Optimal construction of *k*-nearest-neighbor graphs for identifying noisy clusters

Markus Maier [a,*], Matthias Hein [b], Ulrike von Luxburg [a]

[a] *Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany*
[b] *Saarland University, P.O. Box 151150, 66041 Saarbrücken, Germany*

ARTICLE INFO

ABSTRACT

We study clustering algorithms based on neighborhood graphs on a random sample of data points. The question we ask is how such a graph should be constructed in order to obtain optimal clustering results. Which type of neighborhood graph should one choose, mutual *k*-nearest-neighbor or symmetric *k*-nearest-neighbor? What is the optimal parameter *k*? In our setting, clusters are defined as connected components of the *t*-level set of the underlying probability distribution. Clusters are said to be identified in the neighborhood graph if connected components in the graph correspond to the true underlying clusters. Using techniques from random geometric graph theory, we prove bounds on the probability that clusters are identified successfully, both in a noise-free and in a noisy setting. Those bounds lead to several conclusions. First, *k* has to be chosen surprisingly high (rather of the order *n* than of the order log *n*) to maximize the probability of cluster identification. Secondly, the major difference between the mutual and the symmetric *k*-nearest-neighbor graph occurs when one attempts to detect the most significant cluster only.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Using graphs to model real world problems is one of the most widely used techniques in computer science. This approach usually involves two major steps: constructing an appropriate graph which represents the problem in a convenient way, and then constructing an algorithm which solves the problem on the given type of graph. While in some cases there exists an obvious natural graph structure to model the problem, in other cases one has much more choice when constructing the graph. In the latter cases it is an important question how the actual construction of the graph influences the overall result of the graph algorithm.

The kinds of graphs we want to study in the current paper are neighborhood graphs. The vertices of those graphs represent certain "objects", and vertices are connected if the corresponding objects are "close" or "similar". The best-known families of neighborhood graphs are ε-neighborhood graphs and *k*-nearest-neighbor graphs. Given a number of objects and their mutual distances to each other, in the first case each object will be connected to all other objects which have distance smaller than ε, whereas in the second case, each object will be connected to its *k* nearest neighbors (for exact definitions see below). Neighborhood graphs are used for modeling purposes in many areas of computer science: sensor networks and wireless ad hoc networks, machine learning, data mining, percolation theory, clustering, computational geometry, modeling the spread of diseases, modeling connections in the brain, etc.

---

* Corresponding author.
E-mail addresses: markus.maier@tuebingen.mpg.de (M. Maier), hein@cs.uni-sb.de (M. Hein), ulrike.luxburg@tuebingen.mpg.de (U. von Luxburg).

In all those applications one has some freedom in constructing the neighborhood graph, and a fundamental question arises: how exactly should we construct the neighborhood graph in order to obtain the best overall result at the end? Which type of neighborhood graph should we choose? How should we choose its connectivity parameter, for example the parameter k in the k-nearest-neighbor graph? It is obvious that those choices will influence the results we obtain on the neighborhood graph, but often it is completely unclear how.

In this paper, we want to focus on the problem of clustering. We assume that we are given a finite set of data points and pairwise distances or similarities between them. It is very common to model the data points and their distances by a neighborhood graph. Then clustering can be reduced to standard graph algorithms. In the easiest case, one can simply define clusters as connected components of the graph. Alternatively, one can try to construct minimal graph cuts which separate the clusters from each other. An assumption often made in clustering is that the given data points are a finite sample from some larger underlying space. For example, when a company wants to cluster customers based on their shopping profiles, it is clear that the customers in the company's data base are just a sample of a much larger set of possible customers. The customers in the data base are then considered to be a random sample.

In this article, we want to make a first step towards such results in a simple setting we call "cluster identification" (see next section for details). Clusters will be represented by connected components of the level set of the underlying probability density. Given a finite sample from this density, we want to construct a neighborhood graph such that we maximize the probability of cluster identification. To this end, we study different kinds of k-nearest-neighbor graphs (mutual, symmetric) with different choices of k and prove bounds on the probability that the correct clusters can be identified in this graph. One of the first results on the consistency of a clustering method has been derived by Hartigan [8], who proved "fractional consistency" for single linkage clustering.

The question we want to tackle in this paper is how to choose the neighborhood graph in order to obtain optimal clustering results. The mathematical model for building neighborhood graphs on randomly sampled points is a geometric random graph, see [12] for an overview. Such graphs are built by drawing a set of sample points from a probability measure on $\mathbb{R}^d$, and then connecting neighboring points (see below for exact definitions). Note that the random geometric graph model is different from the classical Erdős–Rényi random graph model (cf. [3] for an overview) where vertices do not have a geometric meaning, and edges are chosen independently of the vertices and independently of each other. In the setup outlined above, the choice of parameter is closely related to the question of connectivity of random geometric graphs, which has been extensively studied in the random geometric graph community. Connectivity results are not only important in clustering, but also in many other fields of computer science such as modeling ad hoc networks (e.g., [14,1,10]) or percolation theory [4]. The existing random geometric graph literature mainly focuses on asymptotic statements about connectivity, that is results in the limit for infinitely many data points. Moreover, it is usually assumed that the underlying density is uniform — the exact opposite of the setting we consider in clustering. What we would need in our context are *non-asymptotic* results on the performance of different kinds of graphs on a *finite point set* which has been drawn from highly clustered densities.

Our results on the choice of graph type and the parameter k for cluster identification can be summarized as follows. Concerning the question of the choice of k, we obtain the surprising result that k should be chosen surprisingly high, namely of the order of $O(n)$ instead of $O(\log n)$ (the latter would be the rate one would "guess" from results in standard random geometric graphs). Concerning the types of graphs, it turns out that different graphs have advantages in different situations: if one is only interested in identifying the "most significant" cluster (while some clusters might still not be correctly identified), then the mutual kNN graph should be chosen. If one wants to identify many clusters simultaneously the bounds show no substantial difference between the mutual and the symmetric kNN graph.

2. Main constructions and results

In this section we give a brief overview over the setup and techniques we use in the following. Mathematically exact statements follow in the next sections.

Neighborhood graphs. We always assume that we are given n data points $X_1, \ldots, X_n$ which have been drawn i.i.d. from some probability measure which has a density with respect to the Lebesgue measure in $\mathbb{R}^d$. As distance function between points we use the Euclidean distance, which is denoted by dist. The distance is extended to sets $A, B \subseteq \mathbb{R}^d$ via $\text{dist}(A, B) = \inf\{\text{dist}(x, y) \mid x \in A, y \in B\}$. The data points are used as vertices in an unweighted and undirected graph. By $\text{kNN}(X_j)$ we denote the set of the k nearest neighbors of X_j among $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n$. The different neighborhood graphs are defined as follows:

- ε-*neighborhood graph $G_{eps}(n, \varepsilon)$:* X_i and X_j are connected if $\text{dist}(X_i, X_j) \leq \varepsilon$,
- *symmetric k-nearest-neighbor graph $G_{sym}(n, k)$:*
  X_i and X_j are connected if $X_i \in \text{kNN}(X_j)$ or $X_j \in \text{kNN}(X_i)$,
- *mutual k-nearest-neighbor graph $G_{\mathrm{mut}}(n, k)$:*
  X_i and X_j are connected if $X_i \in \text{kNN}(X_j)$ and $X_j \in \text{kNN}(X_i)$.

Note that the literature does not agree on the names for the different kNN graphs. In particular, the graph we call "symmetric" usually does not have a special name.

Most questions we will study in the following are much easier to solve for ε-neighborhood graphs than for kNN graphs. The reason is that whether two points X_i and X_j are connected in the ε-graph only depends on $\text{dist}(X_i, X_j)$, while in the kNN

graph the existence of an edge between X_i and X_j also depends on the distances of X_i and X_j to all other data points. However, the kNN graph is the one which is mostly used in practice. Hence we decided to focus on kNN graphs. Most of the proofs can easily be adapted for the ε-graph.

The cluster model. There exists an overwhelming amount of different definitions of what clustering is, and the clustering community is far from converging to one point of view. In a sample-based setting most definitions agree on the fact that clusters should represent high density regions of the data space which are separated by low density regions. Then a straight forward way to define clusters is to use level sets of the density. Given the underlying density p of the data space and a parameter $t > 0$, we define the t-level set $L(t)$ as the closure of the set of all points $x \in \mathbb{R}^d$ with $p(x) \geq t$. Clusters are then defined as the connected components of the t-level set (where the term "connected component" is used in its topological sense and not in its graph-theoretic sense).

Note that a different popular model is to define a clustering as a partition of the whole underlying space such that the boundaries of the partition lie in a low density area. In comparison, looking for connected components of t-level sets is a stronger requirement. Even when we are given a complete partition of the underlying space, we do not yet know which part of each of the clusters is just "background noise" and which one really corresponds to "interesting data". This problem is circumvented by the t-level set definition, which not only distinguishes between the different clusters but also separates "foreground" from "background noise". Moreover, the level set approach is much less sensitive to outliers, which often heavily influence the results of partitioning approaches.

The cluster identification problem. Given a finite sample from the underlying distribution, our goal is to identify the sets of points which come from different connected components of the t-level set. We study this problem in two different settings:

The noise-free case. Here we assume that the support of the density consists of several connected components which have a positive distance to each other. Between those components, there is only "empty space" (density 0). Each of the connected components is called a cluster. Given a finite sample $X_1, \ldots, X_n$ from such a density, we construct a neighborhood graph G based on this sample. We say that **a cluster is identified in the graph** if the connected components in the neighborhood graph correspond to the corresponding connected components of the underlying density, that is all points originating in the same underlying cluster are connected in the graph, and they are not connected to points from any other cluster.

The noisy case. Here we no longer assume that the clusters are separated by "empty space", but we allow the underlying density to be supported everywhere. Clusters are defined as the connected components of the t-level set $L(t)$ of the density (for a fixed parameter t chosen by the user), and points not contained in this level set are considered as background noise. A point $x \in \mathbb{R}^d$ is called a *cluster point* if $x \in L(t)$ and *background point* otherwise. As in the previous case we will construct a neighborhood graph G on the given sample. However, we will remove points from this graph which we consider as noise. The remaining graph $\tilde{G}$ will be a subgraph of the graph G, containing fewer vertices and fewer edges than G. As opposed to the noise-free case, we now define two slightly different cluster identification problems. They differ in the way background points are treated. The reason for this more involved construction is that in the noisy case, one cannot guarantee that no additional background points from the neighborhood of the cluster will belong to the graph.

We say that **a cluster is roughly identified** in the remaining graph $\tilde{G}$ if the following properties hold:

- all sample points from a cluster are contained as vertices in the graph, that is, only background points are dropped,
- the vertices belonging to the same cluster are connected in the graph, that is, there exists a path between each two of them, and
- every connected component of the graph contains only points of exactly one cluster (and maybe some additional noise points, but no points of a different cluster).

We say that **a cluster is exactly identified in $\tilde{G}$** if

- it is roughly identified, and
- the ratio of the number of background points and the number of cluster points in the graph $\tilde{G}$ converges almost surely to zero as the sample size approaches infinity.

