

Statistical Shape Analysis using Topological Data Analysis

Part III: Real Applications

Chul Moon

ECSSC 2021
25 July 2021

Outline

Point Cloud Data: Application to Fingerprint Classification

3D Material Image: Application to Porous Materials

2D Medical Image: Application to Tumor Images

Application to Lung Cancer Image

Application to Brain Tumor Images

Time Series Data

Time Series as Morse Function

Point Cloud Embedding

Outline

Point Cloud Data: Application to Fingerprint Classification

3D Material Image: Application to Porous Materials

2D Medical Image: Application to Tumor Images

Application to Lung Cancer Image

Application to Brain Tumor Images

Time Series Data

Time Series as Morse Function

Point Cloud Embedding

Application to Fingerprint Classification

Do we need better methods for fingerprints?

- ▶ Three types of fingerprints by Galton (1892)



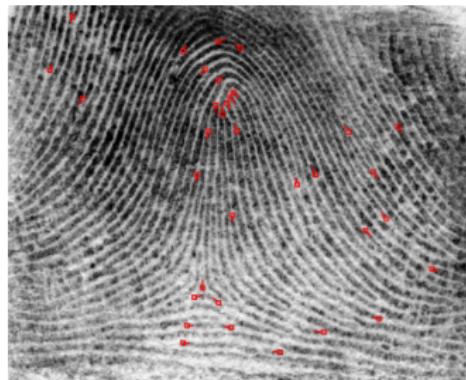
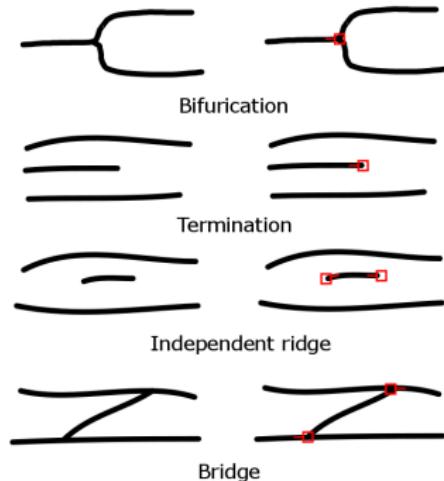
(a) arch

(b) loop

(c) whorl

- ▶ Research objective: classify fingerprints into three classes – arch, loop and whorl – using topological data analysis

Minutiae: Major Features of Fingerprints



- ▶ Minutiae are specific patterns in a fingerprint
- ▶ Minutia point p_i includes location and orientation
 $p_i = (x_i, y_i, \theta_i) \in \mathbb{R}^2 \times S^1$

