Fixed and adaptive landmark sets for finite metric spaces

Jason Cory Brunson Yara Skaf

1 Introduction

Topological data analysis (TDA) is a maturing field in data science at the interface of statistics, computer science, and mathematics. Topology is the discipline at the intersection of geometry (the study of shape) and analysis (the study of continuity) that focuses on geometric properties that are preserved under continuous transformations. TDA consists of the use of computational theories of continuity to investigate or exploit the structure of data. While TDA is most commonly associated with persistent homology (PH) and mapper-like constructions, it can be understood to include such classical and conventional techniques as cluster analysis, network analysis, and nearest neighbors prediction.

TDA methods, including the ~~canonical/recognizable/archetypal (i.e. the most recognizable as a result of historical contingency)~~ persistent homology and mapper (Skaf and Laubenbacher 2022) as well as manifold learning dimension reduction tools (Ivakhno and Armstrong 2007; Reutlinger and Schneider 2012; Aziz et al. 2017; Viswanath et al. 2017; Konstorum et al. 2018; Becht et al. 2019), have been deployed widely in biomedicine. These methods require that data be pre-processed into the form of Euclidean vectors, and for many varieties of biomedical data, including most image and omics data, this is a straightforward requirement. In contrast, analysis tasks in clinical and public health often involve data that have diverse numerical and categorical variable types, high rates and complex patterns of missingness, or only a small number of variables taking finitely many values. A variety of approaches have been taken to apply classical topological methods to clinical and public health data, whether by transformaing records to coordinate vectors ("vector space embeddings") or by using non-metric similarity measures such as those common in ecology (Johnston 1976; Lee, Maslove, and Dubin 2015; Dai, Zhu, and Liu 2020). Extensions of these ideas to the present-day TDA toolkit are needed to bring the ~~potential/power/value~~ of more advanced methods to bear in domains that rely on more challenging data.

One support tool for TDA is the selection of landmarks or witnesses from a data set, which can reduce the conceptual or computational complexity of an analysis. Maxmin is a common such selection procedure, which has been used to expedite the calculation of PH (de Silva and Carlsson 2004) or of mapper (Singh, Mémoli, and Carlsson 2007) and to compute computational-topological representations of data directly. The maxmin sampling procedure is often used for this purpose, as it is deterministic, is computationally efficient, and generates more evenly distributed samples than random selection. In addition to approximating PH, maxmin was used in these cases to reduce the sizes of simplicial complex models of point cloud data for the sake of visualization and exploration.

In this paper, we describe an adaptation of the maxmin procedure to health data that are not naturally represented as Euclidean vectors. We focus on three questions: (1) How is maxmin most naturally modified to sample landmarks from health data? (2) What useful properties do the modified procedure and its samples and covers have? (3) How does the procedure perform at certain analysis tasks on real-world data, and is its performance comparable or superior to that of maxmin? We propose a procedure that is deterministic and based not on a raw similarity or distance measure but on a rank-order of neighboring cases to each index case.

We organize the paper as follows: In the remainder of this section, we introduce mathematical notation used throughout and use a simple example to motivate the procedure as a counterpart to maxmin. We carefully define and prove algorithms and other properties for the procedure in Section 2, after which we describe our implementation and several experiments performed on

real-world data to address our motivating questions. In Section 3 we summarize key definitions, algorithms, and properties, report the results of benchmark tests and robustness checks, and report the results of experiments. We interpret our findings and comment on limitations and ongoing work in Section 4.

1.1 Conventions

(X, d_X) will refer to a finite pseudometric space with point set X and pseudometric $d_X : X \times X \to \mathbb{R}_{\geq 0}$, which by definition satisfies all of the properties of a metric except that $d_X(x, y) = 0$ implies $x = y$. (X, d_X) may be shortened to X, and d_X to d, when clear from context. If $x \neq y$ but $d(x, y) = 0$ then x and y are said to be indistinguishable or co-located. The cardinality of $Y \subseteq X$ (counting multiplicities) is denoted $|Y|$, the set of points co-located with those in Y is denoted $\overline{Y}$, and the set of equivalence classes of Y under co-location is denoted $[Y]$. Throughout, let $N = |X|$. When $Y, Z \subseteq X$, let $Y \setminus Z$ denote the set difference $\{x : x \in Y \wedge x \notin Z\}$. Then $Y \setminus \overline{Z}$ is the set of points in Y (with multiplicities) that are distinguishable from all points in Z. This means that, when defined, $\min_{y \in Y \setminus \overline{Z}, z \in Z} d(y, z) > 0$.

We denote the diameter $D(Y) = \max_{x,y \in Y} d(x, y)$ and write:

$$d(Y, Z) = \min_{y \in Y, z \in Z} d(y, z) \qquad D(Y, Z) = \max_{y \in Y, z \in Z} d(y, z)$$
$$d(x, Y) = d(\{x\}, Y) \qquad D(x, Y) = D(\{x\}, Y)$$

If, for any $x, y, z, w \in X$, $d(x, y) = d(z, w)$ implies $\{x, y\} = \{z, w\}$—that is, if no two pairs of points in X have equal distance—then X is said to be in general position. We also say that X is in *locally general position* if, for any $x, y, z \in X$, $d(x, y) = d(x, z)$ implies $y = z$—a weaker condition, since there may exist $w \in X$ for which $d(x, y) = d(z, w)$ but $\{x, y\} \neq \{z, w\}$. Either condition implies that X is Hausdorff: $d(x, y) = 0$ implies $x = y$. $f : X \to Y$ will denote a morphism of pseudometric spaces, which we take to be a 1-Lipschitz map: $d_X(x, y) \geq d_Y(f(x), f(y))$.

Denote by $\mathcal{P}(X)$ the power set of X and by $\mathcal{Q}(X)$ the set of ordered, non-duplicative sequences from X. We use the ball notation $B_\varepsilon(x) \in \mathcal{P}(X)$ for the set of points less than distance ε from a point x; that is, $B_\varepsilon(x) = \{y : d(x, y) < \varepsilon\}$. We use an overline to also include points exactly distance ε from x: $\overline{B_\varepsilon}(x) = \{y : d(x, y) \leq \varepsilon\}$. If $|\overline{B_\varepsilon}(x)| \geq k$ and $\varepsilon' < \varepsilon \implies |\overline{B_{\varepsilon'}}(x)| < k$, then we call $N_k(x) = \overline{B_\varepsilon}(x)$ the k-nearest neighborhood of x. When X is in locally general position, $|N_k(x)| = k$.

For convenience, we assume $0 \in \mathbb{N}$. For $a, b \in \mathbb{N}$ with $a < b$, we use $[a, b]$ to denote the arithmetic sequence $(a, a + 1, \ldots, b)$. For $a, b \in \mathbb{N}$, we use a^b to denote the sequence $(a, \ldots, a)$ of length b.

1.2 Motivation

In contrast to geometric data analytic tools like principal components analysis that reduce the dimensionality of data, much TDA relies on *cardinality reduction*. As distinguished by Byczkowska-Lipińska and Wosiak (2017), an $n \times p$ data table of n cases (rows) and p variables (columns) can be dimension-reduced to an $n \times q$ table, where $q < p$, or cardinality-reduced to an $m \times p$ table, where $m < n$. The most common cardinality reduction method is data reduction, and unsupervised clustering methods are classical examples. Many popular TDA techniques use cardinality rather than dimension reduction to improve efficiency, often through landmark or witness sampling: de Silva and Carlsson (2004) propose witness complexes, related to alpha complexes (Akkiraju et al.

1995), for the rapid approximation of PH: Given a point cloud, a set of landmark points and their overlapping neighborhoods define a nerve, which stands in for the Vietoris–Rips complex at each scale. They use maxmin as an alternative to selecting landmark points uniformly at random, which ensures that the landmarks are locally separated and roughly evenly distributed. Other uses include the selection of a sample of points from a computationally intractable point cloud for the purpose of downstream topological analysis, as when performing the mapper construction (Singh, Mémoli, and Carlsson 2007); and the heuristic optimization of a fixed-radius ball cover of a point cloud, in the sense of minimizing both the number of balls and their shared radius (Dłotko 2019). However, maxmin comes with its own limitations in the analysis of data that vary greatly in density or have many multiplicities. This is a frequent concern when sparse, heterogeneous, and incomplete data are modeled as finite pseudometric spaces.

Especially in analyses of medical and healthcare data, underlying variables can often only be understood as ordinal, and high-dimensional data sets are commonly analyzed using similarity measures rather than vector space embeddings. Furthermore, because measurements are coarse and often missing, such data often contain indistinguishable entries—cases all of whose measurements are equal and that are therefore represented as multiple instances of the same point. All of these attributes violate the assumptions of the ball cover approach and suggest the need for an ordinal counterpart.

These considerations motivate us to produce a counterpart to the ball cover that we call the *neighborhood cover*, each set of which may have a different radius but (roughly) the same cardinality. The two approaches are visually contrasted in Section 1.3. Later sections will compare their performance on several tasks using real-world data, and specifically data and tasks for which precise sample sizes can be advantageous. From these experiments, then, we want to know both whether cardinality-based methods outperform distance-based methods (superiority), which might be expected, but also whether cardinality-based methods are not outperformed by distance-based methods (non-inferiority), so that they can be recommended when specific sample sizes are preferable for reasons other than performance.

1.3 Examples

We motivate our alternative sampler using two examples. The first suggests a practical setting in which fixed-cardinality cover sets are more desirable for applications. The second provides an abstraction in which they are better able to detect topological properties.

Example 1. Bimodal distribution of risk

Imagine an intensive care unit whose patients fall roughly into three groups: a large, clinically homogeneous, low-risk majority; a smaller, more heterogeneous, higher-risk group; and a minority of highly distinctive patients who cannot be sorted into either group and are at less predictable risk. The top panel of Figure 1 depicts a simple model of this situation in which each group is represented by a Gaussian distribution. <!–The shape approximates an empirical distribution of standard risk scores observed for patients in the MIMIC-III database (see Section 2.3.1). (Imagine that probabilistic risk has been logit-transformed, so that all real values are possible risk estimates.)–> The points X along the abscissa are sampled randomly from this distribution.

The bottom panel of Figure 1 shows how the maxmin and lastfirst procedures generate sequences in X of four landmarks each. Each procedure begins with a seed landmark ℓ_0 selected to be reachable

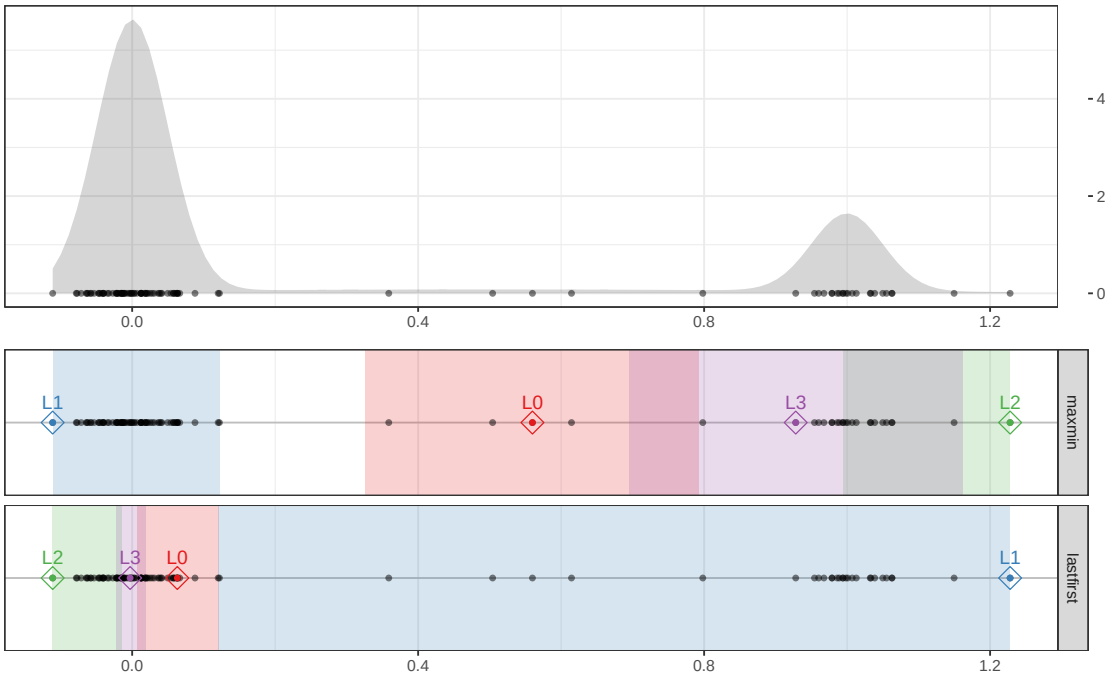


Figure 1. Landmark samples from the imagined ICU sample and their associated covers using two selection procedures.

by minimum-radius balls (maxmin) or by minimum-cardinality neighborhoods (lastfirst) from all other points in X; see the appendix for details. Whereas the maxmin landmarks are roughly equally-spaced across the range of X, the lastfirst landmarks lie at roughly equal quantiles of X. The lastfirst procedure will be advantageous when subpopulations can be discriminated at different resolutions and the ability to do so is limited primarily by sample size.