If all the clusters have been roughly identified, the number of connected components of the graph $\tilde{G}$ is equal to the number of connected components of the level set $L(t)$. However, the graph $\tilde{G}$ might still contain a significant number of background points. In this sense, exact cluster identification is a much stronger problem, as we require that the fraction of background points in the graph has to approach zero. Exact cluster identification is an asymptotic statement, whereas rough cluster identification can be verified on each finite sample. Finally, note that in the noise-free case, rough and exact cluster identification coincide.

The clustering algorithms. To determine the clusters in the finite sample, we proceed as follows. First, we construct a neighborhood graph on the sample. This graph looks different, depending on whether we allow noise or not:

Noise-free case. Given the data, we simply construct the mutual or the symmetric k-nearest-neighbor graph ($G_{\mathrm{mut}}(n, k)$ resp. $G_{\mathrm{sym}}(n, k)$) on the data points, for a certain parameter k, based on the Euclidean distance. Clusters are then the connected components of this graph.

Noisy case. Here we use a more complex procedure:

- As in the noise-free case, construct the mutual (symmetric) kNN graph $G_{\mathrm{mut}}(n, k)$ (resp. $G_{\mathrm{sym}}(n, k)$) on the samples.

- Estimate the density $\hat{p}_n(X_i)$ at every sample point X_i (e.g., by kernel density estimation).
- If $\hat{p}_n(X_i) < t'$, remove the point X_i and its adjacent edges from the graph (where t' is a parameter determined later). The resulting graph is denoted by $G'_{\text{mut}}(n, k, t')$ (resp. $G'_{\text{sym}}(n, k, t')$).
- Determine the connected components of $G'_{\text{mut}}(n, k, t')$ (resp. $G'_{\text{sym}}(n, k, t')$), for example by a simple depth-first search.
- Remove the connected components of the graph that are "too small", that is, which contain less than δn points (where δ is a small parameter determined later).
- The resulting graph is denoted by $\tilde{G}_{\text{mut}}(n, k, t', \delta)$ (resp. $\tilde{G}_{\text{sym}}(n, k, t', \delta)$); its connected components are the clusters of the sample.

Note that by removing the small components in the graph the method becomes very robust against outliers and "fake" clusters (small connected components just arising by random fluctuations).

Main results, intuitively. We would like to outline our results briefly in an intuitive way. Exact statements can be found in the following sections.

Result 1 (*Range of k for Successful Cluster Identification*). *Under mild assumptions, and for n large enough, there exist constants* $c_1, c_2 > 0$ *such that for any* $k \in [c_1 \log n, c_2 n]$, *all clusters are identified with high probability in both the mutual and the symmetric kNN graph. This result holds for cluster identification in the noise-free case as well as for the rough and the exact cluster identification problem (the latter seen as an asymptotic statement) in the noisy case (with different constants* c_1, c_2).

For the noise-free case, the lower bound on k has already been proven in [5], for the noisy case it is new. Importantly, in the exact statement of the result all constants have been worked out more carefully than in [5], which is very important for proving the following statements.

Result 2 (*Optimal k for Cluster Identification*). *Under mild assumptions, and for n large enough, the parameter k which maximizes the probability of successful identification of one cluster in the noise-free case has the form* $k = c_1 n + c_2$, *where* c_1, c_2 *are constants which depend on the geometry of the cluster. This result holds for both the mutual and the symmetric kNN graph, but the convergence rates are different (see Result 3). A similar result holds as well for rough cluster identification in the noisy case, with different constants.*

This result is completely new, both in the noise-free and in the noisy case. In the light of the existing literature, it is rather surprising. So far it has been well known that in many different settings the lower bound for obtaining connected components in a random kNN graph is of the order $k \sim \log n$. However, we now can see that *maximizing the probability* of obtaining connected components on a finite sample leads to a dramatic change: k has to be chosen much higher than $\log n$, namely of the order n itself. Moreover, we were surprised ourselves that this result does not only hold in the noise-free case, but can also be carried over to rough cluster identification in the noisy setting.

For exact cluster identification we did not manage to determine an optimal choice of k due to the very difficult setting. For large values of k, small components which can be discarded will no longer exist. This implies that a lot of background points are attached to the real clusters. On the other hand, for small values of k there will exist several small components around the cluster which are discarded, so that there are less background points attached to the final cluster. However, this tradeoff is very hard to grasp in technical terms. We therefore leave the determination of an optimal value of k for exact cluster identification as an open problem. Moreover, as exact cluster identification concerns the asymptotic case of $n \to \infty$ only, and rough cluster identification is all one can achieve on a finite sample anyway, we are perfectly happy to be able to prove the optimal rate in that case.

Result 3 (*Identification of the Most Significant Cluster*). *For the optimal k as stated in Result 2, the convergence rate (with respect to n) for the identification of one fixed cluster* $C^{(i)}$ *is different for the mutual and the symmetric kNN graph. It depends*

- *only on the properties of the cluster* $C^{(i)}$ *itself in the mutual kNN graph*
- *on the properties of the "least significant", that is the "worst" out of all clusters in the symmetric kNN graph.*

This result shows that if one is interested in identifying the "most significant" clusters only, one is better off using the mutual kNN graph. When the goal is to identify all clusters, then there is not much difference between the two graphs, because both of them have to deal with the "worst" cluster anyway. Note that this result is mainly due to the different between-cluster connectivity properties of the graphs, the within-cluster connectivity results are not so different (using our proof techniques at least).

Proof techniques, intuitively. Given a neighborhood graph on the sample, cluster identification always consists of two main steps: ensuring that points of the same cluster are connected and that points of different clusters are not connected to each other. We call those two events "within-cluster connectedness" and "between-cluster disconnectedness" (or "cluster isolation").

To treat within-cluster connectedness we work with a covering of the true cluster. We cover the whole cluster by balls of a certain radius z. Then we want to ensure that, first, each of the balls contains at least one of the sample points, and second, that points in neighboring balls are always connected in the kNN graph. Those are two contradicting goals. The larger the z is, the easier it is to ensure that each ball contains a sample point. The smaller the z is, the easier it is to ensure that points in neighboring balls will be connected in the graph for a fixed number of neighbors k. So the first part of the proof consists in computing the probability that for a given z both events occur at the same time and finding the optimal z.

Between-cluster connectivity is easier to treat. Given a lower bound on the distance u between two clusters, all we have to do is to make sure that edges in the kNN graph never become longer than u, that is we have to prove bounds on the maximal kNN distance in the sample.

In general, those techniques can be applied with small modifications both in the noise-free and in the noisy case, provided we construct our graphs in the way described above. The complication in the noisy case is that if we just used the standard kNN graph as in the noise-free case, then of course the whole space would be considered as one connected component, and this would also show up in the neighborhood graphs. Thus, one has to artificially reduce the neighborhood graph in order to remove the background component. Only then one can hope to obtain a graph with different connected components corresponding to different clusters. The way we construct the graph $\tilde{G}$ ensures this. First, under the assumption that the error of the density estimator is bounded by ε, we consider the $(t - \varepsilon)$-level set instead of the t-level set we are interested in. This ensures that we do not remove "true cluster points" in our procedure. A second, large complication in the noisy case is that with a naive approach, the radius z of the covering and the accuracy ε of the density estimator would be coupled to each other. We would need to ensure that the parameter ε decreases with a certain rate depending on z. This would lead to complications in the proof as well as very slow convergence rates. The trick by which we can avoid this is to introduce the parameter δ and throw away all connected components which are smaller than δn. Thus, we ensure that no small connected components are left over in the boundary of the $(t - \varepsilon)$-level set of a cluster, and all remaining points which are in this boundary strip will be connected to the main cluster represented by the t-level set. Note, that this construction allows us to estimate the number of clusters even without exact estimation of the density.

Building blocks from the literature. To a certain extent, our proofs follow and combine some of the techniques presented in [5] and [2].

In [5] the authors study the connectivity of random mutual k-nearest-neighbor graphs. However, they are mainly interested in asymptotic results, only consider the noise-free case, and do not attempt to make statements about the optimal choice of k. Their main result is that in the noise-free case, choosing k at least of the order $O(\log n)$ ensures that in the limit for $n \to \infty$, connected components of the mutual k-nearest-neighbor graph correspond to true underlying clusters.

In [2], the authors study the noisy case and define clusters as connected components of the t-level set of the density. As in our case, the authors use density estimation to remove background points from the sample, but then work with an ε-neighborhood graph instead of a k-nearest-neighbor graph on the remaining sample. Connectivity of this kind of graph is much easier to treat than the one of k-nearest-neighbor graphs, as the connectivity of two points in the ε-graph does not depend on any other points in the sample (this is not the case in the k-nearest-neighbor graph). Then, Biau et al. [2] prove asymptotic results for the estimation of the connected components of the level set $L(t)$, but also do not investigate the optimal choice of their graph parameter ε. Moreover, due to our additional step where we remove small components of the graph, we can provide much faster rates for the estimation of the components, since we have a much weaker coupling of the density estimator and the clustering algorithm.

Finally, note that a considerably shorter version of the current paper dealing with the noise-free case only has appeared in [11]. In the current paper we have shortened the proofs significantly at the expense of having slightly worse constants in the noise-free case.

3. General assumptions and notation

Density and clusters. Let p be a bounded probability density with respect to the Lebesgue measure on $\mathbb{R}^d$. The measure on $\mathbb{R}^d$ that is induced by the density p is denoted by μ. Given a fixed level parameter $t > 0$, the t-level set of the density p is defined as

$$L(t) = \overline{\{x \in \mathbb{R}^d \mid p(x) \geq t\}}$$

where the bar denotes the topological closure (note that level sets are closed by assumptions in the noisy case, but this is not necessarily the case in the noise-free setting).

Geometry of the clusters. We define clusters as the connected components of $L(t)$ (where the term "connected component" is used in its topological sense). The number of clusters is denoted by m, and the clusters themselves by $C^{(1)}, \ldots, C^{(m)}$. We set $\beta_{(i)} := \mu(C^{(i)})$, that means, the probability mass in cluster $C^{(i)}$.

We assume that each cluster $C^{(i)}$ ($i = 1, \ldots, m$) is a disjoint, compact and connected subset of $\mathbb{R}^d$, whose boundary $\partial C^{(i)}$ is a smooth $(d - 1)$-dimensional submanifold in $\mathbb{R}^d$ with minimal curvature radius $\kappa^{(i)} > 0$ (the inverse of the largest principal curvature of $\partial C^{(i)}$). For $\nu \leq \kappa^{(i)}$, we define the collar set $Col^{(i)}(\nu) = \{x \in C^{(i)} \mid \text{dist}(x, \partial C^{(i)}) \leq \nu\}$ and the maximal covering radius $\nu_{\max}^{(i)} = \max_{\nu \leq \kappa^{(i)}} \{\nu \mid C^{(i)} \setminus Col^{(i)}(\nu) \text{ connected }\}$. These quantities will be needed for the following reasons: It will be necessary to cover the inner part of each cluster by balls of a certain fixed radius z, and those balls are not supposed to "stick outside". Such a construction is only possible under assumptions on the maximal curvature of the boundary of the cluster. This will be particularly important in the noisy case, where all statements about the density estimator only hold in the inner part of the cluster.

