf{1}^\top + \mathbf{x}^k \mathbf{1}^\top - \mathbf{1}(\mathbf{x}^k)^\top = -(\mathbf{1} - \mathbf{x}^k)(\mathbf{1} + \mathbf{x}^k)^\top \leq \mathbf{0}$$

$$\mathbf{M}^k - \mathbf{1}\mathbf{1}^\top - \mathbf{x}^k \mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top = -(\mathbf{1} + \mathbf{x}^k)(\mathbf{1} - \mathbf{x}^k)^\top \leq \mathbf{0},$$

which ensures that $(\mathbf{x}^k, \mathbf{M}^k) \in \mathcal{C}_{\text{SDP}}(n_k)$ for every k . Finally, the constraint $\sum_{k=1}^K \mathbf{x}^k = (2 - K)\mathbf{1}$ in \mathcal{R}_{SDP} coincides with the last constraint in (3). Thus, $\{(\mathbf{x}^k, \mathbf{M}^k)\}_{k=1}^K$ is feasible in \mathcal{R}_{SDP} . The desired inequality now follows because any feasible point in (3) corresponds to a feasible point in \mathcal{R}_{SDP} with the same objective value. Note that the converse implication is generally false. \square

REMARK 1. In the special case when $K = 2$, we can half the number of variables in \mathcal{R}_{SDP} and \mathcal{R}_{LP} by setting $\mathbf{x}^2 = -\mathbf{x}^1$ and $\mathbf{M}^2 = \mathbf{M}^1$ without loss of generality.

Next, we develop a rounding algorithm that recovers a feasible clustering (and thus an upper bound on \mathcal{P}) from an optimal solution of the relaxed problem \mathcal{R}_{SDP} or \mathcal{R}_{LP} ; see Algorithm 1.

Algorithm 1 Rounding algorithm for cardinality-constrained clustering

- 1: **Input:** $\mathcal{I}_1 = \{1, \dots, N\}$ (data indices), $n_k \in \mathbb{N}$, $k = 1, \dots, K$ (cluster sizes).
- 2: Solve \mathcal{R}_{SDP} or \mathcal{R}_{LP} for the datapoints $\boldsymbol{\xi}_i$, $i \in \mathcal{I}_1$, and record the optimal $\mathbf{x}^1, \dots, \mathbf{x}^K \in \mathbb{R}^N$.
- 3: Solve the linear assignment problem

$$\boldsymbol{\Pi}' \in \underset{\boldsymbol{\Pi}}{\operatorname{argmax}} \left\{ \sum_{i=1}^N \sum_{k=1}^K \pi_i^k x_i^k : \pi_i^k \in \{0, 1\}, \sum_{i=1}^N \pi_i^k = n_k \quad \forall k, \sum_{k=1}^K \pi_i^k = 1 \quad \forall i \right\}.$$

- 4: Set $I'_k \leftarrow \{i : (\pi')_i^k = 1\}$ for all $k = 1, \dots, K$.
- 5: Set $\boldsymbol{\zeta}_k \leftarrow \frac{1}{n_k} \sum_{i \in I'_k} \boldsymbol{\xi}_i$ for all $k = 1, \dots, K$.
- 6: Solve the linear assignment problem

$$\boldsymbol{\Pi}^* \in \underset{\boldsymbol{\Pi}}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \sum_{k=1}^K \pi_i^k \|\boldsymbol{\xi}_i - \boldsymbol{\zeta}_k\|^2 : \pi_i^k \in \{0, 1\}, \sum_{i=1}^N \pi_i^k = n_k \quad \forall k, \sum_{k=1}^K \pi_i^k = 1 \quad \forall i \right\}.$$

- 7: Set $I_k \leftarrow \{i : (\pi^*)_i^k = 1\}$ for all $k = 1, \dots, K$.
 - 8: **Output:** I_1, \dots, I_K .
-

Recall that the continuous variables $\mathbf{x}^k = (x_1^k, \dots, x_N^k)^\top$ in \mathcal{R}_{SDP} and \mathcal{R}_{LP} correspond to the binary variables in (3) with identical names. This correspondence motivates us to solve a linear

assignment problem in Step 3 of Algorithm 1, which seeks a matrix $\mathbf{\Pi} \in \{0, 1\}^{N \times K}$ with $\pi_i^k \approx \frac{1}{2}(x_i^k + 1)$ for all i and k subject to the prescribed cardinality constraints. Note that even though this assignment problem constitutes an MILP, it can be solved in polynomial time because its constraint matrix is totally unimodular, implying that its LP relaxation is exact. Alternatively, one may solve the assignment problem using the Hungarian algorithm; see, *e.g.*, Burkard et al. (2009).

Note that Steps 5–7 of Algorithm 1 are reminiscent of a *single* iteration of Lloyd’s algorithm for cardinality-constrained K -means clustering as described by Bennett et al. (2000). Specifically, Step 5 calculates the cluster centers ζ_k , while Steps 6 and 7 reassign each point to the nearest center while adhering to the cardinality constraints. Algorithm 1 thus follows just one step of Lloyd’s algorithm initialized with an optimizer of \mathcal{R}_{SDP} or \mathcal{R}_{LP} . This refinement step ensures that the output clustering is compatible with a Voronoi partition of \mathbb{R}^d , which is desirable in view of Theorem 2.

3.2. Tighter Relaxations for Balanced Clustering

The computational burden of solving \mathcal{R}_{SDP} and \mathcal{R}_{LP} grows with K . We show in this section that if all clusters share the same size n (*i.e.*, $n_k = n$ for all k), then \mathcal{R}_{SDP} can be replaced by

$$\begin{aligned} & \text{minimize} \quad \frac{1}{8n} \langle \mathbf{D}, \mathbf{M}^1 + \mathbf{1}\mathbf{1}^\top + \mathbf{x}^1\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^1)^\top + (K-1)(\mathbf{M} + \mathbf{1}\mathbf{1}^\top + \mathbf{x}\mathbf{1}^\top + \mathbf{1}\mathbf{x}^\top) \rangle \\ & \text{subject to} \quad (\mathbf{x}^1, \mathbf{M}^1), (\mathbf{x}, \mathbf{M}) \in \mathcal{C}_{\text{SDP}}(n), \quad \mathbf{x}^1 + (K-1)\mathbf{x} = (2-K)\mathbf{1}, \quad x_1^1 = 1, \end{aligned} \quad (\mathcal{R}_{\text{SDP}}^b)$$

whose size no longer scales with K . Similarly, \mathcal{R}_{LP} simplifies to the LP $\mathcal{R}_{\text{LP}}^b$ obtained from $\mathcal{R}_{\text{SDP}}^b$ by replacing $\mathcal{C}_{\text{SDP}}(n)$ with $\mathcal{C}_{\text{LP}}(n)$. This is a manifestation of how symmetry can be exploited to simplify convex programs, a phenomenon which is studied in a more general setting by Gatermann and Parrilo (2004).

COROLLARY 1 (Relaxations for Balanced Clustering). *We have $\min \mathcal{R}_{\text{LP}}^b \leq \min \mathcal{R}_{\text{SDP}}^b \leq \min \mathcal{P}$.*

Proof The inequality $\min \mathcal{R}_{\text{LP}}^b \leq \min \mathcal{R}_{\text{SDP}}^b$ is trivially satisfied. To prove the inequality $\min \mathcal{R}_{\text{SDP}}^b \leq \min \mathcal{P}$, we first add the symmetry breaking constraint $x_1^1 = 1$ to the MILP \mathcal{P} . Note that this constraint does not increase the optimal value of \mathcal{P} . It just requires that the cluster containing the datapoint ξ_1 should be assigned the number $k = 1$. This choice is unrestrictive because all clusters have the same size. By repeating the reasoning that led to Theorem 3, the MILP \mathcal{P} can then be relaxed to a variant of the SDP \mathcal{R}_{SDP} that includes the (linear) symmetry breaking constraint $x_1^1 = 1$. Note that the constraints and the objective function of the resulting SDP are invariant under permutations of the cluster indices $k = 2, \dots, K$ because $n_k = n$ for all k . Note also that the constraints are not invariant under permutations involving $k = 1$ due to the symmetry breaking constraint. Next, consider any feasible solution $\{(\mathbf{x}^k, \mathbf{M}^k)\}_{k=1}^K$ of this SDP, and define

$$\mathbf{x} = \frac{1}{K-1} \sum_{k=2}^K \mathbf{x}^k \quad \text{and} \quad \mathbf{M} = \frac{1}{K-1} \sum_{k=2}^K \mathbf{M}^k.$$

