SIAM J. OPTIM.
Vol. 29, No. 2, pp. 1211–1239

SIZE MATTERS: CARDINALITY-CONSTRAINED CLUSTERING AND OUTLIER DETECTION VIA CONIC OPTIMIZATION[*]

NAPAT RUJEERAPAIBOON[†], KILIAN SCHINDLER[‡], DANIEL KUHN[‡], AND WOLFRAM WIESEMANN[§]

Abstract. Plain vanilla K-means clustering has proven to be successful in practice, yet it suffers from outlier sensitivity and may produce highly unbalanced clusters. To mitigate both shortcomings, we formulate a joint outlier detection and clustering problem, which assigns a prescribed number of data points to an auxiliary outlier cluster and performs cardinality-constrained K-means clustering on the residual data set, treating the cluster cardinalities as a given input. 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

AMS subject classifications. 90C22, 90C05, 62H30

DOI. 10.1137/17M1150670

1. Introduction.

Clustering aims to partition a set of data points into a set of clusters so that data points in the same cluster are more similar to each other than to those in other clusters. Among the myriad of clustering approaches from the literature, K-means clustering stands out for its 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 the task of partitioning N data points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N \in \mathbb{R}^d$ into K clusters $I_1, \ldots, I_K$ of prescribed sizes $n_1, \ldots, n_K$, with $n_1 + \cdots + n_K = N$, so as to minimize the sum of squared intra-cluster distances. We can formalize the cardinality-constrained K-means clustering problem as follows:

$$(1) \qquad \begin{array}{ll} \text{minimize} & \sum_{k=1}^K \sum_{i \in I_k} \|\boldsymbol{\xi}_i - \frac{1}{n_k}(\sum_{j \in I_k} \boldsymbol{\xi}_j)\|^2 \\ \text{subject to} & (I_1, \ldots, I_K) \in \mathfrak{P}(n_1, \ldots, n_K), \end{array}$$

[*] Received by the editors October 5, 2017; accepted for publication (in revised form) January 22, 2019; published electronically April 30, 2019.

http://www.siam.org/journals/siopt/29-2/M115067.html

Funding: This research was supported by the Singapore Ministry of Education Academic Research Fund Tier 1 R-266-000-131-133, the SNSF grant BSCGI0_157733, and the EPSRC grant EP/N020030/1.

[†] Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 117576, Singapore (isenapa@nus.edu.sg).

[‡] Risk Analytics and Optimization Chair, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland (kilian.schindler@epfl.ch, daniel.kuhn@epfl.ch).

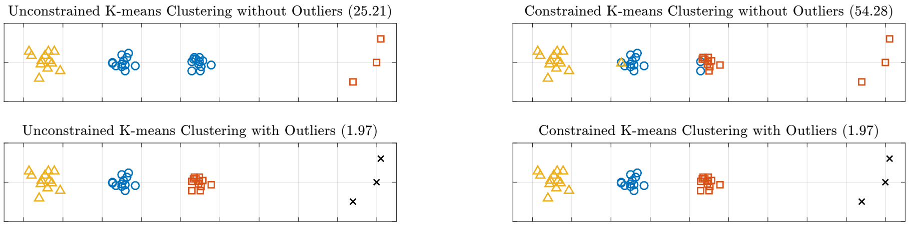
[§] Imperial College Business School, London SW7 2AZ, United Kingdom (ww@imperial.ac.uk).

where

$$\mathfrak{P}(n_1,\ldots,n_K) = \left\{ (I_1,\ldots,I_K) \,:\, \begin{array}{c} |I_k| = n_k \ \ \forall k, \\ \bigcup_{k=1}^{K} I_k = \{1,\ldots,N\}, \\ I_k \cap I_\ell = \emptyset \ \ \forall k \neq \ell \end{array} \right\}$$

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

Our motivation for studying problem (1) is threefold. Firstly, it has been shown by Bennett, Bradley, and Demiriz (2000) and Chen, Zhang, and Ji (2006) that the algorithms commonly employed for the *un*constrained K-means clustering problem frequently produce suboptimal solutions, where some of the clusters contain very few or even no data points. In this context, cardinality constraints can act as a regularizer that avoids local minima of poor quality. Secondly, many application domains require the clusters $I_1,\ldots,I_K$ to be of comparable size. This is the case in, among others, 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, Ehrlich, and Liang (2013). Finally, and perhaps most importantly, K-means clustering is highly sensitive to outliers. To illustrate this, consider the data set 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 data points, shows a similar behavior on this data set (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. A comprehensive and principled treatment of outlier detection methods can be found in the book by Aggarwal (2013).



Unconstrained K-means Clustering without Outliers (25.21)

Constrained K-means Clustering without Outliers (54.28)

Unconstrained K-means Clustering with Outliers (1.97)

Constrained K-means Clustering with Outliers (1.97)

FIG. 1. *Sensitivity of the (un)constrained K-means clustering problem to outliers. Indicated in parentheses next to the panel titles are the respective sums of squared intra-cluster distances achieved.*

To the best of our knowledge, only two solution approaches have been proposed for problem (1) to date. Bennett, Bradley, and Demiriz (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 $n_1, \ldots, n_K$. Banerjee and Ghosh (2006) solve the balanced version of problem (1), where $n_1 = \cdots = n_K$, by sampling a subset of the data points, performing a clustering on this subset, and subsequently populating the resulting clusters with the remaining data points while adhering to the cardinality constraints. Balanced clustering is also considered by Malinen and Fränti (2014) and Costa, Aloise, and Mladenović (2017). Malinen and Fränti (2014) proceed similarly to Bennett, Bradley, and Demiriz (2000) but take explicit advantage of the Hungarian algorithm to speed up the cluster assignment step within the local search heuristic. Costa, Aloise, and Mladenović (2017) propose a variable neighborhood search heuristic that starts from a random partition of the data points into balanced clusters and subsequently searches for better solutions in the neighborhood obtained by an increasing number of data point swaps between two clusters. Although all of these heuristics tend to quickly produce solutions of high quality, they are not known to be polynomial-time algorithms, 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 data points is reduced to $K-1$ (provided 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, Katoh, and Hiroshi (1994), or by approximation methods such as those in Hasegawa et al. (1993). Using either approach, the runtime grows polynomially in the number of data points 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, Sapiro, and Vidal (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., data points 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 data set is characterized by a similarity matrix where the expected pairwise similarities of points in the same cluster are greater 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 so 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 data sets containing outliers.

Table 1
Comparison of recovery guarantees for K-means clustering relaxations.

	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

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 both recover the optimal value of (1) whenever a cluster separation condition is met. The latter requires all cluster diameters to be smaller than the distance between any two distinct clusters and, in the presence of outliers, smaller than the distance between any outlier and any other point. The same condition (in the absence of outliers) was used in Elhamifar, Sapiro, and Vidal (2012) and Awasthi et al. (2015). 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 $n_1, \ldots, n_K$ 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 NK binary variables, as opposed to the standard MILP reformulation that

contains N^2 binary variables, and whose LP relaxation is at least as tight as the LP relaxation of the standard reformulation.

2. We develop lower bounds that 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 the best of our knowledge, we are proposing the first tractable solution scheme for problem (1) with optimality guarantees.
4. We illustrate 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 presents numerical experiments, and section 6 gives concluding remarks. Finally, a detailed description of the heuristic proposed by Bennett, Bradley, and Demiriz (2000) for cardinality-constrained K-means clustering is provided in Appendix A.

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 nonnegative. The notation $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{AB})$ represents the trace inner product of $\mathbf{A}$ and $\mathbf{B}$. Furthermore, we use $\mathrm{diag}(\mathbf{A})$ to denote a vector in $\mathbb{R}^N$ whose entries coincide with those of $\mathbf{A}$'s main diagonal. Finally, for a set of N data points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N$, we use $\mathbf{D} \in \mathbb{S}^N$ to denote the matrix of squared pairwise distances $d_{ij} = \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2$.

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 $\mathbb{R}^d$.

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

LEMMA 1. *For any vectors $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n \in \mathbb{R}^d$, we have*

$$\sum_{i=1}^{n} \left\| \boldsymbol{\xi}_i - \frac{1}{n}\left(\sum_{j=1}^{n} \boldsymbol{\xi}_j \right) \right\|^2 = \frac{1}{2n} \sum_{i,j=1}^{n} \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2.$$

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

Using Lemma 1, Costa, Aloise, and Mladenović (2017) noticed that the K-means objective can be stated as a sum of quadratic terms. In the following proposition, we elaborate on this insight and prove that problem (1) is a specific instance of a quadratic assignment problem.

PROPOSITION 1 (quadratic assignment reformulation). *Clustering problem* (1) *can be cast as the quadratic assignment problem*

$$\underset{\sigma\in\mathfrak{S}^N}{\text{minimize}} \quad \tfrac{1}{2}\langle \mathbf{Q}, \mathbf{P}_\sigma \mathbf{D}\mathbf{P}_\sigma^\top\rangle, \tag{2}$$

where $\mathbf{Q}\in\mathbb{S}^N$ *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* $\mathbf{P}_\sigma\in\mathbb{R}^{N\times N}$ *is defined through* $(\mathbf{P}_\sigma)_{ij}=1$ *if* $\sigma(i)=j$; $(\mathbf{P}_\sigma)_{ij}=0$ *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 $k=1,\dots,K$, and denote its inverse by σ^{-1}. This permutation is feasible in (2), and it achieves the same objective value as $(I_1,\dots,I_K)$ in (1) because

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i\in I_k}\left\|\boldsymbol{\xi}_i-\frac{1}{n_k}\Bigg(\sum_{j\in I_k}\boldsymbol{\xi}_j\Bigg)\right\|^2 &= \frac{1}{2}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,j\in I_k}d_{ij}\\ &= \frac{1}{2}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,j\in\sigma^{-1}(I_k)}d_{\sigma(i)\sigma(j)}\\ &= \frac{1}{2}\langle\mathbf{Q},\mathbf{P}_\sigma\mathbf{D}\mathbf{P}_\sigma^\top\rangle,\end{aligned}$$

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 $\mathbf{Q}$.

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 to that above shows that the partition achieves the same objective value as σ in (2). □

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 (2013, Page 2741). The LP relaxations of these MILPs are, however, known to be weak and give a trivial lower bound of zero; see, e.g., Zhang, Royo, and Ma (2013, Theorem 4.1). In Proposition 2 we show that the intra-cluster permutation symmetry of the data points enables us to give an alternative MILP reformulation containing only $NK \ll \Omega(N^2)$ binary variables. We also mention that the related, yet different, cardinality-constrained exemplar-based clustering problem can be formulated as an MILP containing $\Omega(N^2)$ binary variables; see Mulvey and Beck (1984).