For an arbitrary $\varepsilon > 0$, the connected component of $L(t - \varepsilon)$ which contains the cluster $C^{(i)}$ is denoted by $C_-^{(i)}(\varepsilon)$. Points in the set $C_-^{(i)}(\varepsilon) \setminus C^{(i)}$ will sometimes be referred to as boundary points. To express distances between the clusters, we assume that there exists some $\tilde{\varepsilon} > 0$ such that $\text{dist}(C_-^{(i)}(2\tilde{\varepsilon}), C_-^{(j)}(2\tilde{\varepsilon})) \geq u^{(i)} > 0$ for all $i, j \in \{1, \ldots, m\}$. The numbers $u^{(i)}$
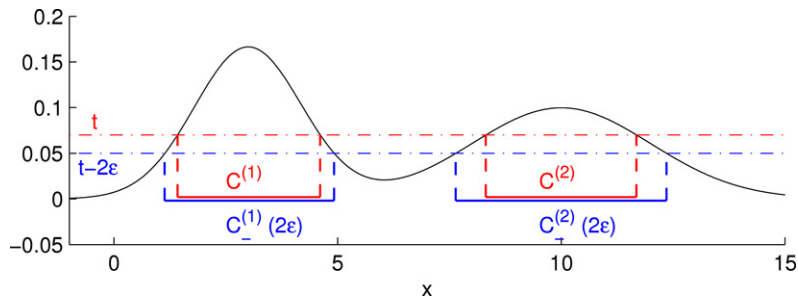
Fig. 1. An example of our cluster definition. The clusters $C^{(1)}, C^{(2)}$ are defined as the connected components of the t-level set of the density (here $t = 0.07$). The clusters are subsets of the sets $C_-^{(1)}(2\varepsilon), C_-^{(2)}(2\varepsilon)$ (here for $\varepsilon = 0.01$).

will represent lower bounds on the distances between cluster $C^{(i)}$ and the remaining clusters. Note that the existence of the $u^{(i)} > 0$ ensures that $C_-^{(i)}(2\varepsilon)$ does not contain any other clusters apart from $C^{(i)}$ for $\varepsilon < \tilde{\varepsilon}$. Analogously to the definition of $\beta_{(i)}$ above we set $\tilde{\beta}_{(i)} = \mu(C_-^{(i)}(2\tilde{\varepsilon}))$, that is the mass of the enlarged set $C_-^{(i)}(2\tilde{\varepsilon})$. These definitions are illustrated in Fig. 1. Furthermore, we introduce a lower bound on the probability mass in balls of radius $u^{(i)}$ around points in $C_-^{(i)}(2\tilde{\varepsilon})$

$$\rho^{(i)} \leq \inf_{x \in C_-^{(i)}(2\tilde{\varepsilon})} \mu\left(B(x, u^{(i)})\right).$$

In particular, under our assumptions on the smoothness of the cluster boundary we can set $\rho^{(i)} = O^{(i)}(u^{(i)})t\eta_d(u^{(i)})^d$ for an *overlap constant*

$$O^{(i)}(u^{(i)}) = \inf_{x \in C_-^{(i)}(2\tilde{\varepsilon})} \left(\text{vol}(B(x, u^{(i)}) \cap C_-^{(i)}(2\tilde{\varepsilon})) / \text{vol}(B(x, u^{(i)}))\right) > 0.$$

The way it is constructed, $\rho^{(i)}$ becomes larger the larger the distance of $C^{(i)}$ to all the other clusters and is upper bounded by the probability mass of the extended cluster $\tilde{\beta}_{(i)}$.

Example in the noisy case. All assumptions on the density and the clusters are satisfied if we assume that the density p is twice continuously differentiable on a neighborhood of $\{p = t\}$, for each $x \in \{p = t\}$ the gradient of p at x is non-zero, and $\text{dist}(C^{(i)}, C^{(j)}) = u' > u^{(i)}$.

Example in the noise-free case. Here we assume that the support of the density p consists of m connected components $C^{(1)}, \ldots, C^{(m)}$ which satisfy the smoothness assumptions above, and such that the densities on the connected components are lower bounded by a positive constant t. Then the noise-free case is a special case of the noisy case.

Sampling. Our n sample points $X_1, \ldots, X_n$ will be sampled i.i.d. from the underlying probability distribution.

Density estimation in the noisy case. In the noisy case we will estimate the density at each data point X_j by some estimate $\hat{p}_n(X_j)$. For convenience, we state some of our results using a standard kernel density estimator, see [6] for background reading. However, our results can easily be rewritten with any other density estimate.

Further notation. The kNN radius of a point X_j is the maximum distance to a point in kNN(X_i). $R_{\min}^{(i)}$ denotes the minimal kNN radius of the sample points in cluster $C^{(i)}$, whereas $\tilde{R}_{\max}^{(i)}$ denotes the maximal kNN radius of the sample points in $C_-^{(i)}(2\tilde{\varepsilon})$. Note here the difference in the point sets that are considered.

Bin(n, p) denotes the binomial distribution with parameters n and p. Probabilistic events will be denoted by curly capital letters $\mathcal{A}, \mathcal{B}, \ldots$, and their complements by $\mathcal{A}^c, \mathcal{B}^c, \ldots$. An overview of the most important notations used in this paper can be found in Table 1.

4. Exact statements of the main results

In this section we are going to state all our main results in a formal way. In the statement of the theorems we need the following conditions. The first one is necessary for both, the noise-free and the noisy case, whereas the second one is needed for the noisy case only.

- *Condition 1:* Lower and upper bounds on the number of neighbors k,

$$k \geq 4^{d+1} \frac{p_{\max}^{(i)}}{t} \log\left(2\, 8^d\, p_{\max}^{(i)}\, \text{vol}(C^{(i)})\, n\right),$$

$$k \leq (n-1) \min\left\{\frac{\rho^{(i)}}{2} - \frac{2\log(\tilde{\beta}_{(i)}n)}{(n-1)}, 2\, 4^d\, \eta_d\, p_{\max}^{(i)} \min\left\{(u^{(i)})^d, (v_{\max}^{(i)})^d\right\}\right\}.$$

- *Condition 2:* The density p is three times continuously differentiable with uniformly bounded derivatives, $\beta_{(i)} > 2\delta$, and ε_n sufficiently small such that $\mu\left(\bigcup_i (C_-^{(i)}(2\varepsilon_n)\backslash C^{(i)})\right) \leq \delta/2$.

Note that in Theorems 1–3 ε_n is considered small but constant and thus we drop the index n there.

Table 1

Table of notations.

$p(x)$	Density
$\hat{p}_n(x)$	Density estimate in point x
t	Density level set parameter
$L(t)$	t-level set of p
$C^{(1)}, \ldots, C^{(m)}$	Clusters, i.e. connected components of $L(t)$
$C_-^{(i)}(\varepsilon)$	Connected component of $L(t - \varepsilon)$ containing $C^{(i)}$
$\beta_{(i)}, \tilde{\beta}_{(i)}$	Probability mass of $C^{(i)}$ and $C_-^{(i)}(2\tilde{\varepsilon})$ Respectively
$p_{\max}^{(i)}$	Maximal density in cluster $C^{(i)}$
$\rho^{(i)}$	Probability of balls of radius $u^{(i)}$ around points in $C_-^{(i)}(2\tilde{\varepsilon})$
$\kappa^{(i)}$	Minimal curvature radius of the boundary $\partial C^{(i)}$
$\nu_{\max}^{(i)}$	Maximal covering radius of cluster $C^{(i)}$
$Col^{(i)}(\nu)$	Collar set for radius ν
$u^{(i)}$	Lower bound on the distances between $C^{(i)}$ and other clusters
$\tilde{\varepsilon}$	Parameter such that $\text{dist}(C_-^{(i)}(2\varepsilon), C_-^{(j)}(2\varepsilon)) \geq u^{(i)}$ for all $\varepsilon \leq \tilde{\varepsilon}$
η_d	Volume of the d-dimensional unit ball
k	Number of neighbors in the construction of the graph

In our first theorem, we present the optimal choice of the parameter k in the mutual kNN graph for the identification of a cluster. This theorem treats both, the noise-free and the noisy case.

Theorem 1 (*Optimal k for Identification of One Cluster in the Mutual* kNN *Graph*). *The optimal choice of k for identification of cluster $C^{(i)}$ in $G_{\text{mut}}(n, k)$ (noise-free case) resp. rough identification in $\tilde{G}_{mut}(n, k, t - \varepsilon, \delta)$ (noisy case) is*

$$k = (n-1)\Gamma^{(i)} + 1, \quad \text{with} \quad \Gamma^{(i)} := \frac{\rho^{(i)}}{2 + \frac{1}{4^d}\frac{t}{p_{\max}^{(i)}}},$$

provided this choice of k fulfills Condition 1.

In the noise-free case we obtain with $\Omega_{\text{noise-free}}^{(i)} = \dfrac{\rho^{(i)}}{2\,4^{d+1}\frac{p_{\max}^{(i)}}{t}+4}$ and for sufficiently large n

$$P\left(\text{Cluster } C^{(i)} \text{ is identified in } G_{\text{mut}}(n, k)\right) \geq 1 - 3e^{-(n-1)\Omega_{\text{noise-free}}^{(i)}}.$$

For the noisy case, assume that additionally Condition 2 holds and let $\hat{p}_n$ be a kernel density estimator with bandwidth h. Then there exist constants C_1, C_2 such that if $h^2 \leq C_1\varepsilon$ we get with

$$\Omega_{\text{noisy}}^{(i)} = \min\left\{\frac{\rho^{(i)}}{2\,4^{d+1}\frac{p_{\max}^{(i)}}{t}+4}, \frac{n}{n-1}\frac{\delta}{8}, \frac{n}{n-1}C_2\,h^d\,\varepsilon^2\right\}$$

and for sufficiently large n

$$P\left(\text{Cluster } C^{(i)} \text{ roughly identified in } \tilde{G}_{mut}(n, k, t - \varepsilon, \delta)\right) \geq 1 - 8e^{-(n-1)\Omega_{noisy}^{(i)}}.$$

This theorem has several remarkable features. First of all, we can see that both in the noise-free and in the noisy case, the optimal choice of k is roughly linear in n. This is pretty surprising, given that the lower bound for cluster connectivity in random geometric graphs is $k \sim \log n$. We will discuss the important consequences of this result in the last section.

Secondly, we can see that for the mutual kNN graph the identification of one cluster $C^{(i)}$ only depends on the properties of the cluster $C^{(i)}$, but not on the ones of any other cluster. This is a unique feature of the mutual kNN graph which comes from the fact that if cluster $C^{(i)}$ is very "dense", then the neighborhood relationship of points in $C^{(i)}$ never links outside of cluster $C^{(i)}$. In the mutual kNN graph this implies that any connections of $C^{(i)}$ to other clusters are prevented. Note that this is not true for the symmetric kNN graph, where another cluster can simply link into $C^{(i)}$, no matter which internal properties $C^{(i)}$ has.

For the mutual graph, it thus makes sense to define the *most significant* cluster as the one with the largest coefficient $\Omega^{(i)}$, since this is the one which can be identified with the fastest rate. In the noise-free case one observes that the coefficient $\Omega^{(i)}$ of cluster $C^{(i)}$ is large given that

- $\rho^{(i)}$ is large, which effectively means a large distance $u^{(i)}$ of $C^{(i)}$ to the closest other cluster,
- $p_{\max}^{(i)}/t$ is small, so that the density is rather uniform inside the cluster $C^{(i)}$.