Moreover, construct a permutation-symmetric solution $\{(\mathbf{x}_s^k, \mathbf{M}_s^k)\}_{k=1}^K$ by setting

$$\begin{aligned}\mathbf{x}_s^1 &= \mathbf{x}^1, & \mathbf{x}_s^k &= \mathbf{x} & \forall k = 2, \dots, K, \\ \mathbf{M}_s^1 &= \mathbf{M}^1, & \mathbf{M}_s^k &= \mathbf{M} & \forall k = 2, \dots, K.\end{aligned}$$

By the convexity and permutation symmetry of the SDP, the symmetrized solution $\{(\mathbf{x}_s^k, \mathbf{M}_s^k)\}_{k=1}^K$ is also feasible in the SDP and attains the same objective value as $\{(\mathbf{x}^k, \mathbf{M}^k)\}_{k=1}^K$. Moreover, as the choice of $\{(\mathbf{x}^k, \mathbf{M}^k)\}_{k=1}^K$ was arbitrary, we may indeed restrict attention to symmetrized solutions with $\mathbf{x}^k = \mathbf{x}^\ell$ and $\mathbf{M}^k = \mathbf{M}^\ell$ for all $k, \ell \in \{2, \dots, K\}$ without increasing the objective value of the SDP. Therefore, the simplified SDP relaxation $\mathcal{R}_{\text{SDP}}^b$ provides a lower bound on \mathcal{P} . \square

If $n_k = n$ for all k , then the SDP and LP relaxations from Section 3.1 admit an optimal solution where both \mathbf{x}^k and \mathbf{M}^k are independent of k , in which case Algorithm 1 performs poorly. This motivates the improved relaxations $\mathcal{R}_{\text{SDP}}^b$ and $\mathcal{R}_{\text{LP}}^b$ involving the symmetry breaking constraint $x_1^1 = 1$, which ensures that—without loss of generality—the cluster harboring the first datapoint ξ_1 is indexed by $k = 1$. As the symmetry between clusters $2, \dots, K$ persists and because any additional symmetry breaking constraint would be restrictive, the optimal solutions of $\mathcal{R}_{\text{SDP}}^b$ and $\mathcal{R}_{\text{LP}}^b$ only facilitate a reliable recovery of cluster 1. To recover *all* clusters, however, we can solve $\mathcal{R}_{\text{SDP}}^b$ or $\mathcal{R}_{\text{LP}}^b$ $K - 1$ times over the yet unassigned datapoints, see Algorithm 2. The resulting clustering could be improved by appending one iteration of Lloyd’s algorithm (akin to Steps 5–7 in Algorithm 1).

Algorithm 2 Rounding algorithm for balanced clustering

- 1: **Input:** $\mathcal{I}_1 = \{1, \dots, N\}$ (data indices), $n \in \mathbb{N}$ (cluster size), $K = N/n \in \mathbb{N}$ (# clusters).
 - 2: **for** $k = 1, \dots, K - 1$ **do**
 - 3: Solve $\mathcal{R}_{\text{SDP}}^b$ or $\mathcal{R}_{\text{LP}}^b$ for the datapoints $\xi_i, i \in \mathcal{I}_k$, and record the optimal $\mathbf{x}^1 \in \mathbb{R}^{|\mathcal{I}_k|}$.
 - 4: Determine a bijection $\rho: \{1, \dots, |\mathcal{I}_k|\} \rightarrow \mathcal{I}_k$ such that $x_{\rho(1)}^1 \geq x_{\rho(2)}^1 \geq \dots \geq x_{\rho(|\mathcal{I}_k|)}^1$.
 - 5: Set $I_k \leftarrow \{\rho(1), \dots, \rho(n)\}$ and $\mathcal{I}_{k+1} \leftarrow \mathcal{I}_k \setminus I_k$.
 - 6: Set $I_K \leftarrow \mathcal{I}_K$.
 - 7: **Output:** I_1, \dots, I_K .
-

3.3. Perfect Recovery Guarantees

We now demonstrate that the relaxations of Section 3.2 are tight and that Algorithm 2 finds the optimal clustering if the clusters are perfectly separated in the sense of the following assumption.

(S) *Perfect Separation:* There exists a balanced partition (J_1, \dots, J_K) of $\{1, \dots, N\}$ where each cluster $k = 1, \dots, K$ has the same cardinality $|J_k| = N/K \in \mathbb{N}$, and

$$\max_{1 \leq k \leq K} \max_{i, j \in J_k} d_{ij} < \min_{1 \leq k_1 < k_2 \leq K} \min_{i \in J_{k_1}, j \in J_{k_2}} d_{ij}.$$

Assumption **(S)** implies that the dataset admits the natural balanced clustering (J_1, \dots, J_K) and that the diameter of each cluster is smaller than the distance between any two distinct clusters.

THEOREM 4. *If Assumption **(S)** holds, then the optimal values of $\mathcal{R}_{\text{LP}}^b$ and \mathcal{P} coincide. Moreover, the clustering (J_1, \dots, J_K) is optimal in \mathcal{P} and is recovered by Algorithm 2.*

Proof Throughout the proof we assume without loss of generality that the clustering (J_1, \dots, J_K) from Assumption **(S)** satisfies $1 \in J_1$, that is, the cluster containing the datapoint ξ_1 is assigned the number $k = 1$. The proof now proceeds in two steps. In the first step, we show that the optimal values of the LP $\mathcal{R}_{\text{LP}}^b$ and the MILP \mathcal{P} are equal and that they both coincide with the sum of squared intra-cluster distances of the clustering (J_1, \dots, J_K) , which amounts to

$$\frac{1}{2n} \sum_{k=1}^K \sum_{i,j \in J_k} d_{ij}.$$

In the second step we demonstrate that the output (I_1, \dots, I_K) of Algorithm 2 coincides with the optimal clustering (J_1, \dots, J_K) from Assumption **(S)**. As the algorithm uses the same procedure K times to recover the clusters one by one, it is actually sufficient to show that the first iteration of the algorithm correctly identifies the first cluster, that is, it suffices to prove that $I_1 = J_1$.

Step 1: For any feasible solution $(\mathbf{x}^1, \mathbf{x}, \mathbf{M}^1, \mathbf{M})$ of $\mathcal{R}_{\text{LP}}^b$, we define $\mathbf{H}, \mathbf{W} \in \mathbb{S}^N$ through

$$\mathbf{H} = \mathbf{M}^1 + \mathbf{1}\mathbf{1}^\top + \mathbf{x}^1\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^1)^\top \quad \text{and} \quad \mathbf{W} = \mathbf{M} + \mathbf{1}\mathbf{1}^\top + \mathbf{x}\mathbf{1}^\top + \mathbf{1}\mathbf{x}^\top. \quad (4)$$

From the definition of $\mathcal{C}_{\text{LP}}(n)$ it is clear that $\mathbf{H}, \mathbf{W} \geq \mathbf{0}$. Moreover, we also have that

$$\begin{aligned} \sum_{i \neq j} h_{ij} &= \sum_{i \neq j} m_{ij}^1 + N(N-1) + 2(N-1)(\mathbf{x}^1)^\top \mathbf{1} \\ &= (2n - N)^2 - N + N(N-1) + 2(N-1)(2n - N) = 4n(n-1). \end{aligned}$$

A similar calculation for \mathbf{W} reveals that $\sum_{i \neq j} w_{ij} = 4n(n-1)$. Next, we consider the objective function of $\mathcal{R}_{\text{LP}}^b$, which can be rewritten in terms of \mathbf{W} and \mathbf{H} as

$$\frac{1}{8n} \langle \mathbf{D}, \mathbf{H} + (K-1)\mathbf{W} \rangle = \frac{1}{8n} \sum_{i \neq j} d_{ij} (h_{ij} + (K-1)w_{ij}). \quad (5)$$

The sum on the right-hand side can be viewed as a weighted average of the squared distances d_{ij} with non-negative weights $h_{ij} + (K-1)w_{ij}$, where the total weight is given by

$$\sum_{i \neq j} (h_{ij} + (K-1)w_{ij}) = 4Kn(n-1).$$

From the definition of $\mathcal{C}_{\text{LP}}(n)$ we also know that

$$\begin{aligned} 2(\mathbf{M}^1 - \mathbf{1}\mathbf{1}^\top) &= (\mathbf{M}^1 - \mathbf{1}\mathbf{1}^\top + \mathbf{x}^1\mathbf{1}^\top - \mathbf{1}(\mathbf{x}^1)^\top) + (\mathbf{M}^1 - \mathbf{1}\mathbf{1}^\top - \mathbf{x}^1\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^1)^\top) \leq \mathbf{0} \implies \mathbf{M}^1 \leq \mathbf{1}\mathbf{1}^\top, \\ 2(\mathbf{M} - \mathbf{1}\mathbf{1}^\top) &= (\mathbf{M} - \mathbf{1}\mathbf{1}^\top + \mathbf{x}\mathbf{1}^\top - \mathbf{1}\mathbf{x}^\top) + (\mathbf{M} - \mathbf{1}\mathbf{1}^\top - \mathbf{x}\mathbf{1}^\top + \mathbf{1}\mathbf{x}^\top) \leq \mathbf{0} \implies \mathbf{M} \leq \mathbf{1}\mathbf{1}^\top. \end{aligned}$$

This further implies that each weight $h_{ij} + (K - 1)w_{ij}$ is bounded above by 4 because

$$\begin{aligned} h_{ij} + (K - 1)w_{ij} &= (m_{ij}^1 + 1 + x_i^1 + x_j^1) + (K - 1)(m_{ij} + 1 + x_i + x_j) \\ &\leq 2K + (x_i^1 + (K - 1)x_i) + (x_j^1 + (K - 1)x_j) = 4, \end{aligned} \quad (6)$$

where the inequality holds because $\mathbf{M}^1, \mathbf{M} \leq \mathbf{1}\mathbf{1}^\top$ and the last equality follows from the constraint $\mathbf{x}^1 + (K - 1)\mathbf{x} = (2 - K)\mathbf{1}$ in $\mathcal{R}_{\text{LP}}^b$.

