Quantifying Centrality for Complex Data

Hang Zhou* UNC Chapel Hill hzh@unc.edu

Yidong Zhou* University of Minnesota ydzhou@umn.edu

Hans-Georg Müller University of California, Davis hgmueller@ucdavis.edu

Abstract

Analyzing complex data across a wide range of scientific fields often requires identifying central or typical elements, distinguishing them from atypical ones, and constructing central or "normal" regions at prespecified levels of centrality. We introduce centrality scores (C-scores) based on distance profiles for random objects taking values in general metric spaces. The proposed C-scores are defined through weighted optimal transport of distance profiles and provide a unified framework for constructing central or "normal" regions for complex data. We show that, for Euclidean data the proposed C-scores asymptotically rank points in the same order as the underlying probability density function. The practical merits of the proposed method are illustrated using gene expression data, distributions of recorded temperatures, handwritten digit recognition, and taxi trip records. These examples demonstrate the effectiveness of C-scores in identifying central and peripheral elements and in yielding meaningful insights for non-Euclidean and high-dimensional data.

1 Introduction

Complex data are increasingly recorded and often defy simple characterizations, including gene expression patterns [Li et al., 2023], social networks [Aral and Walker, 2012], brain connectivity [Zhao et al., 2023] or computer vision [Jordan and Mitchell, 2015]. A fundamental but challenging task for such data is to identify "typical" or "central" elements and to distinguish these from peripheral or outlying data. Quantifying centrality has been a focus of statistical research since the 1950s, with tolerance regions emerging as a popular notion [Fraser and Guttman, 1956, Krishnamoorthy and Mathew, 2009]. However, these traditional approaches are inadequate for complex data that reside in general metric spaces $\mathcal{M}$ with a metric d [Müller, 2016], due to the absence of standard Euclidean geometry, linearity and inner products. Identifying inlying and outlying elements among random objects requires inferring probability measures on complex data spaces.

To address this task, a key tool is the use of distance profiles [Dubey et al., 2024], which represent the distribution of distances between any given element and random objects in the space $\mathcal{M}$ that are distributed according to an underlying probability measure P. Assigning to each element a distribution of its distances to the points in the data, empirical distance profiles provide a flexible framework that is applicable to diverse data types and metrics. Transport ranks and notions of centrality as introduced in Dubey et al. [2024] are well suited for distributions in which the random objects are distributed around the most central point, but less so for multimodal data, for which they lead to larger and coarser centrality regions. This motivates *centrality scores* (*C-scores*), which are based on weighted transports of distance profiles. The resulting *centrality regions* consist of data elements with C-scores above a prespecified threshold that defines the centrality level. The proposed C-scores and the associated centrality regions are effective for both unimodal and multimodal data across various metric spaces.

*The first two authors contributed equally to this work.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Related constructions for metric-valued data include conformal scores based on the transport cost between distance profiles [Zhou and Müller, 2025], metric halfspace depth [Dai et al., 2023], and metric spatial depth [Virta, 2026]. C-scores differ by using a tunable lower-tail weight: they recover the global transport rank at $\tau = 0$ and increasingly reflect local probability concentration as τ grows. Appendix D gives numerical comparisons and discusses settings in which each construction is informative.

We demonstrate the effectiveness and utility of C-scores through simulation studies, including Gaussian mixture distributions in Euclidean space, data on spheres and distributional data in the Wasserstein space, and also for complex real-world data from gene expression studies, temperature recordings, taxi trip records and the MNIST image database. The remainder of the paper is organized as follows. Section 2 introduces random objects and weighted transports of distance profiles. Section 3 presents the proposed C-scores and their theoretical properties. Sections 4 and 5 contain simulation studies and real data applications. Proofs and additional results are deferred to the Appendix.

2 Random objects and distance profiles

Samples of random objects $X_1, \ldots, X_n$ that take values in a general metric space $(\mathcal{M}, d)$ equipped with a probability measure, i.e., a metric measure space [Heinonen, 2001], encompass a multitude of data types. These include multivariate and high-dimensional random vectors in $\mathbb{R}^p$, possibly with large p; functional data valued in infinite-dimensional Hilbert spaces [Yao et al., 2005, Hsing and Eubank, 2015]; networks [Kolaczyk and Csárdi, 2020, Zhou and Müller, 2022]; probability distributions [Petersen et al., 2022, Zhou and Müller, 2024, Zhu and Müller, 2023]; data on Riemannian manifolds [Pennec, 2006, Dai and Müller, 2018, Lin and Yao, 2019]; as well as images and phylogenetic or other tree-structured data [Billera et al., 2001].

We next introduce four representative examples of such complex data that are commonly encountered in practice. Several of these example spaces will be revisited later to illustrate the practical performance of the proposed approach in both simulation studies and real data analyses.

Example 1 (Compositional data)**.** *Compositional data, represented by vectors of nonnegative proportions summing to one, reside in the simplex* $\Delta^{p-1} = \{\mathbf{x} \in \mathbb{R}^p : x_i \geq 0 \text{ for } i = 1, \ldots, p, \sum_{i=1}^p x_i = 1\}$*. A square root transformation* $\mathbf{y} = (x_1^{1/2}, \ldots, x_p^{1/2})^\top$ *[Scealy and Welsh, 2011] maps the compositional data to a subset corresponding to the positive orthant of the unit sphere* $\mathbb{S}^{p-1} = \{\mathbf{y} \in \mathbb{R}^p : y_i \geq 0 \text{ for } i = 1, \ldots, p, \sum_{i=1}^p y_i^2 = 1\}$.

Example 2 (Networks)**.** *Large collections of networks are increasingly available, where each of the networks in the sample may be viewed as a fundamental unit of observation [Ginestet et al., 2017, Kolaczyk et al., 2020, Severn et al., 2022]. Consider simple undirected weighted networks with a fixed number of nodes* m *and bounded edge weights. There is a one-to-one correspondence between such networks and their graph Laplacians. The space of graph Laplacians equipped with the Frobenius or power metric [Dryden et al., 2009] characterizes the space of networks [Zhou and Müller, 2022], which emerges as a bounded and convex subset of* $\mathbb{R}^{m^2}$.

Example 3 (Probability distributions)**.** *Samples of probability distributions have become a common data type [Chen et al., 2023, Zhou et al., 2026]. Sample elements are typically univariate distributions that can be considered as random elements of the 2-Wasserstein space* $\mathcal{W}_2$ *[Panaretos and Zemel, 2020], which is the collection of probability measures on closed intervals* $\mathcal{S}$ *defined as*

$$\mathcal{W}_2 = \left\{ \mu \in \mathcal{P}(\mathcal{S}) : \int_{\mathcal{S}} |x|^2 \mathrm{d}\mu(x) < \infty \right\},$$

where $\mathcal{P}(\mathcal{S})$ *is the set of all probability measures on* $\mathcal{S}$*. The* 2*-Wasserstein distance between two measures* $\mu, \nu \in \mathcal{W}_2$ *is*

$$d^2_{\mathcal{W}}(\mu, \nu) := \inf \left\{ \int |x_1 - x_2|^2 \mathrm{d}\Gamma(x_1, x_2) : \Gamma \in \Gamma(\mu, \nu) \right\},$$

where $\Gamma(\mu, \nu)$ *is the set of joint probability measures on* $\mathcal{S}^2$ *with* μ *and* ν *as marginal measures.*

The Wasserstein space is connected with the optimal transport problem [Ambrosio et al., 2008], which aims to find a non-decreasing map T *on* $\mathcal{S} \mapsto \mathcal{S}$ *that minimizes the transport cost. The optimal transport from* μ *to* ν *has the analytic form* $T_\mu^\nu = F_\nu^{-1} \circ F_\mu$*, where* $F_\mu(\cdot)$ *and* $F_\nu^{-1}(u) = \inf\{x \in \mathbb{R} :$

$F_\nu(x) \geq u\}$ are the cumulative distribution and quantile function of μ and ν, respectively [Gangbo and McCann, 1996].

Example 4 (Functional data)**.** *Functional data are ubiquitous and consist of samples of trajectories that are realizations of a square-integrable and smooth stochastic process [Wang et al., 2016, Horvath and Kokoszka, 2012, Hsing and Eubank, 2015]; sample elements are usually viewed as random elements of $\mathcal{L}^2([0,1]) := \{f : \int_{[0,1]} f^2(x)\mathrm{d}x < \infty\}$.*

The absence of linearity in general metric spaces poses significant challenges. In spaces with Riemannian geometry, such as spheres, transforming data to linear spaces through log maps onto tangent spaces is a standard technique that to some extent can address the challenges of curvature and nonlinearity by local linearization; however, the inverse exp maps are usually only defined on a convex subset of the linear space and this limits the utility of linearization methods [Bigot et al., 2017, Chen et al., 2023, Pegoraro and Beraha, 2022]. More generally, methodologies available for Euclidean and functional data can be extended to random objects by mapping metric spaces to a linear space either by a generic kernel map [Schoenberg, 1938] or by specifically tailored transformation maps such as log quantile density and log hazard transformations for distributional data [Petersen and Müller, 2016]. Linearizing tangent space projections in the 2-Wasserstein space of univariate distributions have been used for principal component analysis [Bigot et al., 2017] and regression analysis [Chen et al., 2023] and also for finite-dimensional Riemannian manifolds [Lin and Yao, 2019]. However, such maps are typically not surjective, and an inverse defined on the whole linear space exists only in very few instances. This non-invertibility leads to difficulties as the elements resulting from operations in the linearized space, including prediction and principal component analysis, may not reside in the image space of the transformation map and so have no viable representation in the underlying metric space. This fundamental problem of non-existence of inverse maps is also shared by kernel methods, which additionally require a tuning parameter.