Note that those properties are the most simple properties one would think of when imagining an "easily detectable" cluster. For the noisy case, a similar analysis still holds as long as one can choose the constants δ, h and ε small enough.

Formally, the result for the identification of clusters in the symmetric kNN graph looks very similar to the one above.

Theorem 2 (*Optimal k for Identification of One Cluster in the Symmetric* kNN *Graph*). *We use the same notation as in Theorem 1 and define $\rho_{\min} = \min_{i=1,\ldots,m} \rho^{(i)}$. Then all statements about the optimal rates for k in Theorem 1 can be carried over to the symmetric* kNN *graph, provided one replaces $\rho^{(i)}$ with $\rho_{\min}$ in the definitions of $\Gamma^{(i)}$, $\Omega_{\text{noise-free}}^{(i)}$ and $\Omega_{\text{noisy}}^{(i)}$. If Condition 1 holds and the condition $k \leq (n-1)\rho_{\min}/2 - 2\log(n)$ replaces the corresponding one in Condition 1, we have in the noise-free case for sufficiently large n*

$$P\big(C^{(i)} \text{ is identified in } G_{sym}(n,k)\big) \geq 1 - (m+2)e^{-(n-1)\Omega_{\text{noise-free}}^{(i)}}.$$

If additionally Condition 2 holds we have in the noisy case for sufficiently large n

$$P\big(C^{(i)} \text{ roughly identified in } \tilde{G}_{\text{sym}}(n,k,t-\varepsilon,\delta)\big) \geq 1 - (m+7)e^{-(n-1)\Omega_{\text{noisy}}^{(i)}}.$$

Observe that the constant $\rho^{(i)}$ has now been replaced by the minimal $\rho^{(j)}$ among all clusters $C^{(j)}$. This means that the rate of convergence for the symmetric kNN graph is governed by the constant $\rho^{(j)}$ of the "worst" cluster, that is the one which is most difficult to identify. Intuitively, this worst cluster is the one which has the smallest distance to its neighboring clusters. In contrast to the results for the mutual kNN graph, the rate for identification of $C^{(i)}$ in the symmetric graph is governed by the worst cluster instead of the cluster $C^{(i)}$ itself. This is a big disadvantage if the goal is to only identify the "most significant" clusters. For this purpose the mutual graph has a clear advantage.

On the other hand we will see in the next theorem that the difference in behavior between the mutual and symmetric graphs vanishes as soon as we attempt to identify *all* clusters.

Theorem 3 (*Optimal k for Identification of All Clusters in the Mutual* kNN *Graph*). *We use the same notation as in Theorem 1 and define $\rho_{\min} = \min_{i=1,\ldots,m} \rho^{(i)}$, $p_{\max} = \max_{i=1,\ldots,m} p_{\max}^{(i)}$. The optimal choice of k for the identification of all clusters in the mutual* kNN *graph in $G_{\text{mut}}(n,k)$ (noise-free case) resp. rough identification of all clusters in $\tilde{G}_{\text{mut}}(n,k,t-\varepsilon,\delta)$ (noisy case) is given by*

$$k = (n-1)\Gamma^{\text{all}} + 1, \quad \text{with} \quad \Gamma^{\text{all}} = \frac{\rho_{\min}}{2 + \frac{1}{4^d}\frac{t}{p_{\max}}},$$

provided this choice of k fulfills Condition 1 for all clusters $C^{(i)}$. In the noise-free case we get the rate

$$\Omega_{\text{noise-free}} = \frac{\rho_{\min}}{2\,4^{d+1}\frac{p_{\max}}{t} + 4},$$

such that for sufficiently large n

$$P\big(\text{All clusters exactly identified in } G_{\text{mut}}(n,k)\big) \geq 1 - 3m\,e^{-(n-1)\Omega_{\text{noise-free}}}.$$

For the noisy case, assume that additionally Condition 2 holds for all clusters and let $\hat{p}_n$ be a kernel density estimator with bandwidth h. Then there exist constants C_1, C_2 such that if $h^2 \leq C_1\varepsilon$ we get with

$$\Omega_{\text{noisy}} = \min\left\{\frac{\rho_{\min}}{2\,4^{d+1}\frac{p_{\max}}{t} + 4},\ \frac{n}{n-1}\frac{\delta}{8},\ \frac{n}{n-1}C_2\,h^d\,\varepsilon^2\right\}$$

and for sufficiently large n

$$P\big(\text{All clusters roughly identified in } \tilde{G}_{\text{mut}}(n,k,t-\varepsilon,\delta)\big) \geq 1 - (3m+5)\,e^{-(n-1)\Omega_{\text{noisy}}}.$$

We can see that as in the previous theorem, the constant which now governs the speed of convergence is the worst case constant among all the $\rho^{(j)}$. In the setting where we want to identify all clusters this is unavoidable. Of course the identification of "insignificant" clusters will be difficult, and the overall behavior will be determined by the most difficult case. This is what is reflected in the above theorem. The corresponding theorem for identification of all clusters in the symmetric kNN graph looks very similar, and we omit it.

So far for the noisy case we mainly considered the case of rough cluster identification. As we have seen, in this setting the results of the noise-free case are very similar to the ones in the noisy case. Now we would like to conclude with a theorem for exact cluster identification in the noisy case.

Theorem 4 (*Exact Identification of Clusters in the Noisy Case*). *Let p be three times continuously differentiable with uniformly bounded derivatives and let $\hat{p}_n$ be a kernel density estimator with bandwidth $h_n = h_0(\log n/n)^{1/(d+4)}$ for some $h_0 > 0$. For a suitable constant $\varepsilon_0 > 0$ set $\varepsilon_n = \varepsilon_0(\log n/n)^{2/(d+4)}$. Then there exist constants c_1, c_2 such that for $n \to \infty$ and $c_1\log n \leq k \leq c_2 n$ we obtain*

Cluster $C^{(i)}$ is exactly identified in $\tilde{G}_{mut}(n,k,t-\varepsilon_n,\delta)$ almost surely.

Note that as opposed to rough cluster identification, which is a statement about a given finite nearest neighbor graph, exact cluster identification is an inherently asymptotic property. The complication in this asymptotic setting is that one has to balance the speed of convergence of the density estimator with the one of the "convergence of the graph". The exact form of the density estimation is not important. Every other density estimator with the same convergence rate would yield the same result. One can even lower the assumptions on the density to $p \in C^1(\mathbb{R}^d)$ (note that differentiability is required elsewhere). Finally, note that since it is technically difficult to grasp the graph after the small components have been discarded, we could not prove what the optimal k in this setting should be.
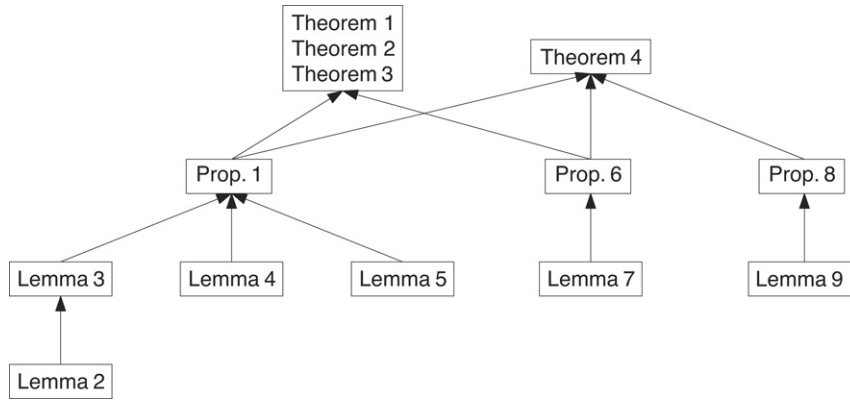
Fig. 2. The structure of our proofs. Proposition 1 deals with within-cluster connectedness and Proposition 6 with between-cluster disconnectedness. Proposition 8 bounds the ratio of background and cluster points for the asymptotic analysis of exact cluster identification.

5. Proofs

The propositions and lemmas containing the major proof steps are presented in Section 5.1. The proofs of the theorems themselves can be found in Section 5.2. An overview of the proof structure can be seen in Fig. 2.

5.1. Main propositions for cluster identification

In Proposition 1 we identify some events whose combination guarantee the connectedness of a cluster in the graph and at the same time that there is not a connected component of the graph that consists of background points only. The probabilities of the events appearing in the proposition are then bounded in Lemmas 2–5. In Proposition 6 and Lemma 7 we examine the probability of connections between clusters. The section concludes with Proposition 8 and Lemma 9, which are used in the exact cluster identification in Theorem 4, and some remarks about the differences between the noise-free and the noisy case.

Proposition 1 (*Connectedness of one Cluster $C^{(i)}$ in the Noisy Case*). *Let $\mathcal{C}_n^{(i)}$ denote the event that in $\tilde{G}_{mut}(n, k, t - \varepsilon_n, \delta)$ (resp. $\tilde{G}_{\text{sym}}(n, k, t - \varepsilon_n, \delta)$) it holds that*

- *all the sample points from $C^{(i)}$ are contained in the graph,*
- *the sample points from $C^{(i)}$ are connected in the graph,*
- *there exists no component of the graph which consists only of sample points from outside $L(t)$.*

Then under the conditions

(1) $\beta_{(i)} > 2\delta$,

(2) ε_n *sufficiently small such that* $\mu\left(\bigcup_i (C_-^{(i)}(2\varepsilon_n) \backslash C^{(i)})\right) \le \delta/2$,

(3) $k \ge 4^{d+1} \frac{p_{\max}^{(i)}}{t} \log\left(2\,8^d\, p_{\max}^{(i)}\, \text{vol}(C^{(i)})\, n\right)$,
   $k \le (n-1)2\,4^d\, \eta_d\, p_{\max}^{(i)} \min\left\{(u^{(i)})^d, (v_{\max}^{(i)})^d\right\}$,

and for sufficiently large n, we obtain

$$\mathrm{P}\left((\mathcal{C}_n^{(i)})^c\right) \le \mathrm{P}\left((\mathcal{A}_n^{(i)})^c\right) + \mathrm{P}\left((\mathcal{B}_n^{(i)})^c\right) + \mathrm{P}(\mathcal{E}_n^c) + \mathrm{P}(\mathcal{D}_n^c)$$

$$\le 2\,\mathrm{e}^{-\frac{k-1}{4^{d+1}}\frac{t}{p_{\max}^{(i)}}} + 2\mathrm{e}^{-n\frac{\delta}{8}} + 2\mathrm{P}(\mathcal{D}_n^c),$$

where the events are defined as follows:

- $\mathcal{A}_n^{(i)}$: *the subgraph consisting of points from $C^{(i)}$ is connected in $G'_{mut}(n, k, t - \varepsilon_n)$ (resp. $G'_{\text{sym}}(n, k, t - \varepsilon_n)$),*
- $\mathcal{B}_n^{(i)}$: *there are more than δn sample points from cluster $C^{(i)}$,*
- $\mathcal{E}_n$: *there are less than δn sample points in the set $\bigcup_i \left(C_-^{(i)}(2\varepsilon_n) \backslash C^{(i)}\right)$, and*
- $\mathcal{D}_n$: $|\hat{p}_n(X_i) - p(X_i)| \le \varepsilon_n$ *for all sample points X_i, $i = 1, \ldots, n$.*

Proof. We bound the probability of $\mathcal{C}_n^{(i)}$ using the observation that $\mathcal{A}_n^{(i)} \cap \mathcal{B}_n^{(i)} \cap \mathcal{E}_n \cap \mathcal{D}_n \subseteq \mathcal{C}_n^{(i)}$ implies

$$\mathrm{P}\left((\mathcal{C}_n^{(i)})^c\right) \le \mathrm{P}\left((\mathcal{A}_n^{(i)})^c\right) + \mathrm{P}\left((\mathcal{B}_n^{(i)})^c\right) + \mathrm{P}(\mathcal{E}_n^c) + \mathrm{P}(\mathcal{D}_n^c). \tag{1}$$

This follows from the following chain of observations. If the event $\mathcal{D}_n$ holds, no point with $p(X_i) \geq t$ is removed, since on this event $p(X_i) - \hat{p}_n(X_i) \leq \varepsilon_n$ and thus $\hat{p}_n(X_i) \geq p(X_i) - \varepsilon_n \geq t - \varepsilon_n$, which is the threshold in the graph $G'(n, k, t - \varepsilon_n)$.