Example 2. Bimodal density on a circle

Consider a probability density function that varies significantly over $\mathbb{S}^1$ and a sample X from its distribution. A paradigmatic goal of TDA would be to detect the 1-dimensional feature of $\mathbb{S}^1$ from the sample. Given a small subset of landmarks $L \subset X$, we would want a witness complex $\mathrm{Wit}(L, X)$ to reliably satisfy $H_1(\mathrm{Wit}(L, X)) = 1$. We would also want to reduce complexity, computational cost, and detection of spurious features.

Figure **??** shows the covers obtained from 12 landmarks generated using the maxmin and lastfirst procedures, in both cases increasing the sizes of the sets twofold from minimality. It can be seen that the maxmin witness complex (the nerve of the cover) fails to detect the 1-feature while the lastfirst witness complex succeeds. As will be shown later in the paper, this superior performance is robust with respect to several parameters governing the distribution and the samplers, and in

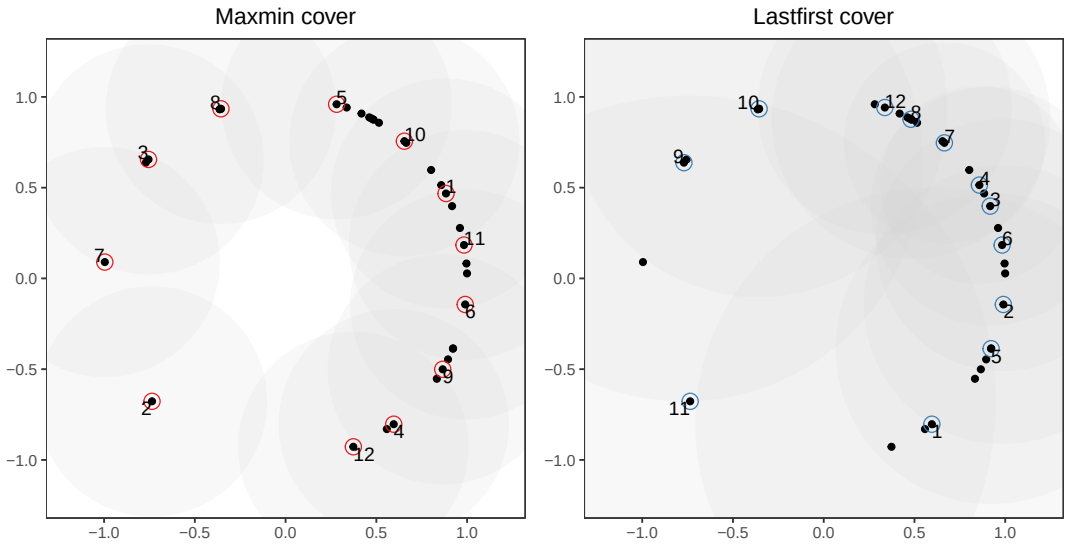


Figure 2. Landmark samples and their covers from the bumpy distribution on the circle using two selection procedures.

particular that the lastfirst witness complex detects the feature over a wider range of sizes of $|L|$.

2 Materials and Methods

This section provides mathematical proofs of algorithms and oher properties of the landmark samplers (Section 2.1), summarizes their implementation as an R package (Section 2.2), and describes several experiments performed on real-world data to address our motivating questions (Sections 2.3 and 3.3).

2.1 Samplers

In this section we define both maxmin and its counterpart lastfirst, provide algorithms for their implementation, and prove some additional properties about them. The concepts and results are summarized in Section **??**.

2.1.1 Maxmin procedure

Definition 3 (maxmin)**.** Given (X, d) and $Y \subset X$, define the *maxmin set*

$$\operatorname{maxmin}(Y) = \operatorname{maxmin}(Y; X) = \{x \in X \setminus \overline{Y} : d(x, Y) = \max_{y \in X \setminus \overline{Y}} d(y, Y)\}$$

consisting of *maxmin points*.

Note that $\text{maxmin}(Y)$ is nonempty when $X \setminus \overline{Y} \neq \varnothing$ and that $|\text{maxmin}(Y)| = 1$ when X is in locally general position.

Algorithm 1 Select a maxmin landmark set.

Require: finite pseudometric space (X, d)
Require: at least one parameter $n \in \mathbb{N}$ or $\varepsilon \geq 0$
Require: seed point $\ell_0 \in X$
Require: selection procedure σ
if ε is not given **then**
  $\varepsilon \leftarrow \infty$
end if
if n is not given **then**
  $n \leftarrow 1$
end if
$L \leftarrow \varnothing$
$i \leftarrow 0$
repeat
  $L \leftarrow L \cup \{\ell_i\}$
  $i \leftarrow i + 1$
  $\ell_i \leftarrow \sigma(\text{maxmin}(L))$
  $d_{\max} \leftarrow d(\ell_i, L)$
until $d_{\max} < \varepsilon$ and $|L| \geq n$
return maxmin landmark set L

The *maxmin procedure* for generating a landmark set $L \subseteq X$ proceeds as follows (see Algorithm 1). Each step receives a proper subset $L \subset X$ and returns a point $x \in X \setminus \overline{L}$.

First, choose a number $n \leq |[X]|$ of landmark points to generate or a radius $\varepsilon \geq 0$ for which to require that the balls $\{\overline{B_\varepsilon}(\ell) : \ell \in L\}$ cover X. Choose a first landmark point $\ell_0 \in X$.[1] Inductively over $i \in \mathbb{N}$, if ever $i \geq n$ or $d(L, X \setminus \overline{L}) \leq \varepsilon$, then stop. Otherwise, when $L = \{\ell_0, \ldots, \ell_{i-1}\}$, choose $\ell_i \in \text{maxmin}(L)$, according to a preferred procedure σ (see the Appendix). If n was prescribed, then set $\varepsilon = \varepsilon(n) = d(L, X \setminus \overline{L})$; if ε was prescribed, then set $n = n(\varepsilon) = |L|$. Write $\mathcal{B}_{n,\varepsilon}(\ell_0) = \{\overline{B_\varepsilon}(\ell_i)\}_{i=0}^{n}$ for the resulting *landmark ball cover of* X.

We will write the elements of landmark sets $L = \{\ell_0, \ldots, \ell_{n-1}\}$ in the order in which they were generated. Note that, if $n = |[X]|$ or $\varepsilon = 0$, then $\overline{L} = X$. When the procedure stops, $X = \bigcup_{i=0}^{n-1} \overline{B_\varepsilon}(\ell_i)$. This cover is not, in general, a minimal cover, but it is a minimal landmark cover in the sense that the removal of $\overline{B_\varepsilon}(\ell_{n-1})$ or any decrease in ε will yield a collection of sets that fails to cover X. A "thickened" cover can be obtained by pre-specifying both n and ε in such a way that $n \geq n(\varepsilon)$ and $\varepsilon \geq \varepsilon(n)$. In Section 2.2, we describe two adaptive parameters implemented in our software package that make these choices easier.

Maxmin is a heuristic, iterative procedure used to select a well-dispersed subset of points from X, where dispersion is understood in terms of the interpoint distances of this subset. At each step, landmarks $L = \{\ell_0, \ldots, \ell_{i-1}\}$ having been selected, the next point ℓ_i is selected to maximize its

[1] This choice may be arbitrary; we specifically consider three selection rules: the first point index in the object representing X, selection at random, and the Chebyshev center $\text{argmin}_{x \in X} D(x, X \setminus \{x\})$.

minimum distance $d(\ell_j, \ell_i)$ from the ℓ_j, $0 \leq j < 1$. Equivalently, $\ell_i \in X \setminus L$ is selected so that the minimum radius ε required for $\ell_i \in \bigcup_{j=0}^{i-1} \overline{B_\varepsilon}(\ell_j)$ is maximized. This is a useful heuristic for constructing a ball cover $\mathcal{B} = \{\overline{B_\varepsilon}(\ell_j) : 0 \leq j < n\}$ centered at a highly mutually distant set of landmarks.

We desire in the next section to construct a neighborhood cover $\mathcal{N} = \{N_k^+(\ell_j) : 0 \leq j < n\}$ whose centers are analogously dispersed. Accordingly, let us redefine the maxmin procedure in terms of balls rather than of distances:

Proposition 4 (maxmin in terms of balls)**.** *Given (X, d) and $Y \subset X$, write $B_\varepsilon(Y) = \bigcup_{y \in Y} B_\varepsilon(y)$ and similarly for closed balls, then let*

$$\mathrm{E}(Y, X) = \min\{\varepsilon : \overline{B_\varepsilon}(Y) = X\}$$

(using a capital epsilon). Then

$$\operatorname{maxmin}(Y; X) = X \setminus B_{\mathrm{E}(Y,X)}(Y)$$

This yields the alternative loop for Algorithm 1, excerpted as Algorithm 2.

Algorithm 2 Select a maxmin landmark set, in terms of balls (loop).

```
repeat
    L ← L ∪ {ℓ_i}
    i ← i + 1
    ε_min ← E(L, X)
    F ← X ∖ B_{ε_min}(L) (maxmin set)
    ℓ_i ← σ(F)
until ε_min < ε and |L| ≥ n
```

2.1.2 Lastfirst procedure

The lastfirst procedure is defined analogously to the maxmin procedure, substituting nearest neighborhoods, parameterized by their cardinality k, for balls, parameterized by their radius ε. Unlike distance, membership in nearest neighborhoods is not a symmetric relation: It may be that $y \in N_k(x)$ while $x \notin N_k(y)$. We therefore introduce a companion concept that reverses this relationship:

Definition 5 (k-neighborhoods)**.** For $x \in X$, define the *k-out-neighborhoods* N_k^+ and *k-in-neighborhoods* N_k^- of x as the sets

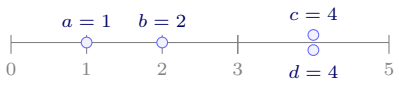
$$N_k^+(x) = \{y \in X : y \in N_k(x)\} = N_k(x)$$
$$N_k^-(x) = \{y \in X : x \in N_k(y)\}$$

Given a subset $Y \subseteq X$, we also define

$$N_k^+(x, Y) = Y \cap N_k^+(x)$$
$$N_k^-(x, Y) = Y \cap N_k^-(x)$$

Note that $[\{x\}] = N_0^{\pm}(x) \subseteq \cdots \subseteq N_{N-1}^{\pm}(x) = X$. The k-out-neighborhoods of x are the k-nearest neighbors of x, i.e. the sets of points in X that have out-rank at most k from x. The k-in-neighborhoods of x are the sets of points in X from which x has out-rank at most k.

Example 6. Consider the simple case $X = \{a, b, c, d\}$, visualized below, equipped with the standard Euclidean metric:



Compute N_k^+ and N_k^- for a and c, using $k = 3$:

$$N_3^+(a) = \{a, b, c, d\}, \quad N_3^+(c) = \{b, c, d\}, \quad N_3^-(a) = \{a, b\}, \quad N_3^-(c) = \{a, b, c, d\}$$

Remark 7. If $|N_k^{\pm}(x)| = n_k$, then $|N_k^{\pm}(x, X \setminus \{x\})| = n_k - 1$.

Example 8. Continuing Example 6, we can compute the other $N_\bullet^+$ and $N_\bullet^-$ for a and c:

$$\begin{aligned} N_\bullet^+(a) &= (|N_0^+(a)|, |N_1^+(a)|, |N_2^+(a)|, |N_3^+(a)|) \\ &= (|\{a\}|, |\{a, b\}|, |\{a, b, c, d\}|, |\{a, b, c, d\}|) \\ &= (1, 2, 4, 4) \end{aligned} \qquad \begin{aligned} N_\bullet^+(c) &= (|N_0^+(c)|, |N_1^+(c)|, |N_2^+(c)|, |N_3^+(c)|) \\ &= (|\{c, d\}|, |\{c, d\}|, |\{b, c, d\}|, |\{a, b, c, d\}|) \\ &= (2, 2, 3, 4) \end{aligned}$$

$$\begin{aligned} N_\bullet^-(a) &= (|N_0^-(a)|, |N_1^-(a)|, |N_2^-(a)|, |N_3^-(a)|) \\ &= (|\{a\}|, |\{a, b\}|, |\{a, b\}|, |\{a, b, c, d\}|) \\ &= (1, 2, 2, 4) \end{aligned} \qquad \begin{aligned} N_\bullet^-(c) &= (|N_0^-(c)|, |N_1^-(c)|, |N_2^-(c)|, |N_3^-(c)|) \\ &= (|\{c, d\}|, |\{c, d\}|, |\{a, b, c, d\}|, |\{a, b, c, d\}|) \\ &= (2, 2, 4, 4) \end{aligned}$$

As with maxmin, a selection procedure σ must be chosen, which may be needed even if X is in general position. A thorough discussion of our preferred procedure, which adapts the Chebyshev center to the neighborhood setting, is in the Appendix. We operationalize it as well as our lastfirst procedure by way of a total ordering on the $N_\bullet^{\pm}$.

Definition 9 (total orders on rank sequences)**.** Let $a_n = (a_1, \ldots, a_M), b_n = (b_1, \ldots, b_M) \in \mathbb{N}^M$. Then $a_n < b_n$ in the reverse lexicographic (revlex) order if $\exists i : a_i > b_i \wedge (\forall j < i : a_j = b_j)$.

