Computers & Graphics (2020)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Vectorization of persistence barcode with applications in pattern classification of porous structures

Zhetong Dong[a], Chuanfeng Hu[a], Chi Zhou[a], Hongwei Lin[a,b,*]

[a]School of Mathematical Sciences, Zhejiang University, Hangzhou, Zhejiang Province, China
[b]State Key Lab. of CAD&CG, Zhejiang University, Hangzhou, Zhejiang Province, China

ARTICLE INFO

Article history:
Received May 3, 2020

Keywords: Computational Topology, Machine Learning, Persistent Homology, Porous Classification

ABSTRACT

Persistence barcode is a topological summary for persistent homology to exhibit topological features with different persistence. Persistence rank function (PRF), derived from persistence barcode, organizes persistence Betti numbers in the form of an integer-valued function. To obtain topological patterns of objects such as point clouds represented by finite-dimensional vectors for machine learning classification tasks, the vectorizing representations of barcodes is generated via decomposing PRF on a system of Haar basis. Theoretically, the generated vectorizing representation is proved to have 1-Wasserstein stability. In practice, to reduce training time and achieve better results, a technique of dimensionality reduction through out-of-sample mapping in supervised manifold learning is used to generate a low-dimensional vector. Experiments demonstrate that the representation is effective for capturing the topological patterns of data sets. Moreover, the classification of porous structures has become an essential problem in the fields such as material science in recent decades. The proposed method is successfully applied to distinguish porous structures on a novel data set of porous models.

1. Introduction

Persistent homology (PH) [1], an effective tool for obtaining topological features of spatial objects, has been used to discover topological patterns of practical data in recent decades. Topological patterns can be understood as representations related to topological invariants, such as connected components in zero-dimension, loops in one-dimension, and spatial voids in two-dimension. Patterns represented by topology and geometry have been widely adopted in numerous studies, for instance, image processing [2], and network analysis [3]. Especially, in computer graphics, the topological features of PH are used to solve problems such as point cloud recognition [4], and surface reconstruction [5].

PH can be used to infer possible 'shape' of a point cloud in Euclidean space. Assume that on each point, a ball with its center to be the point and the initial radius to be zero is assigned. As the radius of each ball increases, simplexes are determined by the intersections of these balls, and a simplicial complex is generated corresponding to the value of the radius. In this dynamical procedure, the homological invariants in different dimension appear (or are born) at a moment when the value of the radius is b, and it might disappear (or die) when some high-dimensional simplex appears at the value of d. Therefore, the pair (b, d) is the life span of the topological invariants, and the persistence is its lifetime $d - b$. One can obtain a set of these pairs in each dimension. A persistence *barcode* [6, 7], shown in Figure 1 as an instance, is ordered linear segments obtained by embedding each pair into 2D Euclidean space. The endpoints of each 'bar' represent the birth time b_i and the death time d_i of the topological invariant, respectively. And a *persistence di-*

*Corresponding author:
e-mail: hwlin@zju.edu.cn (Hongwei Lin)

agram is another topological summary embedding the set into 2D Euclidean space as points. Because the death time is always greater than the birth time, the points are above the diagonal. To measure the similarity of two barcodes or persistence diagrams, p-Wasserstein distance and the bottleneck distance are defined in [8]. Readers can refer to [9] for more details.

In point cloud classification using topological features with machine learning (ML) tools, the problem is how to transform these persistence barcodes into representations compatible with ML tools. Persistence barcode does not satisfy the multifold requirements of practical applications in ML tasks such as classification. It is because (1) 'bars' in a barcode are unordered, and the number of 'bars' is not fixed; (2) to compute the distance between two barcodes is not straightforward. Therefore, to use topological features with ML tools, we need to identify a vectorizing representation of barcodes.

The new representation of barcode should follow three principles: (1) the representation has an explainable meaning; (2) transformation should preserve the information which a barcode contains, and (3) representation should be stable with respect to the metrics of barcode. The persistence rank function (PRF), also known as persistence Betti number function [10][11], is an ideal alternative to match these principles. First, PRF is a bivariate and non-negative integer-valued function that summarizes persistence Betti numbers (ranks) [12]. And the PRF can be induced from barcode [13]. Second, no external information is introduced to generate PRF. Third, it is shown in Section 4 that PRF in L^2 in the sense of a particular measure has 1-Wasserstein stability in the condition of a small perturbation. Consequently, we propose a framework to represent a barcode as a stable finitely dimensional vector by vectorizing PRF.

As an application of the framework for point cloud classification, the vectorizing representation is used to classify porous structures. Porous structure plays an important role in analyzing the function of nanomaterials in material science [14]. In recent decades, as the databases of porous materials create, such as the database in the material genome initiative [15], the classification of porous materials has become a novel and important problem. However, it is shown in [16] that traditional geometric descriptors do not encode enough topological information to detect materials that have similar global porous structures. Therefore, persistent homology is adopted to capture the overall porous features for quantifying similarity of nanoporous materials. Inspired by this, we generate a data set of porous models with category labels designed by triply periodic minimal surfaces (TPMSs). On this data set, we extend the framework of vectorizing barcode with the technique of dimensionality reduction via out-of-sample mapping in supervised manifold learning to speed up the training process of classifiers while maintaining or even improving the classification accuracy.

Our contribution: In this paper, to vectorize persistence barcodes for ML classification tasks of spatial point clouds using topological features, we propose a finitely dimensional vectorizing representation of barcodes based on Haar basis decomposition of PRF in L^2 space with a limited domain, shown in Figure 1. Theoretically, the generated vectorizing representation is proved to be stable with respect to the 1-Wasserstein distance in the condition of a small perturbation. For practical classification tasks, because of the relatively high dimension of the generated vector, a technique of dimensionality reduction via out-of-sample mapping in supervised manifold learning is employed to extend the vectorization framework. On a novel data set of various porous materials, the proposed vectorizing representation is applied to classify the models, and it has the best performance compared with other vectorizing methods.

2. Related Work

In this section, the vectorizing representations of barcode and persistence diagram for ML classification tasks are first introduced. Then, because the proposed framework is applied to classify the generated data set of porous structures, the generation approaches of 3D porous models are presented.

Vectorizing representations: From an historical perspective, size functions [17] were first used for the vectorization of 0-th homology, and the very first work about algebraic representations of 0-th persistence diagrams is in [18]. The vectorizing representations of a barcode or persistence diagram can be divided into non-learnable methods and kernel methods. There are numerous non-learnable representations for practical problems [19, 20, 21, 22, 23, 24]. For example, persistence images were proposed in [25] based on the integral on each mesh patch of a persistence surface, which is produced in the form of summation of weighted Gaussian functions related to the persistence of each point on a persistence diagram. To combine PH with statistics, a topological summary, called persistence landscapes, was proposed in [26]. This summary is a series of functions in a separable Banach space so that the vector space structure can be used to do statistics, such as computing mean values. Moreover, Robins et al. [13] considered PRF in a Hilbert space and in an affine subspace under reasonable conditions, and they performed functional principal component analysis on experimental data from colloids to study spatial point patterns.

Kernel methods are motivated by searching meaningful measurements to construct kernels used in machine learning models. Persistence scale space kernel (PSSK) was proposed in [27] as a multi-scale kernel by considering a heat diffusion problem. PSSK was proved to be 1-Wasserstein stable, while it does not have higher-degree Wasserstein stability. Kernel embedding of measures was used in [28] to transform a persistence diagram into an element of Hilbert space, and a framework of the kernel method, persistence weighted Gaussian kernel (PWGK) was developed. Meanwhile, the kernel was proved to be stable with respect to Hausdorff distance. Additionally, another kernel method was proposed in [29], referred to as sliced Wasserstein kernel (SWK), for machine learning tasks using persistence diagrams. Readers can refer to the survey [30] for other recent kernel methods.

Production of porous models: Cheah [31, 32] investigated and selected various polyhedral shapes suitable for porous structure modeling and created a parametric library of porous structures. Schroeder [33] introduced stochastic geometry theory into porous structure modeling to represent porous structures

with density and porosity. Cai and Xi [34] proposed a porous structure modeling method based on a shape function and hexahedral mesh refinement. An approach of irregular porous structure modeling based on subdivision and non-uniform rational B-splines was developed in [35]. You et al.[36] presented an improved method based on centroidal Voronoi tessellation and B-splines to design porous structures. To overcome the limitation in geometry of pore-making element, easily control the porosity and pore size, and ensure the internal connectivity, several researchers used TPMSs to design porous structures [37, 38, 39]. With the appealing properties of connectivity, smoothness, and geometric representation, TPMS emerged as a significant tool for designing porous structures.

3. Vectorization of Barcode

In this section, we introduce the vectorizing method of barcode based on PRF, as shown in Figure 1. First, the approach of inducing PRF from barcode and the Hilbert space where PRF is considered are summarized and introduced. Second, we present one of our contributions, that is, the method to generate the feature vector by decomposing PRF using Haar basis on the bounded domain. Finally, we mention that, because the feature vector possesses a relatively high dimension, for tasks of classification, a layer of dimension reduction is provided based on out-of-sample mapping in supervised manifold learning. The approach we employed is introduced in Section 5.