PROPOSITION 2 (MILP reformulation). *The clustering problem* (1) *is equivalent to the MILP*

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

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

Proof of Proposition 2. 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

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

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 for all $k = 1, \dots, K$ is feasible in $(\mathcal{P}')$ and attains the same objective value as $(I_1, \dots, I_K)$ in (1) since

$$\begin{aligned} \sum_{k=1}^K \sum_{i \in I_k} \left\| \boldsymbol{\xi}_i - \frac{1}{n_k}\Bigg(\sum_{j \in I_k} \boldsymbol{\xi}_j\Bigg)\right\|^2 &= \frac{1}{2}\sum_{k=1}^K \frac{1}{n_k}\sum_{i,j \in I_k} d_{ij} \\ &= \frac{1}{2}\sum_{k=1}^K \frac{1}{n_k}\sum_{i,j=1}^N d_{ij}\max\{\pi_i^k + \pi_j^k - 1, 0\}. \end{aligned}$$

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 (1) and that attains the same objective value. ∎

Remark 1. Note that zero is a (trivial) lower bound on the objective value of the LP relaxation of the MILP $(\mathcal{P})$. As a consequence, this LP relaxation is at least as tight as the LP relaxation of the Kaufmann and Broeckx exact MILP formulation of problem (2), which always yields a lower bound of zero. It is also possible to construct instances where the LP relaxation of the MILP $(\mathcal{P})$ is strictly tighter.

K-means clustering with cardinality constraints is known to be NP-hard as it is a special case of cardinality-constrained p-norm clustering, which was shown to be NP-hard (for any $p > 1$) by Bertoni et al. (2012). The restriction to the Euclidean norm (i.e., $p = 2$), however, allows for a more concise proof, which is given in the following proposition.

PROPOSITION 3. *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 Proposition 2, 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 $(n_1, n_2) \in \{(1, N-1), (2, N-2), \ldots, (\lfloor N/2 \rfloor, \lceil N/2 \rceil)\}$ and selecting the clustering with the lowest objective value. Thus, in this case, if problem $(\mathcal{P})$ were 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. □

In the context of balanced clustering, similar hardness results have been established by Pyatkin, Aloise, and Mladenović (2017). Specifically, they prove that the balanced K-means clustering problem is NP-complete for $K \geq 2$ and $\frac{N}{K} \geq 3$ (i.e., the shared cardinality of all clusters is greater than or equal to three). In contrast, if $K \geq 2$ and $\frac{N}{K} = 2$ (i.e., each cluster should contain two points), balanced K-means clustering reduces to a minimum-weight perfect matching problem that can be solved in polynomial-time by different algorithms; see Cook and Rohe (1999, Table I) for a review.

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 1 (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 $(I_1, \ldots, I_K)$ of (1) and every $k, \ell \in \{1, \ldots, K\}$, $k < \ell$, 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. Given a cluster I_m for any $m \in \{1, \ldots, K\}$, define its cluster center as $\boldsymbol{\zeta}_m = \frac{1}{n_m} \sum_{i \in I_m} \boldsymbol{\xi}_i$, and let $\boldsymbol{h} = \boldsymbol{\zeta}_k - \boldsymbol{\zeta}_\ell$ be the vector that connects the cluster centers of I_k and I_ℓ. The statement holds if $\boldsymbol{h}^\top (\boldsymbol{\xi}_{i_k} - \boldsymbol{\xi}_{i_\ell}) \geq 0$ for all $i_k \in I_k$ and $i_\ell \in I_\ell$ as $\boldsymbol{h}$ itself determines a separating hyperplane for I_k and I_ℓ in that case. We thus assume that $\boldsymbol{h}^\top (\boldsymbol{\xi}_{i_k} - \boldsymbol{\xi}_{i_\ell}) < 0$ for some $i_k \in I_k$ and $i_\ell \in I_\ell$. However, this contradicts the optimality of the clustering $(I_1, \ldots, I_K)$ because

$$\begin{aligned} \boldsymbol{h}^\top (\boldsymbol{\xi}_{i_k} - \boldsymbol{\xi}_{i_\ell}) < 0 &\iff (\boldsymbol{\zeta}_k - \boldsymbol{\zeta}_\ell)^\top (\boldsymbol{\xi}_{i_k} - \boldsymbol{\xi}_{i_\ell}) < 0 \\ &\iff \boldsymbol{\xi}_{i_k}^\top \boldsymbol{\zeta}_k + \boldsymbol{\xi}_{i_\ell}^\top \boldsymbol{\zeta}_\ell < \boldsymbol{\xi}_{i_k}^\top \boldsymbol{\zeta}_\ell + \boldsymbol{\xi}_{i_\ell}^\top \boldsymbol{\zeta}_k \\ &\iff \|\boldsymbol{\xi}_{i_\ell} - \boldsymbol{\zeta}_k\|^2 + \|\boldsymbol{\xi}_{i_k} - \boldsymbol{\zeta}_\ell\|^2 < \|\boldsymbol{\xi}_{i_k} - \boldsymbol{\zeta}_k\|^2 + \|\boldsymbol{\xi}_{i_\ell} - \boldsymbol{\zeta}_\ell\|^2, \end{aligned}$$

where the last equivalence follows from multiplying both sides of the second inequality with 2 and then completing the squares by adding $\boldsymbol{\xi}_{i_k}^\top \boldsymbol{\xi}_{i_k} + \boldsymbol{\zeta}_k^\top \boldsymbol{\zeta}_k + \boldsymbol{\xi}_{i_\ell}^\top \boldsymbol{\xi}_{i_\ell} + \boldsymbol{\zeta}_\ell^\top \boldsymbol{\zeta}_\ell$ 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

$$\begin{aligned} &\sum_{i \in \tilde{I}_k} \|\boldsymbol{\xi}_i - \boldsymbol{\zeta}_k\|^2 + \sum_{i \in \tilde{I}_\ell} \|\boldsymbol{\xi}_i - \boldsymbol{\zeta}_\ell\|^2 + \sum_{\substack{m=1,\ldots,K \\ m \notin \{k,\ell\}}} \sum_{i \in I_m} \|\boldsymbol{\xi}_i - \boldsymbol{\zeta}_m\|^2 \\ &\quad < \sum_{i \in I_k} \|\boldsymbol{\xi}_i - \boldsymbol{\zeta}_k\|^2 + \sum_{i \in I_\ell} \|\boldsymbol{\xi}_i - \boldsymbol{\zeta}_\ell\|^2 + \sum_{\substack{m=1,\ldots,K \\ m \notin \{k,\ell\}}} \sum_{i \in I_m} \|\boldsymbol{\xi}_i - \boldsymbol{\zeta}_m\|^2. \end{aligned}$$

The left-hand side of the above inequality represents an upper bound on the sum of squared intra-cluster distances attained by the clustering $(I_1, \ldots, \tilde{I}_k, \ldots, \tilde{I}_\ell, \ldots, I_K)$

since $\boldsymbol{\zeta}_k$ and $\boldsymbol{\zeta}_\ell$ may not coincide with the minimizers $\frac{1}{n_k}\sum_{i\in\tilde{I}_k}\boldsymbol{\xi}_i$ and $\frac{1}{n_\ell}\sum_{i\in\tilde{I}_\ell}\boldsymbol{\xi}_i$, respectively. Recall that the cluster centers are chosen so as to minimize the sum of the squared distances from the cluster center to each point in the cluster. We thus conclude that the clustering $(I_1,\ldots,\tilde{I}_k,\ldots,\tilde{I}_\ell,\ldots,I_K)$ attains a strictly lower objective value than $(I_1,\ldots,I_K)$ 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

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

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

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

Here, x_i^k takes the value $+1$ if the ith data point 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 2 we will show that the reformulation (3) of the MILP $(\mathcal{P})$ admits the SDP relaxation

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

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

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

Note that $\mathcal{C}_{\text{SDP}}(n)$ is semidefinite representable because Schur's complement allows us to express the constraint $\mathbf{M}\succeq\boldsymbol{x}\boldsymbol{x}^\top$ as a linear matrix inequality; see, e.g., Boyd and Vandenberghe (2004). Furthermore, we point out that the last four constraints in $\mathcal{C}_{\mathrm{SDP}}(n)$ are also used in the *reformulation-linearization technique* for nonconvex programs, as described by Anstreicher (2009).

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

THEOREM 2 (SDP and LP relaxations). *We have*

$$\min(\mathcal{R}_{\mathrm{LP}}) \leq \min(\mathcal{R}_{\mathrm{SDP}}) \leq \min(\mathcal{P}).$$

Proof. The inequality

$$\min(\mathcal{R}_{\mathrm{LP}}) \leq \min(\mathcal{R}_{\mathrm{SDP}})$$

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

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

and

$$\begin{aligned}
\mathbf{M}^k + \mathbf{1}\mathbf{1}^\top + \boldsymbol{x}^k\mathbf{1}^\top + \mathbf{1}(\boldsymbol{x}^k)^\top &= +(\mathbf{1} + \boldsymbol{x}^k)(\mathbf{1} + \boldsymbol{x}^k)^\top \geq \mathbf{0},\\
\mathbf{M}^k + \mathbf{1}\mathbf{1}^\top - \boldsymbol{x}^k\mathbf{1}^\top - \mathbf{1}(\boldsymbol{x}^k)^\top &= +(\mathbf{1} - \boldsymbol{x}^k)(\mathbf{1} - \boldsymbol{x}^k)^\top \geq \mathbf{0},\\
\mathbf{M}^k - \mathbf{1}\mathbf{1}^\top + \boldsymbol{x}^k\mathbf{1}^\top - \mathbf{1}(\boldsymbol{x}^k)^\top &= -(\mathbf{1} - \boldsymbol{x}^k)(\mathbf{1} + \boldsymbol{x}^k)^\top \leq \mathbf{0},\\
\mathbf{M}^k - \mathbf{1}\mathbf{1}^\top - \boldsymbol{x}^k\mathbf{1}^\top + \mathbf{1}(\boldsymbol{x}^k)^\top &= -(\mathbf{1} + \boldsymbol{x}^k)(\mathbf{1} - \boldsymbol{x}^k)^\top \leq \mathbf{0},
\end{aligned}$$

which ensures that $(\boldsymbol{x}^k, \mathbf{M}^k) \in \mathcal{C}_{\mathrm{SDP}}(n_k)$ for every k. Finally, the remaining constraint

$$\sum_{k=1}^K \boldsymbol{x}^k = (2 - K)\mathbf{1}$$

in $(\mathcal{R}_{\mathrm{SDP}})$ coincides with the last constraint in (3). Thus, $\{(\boldsymbol{x}^k, \mathbf{M}^k)\}_{k=1}^K$ is feasible in $(\mathcal{R}_{\mathrm{SDP}})$. The desired inequality now follows because any feasible point in (3) corresponds to a feasible point in $(\mathcal{R}_{\mathrm{SDP}})$ with the same objective value. Note that the converse implication is generally false. ∎

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

It is possible to show that $(\mathcal{R}_{\mathrm{LP}})$ is at least as tight as the naïve LP relaxation of the MILP $(\mathcal{P})$, which we denote by $(\mathcal{L})$, where the integrality constraints are simply ignored. One can also construct instances where $(\mathcal{R}_{\mathrm{LP}})$ is strictly tighter than $(\mathcal{L})$. We also emphasize that both LP relaxations entail $\mathcal{O}(N^2K)$ variables and $\mathcal{O}(N^2K)$ constraints.

PROPOSITION 4. *We have* $\min(\mathcal{R}_{\mathrm{LP}}) \geq \min(\mathcal{L})$.