Impose the revlex order on the $N_\bullet^+$ and $N_\bullet^-$ to emphasize the sizes of smaller neighborhoods. Sequences with more large values indicate points with lower out-ranks to or from more other points.

We now define our counterpart to the maxmin procedure.

Definition 10 (lastfirst, in terms of neighborhoods)**.** Given (X, d) and $Y \subset X$, write $N_k(Y) = \bigcup_{y \in Y} N_k(y)$, then let

$$K(Y, X) = \min\{k : N_k(Y) = X\}$$

Then define the *lastfirst set*

$$\mathrm{lf}(Y) = \mathrm{lf}(Y; X) = X \setminus N_{K(Y,X)-1}(Y)$$

consisting of *lastfirst points*.

The *lastfirst procedure* proceeds analogously to the maxmin procedure (Algorithm 3). Write $\mathcal{N}_{n,k}(\ell_0) = \{N_k(\ell_i)\}_{i=0}^{n}$ for the resulting *landmark neighborhood cover of* X.

Algorithm 3 Calculate the lastfirst landmark sequence from a seed point.

Require: finite pseudometric space (X, d)
Require: at least one parameter $n \in \mathbb{N}$ or $k \in \mathbb{N}$
Require: seed point $\ell_0 \in X$
Require: selection procedure σ
if k is not given **then**
  $k \leftarrow \infty$
end if
if n is not given **then**
  $n \leftarrow 1$
end if
$L \leftarrow \varnothing$
$i \leftarrow 0$
repeat
  $L \leftarrow L \cup \{\ell_i\}$
  $i \leftarrow i + 1$
  $k_{\min} \leftarrow K(L, X)$
  $F \leftarrow X \smallsetminus N_{k_{\min}-1}(L)$ (lastfirst set)
  $\ell_i \leftarrow \sigma(F)$
until $k_{\min} > k$ and $|L| \geq n$
return lastfirst landmark set L

We defer further proofs to the Appendix, where we rely on additional technical definitions. The results apply to a more general conception of "relative rank" (our term), one of which is induced by the pseudometric on any finite pseudometric space but which need only satisfy the condition that a point is no nearer any other point than itself.

Corollary 11 (lastfirst using relative rank)**.** *Given $Y \subset X$ and a pseudometric d on X with relative rank q,*

$$\mathrm{lf}(Y; X) = \mathrm{maxmin}(Y; X, q_d)$$
$$\mathrm{lf}(X, d) = \mathrm{maxmin}(X, q_d)$$

Proposition 12. *Algorithm 4 returns a lastfirst landmark set. If $n \leq |[X]|$ is given as input and k is not, then $|L| = n$. If n and k are both given, then $|L| \geq n$. Otherwise, L is minimal in the sense that no proper prefix of L gives a cover of X by k-nearest neighborhoods.*

Example 13. Return again to $X = \{a, b, c, d\}$ from Example 17. We calculate an exhaustive lastfirst landmark set, seeded with a point of minimal out-neighborhood sequence:

1. We have

$$N_\bullet^+(a, \{b, c, d\}) = (0, 1, 3, 3) \qquad N_\bullet^+(b, \{a, c, d\}) = (0, 1, 3, 3)$$
$$N_\bullet^+(c, \{a, b, d\}) = (1, 1, 2, 3) \qquad N_\bullet^+(d, \{a, b, c\}) = (1, 1, 2, 3)$$

   Under the revlex order, $\operatorname{argmin}_{x \in X} N_\bullet^+(x, X \setminus \overline{\{x\}}) = \{c, d\}$, and we arbitrarily select $\ell_0 = c$ and $L = \{c\}$.

2. For $x \in X \setminus \overline{L}$ we now have

$$N_\bullet^-(a, \{c\}) = (0, 0, 0, 1) \qquad N_\bullet^-(b, \{c\}) = (0, 0, 1, 1)$$

   Under the revlex order, $\operatorname{argmax}_{x \in X \setminus \overline{\{c\}}} N_\bullet^-(x, \{c\}) = \{a\}$; we select $\ell_1 = a$, so that now $L = \{c, a\}$.

3. Only one point remains in $X \setminus \overline{L} = \{b\}$, so the exhaustive landmark set is $\{c, a, b\}$.

2.2 Implementation

We have implemented both procedures in the R package landmark (Brunson and Skaf 2021), borrowing a maxmin implementation by Piekenbrock (2020). Each procedure is implemented for Euclidean distances in C++ using Rcpp (Eddelbuettel and Francois 2011) and for many other distance metrics and similarity measures in R using the proxy package (Meyer and Buchta 2021). For relative rank–based procedures, the user can choose any tie-handling rule (see the Appendix). The landmark-generating procedures return the indices of the selected landmarks, optionally together with the sets of indices of the points in the cover set centered at each landmark. In addition to the number of landmarks n and either the radius ε of the balls or the cardinality k of the neighborhoods, the user may also specify additive and multiplicative extension factors for n and for ε or k. These will produce additional landmarks (n) and larger cover sets (ε or k) with increased overlaps, in order to construct more overlapping covers.

2.2.1 Validation

We validated the maxmin and lastfirst procedures against manual calculations on several small example data sets, including that of Example 17. We also validated the C++ and R implementations against each other on several larger data sets, including as part of the benchmark tests reported in the next section. We invite readers to experiment with new cases and to request or contribute additional features.

2.2.2 Benchmark tests

We benchmarked the C++ and R implementations on three data sets: uniform samples from the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$ convoluted with Gaussian noise, samples with duplication from the integer lattice $[0, 23] \times [0, 11]$ using the probability mass function $p(a, b) \propto 2^{-ab}$, and patients recorded at each critical care unit in MIMIC-III using RT-transformed data and cosine similarity (Section 2.3). We conducted benchmarks using the bench package (Hester 2020) on the University of Florida high-performance cluster HiPerGator.

2.3 Empirical data

The experiments described in Section **??** make use of the two real-world data sets detailed here.

2.3.1 MIMIC-III

The open-access critical care database MIMIC-III ("Medical Information Mart for Intensive Care"), derived from the administrative and clinical records for 58,976 admissions of 46,520 patients over 12 years and maintained by the MIT Laboratory for Computational Physiology and collaborating groups, has been widely used for education and research (Goldberger et al. 2000; Johnson et al. 2016). For our analyses we included data for patients admitted to five care units: coronary care (CCU), cardiac surgery recovery (CSRU), medical intensive care (MICU), surgical intensive care (SICU), and trauma/surgical intensive care (TSICU).[2] For each patient admission, we extracted the set of ICD-9/10 codes from the patient's record and several categorical demographic variables: age group (18–29, decades 30–39 through 70–79, and 80+), recorded gender (M or F), stated ethnicity (41 values),[3] stated religion,[4] marital status[5], and type of medical insurance[6]. Following Zhong, Loukides, and Gwadera (2020), we transformed these *relational-transaction (RT)* data into a binary case-by-variable matrix $X \in \mathbb{B}^{n \times p}$ suitable for the cosine similarity measure, which was converted to a distance measure by subtraction from 1. Because cosine similarity is monotonically related to the angle metric, our topological results will be the same up to this rescaling, so for simplicity we use cosine similarity in our experiments.

2.3.2 Mexican Department of Health

The Mexican Department of Health (MXDH) has released official open-access data containing an assortment of patient-level clinical variables related to COVID-19 infection and outcomes. These data have been compiled into a database and made freely available on Kaggle[7], a collaborative data science platform. The data we obtained includes information regarding over 724,000 patients confirmed to be COVID-positive via diagnostic laboratory testing. Two main types of information are present for each patient: (1) dates, and (2) categorical or binary variables. The former are dates associated with key moments in the clinical course of infection such as symptom onset, admission to a healthcare institution, and death (if applicable). The categorical and binary fields encode clinical factors likely to be associated with COVID-19 infection, severity, or outcome. These variables include information such as sex, state of patient residence, and intubation status, as well as binary fields

[2] https://mimic.physionet.org/mimictables/transfers/

[3] White, Black/African American, Unknown/Not Specified, Hispanic or Latino, Other, Unable to Obtain, Asian, Patient Declined to Answer, Asian – Chinese, Hispanic Latino – Puerto Rican, Black/Cape Verdean, White – Russian, Multi Race Ethnicity, Black/Haitian, Hispanic/Latino – Dominican, White – Other European, Asian – Asian Indian, Portuguese, White – Brazilian, Asian – Vietnamese, Black/African, Middle Eastern, Hispanic/Latino – Guatemalan, Hispanic/Latino – Cuban, Asian – Filipino, White – Eastern European, American Indian/Alaska Native, Hispanic/Latino – Salvadoran, Asian – Cambodian, Native Hawaiian or Other Pacific Islander, Asian – Korean, Asian – Other, Hispanic/Latino – Mexican, Hispanic/Latino – Central American (Other), Hispanic/Latino – Colombian, Caribbean Island, South American, Asian – Japanese, Hispanic/Latino – Honduran, Asian – Thai, American Indian/Alaska Native Federally Recognized Tribe

[4] Catholic, unspecified/unobtainable/missing, Protestant Quaker, Jewish, other, Episcopalian, Greek Orthodox, Christian Scientist, Buddhist, Muslim, Jehovah's Witness, Unitarian-Universalist, 7th Day Adventist, Hindu, Romanian Eastern Orthodox, Baptist, Hebrew, Methodist, Lutheran

[5] married, single, widowed, divorced, unknown/missing, separated, life partner

[6] Medicare, private, Medicaid, povernment, self pay

[7] https://www.kaggle.com/lalish99/covid19-mx

encoding the presence or absence of a wide variety of comorbidities such as asthma, hypertension, cardiovascular disease. Though these variables are categorical rather than continuous/numeric, there are sufficiently many of them (≈ 50) to potentially distinguish between many patient phenotypes. Further, this data set is very complete in that every patient is required to contain a valid value for every field, which minimizes concerns around missing data.

2.4 Experiments

2.4.1 Covers and nerves

Cardinality reduction techniques can be used to model a large number of cases represented by a large number of variables as a smaller number of clusters with similarity or overlap relations among them. The deterministic maxmin and lastfirst procedures provide fuzzy clusters (cover sets) defined by proximity to the landmark cases and relations defined by their overlap. The clusters obtained by these procedures occupy a middle ground between the regular intervals or quantiles commonly used to cover samples from Euclidean space and the emergent clusters obtained heuristically by penalizing between-cluster similarity and rewarding within-cluster similarity. The maxmin procedure produces cover sets of (roughly) fixed radius, analogous to overlapping intervals of fixed length, while the lastfirst procedure produces cover sets of (roughly) fixed size, analogous to the quantiles of an adaptive cover. This makes them natural solutions to the task of covering an arbitrary finite metric space that may or may not contain important geometric or topological structure (Singh, Mémoli, and Carlsson 2007).

As a practical test of this potential, we loosely followed the approach of Dłotko (2019) to construct covers and their nerves for each care unit of MIMIC-III, using maxmin and lastfirst. We varied the number of landmarks (6, 12, 24, 36, 48, 60, 120) and the multiplicative extension of the cover sets' sizes (0, .1, .2). We evaluated the procedures in three ways:

- **Clustering quality:** Both procedures yield *fuzzy* clusters—that is to say, clusters that allow for some overlap. While clustering quality measures might be useful, most, including almost all that have been proposed for fuzzy clusterings, rely on coordinate-wise calculations, specifically data and cluster centroids (Bouguessa, Wang, and Sun 2006; Wang and Zhang 2007; Falasconi et al. 2010). To our knowledge, the sole exception to have appeared in a comprehensive comparison of such measures is the *modified partition coefficient* (Dave 1996), defined as

$$\mathrm{MPC} = 1 - \frac{k}{k-1}(1 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} {u_{ij}}^2)$$

  where $U = (u_{ij})$ is the $n \times k$ fuzzy partition matrix: u_{ij} encodes the extent of membership of point x_i in cluster c_j, and $\sum_{j=1}^{k} u_{ij} = 1$ for all i. When a point x_i is contained in m cover sets c_j, we equally distribute its membership so that $u_{ij} = \frac{1}{m}$ when $x_i \in c_j$ and $u_{ij} = 0$ otherwise. Thus, the MPC quantifies the extent of overlap between all pairs of clusters. Like the partition coefficient from which it is adapted, the MPC takes the value 1 on crisp partitions and is penalized by membership sharing, but it is standardized so that its range does not depend on k.
- **Discrimination of risk:** For purposes of clinical phenotyping, patient clusters are more useful that better discriminate between low- and high-risk subgroups. We calculate a cover-based risk estimate from individual outcomes y_i as follows: For each cover set $c_j \subset X$, let

$p_j = \frac{1}{|c_j|} \sum_{x_i \in c_j} y_i$ be the incidence of the outcome in that set. Then compute the weighted sum $q_i = \sum_{x_i \in c_j} u_{ij} p_j$ of these incidence rates for each case. We measure how well the cover discriminates risk as the area under the receiver operating characteristic curve (AUROC).

