Foundations of Data Science
Vol. 4, No. 4, December 2022, pp. 537–561
doi:10.3934/fods.2021026

RECONSTRUCTING LINEARLY EMBEDDED GRAPHS: A FIRST STEP TO STRATIFIED SPACE LEARNING

Yossi Bokor*[,1,2], Katharine Turner[1]
and Christopher Williams[1]

[1]Mathematical Sciences Institute
Australian National University
Acton, ACT, 2601, Australia

[2]School of Mathematics and Statistics
The University of Sydney
Camperdown, NSW, 2006, Australia

Abstract. In this paper, we consider the simplest class of stratified spaces – linearly embedded graphs. We present an algorithm that learns the abstract structure of an embedded graph and models the specific embedding from a point cloud sampled from it. We use tools and inspiration from computational geometry, algebraic topology, and topological data analysis and prove the correctness of the identified abstract structure under assumptions on the embedding. The algorithm is implemented in the Julia package Skyler, which we used for the numerical simulations in this paper.

1. Introduction.

Increases in the quantity and complexity of collectable data have lead to the search for new methods for efficiently discovering and modelling their underlying structures. The importance of dimensionality reduction of large amounts of data grows with the embedding dimension. By expanding the class of underlying structures which can be detected and modelled, we aim to address some of the difficulties. To improve dimensionality reduction's efficiency and accuracy, we remove the manifold assumption where the dimension is constant and instead treat it as a stratified space, learning the local dimension in the algorithm. We focus on one-dimensional stratified spaces (i.e. graphs) and here provide a new method for dimensionality reduction and compression.

Manifold learning is a method of detecting and modelling structures underlying data sets. There are numerous algorithms and theorems for learning geometric and topological features of manifolds from (noisy) samples, such as dimension or the manifold itself (see [6], [8], [9]). These algorithms make assumptions about the manifold and the sampling procedure, often in the form of curvature restrictions and conditions on the sample's density and noise. Unfortunately, these assumptions are not satisfied by point clouds arising in many applications, such as geospatial transportation network data of vehicle movement. We move towards resolving this problem by expanding the set of allowable underlying structures to include stratified

2020 *Mathematics Subject Classification.* Primary: 57Z25, 51-08; Secondary: 55N31.

Key words and phrases. Stratified space learning, reconstruction and modelling, embedded spaces, implementation.

* Corresponding author: Yossi Bokor.

spaces. A *stratified space* is a space described by gluing together (manifold) pieces, called strata. There are no restrictions placed upon each stratum's dimension, and the gluing can give rise to a variety of interesting and complex local structures.

Bendich et al. ([3], [2]) describe an algorithm which, under certain conditions, can identify if two points have been sampled from the same stratum of a stratified space. This algorithm does not provide a method for learning the global abstract structure. In related work, Nanda et al. ([11]) present an algorithm for detecting when points have been sampled from two intersecting manifolds, which is a cruder splitting than the splitting into stratified subspaces. They have some experimental verification but no theoretical guarantees.

The closest previous work to this paper is [1], in which Aanjaneya et al. consider reconstructing *metric graphs* to detect branch points and the graph structure. There are a few crucial differences. They focus in on the reconstruction of the metric, with input intrinsic distances on the metric graph (plus noise), and they aim to reconstruct a metric graph that is homeomorphic and close as metrics. This means that the theoretical guarantees are about the lengths of edges in the metric graph instead of geometric conditions on an embedding. Crucially, they do not need to consider vertices of degree 2 as in a metric space setting, these are points on an edge.

In contrast, this paper describes an algorithm for modelling a linear embedding of a simple graph from a point cloud sample and provide theoretical guarantees in terms of the geometric embedding that the graph structure modelled is equivalent to the structure embedded.

Definition 1.1 (Graph)**.** A *graph* G consists of

1. A set of vertices $V = \{v_i\}_{i=1}^{n_v}$,
2. A set of edges $E = \{(v_{j_1}, v_{j_2})\}_{j=1}^{n_e}$.

For any graph G, the *boundary operator* $\partial_G : E \to V \times V$, maps an edge to the two boundary vertices. We can represent ∂_G via the *boundary matrix* B, which is the $n_v \times n_e$ matrix with $B[i,j] = 1$ if $v_i = v_{j_1}$ or $v_i = v_{j_2}$. Edges (v_{j_1}, v_{j_2}) are *open*, and their boundary consists of the two vertices.

Given a graph G, we can embed it into $\mathbb{R}^n$ in numerous ways. We will restrict to linear embeddings, such that at degree 2 vertices, the angle between edges is not π.

Definition 1.2 (Linear embedding)**.** A linearly embedded graph

$$|G| = (G, \phi_G) \subset \mathbb{R}^n$$

is a graph G, and a map $\phi_G : G \to \mathbb{R}^n$, such that

1. On the vertex set V, ϕ_G is injective, and we denote $\phi_G(v)$ by v,
2. On E, ϕ_G is defined by linear interpolation: the embedding of an edge (u, v) is the line segment joining $\phi_G(u)$ and $\phi_G(v)$, denoted $\overline{\phi_G(u)\phi_G(v)} = \overline{uv}$,
3. Embedded edges $\overline{uv}, \overline{u'v'}$ only intersect if they share a boundary vertex, say $v' = v$, and their intersection is $\phi_G(v)$.

We restrict our attention to embedded graphs $|G|$ such that at a degree two vertex v, the embedded edges, say $\overline{uv}$, $\overline{wv}$ form an angle $\alpha \neq \pi$.

Please note that with an abuse of notation we will usually use v to denote both the abstract vertex and the embedded location $\phi_G(v)$, and use $\overline{uv}$ to denote both the abstract edge and the embedded image of that edge by ϕ_G. It should always be clear from context whether we are referring to an element in the abstract structure or to its image in $\mathbb{R}^n$.

Throughout this paper, we use the following conventions. For two points $x, y \in \mathbb{R}^n$, $\|x-y\|$ is the distance between x and y in the standard Euclidean metric on $\mathbb{R}^n$, $\langle x, y\rangle$ is the inner product of x and y. For a point $x \in \mathbb{R}^n$ and a set $Y \subset \mathbb{R}^n$, we set

$$d(x, Y) := \inf_{y \in Y} \|x-y\|,$$

and for two sets $X, Y \subset \mathbb{R}^n$,

$$d(X, Y) := \inf_{x \in X, y \in Y} \|x-y\|.$$

Given a point $x \in |G|$, we can determine if x is on an edge, or is a vertex by considering the intersection of $|G|$ with a small ball around x. Consider $B_r(x)$ for small $r > 0$. If x is a vertex, r is less than $\|x-w\|$ for all vertices $w \neq x$ and there are no edges $\overline{uw}$ within r of x, then $B_r(x) \cap |G|$ is connected, and for each edge containing x, there is a unique point in $\partial B_r(x)$. If x is a degree 2 vertex, let the two points on $\partial B_r(x)$ be p and q, then $\angle pxq < \pi$. Now consider $x \in \overline{uv}$ for some embedded edge $\overline{uv}$, and take $r < \min\{\|x-v\|, \|x-u\|\}$. If there is some edge $\overline{wz}$ with $d(x, \overline{wz}) \leq r$, then $B_r(x) \cap |G|$ is disconnected. Otherwise, $B_r(x) \cap |G|$ is connected, and there are two points q, p in $\partial B_r(x) \cap |G|$, and $\angle pxw = \pi$. This is an adaption of the local homology of $|G|$ at x.

We suppose that we do not have the entire embedded graph $|G|$, but only a finite sample P. Furthermore, we expect noise so that $P \not\subseteq |G|$, and we can only make statements about the distance between P and $|G|$. We restrict to sufficiently dense samples P of $|G|$ with bounded noise. Let $d_H(X, Y)$ be the Hausdorff distance between two subsets X, Y of $\mathbb{R}^n$. We consider *ε-samples* of embedded graphs $|G|$.

Definition 1.3 (ε-sample)**.** Let $|G| \subset \mathbb{R}^n$ be an embedded graph. An *ε-sample* P of $|G|$ is a finite subset of $\mathbb{R}^n$ such that $d_H(|G|, P) \leq \varepsilon$.

We can now state the aim of this paper: given an ε-sample P of a linearly embedded graph $|G|$, we want to 1) detect the graph structure G, and then 2) model ϕ_G. This is a semi-parametric problem: the parameters we need to learn are the number of vertices, the number of edges, and the boundary operator. To do so, we need to decide if p is near a vertex v or far away from all vertices for each $p \in P$. This partitions our sample P into two subsets, which intuitively are P_0 containing samples p which are near a vertex, and P_1 containing samples p which are not near any vertex. We define P_0 and P_1 rigorously in Definition 3.5. In the process of partitioning P, we approximate the local homology at each $p \in P$ using radius r. This requires choosing a scale for approximating $|G|$ from P. The clusters in P_0 and P_1 correspond to vertices and edges in G respectively, and we can use the minimal distance between clusters in P_1 and P_0 to learn the boundary operator. Using this information, we model the embedding ϕ_G.

A necessary but not sufficient condition for a point p to be near a vertex is $B_r(p) \cap |G|$ being connected. If it is disconnected, p is not near any vertex, and if it is, we need to check the number of connected components in $B_r(p) \cap |G|$ to determine if p is near a vertex or not. As p is within ε of $|G|$, r must be greater than ε to ensure $B_r(p) \cap |G|$ is non-empty.

Fix $R > \varepsilon$. We first want to approximate $B_R(x) \cap |G|$, and then $\partial B_R(x) \cap |G|$ from P. We can approximate $B_R(p) \cap |G|$ by considering samples $q \in P$ with $\|p - q\| \leq R$. As P is an ε-sample of $|G|$, we can approximate $\partial B_R(p) \cap |G|$ by considering the samples in a spherical shell $S_{R-\varepsilon}^{R+\varepsilon}(p)$ of inner radius $R - \varepsilon$, outer radius $R + \varepsilon$ around p.

We model ϕ_G by aiming to reconstruct a probability measure ν which is supported on $|G| \subset \mathbb{R}^n$. As recorded data has errors, we cannot directly reconstruct ν, but instead construct an approximating measure ν_δ such that ν_δ is equivalent to the Lebesgue measure, and $\text{supp}(\lim_{\delta \to 0} \nu_\delta) = |G|$. We form ν_δ from a categorical mixture model of measures over the individual strata pieces, with latent variables for strata assignment. We use a Gaussian convolution for each individual strata piece to form our approximation of ν with ν_δ. We derive a log-likelihood function which is maximised through an Expectation-Maximisation algorithm (Algorithm 3).

In Section 2, we present and prove some geometric lemmas used throughout Sections 3 and 4, then in Section 3 we define (R, ε)-local structure, describe the (R, ε)-local structure of a vertex and of an edge, before providing conditions under which we can guarantee what (R, ε)-local structure a sample p has. Section 4 presents the algorithm, relates it to the (R, ε)-local structure, before proving that the abstract graph identified is equivalent to the original one. Finally, Section 5 describes the modelling process used and contains some simulations.

2. **Some geometric lemmas.**

As motivation for the formulas both in the definitions of local structure and the geometric assumptions of the graphs' embedding, we first prove some geometric lemmas. Throughout our process, we consider $\langle x_1 - p, x_2 - p \rangle$ for p, x_1, x_2 samples, and $\|p - x_1\|, \|p - x_2\| \in [R - \varepsilon, R + \varepsilon]$. In particular, if there are two clusters of points in the spherical shell around a sample p, all points (including p) are within ε of an edge $\overline{uv}$, and x_1 and x_2 are from different clusters, we wish to bound $\langle x_1 - p, x_2 - p \rangle$ from above.