If the samples in cluster $C^{(i)}$ are connected in $G'(n, k, t - \varepsilon_n)$ ($\mathcal{A}_n^{(i)}$), and there are more than δn samples in cluster $C^{(i)}$ ($\mathcal{B}_n^{(i)}$), then the resulting component of the graph $G'(n, k, t - \varepsilon_n)$ is not removed in the algorithm and is thus contained in $\tilde{G}(n, k, t - \varepsilon_n, \delta)$.

Conditional on $\mathcal{D}_n$ all remaining samples are contained in $\bigcup_i C_-^{(i)}(2\varepsilon_n)$. Thus all non-cluster samples lie in $\bigcup_i (C_-^{(i)}(2\varepsilon_n) \backslash C^{(i)})$. Given that this set contains less than δn samples, there can exist no connected component only consisting of non-cluster points, which implies that all remaining non-cluster points are connected to one of the clusters.

The probabilities for the complements of the events $\mathcal{A}_n^{(i)}$, $\mathcal{B}_n^{(i)}$ and $\mathcal{E}_n$ are bounded in Lemmas 3–5 below. Plugging in those bounds into Eq. (1) leads to the desired result. $\square$

We make frequent use of the following tail bounds for the binomial distribution introduced by Hoeffding.

Theorem 5 (*Hoeffding, [9]*). *Let $M \sim \mathrm{Bin}(n, p)$ and define $\alpha = k/n$. Then,*

$$\alpha \geq p, \quad \mathrm{P}(M \geq k) \leq \mathrm{e}^{-n K(\alpha \| p)},$$
$$\alpha \leq p, \quad \mathrm{P}(M \leq k) \leq \mathrm{e}^{-n K(\alpha \| p)},$$

where $K(\alpha \| p)$ is the Kullback–Leibler divergence of $(\alpha, 1 - \alpha)$ and $(p, 1 - p)$,

$$K(\alpha \| p) = \alpha \log \left(\frac{\alpha}{p} \right) + (1 - \alpha) \log \left(\frac{1 - \alpha}{1 - p} \right).$$

In the following lemmas we derive bounds for the probabilities of the events introduced in the proposition above.

Lemma 2 (*Within-cluster Connectedness ($\mathcal{A}_n^{(i)}$)*). *As in Proposition 1 let $\mathcal{A}_n^{(i)}$ denote the event that the points of cluster $C^{(i)}$ are connected in $G'_{mut}(n, k, \varepsilon_n)$ (resp. $G'_{sym}(n, k, \varepsilon_n)$). For $z \in (0, 4 \min\{u^{(i)}, v_{max}^{(i)}\})$,*

$$\mathrm{P}\big((\mathcal{A}_n^{(i)})^c\big) \leq n \, \beta_{(i)} \, \mathrm{P}(M \geq k) + N \left(1 - t \, \eta_d \frac{z^d}{4^d} \right)^n + \mathrm{P}(\mathcal{D}_n^c),$$

where M is a $\mathrm{Bin}(n - 1, p_{max}^{(i)} \eta_d z^d)$-distributed random variable and $N \leq (8^d \, \mathrm{vol}(C^{(i)}))/(z^d \eta_d)$.

Proof. Given that $\mathcal{D}_n$ holds, all samples lying in cluster $C^{(i)}$ are contained in the graph $G'(n, k, \varepsilon_n)$. Suppose that we have a covering of $C^{(i)} \backslash Col^{(i)}(z/4)$ with balls of radius $z/4$. By construction every ball of the covering lies entirely in $C^{(i)}$, so that t is a lower bound for the minimal density in each ball. If every ball of the covering contains at least one sample point and the minimal kNN radius of samples in $C^{(i)}$ is larger than or equal to z, then all samples of $C^{(i)} \backslash Col^{(i)}(z/4)$ are connected in $G'(n, k, \varepsilon_n)$ given that $z \leq 4v_{max}^{(i)}$. Moreover, one can easily check that all samples lying in the collar set $Col^{(i)}(z/4)$ are connected to $C^{(i)} \backslash Col^{(i)}(z/4)$. In total, then all samples points lying in $C^{(i)}$ are connected. Denote by $\mathcal{F}_z^{(i)}$ the event that one ball in the covering with balls of radius $z/4$ contains no sample point. Formally, $\{R_{min}^{(i)} > z\} \cap (\mathcal{F}_z^{(i)})^c$ implies connectedness of the samples lying in $C^{(i)}$ in the graph $G'(n, k, \varepsilon_n)$.

Define $N_s = |\{j \neq s \mid X_j \in B(X_s, z)\}|$ for $1 \leq s \leq n$. Then $\{R_{min}^{(i)} \leq z\} = \bigcup_{s=1}^n \big\{\{N_s \geq k\} \cap \{X_s \in C^{(i)}\}\big\}$. We have

$$\mathrm{P}\big(R_{min}^{(i)} \leq z\big) \leq \sum_{s=1}^n \mathrm{P}\big(N_s \geq k \mid X_s \in C^{(i)}\big) \mathrm{P}\big(X_s \in C^{(i)}\big) \leq n \beta_{(i)} \mathrm{P}(U \geq k),$$

where $U \sim \mathrm{Bin}(n-1, \sup_{x \in C^{(i)}} \mu(B(x, z)))$. The final result is obtained using the upper bound $\sup_{x \in C^{(i)}} \mu(B(x, z)) \leq p_{max}^{(i)} \eta_d z^d$.

For the covering, a standard construction using a $z/4$-packing provides us with the covering. Since $z/4 \leq v_{max}^{(i)}$ we know that balls of radius $z/8$ around the packing centers are subsets of $C^{(i)}$ and disjoint by construction. Thus, the total volume of the N balls is bounded by the volume of $C^{(i)}$ and we get $N(z/8)^d \eta_d \leq \mathrm{vol}(C^{(i)})$. Since we assume that $\mathcal{D}_n$ holds, no sample lying in $C^{(i)}$ has been discarded. Thus the probability for one ball of the covering being empty can be upper bounded by $(1 - t \eta_d z^d/4^d)^n$, where we have used that the balls of the covering are entirely contained in $C^{(i)}$ and thus the density is lower bounded by t. In total, a union bound over all balls in the covering yields,

$$\mathrm{P}\big(\mathcal{F}_z^{(i)}\big) \leq N (1 - t \eta_d z^d/4^d)^n + \mathrm{P}\big(\mathcal{D}_n^c\big).$$

Plugging both results together yields the final result. $\square$

In Lemma 2 we provided a bound on the probability which includes two competing terms for the choice of z. One favors small z whereas the other favors large z. The next lemma will provide a tradeoff optimal choice of the radius z in terms of k.

Lemma 3 (*Choice of k for Within-cluster Connectedness ($\mathcal{A}_n^{(i)}$)*). *If k fulfills Condition (3) of Proposition 1, we have for sufficiently large n*

$$\mathrm{P}\big((\mathcal{A}_n^{(i)})^c\big) \leq 2 \, \mathrm{e}^{-\frac{k-1}{4^{d+1}} \frac{t}{p_{max}^{(i)}}} + \mathrm{P}\big(\mathcal{D}_n^c\big).$$

Proof. The upper bound on the probability of $(\mathcal{A}_n^{(i)})^c$ given in Lemma 2 has two terms dependent on z. The tail bound for the binomial distribution is small if z is chosen to be small, whereas the term from the covering is small given that z is large. Here, we find a choice for z which is close to optimal. Define $p = p_{max}^{(i)} \eta_d z^d$ and $\alpha = k/(n-1)$. Using Theorem 5 we obtain for $M \sim \mathrm{Bin}(n-1, p)$ and a choice of z such that $p < \alpha$,

$$n\beta_{(i)}P(M \geq k) \leq n\beta_{(i)}e^{-(n-1)\left(\alpha \log\left(\frac{\alpha}{p}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-p}\right)\right)}$$
$$\leq n\beta_{(i)}e^{-(n-1)\left(\alpha \log\left(\frac{\alpha}{p}\right) + p - \alpha\right)},$$

where we have used $\log(z) \geq (z-1)/z$ for $z > 0$. Now, introduce $\theta := \eta_d z^d / \alpha$ so that $p = p_{max}^{(i)} \theta \alpha$, where with $p \leq \alpha$ we get, $0 \leq \theta p_{max}^{(i)} \leq 1$. Then,

$$n\beta_{(i)}P(M \geq k) \leq n\beta_{(i)}e^{-k\left(\log\left(\frac{1}{p_{max}^{(i)}\theta}\right) + \theta p_{max}^{(i)} - 1\right)}$$
$$\leq e^{-\frac{k}{2}\left(\log\left(\frac{1}{p_{max}^{(i)}\theta}\right) + \theta p_{max}^{(i)} - 1\right)}, \tag{2}$$

where we used in the last step an upper bound on the term $n\beta_{(i)}$ which holds given $k \geq \left(2\log(\beta_{(i)}n)\right) / \left(\log(1/(\theta p_{max}^{(i)})) + \theta p_{max}^{(i)} - 1\right)$. On the other hand,

$$N(1 - t\,\eta_d z^d/4^d)^n = Ne^{n\log(1 - t\,\eta_d z^d/4^d)} \leq Ne^{-n\,t\,\eta_d z^d/4^d}$$

where we used $\log(1-x) \geq -x$ for $x \leq 1$. With $\eta_d z^d = \theta\alpha$ and the upper bound on N we get using $n/(n-1) \geq 1$,

$$Ne^{-n\,t\,\eta_d z^d/4^d} \leq e^{-\frac{n\,t\,\theta\,\alpha}{4^d} + \log\left(\frac{\mathrm{vol}(C^{(i)})8^d}{\theta\,\alpha}\right)} \leq e^{-k\frac{t\,\theta}{4^d} + \log\left(\frac{\mathrm{vol}(C^{(i)})8^d}{\theta\,\alpha}\right)}$$
$$\leq e^{-k\frac{t\,\theta}{2\,4^d}}, \tag{3}$$

where the last step holds given $k \geq \frac{2\,4^d}{t\theta}\log\left(\frac{n\,\mathrm{vol}(C^{(i)})8^d}{\theta}\right)$. Upper bounding the bound in (2) with the one in (3) requires,

$$\frac{t\theta}{2\,4^d} \leq \frac{1}{2}\left(\log\left(\frac{1}{p_{max}^{(i)}\theta}\right) + \theta p_{max}^{(i)} - 1\right).$$

