Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Adaptive manifold clustering

Franz Besold, Vladimir Spokoiny

submitted: December 18, 2020 (revision: August 25, 2022)

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: franz.besold@wias-berlin.de
vladimir.spokoiny@wias-berlin.de

No. 2800
Berlin 2022

2010 *Mathematics Subject Classification.* 62H30, 62G10.

Key words and phrases. Adaptive weights, likelihood-ratio test, nonparametric clustering, manifold, reach.

Financial support by German Ministry for Education via the Berlin Center for Machine Learning (01IS18037I) is gratefully acknowledged. The research was also supported by the Russian Science Foundation grant No. 18-11-00132.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Adaptive manifold clustering

Franz Besold, Vladimir Spokoiny

Abstract

Clustering methods seek to partition data such that elements are more similar to elements in the same cluster than to elements in different clusters. The main challenge in this task is the lack of a unified definition of a cluster, especially for high-dimensional data. Different methods and approaches have been proposed to address this problem. This paper continues the study originated by [10] where a novel approach to adaptive nonparametric clustering called *Adaptive Weights Clustering* (AWC) was offered. The method allows analyzing high-dimensional data with an unknown number of unbalanced clusters of arbitrary shape under very weak modeling assumptions. The procedure demonstrates a state-of-the-art performance and is very efficient even for large data dimension D. However, the theoretical study in [10] is very limited and did not really address the question of efficiency. This paper makes a significant step in understanding the promising performance of the AWC procedure, particularly in high dimension. The approach is based on combining the ideas of adaptive clustering and manifold learning. The manifold hypothesis means that high-dimensional data can be well approximated by a d-dimensional manifold for small d helping to overcome the *curse of dimensionality* problem and to get sharp bounds on the cluster separation which only depend on the intrinsic dimension d. We also address the problem of parameter tuning. Our general theoretical results are illustrated by some numerical experiments.

1 Introduction

1.1 Manifold Clustering

The task of clustering is often informally described as partitioning a set of objects such that objects in the same group are more similar to each other than to those in other groups. The lack of a unified definition has led to a range of algorithms with different objectives. One of the oldest and best-known procedures are centroid-based methods such as k-means [34]. Other well-known approaches are density-based methods, like DBSCAN [12] or spectral methods [23]. For a comprehensive survey of clustering methods, we refer to [39]. A more general task is to obtain a hierarchical collection of clusters, the so-called *density cluster tree* [17]. This problem has been studied thoroughly, see e.g. [8], [21], [11] and [4] for more recent work. Allthough this approach avoids the choice of a scale parameter, it utilizes a specific definition of clusters beeing connected components of superlevel sets of the underlying density. In this paper, we study a nonparametric clustering algorithm originated from [10] and called *Adaptive Weights Clustering (AWC)*. It is *adaptive* as it does not require the user to specify the number of clusters, and it is able to recover clusters of different size, level of density and shape, including non-convex clusters. The cluster structure of the data is represented by an adjacency matrix containing binary entries, so-called *weights*, hence the name. The adjacency matrix is not guaranteed to correspond to a partition of the data, but rather will give information about local clusters for each data point. Informally speaking, the objective of the algorithm is to find maximal subsets of

DOI 10.20347/WIAS.PREPRINT.2800 Berlin, December 18, 2020/rev. August 25, 2022

the data without any significant gap, that is a region within the cluster adjoining two areas in opposite direction of relatively larger density. This novel objective is in fact the reason for the high adaptivity of AWC to clusters with very different structural properties.

This paper focuses on a theoretical study of the algorithm, as [10] already provides a comprehensive comparative numerical study. In particular, we want to address the challenges that arise from high-dimensional data that does not concentrate on lower-dimensional linear subspaces and where the PCA analysis does not yield a significant spectral gap. We are therefore interested in the case of high-dimensional data lying close to a lower-dimensional submanifold $\mathcal{M}$. This setup has already been studied for other clustering algorithms, e.g. in [4] and [19]. Moreover, it appears in the context of homology inference [5]. It has been shown that this is a realistic model for various data, e.g. for images which are represented in a patch space [28, 26] and a wide range of algorithms have been proposed to deal with the problem of non-linear dimension reduction [40], e.g. multidimensional scaling (MDS), kernel PCA, Isomap, Laplacian eigenmaps, self-organizing maps (SOM), locally-linear embeddings and autoencoders [33]. In this work, we will not rely on any of these techniques, however, we recommend using a manifold denoising algorithm in practice such as [30] as an additional preprocessing step in order to reduce the magnitude of the noise.

1.2 Submanifolds with positive reach

As regularity condition for the manifold we assume a positive *reach*, see Definition 1.

Definition 1. *For $\epsilon > 0$ and a set $S \subset \mathbb{R}^D$, let us denote the ϵ-offset of S by*

$$S^\epsilon = \{y \in \mathbb{R}^D : \exists x \in S \text{ with } \|x - y\| \leq \epsilon\}$$

and define the reach *of S to be*

$$\text{reach}(S) := \sup\{r \geq 0 : \forall y \in S^r \text{ there exists a unique } x \in S \text{ nearest to } y\}.$$

Originally introduced by [13], a positive reach has proven to be a widely used minimal condition in geometric and topological inference, c.f. [7]. This includes in particular the topics of manifold estimation [15], [2] as well as homology inference [24], [5]. The latter can in fact be seen as a generalization of the clustering problem.

If a set has a positive reach $\frac{1}{\kappa}$, it is also $\frac{1}{\kappa}$-convex and one can freely roll a ball of radius $r < \frac{1}{\kappa}$ around it [9]. The reach provides information about the local and the global structure of the manifold at the same time [1]: Any unit speed geodesic of a compact smooth submanifold $\mathcal{M}$ without boundary with $\text{reach}(\mathcal{M}) \geq \frac{1}{\kappa} > 0$ has a curvature bounded by κ and also any so-called bottleneck, i.e. a point on the manifold that has two distinct projections onto the manifold in exactly opposite directions, has a distance of at least $\frac{1}{\kappa}$ to $\mathcal{M}$. More precisely, it can be shown that the reach is either attained by the curvature of a unit speed geodesic or is equal to the distance of a bottleneck to the manifold. See Figure 1 for a visualization. Moreover, $\mathcal{M}$ has a local Lipschitz continuous parametrization in terms of the tangent plane, see Lemma 4. We exploit this property, using that any L -Lipschitz function changes the d -dimensional Lebesgue volume at most by a factor L^d, see Lemma 3. For a survey on sets with positive reach see [35].

1.3 AWC revisited

The key ingredient of the AWC procedure is a so-called *test of no gap*, which is based on a likelihood-ratio test for local homogeneity from [29]. Given a sequence of radii $0 < h_0 < \cdots < h_K$ in addition to

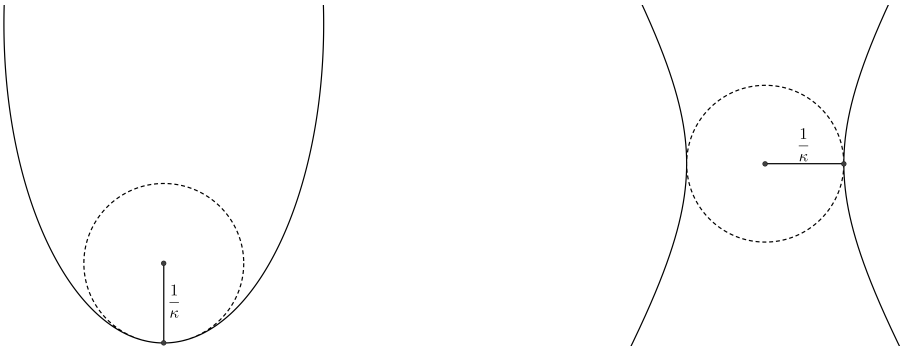


Figure 1: The reach of a manifold can be either attained by the curvature radius of a geodesic (left) or the distance to a bottleneck (right)

Figure 2: For locally homogeneous data we observe $\theta_{ij}^{(k)} \approx q_{ij}^{(k)}$ (left), whereas a significant gap is characterized by $\theta_{ij}^{(k)} \ll q_{ij}^{(k)}$ (right)

our data $X_1, \dots, X_n \in \mathbb{R}^D$ and using the test of no gap, the algorithm successively screens subsets of increasing diameters. Using information from previous steps, AWC defines at each step k around each point X_i a so-called *local cluster* $C_i^{(k)}$ that is supposed to be a maximal subset of the data in a vicinity of the given radius h_k satisfying the no gap objective.

In the following, let us explain the main idea of the algorithm more formally. An exact description via pseudocode is given in Algorithm 1. By $\|\cdot\|$ we denote the euclidean norm, λ denotes the D-dimensional Lebesgue measure and $B(\cdot,\cdot)$ is the usual notation for a closed euclidean Ball in $\mathbb{R}^D$ with given center and radius. Suppose our data $X_1, \dots, X_n \in \mathbb{R}^D$ is sampled independently from a common probability distribution $\mathbb{P}$. Using regular conditional distributions, let us treat X_i and X_j as deterministic for some $i \neq j$. From a given sequence of radii $h_0 < h_1 < \cdots < h_K$ s.t. $\frac{h_{l+1}}{h_l} < 2$ we choose h_k such that $\|X_i - X_j\| < h_k$ and define the so-called *gap coefficient*

$$\theta_{ij}^{(k)} = \frac{\mathbb{P}\left(B(X_i, h_{k-1}) \cap B(X_j, h_{k-1})\right)}{\mathbb{P}\left(B(X_i, h_{k-1}) \cup B(X_j, h_{k-1})\right)}.$$

In case of our distribution being uniform on a neighborhood of $B(X_i, h_k) \cup B(X_j, h_k)$, or more generally, having a linear density, the gap coefficient coincides with the so-called *volume coefficient*

$$q_{ij}^{(k)} = \frac{\lambda\left(B(X_i, h_{k-1}) \cap B(X_j, h_{k-1})\right)}{\lambda\left(B(X_i, h_{k-1}) \cup B(X_j, h_{k-1})\right)}.$$

In Figure 2, we visualize the relationship between those two quantities. The idea of a significant gap is formalized using a likelihood-ratio test of the null hypothesis

$$H_0 : \theta_{ij}^{(k)} \geq q_{ij}^{(k)}$$

against the alternative

$$H_1 : \theta_{ij}^{(k)} < q_{ij}^{(k)}.$$

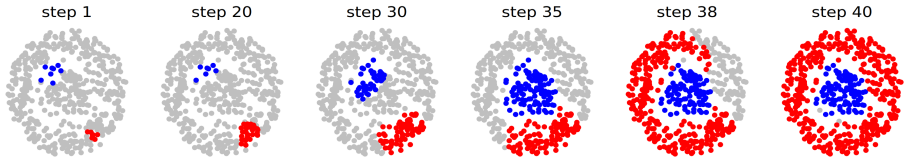


Figure 3: Local clusters during different steps of the AWC algorithm

Suppose we are given binary weights $w_{ij}^{(k-1)} = \mathbb{1}(\|X_i - X_j\| \leq h_{k-1})$ and let us denote the local cluster around X_i of radius h_{k-1} by $\mathcal{C}_i^{(k-1)} = \{X_j : w_{ij}^{(k-1)} = 1\}$. Then the corresponding test statistic can be written as

$$T_{ij}^{(k)} = N_{i\vee j}^{(k)} \mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)}) \left(\mathbb{1}(\widetilde{\theta}_{ij}^{(k)} < q_{ij}^{(k)}) - \mathbb{1}(\widetilde{\theta}_{ij}^{(k)} \geq q_{ij}^{(k)}) \right), \tag{1}$$

where

$$N_{i\vee j}^{(k)} = \sum_{l \neq i,j} \mathbb{1}(X_l \in \mathcal{C}_i^{(k-1)} \cup C_j^{(k-1)})$$

denotes the *empirical mass of the union*, $\mathcal{K}(\alpha, \beta)$ denotes the Kullback-Leibler divergence of two Bernoulli variables with means α and β and

$$\widetilde{\theta}_{ij}^{(k)} = \frac{\sum_{l \neq i,j} \mathbb{1}(X_l \in \mathcal{C}_i^{(k-1)} \cap C_j^{(k-1)})}{N_{i\vee j}^{(k)}}$$

is an estimator for the gap coefficient. In the AWC algorithm, the assumption of the weights being of the non-adaptive form $w_{ij}^{(k-1)} = \mathbb{1}(\|X_i - X_j\| \leq h_{k-1})$ will only be guaranteed for the first step, as the weights are successively updated as

$$w_{ij}^{(k)} = \mathbb{1}(d(X_i, X_j) \leq h_k) \mathbb{1}(T_{ij}^{(k)} \leq \lambda)$$

for some parameter $\lambda \in \mathbb{R}$. That is, the so-called *test of no gap* given in (1) that is used in the procedure does not necessarily coincide with the likelihood-ratio test, complicating the theoretical study. However, those successive updates allow the weights to carry information from all previous steps and enable the algorithm to detect gaps at any scale, in particular at a significantly smaller scale than the size of the final clusters.

The output of the algorithm will be a weight matrix $\left(w_{ij}^{(K)}\right)_{i,j=1}^{n}$. Experiments have shown this matrix to carry relevant information about the cluster structure of the data. In fact, AWC performs well on artificial and real-live data benchmarks. However, there is no theoretical guarantee, that these weights actually describe the edge-disjoint union of fully connected graphs. The lack of a well-defined global cluster objective of AWC distinguishes it from most other methods and can be seen as a disadvantage from a comparative point of view. But from a practical point of view, this allows the algorithm to adapt well to a very inhomogeneous and unknown cluster structure. Moreover, the local cluster structure can also be seen as an advantage as it allows for overlapping clusters.

The idea of the no gap test seems similar to a density-based method such as DBSCAN. This is in fact true on a local level in most situations. However, the absolute density levels are irrelevant for the local decisions of the AWC procedure. Thus, the results on a global level differ significantly from those obtained at a certain level of a density level tree, c.f. figure 5.

Currently, there is a significant gap between practical and theoretical results on AWC. Experiments have shown the algorithm to deliver state-of-the-art performance on a wide range of artificial and real-life examples. Some artificial examples are shown in Figure 4. Theoretical results are fairly limited:

Algorithm 1 Adaptive Weights Clustering (AWC)

input: data $X_1, \dots, X_n \in \mathbb{R}^D$, a sequence of bandwidths $0 < h_0 < \cdots < h_K$ and a threshold $\lambda \in \mathbb{R}$ for the likelihood-ratio test

initialize the weights $w_{ij}^{(0)} = \mathbb{1}(\|X_i - X_j\| \le h_0)$, $1 \le i, j \le n$

for k from 1 to K **do**

  for $i \ne j$ s.t. $\|X_i - X_j\| \le h_k$ **do**

    compute the empirical mass of the union

$$N_{i \vee j}^{(k)} = \sum_{l \ne i,j} \mathbb{1}(X_l \in \mathcal{C}_i^{(k-1)} \cup C_j^{(k-1)})$$

    where $\mathcal{C}_i^{(k-1)} := \{X_j : w_{ij}^{(k-1)} = 1\}$.

    compute the estimation of the gap coefficient

$$\widetilde{\theta}_{ij}^{(k)} = \frac{\sum_{l \ne i,j} \mathbb{1}(X_l \in \mathcal{C}_i^{(k-1)} \cap C_j^{(k-1)})}{N_{i \vee j}^{(k)}}$$

    compute the likelihood-ratio test statistic

$$T_{ij}^{(k)} = N_{i \vee j}^{(k)} \mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)}) \left(\mathbb{1}(\widetilde{\theta}_{ij}^{(k)} < q_{ij}^{(k)}) - \mathbb{1}(\widetilde{\theta}_{ij}^{(k)} \ge q_{ij}^{(k)}) \right)$$

    where $\mathcal{K}(\alpha, \beta) = \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1-\alpha}{1-\beta}$ and

$$q_{ij}^{(k)} = \left(2 \frac{\mathcal{B}\left(\frac{D+1}{2}, \frac{1}{2}\right)}{\mathcal{B}\left(1 - \frac{\|X_i - X_j\|^2}{4h_{k-1}^2}, \frac{D+1}{2}, \frac{1}{2}\right)} - 1 \right)^{-1}$$

    with $\mathcal{B}(\cdot, \cdot, \cdot)$ denoting the incomplete beta function and $\mathcal{B}(\cdot, \cdot) = \mathcal{B}(1, \cdot, \cdot)$ denoting the usual beta function

  end for

  update the weights

$$w_{ij}^{(k)} = \begin{cases} \mathbb{1}(\|X_i - X_j\| \le h_k) \mathbb{1}(T_{ij}^{(k)} \le \lambda) & \text{for } 1 \le i \ne j \le n \\ 1 & \text{for } 1 \le i = j \le n \end{cases}$$

end for

output: matrix of weights $\left(w_{ij}^{(K)} \right)_{i,j=1}^{n}$

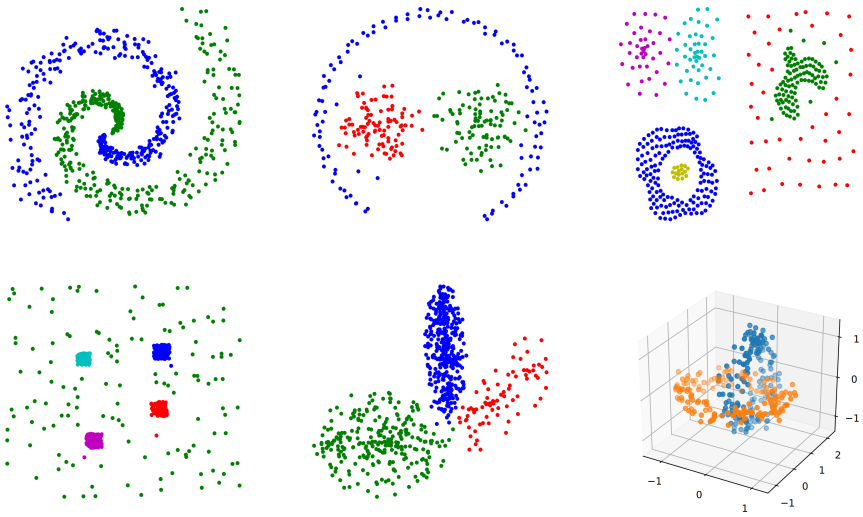
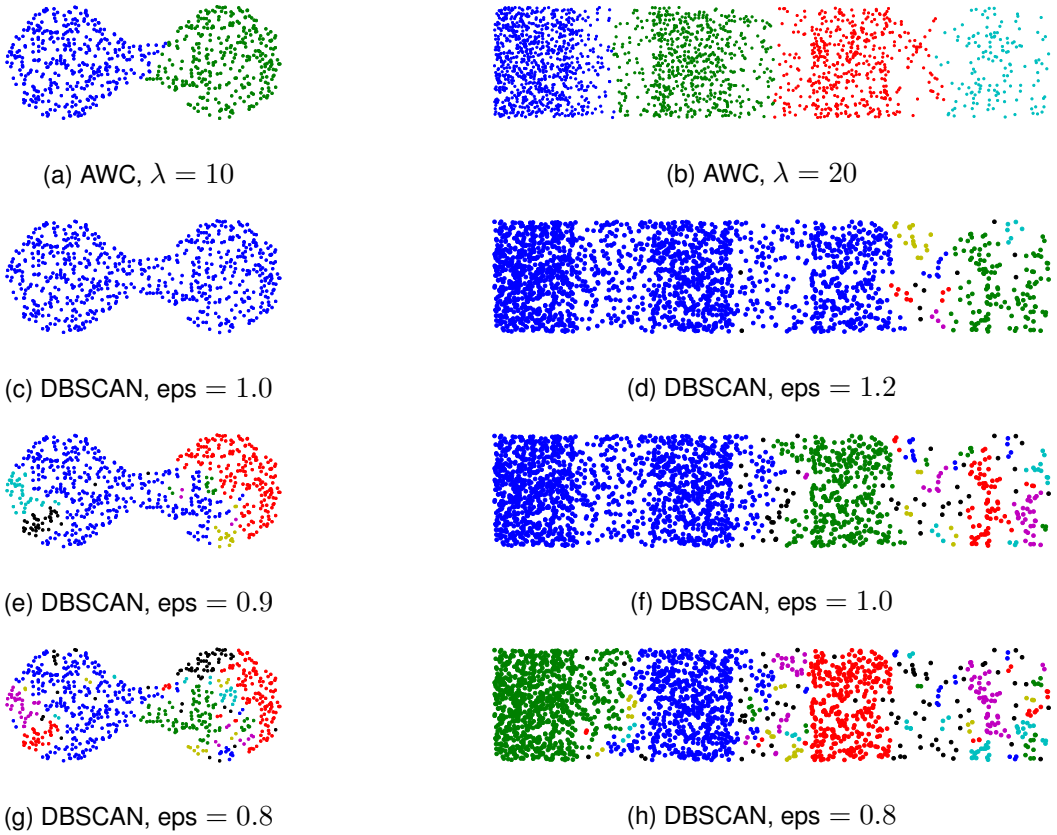


Figure 4: Six artificial examples demonstrate the adaptivity of AWC w.r.t. clusters of different size and density, non-convex shapes and clusters with manifold structure. The top left and the bottom right examples are original data sets, the rest are taken from [6].