Defining a Metric Space for Minutiae Points

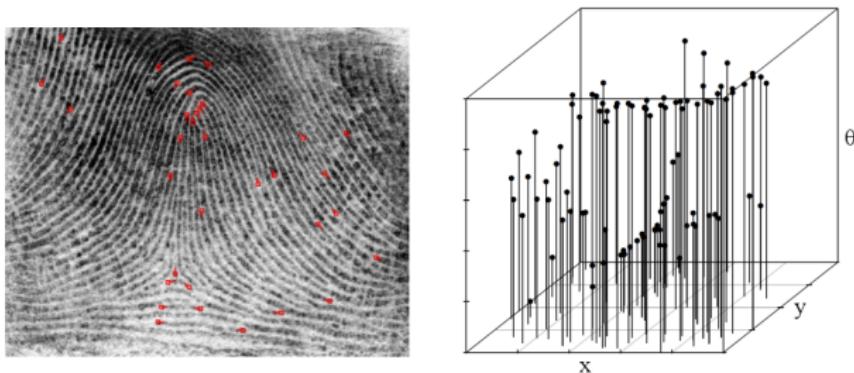


Figure: Minutiae points on $\mathbb{R}^2 \times S^1$ space

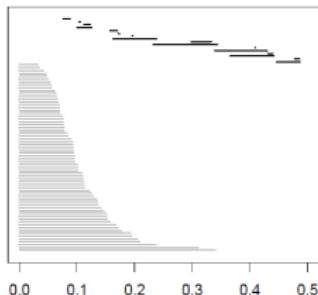
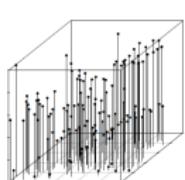
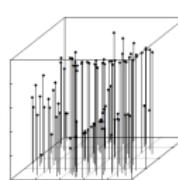
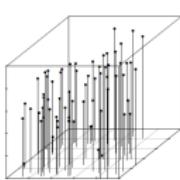
- ▶ Define a metric space considering both location and orientation

$$d(p_i, p_j) = \lambda * \left(\frac{1}{\sqrt{2}} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \right) + (1 - \lambda) \theta_{ij},$$

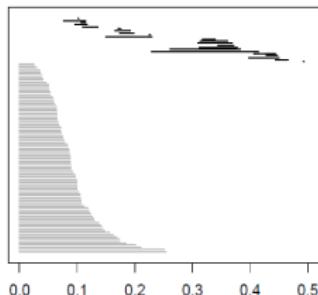
where $\theta_{ij} = \frac{1}{2\pi} |\theta_i - \theta_j|$

- ▶ Compute persistent homology for the metric space

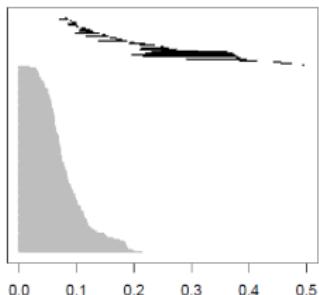
Persistent Homology Computation Result



(a) arch



(b) loop



(c) whorl

Feature Extraction: Intervals to Vector

- ▶ To be used in classification algorithms, multisets of intervals need to be converted into vector format
- ▶ Suggest features for interval [Birth, Death]

Compute features:

1. Statistical features
2. Polynomial features
3. Regression coefficients

⇒ We obtain 552 features for one fingerprint image

Fingerprint Classification Result

Results

- ▶ NIST database SD-27 including 258 fingerprints
- ▶ Feature selection using correlation and backward elimination
- ▶ Train a linear discriminant analysis (LDA) classifier using the leave-one-out-cross-validation (LOOCV)
- ▶ Highest classification rate is **93.1%** (32 features used)

Conclusion

- ▶ Achieve near state-of-the-art classification accuracy rates by using minimum pre-processing processes and a stricter success rule
- ▶ The first paper that applies topological data analysis to fingerprint data

Outline

Point Cloud Data: Application to Fingerprint Classification

3D Material Image: Application to Porous Materials

2D Medical Image: Application to Tumor Images

Application to Lung Cancer Image

Application to Brain Tumor Images

Time Series Data

Time Series as Morse Function

Point Cloud Embedding

Study of Rocks Using Rock Image Data

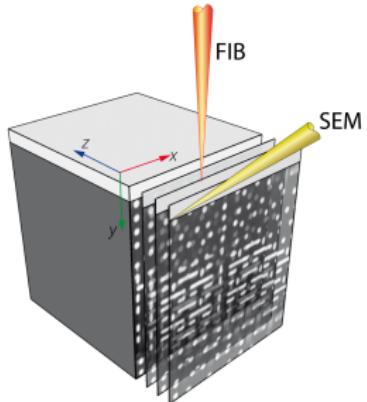


Figure: Figures from Stelma (2004); Weller (2004); Davami (2015)

Data of Interest: Binarized Material Images

- ▶ Original scanned images are grayscale images
- ▶ We only consider binary images in this presentation; applications are not limited to binary images