We hypothesized that lastfirst covers would exhibit less overlap than maxmin covers by virtue of their greater sensitivity to local density, and that they would outperform maxmin covers at risk prediction by reducing the sizes of cover sets in denser regions of the data (taking advantage of more homogeneous patient cohorts).

2.4.2 Interpolative nearest neighbors prediction

Landmark points may also be used to trade accuracy for memory in neighborhood-based prediction modeling. Consider the following approach: A modeling process involves predictor data $X \in \mathbb{R}^{n \times p}$ and response data $y \in \mathbb{R}^{n \times 1}$, partitioned into training and testing sets X_0, X_1 and y_0, y_1 according to a partition $I_0 \sqcup I_1 = \{1, \ldots, n\}$ of the index set. Given $x \in X_1$, a nearest neighbors model computes the prediction $p(x) = \frac{1}{k} \sum_{q(x,x_i) \leq k} y_i$ by averaging the responses for the k^{th} nearest neighbors of x in X_0. By selecting a landmark set $L \subset X_0$, a researcher can reduce the computational cost of the model as follows: For each $\ell \in L$, calculate $p(\ell)$ as above. Then, for each $x \in X_1$, calculate $p_L(x) = \sum_{\ell \in L} w(d(x,\ell)) p(\ell) / \sum_{\ell \in L} w(d(x,\ell))$, where $w : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a weighting function (for example, $w(d) = d^{-1}$). The nearest neighbor predictions for L thus serve as proxies for the responses associated with X_0.

We took this approach to the prediction of in-hospital mortality for patients with records in each critical care unit of MIMIC-III. We then implemented the following procedure:

1. Determine a nested 6×6–fold split for train–tune–test cross-validation. That is, partition $[n] = \bigsqcup_{i=1}^{6} I_i$ into roughly equal parts, and partition each $[n] \smallsetminus \overline{I_i} = \bigsqcup_{j=1}^{6} J_{ij}$ into roughly equal parts.
2. Iterate the following steps over each i, j:
   a) Generate a sequence L of landmarks from the points $X_{([n] \smallsetminus \overline{I_i}) \smallsetminus \overline{J_{ij}}}$.
   b) Identify the 180 nearest neighbors $N_{180}^{+}(\ell)$ of each landmark ℓ. This was a fixed parameter, chosen for being slightly larger than the optimal neighborhood size in a previous study of individualized models (Lee, Maslove, and Dubin 2015).
   c) Find the value of $k \in [180]$ and the weighting function w (among those available) for which the predictions $p_L : X_{J_{ij}} \to [0,1]$ maximize the AUROC.
   d) Use the AUROC to evaluate the performance of the predictions $p_L : X_{I_i} \to [0,1]$ using these k and w.

We replicated the experiment for each combination of procedure (random, maxmin, lastfirst), number of landmarks ($|L| = 36, 60, 180, 360$), and each of several weighting functions (integer rank, triangle, inverse distance, Gaussian). We hypothesized that, as measured by overall accuracy of the resulting predictive model, the maxmin and lastfirst procedures would outperform random selection, and that lastfirst would outperform maxmin, for similar reasons to those in the previous section.

3 Results

3.1 Definition

We introduce the following definitions in Section 2:

Definition 14 (sampling procedures in terms of cover sets)**.** Given (X, d) and $Y \subset X$, write

$$\overline{B_\varepsilon(Y)} = \bigcup_{y \in Y} \overline{B_\varepsilon(y)}$$
$$N_k(Y) = \bigcup_{y \in Y} N_k(y)$$

Take

$$\mathrm{E}(Y, X) = \min\{\varepsilon : \overline{B_\varepsilon(Y)} = X\}$$
$$K(Y, X) = \min\{k : N_k(Y) = X\}$$

Then define the *maxmin* and *lastfirst sets*

$$\mathrm{maxmin}(Y) = \mathrm{maxmin}(Y; X) = X \setminus B_{\mathrm{E}(Y,X)}(Y)$$
$$\mathrm{lf}(Y) = \mathrm{lf}(Y; X) = X \setminus N_{K(Y,X)-1}(Y)$$

consisting of *maxmin* and *lastfirst points*, respectively.

As suggested by their definitions, these procedures arise from the construction of conditionally minimal covers of X, where minimality is defined in terms of the common radius (maxmin) or cardinality (lastfirst) of the sets. The centers of the cover sets comprise the landmark points, beginning from an arbitrarily selected first landmark. In practice, we suggest a Chebyshev center $\mathrm{argmin}_{x \in X} \min\{\varepsilon : \overline{B_\varepsilon(x)} = X\}$ as a starting landmark for the maxmin procedure, in that it provides an (unconditionally) minimal one-set cover.[8]

3.2 Implementation

The definition of our lastfirst procedure is analogous to that of maxmin, substituting ranks in the role of distances. In this way, lastfirst is an alternative to maxmin that is adaptive to the local density of the data, similar to the use of fixed quantiles in place of fixed-length intervals. The maxmin and lastfirst procedures implicitly construct a minimal cover whose sets are centered at the selected landmarks, and the fixed-radius balls of maxmin correspond to the fixed-cardinality neighborhoods of lastfirst. The rank-based procedures are more combinatorially complex and computationally expensive, primarily because relative ranks are asymmetric, which doubles (in the best case) or squares (in the worst case) the number of distances that must be calculated. Nevertheless, the procedure can be performed in a reasonable time for many real-world uses.

3.2.1 Illustration

Consider the "necklace" data set adapted from Yoon and Ghrist (2020). The points are sampled in the plane from a large, low-density circular region (the "string") and from several smaller, higher-density circular regions (the "beads") evenly spaced along the string. Only a method of detecting

[8] An analogously defined point $\mathrm{argmin}_{x \in X} \min\{k : N_k(x) = X\}$ might be used for the lastfirst procedure; when X is not in locally general position, it locates a landmark that is equidistant to as many farthest neighbors as possible, i.e. the center of a maximally populated circumcenter of X.

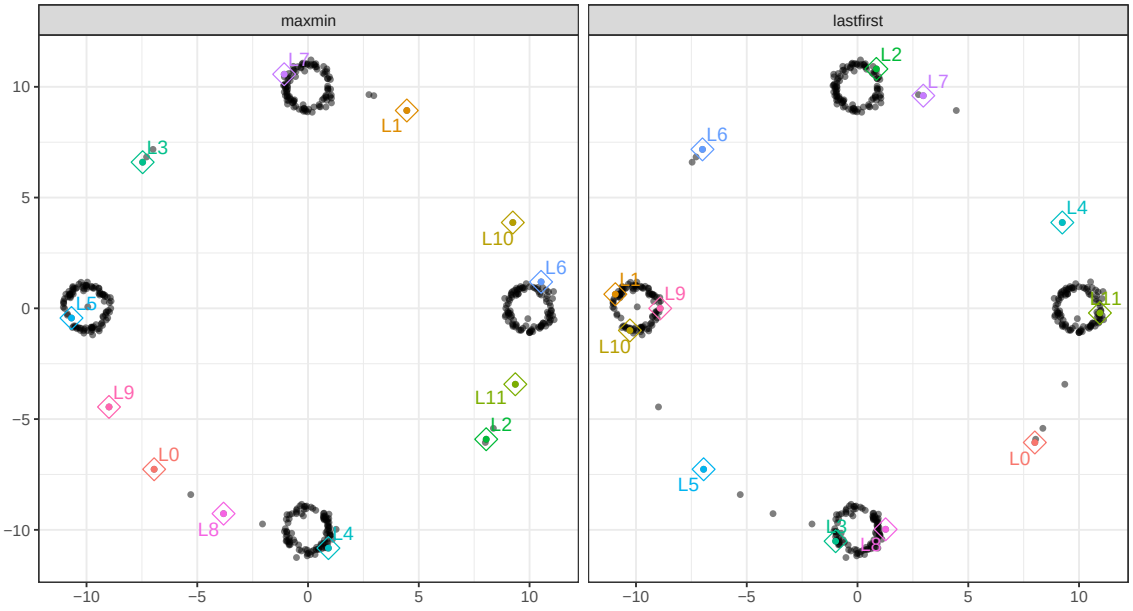


Figure 3. Landmark samples of size 12 from a necklace data set using two selection procedures.

topology that is adaptive to the different scales of the string and of the beads will recover both the large feature and the smaller features of dimension 1. We use small landmark samples to illustrate the differences between the maxmin and lastfirst selection procedures. Publicly available software tools are not yet available to compute persistent homology for arbitrary sequences of simplicial maps, but we suggest that the relative strengths of maxmin and lastfirst would be revealed by a comparison of the persistent features obtained by sequences of covers obtained by both procedures on this data set.

3.2.2 Benchmark tests

Benchmark results are reported in Figure 4. The R implementation of maxmin used orders of magnitude more memory and took slightly longer than the C++ implementation. They appeared to scale slightly better in terms of time than the lastfirst implementations. The additional calculations required for the lastfirst procedure increase runtimes by a median factor of 2.5 in our R implementations. The C++ implementation of lastfirst is based on combinatorial definitions and not optimized for speed, and as a result took much longer and failed to complete in many of our tests.

3.3 Experiments

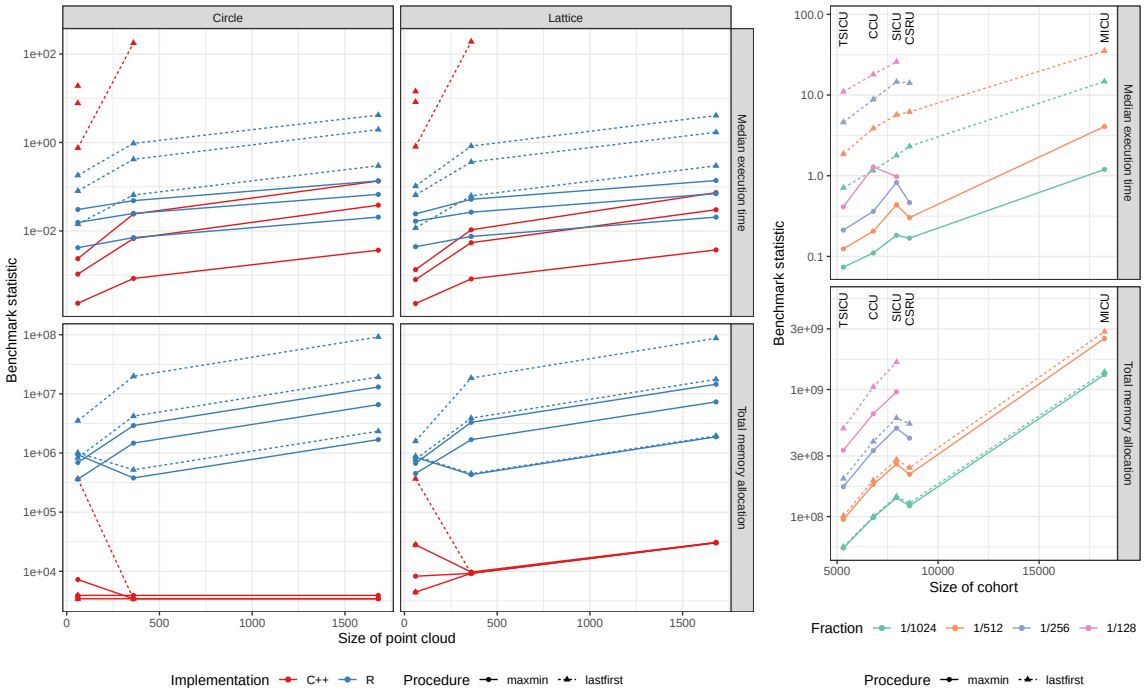


Figure 4. Benchmark results for computing landmarks on two families of artificial data (circle and lattice) and one collection of empirical data (RT-similarity space of critical care units in MIMIC-III). Some points are missing because benchmark tests did not complete within 1 hour.

3.3.1 Covers and nerves

Figure 5 presents, for the analysis of the five MIMIC-III care units, the sizes of the nerves of the covers and the two evaluation statistics as functions of the number of landmarks. The numbers of 1- and of 2-simplices grew at most roughly quadratically and roughly cubically, respectively. This suggests that the densities of the simplicial complex models were at most roughly constant, regardless of the number of landmarks. Landmark covers grew fuzzier and generated more accurate predictions until the number of landmarks reached around 60, beyond which point most covers grew crisper while performance increased more slowly (and in one case decreased). This pattern held for covers with any fixed multiplicative extension. Naturally, these extensions produced fuzzier clusters, but they also reduced the overall accuracy of the predictive model. Independently of these patterns, models fitted to smaller care units tended to outperform those fitted to larger care units. Contrary to expectations, unextended maxmin covers were usually crisper than their lastfirst counterparts and yielded more accurate predictions, though extensions reduced the crispness of maxmin covers more dramatically than of lastfirst covers. The same patterns were observed in the risk discrimination of maxmin versus lastfirst covers, with maxmin covers yielding the most accurate predictions when unextended but lastfirst covers retaining more accuracy after extension.

Figure 6 presents the same evaluations for covers of the MXDH data. In contrast to the MIMIC experiments, lastfirst-based nerves of the MXDH data grew sub-polynomially and were significantly sparser than maxmin-based nerves. Lastfirst covers tended to be crisper, especially as the number of landmarks and the extension factors increased. This indicates that the nearest neighborhoods formed a more parsimonious cover of the data than the centered balls. The predictive accuracies of the cover set–based models converged with increasing numbers of landmarks, though for smaller numbers different selection procedures performed best for different outcomes.

