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 interface 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 in the use of computational theories of continuity to investigate the shape or structure of data. While TDA is most commonly associated with persistent homology (PH) and mapper, it can be understood to encompass and generalize many conventional and even classical techniques, including cluster analysis, network analysis, and nearest neighbors prediction.

Two basic maneuvers in TDA are locality preservation and cardinality reduction. *Locality preservation* is the property of some functions, such as projections or hashes, that nearby cases or points in the domain have nearby images in the codomain. (Continuity, as defined in analysis and topology, is a type of locality preservation.) This is the defining property of many non-linear dimensionality reduction techniques, including t-SNE and UMAP, but also an asymptotic property of nearest neighbors prediction and the locally-sensitive hashing used in its implementations. We use the term *cardinality reduction*[1], in contrast to dimension reduction, to describe techniques that produce more tractable or comprehensible representations of complex data by reducing the number of cases or points rather than the number of variables or dimensions used to represent them. Cardinality reduction describes cluster analysis and association (co-occurrence or correlation) network analysis, for example.

More elaborate TDA techniques combine both maneuvers, often through the use of covering methods, as with the approximation of PH through witness complexes (de Silva and Carlsson 2004) or in the mapper construction (Singh, Mémoli, and Carlsson 2007). Covering methods, in turn, can be enhanced by strategic sampling from large data sets [@], as can other proximity-based techniques like nearest neighbors. The maxmin procedure is often used for this purpose, as it is deterministic, is computationally efficient, and generates more evenly distributed samples than random selection. However, maxmin comes with its own potential limitations in the analysis of data that vary greatly in density or have multiplicities. This is a frequent concern when sparse, heterogeneous, and incomplete data are modeled as finite pseudometric spaces.

In this paper, we develop a landmark sampling procedure complementary to maxmin, based on the rankings of points by a distance or similarity measure rather than on the raw values of such a measure. In the remainder of this section, we motivate the procedure, which we call lastfirst, as a counterpart to maxmin obtained by adapting this procedure from the use of fixed-radius balls to the use of fixed-cardinality neighborhoods. We then formally describe the procedure and prove some of its basic properties in Section 2. In Section 3 we report the results of benchmark tests and robustness checks of lastfirst on simulated and empirical data sets. We describe some basic and novel applications to real-world data in Section 4.

[1]The term "cardinality reduction" takes different meanings in the data analysis literature, including the combining of different values of a categorical variable (Micci-Barreca 2001; Refaat 2010) or of time series (Hu et al. 2011) (also "numerosity reduction" (Lin et al. 2007)). Our meaning is that of 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.

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 $|Y|$, and the set of distinguishable points of Y—or of equivalence classes under co-location—is denoted $[Y]$. When $Y, Z \subseteq X$, let $Y \smallsetminus Z$ denote the set of points in Y (with multiplicities) that are distinguishable from all points in Z. This means that, when defined, $\min_{y \in Y \smallsetminus Z, z \in Z} d(y, z) > 0$.

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

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

If $d(x, y) = d(x, z) \implies y = z$, then X is considered to be in *ego-general position*, even if $d(x, y) = d(z, w) \not\implies \{x, y\} = \{z, w\}$. This is condition implies that X is Hausdorff: $d(x, y) = 0 \implies x = y$. $f : X \to Y$ will denote a morphism of pseudometric spaces, meaning a function (or morphism of sets) that preserves locality in the sense that $d_X(x, y) \geq d_Y(f(x), f(y))$.

We use the ball notation $B_\varepsilon(x)$ for the set of points less than distance ε from a point x; that is, $B_\varepsilon(x) = \{y \mid d(x, y) < \varepsilon\}$. We use an overline to also include points exactly distance ε from x: $\overline{B_\varepsilon}(x) = \{y \mid d(x, y) \leq \varepsilon\}$. (While these connote openness and closedness in the discrete topology, every such ball is both open and closed.) If $|\overline{B_\varepsilon}(x)| \geq k$ and $\varepsilon' < \varepsilon \implies |\overline{B'_\varepsilon}(x)| < k$, then we say that $N_k^+(x) = \overline{B_\varepsilon}(x)$ is the k-nearest neighborhood of x. When X is in general position, therefore, $|N_k(x)| = k$.

Throughout, let $N = |X|$. For convenience, we assume $0 \in \mathbb{N}$. We use the notation $[a, b]$ with $a < b$ as shorthand for the arithmetic sequence $(a, a + 1, \ldots, b)$ and a^b with $b \in \mathbb{N}$ for the sequence $(a, \ldots, a)$ of length b.

1.2 Background

We designed the lastfirst procedure to addresses an issue with the maxmin procedure that arises when, due to the use of binary or categorical variables or to limits on measurement resolution, a data set includes many duplicate or otherwise indistinguishable cases. These render the finite metric space representation of the data non-Hausdorff. While these issues may be negligible when such points are rare, they raise computational and interpretative concerns when they are common. Because our procedure is motivated by the same practical needs as the maxmin procedure, we begin with a discussion of those needs.

The earliest appearance of the maxmin[2] procedure of which we are aware is by de Silva and Carlsson (2004). The authors propose witness complexes, later generalized to alpha complexes [@], for the rapid approximation of persistent homology: 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

[2]This procedure is distinct from many other uses of "maxmin" and related terms.

each scale. They use the maxmin procedure, which we define in Section 2.1, as an alternative to selecting landmark points uniformly at random. The procedure ensures that the landmarks are locally separated and roughly evenly distributed. While the procedure improved little upon uniform random selection in most use cases, on some tasks it far outperformed.

Subsequent uses of maxmin 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 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). In addition to approximating persistent homology (de Silva and Carlsson 2004; Dłotko 2019), maxmin has been used to reduce the sizes of simplicial complex models of point cloud data for the sake of visualization and exploration (Singh, Mémoli, and Carlsson 2007; Dłotko 2019).