Proof. Consider a feasible solution $\{(\boldsymbol{x}^k, \mathbf{M}^k)\}_{k=1}^K$ of $(\mathcal{R}_{\mathrm{LP}})$. Its feasibility implies that

(a) $\sum_{k=1}^K x_i^k = 2 - K\ \forall i$, (b) $\sum_{i=1}^N x_i^k = 2n_k - N\ \forall k$, (c) $m_{ij}^k - x_i^k - x_j^k + 1 \geq 0\ \forall i, j, k$.

Next, set $\pi_i^k = (x_i^k + 1)/2$ and $\eta_{ij}^k = \frac{1}{4}(m_{ij}^k + x_i^k + x_j^k + 1)$ for all i, j, k. Then,

$$\text{(a}'\text{)}\ \textstyle\sum_{k=1}^K \pi_i^k = 1\ \forall i,\ \text{(b}'\text{)}\ \sum_{i=1}^N \pi_i^k = n_k\ \forall k,\ \text{(c}'\text{)}\ \eta_{ij}^k \geq \pi_i^k + \pi_j^k - 1\ \forall i, j, k.$$

Hence, this solution is feasible in $(\mathcal{L})$. A direct calculation also reveals that both solutions attain the same objective value in their respective optimization problems. This confirms that $(\mathcal{R}_{\text{LP}})$ is a relaxation that is at least as tight as $(\mathcal{L})$. □

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 $(\mathcal{R}_{\text{SDP}})$ or $(\mathcal{R}_{\text{LP}})$; 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 $(\mathcal{R}_{\text{SDP}})$ or $(\mathcal{R}_{\text{LP}})$ for the data points $\boldsymbol{\xi}_i$, $i \in \mathcal{I}_1$, and record the optimal
   $\boldsymbol{x}^1, \dots, \boldsymbol{x}^K \in \mathbb{R}^N$.
3: Solve the linear assignment problem
   $$\boldsymbol{\Pi}' \in \operatorname*{argmax}_{\boldsymbol{\Pi}} \left\{ \sum_{i=1}^N \sum_{k=1}^K \pi_i^k x_i^k : \begin{array}{l} \pi_i^k \in \{0,1\}, \\ \sum_{i=1}^N \pi_i^k = n_k \ \ \forall k, \\ \sum_{k=1}^K \pi_i^k = 1 \ \ \forall i \end{array} \right\}.$$
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
   $$\boldsymbol{\Pi}^\star \in \operatorname*{argmin}_{\boldsymbol{\Pi}} \left\{ \sum_{i=1}^N \sum_{k=1}^K \pi_i^k \|\boldsymbol{\xi}_i - \boldsymbol{\zeta}_k\|^2 : \begin{array}{l} \pi_i^k \in \{0,1\}, \\ \sum_{i=1}^N \pi_i^k = n_k \ \ \forall k, \\ \sum_{k=1}^K \pi_i^k = 1 \ \ \forall i \end{array} \right\}.$$
7: Set $I_k \leftarrow \{i : (\pi^\star)_i^k = 1\}$ for all $k = 1, \dots, K$.
8: Output: $I_1, \dots, I_K$.
```

Recall that the continuous variables $\boldsymbol{x}^k = (x_1^k, \dots, x_N^k)^\top$ in $(\mathcal{R}_{\text{SDP}})$ and $(\mathcal{R}_{\text{LP}})$ 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 $\boldsymbol{\Pi} \in \{0,1\}^{N \times K}$ 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, Dell'Amico, and Martello (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, Bradley, and Demiriz (2000). Specifically, step 5 calculates the cluster centers $\boldsymbol{\zeta}_k$, 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 $(\mathcal{R}_{\text{SDP}})$ or $(\mathcal{R}_{\text{LP}})$. 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 1.

3.2. Tighter relaxations for balanced clustering. The computational burden of solving ($\mathcal{R}_{\text{SDP}}$) and ($\mathcal{R}_{\text{LP}}$) 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 ($\mathcal{R}_{\text{SDP}}$) can be replaced by

$$(\mathcal{R}^{\text{b}}_{\text{SDP}}) \qquad \begin{array}{ll} \text{minimize} & \frac{1}{8n}\langle \mathbf{D}, \mathbf{M}^1 + \mathbf{1}\mathbf{1}^\top + \boldsymbol{x}^1\mathbf{1}^\top + \mathbf{1}(\boldsymbol{x}^1)^\top \\ & \qquad\qquad + (K-1)(\mathbf{M} + \mathbf{1}\mathbf{1}^\top + \boldsymbol{x}\mathbf{1}^\top + \mathbf{1}\boldsymbol{x}^\top)\rangle \\ \text{subject to} & (\boldsymbol{x}^1, \mathbf{M}^1), (\boldsymbol{x}, \mathbf{M}) \in \mathcal{C}_{\text{SDP}}(n), \\ & \boldsymbol{x}^1 + (K-1)\boldsymbol{x} = (2-K)\mathbf{1}, \\ & x^1_1 = 1, \end{array}$$

whose size no longer scales with K. Similarly, ($\mathcal{R}_{\text{LP}}$) simplifies to the LP ($\mathcal{R}^{\text{b}}_{\text{LP}}$) obtained from ($\mathcal{R}^{\text{b}}_{\text{SDP}}$) 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*

$$\min(\mathcal{R}^{\text{b}}_{\text{LP}}) \le \min(\mathcal{R}^{\text{b}}_{\text{SDP}}) \le \min(\mathcal{P}).$$

Proof. The inequality $\min(\mathcal{R}^{\text{b}}_{\text{LP}}) \le \min(\mathcal{R}^{\text{b}}_{\text{SDP}})$ is trivially satisfied. To prove the inequality $\min(\mathcal{R}^{\text{b}}_{\text{SDP}}) \le \min(\mathcal{P})$, 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 data point $\boldsymbol{\xi}_1$ 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 2, the MILP ($\mathcal{P}$) can then be relaxed to a variant of the SDP ($\mathcal{R}_{\text{SDP}}$) 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 $k = 2, \ldots, K$ 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 $\{(\boldsymbol{x}^k, \mathbf{M}^k)\}_{k=1}^K$ of this SDP, and define

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

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

$$\begin{array}{lll} \boldsymbol{x}^1_{\text{s}} = \boldsymbol{x}^1, & \boldsymbol{x}^k_{\text{s}} = \boldsymbol{x} & \forall k = 2, \ldots, K, \\ \mathbf{M}^1_{\text{s}} = \mathbf{M}^1, & \mathbf{M}^k_{\text{s}} = \mathbf{M} & \forall k = 2, \ldots, K. \end{array}$$

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

If $n_k = n$ for all k, then the SDP and LP relaxations from section 3.1 admit an optimal solution where both $\boldsymbol{x}^k$ and $\mathbf{M}^k$ are independent of k, in which case Algorithm 1 performs poorly. This motivates the improved relaxations ($\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}}$) and ($\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}}$) involving the symmetry-breaking constraint $x^1_1 = 1$, which ensures that—without loss of generality—the cluster harboring the first data point $\boldsymbol{\xi}_1$ is indexed by $k = 1$. As the symmetry between clusters $2, \ldots, K$ persists and because any additional symmetry-breaking constraint would be restrictive, the optimal solutions of ($\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}}$) and ($\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}}$) only facilitate a reliable recovery of cluster 1. To recover *all* clusters, however, we can solve ($\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}}$) or ($\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}}$) $K-1$ times over the as yet unassigned data points; 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).

In contrast, the naïve relaxation ($\mathcal{L}$) of ($\mathcal{P}$) becomes significantly weaker when all cardinalities are equal. To see this, we note that a solution $\pi^k_i = 1/K$ and $\eta^k_{ij} = 0$ for all $i, j = 1, \ldots, N$ and for all $k = 1, \ldots, K$ is feasible in ($\mathcal{L}$) (i.e., it satisfies all constraints in problem ($\mathcal{P}$) except for the integrality constraints that are imposed on π^k_i) whenever $K \geq 2$. Hence, the optimal objective value of ($\mathcal{L}$) is zero. This could be avoided by adding a symmetry-breaking constraint $\pi^1_1 = 1$ to problem ($\mathcal{L}$) to ensure that the cluster containing the first data point $\boldsymbol{\xi}_1$ is indexed by $k = 1$. However, the improvement appears to be marginal.

Algorithm 2 Rounding algorithm for balanced clustering.