Distance profiles [Dubey et al., 2024] correspond to cumulative distribution functions that characterize the distribution of distances from a fixed point to the random object X in the metric space $\mathcal{M}$. For any $\omega \in \mathcal{M}$, the distance profile at ω is defined as

$$F_\omega(t) = \mathbb{P}(d(\omega, X) \leq t), \text{ for all } t \in \mathbb{R}^+. \tag{1}$$

Figure 1 illustrates distance profiles and their corresponding density functions for three selected points in $\mathcal{M} = \mathbb{R}^2$. For ω close to the center of the distribution of X the distance profile $F_\omega(t)$ has a steeper increase near $t = 0$. Distance profiles are determined by both the metric of the object space and the underlying probability measure.

Note that the distance profile $F_X(t) = \mathbb{P}(d(X, X') \leq t)$ for the random object X can be regarded as a random element in the Wasserstein space, where X' is an independent copy of X. For each $\omega \in \mathcal{M}$ with continuous F_ω, the optimal transport map moving from F_ω to F_X is given by $F_X^{-1} \circ F_\omega$. The last row of Figure 1 illustrates the transport maps between each pair of selected data points. The integral

$$H_X(\omega) = \int \left\{F_X^{-1} \circ F_\omega(u) - u\right\} \mathrm{d}F_\omega(u) = \int_0^1 \left\{F_X^{-1}(u) - F_\omega^{-1}(u)\right\} \mathrm{d}u \tag{2}$$

quantifies the overall mass transfer from F_ω to F_X. If ω is centrally located with respect to the distribution of X, we expect the mass transfer from F_ω to F_X to be predominantly from left to right, and $H_X(\omega)$ will be a positive value with high probability that quantifies the difference in outlyingness between ω and the random object X. As was observed previously [Dubey et al., 2024], one can use the expectation of $H_X(\omega)$ to quantify the centrality of object data.

3 Centrality scores and centrality regions for random objects

A limitation of the aforementioned centrality measure is that it is efficient only for distributions centered at a single point in the metric space, analogous to the case of a unimodal distribution with monotone decreasing density away from the mode. This is because $H_X(\omega)$ accounts for the directionality of optimal transports and reflects the difference between rightward and leftward moving mass, rather than the total transport cost, which the latter has been recently utilized for conformal prediction [Zhou and Müller, 2025]. When the distribution functions F_ω and F_X intersect, there is a sign difference in the transport before and after the intersection point. In this case $H_X(\omega)$ reflects

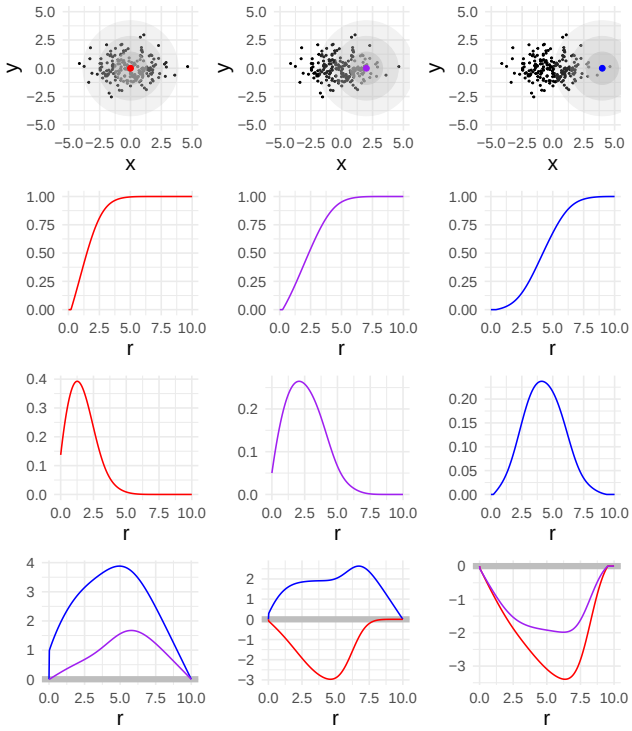


Figure 1: Distance profiles for a Gaussian distribution in $\mathcal{M} = \mathbb{R}^2$. First row: 200 random samples from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (2, 0; 0, 0.5)$, with three randomly selected points: One point close to the center (red, left column), one far from the center (blue, right column), and one between the center and the distant point (purple, middle column). Gray circles with varying radii indicate balls around each selected point. Second and third rows: Estimated distance profiles and their corresponding probability density functions for the red (left), purple (middle), and blue (right) points, respectively. Last row: Optimal transport maps for the selected data points, with red, purple, and blue points as starting locations. Positive (negative) values indicate transport to the right (left), with the magnitude reflecting the distance a point is transported from the uniform distribution. In each panel, curve colors correspond to the transport endpoints. For example, in the bottom-left panel, the blue and purple curves represent transports from the red point to the blue and purple points, respectively.

the difference between rightward and leftward moving mass in the transport. Such intersecting distribution functions occur, for example, when the underlying distribution of X does not have a single mode around which it is centered.

As discussed in Section 2, $F_\omega(t)$ represents the probability that $d(\omega, X) \leq t$. When t is small, $F_\omega(t)$ should exhibit a sharp increase if ω is centrally located with respect to the distribution of X, and we expect it to have a relatively flat pattern if ω is far from the center. In contrast, when t is large, the difference between $F_\omega(t)$ for different ω becomes less pronounced because $F_\omega(t)$ approaches 1 for all ω for sufficiently large t. This puts the focus on the behavior of $F_\omega(t)$ around $t = 0^+$ and motivates adding a penalty with respect to u in the integral $H_X(\omega)$. Incorporating such a penalty leads to the *centrality score*, defined as

$$R_\tau(\omega) = g\left(\mathbb{E}\left[\int_0^1 \left\{F_X^{-1}(u) - F_\omega^{-1}(u)\right\} \, e^{-\tau u} \mathrm{d}u\right]\right), \tag{3}$$

where $\tau \geq 0$ is the tuning parameter for the penalty and g is a strictly increasing function mapping the argument to $(0, 1)$, where the default is the expit function $\text{expit}(x) = e^x/(1 + e^x)$. This ensures that the proposed scores are scaled to lie in $(0, 1)$; see also Dubey et al. [2024]. The choice of exponential weights is not mandatory for the construction or finite-sample coverage: Theorem 1 applies to any integrable weight that is nonzero almost everywhere. We use $w_\tau(u) = e^{-\tau u}$ because it defines a simple one-parameter family of monotone declining weights. At $\tau = 0$, $w_0 \equiv 1$ and the C-score

Algorithm 1 Centrality region for random objects

Input: Observations $\{X_i\}_{i=1}^n$, level α, tuning parameter τ
Randomly split $\{X_i\}_{i=1}^n$ into a training set $\mathcal{D}_{\text{train}}$ and a calibration set $\mathcal{D}_{\text{cal}}$
Estimate $\hat{F}_\omega(t)$ using the training set $\mathcal{D}_{\text{train}}$ via (4)
for each $X_j \in \mathcal{D}_{\text{cal}}$ **do**
  Compute the C-score $\hat{R}_\tau(X_j)$
end for
Let $k = \lfloor (|\mathcal{D}_{\text{cal}}|+1)\alpha \rfloor$. Set $\hat{Q}_\alpha = -\infty$ if $k = 0$; otherwise let $\hat{Q}_\alpha$ be the kth smallest calibration C-score.
Output: Centrality region -0.05in

$$\hat{C}_\alpha = \left\{ X \in \mathcal{M} : \hat{R}_\tau(X) \geq \hat{Q}_\alpha \right\}$$

reduces to the transport rank of Dubey et al. [2024]; as τ increases, the weight concentrates on small values of u, corresponding to small distances, thus emphasizing local small-ball structure.

We then construct centrality regions using C-scores. The idea is to split the observed dataset $\mathcal{D} = \{X_i\}_{i=1}^n$ into a training set $\mathcal{D}_{\text{train}}$ and a calibration set $\mathcal{D}_{\text{cal}}$. The empirical estimates $\hat{F}_\omega(t)$ are obtained based on the data in $\mathcal{D}_{\text{train}}$:

$$\hat{F}_\omega(t) = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{X_i \in \mathcal{D}_{\text{train}}} \mathbf{1}_{\{d(\omega, X_i) \leq t\}}, \quad t \geq 0, \tag{4}$$

where $\mathbf{1}_{\{A\}}$ is the indicator function for an event A. We estimate the C-score for each $X_j \in \mathcal{D}_{\text{cal}}$ by

$$\hat{R}_\tau(X_j) = g\left(\frac{1}{|\mathcal{D}_{\text{train}}|} \times \sum_{X_i \in \mathcal{D}_{\text{train}}} \left[\int_0^1 \left\{ \hat{F}_{X_i}^{-1}(u) - \hat{F}_{X_j}^{-1}(u) \right\} e^{-\tau u} \mathrm{d}u \right] \right). \tag{5}$$

For a new observation X_{n+1} lies in the $(1-\alpha)$ centrality region $\hat{C}_\alpha$ if its C-score $\hat{R}_\tau(X_{n+1}) \geq \hat{Q}_\alpha$, where $\hat{Q}_\alpha$ denotes the $\lfloor (|\mathcal{D}_{\text{cal}}|+1)\alpha \rfloor / |\mathcal{D}_{\text{cal}}|$-th empirical quantile of $\{\hat{R}_\tau(X_j) : X_j \in \mathcal{D}_{\text{cal}}\}$ and $(1-\alpha)$ is the target coverage probability with a default choice $\alpha = 0.05$. The outline of the proposed method is summarized in Algorithm 1. Conditional on the training set, the $n_{\text{cal}} = |\mathcal{D}_{\text{cal}}|$ calibration scores and the test score are exchangeable, so the finite-sample guarantee holds for any size of the calibration set. The attainable miscoverage levels have granularity $1/(n_{\text{cal}}+1)$. For $\alpha = 0.05$, at least 19 calibration observations are required for a nontrivial finite threshold; if $n_{\text{cal}} < 19$, the region is conservatively the entire space. A larger calibration set yields finer and less variable quantile calibration, whereas a larger training set improves estimation of the distance profiles and C-scores. We therefore recommend a roughly balanced split, with n_{cal} comfortably above 19 whenever sample size permits this.

The following theorem guarantees the coverage level for the centrality region obtained by Algorithm 1, which is a standard result in conformal prediction.

