Comparison-based centrality measures

Luca Rendsburg, Damien Garreau

International Journal of Data Science and Analytics (2021) 11:243–259
https://doi.org/10.1007/s41060-021-00254-4

REGULAR PAPER

Comparison-based centrality measures

Luca Rendsburg[1] · **Damien Garreau**[2]

Received: 24 September 2020 / Accepted: 10 March 2021 / Published online: 12 April 2021

Abstract
Recently, learning only from ordinal information of the type "item x is closer to item y than to item z" has received increasing attention in the machine learning community. Such triplet comparisons are particularly well suited for learning from crowdsourced human intelligence tasks, in which workers make statements about the relative distances in a triplet of items. In this paper, we systematically investigate comparison-based centrality measures on triplets and theoretically analyze their underlying Euclidean notion of centrality. Two such measures already appear in the literature under opposing approaches, and we propose a third measure, which is a natural compromise between these two. We further discuss their relation to statistical depth functions, which comprise desirable properties for centrality measures, and conclude with experiments on real and synthetic datasets for medoid estimation and outlier detection.

Keywords Pairwise comparisons · Ordinal information · Triplets · Centrality measures · Statistical depth functions

1 Introduction

Assume we are given a finite dataset $\mathcal{D} = \{x_1, \dots, x_n\}$ in a metric space $(\mathcal{X}, d)$, but do not have access to an explicit representation or the pairwise distances. Instead, the only information available for any triplet (x, y, z) in $\mathcal{D}$ is the answer to the *triplet comparison*

$$d(x, y) \overset{?}{<} d(x, z)\,. \tag{1}$$

In many applications of data analysis, such as human intelligence tasks for crowdsourcing, only ordinal information as in Eq. (1) is naturally available. It is difficult for humans to determine the distance between items in absolute terms, and their answers will have a large variance. Relative statements, however, are easier to make and more consistent. For example in Fig. 1, image x is obviously central in the presented triplet, because it contains both snow and trees, but it is hard to quantify this relationship.

✉ Luca Rendsburg
luca.rendsburg@uni-tuebingen.de

Damien Garreau
damien.garreau@unice.fr

[1] Department of Computer Science, University of Tübingen, Tübingen, Germany

[2] Laboratoire J. A. Dieudonné & Inria Maasai Project-Team, Université Nice Côte d'Azur, Nice, France

Lately, comparison-based settings receive increasing attention in the machine learning community; for an overview, see Kleindessner and von Luxburg [18, Section 5.1].

An important task in statistics is to summarize the distribution of the data into a few descriptive values. In this paper, we focus on measures of centrality, the most prominent examples in a standard setting being the mean, the median, and the mode. While those summary statistics are readily computed given an explicit representation of the dataset, it is much less obvious how to do so when only ordinal information is available, or whether it is possible at all.

Learning a centrality measure based on triplet comparisons already appears in the literature under two opposing approaches: Heikinheimo and Ukkonen [12] propose a score that penalizes points for being the outlier in a triplet, while Kleindessner and von Luxburg [18] reward points for being central in a triplet. The latter motivate their central-based score by relating it to k-relative neighborhood graphs and a statistical depth function, which cannot be done for the outlier-based score. Figure 1 shows a triplet with its corresponding central point and outlier. Both scores are intuitively appealing as a measure of centrality, but it is not obvious how they relate to other known centrality notions. Furthermore, there is a third kind of point in a triplet next to central point and outlier, namely the remaining point that opposes the middle side (compare y in Fig. 1). This point scores 0 for both

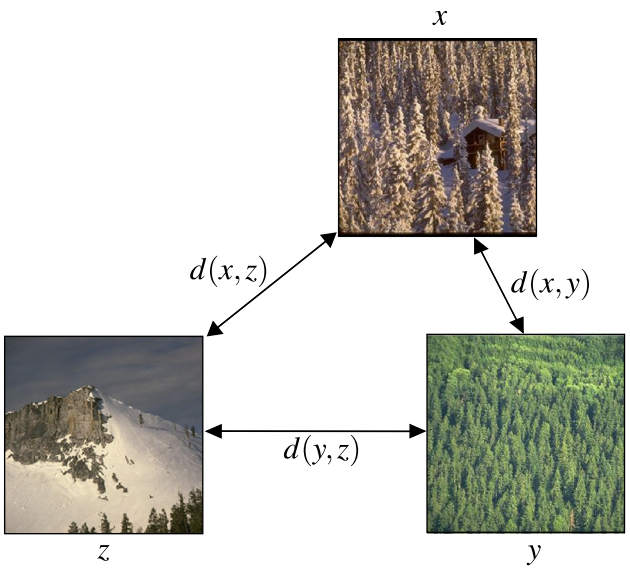


Fig. 1 A triplet (x, y, z) of images with distances $d(x, y) < d(x, z) < d(y, z)$. The central image x opposes the largest side $d(y, z)$, and the outlier z opposes the smallest side $d(x, y)$

scores, because they only use a binary distinction with respect to outlier or central point.

Our main contributions are summarized as follows:

- **Introduction of rank score.** We propose the rank *score*, a natural compromise between outlier and central score that considers all three possible scenarios in a triplet.
- **Characterization of central points.** We establish a theoretical connection between the three scores and their corresponding scoring probabilities. Based on this connection, we characterize the most central point with respect to each score for one-dimensional distributions. In doing so, we improve on a result from Heikinheimo and Ukkonen [12] for the outlier score.
- **Statistical depth functions.** We show that the rank score is not a statistical depth function, but nevertheless satisfies a weaker version of the violated properties.
- **Experiments.** On two image datasets, we evaluate the three scores for the tasks of medoid estimation and outlier identification and find that their estimations are similar to the ones given by the average Euclidean distance. On synthetic datasets, we highlight two respective weaknesses of outlier and central score, which are mitigated by the rank score.

After a brief summary of related work in Sect. 2, we define the scores in Sect. 3 and establish the connection to their scoring probabilities. In Sect. 4, we characterize the central points for one-dimensional distributions, and then in Sect. 5 investigate the relation between our proposed rank score and statistical depth functions. We conclude with experiments on real and synthetic datasets in Sect. 6.

2 Related work

There exist various different approaches for learning from triplet comparisons on a multitude of tasks. An indirect approach is to first learn a Euclidean embedding for the dataset that satisfies the ordinal constraints and then use standard machine learning algorithms on the explicit representation. This can be done by a max-margin approach [1] or by maximizing the probabilities that all triplet constraints are satisfied under a stochastic selection rule using a Student-t kernel [30]. Other variants of this problem include the active setting [26] and learning hidden attributes in multiple maps [2]. There exists some theory for the consistency [3,14,15] of such embeddings, also in the case of noisy [13] and local [27] comparisons. This approach has a number of drawbacks: existing embedding algorithms do not scale well, they make the implicit assumption that the points lie in a Euclidean space, the quality of the embedding depends on the embedding dimension, and this intermediate step in the learning procedure introduces additional distortion; for a discussion, see Kleindessner and von Luxburg [18, Section 5.1.2]. For instance, our problem of determining the mean (or similarly the median or the mode) of the underlying distribution of a dataset could be tackled this way: we embed the points using one of the methods above, compute the mean in the Euclidean space, and then return the datapoint whose embedding is closest.

A more direct approach than embedding the dataset is the quantization of the ordinal information by learning a distance metric from a parametrized family with a max-margin approach [24] or by using kernels that only depend on triplet comparisons [17]. Lately, however, the trend is to avoid any such intermediate steps: Heikinheimo and Ukkonen [12] and Kleindessner and von Luxburg [18], which we consider in this work, find the medoid with score functions that are computed directly from triplet comparisons. Ukkonen et al. [29] build on this idea for nonparametric density estimation. As an alternative to a generic embedding approach, both scores provide parallelizable direct approaches to learning from triplet comparisons that are tractable on medium-sized datasets, where the embedding approach is not. For clustering, Ukkonen [28] formulates and approximates a variant of correlation-clustering [4] that takes a set of relative comparisons as input; hierarchical clustering is addressed in a setting, where one can actively query ordinal comparisons [7,33], and also in the passive setting [9]. To estimate the intrinsic dimension of the dataset, Kleindessner and Luxburg [16] propose two statistically consistent estimators that only require the k nearest neighbors for each point. Haghiri et al.

[10] construct decision trees based on triplet comparisons to find those nearest neighbors and then extend this idea for classification and regression [11]. Another approach to classification is to aggregate triplet comparisons into weak classifiers and boosting them to a strong classifier [23].

3 Preliminaries

Let $(\mathcal{X}, d)$ be a metric space and $\mathcal{D} = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ a finite set of $n \in \mathbb{N}$ points. For $x \in \mathcal{X}$, let $\mathcal{T}_x = \{(x, y, z) \mid y, z \in \mathcal{D},\ x, y, z \text{ distinct}\}$ denote the set of all ordered triplets in $\mathcal{D}$ with x in the first position. For simplicity, we assume all distances between points in $\mathcal{D}$ to be distinct. We say that x is an *outlier* in $(x, y, z) \in \mathcal{T}_x$, if it lies opposite the shortest side in the triangle given by the three datapoints, that is,

$$d(x, y) > d(y, z) \quad \text{and} \quad d(x, z) > d(y, z) .$$

Similarly, x is *central* in $(x, y, z) \in \mathcal{T}_x$, if it lies opposite the longest side in the triangle, that is,

$$d(x, y) < d(y, z) \quad \text{and} \quad d(x, z) < d(y, z) .$$

We consider three different score functions that aim to measure the centrality of any datapoint with respect to a given dataset.

Definition 1 (*Scores*) Let $x \in \mathcal{X}$ be any point in the input space. With respect to a finite dataset $\mathcal{D} = \{x_1, \ldots, x_n\} \subset \mathcal{X}$, we define the *outlier score* S_O, the *central score* S_C, and the *rank score* S_R as

$$S_O(x) \coloneqq \frac{1}{n(n-1)} \sum_{(x,y,z)\in\mathcal{T}_x} \mathbb{1}_{\{d(x,y)>d(y,z)\}} \mathbb{1}_{\{d(x,z)>d(y,z)\}} ,$$

$$S_C(x) \coloneqq \frac{1}{n(n-1)} \sum_{(x,y,z)\in\mathcal{T}_x} \mathbb{1}_{\{d(x,y)<d(y,z)\}} \mathbb{1}_{\{d(x,z)<d(y,z)\}} ,$$

$$S_R(x) \coloneqq \frac{1}{n(n-1)} \sum_{(x,y,z)\in\mathcal{T}_x} \mathbb{1}_{\{d(x,y)<d(y,z)\}} .$$

The outlier score S_O is a normalized version of the score considered in Heikinheimo and Ukkonen [12] and counts the number of triplets $(x, y, z) \in \mathcal{T}_x$ in which x is the outlier; similarly, the central score S_C is a normalized version of the score considered in Kleindessner and von Luxburg [18] and counts the number of triplets $(x, y, z) \in \mathcal{T}_x$ in which x is the central point. The rank score S_R, which we propose, compares the distance $d(x, y)$ between the fixed point x and one test point y to the distance $d(y, z)$ between both test points y and z. A corresponding crowdsourcing task would present the workers with two items A and B and then let them assign a third item C to the one which is more similar.

