Operations Research Letters 50 (2022) 205–212

Contents lists available at ScienceDirect

Operations Research Letters

www.elsevier.com/locate/orl

Optimality-based clustering: An inverse optimization approach

Zahed Shahmoradi, Taewoo Lee *

Department of Industrial Engineering, University of Houston, Houston, TX 77204, USA

ARTICLE INFO

Article history:
Received 23 July 2021
Received in revised form 19 November 2021
Accepted 26 December 2021
Available online 3 January 2022

Keywords:
Inverse optimization
Inverse linear programming
Clustering

ABSTRACT

We propose a new clustering approach, called optimality-based clustering, that clusters data points based on their latent decision-making preferences. We assume that each data point is a decision generated by a decision-maker who (approximately) solves an optimization problem and cluster the data points by identifying a common objective function of the optimization problems for each cluster such that the worst-case optimality error is minimized. We propose three different clustering models and test them in the diet recommendation application.

1. Introduction

Clustering is a technique that groups objects (e.g., data points) into clusters such that the objects in the same cluster are more similar to each another than to those in other clusters based on some similarity measure [9]. Most clustering approaches fall into one of the following categories: centroid-based clustering, distribution-based clustering, and density-based clustering. In centroid-based clustering, each object is assigned to a cluster based on its similarity to a representative object called a centroid (e.g., K-means clustering) [10,13,17]. In density-based clustering, a density measure, e.g., the number of objects within a certain distance, is used to detect areas with high density, in which the objects are grouped into the same cluster [7,12]. Distribution-based clustering groups the objects based on whether or not they belong to the same distribution [18].

Often, data points correspond to decisions generated by decision-makers (DMs) who are assumed to solve some kind of decision-making problems (DMPs). Although traditional clustering approaches for such decision data may indicate which decisions are similar to each other, this similarity does not necessarily imply that the DMs whose decisions are in the same cluster have similar preferences. For example, suppose the DMPs can be formulated as optimization problems where the DM's preferences are encoded in the objective function parameters. Even when two DMs' decisions are geometrically close to each other, they might have been generated by two DMPs with completely different objective function parameters under different feasible regions, which traditional clustering cannot capture. The focus of this paper is to cluster decision data based on the similarity in the DM's decision-making preferences, captured by parameters in their underlying DMPs.

Clustering based on decision-making preferences can help create targeted, group-based decision support tools. For example, by clustering patients based on their health-related preferences (e.g., health benefit vs. cost saving) using their past disease screening decisions, one can create a group-based yet easily implementable screening guideline that is consistent with the patients' preferences (e.g., increased use of telemedicine for a specific group of patients). Similarly, when developing a diet recommendation system, clustering individuals based on their food preferences and inferring a common objective function for each cluster can help create a group-specific diet recommendation framework. A post-hoc analysis can be done to further identify association of the preference clusters with other factors such as health conditions and socio-demographic factors.

Since this clustering problem requires inferring objective function parameters of the DMPs from decision data, it inherently involves inverse optimization. Given an observed decision from a DM who solves an optimization problem, inverse optimization infers parameters of the problem that make the decision as optimal as possible (e.g., [1,2,4–6,11]). Solving the DM's optimization problem with these inferred parameters then leads to a decision that is close to the observed one. Previous inverse optimization models assume that decision

* Corresponding author.
E-mail address: tlee6@uh.edu (T. Lee).

https://doi.org/10.1016/j.orl.2021.12.012
0167-6377/

data is collected from either a single DM or a group of DMs whose preferences are known to be similar, for which the same, single set of parameters is inferred [2,3,6,15].

In this paper, we develop a new clustering approach that clusters decision data (hence DMs) based on their latent decision-making preferences. In particular, inspired by inverse optimization, we propose the clustering problem that simultaneously groups observed decisions into clusters and finds an objective function for each cluster such that the decisions in the same cluster are rendered as optimal as possible for the assumed DMPs. We use optimality errors associated with the decisions with respect to the inferred objective function as a measure of similarity; hence we call this problem "optimality-based clustering." We further enhance the problem by incorporating the notion of cluster stability, measured for each cluster by the worst-case distance between the decision data in the cluster and optimal decisions achieved by the DMPs using the inferred objective function for the cluster. The stability-driven, optimality-based clustering problem is computationally challenging. We derive mixed-integer programs (MIPs) that provide upper and lower bound solutions for the true clustering problem as well as heuristics that approximately solve this problem. Finally, we demonstrate the proposed clustering approach in the diet recommendation application to cluster individuals based on their food preferences. Unless otherwise stated, proofs are in the appendix.

2. Preliminaries

In this section, we present an initial formulation for the optimality-based clustering problem and a simple example to demonstrate the idea. We then define the notion of cluster stability in the context of optimality-based clustering, which we later use to propose an enhanced clustering formulation.

2.1. A general clustering problem

We focus on a centroid-based clustering problem where the similarity of a data point to a cluster is assessed by the distance between the data point and a centroid of the cluster. Given a dataset $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}^1, \dots, \hat{\mathbf{x}}^K\}$ with the index set $\mathcal{K} = \{1, \dots, K\}$, let $\{\mathcal{G}^\ell\}_{\ell\in\mathcal{L}}$ be a collection of L clusters where $\mathcal{G}^\ell \subseteq \mathcal{K}$ and $\mathcal{L} = \{1, \dots, L\}$. For each cluster $\mathcal{G}^\ell$, the *dissimilarity* among the members of the cluster is measured by $\sum_{k\in\mathcal{G}^\ell} d(\hat{\mathbf{x}}^k, \mathbf{x}^\ell)$, where $\mathbf{x}^\ell$ denotes the centroid of the cluster and $d(\hat{\mathbf{x}}^k, \mathbf{x}^\ell)$ represents the distance between observation $\hat{\mathbf{x}}^k$ and its cluster centroid $\mathbf{x}^\ell$, e.g., $d(\hat{\mathbf{x}}^k, \mathbf{x}^\ell) = \|\hat{\mathbf{x}}^k - \mathbf{x}^\ell\|_r$ for some $r \geq 1$. Based on the above definition, a centroid-based clustering problem seeks clusters $\{\mathcal{G}^\ell\}_{\ell\in\mathcal{L}}$ such that the sum of dissimilarities over all clusters is minimized, i.e.,

$$\underset{\{\mathcal{G}^\ell\}_{\ell\in\mathcal{L}}, \{\mathbf{x}^\ell\}_{\ell\in\mathcal{L}}}{\text{minimize}} \quad \sum_{\ell\in\mathcal{L}} \sum_{k\in\mathcal{G}^\ell} d(\hat{\mathbf{x}}^k, \mathbf{x}^\ell).$$

