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

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

No. 2800
Berlin 2020

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 [6] 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 [6] is very limited and did not really address the question of efficiency. This paper makes a significant step in understanding the remarkable 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 [22]. Other well-known approaches are density-based methods like DBSCAN [7] or spectral methods [13]. For a comprehensive survey of clustering methods, we refer to [26]. In this paper, we study a nonparametric clustering algorithm originated from [6] 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. Informally speaking, the objective of the algorithm is to find maximal subsets of 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 [6] 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

DOI 10.20347/WIAS.PREPRINT.2800 Berlin 2020

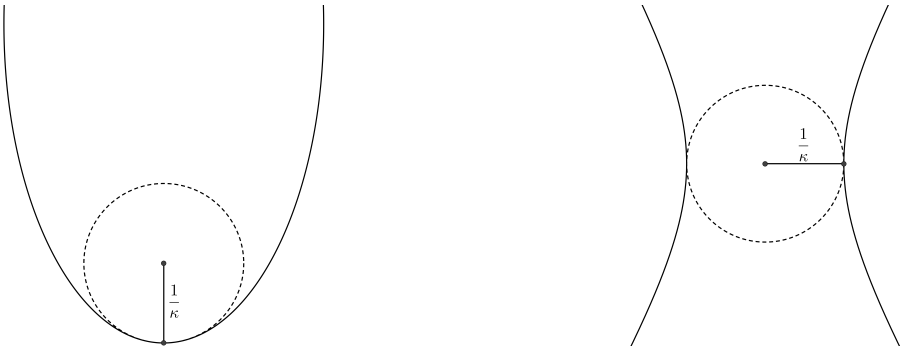

Figure 1: The reach of a manifold can be either attained by the curvature radius of a geodesic (left) or the distance to a bottleneck (right)

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 $\mathcal{M}$. It has been shown that this is a realistic model for various data, e.g. for images which are represented in a patch space [16, 14] and a wide range of algorithms have been proposed to deal with the problem of non-linear dimension reduction [27], e.g. multidimensional scaling (MDS), kernel PCA, Isomap, Laplacian eigenmaps, self-organizing maps (SOM), locally-linear embeddings and autoencoders [21]. In this work, we will not rely on any of these techniques, however, we recommend using a manifold denoising algorithm in practice such as [18] 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*

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

Originally introduced by [8], a positive reach has proven to be a widely used minimal condition in geometric and topological inference [4]. 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 [5]. 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 [23].

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

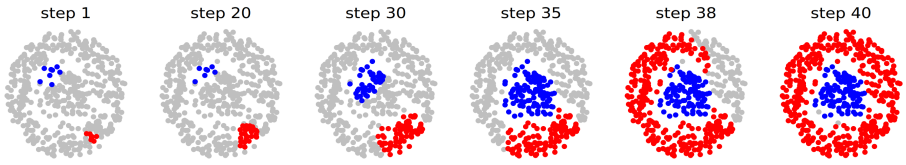

Figure 3: Local clusters during different steps of the AWC algorithm

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 [17]. Given a sequence of radii $0 < h_0 < \cdots < h_K$ in addition to our data $X_1, \ldots, X_n \in \mathbb{R}^D$ 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* $\mathcal{C}_i^{(k)}$ 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 $X_1, \ldots, X_n \in \mathbb{R}^D$ 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

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

against the alternative

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

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 writte 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. 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 cluster objective of AWC can be seen as a disadvantage from a theoretical 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.