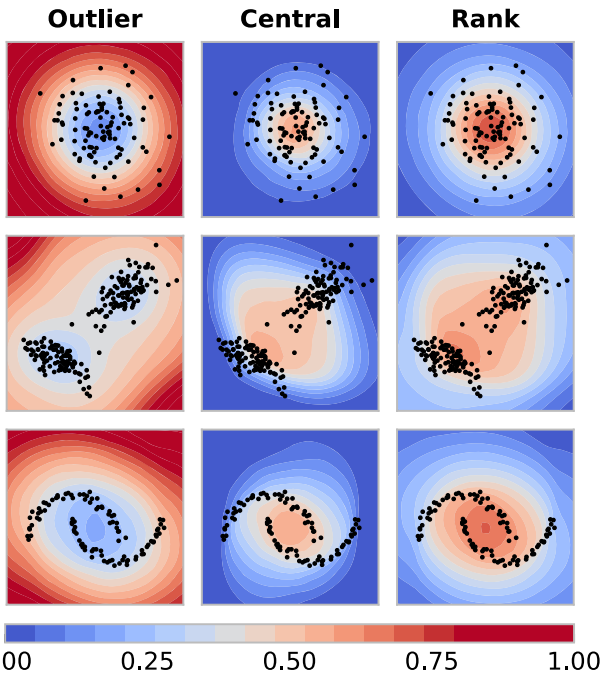


Fig. 2 Scores (columns) as a function of $x \in \mathbb{R}^2$ with respect to three datasets: 100 points from a standard Gaussian (first row), 200 points from a mixture of Gaussians (second row), and 100 points from the moons dataset (third row)

From the perspective of an unordered triplet $\{x, y, z\}$, a point x scores 0 for being the outlier, 2 for being the central point (once in (x, y, z) and once in (x, z, y)), and 1 otherwise. With that, S_R is the only of the three scores that depends on the order of the triple and distinguishes between all three possible cases. The name rank score was chosen because the score of x in a triplet $\{x, y, z\}$ corresponds to the rank of $d(y, z)$.

To get an intuition for the behavior of the three scores, we plot them as a function of $x \in \mathbb{R}^2$ with respect to three different two-dimensional datasets in Fig. 2. As expected, points in the middle of the dataset score *low* for the outlier score S_O and *high* for the central score S_C. The rank score S_R behaves similarly to the central score S_C, but transitions more softly from high to low scores.

The rank score is closely related to the outlier and the central score, as they are all based on triplets. The following proposition shows that it can be viewed as a compromise between the two: points are rewarded for being the central point in a triplet, but also penalized for being the outlier.

Proposition 1 (Relation between scores) *For any $x \in \mathcal{X}$, it holds that*

$$2S_R(x) = 1 + S_C(x) - S_O(x) .$$

Proof Let $x \in \mathcal{X}$ and $(x, y, z) \in \mathcal{T}_x$. By checking all possible strict orders on $\{d(x, y), d(x, z), d(y, z)\}$, one easily

verifies

$$\begin{aligned}&\mathbb{1}_{\{d(x,y)<d(y,z)\}} + \mathbb{1}_{\{d(x,z)<d(y,z)\}}\\ &\quad = 1 + \mathbb{1}_{\{d(x,y)<d(y,z)\}}\mathbb{1}_{\{d(x,z)<d(y,z)\}}\\ &\qquad -\mathbb{1}_{\{d(x,y)>d(y,z)\}}\mathbb{1}_{\{d(x,z)>d(y,z)\}}\,.\end{aligned}$$

This yields the claimed equality

$$\begin{aligned}2S_R(x) &= \frac{1}{n(n-1)}\sum_{(x,y,z)\in\mathcal{T}_x}\left(\mathbb{1}_{\{d(x,y)<d(y,z)\}} + \mathbb{1}_{\{d(x,z)<d(y,z)\}}\right)\\ &= \frac{1}{n(n-1)}\sum_{(x,y,z)\in\mathcal{T}_x}\left(1 + \mathbb{1}_{\{d(x,y)<d(y,z)\}}\mathbb{1}_{\{d(x,z)<d(y,z)\}}\right.\\ &\qquad \left.-\mathbb{1}_{\{d(x,y)>d(y,z)\}}\mathbb{1}_{\{d(x,z)>d(y,z)\}}\right)\\ 2S_R(x) &= 1 + S_C(x) - S_O(x)\,.\end{aligned}$$

□

From now on, we will assume $\mathcal{D} = \{X_1, \ldots, X_n\}$ to consist of i.i.d. (independent and identically distributed) samples drawn from some distribution P on $\mathcal{X}$. That is, the samples are jointly independent and every X_i is distributed as P. This allows us to treat the scores on $\mathcal{D}$ as random variables in order to analyze their statistical properties and relate them to other centrality measures such as mean, median, or mode. In our running example of a crowdsourcing task, the workers could be presented with random, independent samples from a large database of images $\mathcal{X}$.

The i.i.d. assumption is strong, but central in statistical learning theory, because it allows for strong consistency results. For example in classification, i.i.d. data ensure that nearest neighbor classifiers and support vector machines asymptotically achieve the lowest possible risk [25,31].

An important quantity for our analysis is the probability of a fixed point scoring in a random triplet:

Definition 2 (*Scoring probabilities*) Let $x \in \mathcal{X}$ and $Y, Z \sim P$ be two independent random variables distributed according to P. Then, for each score, we define the corresponding scoring probabilities

$$\begin{aligned}q_O(x) &:= \mathbb{P}(x \text{ outlier in } (x, Y, Z))\\ &= \mathbb{P}(d(x, Y) > d(Y, Z), d(x, Z) > d(Y, Z))\,,\\ q_C(x) &:= \mathbb{P}(x \text{ central in } (x, Y, Z))\\ &= \mathbb{P}(d(x, Y) < d(Y, Z), d(x, Z) < d(Y, Z))\,,\\ q_R(x) &:= \mathbb{P}(d(x, Y) < d(Y, Z))\,.\end{aligned}$$

Because of the normalization, the scoring probabilities are equal to the expected score in a random set.

Proposition 2 (Expected score) *For a fixed $x \in \mathcal{X}$ and a set of i.i.d. random variables $\mathrm{D} = \{X_1, \ldots, X_n\}$, it holds that*

$$\mathbb{E}[S_O(x)] = q_O(x)\,,\quad \mathbb{E}[S_C(x)] = q_C(x)\,,$$

$$\mathbb{E}[S_R(x)] = q_R(x)\,,$$

where the expectation is taken with respect to D. *Furthermore,*

$$2q_R(x) = 1 + q_C(x) - q_O(x)\,.$$

Proof Follows directly from the linearity of the expectation and Proposition 1. □

The following theorem shows that all three scores concentrate around their mean in the large sample size limit.

Theorem 1 *(Concentration inequality for scores) Let $S \in \{S_O, S_C, S_R\}$ be any score on a set of i.i.d. random variables $\mathrm{D} = \{X_1, \ldots, X_n\}$ and $q \in \{q_O, q_C, q_R\}$ the corresponding scoring probability. Then, for $n \geq 2$, $\varepsilon > 0$, and any $x \in \mathcal{X}$, it holds that*

$$\mathbb{P}(|S(x) - q(x)| > \varepsilon) \leq \frac{4}{\varepsilon^2 n}\,. \tag{2}$$

Equivalently, with probability greater than $1-\delta$ it holds that

$$|S(x) - q(x)| \leq \sqrt{\frac{4}{\delta n}}\,. \tag{3}$$

Proof We obtain Eq. (2) by using Chebyshev's inequality and bounding the variance with simple combinatorial arguments, and Eq. (3) is merely a reformulation. We refer to Section A of Appendix for a full version of the proof. □

Our analysis of the scores as centrality measures requires that we restrict our attention to rotationally invariant (symmetric) distributions P. Intuitively, a rotationally invariant distribution has a center such that rotations around this center do not change the distribution. For example, the standard Gaussian distribution $\mathcal{N}(0, I_d)$ on $\mathbb{R}^d$ is rotationally invariant around the origin. Section 4 requires this additional assumption to make the analysis of outlier and rank score tractable (yet still generalizes previous results from Heikinheimo and Ukkonen [12], which only considers univariate Gaussians); Sect. 5 investigates whether the scores are statistical depth functions, whose definition requires rotational invariance. Note, however, that this assumption is only needed for the theoretical analysis, which justifies the scores as centrality measures in simple settings. The scores can also produce meaningful results on more general data such as images, see Sect. 6.

Definition 3 (*Rotational invariance*) Let $X \sim P$ be a random variable on $\mathbb{R}^d$ for some $d \in \mathbb{N}$. We say that P (or X) is *rotationally invariant around c*, if $X \sim \mathrm{R}X$ for any rotation R around c, where "$\sim$" denotes "equal in distribution." A function $f : \mathbb{R}^d \to \mathbb{R}^d$ is *rotationally invariant around c*, if $f = f \circ \mathrm{R}$ for any rotation R around c.

Since the scoring probabilities are based on distances, it is not surprising that they inherit the rotational invariance from rotationally invariant distributions. Without loss of generality, the center of rotation is assumed to be the origin, because distances are invariant under translations.

Proposition 3 (Rotational invariance of scoring probabilities) *Let P be a distribution on $\mathbb{R}^d$ for some $d \in \mathbb{N}$, which is rotationally invariant around the origin according to Definition 3. Then, the scoring probabilities q_O, q_C, and q_R are also rotationally invariant around the origin.*

Proof Follows directly from the invariance of distances under rotations and the rotational invariance of P. For example, consider q_O and any rotation R. For fixed $x \in \mathbb{R}^d$ and Y, Z independently drawn from P, it holds that

$$\mathrm{R}x \text{ outlier in } \{\mathrm{R}x, Y, Z\} \Leftrightarrow x \text{ outlier in } \{x, \mathrm{R}^{-1}Y, \mathrm{R}^{-1}Z\}\,.$$

By the rotational invariance of P, we have $\mathrm{R}^{-1}Y \overset{d}{=} Y$ and $\mathrm{R}^{-1}Z \overset{d}{=} Z$. Together, this yields

$$\begin{aligned} q_O(\mathrm{R}x) &= \mathbb{P}(\mathrm{R}x \text{ outlier in } \{\mathrm{R}x, Y, Z\}) \\ &= \mathbb{P}(x \text{ outlier in } \{x, \mathrm{R}^{-1}Y, \mathrm{R}^{-1}Z\}) \\ &= \mathbb{P}(x \text{ outlier in } \{x, Y, Z\}) \\ q_O(\mathrm{R}x) &= q_O(x)\,. \end{aligned}$$

□

4 Characterization of central points

The goal of this section is to understand the behavior of the score functions with respect to the underlying distribution of the data. In particular, we characterize the points that are declared as most central by the scores. For the outlier score, the most central point is the one that is penalized least often, that is, $\arg\min_{x\in\mathcal{D}} S_O(x)$. For central and rank score, this point is the one that is rewarded most often, that is, $\arg\max_{x\in\mathcal{D}} S_C(x)$ and $\arg\max_{x\in\mathcal{D}} S_R(x)$. Theorem 1 allows us to do the analysis for the corresponding scoring probabilities instead of the scores. Unfortunately, the scoring probabilities are tractable enough to do so only for one-dimensional distributions.