Introduce $\gamma = \theta p_{max}^{(i)}$, then this is equivalent to $\gamma t/(4^d p_{max}^{(i)}) \leq (-\log(\gamma) + \gamma - 1)$. Note, that $t/(4^d p_{max}^{(i)}) \leq 1/4$. Thus, the above inequality holds for all $d \geq 1$ given that $-\log(\gamma) \geq 1 - 3\gamma/4$. A simple choice is $\gamma = 1/2$ and thus $\theta = 1/(2p_{max}^{(i)})$, which fulfills $\theta p_{max}^{(i)} \leq 1$. In total, we obtain with the result from Lemma 2,

$$P\left((\mathcal{A}_n^{(i)})^c\right) \leq 2\,e^{-\frac{k}{4^{d+1}}\frac{t}{p_{max}^{(i)}}} + P\left(\mathcal{D}_n^c\right) \leq 2\,e^{-\frac{k-1}{4^{d+1}}\frac{t}{p_{max}^{(i)}}} + P\left(\mathcal{D}_n^c\right).$$

We plug in the choice of θ into the lower bounds on k. One can easily find an upper bound for the maximum of the two lower bounds which gives,

$$k \geq 4^{d+1}\frac{p_{max}^{(i)}}{t}\log\left(2\,8^d\,p_{max}^{(i)}\,\mathrm{vol}(C^{(i)})\,n\right).$$

The upper bound, $z \leq 4\min\{u^{(i)}, v_{max}^{(i)}\}$, translates into the following upper bound on k, $k \leq (n-1)2\,4^d\,\eta_d\,p_{max}^{(i)}\min\left\{(u^{(i)})^d, (v_{max}^{(i)})^d\right\}$. $\square$

The result of this lemma means that if we choose $k \geq c_1 + c_2\log n$ with two constants c_1, c_2 that depend on the geometry of the cluster and the respective density, then the probability that the cluster is disconnected approaches zero exponentially in k.

Note that due to the constraints on the covering radius, we have to introduce an upper bound on k which depends linearly on n. However, as the probability of connectedness is monotonically increasing in k, the value of the within-connectedness bound for this value of k is a lower bound for all larger k as well. Since the lower bound on k grows with $\log n$ and the upper bound grows with n, there exists a feasible region for k if n is large enough.

Lemma 4 (*Event $\mathcal{B}_n^{(i)}$*). *As in Proposition 1 let $\mathcal{B}_n^{(i)}$ denote the event that there are more than δn sample points from cluster $C^{(i)}$. If $\beta_{(i)} > \delta$ then*

$$P\left((\mathcal{B}_n^{(i)})^c\right) \leq \exp\left(-\frac{1}{2}n\beta_{(i)}\left(\frac{\beta_{(i)} - \delta}{\beta_{(i)}}\right)^2\right).$$

Proof. Let $M^{(i)}$ be the number of samples in cluster $C^{(i)}$. Then,

$$P\left(M^{(i)} < \delta n\right) \leq P\left(M^{(i)} < \frac{\delta}{\beta_{(i)}}\beta_{(i)}n\right) \leq \exp\left(-\frac{1}{2}n\beta_{(i)}\left(\frac{\beta_{(i)} - \delta}{\beta_{(i)}}\right)^2\right),$$

where we used $M^{(i)} \sim \mathrm{Bin}(n, \beta_{(i)})$ and a Chernoff bound. $\square$

Lemma 5 (*Event $\mathcal{E}_n$*). *As in Proposition 1 let $\mathcal{E}_n$ denote the event that there are less than δn sample points in all the boundary sets $C_-^{(j)}(2\varepsilon_n) \setminus C^{(j)}$ together. If $\sum_{j=1}^{m} \mu\big(C_-^{(j)}(2\varepsilon_n) \setminus C^{(j)}\big) < \delta/2$, we have $P(\mathcal{E}_n^c) \leq \exp(-\delta n/8)$.*

Proof. By assumption, for the probability mass in the boundary strips we have $\sum_{j=1}^{m} \mu(C_-^{(j)}(2\varepsilon_n) \setminus C^{(j)}) < \delta/2$. Then the probability that there are at least δn points in the boundary strips can be bounded by the probability that a $\mathrm{Bin}(n, \delta/2)$-distributed random variable V exceeds δn. Using a Chernoff bound we obtain $P(V > \delta n) \leq \exp(-\delta n/8)$. □

The proposition and the lemmas above are used in the analysis of within-cluster connectedness. The following proposition deals with between-cluster disconnectedness.

We say that a cluster $C^{(i)}$ is *isolated* if the subgraph of $\tilde{G}_{\mathrm{mut}}(n, k, t - \varepsilon_n, \delta)$ (resp. $\tilde{G}_{\mathrm{sym}}(n, k, t - \varepsilon_n, \delta)$) corresponding to cluster $C^{(i)}$ is not connected to another subgraph corresponding to any other cluster $C^{(j)}$ with $j \neq i$. Note, that we assume $\min_{j=1,\ldots,m} \mathrm{dist}(C_-^{(i)}(2\varepsilon_n), C_-^{(j)}(2\varepsilon_n)) \geq u^{(i)}$ for all $\varepsilon_n \leq \tilde{\varepsilon}$. The following proposition bounds the probability for cluster isolation. This bound involves the probability that the maximal k-nearest-neighbor radius is greater than some threshold. Therefore in Lemma 7 we derive a bound for this probability. Note that our previous paper [11] contained an error in the result corresponding to Lemma 7, which changed some constants but did not affect the main results.

Proposition 6 (*Cluster Isolation*). *Let $\mathcal{I}_n^{(i)}$ denote the event that the subgraph of the samples in $C_-^{(i)}(2\varepsilon_n)$ is isolated in $\tilde{G}_{mut}(n, k, t - \varepsilon_n, \delta)$. Then given that $\varepsilon_n \leq \tilde{\varepsilon}$, $k < \rho^{(i)} n/2 - 2\log(\tilde{\beta}_{(i)} n)$, we obtain*

$$P\big((\mathcal{I}_n^{(i)})^c\big) \leq P\big(\tilde{R}_{\max}^{(i)} \geq u^{(i)}\big) + P\big(\mathcal{D}_n^c\big) \leq e^{-\frac{n-1}{2}\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)} + P\big(\mathcal{D}_n^c\big).$$

Let $\hat{\mathcal{I}}_n^{(i)}$ be the event that the subgraph of samples in $C_-^{(i)}(2\varepsilon_n)$ is isolated in $\tilde{G}_{\mathrm{sym}}(n, k, t - \varepsilon_n, \delta)$. Define $\rho_{\min} = \min_{i=1,\ldots,m} \rho^{(i)}$ and $\tilde{\beta}_{\max} = \max_{i=1,\ldots,m} \tilde{\beta}_{(i)}$. Then for $\varepsilon_n \leq \tilde{\varepsilon}_n$, $k < \rho_{\min} n/2 - 2\log(\tilde{\beta}_{\max} n)$, we obtain

$$P\big((\hat{\mathcal{I}}_n^{(i)})^c\big) \leq \sum_{j=1}^{m} P\big(\tilde{R}_{\max}^{(j)} \geq u^{(j)}\big) + P\big(\mathcal{D}_n^c\big) \leq m\, e^{-\frac{n-1}{2}\left(\frac{\rho_{\min}}{2} - \frac{k-1}{n-1}\right)} + P\big(\mathcal{D}_n^c\big).$$

Proof. We have $P((\mathcal{I}_n^{(i)})^c) \leq P((\mathcal{I}_n^{(i)})^c \mid \mathcal{D}_n) + P(\mathcal{D}_n^c)$. Given the event $\mathcal{D}_n$, the remaining points in $\tilde{G}_{\mathrm{mut}}(n, k, t - \varepsilon_n, \delta)$ are samples from $C_-^{(j)}(2\varepsilon_n)$ ($j = 1, \ldots, m$). By assumption we have for $\varepsilon_n \leq \tilde{\varepsilon}$ that $\min_{j \neq i} \mathrm{dist}(C_-^{(i)}(2\varepsilon_n), C_-^{(j)}(2\varepsilon_n)) \geq u^{(i)}$. In order to have edges from samples in $C_-^{(i)}(2\varepsilon_n)$ to any other part in $\tilde{G}_{\mathrm{mut}}(n, k, t - \varepsilon_n, \delta)$, it is necessary that $\tilde{R}_{\max}^{(i)} \geq u^{(i)}$. Using Lemma 7 we can lower bound the probability of this event. For the symmetric kNN graph there can be additional edges from samples in $C_-^{(i)}(2\varepsilon_n)$ to other parts in the graph if samples lying in $C_-^{(i)}(2\varepsilon_n)$ are among the kNN-neighbors of samples in $C_-^{(j)}(2\varepsilon_n), j \neq i$. Let u^{ij} be the distance between $C_-^{(i)}(2\tilde{\varepsilon})$ and $C_-^{(j)}(2\tilde{\varepsilon})$. There can be edges from samples in $C_-^{(i)}(2\varepsilon_n)$ to any other part in $\tilde{G}_{\mathrm{sym}}(n, k, \varepsilon_n, \delta)$ if the following event holds: $\{\tilde{R}_{\max}^{(i)} \geq u^{(i)}\} \cup \{\cup_{j \neq i}\{\tilde{R}_{\max}^{(j)} \geq u^{ij}\}\}$. Using a union bound we obtain,

$$P\big((\hat{\mathcal{I}}_n^{(i)})^c \mid \mathcal{D}_n\big) \leq P\big(\tilde{R}_{\max}^{(i)} \geq u^{(i)}\big) + \sum_{j \neq i} P\big(\tilde{R}_{\max}^{(j)} \geq u^{ij}\big).$$

With $u^{(j)} \leq u^{ij}$ and Lemma 7 we obtain the result for $\tilde{G}_{\mathrm{sym}}(n, k, \varepsilon_n, \delta)$. □

The following lemma states the upper bound for the probability that the maximum k-nearest-neighbor radius $\tilde{R}_{\max}^{(i)}$ of samples in $C_-^{(i)}(2\varepsilon_n)$ used in the proof of Proposition 6.