3.3.2 Interpolative nearest neighbors prediction

Boxplots of the AUROCs for each cross-validation step are presented in Figure 7. While both landmark procedures yielded stronger results than random selection, lastfirst performed on average slightly worse than maxmin on each data set. Importantly, both landmark procedures also yielded more accurate predictions than a basic unweighted nearest-neighbors model, lending support to the modeling approach itself. Interestingly, only on the largest data set (the MICU) did increasing the number of landmarks from 36 to 360 appreciably improve predictive accuracy (using all three selection procedures).

Over the course of the COVID-19 pandemic, hospitals and other facilities experienced periods of overburden and resource depletion, and best practices were continually learned and disseminated. As a result, outcomes in the MXDH data reflect institutional- as well as population-level factors. We took advantage of the rapid learning process in particular by adapting the nested CV approach above to a temporal CV approach (Major, Jethani, and Aphinyanaphongs 2020): We partitioned the data by week, beginning with Week 11 (March 11–17) and ending with Week 19 (May 6–9, the last dates for which data were available). For each week i, $11 < i \leq 19$, we trained prediction models on the data from Week $i - 1$. We then randomly partitioned Week i into six roughly equal parts and optimized and evaluated the models as above. (For this analysis, we only considered Gaussian weighting.)

Line plots of model performance are presented in Figure 8, with one curve (across numbers of

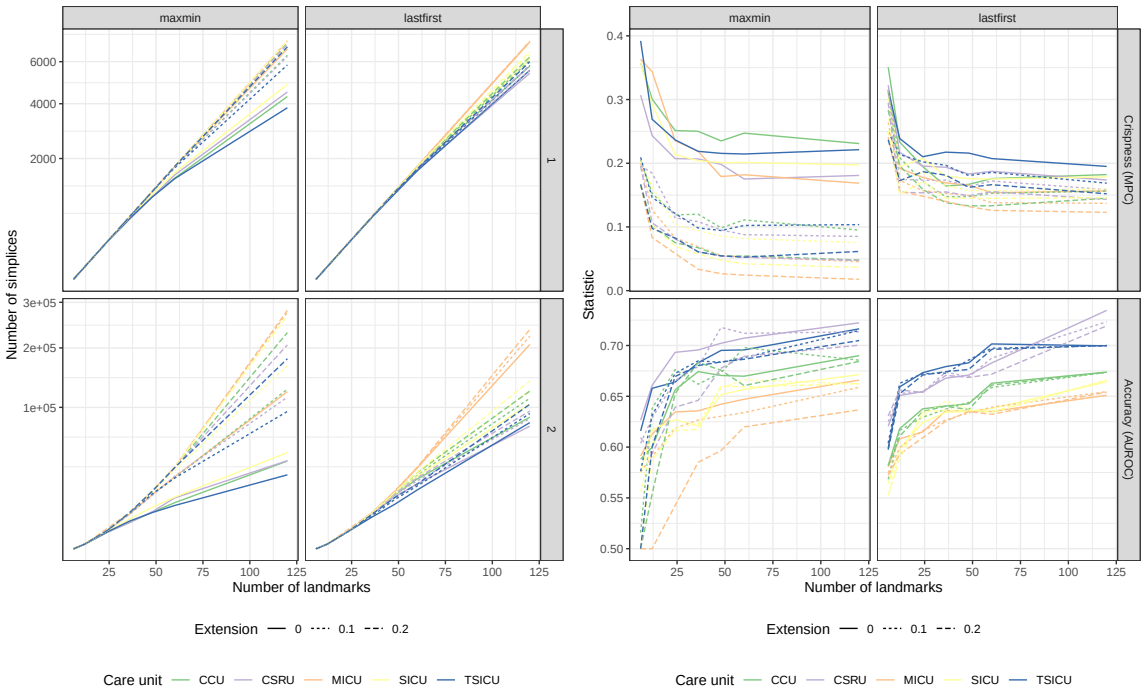


Figure 5. Summary and evaluation statistics versus number of 0-simplices (landmarks) for the covers generated using the maxmin and lastfirst procedures, with three multiplicative extensions in their size. Left: the sizes of their nerves, as numbers of 1- and 2-simplices, using a square root–transformed vertical scale. Right: the modified partition coefficient (MPC) and the c-statistic of the risk prediction model based on the cover sets (AUROC).

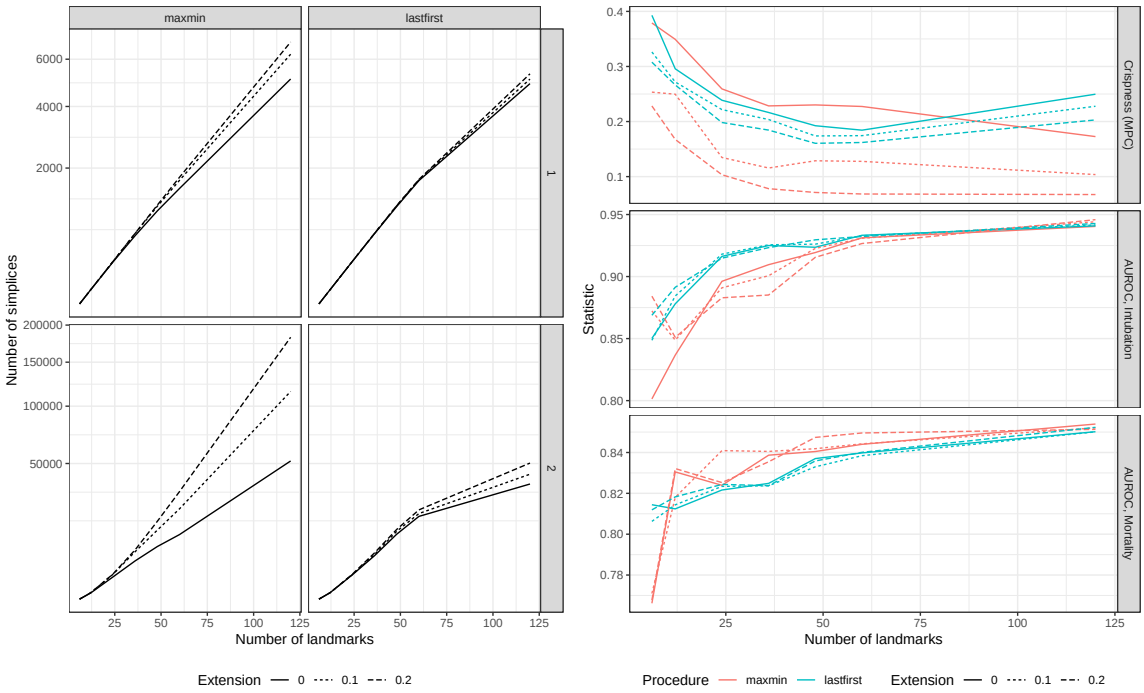


Figure 6. Summary and evaluation statistics versus number of 0-simplices (landmarks) for the covers generated using the maxmin and lastfirst procedures, with three multiplicative extensions in their size. Left: the sizes of their nerves, as numbers of 1- and 2-simplices, using a square root–transformed vertical scale. Right: the modified partition coefficient (MPC) and the c-statistics of the risk prediction models based on the cover sets (AUROC).

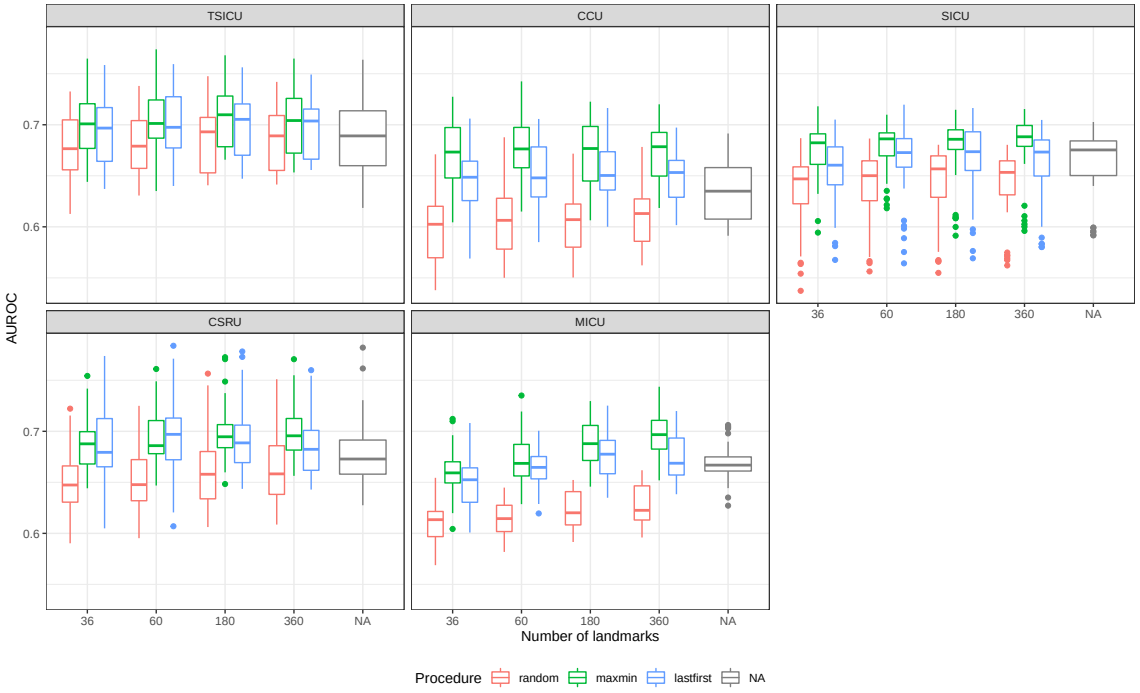


Figure 7. AUROCs of the interpolative predictive models of mortality in five MIMIC-III care units based on covers constructed using random, maxmin, and lastfirst procedures to generate landmarks. Each boxplot summarizes AUROCs from $6 \times 6 = 36$ models, one for each combination of outer and inner fold. AUROCs of simple nearest-neighbor predictive models are included for comparison.

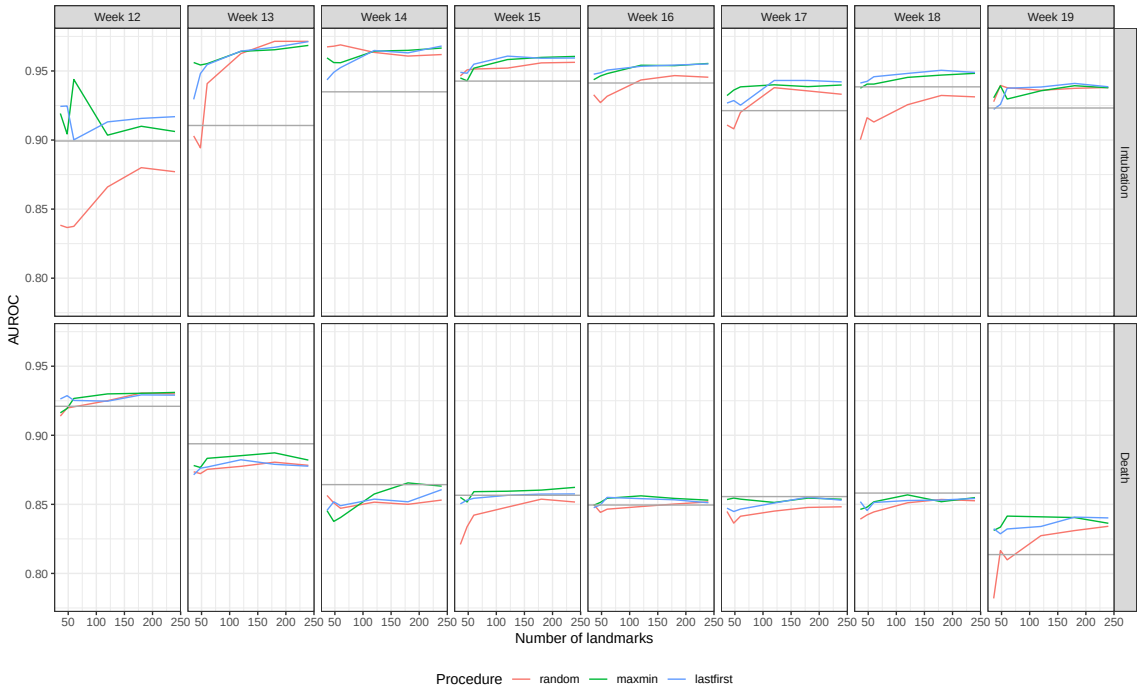


Figure 8. AUROCs of the sliding-window interpolative predictive models of intubation and mortality in the MXDH data based on covers constructed using random, maxmin, and lastfirst procedures to generate landmarks.

landmarks) per selection procedure, outcome, and week. Again both landmark selection procedures yielded stronger results than random selection. Interestingly, this was more pronounced in later weeks, as the pandemic progressed, even as overall predictive accuracy declined.[9] Overall, performance improved slightly as the number of landmarks increased from 50 to 150 but either plateaued or declined from 150 to 250.