The central scoring probability q_C can be computed in closed form and analyzed without any further assumptions:

Proposition 4 (Central scoring probability) *Let P be a distribution with density function $p > 0$, F its cumulative distribution function, and m its median,* i.e., *$F(m) = 1/2$. Then, it holds that*

$$q_C(x) = 2F(x)(1 - F(x))\,, \tag{4}$$

q_C is increasing on $(-\infty, m]$, decreasing on $[m, \infty)$, and m is the unique global maximum of q_C.

Proof By definition, it is $q_C(x) = \mathbb{P}(x \text{ central in } \{x, Y, Z\})$. In our particular case of $d = 1$, this can be reformulated to

$$\begin{aligned} q_C(x) &= \mathbb{P}(x \text{ is between } Y \text{ and } Z) \\ &= \mathbb{P}(Y < x < Z) + \mathbb{P}(Z < x < Y)\,, \end{aligned}$$

and since Y and Z are i.i.d. with cumulative distribution function F, we obtain Eq. (4). The derivative is given by

$$q_C'(x) = 2p(x)(1 - 2F(x))\,,$$

and since $p > 0$ and F is increasing with $F(m) = 1/2$, the sign is

$$\operatorname{sgn}\left(q_C'(x)\right) = \begin{cases} 1, & \text{if } x < m, \\ 0, & \text{if } x = m, \\ -1, & \text{if } x > m. \end{cases}$$

In particular, q_C is increasing on $(-\infty, m]$ and decreasing on $[m, \infty)$, wherefore the unique global maximum of q_C is given by m as claimed. □

Our next result covers the outlier scoring probability q_O in a more restrictive setting, yet generalizes Theorem 1 in Heikinheimo and Ukkonen [12], which only treats univariate Gaussian distributions.

Proposition 5 (Outlier scoring probability) *Let P be a distribution with density function $p > 0$, which is symmetric around the median m and increasing on $(-\infty, m]$. Then, q_O is symmetric around m and decreasing on $(-\infty, m]$. In particular, m is the unique global minimum of q_O.*

Proof Without loss of generality, we can assume $m = 0$ as argued before Proposition 3, which then yields the symmetry of q_O. As Heikinheimo and Ukkonen [12] derived, the derivative of q_O for fixed $x \in \mathbb{R}$ is given by

$$\begin{aligned} q_O'(x) = \frac{1}{2}&\left[\int_{-\infty}^{x} - \int_{x}^{\infty}\right] p(u)\, p\left(\frac{x+u}{2}\right) \mathrm{d}u \\ &+ \left[\int_{-\infty}^{x} - \int_{x}^{\infty}\right] p(u)\, p(2u - x)\, \mathrm{d}u \end{aligned}$$

where $[\int_{-\infty}^{x} - \int_{x}^{\infty}] f(u)\,\mathrm{d}u$ is shorthand for $\int_{-\infty}^{x} f(u)\,\mathrm{d}u - \int_{x}^{\infty} f(u)\,\mathrm{d}u$. Substituting $v = (x + u)/2$ in the first two integrals yields

$$q_O'(x) = 2\left[\int_{-\infty}^{x} - \int_{x}^{\infty}\right] p(u)\, p(2u - x)\, \mathrm{d}u\,. \tag{5}$$

For $x < 0$, we can lower bound a part of the second integral $\Lambda := \int_0^\infty p(u)p(2u-x)\,\mathrm{d}u$ in Eq. (5) using the monotonicity of p to obtain

$$\Lambda \geq \int_0^\infty p(u-x)p(2u-x)\,\mathrm{d}u\,.$$

Substituting $w = x - u$ on the right hand side yields

$$\Lambda \geq \int_{-\infty}^x p(-w)p(x-2w)\,\mathrm{d}w\,,$$

and with the symmetry of p around 0, we get that

$$\Lambda \geq \int_{-\infty}^x p(w)p(2w-x)\,\mathrm{d}w\,.$$

Using this lower bound in Eq. (5) yields

$$q'_O(x) \leq -2\int_x^0 p(u)p(2u-x)\,\mathrm{d}u < 0\,,$$

thus q_O is decreasing on $(-\infty, 0]$. By the symmetry of q_O around 0, it also has to be increasing on $[0, \infty)$ and thus 0 is the unique global minimum as claimed. □

Using Proposition 2, the previous two propositions can be combined to a corresponding statement about the rank scoring probability.

Corollary 1 (Rank scoring probability) *Under the conditions of Proposition 5, the rank scoring probability q_R is symmetric around the median m and increasing on $(-\infty, m]$. In particular, m is the unique global maximum.*

Proof Follows directly from the decomposition of q_R in Proposition 2 and the results on the other scoring probabilities q_O and q_C in Proposition 4 and Proposition 5. □

We tried extending this approach of directly computing the scoring probabilities by definition to distributions on $\mathbb{R}^d$; although closed-form representations for the rank scoring probability q_R and its derivative for rotationally invariant distributions P are available, they are not tractable enough to infer on the monotonicity or global maximum of q_R.

5 The rank score and statistical depth functions

The main motivation for the central score is its relation to the lens depth function [21], which is assumed to be a statistical depth function. It is important to note that this has yet to be proven, as Kleindessner and von Luxburg [18, Section 5.2] pointed out a mistake in the proof of properties P2 and P3 (defined below). The outlier score, on the other hand, is provably not related to a statistical depth function, because its scoring probability does not satisfy those two properties. A counterexample for this was already given by Heikinheimo and Ukkonen [12] for symmetric bimodal distributions in one dimension.

In this section, we show that the rank scoring probability also does not satisfy P2 and P3. However, Proposition 6 shows that it satisfies a weaker version for rotationally invariant distributions, which extends to the rank score in Theorem 2. For the remainder of this paper, we only consider $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$.

5.1 Statistical depth functions

Denote by $\mathcal{F}$ the class of distributions on the Borel sets of $\mathbb{R}^d$ and by F_ξ the distribution of a given random vector ξ. The four desirable properties that an ideal depth function $D\colon \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}$ should possess as defined in Zuo and Sering [35] are

P1. *Affine invariance*. The depth of a point $x \in \mathbb{R}^d$ should not depend on the underlying coordinate system or, in particular, on the scales of the underlying measurements. For any random vector X in $\mathbb{R}^d$, any non-singular $A \in \mathbb{R}^{d\times d}$ and any $b \in \mathbb{R}^d$, it should hold that $D(Ax+b, F_{Ax+b}) = D(x, F_X)$.

P2. *Maximality at center.* For a distribution having a uniquely defined "center" (*e. g.*, the point of symmetry with respect to some notion of symmetry), the depth function should attain maximum value at this center. This means $D(\theta, F) = \sup_{x\in\mathbb{R}^d} D(x, F)$ holds for any $F \in \mathcal{F}$ having center θ.

P3. *Monotonicity relative to deepest point.* As a point $x \in \mathbb{R}^d$ moves away from the "deepest point" (the point at which the depth function attains maximum value; in particular, for a symmetric distribution, the center) along any fixed ray through the center, the depth at x should decrease monotonically. This means for any $F \in \mathcal{F}$ having deepest point θ, $D(x, F) \leq D(\theta + \alpha(x-\theta), F)$ holds for any $\alpha \in [0, 1]$.

P4. *Vanishing at infinity.* The depth of a point x should approach zero as $\|x\|$ approaches infinity. This means that $D(x, F) \to 0$ as $\|x\| \to \infty$ for each $F \in \mathcal{F}$.

Property P1 is not satisfied by any of the scores in this general form, but it holds for similarity transformations, because they preserve inequalities between distances. That is, whenever $A = rQ$ for a scalar $r \in \mathbb{R}_+$ and an orthogonal matrix $Q \in \mathbb{R}^{d\times d}$. P1 also holds for non-singular $A \in \mathbb{R}^{d\times d}$ with respect to the Mahalanobis distance, which itself depends on the distribution [21, Theorem 1]. In this case, the transformation preserves the distances.

The rank scoring probability does not satisfy P2 and P3 and is therefore not a statistical depth function. As a counterexample, consider the uniform distribution on $\{-5, -3, 3, 5\}$, which is rotationally invariant around 0 according to Definition 3. For this distribution, it is $q_R(0) = 8/16 < 9/16 = q_R(2)$, which violates P2, because q_R is not maximal at the center of rotation 0, and P3, because it is non-decreasing on lines away from 0.

The rank scoring probability satisfies P4 by Lebesgue's dominated convergence theorem and $\mathbb{1}_{\{d(x,y)<d(y,z)\}} \to 0$ as $\|x\| \to \infty$ for any $y, z \in \mathbb{R}^d$.

5.2 Rank score is approximately decreasing

We now show that the rank scoring probability at least satisfies a weaker version of P2 and P3 in Proposition 6, which translates to the rank score itself in Theorem 2. First, we give an alternative formula for the rank scoring probability:

Lemma 1 *(Formula for rank scoring probability) Let $x \in \mathbb{R}^d$, Y a random variable with distribution P, which is rotationally invariant around the origin, and q_R defined on P. Let S and R_x be two independent random variables, distributed as $S \sim Y_1$ and $R_x \sim \left(\|Y\|^2 - \|x\|^2\right) / (2\|Y - x\|)$, where Y_1 denotes the first coordinate of Y. Then,*

$$q_R(x) = \mathbb{P}(S \leq R_x)\,. \tag{6}$$

Proof Here, we provide a sketch of the proof; for a full version of the proof, we refer to Section B in Appendix. The idea is to use the definition of q_R and rotate the appearing areas of integration such that they can be described in terms of the marginal distribution Y_1 alone. Because P is rotationally invariant by assumption, doing so does not change the value of q_R. Analyzing the involved rotations, which determine the distribution of R_x, concludes the proof. □

Our next result is a weak version of properties P2 and P3 from Sect. 5.1 and states that the rank scoring probability q_R is at least approximately decreasing away from the center of rotation.