2.2. Optimality-based clustering: the initial model

We assume that each data point $\hat{\mathbf{x}}^k \in \hat{\mathcal{X}}$ is an observed decision from DM k (denoted by DM_k) who approximately solves the following optimization problem as a decision-making problem (DMP_k):

$$\mathbf{DMP}_k(\mathbf{c}): \quad \underset{\mathbf{x}}{\text{minimize}} \{\mathbf{c}'\mathbf{x} \mid \mathbf{A}^k\mathbf{x} \geq \mathbf{b}^k\},$$

where $\mathbf{c} \in \mathbb{R}^n, \mathbf{x} \in \mathbb{R}^n$, $\mathbf{A}^k \in \mathbb{R}^{m_k \times n}$, and $\mathbf{b}^k \in \mathbb{R}^{m_k}$, for each $k \in \mathcal{K}$. For each DM $k \in \mathcal{K}$, let $\mathcal{I}^k = \{1, \dots, m_k\}$ and $\mathcal{J} = \{1, \dots, n\}$ index the constraints and variables of DMP_k, and $\mathbf{a}^{ki} \in \mathbb{R}^n$ be a (column) vector corresponding to the i-th row of $\mathbf{A}^k$. We let $\mathcal{X}^k$ be the set of feasible solutions for DMP_k, assumed bounded, full-dimensional, and free of redundant constraints, and $\mathcal{X}^{ki} = \{\mathbf{x} \in \mathcal{X}^k \mid \mathbf{a}^{ki\prime}\mathbf{x} = b_i^k\}$, $i \in \mathcal{I}^k$. Let $\mathcal{X}'^{k*}(\mathbf{c}) = \operatorname{argmin} \mathbf{DMP}_k(\mathbf{c})$. Without loss of generality, we assume that each $\mathbf{a}^{ki}$ is normalized *a priori* such that $\|\mathbf{a}^{ki}\|_1 = 1$.

Given a set of observed decisions $\hat{\mathcal{X}}$, the goal of optimality-based clustering is to group the observations into $L < K$ clusters $\{\mathcal{G}^\ell\}_{\ell\in\mathcal{L}}$ and find a cost vector $\mathbf{c}^\ell$ for each cluster ℓ such that each observation $\hat{\mathbf{x}}^k$ in cluster ℓ (i.e., for $k \in \mathcal{G}^\ell$) is as close as possible to an optimal solution to $\text{DMP}_k(\mathbf{c}^\ell)$. This problem can be formulated as follows:

$$\underset{\{\mathbf{x}^k\}_{k\in\mathcal{K}}, \{(\mathbf{c}^\ell, \mathcal{G}^\ell)\}_{\ell\in\mathcal{L}}}{\text{minimize}} \quad d(\hat{\mathcal{X}}, \{\mathbf{x}^k\}_{k\in\mathcal{K}}) \tag{1a}$$

$$\text{subject to} \quad \mathbf{x}^k \in \mathcal{X}'^{k*}(\mathbf{c}^\ell), \quad \forall k \in \mathcal{G}^\ell, \ell \in \mathcal{L}, \tag{1b}$$

$$\|\mathbf{c}^\ell\|_1 = 1, \quad \forall \ell \in \mathcal{L}. \tag{1c}$$

The objective of the above problem is to minimize the distance between the observations $\hat{\mathcal{X}}$ and solutions $\mathbf{x}^k$'s that are optimal for their respective DMPs with respect to $\mathbf{c}^\ell$ for $k \in \mathcal{G}^\ell$; e.g., $d(\hat{\mathcal{X}}, \{\mathbf{x}^k\}_{k\in\mathcal{K}}) = \sum_{\ell\in\mathcal{L}} \sum_{k\in\mathcal{G}^\ell} \|\hat{\mathbf{x}}^k - \mathbf{x}^k\|$. Constraint (1c) prevents the trivial solution $\mathbf{c}^\ell = \mathbf{0}$ from being feasible. Note that the above problem is analogous to centroid-based clustering problems in that $\mathbf{c}^\ell$ can be seen as the centroid of cluster ℓ, representing the shared decision preference of the observations assigned to cluster ℓ. We use the following simple example to demonstrate the idea.

Example 1. Suppose three DMs solve the following problem with their own objective functions:

$$\underset{x_1, x_2}{\text{maximize}} \{c_1x_1 + c_2x_2 \mid x_1 \leq b_1,\ x_2 \leq b_2,\ x_1, x_2 \geq 0\}.$$

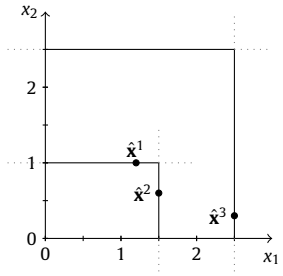

Fig. 1. Observations from DMs $k = 1, 2$, and 3 and their respective feasible regions.

Let $(b_1, b_2) = (1.5, 1)$ for DMs $k = 1, 2$ and $(b_1, b_2) = (2.5, 2.5)$ for DM $k = 3$ (see Fig. 1 for the feasible regions). We assume the following decisions are observed from the DMs: $\hat{\mathbf{x}}^1 = \begin{bmatrix} 1.2 \\ 1 \end{bmatrix}$, $\hat{\mathbf{x}}^2 = \begin{bmatrix} 1.5 \\ 0.6 \end{bmatrix}$, and $\hat{\mathbf{x}}^3 = \begin{bmatrix} 2.5 \\ 0.3 \end{bmatrix}$ (see Fig. 1). If the desired number of clusters is two (i.e., $L = 2$), traditional K-means clustering based on the Euclidean distance finds $\{\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2\}$ and $\{\hat{\mathbf{x}}^3\}$ to be optimal clusters. However, if the goal is to group the decisions based on the preferences encoded in the corresponding DMPs, clustering should be done differently. In particular, given their respective feasible regions, $\hat{\mathbf{x}}^2$ and $\hat{\mathbf{x}}^3$ share the same preference as they are optimal for their respective DMPs based on the same cost vector $\mathbf{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$; on the other hand, $\hat{\mathbf{x}}^1$ is optimal to the DMP with respect to a completely different cost vector $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. As a result, an optimal clustering is $\{\hat{\mathbf{x}}^2, \hat{\mathbf{x}}^3\}$ and $\{\hat{\mathbf{x}}^1\}$.

2.3. Cluster instability

In this subsection, we show that the initial model (1) is often subject to an instability issue due to the structure of the DMP formulation and propose a measure of instability in the context of optimality-based clustering. Given an optimal cost vector $\mathbf{c}^{\ell *}$ for some cluster ℓ achieved by model (1), we note that $\mathbf{DMP}_k(\mathbf{c}^{\ell *})$ often leads to $\mathbf{x}^k \in \mathcal{X}^{k*}(\mathbf{c}^\ell)$ that is far from the observations assigned to the cluster. For illustration, consider the same example in Fig. 1, where model (1) finds $\mathcal{G}_2 = \{\hat{\mathbf{x}}^1\}$ (i.e., $\hat{\mathbf{x}}^1$ assigned to cluster $\ell = 2$) and $\mathbf{c}^{2*} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. While the desirable forward optimal solution with respect to this cost vector is supposed to be close to $\hat{\mathbf{x}}^1$, solving $\text{DMP}_1(\mathbf{c}^{2*})$ can lead to an optimal solution $\mathbf{x}^* = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, which is far from $\hat{\mathbf{x}}^1$. Note that this cluster instability issue is different from the cluster assignment instability issues considered in the traditional clustering literature [14,16]; it is rather associated with the argmin set of the DMP for a certain cost vector. This type of instability is also discussed in Shahmoradi and Lee [15] in the context of inverse linear programming.