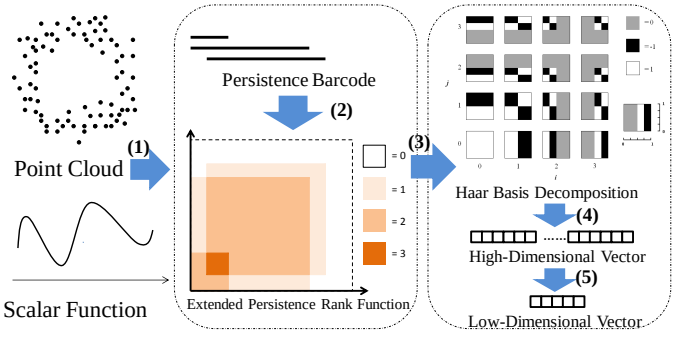


Fig. 1. Pipeline of generating low-dimensional feature vectors from barcodes: (1) Barcodes are computed from a filtration built up on the objects; (2) Extended persistence rank function (PRF) is induced from a barcode in a certain dimension; (3) A finite number of Haar basis functions, of which the first 16 functions are shown in the figure, are used to decompose the extended PRF on a bounded domain; (4) A feature vector with a relatively high dimension is generated by concatenating the coefficients of the Haar basis functions into a vector; (5) A layer of dimensionality reduction is used to produce low-dimensional vectors.

3.1. PRF and Functional Space

In real data sets, the number of bars in a barcode is finite, and most of the bars are distributed in a bounded domain. Therefore, it is reasonable to consider PRF within a limited domain. Given a kth persistence barcode $\{(b_i, d_i) \mid 0 \le b_i \le d_i, i \in I, |I| < \infty\}$, the kth PRF of a filtration V can be induced by the barcode and is defined as:

$$r_{k,V}(s,t) = \sum_{i \in I} r_i(s,t), \tag{1}$$

where

$$r_i(s,t) = \begin{cases} 1, & b_i \le s \le t \le d_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

At the coordinate (s,t) where $s \le t$, the integer value is exactly the persistence Betti number, which, intuitively, means the number of k-dimensional 'holes' existing in the time interval $[s,t]$. And the function in Equation (2) draws a right triangle with its hypotenuse coincident with the diagonal.

We denote $r_{k,V}(s,t)$ as $r(s,t)$ if no confusion occurs. The definition indicates that PRF is a non-negative integer-valued function, precisely, a piecewise constant valued function. Further, PRF is defined on the domain $\mathbb{R}^2_{\Delta+} := \{(s,t) \mid 0 \le s \le t\}$, which maps $\mathbb{R}^2_{\Delta+}$ into $\mathbb{Z} \cup 0 \subset \mathbb{R}$.

To evaluate the distance between two PRFs, the space $L^2(\mathbb{R}^2_{\Delta+}, \mu)$ is chosen here with standard inner product, where μ is a measure. To make a PRF be an element of the Hilbert space $L^2(\mathbb{R}^2_{\Delta+}, \mu)$, the measure μ is supposed to be carefully designed because there may be an infinite homology feature $[b, \infty)$ in a persistence barcode, which makes PRF not be finitely supported in $\mathbb{R}^2_{\Delta+}$. In [13], the measure μ can be obtained by a weighted function $\phi(t-s)$, i.e., $d\mu = \phi(t-s)dsdt$, where $t-s$ has the meaning of persistence of a homology class. It is necessary to show that PRFs are elements of $L^2(\mathbb{R}^2_{\Delta+}, \mu)$. For $r_{k,V}, r_{k,W}$ generated by filtrations V, W, respectively, suppose that V_∞, W_∞ are finite simplicial complexes with $H_k(V_\infty) = H_k(W_\infty)$, and it is proved in [13] that $\|r_{k,V} - r_{k,W}\|_2 < \infty$ if $\int_0^{+\infty} \phi(x)dx < \infty$. Simply, the weighted function $\phi(s,t)$ in the measure μ can be chosen to be

$$\phi(s,t) = \begin{cases} 1, & 0 \le s,t \le B, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where B is a finite number such that the distance is considered in the bounded domain $\Omega = [0,B] \times [0,B]$. The choice of B will be discussed in Section 5.

Finally, to extend the domain of PRF from $\mathbb{R}^2_{\Delta+}$ to the domain $\mathbb{R}^2_+ = \{(s,t) \mid s,t \ge 0\}$ to make the function decomposed by a basis defined on $\mathbb{R}^2_+$, the extended PRF is defined by making PRF be symmetric with respect to the diagonal, that is,

$$\widetilde{r}_{k,V}(s,t) = \sum_{i \in I} \widetilde{r}_i(s,t), \tag{4}$$

where

$$\widetilde{r}_i(s,t) = \begin{cases} 1, & b_i \le s,t \le d_i, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Intuitively, an example of extended PRF is given in Figure 1, and it draws several square steps along the diagonal. The extended PRF does not introduce any extra information. In the context, $\widetilde{r}_{k,V}(s,t)$ is denoted as $\widetilde{r}(s,t)$ if no confusion occurs.

3.2. Vector Generation by Haar Decomposition

Here, we present the method to generate vectorizing representations. To extract features of PRF with different scales, one idea is to decompose a PRF on a system of basis that captures both local and global characteristics. Note that extended PRF has a structure of square steps along the diagonal because extended PRF is a piecewise constant-valued bivariant function.

This nature indicates an appropriate decomposition based on a system of non-continuous orthonormal basis, similar to Fourier decomposition. The system of Haar basis, a series of non-continuous orthonormal functions, is exactly a proper alternative to transform a PRF into a vector consisting of the coefficients of the basis.

Haar basis is a complete orthonormal basis in $L^2[0,1]$ [40]. The definition of the system of Haar functions is given as follows:

$$\begin{aligned} &\mathrm{har}_0(0,t) = 1, \quad 0 \le t \le 1, \\ &\mathrm{har}_n(k,t) = \begin{cases} \sqrt{2^{n-1}}, & \frac{2k-2}{2^n} \le t < \frac{2k-1}{2^n}, \\ -\sqrt{2^{n-1}}, & \frac{2k-1}{2^n} \le t \le \frac{2k}{2^n}, \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{6}$$

where $n = 1, 2, \cdots$, $k = 1, 2, 3, \cdots, 2^{n-1}$. The Haar basis can be extended to 2D in the sense of tensor product. Without losing generality, set $\Omega = [0,1]^2$ and arrange an order on the Haar functions so that $\{\mathrm{har}_n(k,t)\}_{n,k}$ is denoted as $\{\mathrm{har}_i(t)\}_i$ where $i = 0, 1, 2, \cdots, N$. The 2D Haar basis in the sense of tensor product is given by

$$\mathrm{har}_i(s,t) = \mathrm{har}_j(s)\mathrm{har}_k(t), \qquad j,k = 1,2,3,\cdots,N. \tag{7}$$

The 2D Haar system is a complete standard orthonormal basis on $L^2[0,1]^2$. When n in Equation (6) is relatively small, the corresponding Haar functions extract features with a large scale. And the local features are extracted as n increases. When n is large enough, subtle features can be captured. Therefore, it is reasonable to approximate an extended PRF by a finite number of Haar basis functions denoted as $\{\mathrm{har}_i(s,t)\}_{i=1}^{N^2}$ generated sequentially by Equation (7). We truncate extended PRF in the domain Ω, and normalize the domain to be $[0,1]^2$ so that extended PRF can be decomposed by Haar basis. The coefficients λ_i are obtained by

$$\lambda_i = \langle \widetilde{r}(s,t), \mathrm{har}_i(s,t) \rangle, \qquad i = 1, 2, \cdots, N^2. \tag{8}$$

And then, the coefficients λ_i are concatenated to be a vector in $\mathbb{R}^{N^2}$, i.e., $v = (\lambda_1, \lambda_2, \cdots, \lambda_{N^2})$, called the *feature vector* based on Haar basis decomposition. The algorithm to generate the feature vector from extended PRF is given in Algorithm 1. In the algorithm, the operator $.*$ means to multiply each entry of a matrix with the corresponding entry of the other matrix. The double *sum* means to add up all entries of the computed matrix. The time complexity of the algorithm is $O(2^{2(n+N)+1})$, where n is given in Equation (6), and the interval $[0,1]$ is divided into 2^N shares. In practice, we set $N = 8$ (256 shares) and $n = 5$ (1024 basis functions) to fast compute the feature vector.

Note that it is inevitable to lose some subtle information of a PRF, particularly the information close to the diagonal, because a limited number of Haar basis functions are used to decompose a PRF. Fortunately, this information does not represent prominent topological features.

Practically, feature vectors are sparse vectors with a rather high dimension. It is because, intuitively, high-frequency information is mainly concentrated near the diagonal, and a large number of high-frequency basis functions do not capture any information. Before feeding data to classifiers, one can use the technique of out-of-sample mapping in supervised manifold learning to obtain low dimensional vectors such that the training time is saved. Therefore, a layer of dimensionality reduction is adopted in the classification framework exhibited in Section 5.

Algorithm 1: Generation of the Feature Vector

Input: A barcode $\{b_i, d_i\}_{i\in I}$, the bound B, an integer n for Haar basis functions, and an integer N to split $[0,1]$ into 2^N shares.

