The utility of fixed and adaptive landmark algorithms for healthcare informatics

Jason Cory Brunson   Yara Skaf

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 [@deSilva2004] or in the mapper construction [@Singh2007]. 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. One example of this scenario is the problem of clinical outcome prediction on a diverse population of patients, many of whom are at similarly high risk for an outcome such as mortality. This particular use case contributed to our initial motivation to develop the alternative approach discussed in this paper.

In Section , we briefly describe a landmark sampling procedure we developed to complement maxmin, describe some of its basic properties, and report the results of our validation studies and benchmark tests on simulated and empirical data. We then describe some basic and novel applications to real-world data in Section in an effort to showcase the potential utility of such algorithms in the setting of biomedical informatics.

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 @deSilva2004, where it is used 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

---

[1]The term "cardinality reduction" takes different meanings in the data analysis literature, including the combining of different values of a categorical variable [@MicciBarreca2001; @Refaat2010] or of time series [@Hu2011] (also "numerosity reduction" [@Lin2007]). Our meaning is that of @ByczkowskaLipinska2017: 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.

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

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 [@Singh2007]; 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 [@Dlotko2019]. In addition to approximating persistent homology [@deSilva2004; @Dlotko2019], maxmin has been used to reduce the sizes of simplicial complex models of point cloud data for the sake of visualization and exploration [@Singh2007; @Dlotko2019].

Motivation

The ball covers mentioned above have been proposed as an alternative to mapper [@Dlotko2019], 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 [@Piekenbrock2020]. 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 .)

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 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.

Procedures

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

We restrict ourselves to general, non-technical descriptions of the algorithms here but we plan to include a more mathematically rigorous discussion in the upcoming article [INSERT TITLE].

---

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

Overview of Algorithms

Maxmin procedure

[Fill in non-mathematical description of the algorithm.]

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
    $L \leftarrow L \cup \{\ell_i\}$
    $i \leftarrow i + 1$
    $\ell_i \in \mathrm{maxmin}(L)$
    $d_{\max} \leftarrow d(\ell_i, L)$
  until $d_{\max} < \varepsilon$ and $|L| \geq n$
  return maxmin landmark set L

Lastfirst procedure

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

[Fill in non-mathematical description of the algorithm.]

Implementation

We have implemented all four procedures (firstlast, lastfirst, minmax, and maxmin) in the R package landmark [@Brunson2021a], both in C++ using Rcpp [@Eddelbuettel2011] and separately in R. Users can specify all relevant parameters, including the number and size (meaning radius or cardinality) of the desired cover sets. Additional optional parameters can be used to extend the covers to include more and larger sets with increased overlaps, which may produce more illustrative output in certain scenarios. Any of the algorithms will return the resulting landmarks, optionally together with the sets of points in the cover set (ball or neighborhood) centered at each landmark.

Validation & Benchmarking

We validated the firstlast and lastfirst procedures against several small example data sets as well as on several larger data sets. We invite readers to install the package and experiment with new test cases, as well as to request or write any desired additional features.

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 the RT-similarity measure (Section). We conducted benchmarks using the bench package [@Hester2020] on the University of Florida high-performance cluster HiPerGator.

Algorithm 2 Select a lastfirst landmark set.