We now formally define a notion of *cluster stability*, which we then use to propose an enhanced clustering problem formulation that improves on the initial model (1) in the next section. Given that the instability issue is caused by $\mathbf{x} \in \mathcal{X}^{k*}(\mathbf{c})$ being too far from $\hat{\mathbf{x}}^k$, we assess the instability of a cost vector $\mathbf{c}$ associated with each $\hat{\mathbf{x}}^k$ via the worst-case distance between $\hat{\mathbf{x}}^k$ and $\mathcal{X}^{k*}(\mathbf{c})$:

$$\max\{d(\hat{\mathbf{x}}^k, \mathbf{x}^k) \,|\, \mathbf{x}^k \in \mathcal{X}^{k*}(\mathbf{c})\}.$$

Then, the instability of cluster $\mathcal{G}^\ell$ with its cost vector $\mathbf{c}^\ell$ is measured by the following measure:

$$\max_{k \in \mathcal{G}^\ell} \max\{d(\hat{\mathbf{x}}^k, \mathbf{x}^k) \,|\, \mathbf{x}^k \in \mathcal{X}^{k*}(\mathbf{c}^\ell)\}. \tag{2}$$

In other words, cluster ℓ is said to be more stable if its cost vector $\mathbf{c}^\ell$ leads to a smaller worst-case distance between $\hat{\mathbf{x}}^k$ and the set of optimal solutions for $\text{DMP}_k(\mathbf{c}^\ell)$ over all data points in the cluster. For brevity, from here on out we combine the two max terms in (2) and simply write it as $\max_{k \in \mathcal{G}^\ell}\{d(\hat{\mathbf{x}}^k, \mathbf{x}^k) \,|\, \mathbf{x}^k \in \mathcal{X}^{k*}(\mathbf{c}^\ell)\}$.

3. Models

In this section, we first propose an enhanced optimality-based clustering problem that addresses the cluster instability issue by incorporating the stability measure (2). We also propose two heuristics that approximately solve the problem by separating it into two stages: the clustering stage and the cost vector inference stage. We then analytically compare the performances of these approaches.

3.1. The stability-driven clustering model

To address the cluster instability issue in model (1), we replace its objective function with the stability-incorporated dissimilarity measure in (2). This leads to the following, which we call the stability-driven clustering (SC) model:

$$\mathbf{SC}(\mathcal{K}, L): \quad \underset{\{(\mathbf{c}^\ell, \mathcal{G}^\ell)\}_{\ell \in \mathcal{L}}}{\text{minimize}} \quad \max_{\ell \in \mathcal{L}} \max_{k \in \mathcal{G}^\ell} \{d(\hat{\mathbf{x}}^k, \mathbf{x}^k)\} \tag{3a}$$

$$\text{subject to} \quad \mathbf{x}^k \in \mathcal{X}^{k*}(\mathbf{c}^\ell), \quad \forall \ell \in \mathcal{L}, k \in \mathcal{G}^\ell, \tag{3b}$$

$$\|\mathbf{c}^\ell\|_1 = 1, \quad \forall \ell \in \mathcal{L}, \tag{3c}$$

where now the objective is to maximize stability for all clusters by minimizing the worst-case distance between $\hat{\mathbf{x}}^k$ and the argmin set $\mathcal{X}^{k*}(\mathbf{c}^\ell)$ over all observations and clusters. Since each DMP_k is a linear program (LP), we utilize the LP optimality conditions to reformulate model (3) as follows:

$$\underset{\{(\mathbf{c}^\ell,\mathcal{G}^\ell)\}_{\ell\in\mathcal{L}},\{(\mathbf{x}^k,\mathbf{y}^k)\}_{k\in\mathcal{K}}}{\text{minimize}} \quad \max_{\ell\in\mathcal{L}} \max_{k\in\mathcal{G}^\ell} \{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\} \tag{4a}$$

$$\text{subject to} \quad \mathbf{A}^{k\prime}\mathbf{y}^k = \mathbf{c}^\ell, \quad \forall \ell\in\mathcal{L}, k\in\mathcal{G}^\ell, \tag{4b}$$

$$\mathbf{y}^k \geq \mathbf{0}, \quad \forall k\in\mathcal{K}, \tag{4c}$$

$$\mathbf{A}^k\mathbf{x}^k \geq \mathbf{b}^k, \quad \forall k\in\mathcal{K}, \tag{4d}$$

$$\mathbf{c}^{\ell\prime}\mathbf{x}^k = \mathbf{b}^{k\prime}\mathbf{y}^k, \quad \forall \ell\in\mathcal{L}, k\in\mathcal{G}^\ell, \tag{4e}$$

$$\|\mathbf{c}^\ell\|_1 = 1, \quad \forall \ell\in\mathcal{L}. \tag{4f}$$

Constraints (4b)–(4e) represent the LP optimality conditions for solutions $\{\mathbf{x}^k\}_{k\in\mathcal{G}^\ell}$ with respect to $\mathbf{c}^\ell$ for each cluster $\ell\in\mathcal{L}$: constraints (4b)–(4c) enforce dual feasibility where $\mathbf{y}^k\in\mathbb{R}^{m_k}$ represents the vector of dual variables corresponding to DMP_k, constraint (4d) corresponds to primal feasibility, and constraint (4e) ensures strong duality. Note that problem (4) is non-convex due to its objective function and constraints (4e) and (4f). In Section 4, we analyze its solution structure and propose MIP formulations that provide lower and upper bounds on the optimal objective value of problem (4). Our subsequent analysis for the rest of this paper focuses on $d(\hat{\mathbf{x}},\mathbf{x}) = \|\hat{\mathbf{x}}-\mathbf{x}\|_r$ for $r\geq 1$, though similar analysis can be derived for other distance functions.

3.2. Heuristics: two-stage approaches

While (4) provides an exact reformulation of the SC problem, it is computationally challenging. Instead, one naive view on this problem would be to treat the clustering and cost vector inference parts separately. In this subsection, we propose two heuristics based on this separation idea.

The first algorithm applies traditional K-means clustering to cluster dataset $\hat{\mathcal{X}}$ *a priori* based on some distance function, e.g., Euclidean distance, followed by applying inverse optimization post-hoc to derive a cost vector for each of the predetermined clusters. This approach, which we call the cluster-then-inverse (CI) approach, can be written as follows.

$$\mathbf{CI}(\mathcal{K},L): \begin{cases} \text{Stage 1.} & \text{Find } \{\mathcal{G}^\ell_{\text{CI}}\}_{\ell\in\mathcal{L}} \in \underset{\{\mathcal{G}^\ell\}_{\ell\in\mathcal{L}}}{\text{argmin}} \left\{ \sum_{\ell\in\mathcal{L}}\sum_{k\in\mathcal{G}^\ell} d(\hat{\mathbf{x}}^k,\mathbf{x}^\ell_{\text{cen}}) \,\middle|\, \mathbf{x}^\ell_{\text{cen}} \text{ is the centroid of } \mathcal{G}^\ell \right\} \\ \text{Stage 2.} & \text{Find } \mathbf{c}^\ell_{\text{CI}} \in \underset{\mathbf{c}^\ell}{\text{argmin}} \left\{ \max_{k\in\mathcal{G}^\ell_{\text{CI}}} d(\hat{\mathbf{x}}^k,\mathbf{x}^k) \,\middle|\, \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell), \|\mathbf{c}^\ell\|_1 = 1 \right\}, \ \forall \ell\in\mathcal{L}. \end{cases} \tag{5}$$

In Stage 1, clusters $\{\mathcal{G}^\ell_{\text{CI}}\}_{\ell\in\mathcal{L}}$ are obtained by solving a traditional clustering problem on $\hat{\mathcal{X}}$. Then, Stage 2 finds a cost vector for each of the clusters that minimizes cluster instability. Note that Stage 2 of the CI approach solves a "reduced" version of the SC problem that finds a stability-maximizing cost vector for each $\ell\in\mathcal{L}$ with respect to the observations assigned to cluster ℓ; i.e., $\mathbf{SC}(\mathcal{G}^\ell_{\text{CI}}, L=1)$ where $L=1$ implies that no further clustering happens.