1. Generate the discrete extended PRF according to Equation (4) and (5), truncate it in $[0,B]^2$, and normalize the domain to be $[0,1]^2$. The discrete extended PRF is stored in the form of matrix with its size $2^N \times 2^N$, denoted as z;
2. Generate 2^n discrete 1D Haar basis functions according to Equation (6), stored in a matrix $har1D$ with its size $2^n \times 2^N$;

```
Initiate v = zeros(2^{2n}, 1);
Initiate index = 1;
for i = 1 to 2^n do
    for j = 1 to 2^n do
        har2D = har1D(i,:)^T * har1D(j,:);
        v(index) = (1/2^N)^2 * sum(sum(har2D.*z));
        index = index + 1;
    end
end
```

Output: The feature vector v.

4. Theoretical Guarantee: Stability

In this section, we theoretically prove the stability of the generated vectorizing representation to tiny noise. Intuitively, the core of the stability of the Haar feature vector shows that the perturbation on the generated feature vector is correspondingly small if the perturbation on the source data is small. Due to the stability of barcode in [8], in some conditions, it shows that perturbations on the vector can be controlled by those on the barcode. In mathematical literature, for the norm of the difference between the original feature vector and the one with perturbation, there is an upper bound given by a distance between the original barcode and the perturbed one.

4.1. Distance of Barcodes

To evaluate measure similarity between two barcodes, one of alternative distance is p-Wasserstein distance, and the stability result is also obtain under the Wasserstein distance for a reasonably large class of function in [9], which guarantees the robustness of barcode in the presence of noise. Intuitively, to measure the differences between two barcodes, which can be seen as the sets of intervals, a matching is needed between them. There are two cases for the intervals in a barcode: the matched intervals and the others not being matched.

Given two barcodes denoted as $Bc, Bc^{'}$ together with index sets I, and $I^{'}$, respectively, let l_i be the interval in a barcode, i.e., $l_i := (b_i, d_i)$ and the differences of matched and non-matched intervals are given by

$$\begin{aligned} &\|l_i - l_j\|_\infty := \max\{|b_i - b_j|, |d_i - d_j|\},\ i \in I, j \in I^{'}, \\ &\|l_i\|_\infty := \frac{d_i - b_i}{2}. \end{aligned} \tag{9}$$

A matching φ is a binary relation between Bc and $Bc^{'}$ such that any interval in Bc and $Bc^{'}$ matches at most one pair. A matching can be seen as a subset of $I \times I^{'}$. Define M and $M^{'}$ to be the index set of matched intervals in Bc and $Bc^{'}$, respectively. The $p-$Wasserstein distance $W_p(Bc, Bc^{'})$ between the two barcodes Bc and $Bc^{'}$, which intuitively summarizes all the differences under the matching that makes the differences minimal, is defined as

$$\inf_{\varphi} \left\{ \sum_{i \in M} \|l_i - l^{'}_{\varphi(i)}\|^p_\infty + \sum_{j \in I \setminus M} \|l_j\|^p_\infty + \sum_{k \in I^{'} \setminus M^{'}} \|l^{'}_k\|^p_\infty \right\}^{1/p}, \tag{10}$$

where p is a positive real number.

4.2. 1-Wasserstein Stability

At the beginning of the proof of stability, some assumptions are proposed: (1) the number of intervals (b_i, d_i) is limited, i.e., $|I| \leq L$ where L is a positive integer; (2) the domain Ω is bounded, i.e., $\Omega = [0, B] \times [0, B]$, where B is given by a finite constant; (3) the perturbation on the barcode is small.

Given a barcode Bc, the corresponding PRF $r(s,t)$ and feature vector $v \in \mathbb{R}^{N \times N}$, a small perturbation is exerted on the barcode to produce the barcode $Bc^{'}$, PRF $r^{'}(s,t)$ and feature vector $v^{'}$. The p-Wasserstein distance of Bc and $Bc^{'}$ is given by $W_p(Bc, Bc^{'})$. Therefore, suppose that φ is the exact matching map which reaches the infimum of $W_p(Bc, Bc^{'})$ in Equation (10). Because the perturbation is subtle according to assumption (3), that is, one can assume $W_1(Bc, Bc^{'}) \leq 1$. The following theorem holds.

Theorem 4.1. *Given a barcode Bc and the corresponding perturbed barcode $Bc^{'}$, $r(s,t)$ and $r^{'}(s,t)$ are the PRFs in $L^p(\Omega)$, respectively. If $W_1(Bc, Bc^{'}) \leq 1$, then*

$$\|r - r^{'}\|^p_p \leq 2^{(p-1)(2L-1)}(2B+1)W_1(Bc, Bc^{'}), \tag{11}$$

where $B < \infty$ is given by the bounded domain $\Omega = [0, B] \times [0, B]$, and L is the largest number of intervals contained in a barcode.

Proof. See Appendix. □

For extended PRF, there is a corollary from Theorem 4.1.

Corollary 1. *In the same conditions as Theorem 4.1 given, $\widetilde{r}(s,t)$ and $\widetilde{r}^{'}(s,t)$ are the extended PRFs in $L^p(\Omega)$. Then we have*

$$\|\widetilde{r} - \widetilde{r}^{'}\|^p_p \leq 2^{(p-1)(2L-1)+p}(2B+1)W_1(Bc, Bc^{'}), \tag{12}$$

where the notations L and B follow Theorem 4.1.

Proof. See Appendix. □

And then the 1-Wasserstein stability of the feature vector is true by the following theorem.

Theorem 4.2 (Stability of Feature Vector)**.** *Given a barcode Bc and the corresponding perturbed barcode $Bc^{'}$, v and $v^{'}$ are the feature vectors generated by decomposing the extended PRFs, produced by Bc and $Bc^{'}$, on a finite series of Haar basis $\{har_i(s,t)\}_{i=1}^{N^2}$, respectively. If $W_1(Bc, Bc^{'}) \leq 1$, then*

$$\|v - v^{'}\|^2_2 \leq 2^{2L+1}(2B+1)W_1(Bc, Bc^{'}), \tag{13}$$

where $B < \infty$ is given by the bounded domain $\Omega = [0, B] \times [0, B]$, and L is the largest number of intervals contained in a barcode.

Proof. See Appendix. □

5. Experiments

In this section, we exhibit the technique of dimensionality reduction (DR) through out-of-sample mapping in supervised manifold learning practically used on data sets. And then, we discuss the determination of some parameters in generating the feature vectors, i.e., the bound B in Equation (3), and the appropriate dimension for the DR layer. Then classification experiments were done in a data set of random images of Brownian motion [41], which can be seen as scalar fields on the grid, a data set of a 2D dynamical system [25] and nine time-series data sets of multi-source signals [42], which are essentially the point clouds embedded in Euclidean space. Accuracy on different classifiers was obtained, and the performance of the feature vectors was compared with state-of-the-art kernel methods and persistence images (PIs). Finally, to clarify the effect of vectorizing process and the layer of DR, the ablation study was conducted. In the experiments, the barcodes of Vietoris-Rips filtration were computed via the Python package *Ripser* [43]. And Python package *dionysus 2* [44] was used to compute the 1-Wasserstein distance between two barcodes.

5.1. A Layer of Dimensionality Reduction

To reduce time of training classifiers and preserve or even improve classification accuracy, a dimensionality reduction technique based on supervised manifold learning is adopted to deal with high dimensional data. We employ the method proposed in [45], which is multi-output kernel ridge regression for out-of-sample mapping in supervised manifold learning.

Before choosing the DR method of out-of-sample mapping in supervised manifold learning, we attempted commonly used unsupervised DR methods, such as multidimensional scaling [46], and unsupervised manifold learning methods, like isometric feature mapping [47], but found the classification accuracy after DR dropped a lot, compared with that without DR. Therefore, we attempted supervised DR method based on manifold learning technique, kernel ridge regression for out-of-sample mapping proposed in [45], in which the category labels are used to improve the embedding of the training points.

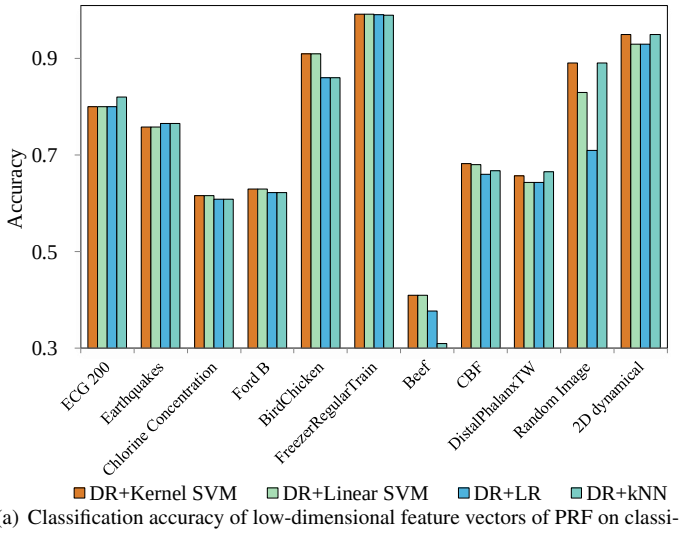
The supervised approach of manifold learning utilizes the labels of training data set to compute a projection of training data

for the subsequent classification task. To increase inter-class dissimilarity and to decrease intra-class dissimilarity, a supervised variant of Isomap based on a hierarchical agglomeration of the components is obtained from training labels. Furthermore, to evaluate the classification accuracy of a classifier on a test data set, an out-of-sample embedding method based on multi-output kernel ridge regression is used, which projects the test data into the embedding space. Note that before classifying, the class labels of the test data remain hidden. More details are introduced in [45].

