**Les Cahiers du GERAD** ISSN: 0711–2440

# On counterfactual explanations for clustering medoids

D. Aloise, C. Rocha

G–2025–58

September 2025



**GERAD** HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

**Tél. : 514 340-6053**
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

# On counterfactual explanations for clustering medoids

**Daniel Aloise** [a, b]
**CAroline Rocha** [c]

[a] *Computer and Software Engineering Department, Polytechnique Montréal, Montréal (Qc), Canada, H3T 1J4*

[b] *GERAD, Montréal (Qc), Canada, H3T 1J4*

[c] *IVADO Labs, Montréal (Qc), Canada, H2S 3J9*

daniel.aloise@polymtl.ca
caroline.rocha@ivadolabs.com

**September 2025**
**Les Cahiers du GERAD**
**G–2025–58**

**Abstract :** As the need for interpretable machine learning continues to grow, we propose a novel post-clustering method to generate counterfactual explanations for clustering results. Specifically, our method answers the question: What is the smallest change to a data point that would make it the medoid of its cluster? These explanations offer valuable insights in domains like healthcare and marketing, where identifying minimal adjustments to align individuals or customers with desirable cluster representative profiles can inform personalized interventions and strategic decision-making. By formulating the problem as a convex optimization model, specifically a second-order cone program, our method guarantees global optimality requiring readily available and effective solvers for practical implementation.

**Keywords :** Counterfactual explanations; clustering interpretability; convex optimization

# 1 Introduction

As machine learning models are increasingly implemented across various applications, the demand for interpretability of the predicted classes or values has been rising significantly. The scientific literature has made a significant effort in increasing the explainability of machine learning models, both in supervised or unsupervised modes (e.g. Ribeiro et al. (2016); Carrizosa et al. (2017, 2022); Rudin et al. (2022); Carrizosa et al. (2025); Randel et al. (2021)).

Explainability models can be based on global and/or local explanation strategies (Ribeiro et al., 2016). While global approaches seek to offer understanding of a model's behaviour over the entire dataset, frequently employing techniques like decision trees (Bertsimas et al., 2021) or rule-based systems (Carrizosa et al., 2023), local explanation methods concentrate on specific predictions or decisions, providing insights into particular outcomes. Counterfactual explanations are a common example of these local methods (Wachter et al., 2017; Contardo et al., 2024; Parmentier and Vidal, 2021), identifying minimal changes needed to alter a specific decision made by the model.

Particularly in cluster analysis, numerous methods have been developed to tackle clustering problems across various applications, with emphasis on the accuracy in identifying hidden cluster structures (Kettenring, 2006). However, less attention is generally given to the ability of these methods in interpreting the resulting clusters. In (Hu et al., 2024), the authors categorize interpretable clustering methods based on when interpretability is introduced in the clustering pipeline. *Pre-clustering* methods focus on selecting or extracting meaningful and human-understandable features before the clustering process begins, ensuring that the input data itself is interpretable. *In-clustering* methods integrate interpretability directly into the clustering algorithm, often by optimizing for both cluster quality and explainability. These methods ensure that the clustering process remains transparent from the start. Finally, *post-clustering* methods are centered on interpreting the outcomes of pre-existing clustering models. This stage is crucial for making results understandable, particularly when working with complex, high-dimensional data where traditional clustering methods lack transparency. By structuring methods across these three stages, researchers can systematically analyze and enhance interpretability in clustering applications.

The use of cluster *medoids*, also called *medians* or *prototypes*, is known to facilitate the interpretation of clustering results (Köhn et al., 2010). A medoid is a data object that serves as the representative of its cluster, being characterized e.g. by its minimal dissimilarity sum to the other data objects within its cluster. There are many advantages on the use of medoids for clustering (e.g. outlier robustness, dissimilarity generalization, etc.). In fact, research on perception indicates that individuals frequently use actual data objects to represent larger subsets of objects (Blanchard et al., 2012).