Lemma 2.1. *Fix $R > 12\varepsilon > 0$ and consider a sample p within ε of an edge $\overline{uv}$. Let H be the hyper-plane through p perpendicular to $\overline{uv}$. Now take x_1, x_2 within ε of edge $\overline{uv}$ such that $\|p - x_1\|, \|p - x_2\| \in [R - \varepsilon, R + \varepsilon]$ and x_1, x_2 are on different sides of H. Then*

$$\langle x_1 - p, x_2 - p \rangle \leq -R^2 + 2R\varepsilon + 7\varepsilon^2.$$

Proof. By assumption $\|x_1 - p\|, \|x_2 - p\| \geq R - \epsilon$. As x_1, p, x_2 are all within ε of $\overline{uv}$ we know that $\angle(x_1 p x_2) \in [\pi - 2\arccos(\frac{2\epsilon}{R-\epsilon}), \pi]$. Together we can bound

$$\begin{aligned}
\langle x_1 - p, x_2 - p \rangle &= \|x_1 - p\| \|x_2 - p\| \cos \angle(x_1 p x_2) \\
&\leq (R - \epsilon)^2 \cos\left(\pi - 2\arccos\left(\frac{2\epsilon}{R - \epsilon}\right)\right) \\
&\leq (R - \epsilon)^2 \left(2\frac{(2\epsilon)^2}{(R - \epsilon)^2} - 1\right) \\
&\leq -R^2 + 2R\varepsilon + 7\varepsilon^2.
\end{aligned}$$

□

We want to distinguish points very close to a vertex of degree 2 as close to a vertex, from points on an edge. This requires an upper bound on the angle at any vertex of degree 2 within our geometric assumptions due to the noise in sampling. The following geometric lemma motivates the upper bound given in the next section.

Lemma 2.2. *Fix $R \geq 12\varepsilon > 0$. Take $u, v, w \in \mathbb{R}^n$, and consider the line segments $\overline{uv}, \overline{wv}$.*

Let $p, x_1, x_2 \in \mathbb{R}^n$ be such that p and x_1 are within ε of $\overline{vw}$, x_2 is within ε of $\overline{uv}$, and $\|x_1 - p\|, \|x_2 - p\| \in [R - \varepsilon, R + \varepsilon]$.

If either

1. *$\|p - v\| < 4\varepsilon$ and*

$$\pi/2 < \angle uvw < \pi - \arctan\left(\frac{R + 3\varepsilon}{6\varepsilon}\right) + \arcsin\left(\frac{R^2 - 4R\varepsilon - 9\varepsilon^2}{(R+\varepsilon)\sqrt{R^2 + 6R\varepsilon + 34\varepsilon^2}}\right),$$

OR

2. *$\|p - v\| < (R - \varepsilon)/2$ and $\angle uvw \leq \pi/2$*

then

$$\langle x_1 - p, x_2 - p\rangle > -R^2 + 2R\varepsilon + 7\varepsilon^2.$$

Proof. Let $\widetilde{p}, \widetilde{x_1}, \widetilde{x_2}$ be the projections of p, x_1, x_2 to $\overline{uv} \cup \overline{wv}$. Without loss of generality, we assume $\widetilde{p}, \widetilde{x_1} \in \overline{wv} \cup v$, and $\widetilde{x_2} \in \overline{uv}$. Then there are $e_p, e_1, e_2 \in \mathbb{R}^n$ with $\|e_q\|, \|e_1\|, \|e_2\| \leq \varepsilon$ and

$$\begin{aligned} p &= \widetilde{p} + e_p \\ x_1 &= \widetilde{x_1} + e_1 \\ x_2 &= \widetilde{x_2} + e_2. \end{aligned}$$

Now consider the vectors $x_1 - p$ and $x_2 - p$, we have:

$$\langle x_1 - p, x_2 - p\rangle = \langle \widetilde{x_1} - \widetilde{p}, \widetilde{x_2} - \widetilde{p}\rangle + \langle \widetilde{x_1} - \widetilde{p}, e_2\rangle - \langle \widetilde{x_1} - \widetilde{p}, e_p\rangle + \langle e_1 - e_p, x_2 - p\rangle \quad (1)$$

We know that e_p is perpendicular to $\overline{vw}$ and thus it is also perpendicular to $\widetilde{x}_1 - \widetilde{p}$ implying $\langle \widetilde{x_1} - \widetilde{p}, e_p\rangle = 0$. Further, we know that $\|\widetilde{x}_1 - \widetilde{p}\| \leq \|x_1 - p\| \leq R + \varepsilon$ as distances can only decrease when projecting onto $\overline{vw}$. Hence, to bound $\langle \widetilde{x_1} - \widetilde{p}, e_2\rangle$ we first split $e_2 = e_2' + e_2''$ where e_2' is the projection of e_2 into the plane spanned by $\overline{vu}$ and $\overline{vw}$. Note that e_2'' is perpendicular to $\widetilde{x_1} - \widetilde{p}$ and hence $\langle \widetilde{x_1} - \widetilde{p}, e_2\rangle = \langle \widetilde{x_1} - \widetilde{p}, e_2'\rangle$. From here, we need to split the proof into the two scenarios.

Assume we are in scenario 1: $\|p - v\| < 4\varepsilon$ and

$$\pi/2 < \angle uvw < \pi - \arctan\left(\frac{R + 3\varepsilon}{6\varepsilon}\right) + \arcsin\left(\frac{R^2 - 4R\varepsilon - 9\varepsilon^2}{(R+\varepsilon)\sqrt{R^2 + 6R\varepsilon + 34\varepsilon^2}}\right).$$

The angle between e_2' and $\widetilde{x_1} - \widetilde{p}$ is either $\angle uvw + \pi/2$ or $\angle uvw - \pi/2$. Recall that we assumed $\angle uvw \in (\pi/2, \pi)$, so $\cos(\angle uvw - \pi/2) > 0 > \cos(\angle uvw + \pi/2)$ and

$$\langle \widetilde{x_1} - \widetilde{p}, e_2\rangle = \langle \widetilde{x_1} - \widetilde{p}, e_2'\rangle \geq \|\widetilde{x_1} - \widetilde{p}\|\|e_2'\| \cos(\angle uvw + \pi/2) \geq -\varepsilon(R + \varepsilon)\sin\angle uvw. \quad (2)$$

Combining (1) and (2) we see

$$\langle x_1 - p, x_2 - p\rangle \geq \langle \widetilde{x_1} - \widetilde{p}, \widetilde{x_2} - \widetilde{p}\rangle - \sin\angle uvw(R + \varepsilon)\varepsilon - (R + \varepsilon)2\varepsilon. \quad (3)$$

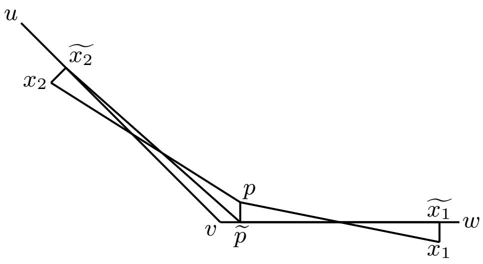


FIGURE 2.1. An example of scenario 1.

To bound $\langle \widetilde{x_1} - \widetilde{p}, \widetilde{x_2} - \widetilde{p}\rangle$ we use that $\angle \widetilde{x_1}\widetilde{p}\widetilde{x_2} = \angle uvw + \angle v\widetilde{x_2}\widetilde{p}$, that the sine rule says $\|\widetilde{x_2} - \widetilde{p}\| \sin(\angle \widetilde{x_2} v \widetilde{p}) = \|v - \widetilde{p}\| \sin \angle uvw$, and that $\cos \angle v\widetilde{x_2}\widetilde{p} > 0$, $\cos \angle uvw < 0$ and $-\sin^2 \angle uvw \leq -\sin \angle uvw$. Together these imply that

$$\begin{aligned}
\langle \widetilde{x_1} - \widetilde{p}, \widetilde{x_2} - \widetilde{p}\rangle &= \|\widetilde{x_1} - \widetilde{p}\|\|\widetilde{x_2} - \widetilde{p}\| \cos(\angle uvw + \angle v\widetilde{x_2}\widetilde{p}) \\
&= \|\widetilde{x_1} - \widetilde{p}\|\|\widetilde{x_2} - \widetilde{p}\| \cos \angle uvw \cos(\angle v\widetilde{x_2}\widetilde{p}) \\
&\quad - \sin^2 \angle uvw \|v - \widetilde{p}\|\|\widetilde{x_1} - \widetilde{p}\| \\
&\geq (R+\varepsilon)(R+3\varepsilon)\cos \angle uvw - \sin \angle uvw \|v - \widetilde{p}\|(R+\varepsilon).
\end{aligned}$$

From the assumptions in this scenario that $\|v - \widetilde{p}\| \leq 4\varepsilon$, we can substitute into (3) to get

$$\begin{aligned}
&\langle x_1 - p, x_2 - p\rangle \\
&\geq (R+\varepsilon)(R+3\varepsilon)\cos \angle uvw - 4\varepsilon(R+\varepsilon)\sin \angle uvw - R\varepsilon(2 + \sin \angle uvw) \\
&\quad - (2 + \sin \angle uvw)\varepsilon^2 \\
&= (R+\varepsilon)\sqrt{R^2 + 6R\varepsilon + 34\varepsilon^2}\sin\left(\angle uvw + \arctan\left(\frac{R+3\varepsilon}{5\varepsilon}\right)\right) - 2\varepsilon R - 2\varepsilon^2.
\end{aligned}$$

From our assumptions in $\angle uvw$

$$\sin\left(\angle uvw + \arctan\left(\frac{R+3\varepsilon}{5\varepsilon}\right)\right) > -\frac{R^2 - 4R\varepsilon + \varepsilon^2}{(R+\varepsilon)\sqrt{R^2 + 6R\varepsilon + 34\varepsilon^2}}.$$

Thus we conclude

$$\begin{aligned}
&\langle x_1 - p, x_2 - p\rangle \\
&> (R+\varepsilon)\sqrt{R^2 + 6R\varepsilon + 34\varepsilon^2}\left(-\frac{R^2 - 4R\varepsilon - 9\varepsilon^2}{(R+\varepsilon)\sqrt{R^2 + 6R\varepsilon + 34\varepsilon^2}}\right) - 2\varepsilon R - 2\varepsilon^2 \\
&= -R^2 + 2R\varepsilon + 7\varepsilon^2.
\end{aligned}$$

Now assume we are in scenario 2: $\|v - p\| < (R - \varepsilon)/2$ and $\angle uvw \leq \pi/2$.

To prove the claim in this scenario, we will need to further split into two cases;

(i) $\angle \widetilde{x_1}\widetilde{p}\widetilde{x_2} \leq \pi/2$, and
(ii) $\angle \widetilde{x_1}\widetilde{p}\widetilde{x_2} > \pi/2$.

In case (i) we have $\langle \widetilde{x_1} - \widetilde{p}, \widetilde{x_2} - \widetilde{p} \rangle \geq 0$ and thus

$$\langle x_1 - p, x_2 - p \rangle \geq -3R\varepsilon - 3\varepsilon^3 > -R^2 + 2R\varepsilon + 7\varepsilon^2$$

as $R > 12\varepsilon$.

In case (ii), thinking of the inner product in terms of the projection of vector $\widetilde{x_2} - \widetilde{p}$ onto $\widetilde{x_1} - \widetilde{p}$ we get

$$\begin{aligned}\langle x_1 - p, x_2 - p \rangle &\geq -\|\widetilde{x_1} - \widetilde{p}\| \|v - \widetilde{p}\| - 3R\varepsilon - 3\varepsilon^3 \\ &\geq -(R + \varepsilon)(R - \varepsilon)/2 - 3R\varepsilon - 3\varepsilon^3 \\ &= -R^2/2 - 3R\varepsilon - 5\varepsilon^2/2 \\ &> -R^2 + 2R\varepsilon + 7\varepsilon^2,\end{aligned}$$

where in the final inequality we use that $R > 12\varepsilon$. □

To find sufficient conditions for detecting when a sample p is near a vertex, we want each edge adjacent to that vertex to correspond to at least one distinct cluster of points in the spherical shell around p. To avoid the clusters around separate edges merging, we assume a lower bound on the angle between the edges as part of our assumptions on the geometric embedding. The following lemma motivates this choice of lower bound.

Lemma 2.3. *Let $u, v, w \in \mathbb{R}^n$, $D > \varepsilon > 0$, and let $x_1, x_2 \in \mathbb{R}^n$ satisfy*

1. *$d(x_1, \overline{uv}), d(x_2, \overline{uw}) < \varepsilon$, and*
2. *$\|x_1 - v\|, \|x_2 - v\| > D$.*

If

$$\angle uvw > \arccos\left(\frac{2D^2 - 9\varepsilon^2}{2D^2}\right) + 2 \arcsin\left(\frac{\varepsilon}{D}\right)$$

then $\|x_1 - x_2\| > 3\varepsilon$.

