Mathematical Optimization Modelling for Group Counterfactual Explanations

Carrizosa, Emilio; Ramírez-Ayerbe, Jasone; Romero Morales, Dolores

Document Version
Final published version

Published in:
European Journal of Operational Research

DOI:
10.1016/j.ejor.2024.01.002

Publication date:
2024

European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa [a,*], Jasone Ramírez-Ayerbe [a], Dolores Romero Morales [b]

[a] Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain
[b] Department of Economics, Copenhagen Business School, Frederiksberg, Denmark

ARTICLE INFO

Keywords:
Machine learning
Interpretability
Mathematical optimization
Counterfactual explanations
Location analysis

ABSTRACT

Counterfactual Analysis has shown to be a powerful tool in the burgeoning field of Explainable Artificial Intelligence. In Supervised Classification, this means associating with each record a so-called counterfactual explanation: an instance that is close to the record and whose probability of being classified in the opposite class by a given classifier is high. While the literature focuses on the problem of finding one counterfactual for one record, in this paper we take a stakeholder perspective, and we address the more general setting in which a group of counterfactual explanations is sought for a group of instances. We introduce some mathematical optimization models as illustration of each possible allocation rule between counterfactuals and instances, and we identify a number of research challenges for the Operations Research community.

1. Introduction

Artificial Intelligence and Machine Learning algorithms are often criticized by their lack of transparency, being seen as black-boxes (Rudin, 2019). Such opaqueness is especially undesirable in high-stakes decision making, involving important decisions to citizens such as social benefits allocation, loan approval, medical diagnosis, or pretrial/parole/sentencing decisions (Azizi, Vayanos, Wilder, Rice, & Tambe, 2018; Baesens, Setiono, Mues, & Vanthienen, 2003; Zeng, Gensheimer, Rubin, Athey, & Shachter, 2022; Zeng, Ustun, & Rudin, 2017), with the danger of yielding unfair outcomes for sensitive groups (Besse, del Barrio, Gordaliza, Loubes, & Risser, 2022; Miron, Tolan, Gómez, & Castillo, 2020, 2021). The importance of this issue has already been recognized by public administrations, such as the European Commission (European Commission, 2020; Goodman & Flaxman, 2017; Hupont, Micheli, Delipetrev, Gómez, & Soler Garrido, 2022). In answer to this need, the field of Explainable Artificial Intelligence (XAI) (Du, Liu, & Hu, 2019; Goethals, Martens, & Evgeniou, 2022; Jung, Concannon, Shroff, Goel, & Goldstein, 2020; Molnar, Casalicchio, & Bischl, 2020; Rudin et al., 2022; Zhang, Song, Sun, Tan, & Udell, 2019) has witnessed an explosion of papers developing novel methods.

In Explainable Artificial Intelligence, supervised classification models are sought to have a good trade-off between prediction accuracy and interpretability. Once the classifier has been trained, it would be convenient to have procedures to identify how records should be changed in their features to being classified in the "good" class, e.g., to be classified as a good payer for a loan, or a healthy person for a medical condition, or an adequate professional in a job selection process, or a defendant not showing recidivism. Such modified solutions, the so-called counterfactual explanations, (Martens & Provost, 2014; Wachter, Mittelstadt, & Russell, 2017), are addressed in this paper. See Artelt and Hammer (2019), Guidotti (2022), Karimi, Barthe, Schölkopf, and Valera (2022), Sokol and Flach (2019), Stepin, Alonso, Catala, and Pereira-Fariña (2021), Verma et al. (2022) for recent surveys on Counterfactual Analysis, and Browne and Swift (2020), Freiesleben (2022), Karimi, Barthe, Schölkopf, and Valera (2022) for related problems under different names, such as inverse classification or adversarial perturbations.

Finding counterfactual explanations amounts to solving a mathematical optimization model, whose ingredients will be detailed below. We search for counterfactuals in a set $\mathcal{X}$. We consider a binary classification problem, with labels in $\{-1, +1\}$ (+1 being considered as the "good" class), and a classifier, identified by a function $P : \mathcal{X} \longrightarrow [0, 1]$ such that $P(\boldsymbol{x})$ is the probability of $\boldsymbol{x}$ being classified as positive, and thus, if a deterministic classification is sought, the classifier labels as positive those $\boldsymbol{x}$ with $P(\boldsymbol{x}) \geq \tau$ for some fixed threshold value τ, (Kanamori, Takagi, Kobayashi, & Arimura, 2020; Parmentier & Vidal, 2021).

The most frequent version of counterfactual analysis found in the literature is the single-instance single-counterfactual case. There, we have at hand just one record $\boldsymbol{x}^0 \in \mathcal{X}$ where the classifier gives a low probability of being classified in the positive class, and we seek a so-called counterfactual instance $\boldsymbol{x}$ in the feasible set $\mathcal{X}(\boldsymbol{x}^0) \subset \mathcal{X}$ such that the cost $C(\boldsymbol{x}^0, \boldsymbol{x})$ to perturb $\boldsymbol{x}^0$ to yield $\boldsymbol{x}$ is low, while its probability

* Corresponding author.
E-mail addresses: ecarrizosa@us.es (E. Carrizosa), mrayerbe@us.es (J. Ramírez-Ayerbe), drm.eco@cbs.dk (D. Romero Morales).

https://doi.org/10.1016/j.ejor.2024.01.002
Received 2 March 2023; Accepted 2 January 2024
Available online 5 January 2024

$P(\boldsymbol{x})$ of being classified as positive is high. In sum, in the single-instance single-counterfactual case, a counterfactual explanation for the instance $\boldsymbol{x}^0$ can be found as an efficient solution of the following bi-objective optimization problem:

$$\min_{\boldsymbol{x}\in\mathcal{X}(\boldsymbol{x}^0)} \quad \left(C(\boldsymbol{x}^0,\boldsymbol{x}), -P(\boldsymbol{x})\right). \tag{CE}$$

Needless to say, the same problem can be solved to change the prediction of a record from positive to negative. In this case, we have a record where the classifier gives a high probability of being classified in the positive class, and we seek a counterfactual explanation whose probability of being classified in the positive class is low, i.e., a high probability of being classified in the negative class.

Different numerical solution approaches can be found in the literature to generate efficient solutions of Problem (CE). These include smooth optimization, e.g., Joshi, Koyejo, Vijitbenjaronk, Kim, and Ghosh (2019), Ramakrishnan, Lee, and Albarghouthi (2020), mixed integer optimization, e.g., Carrizosa, Ramírez-Ayerbe, and Romero Morales (2023), Carrizosa, Ramírez-Ayerbe, and Romero Morales (2024), Cui, Chen, He, and Chen (2015), Fischetti and Jo (2018), Kanamori et al. (2020, 2021), Maragno, Röber, and Birbil (2022), Parmentier and Vidal (2021), Russell (2019), multi-objective optimization, e.g., Dandl, Molnar, Binder, and Bischl (2020), Del Ser, Barredo-Arrieta, Díaz-Rodríguez, Herrera, and Holzinger (2022), Raimundo, Nonato, and Poco (2022), robust optimization, e.g., Maragno et al. (2023), boolean satisfiability (SAT), e.g., Karimi, Barthe, Balle, and Valera (2020), heuristic and metaheuristic approaches, e.g., Guidotti et al. (2019), Poyiadzi, Sokol, Santos-Rodriguez, De Bie, and Flach (2020). In this paper, we study the more general setting in which we are given a group of instances $\boldsymbol{x}_s^0$, $s = 1, 2\ldots, S$, hereafter denoted by $\underline{\boldsymbol{x}}^0 = (\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_S^0)$, to be perturbed to increase their probability of being classified in the positive class. Instead of finding one counterfactual instance for each instance $\boldsymbol{x}_s^0$ individually, we are interested in *group counterfactual analysis* in which we seek for $\underline{\boldsymbol{x}}^0$ a group of R counterfactual instances $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R) \in \underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0) \subset \underline{\mathcal{X}} := \mathcal{X}^R$.

From a stakeholder perspective, there are different reasons to perform a group counterfactual analysis (Carrizosa et al., 2024; Fernández, de Diego, Aceña, Fernández-Isabel, & Moguerza, 2020; Fernández, de Diego, Moguerza, & Herrera, 2022; Rawal & Lakkaraju, 2020; Yousefzadeh & O'Leary, 2020, 2022). First, even if just one counterfactual is sought for each instance $\boldsymbol{x}_s^0$, $s \in \{1, 2, \ldots, S\}$, linking constraints may exist, preventing solving the counterfactual models independently (Carrizosa et al., 2024). Such linking constraints appear, for example, when counterfactuals for records which are close should also be close, or when the statistical distribution of the counterfactuals should resemble the one of the original instances (Slack, Hilgard, Lakkaraju, & Singh, 2021). Second, several counterfactuals may be sought for the same instance, sufficiently far from each other (Wachter et al., 2017), and thus the procedure would yield for each instance a collection of counterfactuals that are hopefully diverse. Third, stakeholders may be in search of just a few counterfactual instances, to be seen as benchmarks for the group $\{\boldsymbol{x}_s^0\}_{s=1}^S$, and hence several instances will share the same counterfactual. To end, stakeholders may want to detect a small subset of features such that perturbing these features can increase the probability $P(\boldsymbol{x}_r^0)$ of being classified in the positive class, for all r, (Eckstein, Bates, Jefferis, & Funke, 2021; Piccialli, Romero Morales, & Salvatore, 2022; Sharma, Henderson, & Ghosh, 2020).

The group counterfactual analysis yields the following optimization problem:

$$\min_{\underline{\boldsymbol{x}}\in\underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0)} \quad \left(\boldsymbol{C}(\underline{\boldsymbol{x}}^0,\underline{\boldsymbol{x}}), -\boldsymbol{P}(\underline{\boldsymbol{x}})\right), \tag{GroupCE}$$

where $\boldsymbol{C}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}})$ measures the cost incurred when $\underline{\boldsymbol{x}}^0$ is perturbed to yield $\underline{\boldsymbol{x}}$ and $\boldsymbol{P}(\underline{\boldsymbol{x}})$ is defined as a componentwise nondecreasing function of the probabilities $P(\boldsymbol{x})$ of the counterfactuals.

To generate efficient solutions of a bi-objective optimization problem such as (GroupCE), two main methods are usually found in the literature of Multi-Objective Optimization, and they differ in the way the second objective $\boldsymbol{P}$ is controlled. If imposed as a hard constraint, i.e., $\boldsymbol{P}$ must be above a threshold value $\nu \in [0, 1]$, we obtain (GroupCEhard),

$$\min_{\underline{\boldsymbol{x}}\in\underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0)} \quad \boldsymbol{C}(\underline{\boldsymbol{x}}^0,\underline{\boldsymbol{x}}) \tag{GroupCEhard}$$

$$\text{s.t.} \quad \boldsymbol{P}(\underline{\boldsymbol{x}}) \geq \nu. \tag{1}$$

Alternatively, we may optimize a convex combination of both objectives, as done in Wachter et al. (2017) for the single-instance single-counterfactual case. However, it is well-known that, as opposed to (GroupCEhard), optimizing weighted sums may not generate all the efficient solutions of Problem (GroupCE), unless both the objective function and the feasible region are convex, assumptions which, as shown below, are unlikely to be fulfilled in this problem. See Carrizosa and Fliege (2002), Carrizosa and Romero Morales (2001), Ogryczak (2001), Ruiz, Luque, and Cabello (2009) and references therein for other ways to obtain solutions to a bi-objective optimization model.

If we leave aside for a moment the $\boldsymbol{P}$ criterion, Problem (GroupCE) as well as its scalarized version (GroupCEhard) strongly resemble facility location problems, (Drezner & Hamacher, 2004; Laporte, Nickel, & Saldanha da Gama, 2020). Indeed, the S instances $\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_S^0$ in $\underline{\boldsymbol{x}}^0$ may be seen as the set of users, and R new facilities, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R$, are sought in a region $\underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0)$ to minimize the transportation costs from the users (the records) to the facilities (counterfactuals). The criterion $\boldsymbol{P}$, concerning the probabilities of the counterfactuals being classified as positive, can be seen as related to the fixed cost of opening facilities. As for facility location problems, different types of optimization problems are obtained, depending on the choices of the ingredients defining Problem (GroupCE). In this paper, we discuss the so-obtained optimization problems, identifying a number of research challenges for the Operations Research community.

The remainder of this paper is organized as follows. In Section 2, we describe the ingredients of the group counterfactual analysis problem (GroupCE). In Section 3, several mathematical optimization models for group counterfactual analysis are illustrated, with different choices of the ingredients. We end the paper in Section 4 with some concluding remarks and avenues for future research.

2. Ingredients

In this section we will describe the different ingredients required to define the mathematical optimization problems arising in counterfactual analysis, namely, the ambient space, i.e., the space where counterfactuals are taken from; the allocation rule, i.e., how counterfactual explanations are assigned to instances; the constraints imposed to build the counterfactuals; the classifier used and how the probabilities given by such classifier to the different counterfactuals to be in the positive class are aggregated; and the cost criterion measuring how difficult is for the instances to be perturbed to yield their counterfactuals.