The literature on interpretable clustering methods based on medoids is scarce. Carrizosa et al. (2022) proposed an post-clustering optimization model to identify medoids from a given clustering solution. The objective of their model is to maximize the count of true positive cases, i.e., data objects which are closer to the identified medoids of their clusters than to other medoids, while minimizing the number of false positive cases, i.e. data objects which are closer to the medoids from other clusters.

In this work, we propose a novel post-clustering interpretable method that provides counterfactual explanations for medoids in clustering solutions. Given the output produced by any clustering algorithm, our method computes minimal numerical feature modifications required to transform a data point into the medoid of its assigned cluster.

The potential applications for our method are numerous. It can explain what can turn data points representative of their cluster — even if they are currently at their peripheries. For example, in healthcare, patients can be clustered based on their medical history, symptoms, or responses to treatments. The medoid of a cluster may represent the ideal patient profile for a specific condition or treatment response. Using our method, healthcare providers could identify minimal changes in a patient's features (e.g., lifestyle modifications, medication adjustments, or exercise routines) that would bring them closer to the prototype of the healthiest or most stable patient in the group. This could enable person-

alized treatment plans aimed at improving patient outcomes, guiding interventions that align patients more closely with the optimal treatment profile for their condition. Likewise, for marketing strategies, businesses can identify small changes in customer behaviors (e.g., purchase frequency, product preferences, or engagement with campaigns) that could make a customer more representative of the prototype costumer of a given segment. This can be used to fine-tune customer retention strategies or to design personalized offers, aiming to move customers toward the core behaviors that drive value for the company. Moreover, the counterfactual path to the medoid highlights which features matter most in defining cluster identity, helping analysts better interpret the structure of the data, possibly refining their clustering approach.

The rest of this article is organized as follows. Section 2 introduces our problem and describes how it is modelled as a convex optimization problem, allowing the use of readily available convex optimization solvers to efficiently compute globally optimal solutions. Section 3 discusses computational experiments for an illustrative example that demonstrates the applicability of the proposed method we developed. Concluding remarks are drawn in the last section.

## 2 Model

Given a set of numerical data points $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$, and an additional point $x_0 \in \mathbb{R}^d$, the objective is to minimally modify $x_0$ so that it becomes the **medoid** of the set $\{x_0, x_1, \ldots, x_n\}$, i.e., the data point whose sum of distances to the other points is minimal. The set $\{x_0, x_1, \ldots, x_n\}$ is supposed to represent a cluster provided by any clustering algorithm executed in a larger dataset. The problem can be mathematically expressed as:

$$\begin{aligned} \min \quad & \|\Delta\| \\ \text{s.t.} \quad & \sum_{\substack{i=1 \\ i \neq j}}^{n} \|x_0 + \Delta - x_i\| \leq \overline{D}_j, \quad \forall j = 1, \ldots, n, \\ & \Delta \in \mathbb{R}^d. \end{aligned} \tag{1}$$

where $\|\cdot\|$ refers to the Euclidean norm, and $\overline{D}_j = \sum_{i=1}^{n} \|x_j - x_i\|$ is a non-negative constant representing the sum of the distance between $x_j$ and all other points. Thus, the constraints in (1) enforce that, for each $j = 1, \ldots, n$, the sum of the distances from all points $x_i \in \{x_1, \ldots, x_n\}$, with $x_i \neq x_j$, to the updated coordinates $x_0 + \Delta$ is less than or equal to $\overline{D}_j$. This condition ensures that $x_0 + \Delta$ corresponds to the medoid of the set $\{x_0, x_1, \ldots, x_n\}$. To guarantee that $x_0 + \Delta$ is the unique medoid, a small $\epsilon$ can be added to the constraints.

We note that problem (1) is convex given that: (i) the objective function is a norm, and therefore, a convex function, and (ii) the left-hand side of its constraints correspond to the sum of norms of affine functions of $\Delta$, i.e., a sum of convex functions. Consequently, the constraints are also convex.