Proof. The distance between x_1 and x_2 is minimised when

$$\|v - x_1\| = D = \|v - x_2\|.$$

Furthermore we can observe that $\angle uvx_1 = \arcsin\left(\frac{d(x_1, \overline{uv})}{\|x_1 - v\|}\right) \leq \arcsin(\varepsilon/D)$. Similarly $\angle uvx_1 \leq \arcsin(\varepsilon/D)$. This implies

$$\angle x_1 v x_2 \geq \angle uvw - \angle uvx_1 - \angle wvx_2 \geq \alpha - 2\arcsin(\varepsilon/D).$$

Combining we conclude

$$\begin{aligned}\|x_1 - x_2\|^2 &\geq \|v - x_1\|^2 + \|v - x_2\|^2 - \|v - x_1\| \|v - x_2\| \cos \angle x_1 v x_2 \\ &\geq 2D^2 - 2D^2 \cos(\alpha - 2\arcsin(\varepsilon/D)) \\ &\geq (3\varepsilon)^2.\end{aligned}$$

□

3. **Determining local structure.** Given an ε-sample P of an embedded graph $|G|$, we want to recover the abstract graph G by approximating the local structure of $|G|$ at each sample $p \in P$. When approximating the local structure at a sample p, we regularly consider the graph on a set of points, with edges (p, q) if $\|p-q\| \leq r$, for some fixed $r \in \mathbb{R}$.

Definition 3.1. Let $P \subset \mathbb{R}^N$ be a finite collection of points, and fix $r > 0$. The *graph at threshold* r *on* P, $\mathfrak{G}_r(P)$, is the graph with vertices $p \in P$, and edges (p, q) if $\|p - q\| \leq r$.

For each $p \in P$, we will consider two graphs on points close to p: the first approximates the connectedness of $|G|$ intersected with a ball around p, the second consists of points in a spherical shell around p. We call this pair of graphs the (R, ε)*-local structure of* P *at* p.

Definition 3.2 ((R, ε)-local structure)**.** Let $P \subset \mathbb{R}^n$ be an ε-sample of an embedded graph $|G|$ and fix $R > 12\varepsilon$. The (R, ε)-local structure of P at a sample $p \in P$ is the pair

$$\left(\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p)), \mathfrak{G}_{3\varepsilon}(P \cap S_{R-\varepsilon}^{R+\varepsilon}(p))\right).$$

We want to use the (R, ε)-local structure to approximate $|G| \cap B_R(p)$ for each $p \in P$, and use this to learn the structure of $|G|$. We will classify samples as being near a vertex or not near a vertex by their (R, ε)-local structure.

We now formalise what the (R, ε)-local structure is for points $p \in P$ not near any vertex $v \in |G|$. That is, points which have (R, ε)-local structure of an edge.

Definition 3.3 (Local structure of an edge)**.** Let P be an ε-sample of a linearly embedded graph $|G|$. A point $p \in P$ has the (R, ε)*-local structure of an edge* if either of the following hold:

1. $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is disconnected,
2. $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected, $\mathfrak{G}_{3\varepsilon}(P \cap S_{R-\varepsilon}^{R+\varepsilon}(p))$ has two connected components c_1, c_2 with average points q_1 and q_2, and
$$\langle q_1 - p, q_2 - p \rangle \leq -R^2 + 2R\varepsilon + 7\varepsilon^2.$$

We now define the (R, ε)*-local structure of a vertex.*

Definition 3.4 (Local structure of a vertex)**.** Let P be an ε-sample of a linearly embedded graph $|G|$. A point $p \in P$ has the (R, ε)*-local structure of a vertex* if either of the following hold:

1. $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected, and the number of connected components in $\mathfrak{G}_{3\varepsilon}(P \cap S_{R-\varepsilon}^{R+\varepsilon}(p))$ is not 2,
2. $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected, $\mathfrak{G}_{3\varepsilon}(P \cap S_{R-\varepsilon}^{R+\varepsilon}(p))$ has two connected components c_1, c_2 with average points q_1 and q_2, and
$$\langle q_1 - p, q_2 - p \rangle > -R^2 + 2R\varepsilon + 7\varepsilon^2.$$

Next, we formally define P_0 and P_1.

Definition 3.5 (P_0 and P_1)**.** Given an ε-sample P of a linearly embedded graph $|G| \subset \mathbb{R}^n$, we define the partitioning sets P_0 and P_1 as follows:

$$\begin{aligned} P_0 &= \{p \in P \mid \ p \text{ has the } (R,\varepsilon)\text{-local structure of a vertex.}\} \\ P_1 &= \{p \in P \mid \ p \text{ has the } (R,\varepsilon)\text{-local structure of an edge.}\} \end{aligned}$$

Remark 3.6. Note that a sample $p \in P$ has either (R,ε)-local structure of a vertex (R,ε)-local structure of an edge. Hence, the partitioning defined in Definition 3.5 is disjoint.

As we use the connected components of $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ and $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$ in the definition of the (R,ε)-local structure of p, we require some assumptions on $|G|$ to ensure that we correctly identify when points are near vertices or not. To ensure $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is not disconnected for points p near some vertex, we assume that the distance between a vertex v and any edge $\overline{uw}$, $u, v \neq v$, is bounded below $d(v, \overline{wv}) > R + \frac{R}{2} + 2\varepsilon$. To ensure that there are samples near edges which are not near any vertex, we additionally assume that for every pair of vertices u, v, $\|u - v\| > \frac{9R}{2} + 6\varepsilon$.

We also place lower and upper bounds on the angles between edges. For ease of notation, we will define two functions for these bounds.

Definition 3.7. Given $R > 12\varepsilon$, we set

$$\begin{aligned} \Psi(R,\varepsilon) &:= \pi - \arctan\left(\frac{R+3\varepsilon}{6\varepsilon}\right) + \arcsin\left(\frac{R^2 - 4R\varepsilon - 9\varepsilon^2}{(R+\varepsilon)\sqrt{R^2 + 6R\varepsilon + 34\varepsilon^2}}\right), \\ \Phi(R,\varepsilon) &:= \arccos\left(\frac{(R-\varepsilon)^2 - 18\varepsilon^2}{(R-\varepsilon)^2}\right) + 2\arcsin\left(\frac{2\varepsilon}{(R-\varepsilon)}\right). \end{aligned}$$

To improve intuition of these functions, Figures 3.2 and 3.3 provide graphs of them. Note they are effectively a function of $\frac{R}{\varepsilon}$ as they are invariant to scaling both R and ε by the same amount.

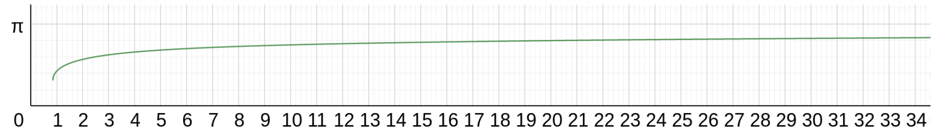


FIGURE 3.2. Graph of $\Psi\left(\frac{R}{\varepsilon}, 1\right)$.

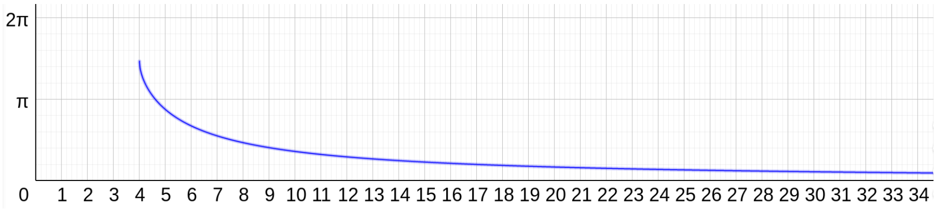


FIGURE 3.3. Graph of $\Phi\left(\frac{R}{\varepsilon}, 1\right)$.

Henceforth, we assume that all embedded graphs $|G|$ satisfy the following assumptions.

Assumption 3.8. *Fix $R \geq 12\varepsilon > 0$. We restrict to embedded graphs $|G| = (G, \phi_G)$ satisfying the following.*

1. *For all vertices u, v, $\|u - v\| > \frac{9R}{2} + 6\varepsilon$.*
2. *For a vertex v and an edge $\overline{uw}$, with $u, w \neq v$, $d(v, \overline{uw}) > \frac{3R}{2} + 4\varepsilon$.*
3. *For any pair of edges $\overline{uv}, \overline{xy}$ with no common vertex, $d(\overline{uv}, \overline{xy}) > 5\varepsilon$.*
4. *For all pairs of edges $\overline{uv}, \overline{wv}$, $\angle uvw \geq \Phi(R, \varepsilon)$.*
5. *For all degree 2 vertices v with edges $\overline{uv}, \overline{wv}$, $\angle uvw \leq \Psi(R, \varepsilon)$.*

The propositions in this section are used to show that the clusters in P_0 and P_1 correspond bijectively with the vertices and edges of $|G|$. The first proposition shows that for all samples p near a vertex v with $\deg(v) \neq 2$, p has the (R, ε)-local structure of a vertex. The second and third prove that samples near degree 2 vertices also have the (R, ε)-local structure of a vertex. The final proposition shows that all samples p not near any vertex have the (R, ε)-local structure of an edge.

Proposition 3.9. *Let v be a vertex of $|G| \subset \mathbb{R}^n$ with $\deg(v) \neq 2$, and let P be an ε-sample of $|G|$. Then for all $p \in P$ with $\|p - v\| \leq \frac{R-\varepsilon}{2}$, p has the (R, ε)-local structure of a vertex.*

Proof. We begin by considering $\mathbf{deg(v) = 0}$. By Assumptions 3.8 (1), $\|p - v\| \leq \varepsilon$, and for all $q \in P \cap B(p, R+\varepsilon)$, $\|q - v\| \leq \varepsilon$. Thus $\mathfrak{G}_{3\varepsilon}\left(P \cap B_{R+\varepsilon}(p)\right)$ is connected. Similarly, $P \cap S^{R+\varepsilon}_{R-\varepsilon}(p) = \emptyset$, and p has the (R, ε)-local structure of a vertex.

Next, assume $\mathbf{deg(v) = 1}$. For the edge $\overline{uv}$, let $t_0, t_1, \ldots, t_m$ be consecutive points along $\overline{uv}$ with $\|t_0 - v\|, \|t_{i+1} - t_i\| \leq \varepsilon$ and $\|p - t_m\| = R + \varepsilon$. Then, there must be $z_0, z_1, \ldots, z_m \in P$ with $\|t_i - z_i\| \leq \varepsilon$. Note, these z_i may not be unique. Since $\|z_i - z_{i+1}\| \leq 3\varepsilon$, and every sample in $P \cap B_{R+\varepsilon}(p)$ is within 3ε of some z_i, $\mathfrak{G}_{3\varepsilon}\left(P \cap B_{R+\varepsilon}(p)\right)$ is connected.

If the number of clusters in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$ is not 2, then p has the (R, ε)-local structure of a vertex. Thus suppose that there are 2 connected components. We will show that inner product condition between their averages will declare that p has the (R, ε)-local structure of a vertex.

Let $x_1, x_2 \in P \cap S^{R+\varepsilon}_{R-\varepsilon}(p)$ be samples in the two connected components c_1 and c_2. Observe that both x_1 and x_2 are within ε of the line segment $\overline{uv}$.

As $\|p - v\| \leq \frac{R-\varepsilon}{2}$, and x_1, x_2 are within ε of the same edge $\overline{uv}$, x_1 and x_2 are contained on the same side of hyper-plane H through p perpendicular to $\overline{vu}$.

We can observe that $\angle x_1 p x_2 \leq 2 \arccos\left(\frac{2\varepsilon}{R-\varepsilon}\right) < \pi/2$, and thus

$$\langle x_1 - p, x_2 - p\rangle > 0 > -R^2 + 2R\varepsilon + 7\varepsilon^2.$$

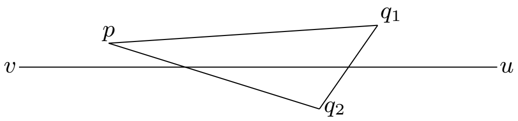


FIGURE 3.4. Both q_1 and q_2 are in the same half-space generated by the hyper-plane through p perpendicular to $\overline{uv}$.