First of all, they are limited to the case where no gaps have been detected in the previous step, as otherwise, the test of no gap does not necessarily coincide with a likelihood-ratio test. Finite sample guarantees on the propagation effect are only given at a local scale under the assumption of homogeneity due to the lack of results concerning the propagation at the boundaries of the clusters. A result about consistent separation is stated for the special case of i.i.d. data $X_1, \dots, X_n$ from a piecewise constant density supported on three neighboring regions of equal cylindrical shape. A sufficient condition that allows consistency is that the density is smaller by a factor $(1 - \epsilon_n)$ on the middle cylinder than on the other two and that $n\epsilon_n^2(\log n)^{\check{}1}$ is large enough. It turns out that this rate is optimal up to the logarithmic factor, more precisely it is impossible for any algorithm to achieve consistent separation if $n\epsilon_n^2 \nrightarrow \infty$. It has also been shown, that AWC adapts asymptotically to a linear submanifold structure of the data if the intrinsic dimension is known. However, specific conditions on the size of the considered deviation from the linear manifold are missing. Moreover, the procedure requires a crucial tuning parameter λ. This parameter has to grow logarithmically in the data size n to ensure both propagation and separation. Unfortunately, these results do not indicate how to scale λ, as no finite sample guarantee is given for the separation case.

In this work, we will significantly improve the current theory for AWC, and also solve some of the open problems mentioned above. First of all, we will consider distributions supported in the vicinity of closed non-linear submanifolds. We propose a slight adjustment of the algorithm in order to take into account the intrinsic dimension as well as local deviations due to the curvature of the manifold and the magnitude of the noise. In addition to generalizing the previous results to this setup, we will give finite sample guarantees both for propagation and separation and propose a theoretically justified choice for λ under rather general assumptions on the structure of the clusters. Moreover, we show that the propagation effect is still valid for points close to the boundary of a homogeneous cluster. This means that the propagation and separation results do no longer need to be stated separately, c.f. Corollary 3. The rest of the paper is organized as follows. In section 2 we present our main results. We start in subsection 2.1 by introducing the manifold hypothesis and studying properties of the gap coefficient. This leads to the introduction of the so-called *adjusted volume coefficient* and a minor modification of the algorithm which will preserve consistency under the manifold hypothesis. In subsection 2.2 we discuss the case of uniform data without any clusters and continue in 2.3 by studying the sensitivity of



(a) AWC, $\lambda = 10$

(b) AWC, $\lambda = 20$

(c) DBSCAN, eps $= 1.0$

(d) DBSCAN, eps $= 1.2$

(e) DBSCAN, eps $= 0.9$

(f) DBSCAN, eps $= 1.0$

(g) DBSCAN, eps $= 0.8$

(h) DBSCAN, eps $= 0.8$

Figure 5: Two datasets and the corresponding clusters obtained via AWC and DBSCAN. The cluster structures obtained via AWC differ from those obtained at a certain level of a density cluster tree. In the left example, DBSCAN is not able to recover the cluster structure because the density is constant, whereas for the right example the density levels of the different clusters and the spaces between them vary too much.

the algorithm w.r.t. local gaps. We will show that the procedure is rate-optimal and discuss the problem of parameter tuning. Finally, we discuss the boundary case in subsection 2.4. In the following section 3 we present numerical results illustrating the main results of section 2. Proofs are collected in section 4.

2 Theoretical results

2.1 Inequalities for the gap coefficient

When the dimension of the data is too large, the curse of dimensionality will cause the AWC procedure to fail. That is why we want to study the case where our data is locally lying approximately on a linear subspace. We start by studying the relationship between two central quantities of the algorithm. The first is the so-called *gap coefficient*

$$q_{\mathbb{P}} := \frac{\int \mathbb{1}_{B(M_1,r)\cap B(M_2,r)} d\mathbb{P}}{\int \mathbb{1}_{B(M_1,r)\cup B(M_2,r)} d\mathbb{P}},$$

where $\mathbb{P}$ is a probability measure on $\mathbb{R}^D$ underlying our data, $r > 0$ is a bandwidth parameter that increases subsequently by a factor $b \in (1,2)$ during the procedure and M_1 and M_2 are two points in $\mathbb{R}^D$. We only need to compute it if $\|M_1 - M_2\| \leq br$. The purpose of this quotient is to measure whether there is a significant *gap* in the data between M_1 and M_2, e.g. a region with a lower density, by comparing it to the *volume coefficient*

$$q := \frac{\int \mathbb{1}_{B(M_1,r)\cap B(M_2,r)} d\lambda}{\int \mathbb{1}_{B(M_1,r)\cup B(M_2,r)} d\lambda},$$

with λ being the Lebegue measure. The volume coefficient in dimension D is a function of $s := \frac{\|M_1 - M_2\|}{r}$ and is given by [10]

$$q = q_D(s) := \left(2 \frac{\mathcal{B}\left(\frac{D+1}{2}, \frac{1}{2}\right)}{\mathcal{B}\left(1 - \frac{s^2}{4}, \frac{D+1}{2}, \frac{1}{2}\right)} - 1 \right)^{-1}, \tag{2}$$

where $\mathcal{B}(\cdot,\cdot,\cdot)$ denotes the incomplete beta function and $\mathcal{B}(\cdot,\cdot) = \mathcal{B}(1,\cdot,\cdot)$ denotes the beta function. As the dimension D increases, the volume coefficient decreases approximately exponentially in D as stated in the following Proposition. This demonstrates the curse of dimensionality, as we need at least an exponential growth in the data size w.r.t. the data dimension to guarantee a reasonable estimation of the gap coefficient, which is a necessity for the AWC algorithm.

Proposition 1. *For $0 < s < 2$, we have*

$$\frac{1}{2} \leq q_D(s) \left(\frac{\left(1 - \frac{s^2}{4}\right)^{\frac{D+1}{2}}}{\Gamma\left(\frac{1}{2}\right)\sqrt{d+1}} \right)^{-1} \leq \frac{2^{\frac{5}{2}}}{s^2}.$$

By considering locally homogeneous data lying close to a lower-dimensional submanifold of dimension d, we show in the second Lemma that the gap coefficient essentially behaves locally as for homogeneous data on a linear subspace of the same dimension. We will use this in the following to prove

theoretical guarantees for the AWC procedure. Let us start by listing all the assumptions on the distribution $\mathbb{P}$ and the tuning parameters of the algorithm that we need - these are mainly a lower bound for the reach of the manifold on which the data is concentrated, an upper bound for the size of the additional noise in terms of the size of the considered vicinity and an upper bound for the radius of the considered vicinity in terms of the reach.

Assumptions A(r_0, r_1):

- $\mathbb{P}$ is the probability distribution of a random variable of the form $X + \xi$, where X follows a density f on a manifold $\mathcal{M}$ and $\|\xi\| \leq r_\xi$
- $\mathcal{M}$ is a connected and compact d-dimensional C^2 submanifold of $\mathbb{R}^D$ without boundary
- reach$(\mathcal{M}) \geq \frac{1}{\kappa}$ for $\kappa > 0$
- $r_\xi \leq \frac{r_0}{\max\{20,5d\}}$
- $r_1 \leq \frac{1}{\max\{120,\sqrt{720d}\}\kappa}$
- $1 < b \leq \frac{b'}{(1+360\kappa^2 r_1^2)\left(1+5\frac{r_\xi}{r_0}\right)}$ for some $b' < 2$

Our assumption of bounded noise is identical to the one in the work of [4] about the cluster density tree on manifolds and is relatively weak. It can be seen as a generalization of the so-called *tubular noise* and *additive noise*, c.f. [5]. Some authors additionally require orthogonality of the noise, c.f. [25] and [30]. Moreover, note that the upper bound for b is not a very restrictive assumption, as it will always be satisfied for $1 < b \leq \frac{3}{2}$. The complexity of the AWC algorithm with respect to b is $\mathcal{O}\left(\frac{1}{\log b}\right)$, so as long as b is bounded away from 1, e.g. as long as $b' \geq \sqrt{2}$, this does not change the overall complexity.

Proposition 2. *Suppose assumptions A(r, r) are satisfied for a constant density f and M_1, M_2 are two points in the support of $\mathbb{P}$ whose distance is at most br. Then*

$$(1+\varepsilon_{\mathcal{M}})^{-1}(1+\varepsilon_\xi)^{-1} \leq \frac{q_{\mathbb{P}}}{q_d(s)} \leq (1+\varepsilon_{\mathcal{M}})(1+\varepsilon_\xi)$$

for

$$\varepsilon_{\mathcal{M}} := \frac{9600(d+1)\kappa^2 r^2}{\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}$$

and

$$\varepsilon_\xi := \frac{80(d+1)\frac{r_\xi}{r}}{\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}.$$

Let us point out that our bound on the deviation of the gap coefficient from the volume coefficient is a product of the form $(1+\mathcal{O}(\kappa^2 r^2))\left(1+\mathcal{O}\left(\frac{r_\xi}{r}\right)\right)$, as long as the intrinsic dimension d is bounded and as long as b' is bounded away from 2. The first factor takes into account the reach of the manifold, whereas the second factor only depends on the size of the noise. In particular, using a manifold denoising algorithm [16, 18, 38, 30], we can preprocess our data in order to reduce noise and expect the second factor to be irrelevant. Thus, it might also be reasonable to study a setup without noise as in the following trivial Corollary.

Corollary 1. *Suppose $r_\xi = 0$ in addition to the assumptions of Proposition 2. Then*

$$(1+\varepsilon_{\mathcal{M}})^{-1} \leq \frac{q_{\mathbb{P}}}{q_d(s)} \leq 1+\varepsilon_{\mathcal{M}}.$$

Recall that the main idea of the AWC algorithm is to distinguish a homogeneous area from a gap between two clusters by estimating and comparing the gap coefficient with the volume coefficient. However, due to the non-linear manifold structure as well as the noise, we cannot establish a strict inequality between the two quantities even for the uniform case. Nevertheless, Proposition 2 guarantees a strict inequality for the homogeneous case if we adjust the volume coefficient by a factor $(1+\varepsilon_{\mathcal{M}})^{-1}(1+\varepsilon_\xi)^{-1}$. Consequently, we will adjust the proposed test of the AWC procedure to

$$T_{ij}^{(k)} := N_{i\vee j}^{(k)} \mathcal{K}\left(\widetilde{\theta}_{ij}^{(k)}, \mathfrak{q}_{ij}^{(k)}\right)\left\{\mathbb{1}\left(\widetilde{\theta}_{ij}^{(k)} < \mathfrak{q}_{ij}^{(k)}\right) - \mathbb{1}\left(\widetilde{\theta}_{ij}^{(k)} \geq \mathfrak{q}_{ij}^{(k)}\right)\right\}$$

by considering an *adjusted volume coefficient*

$$\mathfrak{q}_{ij}^{(k)} := (1+\varepsilon_{\mathcal{M}})^{-1}(1+\varepsilon_\xi)^{-1} q_d\left(\frac{\|X_i - X_j\|}{h_{k-1}}\right).$$

Note that in practice, the parameters d, $\frac{1}{\kappa}$ and r_ξ are unknown. We refer to [20] for an overview of procedures dedicated to estimating the intrinsic dimension. The estimation of the noise is related to the estimation of the manifold and is particularly related to the problem of recovering the projections of the data onto the manifold, see [30]. The estimation of the reach has been studied in [1]. However, the effect of the reach is locally small and can be ignored. Similarly, using a manifold denoising algorithm, we can assume the effect of the noise to be insignificant. In contrast, the effective dimension parameter is crucial for the computation of the test statistic. Following the proofs of theorems 1 and 2, we see that the AWC procedure is still consistent in case of overestimation of d as long as the gap is significant enough. However, we cannot expect the algorithm to be rate optimal in this case. In subsection 3.4 we discuss a simple numerical example, that suggests that the procedure might be stable in practice w.r.t. to over- and underestimation of d.

2.2 Propagation in the uniform case

In the following, we generalize the results from [10] to our considered setup. As expected, the adjusted AWC algorithm consistently propagates homogeneous areas of our data: If the threshold λ of our likelihood-ratio test is of the form $C\log n$, then the accuracy in estimating the weights of the adjacency matrix is of order $1 - \mathcal{O}\left(n^{-(C-3)}\right)$.

Theorem 1. *With high probability, the AWC algorithm does not detect a gap between two points from a distribution that is nearly uniform on a manifold, as long as it did not detect any gaps in the previous step. To be precise, suppose assumptions A(h_{k-1}, h_{k-1}) hold and $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$. We consider a constant density f and assume that the AWC algorithm did not detect any gaps in the previous step. If we choose the threshold $\lambda = C\log n$ for some $C > 0$, then*

$$\mathbb{P}^{\otimes n}\left(T_{ij}^{(k)} > C\log n \middle| \|X_i - X_j\| \leq h_k\right) \leq 2n^{\breve{}C}.$$

Corollary 2. *With high probability, the AWC algorithm does not detect any gaps if our data distribution is close to a uniform distribution on a submanifold of $\mathbb{R}^D$. To be precise, suppose assumptions A(h_0, h_K) hold and $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$. We consider a constant density f. If $K < n$ and we choose the threshold $\lambda = C\log n$ for $C > 3$, then*

$$\mathbb{P}^{\otimes n}\left(w_{ij}^{(K)} = \mathbb{1}(\|X_i - X_j\| \leq h_K)\forall i,j\right) \geq 1 - 2n^{-(C-3)}.$$

Remark 1. *By symmetricity, a linear density also satisfies the no gap condition in the full dimensional case $d = D$. So up to the constants in the terms $\varepsilon_{\mathcal{M}}$ and ε_ξ, Proposition 2 is still valid if the underlying density is of the form $f = \mathbb{1}(\mathcal{M})\mathfrak{f}$ for a linear function $\mathfrak{f}$ on $\mathbb{R}^D$. Consequently, the above results on propagation in the uniform case can also be generalized to this linear model.*

2.3 Separation in the gap case

For the case of a significant gap in the data, we can also generalize the results of [10] to the manifold setup and show that we consistently separate the data achieving nearly rate-optimality. In addition, we give a finite sample guarantee. Together with the previous results for the homogeneous case, this yields a first theoretically justified proposal to choose the parameter λ. Moreover, we do not only generalize from a linear to a smooth subspace structure of our data but also significantly generalize the definition of the considered clusters.

Assumptions B(r):

- First of all, we include assumptions A(r, r)
- Additionally, we consider disjoint subsets $\mathcal{C}_1, \dots, \mathcal{C}_{k_\mathcal{C}}$ of $\mathcal{M}$
- Spatial separation of clusters is ensured by

$$d_\infty(\mathcal{C}_l, \mathcal{C}_m) := \min_{x \in \mathcal{C}_l, y \in \mathcal{C}_m} \|x - y\| \geq r + 2r_\xi \quad \text{for} \quad 1 \leq l \neq m \leq k_\mathcal{C}$$

- Similarly as in [32], we assume a thickness condition on each cluster: We assume there is a constant $f_0 > 0$ s.t. for any $x \in \mathcal{C}_l$ and $r' \in [r - 2r_\xi, r + 2r_\xi]$ we have

$$\int f \mathbb{1}_{B(x,r')} \geq f_0 \int \mathbb{1}_{B(x,r') \cap \mathcal{M}}$$

- Separation of clusters is also ensured by a significant depth of the gap: For $x_1 \in \mathcal{C}_l, x_2 \in \mathcal{C}_m$, $r' \in [r - 2r_\xi, r + 2r_\xi]$ with $l \neq m$ and $\|x_1 - x_2\| \leq br$ we have

$$\int f \mathbb{1}_{B(x_1,r') \cap B(x_2,r')} \leq (1 - \epsilon) f_0 \int \mathbb{1}_{B(x_1,r') \cap B(x_2,r') \cap \mathcal{M}}$$

- The sample size n has to be large enough, i.e. for some $\beta > 0$ we have

$$\frac{n}{\log n} \geq \frac{2\beta}{z_k^2}$$

  where $f_0^{-1} z_k$ denotes the volume of a d-dimensional ball of radius $\frac{7}{8}r$

- The depth $\epsilon < 1$ of must be significant w.r.t. the effect of curvature and noise, and decreases not faster than $(\log n)^{\frac{1}{2}} n^{-\frac{1}{2}}$, i.e. it satisfies the lower bound

$$\epsilon \geq \max \left\{ 7(\varepsilon_{\mathcal{M}} + \varepsilon_\xi + \varepsilon_{\mathcal{M}} \varepsilon_\xi), \sqrt{\frac{2\alpha \log n}{z_k q_d^2(b) n}} \right\}$$

  for some $\alpha > \beta$.