```
Input: $\mathcal{I}_1 = \{1, \ldots, N\}$ (data indices), $n \in \mathbb{N}$ (cluster size), $K = N/n \in \mathbb{N}$ (number of clusters).
for $k = 1, \ldots, K-1$ do
    Solve ($\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}}$) or ($\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}}$) for the data points $\boldsymbol{\xi}_i$, $i \in \mathcal{I}_k$, and record the optimal $\boldsymbol{x}^1 \in \mathbb{R}^{|\mathcal{I}_k|}$.
    Determine a bijection $\rho : \{1, \ldots, |\mathcal{I}_k|\} \to \mathcal{I}_k$ such that $x^1_{\rho(1)} \geq x^1_{\rho(2)} \geq \cdots \geq x^1_{\rho(|\mathcal{I}_k|)}$.
    Set $I_k \leftarrow \{\rho(1), \ldots, \rho(n)\}$ and $\mathcal{I}_{k+1} \leftarrow \mathcal{I}_k \setminus I_k$.
Set $I_K \leftarrow \mathcal{I}_K$.
Output: $I_1, \ldots, I_K$.
```

3.3. Comparison to existing SDP relaxations. We now compare ($\mathcal{R}_{\mathrm{SDP}}$) and ($\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}}$) with existing SDP relaxations from the literature. First, we report the various SDP relaxations proposed by Peng and Wei (2007) and Awasthi et al. (2015). Then, we establish that two of them are equivalent. Finally, we show that ($\mathcal{R}_{\mathrm{SDP}}$) and ($\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}}$) are relaxations that are at least as tight as their corresponding counterparts from the literature. The numerical experiments in section 5 provide evidence that this relation can also be strict.

Peng and Wei (2007) suggest two different SDP relaxations for the *un*constrained K-means clustering problem and an SDP relaxation for the balanced K-means clustering problem. All of them involve a Gram matrix $\mathbf{W} \in \mathbb{S}^N$ with entries $w_{ij} = \boldsymbol{\xi}_i^\top \boldsymbol{\xi}_j$. Their stronger relaxation for the unconstrained K-means clustering problem takes the form

$$(\mathcal{PW}_1) \qquad \begin{array}{ll} \text{minimize} & \langle \mathbf{W}, \mathbb{I} - \mathbf{Z} \rangle \\ \text{subject to} & \mathbf{Z} \in \mathbb{S}^N, \\ & \mathbf{Z} \succeq \mathbf{0},\ \mathbf{Z} \geq \mathbf{0},\ \mathbf{Z}\mathbf{1} = \mathbf{1},\ \mathrm{Tr}(\mathbf{Z}) = K, \end{array}$$

where $\mathbb{I}$ denotes the identity matrix of dimension N. Note that the constraints $\mathbf{Z} \geq \mathbf{0}$ and $\mathbf{Z1} = \mathbf{1}$ ensure that $\mathbf{Z}$ is a stochastic matrix, and hence all of its eigenvalues lie between 0 and 1. Thus, further relaxing the nonnegativity constraints leads to the following weaker relaxation:

$$(\mathcal{PW}_2) \qquad \begin{array}{ll} \text{minimize} & \langle \mathbf{W}, \mathbb{I} - \mathbf{Z} \rangle \\ \text{subject to} & \mathbf{Z} \in \mathbb{S}^N, \\ & \mathbb{I} \succeq \mathbf{Z} \succeq \mathbf{0},\ \mathbf{Z1} = \mathbf{1},\ \mathrm{Tr}(\mathbf{Z}) = K. \end{array}$$

Peng and Wei (2007) also demonstrate that $(\mathcal{PW}_2)$ essentially reduces to an eigenvalue problem, which implies that one can solve $(\mathcal{PW}_2)$ in $\mathcal{O}(KN^2)$ time; see Golub and van Loan (1996). Their SDP relaxation for the *balanced* K-means clustering problem is similar to $(\mathcal{PW}_1)$ and takes the form

$$(\mathcal{PW}_1^{\mathrm{b}}) \qquad \begin{array}{ll} \text{minimize} & \langle \mathbf{W}, \mathbb{I} - \mathbf{Z} \rangle \\ \text{subject to} & \mathbf{Z} \in \mathbb{S}^N, \\ & \mathbf{Z} \succeq \mathbf{0},\ \mathbf{0} \leq \mathbf{Z} \leq (K/N)\mathbf{1}\mathbf{1}^\top,\ \mathbf{Z1} = \mathbf{1},\ \mathrm{Tr}(\mathbf{Z}) = K. \end{array}$$

Awasthi et al. (2015) suggest another SDP relaxation for the unconstrained K-means clustering problem, based on the same matrix of squared pairwise distances $\mathbf{D}$ considered in this paper:

$$(\mathcal{A}) \qquad \begin{array}{ll} \text{minimize} & \langle \mathbf{D}, \mathbf{Z} \rangle \\ \text{subject to} & \mathbf{Z} \in \mathbb{S}^N, \\ & \mathbf{Z} \succeq \mathbf{0},\ \mathbf{Z} \geq \mathbf{0},\ \mathbf{Z1} = \mathbf{1},\ \mathrm{Tr}(\mathbf{Z}) = K. \end{array}$$

The following observation asserts that the stronger relaxation $(\mathcal{PW}_1)$ of Peng and Wei (2007) and the relaxation $(\mathcal{A})$ of Awasthi et al. (2015) are actually equivalent.

Observation 1. The problems $(\mathcal{PW}_1)$ and $(\mathcal{A})$ are equivalent.

Proof. Begin by expressing the objective of $(\mathcal{PW}_1)$ in terms of the pairwise distance matrix $\mathbf{D}$:

$$\begin{aligned} \langle \mathbf{W}, \mathbb{I} - \mathbf{Z} \rangle &= \frac{1}{2}[2\langle \mathbf{W}, \mathbb{I} \rangle - \langle 2\mathbf{W}, \mathbf{Z} \rangle - \langle \mathbf{D}, \mathbf{Z} \rangle] + \frac{1}{2}\langle \mathbf{D}, \mathbf{Z} \rangle \\ &= \frac{1}{2}[2\langle \mathbf{W}, \mathbb{I} \rangle - \langle 2\mathbf{W} + \mathbf{D}, \mathbf{Z} \rangle] + \frac{1}{2}\langle \mathbf{D}, \mathbf{Z} \rangle \\ &\overset{\text{(a)}}{=} \frac{1}{2}[2\langle \mathbf{W}, \mathbb{I} \rangle - \langle \mathbf{1}\,\mathrm{diag}(\mathbf{W})^\top + \mathrm{diag}(\mathbf{W})\,\mathbf{1}^\top, \mathbf{Z} \rangle] + \frac{1}{2}\langle \mathbf{D}, \mathbf{Z} \rangle \\ &= \frac{1}{2}[2\langle \mathbf{W}, \mathbb{I} \rangle - \langle \mathbf{1}\,\mathrm{diag}(\mathbf{W})^\top, \mathbf{Z} \rangle - \langle \mathrm{diag}(\mathbf{W})\,\mathbf{1}^\top, \mathbf{Z} \rangle] + \frac{1}{2}\langle \mathbf{D}, \mathbf{Z} \rangle \\ &= \frac{1}{2}[2\mathrm{Tr}(\mathbf{W}) - \mathrm{Tr}(\mathbf{Z}\,\mathbf{1}\,\mathrm{diag}(\mathbf{W})^\top) - \mathrm{Tr}(\mathrm{diag}(\mathbf{W})\,\mathbf{1}^\top\,\mathbf{Z})] + \frac{1}{2}\langle \mathbf{D}, \mathbf{Z} \rangle \\ &\overset{\text{(b)}}{=} \frac{1}{2}[2\mathrm{Tr}(\mathbf{W}) - \mathrm{Tr}(\mathbf{1}\,\mathrm{diag}(\mathbf{W})^\top) - \mathrm{Tr}(\mathrm{diag}(\mathbf{W})\,\mathbf{1}^\top)] + \frac{1}{2}\langle \mathbf{D}, \mathbf{Z} \rangle \\ &= \frac{1}{2}[2(\mathbf{1}^\top\mathrm{diag}(\mathbf{W})) - \mathbf{1}^\top\mathrm{diag}(\mathbf{W}) - \mathbf{1}^\top\mathrm{diag}(\mathbf{W})] + \frac{1}{2}\langle \mathbf{D}, \mathbf{Z} \rangle \\ &= \frac{1}{2}\langle \mathbf{D}, \mathbf{Z} \rangle. \end{aligned} \tag{4}$$

Here, (a) follows from the observation that the ijth element of the matrix $2\mathbf{W}+\mathbf{D}$ can be written as $2\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_j+\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_j\|^2=\|\boldsymbol{\xi}_i\|^2+\|\boldsymbol{\xi}_j\|^2$, and (b) uses the insights that $\mathbf{Z}\mathbf{1}=\mathbf{1}$ and $\mathbf{1}^\top\mathbf{Z}=\mathbf{1}^\top$. Comparing ($\mathcal{PW}_1$) and ($\mathcal{A}$), identity (4) shows that the two relaxations are equivalent because their objective functions are the same (up to a factor 2), while they share the same feasible set. ◻

Next, we establish that ($\mathcal{R}_{\mathrm{SDP}}$) is at least as tight a relaxation of the cardinality-constrained K-means clustering problem (1) as the stronger relaxation ($\mathcal{PW}_1$) of Peng and Wei (2007).

Proposition 5. *We have* $\min(\mathcal{R}_{\mathrm{SDP}})\geq\min(\mathcal{PW}_1)$.

Note that, through Observation 1, Proposition 5 also implies that ($\mathcal{R}_{\mathrm{SDP}}$) is at least as tight as the relaxation ($\mathcal{A}$) of Awasthi et al. (2015).

Proof of Proposition 5. To prove that ($\mathcal{R}_{\mathrm{SDP}}$) is at least as tight a relaxation as ($\mathcal{PW}_1$), we will argue that for every feasible solution $\{(\boldsymbol{x}^k,\mathbf{M}^k)\}_{k=1}^K$ of ($\mathcal{R}_{\mathrm{SDP}}$) one can construct a solution

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

that is feasible in ($\mathcal{PW}_1$) and achieves the same objective value. We first verify the feasibility of the proposed solution $\overline{\mathbf{Z}}$. Note that $\overline{\mathbf{Z}}$ is symmetric by construction. Next, we can directly verify that $\overline{\mathbf{Z}}$ is positive semidefinite since

$$\begin{aligned}
\overline{\mathbf{Z}}\succeq\mathbf{0}\Longleftarrow\;&\mathbf{M}^k+\mathbf{1}\mathbf{1}^\top+\boldsymbol{x}^k\mathbf{1}^\top+\mathbf{1}(\boldsymbol{x}^k)^\top\succeq\mathbf{0} && \forall k=1,\ldots,K\\
\Longleftrightarrow\;&\boldsymbol{v}^\top(\mathbf{M}^k+\mathbf{1}\mathbf{1}^\top+\boldsymbol{x}^k\mathbf{1}^\top+\mathbf{1}(\boldsymbol{x}^k)^\top)\,\boldsymbol{v}\geq 0 \quad \forall\boldsymbol{v}\in\mathbb{R}^N, && \forall k=1,\ldots,K\\
\Longleftarrow\;&\boldsymbol{v}^\top(\boldsymbol{x}^k(\boldsymbol{x}^k)^\top+\mathbf{1}\mathbf{1}^\top+\boldsymbol{x}^k\mathbf{1}^\top+\mathbf{1}(\boldsymbol{x}^k)^\top)\,\boldsymbol{v}\geq 0 \quad \forall\boldsymbol{v}\in\mathbb{R}^N, && \forall k=1,\ldots,K\\
\Longleftrightarrow\;&(\boldsymbol{v}^\top\boldsymbol{x}^k)^2+(\boldsymbol{v}^\top\mathbf{1})^2+2(\boldsymbol{v}^\top\boldsymbol{x}^k)(\boldsymbol{v}^\top\mathbf{1})\geq 0 \quad \forall\boldsymbol{v}\in\mathbb{R}^N, && \forall k=1,\ldots,K\\
\Longleftrightarrow\;&(\boldsymbol{v}^\top\boldsymbol{x}^k+\boldsymbol{v}^\top\mathbf{1})^2\geq 0 \quad \forall\boldsymbol{v}\in\mathbb{R}^N, && \forall k=1,\ldots,K,
\end{aligned}$$

where the third implication is due to the definition of $\mathcal{C}_{\mathrm{SDP}}(n_k)$, which requires that $\mathbf{M}^k\succeq\boldsymbol{x}^k(\boldsymbol{x}^k)^\top$. The last statement holds trivially because any quadratic form is nonnegative. Next, we can ensure the elementwise nonnegativity of $\overline{\mathbf{Z}}$, again through the definition of $\mathcal{C}_{\mathrm{SDP}}(n_k)$:

$$\overline{\mathbf{Z}}\geq\mathbf{0}\quad\Longleftarrow\quad\mathbf{M}^k+\mathbf{1}\mathbf{1}^\top+\boldsymbol{x}^k\mathbf{1}^\top+\mathbf{1}(\boldsymbol{x}^k)^\top\geq\mathbf{0}\quad\forall k=1,\ldots,K.$$

Furthermore, combining the definition of $\mathcal{C}_{\mathrm{SDP}}(n_k)$ and the constraint

$$\sum_{k=1}^K\boldsymbol{x}^k=(2-K)\mathbf{1}$$

of ($\mathcal{R}_{\mathrm{SDP}}$), we can see that each row of $\overline{\mathbf{Z}}$ indeed sums to 1:

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

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

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

Thus, $\overline{\mathbf{Z}}$ is feasible in $(\mathcal{PW}_1)$, and it remains to prove that it achieves the same objective value as the original solution $\{(\boldsymbol{x}^k, \mathbf{M}^k)\}_{k=1}^K$ in $(\mathcal{R}_{\mathrm{SDP}})$. Invoking relation (4), it is easy to see that

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

The proof is thus concluded. □

Finally, we assert that $(\mathcal{R}_{\mathrm{SDP}}^{\mathrm{b}})$ is at least as tight a relaxation of the balanced K-means clustering problem as the corresponding relaxation $(\mathcal{PW}_1^{\mathrm{b}})$ of Peng and Wei (2007).

PROPOSITION 6. *We have* $\min(\mathcal{R}_{\mathrm{SDP}}^{\mathrm{b}}) \geq \min(\mathcal{PW}_1^{\mathrm{b}})$.

Proof. To show that $(\mathcal{R}_{\mathrm{SDP}}^{\mathrm{b}})$ is at least as tight a relaxation as $(\mathcal{PW}_1^{\mathrm{b}})$, we will again argue that for every feasible solution $\{(\boldsymbol{x}^1, \mathbf{M}^1), (\boldsymbol{x}, \mathbf{M})\}$ of $(\mathcal{R}_{\mathrm{SDP}}^{\mathrm{b}})$ one can construct a solution

$$\overline{\mathbf{Z}} = \frac{K}{4N}((\mathbf{M}^1 + \mathbf{1}\mathbf{1}^\top + \boldsymbol{x}^1\mathbf{1}^\top + \mathbf{1}(\boldsymbol{x}^1)^\top) + (K-1)(\mathbf{M} + \mathbf{1}\mathbf{1}^\top + \boldsymbol{x}\mathbf{1}^\top + \mathbf{1}\boldsymbol{x}^\top))$$

that is feasible in $(\mathcal{PW}_1^{\mathrm{b}})$ and achieves the same objective value. Following similar steps as in the proof of Proposition 5, one can verify that $\overline{\mathbf{Z}}$ indeed satisfies $\overline{\mathbf{Z}} \succeq \mathbf{0}$, $\overline{\mathbf{Z}} \geq \mathbf{0}$, $\overline{\mathbf{Z}}\mathbf{1} = \mathbf{1}$, and $\operatorname{Tr}(\overline{\mathbf{Z}}) = K$. In order to see that $\overline{\mathbf{Z}} \leq (K/N)\mathbf{1}\mathbf{1}^\top$, note from the definition of $\mathcal{C}_{\mathrm{SDP}}(n)$ (where $n = N/K$ denotes the shared cardinality of all clusters) that