1.3 Motivation

The ball covers mentioned above have been proposed as an alternative to mapper (Dłotko 2019), where they exchange complexity for computational cost, and the mapper construction itself relies on a crucial covering step that has received limited theoretical attention. Conventionally, mappers rely on covers consisting either of overlapping intervals (when the lens is one-dimensional) or of their cartesian products (higher-dimensional). For this purpose, we propose that ball covers, heuristically optimized using the maxmin procedure, have a potential advantage over conventional covers, alongside a potential disadvantage.

Conventionally, mappers use low-dimensional lens spaces $\mathbb{R}^p$ and one of two types of cover, based on overlapping intervals either of fixed length or at fixed quantiles (Piekenbrock 2020). We think of this length or quantile as the *resolution* of the cover. When $m > 1$, covers for $Y \subset \mathbb{R}^m$ can be obtained as the cartesian products of those of the coordinate projections of Y—so, if $\pi_1, \pi_2 : \mathbb{R}^2 \to \mathbb{R}$ are the coordinate projections, then interval covers $\{I_\alpha\}$ and $\{J_\beta\}$ of $\pi_1(Y)$ and $\pi_2(Y)$ give rise to the rectangle cover $\{I_\alpha \times J_\beta\}$ for Y. While these cover types are manageable in very few dimensions, the number of sets scales geometrically with p, holding the resolution fixed. Moreover, eventually most of the resulting cover sets will contain no points of X, and additional calculations will be needed to restrict to the non-empty sets.

In contrast, a cover obtained by centering balls at a subset of landmark points in Y will have greater up-front computational cost but will be guaranteed to contain no empty sets, and the number of sets required to capture the topology of Y will increase only with the geometric and topological complexity of Y, not with p. (We test this hypothesis in Section 4.)

Nevertheless, the maxmin cover requires a meaningful distance metric: the dissimilarity of cases x and y is captured by their distance $d(x, y)$, regardless of where x and y are located in X, and the neighborhoods $B_r(x)$ and $B_r(y)$ about landmarks x and y play an equal role in the cover.[3] This means that cover sets centered at landmarks in sparse regions of X will be more numerous and of lower cardinality than those centered in dense regions. The assumption is violated in many real-world settings, including much of biomedicine. For example, in psychometric terms, this would mean that inter-case distance is an *interval*, not only an *ordinal*, variable, so that the distances between cases in a point cloud representation has a definite meaning independent of which cases are

[3] Is there a term for this property, e.g. something being a "universal constant?"

considered. This assumption is often made for convenience, but it generally does not follow from theory.

This motivates 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 (nearly) the same cardinality. Especially in analyses of medical and healthcare data, underlying variables can often only be understood as ordinal. Other representations of high-dimensional data sets are commonly defined by similarity (or dissimilarity) measures such as cosine similarity rather than by coordinates and associated metrics. 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.

2 Procedures

In this section we review the maxmin procedure and introduce the lastfirst procedure as a complement to it.

2.1 Maxmin procedure

Maxmin takes as input a proper subset $Y \subset X$ and returns as output a point $x \in X \setminus Y$. To build intuition, consider its calculation also with reference to X itself:

Given $L \subset X$ and $x \in X \setminus L$, define the *maxmin sets*

$$\operatorname{maxmin}(L) = \{x \in X \setminus L \mid d(x, L) = \max_{y \in X \setminus L} d(y, L)\}$$
$$\operatorname{maxmin}(X) = \{x \in X \mid d(x, X \setminus \{x\}) = \max_{y \in X} d(y, X \setminus \{y\})\}$$

consisting of *maxmin points*. Note that $\operatorname{maxmin}(\,\cdot\,)$ is nonempty and that, when X is in ego-general position, it has cardinality 1.

The *maxmin procedure* for generating a landmark set $L \subseteq X$ proceeds as follows (see Algorithm 1): 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 $\overline{B_\varepsilon}(\ell)$ minimally cover X. Choose a first landmark point $\ell_0 \in X$. This choice may be arbitrary; we specifically consider three selection rules: the first point index in the object representing X, selection at random, and (random selection from) $\operatorname{minmax}(X)$. Inductively over $i \in \mathbb{N}$, if ever $i \geq n$ or $d(L, X \setminus L) \leq \varepsilon$, then stop. Otherwise, when $L = \{\ell_0, \ldots, \ell_{i-1}\}$, choose $\ell_i \in \operatorname{maxmin}(L)$, again according to a preferred rule. If a fixed number n of landmarks was prescribed, then set $\varepsilon = d(L, X \setminus L)$; if ε was prescribed, then set $n = |L|$.

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 $L = [X]$. When the procedure stops, $X = \bigcup_{i=0}^{n-1} \overline{B_\varepsilon}(\ell_i)$. This cover is minimal in the sense that both the removal of any $\overline{B_\varepsilon}(\ell_i)$ and any decrease in ε will yield a collection of sets that fail to cover X.[4] More generally, a non-minimal cover can be obtained by specifying both n and ε in a compatible way. In Section 3, we describe two adaptive parameters implemented in our software package that make these choices easier.

[4] We have not investigated whether or when a landmark cover set is a minimal cover.

Algorithm 1 Select a maxmin landmark set.

Require: finite metric space (X, d)
Require: at least one parameter $\varepsilon \geq 0$ or $n \in \mathbb{N}$
Require: seed point $\ell_0 \in X$
if only ε is given **then**
  $n \leftarrow 1$
end if
if only n is given **then**
  $\varepsilon \leftarrow \infty$
end if
$L \leftarrow \varnothing$
$i \leftarrow 0$
repeat
  $L \leftarrow L \cup \{\ell_i\}$
  $i \leftarrow i + 1$
  $\ell_i \in \text{maxmin}(L)$
  $d_{\max} \leftarrow d(\ell_i, L)$
until $d_{\max} < \varepsilon$ and $|L| \geq n$
return maxmin landmark set L

2.2 Lastfirst procedure

The lastfirst procedure is defined analogously to the maxmin procedure, substituting "ego-rank" for the pseudometric d_X.