2.1. Ambient space

The first modelling decision affects the set $\mathcal{X}$, hereafter the ambient space, containing each of the R counterfactuals. As in the literature in Location Analysis, in which this same question induces a taxonomy between Discrete Location, (Mirchandani & Francis, 1990), and Continuous Location, (Plastria, 1995), here we will have endogenous and exogenous counterfactuals.

The set $\mathcal{X}$ can be a finite collection of observed datapoints, yielding the so-called *endogenous counterfactuals*, as in Ramon, Martens, Provost, and Evgeniou (2020), Wexler et al. (2019). Using endogenous explanations has the advantage that the counterfactual explanations obtained actually exist in reality (Keane & Smyth, 2020). Alternatively, and this is the most popular approach in the literature, (Guidotti, 2022), the counterfactuals can be synthetically built, yielding in this case the so-called *exogenous explanations*. When searching for exogenous

(a) One-for-one (b) Many-for-one (c) One-for-all (d) One-for-many

Fig. 1. Illustration of the different allocation rules between instances (squares) and their counterfactual explanations (circles) in group counterfactual analysis.

explanations, it is common to assume that $\mathcal{X}$ is a finite-dimensional vector space of dimension J, which is the case when the instances are characterized by numerical features. Categorical features are typically represented as a binary vector (in the so-called one-hot encoding, a categorical feature with v categories is represented by a binary vector in $\{0,1\}^{v-1}$), and thus can also be seen as points in a vector space. For more complex types of data, a common practice consists of mapping the data into a vector space via a vector embedding (Van Looveren & Klaise, 2021), such as word embeddings for text data (Tolkachev, Mell, Zdancewic, & Bastani, 2022) or node embeddings for graph data (Prado-Romero, Prenkaj, Stilo, & Giannotti, 2022). While such embeddings are very useful since they allow us to use the powerful machinery of vector analysis, one needs to revert to the original space (of words, graphs, etc.) the counterfactual solution obtained in the embedded vector space. Finally, and to prevent the infinite-dimensional nature of the problems arising in the presence of functional features, the counterfactuals can be built as linear or convex combinations of observed datapoints, yielding thus a finite-dimensional $\mathcal{X}$, making optimization much easier (Ates, Aksar, Leung, & Coskun, 2021; Carrizosa, Ramírez-Ayerbe, & Romero Morales, 2023; Delaney, Greene, & Keane, 2021; Karlsson, Rebane, Papapetrou, & Gionis, 2020).

In terms of the type of mathematical optimization problems that are obtained, finding endogenous counterfactual explanations amounts to solving combinatorial problems, with a very similar structure to discrete facility location problems, such as the classic p-median problem, (Avella, Sassano, & Vasil'ev, 2007; García, Labbé, & Marín, 2011; Grötschel & Wakabayashi, 1989; Marín & Pelegrín, 2019; Mladenović, Brimberg, Hansen, & Moreno-Pérez, 2007), whereas searching for exogenous counterfactual explanations will lead to either continuous or mixed integer optimization problems, such as the classic minimum-sum-of-squared-distances problem, (Aloise, Hansen, & Liberti, 2012; Liberti & Manca, 2022; Piccialli, Sudoso, & Wiegele, 2022), the Weber problem, (Chandrasekaran & Tamir, 1990; Plastria, 1992; Weiszfeld & Plastria, 2009), or its extensions to the multi-facility case, namely, the multisource Weber problem, (Brimberg, Hansen, Mladenović, & Taillard, 2000) and the minisum multifacility Weber problem, (Lefebvre, Michelot, & Plastria, 1991). This means that some of the existing numerical optimization tools developed in Discrete as well as Continuous Location Analysis can be easily tailored to tackle group counterfactual analysis problems successfully, and, conversely, new facility location models are obtained motivated by this emerging field of application.

2.2. Allocation rules

Most of the literature addresses the problem of finding one counterfactual explanation $\boldsymbol{x}$ for one instance $\boldsymbol{x}^0$, as in Problem (CE). Even if it is assumed that there is a tuple of instances $\underline{\boldsymbol{x}}^0 = (\boldsymbol{x}_1^0, \dots, \boldsymbol{x}_S^0)$, and a counterfactual for each individual is sought, Problem (CE) is solved separately for the different individuals. A user perspective is then followed. However, when a stakeholder perspective is followed, and a tuple $\underline{\boldsymbol{x}}$ of counterfactuals is built for the tuple $\underline{\boldsymbol{x}}^0$ of individuals, the question of how to allocate instances to counterfactuals arises.

Consider a bipartite graph as the ones in Fig. 1 , where the origin nodes (squares in red) represent the instances, and the destination nodes (circles in blue) represent the labels for the R counterfactuals. Observe that the value $\boldsymbol{x}_r \in \mathcal{X}$ corresponding to each destination node is a decision variable. Moreover, in some cases the edges of this bipartite graph are given, and in some cases they will also be decision variables. Let us now introduce some notation for the edges of the bipartite graph linking instances to the labels of their counterfactuals. For each $s \in \{1,2,\dots,S\}$, let $\mathcal{R}_s$ be the set of indices $r \in \{1,2,\dots,R\}$ such that counterfactuals $\boldsymbol{x}_r$ are associated with instance $\boldsymbol{x}_s^0$. Conversely, for each $r \in \{1,2,\dots,R\}$, let $\mathcal{S}_r$ be set the set of indices $s \in \{1,2,\dots,S\}$ such that instances $\boldsymbol{x}_s^0$ are associated with counterfactual $\boldsymbol{x}_r$. Observe that, by construction, $r \in \mathcal{R}_s$ iff $s \in \mathcal{S}_r$. Depending on whether the edges in this bipartite graph are fixed or are decision variables, and depending on the geometry of such connections, different models are obtained. We briefly discuss four of them in what follows, which are easy to relate to the Location Analysis literature and are versatile to encompass the existing literature on Counterfactual Analysis. In the first allocation rule under consideration, for each instance $\boldsymbol{x}_s^0$ exactly one counterfactual $\boldsymbol{x}_s$ is sought and vice versa, as in Fig. 1(a), i.e., $\mathcal{R}_s$ and $\mathcal{S}_r$ are both a singleton and known in advance, yielding what we call the *one-for-one* allocation model, in which just the locations of the $R = S$ counterfactuals are sought.

In the second allocation rule, the *many-for-one* allocation model, as in Fig. 1(b), all edges are known in advance too, satisfying that $\mathcal{S}_r$ is a singleton for each $s \in \{1,2,\dots,S\}$ and each $\mathcal{R}_s \neq \emptyset$: each instance $\boldsymbol{x}_s^0$ has associated its tuple $(\boldsymbol{x}_r)_{r\in\mathcal{R}_s}$ of counterfactuals, (Mothilal, Sharma, & Tan, 2020), fixing the concern raised in Wachter et al. (2017), where it is stated that one single counterfactual instance could be too restrictive and not take into account the user's personal circumstances.

In the third allocation rule, the *one-for-all* allocation model, as in Fig. 1(c), $R = 1$ and $\mathcal{S}_1 = \{1,2,\dots,S\}$: all instances $\boldsymbol{x}_s^0$, $s = 1,2,\dots,S$, share the same counterfactual $\boldsymbol{x}_1$, to be seen as a benchmark for the group. Therefore, in this allocation rule all edges are known in advance too.

Finally, in the fourth allocation rule, the *one-for-many* allocation model, as in Fig. 1(d), each $\mathcal{R}_s$ is a singleton and each $\mathcal{S}_r \neq \emptyset$. Here, the edges are decision variables too, and thus, one seeks both the location $\boldsymbol{x}_r$ of the counterfactuals, $r \in \{1,2,\dots,R\}$, and also a partition of $\{1,2,\dots,S\}$ such that each subset $\{\boldsymbol{x}_s^0 : s \in \mathcal{S}_r\}$ shares the same counterfactual.

2.3. Constraints

We discuss in what follows the type of constraints the counterfactual explanations $\underline{\boldsymbol{x}} \in \underline{\mathcal{X}} := \mathcal{X}^R$ for the group $\underline{\boldsymbol{x}}^0$ are expected to satisfy. Most are valid for both models with endogenous and exogenous counterfactuals, whereas some require a certain structure (e.g., a vector space) on the ambient space $\mathcal{X}$. Some can be found in the literature on single-instance single-counterfactual models, and some appear naturally when, as done in this paper, a stakeholder's perspective is followed.

Constraints defining $\underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0)$ model either the interaction between each record $\boldsymbol{x}_s^0$ and the tuple $(\boldsymbol{x}_r)_{r\in\mathcal{R}_s}$ of counterfactual explanations associated with $\boldsymbol{x}_s^0$, or the interaction between counterfactuals. For instance-counterfactual interactions, several constraints can be defined. One may wish to limit the range of the perturbations imposed on $\boldsymbol{x}_s^0$, and thus avoiding impossible or unrealistic perturbations. This can be expressed as

$$d(\boldsymbol{x}_s^0, \boldsymbol{x}_r) \leq \tau \quad \forall r \in \mathcal{R}_s,\ s \in \{1,2,\dots,S\}, \tag{2}$$

for some distance, or with more generality, dissimilarity measure, (Kaufman & Rousseeuw, 2009). If the ambient space $\mathcal{X}$ in which counterfactuals are sought is a finite-dimensional vector space, a natural choice for d is the weighted ℓ_∞ norm, and thus the balls of d are hyperrectangles centred at $\boldsymbol{x}_s^0$, implying that lower and upper bounds are given on the values of each coordinate (feature) of each $\boldsymbol{x}_r$.

While constraint (2) forces counterfactuals not to be too different from their associated instances, one may also force them not to be too different from a cloud of historical data. If a finite set $\mathcal{D} \subset \mathcal{X}$ of data points is considered, e.g., $\mathcal{D} = \{\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_S^0\}$, one may force each counterfactual to be close to some point in $\mathcal{D}$. If $\mathcal{X}$ is a vector space, instead of forcing the counterfactuals to be in the union of balls centred at points in $\mathcal{D}$, one can force, as in, e.g., (Maragno et al., 2022), counterfactuals to be close to $\text{conv}(\mathcal{D})$, the convex hull of $\mathcal{D}$.

Other constraints appear naturally when the instances in $\underline{\boldsymbol{x}}^0$ have features that cannot be moved, yielding constraints of the form $(\boldsymbol{x}_r)_k = (\boldsymbol{x}_s^0)_k \; \forall r \in \mathcal{R}_s$, or when a feature in $\boldsymbol{x}^0$ is categorical, yielding constraints of the form $\sum_{k\in\mathcal{K}}(\boldsymbol{x}_r)_k \leq 1$, where $(\boldsymbol{x}_r)_k$ are binary variables for all $r \in \mathcal{R}_s$, $k \in \mathcal{K}$, and $\mathcal{K}$ is the set of indices of the binary features used in the one-hot encoding of such categorical feature. See Maragno et al. (2022) for further details.

The allocation of counterfactual explanations to instances (and vice versa) may be known in advance – a natural choice in the one-for-one, the many-for-one and the one-for-all models –, but it may also be a decision variable. In the one-for-many model, where each $\mathcal{R}_s$ is a singleton, the S instances $\{\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_S^0\}$ are partitioned into R clusters, namely, the sets $\left\{\boldsymbol{x}_s^0 : s \in S_r\right\}$, $r = 1, 2, \ldots, R$. Must-link (respectively cannot-link) constraints, e.g., Vasilyev and Ushakov (2021), force two instances $\boldsymbol{x}_i^0$, $\boldsymbol{x}_j^0$ to be allocated to the same (respectively different) counterfactual, i.e., $\mathcal{R}_i = \mathcal{R}_j$ (respectively $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$), or cardinality constraints of type $|S_r| \leq \tau$, see e.g., Mulvey and Beck (1984), could be imposed. Observe that having the sets S_r to be decided implies that constraints such as (2) need to be rewritten as indicator constraints, (Belotti et al., 2016; Bomze & Peng, 2023; Han, Gómez, & Atamtürk, 2023; Wei, Gómez, & Küçükyavuz, 2022).

Different types of constraints modelling the interaction between counterfactuals are also natural. For instance, we may want to avoid shifting a specific categorical feature for all instances to the same category, avoiding unrealistic scenarios, such as requiring all individuals to be in the highest income bracket. Alternatively, we may want to ensure that the statistical distribution of the counterfactuals should resemble the one of the original instances by imposing a constraint of the form $W(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) \leq \tau$, where $W(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}})$ denotes a distance, e.g., the Wasserstein distance, (Carrizosa, Halskov, & Romero Morales, 2023; Chen, Kuhn, & Wiesemann, 2022; Peyré & Cuturi, 2019), or divergence between the uniform distributions on $\underline{\boldsymbol{x}}^0$ and $\underline{\boldsymbol{x}}$, e.g., Klafszky, Mayer, and Terlaky (1989). In the same vein, in the one-for-many allocation model in which the set of instances is partitioned into clusters $\{S_r\}_{r=1}^R$, we may impose conditions on the statistical distribution of instances in a given cluster (e.g., to have records with different values of a categorical feature to model diversity within the cluster). Finally, while it seems desirable that similar instances should have similar counterfactuals, it has been noted in the literature that this does not necessarily hold for some existing counterfactual analysis models (Artelt et al., 2021; Slack et al., 2021). One way to fix this is to impose a Lipschitz continuity constraint through a dissimilarity d:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq \tau d(\boldsymbol{x}_k^0, \boldsymbol{x}_l^0), \quad \forall i \in \mathcal{R}_k, j \in \mathcal{R}_l, \forall k, l \in \{1, 2, \ldots, S\}, \tag{3}$$

for some threshold value τ.