Hence, the sum on the right hand side of (5) assigns each squared distance d_{ij} with $i \neq j$ a weight of at most 4, while the total weight equals $4Kn(n - 1)$. A lower bound on the sum is thus obtained by assigning a weight of 4 to the $Kn(n - 1)$ smallest values d_{ij} with $i \neq j$. Thus, we have

$$\begin{aligned} \frac{1}{8n} \langle \mathbf{D}, \mathbf{H} + (K - 1)\mathbf{W} \rangle &\geq \frac{1}{2n} \{\text{sum of the } Kn(n - 1) \text{ smallest entries of } d_{ij} \text{ with } i \neq j\} \\ &= \frac{1}{2n} \sum_{k=1}^K \sum_{i,j \in J_k} d_{ij}, \end{aligned} \quad (7)$$

where the last equality follows from Assumption (S). By Lemma 1, the right-hand side of (7) represents the objective value of the clustering (J_1, \dots, J_K) in the MILP \mathcal{P} . Thus, $\mathcal{R}_{\text{LP}}^b$ provides an upper bound on \mathcal{P} . By Corollary 1, $\mathcal{R}_{\text{LP}}^b$ also provides a lower bound on \mathcal{P} . We may thus conclude that the LP relaxation $\mathcal{R}_{\text{LP}}^b$ is tight and, as a consequence, that the clustering (J_1, \dots, J_K) is indeed optimal in \mathcal{P} .

Step 2: As the inequality in (7) is tight, any optimal solution to $\mathcal{R}_{\text{LP}}^b$ satisfies $h_{ij} + (K - 1)w_{ij} = 4$ whenever $i \neq j$ and $i, j \in J_k$ for some $k = 1, \dots, K$ (i.e., whenever the datapoints ξ_i and ξ_j belong to the same cluster). We will use this insight to show that Algorithm 2 outputs $I_1 = J_1$.

For any $i \in J_1$, the above reasoning and our convention that $1 \in J_1$ imply that $h_{1i} + (K - 1)w_{1i} = 4$. This in turn implies via (6) that $m_{1i}^1 = m_{1i} = 1$ for all $i \in J_1$.

From the definition of $\mathcal{C}_{\text{LP}}(n)$, we know that

$$2(\mathbf{M}^1 + \mathbf{1}\mathbf{1}^\top) = (\mathbf{M}^1 + \mathbf{1}\mathbf{1}^\top + \mathbf{x}^1\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^1)^\top) + (\mathbf{M}^1 + \mathbf{1}\mathbf{1}^\top - \mathbf{x}^1\mathbf{1}^\top - \mathbf{1}(\mathbf{x}^1)^\top) \geq \mathbf{0} \implies \mathbf{M}^1 \geq -\mathbf{1}\mathbf{1}^\top.$$

This allows us to conclude that

$$2n - N = \sum_{i=1}^N m_{1i}^1 = \sum_{i \in J_1} m_{1i}^1 + \sum_{i \notin J_1} m_{1i}^1 \geq n + (N - n)(-1) = 2n - N,$$

where the first equality holds because $\mathbf{M}^1\mathbf{1} = (2n - N)\mathbf{x}^1$, which is one of the constraints in $\mathcal{R}_{\text{LP}}^b$, and because of our convention that $x_1^1 = 1$. Hence, the above inequality must be satisfied as an equality, which in turn implies that $m_{1i}^1 = -1$ for all $i \notin J_1$.

For any $i \notin J_1$, the $(1, i)$ -th entry of the matrix inequality $\mathbf{M}^1 + \mathbf{1}\mathbf{1}^\top - \mathbf{x}^1\mathbf{1}^\top - \mathbf{1}(\mathbf{x}^1)^\top \geq \mathbf{0}$ from the definition of $\mathcal{C}_{\text{LP}}(n)$ can be expressed as

$$0 \leq m_{1i}^1 + 1 - x_1^1 - x_i^1 \quad \forall i = 1, \dots, N \implies x_i^1 \leq -1,$$

where the implication holds because $m_{1i}^1 = -1$ for $i \notin J_1$ and because $x_1^1 = 1$ due to the symmetry breaking constraint in $\mathcal{R}_{\text{LP}}^b$. Similarly, for any $i \in J_1$, the (i, i) -th entry of the matrix inequality $\mathbf{M}^1 + \mathbf{1}\mathbf{1}^\top - \mathbf{x}^1\mathbf{1}^\top - \mathbf{1}(\mathbf{x}^1)^\top \geq \mathbf{0}$ can be rewritten as

$$0 \leq m_{ii}^1 + 1 - 2x_i^1 \quad \forall i = 1, \dots, N \implies x_i^1 \leq 1,$$

where the implication follows from the constraint $\text{diag}(\mathbf{M}^1) = \mathbf{1}$ in $\mathcal{R}_{\text{LP}}^b$.

As $x_i^1 \leq 1$ for all $i \in J_1$ and $x_i^1 \leq -1$ for all $i \notin J_1$, the equality constraint $\mathbf{1}^\top \mathbf{x}^1 = 2n - N$ from the definition of $\mathcal{C}_{\text{LP}}(n)$ can only be satisfied if $x_i^1 = 1$ for all $i \in J_1$ and $x_i^1 = -1$ for all $i \notin J_1$.

Since Algorithm 2 constructs I_1 as the index set of the n largest entries of the vector \mathbf{x}^1 , we conclude that it must output $I_1 = J_1$ and the proof completes. \square

Theorem 4 implies via Corollary 1 that the optimal values of $\mathcal{R}_{\text{SDP}}^b$ and \mathcal{P} are also equal. Thus, both the LP and the SDP relaxation lead to perfect recovery.

In the related literature, Assumption **(S)** has previously been used by Elhamifar et al. (2012) to show that the natural clustering can be recovered in the context of unconstrained exemplar-based clustering whenever a regularization parameter is chosen appropriately. In contrast, our formulation does not rely on regularization parameters. Likewise, Theorem 4 is reminiscent of Theorem 9 by Awasthi et al. (2015) which formalizes the recovery properties of their LP relaxation for the unconstrained K -means clustering problem. Awasthi et al. (2015) assume, however, that the datapoints are drawn independently from a mixture of K isotropic distributions and provide a probabilistic recovery guarantee that improves with N and deteriorates with d . In contrast, our recovery guarantee for constrained clustering is deterministic, model-free and dimension-independent. If Assumption **(S)** holds, simpler algorithms than Algorithm 1 and 2 can be designed to recover the true clusters. It seems however unlikely that such approaches would perform well in a setting where Assumption **(S)** is not satisfied. In contrast, the numerical experiments of Section 5 suggest that Algorithms 1 and 2 perform well even if Assumption **(S)** is violated.

REMARK 2. To our best knowledge, there is no perfect recovery result for the cardinality-constrained K -means clustering algorithm by Bennett et al. (2000), see Appendix B, whose performance depends critically on its initialization. To see that it can be trapped in a local optimum, consider the $N = 4$ two-dimensional datapoints $\xi_1 = (0, 0)$, $\xi_2 = (a, 0)$, $\xi_3 = (a, b)$ and $\xi_4 = (0, b)$ with $0 < a < b$, and assume that we seek two balanced clusters. If the algorithm is initialized with the clustering $\{\{1, 4\}, \{2, 3\}\}$, then this clustering remains unchanged, and the algorithm terminates and reports a suboptimal solution with relative optimality gap $b^2/a^2 - 1$. In contrast, as Assumption **(S)** holds, Algorithm 2 recovers the optimal clustering $\{\{1, 2\}, \{3, 4\}\}$ by Theorem 4.