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: 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 $X_1, \ldots, X_n$ 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 $n\epsilon_n^2 (\log n)^{\breve{}1}$ 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 in 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

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 $\left(w_{ij}^{(K)} \right)_{i,j=1}^{n}$

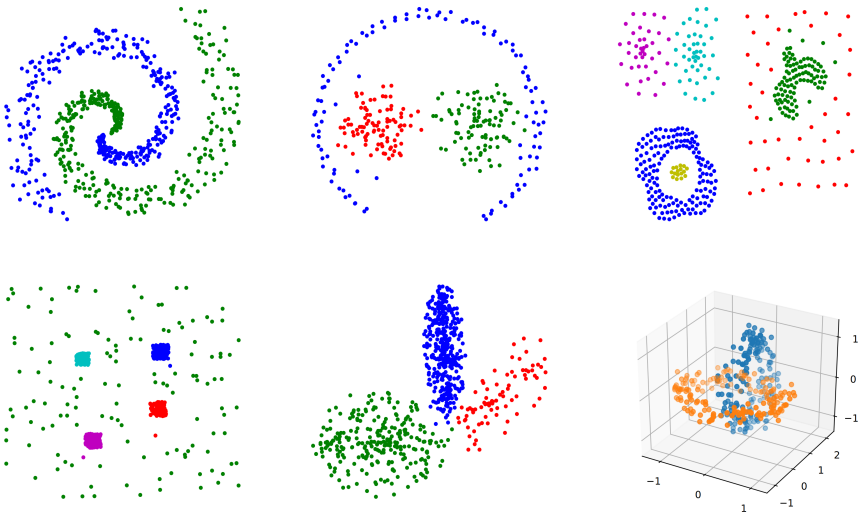

Figure 4: Six artificial examples demonstrate the adaptivity of AWC w.r.t. clusters of different size and density, non-convex shapes and clusters with manifold structure. The top left and the bottom right examples are original data sets, the rest are taken from [3].

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 solving 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.

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 the algorithm w.r.t. local gaps. We will show that the procedure is rate-optimal and discuss the problem of parameter tuning. 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*

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

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 [6]

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

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 Lemma. 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.

Lemma 1. *For $0 \leq s < 2$, we have*

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

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 is uniformly distributed on a manifold $\mathcal{M}$ and $\|\xi\| \leq r_\xi$
- $\mathcal{M}$ is a compact d-dimensional C^2 submanifold of $\mathbb{R}^D$ without boundary
- reach$(\mathcal{M}) \geq \frac{1}{\kappa}$ for $\kappa > 0$
- $r_\xi \leq \frac{r_0}{\max\{20,5d\}}$

- $r_1 \leq \frac{1}{\max\{48,6d\}\kappa}$
- $1 < b \leq \frac{b'}{(1+3\kappa r_1)(1+5\frac{r_\xi}{r})}$ for some $b' < 2$

Note that the upper bound for b is not a very restrictive assumption. 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 $b' > \frac{3}{2}$, this does not change the overall complexity.

Lemma 2. *Suppose assumptions A(r, r) are satisfied 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

$$\varepsilon_{\mathcal{M}} := \frac{84\kappa(d+1)r}{\left(1-\left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}$$

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 $(1+\mathcal{O}(\kappa r))\left(1+\mathcal{O}\left(\frac{r_\xi}{r}\right)\right)$, as long as the intrinsic dimension d is bounded and as long as b' is bounded away from 1. 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 [10, 11, 25, 18], 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 Lemma 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, Lemma 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

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

by considering an *adjusted volume coefficient*

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

Note that in practice, the parameters d, $\frac{1}{\kappa}$ and r_ξ are unknown. We refer to [12] for an overview of procedures dedicated to estimate 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 [18]. 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 estimation of the dimension is crucial and cannot be ignored.