The integral conditions are up to a change of constants a generalization of the simpler separation condition

$$\text{ess sup}_{\mathcal{M}\setminus\cup_i\mathcal{C}_i} f \leq (1-\epsilon)\inf_{\cup_i\mathcal{C}_i} f$$

from [8]. However, the here introduced generalization allows for both smooth f as well as a step function. Moreover, the upper bound on the size of the bounded noise $\epsilon \gtrsim \frac{r_\xi}{r}d$ also appears in the work of [4] (with parameters (θ,σ) instead of (r_ξ, r)).

The assumptions above are designed to be comparable to the framework of other density-based methods. However, AWC does not reconstruct connected superlevel sets of the unterlying density. Conversely, other procedures will in general not find a cluster structure respecting the idea of significant gaps. Moreoever, theoretical guarantees for AWC are only given for local clusters. In general, it is difficult to assign a global partition of the data from this information, as the local clusters might form connected components that are heavily overlapping. This limits the comparability of the presented results to a local level.

Theorem 2. *We consider a distribution on the vicinity of a submanifold of $\mathbb{R}^D$ containing different clusters separated by significant gaps in the density. As long as the AWC algorithm did not detect gaps in the previous step, it will detect the gap between two points from different clusters with high probability. To be precise, consider the assumptions B(h_{k-1}) and $X_1, X_2, \dots, X_{n+2} \overset{i.i.d.}{\sim} \mathbb{P}$. Suppose that the algorithm did not detect any gaps in the previous steps. Then*

$$\mathbb{P}^{\otimes(n+2)}\left(T_{ij}^{(k)} \geq \left(\sqrt{\alpha}-\sqrt{\beta}\right)^2 \log n \Big|_{\exists l\neq m: X_i\in\mathcal{C}_l^{r_\xi}, X_j\in\mathcal{C}_m^{r_\xi}}^{\|X_i-X_j\|\leq h_k}\right) \geq 1-3n^{-\beta}.$$

Remark 2. *Under the previous assumptions, the gap will be consistently detected at the step k where the considered vicinity first exceeds the width of the gap. However, as in the homogeneous case, the speed of convergence depends on the choice of the tuning parameter λ. Theorems 1 and 2 suggest choosing a threshold of the form $\lambda = C\log n$. Moreover, the optimal constant C^* that yields the fastest convergence $w_{ij}^{(k)} \to w_{ij}$ in probability for both discussed cases according to the given lower bounds for the accuracy of the estimation of the weights is given by*

$$\begin{aligned} C^* &= \sup_{\beta\in(0,\alpha)} \min\left\{\left(\sqrt{\alpha}-\sqrt{\beta}\right)^2, \beta\right\} \\ &= \frac{\alpha}{4}. \end{aligned}$$

The corresponding rate of misclassification is for both cases

$$\mathbb{P}^{\otimes n}\left(w_{ij}^{(k)} \neq w_{ij}\right) \leq \mathcal{O}(n^{-\frac{\alpha}{4}}).$$

Remark 3. *We consider a low manifold dimension d as a reasonable assumption and thus consider only asymptotics in n while d is bounded from above. While the rate of the algorithm is essentially (i.e. up the involved constants) independent of d, we have the following dependencies on d:*

- *To guarantee a fixed level of uncertainty, i.e. with fixed β, the lower bound on the sample size n in the list of assumptions increases exponentially in d, demonstrating the curse of dimensionality if the manifold dimension is very large.*
- *For larger d we allow a smaller level of noise $\propto d^{-1}$ and a smaller size of the considered vicinity $\propto d^{-\frac{1}{2}}$.*

2.4 Boundary case

In the previous subsection 2.2 we considered a homogeneous distribution on the manifold. In the presence of non-trivial clusters, this assumption can only be satisfied locally and only for points far enough from the boundaries of the clusters. However, the no gap condition enjoys the remarkable property that is still valid for points close to a locally linear boundary. In fact, the corresponding gap coefficient might only be larger than in the homogeneous case.

Lemma 1. *We assume $M_1 \neq M_2 \in \mathbb{R}^D$ and $r_1, r_2 > 0$. Moreover, suppose that $\mathcal{H}$ is a D-dimensional half-space containing M_1 and M_2. Then*

$$\frac{\lambda(B(M_1,r) \cap B(M_2,r_2))}{\lambda(B(M_1,r) \cup B(M_2,r_2))} \leq \frac{\lambda(\mathcal{H} \cap B(M_1,r) \cap B(M_2,r_2))}{\lambda(\mathcal{H} \cap (B(M_1,r) \cup B(M_2,r_2)))}$$

The proof of Lemma 1 relies on the following result via Fubini's theorem. Again, we assume $D > 0$ and denote the D-dimensional Lebesgue measure by λ.

Lemma 2. *Suppose $M_1 \neq M_2 \in \mathbb{R}^{D+1}$ and $r > 0$. We consider a hyperplane $H \subset \mathbb{R}^{D+1}$ containing $\frac{M_1+M_2}{2}$. Suppose v is vector of norm 1 that is orthogonal to H. Moreover, we define $t_{max} := \sup\{t : (H+tv) \cap (B(M_1,r) \cup B(M_2,r)) \neq \{\}\}$. Then the function $\mathcal{Q} : [0, t_{max}) \to \mathbb{R}_{\geq 0}$,*

$$\mathcal{Q}(t) := \frac{\lambda((H+tv) \cap B(M_1,r) \cap B(M_2,r))}{\lambda((H+tv) \cap (B(M_1,r) \cup B(M_2,r)))}$$

is monotonely decreasing in t.

The quantity $\mathcal{Q}$ in the result above is a generalization of the volume coefficient in a lower dimension: The intersection of the considered hyperplane with each ball is again a ball of a lower dimension - however, the corresponding radii are in general not identical.

Lemma 2 shows in fact more than what is claimed in Lemma 1: As we move the two center points closer to the linear boundary, the volume coefficient starts increasing monotonely as soon as the two balls are not completely contained by the half-space anymore. At some point, the volume coefficient attains its maximum, after which it decreases monotonely. By symmetricity, the volume coefficient has the same value again as in the homogeneous case, when the boundary of the half-space contains $\frac{M_1+M_2}{2}$. If we consider a stepfunction

$$f \propto \mathbb{1}(\mathcal{H} \cap (B(M_1,r) \cup B(M_2,r))) + (1-\epsilon)\mathbb{1}(\mathcal{H}^C \cap (B(M_1,r) \cup B(M_2,r))) \tag{3}$$

as a generalization of the uniform density considered in Lemma 1, we observe the analogue monotonicity, if we move the two center points further away from the half-space $\mathcal{H}$, c.f. Figure 6.

Lemma 1 allows to extend the lower bound of Proposition 2 to the boundary case under an almost identical set of assumptions with an additional cluster structure.

Assumptions C(r):

- First of all, we consider assumptions $A(r,r)$
- Additionally, we consider disjoint clusters $\mathcal{C}_1, \ldots, \mathcal{C}_{k_\mathcal{C}}$ of d_∞-distance at least $r + 2r_\xi$ as submanifolds of $\mathcal{M}$ with boundaries $\partial\mathcal{C}_i$ of reach at least $\frac{1}{\kappa'}$

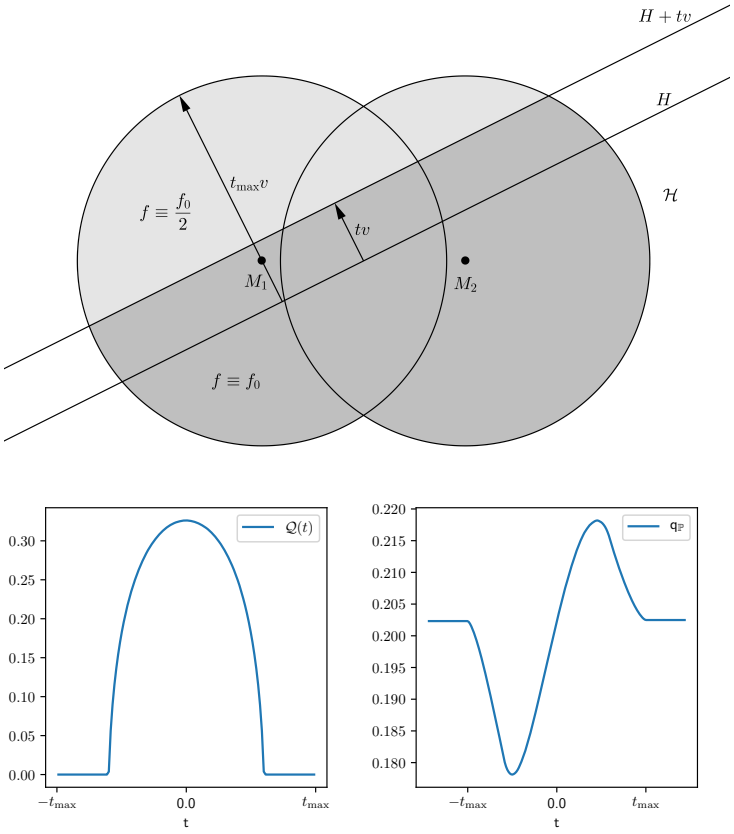


Figure 6: The top sketch illustrates the notation and relation between Lemma 1 and 2: The half-plane $H + tv$ is the boundary of the half-space $\mathcal{H}$. However, the uniform assumption of Lemma 1 is modified to a piecewise constant density as described in (3) with $\epsilon = \frac{1}{2}$. At the bottom, we see a plot of the corresponding function $\mathcal{Q}(t)$ from Lemma 2 (left) as well as the gap coefficient $q_{\mathbb{P}}$ (right). These values were obtained by Monte Carlo integration.

- The density f on $\mathcal{M}$ is constant with value c_0 on $\cup \mathcal{C}_i$ and satisfies

$$\text{ess sup}_{\mathcal{M} \setminus \cup \mathcal{C}_i} f \leq c_0$$

- Outside of the clusters we require the following regularity condition for the density: Any nontrivial intersection of a superlevel set of f with an offset $\mathcal{C}_i^{r+2r_\xi}$ is equal to the intersection of that superlevel set with a submanifold of $\mathcal{M}$ having a boundary of reach at least $\frac{1}{\kappa'}$
- $r \leq \frac{1}{132\kappa'\sqrt{d+1}}$

The last condition together with the upper bound from $A(r, r)$ ensures that both the reach of $\mathcal{M}$ and $\partial\mathcal{C}_i$ are large enough w.r.t. the radius r, such that both the manifold and the boundary of the cluster can be locally approximated by affine subspaces.

Proposition 3. *We consider assumptions $C(r)$. Suppose $M_1, M_2 \in \mathcal{C}_i$ are points of distance at most br. Then*

$$q_{\mathbb{P}} \geq q_d(s) (1 + \epsilon_{\mathcal{M}})^{-1} (1 + \epsilon_\xi)^{-1} (1 + \epsilon_{\partial\mathcal{C}})^{-1},$$

for

$$\epsilon_{\mathcal{M}} = \frac{45360(d+1)\kappa^2 r^2}{\left(1-\frac{b'^2}{4}\right)^{\frac{d+1}{2}}}$$
$$\epsilon_{\xi} = \frac{264(d+1)\frac{r_\xi}{r}}{\left(1-\frac{b'^2}{4}\right)^{\frac{d+1}{2}}}$$
$$\epsilon_{\partial\mathcal{C}} = 132\kappa' r\sqrt{d+1}$$

This inequality is stronger than the lower bound from Proposition 2. Hence, we have to modify the definition of the *adjusted volume coefficient*. For the following, we consider

$$\mathfrak{q}_{ij}^{(k)} := (1+\epsilon_{\mathcal{M}})^{-1}(1+\epsilon_{\xi})^{-1}\,(1+\epsilon_{\partial\mathcal{C}})^{-1}\,q_d\left(\frac{\|X_i - X_j\|}{h_{k-1}}\right)$$

to allow for consistent propagation in the boundary case as stated in the following Theorem. Again, in practice, the implementation of the adjusted volume coefficient might be ignored, c.f. [10]. However, it is important to not underestimate the dimension parameter d. In fact, an overestimation of d might compensate for dropping the first three factors of the adjusted volume coefficient and ensure the propagation of homogeneous areas.

Theorem 3. *We consider a distribution in the vicinity of a manifold and two points inside a homogenous cluster. Then with high probability, the AWC algorithm will not detect a gap between them, even if the points happen to be in close proximity to the boundary of the cluster. To be precise, suppose assumptions C(h_{k-1}) hold and $X_1, X_2, \dots, X_{n+2} \overset{i.i.d.}{\sim} \mathbb{P}$. We assume that the AWC algorithm did not detect any gaps in the previous step. If we choose the threshold $\lambda = C\log n$ for some $C > 0$, then*

$$\mathbb{P}^{\otimes(n+2)}\left(T_{ij}^{(k)} > C\log n \middle| X_i, X_j \in \mathcal{C}^{r_\xi}, \|X_i - X_j\| \le h_k\right) \le 2n^{\breve{}C}.$$

Together with Theorem 2 we are able to cover all the discussed cases at once. In the following corollary, we will use the term *global clusters* to describe the disjoint offsets $\mathcal{C}_i^{r_\xi}$.

Corollary 3. *We consider the conditions $C(h_{k-1})$ and $B(h_{k-1})$ with a slightly stricter lower bound*

$$\varepsilon \ge 7\,(1+\epsilon_{\mathcal{M}})\,(1+\epsilon_{\xi})\,(1+\epsilon_{\partial\mathcal{C}}) - 7.$$

Suppose $X_1, \dots, X_{n+2} \overset{i.i.d.}{\sim} \mathbb{P}$. We assume that the AWC algorithm did not detect any gaps in the previous step. Moreover, we choose the threshold $\lambda = \frac{\alpha}{4}\log n$. Then with probability at least $1 - 3n^{-\frac{\alpha-8}{4}}$, every local cluster $\mathcal{C}_i^{(k)}$ calculated by AWC at step k satisfies the following: If X_i belongs to a global cluster, $\mathcal{C}_i^{(k)}$ contains all points from this cluster of distance at most h_k to X_i, while it does not contain any points from other global clusters.

2.5 Optimality

The lack of a rigorous global cluster objective makes it difficult to compare our theoretical results to previous work. Moreover, we have shown that the algorithm differs significantly from other density-based methods, c.f. Figure 5. However, the local separation considered in Theorem 2 as well as Corollary 3 is very similar to the split of two components in the cluster density tree. Consistent and

rate-optimal estimation of the cluster density tree using a single-linkage clustering algorithm has been established in[8]. Using different notation (i.e. σ instead of r as width of the gap and λ instead of f_0 as density level), the authors show that the optimal rate is (up to logarithmic factors and factors dependent on d) given by

$$\epsilon \gtrsim \sqrt{\frac{1}{nr^d f_0}}$$

In [4] this has been extended to the manifold setup. Further work by [37] shows that, under the assumption of a Hölder smooth density, this rate can be described by only one separation parameter together with the smoothness parameter.

In view of $z_k \propto f_0 r^d$, our lower bound on the depth of the gap

$$\epsilon \geq \sqrt{\frac{2\alpha \log n}{z_k q_d^2(b) n}}$$

achieves in fact the optimal rate given above w.r.t. (n, r, f_0). We verify the optimality w.r.t. n for our setup under very simple conditions, showing that no algorithm can consistently detect the gap if ϵ decreases at the rate $n^{-\frac{1}{2}}$.

Assumptions D:

- $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are disjoint subsets of a manifold $\mathcal{M} \subset \mathbb{R}^D$
- $X_1, \ldots, X_n$ are drawn i.i.d. from a density supported on $\mathcal{M}$ that is constant on $V := \cup \mathcal{C}_i$ with value f_V and constant on $G := \mathcal{M} \setminus V$ with value f_G

Theorem 4. *Let assumptions D be satisfied. We consider the null hypothesis of a uniform distribution on the manifold, i.e.*

$$H_0 : f_G = f_V$$

against the alternative

$$H_1 : f_G = (1 - \delta) f_V$$

for $\delta > 0$. Then no test can separate the two cases consistently if $n\delta^2 \nrightarrow \infty$ as $n \to \infty$.

3 Experimental Results

Although manifold models are considered to be realistic, we still impose some assumptions for our theoretical study that are usually not satisfied in real-life. Most importantly we assume that our data lies on a manifold without boundary and positive reach up to bounded noise. A comprehensive numerical study of the procedure including real-life data by [10] suggested that these assumptions are not necessary in practice and the performance of the algorithm is competitive with state-of-the-art algorithms. Rather, the limiting factor of the algorithm for clustering so-called big data at a global scale seems to be its polynomial complexity. That being said, in this work, we will restrict to some rather simple artificial examples in order to illustrate and verify our theoretical results.