Require: finite metric space (X, d)
Require: at least one parameter $k > 0$ or $n \in \mathbb{N}$
Require: seed point $\ell_0 \in X$
Require: selection procedure `pick`
  if only k is given **then**
    $n \leftarrow 0$
  end if
  if only n is given **then**
    $k \leftarrow |X|$
  end if
  $L \leftarrow \varnothing$
  $i \leftarrow 0$
  repeat
    $L \leftarrow L \cup \{\ell_i\}$
    $i \leftarrow i + 1$
    $F \leftarrow \mathrm{lf}(L)$
    $\ell_i \leftarrow \texttt{pick}(F)$
    $q_{\max} \leftarrow q(\ell_i, L)$
  until $q_{\max} < k$ and $|L| \geq n$
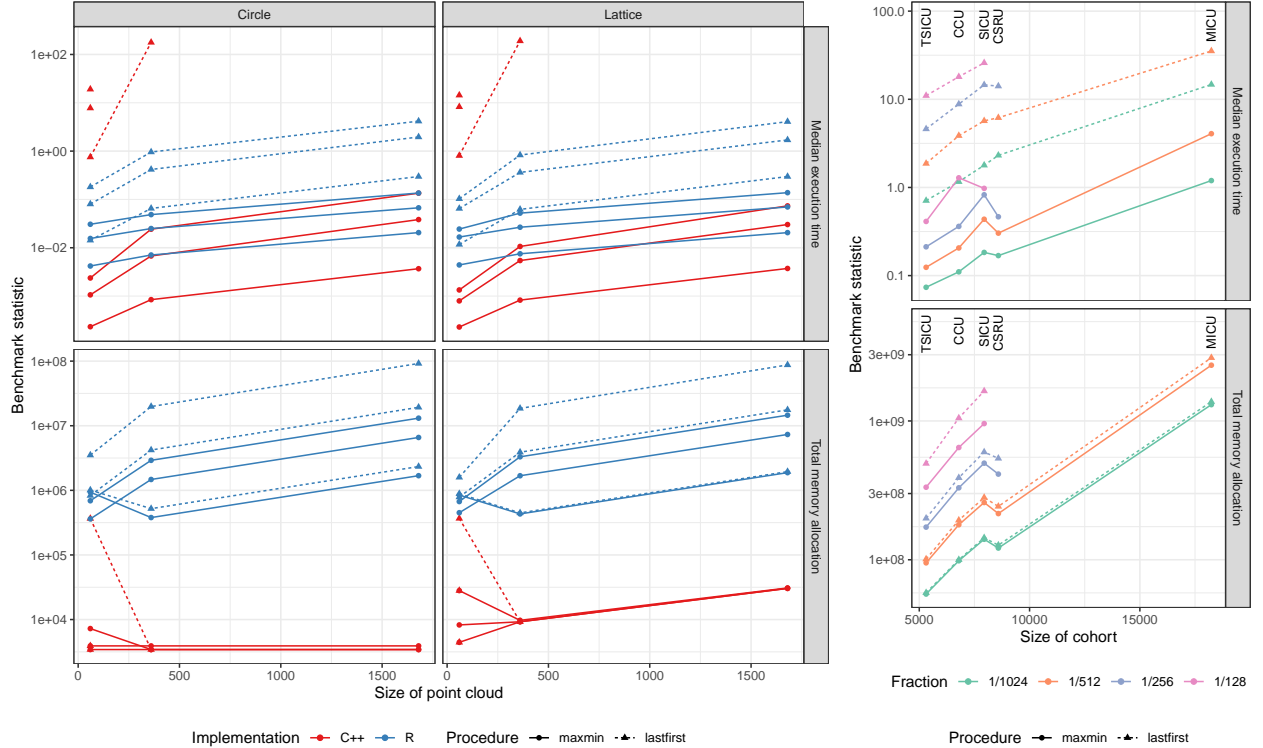 **return** lastfirst landmark set L

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.

Benchmark results are reported in Figure 1. R implementations 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.

Applications

Empirical data

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 [@Goldberger2000; @Johnson2016]. 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).[4] 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),[5] stated religion (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, or Lutheran), marital status (married, single, widowed, divorced, unknown/missing, separated, or life partner), and type of medical insurance (Medicare, private, Medicaid, povernment, or self pay). Following @Zhong2020, 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.

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[6], a collaborative data science platform, where they are maintained and updated regularly. The database 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) temporal data, and (2) categorical or binary variables. The temporal data consist of key dates associated with the clinical course of infection such as date of symptom onset, date of admission to a healthcare institution, and date of death (if applicable). The categorical or 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 distinguish between many phenotypic

---

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

[5] 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

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

subtypes of COVID-19 patients. 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.

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 sampling procedures maxmin and lastfirst 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 [@Singh2007].