Theorem 1. *Let $X_1, \ldots, X_{n+1}$ be i.i.d. Fix g and w, or select them using only $\mathcal{D}_{\text{train}}$. Assume $w(u)$ is a measurable weight function such that $\int_0^1 |w(u)|\,\mathrm{d}u < \infty$ and $\lambda(\{u : w(u) = 0\}) = 0$, where $\lambda(\cdot)$ is the Lebesgue measure. For a new data point X_{n+1} and the centrality region $\hat{C}_\alpha$ obtained by Algorithm 1 with score function*

$$\hat{R}_{g,w}(X_j) = g\left(\frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{X_i \in \mathcal{D}_{\text{train}}} \left[\int_0^1 \left\{ \hat{F}_{X_i}^{-1}(u) - \hat{F}_{X_j}^{-1}(u) \right\} w(u) \mathrm{d}u \right] \right),$$

one has $\mathbb{P}(X_{n+1} \in \hat{C}_\alpha) \geq 1 - \alpha$.

Figure 2 illustrates the centrality regions constructed by Algorithm 1 with $w(u) = e^{-\tau u}$ that result when choosing tuning parameters $\tau = 0$ and $\tau = 4$. When $\tau = 0$, the proposed C-scores coincide with the original transport ranks [Dubey et al., 2024]. The corresponding centrality region is a simply connected subset of $\mathbb{R}^2$ centered at the global center of the mixture distribution, which is clearly too

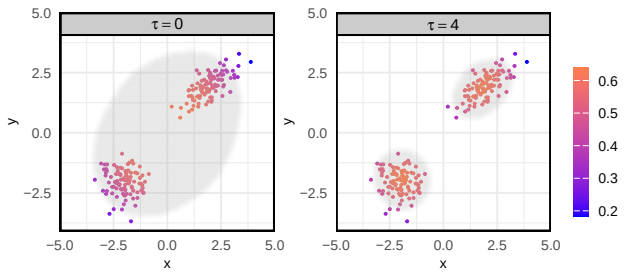


Figure 2: Estimated centrality region using the proposed centrality scores with different tuning parameters τ. The training data are from a 2-dimensional Gaussian mixture, $0.5\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)+0.5\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\mu}_1 = (2,2)^\top$, $\boldsymbol{\mu}_2 = (-2,-2)^\top$, $\boldsymbol{\Sigma}_1 = (0.5, 0.3; 0.3, 0.3)$ and $\boldsymbol{\Sigma}_2 = (0.5, 0; 0, 0.3)$. The transition from low to high C-scores is indicated by a color gradient ranging from blue to coral for each data point. The respective 95% centrality regions are shaded in gray.

large. Choosing a larger τ yields centrality regions that are expected to be smaller as they adapt more flexibly to the underlying structure, and indeed when $\tau = 4$, the resulting centrality regions evolve into a union of two disjoint simply connected subsets that are centered at each mixture component; this centrality region is much smaller and clearly preferable to the one resulting from $\tau = 0$. The increase in adaptivity arises as the weight function $\exp(-\tau u)$ places more emphasis on values of u close to 0^+, and the centrality regions become more concentrated around local centers. For the special case where $X \in \mathbb{R}^d$ has a density f, the following theorem characterizes the limiting behavior of the proposed C-score as $\tau \to \infty$, showing that its ordering asymptotically agrees with that of the probability density function. This theorem not only establishes a key theoretical property of the proposed C-score, but also provides intrinsic insight into why the C-score serves as a meaningful notion of centrality in general metric spaces, and offers practical guidance for choosing τ in implementation.

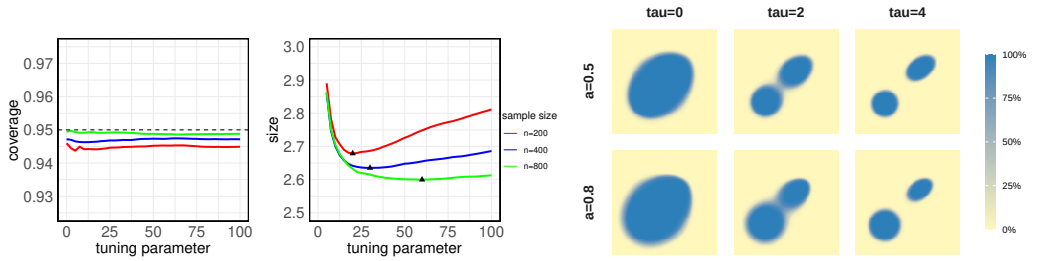
Theorem 2. *Let $X_1, \ldots, X_m \overset{i.i.d.}{\sim} f$ on $\mathbb{R}^d$ with $\mathbb{E}\|X\| < \infty$, where f is a continuous density function. Assume that $\tau = \tau_m$ satisfies $\tau_m \to \infty$ and $\tau_m/m \to 0$ as $m \to \infty$ and denote by $\hat{R}_{\tau_m}(\omega)$ the empirical C-score computed from the training sample with weight $e^{-\tau_m u}$. Then for any fixed pair of elements $\omega_1, \omega_2 \in \mathbb{R}^d$, if $f(\omega_1) > f(\omega_2) > 0$ then $\mathbb{P}\Big(\hat{R}_{\tau_m}(\omega_1) > \hat{R}_{\tau_m}(\omega_2)\Big) \to 1$.*

The proof uses Euclidean structure through the local small-ball expansion $\mathbb{P}\{d(X,\omega) \le r\} = v_d f(\omega) r^d\{1 + o(1)\}$, rather than global linear structure. The same argument therefore extends to a smooth d-dimensional Riemannian manifold without boundary, equipped with geodesic distance and a density f relative to Riemannian volume, assuming $\mathbb{E}d(X,\omega_0) < \infty$ for some fixed ω_0. Since $\mathrm{vol}\{B(\omega, r)\} = v_d r^d\{1 + O(r^2)\}$, continuity and positivity of f at ω yield the same small-ball expansion and limiting density ordering.

Since the choice of the tuning parameter τ affects the size and shape of the resulting centrality regions, selecting τ in practice is a nontrivial problem. However, in general metric spaces, the volume of the centrality region is difficult or even impossible to compute, making it challenging to select τ by directly minimizing the region size. In Appendix B, we propose a simple empirical procedure for choosing τ that is easy to implement and performs well across a wide range of applications. The procedure selects τ by minimizing the proportion of pairwise geodesic midpoints included in the centrality region, coupled with a penalty against excessive localization.

4 Experiments

We illustrate the proposed method using synthetic data across various metric spaces: the Euclidean space $\mathbb{R}^2$, the sphere $\mathbb{S}^2$, and the Wasserstein space $\mathcal{W}_2$. Evaluation criteria are (a) the coverage level and (b) the size of centrality regions. Each simulation is based on 200 Monte Carlo runs and for each run the coverage level is evaluated for a test set comprising 2000 samples, drawn from the same distribution as the training and calibration sets. The size of the centrality region is estimated through a grid search approach, where we generate a mesh grid for each metric space, apply the proposed method, and approximate the size of the centrality region by calculating the proportion of grid points that fall within it. Unless otherwise specified, the target coverage level is set to 95%.



(a) (b)

Figure 3: Simulation results for centrality regions under the Gaussian mixture model (6). (a) Average coverage levels (left panel) and sizes of centrality regions (right panel) over 200 Monte Carlo runs for varying tuning parameters. The mixing coefficient is $a = 0.5$ and the target centrality level is 95%. Red, blue, and green lines correspond to sample sizes $n = 200$, $n = 400$, and $n = 800$, respectively, and black triangles indicate the optimal τ that minimizes the size of the prediction set. (b) Illustration of 95% centrality regions over 200 Monte Carlo runs for varying tuning parameters τ, with mixing coefficients $a = 0.5$ and $a = 0.8$. The sample size is $n = 200$, and heat maps show the average proportion of each mesh grid residing in the centrality region, ranging from high (dark) to low (light).

A. Data from a Gaussian mixture in $\mathbb{R}^2$ with Euclidean distance. The data are generated from a mixture of Gaussian distributions in $\mathbb{R}^2$:

$$X_i \sim (1-a)\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + a\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \tag{6}$$

with mixing coefficient $a \in (0,1)$. Specifically, each X_i is drawn from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ with probability $1-a$, and from $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with probability a with parameters

$$\boldsymbol{\mu}_1 = (2,2)^\top, \ \boldsymbol{\mu}_2 = (-2,-2)^\top,$$

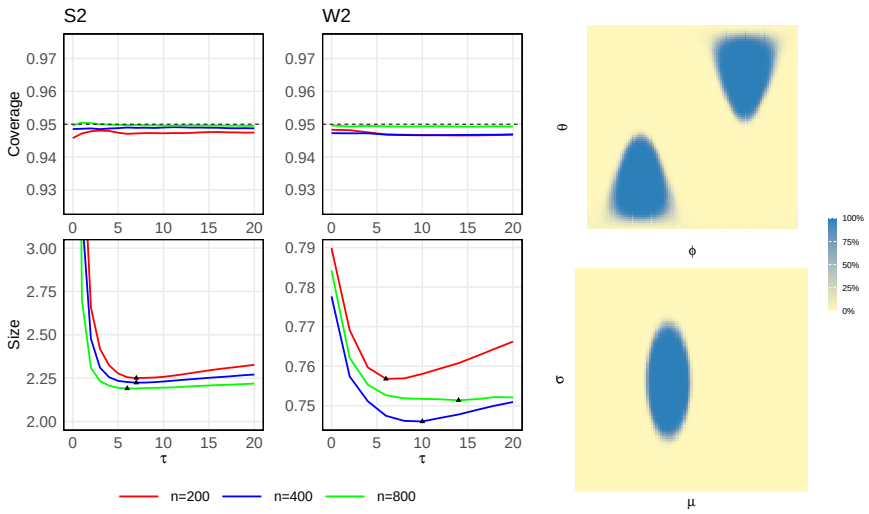
$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.3 \end{pmatrix}, \ \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}$$

and sample sizes $n = 200, 400, 800$. Figure 3(a) displays average coverage levels and sizes of the centrality regions for various tuning parameter values τ.