2.2.1 Ego-ranks

Ego-rank is an adaptive notion of distance with respect to nearest neighborhoods.[5] It relies on the underlying pseudometric d_X but takes values in $\mathbb{N}$ given by the ordinal of one point's distance from another.

Definition 1. (ego-rank) For $x, y \in X$, define the *ego-rank* $q_X : X \times X \longrightarrow \mathbb{N}$ as follows:

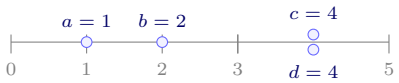
$$q_X(x, y) = |\{z \in X \mid d_X(x, z) < d_X(x, y)\}| + 1$$

As with d, we allow ourselves to write $q = q_X$ when clear from context. Note that $q(x, x) = 1$ and $q(x, y) \leq N$, and that q is not, in general, symmetric (or, therefore, a pseudometric).[6]

Example 2. Consider the simple case $X = \{a, b, c, d\}$, visualized below, equipped with the standard Euclidean metric:

[5] As used here, ego-rank is distinct from the *rank-distance* between permutations, which is used to define rank correlation coefficients, and from the *ordinal distance* proposed by Pattanaik and Xu (2008), a loosening of the concept of pseudometric that dispenses with the triangle inequality.

[6] Yara: These images might be made larger, with less or no transparency.



Lack of symmetry of q is shown by points b and c :

$$\begin{aligned} q(a,c) &= |\{x \in X \mid |x-a| < |c-a| = 2\}| + 1 \\ &= |\{a,b\}| + 1 \\ &= 3 \end{aligned} \qquad \begin{aligned} q(c,a) &= |\{x \in X \mid |x-c| < |a-c| = 2\}| + 1 \\ &= |\{b,c,d\}| + 1 \\ &= 4 \end{aligned}$$

Observe in particular that $\max_{x \in X} q(a,x) = 3 < |X|$; $q(x, \cdot\,)$ will not max out at N when the most distant points from x have multiplicity.

However, $q(x,x) = 1$ whether x has multiplicity or not:

$$\begin{aligned} q(a,a) &= |\{x \in X \mid |x-a| < |a-a| = 0\}| + 1 \\ &= |\varnothing| + 1 \\ &= 1 \end{aligned} \qquad \begin{aligned} q(c,c) &= |\{x \in X \mid |x-c| < |c-c| = 0\}| + 1 \\ &= |\varnothing| + 1 \\ &= 1 \end{aligned}$$

We 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 then be used to define *out-* and *in-neighborhoods* of x.[7]

Definition 3. (k-neighborhoods) For $x \in X$, define the *k-out-neighborhoods* N_k^+ and *k-in-neighborhoods* N_k^- of x as the sets

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

Note that $\varnothing \subseteq N_1^{\pm}(x) \subseteq \cdots \subseteq N_N^{\pm}(x) = X$. The k-out-neighborhoods of x are the sets of points in X that have ego-rank at most k from x. This is equivalent to the k-nearest neighbors of x. The k-in-neighborhoods of x are the sets of points in X from which x has ego-rank at most k. These definitions can be adapted as follows to be relative to a subset $Y \subseteq X$, using the notation q_X to emphasize that the points in $X \setminus (Y \cup \{x\})$ are still involved in the calculation:

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

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

Example 4. Consider the same X as in Example 2. Compute N_k^+ and N_k^- for a and c, using $k = 3$:

$$N_3^+(a) = \{x \in X \mid q(a,x) \leq 3\} = \{a,b,c,d\} \qquad N_3^+(c) = \{x \in X \mid q(c,x) \leq 3\} = \{b,c,d\}$$

$$N_3^-(a) = \{x \in X \mid q(x,a) \leq 3\} = \{a,b\} \qquad N_3^-(c) = \{x \in X \mid q(x,c) \leq 3\} = \{a,b,c,d\}$$

Ego-ranks are not as straightforward to compare among subsets of points. For example, for $y \neq x$, $q(x,y)$ takes integer values between $\overline{B_0}(x) + 1$ and N. To explain and motivate their use, we now intuitively adapt the minmax and maxmin procedures to this setting, then state and prove a formal definition for their analogs.

2.2.2 Rank sequences and landmark selection

Consider the case of choosing the next landmark point given a subset $L = \{\ell_0, \ldots, \ell_{i-1}\} \subset X$ of collected landmark points. For a ball cover, we choose $\ell_i \in X \setminus L$ so that the *minimum radius* ε required for some $B_\varepsilon(\ell_j)$ to contain ℓ_i is *maximized*. If X is in general position, the choice of ℓ_i will be unique; otherwise, ℓ_i can be chosen at random from the subset of X that satisfies this criterion. Analogously, for a neighborhood cover, we want that the *minimum cardinality* k required for some $N_k^+(\ell_j)$ to contain ℓ_i is *maximized*. 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 $\ell \in X \setminus L$ 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.

We formalize this procedure, and a companion procedure without reference to L, by defining a sequence of neighborhood sizes that encodes the optimized cardinalities k_i.

Definition 5. (Rank sequences) For $x \in X$ and $Y \subset X$, define the *out-rank* (Q^+) and *in-rank* (Q^-) *sequences* of k-neighborhood cardinalities:

$$Q^\pm(x) = (|N_k^\pm(x)|)_{k=1}^N \qquad Q^\pm(x,Y) = (|N_k^\pm(x,Y)|)_{k=1}^N$$

Remark 6. If $Q^\pm(x) = (Q_1, \ldots, Q_N)$, then $Q^\pm(x, X \setminus \{x\}) = (Q_1 - 1, \ldots, Q_N - 1)$.