4. Cardinality-Constrained Clustering with Outliers

If the dataset is corrupted by outliers, then the optimal value of (1) may be high, indicating that the dataset admits no natural clustering. Note that the bounds from Section 3 could still be tight, *i.e.*, it is thinkable that the optimal clustering is far from ‘ideal’ even if it can be found with Algorithm 2. If we gradually remove datapoints that are expensive to assign to any cluster, however, we should eventually discover an ‘ideal’ low-cost clustering. In the extreme case, if we omit all but K datapoints, then the optimal value of (1) drops to zero, and Algorithm 2 detects the optimal clustering due to Theorem 4.

We now show that the results of Section 3 (particularly Theorems 1 and 3) extend to situations where n_0 datapoints must be assigned to an auxiliary *outlier cluster* indexed by $k = 0$ ($\sum_{k=0}^K n_k = N$), and where neither the distances between outliers and retained datapoints nor the distances between different outliers contribute to the objective function. In fact, we could equivalently postulate that each of the n_0 outliers forms a trivial singleton cluster. The use of cardinality constraints in integrated clustering and outlier detection has previously been considered by Chawla and Gionis (2013) in the context of local search heuristics. Inspired by this work, we henceforth minimize the sum of squared intra-cluster distances of the $N - n_0$ non-outlier datapoints. We first prove that the joint outlier detection and cardinality-constrained clustering problem admits an exact MILP reformulation.

THEOREM 5 (MILP Reformulation). *The joint outlier detection and cardinality-constrained clustering problem is equivalent to the MILP*

$$\begin{aligned}
& \text{minimize} && \frac{1}{2} \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j=1}^N d_{ij} \eta_{ij}^k \\
& \text{subject to} && \pi_i^k \in \{0, 1\}, \eta_{ij}^k \in \mathbb{R}_+ && i, j = 1, \dots, N, \ k = 0, \dots, K \\
& && \sum_{i=1}^N \pi_i^k = n_k && k = 0, \dots, K \\
& && \sum_{k=0}^K \pi_i^k = 1 && i = 1, \dots, N \\
& && \eta_{ij}^k \geq \pi_i^k + \pi_j^k - 1 && i, j = 1, \dots, N, \ k = 0, \dots, K.
\end{aligned} \tag{P^\circ}$$

Proof This is an immediate extension of Theorem 1 to account for the outlier cluster. \square

In analogy to Section 3.1, one can demonstrate that the MILP \mathcal{P}° admits the SDP relaxation

$$\begin{aligned}
& \text{minimize} && \frac{1}{8} \left\langle \mathbf{D}, \sum_{k=1}^K \frac{1}{n_k} (\mathbf{M}^k + \mathbf{1}\mathbf{1}^\top + \mathbf{x}^k \mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top) \right\rangle \\
& \text{subject to} && (\mathbf{x}^k, \mathbf{M}^k) \in \mathcal{C}_{\text{SDP}}(n_k) \quad k = 0, \dots, K \\
& && \sum_{k=0}^K \mathbf{x}^k = (1 - K)\mathbf{1}.
\end{aligned} \tag{\mathcal{R}_{\text{SDP}}^\circ}$$

Moreover, $\mathcal{R}_{\text{SDP}}^\circ$ can be further relaxed to an LP, henceforth denoted by $\mathcal{R}_{\text{LP}}^\circ$, by replacing the semidefinite representable set $\mathcal{C}_{\text{SDP}}(n_k)$ in $\mathcal{R}_{\text{SDP}}^\circ$ with the polytope $\mathcal{C}_{\text{LP}}(n_k)$ for all $k = 0, \dots, K$.

THEOREM 6 (SDP and LP Relaxations). *We have $\min \mathcal{R}_{\text{LP}}^{\circ} \leq \min \mathcal{R}_{\text{SDP}}^{\circ} \leq \min \mathcal{P}^{\circ}$.*

Proof This result generalizes Theorem 3 to account for the additional outlier cluster. As it requires no fundamentally new ideas, the proof is omitted for brevity. \square

The relaxations $\mathcal{R}_{\text{SDP}}^{\circ}$ and $\mathcal{R}_{\text{LP}}^{\circ}$ not only provide a lower bound on \mathcal{P}° , but they also give rise to a rounding algorithm that recovers a feasible clustering and thus an upper bound on \mathcal{P}° ; see Algorithm 3. Note that this procedure calls the outlier-unaware Algorithm 1 as a subroutine.

Algorithm 3 Rounding algorithm for joint outlier detection and cardinality-constrained clustering

- 1: **Input:** $\mathcal{I}_0 = \{1, \dots, N\}$ (data indices), $n_k \in \mathbb{N}$, $k = 0, \dots, K$ (cluster sizes).
 - 2: Solve $\mathcal{R}_{\text{SDP}}^{\circ}$ or $\mathcal{R}_{\text{LP}}^{\circ}$ for the datapoints ξ_i , $i \in \mathcal{I}_0$, and record the optimal $\mathbf{x}^0 \in \mathbb{R}^N$.
 - 3: Determine a bijection $\rho: \mathcal{I}_0 \rightarrow \mathcal{I}_0$ such that $x_{\rho(1)}^0 \geq x_{\rho(2)}^0 \geq \dots \geq x_{\rho(N)}^0$.
 - 4: Set $I_0 \leftarrow \{\rho(1), \dots, \rho(n_0)\}$ and $\mathcal{I}_1 \leftarrow \mathcal{I}_0 \setminus I_0$.
 - 5: Call Algorithm 1 with input $(\mathcal{I}_1, \{n_k\}_{k=1}^K)$ to obtain I_1, \dots, I_K .
 - 6: **Output:** I_0, \dots, I_K .
-

If all normal clusters are equally sized, *i.e.*, $n_k = n$ for $k = 1, \dots, K$, then $\mathcal{R}_{\text{SDP}}^{\circ}$ can be replaced by

$$\begin{aligned} & \text{minimize} \quad \frac{K}{8n} \langle \mathbf{D}, \mathbf{M} + \mathbf{1}\mathbf{1}^{\top} + \mathbf{x}\mathbf{1}^{\top} + \mathbf{1}\mathbf{x}^{\top} \rangle \\ & \text{subject to} \quad (\mathbf{x}, \mathbf{M}) \in \mathcal{C}_{\text{SDP}}(n), \quad (\mathbf{x}^0, \mathbf{M}^0) \in \mathcal{C}_{\text{SDP}}(n_0), \quad K\mathbf{x} + \mathbf{x}^0 = (1 - K)\mathbf{1}, \end{aligned} \quad (\mathcal{R}_{\text{SDP}}^{\text{ob}})$$

whose size no longer scales with K . Similarly, $\mathcal{R}_{\text{LP}}^{\circ}$ simplifies to the LP $\mathcal{R}_{\text{LP}}^{\text{ob}}$ obtained from $\mathcal{R}_{\text{SDP}}^{\text{ob}}$ by replacing $\mathcal{C}_{\text{SDP}}(n)$ and $\mathcal{C}_{\text{SDP}}(n_0)$ with $\mathcal{C}_{\text{LP}}(n)$ and $\mathcal{C}_{\text{LP}}(n_0)$, respectively. Note that the cardinality $n_0 = N - Kn$ may differ from n .

COROLLARY 2 (Relaxations for Balanced Clustering). *We have $\min \mathcal{R}_{\text{LP}}^{\text{ob}} \leq \min \mathcal{R}_{\text{SDP}}^{\text{ob}} \leq \min \mathcal{P}^{\circ}$.*

Proof This follows from a marginal modification of the argument that led to Corollary 1. \square

If the normal clusters are required to be balanced, then Algorithm 3 should be modified as follows. First, in Step 2 the relaxations $\mathcal{R}_{\text{SDP}}^{\text{ob}}$ or $\mathcal{R}_{\text{LP}}^{\text{ob}}$ can be solved instead of $\mathcal{R}_{\text{SDP}}^{\circ}$ or $\mathcal{R}_{\text{LP}}^{\circ}$, respectively. Moreover, in Step 5 Algorithm 2 must be called as a subroutine instead of Algorithm 1.

In the presence of outliers, the perfect recovery result from Theorem 4 remains valid if the following perfect separation condition is met, which can be viewed as a generalization of Assumption (S).

(S') Perfect Separation: There exists a partition (J_0, J_1, \dots, J_K) of $\{1, \dots, N\}$ where each normal cluster $k = 1, \dots, K$ has the same cardinality $|J_k| = (N - n_0)/K \in \mathbb{N}$, while

$$\max_{1 \leq k \leq K} \max_{i, j \in J_k} d_{ij} < \min_{1 \leq k_1 < k_2 \leq K} \min_{i \in J_{k_1}, j \in J_{k_2}} d_{ij} \quad \text{and} \quad \max_{1 \leq k \leq K} \max_{i, j \in J_k} d_{ij} < \min_{i \in \{1, \dots, N\}, j \in J_0} d_{ij}.$$

Assumption (S') implies that the dataset admits the natural outlier cluster J_0 and the natural normal clusters (J_1, \dots, J_K) . It also postulates that the diameter of each normal cluster is strictly smaller than (i) the distance between any two distinct normal clusters and (ii) the distance between any outlier and any other datapoint. Under this condition, Algorithm 3 correctly identifies the optimal clustering.