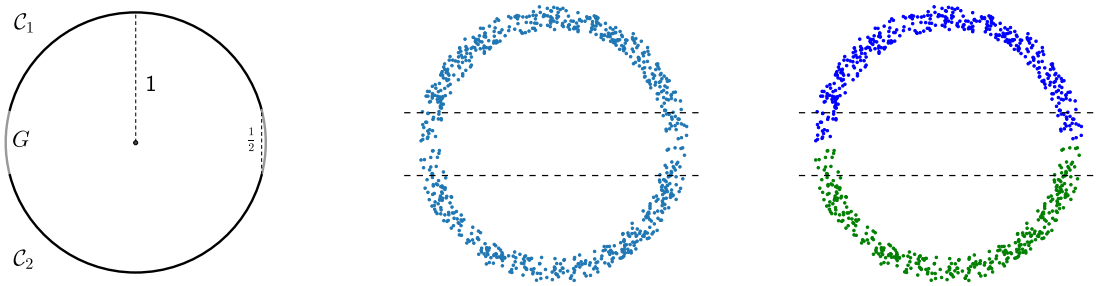


Figure 7: Density f_ϵ (left), i.i.d. sample of size $n = 800$ from $\mathbb{P}_{\frac{1}{2}}^{\mathcal{U}(\frac{1}{10})}$ with two dashed lines highlighting the gap in the data (center) and clusters obtained via AWC (right)

3.1 Consistency

In order to verify the sensitivity of the AWC algorithm w.r.t. local gaps for data lying on non-linear submanifolds and illustrate the main results Theorem 1 and Theorem 2, we will start by studying an artificial example where the embedding dimension is equal to 2 and the intrinsic dimension of the data is 1. We consider a distribution on the vacinity of the unit circle S^1 in $\mathbb{R}^2$ with two clusters

$$\mathcal{C}_1 := \{(x, y) \in S^1 : y > \frac{1}{4}\}$$

and

$$\mathcal{C}_2 := \{(x, y) \in S^1 : y < -\frac{1}{4}\}.$$

By $\mathbb{P}_\epsilon$ we denote the distribution corresponding to the density

$$f_\epsilon := \frac{1}{2\pi}\left(\mathbb{1}_{\mathcal{C}_1\cup\mathcal{C}_2} + (1-\epsilon)\mathbb{1}_{S^1\setminus(\mathcal{C}_1\cup\mathcal{C}_2)}\right).$$

Moreover, by $\mathcal{U}(r)$ we denote the uniform distribution on a 2-dimensional ball of radius r. Then we sample $X_1, \ldots, X_n$ i.i.d. from

$$\mathbb{P}_\epsilon^{\mathcal{U}(\frac{1}{10})} := \mathbb{P}_\epsilon * \mathcal{U}\left(\frac{1}{10}\right),$$

cf. Figure 7. To measure the performance of the algorithm we use a modified version of the Rand index [31]

$$\left(\sum_{\substack{(X_i,X_j)\in(\mathcal{C}_1\cup\mathcal{C}_2)^2\\ 0<||X_i-X_j||<h_K}} 1\right)^{-1}\left(\sum_{\substack{X_i,X_j\in\mathcal{C}_1\\ X_i,X_j\in\mathcal{C}_2\\ 0<||X_i-X_j||<h_K}} w_{ij}^{(K)} + \sum_{\substack{X_i\in\mathcal{C}_1,X_j\in\mathcal{C}_2\\ X_i\in\mathcal{C}_2,X_j\in\mathcal{C}_1\\ ||X_i-X_j||<h_K}}\left(1-w_{ij}^{(K)}\right)\right).$$

For simplicity, we refer to this measure as Rand index. It can also be defined as the accuracy of a subset of the weights $(w_{ij}^{(K)})_{i,j=1}^n$. As our theoretical results only apply at a local scale, we also restrict here to a local scale $h_K = 1$ and fix a series of bandwidths $h_i = 2^{\frac{i}{2}-2}$, $i = 0, \ldots, 4$. We only adjust the gap coefficient with respect to the intrinsic dimension, that is, we assume the reach and the noise magnitude to be zero in the computation of the adjusted volume coefficient. For each sample,

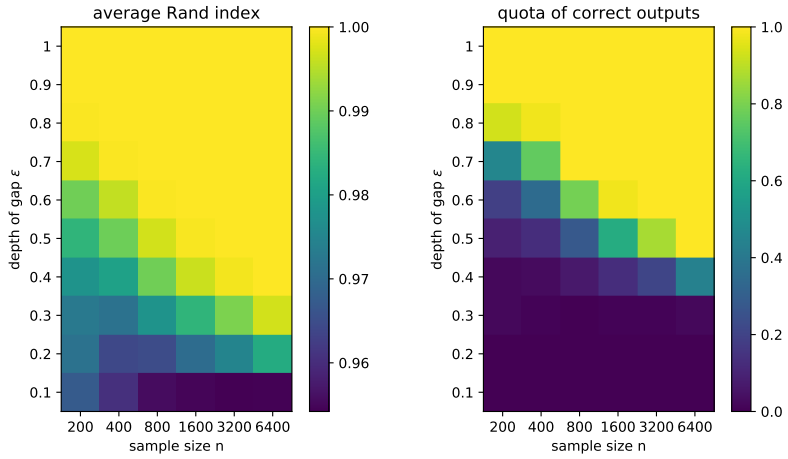


Figure 8: Average rand index (left) and quota of experiments yielding a rand index 1 (right)

we run the algorithm for different λ and consider only the best resulting Rand index, i.e. we overfit λ. Finally, for different values of ϵ, we repeat the experiment 100 times. The resulting average rand index is plotted in Figure 8 on the left. Note that the Rand index is in general quite close to 1, however, this is only due to the imbalance in the considered classification problem. For the evaluation of the results, we are only interested in the relatively large values, e.g. ≥ 0.99. On the right, the quota of experiments is plotted where a rand index of 1 is achieved. This relates to our theoretical results, whereas the average rand index is a more common measure in practice. Our theoretical results show, that the minimal ϵ, for which we can reconstruct the cluster structure with high probability, is up to logarithmic factors of order $\sqrt{\frac{1}{n}}$. The experiment is not exhaustive enough to verify this result. However, the results verify the asymptotics $\epsilon \xrightarrow{n\to\infty} 0$ and indicate that ϵ decreases significantly slower than $\frac{1}{n}$.

A less expected detail in the plot is the fact, that for small values of the depth ϵ, we observe better Rand indices as the sample size n decreases. This can be explained as follows. If ϵ is small, our distribution is very close to a distribution without a gap. Thus, for large n, the empirical distribution will also be close to a uniform distribution, and it will be very difficult for the algorithm to detect the clusters. However, for small n, the distribution may deviate more from the uniform distribution and form random clusters that in some cases do accidentally have similarities to the true cluster structure.

3.2 Scaling of sensitivity parameter λ

In the experiment above, we also computed for each experiment the minimal value of λ that achieved the largest rand index and plotted the resulting average in Figure 9. The results support our proposition that λ should be scaled logarithmically w.r.t. the data size.

3.3 High-dimensional data

In this subsection, we study the effect of the embedding dimension, i.e. the effect of high-dimensional noise. Recall that the presented results are independent of the embedding dimension D of the data. However, as we assume the norm of the noise to be bounded. In the case of centered noise with i.i.d. coordinates this implies that for each coordinate the variance is of order $\mathcal{O}(D^{-1})$. This motivates the study of two different noise distributions. Firstly and corresponding to our theoretical results, we

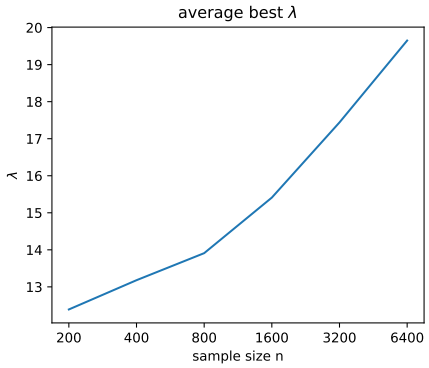


Figure 9: Average minimal lambda with best rand index for $\epsilon = 0.9$

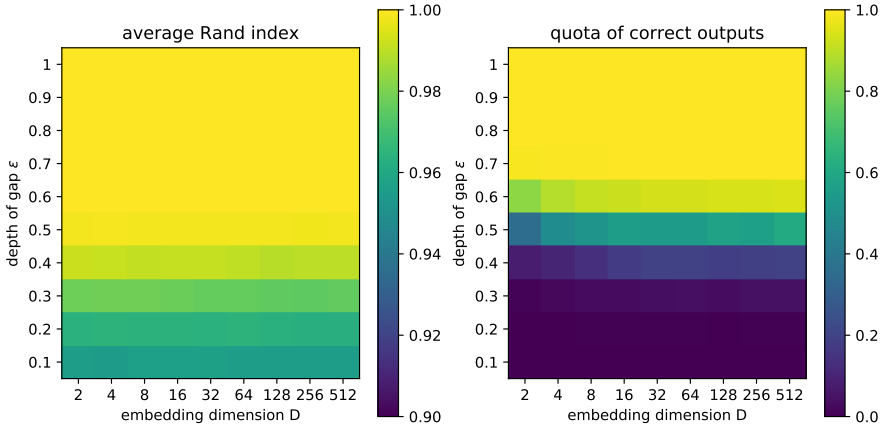


Figure 10: Average rand index (left) and quota of experiments yielding a rand index 1 (right) for uniform noise of norm $\leq \frac{1}{10}$

consider the uniform distribution $\mathcal{U}(r)$ on a centered D -dimensional ball of radius r. Also we want to consider the centered multivariate normal distribution $\mathcal{N}(\sigma^2)$ with covariance matrix $\sigma^2 I_D$. Note that for large $D, \mathcal{N}(\sigma^2)$ is concentrated on a thin annulus around the centered sphere of radius $\sigma\sqrt{D}$, so the two noise distributions mainly differ in the parametrization of the scale.

By $\mathbb{P}_{D,\epsilon}$ we denote an D -dimensional embedding of the distribution $\mathbb{P}_\epsilon$ described in subsection 3.1. Then we draw our sample $X_1, \ldots, X_n$ i.i.d. either from

$$\mathbb{P}_{D,\epsilon}^{\mathcal{U}\left(\frac{1}{10}\right)} := \mathbb{P}_{D,\epsilon} * \mathcal{U}\left(\frac{1}{10}\right)$$

or

$$\mathbb{P}_{D,\epsilon}^{\mathcal{N}\left(\frac{1}{3200}\right)} := \mathbb{P}_{D,\epsilon} * \mathcal{N}\left(\frac{1}{3200}\right).$$

Note that the distribution $\mathbb{P}_\epsilon^{\mathcal{U}\left(\frac{1}{10}\right)}$ used in the above experiments is a special case of $\mathbb{P}_{D,\epsilon}^{\mathcal{U}\left(\frac{1}{10}\right)}$ for $D = 2$. Moreover, for $D = 32$, both distributions concentrate on the proximity of a centered sphere of radius $\frac{1}{10}$. Thus we might expect similar performance of the algorithm for both distributions for $D = 32$. According to our results, the performance should not break down in the uniform case for large D while we expect the performance to decrease with growing embedding dimension for the Gaussian noise as the noise radius increases.

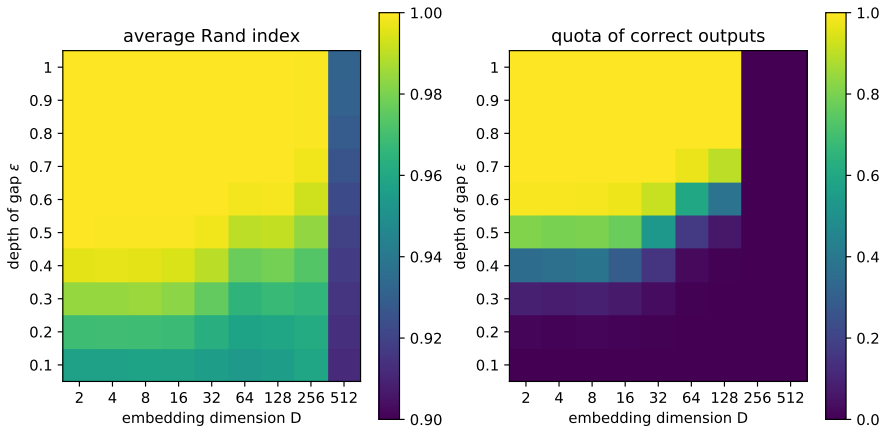


Figure 11: Average rand index (left) and quota of experiments yielding a rand index 1 (right) for Gaussian noise of variance $\frac{1}{3200} I_D$

We fix the sample size $n = 1000$ and proceed otherwise analogously to the first experiment: For each sample, we optimize λ and repeat the experiment 1000 times for each value of ϵ. The resulting average rand indices, as well as the quota of experiments with rand index equal to 1, are presented in Figures 10 and 11 and confirm our expectations. We observe one interesting detail in the quota of correct outputs in the presence of uniform noise on the right plot in Figure 10. For a very small embedding dimension D the performance is slightly worse. A possible explanation is that the high-dimensional noise approximately preserves distances up to a constant summand with large probability. So in this experiment, the separation of the two clusters might be more difficult under smaller embedding dimension D.

3.4 Effect of intrinsic dimension parameter d

Our theoretical results require knowledge of the parameter d of the effective dimension of the data. Otherwise, we cannot expect consistency under the asymptotics $\epsilon \to 0$. In practical applications, the dimension parameter is often unknown and can be estimated [20]. However, under the reasonable assumption that d is not too large, e.g. $d \leq 5$, we can also just run the clustering procedure for the different values of d. In both cases, uncertainty about the true intrinsic dimension remains. Unfortunately, our theoretical study does not provide much insight into the stability of the algorithm with respect to the dimension parameter.

In order to observe the effect of both under- and overstimation of the dimension parameter, we will consider the following simple 2-dimensional example. We consider a distribution on the unit sphere S^2 in $\mathbb{R}^3$ with two clusters

$$\mathcal{C}_1 := \{(x, y, z) \in S^2 : z > \frac{1}{4}\}$$

and

$$\mathcal{C}_2 := \{(x, y, z) \in S^2 : z < -\frac{1}{4}\}.$$

We sample $X_1, \dots, X_n$ i.i.d. from the distribution $\mathbb{P}_\epsilon$ corresponding to the density

$$f_\epsilon \propto \mathbb{1}_{\mathcal{C}_1 \cup \mathcal{C}_2} + (1 - \epsilon)\mathbb{1}_{S^2 \setminus (\mathcal{C}_1 \cup \mathcal{C}_2)},$$

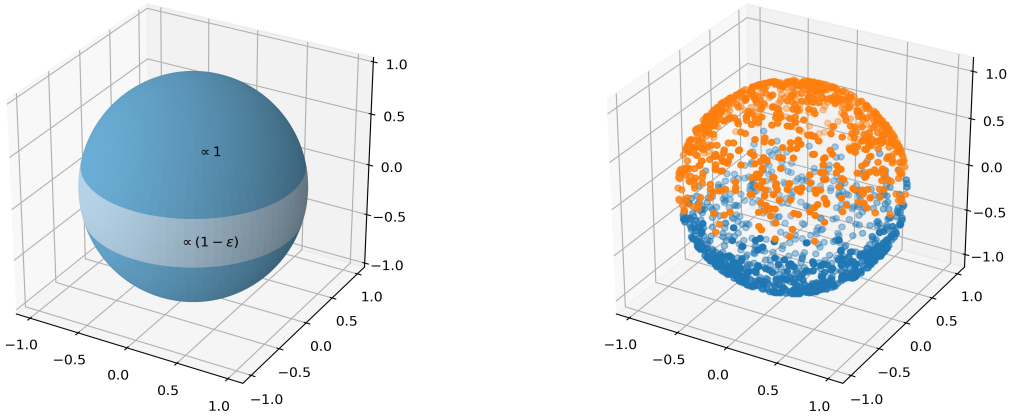


Figure 12: Left: Sketch of density f_ϵ. Right: Obtained clustering from AWC with parameters $d = 1$ and $\lambda = 50$ for a sample of size $n = 1000$ and depth $\epsilon = \frac{2}{3}$

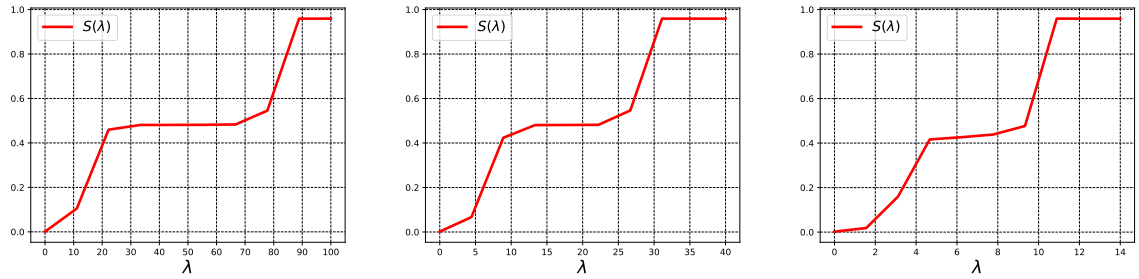


Figure 13: Sum of weights heuristics for the sample in Figure 12 with parameters $d = 1$ (left), $d = 2$ (middle) and $d = 3$ (right)

cf. Figure 12. For a sample of size $n = 1000$ with depth $\epsilon = \frac{2}{3}$, we consider various parameter λ and plot the corresponding sum of weight heuristic S, i.e. the normalized sum of all weigths obtained at the final step of the AWC procedure. This statistic is a possible way to tune λ in practice. One might simply take λ at a plateau of the graph of S, as it is expected for a clear cluster structure that the output of the algorithm is stable with respect to the tuning parameter. The results are shown in figures 12 and 13.

In figure 13 we see for each dimension parameter $d = 1, 2, 3$ a unique plateau at a value around 0.5. The value $S(\lambda) = 0.5$ corresponds to two clusters of equal size. Indeed a plot for the parameters $(d = 1, \lambda = 50)$ in figure 12 verifies that the cluster structure is detected as expected. We ommited plots for $(d = 2, \lambda = 20)$ and $(d = 3, \lambda = 8)$, as the results are nearly identical. Moreover, we observe that the scaling of λ depends on d. A larger dimension parameter requires smaller λ. This can be explained by the fact the the corresponding volume coefficient decreases with an increase of the dimension parameter. So it is harder for the algorithm to detect gaps, while the propagation effect is even stronger. A smaller λ compensates this effect.

The experiment suggests that the AWC procedure is able to detect the cluster structure even if the effective dimension parameter d is over- or underestimated. However, the scaling of λ depends on the choice of d.

4 Proofs

Proof of Proposition 1. The main tool for the bounds will the series representation

$$\mathcal{B}(x,a,b) = x^a \sum_{n=0}^{\infty} \frac{\Gamma(1-b+n)}{\Gamma(1-b)\Gamma(n+1)(a+n)} x^n$$

for the incomplete beta function [27]. Also, we use the logarithmic convexity of the gamma function. For the upper bound we get

$$\begin{aligned}
q_d(t) &= \frac{\mathcal{B}\left(1-\frac{t^2}{4}, \frac{d+1}{2}, \frac{1}{2}\right)}{2\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right) - \mathcal{B}\left(1-\frac{t^2}{4}, \frac{d+1}{2}, \frac{1}{2}\right)} \\
&\leq \frac{\mathcal{B}\left(1-\frac{t^2}{4}, \frac{d+1}{2}, \frac{1}{2}\right)}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} \\
&\leq \frac{\frac{2}{d+1} \sum\limits_{n=0}^{\infty} \left(1-\frac{t^2}{4}\right)^{\frac{d+1}{2}+n}}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} \\
&= \frac{\frac{2}{d+1} \left(1-\frac{t^2}{4}\right)^{\frac{d+1}{2}} \Gamma\left(\frac{d+2}{2}\right)}{\frac{t^2}{4}\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \\
&\leq \frac{\frac{2}{d+1} \left(1-\frac{t^2}{4}\right)^{\frac{d+1}{2}} \Gamma^{\frac{1}{2}}\left(\frac{d+3}{2}\right)}{\frac{t^2}{4}\Gamma^{\frac{1}{2}}\left(\frac{d+1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \\
&= 2^{\frac{5}{2}} t^{-2} \frac{\left(1-\frac{t^2}{4}\right)^{\frac{d+1}{2}}}{(d+1)^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}
\end{aligned}$$

and similarly, we compute the lower bound

$$\begin{aligned}
q_d(t) &\geq \frac{\mathcal{B}\left(1-\frac{t^2}{4}, \frac{d+1}{2}, \frac{1}{2}\right)}{2\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} \\
&\geq \frac{\left(1-\frac{t^2}{4}\right)^{\frac{d+1}{2}}}{(d+1)\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} \\
&= \frac{\left(1-\frac{t^2}{4}\right)^{\frac{d+1}{2}} \Gamma\left(\frac{d+2}{2}\right)}{(d+1)\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \\
&\geq \frac{\left(1-\frac{t^2}{4}\right)^{\frac{d+1}{2}} \Gamma^{\frac{1}{2}}\left(\frac{d+2}{2}\right)}{(d+1)\Gamma^{\frac{1}{2}}\left(\frac{d}{2}\right)\Gamma\left(\frac{1}{2}\right)} \\
&= \frac{d^{\frac{1}{2}}\left(1-\frac{t^2}{4}\right)^{\frac{d+1}{2}}}{2^{\frac{1}{2}}(d+1)\Gamma\left(\frac{1}{2}\right)} \\
&\geq 2^{-1}\frac{\left(1-\frac{t^2}{4}\right)^{\frac{d+1}{2}}}{(d+1)^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}.
\end{aligned}$$

□

For the proof of Proposition 2 we will use the following two auxiliary Lemmas. By $\mathrm{vol}(\cdot)$ we denote the Lebesgue volume on a submanifold of $\mathbb{R}^D$. We will consider different such manifolds and not specify them explicitly, as long as it clear from the context to which manifold we refer.