The left panel of Figure 3(a) shows that the coverage levels remain close to the target level across different τ values. The right panel indicates that the size of the centrality regions initially decreases as τ increases, before eventually increasing again. As expected, when sample size increases, the optimal size of the centrality regions decreases. Figure 3(b) illustrates the estimated average centrality regions over 200 runs for $\tau = 0$, $\tau = 2$, and $\tau = 4$ with mixing coefficient $a = 0.5$ and $a = 0.8$. When $\tau = 0$, the centrality region is centered around a single element, which is suboptimal for a bimodal distribution and leads to an oversized region. As τ increases, the centrality regions gradually evolve into two unconnected sets centered at $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and shrink in size, highlighting the effectiveness of the proposed method.

B. Data on the sphere $\mathbb{S}^2$ with geodesic distance. The unit sphere $\mathbb{S}^2$ is a 2-dimensional Riemannian manifold endowed with the geodesic distance $d(p,q) = \arccos(p^\top q)$. The tangent space at a point p is defined as $T_p := \{y \in \mathbb{R}^3 \mid y^\top p = 0\}$. For all $p \in \mathbb{S}^2$ and $v \in T_p$, the Riemannian exponential map, which projects v onto $\mathbb{S}^2$, is given by $\exp_p(v) = \cos(\|v\|)p + \sin(\|v\|)\|v\|^{-1}v$. The observations are generated by $y_i = a_i \exp_{\boldsymbol{\mu}_1}(\mathbf{V}_i) + (1-a_i)\exp_{\boldsymbol{\mu}_2}(\mathbf{V}_i)$, where $\boldsymbol{\mu}_1 = (0, \sqrt{2}/2, \sqrt{2}/2)^\top$, $\boldsymbol{\mu}_2 = (0, -\sqrt{2}/2, -\sqrt{2}/2)^\top$, and $\mathbf{V}_i = (\epsilon_{i1}, \epsilon_{i2}, -\epsilon_{i2})^\top$. Here, a_i are Bernoulli random variables with success probability 0.5, $\epsilon_{i1} \sim \mathcal{N}(0, 0.2^2)$, and $\epsilon_{i2} \sim \mathcal{N}(0, 0.2^2)$.

We consider three different sample sizes, $n = 200$, $n = 400$ and $n = 800$. Results are visualized in Figure 4. The top-left panel of Figure 4(a) shows the average coverage levels for various tuning parameters and indicates that the coverage levels remain relatively stable as the tuning parameter changes; for larger sample sizes, the coverage level is closer to the target. The bottom-left panel indicates that the changes in the sizes of the centrality regions across different τ values follow a similar pattern as in the $\mathbb{R}^2$ case. The left panel of Figure 4(b) illustrates the average centrality regions over 200 simulations, demonstrating that the proposed method accurately reflects the true distribution.



(a) (b)

Figure 4: Simulation results for centrality regions in non-Euclidean spaces. (a) Average coverage levels (first row) and sizes of the centrality regions (second row) over 200 Monte Carlo runs for varying tuning parameters τ. Results for data on the sphere $\mathbb{S}^2$ are shown in the left column and for distributional data in the Wasserstein space in the right column. Red, blue, and green lines correspond to sample sizes $n = 200, 400$ and 800, respectively, and black triangles indicate the optimal value of τ that minimizes the size of the centrality region. (b) Illustrations of the obtained 95% centrality regions over 200 Monte Carlo runs, with data on $\mathbb{S}^2$ (upper panel) and in the Wasserstein space of univariate distributions (lower panel). The sample size is $n = 200$ and the tuning parameter is set to 10. Heat maps show the average proportion of each mesh grid belonging to the centrality region, ranging from high (blue) to low (orange); in the upper panel, axes represent ϕ and θ in spherical coordinates, while in the lower panel, axes represent the mean and standard deviation of the normal distribution.

C. Distributional data with the 2-Wasserstein distance. For the Wasserstein space, the observations are generated as density functions of normal distributions $\mathcal{N}(\mu_i, \sigma_i^2)$ with random mean μ_i and variance σ_i^2. Here, $\mu_i \sim \mathcal{N}(2, 0.2^2)$ and $\sigma_i \sim |\mathcal{N}(0.5, 0.1^2)|$.

We consider again three sample sizes, $n = 200, 400, 800$. The average coverage levels and changes in the sizes of centrality regions for various tuning parameters are shown in the right column of Figure 4(a), and the average centrality regions over 200 simulations are presented in Figure 4(b). The results are consistent with the $\mathbb{R}^2$ and $\mathbb{S}^2$ cases, demonstrating the efficiency of the proposed method.

5 Data illustrations

5.1 Gene expression in colon tissues

Gene expression data from colon tissues, including data on gene expression levels for 2000 genes, selected based on the highest minimum intensity, across 62 colon tissue samples [Alon et al., 1999], are available at `http://genomics-pubs.princeton.edu/oncology/affydata/index.html`. These samples were collected from 40 patients: 22 patients provided paired normal and tumor tissue samples, while 18 provided only tumor tissue samples. One normal tissue sample was identified as an outlier and the corresponding patient was excluded from the analysis.

Our goal was to identify genes with aberrant expression patterns in tumor tissues compared to normal tissues through C-scores. We focused on the 21 patients with paired normal and tumor samples, analyzing the empirical distributions of gene expression for each of the 2000 genes in both tissue types. For each gene, we computed the optimal transport map from the normal tissue expression distribution to the tumor tissue expression distribution. This map characterizes the differential expression patterns

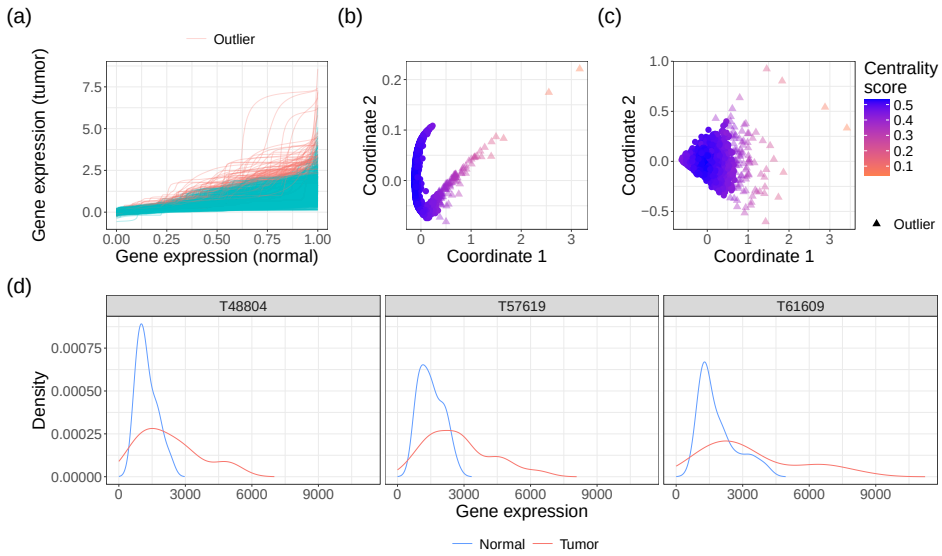


Figure 5: Gene expression in colon tissues with tuning parameter $\tau = 0$. (a) Optimal transport maps from normal tissue expression distribution to tumor tissue expression distribution for 2000 genes where red curves represent outliers. (b) Profile and (c) object MDS for optimal transport maps where triangles denote outliers. (d) Distributions of gene expression represented as densities in normal (blue) and tumor (red) tissues for three outlying genes.

between normal and tumor tissues. Here the optimal transport maps themselves serve as the object data. We view these as elements of the space $\mathcal{L}^2$.

For this application, the tuning parameter τ is selected using the empirical procedure described in Appendix B, yielding $\tau = 0$; the same procedure is used for all subsequent data applications. Figure 5(a) shows the optimal transport maps for all 2,000 genes, with genes whose C-scores fall below the 5th percentile classified as outliers and highlighted. Figure 5(d) displays the distributions of gene expression in the form of densities for normal and tumor tissues for three genes identified as outliers. These genes exhibit a rightward shift in expression density and increased variation in tumor tissues compared to normal tissues, suggesting their potential involvement in colon tumor development; they are homologous to ribosomal proteins and associated with cellular metabolism, including an ATP-synthase component and an elongation factor [Alon et al., 1999], consistent with previous findings [Pogue-Geile et al., 1991].

To visualize these complex data, we employed Multidimensional Scaling (MDS) [Mardia, 1978]. We considered two variants: object MDS based on the original distance d, and profile MDS based on the Wasserstein distance between distance profiles [Dubey et al., 2024]. As shown in Figure 5(b), profile MDS reveals a pronounced horseshoe pattern that provides a more informative representation than object MDS in Figure 5(c). In particular, profile MDS uncovers a clear trajectory from top left to top right of the embedding, along which the outlier genes are clustered on the right-hand side.

5.2 Climate change in Tampa

As anthropogenic activities drive global warming and urban heat islands intensify, numerous cities have experienced unprecedented hot summer nights in recent years [Bhatia and Katz, 2021]. In the summer of 2023, several Florida regions set new temperature records, with the Tampa Bay area emerging as one of the hottest zones. We analyzed temperature data recorded at Tampa International Airport from 1939 to 2023, obtained from the National Climatic Data Center (`https://www.ncdc.noaa.gov/cdo-web/search`), with a focus on the empirical distribution of daily minimum temperatures during the summer months (June through August) for each year.