Alternatively, the second approach finds a cost vector $\mathbf{c}^{k*}$ for each data point $k\in\mathcal{K}$ *a priori* such that $\max\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k) \,|\, \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^{k*})\}$ is minimized. Then, the cost vectors are clustered post-hoc into L groups via traditional clustering. We call this approach inverse-then-cluster (IC):

$$\mathbf{IC}(\mathcal{K},L): \begin{cases} \text{Stage 1.} & \text{Find } \mathbf{c}^{k*} \in \underset{\mathbf{c}^k}{\text{argmin}} \left\{ \max\ d(\hat{\mathbf{x}}^k,\mathbf{x}^k) \,\middle|\, \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^k), \ \|\mathbf{c}^k\|_1 = 1 \right\}, \ \forall k\in\mathcal{K} \\ \text{Stage 2.} & \text{Find } \{\mathcal{G}^\ell_{\text{IC}}\}_{\ell\in\mathcal{L}} \in \underset{\{\mathcal{G}^\ell\}_{\ell\in\mathcal{L}}}{\text{argmin}} \left\{ \sum_{\ell\in\mathcal{L}}\sum_{k\in\mathcal{G}^\ell} d(\mathbf{c}^{k*},\mathbf{c}^\ell_{\text{cen}}) \,\middle|\, \mathbf{c}^\ell_{\text{cen}} \text{ is the centroid of } \mathcal{G}^\ell \right\}. \end{cases} \tag{6}$$

Note that, similarly, Stage 1 of the above IC approach can be seen as solving a reduced version of SC, i.e., $\mathbf{SC}(\{k\}, L=1)$, which finds a "per-observation" cost vector $\mathbf{c}^{k*}$ that maximizes stability associated with each observation $\hat{\mathbf{x}}^k$. However, once the cost vectors are clustered in Stage 2, it is the resulting centroid cost vector, i.e., $\mathbf{c}^\ell_{\text{cen}}$, that represents the preferences for the observations assigned to cluster ℓ, which does not necessarily retain the same level of stability achieved by the per-observation cost vectors (i.e., $\mathbf{c}^{k*}$'s) in Stage 1. To address this, once the clustering is done, one may solve the SC problem for each cluster again to find a "corrected" cost vector; i.e., $\mathbf{SC}(\mathcal{G}^\ell_{\text{IC}}, L=1)$ for each $\ell\in\mathcal{L}$. We denote such a post-processed cost vector by $\mathbf{c}^\ell_{\text{IC}}$, $\ell\in\mathcal{L}$.

3.3. Model comparison

Next, we compare the performance of the SC model (i.e., (3)) and the CI and IC approaches.

Proposition 1. *Given $\hat{\mathcal{X}}$, let $\{\mathcal{G}^\ell_{\text{SC}}, \mathbf{c}^\ell_{\text{SC}}\}_{\ell\in\mathcal{L}}$ denote an optimal solution to model (3), and $\{\mathcal{G}^\ell_{\text{CI}}, \mathbf{c}^\ell_{\text{CI}}\}_{\ell\in\mathcal{L}}$ and $\{\mathcal{G}^\ell_{\text{IC}}, \mathbf{c}^\ell_{\text{IC}}\}_{\ell\in\mathcal{L}}$ be the clusters and corresponding cost vectors achieved by the CI and IC approaches, respectively. Then we have*

(i) $\max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{SC}}} \{d(\hat{\mathbf{x}}^k,\mathbf{x}^k) \,|\, \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{SC}})\} \leq \max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{CI}}} \{d(\hat{\mathbf{x}}^k,\mathbf{x}^k) \,|\, \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{CI}})\}$, *and*

(ii) $\max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{SC}}} \{d(\hat{\mathbf{x}}^k,\mathbf{x}^k) \,|\, \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{SC}})\} \leq \max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{IC}}} \{d(\hat{\mathbf{x}}^k,\mathbf{x}^k) \,|\, \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{IC}})\}$.

Proof. Since $\{\mathcal{G}^\ell_{\text{SC}}, \mathbf{c}^\ell_{\text{SC}}\}_{\ell\in\mathcal{L}}$ is an optimal solution to (3), we have

$$\max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{SC}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{SC}})\} = \min_{\{(\mathbf{c}^\ell,\mathcal{G}^\ell)\}_{\ell\in\mathcal{L}}}\left\{\max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell)\}\,\middle|\, \|\mathbf{c}^\ell\|_1=1\right\},$$

where the right hand side corresponds to model (3).

For part (i), consider $\{\mathcal{G}^\ell_{\text{CI}}, \mathbf{c}^\ell_{\text{CI}}\}_{\ell\in\mathcal{L}}$ generated by CI. Recall from (5) that $\mathbf{c}^\ell_{\text{CI}}\in\operatorname{argmin}_{\mathbf{c}^\ell}\left\{\max_{k\in\mathcal{G}^\ell_{\text{CI}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell)\}\,\middle|\, \|\mathbf{c}^\ell\|_1=1\right\}$ for the given cluster $\mathcal{G}^\ell_{\text{CI}}$ for each ℓ. Thus, $\max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{CI}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{CI}})\} = \max_{\ell\in\mathcal{L}}\min_{\mathbf{c}^\ell}\left\{\max_{k\in\mathcal{G}^\ell_{\text{CI}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell)\}\,\middle|\, \|\mathbf{c}^\ell\|_1=1\right\}$. Then it follows that

$$\begin{aligned}&\max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{SC}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{SC}})\} = \min_{\{(\mathbf{c}^\ell,\mathcal{G}^\ell)\}_{\ell\in\mathcal{L}}}\left\{\max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell)\}\,\middle|\, \|\mathbf{c}^\ell\|_1=1\right\}\\ &\le \max_{\ell\in\mathcal{L}}\min_{\mathbf{c}^\ell}\left\{\max_{k\in\mathcal{G}^\ell_{\text{CI}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell)\}\,\middle|\, \|\mathbf{c}^\ell\|_1=1\right\} = \max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{CI}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{CI}})\},\end{aligned}$$

as desired.