By introducing auxiliary variables $t \in \mathbb{R}$ and $s_{0i} \in \mathbb{R}_+$ for each $i = 1, \ldots, n$, problem (1) can be reformulated as a second-order cone program (SOCP), for which many powerful solvers exist. The resulting SOCP problem is given by:

$$\begin{aligned} \min \quad & t \\ \text{s.t.} \quad & \|\Delta\| \leq t \\ & \|x_0 + \Delta - x_i\| \leq s_{0i} && \forall i = 1, \ldots, n \\ & \sum_{\substack{i=1 \\ i \neq j}}^{n} s_{0i} \leq \overline{D}_j && \forall j = 1, \ldots, n, \\ & t \geq 0, \quad s_{0i} \geq 0 && \forall i = 1, \ldots, n, \\ & \Delta \in \mathbb{R}^d. \end{aligned} \tag{2}$$

Since (1) is a convex problem, any local minimum is also a global optimizer for the problem.

# 3 Practical illustration

We illustrate our model in the dogs size dataset available at `https://data.world/len/dog-size-intelligence-linked`.[1] The data contain lower and upper values for the weights (in pounds) and heights (in inches) of 150 dog breeds. To showcase our method in two dimensions, we processed the data by computing average values from the associated lower and upper limits. A few duplicate records were generated during this process and subsequently removed, resulting in a final dataset of 134 distinct dog breeds.

Our baseline clustering solution is taken from the American Kennel Club (AKC) American Kennel Club (2024) that classifies dog breeds into 5 distinct groups according to their sizes: XLarge, Large, Medium, Small and XSmall. Figure 1 presents the scatter plot of the dogs data, with the provided data clustering.

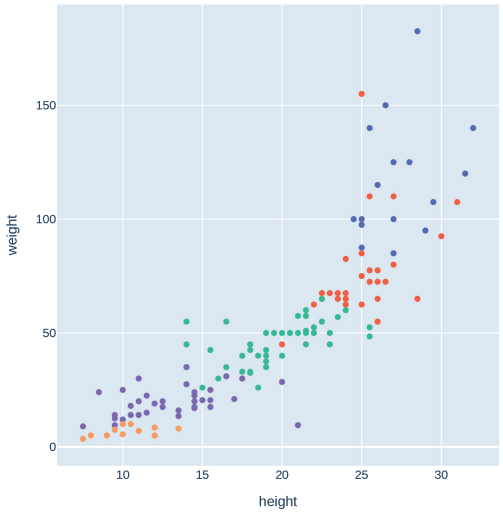


**Figure 1: Scatter plot of dog breeds clustered by height and weight into five groups.**

Our approach considers the underlying clustering to provide counterfactual explanations for medoids in each cluster. We will illustrate our method in the cluster of XSmall dog breeds presented in Figure 2, for which the current medoid is the *Manchester Terrier*. The standardized coordinates (z-scores) for the data points in the XSmall cluster are presented in Table 1.

Our model enables the identification of the minimal modifications required for a given dog breed to transform into the medoid of the analyzed cluster. For the presented experiments, model (2) is solved by the Clarabel solver (Goulart and Chen, 2024) called from the Python library CVXPY. To illustrate its utilization, let us consider the *Maltese* dog breed. By solving the model for $x_0 = (-0.744, -0.873)$ corresponding to the standardized measures of the *Maltese* dog breed, we are able to compute the value of $\Delta = (\Delta_{height}, \Delta_{weight}) = (0.509, 0.616)$ so that $x_0 + \Delta = (-0.235, -0.257)$ becomes a prototype for the cluster of XSmall dog breeds. Figure 3 illustrates the computed modification in the underlying 2D space. Projected back to the original space, this would correspond to increasing the height of the *Maltese* breed in 0.871 inches while increasing the weight in 1.283 pounds.