Employing the Wasserstein metric for these distributional data as discussed in Example 3, we obtained C-scores and centrality regions for these 85 annual temperature distributions. Figure 6(a) illustrates the corresponding density functions, color-coded by their C-scores. Notably, 2022 and 2023 exhibit

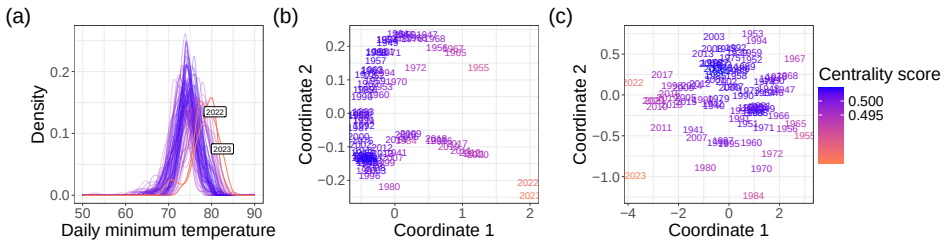


Figure 6: Climate change in Tampa with tuning parameter $\tau = 10$. (a) Density of daily minimum temperatures in summer months from 1939 to 2023 at Tampa. (b) Profile and (c) object MDS.

the lowest centrality and the most atypical temperature profiles among the 85-year record. The strong rightward shift in their density functions indicates that the Tampa Bay area experienced unprecedented heat during this period.

To visualize the complex relationships within these data, we employed profile MDS and object MDS (Figures 6(b) and (c)). Again, profile MDS proves effective in uncovering intrinsic relationships. Of particular interest is that the anomalous years 2022 and 2023 are distinctly positioned in the bottom right of the plot and immediately identifiable as outliers. Traversing the plot from top to bottom reveals a trend from earlier to more recent years, effectively capturing the accelerating impact of global warming.

Additional data applications, including handwriting analysis using MNIST and transport networks in Manhattan, are presented in Appendix C.

6 Limitations

The method depends on a scientifically meaningful object metric. Its ranking resolution can deteriorate when pairwise distances concentrate, the sample is sparse relative to intrinsic dimension, or coarse distances create many ties. The volume of a centrality region may also be undefined in a general metric space, which motivates the midpoint-based tuning criterion in Appendix B. These issues are less pronounced when local small-ball probabilities separate typical from atypical objects, especially for multimodal or heterogeneous populations; in a well-sampled unimodal setting, the global score at $\tau = 0$ may suffice.

For a reference sample of size n, metric evaluation cost T_d, and L_τ candidate values of τ, fitting requires $n(n-1)/2$ unique distances, $\Theta(n^2 T_d)$ metric time, $\Theta(n^2 \log n)$ sorting time, and $\Theta(L_\tau n^2)$ integration time. Storing all distances costs $\Theta(n^2)$ memory, although blocked computation can trade memory for recomputation. One query costs $\Theta(nT_d + n \log n + nL_\tau)$. The all-pairs midpoint tuner evaluates $\binom{n}{2}$ queries and is therefore the principal cubic bottleneck; distance profiles and query evaluations can be parallelized.

7 Concluding remarks

Centrality regions are of importance across statistical applications. Their construction is challenging even for multivariate data and even more so for random objects. We propose here C-scores based on weighted transports of distance profiles and show that the proposed approach leads to efficient and fairly small centrality regions for complex data. The data examples demonstrate that the proposed centrality regions are promising and relevant for a wide range of applications. Future work will explore how these regions can be used for clustering, outlier detection, and other statistical tasks, including applications to emerging complex data types such as networks and large language models. An important feature of the proposed methodology is that it is fully nonparametric and does not rely on parametric modeling assumptions, even for univariate or multivariate responses. For more general random objects, parametric models are often unavailable, making model-free approaches such as the proposed centrality regions the most natural, and sometimes the only, viable option.

Code and data needed to reproduce the simulations and data analyses are available at `https://github.com/hg-zh/centrality-reproducibility`.

Acknowledgments and Disclosure of Funding

Funding: HZ was supported in part by NSF grant DMS-2625746 and the SDSS Computational Data Science & AI Cloud Pilot Program at the University of North Carolina at Chapel Hill. The research of HGM was supported in part by NSF grant DMS-2310450. Computing resources for this work were provided in part by the University of North Carolina at Chapel Hill through its AI Acceleration Program.
Competing interests: The authors declare no competing interests.

References

Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

A Bhatia and J Katz. Why we're experiencing so many unusually hot summer nights. *The New York Times*, A12, 2021.

Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, 53:1–26, 2017.

Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics*, 27:733–767, 2001.

Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, 118(542):869–882, 2023.

Xiongtao Dai and Hans-Georg Müller. Principal component analysis for functional data on Riemannian manifolds and spheres. *Annals of Statistics*, 46(6B):3334 – 3361, 2018.

Xiongtao Dai, Sara Lopez-Pintado, and Alzheimer's Disease Neuroimaging Initiative. Tukey's depth for object data. *Journal of the American Statistical Association*, 118(543):1760–1772, 2023.

Ian L. Dryden, Alexey Koloydenko, and Diwei Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics*, 3(3):1102 – 1123, 2009.

Paromita Dubey, Yaqing Chen, and Hans-Georg Müller. Metric statistics: Exploration and inference for random objects with distance profiles. *Annals of Statistics*, 52(2):757–792, 2024.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231, 1996.

Donald AS Fraser and Irwin Guttman. Tolerance regions. *Annals of Mathematical Statistics*, 27(1): 162–179, 1956.

Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177:113–161, 1996.

Cedric E Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, and Eric D Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *Annals of Applied Statistics*, 11(2):725–750, 2017.

Juha Heinonen. *Lectures on Analysis on Metric Spaces*. Springer Science & Business Media, 2001.

L. Horvath and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.

Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015.

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Eric D Kolaczyk and Gábor Csárdi. *Statistical analysis of network data with R*. Springer Cham, 2nd edition, 2020.

Eric D Kolaczyk, Lizhen Lin, Steven Rosenberg, Jackson Walters, and Jie Xu. Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *Annals of Statistics*, 48(1):514–538, 2020.

Kalimuthu Krishnamoorthy and Thomas Mathew. *Statistical Tolerance Regions: Theory, Applications, and Computation*. John Wiley & Sons, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Xihao Li, Corbin Quick, Hufeng Zhou, Sheila M Gaynor, Yaowu Liu, Han Chen, Margaret Sunitha Selvaraj, Ryan Sun, Rounak Dey, Donna K Arnett, et al. Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, 55(1):154–164, 2023.

Zhenhua Lin and Fang Yao. Intrinsic Riemannian functional data analysis. *Annals of Statistics*, 47 (6):3533 – 3577, 2019.

Kanti V Mardia. Some properties of classical multi-dimensional scaling. *Communications in Statistics-Theory and Methods*, 7(13):1233–1241, 1978.

Hans-Georg Müller. Peter Hall, functional data analysis and random objects. *Annals of Statistics*, 44 (5):1867–1887, 2016.

Victor M. Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer New York, 2020.

Matteo Pegoraro and Mario Beraha. Projected statistical methods for distributional data on the real line with the Wasserstein metric. *Journal of Machine Learning Research*, 23:1686–1744, 2022.

Xavier Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

Alexander Petersen and Hans-Georg Müller. Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics*, 44(1):183–218, 2016.

Alexander Petersen, Chao Zhang, and Piotr Kokoszka. Modeling probability density functions as data objects. *Econometrics and Statistics*, 21:159–178, 2022.

Kay Pogue-Geile, John R Geiser, Min Shu, Carla Miller, Ira G Wool, Arnold I Meisler, and James M Pipas. Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human S3 ribosomal protein. *Molecular and Cellular Biology*, 11(8):3842–3849, 1991.

JL Scealy and AH Welsh. Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3): 351–375, 2011.

Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

Katie E Severn, Ian L Dryden, and Simon P Preston. Manifold valued data analysis of samples of networks, with applications in corpus linguistics. *Annals of Applied Statistics*, 16(1):368–390, 2022.

Joni Virta. Spatial depth for data in metric spaces. *Scandinavian Journal of Statistics*, 53(2):684–711, 2026. doi: 10.1111/sjos.70054.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional Data Analysis. *Annual Review of Statistics and its Application*, 3:257–295, 2016.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

Bingxin Zhao, Tengfei Li, Zirui Fan, Yue Yang, Juan Shu, Xiaochen Yang, Xifeng Wang, Tianyou Luo, Jiarui Tang, Di Xiong, et al. Heart-brain connections: Phenotypic and genetic insights from magnetic resonance images. *Science*, 380(6648):abn6598, 2023.

Hang Zhou and Hans-Georg Müller. Conformal inference for random objects. *Annals of Statistics*, 53(3):1123 – 1149, 2025.

Hang Zhou, Zhenhua Lin, and Fang Yao. Intrinsic correlation analysis for wasserstein functional data. *Statistica Sinica*, 36:627–648, 2026.

Yidong Zhou and Hans-Georg Müller. Network regression with graph Laplacians. *Journal of Machine Learning Research*, 23:1–41, 2022.

Yidong Zhou and Hans-Georg Müller. Wasserstein regression with empirical measures and density estimation for sparse data. *Biometrics*, 80(4):ujae127, 2024.

Changbo Zhu and Hans-Georg Müller. Autoregressive optimal transport models. *Journal of the Royal Statistical Society Series B*, 85(3):1012–1033, 2023.

A Proofs for Section 3

Proof of Theorem 1. Let $n = |\mathcal{D}_{\text{cal}}|$ without loss of generality and relabel the calibration observations as $\mathcal{D}_{\text{cal}} = \{X_1, \ldots, X_n\}$ and the independent test point as X_{n+1}. Fix the random split and condition on the realized training sample $\mathcal{D}_{\text{train}}$ and any independent randomization used for tuning. All probabilities below are conditional on this randomization as well. The choices of g and w are then fixed. Define the calibration scores

$$S_i := \hat{R}_{g,w}(X_i), \quad i = 1, \ldots, n, \quad \text{and} \quad S_{n+1} := \hat{R}_{g,w}(X_{n+1}).$$

Conditional on $\mathcal{D}_{\text{train}}$, the random variables $X_1, \ldots, X_n, X_{n+1}$ are i.i.d.; therefore the scores $S_1, \ldots, S_n, S_{n+1}$ are exchangeable.