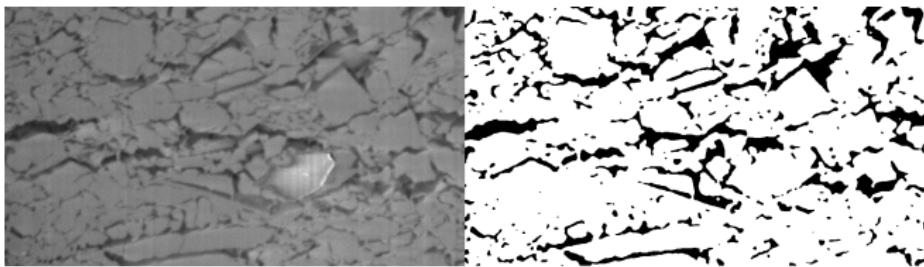


Figure: Original grayscale image and binarized image of Selma Chalk

Traditional Approaches on Rock Physics

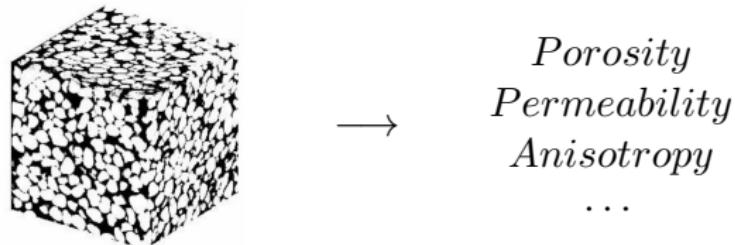


Figure: Figure from Talabi et al. (2009)

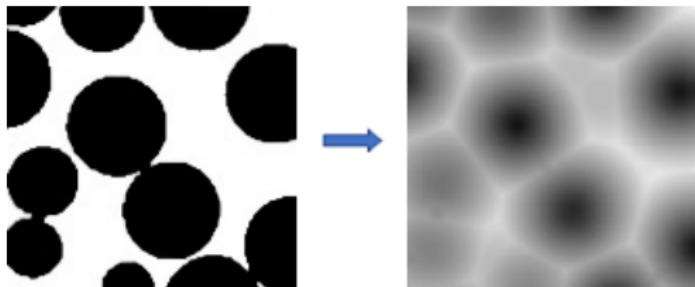
Rock image data – (Simulations) → Estimate geophysical properties

- ▶ Models include limiting geometrical assumptions
- ▶ High computational cost

Topological data analysis provides numerical summary of structure and connectivity of rocks

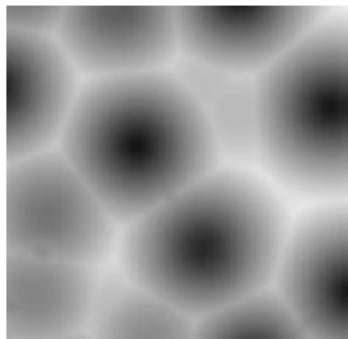
Homology of Rock Image Data?

- ▶ 2-phase binary image → Multiscale image
- ▶ Robins et al. (2016) derive geometric characteristics from binary image using the Signed Euclidean Distance Transform (SEDT)
- ▶ Grain – Positive vs. Pore – Negative
- ▶ Euclidean distance to the opposite phase



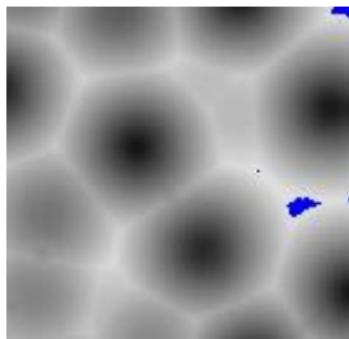
Constructing Complex

- ▶ Multiscale image → Complex
- ▶ Set the level-set filtration on SEDT value
- ▶ Start in the pore phase (negative), moving up to the grain phase (positive)



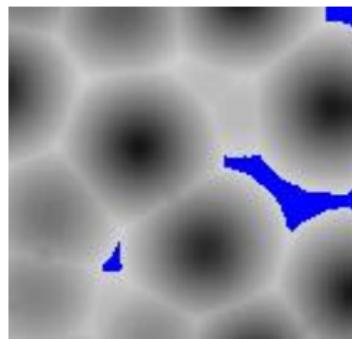
Constructing Complex

- ▶ Multiscale image → Complex
- ▶ Set the level-set filtration on SEDT value
- ▶ Start in the pore phase (negative), moving up to the grain phase (positive)