4 Discussion

In this work, we introduce the lastfirst procedure as a complement to the widely-used maxmin sampler for the selection of landmark points from a data set. Lastfirst applies the same logic to the growing of nearest neighborhoods rather than balls around the sampled points. This results in a heuristic sample of landmarks and an associated cover with analogous properties to those of maxmin. Common data limitations require careful consideration of edge cases, which also apply to maxmin but have not been worked out before, and our choices of solution can be used with either

[9]Do we know what major events in Brazil might have influenced these outcomes? We should at least also plot the number of cases, stratified by outcomes, per day over this period.

procedure. Further, our experiments described in the previous section demonstrate a number of advantages that the lastfirst sampler confers over maxmin.

First of all, the lastfirst procedure is more general than maxmin, in that it can be applied to any set of cases for which directed pairwise distances are available or can be computed. This relaxes the symmetry assumption of maxmin and would allow lastfirst to be used when the relevance of one case to another is not a two-way street. One use case would be to sample dispersed nodes from a transportation network involving one-way streets and other asymmetries such as traffic patterns, which would be modeled as a directed graph with shortest path distances between nodes. However, it imposes costs to computation, as our algorithm requires us to either impute or compute and store all distance ranks from the new landmark at each step. Based on our experiments comparing implementations of the same type, this increases runtime and storage only by a constant factor.

Additionally, lastfirst confers some interpretability benefits in the context of applications. When dealing with certain data types, the meaning of numerical values of distance may not be clear or intuitive. For instance, on patient medical data obtained from electronic health records, custom distance metrics are often used to measure the *relative* clinical relevance between patients, and it may not be obvious what an *absolute* distance of, say, .35 between patient A and patient B might mean. In these types of settings, it can therefore be difficult to select parameters such as a ball radius to use in maxmin or other algorithms, and such choices are often made arbitrarily. This is especially true when minimal literature references are available, as is the case for many non-mathematical applications of TDA. However, in most applications, the notion of neighborhoods remains intuitive: Even when a ball of radius .35 around patient A is hard to conceptualize, a neighborhood of 200 similar patients still has a clear meaning. This increased interpretability makes parameter selection easier and also makes the algorithm more accessible to researchers outside mathematics.

Finally, lastfirst outperformed maxmin in settings tailored to the strengths of the algorithm (and achieved comparable performance in all other settings). We ran several experiments that used landmarks to obtain well-separated clusters of patients with common risk profiles and to more efficiently generate nearest neighbor predictions. Because we designed lastfirst to produce cover sets of equal size despite variation in the density or multiplicity of the data, we expected it to outperform maxmin with respect to the crispness of clusters and to the accuracy of predictions. In particular, we expected that the optimal neighborhood size for outcome prediction would be roughly equal across our data; as a result, by assigning each landmark case an equally-sized cohort of similar cases, we expected predictions based on these cohorts to outperform those based on cohorts using a fixed similarity threshold.

Contrary to expectations, maxmin produced crisper clusterings on average, and in the case of MIMIC-III more accurate predictions. However, when the sets of these covers had their radii or cardinalities extended by a fixed proportion, those of lastfirst better preserved these qualities. Additionally, in the case of MXDH, neither landmark selection procedure produced consistently more accurate predictions. Specifically, when predicting mortality in MIMIC-III, even with the theory-agnostic RT-similarity measure, a patient's prognosis is better-guided by a cohort cutoff at a fixed minimum similarity than by one of a fixed size. This suggests that the use of personalized cohorts to improve predictive modeling, as employed by Lee, Maslove, and Dubin (2015), may be strengthened by optimizing a fixed similarity threshold rather than a fixed cohort size.

It is worth noting that Park, Kim, and Chun (2006), to our knowledge the only other investigators who have compared predictive models based on cohorts bounded by a radius versus a cardinality,

reached a similar conclusion. Notably, however, lastfirst performed better, relative to maxmin, on tasks involving the MXDH data, which had more categorical variables, more indistinguishable records, and fewer variables overall—that is, data for which the limitations of maxmin we described at the outset are more acute. Taken together, these results indicate that lastfirst covers are competitive with maxmin covers for basic analysis tasks and should be considered for more advanced tasks, for example witness complexes or mapper-like constructions, where consistency in the sizes of cover sets or density-based representativeness is advantageous.

One way to encapsulate these results is in terms of a balance between relevance and power, with fixed-radius balls (respectively, fixed-cardinality neighborhoods) providing training cohorts of roughly equal relevance (statistical power) to all test cases. With sufficiently rich data, relevance can be more precisely measured and becomes more important to cohort definition, as with MIMIC. When variables are fewer, as with MXDH, relevance is more difficult to measure, so that larger samples can improve performance even at the expense of such a measure.

5 Appendix

5.1 Relative ranks

This section develops a more general lastfirst procedure and makes rigorous some ideas in the main text.

Relative ranks are a much more general notion of metric that encompasess ranks in nearest neighborhoods.

Definition 15 (relative rank)**.** A *relative rank* on X is a binary relation $q : X \times X \to \mathbb{R}_{\geq 0}$ subject only to the following inequality:

$$\forall x, y \in X : d(x, x) \leq d(x, y) \tag{1}$$

Relative ranks can be used as the basis for a much more general maxmin procedure, taking care in particular to account for asymmetry.

Given a relative rank q on X, write

$$\begin{aligned} q(Y, Z) &= \min_{y \in Y, z \in Z} q(y, z) & Q(Y, Z) &= \max_{y \in Y, z \in Z} q(y, z) \\ q(x, Y) &= q(\{x\}, Y) & Q(x, Y) &= Q(\{x\}, Y) \\ q(Y, x) &= q(Y, \{x\}) & Q(x, Y) &= Q(Y, \{x\}) \end{aligned}$$

A pseudometric d_X induces a relative rank that takes values in $\mathbb{N}$ given by the ordinal of one point's distance from another:

Definition 16 (out-rank and in-rank)**.** For $x, y \in X$ with pseudometric d, define the *out-rank* $q_{X,d} : X \times X \longrightarrow \mathbb{N}$ as follows:

$$q_{X,d}(x, y) = |\{z \in X : d(x, z) < d(x, y)\}| \tag{2}$$

and the *in-rank* $q^{\top}_{X,d}(x, y) = q_{X,d}(y, x)$.

As with d, we allow ourselves to write $q = q_{X,d}$ when clear from context. Note that $q_{X,d}$ is a relative rank with $q(x,x) = 0$, $q(x,y) < N$, and $\forall x, y \in X : q(x,x) \leq q(x,y)$.

Example 17. Recall $X = \{a, b, c, d\}$ from Example 22. Lack of symmetry of q is shown by points b and c :

$$\begin{aligned} q(a,c) &= |\{x \in X : |x-a| < |c-a| = 2\}| \\ &= |\{a,b\}| \\ &= 2 \end{aligned} \qquad \begin{aligned} q(c,a) &= |\{x \in X : |x-c| < |a-c| = 2\}| \\ &= |\{b,c,d\}| \\ &= 3 \end{aligned}$$

Observe in particular that $\max_{x \in X} q(a,x) = 2 < |X| - 1$; $q(x, \cdot\,)$ will not max out at $N-1$ when the most distant points from x have multiplicity.

However, $q(x,x) = 0$ whether x has multiplicity or not:

$$\begin{aligned} q(a,a) &= |\{x \in X : |x-a| < |a-a| = 0\}| \\ &= |\varnothing| \\ &= 0 \end{aligned} \qquad \begin{aligned} q(c,c) &= |\{x \in X : |x-c| < |c-c| = 0\}| \\ &= |\varnothing| \\ &= 0 \end{aligned}$$

We also term the unary rankings $q(x, \cdot\,)$ and $q(\cdot\,, x)$ the *out- (from x)* and *in- (to x) rankings* of X, respectively. These can be used to define *out-* and *in-neighborhoods* of x.[10]

A relative rank q can be used to define k-neighborhoods in greater generality:

Definition 18 (k-neighborhoods using relative rank)**.** For $x \in X$, define the *k-out-neighborhoods* N_k^+ and *k-in-neighborhoods* N_k^- of x as the sets

$$N_k^+(x) = \{y \in X : q(x,y) \leq k\}$$
$$N_k^-(x) = \{y \in X : q(y,x) \leq k\}$$

Given a subset $Y \subseteq X$, we also define

$$N_k^+(x,Y) = \{y \in Y : q_X(x,y) \leq k\}$$
$$N_k^-(x,Y) = \{y \in Y : q_X(y,x) \leq k\}$$

Relative ranks are not as straightforward to compare among subsets of points. For example, for $y \neq x$, $q(x,y)$ takes integer values between $|[\{x\}]|$ and $N-1$. However, they do provide us with a definition of the lastfirst procedure that straightforwardly adapts Definition 3.

Corollary 19 (lastfirst using relative rank)**.** *Given $Y \subset X$ and a pseudometric d on X with relative rank q,*

$$\mathrm{lf}(Y;X) = \mathrm{maxmin}(Y;X,q_d)$$
$$\mathrm{lf}(X,d) = \mathrm{maxmin}(X,q_d)$$

[10]The terminology and notation are adapted from graph theory. These definitions are the same as those for a complete directed graph on X with directed arcs $x \to y$ weighted by $q(x,y)$.

Proof. This follows directly from the observations

$$\operatorname{maxmin}(Y; X, q) = \{x \in X \setminus \overline{Y} : N_\bullet^-(x, Y) = \max_{y \in X \setminus \overline{Y}} N_\bullet^-(y, Y)\}$$
$$\operatorname{maxmin}(X, q) = \{x \in X : N_\bullet^-(x, X \setminus \overline{\{x\}}) = \max_{y \in X} N_\bullet^-(y, X \setminus \overline{\{y\}})\}$$

obtained by adapting Proposition 4 to the relative rank. □

5.1.1 Selection procedures

The choice of $\ell_i \in \operatorname{maxmin}(L)$ is trivial when X is in locally general position, but this study is specifically interested in cases with a high frequency of violations of this property, and indeed with such violations of Hausdorffness that large numbers of points in X may be co-located. When X is not in locally general position, the choice of selection from among the maxmin set is consequential. For convenience, let $\operatorname{maxmin}(L; X)$ take the value $\varnothing$ if $\overline{L} = X$ and the value X if $L = \varnothing$. Then write $\mathrm{G}(\operatorname{maxmin}, X) \subset \mathcal{Q}(X) \times \mathcal{P}(X)$ for the graph of the unary function $\operatorname{maxmin}(\,\cdot\,; X) : \mathcal{Q}(X) \to \mathcal{P}(X)$, so that $(L, \Gamma) \in \mathrm{G}(\operatorname{maxmin}, X)$ if and only if $\operatorname{maxmin}(L; X) = \Gamma$. Then a selection procedure is a function $\sigma : \mathrm{G}(\operatorname{maxmin}, X) \to X$ subject to $\sigma((L, \Gamma)) \in \Gamma = \operatorname{maxmin}(L; X)$. Importantly, σ may depend not only on the maxmin set Γ but also on the ordered sequence L.

We assume the following choice of σ: Take $d_L = \max_{y \in X \setminus \overline{L}} d(y, L)$. For each $y \in \operatorname{maxmin}(L; X)$, choose $\ell_y \in L$ for which $d(y, \ell_y) = d_L$. Then take $\operatorname{maxmin}^{(1)}(L; X) = \{y \in \operatorname{maxmin}(L; X) : d(y, L \setminus \{\ell_y\}) = \max_{z \in \operatorname{maxmin}(L;X)} d(z, L \setminus \{\ell_z\})\}$. While $|\operatorname{maxmin}^{(j)}(L; X)| > 1$, continue in this way until either a singleton is reached or $j = |L| = i$. If the latter, then the choice σ is arbitrary among the remaining y.[11]

Suppose ℓ_i is selected so that the minimum k required for $\ell_i \in \bigcup_{j=0}^{i-1} N_k(\ell_j)$ is maximized. This is equivalent to maximizing the minimum out-rank $q(\ell_j, \ell_i)$ of ℓ_i from any ℓ_j. Switching perspective from out- to in- and reversing the roles of L and ℓ_i, we want $N_k^-(\ell_i, L) = 0$ for the latest (largest) k possible, say k_0^-. When indistinguishable points abound, this may still not uniquely determine ℓ_i, so we may extend the principle: Among those $\ell \in X \setminus \overline{L}$ for which $N_{k_0^-}^-(\ell_i, L) = 0$, choose ℓ_i for which $N_k^-(\ell_i, L) \leq 1$ for the latest k possible, say $k_1^- \geq k_0^-$. (It is possible that $k_1^- = k_0^-$, in which case no $N_k^-(\ell_i, L) = 1$.) Continue this process until only one candidate ℓ remains (up to multiplicity), or until $N_k^-(\ell, L) = |L|$, in which case all remaining candidates may be considered equivalent.

5.1.2 Algorithms

Algorithm 4 calculates a lastfirst set from a seed point, subject to parameters analogous to n and ε in Algorithm 1. The algorithm is tailored to the vectorized arithmetic of R, and Lemma 20 provides a shortcut between Q^- and the more compact way that the relative rank data are stored.