2.4. Score-based models

To build counterfactual explanations, an already trained classifier is given. To formulate the optimization problem, it is necessary to know the inner-workings of the classifier and how to model them. We will do this for a large class of well-known classifiers, namely score-based ones, as in Carrizosa, Ramírez-Ayerbe, and Romero Morales (2023), Carrizosa et al. (2024). In a score-based classifier (Carrizosa, Molero-Río, & Romero Morales, 2021; Carrizosa & Romero Morales, 2013; Gambella, Ghaddar, & Naoum-Sawaya, 2021) one has a score function $f : \mathcal{X} \longrightarrow \mathbb{R}$ and the probability of $\boldsymbol{x}$ being classified in the positive class has the form

$$P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x})), \tag{4}$$

where φ is an increasing function. The simplest case of score-based classifier corresponds to linear classification models, (Ustun, Spangher, & Liu, 2019), where $\mathcal{X}$ is assumed to be a finite-dimensional vector space, and the score function f is defined as $f(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x}+b$. In particular, for logistic regression, the probability of $\boldsymbol{x}$ being classified as positive $\varphi(f(\boldsymbol{x}))$ is obtained when φ is the logistic function

$$\varphi(t) = \frac{1}{1+e^{-t}}, \tag{5}$$

while for (linear) support vector machine (SVM), (Salazar, Denton, & Salleb-Aouissi, 2022), φ has been assumed in Platt (1999) to have a sigmoidal form, with parameters estimated via maximum likelihood from a training sample. The reader is referred to the Supplementary Material for more examples of score-based classifiers, and to e.g., Carrizosa and Romero Morales (2013), Duarte Silva (2017), Gambella et al. (2021), Hastie, Tibshirani, and Friedman (2009), Palagi (2019), Piccialli and Sciandrone (2018) for further details on classification and the role played by Mathematical Optimization in the field.

2.5. Aggregating probabilities: Modelling $\boldsymbol{P}$

For each $\boldsymbol{x} \in \mathcal{X}$, the classifier at hand gives a score $f(\boldsymbol{x})$, yielding a probability $\varphi(f(\boldsymbol{x}))$ of belonging to the positive class. Hence, for a tuple $\underline{\boldsymbol{x}}$ of counterfactuals, one obtains a tuple of probabilities $\left(P(\boldsymbol{x}_1), \ldots, P(\boldsymbol{x}_R)\right) = \left(\varphi(f(\boldsymbol{x}_1)), \ldots, \varphi(f(\boldsymbol{x}_R))\right)$, aggregated into the scalar $\boldsymbol{P}(\underline{\boldsymbol{x}})$, in such a way that the higher the value of $\boldsymbol{P}(\underline{\boldsymbol{x}})$, the more reliable the tuple of counterfactuals $\underline{\boldsymbol{x}}$ is. The function $\boldsymbol{P}(\underline{\boldsymbol{x}})$ can be defined in different ways. One may want to ensure that every counterfactual explanation has a sufficiently high probability of being classified in the positive class, and thus we can take

$$\boldsymbol{P}(\underline{\boldsymbol{x}}) = \min_{1\leq r\leq R} P(\boldsymbol{x}_r), \tag{6}$$

which, when applied to Problem (GroupCEhard), makes constraint (1) take the form

$$f(\boldsymbol{x}_r) \geq \varphi^{-1}(\nu) \quad \forall r = 1, 2, \ldots, R. \tag{7}$$

Some remarks follow. First, observe that, if the probabilistic classifier is made deterministic by labelling as positive all $\boldsymbol{x}_r$ with score $f(\boldsymbol{x}_r)$ above some threshold value, say $\varphi^{-1}(\nu)$, as typically done in the literature on single-instance single-counterfactual models, constraints (7) indicate that all counterfactuals are to be labelled as members of the positive class. Second, since our aim is to generate efficient solutions of the bi-objective problem (GroupCE), different values of the parameter ν are expected to be taken in (GroupCEhard). This is equivalent to take different values of the right hand side in (7), avoiding the estimation process associated with φ. Finally, observe also that these are linear constraints when f is the score function associated with a linear classifier, as, e.g., the logistic regression or the SVM with linear kernel.

Other choices for $\boldsymbol{P}(\underline{\boldsymbol{x}})$ yielding tractable models for some classifiers include taking as $\boldsymbol{P}$ the average of the probabilities of the counterfactuals, $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \frac{1}{R}\sum_{r=1}^R |S_r| P(\boldsymbol{x}_r)$, or their geometric mean, $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \left(\prod_{r=1}^R P(\boldsymbol{x}_r)^{|S_r|}\right)^{1/R}$. The first model yields a linear constraint when P is linear, as happens, for instance, in additive tree models. For the second model, expressing the constraint $\boldsymbol{P}(\underline{\boldsymbol{x}}) \geq \nu$ as $\log(\boldsymbol{P}(\underline{\boldsymbol{x}})) \geq \log(\nu)$, we obtain a concave constraint when P is log-concave, as happens in the logistic regression classifier or if the score function is linear and a log-concave function φ is used to pass from scores to probabilities, as in (5).

2.6. Modelling the cost criterion

To obtain a counterfactual explanation one has to define the cost function. The most simple case is when the cost function reflects the dissimilarity between the instances and the counterfactuals, $\boldsymbol{C}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \text{Dissimilarity}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}})$. For endogenous counterfactuals, we can take $\text{Dissimilarity}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \sum_{s=1}^{S} \sum_{r \in \mathcal{R}_s} d(\boldsymbol{x}_s^0, \boldsymbol{x}_r)$, where each $d(\boldsymbol{x}_s^0, \boldsymbol{x}_r)$ represents the distance or dissimilarity between the two points involved, and are elements of a given distance matrix. Observe that, when the allocations are decision variables, as in the one-for-many allocation model, minimizing such $\boldsymbol{C}$ amounts to solving a discrete R-median location problem eventually with some constraints, such as capacity or budget constraints. If instead of measuring $\text{Dissimilarity}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}})$ by a sum or average of individual costs we take into consideration the largest individual cost, i.e., $\text{Dissimilarity}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \max_{1 \leq s \leq S} \max_{r \in \mathcal{R}_s} d(\boldsymbol{x}_s^0, \boldsymbol{x}_r)$, minimizing $\boldsymbol{C}$ amounts to solving (a version of) the discrete R-center problem, (Çalık, Labbé, & Yaman, 2019; Espejo, Marín, & Rodríguez-Chía, 2015).

Several modelling options appear when exogenous counterfactual are sought. We will consider first the case in which the ambient space $\mathcal{X}$, in which counterfactuals are to be located, is a vector space of dimension J, i.e., we have J numerical features. In this case, a central role is played by distances, or, more generally, *gauges*, (Carrizosa & Plastria, 2008; Plastria, 2019; Plastria & Carrizosa, 2001, 2012). A gauge in $\mathbb{R}^J$ is a function $\sigma : \mathbb{R}^J \longrightarrow \mathbb{R}_+$ which is definite positive ($\sigma(\boldsymbol{u}) \geq 0 \forall \boldsymbol{u}, \sigma(\boldsymbol{u}) = 0$ iff $\boldsymbol{u} = \boldsymbol{0}$), it is positively homogeneous ($\sigma(\tau \boldsymbol{u}) = \tau \sigma(\boldsymbol{u}) \forall \tau \geq 0$) and it is subadditive ($\sigma(\boldsymbol{u} + \boldsymbol{u}') \leq \sigma(\boldsymbol{u}) + \sigma(\boldsymbol{u}')$). The unit ball of σ, i.e., the lower level set $\{\boldsymbol{u} : \sigma(\boldsymbol{u}) \leq 1\}$ is a convex compact set, containing the origin in its interior, and indicates the moves which cost just 1 unit. When σ is absolutely homogeneous ($\sigma(\tau \boldsymbol{u}) = |\tau| \sigma(\boldsymbol{u}) \forall \tau$), i.e., when the unit ball of σ is symmetric with respect to the origin, σ is a norm.

A natural choice for $\text{Dissimilarity}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}})$ is

$$\text{Dissimilarity}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \sum_{s=1}^{S} \sum_{r \in \mathcal{R}_s} \omega_s \pi(\sigma_s(\boldsymbol{x}_r - \boldsymbol{x}_s^0)), \tag{8}$$

where $\omega_s > 0$, π is a convex increasing function in $\mathbb{R}^+$, and σ_s is a gauge in $\mathbb{R}^J$. Under these conditions, (8) is convex in $\underline{\boldsymbol{x}}$, and thus the objective function is convex if $\mathcal{R}_s$ is fixed and not a decision variable, as in the one-for-one, the many-for-one or the one-for-all cases.

The convexity of π implies that movements are penalized more than linearly. Using different weights ω_s for different instances $\boldsymbol{x}_s^0$ allows stakeholders to ask for smaller perturbations for some individuals. This may be relevant in the framework of fairness, in which we may have records split into two groups, namely, those belonging to a sensitive group and the remaining ones, and a higher weight is given to individuals $\boldsymbol{x}_s^0$ in the sensitive group.

With respect to the choice of the gauges σ_s, norms, i.e., symmetric gauges, such as ℓ_p norms, are systematically used in the single-instance single-counterfactual models, (Kanamori et al., 2020; Russell, 2019; Wachter et al., 2017). However, in practice, increasing in δ units one feature may not be as costly as decreasing δ units the very same feature, (Karimi, Schölkopf, & Valera, 2021), and thus there is a need to depart from the state-of-the-art and address models with asymmetric gauges.

Plausible asymmetric gauges for this problem with good structural properties are, nevertheless, easy to build. A well-known family of asymmetric gauges are the so-called *skewed norms*, (Plastria, 1992), namely, gauges σ of the form

$$\sigma(\boldsymbol{u}) = \sigma_0(\boldsymbol{u}) + \boldsymbol{\eta} \boldsymbol{u}, \tag{9}$$

where σ_0 is a norm, and $\boldsymbol{\eta}$ is a fixed vector with $\sigma_0^0(\boldsymbol{\eta}) < 1$, where σ_0^0 is the dual norm of σ_0. Observe that σ is asymmetric unless $\boldsymbol{\eta} = 0$. See Plastria (1992) for a model through a skewed norm of the work expended when moving along an inclined plane, and Drezner and Drezner (2021), Plastria (1992) for the effort of flying under steady wind conditions.

Let us discuss two particular cases of skewed norms, namely the skewed ℓ_1 and ℓ_2, which yield much more realistic models than those obtained with symmetric norms, and are as tractable as their symmetric counterparts. Taking in (9) as σ_0 the ℓ_2 norm in $\mathbb{R}^J$, and $\boldsymbol{\eta}$ such that $\sigma_0^0(\boldsymbol{\eta}) = \sigma_0(\boldsymbol{\eta}) < 1$, the unit ball of σ is an ellipsoid whose center is the origin only for $\boldsymbol{\eta} = 0$, (Plastria, 1992). Another plausible alternative in (9) is to take σ_0 as the ℓ_1 norm in $\mathbb{R}^J$, and $\boldsymbol{\eta}$ with $\sigma_0^0(\boldsymbol{\eta}) = \max_j |\eta_j| < 1$, yielding

$$\sigma(\boldsymbol{u}) = \sum_{j=1}^{J} |u_j| + \sum_{j=1}^{J} \eta_j u_j = \sum_{j=1}^{J} \left((1 + \eta_j) \max(u_j, 0) + (1 - \eta_j) \max(-u_j, 0) \right), \tag{10}$$

and thus, the cost of increasing δ units the feature j is $\delta(1 + \eta_j)$, while the cost of decreasing δ units the same feature is $\delta(1 - \eta_j)$. Observe that (10) is used as loss function in *quantile regression,* (Yu, Lu, & Stander, 2003).

Hence, taking in (8) π affine and all σ_s in the form of (10), $\text{Dissimilarity}(\underline{\boldsymbol{x}}^0, \cdot)$ is piecewise linear, and its optimization can be done, after adding auxiliary variables, by optimizing a linear objective. This way we can address with the same methods as in the weighted ℓ_1 norm, (Russell, 2019; Wachter et al., 2017), a much more realistic case in which perturbations increasing vs decreasing a feature are not equally costly.