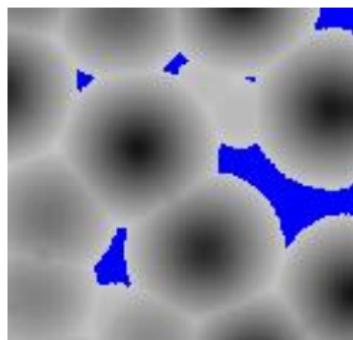
Constructing Complex

- ▶ Multiscale image → Complex
- ▶ Set the level-set filtration on SEDT value
- ▶ Start in the pore phase (negative), moving up to the grain phase (positive)



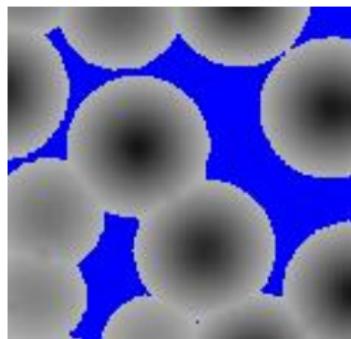
Constructing Complex

- ▶ Multiscale image → Complex
- ▶ Set the level-set filtration on SEDT value
- ▶ Start in the pore phase (negative), moving up to the grain phase (positive)



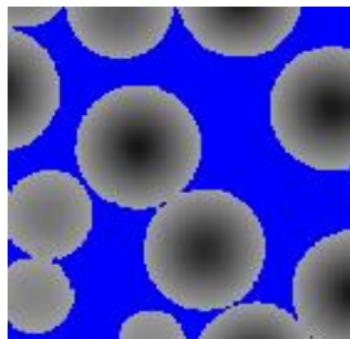
Constructing Complex

- ▶ Multiscale image → Complex
- ▶ Set the level-set filtration on SEDT value
- ▶ Start in the pore phase (negative), moving up to the grain phase (positive)



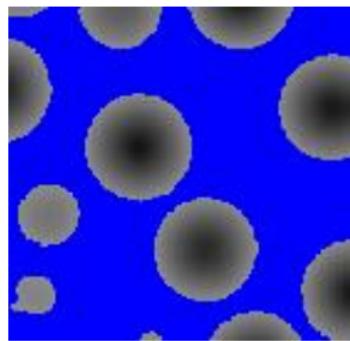
Constructing Complex

- ▶ Multiscale image → Complex
- ▶ Set the level-set filtration on SEDT value
- ▶ Start in the pore phase (negative), moving up to the grain phase (positive)



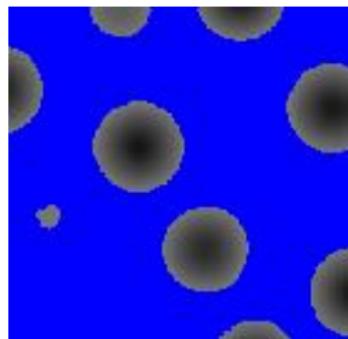
Constructing Complex

- ▶ Multiscale image → Complex
- ▶ Set the level-set filtration on SEDT value
- ▶ Start in the pore phase (negative), moving up to the grain phase (positive)



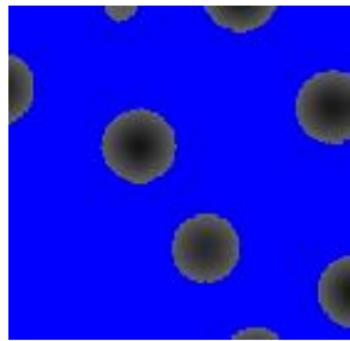
Constructing Complex

- ▶ Multiscale image → Complex
- ▶ Set the level-set filtration on SEDT value
- ▶ Start in the pore phase (negative), moving up to the grain phase (positive)