2.2 Propagation in the uniform case

In the following, we generalize the results from [6] 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. *Suppose assumptions A(h_{k-1}, h_{k-1}) hold and $X_1, X_2, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$. 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*

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

Corollary 2. *Suppose assumptions A(h_0, h_K) hold and $X_1, X_2, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$. If $K < n$ and we choose the threshold $\lambda = C \log n$ for $C > 3$, then*

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

2.3 Separation in the gap case

For the case of a significant gap in the data, we can also generalize the results of [6] 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):

- Suppose assumptions A(r, r) hold except that X does not need to be uniformly distributed on the considered manifold $\mathcal{M}$, but follows some Lebesgue density f whose support is a subset of $\mathcal{M}$
- $C_1, \dots, C_k$, are disjoint subsets of $\mathcal{M}$
- Spatial separation of clusters is ensured by

$$d_\infty(C_{i'}, C_{j'}) := \min_{x \in \mathcal{C}_{i'}, y \in \mathcal{C}_{j'}} \|x - y\| \geq r + 2r_\xi \quad \text{for} \quad 1 \leq i' \neq j' \leq k$$

- Similarly as in [20], we assume a thickness condition on each cluster: We assume there is a constant $f_0 > 0$ s.t. for any $x \in \mathcal{C}_i$ and $r' < r$ 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:

$$\text{ess sup}_{\mathcal{M}\setminus\cup C_i} f \leq (1-\epsilon) f_0$$

- The sample size n has to be large enough, i.e.

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

for $z_k := \mathbb{P}(B(x_i, r) \cup B(x_j, r))$ and some $\beta > 0$.

- 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

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

for some $\alpha > \beta$.

Theorem 2. *Consider the assumptions B(h_{k-1}) and $X_1, X_2, \ldots, X_{n+2} \overset{i.i.d.}{\sim} \mathbb{P}$. Suppose x_i and x_j have a distance of at most h_k and are r_ξ -close to two different clusters. We suppose that the algorithm did not detect any gaps in the previous steps. Then*

$$\mathbb{P}^{\otimes(n+2)}\left(T_{ij}^{(k)} \geq \left(\sqrt{\alpha} - \sqrt{\beta}\right)^2 \log n \middle| X_i = x_i, X_j = x_j\right) \geq 1 - 3n^{-\beta}.$$

Remark 1. *Under the assumptions above, the gap will be consistently detected at 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

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

Theorem 2 guarantees consistent separation as long as $\epsilon^2 \gtrsim \frac{\log n}{n}$. This rate turns out to be nearly optimal if we consider a noiseless setup and a density that is piecewise constant. To be precise, it is impossible for any algorithm to consistently detect the gap if ε decreases at the rate $n^{-\frac{1}{2}}$.

Assumptions C:

- $C_1, \ldots, C_k$, are disjoint subsets of a manifold $\mathcal{M} \subset \mathbb{R}^D$

- $X_1, \dots, X_n$ are drawn i.i.d. from a density supported on $\mathcal{M}$ that is constant on $V := \cup C_i$ with value f_V and constant on $G := \mathcal{M} \setminus V$ with value f_G

Theorem 3. *Let assumptions C 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 [6] 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.

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 $X_1, \dots, X_n$ i.i.d. from

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

cf. Figure 5. To measure the performance of the algorithm we use a modified version of the Rand index

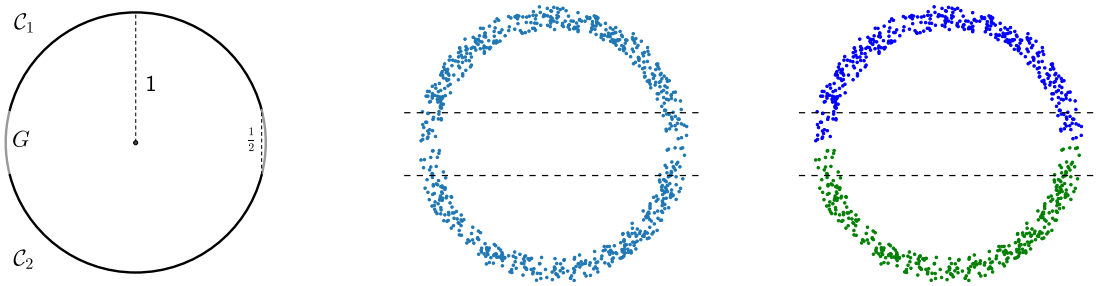

Figure 5: Density f_ϵ (left), i.i.d. sample of size $n = 800$ from $\mathbb{P}_{\frac{1}{2}}^{\mathcal{U}(\frac{1}{10})}$ with two dashed lines highlighting the gap in the data (center) and clusters obtained via AWC (right)