As this holds for all $x_1 \in c_1, x_2 \in c_2$, it also holds for the averages q_1 and q_2. Thus p has the (R,ε)-local structure of a vertex.

Finally, assume $\mathbf{deg(v)} \geq \mathbf{3}$. From analogous arguments as in the degree 1 case we know that $\mathfrak{G}_{3\varepsilon}\,(P \cap B_{R+\varepsilon}(p))$ is connected.

Now consider $S^{R+\varepsilon}_{R-\varepsilon}(p)$. For each edge $\overline{uv}$, there is a sample $x_{\overline{uv}} \in S^{R+\varepsilon}_{R-\varepsilon}(p)$. To show there are at least 3 connected components in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$, we need only check that samples from different edges cannot merge to be in the same connected component in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$. By way of contradiction suppose there were edges $\overline{uv}$ and $\overline{wv}$ and samples $x_u, x_v \in P \cap S^{R+\varepsilon}_{R-\varepsilon}(p)$ within ε of $\overline{uv}$ and $\overline{wv}$ respectively such that $\|x_u - x_w\| \leq 3\varepsilon$. As $\|p - v\| \leq (R-\varepsilon)/2$ and $\|p - x_u\|, \|p - x_v\| \geq R - \varepsilon$ we know $\|v - x_u\|, \|v - x_w\| > (R-\varepsilon)/2$. This contradicts Lemma 2.3 as this implies that $\|x_u - x_v\| > 3\varepsilon$.

We conclude that $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$ has at least as many connected components as the degree of v. Thus, p has the (R,ε)-local structure of a vertex. □

Proposition 3.10. *Let v be a vertex of $|G| \subset \mathbb{R}^n$ with deg$(v) = 2$, with edges $\overline{uv}, \overline{wv}$. Let P be an ε-sample of $|G|$. If $\angle uwv > \frac{\pi}{2}$, then for all $p \in P$ with $\|p - v\| \leq 4\varepsilon$, p has the (R,ε)-local structure of a vertex.*

Proof. As in the proof Proposition 3.9, $\mathfrak{G}_{2\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected.

For both edges $\overline{uv}, \overline{wv}$ there is at least one sample in $S^{R+\varepsilon}_{R-\varepsilon}(P)$, say $q_{\overline{uv}}$ and $q_{\overline{wv}}$. By Lemma 2.3, for all q' in $S^{R+\varepsilon}_{R_\varepsilon} \cap P$, if $d(q', q_{\overline{wv}}) \leq 3\varepsilon$, then $\|q' - q\| > 3\varepsilon$. Hence, each edge contributes at least 1 connected component to $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$.

If there are more than 2, then p has the (R,ε)-local structure of a vertex. We now assume there are 2 connected components c_1, c_2 (one per edge) in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$. Lemma 2.2 gives

$$\langle q_1 - p, q_2 - p \rangle > -R^2 + 2R\varepsilon + 7\varepsilon^2,$$

and p has the (R,ε)-local structure of a vertex. □

Proposition 3.11. *Let v be a vertex of $|G| \subset \mathbb{R}^n$ with deg$(v) = 2$, with edges $\overline{uv}, \overline{wv}$. Let P be an ε-sample of $|G|$. If $\angle uvw \leq \frac{\pi}{2}$, then for all $p \in P$ with $\|p - v\| \leq \frac{R-\varepsilon}{2}$, p has the (R,ε)-local structure of a vertex.*

Proof. As in the proof of Proposition 3.9, $\mathfrak{G}_{2\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected.

For both edges $\overline{uv}, \overline{wv}$ there is at least one sample in $S^{R+\varepsilon}_{R-\varepsilon}(P)$, say $q_{\overline{uv}}$ and $q_{\overline{wv}}$. By Lemma 2.3, for all q' in $S^{R+\varepsilon}_{R_\varepsilon} \cap P$, if $\|q' - q_{\overline{wv}}\| \leq 3\varepsilon$, then $\|q' - q\| > 3\varepsilon$. Hence, each edge contributes at least 1 connected component to $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$.

If there are more than 2, then p has the (R,ε)-local structure of a vertex. We now assume there are 2 connected components c_1, c_2 (one per edge) in $\mathfrak{G}_{3\varepsilon}(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p))$.

Let x_1 and x_2 be points in c_1 and c_2. Without loss of generality, we have $d(x_1, \overline{uv}), d(x_2, \overline{wv}) \leq \varepsilon$.

From Lemma 2.2 we know that $\langle x_1 - p, x_2 - p \rangle < -R^2 + 2R\varepsilon + 7\varepsilon^2$. Since this inequality holds for all pairs x_1, x_2 in the connected components c_1 and c_2 we know it also holds for the averages q_1 and q_2. Thus we conclude p has the (R,ε)-local structure of a vertex.

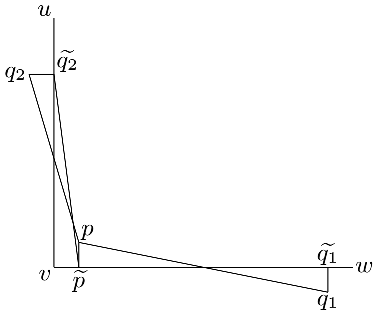


FIGURE 3.5

FIGURE 3.6. The case where $\angle uvw \leq \frac{\pi}{2}$.

□

Proposition 3.12. *Let $p \in P$ be a sample with $\|p - v\| > \frac{3R+\varepsilon}{2}$ for all vertices $v \in |G|$. Then p has the (R, ε)-local structure of an edge.*

Proof. We begin by showing that if there is a sample $q \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$ with $d(q, \overline{uv}) > \varepsilon$, then $\mathfrak{G}_{3\varepsilon}\left(B_{R+\varepsilon}(p) \cap P\right)$ is disconnected. To prove this suppose not. Then there exists $x, y \in B_{R+\varepsilon}(p) \cap P$ such that $d(x, \overline{uv}) < \varepsilon$, $d(y, \overline{uv}) > \varepsilon$ and yet $\|x - y\| < 3\varepsilon$.

This splits into two cases:

(i) $d(y, \overline{wv}) \leq \varepsilon$ for some vertex $w \neq u$ (noting that this case covers an edge $\overline{wu}$ as well),

(ii) $d(y, \overline{wz}) \leq \varepsilon$ for vertices $w, z \neq u, v$.

For case (i), first observe that $\|x - v\|, \|y - v\| > \frac{R-\varepsilon}{2}$. We then get a contradiction via Lemma 2.3 (with $D = \frac{R-\varepsilon}{2}$) using Assumption 3.8 (4).

For case (ii) recall that Assumption 3.8 (3) implies $d(\overline{uv}, \overline{wz}) > 5\varepsilon$. However $d(\overline{uv}, \overline{wz}) < d(\overline{uv}, x) + \|x - y\| + d(y, \overline{wz}) \leq 5\varepsilon$ which is a contradiction.

We thus conclude that if there is some $q \in S^{R+\varepsilon}_{R-\varepsilon}(p) \cap P$ with $d(q, \overline{uv}) > \varepsilon$ then $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is disconnected and p has the (R, ε)-local structure of an edge.

We can now assume that $\mathfrak{G}_{3\varepsilon}(P \cap B_{R+\varepsilon}(p))$ is connected, and for all $q \in P \cap B_{R+\varepsilon}(p)$, $d(q, \overline{uv}) \leq \varepsilon$. We need to show that there are two clusters of samples in $S^{R+\varepsilon}_{R-\varepsilon}(p)$. Let $n \in \overline{uv}$ satisfy $\|p - n\| = R$, and assume that n and q are on the same side of the hyper-plane H through p perpendicular to $\overline{uv}$. Now let $\widetilde{p}, \widetilde{q}$ be the projections of p and q respectively to $\overline{uv}$.

We will split the analysis into the cases where $\|\tilde{p} - \tilde{q}\| \leq \|\tilde{p} - n\|$ and where $\|\tilde{p} - \tilde{q}\| > \|\tilde{p} - n\|$.

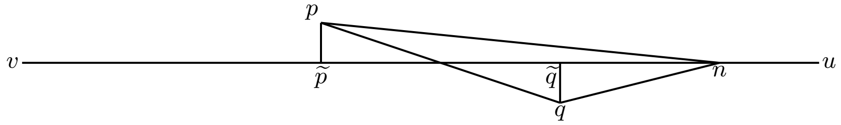


FIGURE 3.7. The case where $\|\widetilde{p} - \widetilde{q}\| < \|\widetilde{p} - n\|$.

Consider $\|\widetilde{p} - \widetilde{q}\| \leq \|\widetilde{p} - n\|$, as in Figure 3.7. Note that $\|\tilde{p} - n\| \leq R$ and $\|\tilde{p} - \tilde{q}\| \geq \sqrt{(R - \varepsilon)^2 - (2\varepsilon)^2}$ which implies

$$\begin{aligned}\|q-n\|^2 &= \|q-\widetilde{q}\|^2 + \big(\|\widetilde{p}-n\| - \|\widetilde{p}-\widetilde{q}\|^2\big)\\ &\leq \varepsilon^2 + \left(R - \sqrt{(R-\varepsilon)^2 - 4\varepsilon^2}\right)^2 . \end{aligned} \tag{4}$$

Now consider $\|\widetilde{p}-n\| < \|\widetilde{p}-\widetilde{q}\|$, such as in Figure 3.8. Here we use the bounds $\|\widetilde{p}-n\| \geq \sqrt{R^2-\varepsilon^2}$ and $\|\widetilde{p}-\widetilde{q}\| \leq R+\varepsilon$ to say

$$\begin{aligned}\|q-n\|^2 &= \|q-\widetilde{q}\|^2 + (\|\widetilde{p}-\widetilde{q}\| - \|\widetilde{p}-n\|)^2\\ &\leq \varepsilon^2 + \left(\sqrt{(R+\varepsilon)^2} - \sqrt{R^2-\varepsilon^2}\right)^2 . \end{aligned} \tag{5}$$

Algebraic manipulation shows that both (4) and (5) are bounded from above by $4\varepsilon^2$ whenever $R > 12\varepsilon$.

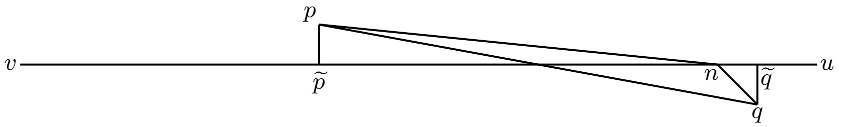


FIGURE 3.8. The case where $\|\widetilde{p}-\widetilde{q}\| > \|\widetilde{p}-n\|$.

Thus, for all q on the same side of H as n with $\|p-q\| \leq R$, we have $\|q-n\| \leq 2\varepsilon$. As $n \in \overline{uv}$, there is a sample $q_n \in P$ with $\|n-q_n\| \leq \varepsilon$. Importantly since $B_\varepsilon(n) \subset S^{R+\varepsilon}_{R-\varepsilon}(p)$ we can say that q_n connects to all the $P \cap S^{R+\varepsilon}_{R-\varepsilon}(p)$ on the same side of H within $\mathfrak{G}_{3\varepsilon}\left(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p)\right)$.

Thus, on each side of H, we have a single cluster of points, which are connected at 3ε. Thus, $\mathfrak{G}_{3\varepsilon}\left(P \cap S^{R+\varepsilon}_{R-\varepsilon}(p)\right)$ has two connected components. Then, Lemma 2.1 implies that p has the (R,ε)-local structure of an edge. □

4. Algorithm and its correctness.

In this section, we present the algorithm from Skyler, and prove that given P an ε-sample of an embedded graph $|G| = (G, \phi_G)$ satisfying Assumptions 3.8, the algorithm returns an isomorphic graph structure. The algorithm partitions P into P_0 and P_1, such that for each $p \in P_0$, p has the (R,ε)-local structure of a vertex, and for each $p \in P_1$, p has the (R,ε)-local structure of an edge. We then detect the number of vertices, the number of edges and the boundary operator. To obtain P_0 and P_1, we use the function $\Delta_{R,\varepsilon} : P \to \{0,1\}$, (Algorithm 1), such that if p has (R,ε)-local structure of a vertex $\Delta_{R,\varepsilon}(p) = 0$ and if p (R,ε)-local structure of an edge, $\Delta_{R,\varepsilon}(p) = 1$. Then, $P_0 = \Delta^{-1}_{R,\varepsilon}(0)$ and $P_1 = \Delta^{-1}_{R,\varepsilon}(1)$.