Figure 4 shows the computed modifications $\Delta$ obtained by solving model (2) for $x_0$ representing each dog breed in the XSmall cluster. Our model reveals that the *Japanese Chin* requires the least modification to serve as the medoid of the cluster, while the *Chihuahua* requires the most substantial transformation. Table 2 reports the computed $\Delta$ values for each dog breed. The computational time required to solve model (2) ranged from 2.66 to 8.25 milliseconds.

[1] login is required

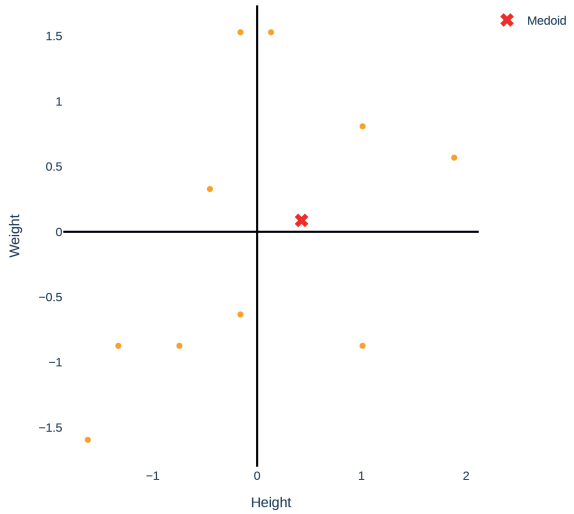


**Figure 2: Standardized height and weight (z-scores) for XSmall dog breeds.**

**Table 1: Normalized coordinates of the dog breeds in cluster XSmall. The coordinates of the cluster medoid (Manchester Terrier) are indicated in red.**

| Breed | height | weight |
|---|---|---|
| Affenpinscher | 0.133 | 1.529 |
| Chihuahua | -1.621 | -1.594 |
| Chinese Crested | 1.001 | 0.808 |
| Italian Greyhound | 1.887 | 0.568 |
| Japanese Chin | -0.452 | 0.328 |
| Maltese | -0.744 | -0.873 |
| Manchester Terrier | 0.425 | 0.087 |
| Pomeranian | 1.010 | -0.874 |
| Poodle Toy | -0.159 | 1.529 |
| Toy Fox Terrier | -0.159 | -0.633 |
| Yorkshire Terrier | -1.328 | -0.874 |

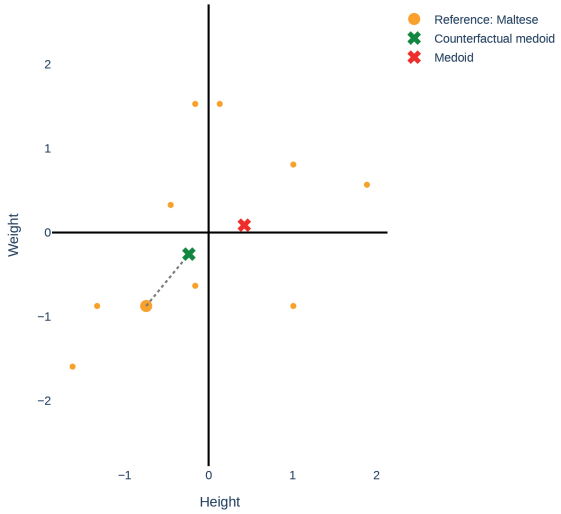


**Figure 3: $\Delta$ modifications required to transform the $Maltese$ dog breed into the medoid of the XSmall cluster (indicated in green).**