As a practical test of this potential, we loosely followed the approach of @Dlotko2019 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 [@Bouguessa2006; @Wang2007; @Falasconi2010]. To our knowledge, the sole exception to have appeared in a comprehensive comparison of such measures is the *modified partition coefficient* [@Dave1996], defined as

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

  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 2 presents 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.
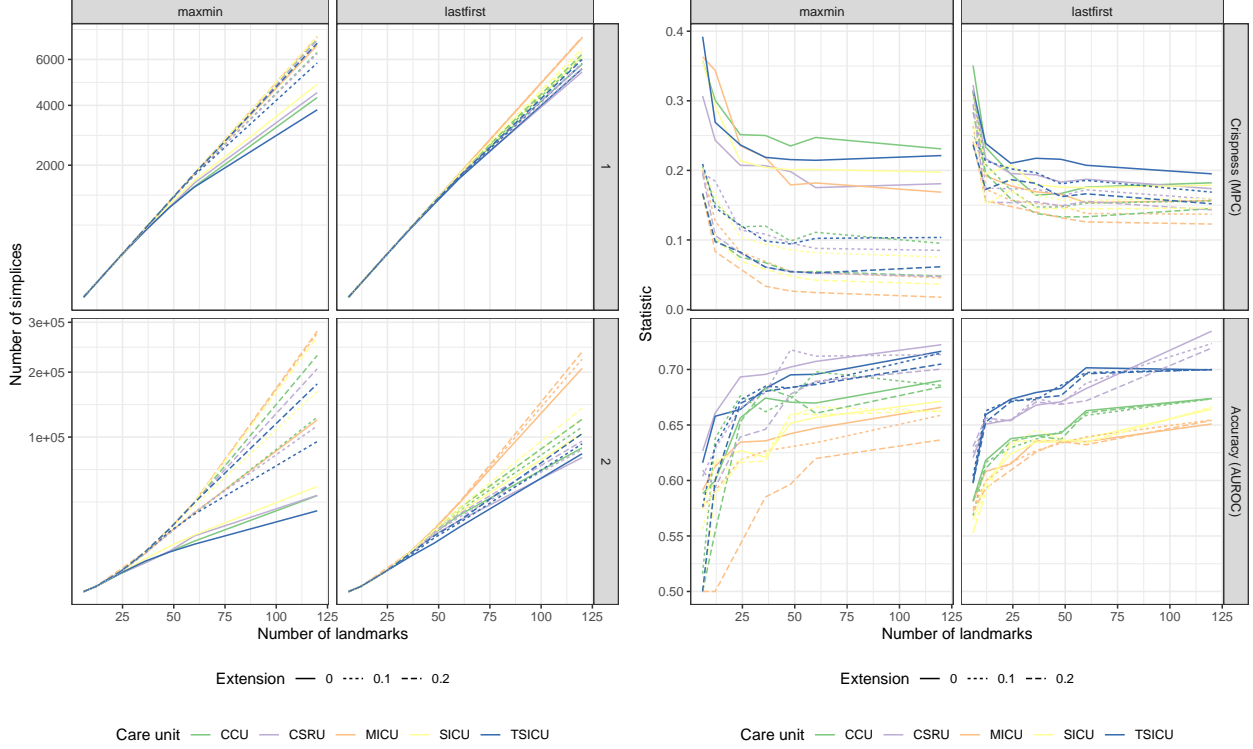

Figure 2.   Summary and evaluation statistics versus number of 0-simplices (landmarks) for the covers generated using the maxmin and lastfirst procedures, with three multiplicative extensions in their size. Left: the sizes of their nerves, as numbers of 1- and 2-simplices, using a square root–transformed 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 the same evaluations for covers of the MXDH data. In contrast to the MIMIC experiments, lastfirst-based nerves of the MXDH data grow sub-polynomially and are significantly sparser than maxmin-based nerves. Lastfirst covers tend to be crisper, especially as the number of landmarks and the extension factors increase. This indicates that the nearest neighborhoods form a more parsimonious cover of the data than the centered balls. The predictive accuracies of the maxmin- and lastfirst-selected cover-based models converge with increasing landmarks[7], though for smaller numbers different selection procedures perform best for different outcomes.