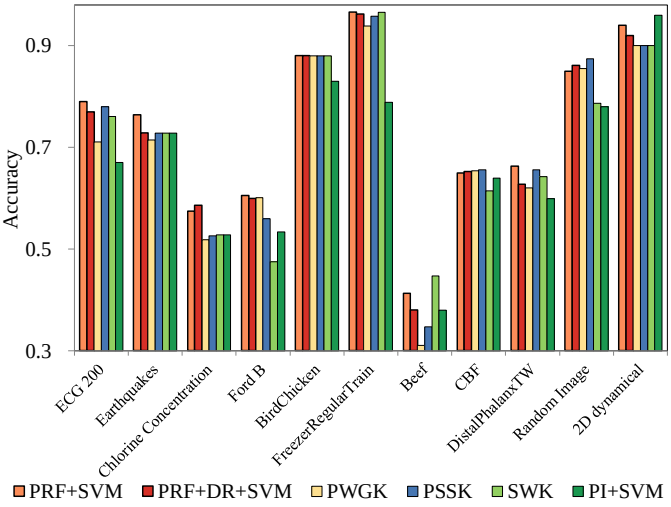
5.2. Evaluation

Determination of Parameters: In practice, the topological features of PH distribute in a bounded domain. PRF is considered as an element in L^2 space by choosing the weighted function $\phi(s,t)$ shown in Equation (3). Equivalently, PRF is limited in the domain $\Omega = [0,B]^2$. The problem arises to select an appropriate bound value B in the classification tasks. A reasonable alternative to determine the bound B is to find the maximum of death indices of barcodes of training data, i.e., $B = \max_{j\in train} \max_{i\in I_j}(d_i^j)$, and then to truncate bars with B in barcodes of test data. However, for few data sets in which a certain bar emerges with a relatively large death index, regarded as an outlier, the bound B is computed unreasonably large, which may dilute detailed features that Haar basis can capture. To avoid this, one can roughly assume that the sequence of birth and death indices follows a certain distribution, such as a normal distribution or an exponential distribution. Precisely, let $\mathbf{x} = (b_1, d_1, \cdots, b_i, d_i, \cdots)$ be a sequence. Fit the sequence $\mathbf{x}$ by a distribution $\varphi(t)$, and B is determined by $\inf_y\{y : \int_{-\infty}^{y} \varphi(t)dt \geq \delta\}$, where δ is a constant in $[0,1]$. In this way, one can obtain a reasonable value of bound B. Note that it is an alternative to determine a reasonable value of bound, not a rigorous assumption that the sequence follows the distribution. By observing experimental barcode data, we computed the bound value B using exponential distributions with $\delta = 0.99$ on a few data sets, and on most of data sets, B was determined by the maximum of death indices of barcodes. To test the effect of the selection of bound B on classification accuracy, we changed the bound in the parameter interval $[0.9B, B]$, where B was determined by the maximum of death indices , and the classification accuracy fluctuates in a small range, about 0.2% to 2.0%. If some of topological features are truncated, the classification accuracy will be affected. And when the bound is set in the interval $[B, 1.1B]$, then the classification accuracy has no significant change.

Because feature vectors are sparse, the basic idea is to look for a dimension in which the information of the vectors can be preserved as much as possible. To determine the appropriate dimension before the process of dimensionality reduction, principal component analysis (PCA) based on singular value decomposition is used. Concretely, in the implementation of PCA, when singular value decomposition is performed on the covariance matrix, the eigenvalue matrix Λ is obtained. Projection error which measures loss of information can be estimated by $\varepsilon(k) = 1 - \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{j=1}^{n}\lambda_j}$, where $\{\lambda_i\}$ is a sequence of eigenvalues with descending order and k is the reduced dimension. Let σ be tolerance. The appropriate dimension k_{app} is given by $\inf\{k \in \mathbb{Z}^+ : \varepsilon(k) \leq \sigma,\ k \leq n\}$. With k_{app} as an estimated reduced dimension, the DR layer introduced in Section 5.1 maps the labeled vectors to a low dimensional vector space. In our experiments, σ was set to be 0.01.



(a) Classification accuracy of low-dimensional feature vectors of PRF on classifiers of kernel SVM, linear SVM, LR, and kNN, respectively.



(b) The classification accuracy of feature vectors with and without DR on kernel SVM classifier, and classification accuracy of PWGK, PSSK, SWK and PI, each of which is adjusted to have the best performance.

Fig. 2. Classification accuracy on eleven data sets: The first nine data sets are of time-series for multi-source signals [42]; the tenth data set is of random images of Brownian motion [41], and the eleventh data set is of a 2D dynamical system in [25].

Comparison: To evaluate the performance of feature vectors for extracting topological patterns, classification tasks were performed on eleven data sets. To verify that the vectors can catch the topological patterns of data sets, state-of-the-art kernel methods for persistence diagrams, including PWGK, PSSK, and SWK, and PIs were used to compare with the feature vectors. To obtain barcodes, the H_0 barcode was computed based on lower-star filtration of the scalar function on a data set of random images. For a time series of multi-source signals, time-

delay embedding in [48] was adopted to transform a time series into a point cloud in Euclidean space, such that Vietoris-Rips (V-R) filtration was generated to compute H_1 barcodes. In the data set of a 2D dynamical system, V-R filtration with respect to the 2D Euclidean metric was produced to obtain H_1 barcodes.

In the pipeline of generating the feature vectors of PRF, 2^{10} Haar basis functions are used to extract the coefficients, i.e., $n = 5$ in Equation (6). Meanwhile, because the feature vectors are suitable to be applied on different classifiers, four classifiers, consisting of k nearest neighbors (kNN), logistic regression (LR), kernel support vector machine (kernel SVM) and linear support vector machine (linear SVM), are used to form a layer for classifying the input feature vectors. The parameters of LR are defaulted, while in kNN, linear SVM, and kernel SVM, the hyperparameters are selected by cross-validation to ensure the best performance. Furthermore, in the process of classification using kernel methods, the hyperparameters of kernels are adjusted to reach the best classification performance. For PIs, images are produced with their resolution 10×10 and Gaussian $\sigma = 10^{-3}$.

For the performance of feature vectors on the classifiers kernel SVM, linear SVM, LR, and kNN, the results are shown in Figure 2(a). The low dimensional feature vectors obtained by DR have good performance on the classifiers of kernel SVM for most of data sets (8 in 11), while LR and kNN may not distinguish those low dimensional vectors well on some data sets. In Figure 2(b), the comparison results, i.e., the performance of the proposed vectorizing representations with and without DR on kernel SVM classifier, the performance of kernel methods for barcode (PWGK, PSSK, and SWK), and the performance of PIs are shown. On the time-series data sets Beef, CBF [42] and random images, the kernel methods perform best. PIs have the best classification result on the data set of 2D dynamical system. On the rest of data sets (7 in 11), feature vectors of PRF (without DR, shown in orange) perform the best. Overall, feature vectors can capture prominent topological features of data such that a nice performance is achieved on the classifiers. And, for the comparison of time cost of the proposed method and its competitors, it costs less time to produce feature vectors for classification than kernel methods. The proposed classification pipeline costs about 2 hours, but kernel methods cost about 15 hours in total eleven data sets for all procedures. Techniques of parallel computing can be adopted to compute the vectors of a collection of data.

Furthermore, we notice that on most of data sets, the performance of the feature vectors without DR on kernel SVM classifier is better than that of the vectors with DR. To clarify the effect of DR layer in both aspects of classification accuracy on the rest of classifiers and the time cost of training classifiers with low and high-dimensional vectors, we do ablation study by removing the DR layer. Moreover, the entire framework will be ablated by using the distance matrix with respect to Wasserstein distance equipped on barcodes as the input of kNN classifier in order to show the effect of classification by using the proposed vectoring representations of barcodes.

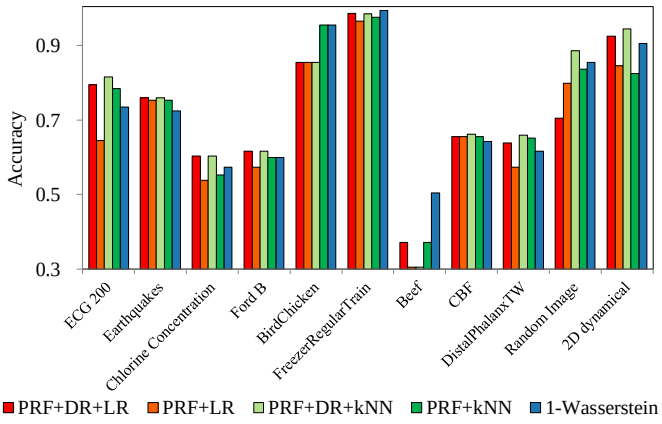


Fig. 3. Classification results for ablation study on eleven data sets: classification accuracies of using feature vectors with DR, feature vectors without DR on LR and kNN classifiers, and a distance matrix of 1-Wasserstein distance on kNN classifier are illustrated.