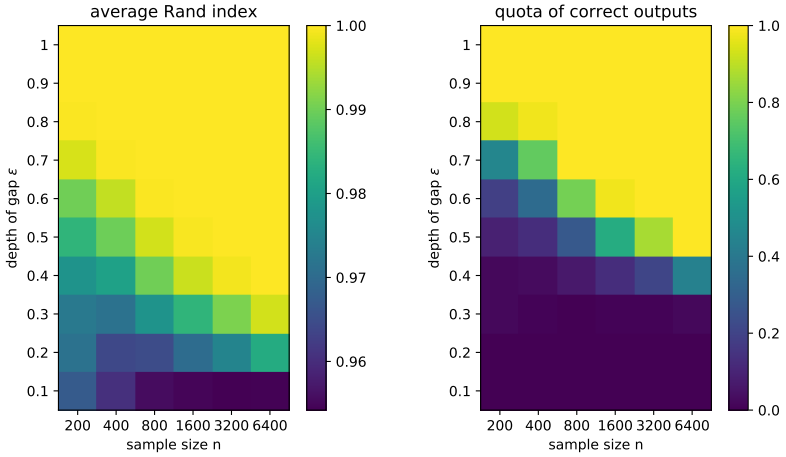

Figure 6: Average rand index (left) and quota of experiments yielding a rand index 1 (right)

[19]

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

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 $h_i = 2^{\frac{i}{2}-2}$, $i = 0, \dots, 4$. 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 gap coefficient. For each sample, 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 6 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 Parameter tuning

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 7. The results support our proposition that λ should be scaled logarithmically w.r.t. the data size.

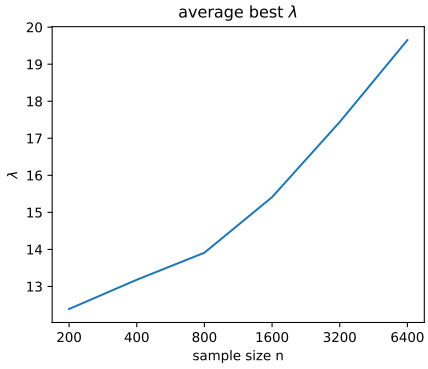

Figure 7: Average minimal lambda with best rand index for $\epsilon = 0.9$

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 from 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 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 $X_1, \dots, X_n$ i.i.d. either from

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

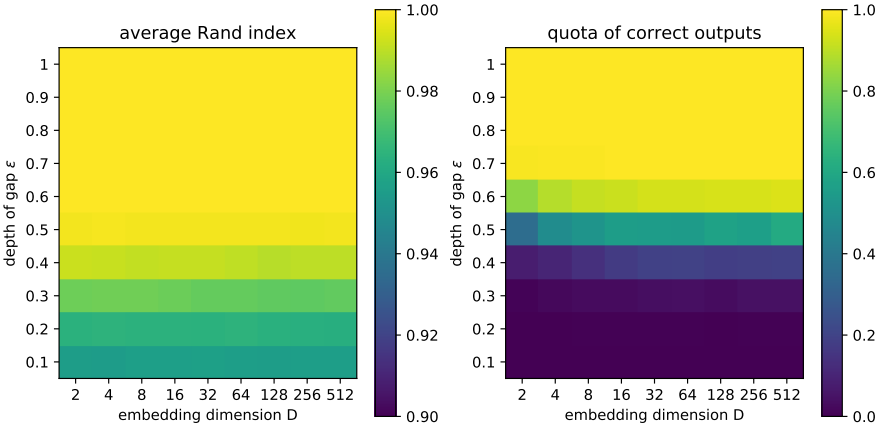

Figure 8: Average rand index (left) and quota of experiments yielding a rand index 1 (right) for uniform noise of norm $\leq \frac{1}{10}$

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.