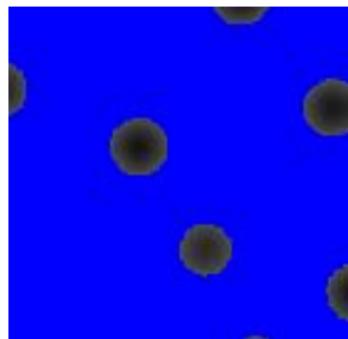
Constructing Complex

- ▶ Multiscale image → Complex
- ▶ Set the level-set filtration on SEDT value
- ▶ Start in the pore phase (negative), moving up to the grain phase (positive)



Constructing Complex

- ▶ Multiscale image → Complex
- ▶ Set the level-set filtration on SEDT value
- ▶ Start in the pore phase (negative), moving up to the grain phase (positive)



Topological Data Analysis Result of Rock Images

- ▶ Binary Image → Grayscale Image → Sequence of Complexes → Persistence diagrams
- ▶ The results represent different characteristics of rock structure and connectivity
 - ▶ Size of pores and grains
 - ▶ Pore throat radius
 - ▶ Shape of pores and grains

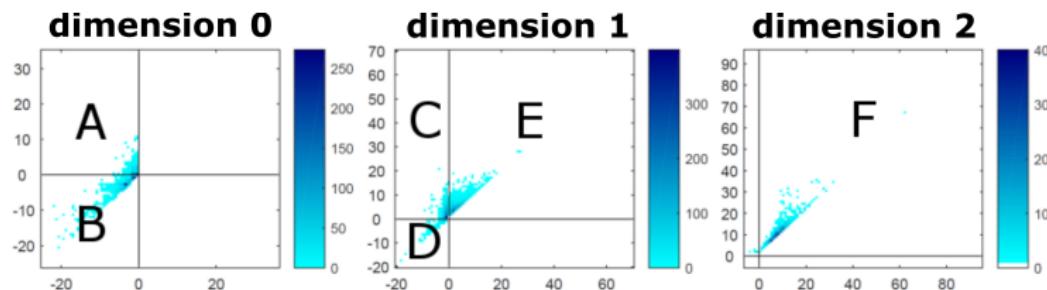


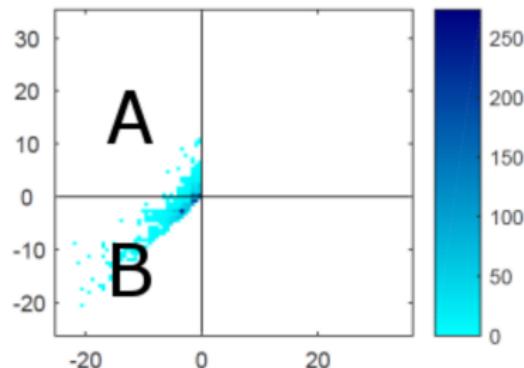
Figure: Computation results summarized as persistence diagrams

Dimension Zero: Size of Pores and Pore Throat Radius

- ▶ A: Size of disconnected pore
 - ▶ Birth (X): size of pore
 - ▶ Death (Y): distance to the nearest pore
- ▶ B: Size of connected pore and pore throat radius
 - ▶ Birth (X): size of pore
 - ▶ Death (Y): pore throat radius in the pore network