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

Using this insight and the constraint $\boldsymbol{x}^1 + (K-1)\boldsymbol{x} = (2-K)\mathbf{1}$ of $(\mathcal{R}_{\mathrm{SDP}}^{\mathrm{b}})$, any arbitrary element $\overline{z}_{ij}$ of $\overline{\mathbf{Z}}$ can be bounded above as desired:

$$\begin{aligned}\overline{z}_{ij} &= \frac{K}{4N}((m_{ij}^1 + 1 + x_i^1 + x_j^1) + (K-1)(m_{ij} + 1 + x_i + x_j))\\ &\leq \frac{K}{4N}(2K + x_i^1 + (K-1)x_i + x_j^1 + (K-1)x_j) = \frac{K}{N}.\end{aligned}$$

Finally, a direct calculation reveals that the objective value of $(\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}})$ evaluated at $\{(\boldsymbol{x}^1,\mathbf{M}^1),(\boldsymbol{x},\mathbf{M})\}$ coincides with the objective of $(\mathcal{PW}^{\mathrm{b}}_1)$ evaluated at $\overline{\mathbf{Z}}$, which, from (4), is equal to $\frac{1}{2}\langle\mathbf{D},\overline{\mathbf{Z}}\rangle$. Hence, $(\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}})$ is at least as tight a relaxation as $(\mathcal{PW}^{\mathrm{b}}_1)$, and the proof is concluded. ∎

Note that, while Propositions 5 and 6 demonstrate that our SDP relaxations $(\mathcal{R}_{\mathrm{SDP}})$ and $(\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}})$ are at least as tight as their respective counterparts by Peng and Wei (2007), similar tightness results cannot be established for our LP relaxations. Indeed, our numerical experiments based on real-world data sets in section 5 show that both $(\mathcal{R}_{\mathrm{LP}})$ and $(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$ can be strictly weaker than $(\mathcal{PW}_1)$ and $(\mathcal{PW}^{\mathrm{b}}_1)$, respectively. Furthermore, it is possible to construct artificial data sets on which even $(\mathcal{PW}_2)$ outperforms $(\mathcal{R}_{\mathrm{LP}})$ and $(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$.

3.4. 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 is a balanced partition $(J_1,\ldots,J_K)$ of $\{1,\ldots,N\}$ where each cluster $k=1,\ldots,K$ has the same cardinality $|J_k|=N/K\in\mathbb{N}$ and

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

Assumption (S) implies that the data set admits the natural balanced clustering $(J_1,\ldots,J_K)$, and that the largest cluster diameter (that is, $\max_{1\le k\le K}\max_{i,j\in J_k} d_{ij}$) is smaller than the smallest distance between any two distinct clusters (that is, $\min_{1\le k_1<k_2\le K}\min_{i\in J_{k_1},j\in J_{k_2}} d_{ij}$).

THEOREM 3. *If assumption* (S) *holds, then the optimal values of* $(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$ *and* $(\mathcal{P})$ *coincide. Moreover, the clustering* $(J_1,\ldots,J_K)$ *is optimal in* $(\mathcal{P})$ *and is recovered by Algorithm* 2.

Put simply, Theorem 3 states that, for data sets whose hidden classes are balanced and well separated, Algorithm 2 will succeed in recovering this hidden, provably optimal clustering.

Proof of Theorem 3. Throughout the proof we assume without loss of generality that the clustering $(J_1,\ldots,J_K)$ from assumption (S) satisfies $1\in J_1$, that is, the cluster containing the data point $\boldsymbol{\xi}_1$ 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 $(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$ and the MILP $(\mathcal{P})$ are equal and that they both coincide with the sum of squared intra-cluster distances of the clustering $(J_1,\ldots,J_K)$, 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 $(I_1,\ldots,I_K)$ of Algorithm 2 coincides with the optimal clustering $(J_1,\ldots,J_K)$ 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 $(\boldsymbol{x}^1,\boldsymbol{x},\mathbf{M}^1,\mathbf{M})$ of $(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$, we define $\mathbf{H},\mathbf{W}\in\mathbb{S}^N$ through

$$\mathbf{H}=\mathbf{M}^1+\mathbf{1}\mathbf{1}^\top+\boldsymbol{x}^1\mathbf{1}^\top+\mathbf{1}(\boldsymbol{x}^1)^\top \quad\text{and}\quad \mathbf{W}=\mathbf{M}+\mathbf{1}\mathbf{1}^\top+\boldsymbol{x}\mathbf{1}^\top+\mathbf{1}\boldsymbol{x}^\top. \tag{5}$$

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

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

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

$$\frac{1}{8n}\langle \mathbf{D}, \mathbf{H} + (K-1)\mathbf{W}\rangle = \frac{1}{8n}\sum_{i\neq j} d_{ij}(h_{ij} + (K-1)w_{ij}). \tag{6}$$

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

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

Furthermore, each weight $h_{ij} + (K-1)w_{ij}$ is bounded above by 4 because

$$\begin{aligned}h_{ij} + (K-1)w_{ij} &= (m^1_{ij} + 1 + x^1_i + x^1_j) + (K-1)(m_{ij} + 1 + x_i + x_j)\\ &\leq 2K + (x^1_i + (K-1)x_i) + (x^1_j + (K-1)x_j) = 4,\end{aligned} \tag{7}$$

where the inequality holds because $\mathbf{M}^1, \mathbf{M} \leq \mathbf{1}\mathbf{1}^\top$ (which we know from the proof of Proposition 6) and the last equality follows from the constraint $\boldsymbol{x}^1 + (K-1)\boldsymbol{x} = (2-K)\mathbf{1}$ in $(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$.

Hence, the sum on the right-hand side of (6) 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

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

where the last equality follows from assumption (S). By Lemma 1, the right-hand side of (8) represents the objective value of the clustering $(J_1, \ldots, J_K)$ in the MILP $(\mathcal{P})$. Thus, $(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$ provides an upper bound on $(\mathcal{P})$. By Corollary 1, $(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$ also provides a lower bound on $(\mathcal{P})$. We may thus conclude that the LP relaxation $(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$ is tight and, as a consequence, that the clustering $(J_1, \ldots, J_K)$ is indeed optimal in $(\mathcal{P})$.

Step 2. As the inequality in (8) is tight, any optimal solution to $(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$ satisfies $h_{ij} + (K-1)w_{ij} = 4$ whenever $i \neq j$ and $i, j \in J_k$ for some $k = 1, \ldots, K$ (i.e., whenever the data points $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ 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 (7) that $m^1_{1i} = m_{1i} = 1$ for all $i \in J_1$.

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

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

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 $\mathbf{M}^1\mathbf{1} = (2n-N)\boldsymbol{x}^1$, which is one of the constraints in $(\mathcal{R}_{\mathrm{LP}}^{\mathrm{b}})$, 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 1ith entry of the matrix inequality $\mathbf{M}^1 + \mathbf{1}\mathbf{1}^\top - \boldsymbol{x}^1\mathbf{1}^\top - \mathbf{1}(\boldsymbol{x}^1)^\top \geq \mathbf{0}$ from the definition of $\mathcal{C}_{\mathrm{LP}}(n)$ can be expressed as

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

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 $(\mathcal{R}_{\mathrm{LP}}^{\mathrm{b}})$. Similarly, for any $i \in J_1$, the iith entry of the matrix inequality

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

can be rewritten as

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

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

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 $\mathbf{1}^\top\boldsymbol{x}^1 = 2n - N$ from the definition of $\mathcal{C}_{\mathrm{LP}}(n)$ 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 $\boldsymbol{x}^1$, we conclude that it must output $I_1 = J_1$, which completes the proof.

Theorem 3 implies via Corollary 1 that the optimal values of $(\mathcal{R}_{\mathrm{SDP}}^{\mathrm{b}})$ and $(\mathcal{P})$ are also equal. Thus, both the LP and the SDP relaxations lead to perfect recovery.

In the related literature, assumption (S) has previously been used by Elhamifar, Sapiro, and Vidal (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 3 is reminiscent of Awasthi et al. (2015, Theorem 9), which formalizes the recovery properties of their LP relaxation for the unconstrained K-means clustering problem. Awasthi et al. (2015) assume, however, that the data points 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. For instance, a simple threshold approach (i.e., assigning data points to the same cluster whenever the distance between them falls below a given threshold) would be able to recover the true clusters whenever assumption (S) holds. It seems unlikely, however, that such approaches would perform well in a setting where assumption (S) is not satisfied. In fact, Awasthi et al. (2015) show that their LP relaxation fails to recover the true clusters with high probability if assumption (S) is violated. In contrast, the numerical experiments of section 5 suggest that Algorithms 1 and 2 perform well even if assumption (S) is violated.

Remark 3. To the best of our knowledge, there is no perfect recovery result for the cardinality-constrained K-means clustering algorithm by Bennett, Bradley, and Demiriz (2000) (see Appendix A), whose performance depends critically on its initialization. To see that it can be trapped in a local optimum, consider the $N = 4$ two-dimensional data points $\boldsymbol{\xi}_1 = (0, 0)$, $\boldsymbol{\xi}_2 = (a, 0)$, $\boldsymbol{\xi}_3 = (a, b)$, and $\boldsymbol{\xi}_4 = (0, b)$ 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 3.

4. Cardinality-constrained clustering with outliers.

If the data set is corrupted by outliers, then the optimal value of (1) may be high, indicating that the data set 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 data points 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 data points, then the optimal value of (1) drops to zero, and Algorithm 2 detects the optimal clustering due to Theorem 3.

We now show that the results of section 3 (particularly Proposition 2 and Theorem 2) extend to situations where n_0 data points 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 data points 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 $N - n_0$ nonoutlier data points. We first prove that the joint outlier detection and cardinality-constrained clustering problem admits an exact MILP reformulation.

PROPOSITION 7 (MILP reformulation). *The joint outlier detection and cardinality-constrained clustering problem is equivalent to the MILP*

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

Proof. The proof is an immediate extension of Proposition 2 to account for the outlier cluster. ∎

In analogy to section 3.1, one can demonstrate that the MILP ($\mathcal{P}^{\rm o}$) admits the SDP relaxation

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

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

THEOREM 4 (SDP and LP relaxations). *We have*

$$\min(\mathcal{R}^{\rm o}_{\rm LP}) \le \min(\mathcal{R}^{\rm o}_{\rm SDP}) \le \min(\mathcal{P}^{\rm o}).$$

Proof. This result generalizes Theorem 2 to account for the additional outlier cluster. As it needs no fundamentally new ideas, the proof is omitted for brevity. □

The relaxations ($\mathcal{R}^{\rm o}_{\rm SDP}$) and ($\mathcal{R}^{\rm o}_{\rm LP}$) not only provide a lower bound on ($\mathcal{P}^{\rm o}$), but also give rise to a rounding algorithm that recovers a feasible clustering and thus an upper bound on ($\mathcal{P}^{\rm o}$); 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.