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 $I_0 \sqcup I_1 = \{1, \dots, n\}$ 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

---

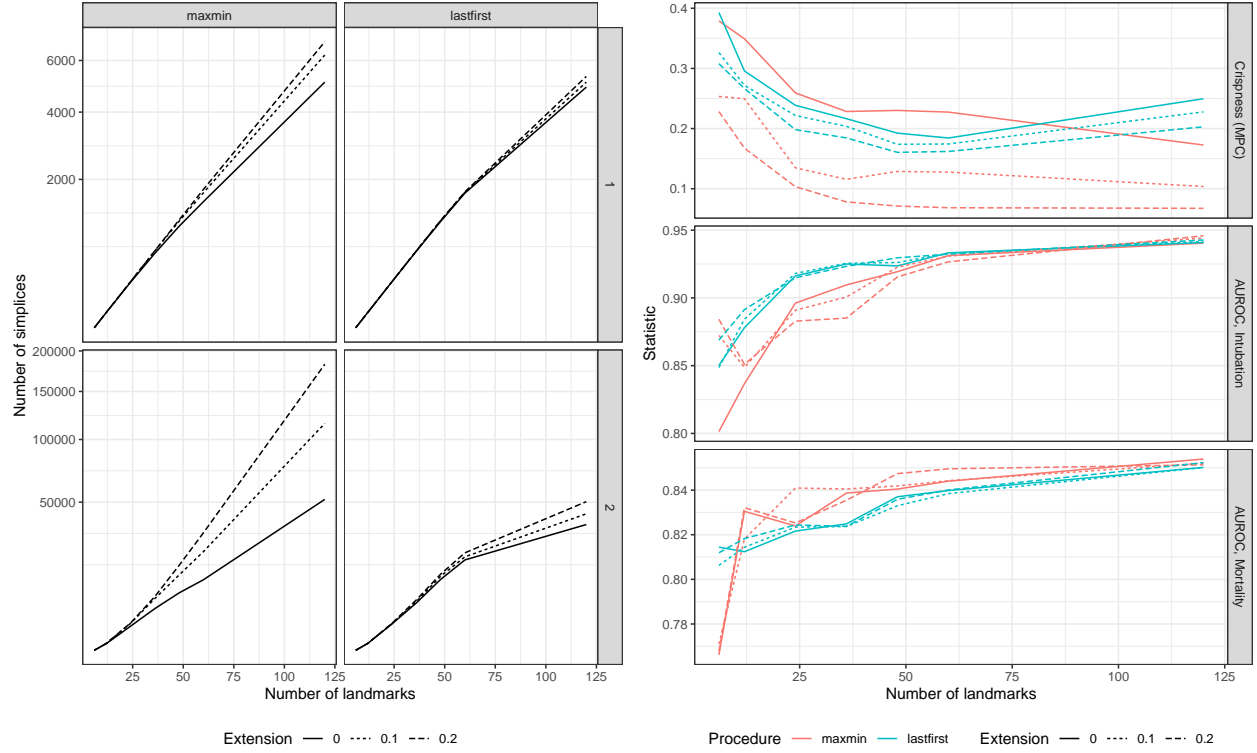[7]We should increase the maximum number of landmarks to be sure.

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-statistics of the risk prediction models based on the cover sets (AUROC).

averaging the responses for the kth 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 $[n] \smallsetminus I_i = \bigsqcup_{j=1}^{6} J_{ij}$ into roughly equal parts.
2. Iterate the following steps over each i, j:
   a) Generate a sequence L of landmarks from the points $X_{([n] \smallsetminus I_i) \smallsetminus J_{ij}}$.
   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 [@Lee2015].
   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 4. 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).
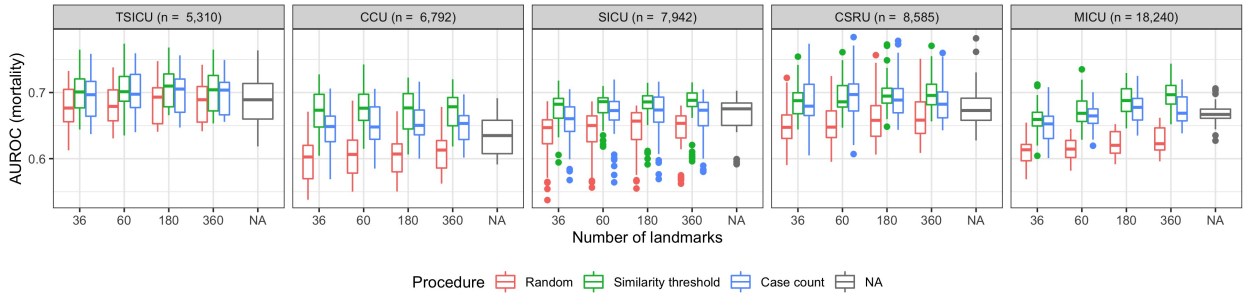

Figure 4. AUROCs of the interpolative predictive models of mortality in five MIMIC-III care units based on covers constructed using random, maxmin, and lastfirst procedures to generate landmarks. Each boxplot summarizes AUROCs from $6 \times 6 = 36$ models, one for each combination of outer and inner fold. AUROCs of simple nearest-neighbor predictive models are included for comparison.