THEOREM 7. *If Assumption (S') holds, then the optimal values of $\mathcal{R}_{\text{LP}}^{\text{ob}}$ and \mathcal{P}° coincide. Moreover, the clustering (J_0, \dots, J_K) is optimal in \mathcal{P}° and is recovered by Algorithm 3.*

Proof The proof parallels that of Theorem 4 and can be divided into two steps. In the first step we show that the LP relaxation $\mathcal{R}_{\text{LP}}^{\text{ob}}$ for balanced clustering and outlier detection is tight, and in the second step we demonstrate that Algorithm 3 correctly identifies the clusters (J_0, \dots, J_K) . As for the second step, it suffices to prove that the algorithm correctly identifies the outlier cluster J_0 . Indeed, once the outliers are removed, the residual dataset satisfies Assumption (S), and Theorem 4 guarantees that the normal clusters (J_1, \dots, J_K) are correctly identified with Algorithm 2.

As a preliminary, note that $(\mathbf{x}, \mathbf{M}) \in \mathcal{C}_{\text{LP}}(n)$ implies

$$\text{diag}(\mathbf{M} + \mathbf{1}\mathbf{1}^\top + \mathbf{x}\mathbf{1}^\top + \mathbf{1}\mathbf{x}^\top) \geq \mathbf{0} \implies \mathbf{x} \geq -\mathbf{1},$$

$$\text{diag}(\mathbf{M} + \mathbf{1}\mathbf{1}^\top - \mathbf{x}\mathbf{1}^\top - \mathbf{1}\mathbf{x}^\top) \geq \mathbf{0} \implies \mathbf{x} \leq +\mathbf{1},$$

where the implications use $\text{diag}(\mathbf{M}) = \mathbf{1}$. Similarly, $(\mathbf{x}^0, \mathbf{M}^0) \in \mathcal{C}_{\text{LP}}(n_0)$ implies $-\mathbf{1} \leq \mathbf{x}^0 \leq +\mathbf{1}$.

Step 1: For any feasible solution $(\mathbf{x}^0, \mathbf{x}, \mathbf{M}^0, \mathbf{M})$ of $\mathcal{R}_{\text{LP}}^{\text{ob}}$, introduce the auxiliary matrix $\mathbf{H} = \mathbf{M} + \mathbf{1}\mathbf{1}^\top + \mathbf{1}\mathbf{x}^\top + \mathbf{x}\mathbf{1}^\top$. Recall from the proof of Theorem 4 that $\mathbf{H} \geq \mathbf{0}$ and

$$\sum_{i \neq j} h_{ij} = 4n(n-1).$$

The constraint $K\mathbf{x} + \mathbf{x}^0 = (1-K)\mathbf{1}$ from $\mathcal{R}_{\text{LP}}^{\text{ob}}$ ensures via the inequality $-\mathbf{1} \leq \mathbf{x}^0$ that $\mathbf{x} \leq (\frac{2}{K} - 1)\mathbf{1}$. Recalling from the proof of Theorem 4 that $\mathbf{M} \leq \mathbf{1}\mathbf{1}^\top$, we then find

$$h_{ij} = m_{ij} + 1 + x_i + x_j \leq 1 + 1 + \left(\frac{2}{K} - 1\right) + \left(\frac{2}{K} - 1\right) = \frac{4}{K} \quad \forall i, j = 1, \dots, N. \quad (8)$$

Similar arguments as in the proof of Theorem 4 reveal that the objective function of the joint outlier detection and (balanced) clustering problem $\mathcal{R}_{\text{LP}}^{\text{ob}}$ can be expressed as

$$\begin{aligned} \frac{K}{8n} \langle \mathbf{D}, \mathbf{H} \rangle &\geq \frac{1}{2n} \{\text{sum of the } Kn(n-1) \text{ smallest entries of } d_{ij} \text{ with } i \neq j\} \\ &= \frac{1}{2n} \sum_{k=1}^K \sum_{i,j \in J_k} d_{ij}, \end{aligned} \quad (9)$$

where the equality follows from Assumption (S'). By Lemma 1, the right-hand side of (9) represents the objective value of the clustering (J_0, \dots, J_K) in the MILP \mathcal{P}° . Thus, $\mathcal{R}_{\text{LP}}^{\text{ob}}$ provides an upper bound on \mathcal{P}° . By Corollary 2, $\mathcal{R}_{\text{LP}}^{\text{ob}}$ also provides a lower bound on \mathcal{P}° . We may thus conclude that the LP relaxation $\mathcal{R}_{\text{LP}}^{\text{ob}}$ is tight and, as a consequence, that the clustering (J_0, \dots, J_K) is indeed optimal in \mathcal{P}° .

Step 2: As the inequality in (9) is tight, any optimal solution to $\mathcal{R}_{\text{LP}}^{\text{ob}}$ satisfies $h_{ij} = \frac{4}{K}$ whenever $i \neq j$ and $i, j \in J_k$ for some $k = 1, \dots, K$ (i.e., whenever ξ_i and ξ_j are *not* outliers and belong to the same cluster). This in turn implies via (8) that $x_i = \frac{2}{K} - 1$ for all $i \in \cup_{k=1}^K J_k$. Furthermore, the constraint $\mathbf{1}^\top \mathbf{x} = 2n - N$ from $\mathcal{C}_{\text{LP}}(n)$ implies

$$2n - N = \sum_{k=1}^K \sum_{i \in J_k} x_i + \sum_{i \in J_0} x_i \geq Kn \left(\frac{2}{K} - 1 \right) + \sum_{i \in J_0} (-1) = 2n - N,$$

where the inequality holds because $-\mathbf{1} \leq \mathbf{x}$. Thus, the above inequality must in fact hold as an equality, which implies that $x_i = -1$ for all $i \in J_0$. The constraint $K\mathbf{x} + \mathbf{x}^0 = (1 - K)\mathbf{1}$ from $\mathcal{R}_{\text{LP}}^{\text{ob}}$ further implies that $x_i^0 = -1$ for all $i \in \cup_{k=1}^K J_k$ and $x_i^0 = +1$ for all $i \in J_0$.

Since Algorithm 3 constructs I_0 as the index set of the $n_0 = N - Kn$ largest entries of the vector \mathbf{x}^0 , we conclude that it must output $I_0 = J_0$ and the proof completes. \square

REMARK 3 (UNKNOWN CLUSTER CARDINALITIES). The joint outlier detection and cardinality-constrained clustering problem \mathcal{P}^o can be used to solve the cardinality-constrained clustering problem without outliers when the cluster cardinalities n_1, \dots, n_K are not precisely known. To this end, we solve \mathcal{P}^o for different values of n_0 and choose the optimal value n_0^* using the elbow method. The natural clusters I_1, \dots, I_K provide estimates of the relative cluster sizes $\varrho_k = |I_k|/(N - n_0^*)$, $k = 1, \dots, K$, which can be used to construct the cardinality estimates $n_k \approx \varrho_k N$ for problem (1).

5. Numerical Experiments

We now investigate the performance of our algorithms on synthetic as well as real-world clustering problems with and without outliers. All LPs and SDPs are solved with CPLEX 12.7.1 and MOSEK 8.0, respectively, using the YALMIP interface on a 3.40GHz i7 computer with 16GB RAM.