Lemma 20. *For $L = \{\ell_0, \ldots, \ell_n\} \subset X$, write $S(x, L) = (q(\ell_{\pi^{-1}(1)}, x) \leq \cdots \leq q(\ell_{\pi^{-1}(n)}, x))$, where π is any suitable permutation on $[n]$. Then $Q^-(x, L) < Q^-(y, L) \Leftrightarrow S(x, L) < S(y, L)$.*

Proof. Write $Q(x) = Q^-(x, L)$ and $Q(y) = Q^-(y, L)$ and suppose that $Q(x) < Q(y)$. This means that $Q_i(x) > Q_i(y)$ for some index $i \in [N]$ while $Q_j(x) = Q_j(y)$ for all $j < i$. There are then, for

[11] Should this be written up as an algorithm?

each $j < i$, equal numbers of $\ell \in L$ for which $q(\ell, x) = j$ and for which $q(\ell, y) = j$; while there are more $\ell \in L$ for which $q(\ell, x) = i$ than for which $q(\ell, y) = i$. When the sets $\{q(\ell, x)\}_{\ell in L}$ and $\{q(\ell, y)\}_{\ell in L}$ are arranged in order to get $S(x, L)$ and $S(y, L)$, therefore, the leftmost position at which they differ is $Q_1(y) + \cdots + Q_i(Y) + 1$, at which $q(\ell, x) = i$ while $q(\ell, y) \geq i + 1$. Thus $S(x, L) <_{\text{lex}} S(y, L)$.

The reverse implication is similarly straightforward. □

Algorithm 4 Calculate the lastfirst landmark sequence from a seed point.

Require: finite pseudometric space (X, d)
Require: seed point $\ell_0 \in X$
Require: number of landmarks $n \in \mathbb{N}$ or cover set cardinality $k \in \mathbb{N}$
Require: selection procedure σ
if n is not given, set $n \leftarrow 0$
if k is not given, set $k \leftarrow \infty$
$L \leftarrow \varnothing$ initial landmark set
$F \leftarrow \{\ell_0\}$ initial lastfirst set
$R \in \mathbb{N}^{N \times 0}$, a 0-dimensional $\mathbb{N}$-valued matrix
for i from 0 to $|[X]| - 1$ **do**
  $\ell_i \leftarrow \sigma(F)$
  $L \leftarrow L \cup \{\ell_i\}$
  $D_i \leftarrow (d_{i1}, \ldots, d_{iN}) \in {\mathbb{R}_{\geq 0}}^N$, where $d_{ir} = d(\ell_i, x_r)$
  $Q_i \leftarrow \texttt{rank}(D_i) \in {\mathbb{N}_{\geq 0}}^N$ (so that $Q = (q(\ell_i, x_1), \ldots, q(\ell_i, x_N)))$
  $R \leftarrow [R, Q_i] \in \mathbb{N}^{N \times (i+1)}$
  $k_{\min} \leftarrow \max_{r=1}^{N} \min_{j=1,i+1} R_{r,j}$ (minimum k for which neighborhoods centered at L cover X)
  if $D(L, X \setminus L) = 0$ **then**
    break
  end if
  if $i \geq n$ and $k_{\min} \leq k$ **then**
    break
  end if
  $R \leftarrow [\texttt{sort}(R_{1,\bullet})^\top \cdots \texttt{sort}(R_{N,\bullet})^\top]^\top \in \mathbb{N}^{N \times (i+1)}$
  $F \leftarrow X \setminus L$
  for j from 1 to i **do**
    $F \leftarrow \{x_r \in F : R_{rj} = \max_{r'} R_{r'j}\}$
    if $|F| = 1$ **then**
      break
    end if
  end for
end for
return L
return lastfirst landmark set L with at least n cover sets of cardinality at most k

Proposition 21. *Algorithm 4 returns a lastfirst landmark set. If $n \leq |[X]|$ is given as input and k is not, then $|L| = n$. If n and k are both given, then $|L| \geq n$. Otherwise, L is minimal in the sense that no proper prefix of L gives a cover of X by k-nearest neighborhoods.*

Proof. Let (X, d) be a finite metric space and $\ell_0 \in X$ be a seed point, as required by Algorithm 4. Note that, for the algorithm to terminate its loop and subsequently return L, either there must be no points in $X \setminus \overline{L}$ distinguishable from L (line 14), or both of two exit conditions must hold (line 17): (1) $k_{\min} \leq k$ and (2) $|L| \geq n$.

Because the seed point is arbitrary, for the main result it is enough to show that, at each step i, $F = \mathrm{lf}(\{\ell_0, \ldots, \ell_{i-1}\})$. When F is calculated on line 22, the rows R_r of R contain the in-ranks $q(\ell_i, x_r)$ of x_r in increasing order. Because $D(L, X \setminus \overline{L}) > 0$ (line 14), F is nonempty. By Lemma 20, then, $Q^-(x_r, L)$ is maximized (in revlex) when R_r is maximized in lex, and this is exactly what the loop that begins on line 21 does.

Suppose first that $n \leq |[X]|$ is given. The loop will only break on line 14 if $D(L, X \setminus \overline{L}) = 0$, which is only possible if $|L| = |[X]| \geq n$. The loop will only break on line 17 if both $|L| = i \geq n$ and $k_{\min} \leq k$. Since these are the only two possible breaks, $|L| \geq n$ is a necessary condition. [12]

If k is not given, then k is set to ∞ on line 2, which means that (1) holds throughout the loop. Then the algorithm terminates as least as soon as (2) is satisfied, when $|L| = n$, and as discussed above it cannot terminate any sooner.

Now suppose n is not given. Then k must be given, and n is set to 0 (line 1). This means that (2) always holds, so the algorithm terminates as soon as (1) is satisfied, i.e. when $k \geq k_{\min}$ with $k_{\min}$ as defined on line 12. We claim that

$$k_{\min} = \max_{x \in X} \min_{\ell \in L} q(\ell, x)$$

which means that every point in x is within a k-neighborhood of some existing landmark $\ell \in L$, i.e. that the k-neighborhoods at L constitute a cover of X. For this to have not been true for L at previous iterations, there must have been $x \in X$ with $q(\ell, x) > k$ for all $\ell \in L$, meaning that the k-neighborhoods at L did not cover X.

To prove the claim, note that the maximum is taken over all rows of R, which at no point in the algorithm are permuted—that is, the entries of row $R_{r,\bullet}$ at each iteration consist of $q(\ell_0, x_r), \ldots, q(\ell_i, x_r)$ in some order. Therefore, $\min_{j=1}^{i} R_{r,j} = \min_{\ell \in L} q(\ell, x_r)$. Because columns 1 through $i-1$ of R were sorted in the previous iteration (line 19), this minimum only needs to be taken over $j = 1, i$, which gives the formula on line 12. □

5.1.3 Tie handling

We might have defined two relative ranks $\check{q}, \hat{q} : X \times X \longrightarrow \mathbb{N}$ ("q-check" and "q-hat") as follows:

$$\check{q}(x, y) = |\{z \in X : d(x, z) < d(x, y)\}|$$
$$\hat{q}(x, y) = |\{z \in X : d(x, z) \leq d(x, y)\}| - 1$$

In this notation, $\check{q} = q$, while $\hat{q}(x, y)$ is the cardinality of the smallest ball centered at x that contains y. Then $\check{N}_1^{\pm}(x) \subseteq \{x\} \subseteq \hat{N}_1^{\pm}(x)$, and $\hat{q}(x, x) > 0$ when x has multiplicity. The two relative ranks derive from two tie-handling schemes for calculating rankings of lists with duplicates. For example, if $a < b =$

[12] Note that $|L| > n$ if and only if $k_{\min} \leq k$ is not satisfied when $|L| = n$, meaning that X would not be covered by k-neighborhoods around n landmark points, so that more landmarks must be chosen to guarantee the algorithm produces a valid k-neighborhood cover.

$c < d$ are the distances from x to y_1, y_2, y_3, y_4, respectively, then $(\check{q}(x, y_1), \check{q}(x, y_2), \check{q}(x, y_3), \check{q}(x, y_4)) = (0, 1, 1, 3)$ and $(\hat{q}(x, y_1), \hat{q}(x, y_2), \hat{q}(x, y_3), \hat{q}(x, y_4)) = (0, 2, 2, 3)$. Indeed, any tie-handling rule could be used, and the choice becomes more consequential with greater multiplicity in the data.

Conceptually, the lastfirst procedure based on $\hat{q}$ would produce landmark sets that yield neighborhood covers with smaller, rather than larger, neighborhoods in regions of high multiplicity. While we do not use these ideas in this study, they may be suitable in some settings or for some purposes, for example when high multiplicity indicates a failure to discriminate between important categories. It is also possible that $\check{q}$- and $\hat{q}$-based covers could be used to produce interveaving sequences of nerves useful for stability analysis.

Example 22. Consider the same X as in Example 6.

The relative rank $\hat{q}$ is also asymmetric:

$$\begin{aligned}\hat{q}(b,c) &= |\{x \in X : |x-b| \leq |c-b| = 2\}| \\ &= |\{a,b,c,d\}| \\ &= 4\end{aligned} \qquad \begin{aligned}\hat{q}(c,b) &= |\{x \in X : |x-c| \leq |b-c| = 2\}| \\ &= |\{b,c,d\}| \\ &= 3\end{aligned}$$

Observe that $\hat{q}(x, x) = 1$ only for distinguishable points $x \in X$:

$$\begin{aligned}\hat{q}(a,a) &= |\{x \in X : |x-a| \leq |a-a| = 0\}| \\ &= |\{a\}| \\ &= 1\end{aligned} \qquad \begin{aligned}\hat{q}(c,c) &= |\{x \in X : |x-c| \leq |c-c| = 0\}| \\ &= |\{c,d\}| \\ &= 2\end{aligned}$$

Finally, observe that $\hat{q}(x, \cdot)$ always maxes out at $|X|$: $\hat{q}(a, c) = \hat{q}(b, c) = \hat{q}(c, a) = \hat{q}(d, a) = |X|$.

Continuing on as in Example 6, we can compute $\hat{N}_2^+$ and $\hat{N}_2^-$ for b and c:

$$\begin{aligned}\hat{N}_2^+(b) &= \{x \in X : \hat{q}(b,x) \leq 2\} \\ &= \{a,b\}\end{aligned} \qquad \begin{aligned}\hat{N}_2^+(c) &= \{x \in X : \hat{q}(c,x) \leq 2\} \\ &= \{c,d\}\end{aligned}$$

$$\begin{aligned}\hat{N}_2^-(b) &= \{x \in X : \hat{q}(x,b) \leq 2\} \\ &= \{a,b\}\end{aligned} \qquad \begin{aligned}\hat{N}_2^-(c) &= \{x \in X : \hat{q}(x,c) \leq 2\} \\ &= \{c,d\}\end{aligned}$$

Similarly, we can compute the other $\hat{N}_k^+$ and $\hat{N}_k^-$ for b and c:

$$\begin{aligned}\hat{Q}^+(b) &= (|\hat{N}_1^+(b)|, |\hat{N}_2^+(b)|, |\hat{N}_3^+(b)|, |\hat{N}_4^+(b)|) \\ &= (|\{b\}|, |\{a,b\}|, |\{a,b\}|, |\{a,b,c,d\}|) \\ &= (1,2,2,4)\end{aligned} \qquad \begin{aligned}\hat{N}_\bullet^+(c) &= (|\hat{N}_1^+(c)|, |\hat{N}_2^+(c)|, |\hat{N}_3^+(c)|, |\hat{N}_4^+(c)|) \\ &= (|\varnothing|, |\{c,d\}|, |\{b,c,d\}|, |\{a,b,c,d\}|) \\ &= (0,2,3,4)\end{aligned}$$

$$\begin{aligned}\hat{N}_\bullet^-(b) &= (|\hat{N}_1^-(b)|, |\hat{N}_2^-(b)|, |\hat{N}_3^-(b)|, |\hat{N}_4^-(b)|) \\ &= (|\{b\}|, |\{a,b\}|, |\{a,b,c,d\}|, |\{a,b,c,d\}|) \\ &= (1,2,4,4)\end{aligned} \qquad \begin{aligned}\hat{N}_\bullet^-(c) &= (|\hat{N}_1^-(c)|, |\hat{N}_2^-(c)|, |\hat{N}_3^-(c)|, |\hat{N}_4^-(c)|) \\ &= (|\varnothing|, |\{c,d\}|, |\{c,d\}|, |\{a,b,c,d\}|) \\ &= (0,2,2,4)\end{aligned}$$

5.2 Homology recovery

We compared the suitability of three landmarking procedures (uniformly random, maxmin, lastfirst) on datasets with varying density and duplication patterns by extending an example of de Silva and Carlsson (2004). Each expriment proceeded as follows: We sampled $n = 540$ points from the sphere $\mathbb{S}^2 \subset \mathbb{R}^3$ and in different experiments selected $k = 12, 36, 60$ landmark points. We then used the landmarks to compute PH and computed the *relative dominance* $(R_1 - R_0)/K_0$ and *absolute dominance* $(R_1 - R_0)/K_1$ of the last interval over which all Betti numbers agreed with those of $\mathbb{S}^2$. These statistics provide an indication of how successfully PH recovered the homology of the manifold from which the points were sampled.