Let $k := \lfloor (n+1)\alpha \rfloor$ and let $S_{(1)} \leq \cdots \leq S_{(n)}$ be the order statistics of $\{S_i\}_{i=1}^n$ (with the convention that $S_{(0)} = -\infty$ if $k = 0$). By the definition of the empirical quantile, $\hat{Q}_\alpha = S_{(k)}$. Hence,

$$\mathbb{P}\Big(X_{n+1} \in \hat{C}_\alpha \mid \mathcal{D}_{\text{train}}\Big) = \mathbb{P}\Big(S_{n+1} \geq \hat{Q}_\alpha \mid \mathcal{D}_{\text{train}}\Big) = 1 - \mathbb{P}\big(S_{n+1} < S_{(k)} \mid \mathcal{D}_{\text{train}}\big)\,.$$

Now consider the conformal p-value $p_{n+1} := (1 + \sum_{i=1}^n \mathbf{1}_{\{S_i \leq S_{n+1}\}})/(n+1)$. By exchangeability of $(S_1, \ldots, S_n, S_{n+1})$ conditional on $\mathcal{D}_{\text{train}}$, p_{n+1} is super-uniform, i.e., for all $t \in [0,1]$, $\mathbb{P}(p_{n+1} \leq t \mid \mathcal{D}_{\text{train}}) \leq t$. Moreover, the event $\{S_{n+1} < S_{(k)}\}$ implies that at most $k-1$ calibration scores are $\leq S_{n+1}$, hence

$$\{S_{n+1} < S_{(k)}\} \subseteq \left\{p_{n+1} \leq \frac{k}{n+1}\right\}.$$

Therefore,

$$\mathbb{P}\big(S_{n+1} < S_{(k)} \mid \mathcal{D}_{\text{train}}\big) \leq \mathbb{P}\left(p_{n+1} \leq \frac{k}{n+1} \mid \mathcal{D}_{\text{train}}\right) \leq \frac{k}{n+1} \leq \alpha,$$

where the last inequality uses $k = \lfloor (n+1)\alpha \rfloor$. Consequently,

$$\mathbb{P}\Big(X_{n+1} \in \hat{C}_\alpha \mid \mathcal{D}_{\text{train}}\Big) \geq 1 - \alpha.$$

Taking expectations over the training sample, split, and tuning randomization yields $\mathbb{P}(X_{n+1} \in \hat{C}_\alpha) \geq 1 - \alpha$. □

Proof of Theorem 2. Let X be a random vector in $\mathbb{R}^d$ with Lebesgue density f. Fix a point $\omega \in \mathbb{R}^d$ and let

$$D_\omega := \|X - \omega\|_2, \quad F_\omega(t) := \mathbb{P}(D_\omega \leq t), \quad t \geq 0.$$

Given a training sample $\mathcal{D}_{\text{train}} = \{X_i\}_{i=1}^m$ of i.i.d. copies of X, the empirical distance profile is

$$\hat{F}_\omega(t) := \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\|X_i - \omega\| \leq t\}}, \quad t \geq 0,$$

and its generalized inverse is $\hat{F}_\omega^{-1}(u) := \inf\{t \geq 0 : \ \hat{F}_\omega(t) \geq u\}, \ u \in (0,1]$. For $\tau > 0$, define the empirical weighted lower-tail functional

$$\hat{J}_\tau(\omega) := \int_0^1 \hat{F}_\omega^{-1}(u)\, e^{-\tau u}\, \mathrm{d}u. \tag{7}$$

Let v_d denote the volume of the unit Euclidean ball in $\mathbb{R}^d$: $v_d = \pi^{d/2}/\Gamma(d/2+1)$. We shall introduce the following lemmas.

Lemma 1. *Assume f is continuous at ω and $f(\omega) > 0$. Then as $t \downarrow 0$,*

$$F_\omega(t) = \mathbb{P}(\|X - \omega\| \leq t) = v_d f(\omega)\, t^d \{1 + o(1)\}.$$

Proof. Write $B(\omega,t)=\{x:\|x-\omega\|\le t\}$. By change of variables $x=\omega+tz$,

$$F_\omega(t)=\int_{B(\omega,t)} f(x)\,\mathrm{d}x = t^d\int_{\|z\|\le 1} f(\omega+tz)\,\mathrm{d}z.$$

By continuity of f at ω, $\sup_{\|z\|\le1}|f(\omega+tz)-f(\omega)|\to 0$ as $t\downarrow 0$. Thus

$$\int_{\|z\|\le1} f(\omega+tz)\,\mathrm{d}z = f(\omega)\int_{\|z\|\le1}\mathrm{d}z\{1+o(1)\} = v_d f(\omega)\{1+o(1)\}.$$

□

Lemma 2. *Let $R_i(\omega):=\|X_i-\omega\|$ and let $R_{(1)}(\omega)\le\cdots\le R_{(m)}(\omega)$ be their order statistics. Assume f is continuous at ω and $f(\omega)>0$. Let $k=k_m$ satisfy $k_m\to\infty$ and $k_m/m\to 0$. Then*

$$R_{(k_m)}(\omega)=\left(\frac{k_m}{mv_d f(\omega)}\right)^{1/d}\{1+o_{\mathbb{P}}(1)\}.$$

Proof. Fix $\varepsilon\in(0,1)$. Define

$$r_m^- := \left(\frac{(1-\varepsilon)k_m}{mv_d f(\omega)}\right)^{1/d},\quad r_m^+ := \left(\frac{(1+\varepsilon)k_m}{mv_d f(\omega)}\right)^{1/d}.$$

Because $k_m/m\to0$, we have $r_m^\pm\to0$. By Lemma 1, for m large enough,

$$F_\omega(r_m^-)\le(1-\varepsilon/2)\frac{k_m}{m},\quad F_\omega(r_m^+)\ge(1+\varepsilon/2)\frac{k_m}{m}.$$

Let $N(r):=\sum_{i=1}^m \mathbf{1}_{\{R_i(\omega)\le r\}}\sim\mathrm{Bin}(m,F_\omega(r))$. Then

$$\{R_{(k_m)}(\omega)\le r_m^-\}=\{N(r_m^-)\ge k_m\},\quad \{R_{(k_m)}(\omega)>r_m^+\}=\{N(r_m^+)\le k_m-1\}.$$

Standard Chernoff bounds for binomials imply

$$\mathbb{P}\big(N(r_m^-)\ge k_m\big)\le\exp(-c_1k_m),\quad \mathbb{P}\big(N(r_m^+)\le k_m-1\big)\le\exp(-c_2k_m),$$

for constants $c_1,c_2>0$ depending only on ε. Since $k_m\to\infty$, both probabilities vanish, and hence $\mathbb{P}(r_m^-\le R_{(k_m)}(\omega)\le r_m^+)\to1$, which completes the proof. □

For the proof of Theorem 2, fix ω with $f(\omega)>0$, and let $R_{(k)}:=R_{(k)}(\omega)$. Since $\hat F_\omega^{-1}(u)=R_{(k)}$ for $u\in((k-1)/m,k/m]$, we can write

$$\hat J_\tau(\omega)=\sum_{k=1}^m R_{(k)}\int_{(k-1)/m}^{k/m} e^{-\tau u}\,\mathrm{d}u=\frac{1}{\tau}\sum_{k=1}^m R_{(k)}\left(e^{-\tau(k-1)/m}-e^{-\tau k/m}\right). \tag{8}$$

Let $a_m:=c\log\tau_m$ with $c>1+1/d$ and set $K_m:=\lceil ma_m/\tau_m\rceil$ and $u_m:=K_m/m\ge a_m/\tau_m$. Then $e^{-\tau_m u}\le e^{-a_m}$ for all $u\in[u_m,1]$, hence

$$0\le\int_{u_m}^1\hat F_\omega^{-1}(u)e^{-\tau_m u}\mathrm{d}u\le e^{-a_m}\int_0^1\hat F_\omega^{-1}(u)\,\mathrm{d}u.$$

For any empirical quantile function, $\int_0^1\hat F_\omega^{-1}(u)\,du=\frac1m\sum_{i=1}^m\|X_i-\omega\|$. By the moment assumption $\mathbb{E}\|X-\omega\|<\infty$, this sample mean is $O_{\mathbb{P}}(1)$. Therefore

$$\tau_m^{1+1/d}\int_{u_m}^1\hat F_\omega^{-1}(u)e^{-\tau_m u}\mathrm{d}u\le\tau_m^{1+1/d}e^{-a_m}\,O_{\mathbb{P}}(1)=\tau_m^{1+1/d-c}\,O_{\mathbb{P}}(1)=o_{\mathbb{P}}(1),$$

since $c>1+1/d$.

Write $\hat J_{\tau_m}(\omega)=A_m+B_m$, where

$$A_m:=\int_0^{u_m}\hat F_\omega^{-1}(u)e^{-\tau_m u}\mathrm{d}u=\sum_{k=1}^{K_m}R_{(k)}\int_{(k-1)/m}^{k/m}e^{-\tau_m u}\mathrm{d}u,$$

and $B_m := \int_{u_m}^1 \hat{F}_\omega^{-1}(u) e^{-\tau_m u} \mathrm{d}u$. By the tail bound above, $\tau_m^{1+1/d} B_m = o_{\mathbb{P}}(1)$.