Cardinality-Constrained K-Means Clustering (Real-World Data): We compare the performance of our algorithms from Section 3 with the algorithm of Bennett et al. (2000), see Appendix A, and with the two (cardinality-ignorant) SDP relaxations proposed by Peng and Wei (2007), see Appendix B, on the classification datasets of the UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/>) with 150–300 datapoints, up to 200 continuous attributes and no missing values. In our experiments, we set the cluster cardinalities to the numbers of true class occurrences in each dataset. Table 2 reports the lower bounds provided by $\mathcal{R}_{\text{LP}}/\mathcal{R}_{\text{LP}}^b$ and $\mathcal{R}_{\text{SDP}}/\mathcal{R}_{\text{SDP}}^b$ (LB), the upper bounds from Algorithms 1 and 2 (UB), the objective value of the best of 10 runs of the algorithm of Bennett et al. (UB), randomly initialized by the cluster centers produced by the *K-means++ algorithm* of Arthur and Vassilvitskii (2007), the coefficient of variation across these 10 runs (CV), and the respective lower bounds (LB1, LB2) obtained from the SDP relaxations of Peng and Wei (2007). The obtained lower bounds of \mathcal{R}_{LP} and \mathcal{R}_{SDP} allow

us to certify that the algorithm of Bennett et al. provides nearly optimal solutions in all instances except for *Urban Land Cover*. Also, both Algorithms 1 and 2 are competitive with the algorithm of Bennett et al. while providing rigorous error bounds. Moreover, for all datasets \mathcal{R}_{SDP} yields better lower bounds than the SDP relaxations of Peng and Wei (2007). In fact, it is also possible to prove this rigorously (see Appendix C). The lower bounds obtained from \mathcal{R}_{LP} are competitive with those provided by the stronger relaxation of Peng and Wei (2007), and they are always better than the lower bounds provided by their weaker relaxation. Peng and Wei (2007) also suggest a procedure to compute a feasible clustering (and thus upper bounds) for the unconstrained K -means clustering problem. However, this procedure relies on an enumeration of all possible Voronoi partitions, which is impractical for $K \geq 3$; see Inaba et al. (1994). Furthermore, it is not clear how to impose cardinality constraints in this setting. The average runtimes are 376s (\mathcal{R}_{LP}), 3,906s (\mathcal{R}_{SDP} ; without *Glass Identification*, which was not solved within three hours), 8s (Bennett et al.) as well as 864s and 0.026s (Peng and Wei).

Dataset	$\mathcal{R}_{\text{LP}}/\mathcal{R}_{\text{LP}}^{\text{b}}$		$\mathcal{R}_{\text{SDP}}/\mathcal{R}_{\text{SDP}}^{\text{b}}$		Bennett et al.		Peng and Wei	
	UB	LB	UB	LB	UB	CV (%)	LB1	LB2
Iris	81.4	78.8	81.4	81.4	81.4	0.0	75.6	15.2
Seeds	620.7	539.0	605.6	605.6	605.6	0.0	572.6	19.0
Planning Relax	325.9	297.0	315.7	315.7	315.8	0.3	299.0	273.7
Connectionist Bench	312.6	259.1	280.6	280.1	280.6	0.3	270.0	246.2
Urban Land Cover	3.61e9	3.17e9	3.54e9	3.44e9	3.64e9	9.2	2.05e9	1.94e8
Parkinsons	1.36e6	1.36e6	1.36e6	1.36e6	1.36e6	15.1	1.11e6	6.31e5
Glass Identification	469.0	377.2	—	—	438.2	28.4	321.9	23.8

Table 2 Performance of \mathcal{R}_{LP} , \mathcal{R}_{SDP} , Bennett et al., and Peng and Wei.

Cardinality-Constrained K -Means Clustering (Synthetic Data): We now randomly generate partitions of 10, 20 and 70 datapoints in \mathbb{R}^2 that are drawn from uniform distributions over $K = 3$ unit balls centered at ζ_1, ζ_2 and ζ_3 , respectively, such that $\|\zeta_1 - \zeta_2\| = \|\zeta_1 - \zeta_3\| = \|\zeta_2 - \zeta_3\| = \delta$. Theorem 4 shows that $\mathcal{R}_{\text{LP}}^{\text{b}}$ is tight and that Algorithm 2 can recover the true clusters whenever $n_1 = n_2 = n_3$ and $\delta \geq 4$. Figure 2 demonstrates that in practice, perfect recovery is often achieved by Algorithm 1 even if $\delta \ll 4$ and $n_1 \neq n_2 \neq n_3$. We also note that \mathcal{R}_{SDP} outperforms \mathcal{R}_{LP} when δ is small, and that the algorithm of Bennett et al. frequently fails to determine the optimal solution even if it is run 10 times. In line with the results from the real-world datasets, \mathcal{R}_{SDP} and \mathcal{R}_{LP} are tighter than the stronger SDP relaxation of Peng and Wei (2007). Furthermore, it can be shown that in this setting the weaker relaxation of Peng and Wei (2007) always yields the trivial lower bound of zero. The average runtimes are 7s (\mathcal{R}_{LP}), 106s (\mathcal{R}_{SDP}), 11s (Bennett et al.) and 15.6s (Peng and Wei).

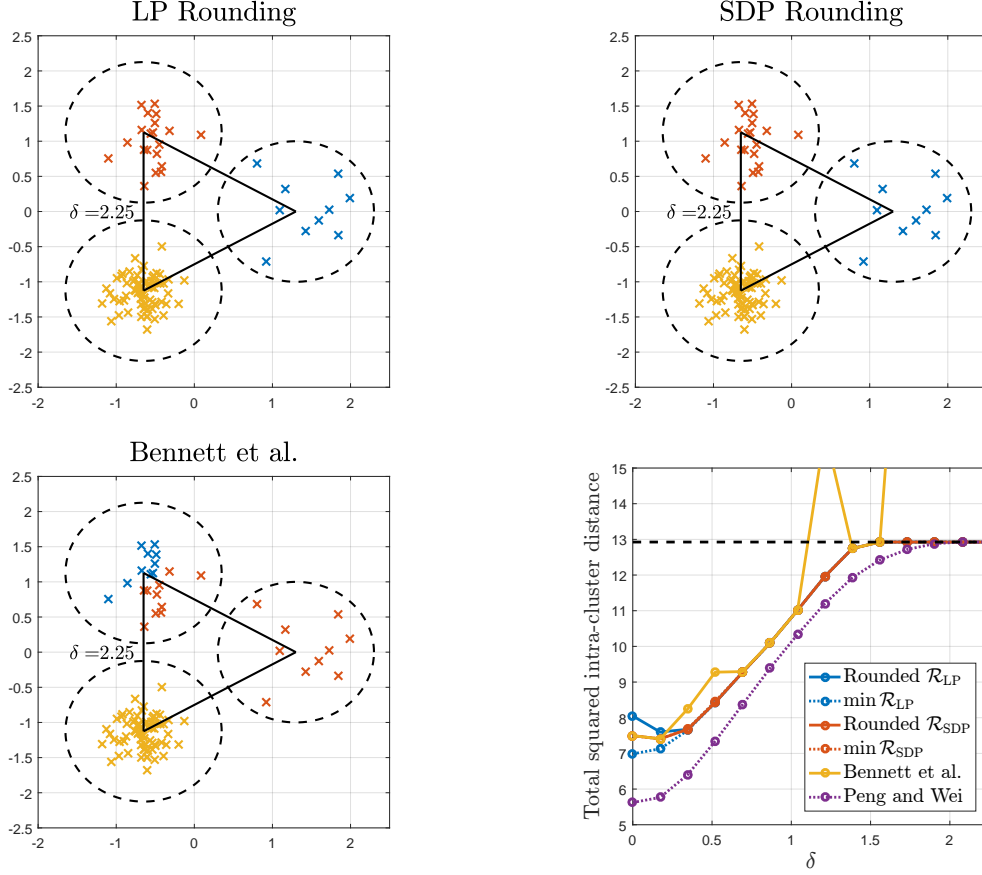


Figure 2 Comparison between different algorithms for (cardinality-constrained) K -means clustering.

Outlier Detection: We use \mathcal{R}_{LP}^o and Algorithm 3 to classify the *Breast cancer Wisconsin (diagnostic)* dataset. The dataset has $d = 30$ numerical features, which we standardize using a Z-score transformation, and it contains 357 benign and 212 malignant cases of breast cancer. We interpret the malignant cases as outliers and thus set $K = 1$. Figure 3 reports the prediction accuracy as well as the false positives (benign cancers classified as malignant) and false negatives (malignant cancers classified as benign) as we increase the number of outliers n_0 from 0 to 400. The figure shows that while setting $n_0 \approx 212$, the true number of malignant cancers, maximizes the prediction accuracy, any choice $n_0 \in [156, 280]$ leads to a competitive prediction accuracy above 80%. Thus, even rough estimates of the number of malignant cancer datapoints can lead to cancer predictors of decent quality. The average runtime is 286s, and the optimality gap is consistently below 3.23% for all values of n_0 .

Acknowledgements: This research was supported by the SNSF grant BSCG10_157733 and the EPSRC grant EP/N020030/1.

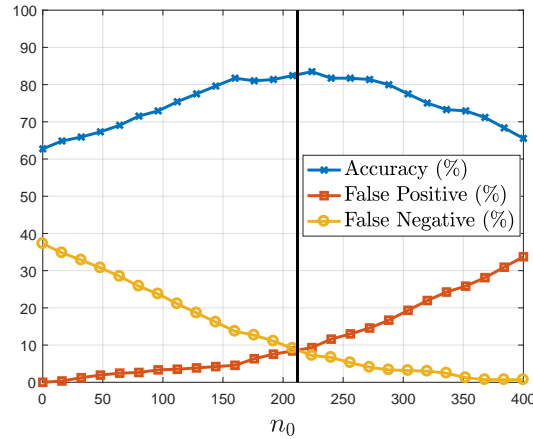


Figure 3 Outlier detection for breast cancer diagnosis.