Lemma 3. *For any d-dimensional C^2 submanifolds $\mathcal{M}_1, \mathcal{M}_2 \in \mathbb{R}^D$, a measurable subset $A \subset \mathcal{M}_1$ and a C-Lipschitz function $f : \mathcal{M}_1 \to \mathcal{M}_2$, we have*

$$\mathrm{vol}(f(A)) \leq C^d \mathrm{vol}(A).$$

Proof. This inequality is also valid for the d -dimensional Hausdorff measure. In this case, it is a simple consequence of the definition of the Hausdorff measure [3]. As the Lebesgue measure is related by a constant factor [14], it also holds for the Lebesgue measure. □

For the second auxiliary Lemma we consider a connected and compact C^2 submanifold $\mathcal{M} \subset \mathbb{R}^D$ with reach $\frac{1}{\kappa} > 0$ and without boundary. For some fixed $x \in \mathcal{M}$ we denote the tangent plane of $\mathcal{M}$ at x by $\mathcal{T}$. Also, we consider the projection $P : \mathbb{R}^D \to \mathcal{T}$ associating each $y \in \mathbb{R}^D$ with the closest point in $\mathcal{T}$.

Lemma 4. *Suppose $0 < r \leq \frac{1}{40\kappa}$. Then the restriction $P|_{\mathcal{M}\cap B(x,r)}$ is a 1-Lipschitz injection and its image contains $\mathcal{T} \cap B(x, r/L)$. Moreover, its inverse is L-Lipschitz for*

$$L := 1 + 40\kappa^2 r^2 \leq 1 + \kappa r.$$

Proof. This Lemma is given in [3] with some unspecified small enough constant instead of $\frac{1}{40}$. Following the corresponding proof, it can be easily verified that this constant is indeed small enough. □

Proof of Proposition 2. Let us denote the uniform measure on the manifold with μ. For $i = 1, 2$, we choose a point M_i' on the manifold $\mathcal{M}$ of distance at most r_ξ to M_i. Because the Euclidean norm of the noise ξ is bounded by r_ξ, we get

$$\begin{aligned} q_l &:= \frac{\int \mathbb{1}_{B(M_1', r-2r_\xi)\cap B(M_2', r-2r_\xi)} d\mu}{\int \mathbb{1}_{B(M_1', r+2r_\xi)\cup B(M_2', r+2r_\xi)} d\mu} \\ &\le q_{\mathbb{P}} \\ &\le \frac{\int \mathbb{1}_{B(M_1', r+2r_\xi)\cap B(M_2', r+2r_\xi)} d\mu}{\int \mathbb{1}_{B(M_1', r-2r_\xi)\cup B(M_2', r-2r_\xi)} d\mu} \\ &=: q_u \end{aligned} \tag{4}$$

Let us denote by $\square$ one of the symbols $\cap$ or $\cup$ and suppose $r' \in [r - 2r_\xi, r + 2r_\xi]$. By P we denote the orthogonal projection onto the tangent plane $\mathcal{T}$ of $\mathcal{M}$ at M_1'. Our assumptions ensure that a ball of radius $3r$ around M_1' contains both $B(M_1', r')$ and $B(M_2', r')$. Since the restriction $P|_{\mathcal{M}\cap B(M_1', 3r)}$ is an injective 1-Lipschitz map with an L -Lipschitz inverse with $L := 1 + 360\kappa^2 r^2$, we conclude (cf. [3])

$$L^{-d} \le \frac{\mathsf{vol}(P(\mathcal{M} \cap (B(M_1', r') \square\, B(M_2', r'))))}{\mathsf{vol}(\mathcal{M} \cap (B(M_1', r') \square\, B(M_2', r')))} \le 1. \tag{5}$$

Moreover, the above Lipschitz constants imply

$$\mathcal{T} \cap B\left(P(M_i'), \frac{r'}{L}\right) \subseteq P(\mathcal{M} \cap B(M_i', r')) \subseteq \mathcal{T} \cap B(P(M_i'), r')$$

for $i = 1, 2$ and therefore

$$\begin{aligned} 1 &\le \frac{\mathsf{vol}\,(\mathcal{T} \cap (B(P(M_1'), r') \square\, B(P(M_2'), r')))}{\mathsf{vol}\,(P(\mathcal{M} \cap (B(M_1', r') \square\, B(M_2', r'))))} \\ &\le \frac{\mathsf{vol}\,(\mathcal{T} \cap (B(P(M_1'), r') \square\, B(P(M_2'), r')))}{\mathsf{vol}\,\left(\mathcal{T} \cap (B(P(M_1'), \frac{r'}{L}) \square\, B(P(M_2'), \frac{r'}{L}))\right)} =: q_{\square\,, r'}. \end{aligned} \tag{6}$$

Note also that according to our assumptions, any intersections encountered so far are nonempty. From (5) and (6) we conclude

$$\begin{aligned} q_{\square\,, r'}^{-1} &\mathsf{vol}\,(\mathcal{T} \cap (B(P(M_1'), r') \square\, B(P(M_2'), r'))) \\ &\le \mathsf{vol}\,(P(\mathcal{M} \cap (B(M_1', r') \square\, B(M_2', r')))) \\ &\le \mathsf{vol}\,(\mathcal{M} \cap (B(M_1', r') \square\, B(M_2', r'))) \\ &\le L^d \mathsf{vol}\,(P(\mathcal{M} \cap (B(M_1', r') \square\, B(M_2', r')))) \\ &\le L^d \mathsf{vol}\,(\mathcal{T} \cap (B(P(M_1'), r') \square\, B(P(M_2'), r'))) \end{aligned}$$

and obtain

$$q_{\square\,, r'}^{-1} \le \frac{\mathsf{vol}\,(\mathcal{M} \cap (B(M_1', r') \square\, B(M_2', r')))}{\mathsf{vol}\,(\mathcal{T} \cap (B(P(M_1'), r') \square\, B(P(M_2'), r')))} \le L^d. \tag{7}$$

In particular, considering $(\square\,, r') = (\cap, r + 2r_\xi)$ and $(\square\,, r') = (\cup, r - 2r_\xi)$ in (7), we get

$$q_u \le q_{\cup, r-2r_\xi} L^d q_\cup q_{r+2r_\xi}, \tag{8}$$

where $q_{r'}$ is defined as

$$q_{r'} := \frac{\mathsf{vol}\,(\mathcal{T} \cap B(P(M_1'), r') \cap B(P(M_2'), r'))}{\mathsf{vol}\,(\mathcal{T} \cap (B(P(M_1'), r') \cup B(P(M_2'), r')))}$$

for $r' \in [r - 2r_\xi, r + 2r_\xi]$ and

$$q_\cup := \frac{\mathsf{vol}\left(\mathcal{T} \cap \left(B(P(M_1'), r + 2r_\xi) \cup B(P(M_2'), r + 2r_\xi)\right)\right)}{\mathsf{vol}\left(\mathcal{T} \cap \left(B(P(M_1'), r - 2r_\xi) \cup B(P(M_2'), r - 2r_\xi)\right)\right)}.$$

For the lower bound, we similarly obtain

$$q_l \geq q_{\cap, r-2r_\xi}^{-1} L^{-d} q_\cup^{-1} q_{r-2r_\xi}. \tag{9}$$

The quotient $q_{r'}$ is exactly the volume coefficient defined in (2) in dimension d at $\frac{\|P(M_1') - P(M_2')\|}{r'}$. The derivative of q_d is given by

$$q_d'(t) = -2\left(1 - \frac{t^2}{4}\right)^{\frac{d-1}{2}} \frac{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}{\left(2\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right) - \mathcal{B}\left(1 - \frac{t^2}{4}, \frac{d+1}{2}, \frac{1}{2}\right)\right)^2}.$$

Its absolute value on $[0, 2)$ is bounded from above by $\frac{2}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}$. For the following we define $s := \frac{\|M_1 - M_2\|}{r}$. Because q_d is a monotonely decreasing function on $[0, 2)$ and

$$\|P(M_1') - P(M_2')\| - 2r_\xi \leq \|M_1 - M_2\| \leq L\|P(M_1') - P(M_2')\| + 2r_\xi,$$

we have

$$\begin{aligned}
q_{r+2r_\xi} &\leq q_d\left(\frac{\max\{0, \|M_1 - M_2\| - 2r_\xi\}}{L(r + 2r_\xi)}\right) \\
&\leq q_d(s) + \frac{2}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\left(s - \frac{\|M_1 - M_2\| - 2r_\xi}{L(r + 2r_\xi)}\right) \\
&= q_d(s) + \frac{2}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\left(\frac{sr(L-1)}{L(r + 2r_\xi)} + \frac{2sr_\xi}{r + 2r_\xi} + \frac{2r_\xi}{L(r + 2r_\xi)}\right) \\
&\leq q_d(s) + \frac{1440\kappa^2 r^2}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} + \frac{12\frac{r_\xi}{r}}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} \\
&\leq q_d(s)\left(1 + \frac{1440\kappa^2 r^2}{q_d(b')\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\right)\left(1 + \frac{12\frac{r_\xi}{r}}{q_d(b')\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\right).
\end{aligned} \tag{10}$$

Similarly, we obtain

$$\begin{aligned}
q_{r-2r_\xi} &\geq q_d\left(\frac{\|M_1 - M_2\| + 2r_\xi}{r - 2r_\xi}\right) \\
&= q_d(s)\left(\frac{q_d(s)}{q_d\left(\frac{\|M_1 - M_2\| + 2r_\xi}{r - 2r_\xi}\right)}\right)^{-1} \\
&\geq q_d(s)\left(\frac{q_d\left(\frac{\|M_1 - M_2\| + 2r_\xi}{r - 2r_\xi}\right) + \frac{2}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\left(\frac{\|M_1 - M_2\| + 2r_\xi}{r - 2r_\xi} - s\right)}{q_d\left(\frac{\|M_1 - M_2\| + 2r_\xi}{r - 2r_\xi}\right)}\right)^{-1} \\
&\geq q_d(s)\left(1 + \frac{2\left(\frac{2sr_\xi}{r - 2r_\xi} + \frac{2r_\xi}{r - 2r_\xi}\right)}{q_d(b')\,\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\right)^{-1} \\
&\geq q_d(s)\left(1 + \frac{12\frac{r_\xi}{r}}{q_d(b')\,\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\right)^{-1}.
\end{aligned} \tag{11}$$

It remains to find upper bounds for $q_\cup$, $c_{\cup,r'}$ and $q_{\cap,r'}$. Firstly, note that for $x \in \mathcal{T}$, we have

$$q_{\cup,r'} \leq \frac{\mathsf{vol}\left(\mathcal{T} \cap B(x, \frac{r'}{L})\right) + 2\mathsf{vol}\left(\mathcal{T} \cap (B(x,r') \setminus B(x, \frac{r'}{L})\right)}{\mathsf{vol}\left(\mathcal{T} \cap B(x, \frac{r'}{L})\right)} = 2L^d - 1. \tag{12}$$

Analogously, using $(1+x)^d < 1 + 2xd$ for $0 \leq x \leq \frac{1}{d}$, we find

$$\begin{aligned} q_\cup &\leq \left(2\left(\frac{r+2r_\xi}{r-2r_\xi}\right)^d - 1\right) \\ &\leq \left(2\left(1+\frac{5r_\xi}{r}\right)^d - 1\right) \\ &\leq \left(1 + \frac{20dr_\xi}{r}\right) \end{aligned} \tag{13}$$

and

$$L^d\left(2L^d - 1\right) \leq 1 + 2880d\kappa^2 r^2. \tag{14}$$

Moreover, for $s' := \frac{\|P(M_1') - P(M_2')\|}{r'}$,

$$\begin{aligned} q_{\cap,r'} &= q_{\cup,r'} \frac{q_d\left(s'\right)}{q_d\left(s'L\right)} \\ &\leq (2L^d - 1) \frac{q_d(s'L) + s'(L-1)\frac{2}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}}{q_d(s'L)} \\ &\leq (2L^d - 1)\left(1 + \frac{1440\kappa^2 r^2}{q_d(b')\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\right). \end{aligned} \tag{15}$$

Finally, we derive a tractable bound for $\frac{1}{q_d(b')\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}$. Using only the first term of the series [27]

$$\mathcal{B}\left(x, a, b\right) = x^a \sum_{n=0}^{\infty} \frac{\Gamma(1-b+n)}{\Gamma(1-b)\Gamma(n+1)(a+n)} x^n,$$

we get

$$\begin{aligned} \frac{1}{q_d(b')\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} &= \frac{2\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right) - \mathcal{B}\left(1 - \left(\frac{b'}{2}\right)^2, \frac{d+1}{2}, \frac{1}{2}\right)}{\mathcal{B}\left(1 - \left(\frac{b'}{2}\right)^2, \frac{d+1}{2}, \frac{1}{2}\right)\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} \\ &\leq \frac{2}{\mathcal{B}\left(1 - \left(\frac{b'}{2}\right)^2, \frac{d+1}{2}, \frac{1}{2}\right)} \\ &\leq \frac{d+1}{\left(1 - \left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}. \end{aligned} \tag{16}$$

Finally, putting (4), (8), (9), (10), (11), (12), (13), (14), (15) and (16) together, we obtain

$$M^{-1} \leq \frac{q_{\mathbb{P}}}{q_d(s)} \leq M$$

for

$$M := \left(1 + 2880 d\kappa^2 r^2\right) \left(1 + \frac{1440(d+1)\kappa^2 r^2}{\left(1 - \left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}\right) \left(1 + 20\frac{dr_\xi}{r}\right) \left(1 + \frac{12(d+1)^{\frac{r_\xi}{r}}}{\left(1 - \left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}\right).$$

According to our assumptions, both $2880 d\kappa^2 r^2$ and $\frac{20dr_\xi}{r}$ are not larger than 4. In particular, M is bounded from above by $(1+\varepsilon_{\mathcal{M}})(1+\varepsilon_\xi)$. □

Proof of Theorem 1. Note that the proof of [10, Theorem 3.1] relies only on the inequality $\theta_{ij}^{(k)} \geq q_{ij}^{(k)}$ for $\|X_i - X_j\| \leq h_k$. However, this is ensured by Proposition 2 and the construction of the adjusted volume coefficient. □

Proof of Corollary 2. This is a simple consequence of Theorem 1 and the union bound. □

Proof of Theorem 2. Suppose $x_i, x_j \in \mathbb{R}^D$ are r_ξ-close to two different clusters and $\|x_i - x_j\| \leq h_k$. To simplify notation, we will implizitely condition on $X_i = x_i$ and $X_j = x_j$ for the remainder of this proof. For $l = i, j$ we choose a point $X_l' \in \mathcal{C}_{k_l}$ for $k_i \neq k_j$ such that $\|X_l' - X_l\| \leq r_\xi$. Our assumptions imply that the density f in the overlap $B(X_i', h_{k-1} + 2r_\xi) \cap B(X_j', h_{k-1} + 2r_\xi) \cap \mathcal{M}$ is bounded from above by $(1-\epsilon)f_0$. Let us denote the uniform measure on the manifold by μ and the distribution with gap and without noise by $\mathbb{P}_\epsilon$. We conclude

$$\begin{aligned}
\theta_{ij}^{(k)} &\leq \frac{\mathbb{P}_\epsilon(B(X_1', r + 2r_\xi) \cap B(X_2', r + 2r_\xi))}{\mathbb{P}_\epsilon(B(X_1', r - 2r_\xi) \cup B(X_2', r - 2r_\xi))} \\
&\leq \frac{(1-\epsilon)f_0 A}{(1-\epsilon)f_0 B + \epsilon f_0 C} \\
&= \frac{A}{B}\left(1 - \frac{\epsilon C}{(1-\epsilon)B + \epsilon C}\right)
\end{aligned}$$

with

$$\begin{aligned}
A &= \mu(B(X_1', r + 2r_\xi) \cap B(X_2', r + 2r_\xi)), \\
B &= \mu(B(X_1', r - 2r_\xi) \cup B(X_2', r - 2r_\xi)) \\
\text{and } C &= \mu(B(X_1', r - 2r_\xi)) + \mu(B(M_2', r - 2r_\xi)).
\end{aligned}$$

The factor $\frac{A}{B}$ is bounded from above by $(1+\varepsilon_{\mathcal{M}})(1+\varepsilon_\xi)q_{ij}^{(k)}$ as shown in the proof of Proposition 2. Moreover, $B < C$ implies that the second factor is bounded from above by $1 - \epsilon$, providing the upper bound

$$\theta_{ij}^{(k)} \leq (1-\epsilon)(1+\varepsilon_{\mathcal{M}})(1+\varepsilon_\xi)q_{ij}^{(k)}.$$

Monotonicity of q_d and the lower bound of the depth ϵ of the gap lead to

$$\begin{aligned}
\mathfrak{q}_{ij}^{(k)} - \theta_{ij}^{(k)} &\geq \left((1+\varepsilon_{\mathcal{M}})^{-1}(1+\varepsilon_\xi)^{-1} - (1-\epsilon)(1+\varepsilon_{\mathcal{M}})(1+\varepsilon_\xi)\right) q_d(b) \\
&\geq \left(\left(1 + \frac{\epsilon}{7}\right)^{-1} - (1-\epsilon)\left(1 + \frac{\epsilon}{7}\right)\right) q_d(b) \\
&\geq \epsilon \frac{q_d(b)}{\sqrt{2}}.
\end{aligned} \tag{17}$$