Remark 7. Taking the $k_i^\pm$ as defined above,

$$Q^+(x,L) = (k_1^+ - 1, k_2^+ - k_1^+, \ldots, k_{|L|}^+ - k_{|L|-1}^+, |L|)$$
$$Q^-(x,L) = (0^{k_0^-}, 1^{k_1^- - k_0^-}, \ldots, (|L|-1)^{k_{|L|-1}^- - k_{|L|-2}^-}, |L|^{N - k_{|L|-1}^-})$$

An example of this computation is shown in Example 8.

Example 8. Continuing Example 4, we can compute the other N_k^+ and N_k^- to obtain Q^+ and Q^- for a and c:

$$\begin{aligned} Q^+(a) &= (|N_1^+(a)|, |N_2^+(a)|, |N_3^+(a)|, |N_4^+(a)|) \\ &= (|\{a\}|, |\{a,b\}|, |\{a,b,c,d\}|, |\{a,b,c,d\}|) \\ &= (1,2,4,4) \end{aligned} \qquad \begin{aligned} Q^+(c) &= (|N_1^+(c)|, |N_2^+(c)|, |N_3^+(c)|, |N_4^+(c)|) \\ &= (|\{c,d\}|, |\{c,d\}|, |\{b,c,d\}|, |\{a,b,c,d\}|) \\ &= (2,2,3,4) \end{aligned}$$

$$\begin{aligned} Q^-(a) &= (|N_1^-(a)|, |N_2^-(a)|, |N_3^-(a)|, |N_4^-(a)|) \\ &= (|\{a\}|, |\{a,b\}|, |\{a,b\}|, |\{a,b,c,d\}|) \\ &= (1,2,2,4) \end{aligned} \qquad \begin{aligned} Q^-(c) &= (|N_1^-(c)|, |N_2^-(c)|, |N_3^-(c)|, |N_4^-(c)|) \\ &= (|\{c,d\}|, |\{c,d\}|, |\{a,b,c,d\}|, |\{a,b,c,d\}|) \\ &= (2,2,4,4) \end{aligned}$$

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 Q^+ and Q^- to emphasize the sizes of smaller neighborhoods. Sequences with more large values indicate points with lower ego-ranks to or from more other points.

We now define counterparts to the minmax and maxmin procedures using these totally ordered sequences.

Definition 10. (Firstlast and lastfirst) Given $Y \subset X$, define the *lastfirst sets*

$$\begin{aligned} \mathrm{lf}(Y; X) = \mathrm{lf}(Y) &= \{x \in X \setminus Y \mid Q^-(x, Y) = \max_{y \in X \setminus Y} Q^-(y, Y)\} \\ \mathrm{lf}(X) &= \{x \in X \mid Q^-(x, X \setminus \{x\}) = \max_{y \in X} Q^-(y, X \setminus \{y\})\} \end{aligned}$$

consisting of *lastfirst points*.

Example 11. Return again to $X = \{a, b, c, d\}$ from Example 2. We calculate an exhaustive lastfirst landmark set, seeded with a point of minimal out-rank sequence:

1. We have

   $$\begin{aligned} Q^+(a, \{b,c,d\}) &= (0,1,3,3) & Q^+(b, \{a,c,d\}) &= (0,1,3,3) \\ Q^+(c, \{a,b,d\}) &= (1,1,2,3) & Q^+(d, \{a,b,c\}) &= (1,1,2,3) \end{aligned}$$

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

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

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

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

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

2.2.3 Algorithms

Algorithm 2 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 12 provides a shortcut between Q^- and the more compact way that the ego-rank data are stored.

Lemma 12. *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 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. □

Proposition 13. *Algorithm 2 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 (**subset?**) 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 2. Note that, for the algorithm to terminate its loop and subsequently return L, either there must be no points in $X \setminus L$ 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, $F = \text{lf}(\{\ell_0, \ldots, \ell_{i-1}\})$. When F is calculated on line 22, the rows R_r of R contain the ego-ranks $q(\ell_i, x_r)$ in increasing order. Because $D(L, X \setminus L) > 0$ (line 14), F is nonempty. By Lemma 12, 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 $D(L, X \setminus L) = 0$, 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. [8]

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

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

[8] 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.

Algorithm 2 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 `pick`
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**
  $\ell_i \leftarrow \texttt{pick}(F)$
  $L \leftarrow L \cup \{\ell_i\}$
  $D_i \leftarrow (d_{i1}, \ldots, d_{iN}) \in {\mathbb{R}_{\geq 0}}^N$, 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**
    $F \leftarrow \{x_r \in F \mid R_{rj} = \max_{r'} R_{r'j}\}$
    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

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. □

2.2.4 Tie handling

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

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

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 $\hat{q}(x, x) > 1$ when x has multiplicity. The two ego-ranks derive from two tie-handling schemes for calculating rankings of lists with duplicates. For example, if $a < b = c < d$ are the distances from x to y_1, y_2, y_3, y_4, respectively, then $(\check{q}(x, y_1), \check{q}(x, y_2), \check{q}(x, y_3), \check{q}(x, y_4)) = (1, 2, 2, 4)$ and $(\hat{q}(x, y_1), \hat{q}(x, y_2), \hat{q}(x, y_3), \hat{q}(x, y_4)) = (1, 3, 3, 4)$. 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 14. Recall $X = \{a, b, c, d\}$ from Example 2. The ego-rank $\hat{q}$ is also asymmetric:

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

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

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

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 4, we can compute $\hat{N}_2^+$ and $\hat{N}_2^-$ for b and c:

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

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

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

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

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

3 Implementation

We have implemented the lastfirst procedure, together with maxmin, in the R package landmark (Brunson and Skaf 2021). 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 ego-rank–based procedures, the user can choose any tie-handling rule. 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 redundant covers.