The points $x = (r, \theta, \phi)$ were sampled using four procedures: uniform sampling, skewed sampling, uniform sampling with skewed boosting, and skewed sampling with skewed boosting. The first procedure was used by de Silva and Carlsson (2004) and here serves as a baseline. For a sample S (with multiplicities) generated from each of the other three procedures, the expected density $\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} |\{x = (r, \theta, \phi) \in S : \alpha - \varepsilon < \phi < \alpha + \varepsilon\}|$ of points near a given latitude $\alpha \in [0, \pi]$ is proportional to the quartic function $p : [0, 1] \to [0, 1]$ defined by $p(x) = (\frac{\phi}{\pi})^4$.[13] Skewed sampling is performed via rejection sampling: Points $x_i = (r_i, \theta_i, \phi_i)$ are sampled uniformly and rejected at random if a uniform random variable $t_i \in [0, 1]$ satisfies $(\frac{\phi_i}{\pi})^\alpha < t_i$ until n points have been kept (Diaconis, Holmes, and Shahshahani 2013). Skewed boosting is performed by first obtaining a (uniform or skewed) sample T of size a fraction $\frac{n}{6}$ of the total, then sampling n points (with replacement) from T using the probability mass function satisfying $P(x_i) \propto (\frac{\phi_i}{\pi})^\beta$. When performed separately, skewed sampling and skewed boosting use $\alpha = \beta = 4$; when performed in sequence, they use $\alpha = \beta = 2$.

The landmark points were selected in three ways: uniform random selection (without replacement), the maxmin procedure, and the lastfirst procedure. We computed PH in Python GUDHI, using three implementations: Vietoris–Rips (VR) filtrations on the landmarks, alpha complexes on the landmarks, and witness complexes on the landmarks with the full sample as witnesses (Maria et al. 2021; Rouvreau 2021; Kachanovich 2021).

The skewed data sets are dense at the south pole and sparse at the north pole. We expect lastfirst to be more sensitive to this variation and place more landmarks toward the south. As measured by dominance, therefore, we hypothesized that lastfirst would be competitive with maxmin when samples are uniform and inferior to maxmin when samples are skewed. Put differently, we expected lastfirst to better reject the homology of $\mathbb{S}^2$, i.e. to detect the statistical void at the north pole.

Figure 9 compares the relative dominance of the spherical homology groups in each case. When PH is computed using VR or alpha complexes, maxmin better recovers the homology of the sphere except on uniform samples, while lastfirst and random selection better detect the void. Random selection is usually better than lastfirst selection at detecting this void when samples are non-uniform, which indicates that lastfirst selection still oversamples from less dense regions. Lastfirst and maxmin perform similarly when PH is computed using witness complexes.

6 References

[13] Should this be analytically shown or confirmed using large samples?

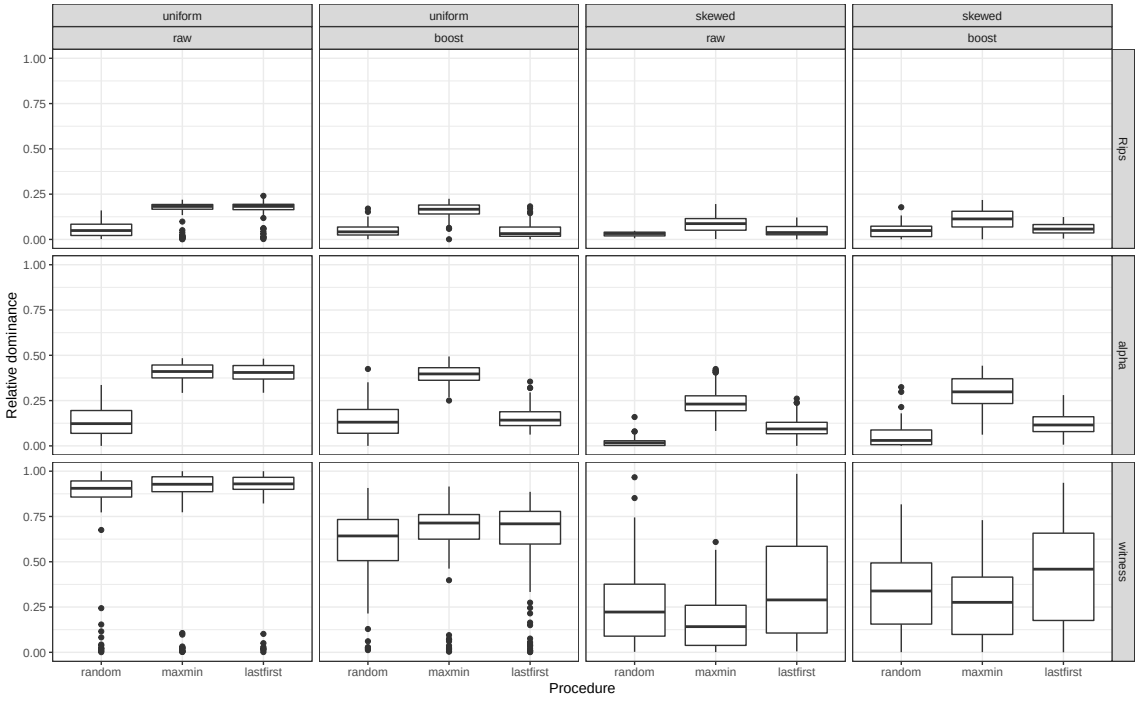


Figure 9. Relative dominance of the spherical homology groups in the persistent homology of four samples from the sphere, using each of three landmark procedures and three persistence computations. Similar plots of absolute dominance (not shown) tell a consistent story, but the distributions are more skewed so the comparisons are less clear.

Akkiraju, Nataraj, Herbert Edelsbrunner, Michael Facello, Ping Fu, Ernst P. Mücke, and Carlos Varela. 1995. "Alpha Shapes: Definition and Software." In.

Aziz, Rabia, C. K. Verma, Namita Srivastava, Rabia Aziz, C. K. Verma, and Namita Srivastava. 2017. "Dimension Reduction Methods for Microarray Data: A Review." *AIMSBOA* 4 (1): 179–97. https://doi.org/10.3934/bioeng.2017.1.179.

Becht, Etienne, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2019. "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP." *Nat Biotechnol* 37 (1): 38–44. https://doi.org/10.1038/nbt.4314.

Bouguessa, Mohamed, Shengrui Wang, and Haojun Sun. 2006. "An Objective Approach to Cluster Validation." *Pattern Recognition Letters* 27 (13): 1419–30. https://doi.org/10.1016/j.patrec.2006.01.015.

Brunson, Jason Cory, and Yara Skaf. 2021. "landmark: Procedures to Generate Landmark Sets from Finite Metric Spaces."

Byczkowska-Lipińska, Liliana, and Agnieszka Wosiak. 2017. "Redukcja Strumienia Danych Pozyskiwanych z Urządzeń Diagnostyki Medycznej Za Pomocą Technik Selekcji Przypadków." *Przegląd Elektrotechniczny* 93 (12): 115–18. https://doi.org/10.15199/48.2017.12.29.

Dai, Leyu, He Zhu, and Dianbo Liu. 2020. "Patient Similarity: Methods and Applications." *arXiv:2012.01976 [Cs]*, December. https://arxiv.org/abs/2012.01976.

Dave, Rajesh N. 1996. "Validating Fuzzy Partitions Obtained Through c-Shells Clustering." *Pattern Recognition Letters* 17 (6): 613–23. https://doi.org/10.1016/0167-8655(96)00026-8.

de Silva, Vin, and Gunnar Carlsson. 2004. "Topological Estimation Using Witness Complexes." In *SPBG'04 Symposium on Point-Based Graphics 2004*, edited by Markus Gross, Hanspeter Pfister, Marc Alexa, and Szymon Rusinkiewicz, 157–66. The Eurographics Association.

Diaconis, Persi, Susan Holmes, and Mehrdad Shahshahani. 2013. "Sampling from a Manifold." In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, 10:102–25. Institute of Mathematical Statistics Collections. Institute of Mathematical Statistics.

Dłotko, Paweł. 2019. "Ball Mapper: A Shape Summary for Topological Data Analysis," January.

Eddelbuettel, Dirk, and Romain Francois. 2011. "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software* 40 (8): 1–18. https://doi.org/10.18637/jss.v040.i08.

Falasconi, M., A. Gutierrez, M. Pardo, G. Sberveglieri, and S. Marco. 2010. "A Stability Based Validity Method for Fuzzy Clustering." *Pattern Recognition* 43 (4): 1292–1305. https://doi.org/10.1016/j.patcog.2009.10.001.

Goldberger, Ary L., Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation* 101 (23). https://doi.org/10.1161/01.CIR.101.23.e215.

Hester, Jim. 2020. "bench: High Precision Timing of R Expressions."

Ivakhno, Sergii, and J. Douglas Armstrong. 2007. "Non-Linear Dimensionality Reduction of Signaling Networks." *BMC Syst Biol* 1 (1): 27. https://doi.org/10.1186/1752-0509-1-27.

Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. "MIMIC-III, a Freely Accessible Critical Care Database." *Scientific Data* 3 (1). https://doi.org/10.1038/sdata.2016.35.

Johnston, J. W. 1976. "Similarity Indices I: What Do They Measure." United States.

Kachanovich, Siargey. 2021. "Witness Complex." In *GUDHI User and Reference Manual*, 3.4.1 ed.

GUDHI Editorial Board.

Konstorum, Anna, Nathan Jekel, Emily Vidal, and Reinhard Laubenbacher. 2018. "Comparative Analysis of Linear and Nonlinear Dimension Reduction Techniques on Mass Cytometry Data." Preprint. https://doi.org/10.1101/273862.

Lee, Joon, David M. Maslove, and Joel A. Dubin. 2015. "Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric." Edited by Frank Emmert-Streib. *PLOS ONE* 10 (5): e0127428. https://doi.org/10.1371/journal.pone.0127428.

Major, Vincent J, Neil Jethani, and Yindalon Aphinyanaphongs. 2020. "Estimating Real-World Performance of a Predictive Model: A Case-Study in Predicting Mortality." *JAMIA Open* 3 (2): 243–51. https://doi.org/10.1093/jamiaopen/ooaa008.

Maria, Clément, Paweł Dłotko, Vincent Rouvreau, and Marc Glisse. 2021. "Rips Complex." In *GUDHI User and Reference Manual*, 3.4.1 ed. GUDHI Editorial Board.

Meyer, David, and Christian Buchta. 2021. "proxy: Distance and Similarity Measures."

Park, Yoon-Joo, Byung-Chun Kim, and Se-Hak Chun. 2006. "New Knowledge Extraction Technique Using Probability for Case-Based Reasoning: Application to Medical Diagnosis." *Expert Systems* 23 (1): 2–20. https://doi.org/10.1111/j.1468-0394.2006.00321.x.

Piekenbrock, Matt. 2020. "simplextree: Provides Tools for Working with General Simplicial Complexes."

Reutlinger, Michael, and Gisbert Schneider. 2012. "Nonlinear Dimensionality Reduction and Mapping of Compound Libraries for Drug Discovery." *Journal of Molecular Graphics and Modelling* 34 (April): 108–17. https://doi.org/10.1016/j.jmgm.2011.12.006.

Rouvreau, Vincent. 2021. "Alpha Complex." In *GUDHI User and Reference Manual*, 3.4.1 ed. GUDHI Editorial Board.

Singh, Gurjeet, Facundo Mémoli, and Gunner Carlsson. 2007. "Topological Methods for the Analysis of High Dimensional Data Sets and 3d Object Recognition." In *Eurographics Symposium on Point-Based Graphics*, edited by M. Botsch, R. Pajarola, B. Chen, and M. Zwicker. The Eurographics Association. https://doi.org/10.2312/SPBG/SPBG07/091-100.

Skaf, Yara, and Reinhard Laubenbacher. 2022. "Topological Data Analysis in Biomedicine: A Review." *Journal of Biomedical Informatics* 130 (June): 104082. https://doi.org/10.1016/j.jbi.2022.104082.

Viswanath, Satish E., Pallavi Tiwari, George Lee, Anant Madabhushi, and for the Alzheimer's Disease Neuroimaging Initiative. 2017. "Dimensionality Reduction-Based Fusion Approaches for Imaging and Non-Imaging Biomedical Data: Concepts, Workflow, and Use-Cases." *BMC Med Imaging* 17 (1): 2. https://doi.org/10.1186/s12880-016-0172-6.

Wang, Weina, and Yunjie Zhang. 2007. "On Fuzzy Cluster Validity Indices." *Fuzzy Sets and Systems* 158 (19): 2095–2117. https://doi.org/10.1016/j.fss.2007.03.004.

Yoon, Hee Rhang, and Robert Ghrist. 2020. "Persistence by Parts: Multiscale Feature Detection via Distributed Persistent Homology." arXiv. https://doi.org/10.48550/arXiv.2001.01623.

Zhong, Haodi, Grigorios Loukides, and Robert Gwadera. 2020. "Clustering Datasets with Demographics and Diagnosis Codes." *Journal of Biomedical Informatics* 102 (February): 103360. https://doi.org/10.1016/j.jbi.2019.103360.