5.3. Ablation Study

To investigate the effects of these two components in the process, an ablation study was done. First, the layer of DR is ablated to show its effectiveness for classifying data with different labels. Then, the whole layer of vectorization is ablated to validate the effectiveness of feature vectors for classification tasks. The classification after ablation is implemented by kNN with the distance matrix with respect to the 1-Wasserstein distance of barcode.

The purpose of DR is to save time of classification and to preserve and even improve classification accuracy. In the experiments, the appropriate reduced dimension was estimated via PCA. As shown in Figure 3, with the layer of DR, the classification accuracies of eight data sets on kNN and LR classifier are improved. For instance, noticeable improvement occurs on the data sets of ECG200 and of 2D dynamical system. This means that it is feasible to improve the accuracy of classification in some data sets as well as to obtain dense low-dimensional feature vectors.

Table 1. Total time cost of classification using feature vectors with and without DR: Four data sets with a relatively large training and test set are chosen to measure the total time cost of classifying the vectors with reduced dimension and the vectors with 1024 dimensions on four classifiers. The time of selecting optimal hyperparameters is included.

Data set	Training/test data	Reduced dimension	Time cost (s) with or without DR
FordB	3636/810	34	102/891
Chlorine Concentration	467/3840	23	15/143
Distal PhalanxTW	600/276	29	5/24
Random Image	800/160	4	5/89

In addition, comparing the proposed method with and with-

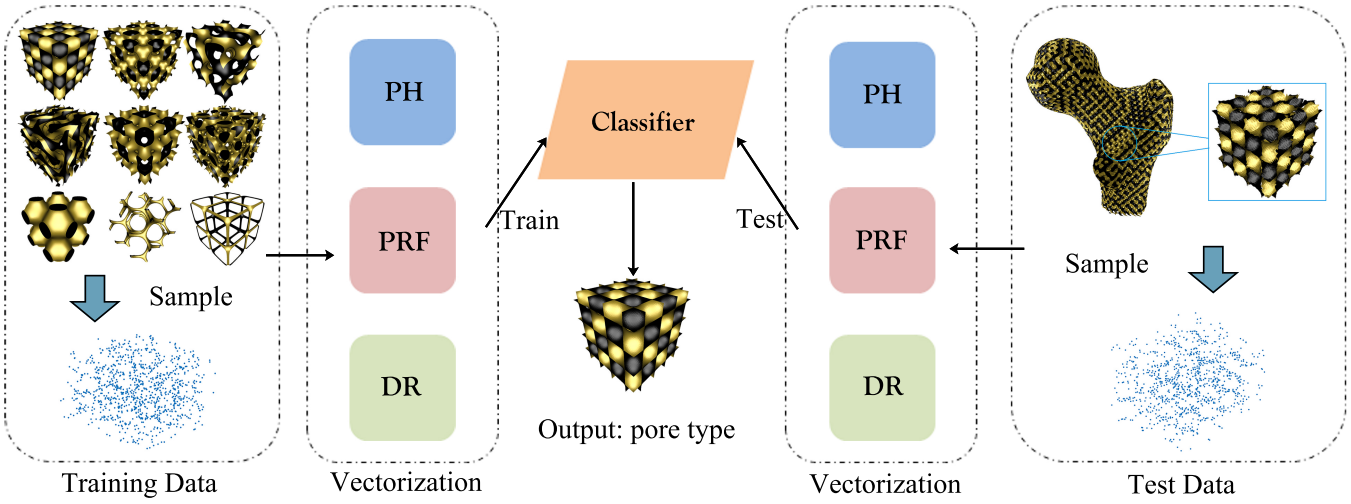


Fig. 4. Pipeline of classifying data set of porous models using feature vectors with the DR technique: an example of nine types of TPMS porous models are exhibited in the column *training data*, and an example of test data sampled from a porous model of bone is shown in the column *test data*.

out DR on kNN classifiers and the method of directly classifying barcodes using 1-Wasserstein distance matrix, Figure 3 also shows the improvement of the proposed method with and without DR on most of data sets for the extraction of topological patterns. On eight data sets, it improves the classification accuracy by using the proposed vectorizing representation and the DR layer. On five data sets such as ECG200 and Earthquakes, the classification accuracy improves by using vectorizing representation in the classification on kNN classifier. However, there exist a few data sets, such as Beef, on which using 1-Wasserstein distance matrix on kNN classifier has a better performance. However, it spent much more time to compute 1-Wasserstein distance even through the efficient method of using geometry of persistence diagrams proposed in [49].

Although, as shown by bars in orange and in red in Figure 2(b), there is some loss of accuracy by using DR technique on SVM classifier, the advantage of DR layer is to reduce the training time of the classifiers, especially when the data set is large. Apparently, it costs much more time to train the classifiers with high dimensional vectors. As shown in Table 1, four data sets with a large training and test set are chosen to show the time saving after DR process. And the time cost was measured on a PC with Intel(R) Core(TM) i7-4790 CPU@3.60GHz×8. After DR process, the total time cost of classification is reduced by 5 to 15 times.

6. Application: Classification of Porous Structures

Porous scaffolds are widely used to engineer various human tissues virtually in tissue engineering and biomaterials [39]. To extract informative patterns from porous structures for the tasks of classification is an important and novel problem for material design. In this section, the generating method of porous structures, the details of training, test data and the approach of classification are introduced. Finally, the classification accuracy and the time cost of the proposed framework and other state-of-the-art methods are analyzed.

6.1. Porous Data Set

In this application, a data set of 3D porous models was produced by using TPMS to generate nine types of different porous structures. TPMS is a minimal surface with a mean curvature of zero, with periodicity in each direction of 3D space, which is popular for engineering porous models.

The Generation of TPMS in B-Spline Solid: We approximated the TPMS using a periodic nodal surface defined by a Fourier series [50],

$$\psi(\mathbf{r}) = \sum_k A_k cos[2\pi(\mathbf{h}_k \cdot \mathbf{r})/\lambda_k - P_k] = C, \tag{14}$$

where $\mathbf{r}$ is the location vector in the Euclidean space, A_k is the amplitude, $\mathbf{h}_k$ is the k^{th} lattice vector in the reciprocal space, λ_k is the wavelength of the period, P_k is the phase shift, and C is the threshold constant. And we set $C = 0$ in this application. In this application, marching tetrahedra (MT) algorithm is employed to extract the TPMS. In consideration of accuracy and storage of TPMS, the physical domain is divided into $100 \times 100 \times 100$ hexahedrons, and each hexahedron is further divided into 6 tetrahedrons. And then all of intersection triangles constitute a mesh approximating the iso-surface.

To generate a TPMS in the 3D mesh model, the B-spline solid needs to be approximated by hexahedral mesh model. We use a set of sampling points along each parametric direction according to the preset resolution. All these sampling points are calculated and mapped into the Cartesian space by the trivariate B-spline function. These sampling evaluation points constitute numerous hexahedron elements to approximately represent the B-spline solid. Every direction of parametric domain is evenly sampled to generate parametric coordinates of evaluation points. And then, the Cartesian coordinates of the 3D mesh vertices can be calculated by the trivariate B-spline function. The B-spline solid can be approximately represented by

hexahedron elements. Therefore, each hexahedron is further divided into 6 tetrahedrons, and the TPMS can be extracted by the MT algorithm mentioned above.

Training Data: The training data are produced by the following operations. First, the physical domain is set to $[x, x+2] \times [y, y+2] \times [z, z+2]$, and the variables x, y, z are sampled from $[-1, 1]$ using random sampling algorithm. Then, with different parameters in Equation (14), the discrete hexahedral physical domain and the iso-surface $\psi(x, y, z) = 0$ are determined, and the nine types of TPMS can be extracted by the MT algorithm. Finally, 100 models with its size $2\times2\times2$ are generated for each TPMS type. For each model, 1000 points are randomly sampled twice to produce a point cloud. In total, 1800 point clouds are generated as training data with labels.

Test Data: As for the test data, five porous models with nine types of TPMS are produced, namely, bone (balljoint), isis, moai, tooth, and venus (45 models in total). From each of the models, three cubes with their size $2 \times 2 \times 2$ are sampled. The samples do not contain the boundary of the models. And for each cube, three point clouds are randomly sampled with 1000 points. Therefore, 405 point clouds are obtained as test data. The labels (TPMS types) keep hidden before classification.

6.2. Classification and Results

The pipeline of classifying porous structures is given in Figure 4. For each point cloud in both training and test data set, the V-R filtration is built on each point cloud in 3D Euclidean space to compute a 1-dimensional persistence barcode via PH. Then, the vectorizing representation based on PRF is computed to obtain the feature vector. And alternatively, the layer of DR can be used to obtain the low-dimensional vector. Finally, the classifiers are trained with cross-validation and classification accuracy is obtained.

Table 2. Classification accuracy and time cost on the data set of porous models: 9-dimensional vectors were obtained through DR, and the hyperparameter C of linear SVM was selected to be 0.01 via cross validation. PIs was obtained by setting the resolution of 10×10 and $\sigma = 10^{-3}$ to achieve the best performance. Time cost for our methods and PIs consists of vectorization, training classifier, and testing. Time cost for 1-Wasserstein+kNN consists of computing distance matrix, training classifier, and testing. And time cost of kernel methods consists of training and testing.