Proposition 6 (q_R is approximately decreasing) *Let Y_1 be the first coordinate of Y, which is distributed as an around the origin rotationally invariant distribution P. Then, for any $x, x' \in \mathbb{R}^d$ with $\|x'\| \leq \|x\|$, it holds that*

$$q_R(x) \leq q_R(x') + \mathbb{P}(Y_1 > \|x\|)\,. \tag{7}$$

Proof To derive the upper bound (7), we use the reformulation of q_R provided in Eq. (6) in combination with bounds for the auxiliary function R_x. For a complete proof, we refer to Appendix, Section C. □

The bound in Proposition 6 uses the tail of the marginal distribution of P. To control these tail probabilities, we consider a special class of distributions:

Definition 4 (*sub-Gaussian*, [32, Section 3.4]) Let $X \sim P$ be a random variable on $\mathbb{R}^d$ for some $d \in \mathbb{N}$. We say that P (or X) is *sub-Gaussian*, if all one-dimensional marginals $\langle X, x\rangle$ for $x \in \mathbb{R}^d$ satisfy

$$\mathbb{P}\left(|\langle X, x\rangle| > t\right) \leq 2e^{-c_x t^2}$$

for all $t \geq 0$ and suitable $c_x > 0$.

Under the assumption that P is sub-Gaussian, we obtain a more specific bound for the rank scoring probability q_R.

Corollary 2 *(sub-Gaussian bound for q_R) Let Y_1 be the first coordinate of Y, which is distributed as an around the origin rotationally invariant and sub-Gaussian distribution P. Then, for any $x, x' \in \mathbb{R}^d$ with $\|x'\| \leq \|x\|$, it holds for suitable $c > 0$ that*

$$q_R(x) \leq q_R(x') + e^{-c\|x\|^2}\,.$$

Proof Since Y is sub-Gaussian, all one-dimensional marginals $\langle Y, y\rangle$ for $y \in \mathbb{R}^d$ satisfy

$$\mathbb{P}\left(|\langle Y, y\rangle| > t\right) \leq 2e^{-c_y t^2} \tag{8}$$

for all $t \geq 0$ and suitable $c_y > 0$. Under our assumption on P to be rotationally invariant around the origin, Y being sub-Gaussian is even equivalent to only requiring condition Eq. (8) for $y = e_1$, because the distribution of any one-dimensional marginal depends only on $\|y\|$. Furthermore, Y_1 is necessarily symmetric and we obtain

$$\begin{aligned}\mathbb{P}(Y_1 > t) &= \frac{1}{2}\left(\mathbb{P}(Y_1 > t) + \mathbb{P}(Y_1 < -t)\right)\\ &= \frac{1}{2}\,\mathbb{P}\left(|\langle Y, e_1\rangle| > t\right)\,.\end{aligned}$$

Denoting $c := c_{e_1}$ in Eq. (8) yields

$$\mathbb{P}(Y_1 > t) \leq e^{-ct^2}\,.$$

We use this upper bound in Eq. (7) to complete the proof. □

Combining Corollary 2 and Theorem 1 yields the main result of this section, which shows that under reasonable assumptions on the distribution P, the rank score S_R is approximately decreasing for points x far away from the center of symmetry.

Theorem 2 *(S_R is approximately decreasing) Let P be an around the origin rotationally invariant and sub-Gaussian distribution on $\mathbb{R}^d$. Then, for suitable $c > 0$, any $0 < \delta < 1$ and $x, x' \in \mathbb{R}^d$ with $\|x'\| \leq \|x\|$, the rank score S_R, defined on a dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$ of $n \geq 2$ i.i.d. samples from P, satisfies with probability greater than $1 - \delta$ that*

$$S_R(x) \leq S_R(x') + e^{-c\|x\|^2} + \sqrt{\frac{40}{\delta n}}\,. \tag{9}$$

Proof Denote $A_\chi := \left\{|S_R(\chi) - q_R(\chi)| \leq \sqrt{10/(\delta n)}\right\}$ for $\chi \in \{x, x'\}$. Using Eq. (3) in Theorem 1 with $\delta/2$ yields

$$\begin{aligned}\mathbb{P}(A_x, A_{x'}) = 1 - \mathbb{P}\left(A_x^{\mathsf{c}} \cup A_{x'}^{\mathsf{c}}\right) &\geq 1 - \mathbb{P}\left(A_x^{\mathsf{c}}\right) - \mathbb{P}\left(A_{x'}^{\mathsf{c}}\right) \\ &\geq 1 - \frac{\delta}{2} - \frac{\delta}{2} = 1 - \delta\,.\end{aligned}$$

Therefore, with probability greater than $1 - \delta$ (on the set $A_x \cap A_{x'}$), it holds that

$$S_R(x) \leq q_R(x) + \sqrt{\frac{10}{\delta n}} \tag{A_x}$$

$$\leq q_R(x') + e^{-c\|x\|^2} + \sqrt{\frac{10}{\delta n}} \tag{Corollary 2}$$

$$S_R(x) \leq S_R(x') + e^{-c\|x\|^2} + \sqrt{\frac{40}{\delta n}}\,, \tag{$A_{x'}$}$$

as claimed. □

The bound in Eq. (9) depends on $\|x\|$ and n, which are decoupled in two terms $\exp\left(-c\,\|x\|^2\right)$ and $\sqrt{40/(\delta n)}$. The latter goes to 0 as $n^{-0.5}$ in the large sample size limit and accounts for the error made by using a finite set of samples $\mathcal{D}$. Although we would like the first term to be 0, the counterexample in Sect. 5.1 shows that q_R is not always decreasing as a function of the norm, wherefore some dependency on $\|x\|$ is necessary. It vanishes exponentially for datapoints of large norm; in this case, however, we already have the statement $S_R(x) \to 0$ by the property P4 ("vanishing at infinity") of q_R and Theorem 1. There is no explicit dependency on the other datapoint x' besides $\|x'\| \leq \|x\|$, because this is sufficient for bounding the auxiliary function R_x in Lemma 2 in Appendix.

As a consequence of Theorem 2, the score S_R cannot blow up far away from the center of rotation with high probability. Since it uses high scores to infer centrality, this means that no obvious outliers are estimated to be central. A decreasing score S_R would imply that our score always recovers the center of rotation as the most central point; in that sense, this proposition controls the distance of our estimation to the center.

We would like to improve the proposition by removing the $\|x\|$ term for dimensions $d > 1$. As mentioned above, this term is necessary for $d = 1$, but simulations in higher dimensions suggest that the scoring probability q_R is actually always decreasing.

6 Experiments

We have shown in Sect. 4 that the scores recover the median for one-dimensional symmetric distributions. But to what notion of centrality do the scores correspond in more general settings? We also observed in Sect. 5 that only the central score comes with the desirable property of being a statistical depth function. But does this imply that it performs better in practice? To answer these questions, we evaluate the scores on medoid estimation and outlier detection tasks on real and synthetic datasets. Medoid estimation aims to find the most central point in a dataset, which minimizes the average distance to other points. Outlier detection has the opposing goal of identifying points that are considerably different from the majority of the dataset. The experiments on real datasets in Sect. 6.1 show no notable difference between the three scores and suggest that they recover the Euclidean notion of centrality given by the average distance to the dataset. We then highlight some of the more subtle weaknesses of and differences between the scores on synthetic datasets in Sect. 6.2.

6.1 Image datasets

Datasets and preprocessing. Our first dataset is the well-known MNIST database of handwritten digits [20] reduced to 300 randomly chosen images per digit. Since these random subsets are unlikely to include clearly visible outliers, each subset was expanded by the 5 images with largest average Euclidean distance within each class. Our other dataset NATURE [22] consists of outdoor scene photographs for 8 landscape and urban categories (COAST, FOREST, HIGHWAY, INSIDE CITY, MOUNTAIN, OPEN COUNTRY, STREET, TALL BUILDING). The number of images per category ranges from 260 to 410.

Ideally, we would like to compute the scores within each category based on triplets obtained in a crowdsourcing task. However, the cubic number of triplets makes this infeasible even for medium-sized datasets (300 images yield ≈ 4,500,000 triplets). Because we want to avoid any finite sampling effects induced by using only a subset of triplets, we use two rough proxies for triplets labeled by humans: triplets are computed based on the Euclidean distances (1) in the pixel space and (2) for a feature embedding given by a neural network. For the latter, we use the AlexNet architecture [19]

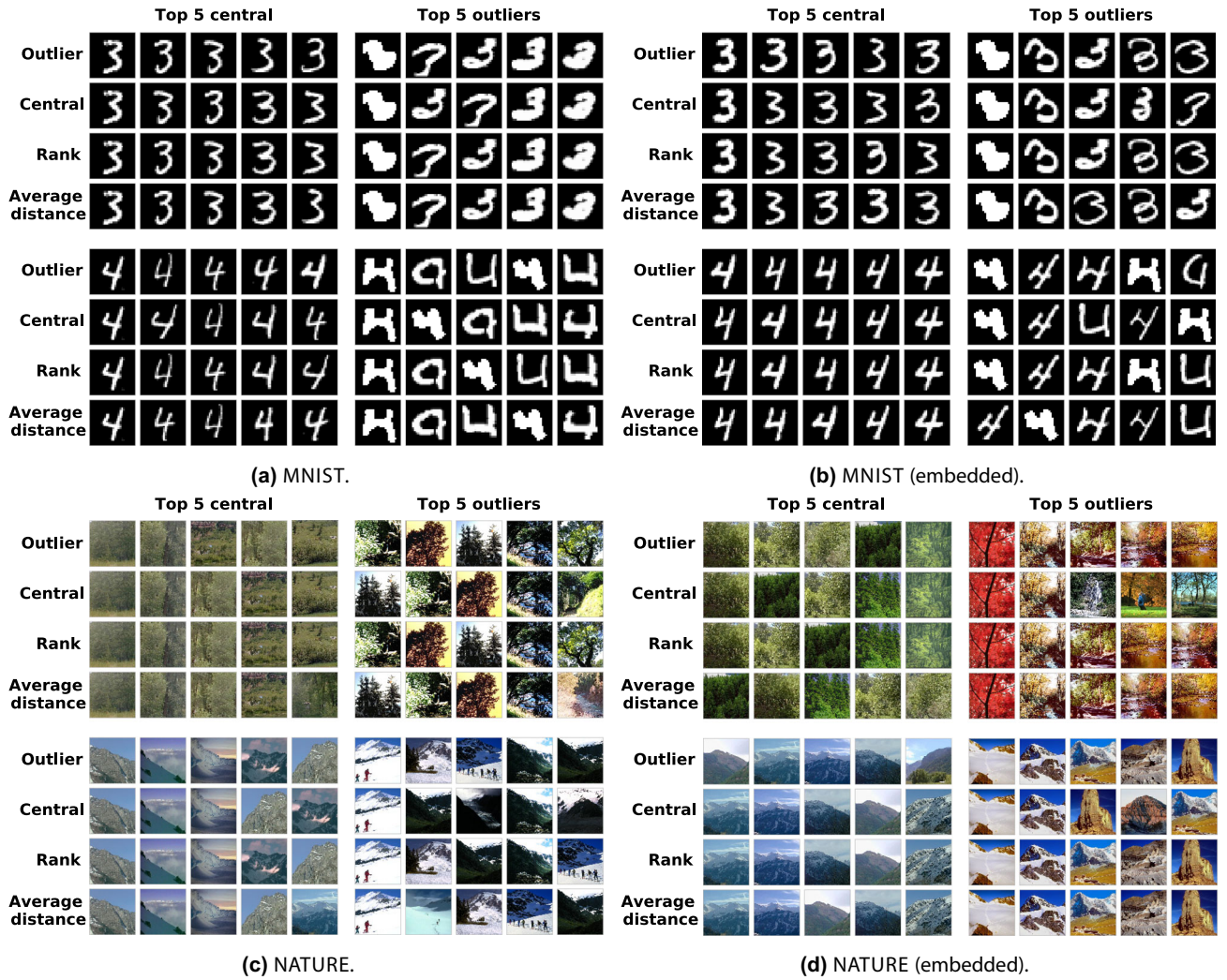


Fig. 3 Top 5 rankings on MNIST classes 3 and 4, and NATURE categories FOREST and MOUNTAIN with and without AlexNet feature embedding. Each quadrant shows top 5 central points and outliers of the respective dataset as predicted by the three scores and the average Euclidean distance baseline

pre-trained on IMAGENET [6]. The feature embedding is then given by the network after removing the last three fully connected and softmax layers. This transfer-learning approach is motivated by the generality of representations learned in the first few network layers across different domains [5,34]. Note that although a Euclidean representation of the data is available, the scores can still only use the resulting triplet information to access the data.