Proof for part (ii) is similar. Consider $\{\mathcal{G}^\ell_{\text{IC}}, \mathbf{c}^\ell_{\text{IC}}\}_{\ell\in\mathcal{L}}$ generated by IC. Recall from (6) and its post-processing step that $\mathbf{c}^\ell_{\text{IC}}\in\operatorname{argmin}_{\mathbf{c}^\ell}\left\{\max_{k\in\mathcal{G}^\ell_{\text{IC}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell)\}\,\middle|\, \|\mathbf{c}^\ell\|_1=1\right\}$. That is, $\max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{IC}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{IC}})\} = \max_{\ell\in\mathcal{L}}\min_{\mathbf{c}^\ell}\left\{\max_{k\in\mathcal{G}^\ell_{\text{IC}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell)\}\,\middle|\, \|\mathbf{c}^\ell\|_1=1\right\}$. Thus, we have

$$\begin{aligned}&\max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{SC}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{SC}})\} = \min_{\{(\mathbf{c}^\ell,\mathcal{G}^\ell)\}_{\ell\in\mathcal{L}}}\left\{\max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell)\}\,\middle|\, \|\mathbf{c}^\ell\|_1=1\right\}\\ &\le \max_{\ell\in\mathcal{L}}\min_{\mathbf{c}^\ell}\left\{\max_{k\in\mathcal{G}^\ell_{\text{IC}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell)\}\,\middle|\, \|\mathbf{c}^\ell\|_1=1\right\} = \max_{\ell\in\mathcal{L},k\in\mathcal{G}^\ell_{\text{IC}}}\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c}^\ell_{\text{IC}})\},\end{aligned}$$

as desired. □

While Proposition 1 implies that the SC model performs at least as well as CI and IC in terms of stability, the SC model is typically computationally more challenging than CI and IC. In the next section, we analyze the solution structure of the SC model, which we use to derive MIP formulations that provide lower and upper bounds on the optimal value of the SC model.

4. Solution structure and bounds

The reformulation of the SC model (i.e., (4)) is non-convex due to the normalization constraint (4f) as well as the objective function: for a given $k\in\mathcal{K}$ and arbitrary $\mathbf{c}$, $\max\{d(\hat{\mathbf{x}}^k,\mathbf{x}^k)\mid \mathbf{x}^k\in\mathcal{X}^{k*}(\mathbf{c})\}$ is a maximization of the convex function d over the convex region $\mathcal{X}^{k*}(\mathbf{c})$. Both the CI and IC approaches also face the same computational challenges because they also involve solving the SC formulations albeit of smaller size; i.e., Stage 2 of the CI approach solves $\mathbf{SC}(\mathcal{G}^\ell_{\text{CI}}, L=1)$ for each $\ell\in\mathcal{L}$ and Stage 1 of the IC approach solves $\mathbf{SC}(\{k\}, L=1)$ for each $k\in\mathcal{K}$. In this section, we analyze the solution structure of the SC model, which leads to MIP formulations that provide lower and upper bound solutions for the SC problem.

Theorem 2. *There exists an optimal solution* $\left(\{(\mathbf{c}^{\ell*},\mathcal{G}^{\ell*})\}_{\ell\in\mathcal{L}}, \{(\mathbf{x}^{k*},\mathbf{y}^{k*})\}_{k\in\mathcal{K}}\right)$ *to* (4) *such that for each cluster* $\ell\in\mathcal{L}$*:*

(i) $\mathbf{a}^{ki\prime}\mathbf{x}^{k*} = b^k_i$ *for* $i\in\mathcal{I}^{k*}\subseteq\mathcal{I}^k$ *where* $|\mathcal{I}^{k*}| = n$ *for all* $k\in\mathcal{G}^{\ell*}$*, and*
(ii) $\mathbf{c}^{\ell*}\in\text{cone}(\{\mathbf{a}^{ki}\}_{i\in\mathcal{I}^{k*}})$ *for all* $k\in\mathcal{G}^\ell$ *where* $\text{cone}(\cdot)$ *denotes the conic hull of the given vectors, i.e.,* $\text{cone}(\{\mathbf{a}^{ki}\}_{i\in\mathcal{I}^{k*}}) = \{\sum_{i\in\mathcal{I}^{k*}}\gamma_i\mathbf{a}^{ki}\mid \gamma_i\ge 0\}$.

Proof. Consider an optimal solution $\left(\{(\mathbf{c}^{\ell*},\mathcal{G}^{\ell*})\}_{\ell\in\mathcal{L}}, \{(\mathbf{x}^{k*},\mathbf{y}^{k*})\}_{k\in\mathcal{K}}\right)$ to (4). Due to constraints (4b)–(4e), we have $\mathbf{x}^{k*}\in\mathcal{X}^{k*}(\mathbf{c}^{\ell*})$ for each $\ell\in\mathcal{L}$ and $k\in\mathcal{G}^{\ell*}$. Note that any point in $\mathcal{X}^{k*}(\mathbf{c}^{\ell*})$ can be represented by a convex combination of extreme points of $\mathcal{X}^{k*}(\mathbf{c}^{\ell*})$. Let $\text{ext}(\mathcal{X}^{k*}(\mathbf{c}^{\ell*}))$ be the set of extreme points of $\mathcal{X}^{k*}(\mathbf{c}^{\ell*})$, $Q_k = \left|\text{ext}(\mathcal{X}^{k*}(\mathbf{c}^{\ell*}))\right|$, and $\mathcal{Q}^k = \{1,\dots,Q_k\}$, i.e., $\text{ext}(\mathcal{X}^{k*}(\mathbf{c}^{\ell*})) = \{\bar{\mathbf{x}}^1,\dots,\bar{\mathbf{x}}^{Q_k}\}$, for each $k\in\mathcal{K}$. Then, there exists $\bar{\boldsymbol{\lambda}}\in\mathbb{R}^{Q_k}_+$ such that $\mathbf{x}^{k*} = \sum_{q_k\in\mathcal{Q}^k}\bar{\lambda}_{q_k}\bar{\mathbf{x}}^{q_k}$ and $\sum_{q_k\in\mathcal{Q}^k}\bar{\lambda}_{q_k} = 1$.

Now we prove part (i). Let $q^*_k\in\operatorname{argmax}_{q_k\in\mathcal{Q}^k}\{\|\hat{\mathbf{x}}^k-\bar{\mathbf{x}}^{q_k}\|_r\}$. That is, we have $\|\hat{\mathbf{x}}^k-\bar{\mathbf{x}}^{q^*_k}\|_r\ge\|\hat{\mathbf{x}}^k-\bar{\mathbf{x}}^{q_k}\|_r$ for all $q_k\in\mathcal{Q}^k$. Multiplying both sides of the inequality by $\bar{\lambda}_{q_k}$ yields $\bar{\lambda}_{q_k}\|\hat{\mathbf{x}}^k-\bar{\mathbf{x}}^{q^*_k}\|_r\ge\bar{\lambda}_{q_k}\|\hat{\mathbf{x}}^k-\bar{\mathbf{x}}^{q_k}\|_r$ for all $q_k\in\mathcal{Q}^k$, and thus $\sum_{q_k\in\mathcal{Q}^k}\bar{\lambda}_{q_k}\|\hat{\mathbf{x}}^k-\bar{\mathbf{x}}^{q^*_k}\|_r\ge\sum_{q_k\in\mathcal{Q}^k}\bar{\lambda}_{q_k}\|\hat{\mathbf{x}}^k-\bar{\mathbf{x}}^{q_k}\|_r$. Note that, from $\sum_{q_k\in\mathcal{Q}^k}\bar{\lambda}_{q_k}=1$ we have $\sum_{q_k\in\mathcal{Q}^k}\bar{\lambda}_{q_k}\|\hat{\mathbf{x}}^k-\bar{\mathbf{x}}^{q^*_k}\|_r = \|\hat{\mathbf{x}}^k-\bar{\mathbf{x}}^{q^*_k}\|_r$. This leads to