Using Pinsker's inequality, we get

$$\mathcal{K}\left(\mathfrak{q}_{ij}^{(k)}, \theta_{ij}^{(k)}\right) \geq \epsilon^2 q_d(b)^2. \tag{18}$$

As $\frac{n}{\log n} \geq \frac{2\beta}{z_k^2}$, we can choose some $\delta > 0$ satisfying the inequalities

$$2\delta^2 n \geq \beta \log n \tag{19}$$

$$\text{and } \delta n \leq \frac{z_k n}{2}. \tag{20}$$

Note that $z_k \leq \mathbb{P}\left(B(X_i, h_{k-1}) \cup B(X_j, h_{k-1})\right)$. Hoeffding's inequality implies in view of (19)

$$N_{i\vee j}^{(k)} \geq (z_k - \delta)n$$

with probability at least $1 - n^{-\beta}$. This implies together with (20)

$$N_{i\vee j}^{(k)} \geq \frac{z_k n}{2} \tag{21}$$

with probability at least $1 - n^{-\beta}$. On the other hand, by [10, Lemma 5.1] we have

$$\mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)}) < \frac{\beta \log n}{N_{i\vee j}^{(k)}} \tag{22}$$

with probability at least $1 - 2n^{-\beta}$. By the union bound, there exists an event E of probability at least $1 - 3n^{-\beta}$ on which both (21) and (22) hold. In the following let us fix an outcome of the event E. Then (21) and (22) imply

$$\mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)}) < \frac{2\beta \log n}{z_k n}$$

The assumption $\frac{\epsilon^2 n}{\log n} \geq 2\alpha z_k^{-1} q_d(b)^{-2}$, $\alpha > \beta > 0$, implies

$$\mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)}) < \frac{\beta}{\alpha}\epsilon^2 q_d(b)^2. \tag{23}$$

Note that (17) implies in particular $\mathfrak{q}_{ij}^{(k)} > \theta_{ij}^{(k)}$. Since the function $\mathcal{K}(\cdot, \theta)$ is strictly monotone on the interval $[\theta, 1)$ and considering $\frac{\beta}{\alpha} < 1$, we conclude from (18) and (23)

$$\widetilde{\theta}_{ij}^{(k)} < \mathfrak{q}_{ij}^{(k)}. \tag{24}$$

The triangle inequality and Pinsker's inequality yield

$$\begin{aligned} |\widetilde{\theta}_{ij}^{(k)} - \mathfrak{q}_{ij}^{(k)}| &\geq |\theta_{ij}^{(k)} - \mathfrak{q}_{ij}^{(k)}| - |\widetilde{\theta}_{ij}^{(k)} - \theta_{ij}^{(k)}| \\ &\geq \epsilon \frac{q_d(b)}{\sqrt{2}} - \sqrt{\frac{1}{2}\mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)})} \\ &\overset{(23)}{\geq} \epsilon \frac{q_d(b)}{\sqrt{2}} \left(1 - \sqrt{\frac{\beta}{\alpha}}\right) \end{aligned} \tag{25}$$

From Pinsker's inequality and the assumption $\frac{\epsilon^2 n}{\log n} \geq 2\alpha z_k^{-1} q_d(b)^{-2}$ we deduce

$$\begin{aligned}\mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \mathfrak{q}_{ij}^{(k)}) &\geq 2\left(\widetilde{\theta}_{ij}^{(k)} - \mathfrak{q}_{ij}^{(k)}\right)^2 \\ &\overset{(25)}{\geq} \epsilon^2 q_d(b)^2 \left(1 - \sqrt{\frac{\beta}{\alpha}}\right)^2 \\ &\geq \frac{\log n}{z_k n} 2\alpha \left(1 - \sqrt{\frac{\beta}{\alpha}}\right)^2 \\ &\overset{(21)}{\geq} \frac{\log n}{N_{i\vee j}^{(k)}} \left(\sqrt{\alpha} - \sqrt{\beta}\right)^2 \end{aligned} \tag{26}$$

Finally, putting together (24) and (26), we conclude that any outcome of the event E satisfies

$$\begin{aligned}T_{ij}^{(k)} &= N_{i\vee j}^{(k)} \mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \mathfrak{q}_{ij}^{(k)}) \{\mathbb{1}(\widetilde{\theta}_{ij}^{(k)} < \mathfrak{q}_{ij}^{(k)}) - \mathbb{1}(\widetilde{\theta}_{ij}^{(k)} \geq \mathfrak{q}_{ij}^{(k)})\} \\ &\geq \left(\sqrt{\alpha} - \sqrt{\beta}\right)^2 \log n.\end{aligned}$$

The choice of x_i and x_j is irrelevant for this result, so it is also valid in the unconditional form. □

Proof of Theorem 4. Let us denote the value of the constant density under the null hypothesis by f_0 and the Kullback-Leibler divergence by $\mathcal{D}_{\mathsf{KL}}(\cdot,\cdot)$. Using $1 = f_G|G| + f_V|V|$, we compute

$$\begin{aligned}f_V &= \frac{1}{|G| + |V| - \delta|G|} \text{ and} \\ f_G &= \frac{1-\delta}{|G| + |V| - \delta|G|}.\end{aligned}$$

Additivity of the Kullback-Leibler divergence and $f_0 = \frac{1}{|V|+|G|}$ yields

$$\begin{aligned}n^{-1}\mathcal{D}_{\mathsf{KL}}(\mathbb{P}_0, \mathbb{P}_1) &= f_0|G| \log \frac{f_0}{f_G} + f_0|V| \log \frac{f_0}{f_V} \\ &= \log\left(1 - \delta \frac{|G|}{|G|+|V|}\right) - \frac{|G|}{|G|+|V|} \log(1-\delta) \\ &= \frac{\delta^2}{2} \frac{|G|}{|G|+|V|} \left(1 + \frac{|G|}{|G|+|V|}\right) + o(\delta^2),\end{aligned}$$

the latter follows from the Taylor expansion. As $\mathcal{D}_{\mathsf{KL}}(\mathbb{P}_0, \mathbb{P}_1) \to \infty$ is a necessary condition for consistent testing [36, Section 2.4.2], we deduce that no test is able to separate the two cases consistently provided that $n\delta^2 \nrightarrow \infty$ as $n \to \infty$. □

Before we prove Lemma 2, let us introduce the so-called *general volume coefficient*.

Definition 2. *Suppose $r_1, r_2 > 0$, $D \in \mathbb{Z}_{>0}$, $M_1 = (0, \ldots, 0) \in \mathbb{R}^D$ and $M_2 = (1, 0, \ldots, 0) \in \mathbb{R}^D$. By λ_D we denote the D-dimensional Lebesgue measure and $B_D(\cdot,\cdot)$ denotes an euclidean Ball in $\mathbb{R}^D$ with given center and radius. We define the D-*dimensional general volume coefficient *by*

$$q_D(r_1, r_2) := \frac{\lambda_D\left(B_D(M_1, r_1) \cap B_D(M_2, r_2)\right)}{\lambda_D\left(B_D(M_1, r_1) \cup B_D(M_2, r_2)\right)}$$

Lemma 5. *For $M_1 \neq M_2 \in \mathbb{R}^D$ and $r_1, r_2 > 0$ we have*

$$\frac{\lambda_D(B_D(M_1, r_1) \cap B_D(M_2, r_2))}{\lambda_D(B_D(M_1, r_1) \cup B_D(M_2, r_2))} = q_D\left(\frac{r_1}{\|M_1 - M_2\|}, \frac{r_2}{\|M_1 - M_2\|}\right)$$

Proof. This follows from the invariance of the quotient of two D-dimensional volumes under rotation, translation and uniform scaling. □

Lemma 6. *Suppose $r_1, r_2 > 0$. Using the usual order of arguments we denote the regularized incomplete beta function by $I_\cdot(\cdot, \cdot)$. Then*

$$q_D(r_1, r_2) = \begin{cases} 0 & , r_1 + r_2 \leq 1 \\ \left(\frac{r_j}{r_i}\right)^D & , r_i - r_j \geq 1 \\ \frac{r_1 + r_2 - 1}{r_1 + r_2 + 1} & , D = 1 \text{ and } r_1 + r_2 > 1 \text{ and } |r_1 - r_2| < 1 \\ \frac{V_D^{cap}(r_1, r_2) + V_D^{cap}(r_2, r_1)}{V_D^{ball}(r_1) + V_D^{ball}(r_2) - V_D^{cap}(r_1, r_2) - V_D^{cap}(r_2, r_1)} & , \text{otherwise} \end{cases}$$

with

$$V_D^{ball}(r_i) = 2r_i^D$$

$$V_D^{cap}(r_i, r_j) = \begin{cases} r_i^D I_{1 - \left(\frac{1 + r_i^2 - r_j^2}{2r_i}\right)^2}\left(\frac{D+1}{2}, \frac{1}{2}\right) & , r_j^2 - r_i^2 \leq 1 \\ 2r_i^D - r_i^D I_{1 - \left(\frac{1 + r_i^2 - r_j^2}{2r_i}\right)^2}\left(\frac{D+1}{2}, \frac{1}{2}\right) & , r_j^2 - r_i^2 > 1 \end{cases}$$

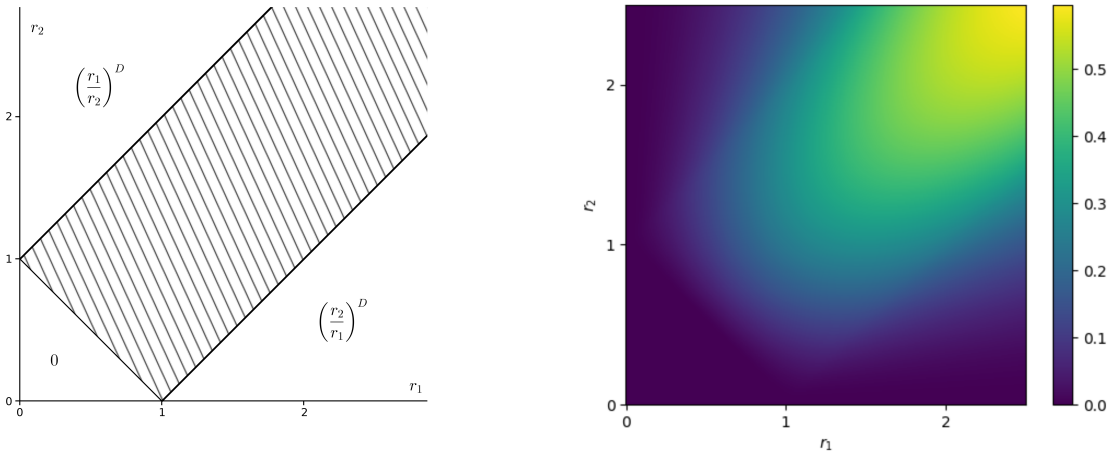


Figure 14: Left: Different regimes for formula of $q_D(r_1, r_2)$ given in Lemma 6.
Right: Plot of $q_2(r_1, r_2)$

Proof. We only discuss the nontrivial regime where $r_1 + r_2 > 1$ and $|r_2 - r_1| < 1$ for $D > 1$. Then the overlap of the two corresponding spheres with radii r_1 and r_2 around $M_1 = (0, \ldots, 0)$ and $M_2 = (1, 0, \ldots, 0)$ contains two points of the form $(x, \pm y, 0, \ldots, 0)$. The coordinate equations of the two spheres yield

$$x^2 + y^2 = r_1^2$$
$$(x - 1)^2 + y^2 = r_2^2$$

implying

$$x = \frac{1 + r_1^2 - r_2^2}{2}$$
$$y = \pm r_1 \sqrt{1 - \left(\frac{1 + r_1^2 - r_2^2}{2r_1}\right)^2}$$

We denote the smaller angle between the x-axis and the line through M_1 and (x, y) by ϕ_1. Analogously we define ϕ_2, c.f. Figure 15.

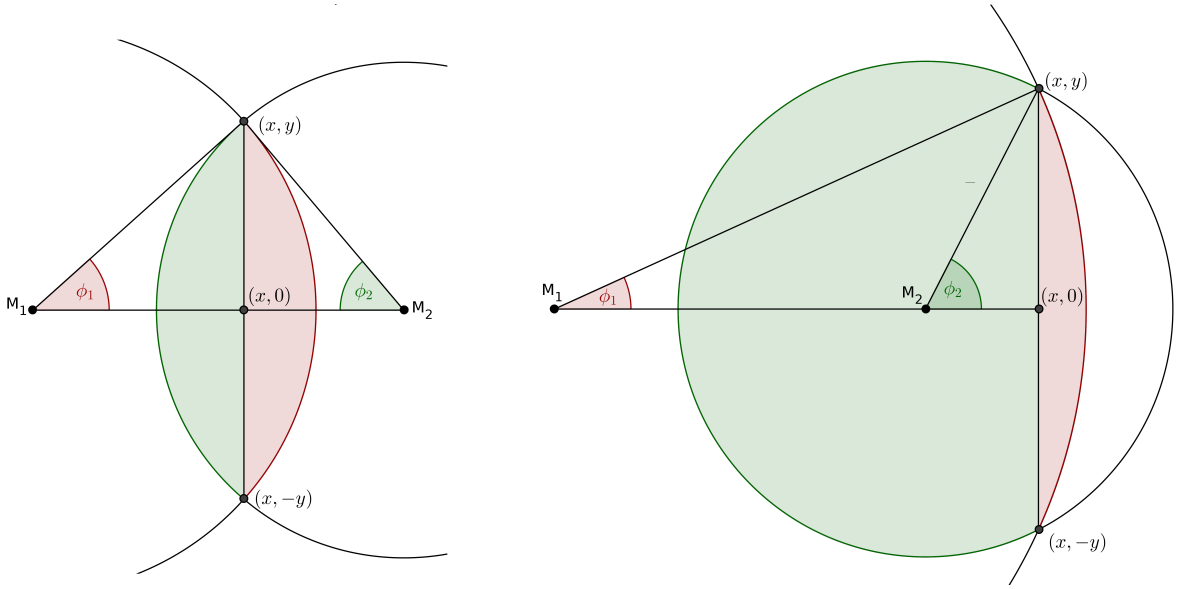


Figure 15: The volume of the overlap of two balls is the sum of the volumes of two caps that are shown in green and red. The corresponding angles used in the formulas of these volumes are highlighted in the same colour. On the left, we see the case $0 < x < 1$, whereas on the right $x > 1$.

We conclude

$$\begin{aligned}\sin^2 \phi_1 &= \left(\frac{|y|}{r_1}\right)^2 \\ &= 1 - \left(\frac{1 + r_1^2 - r_2^2}{2r_1}\right)^2\end{aligned}$$

and

$$\begin{aligned}\sin^2 \phi_2 &= \left(\frac{|y|}{r_2}\right)^2 \\ &= \frac{r_1^2}{r_2^2} - \left(\frac{1 + r_1^2 - r_2^2}{2r_2}\right)^2 \\ &= 1 - \left(\frac{1 + r_2^2 - r_1^2}{2r_2}\right)^2\end{aligned}$$

Note that $x < 1$ is equivalent to $r_1^2 - r_2^2 < 1$ and $x > 0$ is equivalent to $r_2^2 - r_1^2 < 1$. Using the

formula for the volume of a hyperspherical cap given in [22], we conclude

$$\lambda_D\left(B_D(M_1,r_1)\cap B_D(M_2,r_2)\right) = \frac{\pi^{\frac{D}{2}}}{2\Gamma\left(\frac{D}{2}+1\right)}\left(V_D^{\text{cap}}(r_1,r_2)+V_D^{\text{cap}}(r_2,r_1)\right)$$

$$\lambda_D\left(B_D(M_1,r_1)\cup B_D(M_2,r_2)\right) = \frac{\pi^{\frac{D}{2}}}{2\Gamma\left(\frac{D}{2}+1\right)}\left(V_D^{\text{ball}}(r_1)+V_D^{\text{ball}}(r_2)-V_D^{\text{cap}}(r_1,r_2)-V_D^{\text{cap}}(r_2,r_1)\right)$$

□

Proof of Lemma 2. Since $\mathcal{Q}$ is contineous and $\mathcal{Q}(t) = 0$ for $t > t'_{\text{max}} := \sup\{t : (H+tv)\cap B(M_1,r)\cap B(M_2,r)\neq\{\}\}$, we only need to dicuss the case $0 < t < t'_{\text{max}}$. As $\mathcal{Q}(t)$ is contineous w.r.t. rotation of H around the point $\frac{M_1+M_2}{2}$, we can w.l.o.g. assume that the vector $M_1 - M_2$ is neither parallel nor orthogonal to H. Moreoever, we assume w.l.o.g. that there exists $d > 0$ such that $M_1 \in H + dv$. As a nontrivial intersection of a ball in $\mathbb{R}^{D+1}$ with a hyperplane is a D-dimensional ball, we can rewrite $\mathcal{Q}(t)$ using Lemma 5. The corresponding radii can be easily computed using Pythagoras' theorem, c.f. Figure 16. We get

$$\mathcal{Q}(t) = q_D\left(r_1(t), r_2(t)\right)$$

$$\text{with } r_1(t) = \sqrt{\frac{1-(t-d)^2}{\|M_1-M_2\|^2-4d^2}}$$

$$\text{and } r_2(t) = \sqrt{\frac{1-(t+d)^2}{\|M_1-M_2\|^2-4d^2}}$$

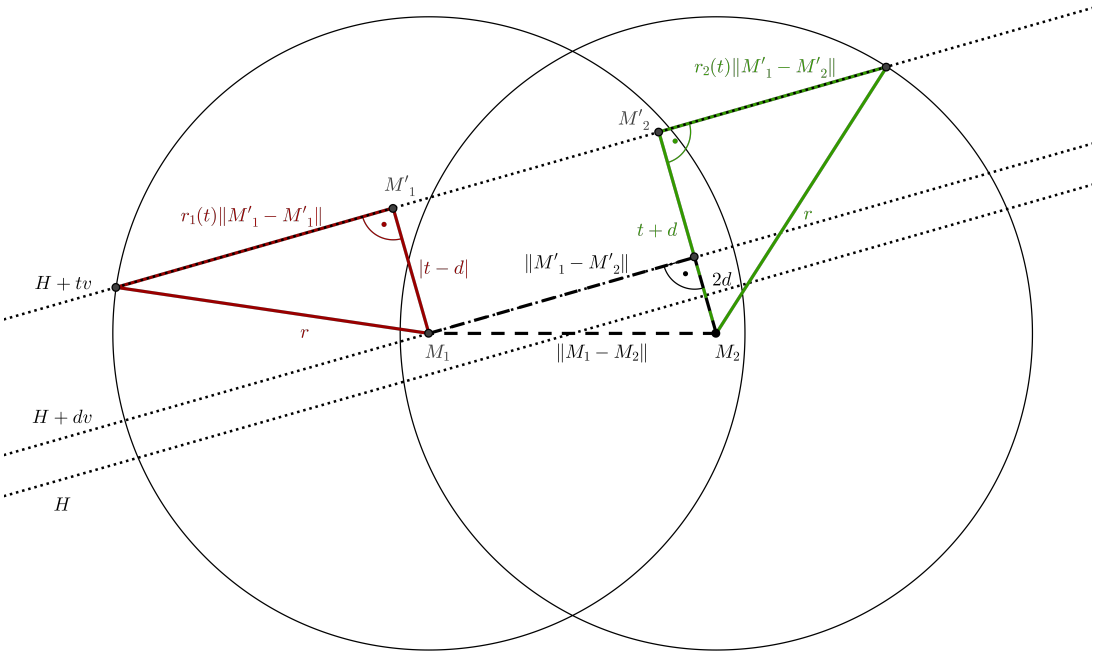


Figure 16: We denote $M_1+(t-d)v$ by M'_1 and $M_2+(t+d)v$ by M'_2. These are the center points of the two D-dimensional balls that form the intersection of the original $(D+1)$-dimensional balls with $H+tv$. According to Pythagoras' theorem, their radii are given by $(r^2-(t-d)^2)^{1/2}$ and $(r^2-(t+d)^2)^{1/2}$, whereas the distance between the center points is $\|M'_1-M'_2\| = (\|M_1-M_2\|^2-4d^2)^{1/2}$.