Experiment setup and evaluation metric. We compare the scores in medoid estimation and outlier detection tasks for each class of both datasets and for both Euclidean spaces. For a fixed class and Euclidean space, let $\mathcal{I} = \{x_1, \ldots, x_n\}$ denote the corresponding set of image representations. First, all three scores are computed for every image in $\mathcal{I}$ based on the triplets given by the Euclidean distances. As a baseline measure of centrality, we additionally consider the average Euclidean distance from $x \in \mathcal{I}$ to the other elements, which is given by $(1/|\mathcal{I}|) \sum_{y \in \mathcal{I}} d(x, y)$. Each of the four centrality measures is then used to produce a ranking of the top 5 central points by choosing the points with highest values for central score and rank score, and the points with lowest values for outlier score and average distance. Similarly, a ranking of the top 5 outliers is obtained by choosing the points at the opposite tail end of the centrality measures.

To quantify the distance between the top 5 rankings, we use a normalized version of the averaging footrule distance F_{avg} proposed by Fagin et al. [8], where low distance implies that the rankings are similar. This distance generalizes Spearman's footrule, which is the L_1-distance between two permutations, to top k lists. It considers both the order and the number of shared elements, and values range from 0 for identical rankings to 1 for rankings on disjoint elements. As an example, it is $F_{\text{avg}}([1, 2, 3, 4, 5], [1, 3, 4, 5, 8]) = 0.16$,

Table 1 Average F_{avg} distances on top 5 rankings (a) between outlier score, central score, and rank score, and (b) from scores to average Euclidean distance baseline. Rankings are computed across all classes for MNIST and NATURE, both feature representations, and both top 5 central points and top 5 outliers

	Pixel space		AlexNet embedding	
	Top 5 Central	Top 5 Outliers	Top 5 Central	Top 5 Outliers
(a) Average F_{avg} distance on top 5 rankings between scores				
Digit 0	0.11	**0.09**	**0.08**	0.16
— 1	0.32	**0.11**	0.27	0.16
— 2	0.15	**0.08**	0.32	0.31
— 3	0.13	**0.03**	*0.39*	0.11
— 4	0.19	**0.11**	**0.03**	0.15
— 5	0.37	0.16	**0.08**	**0.16**
— 6	0.37	**0.05**	0.16	**0.05**
— 7	**0.11**	**0.05**	**0.11**	0.21
— 8	0.24	**0.03**	**0.05**	0.15
— 9	0.20	**0.05**	**0.15**	0.24
Coast	0.13	0.27	**0.03**	**0.08**
Forest	**0.03**	**0.08**	**0.11**	0.25
Highway	0.17	**0.17**	**0.07**	**0.05**
Inside city	**0.03**	0.17	**0.03**	**0.05**
Mountain	**0.03**	0.21	**0.11**	**0.11**
Open country	**0.05**	**0.12**	**0.08**	**0.08**
Street	**0.11**	**0.08**	*0.49*	**0.03**
Tall building	**0.15**	0.27	0.15	**0.08**
(b) Average F_{avg} distance on top 5 rankings from scores to average Euclidean distance baseline				
Digit 0	0.16	**0.05**	**0.05**	0.09
— 1	0.20	**0.12**	0.13	0.09
— 2	0.11	**0.07**	0.19	0.21
— 3	0.07	**0.01**	*0.27*	0.13
— 4	0.20	**0.11**	**0.05**	0.16
— 5	0.25	0.12	**0.11**	**0.17**
— 6	0.32	**0.04**	0.15	**0.05**
— 7	**0.09**	**0.16**	**0.07**	0.19
— 8	0.24	**0.01**	**0.07**	0.11
— 9	0.13	**0.05**	**0.15**	0.16
Coast	0.07	0.15	**0.04**	**0.11**
Forest	**0.05**	**0.09**	0.17	0.13
Highway	0.17	0.19	0.07	**0.04**
Inside city	**0.08**	0.19	**0.03**	**0.07**
Mountain	**0.07**	0.28	**0.07**	**0.05**
Open country	**0.11**	**0.11**	**0.09**	**0.12**
Street	**0.05**	**0.19**	*0.32*	**0.01**
Tall building	0.43	0.23	0.07	**0.04**

Numbers are bold, if the respective unordered rankings agree on at least 4 out of the 5 images, and italic, if they agree on at most one. In every case, the rankings agree on at least one image. The low values in this table imply that all rankings are very similar

and the expected distance between any fixed top 5 ranking and a random one on a domain with 300 elements is ≈ 0.98.

Evaluation. Figure 3 shows the results on two classes for each dataset. The most striking observation is that the predictions for central points and outliers are almost consistent across all four centrality measures. This shows that **all three scores have comparable performance in medoid estimation and outlier detection tasks**, despite their differences in theoretical guarantees. Since the score-induced rankings agree with those of the average Euclidean distance baseline, this observation also suggests that **the scores recover the Euclidean notion of centrality**. This is in line with our one-dimensional findings of Sect. 4: the scores recover the median of a distribution P as the most central point, and the median of a continuous distribution minimizes the expected distance $\mathbb{E}_{X\sim P}\,[|x - X|]$ over all $x \in \mathbb{R}$. These findings are supported by Table 1, which quantifies the distance between the rankings for all datasets. The small distances in Table 1a show that the top 5 rankings produced by the three scores are very similar, and the small distances in Table 1b show that these rankings agree with those given by the average Euclidean distance.

A second observation is that these findings are **consistent across both feature representations** of the datasets, the pixel space and the AlexNet embedding. Therefore, they cannot be dismissed as an artifact of the pixel space. Since triplets based on the AlexNet embedding might resemble the human notion of similarity more closely, it is possible that rankings on crowdsourced triplets display similar behavior.

The third observation concerns the quality of the estimated medoids and outliers and has already been made for the outlier score [12] and the central score [18]: **estimated medoids are generic images of a class, while estimated outliers are more diverse.** This also holds true for the rank score. For example, in Fig. 3d for the class FOREST, the central images are thematically homogeneous and all show green trees. The outliers contain different colors like green, red, and yellow and include diverse motifs like a person, trees, and water. The notion of centrality captured by the scores changes with the feature space, but the general concept of generic and diverse stays the same. For example, in Fig. 3a, the pixel space representation identifies almost exclusively thick digits as outliers, while the AlexNet embedding in Fig. 3b also includes thin, skewed digits.

6.2 Synthetic datasets

In this section, we partly repeat experiments from Kleindessner and von Luxburg [18] on two simple synthetic datasets for which we additionally include the rank score. First, we demonstrate the relation between the scores and the average distance. We then highlight some of their more subtle differences for the tasks of medoid estimation and outlier identification.

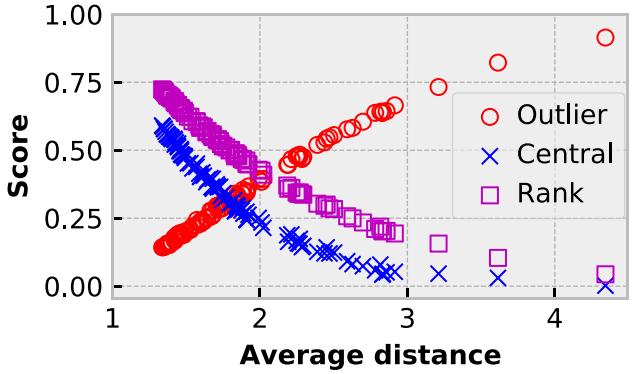
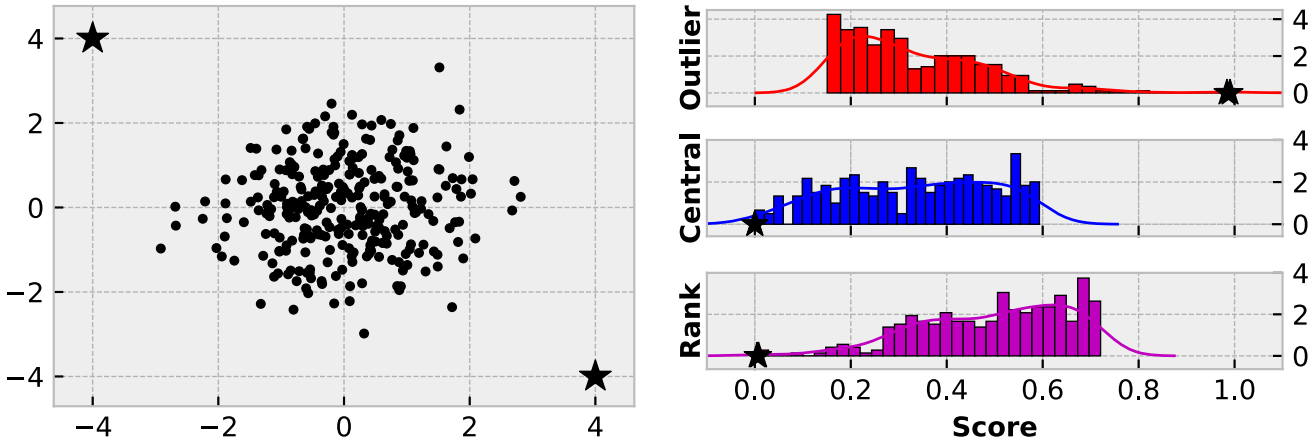


Fig. 4 Scores on a set $\mathcal{X}$ of 100 points from a two-dimensional standard Gaussian, plotted against the average Euclidean distance $(1/|\mathcal{X}|)\sum_{y\in\mathcal{X}} d(x, y)$

Relation to average distance. Figure 4 shows the scores for a Gaussian dataset plotted against the average distance. As expected, the outlier score is roughly increasing as a function of the average distance while central and rank score are roughly decreasing. Hence the scores serve as a proxy for the average distance, which was already observed in Sect. 6.1. We can quantify this relationship with Spearman's rank correlation coefficient, which describes the monotonicity between two variables. The respective values for outlier, central, and rank score are 0.997, -0.995, and -0.999, which implies an almost perfect monotonic relationship. The corresponding tail ends of the scores therefore contain candidates for medoid and outliers. However, this does not tell us *how many* outliers there are, if any at all, because this would require to identify gaps in the average distance. The scores are less sensitive to such gaps, because they are based on relative rather than absolute information. This can be observed in Fig. 4 for the two rightmost average distance values, whose gap is not accompanied by a corresponding gap in the scores. This behavior is further discussed in the next paragraph under outlier identification.