$$\|\hat{\mathbf{x}}^k - \bar{\mathbf{x}}^{q_k^*}\|_r \geq \sum_{q_k \in \mathcal{Q}^k} \bar{\lambda}_{q_k} \|\hat{\mathbf{x}}^k - \bar{\mathbf{x}}^{q_k}\|_r = \sum_{q_k \in \mathcal{Q}^k} \|\bar{\lambda}_{q_k}\hat{\mathbf{x}}^k - \bar{\lambda}_{q_k}\bar{\mathbf{x}}^{q_k}\|_r$$
$$\geq \|\sum_{q_k \in \mathcal{Q}^k} \bar{\lambda}_{q_k}\hat{\mathbf{x}}^k - \sum_{q_k \in \mathcal{Q}^k} \bar{\lambda}_{q_k}\bar{\mathbf{x}}^{q_k}\|_r = \|\hat{\mathbf{x}}^k - \sum_{q_k \in \mathcal{Q}^k} \bar{\lambda}_{q_k}\bar{\mathbf{x}}^{q_k}\|_r = \|\hat{\mathbf{x}}^k - \mathbf{x}^{k*}\|_r,$$

where the second inequality holds due to Minkowski inequalities. Also, from the optimality of $\mathbf{x}^{k*}$, we have $\|\hat{\mathbf{x}}^k - \bar{\mathbf{x}}^{q_k^*}\|_r \leq \|\hat{\mathbf{x}}^k - \mathbf{x}^{k*}\|_r$. Thus, it must be that $\|\hat{\mathbf{x}}^k - \bar{\mathbf{x}}^{q_k^*}\|_r = \|\hat{\mathbf{x}}^k - \mathbf{x}^{k*}\|_r$. This means that the solution $\big(\{(\mathbf{c}^{\ell*}, \mathcal{G}^{\ell*})\}_{\ell\in\mathcal{L}}, \{(\bar{\mathbf{x}}^{q_k^*}, \mathbf{y}^{k*})\}_{k\in\mathcal{K}}\big)$ is also optimal to (4). Since $\bar{\mathbf{x}}^{q_k^*}$ is an extreme point, there must exist $\mathcal{I}^{k*} \subseteq \mathcal{I}^k$ such that $\mathbf{a}^{ki\prime}\bar{\mathbf{x}}^{q_k^*} = b_i^k$ for all $i \in \mathcal{I}^{k*}$ and $|\mathcal{I}^{k*}| = n$.

We prove part (ii) using the same above optimal solution $\big(\{(\mathbf{c}^{\ell*}, \mathcal{G}^{\ell*})\}_{\ell\in\mathcal{L}}, \{(\bar{\mathbf{x}}^{q_k^*}, \mathbf{y}^{k*})\}_{k\in\mathcal{K}}\big)$ to (4). First, note that for each $\ell \in \mathcal{L}$, $\mathbf{c}^{\ell*}$ satisfies (4f), which means for all $k \in \mathcal{G}^{\ell*}$ there exists at least one $i \in \mathcal{I}^{k*}$ for which $y_i^{k*} > 0$. Moreover, because $\mathbf{a}^{ki\prime}\bar{\mathbf{x}}^{q_k^*} > b_i$ for $i \in \mathcal{I}^k \setminus \mathcal{I}^{k*}$ and y_i^{k*} is the associated dual variable, it must be that $y_i^{k*} = 0$ for all $i \in \mathcal{I}^k \setminus \mathcal{I}^{k*}$. Thus, from (4b) we have $\mathbf{c}^{\ell*} = \sum_{i\in\mathcal{I}^k} y_i^{k*}\mathbf{a}^{ki} = \sum_{i\in\mathcal{I}^{k*}} y_i^{k*}\mathbf{a}^{ki}$, or equivalently $\mathbf{c}^{\ell*} \in \text{cone}(\{\mathbf{a}^{ki}\}_{i\in\mathcal{I}^{k*}})$, for all $k \in \mathcal{G}^{\ell*}$. □

The following result characterizes the solution structure of the SC model under the special case where all DMs solve the same DMP.

Corollary 3. *Assume $\mathbf{A}^k = \mathbf{A}$ and $\mathbf{b}^k = \mathbf{b}$ for all $k \in \mathcal{K}$ and let $\mathcal{I}$ be the index set for rows of $\mathbf{A}$. Then there exists an optimal solution $\big(\{(\mathbf{c}^{\ell*}, \mathcal{G}^{\ell*})\}_{\ell\in\mathcal{L}}, \{(\mathbf{x}^{k*}, \mathbf{y}^{k*})\}_{k\in\mathcal{K}}\big)$ to (4) such that $\mathbf{c}^{\ell*} \in \text{cone}_+(\{\mathbf{a}^i\}_{i\in\mathcal{I}^*})$ for each cluster $\ell \in \mathcal{L}$, where $\mathcal{I}^* \subseteq \mathcal{I}$, $|\mathcal{I}^*| = n$, and $\text{cone}_+(\cdot)$ denotes the interior of the conic hull of given vectors, i.e., $\text{cone}_+(\{\mathbf{a}^i\}_{i\in\mathcal{I}^*}) = \{\sum_{i\in\mathcal{I}^*} \lambda_i\mathbf{a}^i \mid \lambda_i > 0, \forall i \in \mathcal{I}^*\}$.*

4.1. Lower bound formulation

Theorem 2 states that there exists an optimal solution to the SC model where $\mathbf{x}^{k*}$ is an extreme point of $\mathcal{X}^k$ for all $k \in \mathcal{K}$. Also, if $k \in \mathcal{G}^{\ell*}$ then $\mathbf{c}^{\ell*}$ must be a conic combination of $\mathbf{a}^{ki}$'s for i such that $\mathbf{a}^{ki\prime}\mathbf{x}^{k*} = b_i^k$. Based on this observation, we propose an MIP formulation that explicitly finds an extreme point $\mathbf{x}^{k*}$ for each $\mathcal{X}^k$, clusters the data points, and constructs $\mathbf{c}^{\ell*}$ for cluster ℓ as a conic combination of $\mathbf{a}^{ki}$'s for k assigned to cluster ℓ and for i such that $\mathbf{a}^{ki\prime}\mathbf{x}^{k*} = b_i^k$. We then show that the optimal value of this MIP is a lower bound on that of the SC problem:

$$\textbf{SC-LB}(\mathcal{K},\mathcal{L}): \quad \underset{\substack{\{(\mathbf{x}^k,\mathbf{v}^k,\boldsymbol{\lambda}^k)\}_{k\in\mathcal{K}},\\ \{(\mathbf{c}^\ell,\mathbf{c}^{\ell+},\mathbf{c}^{\ell-},\mathbf{z}^\ell)\}_{\ell\in\mathcal{L}},\mathbf{u}}}{\text{minimize}} \quad \max_{k\in\mathcal{K}}\{d(\hat{\mathbf{x}}^k, \mathbf{x}^k)\} \tag{7a}$$