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 8 and 9 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 8. 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.

4 Proofs

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

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

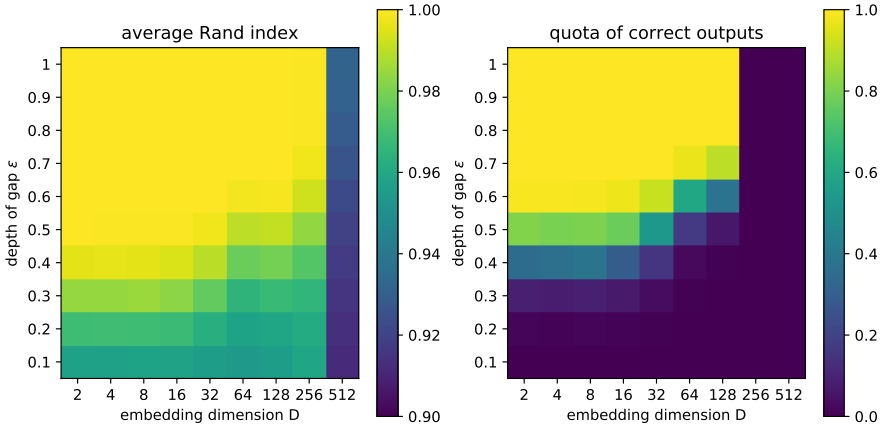

Figure 9: Average rand index (left) and quota of experiments yielding a rand index 1 (right) for Gaussian noise of variance $\frac{1}{3200} I_D$

for the incomplete beta function [15]. Also, we use the logarithmic convexity of the gamma function. For the upper bound we get

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

and similarly, we compute the lower bound

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

□

For the proof of Lemma 2 we will use the following two auxiliary Lemmas. By 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*

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

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 [2]. As the Lebesgue measure is related by a constant factor [9], it also holds for the Lebesgue measure. □

For the second auxiliary Lemma we consider a C^2 submanifold $\mathcal{M} \in \mathbb{R}^D$ with reach $\frac{1}{\kappa} > 0$ and for some fixed $x \in \mathcal{M}$ we denote the tangent plane of $\mathcal{M}$ at x by $\mathcal{T}$.

Lemma 4. *For all $0 < r \leq \frac{1}{16\kappa}$ the restriction of the projection $P : \mathbb{R}^D \to \mathcal{T}$ associating each $y \in \mathbb{R}^D$ with the closest point in $\mathcal{T}$ to $\mathcal{M} \cap B(x, r)$ is a 1-Lipschitz injection and its image contains $\mathcal{T} \cap B(x, r/L)$, whereas its inverse is L -Lipschitz for $L := 1 + \kappa r$.*

Proof. This lemma is given in [2] with some unspecified small enough constant instead of $\frac{1}{16}$. Following the corresponding proof, it can be easily verified that this constant is indeed small enough. □

Proof of Lemma 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

$$\begin{aligned} q_l &:= \frac{\int \mathbb{1}_{B(M_1', r-2r_\xi)\cap B(M_2', r-2r_\xi)} d\mu}{\int \mathbb{1}_{B(M_1', r+2r_\xi)\cup B(M_2', r+2r_\xi)} d\mu} \\ &\leq q_{\mathbb{P}} \\ &\leq \frac{\int \mathbb{1}_{B(M_1', r+2r_\xi)\cap B(M_2', r+2r_\xi)} d\mu}{\int \mathbb{1}_{B(M_1', r-2r_\xi)\cup B(M_2', r-2r_\xi)} d\mu} \\ &=: q_u \end{aligned} \tag{3}$$