Medoid estimation and outlier identification. This paragraph highlights two respective weaknesses of outlier and central score. The rank score shows neither of them, as it is a compromise between outlier and central score.

For the first task of medoid estimation, we consider the circular dataset shown in Fig. 6. Each score tries to predict the medoid by returning the point at its corresponding tail end. That is, the outlier score returns the point with lowest score, whereas central and rank scores return the point with highest score. Central and rank scores correctly predict the origin, but the outlier score predicts the point indicated by the red circle. The heatmap for the outlier score shows that the area around the medoid is in fact a local maximum instead of the global minimum. This effect was already observed in one



(a) Dataset. **(b)** Histogram and kernel density estimation of scores.

Fig. 5 300 points drawn from a standard Gaussian and 2 outliers (stars) added by hand. Histograms with kernel density estimate show the three scores, for which the outliers are indicated with stars

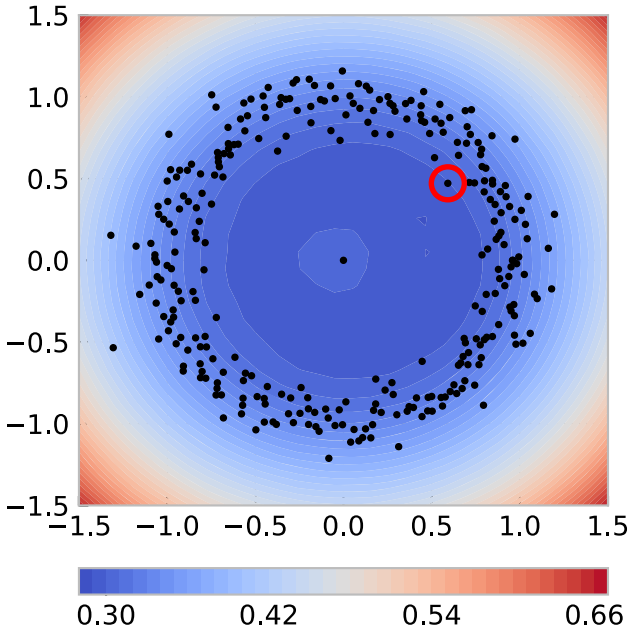


Fig. 6 300 points drawn from a rotationally symmetric distribution with norm distributed as $\mathcal{N}(1, 0.01)$ and medoid at origin added by hand. The heatmap shows the outlier score of $x \in \mathbb{R}^2$ with respect to this dataset, and the red circle indicates the point in the dataset with lowest outlier score

dimension for bimodal symmetric distributions as discussed in Sect. 5 and attests to the fact that the outlier score is not a statistical depth function. A statistical depth function would declare points at heart of the dataset as central, even if this region is sparse, because it ignores multimodal aspects of distributions. Another example in which the outlier score respects the multimodality of a dataset is given by the mixture of Gaussians in Fig. 2.

For the second task of outlier identification, Fig. 5a shows points from a standard Gaussian with two clearly visible outliers added by hand. Similar to medoid estimation, each score now proposes candidates for outliers at its other tail end, that is, highest points for the outlier score and lowest points for central and rank score. Figure 5b shows that both outliers are correctly placed at the corresponding tail end for all three scores. However, only outlier and rank score show a clear gap between the scores of the outliers and the other points. The lack of such a gap makes it hard for the central score to estimate the amount of outliers.

7 Conclusion and future work

In this paper, we consider three comparison-based centrality measures on triplets, one of which we propose as a natural compromise between the other two. We provide a theoretical analysis of the scores for one-dimensional distributions to characterize their most central points and investigate their connection to statistical depth functions. We conclude with experiments for the tasks of medoid estimation and outlier detection: on image datasets, we demonstrate the behavior of the three scores and hint toward their connection to the average Euclidean distance. On synthetic datasets, we highlight two respective weaknesses of the existing two scores, which are mitigated by our proposed score.

As for future work, it is an open question whether the rank score is a statistical depth function for dimensions $d \geq 2$. Similarly, fixing the proof for the lens depth function would solidify the motivation for the central score. We further plan to investigate and formalize the connection between the scores and the average Euclidean distance.

Acknowledgements The authors thank Ulrike von Luxburg, Debarghya Goshdastidar, and Michaël Perrot for fruitful discussions. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Luca Rendsburg.

Funding Open Access funding enabled and organized by Projekt DEAL. This work has been supported by the German Research Foundation through the Cluster of Excellence "Machine Learning—New Perspectives for Science" (EXC 2064/1 No. 390727645) and the BMBF Tübingen AI Center (FKZ: 01IS18039A).

Availability of data and material The datasets are available as referenced in the main paper.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availability There is no code available.

Appendix

A Proof of Theorem 1

In this section, we prove the convergence of the scores to their corresponding scoring probabilities.

Theorem 1 *(Concentration inequality for scores) Let* $S \in \{S_O, S_C, S_R\}$ *be any score on a set of i.i.d. random variables* $\mathrm{D} = \{X_1, \ldots, X_n\}$ *and* $q \in \{q_O, q_C, q_R\}$ *the corresponding scoring probability. Then, for* $n \geq 2$, $\varepsilon > 0$, *and any* $x \in \mathcal{X}$, *it holds that*

$$\mathbb{P}\left(|S(x) - q(x)| > \varepsilon\right) \leq \frac{4}{\varepsilon^2 n}\,. \tag{2}$$

Equivalently, with probability greater than $1 - \delta$ *it holds that*

$$|S(x) - q(x)| \leq \sqrt{\frac{4}{\delta n}}\,. \tag{3}$$

Proof All three scores are of the generic form

$$S(x) = \frac{1}{n(n-1)} \sum_{\substack{i,j=1,\\ i\neq j}}^{n} I(x, X_i, X_j)\,,$$

where $I(x, X_i, X_j)$ is the corresponding indicator function taking values in $\{0, 1\}$. In order to use Chebyshev's inequality, we first upper bound the variance of $S(x)$ using the known formula

$$\mathrm{Var}(S(x)) = \frac{1}{(n(n-1))^2} \left(\sum_{\substack{i,j=1,\\ i\neq j}}^{n} \mathrm{Var}\left(I(x, X_i, X_j)\right) + \sum_{\substack{i,j=1,\\ i\neq j}}^{n} \sum_{\substack{i',j'=1,\\ i'\neq j',\\ (i,j)\neq(i',j')}}^{n} \mathrm{Cov}\left(I(x, X_i, X_j), I(x, X_{i'}, X_{j'})\right) \right).$$

The variance and covariances in the formula above are upper bounded by 1, because the involved random variables only take values in $\{0, 1\}$. Since $\mathcal{D}$ consists of independent samples, the covariance is 0, if i, i', j, j' are distinct. Therefore, the covariances are only summed over the set A of all remaining pairs. Denoting $B := \left\{((i, j) \in \{1, \ldots, n\}^2 \mid i \neq j\right\}^2$ the set of all possible pairs, we have

$$A = B \setminus \Big(\left\{((i, j), (i', j')) \in B \mid (i, j) = (i', j')\right\} \uplus \left\{((i, j), (i', j')) \in B \mid i, i', j, j' \text{ distinct}\right\} \Big)$$

and thus

$$\begin{aligned} |A| &= (n(n-1))^2 - (n(n-1) + n(n-1)(n-2)(n-3)) \\ &= n(n-1)(4n-7)\,. \end{aligned}$$

This yields

$$\begin{aligned} \mathrm{Var}(S(x)) &\leq \frac{1}{(n(n-1))^2} \left(n(n-1) + n(n-1)(4n-7)\right) \\ &\leq \frac{4}{n}\,. \end{aligned}$$

As shown in Proposition 2, it is $\mathbb{E}[S(x)] = q(x)$ and therefore Chebyshev's inequality completes the first part of the proof. The reformulation follows by choosing $\delta := 4/(\varepsilon^2 n)$. □

B Proof of Lemma 1

Lemma 1 (Formula for *rank* scoring probability) *Let $x \in \mathbb{R}^d$, Y a random variable with distribution P, which is rotationally invariant around the origin, and q_R defined on P. Let S and R_x be two independent random variables, distributed as $S \sim Y_1$ and $R_x \sim \left(\|Y\|^2 - \|x\|^2\right) / \left(2\,\|Y - x\|\right)$, where Y_1 denotes the first coordinate of Y. Then,*

$$q_R(x) = \mathbb{P}(S \leq R_x)\,. \tag{6}$$

Proof Let $x \in \mathbb{R}^d$ and denote the marginal density of Y_1 by p_1. For clarity, we explicitly indicate the corresponding random variable at the density as in $p = p_Y$. By definition, it is $q_R(x) = \mathbb{P}(d(x, Y) < d(Y, Z))$. Since Y and Z are i.i.d. and δ is symmetric in its arguments, the right hand side is equal to $\mathbb{P}(d(x, Y) < d(Y, Z))$, and marginalizing over Y yields

$$q_R(x) = \int_{\mathbb{R}^d} p_Y(y) \int_{\{z \in \mathbb{R}^d \mid d(x,z) < d(y,z)\}} p_Z(z)\,\mathrm{d}z\,\mathrm{d}y\,.$$

Because P is rotationally invariant around the origin, we can rotate the half-space $H_x(y) \coloneqq \left\{z \in \mathbb{R}^d \mid d(x, z) < d(y, z)\right\}$ for fixed $y \in \mathbb{R}^d$, the area of integration for the inner integral, without changing its mass under P. In order to express $q_R(x)$ solely by a marginal distribution of P, we rotate the half-space with the unique rotation R that yields the rotated set $\left\{z \in \mathbb{R}^d \mid z_1 < R_x(y)\right\}$ for an appropriate value of $R_x(y)$. By doing so, we obtain

$$\begin{aligned} q_R(x) &= \int_{\mathbb{R}^d} p_Y(y) \int_{\{z \in \mathbb{R}^d \mid z_1 < R_x(y)\}} p_Z(z)\,\mathrm{d}z\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} p_Y(y) \int_{-\infty}^{R_x(y)} p_1(s)\,\mathrm{d}s\,\mathrm{d}y\,. \end{aligned}$$