Another class of (asymmetric) gauges with good structural properties but seemingly unexplored in the literature of counterfactual analysis is obtained as an extension of (10), and called hereafter *quantile gauges*. Let us recall that a norm σ_0 in $\mathbb{R}^J$ is said to be absolute if, for all $\boldsymbol{u} \in \mathbb{R}^J$, one has $\sigma_0(\boldsymbol{u}) = \sigma_0(|\boldsymbol{u}|)$, where $|\boldsymbol{u}|$ is the vector $(|u_1|, \ldots, |u_J|)$, (Bauer, Stoer, & Witzgall, 1961). Observe that ℓ_p norms (and their convex combinations) are absolute norms. Given an absolute norm σ_0 in $\mathbb{R}^J$ and a vector $\boldsymbol{\eta} \in \mathbb{R}^J$ with $\max_j |\eta_j| < 1$, define $\sigma_{\boldsymbol{\eta}}$ as

$$\sigma_{\boldsymbol{\eta}}(\boldsymbol{u}) = \sigma_0 \left(|\boldsymbol{u}| + D_{\boldsymbol{\eta}} \boldsymbol{u} \right), \tag{11}$$

where $D_{\boldsymbol{\eta}}$ is the diagonal matrix with $\boldsymbol{\eta}$ in its diagonal. As before, observe that $\sigma_{\boldsymbol{\eta}}$ is asymmetric unless $\boldsymbol{\eta} = 0$, and also it is a gauge. Note also that, if we take as σ_0 the ℓ_1 norm, we obtain the skewed norm in (10).

Using in (8) each σ_s as a quantile gauge as (11), $\text{Dissimilarity}(\underline{\boldsymbol{x}}^0, \cdot)$ is a sum of (increasing convex functions of) gauges, its minimization being a slight variant of the problem of minimizing a sum of ℓ_p norms, for which (Xue & Ye, 1997, 2000) give polynomial time procedures based on the construction of logarithmically homogeneous self-concordant barrier functions. Moreover, if we take as σ_0 the ℓ_2 norm and $\pi(t) = t^2$, one obtains

$$\begin{aligned} &\text{Dissimilarity}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \\ &= \sum_{s=1}^{S} \sum_{r \in \mathcal{R}_s} \omega_s \sum_{j=1}^{J} \left((1 + \eta_j) \max(x_{rj} - x_{sj}^0, 0) + (1 - \eta_j) \max(x_{sj}^0 - x_{rj}, 0) \right)^2. \end{aligned} \tag{12}$$

If the allocations defining $\mathcal{R}_s$ are fixed, as in the one-for-one, the many-for-one or the one-for-all models, the function in (12) is convex piecewise quadratic, and can be optimized as the squared Euclidean distance ℓ_2^2 has been in the single-instance single-counterfactual case in the literature, but now addressing the asymmetry which penalizes differently the increase and decrease of the different features. Moreover, in the one-for-many rule, in which allocations (and thus $\mathcal{R}_s$) are decision variables, one obtains (12) is a slight generalization of the minimum-sum-of-squares-clustering problem, (Aloise et al., 2012; Liberti & Manca, 2022; Piccialli, Sudoso, & Wiegele, 2022).

As a final remark on our discussion on quantile gauges, we should mention that they can be easily extended to accommodate causality relations between features (Karimi, von Kügelgen, Schölkopf, & Valera,

2022; Mahajan, Tan, & Sharma, 2019; Pearl, 2009). In its simpler form, assume that, if feature i is perturbed from x_i to $x_i + \delta$, then feature j is perturbed automatically (for free) from x_j to $x_j + h_{ij}\delta$. In other words, paying for a perturbation $\boldsymbol{\delta}$ would move the coordinates from $\boldsymbol{x}^0$ to $\boldsymbol{x}^0 + \boldsymbol{\delta}(I + H)$, where I denotes the identity matrix. Assuming $I + H$ is regular, this implies paying for $\boldsymbol{\delta} := (\boldsymbol{x} - \boldsymbol{x}^0)(I + H)^{-1}$ would move the point from $\boldsymbol{x}^0$ to $\boldsymbol{x}$, and thus in the cost function one should replace quantile gauges σ_s by σ_{H_s}, defined as

$$\sigma_{H_s}(\boldsymbol{u}) = \sigma_s\left(\boldsymbol{u}(I + H_s)^{-1}\right). \tag{13}$$

The discussion above applies for the common case in which the ambient space $\mathcal{X}$ is a J-dimensional vector space. When $\mathcal{X}$ has a different nature, other distance measures are to be used, see Kaufman and Rousseeuw (2009). In the presence of mixed data, including numerical as well as categorical data, Gower distance, (Gower, 1971), a weighted sum of a distance measure on the numerical features and Hamming-type measures for categorical features is the most popular choice, see also (Brughmans, Leyman, & Martens, 2021; Wilson & Martinez, 1997). When $\mathcal{X}$ consists of time series and functional data, (Esling & Agon, 2012; Xing, Pei, & Keogh, 2010), the dissimilarity between functions can be measured by some integrated Euclidean distance (easily extended to integrated quantile gauges), and, in case functions are inspected at a different speed, by using the Dynamic Time Warping distance, (Carrizosa, Ramírez-Ayerbe, & Romero Morales, 2023). If features represent frequencies of a discrete event, dissimilarities such as the chi-squared distance, (Carrizosa, Guerrero, & Romero Morales, 2023), can be used. When dealing with text data (Ramon et al., 2020; Tolkachev et al., 2022), an encoder may be used to map the instances into $\mathbb{R}^J$, where counterfactuals are sought. The models of norms or gauges discussed above for data in $\mathbb{R}^J$ may be no longer meaningful to measure dissimilarities between mapped instances and counterfactuals, and instead proximity measures such as the cosine similarity can be used. Once the counterfactual has been calculated in the embedding space, a decoder is needed in order to obtain the final counterfactual instance in the original space. Finally, for image data, see Vermeire, Brughmans, Goethals, de Oliveira, and Martens (2022) for an extensive survey.

In addition to the dissimilarity, the cost function $\boldsymbol{C}$ can also capture how complex the perturbation to move from $\underline{\boldsymbol{x}}^0$ to $\underline{\boldsymbol{x}}$ is. In this case, the cost function may have the form:

$$\boldsymbol{C}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \text{Dissimilarity}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) + \lambda_c \text{Complexity}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}), \tag{14}$$

with $\lambda_c > 0$.

When the ambient space $\mathcal{X}$ is a J-dimensional vector space, complexity is usually measured through the number of features one needs to perturb, and it is therefore the complement of sparsity. Sparsity ensures more interpretability and it has been argued that people prefer explanations where fewer features are changed (Miller, 2019).

Sparsity can be modelled at instance level or at group level. For a given instance $\boldsymbol{x}^0$, this can be done directly through the ℓ_0 norm, $\|\boldsymbol{x}^0 - \boldsymbol{x}\|_0$, counting the number of features changed, or, in order to retain convexity, one can use the ℓ_1 norm instead. If sparsity at the group level is sought, one may minimize the total number of features changed for at least one instance in $\underline{\boldsymbol{x}}^0$. In other words, we may minimize γ_0 defined as:

$$\gamma_0(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \left\| \left(\max_i |x_{ij}^0 - x_{ij}| \right)_{j=1}^{J} \right\|_0, \tag{15}$$

where x_{ij} denotes the value of feature j in $\boldsymbol{x}_i$. Notice that $\left(\max_i |x_{ij}^0 - x_{ij}|\right)_{j=1}^{J}$ is a vector of J components, where each component is the maximum change in feature j across all the instances. Then, we count the number of features changed globally with the ℓ_0 norm.

Note that more sophisticated forms of sparsity may be required in the presence of complex data, as in Carrizosa, Galvis Restrepo, and Romero Morales (2021), Carrizosa, Mortensen, Romero Morales, and Sillero-Denamiel (2022), Carrizosa, Nogales-Gómez, and Romero Morales (2017) for categorical features.

When $\mathcal{X}$ is defined by linear or convex combinations of observed datapoints, sparsity can be measured as the ℓ_0 norm of the vector of these coefficients, as in Carrizosa, Ramírez-Ayerbe, and Romero Morales (2023). In the extreme case, the counterfactuals built in this way are observed datapoints themselves.

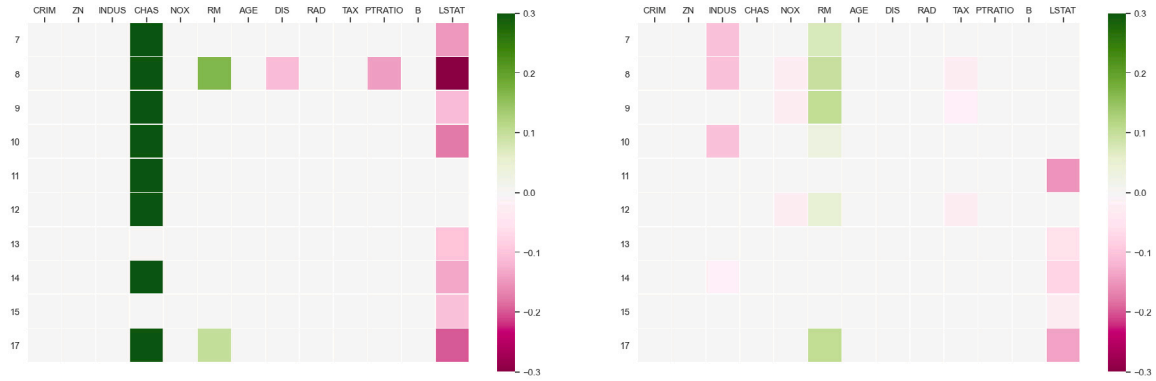
To have a smooth formulation of these measures of complexity, tractable with mixed integer optimization solvers, binary decision variables are introduced indicating whether a given feature can be perturbed, and constraints are expressed either via the usual big-M method or other strategies to address indicator constraints such as those in Belotti et al. (2016), Bomze and Peng (2023), Han et al. (2023), Wei et al. (2022).

A third term is to be added to the cost function $\boldsymbol{C}$ in the many-for-one model, if several counterfactuals are to be built for an instance $\boldsymbol{x}^0$, (Mothilal et al., 2020), and maximal diversity is sought. Different notions of diversity have been introduced in the Operations Research literature, and seem adequate for the problem at hand. The most popular criteria are the maxsum and maxmin, (Erkut & Neuman, 1989; Landete, Peiró, & Yaman, 2023; Lozano-Osorio, Martínez-Gavara, Martí, & Duarte, 2022; Parreño, Álvarez-Valdés, & Martí, 2021; Pisinger, 2006), which, respectively, seek the maximization of the average and minimum distance between the counterfactuals, see Martí, Martínez-Gavara, Pérez-Peló, and Sánchez-Oro (2022), Parreño et al. (2021) for recent surveys. In the case of exogenous counterfactuals having as ambient space $\mathcal{X}$ a finite-dimensional vector space, the distances used in the abovementioned models can be taken naturally as norms, such as the Euclidean distance or the Mahalanobis distance if correlations are taken into account. Since $\boldsymbol{C}(\underline{\boldsymbol{x}}^0, \cdot)$ is not convex, the literature has mostly focused on heuristic approaches, based on solving sequentially a collection of single-instance single-counterfactual problems, (Karimi et al., 2020; Russell, 2019; Ustun et al., 2019; Wachter et al., 2017). Nonetheless, optimizing $\boldsymbol{C}$ can be expressed as a problem of optimizing the difference of convex functions, and thus the machinery of difference of convex (d.c.) optimization, (Le Thi & Pham Dinh, 2018, 2023), is applicable.

3. Numerical illustrations

In this section, a collection of optimization models for group counterfactual analysis are illustrated, covering all the possible allocation rules between counterfactuals and instances. We have made different choices for the ingredients defining Problem (GroupCEhard), namely, we model exogenous explanations in a finite-dimensional space, with and without linking constraints between the counterfactuals, for both linear classifiers and Additive Tree Models (ATM), and different cost criteria. We will use the probability criterion (6) as described in Section 2.5, to ensure that each counterfactual explanation has a high enough probability of being classified in the positive class. All the numerical illustrations are done using a real-world dataset, namely the `Boston housing` dataset (Harrison & Rubinfeld, 1978), which can be accessed, e.g., from the scikit-learn library (Pedregosa et al., 2011). There are 506 instances corresponding to houses and $J = 13$ features, of which 12 are numerical and 1 is binary. The description of the dataset can be found in the Supplementary Material of this paper. All features are normalized, so that all features share the scale $[0, 1]$. The positive class consists of houses which have a high price.

All optimization models have been implemented using Python 3.8 and as a solver Gurobi 9.0 (Gurobi Optimization, 2021). Our numerical experiments have been conducted on a PC, with an Intel R CoreTM i7-1065G7 CPU @ 1.30 GHz 1.50 GHz processor and 16 GB RAM. The operating system is 64 bits. The source code and the data to reproduce all results, as well as all Figures in full size, are available at https://github.com/jasoneramirez/GroupCE.



(a) Logistic regression (b) Random forest

Fig. 2. One-for-one counterfactual explanations for instances $\boldsymbol{x}_s^0$ in Table 2 in the Supplementary Material for two classifiers, the logistic regression and the random forest. The explanations have been calculated solving model (16)–(17) with $\lambda_{ind} = 0.01$, $\lambda_{glob} = 0$ and $\nu = 0.5$. The feature perturbations are displayed.

3.1. The one-for-one allocation model

In this section, we focus on the one-for-one allocation model for two types of classifiers, namely, the logistic regression and the random forest. We will show illustrations for which there are no linking constraints between the counterfactuals, i.e., the cost function is separable and $\underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0) = \prod_{s=1}^{S} \mathcal{X}(\boldsymbol{x}_s^0)$, and thus, each counterfactual is found by solving a single-instance single-counterfactual problem, as well as others in which the counterfactuals are linked to ensure low complexity of the explanations, change of few features, or similarity between counterfactuals of similar instances.