Lemma 7 (*Maximal kNN radius*). *Let $k < \rho^{(i)} n/2 - 2\log(\tilde{\beta}_{(i)} n)$. Then*

$$P\big(\tilde{R}_{\max}^{(i)} \geq u^{(i)}\big) \leq e^{-\frac{n-1}{2}\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)}.$$

Proof. Define $N_s = |\{j \neq s \mid X_j \in B(X_s, u^{(i)})\}|$ for $1 \leq s \leq n$. Then $\{\tilde{R}_{\max}^{(i)} \geq u^{(i)}\} = \bigcup_{s=1}^{n}\{N_s \leq k - 1 \cap X_s \in C_-^{(i)}(2\tilde{\varepsilon})\}$. Thus,

$$P\big(\tilde{R}_{\max}^{(i)} \geq u^{(i)}\big) \leq \sum_{s=1}^{n} P\big(N_s \leq k - 1 \mid X_s \in C_-^{(i)}(2\tilde{\varepsilon})\big) P\big(X_s \in C_-^{(i)}(2\tilde{\varepsilon})\big).$$

Let $M \sim \mathrm{Bin}(n - 1, \rho^{(i)})$. Then $P(N_s \leq k - 1 \mid X_s \in C_-^{(i)}(2\tilde{\varepsilon})) \leq P(M \leq k - 1)$. Using the tail bound from Theorem 5 we obtain for $k - 1 < \rho^{(i)}(n - 1)$,

$$P\big(\tilde{R}_{\max}^{(i)} \geq u^{(i)}\big) \leq n\,\tilde{\beta}_{(i)}\, P(M \leq k - 1)$$

$$\leq n\tilde{\beta}_{(i)}\, e^{-(n-1)\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)} \leq e^{-\frac{n-1}{2}\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)}$$

where we use that $\log(x) \geq (x - 1)/x$, that $-w/e$ is the minimum of $x\log(x/w)$ attained at $x = w/e$ and $(1 - 1/e) \geq 1/2$. Finally, we use that under the stated condition on k we have $\log(n\tilde{\beta}_{(i)}) \leq [(n - 1)\rho^{(i)}/2 - (k - 1)]/2$. □

The following proposition quantifies the rate of *exact cluster identification*, that means how fast the fraction of points from outside the level set $L(t)$ approaches zero.

Proposition 8 (*Ratio of Boundary and Cluster Points*). *Let N_{Cluster} and $N_{\text{NoCluster}}$ be the number of cluster points and background points in $\tilde{G}_{mut}(n, k, t - \varepsilon_n, \delta)$ (resp. $\tilde{G}_{sym}(n, k, t - \varepsilon_n, \delta)$) and let $\mathcal{C}_n^{\text{all}}$ denote the event that the points of each cluster form a connected component of the graph. Let $\varepsilon_n \to 0$ for $n \to \infty$ and define $\beta = \sum_{i=1}^m \beta_{(i)}$. Then there exists a constant $\bar{D} > 0$ such that for sufficiently large n,*

$$P\left(N_{\text{NoCluster}}/N_{\text{Cluster}} > 4\frac{\bar{D}}{\beta}\varepsilon_n \mid \mathcal{C}_n^{\text{all}}\right) \leq e^{-\frac{1}{4}\bar{D}\varepsilon_n n} + e^{-n\frac{\beta}{8}} + P(\mathcal{D}_n^c).$$

Proof. According to Lemma 9 we can find constants $\bar{D}^{(i)} > 0$ such that $\mu(C_-^{(i)}(2\varepsilon_n)\backslash C^{(i)}) \leq \bar{D}^{(i)}\varepsilon_n$ for n sufficiently large, and set $\bar{D} = \sum_{i=1}^m \bar{D}^{(i)}$. Suppose that $\mathcal{D}_n$ holds. Then the only points which do not belong to a cluster lie in the set $\cup_{i=1}^m C_-^{(i)}(2\varepsilon_n)\backslash C^{(i)}$. Some of them might be discarded, but since we are interested in proving an upper bound on $N_{\text{NoCluster}}$ that does not matter. Then with $p = \mathbb{E}\, N_{\text{NoCluster}}/n \leq \bar{D}\varepsilon_n$ and $\alpha = 2\bar{D}\varepsilon_n$ we obtain with Theorem 5 and for sufficiently small ε_n,

$$P(N_{\text{NoCluster}} \geq 2\bar{D}\varepsilon_n n \mid \mathcal{C}_n^{\text{all}}, \mathcal{D}_n) \leq e^{-nK(\alpha||p)} \leq e^{-n\varepsilon_n \bar{D}(2\log(2)-1)},$$

where K denotes the Kullback–Leibler divergence. Here we used that for $p \leq \bar{D}\varepsilon_n$ we have $K(\alpha||p) \geq K(\alpha||\bar{D}\varepsilon_n)$ and with $\log(1+x) \geq x/(1+x)$ for $x > -1$ we have $K(2\bar{D}\varepsilon_n||\bar{D}\varepsilon_n) \geq \bar{D}\varepsilon_n(2\log 2 - 1) \geq \bar{D}\varepsilon_n/4$. Given that $\mathcal{D}_n$ holds and the points of each cluster are a connected component of the graph, we know that all cluster points remain in the graph and we have

$$P\left(N_{\text{Cluster}} \leq \frac{\beta n}{2} \mid \mathcal{C}_n^{\text{all}}, \mathcal{D}_n\right) \leq e^{-n\frac{\beta}{8}}$$

using Theorem 5 and similar arguments as above. □

Lemma 9. *Assume that $p \in C^2(\mathbb{R}^d)$ with $\|p\|_\infty = p_{\max}$ and that for each x in a neighborhood of $\{p = t\}$ the gradient of p at x is non-zero, then there exists a constant $\bar{D}^{(i)} > 0$ such that for ε_n sufficiently small,*

$$\mu\left(C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)}\right) \leq \bar{D}^{(i)}\varepsilon_n.$$

Proof. Under the conditions on the gradient and ε_n small enough, one has $C_-^{(i)}(2\varepsilon_n) \subseteq C^{(i)} + C_1\varepsilon_n B(0, 1)$ for some constant C_1. Here "+" denotes set addition, that is for sets A and B we define $A + B = \{a + b \mid a \in A, b \in B\}$. Since the boundary $\partial C^{(i)}$ is a smooth $(d - 1)$-dimensional submanifold in $\mathbb{R}^d$ with a minimal curvature radius $\kappa^{(i)} > 0$, there exists $\gamma_1 > 0$ and a constant C_2 such that $\text{vol}(C^{(i)} + \varepsilon_n B(0, 1)) \leq \text{vol}(C^{(i)}) + C_2\varepsilon_n \text{vol}(\partial C^{(i)})$ for $\varepsilon_n < \gamma_1$ (see Theorem 3.3.39 in [7]). Thus, by the additivity of the volume,

$$\begin{aligned}
\text{vol}\left(C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)}\right) &\leq \text{vol}\left(C^{(i)} + C_1\varepsilon_n B(0, 1)\right) - \text{vol}\left(C^{(i)}\right) \\
&= C_1 C_2 \,\text{vol}\left(\partial C^{(i)}\right)\varepsilon_n.
\end{aligned}$$

Since p is bounded, we obtain, $\mu(C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)}) \leq C_1 C_2 \,\text{vol}(\partial C^{(i)}) p_{\max} \varepsilon_n$, for ε_n small enough. Setting $\bar{D}^{(i)} = C_1 C_2 \,\text{vol}(\partial C^{(i)}) p_{\max}$ the result follows. □

Noise-free case as special case of the noisy one. In the noise-free case, by definition all sample points belong to a cluster. That means

- we can omit the density estimation step, which was used to remove background points from the graph, and drop the event $\mathcal{D}_n$ everywhere,
- we work with $L(t)$ directly instead of $L(t - \varepsilon)$,
- we do not need to remove the small components of size smaller than δn, which was needed to get a grip on the "boundary" of $L(t - \varepsilon) \setminus L(t)$.

In particular, setting $\delta = 0$ we trivially have $P((\mathcal{B}_n^{(i)})^c) = 0$ and $P(\mathcal{E}_n^c) = 0$ for all $i = 1, \ldots, m$ and all $n \in \mathbb{N}$.

As a consequence, we can directly work on the graphs $G_{\text{mut}}(n, k)$ and $G_{\text{sym}}(n, k)$, respectively. Therefore, the bounds we gave in the previous sections also hold in the simpler noise-free case and can be simplified in this setting.

5.2. Proofs of the main theorems

Proof of Theorem 1. Given we work on the complement of the event $\mathcal{J}_n^{(i)}$ of Proposition 6, there are no connections in $\tilde{G}_{\text{mut}}(n, k, t - \varepsilon, \delta)$ between the subgraph containing the points of cluster $C^{(i)}$ and points from any other cluster. Moreover, by Proposition 1 we know that the event $\mathcal{C}_n^{(i)} = \mathcal{A}_n^{(i)} \cap \mathcal{B}_n^{(i)} \cap \mathcal{E}_n \cap \mathcal{D}_n$ implies that the subgraph of all the sample points lying in cluster $C^{(i)}$ is connected and all other sample points lying not in the cluster $C^{(i)}$ are either discarded or are connected to the subgraph containing all cluster points. That means we have identified cluster $C^{(i)}$. Collecting the bounds from Propositions 1

and 6, we obtain

$$P\big(\text{Cluster } C^{(i)} \text{ not roughly identified in } \tilde{G}_{\text{mut}}(n, k, t - \varepsilon, \delta)\big)$$

$$\leq P\big((\mathcal{I}_n^{(i)})^c\big) + P\big((\mathcal{C}_n^{(i)})^c\big)$$

$$\leq P\big((\mathcal{I}_n^{(i)})^c\big) + P\big((\mathcal{A}_n^{(i)})^c\big) + P\big((\mathcal{B}_n^{(i)})^c\big) + P(\mathcal{E}_n^c) + P(\mathcal{D}_n^c)$$

$$\leq e^{-\frac{n-1}{2}\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)} + 2\,e^{-\frac{k-1}{4^{d+1}}\frac{t}{p_{\max}^{(i)}}} + 2e^{-n\frac{\delta}{8}} + 3P(\mathcal{D}_n^c).$$

In the noise-free case the events $\mathcal{B}_n^{(i)}$, $\mathcal{E}_n$ and $\mathcal{D}_n$ can be ignored. The optimal choice for k follows by equating the exponents of the bounds for $(\mathcal{I}_n^{(i)})^c$ and $(\mathcal{A}_n^{(i)})^c$ and solving for k. One gets for the optimal k,

$$k = (n-1)\frac{\rho^{(i)}}{2 + \frac{1}{4^d}\frac{t}{p_{\max}^{(i)}}} + 1, \text{ and a rate of } (n-1)\frac{\rho^{(i)}}{2\,4^{d+1}\frac{p_{\max}^{(i)}}{t} + 4}.$$

In the noisy case, we know that for n sufficiently large we can take ε small enough (ε is small and fixed) such that the condition $\sum_{j=1}^m \mu(C_-^{(j)}(2\varepsilon) \setminus C^{(j)}) < \delta/2$ holds. It is well known that under our conditions on p there exist constants C_1, C_2 such that $P(\mathcal{D}_n^c) \leq e^{-C_2 nh^d \varepsilon^2}$ given $h^2 \leq C_1\varepsilon$ (cf. [13]). Plugging this result into the bounds above, the rate of convergence is determined by the worst exponent,

$$\min\left\{\frac{(n-1)\rho^{(i)}}{4} - \frac{k-1}{2}, \frac{k-1}{4^{d+1}}\frac{t}{p_{\max}^{(i)}}, n\frac{\delta}{8}, C_2 nh^d\varepsilon^2\right\}.$$

However, since the other bounds do not depend on k the optimal choice for k remains the same. $\square$