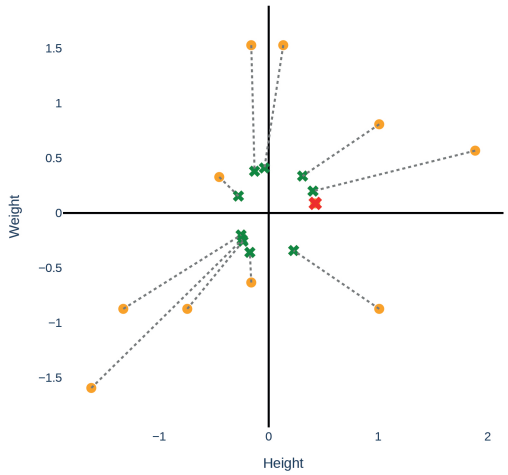


**Figure 4: Required transformations for each dog breed in the XSmall cluster to become the medoid.**

**Table 2: $\Delta$ modifications computed by our model to make each dog breed the medoid of cluster XSmall.**

| Breed | $\Delta_{height}$ | $\Delta_{weight}$ | $\|\Delta\|$ |
|---|---|---|---|
| Affenpinscher | -0.172 | -1.119 | 1.132 |
| Chihuahua | 1.379 | 1.352 | 1.931 |
| Chinese Crested | -0.700 | -0.472 | 0.844 |
| Italian Greyhound | -1.482 | -0.368 | 1.528 |
| Japanese Chin | 0.175 | -0.175 | 0.247 |
| Maltese | 0.509 | 0.617 | 0.800 |
| Pomeranian | -0.782 | 0.531 | 0.945 |
| Poodle Toy | 0.030 | -1.149 | 1.149 |
| Toy Fox Terrier | -0.013 | 0.274 | 0.275 |
| Yorkshire Terrier | 1.076 | 0.673 | 1.269 |

We observe that our model can be readily extended to handle cases where the $\Delta$ modifications are weighted differently across dimensions. For instance, in the context of dog breeds, increasing a breeds's weight over time is more feasible than altering its height. To account for such differences, a weighted norm $\|\Delta_w\|$ can also be optimized in model (2). Figure 5 shows the $\Delta$ values computed by (2) for each dog breed, considering $w_{height} = 0.75$ and $w_{weight} = 0.25$.

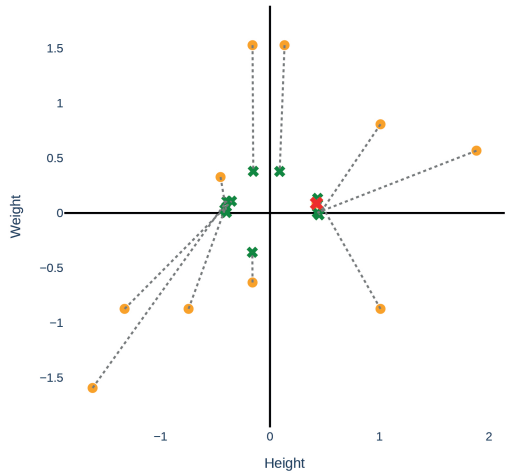


**Figure 5: Required transformations for each dog breed in the XSmall cluster to become the medoid for a weighted norm with $w_{height} = 0.75$ and $w_{weight} = 0.25$.**

We observe in the figure that vertical $\Delta_{weight}$ transformations are now preferred, as they incur a lower cost in the model's objective function due to the smaller weight assigned to that dimension. With weights set to $w_{height} = 0.75$ and $w_{weight} = 0.25$, the *Toy Fox Terrier* requires the least costly transformation to become the medoid of cluster XSmall, while the *Italian Greyhound* incurs the highest cost transformation. The computed $\Delta$ transformations for the illustrated weighted case are shown in Table 3.

**Table 3: $\Delta$ modifications and weighted norm $\|\Delta_w\|$ computed by our model to make each dog breed the medoid of cluster XSmall.**

| Breed | $\Delta_{height}$ | $\Delta_{weight}$ | $\|\Delta_w\|$ |
|---|---|---|---|
| Affenpinscher | -0.044 | -1.150 | 0.289 |
| Chihuahua | 1.226 | 1.702 | 1.013 |
| Chinese Crested | -0.561 | -0.825 | 0.469 |
| Italian Greyhound | -1.454 | -0.565 | 1.100 |
| Japanese Chin | 0.054 | -0.323 | 0.090 |
| Maltese | 0.331 | 0.906 | 0.336 |
| Pomeranian | -0.575 | 1.007 | 0.499 |
| Poodle Toy | 0.007 | -1.149 | 0.287 |
| Toy Fox Terrier | -0.002 | 0.275 | 0.069 |
| Yorkshire Terrier | 0.977 | 0.981 | 0.773 |