```
Input: $\mathcal{I}_0 = \{1, \dots, N\}$ (data indices), $n_k \in \mathbb{N}$, $k = 0, \dots, K$ (cluster sizes).
Solve ($\mathcal{R}^{\rm o}_{\rm SDP}$) or ($\mathcal{R}^{\rm o}_{\rm LP}$) for the data points $\boldsymbol{\xi}_i$, $i \in \mathcal{I}_0$, and record the optimal
    $\boldsymbol{x}^0 \in \mathbb{R}^N$.
Determine a bijection $\rho : \mathcal{I}_0 \to \mathcal{I}_0$ such that $x^0_{\rho(1)} \ge x^0_{\rho(2)} \ge \cdots \ge x^0_{\rho(N)}$.
Set $I_0 \leftarrow \{\rho(1), \dots, \rho(n_0)\}$ and $\mathcal{I}_1 \leftarrow \mathcal{I}_0 \setminus I_0$.
Call Algorithm 1 with input $(\mathcal{I}_1, \{n_k\}_{k=1}^K)$ to obtain $I_1, \dots, I_K$.
Output: $I_0, \dots, I_K$.
```

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

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

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

COROLLARY 2 (relaxations for balanced clustering). *We have*

$$\min(\mathcal{R}^{\rm ob}_{\rm LP}) \le \min(\mathcal{R}^{\rm ob}_{\rm SDP}) \le \min(\mathcal{P}^{\rm o}).$$

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

If the normal clusters are required to be balanced, then Algorithm 3 should be modified as follows. First, in step 2 the relaxations ($\mathcal{R}^{\text{ob}}_{\text{SDP}}$) or ($\mathcal{R}^{\text{ob}}_{\text{LP}}$) can be solved instead of ($\mathcal{R}^{\text{o}}_{\text{SDP}}$) or ($\mathcal{R}^{\text{o}}_{\text{LP}}$), 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 3 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 $(J_0, J_1, \ldots, J_K)$ of $\{1, \ldots, N\}$ where each normal cluster $k = 1, \ldots, K$ has the same cardinality $|J_k| = (N - n_0)/K \in \mathbb{N}$, while

$$\max_{1\le k\le K}\max_{i,j\in J_k} d_{ij} < \min_{1\le k_1<k_2\le K}\;\min_{i\in J_{k_1},\, j\in J_{k_2}} d_{ij} \text{ and } \max_{1\le k\le K}\max_{i,j\in J_k} d_{ij} < \min_{\substack{i\in J_0,\\ j\in\{1,\ldots,N\}\setminus\{i\}}} d_{ij}.$$

Assumption (S$'$) implies that the data set admits the natural outlier cluster J_0 and the natural normal clusters $(J_1, \ldots, J_K)$. 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 data point. Under this condition, Algorithm 3 correctly identifies the optimal clustering.

THEOREM 5. *If assumption* (S$'$) *holds, then the optimal values of* ($\mathcal{R}^{\text{ob}}_{\text{LP}}$) *and* ($\mathcal{P}^{\text{o}}$) *coincide. Moreover, the clustering* $(J_0, \ldots, J_K)$ *is optimal in* ($\mathcal{P}^{\text{o}}$) *and is recovered by Algorithm* 3.

Proof. The proof parallels that of Theorem 3 and can be divided into two steps. In the first step we show that the LP relaxation ($\mathcal{R}^{\text{ob}}_{\text{LP}}$) for balanced clustering and outlier detection is tight, and in the second step we demonstrate that Algorithm 3 correctly identifies the clusters $(J_0, \ldots, J_K)$. 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 data set satisfies assumption (S), and Theorem 3 guarantees that the normal clusters $(J_1, \ldots, J_K)$ are correctly identified with Algorithm 2.

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

$$\begin{aligned}\operatorname{diag}(\mathbf{M} + \mathbf{1}\mathbf{1}^\top + \boldsymbol{x}\mathbf{1}^\top + \mathbf{1}\boldsymbol{x}^\top) \ge \mathbf{0} &\implies \boldsymbol{x} \ge -\mathbf{1},\\ \operatorname{diag}(\mathbf{M} + \mathbf{1}\mathbf{1}^\top - \boldsymbol{x}\mathbf{1}^\top - \mathbf{1}\boldsymbol{x}^\top) \ge \mathbf{0} &\implies \boldsymbol{x} \le +\mathbf{1},\end{aligned}$$

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

Step 1. For any feasible solution $(\boldsymbol{x}^0, \boldsymbol{x}, \mathbf{M}^0, \mathbf{M})$ of ($\mathcal{R}^{\text{ob}}_{\text{LP}}$), introduce the auxiliary matrix $\mathbf{H} = \mathbf{M} + \mathbf{1}\mathbf{1}^\top + \mathbf{1}\boldsymbol{x}^\top + \boldsymbol{x}\mathbf{1}^\top$. Recall from the proof of Theorem 3 that $\mathbf{H} \ge \mathbf{0}$ and

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

The constraint $K\boldsymbol{x} + \boldsymbol{x}^0 = (1-K)\mathbf{1}$ from ($\mathcal{R}^{\text{ob}}_{\text{LP}}$) ensures via the inequality $-\mathbf{1} \le \boldsymbol{x}^0$ that $\boldsymbol{x} \le (\frac{2}{K} - 1)\mathbf{1}$. Recalling from the proof of Theorem 3 that $\mathbf{M} \le \mathbf{1}\mathbf{1}^\top$, we then find

$$h_{ij} = m_{ij} + 1 + x_i + x_j \le 1 + 1 + \left(\frac{2}{K} - 1\right) + \left(\frac{2}{K} - 1\right) = \frac{4}{K} \quad \forall i, j = 1, \ldots, N. \tag{9}$$

Similar arguments to those in the proof of Theorem 3 reveal that the objective function of the joint outlier detection and (balanced) clustering problem ($\mathcal{R}_{\rm LP}^{\rm ob}$) can be expressed as

$$\begin{aligned}\frac{K}{8n}\langle \mathbf{D}, \mathbf{H}\rangle &\geq \frac{1}{2n}\left\{\text{sum of the } Kn(n-1) \text{ smallest entries of } d_{ij} \text{ with } i \neq j\right\}\\ &= \frac{1}{2n}\sum_{k=1}^{K}\sum_{i,j\in J_k} d_{ij},\end{aligned} \tag{10}$$

where the equality follows from assumption (S′). By Lemma 1, the right-hand side of (10) represents the objective value of the clustering $(J_0, \ldots, J_K)$ in the MILP ($\mathcal{P}^{\rm o}$). Thus, ($\mathcal{R}_{\rm LP}^{\rm ob}$) provides an upper bound on ($\mathcal{P}^{\rm o}$). By Corollary 2, ($\mathcal{R}_{\rm LP}^{\rm ob}$) also provides a lower bound on ($\mathcal{P}^{\rm o}$). We may thus conclude that the LP relaxation ($\mathcal{R}_{\rm LP}^{\rm ob}$) is tight and, as a consequence, that the clustering $(J_0, \ldots, J_K)$ is indeed optimal in ($\mathcal{P}^{\rm o}$).

Step 2. As the inequality in (10) is tight, any optimal solution to ($\mathcal{R}_{\rm LP}^{\rm ob}$) satisfies $h_{ij} = \frac{4}{K}$ whenever $i \neq j$ and $i, j \in J_k$ for some $k = 1, \ldots, K$ (i.e., whenever $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ are *not* outliers and belong to the same cluster). This in turn implies via (9) that $x_i = \frac{2}{K} - 1$ for all $i \in \cup_{k=1}^{K} J_k$. Furthermore, the constraint $\mathbf{1}^\top \boldsymbol{x} = 2n - N$ from $\mathcal{C}_{\rm LP}(n)$ 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 $-\mathbf{1} \leq \boldsymbol{x}$. 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 $K\boldsymbol{x} + \boldsymbol{x}^0 = (1-K)\mathbf{1}$ from ($\mathcal{R}_{\rm LP}^{\rm ob}$) 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 $\boldsymbol{x}^0$, we conclude that it must output $I_0 = J_0$, which completes the proof. □

Remark 4 (unknown cluster cardinalities). The joint outlier detection and cardinality-constrained clustering problem ($\mathcal{P}^{\rm o}$) can also be used when the number of outliers is not precisely known and only an estimate of the relative size (as opposed to the exact cardinality) of the clusters is available. To this end, we solve ($\mathcal{P}^{\rm o}$) for different values of n_0, respectively assigning the remaining $N - n_0$ data points to clusters whose relative sizes respect the available estimates. The value $n_0^\star$ representing the most reasonable number of outliers to be removed from the data set can then be determined using the elbow method; see, e.g., Gareth et al. (2017, Chapter 10).

As an illustration, consider again the data set depicted in Figure 1, which showcases the crux of outlier detection in the context of cardinality-constrained clustering. In section 1, we inadvertently assumed we had the knowledge that the data set under consideration was contaminated by three outliers. To demonstrate the practical usefulness of our approach, we will now employ the elbow method to determine the number of outliers n_0 without making any assumptions about the data set. As elucidated in Remark 4, the ideal value of n_0 can be determined by solving problem ($\mathcal{P}^{\rm o}$) repeatedly. However, as ($\mathcal{P}^{\rm o}$) constitutes an intractable optimization problem, we solve its convex relaxations ($\mathcal{R}_{\rm LP}^{\rm ob}$) and ($\mathcal{R}_{\rm SDP}^{\rm ob}$) instead and plot the resulting objective values in logarithmic scale in Figure 2. It becomes apparent that $n_0^\star = 3$ is most appropriate as it marks the transition from the initially steep decline pattern of the objective value to a substantially flatter decline pattern. Note that n_0 needs to be a multiple of $K = 3$ to allow for balanced clustering.

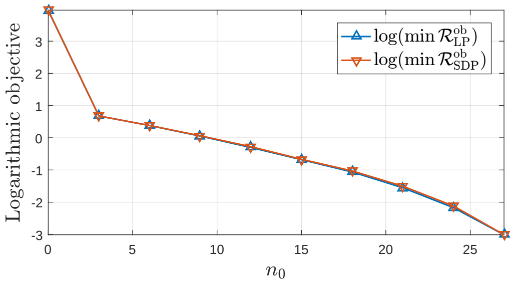


FIG. 2. *Elbow plot for the data set depicted in Figure* 1.

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 3.40 GHz i7 computer with 16 GB RAM.

5.1. Cardinality-constrained K-means clustering (real-world data).

We compare the performance of our algorithms from section 3 with the algorithm of Bennett, Bradley, and Demiriz (2000) (see Appendix A) and with the two SDP relaxations proposed by Peng and Wei (2007) on the classification data sets of the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) with 150–300 data points, up to 200 continuous attributes, and no missing values. Table 2 reports the main characteristics of these data sets. In our experiments, we set the cluster cardinalities to the numbers of true class occurrences in each data set. It should be emphasized that, in contrast to the other methods, and with the exception of the two balanced data sets, the SDP relaxations of Peng and Wei (2007) do not have access to the cluster cardinalities. They should thus be seen as a baseline for the performance of the other methods. Furthermore, we remark that all data sets severely violate assumption (S). Indeed, the ratios of the largest cluster diameter to the smallest distance between clusters (when the clusters are determined by the true labels) vary from 7 to 149, while they should be less than 1 in order to satisfy assumption (S). Also, only two data sets actually entail balanced clusters.