For $k \le K_m$,

$$\int_{(k-1)/m}^{k/m} e^{-\tau_m u} \mathrm{d}u = \frac{1}{\tau_m} \left(e^{-\tau_m (k-1)/m} - e^{-\tau_m k/m} \right) = \frac{1}{m} e^{-\tau_m k/m} \cdot \frac{e^{\tau_m/m} - 1}{\tau_m/m}.$$

Since $\tau_m/m \to 0$, we have $(e^{\tau_m/m} - 1)/(\tau_m/m) = 1 + o(1)$. Thus uniformly over $k \le K_m$,

$$\int_{(k-1)/m}^{k/m} e^{-\tau_m u} \mathrm{d}u = \frac{1}{m} e^{-\tau_m k/m} \{1 + o(1)\}. \tag{9}$$

Because $\tau_m/m \to 0$ and $a_m \to \infty$, we have

$$K_m = \left\lceil \frac{m a_m}{\tau_m} \right\rceil \to \infty \quad \text{and} \quad \frac{K_m}{m} = \frac{a_m}{\tau_m} + O(m^{-1}) \to 0.$$

Define $k_{0,m} := \lfloor (m/\tau_m)^{1/2} \rfloor$, so that $k_{0,m} \to \infty$ and $k_{0,m}/K_m \to 0$. First, the contribution of $k < k_{0,m}$ is negligible after scaling:

$$0 \le \sum_{k=1}^{k_{0,m}-1} R_{(k)} \int_{(k-1)/m}^{k/m} e^{-\tau_m u} \mathrm{d}u \le \int_0^{k_{0,m}/m} \hat{F}_\omega^{-1}(u) \, \mathrm{d}u \le \frac{k_{0,m}}{m} R_{(k_{0,m})}.$$

By Lemma 2 applied with $k_m = k_{0,m}$, $R_{(k_{0,m})} = \{k_{0,m}/(m v_d f(\omega))\}^{1/d} \{1 + o_{\mathbb{P}}(1)\}$, hence

$$\begin{aligned} \tau_m^{1+1/d} \cdot \frac{k_{0,m}}{m} R_{(k_{0,m})} &= (v_d f(\omega))^{-1/d} \tau_m^{1+1/d} \left(\frac{k_{0,m}}{m} \right)^{1+1/d} \{1 + o_{\mathbb{P}}(1)\} \\ &= (v_d f(\omega))^{-1/d} \left(\frac{k_{0,m}}{m/\tau_m} \right)^{1+1/d} \{1 + o_{\mathbb{P}}(1)\} = o_{\mathbb{P}}(1). \end{aligned}$$

Therefore it suffices to consider the sum over $k \in [k_{0,m}, K_m]$. For each such k, Lemma 2 gives

$$R_{(k)} = \left(\frac{k}{m v_d f(\omega)} \right)^{1/d} \{1 + o_{\mathbb{P}}(1)\}, \quad k \in [k_{0,m}, K_m], \tag{10}$$

with an error uniform over this range. Indeed, fix $\varepsilon \in (0,1)$ and set $r_{m,k}^{\pm} := \{(1 \pm \varepsilon) k/(m v_d f(\omega))\}^{1/d}$. Since $K_m/m \to 0$, the small-ball bounds in Lemma 2 hold simultaneously for all $k_{0,m} \le k \le K_m$, with the same Chernoff constants. A union bound therefore gives, for some $c_\varepsilon > 0$,

$$\mathbb{P}\Big(\exists k \in \{k_{0,m}, \dots, K_m\} : \, R_{(k)} \notin [r_{m,k}^-, r_{m,k}^+] \Big) \le 2 \sum_{k=k_{0,m}}^{K_m} e^{-c_\varepsilon k} \le \frac{2 e^{-c_\varepsilon k_{0,m}}}{1 - e^{-c_\varepsilon}} \to 0.$$

As ε is arbitrary, this proves the claimed uniform approximation.

Combining (9)–(10) yields

$$A_m = (v_d f(\omega))^{-1/d} \{1 + o_{\mathbb{P}}(1)\} \cdot \frac{1}{m} \sum_{k=k_{0,m}}^{K_m} \left(\frac{k}{m} \right)^{1/d} e^{-\tau_m k/m} + o_{\mathbb{P}}(\tau_m^{-(1+1/d)}).$$

Let $\Delta t := \tau_m/m$ and $t_k := \tau_m k/m$. Then $\Delta t \to 0$ and

$$\frac{1}{m} \sum_{k=k_{0,m}}^{K_m} \left(\frac{k}{m} \right)^{1/d} e^{-\tau_m k/m} = \tau_m^{-(1+1/d)} \sum_{k=k_{0,m}}^{K_m} t_k^{1/d} e^{-t_k} \, \Delta t.$$

Since $t_{k_{0,m}} = \tau_m k_{0,m}/m \le \sqrt{\tau_m/m} \to 0$ and $t_{K_m} = a_m + O(\Delta t)$, the sum is a Riemann approximation of $\int_0^{a_m} t^{1/d} e^{-t} \mathrm{d}t$ with vanishing mesh Δt. Moreover, $a_m \to \infty$ and $\int_0^\infty t^{1/d} e^{-t} \mathrm{d}t = \Gamma(1 + 1/d)$. Thus

$$\tau_m^{1+1/d} A_m = (v_d f(\omega))^{-1/d} \{1 + o_{\mathbb{P}}(1)\} \int_0^{a_m} t^{1/d} e^{-t} \mathrm{d}t \xrightarrow{\mathbb{P}} (v_d f(\omega))^{-1/d} \Gamma\left(1 + \frac{1}{d}\right).$$

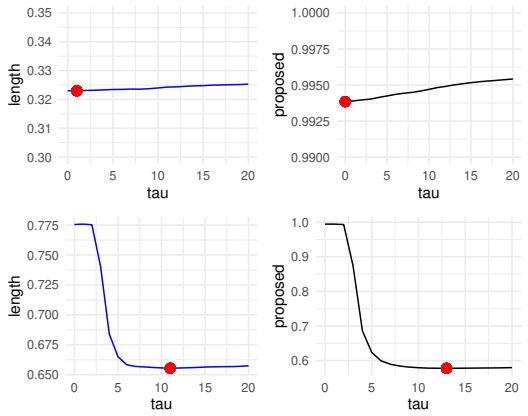


Figure 7: Length of centrality region (left column, blue lines) and the selection criterion for τ (see equation (11)) as a function of τ (black lines). The minimizing value is the selected value for τ and is indicated as a red dot. The data in the first row are generated from $\mathcal{N}(3,1)$, and the data in the second row are generated from a univariate Gaussian mixture $0.5\mathcal{N}(-3,1)+0.5\mathcal{N}(3,1)$. The sample size is $n=200$ for both unimodal and bimodal settings.

Finally, $\tau_m^{1+1/d}B_m = o_{\mathbb{P}}(1)$ by the tail bound above, so

$$\tau_m^{1+1/d}\,\hat{J}_{\tau_m}(\omega) \xrightarrow{\mathbb{P}} \Gamma\left(1+\frac{1}{d}\right)(v_d f(\omega))^{-1/d}.$$

The density-ordering consequence follows immediately since $f(\omega_1) > f(\omega_2)$ implies $(v_d f(\omega_1))^{-1/d} < (v_d f(\omega_2))^{-1/d}$. Note that

$$\hat{R}_\tau(\omega) = g\left(\frac{1}{m}\sum_{i=1}^{m}\hat{J}_\tau(X_i) - \hat{J}_\tau(\omega)\right).$$

The argument of g differs from $-\hat{J}_\tau(\omega)$ by an ω-independent constant. Since the strictly increasing function g preserves order, this completes the proof. □

B Tuning parameter selection

As discussed in Section 3, the selection of the tuning parameter τ significantly affects the size and shape of the obtained centrality regions. In practice, the volume of the centrality regions can be used as a criterion to select τ for finite-dimensional spaces $\mathcal{M}$, where volumes are well defined and evaluating them is straightforward. For Euclidean space $\mathbb{R}^d$, one can generate an equally spaced mesh grid on the cube $[-M,M]^d \subset \mathbb{R}^d$, with large enough M to contain all the observed data, and test whether each point in this mesh grid lies in the centrality region. The volume of the centrality region can then be approximated by the proportion of points in the centrality region over the total number of mesh grid points times the volume of the mesh grid. However, this approach is only feasible for spaces $\mathcal{M}$ for which volumes are well defined and can be easily evaluated. For infinite-dimensional spaces like $\mathcal{W}_2$ and $\mathcal{L}^2([0,1])$, one does not have a Lebesgue measure and volumes are not defined. Moreover, generating a mesh grid and evaluating the volume of the centrality region would be challenging and computationally expensive and in the end is not feasible.

Therefore, we propose an empirical method for the choice of τ that is easy to implement. To retain Theorem 1's guarantee, perform this selection entirely within $\mathcal{D}_{\text{train}}$, using an internal split to construct provisional regions; then refit the score on $\mathcal{D}_{\text{train}}$ and use $\mathcal{D}_{\text{cal}}$ only for the final threshold. For this, consider the $\binom{|\mathcal{D}_{\text{train}}|}{2}$ pairs of training points and determine whether the midpoint of the geodesic connecting each pair lies in the centrality region for a specific τ. This evaluation is feasible as long as $\mathcal{M}$ is a geodesic space, which includes Riemannian manifolds, Wasserstein space and many other spaces. Then τ is selected to minimize

$$\text{Prop}(\tau) + \lambda\tau, \tag{11}$$

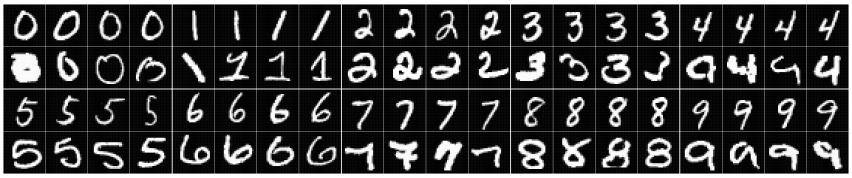


Figure 8: Handwriting quality assessment using the MNIST database with tuning parameter $\tau = 10$. The first and third rows illustrate the four most representative (inlying) handwritten samples for each digit, while the second and fourth rows demonstrate the four most atypical (outlying) handwritten samples.

where $\mathrm{Prop}(\tau)$ is the proportion of the midpoints lying in the centrality region with tuning parameter τ and the term $\lambda\tau$ penalizes against too large τ. A generic choice of $\lambda = 0.0001$, while ad hoc, was found to result in universally small sized centrality regions. Figure 7 demonstrates the comparison between the volume curves as a function of τ and the proposed selection criterion for unimodal and bimodal settings in $\mathcal{M} = \mathbb{R}^1$. This shows that the proposed method selects a reasonable tuning parameter τ in this scenario.