We will use the cost criterion combining the ℓ_2^2, ℓ_0 and γ_0 as in (15), yielding

$$\min_{\underline{\boldsymbol{x}} \in \underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0)} \quad \sum_{s=1}^{S} \|\boldsymbol{x}_s^0 - \boldsymbol{x}_s\|_2^2 + \lambda_{ind} \sum_{s=1}^{S} \|\boldsymbol{x}_s^0 - \boldsymbol{x}_s\|_0 + \lambda_{glob} \gamma_0(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) \tag{16}$$

$$\text{s.t.} \quad f(\boldsymbol{x}_s) \geq \varphi^{-1}(\nu) \quad \forall s = 1, 2, \ldots, S, \tag{17}$$

with $\lambda_{ind}, \lambda_{glob} \geq 0$.

First, we consider the most simple case, the separable one. This is the case in which the literature is mostly focused (Kanamori et al., 2020; Parmentier & Vidal, 2021), where the cost function is separable on the instances, i.e., $\lambda_{glob} = 0$, and there are no other linking constraints between the counterfactuals $\boldsymbol{x}_s$ and $\boldsymbol{x}_{s'}$ with $s \neq s'$, i.e., $\underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0) = \prod_{s=1}^{S} \mathcal{X}(\boldsymbol{x}_s^0)$. In such case, Problem (16)–(17) is equivalent to solving S optimization problems, one per instance, yielding:

$$\min_{\boldsymbol{x}_s \in \mathcal{X}(\boldsymbol{x}_s^0)} \quad \|\boldsymbol{x}_s^0 - \boldsymbol{x}_s\|_2^2 + \lambda_{ind} \|\boldsymbol{x}_s^0 - \boldsymbol{x}_s\|_0 \tag{18}$$

$$\text{s.t.} \quad f(\boldsymbol{x}_s) \geq \varphi^{-1}(\nu). \tag{19}$$

For (piecewise) linear score-based classifiers described in Section 2 and assuming that $\mathcal{X}(\boldsymbol{x}_s^0)$ is a polyhedron with eventually some integer decision variables, Problem (18)–(19) above is a mixed integer convex quadratic problem with linear constraints. Specifically, for a logistic regression model with φ defined as (5), constraint (19) takes the form $\boldsymbol{w}\boldsymbol{x}_s + b \geq -\log\left(\frac{1-\nu}{\nu}\right)$. For ATM models, such as random forests, constraint (19) can be modelled through additional binary decision variables and a set of linear constraints, see Carrizosa et al. (2024).

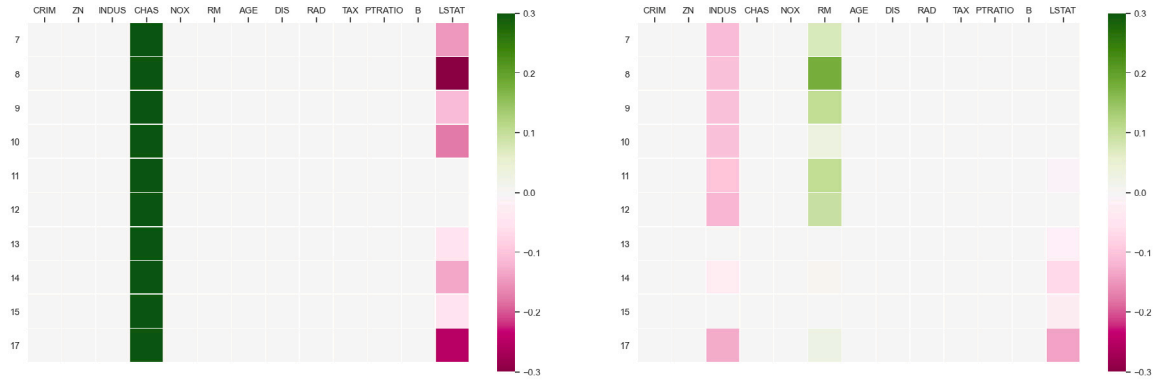
The presence of linking constraints, such as those modelling global sparsity, i.e., considering $\lambda_{glob} > 0$ in the cost function, destroys the separability of Problem (16)–(17), which thus needs to be considered as a whole. The separability is also destroyed when the Lipschitz continuity constraint (3) is added in order to impose continuity.

In the following, we provide some illustrations using the `Boston housing` dataset for the logistic regression model and a random forest with $T = 100$ trees, maximum depth 3 and $w^t = \frac{1}{T}$, for all $t = 1, \ldots, 100$.

First, we consider the simplest case for the one-for-one allocation model (16)–(17), i.e., the separable case, where $\lambda_{glob} = 0$ and $\underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0) = \prod_{s=1}^{S} \mathcal{X}(\boldsymbol{x}_s^0)$, implying the resolution of (18)–(19) for each $s = 1, 2, \ldots, S$. We consider $\lambda_{ind} = 0.01$, $\nu = 0.5$ and in $\mathcal{X}(\boldsymbol{x}_s^0)$ we only impose the binary nature of variable CHAS and lower and upper bounds of each feature to be the minimum and maximum value observed across the 506 observations, respectively. We calculate the counterfactual explanations for 10 instances $\boldsymbol{x}_s^0$, $s = 1, \ldots, 10$, that were given a probability below 0.5 by both the logistic regression and the random forest, i.e., $\varphi(f(\boldsymbol{x}_s^0)) < 0.5$, $s = 1, \ldots, 10$, specifically the instances of the dataset whose values are detailed in Table 2 in the Supplementary Material. The feature perturbations of the explanations are shown in Fig. 2. Obviously, the counterfactual explanations differ across classifiers, but in both cases LSTAT is a feature that it is often perturbed, as well as RM.

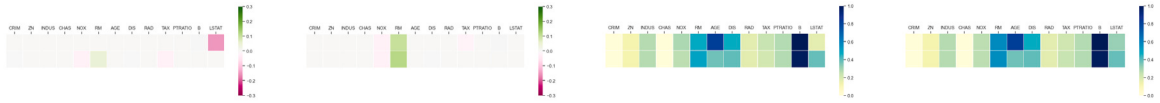
Second, we consider the case where the cost function of Problem (16)–(17) is not separable, as we aim to maximize the global sparsity. Specifically, we consider $\lambda_{ind} = 0$ and $\lambda_{glob} = 0.1$. For the same classifiers, the logistic regression model and the random forest, the feature perturbations for the counterfactual explanations for the same instances as before are shown in Fig. 3. Notice the difference between Figs. 2(a) and 3(a), where in the first case 5 features had to be changed globally in order to obtain the 10 explanations, whereas in the second case, only two features are changed in total. We now consider the case where the cost function is separable, with $\lambda_{ind} = 0.01$ and $\lambda_{glob} = 0$, but the Lipschitz continuity constraint (3) with $\tau = 10$ is imposed, linking the counterfactuals. This is calculated for instances $\boldsymbol{x}_5^0$ and $\boldsymbol{x}_6^0$ from Table 2 in the Supplementary Material, and for the random forest. The difference between imposing or not the Lipschitz continuity constraint is shown in Fig. 4, where both the perturbations and feature values are displayed. The distance between the counterfactual explanations without imposing the Lipschitz continuity constraint is 0.51, whereas when the constraint is imposed the distance between explanations reduces to 0.46.

To end, we illustrate how the lower bound ν imposed on the probabilities $P(\boldsymbol{x}_s)$ affects the choice of the counterfactuals. In all the counterfactuals calculated before, the value of ν imposed has been 0.5. We show in Fig. 5(a) the Pareto frontier when the value of ν changes for the case where $\lambda_{ind} = 0$, $\lambda_{glob} = 0.1$, the classifier is the logistic regression, and no other linking constraints have been imposed, i.e., the



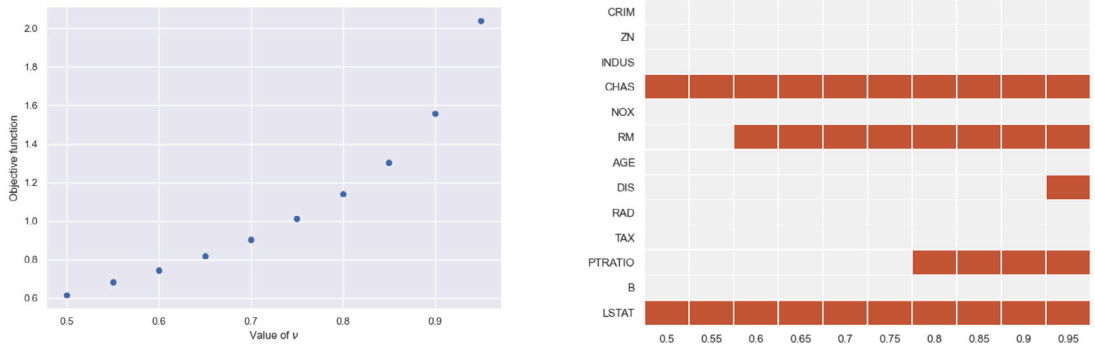
(a) Logistic regression (b) Random forest

Fig. 3. One-for-one counterfactual explanations for instances x_s^0 in Table 2 in the Supplementary Material for two classifiers, the logistic regression and the random forest. The explanations have been calculated solving model (16)–(17) with $\lambda_{ind} = 0$, $\lambda_{glob} = 0.1$ and $\nu = 0.5$. The feature perturbations are displayed.



(a) Perturbations without (3) (b) Perturbations with (3) (c) Feature values without (3) (d) Feature values with (3)

Fig. 4. One-for-one counterfactual explanations for instances x_5^0 and x_6^0 in Table 2 in the Supplementary Material where the classifier is the random forest. The explanations have been calculated solving model (16)–(17) with $\lambda_{ind} = 0.01$, $\lambda_{glob} = 0$ and $\nu = 0.5$. Features perturbations are displayed on the two pictures on the left, with the Lipschitz continuity constraint (3) for $\tau = 10$ and without this constraint, respectively, whereas in the two right pictures the corresponding features values are displayed.



(a) Pareto frontier (b) Features perturbed

Fig. 5. On the left, the objective function of Problem (16)–(17) vs threshold value ν, when $\lambda_{ind} = 0$, $\lambda_{glob} = 0.1$, for the logistic regression model. On the right, the features used in the counterfactual explanations.

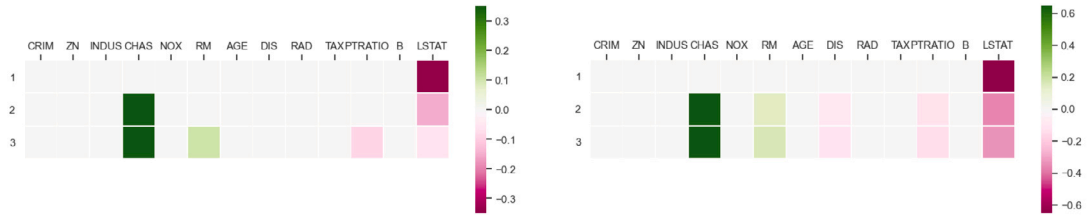
same case as in Fig. 3(a). The counterfactuals have been calculated for the 10 instances in Table 2 in the Supplementary Material. Fig. 5(b) shows the features that need to be changed for the different values of ν. We can see how the stricter constraint (17) is, i.e., the larger the value of ν is, the more features we perturb.

3.2. The many-for-one allocation model

In this section, we consider the many-for-one allocation rule when there are no linking constraints between counterfactuals associated with different instances. Thus, it boils down to the problem of calculating the counterfactuals for each instance, separately. When considering more than one explanation to an instance, some type of diversity is sought between the different explanations. We will illustrate one possible choice of diversity. Considering a specific feature of interest, we will impose that each of the counterfactuals have different values on this feature. In this way, we ensure that we have different options at hand to increase the probability, without relying on this feature to be modified to a specific value, that may be hard to obtain.

The values allowed in each case for the considered feature are imposed in the set $\mathcal{X}(x_s^0)$, that changes for each explanation. Thus solving the problem of finding R counterfactuals explanations for an instance, is equivalent to solving R optimization problems as follows:

$$\min_{x_r \in \mathcal{X}^r(x_s^0)} \quad \|x_s^0 - x_r\|_2^2 + \lambda_{ind}\|x_s^0 - x_r\|_0 \tag{20}$$



(a) Perturbations for $\boldsymbol{x}_1^0$ (b) Perturbations for $\boldsymbol{x}_2^0$

Fig. 6. Many-for-one counterfactual explanations with $R = 3$ for instances $\boldsymbol{x}_1^0$ and $\boldsymbol{x}_2^0$ in Table 2 in the Supplementary Material where the classifier is the logistic regression model. The explanations have been calculated solving model (20)–(21) with $\lambda_{ind} = 0.01$ and $\nu = 0.5$. Each row is an explanation, calculated by adding one constraint from (22) in $\mathcal{X}^r(\boldsymbol{x}_s^0)$. Features perturbations are displayed.

$$\text{s.t.} \quad f(\boldsymbol{x}_r) \geq \varphi^{-1}(\nu), \tag{21}$$

where $\mathcal{X}^r(\boldsymbol{x}_s^0)$ corresponds to the feasible set for the rth explanation associated to $\boldsymbol{x}_s^0$.