Methods	Accuracy	Time Cost
PRF+DR+Linear SVM	**81.6%**	5.2min/3.0s/0.1s
PRF+Linear SVM	79.4%	5.2min/12.2s/0.2s
PWGK	64.9%	1.5h/0.6h
PSSK	71.6%	2.7h/1.1h
SWK	54.7%	2.5h/1.0h
1-Wasserstein+kNN	72.6%	12.0h/13.5s/0.2s
PIs+kernel SVM	66.4%	4.9min/5.6s/0.1s

As shown in Table 2, feature vectors with DR technique on the linear SVM classifier have the best performance. The layer of DR helps improve classification accuracy, and 9-dimensional vectors are obtained to represent topological patterns of porous models. Kernel methods and PIs do not perform well because it is likely to be unable to capture detailed topological features but to be disturbed by subtle topological noises in the test data. For time cost, the calculation of the proposed feature vectors is nearly as efficient as the generation of PIs. And we mention that it costs 1.0 min in DR procedure in our method. The kernel methods cost a large mount of time to train the classifier, and it is unacceptable to use the traditional 1-Wasserstein distance on kNN classifier for this classification task. To eliminate the concern that the nice performance of feature vectors with DR occurs because the decomposition based on a few number of Haar basis removes the topological noise of data, 2^{12} Haar basis functions ($n = 6$ in Equation 6) are adopted, compared with 2^{10} functions ($n = 5$). The classification accuracy is 80.6% on the classifier of linear SVM, which remains high. In general, the results on data sets of porous models show the potential to use PRF to extract topological patterns on data sets with prominent topological features for classification tasks.

7. Discussion and Conclusion

This paper introduces a novel vectorizing representation of PRF based on Haar basis decomposition on a bounded domain to extract topological patterns for the classification tasks of point clouds, which is one of the issues of interest to the computer graphics community. The generated vectorizing representation is proved to have 1-Wasserstein stability, which provides theoretical guarantee of the proposed method to deal with data with noise. Classification experiments in different data sets show its effectiveness. Meanwhile, it is shown that on the data set of porous models with topological noises, feature vectors of PRF perform the best. In the classification pipeline, we employed the DR technique of out-of-sample mapping in supervised manifold learning to reduce time of training classfiers and preserve or even improve classification accuracy in practice. It is more effective than the unsupervised DR methods we attempted. One can also attempt other DR methods that match the practical needs for specific data sets.

In essence, the proposed vectorizing representation of barcodes is used for classification tasks of point clouds, and the classification of porous models exhibited in Section 6 is essentially a task of 3D point cloud classification. In deep learning, the designed networks, such as PointNet [51], usually have stunning performance for point cloud classification, especially in the scenarios involving model semantics. However, the interpretability of the networks is still an urgent issue. In machine learning, kernel methods based on persistence diagrams, such as those we used to compare with the proposed methods, transform the classification of persistence diagrams into a classic non-linear classification problem by designing kernels. These methods have strong generalization ability. But finite-dimensional vectors can not be explicitly given, and it costs a large amount of time in practice to train the classifiers. As a hand-crafted feature extraction, the proposed method captures the informative features of PRF, and can be clearly explained. Meanwhile, the proposed method is shown to have good performance on the data sets with significant topological features, such as porous model data set. Unlike kernel methods, the fea-

ture vectors computed by the proposed method are directly input into classifiers, and it may reduce the generalizing power of the classification pipeline.

For future work, the pore categories on real-world porous material data sets are usually unknown. The proposed method transforms a point cloud of porous material into a finite-dimensional vector containing its topological information, such as connectivity and loops. The generated vectors can be used as a topological descriptor for material retrieval and classification. It will help researchers find porous materials with similar topology in application scenarios.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant Nos. 61872316, 61932018.

References

[1] Edelsbrunner, H, Letscher, D, Zomorodian, A. Topological persistence and simplification. Foundations of Computer Science 2000;28(4):454–463.

[2] Robins, V, Wood, PJ, Sheppard, A. Theory and algorithms for constructing discrete morse complexes from grayscale digital images. IEEE Transactions on Pattern Analysis and Machine Intelligence 2011;33(8):1646–1658.

[3] Rieck, B, Fugacci, U, Lukasczyk, J, Leitte, H. Clique community persistence: A topological visual analysis approach for complex networks. IEEE Transactions on Visualization and Computer Graphics 2018;24(1):822–831.

[4] Carlsson, GE. Topological pattern recognition for point cloud data. Acta Numerica 2014;23:289–368.

[5] Gabrielsson, RB, Ganapathisubramanian, V, Skraba, P, Guibas, LJ. Topology-aware surface reconstruction for point clouds. arXiv: Computational Geometry 2018;.

[6] Collins, AD, Zomorodian, A, Carlsson, GE, Guibas, LJ. A barcode shape descriptor for curve point cloud data. Computers & Graphics 2004;28(6):881–894.

[7] Ghrist, R. Barcodes: The persistent topology of data. Bulletin of the American Mathematical Society 2008;45(1):61–75.

[8] Cohen-Steiner, D, Edelsbrunner, H, Harer, J. Stability of persistence diagrams. Discrete & Computational Geometry 2007;37(1):103–120.

[9] Edelsbrunner, H, Harer, J. Computational Topology: An Introduction. American Mathematical Soc.; 2010.

[10] Cerri, A, Fabio, BD, Ferri, M, Frosini, P, Landi, C. Betti numbers in multidimensional persistent homology are stable functions. Mathematical Methods in The Applied Sciences 2013;36(12):1543–1557.

[11] Frosini, P, Landi, C. Persistent betti numbers for a noise tolerant shape-based approach to image retrieval. Pattern Recognition Letters 2013;34(8):863–872.

[12] Zomorodian, A, Carlsson, GE. Computing persistent homology. Discrete & Computational Geometry 2005;33(2):249–274.

[13] Robins, V, Turner, K. Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. Physica D: Nonlinear Phenomena 2016;334:99–117.

[14] Simon, CM, Kim, J, Lin, L, Martin, RL, Haranczyk, M, Smit, B. Optimizing nanoporous materials for gas storage. Physical Chemistry Chemical Physics 2014;16(12):5499–5513.

[15] Kalil, T, Wadia, C. Materials genome initiative for global competitiveness. https://www.mgi.gov/; 2011.

[16] Lee, Y, Barthel, S, Dlotko, P, Moosavi, SM, Hess, K, Smit, B. Quantifying similarity of pore-geometry in nanoporous materials. Nature Communications 2017;8(1):15396.

[17] Donatini, P, Frosini, P, Lovato, A. Size functions for signature recognition. Proceedings of Spie the International Society for Optical Engineering 1998;3454.

[18] Frosini, P, Landi, C. Size functions and formal series. Applicable Algebra in Engineering, Communication and Computing 2001;12(4):327–349.

[19] Pachauri, D, Hinrichs, C, Chung, MK, Johnson, SC, Singh, V. Topology-based kernels with application to inference problems in Alzheimer's disease. IEEE Transactions on Medical Imaging 2011;30(10):1760–1770.

[20] Carriere, M, Oudot, SY, Ovsjanikov, M. Stable topological signatures for points on 3D shapes. In: Computer Graphics Forum; vol. 34. Wiley Online Library; 2015, p. 1–12.

[21] Di Fabio, B, Ferri, M. Comparing persistence diagrams through complex vectors. In: International Conference on Image Analysis and Processing. Springer; 2015, p. 294–305.

[22] Adcock, A, Carlsson, E, Carlsson, G. The ring of algebraic functions on persistence bar codes. Homology, Homotopy and Applications 2016;18(1):381–402.

[23] Kališnik, S. Tropical coordinates on the space of persistence barcodes. Foundations of Computational Mathematics 2018;:1–29.

[24] Dong, Z, Lin, H, Zhou, C. Persistence B-spline grids: Stable vector representation of persistence diagrams based on data fitting. arXiv preprint arXiv:190908417 2019;.

[25] Adams, H, Emerson, T, Kirby, M, Neville, R, Peterson, C, Shipman, P, et al. Persistence images: A stable vector representation of persistent homology. Journal of Machine Learning Research 2017;18(1):218–252.

[26] Bubenik, P. Statistical topological data analysis using persistence landscapes. Journal of Machine Learning Research 2015;16(1):77–102.

[27] Reininghaus, J, Huber, S, Bauer, U, Kwitt, R. A stable multi-scale kernel for topological machine learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015, p. 4741–4748.

[28] Kusano, G, Fukumizu, K, Hiraoka, Y. Kernel method for persistence diagrams via kernel embedding and weight factor. Journal of Machine Learning Research 2018;18(189):1–41.

[29] Carriere, M, Cuturi, M, Oudot, S. Sliced wasserstein kernel for persistence diagrams. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org; 2017, p. 664–673.

[30] Pun, CS, Xia, K, Lee, SX. Persistent-homology-based machine learning and its applications – a survey. arXiv: Algebraic Topology 2018;.