References

- Aloise, D., A. Deshpande, P. Hansen, P. Popat. 2009. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning* **75**(2) 245–248.
- Ames, B. 2014. Guaranteed clustering and biclustering via semidefinite programming. *Mathematical Programming* **147**(1–2) 429–465.
- Arthur, D., S. Vassilvitskii. 2007. k -means++: The advantages of careful seeding. *ACM-SIAM Symposium on Discrete Algorithms* **18**. 1027–1035.
- Awasthi, P., A. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, R. Ward. 2015. Relax, no need to round: Integrality of clustering formulations. *Conference on Innovations in Theoretical Computer Science* **6**. 191–200.
- Balcan, M., S. Ehrlich, Y. Liang. 2013. Distributed k -means and k -median clustering on general topologies. *Advances in Neural Information Processing Systems* **26**. 1995–2003.
- Banerjee, A., J. Ghosh. 2006. Scalable clustering algorithms with balancing constraints. *Data Mining and Knowledge Discovery* **13**(3) 365–395.
- Bennett, K., P. Bradley, A. Demiriz. 2000. Constrained K-means clustering. Technical Report, Microsoft Research.
- Boyd, S., L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.
- Burkard, R., E. Çela, P. Pardalos, L. Pitsoulis. 1998. The quadratic assignment problem. D. Du, P. Pardalos, eds., *Handbook of Combinatorial Optimization*. Kluwer, 241–337.
- Burkard, R., M. Dell’Amico, S. Martello. 2009. *Assignment Problems*. SIAM.
- Chawla, S., A. Gionis. 2013. k -means--: A unified approach to clustering and outlier detection. *SIAM International Conference on Data Mining* **13**. 189–197.

- Chen, Y., Y. Zhang, X. Ji. 2006. Size regularized cut for data clustering. *Advances in Neural Information Processing Systems 18*. 211–218.
- Elhamifar, E., G. Sapiro, R. Vidal. 2012. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. *Advances in Neural Information Processing Systems 25*. 19–27.
- Gatermann, K., P. Parrilo. 2004. Symmetry groups, semidefinite programs, and sums of squares. *Journal of Pure and Applied Algebra* **192**(1–3) 95–128.
- Golub, G., C. Loan. 1996. *Matrix Computations*. John Hopkins University Press.
- Hasegawa, S., H. Imai, M. Inaba, N. Katoh, J. Nakano. 1993. Efficient algorithms for variance-based k -clustering. *Pacific Conference on Computer Graphics and Applications 1*. 75–89.
- Iguchi, T., D. Mixon, J. Peterson, S. Villar. 2017. Probably certifiably correct k -means clustering. *Mathematical Programming* **165**(2) 605–642.
- Inaba, M., N. Katoh, I. Hiroshi. 1994. Applications of weighted Voronoi diagrams and randomization to variance-based k -clustering. *Symposium on Computational Geometry 10*. 332–339.
- Jain, A. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* **31**(8) 651–666.
- Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**(2) 129–137.
- Mulvey, J., M. Beck. 1984. Solving capacitated clustering problems. *European Journal of Operational Research* **18**(3) 339–348.
- Nellore, A., R. Ward. 2015. Recovery guarantees for exemplar-based clustering. *Information and Computation* **245** 165–180.
- Peng, J., Y. Wei. 2007. Approximating K-means-type clustering via semidefinite programming. *SIAM Journal on Optimization* **18**(1) 186–205.
- Vinayak, R., B. Hassibi. 2016. Similarity clustering in the presence of outliers: Exact recovery via convex program. *IEEE International Symposium on Information Theory*. 91–95.
- Zha, H., X. He, C. Ding, H. Simon, M. Gu. 2002. Spectral relaxation for K-means clustering. *Advances in Neural Information Processing Systems 14*. 1057–1064.

Appendix A: Algorithm of Bennett et al. (2000)

The algorithm of Bennett et al. (2000) is designed for a different variant of problem (1), where only the lower bounds on clusters’ cardinalities are imposed. This algorithm has a natural extension to our cardinality-constrained clustering problem (1) as follows.

Algorithm 4 Algorithm of Bennett et al. for cardinality-constrained clustering

- 1: **Input:** $\mathcal{I}_1 = \{1, \dots, N\}$ (data indices), $n_k \in \mathbb{N}, k = 1, \dots, K$ (cluster sizes).
- 2: Initialization: generate the cluster centers $\zeta_1, \dots, \zeta_K \in \mathbb{R}^d$.
- 3: Assignment: solve the linear assignment problem

$$\Pi^* = \underset{\Pi}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \sum_{k=1}^K \pi_i^k \|\xi_i - \zeta_k\|^2 : \pi_i^k \in \{0, 1\}, \sum_{i=1}^N \pi_i^k = n_k \forall k, \sum_{k=1}^K \pi_i^k = 1 \forall i \right\}.$$

- 4: Set $I_k \leftarrow \{i : (\pi^*)_i^k = 1\}$ for all $k = 1, \dots, K$.
- 5: Set $\zeta_k \leftarrow \frac{1}{n_k} \sum_{i \in I_k} \xi_i$ for all $k = 1, \dots, K$.
- 6: Repeat Steps 3–5 until there are no more changes in ζ_1, \dots, ζ_K .
- 7: **Output:** I_1, \dots, I_K .

Algorithm 4 adapts a classical local search heuristic for the unconstrained K -means clustering problem due to Lloyd (1982) to problem (1). At initialization, it generates random cluster centers ζ_k , $k = 1, \dots, K$. Each subsequent iteration of the algorithm consists of two steps. The first step assigns every datapoint ξ_i to the nearest cluster center while adhering to the prescribed cluster cardinalities, whereas the second step replaces each center ζ_k with the mean of the datapoints that have been assigned to cluster k . The algorithm terminates when the cluster centers ζ_1, \dots, ζ_K no longer change.

Appendix B: Reformulations of Peng and Wei (2007)

Peng and Wei (2007) suggest two different SDP relaxations of the unconstrained K -means clustering problem. Both of them involve a Gram matrix $\mathbf{W} \in \mathbb{S}^N$ with entries $w_{ij} = \xi_i^\top \xi_j$.

The stronger relaxation of Peng and Wei (2007) takes the form

$$\begin{aligned} & \text{minimize} && \langle \mathbf{W}, \mathbb{I} - \mathbf{Z} \rangle \\ & \text{subject to} && \mathbf{Z} \in \mathbb{S}^N \\ & && \mathbf{Z} \succeq \mathbf{0}, \mathbf{Z} \geq \mathbf{0}, \mathbf{Z}\mathbf{1} = \mathbf{1}, \operatorname{Tr}(\mathbf{Z}) = K, \end{aligned} \tag{PW-1}$$

where \mathbb{I} denotes the identity matrix of dimension N . Further relaxing the non-negativity constraints leads to the following weaker relaxation:

$$\begin{aligned} & \text{minimize} && \langle \mathbf{W}, \mathbb{I} - \mathbf{Z} \rangle \\ & \text{subject to} && \mathbf{Z} \in \mathbb{S}^N \\ & && \mathbb{I} \succeq \mathbf{Z} \succeq \mathbf{0}, \mathbf{Z}\mathbf{1} = \mathbf{1}, \operatorname{Tr}(\mathbf{Z}) = K. \end{aligned} \tag{PW-2}$$

Peng and Wei (2007) also demonstrate that (PW-2) essentially reduces to an eigenvalue problem, which means that one can solve (PW-2) in $\mathcal{O}(KN^2)$ time (Golub and Loan 1996).

Appendix C: Relationship between \mathcal{R}_{SDP} , Peng and Wei (2007) and Awasthi et al. (2015)

It is possible to show that \mathcal{R}_{SDP} is at least as tight a relaxation of the cardinality-constrained K -means clustering problem (1) as the stronger relaxation (PW-1) of Peng and Wei (2007). Furthermore, one of the insights gained along the way allows us to prove that the SDP relaxations of the unconstrained K -means clustering problems by Peng and Wei (2007) and Awasthi et al. (2015) are in fact identical. As a consequence, whenever \mathcal{R}_{SDP} and (PW-1) are compared, the obtained insights also apply with respect to the relaxation by Awasthi et al. (2015).