C Additional data applications

C.1 Handwriting quality assessment using MNIST

The MNIST database is a widely used dataset of handwritten digits, consisting of 10 classes (0–9) [LeCun et al., 1998]. Each sample in this dataset is represented as a 28×28 pixel grayscale image. The objective was to apply C-scores to identify atypical or poor-quality handwriting samples within this dataset. We employed the Euclidean metric to quantify the dissimilarity between these 28×28-dimensional image objects.

For this analysis, we randomly selected a subset of 200 samples for each of the 10 digit classes. We then computed C-scores separately for each digit class to identify atypical handwritten samples using metric centrality. The results are shown in Figure 8, contrasting typical and atypical samples for each digit. The first and third rows display the four most representative (inlying) samples for each digit, while the second and fourth rows show the four most atypical (outlying) samples. Figure 8 demonstrates that the atypical samples identified through C-scores indeed correspond to unusual writing styles or poorly formed digits. These outliers often exhibit characteristics such as extreme slanting, incomplete formation, or exaggerated features that deviate significantly from the norm for each digit class.

C.2 Manhattan transport networks

We analyze yellow taxi trip records in Manhattan, excluding islands, based on data available at `https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`. The 66 taxi zones are grouped into 13 regions following the preprocessing outlined in Zhou and Müller [2022]. Our main focus is the transport network that captures daily passenger movements between these 13 regions throughout 2019. To model this, we construct daily undirected networks, where the nodes represent the 13 regions and the edge weights correspond to the number of passengers traveling between them. These edge weights are normalized by the maximum edge weight of the day, ensuring they fall within $[0, 1]$. Each network is represented by a 13×13 graph Laplacian.

Using the space of graph Laplacians with the Frobenius metric, as detailed in Example 2, we computed C-scores for the 365 daily transport networks. Figure 9 presents the corresponding profile MDS, revealing a horseshoe shape similar to Figures 5(b) and 6(b), indicating a nearly one-dimensional ordering of the distance profiles. Notably, the profile MDS effectively distinguishes between weekday and weekend transport patterns in Manhattan.

This application ranks whole networks across days and therefore differs from degree, closeness, betweenness, eigenvector centrality, and PageRank, which rank nodes within one network. To examine whether adaptive weighting preserves the weekday and weekend regimes, we sampled 5,000 geodesic midpoint networks from each of the weekday–weekday, weekend–weekend, and

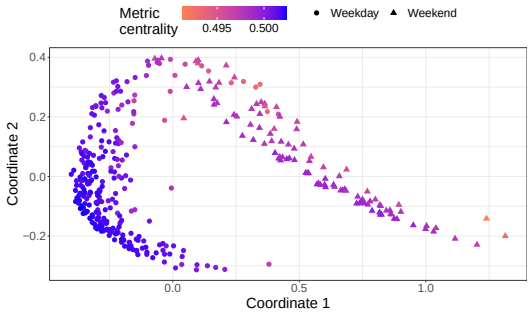


Figure 9: Profile MDS for Manhattan transport networks in 2019 with tuning parameter $\tau = 10$.

Table 1: Clustering for Manhattan transport networks in 2019 using DBSCAN.

Cluster #	Mon	Tue	Wed	Thu	Fri	Sat	Sun
DBSCAN using distance profiles							
0	50	52	51	48	49	2	3
1	0	0	0	1	0	48	44
2	1	0	0	3	3	2	5
3	1	1	1	0	0	0	0
DBSCAN using transport networks							
0	49	51	50	48	50	52	48
1	2	1	1	2	0	0	4
2	1	1	1	1	0	0	0
3	0	0	0	1	2	0	0

weekday–weekend pair groups. We then recorded how often a randomly selected within-regime midpoint was ranked above a weekday–weekend midpoint. This rate increased from 43.3% for the unweighted transport rank ($\tau = 0$) to 75.4% for the C-score with $\tau = 10$; 50% indicates no systematic ordering. For weekend midpoints the rate increased from 1.8% to 53.2%, and for weekday midpoints it increased from 84.8% to 97.6%. Thus lower-tail weighting reduces the strong disadvantage of the smaller weekend regime while sharpening separation for the weekday regime.

We also compared the C-score with metric spatial depth [Virta, 2026], using its closed-form expression for graph Laplacians under the Frobenius metric. A within-regime midpoint was ranked above a weekday–weekend midpoint in 75.4% of comparisons for the C-score, compared with 27.2% for metric spatial depth. The latter more often favors networks lying globally between the weekday and weekend regimes, whereas the adaptive C-score more closely preserves the two-regime structure.

To demonstrate the potential of the distance profiles in clustering tasks, we applied DBSCAN [Ester et al., 1996] using the Wasserstein metric on the distance profiles of the 365 transport networks. This analysis evaluates the distance-profile representation; C-scores themselves are centrality measures rather than a clustering algorithm. The results are summarized in Table 1. For comparison, we also applied DBSCAN directly to the transport networks using the Frobenius metric. The distance profiles provided clearer clustering, with two main clusters corresponding to weekdays and weekends. Upon closer inspection, the only Thursday grouped with the weekend cluster was July 4, Independence Day, which shows a weekend-like pattern. The third cluster consists of days around holidays, such as July 5, September 1, and November 29, while the last cluster includes September 23–25, designated as "gridlock alert days" by New York City due to the UN General Assembly meetings. In contrast, clustering based on the transport networks directly combined weekdays and weekends into a single cluster, failing to capture this distinction.

D Comparison with recent centrality measures

We compare the C-score with four centrality constructions for metric-valued data, orienting each score so that larger values indicate greater centrality. Transport cost uses the expected one-dimensional Wasserstein discrepancy $\mathbb{E}\int_0^1 |F_X^{-1}(u) - F_\omega^{-1}(u)|\,\mathrm{d}u$ [Zhou and Müller, 2025]; transport rank is

the $\tau = 0$ member of the C-score family [Dubey et al., 2024]; metric halfspace depth uses metric halfspaces $H_{a,b} = \{y : d(y, a) \leq d(y, b)\}$ [Dai et al., 2023]; and metric spatial depth replaces Euclidean spatial signs by a distance-defined metric cosine [Virta, 2026].

We studied two equally weighted Gaussian mixtures: Scenario A, $\frac{1}{2}N(-0.75, 1) + \frac{1}{2}N(0.75, 1)$, is unimodal with mode zero; Scenario B, $\frac{1}{2}N(-2.5, 1) + \frac{1}{2}N(2.5, 1)$, has two separated modes. For each of 200 independent replications, $n_{\text{train}} = n_{\text{cal}} = 200$ and coverage was evaluated on 2,000 independent test observations at a 95% target level. The C-score selected $\tau \in \{0, \ldots, 20\}$ using Appendix B; all methods used the same samples, splits, and inclusive thresholds. Parentheses in Tables 2 and 3 are Monte Carlo standard errors. Total length is the sum of the Lebesgue lengths of all connected components.

Table 2: Comparison for the unimodal mixture in Scenario A.

Method	Coverage	Total length ↓	Mode included ↑
C-score	0.950 (0.001)	4.862 (0.022)	1.000 (0.000)
Transport rank	0.951 (0.001)	4.882 (0.021)	1.000 (0.000)
Transport cost	0.951 (0.001)	4.882 (0.021)	1.000 (0.000)
Metric spatial/halfspace depth	0.958 (0.001)	∞; bounded: 5.077 (0.029)	1.000 (0.000)

Table 3: Comparison for the separated bimodal mixture in Scenario B.

Method	Coverage	Total length ↓	Both modes ↑	Valley excluded ↑
C-score	0.950 (0.001)	7.951 (0.036)	1.000 (0.000)	0.905 (0.021)
Transport rank	0.952 (0.001)	8.412 (0.020)	1.000 (0.000)	0.000 (0.000)
Transport cost	0.952 (0.001)	8.224 (0.032)	1.000 (0.000)	0.435 (0.035)
Metric spatial/halfspace depth	0.960 (0.001)	∞; bounded: 8.601 (0.027)	1.000 (0.000)	0.000 (0.000)

In Scenario A, the C-score region was shorter than the transport-rank and transport-cost regions by 0.019 (paired Monte Carlo standard error 0.003), while every method included the mode in every replication. In Scenario B, every method included both modes, but the C-score was shorter than transport rank by 0.461 (0.030) and transport cost by 0.273 (0.029). It excluded the valley in 90.5% of replications, compared with 0% for transport rank and the two metric depths and 43.5% for transport cost. The depth regions were unbounded in 1.0% of Scenario A replications and 0.5% of Scenario B replications because an inclusive calibration cutoff of zero accepts all of $\mathbb{R}$; the reported finite lengths condition on bounded regions. These results show that the adaptive score retains the global behavior of the competing methods in the unimodal setting while better preserving separated modal structure.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe the proposed C-scores, their theoretical properties (Section 3), and empirical performance (Sections 4–5), which are consistent with the results presented.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the proposed method, including dependence on the object metric, distance concentration and ties, the absence of a natural volume measure, and computational scalability, are discussed in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results (Theorems 1 and 2) are clearly stated with assumptions, and complete proofs are provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The simulation setup, data-generating mechanisms, evaluation metrics, parameter choices, and competitor implementations are specified in Section 4 and Appendix D, and the accompanying repository provides the code and data.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code, data, and instructions for reproducing the simulations and real-data analyses are available at `https://github.com/hg-zh/centrality-reproducibility`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies data generation, sample sizes, tuning parameters, and evaluation procedures for all experiments (Section 4 and Section 5).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The competitor study in Appendix D reports Monte Carlo standard errors over 200 independent replications and paired Monte Carlo standard errors for the principal method comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not report computational resources, as experiments are lightweight simulations and standard data analyses.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics and does not involve any unethical data collection or use.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: This work focuses on methodological developments in statistical theory and does not target specific real-world deployment scenarios; therefore, no direct societal impacts are identified. Potential downstream applications are broad and context-dependent.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release models or datasets that pose a risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets used are publicly available and properly cited in the references and text.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets, models, or software artifacts.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve human subject experiments or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subject research requiring IRB approval is conducted.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as part of the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.