A change of variables under the function $y \mapsto R_x(y)$ yields

$$\begin{aligned} q_R(x) &= \int_{\mathbb{R}} p_{R_x(Y)}(r) \int_{-\infty}^{r} p_1(s)\,\mathrm{d}s\,\mathrm{d}r \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{1}_{\{s \leq r\}} p_{R_x(Y)}(r)\, p_1(s)\,\mathrm{d}s\,\mathrm{d}r\,, \end{aligned}$$

which we describe in terms of independent random variables $S \sim Y_1$ and R_x as

$$q_R(x) = \mathbb{P}\left(S \leq R_x\right)\,.$$

It is left to prove $R_x \sim \left(\|Y\|^2 - \|x\|^2\right) / \left(2\,\|Y - x\|\right)$, which we achieve by determining a closed form for $R_x(y)$. Let $\xi_x(y)$ denote the unique point on $\partial H_x(y)$ that satisfies $\mathrm{R}\left(\xi_x(y)\right) = R_x(y)e_1$; this situation is depicted in Fig. 7. Because $(y - x)/\|y - x\|$ is the outward pointing normal vector for $H_x(y)$ before rotating and e_1 is the outward pointing normal vector after rotating, we get that

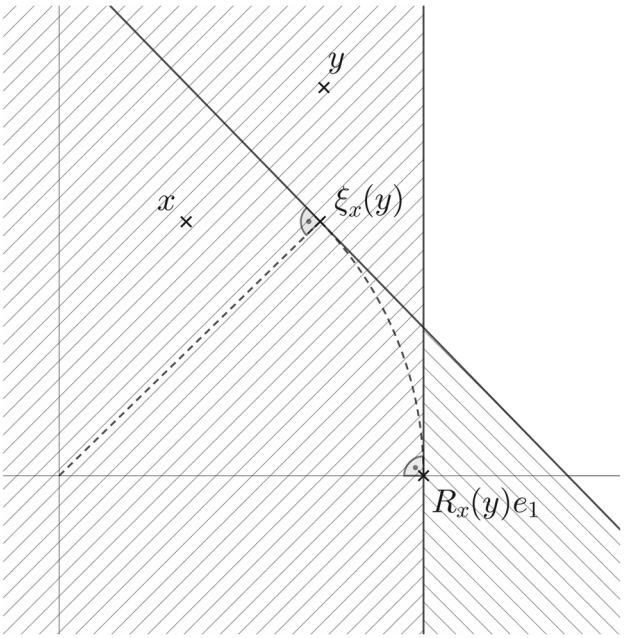


Fig. 7 Illustration of situation in proof of Lemma 1. Rotation of point $\xi_x(y) \in \partial H_x(y)$ with smallest distance to origin onto the x-axis for $x = (1, 2)^\top$ and $y = (2, 3)^\top$ results in point $R_x(y)e_1$. Half-space $H_x(y)$ (shaded area) is rotated with same rotation (resulting in hatched area)

$$\mathrm{R}\left(R_x(y)\frac{y - x}{\|y - x\|}\right) = R_x(y)\mathrm{R}\left(\frac{y - x}{\|y - x\|}\right) = R_x(y)e_1\,.$$

Because $\xi_x(y)$ is the unique point that satisfies $\mathrm{R}\left(\xi_x(y)\right) = R_x(y)e_1$, we obtain

$$\xi_x(y) = R_x(y)\frac{y - x}{\|y - x\|}\,. \tag{10}$$

Since $\xi_x(y), (x + y)/2 \in \partial H_x(y)$ are both points on the dividing hyperplane, they have to satisfy the equation $0 = \langle \xi_x(y) - (x + y)/2, y - x\rangle$. Combined with Eq. (10), this yields

$$\begin{aligned} 0 &= \left\langle \xi_x(y) - \frac{x + y}{2}, y - x \right\rangle \\ &= \left\langle R_x(y)\frac{y - x}{\|y - x\|} - \frac{x + y}{2}, y - x \right\rangle \\ &= R_x(y)\,\|y - x\| - \frac{1}{2}\left(\|y\|^2 - \|x\|^2\right)\,, \end{aligned}$$

or equivalently

$$R_x(y) = \frac{\|y\|^2 - \|x\|^2}{2\,\|y - x\|}\,,$$

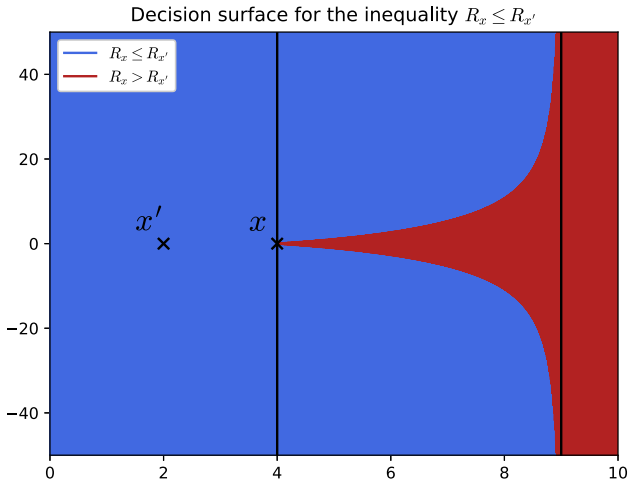


Fig. 8 Decision surface for inequality $R_x(y) \leq R_{x'}(y)$ as function of y with $x = (4, 0)$ and $x' = (2, 0)$, corresponding to $\alpha = 0.5$ in Lemma 2. Left side of line at $\langle x, \cdot\rangle = \langle x, x\rangle$ corresponds to the first case, and right side of line at $\langle x, \cdot\rangle = 3/2(1+\alpha)\,\langle x, x\rangle$ to the second case

which concludes the proof. □

C Proof of Proposition 6

The proof of Proposition 6 is based on the reformulation Eq. (6) of q_R, which uses an auxiliary function R_x. In order to bound q_R, we first provide bounds for R_x in the following lemma:

Lemma 2 *(Bounds for R_x) Let $x, x' \in \mathbb{R}^d$ with $x' = \alpha x$ for some $0 \leq \alpha \leq 1$ and $y \in \mathbb{R}^d$ with $y \neq x, x'$. Then, it holds that*

$$\langle x, y\rangle \leq \langle x, x\rangle \quad \Rightarrow \quad R_x(y) \leq R_{x'}(y)\,,$$
$$\langle x, y\rangle \geq \frac{3}{2}(1+\alpha)\,\langle x, x\rangle \quad \Rightarrow \quad R_x(y) > R_{x'}(y)\,.$$

The first part of this lemma yields a tractable bound for the set $\{R_x(Y) \leq R_{x'}(Y)\}$, whereas the second part tells us that the bound cannot be improved substantially.

Proof For given x and y, we abbreviate $c := \langle x, x\rangle$ and $d := \langle y, y\rangle$. By definition of R_x, we have

$$R_x(y) \leq R_{\alpha x}(y) \Leftrightarrow \frac{\|y\|^2 - \|x\|^2}{2\,\|y - x\|} \leq \frac{\|y\|^2 - \|\alpha x\|^2}{2\,\|y - \alpha x\|}$$
$$\Leftrightarrow \frac{d - c}{\|y - x\|} \leq \frac{d - \alpha^2 c}{\|y - \alpha x\|}\,. \tag{11}$$

We distinguish between three different cases depending on the signs of both sides.
Case I: $\alpha^2 c < d < c$. Here, the left hand side in Eq. (11) is negative, whereas the right hand side is positive, and the inequality holds trivially.

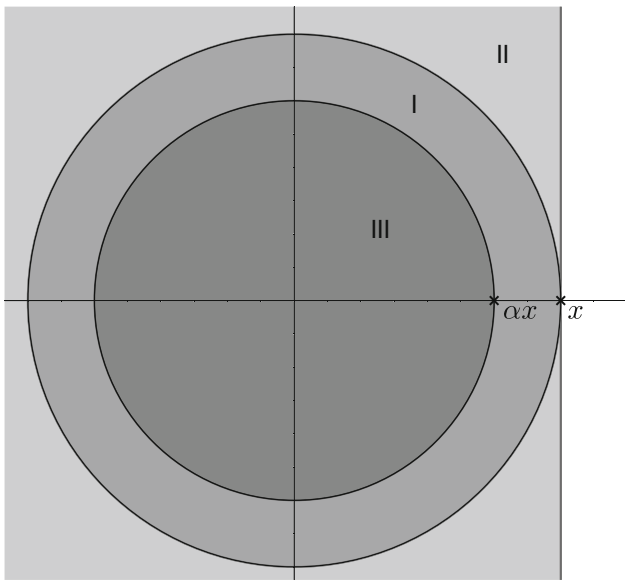


Fig. 9 Three different cases in proof of Lemma 2 for $x = (4, 0)^\top$ and $\alpha = 3/4$. Case I $\left\{y \in \mathbb{R}^2 \mid \alpha\,\|x\| < \|y\| < \|x\|\right\}$, case II $\left\{y \in \mathbb{R}^2 \mid \|x\| < \|y\|\,,\ \langle x, y\rangle \leq \langle x, x\rangle\right\}$, and case III $\left\{y \in \mathbb{R}^2 \mid \|y\| < \alpha\,\|x\|\right\}$

Case II: $c \leq d$. Next, we consider the case where both sides in Eq. (11) are nonnegative. Squaring both sides and rearranging the inequality yields

$$R_x(y) \leq R_{\alpha x}(y)$$
$$\Leftrightarrow\ f(\alpha, x, y) := \|y - \alpha x\|^2\,(d - c)^2 - \|y - x\|^2\left(d - \alpha^2 c\right)^2 \leq 0\,. \tag{12}$$

By expanding and regrouping all products we obtain

$$f(\alpha, x, y) = -3\left(1 - \alpha^2\right)cd^2 + (1 - \alpha^4)c^2 d + \alpha^2\left(1 - \alpha^2\right)c^3$$
$$+ 2(1 - \alpha)\underbrace{\left(d^2 + 2\alpha cd - \alpha\left(1 + \alpha + \alpha^2\right)c^2\right)}_{=:g(\alpha, c, d)}\langle x, y\rangle\,. \tag{13}$$

Next, we want to use the inequality $\langle x, y\rangle \leq c$. We have

$$g(\alpha, c, d) = d^2 + 2\alpha cd - \alpha\left(1 + \alpha + \alpha^2\right)c^2 \qquad (c \leq d)$$
$$\geq c^2 + 2\alpha c^2 - \alpha\left(1 + \alpha + \alpha^2\right)c^2$$
$$= c^2(1 - \alpha)(1 + \alpha)^2$$
$$g(\alpha, c, d) \geq 0\,,$$

and therefore $\langle x, y\rangle \leq c$ yields

$$f(\alpha, x, y) \leq \quad c \leq d$$
$$-3\left(1 - \alpha^2\right)cd^2 + (1 - \alpha^4)c^2 d + \alpha^2\left(1 - \alpha^2\right)c^3$$

$$
\begin{aligned}
&+ 2(1-\alpha)\left(d^2 + 2\alpha cd - \alpha\left(1+\alpha+\alpha^2\right)c^2\right)c \\
&= -c(1-\alpha)(d-c)\left((1+3\alpha)d - \alpha(2+\alpha+\alpha^2)c\right) \\
&\leq 0,
\end{aligned}
$$

which completes this case.