For each vertex $v \in |G|$, if $\deg(v) \neq 2$, Proposition 3.9 implies that for all $p \in P$ with $\|p-v\| \leq \frac{R}{2}$, $\Delta_{R,\varepsilon}(p) = 0$, while if $\deg(v) = 2$, Propositions 3.10 and 3.11 imply that $\Delta_{R,\varepsilon}(p) = 0$, and Proposition 3.12 implies that if $\|p-v\| > \frac{3R}{2} + 2\varepsilon$, $\Delta_{R,\varepsilon}(p) = 1$.

Lemma 4.1. *Let $x \in P_0$ and $\|x-v\| < \frac{3R}{2} + \varepsilon$ for vertex v. Then $y \in P_0$ is in the same connected component as x in $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$ if and only if $\|y-v\| < \frac{3R}{2} + \varepsilon$.*

Proof. By Proposition 3.12 $P_0 \subset P \cap \left\{\bigcup_{v\in V} B_{\frac{3R}{2}+\varepsilon}(v)\right\}$. Our embedding assumptions require that for vertices $v \neq v'$ we have $\|v-v'\| > \frac{9R}{2} + 3\varepsilon$ and hence no points in $P \cap B_{\frac{3R}{2}+\varepsilon}(v')$ are within $\frac{3R}{2}+\varepsilon$ of those in $B_{\frac{3R}{2}+\varepsilon}(v')$. This means they can not be connected in $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$. This implies that the entire connected component containing x must lie in $B_{\frac{3R}{2}+\varepsilon}(v)$. If $\|y-v\| > \frac{3R}{2}+\varepsilon$ then it cannot be in the same connected component as x.

We finally wish to show that $\|y-v\| < \frac{3R}{2}+\varepsilon$ implies that x and y are in the same connected component. Choose vertices u_y and u_x such that $d(y, \overline{u_y v}) < \varepsilon$ and $d(x, \overline{u_x v}) < \varepsilon$. Now let $z_y \in \overline{uv}$ be the point 3ε from v. We analogously define z_x. As P is an ε-sample of $|G|$ we have samples p_y and p_x such that $\|p_y - z_y\| < \varepsilon$ and $\|p_x - z_x\| < \varepsilon$. Note that $p_y, p_x \in P \cap B_{4\varepsilon}(v)$ and hence by Propositions 3.9 and 3.10 we know that $p_y, p_x \in P_0$. By construction $\|y-p_y\|, \|p_y - p_x\|$ and $\|p_x - x\|$ are all less that $\frac{3R}{2}+\varepsilon$ and hence y and x are in the same connected component in $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$. □

The above lemma shows the correspondence between vertices in G and connected components in $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$. Unfortunately, the situation is less clean for the connected components of $\mathfrak{G}_{3\varepsilon}(P_1)$. Around each vertex v there is a 'grey area', in which samples p can be placed in either P_0 or P_1. Due to the size of this spherical shell, it is possible to obtain connected components in $\mathfrak{G}_{3\varepsilon}(P_1)$ which contain points only within such a grey area. We devote the next few results to characterising the connected components of $\mathfrak{G}_{3\varepsilon}(P_1)$. We first show that every connected component of $\mathfrak{G}_{3\varepsilon}(P_1)$ is close to only one edge.

Proposition 4.2. *Let $[x]$ be a connected component of $\mathfrak{G}_{3\varepsilon}(P_1)$. Then there exists an edge $\overline{uv}$ such that $d(y, \overline{uv}) < \varepsilon$ for all $y \in [x]$.*

Proof. Since every sample in P is within ε of some edge it is sufficient to show that if $p, q \in P_1$ with $d(p, \overline{uv}) \leq \varepsilon$ and $\|p-q\| \leq 3\varepsilon$ then $d(q, \overline{uv}) < \varepsilon$.

As $p \in P_1$, Propositions 3.9, 3.10 and 3.11 imply

1. For all vertices $w \in |G|$ with $\deg(w) \neq 2$, $\|p-w\| > \frac{R-\varepsilon}{2}$,
2. For all vertices w with $\deg(w) = 2$, $\|p-w\| \geq 4\varepsilon$.

Without loss of generality, assume $\|p-v\| \leq \|p-u\|$. By Assumptions 3.8 (3) for all edges $\overline{xy}$ with x, y distinct from u, v, $d(\overline{uv}, \overline{xy}) > 5\varepsilon$. Hence, $d(p, \overline{xy}) > 4\varepsilon$, and for any sample q with $d(q, \overline{xy}) \leq \varepsilon$, $\|p-q\| > 3\varepsilon$. If $\deg(v) \neq 2$, then $\|p-v\| > \frac{R-\varepsilon}{2}$, and as $|G|$ satisfies Assumptions 3.8 (4), Lemma 2.3 implies $\|p-q\| > 3\varepsilon$ for all $q \in P_1$ with $d(q, \overline{uv}) > \varepsilon$.

Now assume that $\deg(v) = 2$, and consider another edge $\overline{wv}$. For $\Phi(R, \varepsilon) \leq \angle uvw < \frac{\pi}{2}$ we can apply Lemma 2.3 with $D = \frac{R-\varepsilon}{2}$ to see that for all $q \in P_1$ with $d(q, \overline{wv}) \leq \varepsilon$, $\|q, -p\| > 3\varepsilon$. For $\frac{\pi}{2} \leq \angle uvw \leq \Psi(R, \varepsilon)$, we apply Lemma 2.3 with $D = 4\varepsilon$ and observe that $\pi/2 > \arccos(23/32) + 2\arcsin(1/4)$ to conclude that $d(q, \overline{wv}) \leq \varepsilon$, $\|q-p\| > 3\varepsilon$. □

There can be multiple connected components in $\mathfrak{G}_{3\varepsilon}(P_1)$ near the same edge. However, there will only be one which contains a sample near the midpoint of the edge. We wish to treat these differently, and so we will give them a name.

Definition 4.3. We say that the connected component of $\mathfrak{G}_{3\varepsilon}(P_1)$ *spans* the edge $\overline{uv}$ if it contains a point within ε of the midpoint of $\overline{uv}$. Without reference to the specific edge $\overline{uv}$ we say that the component is *spanning*.

Proposition 4.4. *Let $\overline{uv}$ be an edge in G. There exists a unique connected component $A_{\overline{uv}}$ which spans $\overline{uv}$. $A_{\overline{uv}}$ contains samples in both $B_{\frac{3R+5\varepsilon}{2}}(u)$ and $B_{\frac{3R+5\varepsilon}{2}}(v)$.*

If $[x] \neq A_{\overline{uv}}$ is a connected component in $\mathfrak{G}_{3\varepsilon}(P_1)$ within ε of $\overline{uv}$ then either $[x] \subset B_{\frac{3R+\varepsilon}{2}}(u)$ or $[x] \subset B_{\frac{3R+\varepsilon}{2}}(v)$.

Proof. Let m denote the midpoint of $\overline{uv}$.

Let $t_0, t_1, \dots t_{2M}$ be consecutive points along $\overline{uv}$ with $\|t_i - t_{i+1}\| < \varepsilon$, $\|t_0 - u\| = \frac{3R+3\varepsilon}{2}$, $t_M = m$, and $\|t_{2M} - v\| = \frac{3R+3\varepsilon}{2}$. There must be $z_0, z_1 z_2, \dots z_M \in P$ such that $\|t_i - z_i\| < \varepsilon$. Observe that $\|z_i - u\| > \frac{3R+\varepsilon}{2}$ and $\|z_i - v\| > \frac{3R+\varepsilon}{2}$ and so by Proposition 3.12 all the z_i are in P_1. Since $\|z_i - z_{i+1}\| < 3\varepsilon$ we know that all the z_i lie in the same connected component of $\mathfrak{G}_{3\varepsilon}(P_1)$ which spans $\overline{uv}$ as z_M is within ε of m.

To see this connected component is unique, we need only observe that any pair of samples in P_1 both within ε of m are within 3ε of each other and hence lie in the same connected component. Denote this unique connected component by $A_{\overline{uv}}$.

Observe that $\|u - z_0\| < \frac{3R+3\varepsilon}{2} + \varepsilon$ and $\|v - z_{2M}\| < \frac{3R+3\varepsilon}{2} + \varepsilon$.

Now suppose that $[x] \neq A_{\overline{uv}}$ is a connected component in $\mathfrak{G}_{3\varepsilon}(P_1)$ within ε of $\overline{uv}$. Since $[x] \neq A_{\overline{uv}}$, we have $d([x], t_i) > 2\varepsilon$ for all i and hence

$$[x] \subset B_{\frac{3R+\varepsilon}{2}}(u) \cup B_{\frac{3R+\varepsilon}{2}}(v).$$

As $\|u - v\| > \frac{3R+\varepsilon}{2} + \frac{3R+\varepsilon}{2} + 3\varepsilon$ we further conclude that $[x]$ is contained in only one of $B_{\frac{3R+\varepsilon}{2}}(u)$ or $B_{\frac{3R+\varepsilon}{2}}(v)$. □

In light of Proposition 4.4 we modify our partition of P, into $\widetilde{P_0}$ and $\widetilde{P_1}$, see Definition 4.5 and Algorithm 2. We effectively want to move any points in P_1 that are not contained in a spanning connected component into P_0.

Definition 4.5 ($\widetilde{P_0}$ and $\widetilde{P_1}$)**.** Let P be an ε-sample of an embedded graph $|G|$ satisfying Assumptions 3.8, and consider the sets P_0 and P_1 from Definition 3.5. Let Q_0 be the connected components of $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$, and Q_1 the connected components of $\mathfrak{G}_{3\varepsilon}(P_1)$, and define $f : Q_1 \to \{0, 1\}$ by $f([q]) = 0$ when there is only a single connected component $[p] \in Q_0$ such that $d([p], [q]) < 3\varepsilon$, and $f([q]) = 1$ otherwise.

We define $\widetilde{P_0} := P_0 \cup \left(\bigcup_{f([x])=0}[x]\right)$ and $\widetilde{P_1} := \left(\bigcup_{f([x])=1}[x]\right)$.

Lemma 4.6. *Let $[x] \in Q_1$. Then $f([x]) = 1$ if and only is $[x]$ spans an edge, and $f([x]) = 0$ if and only if $[x] \subset B_{\frac{3R+\varepsilon}{2}}(v)$ for some vertex v.*

Proof. If $[x]$ spans an edge $\overline{uv}$ then by Proposition 4.2 we know that $[x]$ contains samples in both $B_{\frac{3R+5\varepsilon}{2}}(u)$ and $B_{\frac{3R+5\varepsilon}{2}}(v)$. Let $x_u \in [x]$ be the sample closest to u. Note that $\|x_u - u\| \leq \frac{3R+5\varepsilon}{2}$. There must be some sample $p_u \in P$ with $\|p - u\| \in< \|u - x_u\|$ and $\|p - x_u\| < 3\varepsilon$. Now $p \in P_0$ as otherwise it contradicts x_u being the closest sample to u inside $[x]$. By Lemma 4.1, $[p_u] \in Q_0$ is contained in $B_{\frac{3R+\varepsilon}{2}}(u)$.

Similarly we can show that there some $x_v \in [x]$ and $p_v \in P_0$ with $\|x_v - p_v\| \leq 3\varepsilon$ and $[p_v] \in Q_0$ contained in $B_{\frac{3R+\varepsilon}{2}}(v)$. By Lemma 4.1 $[p_u]$ and $[p_v]$ are distinct and hence $f([x]) = 1$.

If $[x]$ does not span any edge then by Proposition 4.2 we know there is a vertex v such that $[x] \subset B_{\frac{3R+\varepsilon}{2}}(v)$. We then can appeal to Lemma 4.1 to say that there is only one connected component in Q_0 within 3ε of $[x]$. □

Let $\widetilde{Q_0}$ denote the connected components of $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(\widetilde{P_0})$ and let $\widetilde{Q_1}$ denote the connected components of $\mathfrak{G}_{3\varepsilon}(\widetilde{P_1})$. We will see that characterisation of the elements of $\widetilde{Q_0}$ is the same as that of Q_0. The elements of $\widetilde{Q_1}$ are exactly those connected components that span some edge.