Proof of Theorem 2. Compared to the proof for cluster identification in the mutual kNN graph in Theorem 1 the only part which changes is the connectivity event. Here we have to replace the bound on $P((\mathcal{I}_n^{(i)})^c)$ by the bound on $P((\hat{\mathcal{I}}_n^{(i)})^c)$ from Proposition 6. With $\rho_{\min} = \min_{i=1,\dots,m} \rho^{(i)}$ we obtain

$$P\big((\hat{\mathcal{I}}_n^{(i)})^c\big) \leq m\,e^{-\frac{n-1}{2}\left(\frac{\rho_{\min}}{2} - \frac{k-1}{n-1}\right)} + P(\mathcal{D}_n^c).$$

Following the same procedure as in the proof of Theorem 1 provides the result (for both, the noise-free and the noisy case). $\square$

Proof of Theorem 3. We set $\mathcal{C}_n^{\text{all}} = \bigcap_{i=1}^m \mathcal{C}_n^{(i)}$ and $\mathcal{I}_n^{\text{all}} = \bigcap_{i=1}^m \mathcal{I}_n^{(i)}$. By a slight modification of the proof of Proposition 1 and $p_{\max} = \max_{i=1,\dots,m} p_{\max}^{(i)}$

$$P\big((\mathcal{C}_n^{\text{all}})^c\big) \leq 2\sum_{i=1}^m e^{-\frac{k-1}{4^{d+1}}\frac{t}{p_{\max}^{(i)}}} + 2e^{-n\frac{\delta}{8}} + 2P(\mathcal{D}_n^c)$$

$$\leq 2\,m\,e^{-\frac{k-1}{4^{d+1}}\frac{t}{p_{\max}}} + 2e^{-n\frac{\delta}{8}} + 2P(\mathcal{D}_n^c).$$

By a slight modification of the proof of Proposition 6 with $\rho_{\min} = \min_{i=1,\dots,m} \rho^{(i)}$,

$$P\big((\mathcal{I}_n^{\text{all}})^c\big) \leq \sum_{i=1}^m e^{-\frac{n-1}{2}\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)} + P(\mathcal{D}_n^c) \leq m\,e^{-\frac{n-1}{2}\left(\frac{\rho_{\min}}{2} - \frac{k-1}{n-1}\right)} + P(\mathcal{D}_n^c).$$

Combining these results we obtain

$$P\big(\text{Not all Clusters } C^{(i)} \text{ roughly identified in } \tilde{G}_{\text{mut}}(n, k, t - \varepsilon, \delta)\big)$$

$$\leq m\,e^{-\frac{n-1}{2}\left(\frac{\rho_{\min}}{2} - \frac{k-1}{n-1}\right)} + 3P(\mathcal{D}_n^c) + 2\,m\,e^{-\frac{k-1}{4^{d+1}}\frac{t}{p_{\max}}} + 2e^{-n\frac{\delta}{8}}.$$

The result follows with a similar argumentation to the proof of Theorem 1. $\square$

Proof of Theorem 4. Clearly we can choose $\varepsilon_0 > 0$ such that $h_n^2 \leq C\varepsilon_n$ for a suitable constant $C > 0$. Then there exists a constant $C_2 > 0$ with $P(\mathcal{D}_n^c) \leq e^{-C_2 nh_n^d \varepsilon_n^2}$. Since

$$nh_n^d\varepsilon_n^2 = h_0^d\varepsilon_0^2 n\left(\frac{\log n}{n}\right)^{\frac{d}{d+4}}\left(\frac{\log n}{n}\right)^{\frac{4}{d+4}} = h_0^d\varepsilon_0^2 \log n$$

we have $\sum_{n=1}^\infty P(\mathcal{D}_n^c) < \infty$. Moreover, let $\mathcal{C}_n^{\text{all}}$ denote the event that the points of each cluster form a connected component of the graph. Then it can be easily checked with Proposition 8 that we have $\sum_{n=1}^\infty P(N_{\text{NoCluster}}/N_{\text{Cluster}} > 4\bar{D}\varepsilon_n/\beta \mid \mathcal{C}_n^{\text{all}}) < \infty$. Moreover, similar to the proof of Theorem 3 one can show that there are constants c_1, $c_2 > 0$ such that for $c_1 \log n \leq k \leq c_2 n$ cluster $C^{(i)}$ will be roughly identified almost surely as $n \to \infty$. (Note here that the bounds on k for which our probability bounds hold are also logarithmic and linear, respectively, in n.) Thus, the event $\mathcal{C}_n^{\text{all}}$ occurs almost surely and consequently $N_{\text{NoCluster}}/N_{\text{Cluster}} \to 0$ almost surely. $\square$

6. Discussion

In this paper we studied the problem of cluster identification in kNN graphs. As opposed to earlier work [5,2] which was only concerned with establishing connectivity results for a certain choice of k (resp. ε in case of an ε-neighborhood graph), our goal was to determine for which value of k the probability of cluster identification is maximized. Our work goes considerably beyond [5,2], concerning both the results and the proof techniques. In the noise-free case we come to the surprising conclusion that the optimal k is rather of the order of $c \cdot n$ than of the order of $\log n$ as many people had suspected, both for mutual and symmetric kNN graphs. A similar result also holds for rough cluster identification in the noisy case. Both results were quite surprising to us — our first naive expectation based on the standard random geometric graph literature had been that $k \sim \log n$ would be optimal. In hindsight, our results perfectly make sense. The minimal k to achieve within-cluster connectedness is indeed of the order $\log n$. However, clusters can be more easily identified the tighter they are connected. In an extreme case where clusters have a very large distance to each other, increasing k only increases the within-cluster connectedness. Only when the cluster is fully connected (that is, k coincides with the number of points in the cluster, that is k is a positive fraction of n), connections to other clusters start to arise. Then the cluster will not be identified any more. Of course, the standard situation will not be as extreme as this one, but our proofs show that the tendency is the same.

While our results on the optimal choice of k are nice in theory, in practical application they are often hard to realize. The higher the constant k in the kNN graph is chosen, the less sparse the neighborhood graph becomes, and the more resources we need to compute the kNN graph and to run algorithms on it. This means that one has to make a tradeoff: even if in many applications it is impossible to choose k of the order of $c \cdot n$ for computational restrictions, one should attempt to choose k as large as one can afford, in order to obtain the most reliable clustering results.

When comparing the symmetric and the mutual kNN graph, in terms of the within-cluster connectedness both graphs behave similarly. But note that this might be an artifact of our proof techniques, which are very similar in both cases and do not really make use of the different structure of the graphs. Concerning the between-cluster disconnectedness, however, both graphs behave very differently. To ensure disconnectedness of one cluster $C^{(i)}$ from the other clusters in the mutual kNN graph, it is enough to make sure that the nearest neighbor of all points of $C^{(i)}$ are again elements of $C^{(i)}$. In this sense, the between-cluster disconnectedness of an individual cluster in the mutual graph can be expressed in terms of properties of this cluster only. In the symmetric kNN graph this is different. Here it can happen that some other cluster $C^{(j)}$ links inside $C^{(i)}$, no matter how nicely connected $C^{(i)}$ is. In particular, this affects the setting where the goal is to identify the most significant cluster only. While this is easy in the mutual kNN graph, in the symmetric kNN graph it is not easier than identifying all clusters as the between-cluster disconnectedness is governed by the worst case.

From a technical point of view there are some aspects about our work which could be improved. First, we believe that the geometry of the clusters does not influence our bounds in a satisfactory manner. The main geometric quantities which enter our bounds are simple things like the distance of the clusters to each other, the minimal and maximal density on the cluster, and so on. However, intuitively it seems plausible that cluster identification depends on other quantities as well, such as the shapes of the clusters and the relation of those shapes to each other. For example, we would expect cluster identification to be more difficult if the clusters are in the form of concentric rings than if they are rings with different centers aligned next to each other. Currently we cannot deal with such differences. Secondly, the covering techniques we use for proving our bounds are not well adapted to small sample sizes. We first cover all clusters completely by small balls, and then require that there is at least one sample point in each of those balls. This leads to the unpleasant side effect that our results are not valid for very small sample size n. However, we did not find a way to circumvent this construction. The reason is that as soon as one has to prove connectedness of a small sample of cluster points, one would have to explicitly construct a path connecting each two points. While some techniques from percolation theory might be used for this purpose in the two-dimensional setting, we did not see any way to solve this problem in high-dimensional spaces.

In the current paper, we mainly worked with the cluster definition used in the statistics community, namely the connected components of t-level sets. In practice, most people try to avoid performing clustering by first applying density estimation — density estimation is inherently difficult on small samples, in particular in high-dimensional spaces. On the other hand, we have already explained earlier that this inherent complexity of the problem also pays off. In the end, not only have we detected where the clusters are, but we also know where the data only consists of background noise.

In the computer science community, clustering is often solved via partitioning algorithms such as mincuts or balanced cuts. Now we have treated the case of the level sets in this paper, discussing the graph partitioning case will be the next logical step. Technically, this is a more advanced setting. The ingredients are no longer simple yes/no events (such as "cluster is connected" or "clusters are not connected to each other"). Instead, one has to carefully "count" how many edges one has in different areas of the graph. In future work we hope to prove the results on the optimal choice of k for such a graph partitioning setting.

References

[1] C. Bettstetter, On the minimum node degree and connectivity of a wireless multihop network, in: MobiHoc'02: Proceedings of the 3rd ACM International Symposium on Mobile ad hoc Networking & Computing, ACM, New York, NY, USA, 2002, pp. 80–91.
[2] G. Biau, B. Cadre, B. Pelletier, A graph-based estimator of the number of clusters, ESIAM Probab. Statist. 11 (2007) 272–280.

[3] B. Bollobas, Random Graphs, Cambridge University Press, Cambridge, 2001.
[4] B. Bollobas, O. Riordan, Percolation, Cambridge University Press, Cambridge, 2006.
[5] M. Brito, E. Chavez, A. Quiroz, J. Yukich, Connectivity of the mutual *k*-nearest-neighbor graph in clustering and outlier detection, Statist. Probab. Lett. 35 (1997) 33–42.
[6] L. Devroye, G. Lugosi, Combinatorial Methods in Density Estimation, Springer, New York, 2001.
[7] H. Federer, Geometric Measure Theory, in: Die Grundlehren der mathematischen Wissenschaften, vol. 153, Springer-Verlag, 1969.
[8] J. Hartigan, Consistency of single linkage for high-density clusters, J. Amer. Statist. Assoc. 76 (1981) 388–394.
[9] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (1963) 13–30.
[10] S.S. Kunniyur, S.S. Venkatesh, Threshold functions, node isolation, and emergent lacunae in sensor networks, IEEE Trans. Inform. Theory 52 (12) (2006) 5352–5372.
[11] M. Maier, M. Hein, U. von Luxburg, Cluster identification in nearest-neighbor graphs, in: M. Hutter, R. Servedio, E. Takimoto (Eds.), Proc. of the 18th Conf. on Algorithmic Learning Theory, ALT, Springer, Berlin, 2007, pp. 196–210.
[12] M. Penrose, Random Geometric Graphs, Oxford University Press, Oxford, 2003.
[13] B.L.S.P. Rao, Nonparametric Functional Estimation, Academic Press, New York, 1983.
[14] P. Santi, D. Blough, The critical transmitting range for connectivity in sparse wireless ad hoc networks, IEEE Trans. Mobile Computing 02 (1) (2003) 25–39.