3.1 Validation

We validated the firstlast and lastfirst procedures against several small example data sets, including that of Example 2. 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.

3.2 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

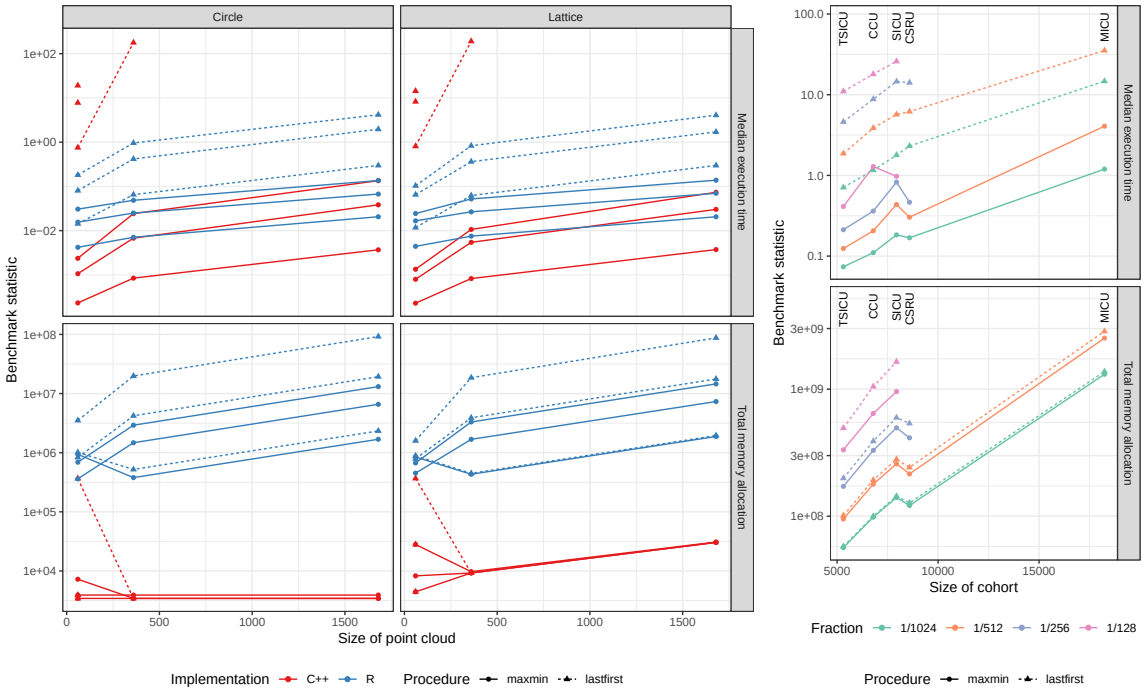


Figure 1. Benchmark results for computing landmarks on two families of artificial data (circle and lattice) and one collection of empirical data (RT-similarity space of critical care units in MIMIC-III). Some points are missing because benchmark tests did not complete within 1 hour.

each critical care unit in MIMIC-III using RT-transformed data and cosine similarity (Section 4.1). We conducted benchmarks using the bench package (Hester 2020) on the University of Florida high-performance cluster HiPerGator.

Benchmark results are reported in Figure 1. The R implementation of maxmin used orders of magnitude more memory and took slightly longer. They appeared to scale slightly better in terms of time and slightly worse in terms of memory. 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 takes much longer—a median factor of almost 2000 relative to maxmin in C++—and failed to complete in many of our tests.

4 Experiments

4.1 Empirical data

4.1.1 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).[9] 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),[10] stated religion,[11] marital status[12], and type of medical insurance[13]. 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.

4.1.2 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[14], 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 encoding the presence or absence of a wide variety of comorbidities such as asthma, hypertension, cardiovascular disease. (For a full description of each field included in the data set, see Kaggle.*) 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

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

[10] 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

[11] 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

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

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

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

set is very complete in that every patient is required to contain a valid value for every field, which minimizes concerns around missing data.

4.2 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 $\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} |\{x = (r, \theta, \phi) \in S \mid \alpha - \varepsilon < \phi < \alpha + \varepsilon\}|$ of points near a given latitude $\alpha \in [0, \pi]$ is proportional to the quartic function $p : [0, 1] \to [0, 1]$ defined by $p(x) = (\frac{\phi}{\pi})^4$.[15] 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 2 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.