Theorem 4.7. *For each vertex v there exists a unique connected component $[x] \in \mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(\widetilde{P_0})$ such that $[x] \subset B_{\frac{3R}{2}+2\varepsilon}(v)$. Every connected component of $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(\widetilde{P_0})$ is of this form.*

For each edge $\overline{uv}$ there exists a unique connected component $[x] \in \mathfrak{G}_{3\varepsilon}(\widetilde{P_1})$ such that $[x]$ spans $\overline{uv}$. Furthermore every connected component of $\mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(\widetilde{P_1})$ is of this form.

Proof. From Proposition 3.12 and Lemma 4.6 we know that $\widetilde{P_0} \subset \bigcup_v B_{\frac{3R+\varepsilon}{2}}(v)$. We can then effectively repeat the proof of Lemma 4.1 to show the analogous result for $\widetilde{P_0}$.

To see the bijection between the vertices of G and $\widetilde{Q_0}$ observe that every sample within 4ε of some vertex is in $P_0 \subset \widetilde{P_0}$ and hence every vertex corresponds to some connected component, and observe that by Lemma 4.6 all points in $\widetilde{P_0}$ lie within $\frac{3R+\varepsilon}{2}$ of some vertex.

The characterisation for connected components of $\mathfrak{G}_{3\varepsilon}(\widetilde{P_1})$ follows directly from Proposition 4.2 and Lemma 4.6. □

Define the map $F_0 : \widetilde{Q_0} \to V$ by $F_0([x]) = \operatorname{argmin}_{v\in V}\{d([x], v)\}$ and $F_1 : \widetilde{Q_1} \to E$ by $F_1([x]) = \operatorname{argmin}_{\overline{uv}\in E}\{d([x], \operatorname{midpt}(\overline{uv})\}$.

That F_0 and F_1 are well defined bijections follows directly from Theorem 4.7. From Proposition 4.2 we further can say that if $[q] \in \widetilde{Q_1}$ and $[x] \in \widetilde{Q_0}$ then the single linkage distance between $[q]$ and $[x]$ is less than 3ε if and only if $F_0([x]) \in \partial_G(F_1([q]))$.

Algorithm 1: $\Delta_{R,\varepsilon}(p)$

```
Data: An $\varepsilon$-dense sample $P$ of an embedded graph $|G|$, a point $p \in P$.
Result: 0 if $p$ has local structure of a vertex, 1 if $p$ has local structure of an
        edge.
begin
    $\mathcal{G}_p \longleftarrow \{q \in P \mid \|p - q\| \leq R + \varepsilon\}$;
    connect $q, q' \in \mathcal{G}_p$ if $\|q - q'\| \leq 3\varepsilon$;
    if $\mathcal{G}_p$ is disconnected then
        return 1
    else
        remove $q \in \mathcal{G}_p$ if $\|p - q\| \leq R - \varepsilon$;
        if number of connected components in $\mathcal{G}_p$ is not 2 then
            return 0
        else
            find the midpoints $q_1, q_2$ of the connected components $c_1$ and $c_2$;
            if $\langle q_1 - p, q_2 - p\rangle > -R^2 + 2R\varepsilon - 7\varepsilon^2$ then
                return 0
            else
                return 1
```

Algorithm 2: Abstract Structure

```
Data: Partition of $P$ into $P_0$ and $P_1$.
Result: Partitions $\widetilde{P_0}, \widetilde{P_1}$, abstract graph $G = (E, V)$.
begin
    $E \longleftarrow \emptyset$;
    $V \longleftarrow \emptyset$;
    $\widetilde{P_0} \longleftarrow P_0$;
    $\widetilde{P_1} \longleftarrow P_1$;
    for connected components $[p] \in \mathfrak{G}_{\frac{3R}{2}+2\varepsilon}(P_0)$ do
        add $[p]$ to $V$
    for connected components $[q] \in \mathfrak{G}_{3\varepsilon}(P_1)$ do
        $B_q \longleftarrow \emptyset$;
        for $[p] \in V$ do
            if $\min_{p' \in [p], q' \in [q]} \|p' - q'\| \leq 3\varepsilon$ then
                add $[p]$ to $B_q$
    if $size(B_q) = 1$ then
        add all $q' \in [q]$ to $\widetilde{P_0}$ and remove them from $\widetilde{P_1}$
    else
        add $B_q$ to $E$
return $\widetilde{P_0}, \widetilde{P_1}, V, E$
```

Algorithm 3: Expectation Maximisation for Vertex Location Prediction

```
Data: $|P|$ data points in $n$ dimensions, $N_0 + N_1 = N$ many strata pieces.
Result: Predicted embedded graph vertex locations.
Input: Abstract graph structure.
begin
    Initialise vertex locations $V$ ;
    Initialise $|P| \times N$ strata assignment matrix $A$ ;
    for $s_i$ in strata pieces $S = V \cup E$, $x_j$ in data points do
        if $x_j \in s_i$ then
            $A_{i,j} \longleftarrow 1$
        else
            $A_{i,j} \longleftarrow 0$
    assign an error threshold $\sigma \in \mathbb{R}_+$;
    Initialise $\pi_i = \frac{\sum_i A_{i,j}}{\sum_{i,j} A_{i,j}}$;
    for iterations in EM-iterations do
        for $s_i$ in strata pieces $S = V \cup E$, $x_j$ in data points do
            assign $A_{i,j} = \mathbb{E}(1_{Z_j=1} | X_j = x_j)$ through (10) ;
        assign $\pi_i = \frac{\sum_i A_{i,j}}{\sum_{i,j} A_{i,j}}$;
        assign $V = \arg\min_V V \to C(V, \Pi; \sigma)$ (9) through a hill climbing
          optimiser such as gradient-descent;
```

5. Vertex prediction.

Thus far, the focus has been on finding the abstract structure of an embedded graph $|G|$. We now aim to form a numerical scheme to estimate the vertex locations of $|G| \subset \mathbb{R}^n$. In [4], a non-linear least-squares method was proposed and used for embedded graph reconstruction. Empirical observation of this method showed vertex predictions were often not within ε of the *true* embedded graph. A point of difficulty here was that data that should belong to a one-dimensional strata piece was often assigned to a zero-dimensional strata when nearby a vertex location. We utilise an Expectation-Maximisation (EM) algorithm, which updates both the predicted vertex locations, and their strata assignments to correct this issue. To do this, we design a likelihood function with latent variables for strata assignment so that we may reconstruct a probability measure over the embedded graph from which our data is sampled. Ideally, we would reconstruct a measure ν whose support is the embedded graph. Recorded data has errors and makes it computationally infeasible to reconstruct ν directly. Instead, we will formulate an approximating measure ν_δ which satisfies:

1. ν_δ is equivalent to Lebesgue measure,
2. $\operatorname{supp}(\lim_{\delta\to 0} \nu_\delta) = |G|$,

where the limit is meant in the weak sense. The first assumption gives robustness to measurement errors, and the second ensures that in ideal circumstances, we form a measure that is supported on the embedded graph. There are many measures which obey these conditions, we choose a Gaussian convolution model for each strata piece and combine all the strata pieces together through a categorical mixture model.

5.1. Embedded graph model.

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, that is Ω is a set, $\mathcal{F}$ is a σ-algebra of sets from Ω, and $\mu : \mathcal{F} \mapsto [0,1]$ is a normalised measure.

Definition 5.1. Given a probability space $(\Omega, \mathcal{F}, \mu)$ and a field with a σ-algebra $\mathcal{B}$, a measurable function $f : (\Omega, \mathcal{F}, \mu) \mapsto (\mathbb{F}, \mathcal{B})$ is a *random variable.* A vector valued random element is a vector valued measurable function $\tilde{f} : (\Omega, \mathcal{F}, \mu) \mapsto (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ given through $\tilde{f} = (f_1, \ldots, f_n)$ where each of the f_i are random variables.

The *expectation* of a random variable is the integral, $\mathbb{E}(f) \coloneqq \int_\Omega f d\mu$. Given a sub-$\sigma$-algebra $C \subset \mathcal{F}$, the *conditional expectation* of a random variable f, $\mathbb{E}(f|C) \in L^2(\Omega, \mathcal{F}, \mu)$, is the unique function that satisfies

$$\int_B \mathbb{E}(f|C) d\mu = \int_B f d\mu,$$

for all $B \in C$. The expectation and conditional expectation of a vector valued random element $\tilde{f} = (f_1, \ldots, f_n)$ is defined component-wise through each of the random variables f_i, that is

$$\mathbb{E}(\tilde{f}|C) \coloneqq (\mathbb{E}(f_1|C), \ldots, \mathbb{E}(f_n|C)) \tag{6}$$

for all $C \in \mathcal{B}(\mathbb{R}^n)$.

Above, we have adopted the standard notation $\mathcal{B}(\mathbb{R}^n)$ for the Borel-σ-algebra generated by the open sets in the standard topology on $\mathbb{R}^n$. Let $X_j : (\Omega, \mathcal{F}, \mu) \mapsto (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ be vector valued random elements and $Z_j : (\Omega, \mathcal{F}, \mu) \mapsto ([N], 2^{[N]})$ be random variables for $j \in \{1, \ldots, |P|\}$, where $[N] \coloneqq \{1, \ldots, N\}$, n is the dimension of the space to which the graph is embedded, $|P|$ is the amount of recorded data points, and N the number of strata in $|G|$. Let $N_0, N_1 \in \mathbb{N}_0$ be N_0 and N_1 are the number of zero and one dimensional strata respectively, and so $N = N_0 + N_1$.

Enumerate the set of vertex locations as $V := \{v_i\}_{i=1}^{N_0}$. For each $i \in \{N_0 + 1, \ldots, N\}$ assign the pairing $v_{i_1}, v_{i_2} \in \{v_i\}_{i=1}^{N_0}$ to be the vertices that form the boundary i^{th} strata piece. Assume that for each j the Z_j are independent and identically distributed, that each X_j is independent of X_i and Z_i for $i \neq j$.

We place the following constraints on the random variables:

1. $Z_j \sim \text{Categorical}(\Pi)$ with parameters $\Pi := (\pi_1, \ldots, \pi_N)$,
2. $\mathbb{E}(X_j | Z_j = i) \sim$ Normal(v_j, σ_j) for $j \in \{1, \ldots, N_0\}$,
3. $\mathbb{E}(X_j | Z_j = i) = t_j v_{i_1} + (1 - t_j) v_{i_2} + \varepsilon$ where $t_j \sim \text{Uniform}([0,1])$ and $\varepsilon_j \sim \text{Normal}(0, \sigma_i)$ for $i \in \{N_0 + 1, \ldots, N\}$.

The categorical random variables Z_j represent which stratum a random element X_j belongs to. The categorical distribution is defined on N many categories, with the i^{th} category having a probability of π_i of being observed. In our case, each π_i represents approximately how many data points belong to the i^{th} stratum.

The distribution of $\mathbb{E}(X_j | Z_j = i \in \{N_0 + 1, \ldots, N\}) = t v_{j_1} + (1 - t) v_{j_2} + \varepsilon$ is

$$\rho_{v_{i_1}, v_{i_2}}(x; \sigma_i) := \frac{1}{(2\pi\sigma_i^2)^{n/2}} \int_0^1 e^{-\|x - (t v_{i_1} + (1-t) v_{i_2})\|_2^2 / 2\sigma_i^2} dt, \tag{7}$$

where $\rho(\,\cdot\,; 0, \sigma_i)$ is a normal density in n dimensions with zero mean and variance σ_i^2. This can be obtained through noting that if $\nu_{v_{i_1}, v_{i_2}}$ is uniform measure on $\mathcal{L}_{v_{i_1}, v_{i_2}} := \{y \mid y = t v_{i_1} + (1 - t) v_{i_2},\ t \in [0,1]\}$, then the measure

$$\nu_{\sigma_i, v_{i_1}, v_{i_2}} = \rho_{v_{i_1}, v_{i_2}}(x; \sigma_i) dx$$

is given through

$$\begin{aligned} \nu_{\sigma_i, v_{i_1}, v_{i_2}} &= \rho(x; 0, \sigma_i) dx * \nu_{v_{i_1}, v_{i_2}} \\ &= \frac{1}{(2\pi\sigma_i^2)^{n/2}} \int_0^1 e^{-\|x - (t v_{i_1} + (1-t) v_{i_2})\|_2^2 / 2\sigma_i^2} dt dx, \end{aligned}$$

where $*$ represents the convolution operation over measures. Through this convolution construction, we have the following proposition.