Table 3 reports the lower bounds provided by $(\mathcal{R}_{\mathrm{LP}})/(\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}})$ and $(\mathcal{R}_{\mathrm{SDP}})/(\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}})$ (LB), the upper bounds from Algorithms 1 and 2 (UB), the objective value of the best of 10 runs of the algorithm of Bennett, Bradley, and Demiriz (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), the respective lower bounds (LB) obtained from the SDP relaxations $(\mathcal{PW}_1)/(\mathcal{PW}^{\mathrm{b}}_1)$ and $(\mathcal{PW}_2)$ of Peng and Wei (2007), and the solution times for each of these methods. The latter was limited to a maximum of three hours, and in one case (namely, "glass identification"), $(\mathcal{R}_{\mathrm{SDP}})$ did not terminate within this limit. The "—" in Table 3 indicate this occurrence.

The obtained lower bounds of $(\mathcal{R}_{\mathrm{SDP}})/(\mathcal{R}^{\mathrm{b}}_{\mathrm{SDP}})$ allow us to certify that the algorithm of Bennett, Bradley, and Demiriz (2000) provides nearly optimal solutions in almost all instances. Also, both Algorithms 1 and 2 are competitive with the

TABLE 2
Overview of the main characteristics of the relevant data sets.

ID	Data set name	N (# data points)	d (# dimensions)	K (# clusters)	n_k (cardinalities)	Balanced
1	Iris	150	4	3	50, 50, 50	yes
2	Seeds	210	7	3	70, 70, 70	yes
3	Planning relax	182	12	2	130, 52	no
4	Connectionist bench	208	60	2	111, 97	no
5	Urban land cover	168	147	9	23, 29, 14, 15, 17, 25, 16, 14, 15	no
6	Parkinsons	195	22	2	48, 147	no
7	Glass identification	214	9	6	70, 76, 17, 13, 9, 29	no

TABLE 3
Performance of $(\mathcal{R}_{\rm LP})$, $(\mathcal{R}_{\rm SDP})$, Bennett, Bradley, and Demiriz (BBD), and Peng and Wei. The "—" indicate that the problem instance could not be solved within a time limit of three hours.

	$(\mathcal{R}_{\rm LP})/(\mathcal{R}^{\rm b}_{\rm LP})$			$(\mathcal{R}_{\rm SDP})/(\mathcal{R}^{\rm b}_{\rm SDP})$			BBD			$(\mathcal{PW}_1)/(\mathcal{PW}^{\rm b}_1)$		$(\mathcal{PW}_2)$	
ID	UB	LB	Time [s]	UB	LB	Time [s]	UB	CV [%]	Time [s]	LB	Time [s]	LB	Time [s]
1	81.4	78.8	17	81.4	81.4	584	81.4	0.0	6	81.4	154	15.2	0.02
2	620.7	539.0	46	605.6	605.6	3,823	605.6	0.0	7	604.5	1,320	19.0	0.03
3	325.9	297.0	24	315.7	315.7	2,637	315.8	0.3	9	299.0	510	273.7	0.02
4	312.6	259.1	49	280.6	280.1	3,638	280.6	0.4	6	270.0	1,376	246.2	0.04
5	3.61e9	3.17e9	2,241	3.54e9	3.44e9	10,754	3.64e9	9.2	13	2.05e9	460	1.94e8	0.02
6	1.36e6	1.36e6	22	1.36e6	1.36e6	2,000	1.36e6	15.1	7	1.11e6	777	6.31e5	0.02
7	469.0	377.2	232	—	—	—	438.2	28.4	13	321.9	1,500	23.8	0.03

algorithm of Bennett, Bradley, and Demiriz (2000) in terms of solution quality while providing rigorous error bounds. Moreover, as expected in view of Propositions 5 and 6, for all data sets, $(\mathcal{R}_{\rm SDP})/(\mathcal{R}^{\rm b}_{\rm SDP})$ yield better lower bounds than the SDP relaxations $(\mathcal{PW}_1)/(\mathcal{PW}^{\rm b}_1)$ and $(\mathcal{PW}_2)$ of Peng and Wei (2007). The lower bounds obtained from $(\mathcal{R}_{\rm LP})/(\mathcal{R}^{\rm b}_{\rm LP})$ are competitive with those provided by the relaxations $(\mathcal{PW}_1)/(\mathcal{PW}^{\rm b}_1)$, and they are always better than the lower bounds provided by their relaxation $(\mathcal{PW}_2)$. It should be mentioned, however, that one can construct instances where the situation is reversed, i.e., both $(\mathcal{PW}_1)/(\mathcal{PW}^{\rm b}_1)$ and $(\mathcal{PW}_2)$ are tighter than $(\mathcal{R}_{\rm LP})/(\mathcal{R}^{\rm b}_{\rm LP})$. 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, Katoh, and Hiroshi (1994). Furthermore, it is not clear how to impose cardinality constraints in this setting.

Specifically, in the context of the two balanced data sets (i.e., "iris" and "seeds"), we can enrich the preceding comparison with the heuristics proposed by Costa, Aloise, and Mladenović (2017) and Malinen and Fränti (2014). As for the variable neighborhood search method of Costa, Aloise, and Mladenović (2017), we were provided with the executables of the *C++* implementation used in that paper. For the "iris" data set, the best objective value out of 10 independent runs of this method was 81.4 (which is provably optimal thanks to the lower bounds provided by $(\mathcal{R}_{\rm SDP})$ and $(\mathcal{PW}^{\rm b}_1)$) and the time to execute all runs was 0.12 seconds. For the "seeds" data set, the best objective value out of 10 independent runs was 605.6 (again, provably optimal in view of the lower bound provided by $(\mathcal{R}_{\rm SDP})$) and the overall runtime was 0.53 seconds. The algorithm of Malinen and Fränti (2014) follows the same steps as the one of Bennett, Bradley, and Demiriz (2000), with the improvement that the cluster assignment step is solved by the Hungarian algorithm, which provides better runtime guarantees and typically solves faster than interior-point methods for LPs. For this reason, the upper bounds of Malinen and Fränti (2014) for the "iris" and "seeds" data sets coincide with those of Bennett, Bradley, and Demiriz (2000), while their algorithm can be expected to terminate faster. A direct comparison of the time complexity of these two methods can be found in Malinen and Fränti (2014).

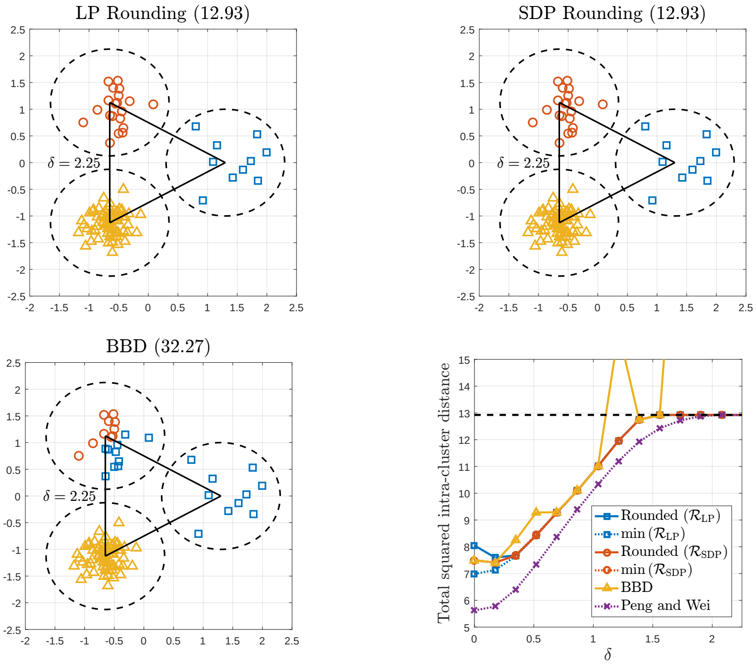


FIG. 3. *Comparison between different algorithms for (cardinality-constrained) K-means clustering for* 100 *data points, where the cardinalities are given by* $(n_1, n_2, n_3) = (10, 20, 70)$. *Indicated in parentheses next to the panel titles are the respective sums of squared intra-cluster distances achieved.*

5.2. Cardinality-constrained K-means clustering (synthetic data). We now randomly generate partitions of 10, 20, and 70 data points in $\mathbb{R}^2$ that are drawn from uniform distributions over $K = 3$ unit balls centered at $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$, and $\boldsymbol{\zeta}_3$, respectively, such that $\|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\| = \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_3\| = \|\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_3\| = \delta$. Theorem 3 shows that ($\mathcal{R}^{\mathrm{b}}_{\mathrm{LP}}$) is tight and that Algorithm 2 can recover the true clusters whenever $n_1 = n_2 = n_3$ and $\delta \geq 4$. Figure 3 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 ($\mathcal{R}_{\mathrm{SDP}}$) outperforms ($\mathcal{R}_{\mathrm{LP}}$) when δ is small, and that the algorithm of Bennett, Bradley, and Demiriz frequently fails to determine the optimal solution even if it is run 10 times. In line with the results from the real-world data sets, ($\mathcal{R}_{\mathrm{SDP}}$) and ($\mathcal{R}_{\mathrm{LP}}$) 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 7 seconds ($\mathcal{R}_{\mathrm{LP}}$), 106 seconds ($\mathcal{R}_{\mathrm{SDP}}$), 11 seconds (Bennett, Bradley, and Demiriz), and 16 seconds (Peng and Wei).

5.3. Outlier detection. We use ($\mathcal{R}^{\mathrm{o}}_{\mathrm{LP}}$) and Algorithm 3 to classify the *Breast Cancer Wisconsin (Diagnostic)* data set. The data set has $d = 30$ numerical features, which we standardize using a Z-score transformation, and it contains 357 benign and

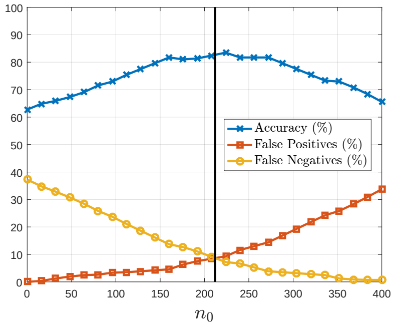


FIG. 4. *Outlier detection for breast cancer diagnosis.*

212 malignant cases of breast cancer. We interpret the malignant cases as outliers and thus set $K = 1$. Figure 4 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 data points can lead to cancer predictors of decent quality. The average runtime is 286 seconds, and the optimality gap is consistently below 3.23% for all values of n_0.

6. Conclusion. Clustering is a hard combinatorial optimization problem. For decades, it has almost exclusively been addressed by heuristic approaches. Many of these heuristics have proven to be very successful in practice as they often provide solutions of high, or at least satisfactory, quality within attractive runtimes. The common drawback of these methods is that there is typically no way of certifying the optimality of the provided solutions nor of giving guaranteed bounds on their suboptimality.

Maybe precisely because of this shortcoming, more recently, convex optimization approaches have been proposed for solving relaxed versions of the clustering problem. These conic programs are polynomial-time solvable and offer bounds on the suboptimality of a given solution. Furthermore, the solutions of these conic relaxations can be "rounded" to obtain actually feasible solutions to the original clustering problem, which results in a new class of heuristic methods.

The results presented in this paper follow precisely this recent paradigm. Combined, conic relaxations and (rounding) heuristics offer solutions to the clustering problem together with a posteriori guarantees on their optimality. Naturally, one would also wish for attractive a priori guarantees on the performance of these combined methods. The conditions required to derive such a priori guarantees are still quite restrictive, but the strong performance of these methods in practical instances makes us confident that this is a promising avenue for future research.

Appendix A. Bennett, Bradley, and Demiriz's algorithm. The algorithm of Bennett, Bradley, and Demiriz (2000) is designed for a variant of problem (1), where only lower bounds on the clusters' cardinalities are imposed. This algorithm has the following natural extension to our cardinality-constrained clustering problem (1).

Algorithm 4 The algorithm of Bennett, Bradley, and Demiriz for cardinality-constrained clustering.