We begin by expressing the objective of (PW-1) in terms of the pairwise distance matrix \mathbf{D} :

$$\begin{aligned}
 \langle \mathbf{W}, \mathbb{I} - \mathbf{Z} \rangle &= \frac{1}{2} \left[2\langle \mathbf{W}, \mathbb{I} \rangle - \langle 2\mathbf{W}, \mathbf{Z} \rangle - \langle \mathbf{D}, \mathbf{Z} \rangle \right] + \frac{1}{2} \langle \mathbf{D}, \mathbf{Z} \rangle \\
 &= \frac{1}{2} \left[2\langle \mathbf{W}, \mathbb{I} \rangle - \langle 2\mathbf{W} + \mathbf{D}, \mathbf{Z} \rangle \right] + \frac{1}{2} \langle \mathbf{D}, \mathbf{Z} \rangle \\
 &\stackrel{(a)}{=} \frac{1}{2} \left[2\langle \mathbf{W}, \mathbb{I} \rangle - \langle \mathbf{1} \text{diag}(\mathbf{W})^\top + \text{diag}(\mathbf{W}) \mathbf{1}^\top, \mathbf{Z} \rangle \right] + \frac{1}{2} \langle \mathbf{D}, \mathbf{Z} \rangle \\
 &= \frac{1}{2} \left[2\langle \mathbf{W}, \mathbb{I} \rangle - \langle \mathbf{1} \text{diag}(\mathbf{W})^\top, \mathbf{Z} \rangle - \langle \text{diag}(\mathbf{W}) \mathbf{1}^\top, \mathbf{Z} \rangle \right] + \frac{1}{2} \langle \mathbf{D}, \mathbf{Z} \rangle \\
 &= \frac{1}{2} \left[2\text{Tr}(\mathbf{W}) - \text{Tr}(\mathbf{Z} \mathbf{1} \text{diag}(\mathbf{W})^\top) - \text{Tr}(\text{diag}(\mathbf{W}) \mathbf{1}^\top \mathbf{Z}) \right] + \frac{1}{2} \langle \mathbf{D}, \mathbf{Z} \rangle \\
 &\stackrel{(b)}{=} \frac{1}{2} \left[2\text{Tr}(\mathbf{W}) - \text{Tr}(\mathbf{1} \text{diag}(\mathbf{W})^\top) - \text{Tr}(\text{diag}(\mathbf{W}) \mathbf{1}^\top) \right] + \frac{1}{2} \langle \mathbf{D}, \mathbf{Z} \rangle \\
 &= \frac{1}{2} \left[2(\mathbf{1}^\top \text{diag}(\mathbf{W})) - \mathbf{1}^\top \text{diag}(\mathbf{W}) - \mathbf{1}^\top \text{diag}(\mathbf{W}) \right] + \frac{1}{2} \langle \mathbf{D}, \mathbf{Z} \rangle \\
 &= \frac{1}{2} \langle \mathbf{D}, \mathbf{Z} \rangle.
 \end{aligned}$$

Here, (a) follows from the observation that the ij -th element of the matrix $2\mathbf{W} + \mathbf{D}$ can be written as $2\xi_i^\top \xi_j + \|\xi_i - \xi_j\|^2 = \|\xi_i\|^2 + \|\xi_j\|^2$, and (b) uses the insights that $\mathbf{Z}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top \mathbf{Z} = \mathbf{1}^\top$. Comparing the SDP relaxation presented in equation (17) of Awasthi et al. (2015) with the SDP relaxation (PW-1) of Peng and Wei (2007), the above identity shows that the two relaxations are identical.

To prove that \mathcal{R}_{SDP} is at least as tight a relaxation as (PW-1), we will make the argument that for every feasible solution $\{(\mathbf{x}^k, \mathbf{M}^k)\}_{k=1}^K$ of \mathcal{R}_{SDP} one can construct a solution

$$\bar{\mathbf{Z}} = \frac{1}{4} \sum_{k=1}^K \frac{1}{n_k} (\mathbf{M}^k + \mathbf{1}\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top + \mathbf{x}^k \mathbf{1}^\top)$$

which is feasible in (PW-1) and achieves the same objective value. We first verify the feasibility of the proposed solution $\bar{\mathbf{Z}}$. Note that $\bar{\mathbf{Z}}$ is symmetric by construction. Next, we can directly verify

that $\bar{\mathbf{Z}}$ is positive semidefinite since

$$\begin{aligned}
\bar{\mathbf{Z}} \succeq 0 &\iff \mathbf{M}^k + \mathbf{1}\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top + \mathbf{x}^k\mathbf{1}^\top \succeq \mathbf{0} && \forall k = 1, \dots, K \\
&\iff \mathbf{v}^\top (\mathbf{M}^k + \mathbf{1}\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top + \mathbf{x}^k\mathbf{1}^\top) \mathbf{v} \geq 0 && \forall \mathbf{v} \in \mathbb{R}^N \quad \forall k = 1, \dots, K \\
&\iff \mathbf{v}^\top (\mathbf{x}^k(\mathbf{x}^k)^\top + \mathbf{1}\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top + \mathbf{x}^k\mathbf{1}^\top) \mathbf{v} \geq 0 && \forall \mathbf{v} \in \mathbb{R}^N \quad \forall k = 1, \dots, K \\
&\iff (\mathbf{v}^\top \mathbf{x}^k)^2 + (\mathbf{v}^\top \mathbf{1})^2 + 2(\mathbf{v}^\top \mathbf{x}^k)(\mathbf{v}^\top \mathbf{1}) \geq 0 && \forall \mathbf{v} \in \mathbb{R}^N \quad \forall k = 1, \dots, K \\
&\iff (\mathbf{v}^\top \mathbf{x}^k + \mathbf{v}^\top \mathbf{1})^2 \geq 0 && \forall \mathbf{v} \in \mathbb{R}^N \quad \forall k = 1, \dots, K,
\end{aligned}$$

where the third implication is due to the definition of $\mathcal{C}_{\text{SDP}}(n_k)$, which requires that $\mathbf{M}^k \succeq \mathbf{x}^k(\mathbf{x}^k)^\top$. The last statement holds trivially because any quadratic form is non-negative. Next, we can ensure the element-wise non-negativity of $\bar{\mathbf{Z}}$, again through the definition of $\mathcal{C}_{\text{SDP}}(n_k)$:

$$\bar{\mathbf{Z}} \succeq \mathbf{0} \iff \mathbf{M}^k + \mathbf{1}\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top + \mathbf{x}^k\mathbf{1}^\top \succeq \mathbf{0} \quad \forall k = 1, \dots, K.$$

Furthermore, combining the definition of $\mathcal{C}_{\text{SDP}}(n_k)$ and the constraint $\sum_{k=1}^K \mathbf{x}^k = (2-K)\mathbf{1}$ of \mathcal{R}_{SDP} , we can see that each row of $\bar{\mathbf{Z}}$ indeed sums up to one:

$$\begin{aligned}
\bar{\mathbf{Z}}\mathbf{1} &= \frac{1}{4} \sum_{k=1}^K \frac{1}{n_k} (\mathbf{M}^k\mathbf{1} + \mathbf{1}\mathbf{1}^\top\mathbf{1} + \mathbf{1}(\mathbf{x}^k)^\top\mathbf{1} + \mathbf{x}^k\mathbf{1}^\top\mathbf{1}) \\
&= \frac{1}{4} \sum_{k=1}^K \frac{1}{n_k} ((2n_k - N)\mathbf{x}^k + N\mathbf{1} + (2n_k - N)\mathbf{1} + N\mathbf{x}^k) \\
&= \frac{1}{2} \sum_{k=1}^K (\mathbf{x}^k + \mathbf{1}) = \mathbf{1}.
\end{aligned}$$

Finally, the trace of $\bar{\mathbf{Z}}$ is uniquely determined as follows:

$$\begin{aligned}
\text{Tr}(\bar{\mathbf{Z}}) &= \frac{1}{4} \sum_{k=1}^K \frac{1}{n_k} \text{Tr}(\mathbf{M}^k + \mathbf{1}\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top + \mathbf{x}^k\mathbf{1}^\top) \\
&= \frac{1}{4} \sum_{k=1}^K \frac{1}{n_k} (2N + 2(\mathbf{1}^\top \mathbf{x}^k)) \\
&= \frac{1}{4} \sum_{k=1}^K \frac{1}{n_k} (2N + 2(2n_k - N)) = K.
\end{aligned}$$

Thus, $\bar{\mathbf{Z}}$ is feasible in (PW-1), and it remains to prove that it achieves the same objective value as the original solution $\{(\mathbf{x}^k, \mathbf{M}^k)\}_{k=1}^K$ in \mathcal{R}_{SDP} . Invoking the relation derived at the beginning of this section, it is easy to see that

$$\langle \mathbf{W}, \mathbb{I} - \bar{\mathbf{Z}} \rangle = \frac{1}{2} \langle \mathbf{D}, \bar{\mathbf{Z}} \rangle = \frac{1}{8} \left\langle \mathbf{D}, \sum_{k=1}^K \frac{1}{n_k} (\mathbf{M}^k + \mathbf{1}\mathbf{1}^\top + \mathbf{1}(\mathbf{x}^k)^\top + \mathbf{x}^k\mathbf{1}^\top) \right\rangle.$$

The proof thus concludes. \square