Proposition 5.2. *Let $\rho_{v_{i_1}, v_{i_2}}$ and $\nu_{v_{i_1}, v_{i_2}}$ be as given above, then:*

1. $\rho_{v_{i_1}, v_{i_2}} \in C^\infty(\mathbb{R}^n)$,
2. $\rho_{v_{i_1}, v_{i_2}} dx$ *is equivalent to Lebesgue measure,*
3. *and* $\nu_{\sigma_i, v_{i_1}, v_{i_2}} \xrightarrow{\sigma_i \to 0} \nu_{v_{i_1}, v_{i_2}}$ *weakly.*

The first two claims follow from Equation 7. The third is a result from mollifier approximation theory, see [10] for details.

Corollary 5.3. *Define $\sigma := \max_i \sigma_i$ and let $\nu_\sigma := \mu(X_j^{-1})$ be the push-forward measure of μ through X_j, then*

1. $\nu_\sigma \sim dx$,
2. *supp*$(\lim_{\sigma \to 0} \nu_\sigma) = |G|$,
3. $\lim_{\sigma \to 0} \nu_\sigma(|G|) = 1$,

where $|G|$ is the embedded graph in $\mathbb{R}^n$.

Proof. Write ν_σ through

$$\begin{aligned}\nu_\sigma &= \sum_{i=1}^{N_0} \pi_i \rho(x; v_i, \sigma_i)dx + \sum_{i=N_0+1}^{N} \pi_i \rho_{v_{j_1}, v_{j_2}}(x; \sigma_i)dx \\ &= \sum_{i=1}^{N_0} \pi_i(\delta_{v_i} * \rho(x; 0, \sigma_i)dx) + \sum_{i=N_0+1}^{N} \pi_i(\nu_{v_{i_1}, v_{i_2}} * \rho(x; 0, \sigma_i)dx)\end{aligned}$$

where δ_{v_i} is the normalised measure: $\delta_{v_i}(U) = 1$ if $v_i \in U$ and zero otherwise. Let $|G|$ be the embedded graph and define $|G|_r := \{x \mid x \in B_r(y),\ y \in |G|\}$, then

$$\begin{aligned}\nu_\sigma(\mathbb{R}^n \setminus |G|_r) \leq \int_{\mathbb{R}^n \setminus |G|_r} \left(\sum_{i=1}^{N_0} \pi_i \delta_{v_i} + \sum_{i=N_0+1}^{N} \pi_i \nu_{v_{i_1}, v_{i_2}} \right) * \rho(x; 0, \sigma)dx \\ \xrightarrow{\sigma \to 0} 0 \quad \text{for all } r > 0.\end{aligned}$$

□

Corollary 5.3 shows the push-forward measure ν has our desired properties for modelling an embedded graph $|G|$.

5.2. **Parameter re-estimation.** We now form an Expectation Maximisation (EM) algorithm to find Maximum Likelihood Estimates (MLEs) for the embedded graph's vertex locations. Let $\widetilde{\mathbb{P}}(\Omega)$ be the space of probability measures over Ω. We are interested in reconstructing the measure μ given evaluations of X_j and Z_j for every $j \in \{1, \dots, n\}$. This forms the following likelihood optimisation problem:

$$\mu^* := \operatorname{argsup}_{\eta \in \widetilde{\mathbb{P}}(\Omega)} \eta \left(\bigcap_{j \in \{1, \dots, |P|\}} X_j^{-1}(B_h(x_j)) \cap Z_j^{-1}(i) \right)$$

for some small $h > 0$. For a single recorded datum:

$$\begin{aligned}&\eta(X_j^{-1}(B_h(x_j)) \cap Z_j^{-1}(i)) = \mathbb{P}(X_j \in B_h(x_j) \mid Z_j = i)\mathbb{P}(Z_j = i) \\ &= \prod_{i=1}^{N_0} \left(\pi_i \int_{B_h(x_j)} \rho(x; v_i, \sigma_i)dx \right)^{1_{Z_j = i}} \prod_{i=N_0+1}^{N} \left(\pi_i \int_{B_h(x_j)} \rho_{v_{j_1}, v_{j_2}}(x; \sigma_i)dx \right)^{1_{Z_j = i}}.\end{aligned}$$

Intersecting over all such data points, taking a logarithm, and evaluating the limit as $h \to 0$ for the argument supremum yields the equivalent optimisation:

$$\begin{aligned}\operatorname{argsup}_{\pi_i \in [0,1],\ v_i \in \mathbb{R}^n} \sum_{j=1}^{|P|} \Big(\sum_{i=1}^{N_0} 1_{Z_j = i} (\log(\rho(x_j; v_i, \sigma_i)) + \log(\pi_i)) + \\ \sum_{i=N_0+1}^{N} 1_{Z_j = i} (\log(\rho_{v_{i_1}, v_{i_2}}(x_j; \sigma_i)) + \log(\pi_i)) \Big).\end{aligned} \tag{8}$$

We cannot observe accurately Z_j for a recorded datum, although the work in estimating the abstract graph structure gives an initial estimate for this value. To dynamically update the prediction of this value, we will utilise an EM-algorithm. Projection to the sub-σ-algebra $\sigma(X_1, \dots, X_n)$ and making the assumption $Z_j \perp X_{\widetilde{j}}$ for $\widetilde{j} \neq j$ gives the following log-likelihood function, which we aim to maximise:

$$
\begin{aligned}
&\mathcal{L}(V, \Pi; \sigma) \coloneqq \qquad (9)\\
&\frac{1}{|P|}\sum_{j=1}^{|P|}\Big(\sum_{i=1}^{N_0}\mathbb{E}(1_{Z_j=i}|X_j \in B_{h'}(x_j))(\log(\rho(x_j; v_i, \sigma_i)) + \log(\pi_i)) + \\
&\qquad \sum_{i=N_0+1}^{N}\mathbb{E}(1_{Z_j=i}|X_j \in B_{h'}(x_j))(\log(\rho_{v_{i_1},v_{i_2}}(x_j; \sigma_i)) + \log(\pi_i))\Big),
\end{aligned}
$$

where we currently view $\mathcal{L}$ as a function of the vertex locations V and assignment weights Π, with σ being a fixed value. Let the densities for each $k \in \{1, \dots, N\}$ strata be enumerated as $\{\rho_k\}_{k=1}^N$. The individual terms of the cost function are

$$
\begin{aligned}
\lim_{h' \to 0}\mathbb{E}(1_{Z_j=i}|X_j \in B_{h'}(x_j)) &= \frac{\pi_i\rho_i(x_j)}{\sum_{k=1}^N \pi_k\rho_k(x_j)} \qquad (10)\\
\log(\rho(x; v_i, \sigma_i)) &= -\frac{d}{2}\log(2\pi\sigma_i) - \|x - v_i\|^2/2\sigma_i.
\end{aligned}
$$

$$
\begin{aligned}
\log(\rho_{v_{i_1},v_{i_2}}(x; \sigma_i)) = \log\Bigg(&\mathrm{erf}\left(\frac{\langle v_{i_1} - v_{i_2}, v_{i_1} + v_{i_2} - 2x\rangle + \|v_{i_1} - v_{i_2}\|_2^2}{2\sqrt{2}\|v_{i_1} - v_{i_2}\|_2\sigma_i}\right)\\
&- \mathrm{erf}\left(\frac{\langle v_{i_1} - v_{i_2}, v_{i_1} + v_{i_2} - 2x\rangle - \|v_{i_1} - v_{i_2}\|_2^2}{2\sqrt{2}\|v_{i_1} - v_{i_2}\|_2\sigma_i}\right)\Bigg)\\
&+ \frac{\langle v_{i_1} - v_{i_2}, v_{i_1} + v_{i_2} - 2x\rangle^2 - 4\|v_{i_1} - v_{i_2}\|_2^2\|(v_{i_1} + v_{i_2})/2 - x\|_2^2}{8\|v_{i_1} - v_{i_2}\|_2^2\sigma_i^2}\\
&- \log(\|v_{i_1} - v_{i_2}\|_2) + \log\left(2^{\frac{1}{2}(-d-1)}\pi^{\frac{1}{2}-\frac{d}{2}}\sigma_i^{1-d}\right).
\end{aligned}
$$

Above, $\mathrm{erf} : \mathbb{R} \mapsto \mathbb{R}$ is the standard error function given through

$$
\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt.
$$

In Skyler, the analytic gradients of the log-likelihood function $\mathcal{L}$ are given. Gradient clipping is used to bound our computations within machine accuracy for when σ_i or the evaluation of $x \mapsto \rho_{v_{i_1},v_{i_2}}(x; \sigma_i)$ is close to machine precision. Our log-likelihood function is often not concave, for instance the function $(v_{i_1}, v_{i_2}) \mapsto \rho_{v_{i_1},v_{i_2}}(x; \sigma_i)$ obeys $\rho_{v_{i_1},v_{i_2}}(x; \sigma_i) = \rho_{v_{i_2},v_{i_1}}(x; \sigma_i)$. It is necessary to have a good initialisation for the embedded graph modelling to find an acceptable local optimum value for vertex prediction. In our computations, we have found that the initial vertex modelling given by the abstract graph structure yields vertex predictions with an error less than the noise of the data, correcting the issue observed in [4]. We can complete Algorithm 3 by noting that if $A_{i,j} \coloneqq \lim_{h' \to 0}\mathbb{E}(1_{Z_j=i}|X_j \in B_{h'}(x_j))$, then the function $\Pi \to \mathcal{L}(V, \Pi; \sigma)$ is concave and has a unique maximum value at $\pi_i^* = \frac{\sum_j A_{i,j}}{\sum_{i,j} A_{i,j}}$. It can be seen that our model is a higher-dimension version of Gaussian clustering as Algorithm 3 degenerates to this when $N = N_0$.

Fixing a noise tolerance σ and solving the optimisation in Equation 8 by minimising the function $(V, \Pi) \to \mathcal{L}(V, \sigma, \Pi)$ through an EM-algorithm [7] gives Algorithm 3.

Ratio	Correct	Log Likelihood	v_1	v_2	v_3	v_4	v_5
R/ε	structure	(Equation 9)	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}4.6\\6.24\\0\end{pmatrix}$	$\begin{pmatrix}4.86\\0.51\\3.47\end{pmatrix}$	$\begin{pmatrix}-1.32\\6.29\\4\end{pmatrix}$	$\begin{pmatrix}-4.23\\-3.48\\-3\end{pmatrix}$
4	No	-	-	-	-	-	-
6	Yes	-33.183	$\begin{pmatrix}0.00\\0.03\\0.01\end{pmatrix}$	$\begin{pmatrix}4.59\\6.23\\-0.02\end{pmatrix}$	$\begin{pmatrix}4.56\\0.56\\3.43\end{pmatrix}$	$\begin{pmatrix}-1.30\\6.26\\3.96\end{pmatrix}$	$\begin{pmatrix}-4.24\\-3.46\\-3.02\end{pmatrix}$
8	Yes	-32.97	$\begin{pmatrix}0.00\\0.02\\0.01\end{pmatrix}$	$\begin{pmatrix}4.59\\6.23\\-.02\end{pmatrix}$	$\begin{pmatrix}4.86\\0.56\\3.42\end{pmatrix}$	$\begin{pmatrix}-1.29\\6.26\\3.96\end{pmatrix}$	$\begin{pmatrix}-4.22\\-3.45\\-3.01\end{pmatrix}$
10	Yes	-33.33	$\begin{pmatrix}0.01\\0.03\\0.01\end{pmatrix}$	$\begin{pmatrix}4.59\\6.22\\-0.02\end{pmatrix}$	$\begin{pmatrix}4.86\\0.56\\3.42\end{pmatrix}$	$\begin{pmatrix}-1.26\\6.24\\3.95\end{pmatrix}$	$\begin{pmatrix}-4.19\\3.42\\-2.99\end{pmatrix}$
12	Yes	-33.84	$\begin{pmatrix}0.01\\0.03\\0.01\end{pmatrix}$	$\begin{pmatrix}4.59\\6.23\\-0.02\end{pmatrix}$	$\begin{pmatrix}4.86\\0.56\\3.42\end{pmatrix}$	$\begin{pmatrix}-1.26\\6.24\\3.95\end{pmatrix}$	$\begin{pmatrix}-4.14\\-3.38\\-2.96\end{pmatrix}$
14	Yes	-36.61	$\begin{pmatrix}0.01\\0.03\\0.01\end{pmatrix}$	$\begin{pmatrix}4.59\\6.26\\-0.03\end{pmatrix}$	$\begin{pmatrix}4.86\\0.56\\3.43\end{pmatrix}$	$\begin{pmatrix}-1.26\\6.23\\3.95\end{pmatrix}$	$\begin{pmatrix}-4.00\\-3.27\\-2.56\end{pmatrix}$
16	Yes	-45.30	$\begin{pmatrix}0.02\\0.03\\0.01\end{pmatrix}$	$\begin{pmatrix}4.58\\6.27\\-0.05\end{pmatrix}$	$\begin{pmatrix}4.56\\0.56\\3.43\end{pmatrix}$	$\begin{pmatrix}-0.70\\3.70\\2.33\end{pmatrix}$	$\begin{pmatrix}-3.96\\-3.22\\-2.81\end{pmatrix}$