Let us denote by $\pm$ 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 $L := 1 + 3\kappa r$, we conclude (cf. [2])

$$L^{-d} \leq \frac{\mathsf{vol}(P(\mathcal{M} \cap (B(M_1', r') \pm B(M_2', r'))))}{\mathsf{vol}(\mathcal{M} \cap (B(M_1', r') \pm B(M_2', r')))} \leq 1. \tag{4}$$

Moreover, the above Lipschitz constants imply

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

for $i = 1, 2$ and therefore

$$\begin{aligned} 1 &\leq \frac{\mathsf{vol}\left(\mathcal{T} \cap (B(P(M_1'), r') \pm B(P(M_2'), r'))\right)}{\mathsf{vol}\left(P(\mathcal{M} \cap (B(M_1', r') \pm B(M_2', r')))\right)} \\ &\leq \frac{\mathsf{vol}\left(\mathcal{T} \cap (B(P(M_1'), r') \pm B(P(M_2'), r'))\right)}{\mathsf{vol}\left(\mathcal{T} \cap (B(P(M_1'), \frac{r'}{L}) \pm B(P(M_2'), \frac{r'}{L}))\right)} =: q_{\pm, r'}. \end{aligned} \tag{5}$$

Note also that according to our assumptions, any intersections encountered so far are nonempty. From (4) and (5) we conclude

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

and obtain

$$q_{\pm,r'}^{-1} \leq \frac{\mathsf{vol}\left(\mathcal{M} \cap (B(M_1', r') \pm B(M_2', r'))\right)}{\mathsf{vol}\left(\mathcal{T} \cap (B(P(M_1'), r') \pm B(P(M_2'), r'))\right)} \leq L^d. \tag{6}$$

In particular, considering $(\pm, r') = (\cap, r + 2r_\xi)$ and $(\pm, r') = (\cup, r - 2r_\xi)$ in (6), we get

$$q_u \leq q_{\cup, r-2r_\xi} L^d q_\cup q_{r+2r_\xi}, \tag{7}$$

where $q_{r'}$ is defined as

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

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

$$q_l \geq q_{\cap, r-2r_\xi}^{-1} L^{-d} q_\cup^{-1} q_{r-2r_\xi}. \tag{8}$$

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

$$\begin{aligned}
q_{r+2r_\xi} &\leq q_d\left(\frac{\min\{0, \|M_1 - M_2\| - 2r_\xi\}}{L(r + 2r_\xi)}\right) \\
&\leq q_d(s) + \frac{2}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\left(s - \frac{\|M_1 - M_2\| - 2r_\xi}{L(r + 2r_\xi)}\right) \\
&= q_d(s) + \frac{2}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\left(\frac{sr(L-1)}{L(r + r2_\xi)} + \frac{2sr_\xi}{r + 2r_\xi} + \frac{2r_\xi}{L(r + 2r_\xi)}\right) \\
&\leq q_d(s) + \frac{12\kappa r}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} + \frac{12\frac{r_\xi}{r}}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} \\
&\leq q_d(s)\left(1 + \frac{12\kappa r}{q_d(b')\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\right)\left(1 + \frac{12\frac{r_\xi}{r}}{q_d(b')\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\right).
\end{aligned} \tag{9}$$

Similarly, we obtain

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

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

$$q_{\cup,r'} \leq \frac{\mathsf{vol}\left(\mathcal{T} \cap B(x, \frac{r'}{L})\right) + 2\mathsf{vol}\left(\mathcal{T} \cap (B(x,r') \setminus B(x, \frac{r'}{L})\right)}{\mathsf{vol}\left(\mathcal{T} \cap B(x, \frac{r'}{L})\right)} = 2L^d - 1. \tag{11}$$