## 4 Concluding remarks

In this work, we presented an optimization model that provides counterfactual explanations to the following question: *What is the smallest change to a data point that would make it the medoid of its cluster?*

One important characteristic of our model is that it is convex, and thus solved to global optimality once a local minimum is found. Through its reformulation as a second-order cone program, the model can be solved by powerful existing optimization solvers.

We illustrate our model by generating counterfactual explanations for dog breeds data, using it as a post-clustering interpretability method that computes minimal transformations required to make each breed the medoid of its cluster. The example is presented in two dimensions for illustrative purposes, but the approach naturally extends to higher-dimensional settings.

Finally, it is important to note that if any dimension of the data points is binary or discretely ordered, the problem becomes non-convex, as the transformation vector $\Delta$ must then be constrained to a discrete set of values. Addressing this setting is left for future work, as it requires alternative solution techniques beyond convex optimization.

## References

American Kennel Club. Dog breeds - complete list of dog breeds, 2024. URL https://www.akc.org/dog-breeds/. Accessed: 2025-04-22.

Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. Interpretable clustering: an optimization approach. Machine Learning, 110(1):89–138, 2021.

Simon J Blanchard, Daniel Aloise, and Wayne S DeSarbo. The heterogeneous p-median problem for categorization based clustering. Psychometrika, 77(4):741–762, 2012.

Emilio Carrizosa, Amaya Nogales-Gómez, and Dolores Romero Morales. Clustering categories in support vector machines. Omega, 66:28–37, 2017.

Emilio Carrizosa, Kseniia Kurishchenko, Alfredo Marín, and Dolores Romero Morales. Interpreting clusters via prototype optimization. Omega, 107:102543, 2022.

Emilio Carrizosa, Kseniia Kurishchenko, Alfredo Marín, and Dolores Romero Morales. On clustering and interpreting with rules by means of mathematical optimization. Computers & Operations Research, 154: 106180, 2023.

Emilio Carrizosa, Kseniia Kurishchenko, and Dolores Romero Morales. On enhancing the explainability and fairness of tree ensembles. European Journal of Operational Research, 2025.

Claudio Contardo, Ricardo Fukasawa, Louis-Martin Rousseau, and Thibaut Vidal. Optimal counterfactual explanations for k-nearest neighbors using mathematical optimization and constraint programming. In International Symposium on Combinatorial Optimization, pages 318–331. Springer, 2024.

Paul J Goulart and Yuwen Chen. Clarabel: An interior-point solver for conic programs with quadratic objectives. arXiv preprint arXiv:2405.12762, 2024.

Lianyu Hu, Mudi Jiang, Junjie Dong, Xinying Liu, and Zengyou He. Interpretable clustering: A survey. arXiv preprint arXiv:2409.00743, 2024.

Jon R Kettenring. The practice of cluster analysis. Journal of classification, 23(1):3–30, 2006.

Hans-Friedrich Köhn, Douglas Steinley, and Michael J Brusco. The p-median model as a tool for clustering psychological data. Psychological methods, 15(1):87, 2010.

Axel Parmentier and Thibaut Vidal. Optimal counterfactual explanations in tree ensembles. In International conference on machine learning, pages 8422–8431. PMLR, 2021.

Rodrigo Randel, Daniel Aloise, Simon J Blanchard, and Alain Hertz. A lagrangian-based score for assessing the quality of pairwise constraints in semi-supervised clustering. Data Mining and Knowledge Discovery, 35:2341–2368, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys, 16:1–85, 2022.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech., 31:841, 2017.