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 5, with one curve (across numbers of 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.[8] Overall, performance improved slightly as the number of landmarks increased from 50 to 150 but either plateaued or declined from 150 to 250.
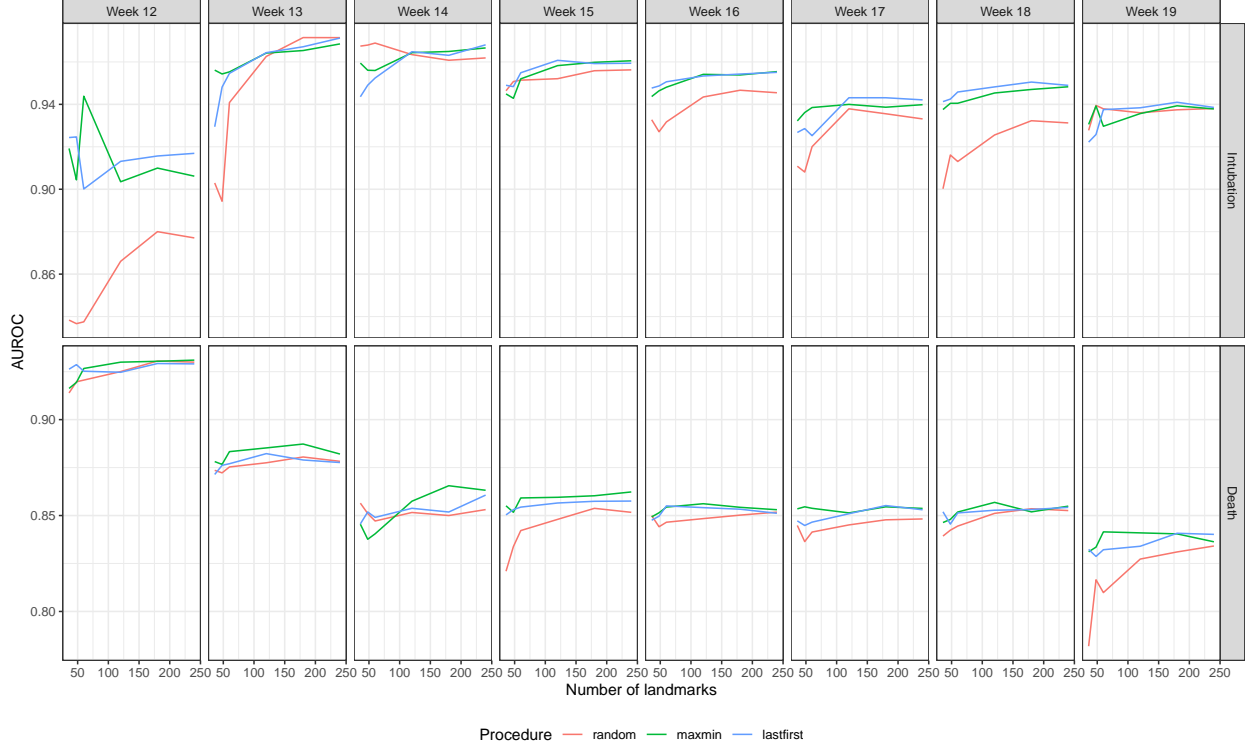

Figure 5. AUROCs of the sliding-window interpolative predictive models of intubation and mortality in the MXDH data based on covers constructed using random, maxmin, and lastfirst procedures to generate landmarks.

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 rank-distances 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

---

[8]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.

homology, extending an experiment of @deSilva2004. 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 rank-distances between landmarks (and other points as witnesses). Yet, compared to random sampling, lastfirst still oversampled from less dense 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 @Lee2015, may be strengthened by optimizing a fixed similarity threshold rather than a fixed cohort size. It is worth noting that @Park2006 reached a similar conclusion. In contrast, this stronger performance was not evident on MXDH, which contained fewer variables and as a result exhibited many more occurrences of duplication.

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.

References