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

$$\begin{aligned} q_\cup &\leq \left(2\left(\frac{r+2r_\xi}{r-2r_\xi}\right)^d - 1\right) \\ &\leq \left(2\left(1+\frac{5r_\xi}{r}\right)^d - 1\right) \\ &\leq \left(1 + \frac{20dr_\xi}{r}\right) \end{aligned} \tag{12}$$

and

$$L^d\left(2L^d - 1\right) \leq 1 + 24\kappa r d. \tag{13}$$

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

$$\begin{aligned} q_{\cap,r'} &= q_{\cup,r'} \frac{q_d\left(s'\right)}{q_d\left(s'L\right)} \\ &\leq (2L^d - 1) \frac{q_d(s'L) + s'(L-1)\frac{2}{\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}}{q_d(s'L)} \\ &\leq (2L^d - 1)\left(1 + \frac{12\kappa r}{q_d(b')\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)}\right). \end{aligned} \tag{14}$$

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 [15]

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

we get

$$\begin{aligned} \frac{1}{q_d(b')\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} &= \frac{2\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right) - \mathcal{B}\left(1 - \left(\frac{b'}{2}\right)^2, \frac{d+1}{2}, \frac{1}{2}\right)}{\mathcal{B}\left(1 - \left(\frac{b'}{2}\right)^2, \frac{d+1}{2}, \frac{1}{2}\right)\mathcal{B}\left(\frac{d+1}{2}, \frac{1}{2}\right)} \\ &\leq \frac{2}{\mathcal{B}\left(1 - \left(\frac{b'}{2}\right)^2, \frac{d+1}{2}, \frac{1}{2}\right)} \\ &\leq \frac{d+1}{\left(1 - \left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}. \end{aligned} \tag{15}$$

Finally, putting (3), (7), (8), (9), (10), (11), (12), (13), (14) and (15) together, we obtain

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

for

$$M := (1 + 24\kappa dr)\left(1 + \frac{12\kappa(d+1)r}{\left(1 - \left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}\right)\left(1 + 20\frac{dr_\xi}{r}\right)\left(1 + \frac{12(d+1)\frac{r_\xi}{r}}{\left(1 - \left(\frac{b'}{2}\right)^2\right)^{\frac{d+1}{2}}}\right).$$

According to our assumptions, both $24\kappa dr$ 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 [6, 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 Theorem 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. For $l = i, j$ we choose a point X_l' of distance of most r_ξ to a cluster $\mathcal{C}_{k_l}$ for $k_i \neq k_j$. Our assumptions imply that the density in the overlap $B(X_i', h_{k-1}+2r_\xi) \cap B(X_j', h_{k-1}+2r_\xi) \cap S$ 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 $\mathbb{P}_\epsilon$. Hence,

$$\begin{aligned}
\theta_{ij}^{(k)} &\leq \frac{\mathbb{P}_\epsilon(B(M_1', r + 2r_\xi) \cap B(M_2', r + 2r_\xi))}{\mathbb{P}_\epsilon(B(M_1', r - 2r_\xi) \cup B(M_2', r - 2r_\xi))} \\
&\leq \frac{(1-\epsilon)f_0 A}{(1-\epsilon)f_0 B + \epsilon f_0 C} \\
&= \frac{A}{B}\left(1 - \frac{\epsilon C}{(1-\epsilon)B + \epsilon C}\right)
\end{aligned}$$

with

$$\begin{aligned}
A &= \mu(B(M_1', r + 2r_\xi) \cap B(M_2', r + r2_\xi)), \\
B &= \mu(B(M_1', r - 2r_\xi) \cup B(M_2', r - r2_\xi)) \\
\text{and } C &= \mu(B(M_1', r - 2r_\xi)) + \mu(B(M_2', r - 2r_\xi)).
\end{aligned}$$