[31] Cheah, CM, Chua, C, Leong, K, Chua, S. Development of a tissue engineering scaffold structure library for rapid prototyping. part 1: Investigation and classification. International Journal of Advanced Manufacturing Technology 2003;21(4):291–301.

[32] Cheah, CM, Chua, C, Leong, K, Chua, S. Development of a tissue engineering scaffold structure library for rapid prototyping. part 2: Parametric library and assembly program. International Journal of Advanced Manufacturing Technology 2003;21(4):302–312.

[33] Schroeder, C, Regli, WC, Shokoufandeh, A, Sun, W. Computer-aided design of porous artifacts 2005;37(3):339–353.

[34] Cai, S, Xi, J. A control approach for pore size distribution in the bone scaffold based on the hexahedral mesh refinement. Computer-Aided Design 2008;40(10-11):1040–1050.

[35] Kou, ST, Tan, ST. An approach of irregular porous structure modeling based on subdivision and nurbs. Computer-Aided Design and Applications 2013;10(2):355–369.

[36] You, Y, Kou, S, Tan, S. A new approach for irregular porous structure modeling based on centroidal voronoi tessellation and B-spline. Computer-Aided Design and Applications 2016;13(4):484–489.

[37] Feng, J, Fu, J, Shang, C, Lin, Z, Li, B. Porous scaffold design by solid T-splines and triply periodic minimal surfaces. Computer Methods in Applied Mechanics & Engineering 2018;336:333–352.

[38] Yang, N, Zhou, K. Effective method for multi-scale gradient porous scaffold design and fabrication. Materials Science & Engineering C 2014;43:502–505.

[39] Yoo, D. Porous scaffold design using the distance field and triply periodic minimal surface models. Biomaterials 2011;32(31):7741–7754.

[40] Lagarias, JC, Wang, Y. Haar type orthonormal wavelet bases in R2. Journal of Fourier Analysis and Applications 1995;2(1):1–14.

[41] Kimura, M, Obayashi, I, Takeichi, Y, Murao, R, Hiraoka, Y. Non-empirical identification of trigger sites in heterogeneous processes using persistent homology. Scientific Reports 2018;8(1):3553–3553.

[42] Dau, HA, Keogh, E, Kamgar, K, Yeh, CCM, Zhu, Y, Gharghabi, S, et al. The ucr time series classification archive. 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

[43] Tralie, C, Saul, N, Bar-On, R. Ripser.py: A lean persistent homology library for python. The Journal of Open Source Software 2018;3(29):925. URL: https://doi.org/10.21105/joss.00925. doi:10.21105/joss.00925.
[44] Morozov, D. Dionysus 2 documentation. https://www.mrzv.org/software/dionysus2/; 2019.
[45] Orsenigo, C, Vercellis, C. Kernel ridge regression for out-of-sample mapping in supervised manifold learning. Expert Systems With Applications 2012;39(9):7757–7762.
[46] Borg, I, Groenen, PJF. Modern multidimensional scaling: Theory and applications. Journal of Educational Measurement 2003;40(3):277–280.
[47] Tenenbaum, JB, De Silva, V, Langford, J. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290(5500):2319–2323.
[48] Seversky, LM, Davis, S, Berger, M. On time-series topological data analysis: New data and opportunities. In: Computer Vision & Pattern Recognition Workshops. 2016,.
[49] Kerber, M, Morozov, D, Nigmetov, A. Geometry helps to compare persistence diagrams. Journal of Experimental Algorithmics (JEA) 2017;22:1–4.
[50] Gandy, PJF, Bardhan, S, Mackay, AL, Klinowski, J. Nodal surface approximations to the ja:math and I-WP triply periodic minimal surfaces. Chemical Physics Letters 2001;336(3):187–195.
[51] Charles, RQ, Su, H, Kaichun, M, Guibas, LJ. Pointnet: Deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, p. 77–85.
[52] Rudin, W, et al. Principles of mathematical analysis; vol. 3. McGraw-hill New York; 1976.

Appendix A: Proof of Theorem 4.1

At first, two inequations in $L^p(\mathbb{R}^2)$ are proved in Lemma 1.

Lemma 1. *1. Given $f, g \in L^p(\mathbb{R}^2)$, we have*

$$\|f+g\|_p^p \le 2^{p-1}\left(\|f\|_p^p + \|g\|_p^p\right), \text{ where } p \ge 1. \tag{15}$$

2. Given $f_1, f_2, \cdots, f_n \in L^p(\mathbb{R}^2)$, where $n < \infty$, it follows that

$$\|\sum_{i=1}^{n} f_i\|_p^p \le 2^{(p-1)(n-1)} \sum_{i=1}^{n} \|f_i\|_p^p, \text{ where } p \ge 1. \tag{16}$$

Proof. To show Equation (15), that is,

$$\int_{\mathbb{R}^2} |f+g|^p \ dxdy \le 2^{p-1}\left(\int_{\mathbb{R}^2} |f|^p \ dxdy + \int_{\mathbb{R}^2} |g|^p \ dxdy\right) \tag{17}$$

holds, we show that for $p \ge 1$,

$$|f+g|^p \le 2^{p-1}(|f|^p + |g|^p). \tag{18}$$

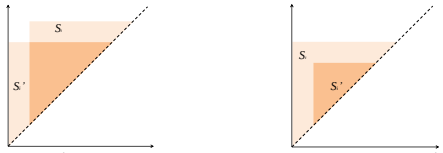
It is obvious when $p = 1$. Then, it is easy to show that $y(x) = x^p$ is convex over $\mathbb{R}^+$, i.e., for $0 < x_1 \le x_2$, we have $y((x_1+x_2)/2) \le (y(x_1) + y(x_2))/2$. Therefore, for $p > 1$, we have

$$\left|\frac{1}{2}f + \frac{1}{2}g\right|^p \le \left|\frac{1}{2}|f| + \frac{1}{2}|g|\right|^p \le \frac{1}{2}\left(|f|^p + |g|^p\right). \tag{19}$$

It shows that Equation (18) holds. Since $f, g \in L^p(\mathbb{R}^2)$, Equation (15) holds by integrating on both sides, which suggests that $f + g \in L^p(\mathbb{R}^2)$.

Moreover, Equation (16) follows by repeatedly using the Equation (15). □

Proof of Theorem 4.1:



(a) S_i and $S_i^{'}$ overlap but one does not contain the other. (b) S_i contains $S_i^{'}$.

Fig. 5. Illustraion of two cases of the relationship between S_i and $S_i^{'}$.

Proof. For $r(s,t), r^{'}(s,t) \in L^p(\Omega)$, the p-norm of the difference of $r(s,t)$ and $r^{'}(s,t)$ is to be evaluated. It follows that

$$\|r - r^{'}\|_p^p = \int_\Omega \left|\sum_{i\in I} r_i - \sum_{j\in I^{'}} r_j^{'}\right|^p \ dsdt. \tag{20}$$

r_i's and $r_j^{'}$'s are paired according to the matching φ. We have

$$\begin{aligned}\|r - r^{'}\|_p^p &= \int_\Omega \left|\sum_{i\in M} r_i - r_{\varphi(i)}^{'} + \sum_{j\in I\setminus M} r_j + \sum_{k\in I^{'}\setminus M^{'}} r_k^{'}\right|^p \ dsdt \\ &\le \int_\Omega \left(\sum_{i\in M} |r_i - r_{\varphi(i)}^{'}| + \sum_{j\in I\setminus M} |r_j| + \sum_{k\in I^{'}\setminus M^{'}} |r_k^{'}|\right)^p \ dsdt.\end{aligned} \tag{21}$$

According to the assumption (2) that the number of intervals in Bc and $Bc^{'}$ is less than L, the total number of pairs and single elements determined by φ is less than $2L$, that is, $|M| + |I\setminus M| + |I^{'}\setminus M^{'}| \le 2L$. Hence, it follows by Equation (16) in Lemma 1 that

$$\begin{aligned}\|r - r^{'}\|_p^p &\le 2^{(p-1)(2L-1)}\left(\sum_{i\in M}\int_\Omega |r_i - r_{\varphi(i)}^{'}|^p \ dsdt\right. \\ &\quad \left. + \sum_{j\in I\setminus M}\int_\Omega |r_j|^p \ dsdt + \sum_{k\in I^{'}\setminus M^{'}}\int_\Omega |r_k^{'}|^p \ dsdt\right) \\ &= 2^{(p-1)(2L-1)}\left(\sum_{i\in M}\|r_i - r_{\varphi(i)}^{'}\|_p^p + \sum_{j\in I\setminus M}\|r_j\|_p^p + \sum_{k\in I^{'}\setminus M^{'}}\|r_k^{'}\|_p^p\right).\end{aligned} \tag{22}$$

Let $\varepsilon_i = \max\{|b_i - b_{\varphi(i)}^{'}|, |d_i - d_{\varphi(i)}^{'}|\}$ if the interval is paired by the matching φ to achieve the infimum, $\varepsilon_j = \|l_j\|_\infty$, and $\varepsilon_k = \|l_k^{'}\|_\infty$. Hence, the 1-Wasserstein distance is able to be expressed by

$$W_1(Bc, Bc^{'}) = \left\{\sum_{i\in M}\varepsilon_i + \sum_{j\in I\setminus M}\varepsilon_j + \sum_{k\in I^{'}\setminus M^{'}}\varepsilon_k\right\}. \tag{23}$$

The persistence of (b_i, d_i) is $d_i - b_i \ge 0$. We denote it as per_i. Since $W_1(Bc, Bc^{'}) \le 1$, it follows by Equation (11) in the paper that $\varepsilon_* \le 1$, where $*$ represents subscripts i, j and k given above.