$$\text{subject to} \quad \mathbf{A}^{k\prime}\boldsymbol{\lambda}^k - M_1(1-u_{k\ell}) \leq \mathbf{c}^\ell$$
$$\leq \mathbf{A}^{k\prime}\boldsymbol{\lambda}^k + M_1(1-u_{k\ell}), \quad \forall \ell \in \mathcal{L}, k \in \mathcal{K}, \tag{7b}$$
$$\lambda_i^k \leq M_2 v_i^k, \quad \forall k \in \mathcal{K}, i \in \mathcal{I}^k, \tag{7c}$$
$$b_i^k \leq \mathbf{a}^{ki\prime}\mathbf{x}^k \leq b_i^k + M_3(1-v_i^k), \quad \forall k \in \mathcal{K}, i \in \mathcal{I}^k, \tag{7d}$$
$$\sum_{i\in\mathcal{I}^k} v_i^k = n, \quad \forall k \in \mathcal{K}, \tag{7e}$$
$$\sum_{\ell\in\mathcal{L}} u_{k\ell} = 1, \quad \forall k \in \mathcal{K}, \tag{7f}$$
$$\mathbf{c}^\ell = \mathbf{c}^{\ell+} - \mathbf{c}^{\ell-}, \quad \forall \ell \in \mathcal{L}, \tag{7g}$$
$$\mathbf{c}^{\ell+} \leq \mathbf{z}^\ell, \quad \forall \ell \in \mathcal{L}, \tag{7h}$$
$$\mathbf{c}^{\ell-} \leq \mathbf{e} - \mathbf{z}^\ell, \quad \forall \ell \in \mathcal{L}, \tag{7i}$$
$$\mathbf{e}'(\mathbf{c}^{\ell+} + \mathbf{c}^{\ell-}) = 1, \quad \forall \ell \in \mathcal{L}, \tag{7j}$$
$$\mathbf{v}^k \in \{0,1\}^n, \mathbf{u} \in \{0,1\}^{K\times L}, \mathbf{z}^\ell \in \{0,1\}^n, \quad \forall k \in \mathcal{K}, \ell \in \mathcal{L}, \tag{7k}$$
$$\boldsymbol{\lambda}^k, \mathbf{c}^{\ell+}, \mathbf{c}^{\ell-} \geq \mathbf{0}, \quad \forall k \in \mathcal{K}, \forall \ell \in \mathcal{L}, \tag{7l}$$

where parameters M_1, M_2, and M_3 are sufficiently large positive constants. Using the result of Theorem 2, constraints (7b)–(7c) enforce each $\mathbf{c}^\ell$ to be a conic combination of some $\mathbf{a}^{ki}$'s; which $\mathbf{a}^{ki}$ is selected is dictated by binary variables v_i^k and $u_{k\ell}$. If $u_{k\ell} = 1$, data point $\hat{\mathbf{x}}^k$ is assigned to cluster ℓ and (7b) holds with equality. The variables λ_i^k in (7b) are then controlled by (7c) using binary variable v_i^k, i.e., if $v_i^k = 0$ then $\lambda_i^k = 0$ and thus preventing $\mathbf{a}^{ki}$ from being a basis vector for the conic hull constructing $\mathbf{c}^\ell$. Constraints (7d)–(7e) enforce each $\mathbf{x}^k$ to be an extreme point of $\mathcal{X}^k$, i.e., satisfying $\mathbf{a}^{ki\prime}\mathbf{x}^k \geq b_i^k$ with equality for n number of i's ensured by (7e). Constraint (7f) ensures that each observation is assigned to only one cluster. Finally, constraints (7g)–(7j) replace the non-convex normalization constraint (4f). The following result shows that the optimal value of the above problem is a lower bound on the optimal value of the SC problem, i.e., (4).

Proposition 4. *Let ρ^* and β^* denote the optimal objective values of problems (4) and (7), respectively. Then, we have* (i) *$\beta^* \leq \rho^*$, and* (ii) *$\beta^* = \rho^*$ if there exists $(\{\tilde{\mathbf{v}}^k\}_{k\in\mathcal{K}}, \tilde{\mathbf{u}}, \{\tilde{\mathbf{c}}^\ell\}_{\ell\in\mathcal{L}})$ optimal for (7) such that $\tilde{\mathbf{c}}^\ell \in \text{cone}_+(\{\mathbf{a}^{ki}\}_{i,k:\tilde{v}_i^k=1,\tilde{u}_{k\ell}=1})$ for each $\ell \in \mathcal{L}$.*

Table 1
Performance of IC, CI, and SC approximated by upper and lower bounds.

		Worst-case distance				Time (s)			
(n, m)	K	IC	CI	UB	LB	IC	CI	UB	LB
(10,30)	30	14.71	11.67	1.92	1.92	36.48	39.01	70.77	13.36
	40	13.09	13.97	1.78	1.77	112.55	174.39	273.01	20.31
	50	14.58	14.26	2.01	1.86	147.84	216.58	5707.04	27.75
(10,40)	30	9.51	9.57	1.97	1.97	74.84	94.52	238.45	16.43
	40	10.81	11.17	1.50	1.50	158.76	119.01	470.01	26.47
	50	12.27	12.15	2.09	1.99	575.92	321.21	1667.48	40.78
(20,60)	100	17.04	15.40	2.03	1.97	675.62	443.55	4441.23	426.37
	115	16.13	19.40	2.08	1.98	789.73	517.47	3417.01	589.38
	130	17.86	16.27	2.01	1.97	917.40	549.12	4323.03	691.85
(20,80)	100	15.20	13.31	2.03	1.85	833.62	609.72	2874.12	807.51
	115	14.98	17.41	2.11	1.92	885.66	739.27	4404.16	991.61
	130	14.50	13.85	2.06	2.00	1667.74	1165.07	5720.49	1141.38

Proposition 4 suggests that the optimal value of model (7) is a lower bound on that of the SC model. Proposition 4 also implies that once model (7) is solved, we can check the condition in Proposition 4(ii) to determine whether (7) achieves the exact optimal value of the SC model.

4.2. Upper bound formulation

Proposition 4 states that if formulation (7) finds a solution such that each $\mathbf{c}^\ell$, $\ell \in \mathcal{L}$, is a strict conic combination of the selected $\mathbf{a}^{ki}$ vectors, then its optimal value is equal to that of the SC model. Based on this observation, we add a constraint to (7) that enforces this condition and show that the following modified problem provides an upper bound on the optimal value of the SC model:

$$\textbf{SC-UB}(\mathcal{K}, \mathcal{L}): \quad \underset{\substack{\{(\mathbf{x}^k, \mathbf{v}^k, \boldsymbol{\lambda}^k)\}_{k\in\mathcal{K}},\\ \{(\mathbf{c}^\ell, \mathbf{c}^{\ell+}, \mathbf{c}^{\ell-}, \mathbf{z}^\ell)\}_{\ell\in\mathcal{L}}, \mathbf{u}}}{\text{minimize}} \quad \max_{k\in\mathcal{K}}\{d(\hat{\mathbf{x}}^k, \mathbf{x}^k)\} \tag{8a}$$

$$\text{subject to} \quad (7b) - (7l), \tag{8b}$$

$$\lambda_i^k \geq v_i^k \hat{\alpha}, \quad \forall k \in \mathcal{K}, i \in \mathcal{I}^k, \tag{8c}$$

where $\hat{\alpha}$ is a small positive constant. For each $\ell \in \mathcal{L}$, $k \in \mathcal{G}^\ell$, and $i \in \mathcal{I}^k$, if $v_i^k = 1$ then $\lambda_i^k \geq \hat{\alpha} > 0$, which enforces $\mathbf{c}^\ell$ to be a strict conic combination of n selected $\mathbf{a}^{ki}$ vectors (i.e., for which $v_i^k = 1$; see (7d)–(7e)). If there exists an optimal solution to the SC problem whose $\mathbf{c}^\ell$ vectors satisfy the strict conic combination condition then (8) with an appropriate $\hat{\alpha}$ generates the optimal solution for the SC model; otherwise, the optimal value of (8) is an upper bound on the optimal value of the SC model. We formalize this in the following result.