```
1: **Input:** $\mathcal{I}_1 = \{1, \ldots, N\}$ (data indices), $n_k \in \mathbb{N}$, $k = 1, \ldots, K$ (cluster sizes).
2: Generate the cluster centers $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_K \in \mathbb{R}^d$.
3: Solve the linear assignment problem
       $\boldsymbol{\Pi}^\star \in \underset{\boldsymbol{\Pi}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \sum_{k=1}^{K} \pi_i^k \|\boldsymbol{\xi}_i - \boldsymbol{\zeta}_k\|^2 : \begin{array}{l} \pi_i^k \in \{0,1\}, \\ \sum_{i=1}^{N} \pi_i^k = n_k \ \forall k, \\ \sum_{k=1}^{K} \pi_i^k = 1 \ \forall i \end{array} \right\}.$
4: Set $I_k \leftarrow \{i : (\pi^\star)_i^k = 1\}$ for all $k = 1, \ldots, K$.
5: Set $\boldsymbol{\zeta}_k \leftarrow \frac{1}{n_k} \sum_{i \in I_k} \boldsymbol{\xi}_i$ for all $k = 1, \ldots, K$.
6: Repeat steps 3–5 until there are no more changes in $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_K$.
7: **Output:** $I_1, \ldots, I_K$.
```

Algorithm 4 adapts to problem (1) a classical local search heuristic for the unconstrained K-means clustering problem due to Lloyd (1982). At initialization, it generates random cluster centers $\boldsymbol{\zeta}_k$, $k = 1, \ldots, K$. Each subsequent iteration of the algorithm consists of two steps. The first step assigns every data point $\boldsymbol{\xi}_i$ to the nearest cluster center while adhering to the prescribed cluster cardinalities, whereas the second step replaces each center $\boldsymbol{\zeta}_k$ with the mean of the data points that have been assigned to cluster k. The algorithm terminates when the cluster centers $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_K$ no longer change.

Acknowledgments. We are grateful to Leandro Costa and Daniel Aloise for sharing with us the executables of the variable neighborhood search heuristic proposed in Costa, Aloise, and Mladenović (2017).

REFERENCES

C. Aggarwal (2013), *Outlier Analysis*, Springer, Berlin.

D. Aloise, A. Deshpande, P. Hansen, and P. Popat, (2009), *NP-hardness of Euclidean sum-of-squares clustering*, Mach. Learn., 75, pp. 245–248.

B. Ames (2014), *Guaranteed clustering and biclustering via semidefinite programming*, Math. Program., 147, pp. 429–465.

K. Anstreicher (2009), *Semidefinite programming versus the reformulation-linearization technique for nonconvex quadratically constrained quadratic programming*, J. Global Optim., 43, pp. 471–484.

D. Arthur and S. Vassilvitskii (2007), *k-means++: The advantages of careful seeding*, in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, PA, pp. 1027–1035.

P. Awasthi, A. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward (2015), *Relax, no need to round: Integrality of clustering formulations*, in Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, Association for Computing Machinery, New York, pp. 191–200.

M. Balcan, S. Ehrlich, and Y. Liang (2013), *Distributed k-means and k-median clustering on general topologies*, in Adv. Neural Inf. Process. Syst. 26, Curran Associates, Red Hook, NY, pp. 1995–2003.

A. Banerjee and J. Ghosh (2006), *Scalable clustering algorithms with balancing constraints*, in Chapman & Hall/CRC Data Min. Knowl. Discov. Ser. 13, CRC Press, Boca Raton, FL, pp. 365–395.

K. Bennett, P. Bradley, and A. Demiriz (2000), *Constrained K-means Clustering*, Technical Report, Microsoft Research.

A. Bertoni, M. Goldwurm, J. Lin, and F Saccà (2012), *Size constrained distance clustering: Separation properties and some complexity results*, Fund. Inform., 115, pp. 125–139.

S. Boyd and L. Vandenberghe (2004), *Convex Optimization*, Cambridge University Press, Cambridge.

R. Burkard (2013), *The quadratic assignment problem*, in Handbook of Combinatorial Optimization, P. Pardalos, D. Du, R. Graham, eds., Springer, New York, pp. 2741–2814.

R. Burkard, M. Dell'Amico, and S. Martello (2009), *Assignment Problems*, SIAM, Philadelphia, PA.

S. Chawla and A. Gionis (2013), *k-means−−: A unified approach to clustering and outlier detection*, in Proceedings of the 2013 SIAM International Conference on Data Mining, SIAM, Philadelphia, PA, pp. 189–197.

Y. Chen, Y. Zhang, and X. Ji (2006), *Size regularized cut for data clustering*, in Adv. Neural Inf. Process. Syst. 18, Curran Associates, MIT Press, Cambridge, MA, pp. 211–218.

W. Cook and A. Rohe (1999), *Computing minimum-weight perfect matchings*, INFORMS J. Comput., 11, pp. 138–148.

L. Costa, D. Aloise, and N. Mladenović (2017), *Less is more: Basic variable neighborhood search heuristic for balanced minimum sum-of-squares clustering*, Inform. Sci., 415, pp. 247–253.

E. Elhamifar, G. Sapiro, and R. Vidal (2012), *Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery*, in Adv. Neural Inf. Process. Syst. 25, Curran Associates, Red Hook, NY, pp. 19–27.

J. Gareth, D. Witten, T. Hastie, and R. Tibshirani (2017), *An Introduction to Statistical Learning with Applications in R*, Springer Texts Statist. 103, Springer, Berlin.

K. Gatermann and P. Parrilo (2004), *Symmetry groups, semidefinite programs, and sums of squares*, J. Pure Appl. Algebra, 192, pp. 95–128.

G. Golub and C. van Loan (1996), *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD.

S. Hasegawa, H. Imai, M. Inaba, N. Katoh, and J. Nakano (1993), *Efficient algorithms for variance-based k-clustering*, in Proceedings of the First Pacific Conference on Computer Graphics and Applications, S. Y. Shin and L. Kunii Tosiyasu, eds., World Scientific, River Edge, NJ, pp. 75–89.

T. Iguchi, D. Mixon, J. Peterson, and S. Villar (2017), *Probably certifiably correct k-means clustering*, Math. Program., 165, pp. 605–642.

M. Inaba, N. Katoh, and I. Hiroshi (1994), *Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering*, in Proceedings of the Tenth Annual Symposium on Computational Geometry, Association for Computing Machinery, New York, pp. 332–339.

A. Jain (2010), *Data clustering:* 50 *years beyond K-means*, Pattern Recogn. Lett., 31, pp. 651–666.

S. Lloyd (1982), *Least squares quantization in PCM*, IEEE Trans. Inf. Theory, 28, pp. 129–137.

M. Malinen and P. Fränti (2014), *Balanced K-means for clustering*, in Structural, Syntactic, and Statistical Pattern Recognition, P. Fränti, G. Brown, M. Loog, F. Escolano, and M. Pelillo, eds, Lecture Notes in Comput. Sci. 8621, Springer, Cham, pp. 32–41.

J. Mulvey and M. Beck (1984), *Solving capacitated clustering problems*, European J. Oper. Res., 18, pp. 339–348.

A. Nellore and R. Ward, (2015), *Recovery guarantees for exemplar-based clustering*, Inform. and Comput., 245, pp. 165–180.

J. Peng and Y. Wei (2007), *Approximating K-means-type clustering via semidefinite programming*, SIAM J. Optim., 18, pp. 186–205.

A. Pyatkin, D. Aloise, and N. Mladenović (2017), *NP-hardness of balanced minimum sum-of-squares clustering*, Pattern Recogn. Lett., 97, pp. 44–45.

R. Vinayak and B. Hassibi (2016), *Similarity clustering in the presence of outliers: Exact recovery via convex program*, in Proceedings of the 2016 IEEE International Symposium on Information Theory, IEEE Press, Piscataway, NJ, pp. 91–95.

H. Zha, X. He, C. Ding, H. Simon, and M. Gu (2002), *Spectral relaxation for K-means clustering*, in Adv. Neural Inf. Process. Syst. 14, Curran Associates, MIT Press, Cambridge, MA, pp. 1057–1064.

H. Zhang, C. Royo, and L. Ma (2013), *Solving the quadratic assignment problem by means of general purpose mixed integer linear programming solvers*, Ann. Oper. Res., 207, pp. 261–278.