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

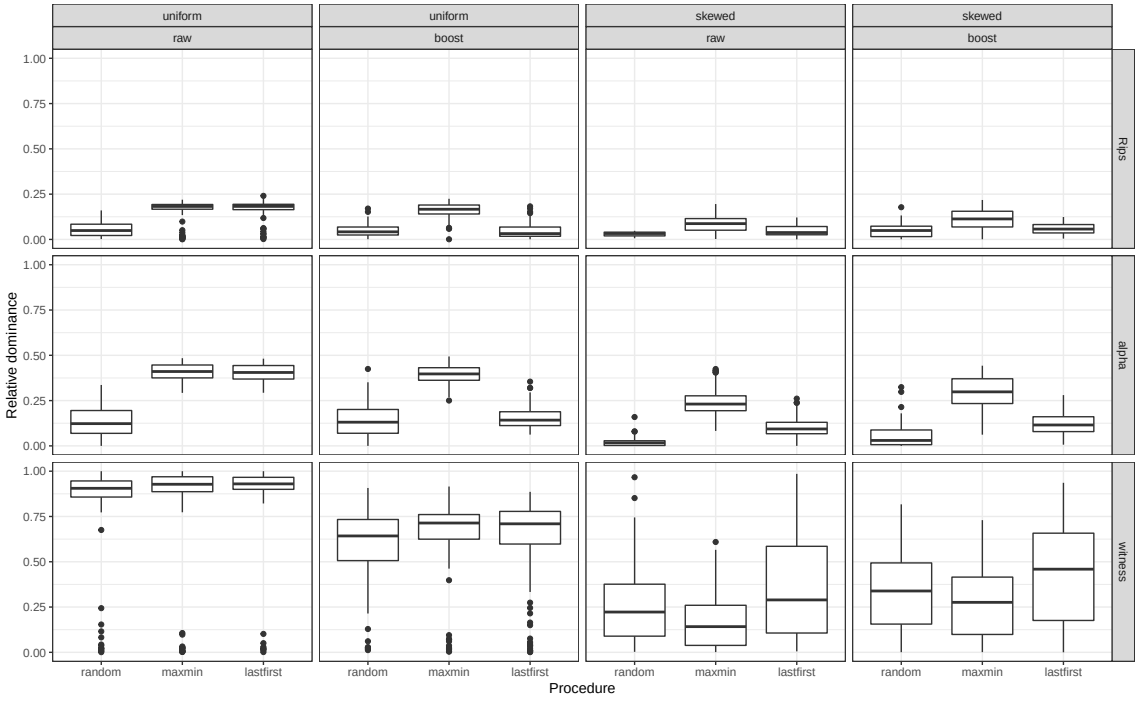


Figure 2. Relative dominance of the spherical homology groups in the persistent homology of four samples from the sphere, using each of three landmark procedures and three persistence computations. Similar plots of absolute dominance (not shown) tell a consistent story, but the distributions are more skewed so the comparisons are less clear.

4.3 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 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 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).

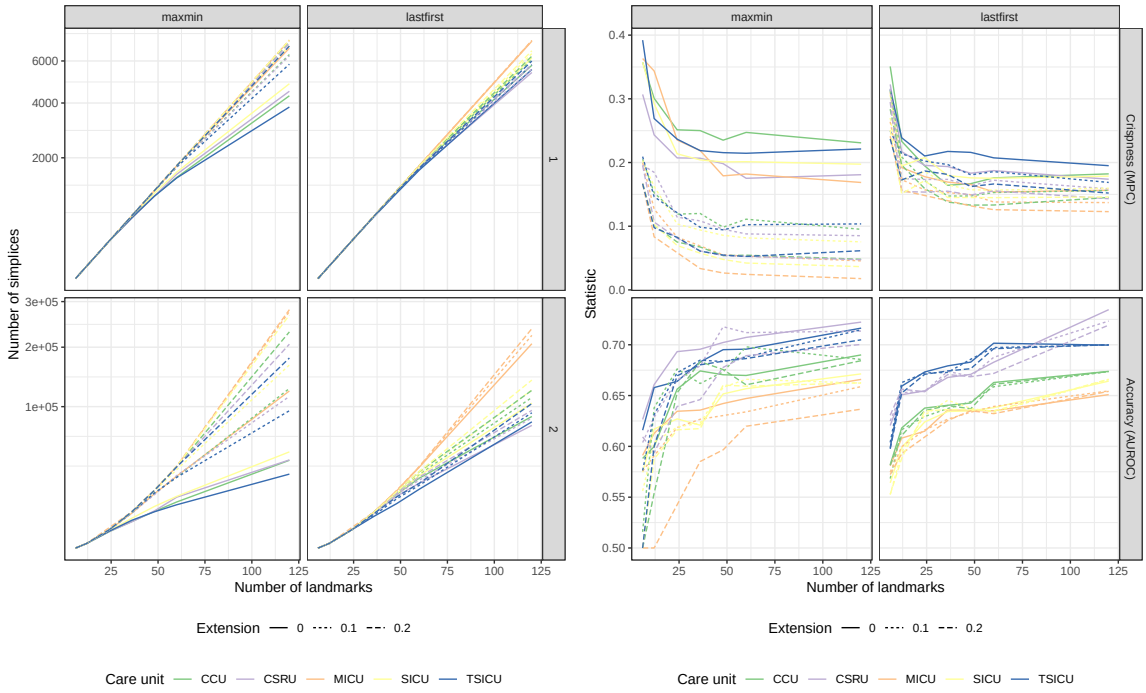


Figure 3. Summary and evaluation statistics versus number of 0-simplices (landmarks) for the covers generated using the maxmin and lastfirst procedures, with three multiplicative extensions in their size. Left: the sizes of their nerves, as numbers of 1- and 2-simplices, using a square root–transformed vertical scale. Right: the modified partition coefficient (MPC) and the c-statistic of the risk prediction model based on the cover sets (AUROC).

Figure 3 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. In addition to (i.e. 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.