Proposition 5. *Given $\hat{\alpha}$, let ρ^* and β^* denote the optimal objective values of problems (4) and (8), respectively. Then, we have $\rho^* \leq \beta^*$.*

While **SC-LB**$(\mathcal{K}, \mathcal{L})$ and **SC-UB**$(\mathcal{K}, \mathcal{L})$ provide bounds for the SC problem, i.e., **SC**$(\mathcal{K}, \mathcal{L})$, these problems are typically large-scale MIPs and thus can be computationally challenging. The online appendix shows how the CI and IC approaches can be used to create initial feasible solutions for these MIPs and reduce the computational burden.

5. Numerical results

In this section, we first examine the performance of the proposed clustering approach using various-sized instances and discuss the computational benefits and limitations. We then present the results of the application of the proposed approach in the diet recommendation context to cluster DMs based on the similarity of their food preferences, which can be found in the online appendix.

We use various-sized randomly generated instances to demonstrate the CI, IC, and SC approaches. For small instances we chose $K \in \{30, 40, 50\}$ and generated LP instances with $n = 10$ and $m^k = m \in \{30, 40\}, \forall k = 1, \ldots, K$. For large instances we used $K \in \{100, 115, 130\}$, $n = 20$, and $m^k = m \in \{60, 80\}, \forall k = 1, \ldots, K$. To generate dataset $\hat{\mathcal{X}}$ for each instance, we generated K random cost vectors, solved K DMPs to generate optimal solutions, and added random noise to the solutions. All optimization problems were solved by Gurobi 9.1 [8] with a 16-core 2.9 GHz processor and 512 GB memory.

Table 1 shows the worst-case distances and solution times achieved by the IC and CI approaches as well as the upper and lower bound formulations for the SC problem (i.e., **SC-UB**$(\mathcal{K}, \mathcal{L})$ and **SC-LB**$(\mathcal{K}, \mathcal{L})$, respectively). Although the IC and CI problems involved solving smaller versions the SC problem, which were approximated by solving their respective smaller versions of both SC-UB and SC-LB problems, for brevity Table 1 only presents the IC and CI results approximated by the smaller version of SC-UB. Columns labeled UB in Table 1 show the results for **SC-UB**$(\mathcal{K}, \mathcal{L})$, which were obtained using an initial solution achieved by the IC and CI results presented in this table (see the appendix); thus, the solution time for UB is the time for finding an initial solution via either IC or CI (whichever gives a smaller worst-case distance) plus the time for the solver to improve the initial solution and find an optimal solution. Columns labeled LB show the results for **SC-LB**$(\mathcal{K}, \mathcal{L})$. For each instance (n, m, K), the reported worst-case distance values and times in the table were averaged over two sub-instances with $L = 3$ and $L = 5$. For all instances, we can see that the UB and LB values were close to each other or identical, indicating that the solutions from both the upper and lower bound formulations are close to the optimal solutions to the SC model. Since

the SC model considers a minimization of the worst-case distance, our suggestion is to use the clusters and cost vectors achieved by the upper bound formulation so as not to underestimate the true cluster instability. Additional results on the performance of IC, CI, and SC can be found in the online appendix.

6. Conclusion

In this paper we introduced a new clustering approach, called optimality-based clustering, that clusters DMs based on similarity of their decision preferences. We formulated the clustering problem as a non-convex optimization problem and proposed MIP formulations that provide lower and upper bounds. We also proposed two heuristics that can be efficient in large instances and perform comparably to solving the problem exactly in certain instances. We used the proposed clustering models in the diet recommendation context to cluster DMs based of their food preferences. The future research includes extending the idea of optimality-based clustering to other types of DMPs such as non-linear, mixed-integer, and multi-objective optimization problems.

Acknowledgement

This research was partially supported by National Science Foundation grant CMMI#1908244.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.orl.2021.12.012.

References

[1] R.K. Ahuja, J.B. Orlin, Inverse optimization, Oper. Res. 49 (5) (2001) 771–783.
[2] A. Aswani, Z.-J. Shen, A. Siddiq, Inverse optimization with noisy data, Oper. Res. 66 (3) (2018) 870–892.
[3] A. Babier, T.C. Chan, T. Lee, R. Mahmood, D. Terekhov, An ensemble learning framework for model fitting and evaluation in inverse linear optimization, Inf. J. Optim. 3 (2) (2021) 119–138.
[4] D. Bertsimas, V. Gupta, I.C. Paschalidis, Data-driven estimation in equilibrium using inverse optimization, Math. Program. 153 (2) (2015) 595–633.
[5] T.C. Chan, T. Craig, T. Lee, M.B. Sharpe, Generalized inverse multiobjective optimization with application to cancer therapy, Oper. Res. 62 (3) (2014) 680–695.
[6] P.M. Esfahani, S. Shafieezadeh-Abadeh, G.A. Hanasusanto, D. Kuhn, Data-driven inverse optimization with imperfect information, Math. Program. 167 (1) (2018) 191–234.
[7] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1996, pp. 226–231.
[8] L. Gurobi, Optimization. Gurobi optimizer reference manual, http://www.gurobi.com, 2020.
[9] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Science & Business Media, 2009.
[10] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.
[11] A. Keshavarz, Y. Wang, S. Boyd, Imputing a convex objective function, in: 2011 IEEE International Symposium on Intelligent Control, IEEE, 2011, pp. 613–619.
[12] H.-P. Kriegel, P. Kröger, J. Sander, A. Zimek, Density-based clustering, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1 (3) (2011) 231–240.
[13] A. Likas, N. Vlassis, J.J. Verbeek, The global k-means clustering algorithm, Pattern Recognit. 36 (2) (2003) 451–461.
[14] A. Rakhlin, A. Caponnetto, Stability of k-means clustering, Adv. Neural Inf. Process. Syst. 19 (2007) 1121.
[15] Z. Shahmoradi, T. Lee, Quantile inverse optimization: improving stability in inverse linear programming, Oper. Res. (2021), https://doi.org/10.1287/opre.2021.2143.
[16] U. Von Luxburg, Clustering Stability: an Overview, 2010.
[17] R. Xu, D. Wunsch, Clustering, Vol. 10, John Wiley & Sons, 2008.
[18] X. Xu, M. Ester, H.-P. Kriegel, J. Sander, A distribution-based clustering algorithm for mining in large spatial databases, in: Proceedings 14th International Conference on Data Engineering, IEEE, 1998, pp. 324–331.