For the `Boston housing` dataset, we choose LSTAT as the feature of interest. We will calculate for each instance $R = 3$ counterfactuals, imposing LSTAT to be in the first quartile, between the first and third quartile, and above the third quartile, respectively. Specifically, we add to $\mathcal{X}^r(\boldsymbol{x}_s^0)$ one of the following constraints:

$$(\boldsymbol{x}_r)_{\text{LSTAT}} \leq Q_1 \quad \text{or} \quad Q_1 < (\boldsymbol{x}_r)_{\text{LSTAT}} \leq Q_3 \quad \text{or} \quad (\boldsymbol{x}_r)_{\text{LSTAT}} > Q_3. \tag{22}$$

We illustrate this for the logistic regression model, the first two instances in Table 2 in the Supplementary Material and $R = 3$. In Fig. 6(a) the three counterfactuals for instance $\boldsymbol{x}_1^0$ are displayed. One can see that imposing in the third explanation a high value for LSTAT, i.e., not allowing this feature to be decreased too much, results then in the necessity of moving other features to reach the desired probability of being classified in the positive class. The same happens for the counterfactuals for instance $\boldsymbol{x}_2^0$, displayed in Fig. 6(b).

3.3. The one-for-many and one-for-all allocation models

In this section a model for the one-for-many allocation rule and another for the one-for-all one are presented and illustrated. Whereas in the later the only decision variable is the location of the counterfactual, in the former, both the location of counterfactuals and the assignment of counterfactuals to instances are to be decided. In this case, the constraints imposed on the explanations cannot depend on specific instances $\underline{\boldsymbol{x}}^0$, thus $\underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0) = \underline{\mathcal{X}}$. Of course, there may be constraints imposed on all counterfactuals generally, and also, cannot link constraints could be added as well. As before, one way to define $\boldsymbol{P}$ is as (6), ensuring in this way that each counterfactual explanation has a high enough probability. As cost function we take the ℓ_2^2.

To formulate the one-for-many allocation rule, binary variables y_{sr} are introduced, where $y_{sr} = 1$ if instance $\boldsymbol{x}_s^0$ is assigned to counterfactual $\boldsymbol{x}_r$ and 0 otherwise. For a linear classifier such as the logistic regression model, we can formulate Problem (GroupCEhard) as follows:

$$\min_{\underline{\boldsymbol{x}} \in \underline{\mathcal{X}}, \boldsymbol{y}} \quad \sum_{r=1}^{R} \sum_{s=1}^{S} y_{sr} \|\boldsymbol{x}_s^0 - \boldsymbol{x}_r\|_2^2 \tag{23}$$

$$\text{s.t.} \quad \boldsymbol{w}\boldsymbol{x}_r + b \geq \varphi^{-1}(\nu) \qquad \forall r = 1, 2, \ldots, R \tag{24}$$

$$\sum_{r=1}^{R} y_{sr} = 1 \qquad \forall s = 1, 2, \ldots, S \tag{25}$$

$$y_{sr} \in \{0, 1\} \qquad \forall s = 1, 2, \ldots, S \quad \forall r = 1, 2, \ldots, R. \tag{26}$$

Constraint (24) guarantees that each counterfactual explanation has at least a probability of being classified in the positive class of ν, while constraints (25) and (26) ensure that each instance is assigned to exactly one counterfactual explanation.

To solve this problem an alternating algorithm can be used, similar to Lloyd's algorithm (Lloyd, 1982), where two phases arise: the allocation of the instances to the counterfactual instance that minimizes the cost function, and the location of the explanations, where the counterfactual explanation is calculated in the case where the clusters are already known. Since we are dealing with exogenous counterfactuals, the chosen cost function is the squared distance and we are considering a logistic regression model, this problem has a very similar structure to the classical minimum-sum-of-squared-distances problem, with the addition of linear constraints (24). The problem becomes more complex when, for instance, other classifiers are chosen, such as an additive tree model.

Setting $R = 1$ in (23)–(26), one obtains the model for the one-for-all allocation rule. In such case, the variables y_{sr} are not needed, as all the instances are assigned to the same counterfactual explanation, and there is no allocation phase.

To visualize the output of model (23)–(26), we consider $\boldsymbol{x}_s^0$, $s = 1, \ldots, 295$, to be all the instances in the `Boston housing` dataset that were given by the logistic regression model a probability of belonging to the positive class below 0.5, i.e.,

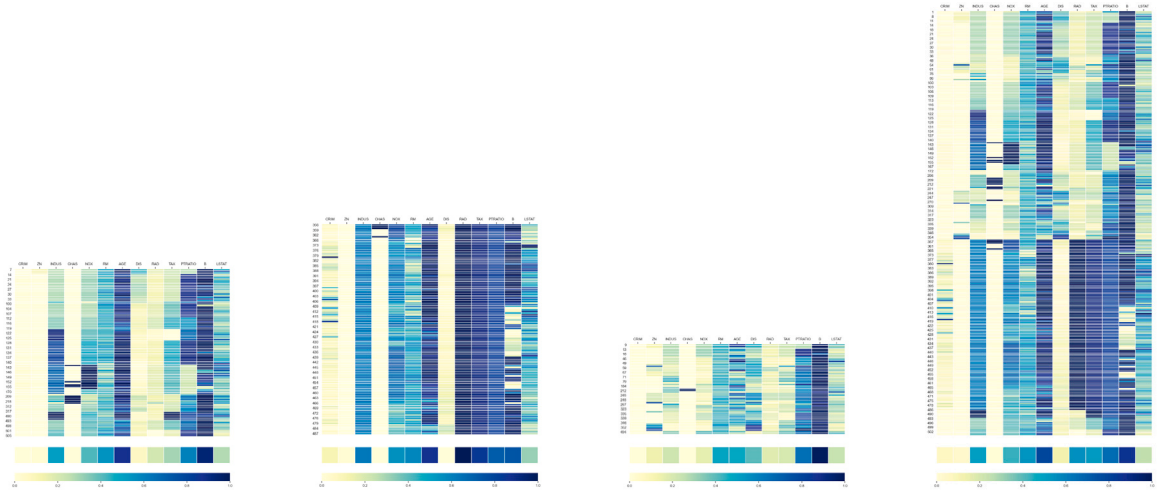
$$\varphi(f(\boldsymbol{x}_s^0)) < 0.5, \quad \forall s = 1, 2, \ldots, S. \tag{27}$$

We calculate for them $R = 3$ counterfactual explanations in the many-for-one case, and one single explanation in the all-for-one case. We impose $\nu = 0.5$. Fig. 7 displays the feature values of the clusters and the explanations.

For the many-for-one allocation rule, we note that, the first cluster is characterized by high values of the feature RM and AGE and lower values of features DIS, RAD and TAX, whereas cluster 2 is characterized by high values of RAD, TAX and PTRATIO. Cluster 3 can be characterized by lower values of most features except B. For the one-for-all allocation rule, the explanation that defines the benchmark for the positive class resembles the counterfactual associated with the first cluster in the case $R = 3$, but with higher values of the features RAD and TAX. Similar conclusions can be derived for other choices of ν, as illustrated in the Supplementary Material.

4. Conclusions

In this paper we have focused on counterfactual explanations, an important class of explanations in Supervised Classification, following a stakeholders perspective. We have formulated the group counterfactual problem as the bi-objective model (GroupCE). We have provided a critical discussion on how the different ingredients defining this problem. In general, finding efficient solutions to this problem calls for solving Mixed Integer Nonlinear Programming formulations. We have related those to classic problems in the Continuous and Discrete Location Analysis literature, such as the p-median, the p-center, or the minimum-sum-of-squared-distances problems, and have highlighted the novel elements, such as constraints to ensure that the



(a) Cluster 1 and its counterfactual

(b) Cluster 2 and its counterfactual

(c) Cluster 3 and its counterfactual

(d) All instances and its counterfactual

Fig. 7. One-for-many and one-for-all analysis for $\underline{x}^0$ as in (27) for the logistic regression and $R = 3$ and $R = 1$ respectively. We consider Problem (23)–(26) with $\nu = 0.5$. The plot represents the feature values of the counterfactual explanations (below) and their associated clusters (above).

probability of being classified in the positive class is high enough. The link established between Counterfactual Analysis and Location Analysis may allow to transfer to the world of Counterfactual Analysis the well developed algorithmic machinery of Location Analysis. Conversely, unexplored versions of Location Analysis problems appear motivated by their counterparts in Counterfactual Analysis.

On top of exploiting the relationship between these two fields, of rather different scientific maturity, this paper poses a number of lines for future research for the Operations Research community, some outlined below.

The practical success of Counterfactual Analysis in Supervised Classification as developed in this paper will be conditioned to three premises: the counterfactuals $\underline{x}$ built for the tuple of instances $\underline{x}^0$ should be feasible, the perturbations suggested on $\underline{x}^0$ to yield $\underline{x}$ should be doable, and counterfactuals, used as benchmarks for records labelled by the classifier used in the negative class should be indeed in the positive class -a condition which may be differ from being classified as positive by the classifier-. Section 2 has given an overview of different judicious choices for the model ingredients to fulfil the three abovementioned premises. However, if robustness of the counterfactuals generation procedure is an issue, an extra effort can be done, at the expense of possibly making the resulting problems less tractable from the computational viewpoint. Let us analyse the three premises.

Concerning feasibility, obtaining counterfactuals which correspond to data that have already been observed, and are thus realistic, is done if one restricts the ambient space $\mathcal{X}$ to a training set, and thus one is considering endogenous counterfactuals. When counterfactuals are exogenous, constraints in Section 2.3 force the counterfactuals to be, if not close to an element of a set D of observed points, close to some convex combination of such points.

Making perturbations doable can be controlled via (2), which allows the stakeholder to define upper bounds on the magnitude $d(\boldsymbol{x}_s^0, \boldsymbol{x}_r)$ of each individual perturbation. However, costs may be excessively underestimated if the weights ω_s in (8), the asymmetry coefficients $\boldsymbol{\eta}$ in (11), or the causality matrices H_s in (13) are not accurate. A robust approach, in which $\boldsymbol{\omega}$, $\boldsymbol{\eta}$ and H_s are assumed to be uncertainty sets, e.g., ellipsoidal or polyhedral (Ben-Tal & Nemirovski, 1999; Bertsimas, Brown, & Caramanis, 2011; El Ghaoui & Lebret, 1997), may be then advisable. The resulting optimization problems, even for the simplest case of the single-instance single-counterfactual remain unexplored.

The most critical premise to make counterfactual analysis reliable is that counterfactuals should be members of the positive class. Most of the literature on counterfactual analysis focuses on finding a solution to the single-objective problem in which the cost is minimized, while the counterfactual $\boldsymbol{x}$ is deterministically classified in the positive class, by imposing $P(\boldsymbol{x}) \geq \tau$ for a given threshold τ, say $\tau = 0.5$. Although there are studies in which two objectives are considered and combined through a scalar, they do not model the probability of positive associated with the classifier but the value of the score function. With our approach we have a direct control on the probability $P(\boldsymbol{x})$, and thus, on function $\boldsymbol{P}$. Therefore, by generating the Pareto frontier of Problem (GroupCE), the stakeholder will have the chance to trade off costs and the probability of positive associated with the classifier. A deeper algorithmic analysis of the bi-objective problem for the different allocation rules discussed in Section 2.2 is worthwhile.

On top of this, we are assuming in Section 2.4 that the classifier is given by a fixed score function f. It is expected that, if the classifier is retrained, e.g., when the process is used over time and new data become available, the score f and therefore the probability $P(\boldsymbol{x})$ will change (Dutta, Long, Mishra, Tilli, & Magazzeni, 2022; Ferrario & Loi, 2022; Forel, Parmentier, & Vidal, 2022; Upadhyay, Joshi, & Lakkaraju, 2021). This means that we may have some uncertainty on f, and $P(\boldsymbol{x})$ in (4) needs to account for this. For instance, we can replace (4) by $P(\boldsymbol{x}) = \min_{f \in \mathcal{F}} \varphi(f(\boldsymbol{x})) = \varphi(\min_{f \in \mathcal{F}} f(\boldsymbol{x}))$ for some uncertainty set of score functions $\mathcal{F}$. This leads to tractable models when f is affine as in $f(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x} + b$, and $\mathcal{F}$ is an ellipsoid, or, for more general types of scoring functions, when $\mathcal{F}$ is a finite set, implying that we have not one score function but several. This situation correspond to, for instance, different runs of a random forest, to a nonlinear SVM with different kernels or parameters, or the score functions corresponding to different classification methods.

Finally, if one suspects that, if the counterfactual $\boldsymbol{x}$ is chosen, instead of $\boldsymbol{x}$ some $\boldsymbol{z}$ will be implemented, where $\boldsymbol{z}$ is in a neighbourhood of $\boldsymbol{x}$, the approach is made robust if in the definition of P one uses $P(\boldsymbol{x}) = \min_{\boldsymbol{z} \in B_\varepsilon(\boldsymbol{x})} \varphi(f(\boldsymbol{z}))$, where $B_\varepsilon(\boldsymbol{x}) = \{\boldsymbol{z} \in \mathcal{X} : d(\boldsymbol{x}, \boldsymbol{z}) \leq \varepsilon\}$, for some distance or dissimilarity function d and some $\varepsilon > 0$, see Maragno et al. (2023) for further details.