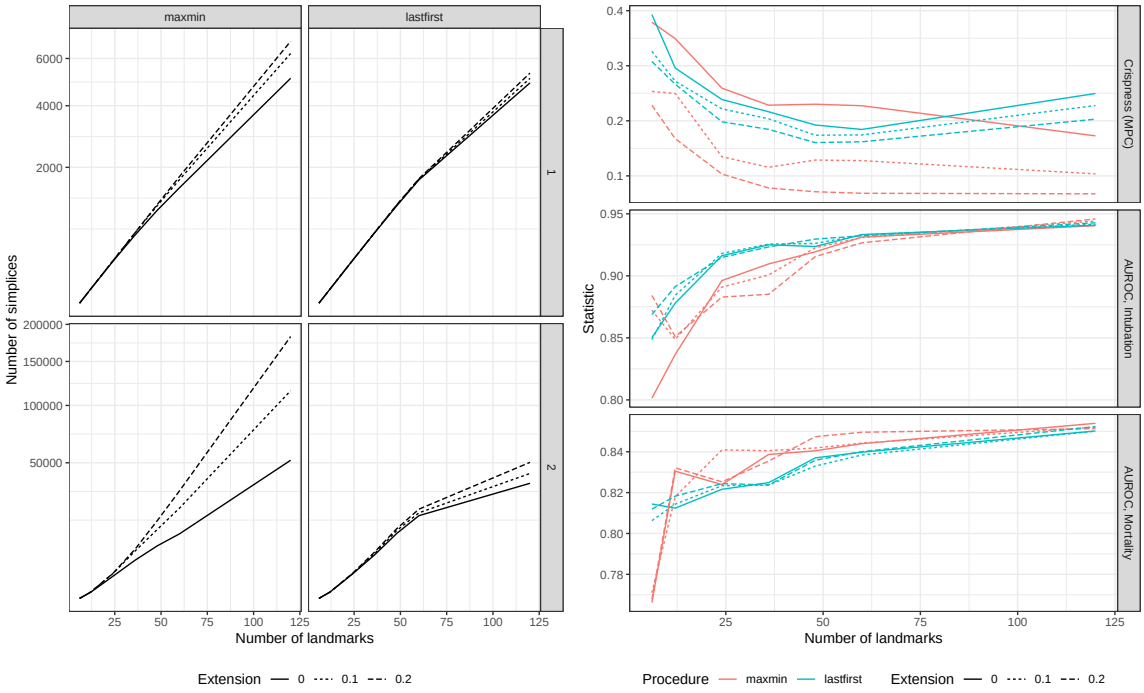


Figure 4. Summary and evaluation statistics versus number of 0-simplices (landmarks) for the covers generated using the maxmin and lastfirst procedures, with three multiplicative extensions in their size. Left: the sizes of their nerves, as numbers of 1- and 2-simplices, using a square root–transformed vertical scale. Right: the modified partition coefficient (MPC) and the c-statistics of the risk prediction models based on the cover sets (AUROC).

Figure 4 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[16], though for smaller numbers different selection procedures performed best for different outcomes.

4.4 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

[16] We should increase the maximum number of landmarks to be sure.

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–fit–test cross-validation. That is, partition $[n] = \bigsqcup_{i=1}^{6} I_i$ into roughly equal parts, and partition each $[n] \setminus I_i = \bigsqcup_{j=1}^{6} J_{ij}$ into roughly equal parts.
2. Iterate the following steps over each i, j:
   a) Generate a sequence L of landmarks from the points $X_{([n] \setminus I_i) \setminus J_{ij}}$.
   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) and number of landmarks ($|L| = 36, 60, 180, 360$). 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.

Boxplots of the AUROCs for each cross-validation step are presented in Figure 5. 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 nested-longitudinal CV approach: 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 6, with one curve (across numbers of landmarks) per selection procedure, outcome, and week. Again both landmark selection procedures

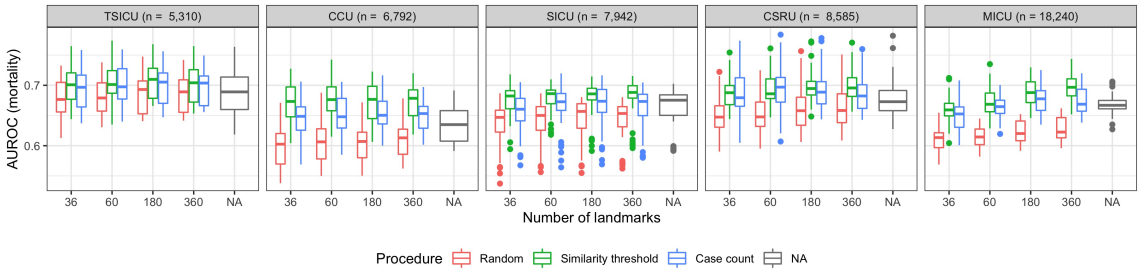


Figure 5. AUROCs of the interpolative predictive models of mortality in five MIMIC-III care units based on covers constructed using random, maxmin, and lastfirst procedures to generate landmarks. Each boxplot summarizes AUROCs from $6 \times 6 = 36$ models, one for each combination of outer and inner fold. AUROCs of simple nearest-neighbor predictive models are included for comparison.

yielded stronger results than random selection. Interestingly, this was more pronounced in later weeks, as the pandemic progressed, even as overall predictive accuracy declined.[17] Overall, performance improved slightly as the number of landmarks increased from 50 to 150 but either plateaued or declined from 150 to 250.

5 Discussion

The definitions of our lastfirst and firstlast procedures are analogous to those of maxmin and minmax, 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 ego-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.

We ran an experiment to compare maxmin and lastfirst landmarks at expediting the computation of persistent homology, extending an experiment of de Silva and Carlsson (2004). In addition to a uniform sample from a sphere, we drew skewed and bootstrapped samples in order to simulate data sets with variable density and multiplicity, in this case exhibiting a statistical void at one pole of the sphere, opposite a concentration at the other pole. Whereas the maxmin procedure would sample more uniformly across the sphere despite this skew, the lastfirst procedure would concentrate landmarks toward the south. Classical persistent homology is notoriously sensitive to outliers, and maxmin better recovered spherical homology than lastfirst, as expected. This is likely due in part to the filtration itself being based on distances rather than ego-ranks between landmarks (and other points as witnesses). Yet, compared to random sampling, lastfirst still oversampled from less dense

[17] 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.