Case III: $d \leq \alpha^2 c$. For the remaining case, we consider the function $\beta \mapsto R_{\beta x}(y)$ and show that it is decreasing on $\left[\sqrt{d/c}, \infty\right)$. Since $d \leq \alpha^2 c$ implies $\sqrt{d/c} \leq \alpha$ and by assumption it is $\alpha \leq 1$, this yields the desired inequality $R_{1x}(y) \leq R_{\alpha x}(y)$. We have

$$
\begin{aligned}
\frac{\partial}{\partial\beta} R_{\beta x}(y) &= \frac{\partial}{\partial\beta} \frac{d - \beta^2 c}{2\,\|y - \beta x\|} \\
&= \frac{-2\beta c\,\|y-\beta x\|^2 - \left(d-\beta^2 c\right)\langle x, \beta x - y\rangle}{2\,\|y-\beta x\|^3}
\end{aligned}
$$

and therefore

$$
\begin{aligned}
&\frac{\partial}{\partial\beta} R_{\beta x}(y) \leq 0 \\
\Leftrightarrow\ &\tilde{f}(\beta, x, y) \coloneqq -2\beta c\,\|y-\beta x\|^2 - \left(d-\beta^2 c\right)\langle x, \beta x - y\rangle \\
&\qquad \leq 0\,. \qquad (14)
\end{aligned}
$$

Expanding the terms in $\tilde{f}$ yields

$$
\begin{aligned}
\tilde{f}(\beta, x, y) &= -\beta^3 c^2 + (3\beta^2 c + d)\langle x, y\rangle - 3\beta cd \\
&\qquad \text{(Cauchy–Schwarz)} \\
&\leq -\beta^3 c^2 + (3\beta^2 c + d)\sqrt{cd} - 3\beta cd \\
&= \sqrt{c}\left(\sqrt{d} - \beta\sqrt{c}\right)^3 \quad \beta \geq \sqrt{d/c} \\
\tilde{f}(\beta, x, y) &\leq 0.
\end{aligned}
$$

Therefore, by Eq. (14), the function $\beta \mapsto R_{\beta x}(y)$ is decreasing on $\left[\sqrt{d/c}, \infty\right)$, which completes the proof of this remaining case, and thus the first implication.

For the second implication $\langle x, y\rangle \geq 3/2(1+\alpha)\langle x, x\rangle \Rightarrow R_x(y) > R_{x'}(y)$, we go back to Eq. (13) and show $f(\alpha, x, y) \geq 0$. As before, we are in the case

$$
\begin{aligned}
d &\geq \frac{\langle x, y\rangle^2}{c} \quad \text{(Cauchy-Schwarz)} \\
&\geq \frac{9}{4}(1+\alpha)^2 c \quad \langle x, y\rangle \geq \frac{3}{2}(1+\alpha)c \\
d &\geq c\,,
\end{aligned}
$$

thus it is $g(\alpha, c, d) \geq 0$ and using the inequality $\langle x, y\rangle \geq 3/2\,(1+\alpha)\,c$ in Eq. (13) yields

$$
\begin{aligned}
&f(\alpha, x, y) \geq \\
&\quad -3\left(1-\alpha^2\right)cd^2 + (1-\alpha^4)c^2 d + \alpha^2\left(1-\alpha^2\right)c^3 \\
&\quad + 2(1-\alpha)\left(d^2 + 2\alpha cd - \alpha\left(1+\alpha+\alpha^2\right)c^2\right)\frac{3}{2}(1+\alpha)c \\
&= c^2(1-\alpha)(1+\alpha)\left((\alpha^2+6\alpha+1)d - \alpha(3\alpha^2+2\alpha+3)c\right) \\
&\overset{(*)}{\geq} \frac{1}{4}c^3(1-\alpha)(1+\alpha)(3+\alpha)^2(1+3\alpha)^2 \\
&\geq 0\,,
\end{aligned}
$$

where the inequality at $(*)$ holds, because $d \geq \frac{9}{4}(1+\alpha)^2 c$. By Eq. (12), this completes the proof. □

With these bounds, we are now prepared to proof Proposition 6.

Proposition 6 (q_R is approximately decreasing) *Let Y_1 be the first coordinate of Y, which is distributed as an around the origin rotationally invariant distribution P. Then for any $x, x' \in \mathbb{R}^d$ with $\|x'\| \leq \|x\|$, it holds that*

$$
q_R(x) \leq q_R(x') + \mathbb{P}(Y_1 > \|x\|)\,. \qquad (7)
$$

Proof Let $x \in \mathbb{R}^d$ and S, R_x, and R_0 as in Lemma 1, where R_x and R_0 are coupled *via* Y. Define $E_x \coloneqq \{R_x \leq R_{x'}\}$. By Lemma 1, it holds that

$$
\begin{aligned}
q_R(x) = \mathbb{P}(S \leq R_x) &= \mathbb{P}(S \leq R_x, E_x) + \mathbb{P}(S \leq R_x, E_x^{\mathsf{c}}) \\
&\leq \mathbb{P}(S \leq R_{x'}) + \mathbb{P}(E_x^{\mathsf{c}}) \\
&= q_R(x') + \mathbb{P}(E_x^{\mathsf{c}})\,.
\end{aligned}
$$

By the first inequality of Lemma 2, we can upper bound $\mathbb{P}(E_x^{\mathsf{c}}) \leq \mathbb{P}(\langle x, Y\rangle > \langle x, x\rangle)$. Lastly, Proposition 3 allows replacing x by $\|x\|\,e_1$ to obtain

$$
\begin{aligned}
q_R(x) &= q_R(\|x\|\,e_1) \\
&\leq q_R(x') + \mathbb{P}(E_{\|x\|e_1}^{\mathsf{c}}) \\
&\leq q_R(x') + \mathbb{P}(\langle \|x\|\,e_1, Y\rangle > \langle \|x\|\,e_1, \|x\|\,e_1\rangle) \\
&= q_R(x') + \mathbb{P}(Y_1 > \|x\|)\,,
\end{aligned}
$$

which completes the proof. □

References

1. Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., Belongie, S.: Generalized non-metric multidimensional scaling. In: AISTATS, pp. 11–18 (2007)
2. Amid, E., Ukkonen, A.: Multiview triplet embedding: learning attributes in multiple maps. In: ICML, pp. 1472–1480 (2015)

3. Arias-Castro, E.: Some theory for ordinal embedding. Bernoulli Soc. Math. Stat. Probab. **23**(3), 1663–1693 (2017)
4. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. Mach. Learn. **56**(1), 89–113 (2004)
5. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp. 17–36 (2012)
6. Deng, J., Dong, W., Socher, R., Li, L., Kai L., Li F.-F.: Imagenet: a large-scale hierarchical image database. In: CVPR, pp. 248–255 (2009)
7. Emamjomeh-Zadeh, E., Kempe, D.: Adaptive hierarchical clustering using ordinal queries. In: SODA, pp. 415–429 (2018)
8. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. SODA **17**(1), 134–160 (2003)
9. Ghoshdastidar, D., Perrot, M., von Luxburg, U.: Foundations of comparison-based hierarchical clustering. In: NIPS, pp. 7456–7466 (2019)
10. Haghiri, S., Ghoshdastidar, D., von Luxburg, U.: Comparison-Based Nearest Neighbor Search. In: AISTATS, vol. 54, pp. 851–859 (2017)
11. Haghiri, S., Garreau, D., von Luxburg, U.: Comparison-based random forests. In: ICML, vol. 80, pp. 1871–1880 (2018)
12. Heikinheimo, H., Ukkonen, A.: The crowd-median algorithm. In: HCOMP, pp. 69–77 (2013)
13. Jain, L., Jamieson, K.G., Nowak, R.: Finite sample prediction and recovery bounds for ordinal embedding. In: NIPS, pp. 2711–2719 (2016)
14. Jamieson, K.G., Nowak, R.D.: Low-dimensional embedding using adaptively selected ordinal data. In: 49th Annual Allerton Conference on Communication, Control, and Computing, pp. 1077–1084 (2011)
15. Kleindessner, M., Luxburg, U.: Uniqueness of ordinal embedding. In: COLT, pp. 40–67 (2014)
16. Kleindessner, M., Luxburg, U.: Dimensionality estimation without distances. In: AISTATS, pp. 471–479 (2015)
17. Kleindessner, M., von Luxburg, U.: Kernel functions based on triplet comparisons. In: NIPS, pp. 6807–6817 (2017a)
18. Kleindessner, M., von Luxburg, U.: Lens depth function and k-relative neighborhood graph: versatile tools for ordinal data analysis. In: JMLR, pp. 1–52 (2017b)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
20. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
21. Liu, Z., Modarres, R.: Lens data depth and median. J. Nonparametric Stat. **23**(4), 1063–1074 (2011)
22. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV **42**(3), 145–175 (2001)
23. Perrot, M., von Luxburg, U.: Boosting for comparison-based learning. In: IJCAI-19, pp. 1844–1850 (2019). https://doi.org/10.24963/ijcai.2019/255
24. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. In: NIPS, pp. 41–48 (2004)
25. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
26. Tamuz, O., Liu, C., Belongie, S., Shamir, O., Kalai, A.T.: Adaptively learning the crowd kernel. In: ICML, pp. 673–680 (2011)
27. Terada, Y., Luxburg, U.: Local ordinal embedding. In: ICML, vol. 32, pp. 847–855 (2014)
28. Ukkonen, A.: Crowdsourced correlation clustering with relative distance comparisons. In: ICDM, pp. 1117–1122 (2017). https://doi.org/10.1109/ICDM.2017.148
29. Ukkonen, A., Derakhshan, B., Heikinheimo, H.: Crowdsourced nonparametric density estimation using relative distances. In: HCOMP (2015)
30. Van Der Maaten, L., Weinberger, K.: Stochastic triplet embedding. In: MLSP, pp. 1–6 (2012)
31. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
32. Vershynin, R.: High-Dimensional Probability: An Introduction with Applications in Data Science, vol. 47. Cambridge University Press, Cambridge (2018)
33. Vikram, S., Dasgupta, S.: Interactive bayesian hierarchical clustering. In: ICML, pp. 2081–2090 (2016)
34. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NIPS, pp. 3320–3328 (2014)
35. Zuo, Y., Sering, R.: General notions of statistical depth function. Ann. Stat. (2000). https://doi.org/10.1214/aos/1016218226