In addition to enhancing the explainability of the output of machine learning algorithms, i.e., detecting the important features to the classification task and how to change features to enhance outcomes, there is also the urgent need to enhance the fairness of these algorithms (Goethals, Martens, & Calders, 2023; Gupta, Nokhiz, Roy, & Venkatasubramanian, 2019; Haldar, Cunningham, & Ferhatosmanoglu, 2022; Kusner, Loftus, Russell, & Silva, 2017; Von Kügelgen et al.,

2022; Zafar, Valera, Gomez-Rodriguez, & Gummadi, 2019). The data available to train algorithms may suffer from bias against a sensitive group, defined, e.g., by gender (females) or income (low income). To reduce the danger that the algorithm amplifies the bias seen in historical data, a plethora of methodologies have been developed in recent years, see Mehrabi, Morstatter, Saxena, Lerman, and Galstyan (2022), Mitchell, Potash, Barocas, D'Amour, and Lum (2021), Pessach and Shmueli (2022) for recent reviews. Unfairness has also been reported for the single-instance and single-counterfactual model. For instance, disparity in the costs $C(\boldsymbol{x}^0, \boldsymbol{x})$ to perturb $\boldsymbol{x}^0$ to yield $\boldsymbol{x}$ is reported for the `Adult` dataset in Ustun et al. (2019), which is a classic dataset in the fairness literature (Fabris, Messina, Silvello, & Susto, 2022), where females incur on higher average costs than males. In the framework proposed in this paper, we are able to have a more direct control on this disparity since we build the counterfactuals for both the sensitive and the non-sensitive groups simultaneously using a single model. Indeed, we can penalize the costs associated with the sensitive groups much higher than for the non-sensitive group with the ω_s weight in (8). An even more direct control, at the expense of making the optimization problems much less tractable, is to impose that the distribution of costs of the sensitive and the nonsensitive groups are *similar*, where their dissimilarity can be measured with the Wasserstein distance. In the latter case, the structure of the problem is affected having a constraint where the left hand side is defined by another optimization problem, the so-called optimal transport problem.

The concepts of explainability and fairness are expanding beyond Supervised Classification (Barocas, Selbst, & Raghavan, 2020; De-Arteaga, Feuerriegel, & Saar-Tsechansky, 2022; Korikov & Beck, 2021; Korikov, Shleyfman, & Beck, 2021; Olson, Khanna, Neal, Li, & Wong, 2021), although the examples are still very scarce (Verma et al., 2022). Extending group counterfactual explanations beyond classification is a nontrivial task, even in the single-instance single-counterfactual case for which bilevel programs arise (Bogetoft, Ramírez-Ayerbe, & Romero Morales, 2024). This, as the other challenges discussed in this paper, deserves further attention from the Operations Research community, whose expertise will strongly improve the algorithmic part of the burgeoning field of Counterfactual Analysis.

Acknowledgements

This research has been financed in part by research projects EC H2020 MSCA RISE NeEDS (Grant agreement ID: 822214), FQM-329, P18-FR-2369 and US-1381178 (Junta de Andalucía), and PID2019-110886RB-I00 and PID2022-137818OB-I00 (Ministerio de Ciencia, Innovación y Universidades, Spain). This support is gratefully acknowledged.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejor.2024.01.002.

References

Aloise, D., Hansen, P., & Liberti, L. (2012). An improved column generation algorithm for minimum sum-of-squares clustering. *Mathematical Programming*, *131*, 195–220.

Artelt, A., & Hammer, B. (2019). On the computation of counterfactual explanations–a survey. arXiv preprint arXiv:1911.07749.

Artelt, A., Vaquet, V., Velioglu, R., Hinder, F., Brinkrolf, J., Schilling, M., et al. (2021). Evaluating robustness of counterfactual explanations. In *2021 IEEE symposium series on computational intelligence* (pp. 1–9). IEEE.

Ates, E., Aksar, B., Leung, V., & Coskun, A. (2021). Counterfactual explanations for multivariate time series. In *2021 International conference on applied artificial intelligence* (pp. 1–8). IEEE.

Avella, P., Sassano, A., & Vasil'ev, I. (2007). Computational study of large-scale *p*-Median problems. *Mathematical Programming*, *109*, 89–114.

Azizi, M., Vayanos, P., Wilder, B., Rice, E., & Tambe, M. (2018). Designing fair, efficient, and interpretable policies for prioritizing homeless youth for housing resources. In *CPAIOR 2018, Delft, the Netherlands, June 26–29, 2018, Proceedings, Vol. 15* (pp. 35–51). Springer.

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*, *49*(3), 312–329.

Barocas, S., Selbst, A., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 80–89).

Bauer, F., Stoer, J., & Witzgall, C. (1961). Absolute and monotonic norms. *Numerische Mathematik*, *3*(1), 257–264.

Belotti, P., Bonami, P., Fischetti, M., Lodi, A., Monaci, M., Nogales-Gómez, A., et al. (2016). On handling indicator constraints in mixed integer programming. *Computational Optimization and Applications*, *65*(3), 545–566.

Ben-Tal, A., & Nemirovski, A. (1999). Robust solutions of uncertain linear programs. *Operations Research Letters*, *25*(1), 1–13.

Bertsimas, D., Brown, D., & Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Review*, *53*(3), 464–501.

Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.-M., & Risser, L. (2022). A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, *76*(2), 188–198.

Bogetoft, P., Ramírez-Ayerbe, J., & Romero Morales, D. (2024). Counterfactual analysis and target setting in benchmarking. *European Journal of Operational Research*, http://dx.doi.org/10.1016/j.ejor.2024.01.005, (in press).

Bomze, I., & Peng, B. (2023). Conic formulation of QPCCs applied to truly sparse QPs. *Computational Optimization and Applications*, *84*(3), 703–735.

Brimberg, J., Hansen, P., Mladenović, N., & Taillard, E. (2000). Improvements and comparison of heuristics for solving the uncapacitated multisource Weber problem. *Operations Research*, *48*(3), 444–460.

Browne, K., & Swift, B. (2020). Semantics and explanation: why counterfactual explanations produce adversarial examples in deep neural networks. arXiv preprint arXiv:2012.10076.

Brughmans, D., Leyman, P., & Martens, D. (2021). Nice: an algorithm for nearest instance counterfactual explanations. arXiv preprint arXiv:2104.07411.

Çalık, H., Labbé, M., & Yaman, H. (2019). *p*-Center problems. In G. Laporte, S. Nickel, & F. Saldanha da Gama (Eds.), *Location science* (pp. 51–65). Cham: Springer International Publishing.

Carrizosa, E., & Fliege, J. (2002). Generalized goal programming: Polynomial methods and applications. *Mathematical Programming*, *93*, 281–303.

Carrizosa, E., Galvis Restrepo, M., & Romero Morales, D. (2021). On clustering categories of categorical predictors in generalized linear models. *Experts Systems with Applications*, *182*, Article 115245.

Carrizosa, E., Guerrero, V., & Romero Morales, D. (2023). On mathematical optimization for clustering categories in contingency tables. *Advances in Data Analysis and Classification*, *17*(2), 407–429.

Carrizosa, E., Halskov, T., & Romero Morales, D. (2023). *Wasserstein SVM: Support vector machines made fair*: *Technical report*, Denmark: Copenhagen Business School, https://www.researchgate.net/publication/371857277_Wasserstein_SVM_Support_Vector_Machines_Made_Fair.

Carrizosa, E., Molero-Río, C., & Romero Morales, D. (2021). Mathematical optimization in classification and regression trees. *TOP*, *29*, 5–33.

Carrizosa, E., Mortensen, L., Romero Morales, D., & Sillero-Denamiel, M. (2022). The tree based linear regression model for hierarchical categorical variables. *Expert Systems with Applications*, *203*(7), Article 117423.

Carrizosa, E., Nogales-Gómez, A., & Romero Morales, D. (2017). Clustering categories in support vector machines. *Omega*, *66*, 28–37.

Carrizosa, E., & Plastria, F. (2008). Optimal expected-distance separating halfspace. *Mathematics of Operations Research*, *33*(3), 662–677.

Carrizosa, E., Ramírez-Ayerbe, J., & Romero Morales, D. (2023). A new model for counterfactual analysis for functional data. *Advances in Data Analysis and Classification*, http://dx.doi.org/10.1007/s11634-023-00563-5, (in press).

Carrizosa, E., Ramírez-Ayerbe, J., & Romero Morales, D. (2024). Generating collective counterfactual explanations in score-based classification via mathematical optimization. *Expert Systems with Applications*, *238*, Article 121954.

Carrizosa, E., & Romero Morales, D. (2001). Combining minsum and minmax: A goal programming approach. *Operations Research*, *49*(1), 169–174.

Carrizosa, E., & Romero Morales, D. (2013). Supervised classification and mathematical optimization. *Computers & Operations Research*, *40*(1), 150–165.

Chandrasekaran, R., & Tamir, A. (1990). Algebraic optimization: The Fermat-Weber location problem. *Mathematical Programming*, *46*(1), 219–224.

Chen, Z., Kuhn, D., & Wiesemann, W. (2022). Data-driven chance constrained programs over Wasserstein balls. *Operations Research*, (in press).

Cui, Z., Chen, W., He, Y., & Chen, Y. (2015). Optimal action extraction for random forests and boosted trees. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 179–188).

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. In *International conference on parallel problem solving from nature* (pp. 448–469). Springer.

De-Arteaga, M., Feuerriegel, S., & Saar-Tsechansky, M. (2022). Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management*, *31*(10), 3749–3770.

Del Ser, J., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., & Holzinger, A. (2022). Exploring the trade-off between plausibility, change intensity and adversarial power in counterfactual explanations using multi-objective optimization. arXiv preprint arXiv:2205.10232.

Delaney, E., Greene, D., & Keane, M. (2021). Instance-based counterfactual explanations for time series classification. In *International conference on case-based reasoning* (pp. 32–47). Springer.

Drezner, T., & Drezner, Z. (2021). Asymmetric distance location model. *INFOR: Information Systems and Operational Research*, *59*(1), 102–110.

Drezner, Z., & Hamacher, H. (2004). *Facility location: applications and theory*. Springer Science & Business Media.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77.

Duarte Silva, A. (2017). Optimization approaches to supervised classification. *European Journal of Operational Research*, *261*(2), 772–788.

Dutta, S., Long, J., Mishra, S., Tilli, C., & Magazzeni, D. (2022). Robust counterfactual explanations for tree-based ensembles. In *International conference on machine learning* (pp. 5742–5756). PMLR.

Eckstein, N., Bates, A., Jefferis, G., & Funke, J. (2021). Discriminative attribution from counterfactuals. arXiv preprint arXiv:2109.13412.

El Ghaoui, L., & Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, *18*(4), 1035–1064.

Erkut, E., & Neuman, S. (1989). Analytical models for locating undesirable facilities. *European Journal of Operational Research*, *40*(3), 275–291.

Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys*, *45*(1), 1–34.

Espejo, I., Marín, A., & Rodríguez-Chía, A. (2015). Capacitated p-center problem with failure foresight. *European Journal of Operational Research*, *247*(1), 229–244.

European Commission (2020). White Paper on Artificial Intelligence : a European approach to excellence and trust. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

Fabris, A., Messina, S., Silvello, G., & Susto, G. (2022). Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, *36*(6), 2074–2152.

Fernández, R., de Diego, I., Aceña, V., Fernández-Isabel, A., & Moguerza, J. (2020). Random forest explainability using counterfactual sets. *Information Fusion*, *63*, 196–207.

Fernández, R., de Diego, I., Moguerza, J., & Herrera, F. (2022). Explanation sets: A general framework for machine learning explainability. *Information Sciences*, *617*, 464–481.

Ferrario, A., & Loi, M. (2022). The robustness of counterfactual explanations over time. *IEEE Access*, *10*, 82736–82750.

Fischetti, M., & Jo, J. (2018). Deep neural networks and mixed integer linear optimization. *Constraints*, *23*(3), 296–309.

Forel, A., Parmentier, A., & Vidal, T. (2022). Robust counterfactual explanations for random forests. arXiv preprint arXiv:2205.14116.

Freiesleben, T. (2022). The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, *32*(1), 77–109.

Gambella, C., Ghaddar, B., & Naoum-Sawaya, J. (2021). Optimization models for machine learning: A survey. *European Journal of Operational Research*, *290*(3), 807–828.

García, S., Labbé, M., & Marín, A. (2011). Solving large p-median problems with a radius formulation. *INFORMS Journal on Computing*, *23*(4), 546–556.

Goethals, S., Martens, D., & Calders, T. (2023). PreCoF: Counterfactual explanations for fairness. *Machine Learning*.

Goethals, S., Martens, D., & Evgeniou, T. (2022). The non-linear nature of the cost of comprehensibility. *Journal of Big Data*, *9*(1), 1–23.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57.

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

Grötschel, M., & Wakabayashi, Y. (1989). A cutting plane algorithm for a clustering problem. *Mathematical Programming*, *45*(1), 59–96.

Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, (in press).

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, *34*(6), 14–23.

Gupta, V., Nokhiz, P., Roy, C., & Venkatasubramanian, S. (2019). Equalizing recourse across groups. arXiv preprint arXiv:1909.03166.

Gurobi Optimization, L. (2021). Gurobi optimizer reference manual.

Haldar, A., Cunningham, T., & Ferhatosmanoglu, H. (2022). RAGUEL: Recourse-aware group unfairness elimination. In *Proceedings of the 31st ACM international conference on information & knowledge management* (pp. 666–675).

Han, S., Gómez, A., & Atamtürk, A. (2023). 2×2-Convexifications for convex quadratic optimization with indicator variables. *Mathematical Programming*, *202*, 95–134. http://dx.doi.org/10.1007/s10107-023-01924-w.

Harrison, D., Jr., & Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (second ed.). New York: Springer.

Hupont, I., Micheli, M., Delipetrev, B., Gómez, E., & Soler Garrido, J. (2022). Documenting high-risk AI: an European regulatory perspective. TechRxiv preprint https://doi.org/10.36227/techrxiv.20291046.v1.

Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., & Ghosh, J. (2019). Towards realistic individual recourse and actionable explanations in black-box decision making systems. arXiv preprint arXiv:1907.09615.

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. (2020). Simple rules to guide expert classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(3), 771–800.

Kanamori, K., Takagi, T., Kobayashi, K., & Arimura, H. (2020). DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20* (pp. 2855–2862).

Kanamori, K., Takagi, T., Kobayashi, K., Ike, Y., Uemura, K., & Arimura, H. (2021). Ordered counterfactual explanation by mixed-integer linear optimization. In *Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 13* (pp. 11564–11574).

Karimi, A.-H., Barthe, G., Balle, B., & Valera, I. (2020). Model-agnostic counterfactual explanations for consequential decisions. In *International conference on artificial intelligence and statistics* (pp. 895–905). PMLR.

Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2022). A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, *55*(5), 1–29.

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., & Valera, I. (2022). Towards causal algorithmic recourse. In *International workshop on extending explainable AI beyond deep models and classifiers* (pp. 139–166). Springer.

Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 353–362).

Karlsson, I., Rebane, J., Papapetrou, P., & Gionis, A. (2020). Locally and globally explainable time series tweaking. *Knowledge and Information Systems*, *62*(5), 1671–1700.

Kaufman, L., & Rousseeuw, P. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Keane, M., & Smyth, B. (2020). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In *International conference on case-based reasoning* (pp. 163–178). Springer.

Klafszky, E., Mayer, J., & Terlaky, T. (1989). Linearly constrained estimation by mathematical programming. *European Journal of Operational Research*, *42*(3), 254–267.

Korikov, A., & Beck, J. (2021). Counterfactual explanations via inverse constraint programming. In *27th International conference on principles and practice of constraint programming*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Korikov, A., Shleyfman, A., & Beck, C. (2021). Counterfactual explanations for optimization-based decisions in the context of the GDPR. In *ICAPS 2021 workshop on explainable AI planning*.

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, *30*, 4066–4076.

Landete, M., Peiró, J., & Yaman, H. (2023). Formulations and valid inequalities for the capacitated dispersion problem. *Networks*, *81*(2), 294–315. http://dx.doi.org/10.1002/net.22132.

Laporte, G., Nickel, S., & Saldanha da Gama, F. (2020). *Location science*. Springer Nature.

Le Thi, H., & Pham Dinh, T. (2018). DC programming and DCA: thirty years of developments. *Mathematical Programming*, *169*(1), 5–68.

Le Thi, H., & Pham Dinh, T. (2023). Open issues and recent advances in DC programming and DCA. *Journal of Global Optimization*, http://dx.doi.org/10.1007/s10898-023-01272-1, (in press).

Lefebvre, O., Michelot, C., & Plastria, F. (1991). Sufficient conditions for coincidence in minisum multifacility location problems with a general metric. *Operations Research*, *39*(3), 437–442.

Liberti, L., & Manca, B. (2022). Side-constrained minimum sum-of-squares clustering: mathematical programming and random projections. *Journal of Global Optimization*, *83*, 83–118.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137.

Lozano-Osorio, I., Martínez-Gavara, A., Martí, R., & Duarte, A. (2022). Max–min dispersion with capacity and cost for a practical location problem. *Expert Systems with Applications*, *200*, Article 116899.

Mahajan, D., Tan, C., & Sharma, A. (2019). Preserving causal constraints in counterfactual explanations for machine learning classifiers. arXiv preprint arXiv:1912.03277.

Maragno, D., Kurtz, J., Röber, T., Goedhart, R., Birbil, Ş., & den Hertog, D. (2023). Finding regions of counterfactual explanations via robust optimization. arXiv preprint arXiv:2301.11113.

Maragno, D., Röber, T. E., & Birbil, I. (2022). Counterfactual explanations using optimization with constraint learning. arXiv preprint arXiv:2209.10997.

Marín, A., & Pelegrín, M. (2019). *p*-Median problems. In G. Laporte, S. Nickel, & F. Saldanha da Gama (Eds.), *Location science* (pp. 25–50). Cham: Springer International Publishing.

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *MIS Quarterly*, *38*(1), 73–99.

Martí, R., Martínez-Gavara, A., Pérez-Peló, S., & Sánchez-Oro, J. (2022). A review on discrete diversity and dispersion maximization from an OR perspective. *European Journal of Operational Research*, *299*(3), 795–813.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*, 1–35.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Mirchandani, P., & Francis, R. (1990). *Discrete location theory*.

Miron, M., Tolan, S., Gómez, E., & Castillo, C. (2020). Addressing multiple metrics of group fairness in data-driven decision making. arXiv preprint arXiv:2003.04794.

Miron, M., Tolan, S., Gómez, E., & Castillo, C. (2021). Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law*, *29*(2), 111–147.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163.

Mladenović, N., Brimberg, J., Hansen, P., & Moreno-Pérez, J. (2007). The *p*-median problem: A survey of metaheuristic approaches. *European Journal of Operational Research*, *179*(3), 927–939.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 417–431). Springer.

Mothilal, R., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 607–617).

Mulvey, J., & Beck, M. (1984). Solving capacitated clustering problems. *European Journal of Operational Research*, *18*(3), 339–348.

Ogryczak, W. (2001). Comments on properties of the minmax solutions in goal programming. *European Journal of Operational Research*, *132*(1), 17–21.

Olson, M., Khanna, R., Neal, L., Li, F., & Wong, W.-K. (2021). Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artificial Intelligence*, *295*, Article 103455.

Palagi, L. (2019). Global optimization issues in deep network regression: an overview. *Journal of Global Optimization*, *73*(2), 239–277.

Parmentier, A., & Vidal, T. (2021). Optimal counterfactual explanations in tree ensembles. In *International conference on machine learning* (pp. 8422–8431). PMLR.

Parreño, F., Álvarez-Valdés, R., & Martí, R. (2021). Measuring diversity. A review and an empirical analysis. *European Journal of Operational Research*, *289*(2), 515–532.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys*, *55*(3), 1–44.

Peyré, G., & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, *11*(5–6), 355–607.

Piccialli, V., Romero Morales, D., & Salvatore, C. (2022). Features compression based on counterfactual analysis. arXiv preprint arXiv:2211.09894.

Piccialli, V., & Sciandrone, M. (2018). Nonlinear optimization and support vector machines. *4OR*, *16*(2), 111–149.

Piccialli, V., Sudoso, A., & Wiegele, A. (2022). SOS-SDP: An exact solver for minimum sum-of-squares clustering. *INFORMS Journal on Computing*, *34*, 2144–2162.

Pisinger, D. (2006). Upper bounds and exact algorithms for *p*-dispersion problems. *Computers & Operations Research*, *33*(5), 1380–1398.

Plastria, F. (1992). On destination optimality in asymmetric distance Fermat-Weber problems. *Annals of Operations Research*, *40*, 355–369.

Plastria, F. (1995). In Z. Drezner (Ed.), *Continuous location problems*. New York: Springer-Verlag.

Plastria, F. (2019). Pasting gauges I: Shortest paths across a hyperplane. *Discrete Applied Mathematics*, *256*, 105–137.

Plastria, F., & Carrizosa, E. (2001). Gauge distances and median hyperplanes. *Journal of Optimization Theory and Applications*, *110*, 173–182.

Plastria, F., & Carrizosa, E. (2012). Minmax-distance approximation and separation problems: geometrical properties. *Mathematical Programming*, *132*(1–2), 153–177.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, *10*(3), 61–74.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 344–350).

Prado-Romero, M., Prenkaj, B., Stilo, G., & Giannotti, F. (2022). A survey on graph counterfactual explanations: Definitions, methods, evaluation. arXiv preprint arXiv:2210.12089.

Raimundo, M., Nonato, L., & Poco, J. (2022). Mining Pareto-optimal counterfactual antecedents with a branch-and-bound model-agnostic algorithm. *Data Mining and Knowledge Discovery*, (in press).

Ramakrishnan, G., Lee, Y., & Albarghouthi, A. (2020). Synthesizing action sequences for modifying model decisions. In *Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 04* (pp. 5462–5469).

Ramon, Y., Martens, D., Provost, F., & Evgeniou, T. (2020). A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C. *Advances in Data Analysis and Classification*, *14*, 801–819.

Rawal, K., & Lakkaraju, H. (2020). Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems*, *33*, 12187–12198.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*, 1–85.

Ruiz, F., Luque, M., & Cabello, J. (2009). A classification of the weighting schemes in reference point procedures for multiobjective programming. *Journal of the Operational Research Society*, *60*(4), 544–553.

Russell, C. (2019). Efficient search for diverse coherent explanations. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 20–28).

Salazar, S., Denton, S., & Salleb-Aouissi, A. (2022). Counterfactual explanations for support vector machine models. arXiv preprint arXiv:2212.07432.

Sharma, S., Henderson, J., & Ghosh, J. (2020). CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 166–172).

Slack, D., Hilgard, A., Lakkaraju, H., & Singh, S. (2021). Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems*, *34*, 62–75.

Sokol, K., & Flach, P. (2019). Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. In *SafeAI @ AAAI*.

Stepin, I., Alonso, J., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, *9*, 11974–12001.

Tolkachev, G., Mell, S., Zdancewic, S., & Bastani, O. (2022). Counterfactual explanations for natural language interfaces. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 113–118).

Upadhyay, S., Joshi, S., & Lakkaraju, H. (2021). Towards robust and reliable algorithmic recourse. *Advances in Neural Information Processing Systems*, *34*, 16926–16937.

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 10–19).

Van Looveren, A., & Klaise, J. (2021). Interpretable counterfactual explanations guided by prototypes. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 650–665). Springer.

Vasilyev, I., & Ushakov, A. (2021). Discrete facility location in machine learning. *Journal of Applied and Industrial Mathematics*, *15*(4), 686–710.

Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., & Shah, C. (2022). Counterfactual explanations and algorithmic recourses for machine learning: A review. arXiv preprint arXiv:2010.10596.

Vermeire, T., Brughmans, D., Goethals, S., de Oliveira, R., & Martens, D. (2022). Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, *25*, 315–335.

Von Kügelgen, J., Karimi, A.-H., Bhatt, U., Valera, I., Weller, A., & Schölkopf, B. (2022). On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI conference on artificial intelligence. Vol. 36. No. 9* (pp. 9584–9594).

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*, 841–887.

Wei, L., Gómez, A., & Küçükyavuz, S. (2022). Ideal formulations for constrained convex optimization problems with indicator variables. *Mathematical Programming*, *192*(1–2), 57–88.

Weiszfeld, E., & Plastria, F. (2009). On the point for which the sum of the distances to *n* given points is minimum. *Annals of Operations Research*, *167*(1), 7–41.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2019). The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 56–65.

Wilson, D., & Martinez, T. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, *6*, 1–34.

Xing, Z., Pei, J., & Keogh, E. (2010). A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, *12*(1), 40–48.

Xue, G., & Ye, Y. (1997). An efficient algorithm for minimizing a sum of euclidean norms with applications. *SIAM Journal on Optimization*, *7*(4), 1017–1036.

Xue, G., & Ye, Y. (2000). An efficient algorithm for minimizing a sum of *p*-norms. *SIAM Journal on Optimization*, *10*(2), 551–579.

Yousefzadeh, R., & O'Leary, D. (2020). Deep learning interpretation: Flip points and homotopy methods. In *Mathematical and scientific machine learning* (pp. 1–26). PMLR.

Yousefzadeh, R., & O'Leary, D. (2022). Auditing and debugging deep learning models via decision boundaries: Individual-level and group-level analysis. *La Matematica*, *1*, 19–52.

Yu, K., Lu, Z., & Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (the Statistician)*, *52*(3), 331–350.

Zafar, M., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. (2019). Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, *20*(1), 2737–2778.

Zeng, J., Gensheimer, M., Rubin, D., Athey, S., & Shachter, R. (2022). Uncovering interpretable potential confounders in electronic medical records. *Nature Communications*, *13*(1), 1014.

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A*, *180*(3), 689–722.

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). "Why should you trust my explanation?" understanding uncertainty in LIME explanations. arXiv preprint arXiv:1904.12991.