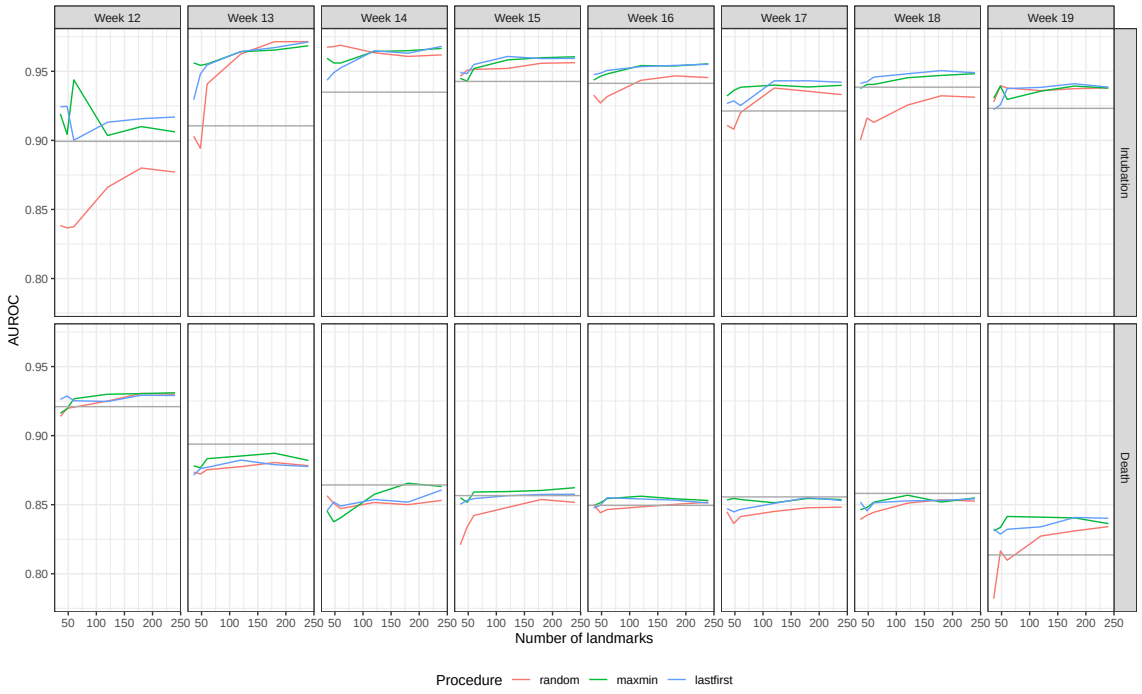


Figure 6. AUROCs of the sliding-window interpolative predictive models of intubation and mortality in the MXDH data based on covers constructed using random, maxmin, and lastfirst procedures to generate landmarks.

regions. This is desirable for settings, such as healthcare, in which regions of missingness often indicate limitations of the data collection rather than rarity of case types in the population.

We also 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.

A possible explanation for the stronger performance of maxmin on MIMIC is that the data did not exhibit very strongly the patterns for which the lastfirst procedure is designed to account, namely variation in density and multiplicity. As a result, the RT-similarity measure is in fact meaningful across the population: Whatever the baseline presentation of a patient, rather than a cohort of similar patients of some fixed size, their prognosis would be better guided by a cohort cut off at a fixed minimum similarity (or maximum distance). 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. In contrast, this stronger performance was not evident on MXDH, which contained fewer variables and as a result exhibited many more occurrences of duplication. 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.

Another way to think about these results is in terms of a balance between relevance and pwoer, 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.

6 References

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* (version 0.0.0.9000). https://github.com/corybrunson/landmark.

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.

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.

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. http://projecteuclid.org/euclid.imsc/1379942050.

Dłotko, Paweł. 2019. "Ball Mapper: A Shape Summary for Topological Data Analysis," January. http://arxiv.org/abs/1901.07410.

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* (version 1.11.1). https://CRAN.R-project.org/package=bench.

Hu, Bing, Thanawin Rakthanmanon, Yuan Hao, Scott Evans, Stefano Lonardi, and Eamonn Keogh. 2011. "Discovering the Intrinsic Cardinality and Dimensionality of Time Series Using MDL." In *Proceedings - IEEE International Conference on Data Mining, ICDM*, 1086–91. https://doi.org/10.1109/ICDM.2011.54.

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.

Kachanovich, Siargey. 2021. "Witness Complex." In *GUDHI User and Reference Manual*, 3.4.1 ed. GUDHI Editorial Board. https://gudhi.inria.fr/doc/3.4.1/group___witness___complex.html.

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.

Lin, Jessica, Eamonn Keogh, Li Wei, and Stefano Lonardi. 2007. "Experiencing SAX: A Novel Symbolic Representation of Time Series." *Data Mining and Knowledge Discovery* 15 (2): 107–44. https://doi.org/10.1007/s10618-007-0064-z.

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. https://gudhi.inria.fr/doc/3.4.1/group___rips___complex.html.

Meyer, David, and Christian Buchta. 2021. *proxy: Distance and Similarity Measures* (version 0.4-25). https://CRAN.R-project.org/package=proxy.

Micci-Barreca, Daniele. 2001. "A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems." *ACM SIGKDD Explorations Newsletter* 3 (1): 27–27. https://doi.org/10.1145/507533.507538.

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* (version 1.0.1). https://github.com/peekxc/simplextree/.

Refaat, Mamdouh. 2010. "Cardinality Reduction." In *Data Preparation for Data Mining Using SAS*, 142–57. Elsevier. https://books.google.com/books?id=FkH9gjihLqMC.

Rouvreau, Vincent. 2021. "Alpha Complex." In *GUDHI User and Reference Manual*, 3.4.1 ed. GUDHI Editorial Board. https://gudhi.inria.fr/doc/3.4.1/group___alpha___complex.html.

Silva, Vin de, 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. http://diglib.eg.org/handle/10.2312/SPBG.SPBG04.157-166.

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.

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.

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.