The first factor of the latter $\frac{A}{B}$ is bounded from above by $Mq_{ij}^{(k)}$ by Theorem 2. Moreover, $B < C$ implies that the second factor is bound from above by $1 - \epsilon$, providing the upper bound

$$\theta_{ij}^{(k)} \leq (1-\epsilon)Mq_{ij}^{(k)}.$$

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

$$\begin{aligned}
q_{ij}^{(k)} - \theta_{ij}^{(k)} &\geq ((1 + \varepsilon_{\mathcal{M}})^{-1}(1 + \varepsilon_\xi)^{-1} - (1-\epsilon)(1 + \varepsilon_{\mathcal{M}})(1 + \varepsilon_\xi))q_d(b) \\
&\geq \left(\left(1 + \frac{\epsilon}{7}\right)^{-1} - (1-\epsilon)\left(1 + \frac{\epsilon}{7}\right)\right)q_d(b) \\
&\geq \epsilon\frac{q_d(b)}{\sqrt{2}}. \qquad (16)
\end{aligned}$$

Using Pinsker's inequality, we get

$$\mathcal{K}\left(q_{ij}^{(k)}, \theta_{ij}^{(k)}\right) \geq \epsilon^2 q_d(b)^2. \qquad (17)$$

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

$$2\delta^2 n \geq \beta \log n \tag{18}$$

$$\text{and } \delta n \leq \frac{z_k n}{2}. \tag{19}$$

For the following we always assume implicitely that we condition on $X_i = x_i$ and $X_j = x_j$. By Hoeffding's inequality and (18) we know

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

with probability at least $1 - n^{-\beta}$. This implies together with (19)

$$N_{i\vee j}^{(k)} \geq \frac{z_k n}{2} \tag{20}$$

with probability at least $1 - n^{-\beta}$. On the other hand, by [6, Lemma 5.1] we have

$$\mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)}) < \frac{\beta \log n}{N_{i\vee j}^{(k)}} \tag{21}$$

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 (20) and (21) hold. In the following let us fix an outcome of the event E. Then (20) and (21) imply

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

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

$$\mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)}) < \frac{\beta}{\alpha}\epsilon^2 q_d(b)^2. \tag{22}$$

Note that (16) implies in particular $q_{ij}^{(k)} > \theta_{ij}^{(k)}$. Since the function $\mathcal{K}(\cdot, \theta)$ is strictly monotone on the interval $[\theta, 1)$ and considering $\frac{\beta}{\alpha} < 1$, we conclude from (17) and (22)

$$\widetilde{\theta}_{ij}^{(k)} < q_{ij}^{(k)}. \tag{23}$$

The triangle inequality and Pinsker's inequality yield

$$\begin{aligned} |\widetilde{\theta}_{ij}^{(k)} - q_{ij}^{(k)}| &\geq |\theta_{ij}^{(k)} - q_{ij}^{(k)}| - |\widetilde{\theta}_{ij}^{(k)} - \theta_{ij}^{(k)}| \\ &\geq \epsilon \frac{q_d(b)}{\sqrt{2}} - \sqrt{\frac{1}{2}\mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, \theta_{ij}^{(k)})} \\ &\overset{(22)}{\geq} \epsilon \frac{q_d(b)}{\sqrt{2}} \left(1 - \sqrt{\frac{\beta}{\alpha}}\right) \end{aligned} \tag{24}$$

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

$$\begin{aligned} \mathcal{K}(\widetilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)}) &\geq 2(\widetilde{\theta}_{ij}^{(k)} - q_{ij}^{(k)})^2 \\ &\overset{(24)}{\geq} \epsilon^2 q_d(b)^2 \left(1 - \sqrt{\frac{\beta}{\alpha}}\right)^2 \\ &\geq \frac{\log n}{N} 2\alpha \left(1 - \sqrt{\frac{\beta}{\alpha}}\right)^2 \\ &\overset{(20)}{\geq} \frac{\log n}{N_{i\vee j}^{(k)}} \left(\sqrt{\alpha} - \sqrt{\beta}\right)^2 \end{aligned} \tag{25}$$

Finally, putting together (23) and (25), we conclude that any outcome of the event E satisfies

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

□

Proof of Theorem 3. 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 [24, 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 $n \to \infty$. □

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

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

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

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] 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.

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

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

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

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

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

[21] 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.

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

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

[24] 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.

[25] 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.

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

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