For the intervals paired according to the matching φ, the nonzero valued region S_i of $r_i(s,t)$, induced by $(b_i, d_i) \in Bc$, has different relevant relation with the region $S_i^{'}$ of $r^{'}(s,t)$, induced by $(b_{\varphi(i)}^{'}, d_{\varphi(i)}^{'})$. In general, three cases are taken into consideration.

Case 1: S_i and $S_i^{'}$ overlap but one does not contain the other.

That is, we have

$$\|r_i - r_{\varphi(i)}^{'}\|_p^p = \int_{T_1\cup T_2} |1|^p \ dsdt = \text{Area}(T_1 \cup T_2), \tag{24}$$

where T_1 and T_2 are two trapezoids formed by $(S_i \cup S_i^{'})\backslash(S_i \cap S_i^{'})$. Without loss of generality, Area$(T_1 \cup T_2)$ is computed according to the case shown in Figure 5(a). It is obtained by directly computing the areas that

$$\text{Area}(T_1) = \frac{1}{2}\left[per_{\varphi(i)}^{'} + per_i - (d_{\varphi(i)}^{'} - d_i)\right](b_{\varphi(i)}^{'} - b_i), \tag{25}$$

and

$$\text{Area}(T_2) = \frac{1}{2}\left[per_{\varphi(i)}^{'} + per_i - (b_{\varphi(i)}^{'} - b_i)\right](d_{\varphi(i)}^{'} - d_i). \tag{26}$$

And then we have

$$\begin{aligned}
&\text{Area}(T_1 \cup T_2) = \text{Area}(T_1) + \text{Area}(T_2)\\
&\le \frac{1}{2}\varepsilon_i\left[2per_{\varphi(i)}^{'} + 2per_i - (d_{\varphi(i)}^{'} - d_i) - (b_{\varphi(i)}^{'} - b_i)\right]\\
&\le \frac{1}{2}\varepsilon_i\left|2per_{\varphi(i)}^{'} + 2per_i + d_i - d_{\varphi(i)}^{'} + b_i - b_{\varphi(i)}^{'}\right|\\
&\le \frac{1}{2}\varepsilon_i\left(2per_{\varphi(i)}^{'} + 2per_i + |d_i - d_{\varphi(i)}^{'}| + |b_i - b_{\varphi(i)}^{'}|\right)\\
&\le (2per + \varepsilon_i)\,\varepsilon_i,
\end{aligned} \tag{27}$$

where per represents the largest persistence in Bc and $Bc^{'}$, i.e., $per = \max\{\max_{i\in I}(d_i - b_i), \max_{j\in I'}(d_i^{'} - b_i^{'})\}$. Since $\varepsilon_i \le 1$ mentioned above,

$$\|r_i - r_{\varphi(i)}^{'}\|_p^p \le 2per\cdot\varepsilon_i + \varepsilon_i^2 \le (2per+1)\varepsilon_i. \tag{28}$$

Case 2: For S_i and $S_i^{'}$, one contains the other.

Without loss of generality, the case is considered as shown in Figure 5(b).

$$\begin{aligned}
\|r_i - r_{\varphi(i)}^{'}\|_p^p &= \int_{S_i\backslash S_i^{'}} |1|^p\, dsdt\\
&= \text{Area}(S_i\backslash S_i^{'}) = \frac{1}{2}(per_i^2 - per_{\varphi(i)}^{'2})\\
&= \frac{1}{2}(per_i + per_{\varphi(i)}^{'})(per_i - per_{\varphi(i)})\\
&\le \frac{1}{2}(per_i + per_{\varphi(i)}^{'})\left|d_i - d_{\varphi(i)}^{'} + b_{\varphi(i)}^{'} - b_i\right|\\
&\le \frac{1}{2}(per_i + per_{\varphi(i)}^{'})\left(|d_i - d_{\varphi(i)}^{'}| + |b_{\varphi(i)}^{'} - b_i|\right)\\
&\le 2per\cdot\varepsilon_i \le (2per+1)\varepsilon_i.
\end{aligned} \tag{29}$$

where per has the same meaning as Equation (27).

Case 3: S_i and $S_i^{'}$ do not overlap.

Notice that in this case we have

$$\begin{aligned}
\max\{|b_i - b_{\varphi(i)}^{'}|, |d_i - d_{\varphi(i)}^{'}|\} &\ge \max\{per_i, per_{\varphi(i)}^{'}\}\\
&\ge \frac{d_i - b_i}{2} + \frac{d_{\varphi(i)}^{'} - b_{\varphi(i)}^{'}}{2},
\end{aligned} \tag{30}$$

i.e.,

$$\|l_i - l_{\varphi(i)}^{'}\|_\infty \ge \|l_i\|_\infty + \|l_{\varphi(i)}^{'}\|_\infty, \tag{31}$$

which means that it is better if l_i and $l_{\varphi(i)}^{'}$ are not paired. Therefore, case 3 does not happen if the matching φ is the one to reach the infimum.

For the intervals not paired via the matching φ, it follows that

$$\int_\Omega |r_*|^p\, dsdt = \int_{S_*} |1|^p\, dsdt = \frac{1}{2}per_*^2, \tag{32}$$

As defined in Equation (9), it follows by $per_* = 2\varepsilon_*$ that

$$\int_\Omega |r_*|^p\, dsdt = per_*\cdot\varepsilon_* \le (2per+1)\varepsilon_*, \tag{33}$$

where $*$ represents the subscripts $j \in I\backslash M$ and $k \in I^{'}\backslash M^{'}$ in the corresponding situation, and per is the largest persistence in the barcodes.

Eventually, it follows by Equation (22) that

$$\begin{aligned}
\|r - r^{'}\|_p^p &\le 2^{(p-1)(2L-1)}\left(\sum_{i\in M}\|r_i - r_{\varphi(i)}^{'}\|_p^p + \sum_{j\in I\backslash M}\|r_j\|_p^p + \sum_{k\in I^{'}\backslash M^{'}}\|r_k^{'}\|_p^p\right)\\
&\le 2^{(p-1)(2L-1)}(2per+1)\left(\sum_{i\in M}\varepsilon_i + \sum_{j\in I\backslash M}\varepsilon_j + \sum_{k\in I^{'}\backslash M^{'}}\varepsilon_k\right)\\
&= 2^{(p-1)(2L-1)}(2per+1)W_1(Bc, Bc^{'}).
\end{aligned} \tag{34}$$

Because the domain is restricted on $[0,B]\times[0,B]$, in a barcode, it is equivalent to consider the intervals satisfying $0 \le b_i \le d_i \le B$, to truncate the intervals in which $b_i < B$ and $d_i > B$ into (b_i, B), and to neglect the intervals in which $b_i > B$. Whence, it is equivalent to assume that $per = \max_{i\in I}(d_i - b_i) \le B$ for any barcode on the domain, which makes Equation (11) hold. □

Appendix B: Proof of Corollary 1

Proof. Let r_{upper} represent the function restricting $\widetilde{r}$ on the domain above the diagonal and r_{down} represent the function below the diagonal. Then the equation $\widetilde{r} = r_{upper} + r_{down}$ holds. Because of the definition of extended PRF given in Equation (4) and Equation (5), we have

$$\|r - r^{'}\|_p = \|r_{upper} - r_{upper}^{'}\|_p = \|r_{down} - r_{down}^{'}\|_p. \tag{35}$$

Therefore, it follows by Equation (15) that

$$\|\widetilde{r} - \widetilde{r}^{'}\|_p^p = \|(r_{upper} - r_{upper}^{'}) + (r_{down} - r_{down}^{'})\|_p^p \le 2^p\|r - r^{'}\|_p^p. \tag{36}$$

By using the result in Theorem 4.1, Equation (12) holds. □

Appendix C: Proof of Theorem 4.2

Before proving the final stability theorem of the feature vectors, a lemma is provided as follows.

Lemma 2. *For 2D Haar basis $\{har_i(s,t)\}_{i=1}^\infty$ and $f \in L^2(\Omega)$, the Parseval identity [52] holds, that is,*

$$\|f\|_2^2 = \sum_{i=1}^\infty \langle f, har_i\rangle^2 = \sum_{i=1}^\infty \lambda_i^2. \tag{37}$$

Proof. It follows the fact that Haar basis is a complete and orthonormal basis in $L^2(\Omega)$ [40]. □

Eventually, we give the proof of Theorem 4.2.

Proof. It is left to show that $\|v - v^{'}\|_2^2 \le \|\widetilde{r} - \widetilde{r}\|_2^2$. According to the definition of feature vector given in Equation (8) and Equation 37 in Lemma 2, we have

$$\|v - v^{'}\|_2^2 = \sum_{i=1}^{N^2}(\lambda_i - \lambda_i^{'})^2 \le \|\widetilde{r} - \widetilde{r}\|_2^2. \tag{38}$$

Then, it follows by Equation (12) that when $p = 2$, Equation (13) holds. □