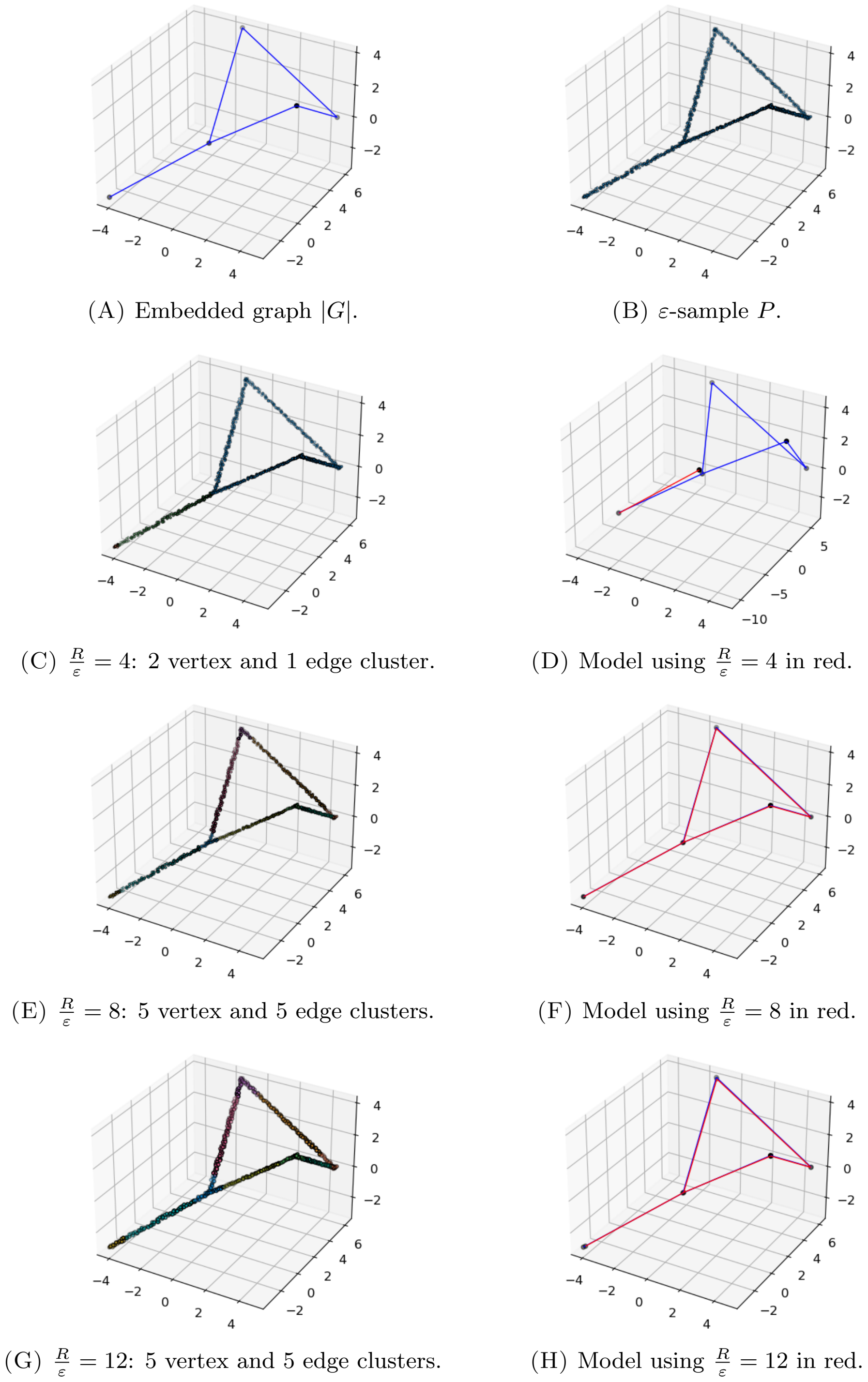
TABLE 1. Summary of the output of the algorithm for various ratios $\frac{R}{\varepsilon}$. Recall we wish to maximise Equation 9. The last 5 columns are the vertex locations obtained.

5.3. **Numerical simulations.** The conditions in Assumption 3.8 are not the sharpest bounds, and other ratios of R and ε can also detect the correct graph structure. We present the results of a few different ratios, for the same 0.1-sample P (Figure 5.9B) of the embedded graph $(G, \phi_G) \subset \mathbb{R}^3$ (Figure 5.9A). There are 705 samples in P, and G has 5 vertices embedded as 1: $(0, 0, 0)$, 2: $(4.6, 6.24, 0)$ 3: $(4.86, 0.51, 3.47)$, 4: $(-1.32, 6.29, 4)$, and 5: $(-4.23, -3.48, -3)$, and edges $E = \{(1, 5), (1, 3), (1, 4), (2, 4), (2, 3)\}$.

Table 1 shows the results with varying choices of ratio $\frac{R}{\varepsilon}$. Comparing the log-likelihood of the models obtained using $\frac{R}{\varepsilon} = 8$ (-2.3712314714356437) and $\frac{R}{\varepsilon} = 12$ (-2.783827546761547), we see that while we have shown that $R \geq 12\varepsilon$ is sufficient to prove correctness of the algorithm, smaller ratios can also identify an isomorphic graph structure, and result in a higher log-likelihood model. In practice, this suggests that we can improve the process by first using $R \geq 12\varepsilon$ to obtain the correct structure, and then decreasing the ratio to model the graph, stopping when we still obtain the correct graph structure and maximise the log-likelihood.

6. **Future directions.** The algorithm presented in this paper focuses on recovering and modelling an embedded graph (G, ϕ_G) given an ε-sample P. Stratified spaces, however, are not restricted to consisting of 0- and 1-dimensional pieces, nor are they restricted to being simplicial complexes. We can consider embeddings of CW complexes, where a stratum is embedded as a semi-algebraic set.

While the algorithm in this paper does not naively extend to higher simplicial complexes or CW complexes, it provides a foundation on which other algorithms can be based, and hence moves towards learning general stratified spaces. The



(A) Embedded graph $|G|$.

(B) ε-sample P.

(C) $\frac{R}{\varepsilon} = 4$: 2 vertex and 1 edge cluster.

(D) Model using $\frac{R}{\varepsilon} = 4$ in red.

(E) $\frac{R}{\varepsilon} = 8$: 5 vertex and 5 edge clusters.

(F) Model using $\frac{R}{\varepsilon} = 8$ in red.

(G) $\frac{R}{\varepsilon} = 12$: 5 vertex and 5 edge clusters.

(H) Model using $\frac{R}{\varepsilon} = 12$ in red.

FIGURE 5.9

algorithm can be adapted to other cases and assumptions. For example, it can be adapted to learn the abstract structure of a graph with non-linear edges and no degree 2 vertices. In particular, to recover embedded CW complexes, we need to remove the assumption that strata are embedded as convex hulls (linearity). Hence, there are two distinct paths forward:

1. Develop an algorithm which identifies the abstract structure of simplicial complexes with 2-dimensional simplices,
2. Explore methods for removing the linearity assumption (even for graphs).

Focusing on increasing the dimension of the cells in the simplicial complex, the next step is to allow 2-simplicies and partition an ε-sample P into three parts P_0, P_1, and P_2. One approach is a peeling argument: first we determine the points in P_2, and then apply the current algorithm to $P \setminus P_2$ to obtain P_1 and P_0. Complications with this include ensuring that points are not over-assigned to $P \setminus P_2$, as this can result in $P \setminus P_2$ not being suitable as input for the current algorithm. To appropriately partition P, we hope to exploit the relationship between (R, ε)-local structure and local homology. For graphs, we saw that the dimension 1 local homology at a point x contains topological information, which corresponds to the number of points in the intersection of the $|G|$ with a ball of small radius r around x, and if there are 2 points, their relative geometry providing more information. By generalising the (R, ε)-local structure appropriately, we hope to see a correspondence with the information contained in higher homology groups and augment this with other geometrical information.

To remove the linearity assumption, we need to address a long standing problem in computational algebraic geometry: learning algebraic varieties from noisy samples. In [5], Breiding et al. develop an algorithm which is robust to machine error but not sampling noise. The algorithm has also been found to fail when given large data sets sampled from simple varieties. These issues need to be overcome before we can remove the linearity assumption.

Acknowledgments. The first and third authors are supported by Australian Government Research Training Program (RTP) Fee Offset Scholarships. The second author is the recipient of an Australian Research Council Discovery Early Career Award (project number DE200100056) funded by the Australian Government. The authors would also like to thank Jane Tan and Martin Helmer for detailed comments on early drafts, and Cale Rankin and Ivo Vekemans for insightful discussions.

The authors are sincerely grateful to the reviewer for suggesting improvements to the paper.

REFERENCES

[1] M. Aanjaneya, F. Chazal, D. Chen, M. Glisse, L. Guibas and D. Morozov, Metric graph reconstruction from noisy data, *Internat. J. Comput. Geom. Appl.*, **22** (2012), 305–325.

[2] P. Bendich, B. Wang and S. Mukherjee, Local homology transfer and stratification learning, in *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM, New York, 2012, 1355–1370.

[3] P. Bendich, B. Wang and S. Mukherjee, Towards stratification learning through homology inference, *AAAI Fall Symposium on Manifold Learning and its Applications (AAAI)*, 2010. Available from: https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/download/2273/2714.

[4] Y. Bokor, D. Grixti-Cheng, M. Hegland, S. Roberts, and K. Turner, Stratified space learning: Reconstructing embedded graphs, *23rd International Congress on Modelling and Simulation*, Australia, 2019. Available from: https://mssanz.org.au/modsim2019/A3/bokor.pdf.

[5] P. Breiding, S. Kališnik, B. Sturmfels and M. Weinstein, Learning algebraic varieties from samples, *Rev. Mat. Complut.*, **31** (2018), 545–593.

[6] S.-W. Cheng, T. K. Dey and E. A. Ramos, Manifold reconstruction from point samples, *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM, New York, 2005, 1018–1027.

[7] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. Ser. B*, **39** (1977), 1–38.

[8] T. K. Dey, *Curve and Surface Reconstruction: Algorithms with Mathematical Analysis*, Cambridge Monographs on Applied and Computational Mathematics, 23, Cambridge University Press, Cambridge, 2007.

[9] T. K. Dey, F. Fan and Y. Wang, Dimension detection with local homology, 26th Canadian Conference on Computational Geometry, Nova Scotia, 2014. Available from: http://www.cccg.ca/proceedings/2014/papers/paper40.pdf.

[10] E. M. Stein and R. Shakarchi, *Real Analysis. Measure Theory, Integration, and Hilbert Spaces*, Princeton Lectures in Analysis, 3, Princeton University Press, Princeton, NJ, 2005.

[11] B. J. Stolz, J. Tanner, H. A. Harrington and V. Nanda, Geometric anomaly detection in data, *Proc. Natl. Acad. Sci. USA*, **117** (2020), 19664–19669.

Received January 2021; revised June 2021; early access September 2021.

E-mail address: yossi.bokor@anu.edu.au
E-mail address: katharine.turner@anu.edu.au
E-mail address: christopher.williams@anu.edu.au