We have

$$\frac{d}{dt}r_1(t) = \frac{-t+d}{\sqrt{1-(t-d)^2}\sqrt{\|M_1-M_2\|^2-(2d)^2}}$$
$$\frac{d}{dt}r_2(t) = \frac{-t-d}{\sqrt{1-(t+d)^2}\sqrt{\|M_1-M_2\|^2-(2d)^2}}$$

We oberserve the following relations between r_1 and r_2:

$$r_1(t) > r_2(t) \tag{27}$$
$$\frac{d}{dt}r_2(t) < 0 \tag{28}$$
$$\left|\frac{d}{dt}r_1(t)\right| < \left|\frac{d}{dt}r_2(t)\right| \tag{29}$$

First, let us discuss the case when there exists an open environment I containing t, such that $\mathcal{Q}(t') = r_1(t')^{-D}r_2(t')^D$ for all $t' \in I$. We conclude from (27), (28) and (29)

$$\begin{aligned}\frac{d}{dt}\mathcal{Q}(t) &= D\left(\frac{r_2(t)}{r_1(t)}\right)^{D-1}\frac{r_1(t)\left(\frac{d}{dt}r_2(t)\right)-\left(\frac{d}{dt}r_1(t)\right)r_2(t)}{r_1(t)^2}\\ &< 0\end{aligned}$$

Next, let us consider that case where $\mathcal{Q} = (r_1+r_2-1)(r_1+r_2+1)^{-1}$ on an open interval containing t. Again, we conclude from (28) and (29)

$$\begin{aligned}\frac{d}{dt}\mathcal{Q}(t) &= 2\frac{\frac{d}{dt}r_1(t)+\frac{d}{dt}r_2(t)}{(r_1+r_2+1)^2}\\ &< 0\end{aligned}$$

Finally, consider the case where $D > 2$ and on an open environment around t we have

$$\begin{aligned}\mathcal{Q} &= q_D(r_1,r_2)\\ &= \frac{V_D^{\mathsf{cap}}(r_1,r_2)+V_D^{\mathsf{cap}}(r_2,r_1)}{V_D^{\mathsf{ball}}(r_1)+V_D^{\mathsf{ball}}(r_2)-V_D^{\mathsf{cap}}(r_1,r_2)-V_D^{\mathsf{cap}}(r_2,r_1)}\end{aligned} \tag{30}$$

for $V_D^{\mathsf{ball}}(\cdot)$ and $V_D^{\mathsf{cap}}(\cdot,\cdot)$ defined as in Lemma 6. The terms $V_D^{\mathsf{ball}}(\cdot)$ and $V_D^{\mathsf{cap}}(\cdot,\cdot)$ denote the volume of the respective balls and caps up to the constant

$$c = \frac{2\Gamma\left(\frac{D}{2}+1\right)}{\pi^{\frac{D}{2}}}$$

Recall that the derivative of the volume of a ball w.r.t. its radius is given by the surface area of the corresponding sphere. In particular, we have

$$\frac{d}{dr_i}V_D^{\mathsf{ball}}(r_j) = \begin{cases} cA_D^{\mathsf{sphere}}(r_i) & , i=j\\ 0 & , i\neq j\end{cases}$$

with $A_D^{\mathsf{sphere}}(\cdot)$ denoting the surface area of a D-dimensional sphere with given radius. Similarly, it can be shown that the partial derivatives of the volume of the overlap $c^{-1}(V_D^{\mathsf{cap}}(r_1,r_2)+V_D^{\mathsf{cap}}(r_2,r_1))$ w.r.t. r_1 and r_2 are up to the same constant given by the surface areas $A_D^{\mathsf{cap}}(r_1,r_2)$ and $A_D^{\mathsf{cap}}(r_2,r_1)$ of

the the corresponding hyperspherical caps that form together the boundary of the overlap, c.f. Figure 17.

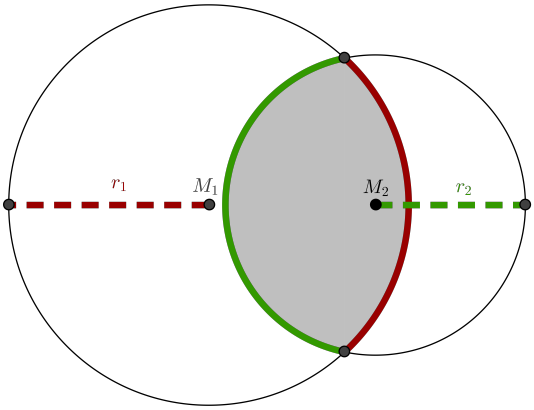


Figure 17: Case $D = 2$: The derivative of the area of the intersection of the two balls (gray) w.r.t. r_1 (r_2) is given by the red (green) arc length

Exact formulas for $A_D^{\text{cap}}(\cdot,\cdot)$ are given in [22]. However, those are not needed for this proof. It is enough to observe the following relation

$$A_D^{\text{cap}}(r_1, r_2) < A_D^{\text{cap}}(r_2, r_1)$$

as a consequence of (27). Let us introduce the notations

$$S^{\text{ball}} := V_D^{\text{ball}}(r_1) + V_D^{\text{ball}}(r_2)$$
$$S^{\text{cap}} := V_D^{\text{cap}}(r_1, r_2) + V_D^{\text{cap}}(r_2, r_1)$$

We conclude

$$\text{and } \frac{d}{dr_1} S^{\text{ball}} > \frac{d}{dr_2} S^{\text{ball}} \tag{31}$$
$$\frac{d}{dr_1} S^{\text{cap}} < \frac{d}{dr_2} S^{\text{cap}} \tag{32}$$

In view of

$$\frac{d}{dr_i} q_D(r_1, r_2) = \frac{\left(\frac{d}{dr_i} S^{\text{cap}}\right) S^{\text{ball}} - S^{\text{cap}} \left(\frac{d}{dr_i} S^{\text{ball}}\right)}{\left(S^{\text{ball}} - S^{\text{cap}}\right)^2}$$

we conclude from (31) and (32)

$$\frac{d}{dr_1} q_D < \frac{d}{dr_2} q_D \tag{33}$$

Note that increasing the radii r_1 and r_2 by a common factor $C > 1$ has the same effect on the coefficient of the volumes of the intersection and the union of the two corresponding balls as when moving the center point M_2 such that $\|M_1 - M_2\|$ decreases by a factor C^{-1}. Considering Lemma 5, we observe

$$q_D(Cr_1, Cr_2) > q_D(r_1, r_2)$$

for $C > 1$. This implies

$$r_1 \frac{d}{dr_1} q_D + r_2 \frac{d}{dr_2} q_D > 0 \tag{34}$$

From (33) and (34) we deduce

$$0 < \frac{d}{dr_2} q_D \tag{35}$$

$$\text{and } \left| \frac{d}{dr_1} q_D \right| < \frac{d}{dr_2} q_D \tag{36}$$

Note that $q_D(r_1, r_2)$ is differentiable at $(r_1(t), r_2(t))$ as the formula given in (30) is valid on open environment. From (28), (29), (35) and (36) we conclude

$$\begin{aligned} \frac{d}{dt}\mathcal{Q}(t) &= \frac{d}{dt} r_1(t) \frac{d}{dr_1} q_D(r_1(t), r_2(t)) + \frac{d}{dt} r_2(t) \frac{d}{dr_2} q_D(r_1(t), r_2(t)) \\ &< 0 \end{aligned}$$

Lastly, $r_1(t) - r_2(t)$ is strictly monotonely increasing on $(0, t'_{\text{max}})$. In view of Lemma 6, this implies that $\mathcal{Q}(t)$ is differentiable on $(0, t'_{\text{max}}) \setminus S$ with a negative derivative for some finite subset $S \subset (0, t'_{\text{max}})$. The function $\mathcal{Q}$ is contineous on $[0, t_{\text{max}})$. Consequently, it is also monotonely decreasing. □

Proof of Lemma 1. The case $D = 1$ is trivial. Let us assume $D > 1$. We prove the lemma by contradiction, i.e. we assume that there exists a counterexample such that

$$\frac{\lambda(\mathcal{H} \cap B(M_1, r) \cap B(M_2, r))}{\lambda(\mathcal{H} \cap (B(M_1, r) \cup B(M_2, r)))} < \frac{\lambda(B(M_1, r) \cap B(M_2, r))}{\lambda(B(M_1, r) \cup B(M_2, r))} \tag{37}$$

We can choose $\mathcal{H}$ such that for any other half-space of the form $\mathcal{H}' = \mathcal{H} + v'$ for some $v' \in \mathbb{R}^D$ containing M_1 and M_2 we have

$$\frac{\lambda\left(\mathcal{H} \cap B(M_1, r) \cap B(M_2, r)\right)}{\lambda(\mathcal{H} \cap (B(M_1, r) \cup B(M_2, r)))} \leq \frac{\lambda\left(\mathcal{H}' \cap B(M_1, r) \cap B(M_2, r)\right)}{\lambda(\mathcal{H}' \cap (B(M_1, r) \cup B(M_2, r)))} \tag{38}$$

There exists a unique half-space $\mathcal{H}_0$ whose boundary H_0 contains $\frac{M_1+M_2}{2}$ and is parallel to the boundary of $\mathcal{H}$. Note that by symmetricity,

$$\frac{\lambda\left(\mathcal{H}_0 \cap B(M_1, r) \cap B(M_2, r)\right)}{\lambda\left(\mathcal{H}_0 \cap (B(M_1, r) \cup B(M_2, r))\right)} = \frac{\lambda\left(B(M_1, r) \cap B(M_2, r)\right)}{\lambda\left(B(M_1, r) \cup B(M_2, r)\right)} \tag{39}$$

There exists a unique vector v of norm 1 that is orthogonal to H_0 such that $\frac{M_1+M_2}{2} \in \mathcal{H}_0 + v$. Moreover, for $t_{\text{max}} := \sup\{t : (H_0 + tv) \cap (B(M_1, r) \cup B(M_2, r)) \neq \{\}\}$, there exists a unique $t_{\mathcal{H}} \in (0, t_{\text{max}})$ such that $\mathcal{H} = \mathcal{H}_0 + t_{\mathcal{H}} v$. Let us denote the $(D-1)$-dimensional Lebesgue measure by λ_{D-1}. According to Fubini's theorem we have

$$\begin{aligned} &\frac{\lambda\left(\mathcal{H} \cap B(M_1, r) \cap B(M_2, r)\right)}{\lambda\left(\mathcal{H} \cap (B(M_1, r) \cup B(M_2, r))\right)} \\ =&\frac{\lambda\left(\mathcal{H}_0 \cap B(M_1, r) \cap B(M_2, r)\right) + \int_0^{t_{\mathcal{H}}} \lambda_{D-1}((H_0 + tv) \cap B(M_1, r) \cap B(M_2, r)) dt}{\lambda\left(\mathcal{H}_0 \cap (B(M_1, r) \cup B(M_2, r))\right) + \int_0^{t_{\mathcal{H}}} \lambda_{D-1}((H_0 + tv) \cap (B(M_1, r) \cup B(M_2, r))) dt} \end{aligned} \tag{40}$$

From (37), (39), (40) and the monotonicity described in Lemma 2, we conclude

$$\frac{\lambda_{D-1}((H_0 + t_{\mathcal{H}} v) \cap B(M_1, r) \cap B(M_2, r))}{\lambda_{D-1}((H_0 + t_{\mathcal{H}} v) \cap (B(M_1, r) \cup B(M_2, r)))} < \frac{\lambda\left(\mathcal{H} \cap B(M_1, r) \cap B(M_2, r)\right)}{\lambda\left(\mathcal{H} \cap (B(M_1, r) \cup B(M_2, r))\right)} \tag{41}$$

Suppose $t' \in (t_{\mathcal{H}}, t_{\mathsf{max}})$. Then

$$\begin{aligned} &\frac{\lambda\left((\mathcal{H}_0 + t' v) \cap B(M_1, r) \cap B(M_2, r)\right)}{\lambda\left((\mathcal{H}_0 + t' v) \cap (B(M_1, r) \cup B(M_2, r))\right)} \\ =&\frac{\lambda\left(\mathcal{H} \cap B(M_1, r) \cap B(M_2, r)\right) + \int_{t_{\mathcal{H}}}^{t'} \lambda_{D-1}((H_0 + tv) \cap B(M_1, r) \cap B(M_2, r)) dt}{\lambda\left(\mathcal{H} \cap (B(M_1, r) \cup B(M_2, r))\right) + \int_{t_{\mathcal{H}}}^{t'} \lambda_{D-1}((H_0 + tv) \cap (B(M_1, r) \cup B(M_2, r))) dt} \end{aligned} \tag{42}$$

From (41), (42) and Lemma 2 we deduce

$$\frac{\lambda\left((\mathcal{H}_0 + t' v) \cap B(M_1, r) \cap B(M_2, r)\right)}{\lambda\left((\mathcal{H}_0 + t' v) \cap (B(M_1, r) \cup B(M_2, r))\right)} < \frac{\lambda\left(\mathcal{H} \cap B(M_1, r) \cap B(M_2, r)\right)}{\lambda\left(\mathcal{H} \cap (B(M_1, r) \cup B(M_2, r))\right)}$$

This is a contradiction to (38). □

Before proving Proposition 3, we state the following generalization of Lemma 3. We denote the Lebesgue measure on a submanifold of $\mathbb{R}^D$ by λ.

Lemma 7. *For a C-Lipschitz function $f_1 : \mathcal{M}_1 \to \mathcal{M}_2$ between two d-dimensional submanifolds of $\mathbb{R}^D$ and a measurable function f_2 on $\mathcal{M}_2$ we have*

$$\int_{\mathcal{M}_2} f_2 d\lambda \leq C^d \int_{\mathcal{M}_1} f_2 \circ f_1 d\lambda$$

Proof. This follows from Lemma 3 together with the definition of the Lebesgue integral of a positive function as a supremum of integrals of step functions. □

Proof of Proposition 3. W.l.o.g. we consider only one cluster $\mathcal{C} = \mathcal{C}_1$ and assume $f \propto \mathbb{1}(\mathcal{C})$. If the set of all possible superlevel sets is finite, the general result follows by summation. In case that this set is infinite, e.g. if f is smooth and not constant, f can be constructed as the limit of discrete functions.

Moving on, consider $M_i' \in \mathcal{C}$ of distance at most r_ξ to M_i. Moreover, let us denote the projection on the tangent plane $\mathcal{T}$ to $\mathcal{M}$ at M_1' by P. Depending on the context, we denote by λ either the Lebesgue measure on $\mathcal{M}$ or a linear space such as the tangent space. We apply Lemma 4. For $r\kappa \leq (120)^{-1}$, the projection P is injective on the Ball $B(M_1', 3r)$ with an inverse that is Lipschitz with constant $L := 1 + 360\kappa^2 r^2$. Note that this ball contains $B(M_1', r + 2r_\xi) \cup B(M_2', r + 2r_\xi)$. From Lemma 7 we conclude

$$\begin{aligned} q_{\mathbb{P}} = \frac{\mathbb{P}(B(M_1, r) \cap B(M_2, r))}{\mathbb{P}(B(M_1, r) \cup B(M_2, r))} &\geq \frac{\mathbb{P}_0(B(M_1', r - 2r_\xi) \cap B(M_2', r - 2r_\xi))}{\mathbb{P}_0(B(M_1', r + 2r_\xi) \cup B(M_2', r + 2r_\xi))} \\ &\geq L^{-d} \frac{\int_{T \cap B(P(M_1'), \frac{r-2r_\xi}{L}) \cap B(P(M_2'), \frac{r-2r_\xi}{L})} f \circ P^{-1} d\lambda}{\int_{T \cap (B(P(M_1'), r+2r_\xi) \cup B(P(M_2'), r+2r_\xi))} f \circ P^{-1} d\lambda}, \end{aligned}$$

where $\mathbb{P}_0$ denotes the noiseless distribution. Moreover, we can rewrite the integral using the push-forward measure $(P|^{-1}_{B(M_1', 3r)})_*(\mathbb{P}_0)$. For simplicity we just use the notation $P_*^{-1}\mathbb{P}_0$ as well as $Z_i :=$

$P(M_i')$. We get the lower bound

$$\begin{aligned}
q_{\mathbb{P}} &\geq L^{-d}AB \\
A &= \frac{P_*^{-1}\mathbb{P}_0\left(T \cap B(Z_1, \frac{r-2r_\xi}{L}) \cap B(Z_2, \frac{r-2r_\xi}{L})\right)}{P_*^{-1}\mathbb{P}_0\left(T \cap (B(Z_1, \frac{r-2r_\xi}{L}) \cup B(Z_2, \frac{r-2r_\xi}{L}))\right)} \\
B &= \frac{P_*^{-1}\mathbb{P}_0\left(T \cap (B(Z_1, \frac{r-2r_\xi}{L}) \cup B(Z_2, \frac{r-2r_\xi}{L}))\right)}{P_*^{-1}\mathbb{P}_0\left(T \cap (B(Z_1, r+2r_\xi) \cup B(Z_2, r+2r_\xi))\right)}
\end{aligned}$$

WLOG we assume that $P(\mathcal{C})$ does not fully contain the intersection in term A. Then there exists $p \in P(\partial\mathcal{C}) \cap B(Z_1, \frac{r-2r_\xi}{L}) \cap B(Z_2, \frac{r-2r_\xi}{L})$. Consider a ball of radius $2r$ around $P^{-1}(p)$ and let's denote by $\mathcal{T}'$ the tangent plane of dimension $d-1$ to $\partial\mathcal{C}$ at $P^{-1}(p)$. If $\kappa' r \leq 80^{-1}$ the inverse of the restriction (to the ball around $P^{-1}(p)$) of the projection of $\partial\mathcal{C}$ to $\mathcal{T}'$ is $L_{\mathcal{C}} := 1 + 160(\kappa' r)^2$-Lipschitz. By Pythagoras theorem the distance of $\partial\mathcal{C}$ to $\mathcal{T}'$ inside the considered Ball is bounded from above by

$$\begin{aligned}
2r\sqrt{L_{\mathcal{C}}^2 - 1} &= 2r\sqrt{320(\kappa' r)^2 + 160^2(\kappa' r)^4} \\
&\leq 2\sqrt{324}\kappa' r^2 \\
&= 36\kappa' r^2
\end{aligned}$$

As the projection onto $\mathcal{T}$ is 1-Lipschitz, also the distance of $P(\partial\mathcal{C}) \cap B(p, \frac{2r}{L})$ to $P(\mathcal{T}')$ is bounded by the same term. I. p. there exists half-planes $H_2 \subset H_1$ of dimension d in $\mathcal{T}$ whos boundaries are parallel at a distance $72\kappa' r^2$ and

$$H_2 \cap B(p, \frac{2r}{L}) \subset P(\mathcal{C}) \cap B(p, \frac{2r}{L}) \subset H_1 \cap B(p, \frac{2r}{L})$$

For the denominator of A we get

$$\begin{aligned}
&P_*^{-1}\mathbb{P}_0\left(T \cap \left(B(Z_1, \frac{r-2r_\xi}{L}) \cup B(Z_2, \frac{r-2r_\xi}{L})\right)\right) \\
&\leq \frac{1}{\lambda(\mathcal{M})}\lambda\left(H_1 \cap \left(B(Z_1, \frac{r-2r_\xi}{L}) \cup B(Z_2, \frac{r-2r_\xi}{L})\right)\right)
\end{aligned} \tag{43}$$

whereas for the nominator we get

$$\begin{aligned}
&P_*^{-1}\mathbb{P}_0\left(\mathcal{T} \cap B(Z_1, \frac{r-2r_\xi}{L}) \cap B(Z_2, \frac{r-2r_\xi}{L})\right) \\
&\geq \frac{1}{\lambda(\mathcal{M})}\lambda\left(H_2 \cap B(Z_1, \frac{r-2r_\xi}{L}) \cap B(Z_2, \frac{r-2r_\xi}{L})\right) \\
&\geq \frac{1}{\lambda(\mathcal{M})}\left[\lambda\left(H_1 \cap B(Z_1, \frac{r-2r_\xi}{L}) \cap B(Z_2, \frac{r-2r_\xi}{L})\right) - \lambda\left((H_1 \setminus H_2) \cap B\left(Z_1, \frac{r-2r_\xi}{L}\right)\right)\right] \\
&\geq \frac{1}{\lambda(\mathcal{M})}\left[\lambda\left(H_1 \cap B(Z_1, \frac{r-2r_\xi}{L}) \cap B(Z_2, \frac{r-2r_\xi}{L})\right) - 72\kappa' r^2 \lambda_{d-1}\left(B_{d-1}\left(\cdot, \frac{r-2r_\xi}{L}\right)\right)\right]
\end{aligned} \tag{44}$$

In the above, we denote by $\lambda_{d-1}(B_{d-1}(\cdot, r'))$ the volume of a $(d-1)$-dimensional ball of radius r'.

We have

$$\begin{aligned}\frac{72\kappa' r^2 \lambda_{d-1}\left(B_{d-1}\left(\cdot, \frac{r-2r_\xi}{L}\right)\right)}{\lambda\left(H_1 \cap (B(Z_1, \frac{r-2r_\xi}{L}) \cup B(Z_2, \frac{r-2r_\xi}{L})\right)} &\leq 144\kappa' r^2 \frac{\lambda_{d-1}\left(B_{d-1}\left(\cdot, \frac{r-2r_\xi}{L}\right)\right)}{\lambda_d\left(B_d\left(\cdot, \frac{r-2r_\xi}{L}\right)\right)} \\ &\leq 144\pi^{-\frac{1}{2}}\kappa' r L \frac{r}{r-2r_\xi} \frac{\Gamma\left(\frac{d+2}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)} \end{aligned} \tag{45}$$

Due to the upper bound assumption on r_ξ we have

$$\frac{r}{r-2r_\xi} \leq \frac{10}{9}$$

Moreoever, the upper bound assumption on r with respect to the reach implies

$$L \leq \frac{41}{40}$$

The last factor can be upper bounded utilizing the logarithmic convexity of the gamma function

$$\begin{aligned}\frac{\Gamma\left(\frac{d+2}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)} &\leq \sqrt{\frac{\Gamma\left(\frac{d+3}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)}} \\ &= \sqrt{\frac{d+1}{2}}\end{aligned}$$

Together, we conclude from (45)

$$\begin{aligned}\frac{72\kappa' r^2 \lambda_{d-1}\left(B_{d-1}\left(\cdot, \frac{r-2r_\xi}{L}\right)\right)}{\lambda\left(H_1 \cap (B(Z_1, \frac{r-2r_\xi}{L}) \cup B(Z_2, \frac{r-2r_\xi}{L})\right)} &\leq 66\kappa' r\sqrt{d+1} \\ &=: \delta \end{aligned} \tag{46}$$

Our assumptions ensure $\delta \leq \frac{1}{2}$, i.p. $(1-\delta) \geq (1+2\delta)^{-1}$. Using Lemma 1, we conclude from (43), (44) and (46)

$$\begin{aligned}A &\geq q_d\left(L\frac{\|Z_1 - Z_2\|}{r-2r_\xi}\right)(1+2\delta)^{-1} \\ &\geq q_d\left(L\frac{\|M_1 - M_2\| + 2r_\xi}{r-2r_\xi}\right)(1+2\delta)^{-1} \\ &= q_d(s)\left(\frac{q_d(s)}{q_d(L\frac{sr+2r_\xi}{r-2r_\xi})}\right)^{-1}(1+2\delta)^{-1}\end{aligned}$$

Using that the absolute value of the derivative of q_d is bounded by $\frac{2}{\mathcal{B}\left(\frac{d+1}{2},\frac{1}{2}\right)}$, we get

$$\begin{aligned}
\frac{q_d(s)}{q_d(L\frac{sr+2r_\xi}{r-2r_\xi})} &\leq 1+\frac{\frac{2}{\mathcal{B}\left(\frac{d+1}{2},\frac{1}{2}\right)}\left(L\frac{sr+2r_\xi}{r-2r_\xi}-s\right)}{q_d(L\frac{sr+2r_\xi}{r-2r_\xi})}\\
&\leq 1+2\frac{L\frac{sr+2r_\xi}{r-2r_\xi}-s}{q_d(b')\mathcal{B}\left(\frac{d+1}{2},\frac{1}{2}\right)}\\
&= 1+\frac{2}{q_d(b')\mathcal{B}\left(\frac{d+1}{2},\frac{1}{2}\right)}\frac{(L-1)sr+2Lr_\xi+2sr_\xi}{r-2r_\xi}\\
&\leq 1+\frac{2}{q_d(b')\mathcal{B}\left(\frac{d+1}{2},\frac{1}{2}\right)}\left(6\frac{r_\xi}{r}+3(L-1)\right)\\
&\overset{(16)}{\leq} 1+\frac{2(d+1)}{\left(1-\frac{b'^2}{4}\right)^{\frac{d+1}{2}}}\left(6\frac{r_\xi}{r}+3(L-1)\right)
\end{aligned}$$

Next, let us consider B. According to the upper bound (43) we have

$$P_*^{-1}\mathbb{P}_0\left(\mathcal{T}\cap\left(B(Z_1,\frac{r-2r_\xi}{L}\right)\cup B\left(Z_2,\frac{r-2r_\xi}{L}\right)\right)\geq\frac{1}{4\lambda(\mathcal{M})}\lambda\left(B(\cdot,\frac{r-2r_\xi}{L})\right)$$

Consequently,

$$\begin{aligned}
B &= \frac{P_*^{-1}\mathbb{P}_0\left(\mathcal{T}\cap(B(Z_1,\frac{r-2r_\xi}{L})\cup B(Z_2,\frac{r-2r_\xi}{L}))\right)}{P_*^{-1}\mathbb{P}_0\left(\mathcal{T}\cap(B(Z_1,r+2r_\xi)\cup B(Z_2,r+2r_\xi))\right)}\\
&\geq \left(1+8\frac{\lambda(B(\cdot,r+2r_\xi)\setminus B(\cdot,\frac{r-2r_\xi}{L}))}{\lambda(B(\cdot,\frac{r-2r_\xi}{L}))}\right)^{-1}\\
&= \left(1+8\frac{(r+2r_\xi)^d-\left(\frac{r-2r_\xi}{L}\right)^d}{\left(\frac{r-2r_\xi}{L}\right)^d}\right)^{-1}\\
&= \left(1+8\left(\frac{L(r+2r_\xi)}{r-2r_\xi}\right)^d-8\right)^{-1}
\end{aligned}$$

Putting everything together, we end up

$$\begin{aligned}
q_{\mathbb{P}} &\geq L^{-d}AB\\
&\geq q_d(s)L^{-d}(1+2\delta)^{-1}\left(1+\frac{2(d+1)}{\left(1-\frac{b'^2}{4}\right)^{\frac{d+1}{2}}}\left(6\frac{r_\xi}{r}+3(L-1)\right)\right)^{-1}\left(1+8\left(\frac{L(r+2r_\xi)}{r-2r_\xi}\right)^d-8\right)^{-1}
\end{aligned}$$

where

$$\delta = 66\kappa' r\sqrt{d+1}$$

The last two factors can be lower bounded as follows

$$\left(1+\frac{2(d+1)}{\left(1-\frac{b'^2}{4}\right)^{\frac{d+1}{2}}}\left(6\frac{r_\xi}{r}+3(L-1)\right)\right)^{-1} \geq \left(1+\frac{12(d+1)\frac{r_\xi}{r}}{\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}\right)^{-1}\left(1+\frac{2160(d+1)(\kappa r)^2}{\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}\right)^{-1}$$

$$\left(1+8\left(\frac{L(r+2r_\xi)}{r-2r_\xi}\right)^d-8\right)^{-1} \geq \left(1+8(L^d-1)\right)^{-1}\left(1+8\left(\left(\frac{r+2r_\xi}{r-2r_\xi}\right)^d-1\right)\right)^{-1}$$

We reorder the factors of the resulting lower bound by variables and get

$$q_{\mathbb{P}} \geq q_d(s) A_{\mathcal{M}} A_{\partial\mathcal{C}} A_\xi$$

with

$$A_{\mathcal{M}} = L^{-d}\left(1+8(L^d-1)\right)^{-1}\left(1+\frac{2160(d+1)(\kappa r)^2}{\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}\right)^{-1}$$

$$A_{\partial\mathcal{C}} = \left(1+132\kappa' r\sqrt{d+1}\right)^{-1}$$

$$A_\xi = \left(1+8\left(\left(\frac{r+2r_\xi}{r-2r_\xi}\right)^d-1\right)\right)^{-1}\left(1+\frac{12(d+1)\frac{r_\xi}{r}}{\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}\right)^{-1}$$

Using the inequality $(1+x)^d \leq 1+2xd$ for $0<x\leq\frac{1}{d}$, we get

$$L^{-d}\left(1+8(L^d-1)\right)^{-1} \geq 1+11520d\kappa^2 r^2$$

Using the inequalities $760\kappa^2 r^2(d+1)\leq 1$ and $\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}} \leq \frac{3}{4}$ we can simplifify

$$A_{\mathcal{M}} \geq \left(1+\frac{45360(d+1)(\kappa r)^2}{\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}\right)^{-1}$$

Next, we discuss the term A_ξ. Since $\frac{r_\xi}{r}\leq\frac{1}{10}$, we have $\frac{r+2r_\xi}{r-2r_\xi}\leq 1+5\frac{r_\xi}{r}$. In view of $\frac{r_\xi}{r}\leq\frac{1}{5d}$ this implies analogously

$$\left(1+8\left(\left(\frac{r+2r_\xi}{r-2r_\xi}\right)^d-1\right)\right)^{-1} \geq \left(1+80d\frac{r_\xi}{r}\right)^{-1}$$

Using again $\frac{r_\xi}{r}\leq\frac{1}{5d}$ and $\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}\leq\frac{3}{4}$, we simplify

$$A_{\partial\mathcal{C}} \geq \left(1+\frac{264(d+1)\frac{r_\xi}{r}}{\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}\right)^{-1}$$

The final result is

$$q_{\mathbb{P}} \geq q_d(s)\left(1+\epsilon_{\mathcal{M}}\right)^{-1}\left(1+\epsilon_\xi\right)^{-1}\left(1+\epsilon_{\partial\mathcal{C}}\right)^{-1}$$

□

Proof of Theorem 3. Again, we can follow the proof of [10, Theorem 3.1]. It relies only on the inequality $\theta_{ij}^{(k)} \geq q_{ij}^{(k)}$ for $\|X_i - X_j\| \leq h_k$. This is ensured by Proposition 3 and the construction of the adjusted volume coefficient. □

Proof of Corollary 3. This result combines Theorem 2 and Theorem 3. Note that for the proof of Theorem 2, we also need to consider the modification of the adjusted volume coefficient from

$$\mathfrak{q}_{ij}^{(k)} = (1+\varepsilon_{\mathcal{M}})^{-1}(1+\varepsilon_{\xi})^{-1} q_d\left(\frac{\|X_i - X_j\|}{h_{k-1}}\right)$$

to

$$\mathfrak{q}_{ij}^{(k)} = (1+\epsilon_{\mathcal{M}})^{-1}(1+\epsilon_{\xi})^{-1}\,(1+\epsilon_{\partial\mathcal{C}})^{-1}\, q_d\left(\frac{\|X_i - X_j\|}{h_{k-1}}\right).$$

However, our assumption

$$\varepsilon \geq 7\,(1+\epsilon_{\mathcal{M}})\,(1+\epsilon_{\xi})\,(1+\epsilon_{\partial\mathcal{C}}) - 7$$

ensures that inequality (17) is still valid. So the results from both theorems are valid under the considered assumptions. Application of the union bound leads to the final result. □

References

[1] E. Aamari, J. Kim, F. Chazal, B. Michel, A. Rinaldo, and L. Wasserman. Estimating the reach of a manifold. *Electron. J. Stat.*, 13(1):1359–1399, 2019.

[2] E. Aamari and C. Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177 – 204, 2019.

[3] E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local PCA. *J. Mach. Learn. Res.*, 18:Paper No. 9, 57, 2017.

[4] S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, and L. Wasserman. Cluster trees on manifolds. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[5] S. Balakrishnan, A. Rinaldo, D. Sheehy, A. Singh, and L. Wasserman. Minimax rates for homology inference. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 64–72, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[6] T. Barton. Clustering benchmarks, 5th November 2019.

[7] J.-D. Boissonnat, F. Chazal, and M. Yvinec. *Geometric and topological inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2018.

[8] K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[9] A. Cuevas, R. Fraiman, and B. Pateiro-López. On statistical properties of sets fulfilling rolling-type conditions. *Adv. in Appl. Probab.*, 44(2):311–329, 2012.

[10] K. Efimov, L. Adamyan, and V. Spokoiny. Adaptive nonparametric clustering. *IEEE Trans. Inform. Theory*, 65(8):4875–4892, 2019.

[11] J. Eldridge, M. Belkin, and Y. Wang. Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 588–606, Paris, France, 03–06 Jul 2015. PMLR.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[13] H. Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

[14] G. B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999. Modern techniques and their applications, A Wiley-Interscience Publication.

[15] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *J. Mach. Learn. Res.*, 13(1):1263–1291, may 2012.

[16] D. Gong, F. Sha, and G. Medioni. Locally linear denoising on image manifolds. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 265–272, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[17] J. A. Hartigan. Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.*, 76(374):388–394, 1981.

[18] M. Hein and M. Maier. Manifold denoising. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 561–568. MIT Press, 2007.

[19] H. Jiang. Density level set estimation on manifolds with DBSCAN. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1684–1693. PMLR, 06–11 Aug 2017.

[20] J. Kim, A. Rinaldo, and L. A. Wasserman. Minimax rates for estimating the dimension of a manifold. *JoCG*, 10:42–95, 2016.

[21] S. Kpotufe and U. von Luxburg. Pruning nearest neighbor cluster trees. pages 225–232, Madison, WI, USA, July 2011. International Machine Learning Society.

[22] S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian J. Math. Stat.*, 4(1):66–70, 2011.

[23] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 849–856, Cambridge, MA, USA, 2001. MIT Press.

[24] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 39(1-3):419–441, Mar. 2008.

[25] P. Niyogi, S. Smale, and S. Weinberger. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40(3):646–663, 2011.

[26] S. Osher, Z. Shi, and W. Zhu. Low dimensional manifold model for image processing. *SIAM J. Imaging Sci.*, 10(4):1669–1690, 2017.

[27] K. Pearson. *Tables of the incomplete beta-function*. Originally prepared under the direction of and edited by Karl Pearson. Second edition with a new introduction by E. S. Pearson and N. L. Johnson. Published for the Biometrika Trustees at the Cambridge University Press, London, 1968.

[28] G. Peyré. Manifold models for signals and images. *Comput. Vis. Image Underst.*, 113(2):249–260, Feb. 2009.

[29] J. Polzehl and V. Spokoiny. Propagation-separation approach for local likelihood estimation. *Probab. Theory Related Fields*, 135(3):335–362, 2006.

[30] N. Puchkin and V. Spokoiny. Structure-adaptive manifold estimation, 2019.

[31] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[32] P. Rigollet. Generalized error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, 8:1369–1392, 2007.

[33] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Learning Internal Representations by Error Propagation, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[34] H. Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci. Cl. III.*, 4:801–804 (1957), 1956.

[35] C. Thäle. 50 years sets with positive reach—a survey. *Surv. Math. Appl.*, 3:123–165, 2008.

[36] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[37] D. Wang, X. Lu, and A. Rinaldo. Dbscan: Optimal rates for density-based cluster estimation. *Journal of Machine Learning Research*, 20(170):1–50, 2019.

[38] W. Wang and M. Á. Carreira-Perpiñán. Manifold blurring mean shift algorithms for manifold denoising. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1759–1766, 2010.

[39] D. Xu and Y. Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, Jun 2015.

[40] H. Yin. Nonlinear dimensionality reduction and data visualization: A review. *International Journal of Automation and Computing*, 4:294–303, 2007.