dimension 0



Dimension One: Shape of Pores and Grains



Figure: Grain contacts



Figure:
Non-convex pore
structure

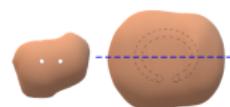
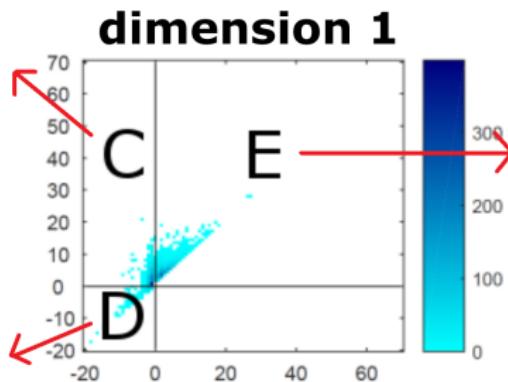
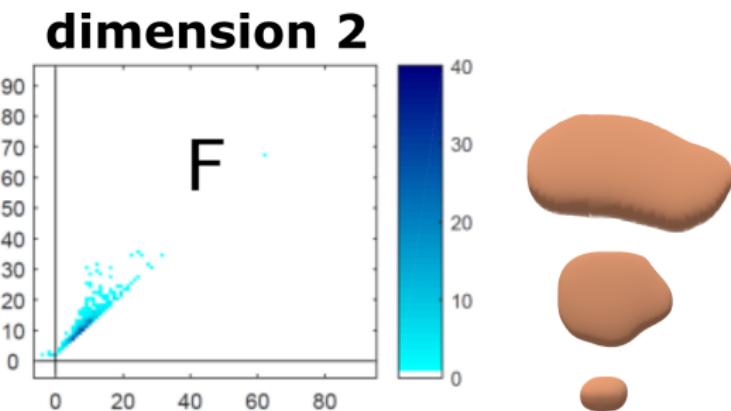


Figure:
Non-convex
grain
structure

Dimension Two: Size of Grain



Simple Application: Strain Level Simulations

- ▶ Research question: “Can persistent homology describe deformation?”
- ▶ Use the Bentheimer strain simulation data (Moon and Andrew, 2019a)
- ▶ GeoDict (2017) finite volume simulations created a series of deformations of the Bentheimer Sandstone
- ▶ The rock was subjected to strain levels of 2%, 4%, 6%, 8%, 10%, 20%, 30%, 40%, and 50%

Rock Type	Modality	Resolution (m)	Sample	Subsamples
Bentheimer	simulation	8.9×10^{-6}	1024^3	$700 \times 600 \times 700$

Persistent Homology Describes Deformation

- ▶ zero-dimensional outputs reflect decreasing pore size
- ▶ one-dimensional outputs describe structural changes in the pores and grains
- ▶ two-dimensional outputs imply that grain sizes increase with pressure

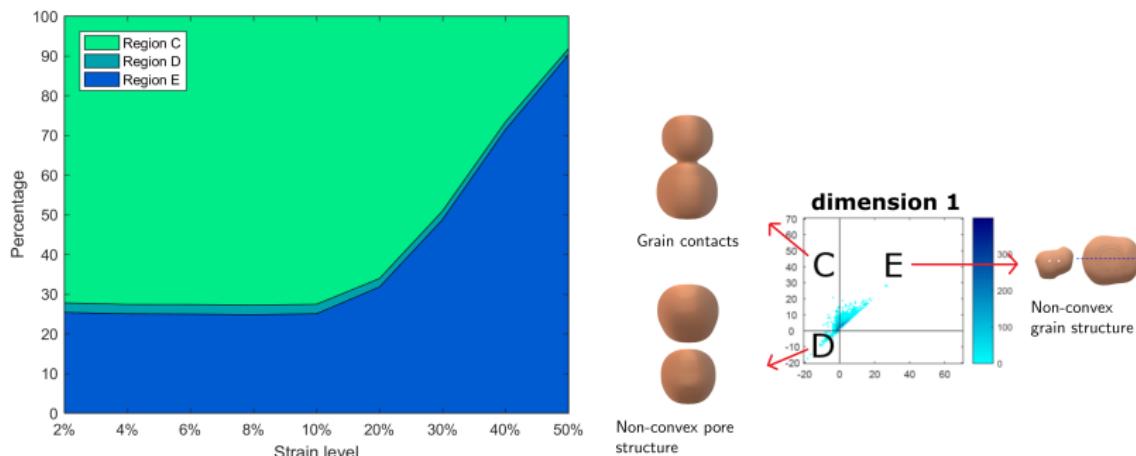


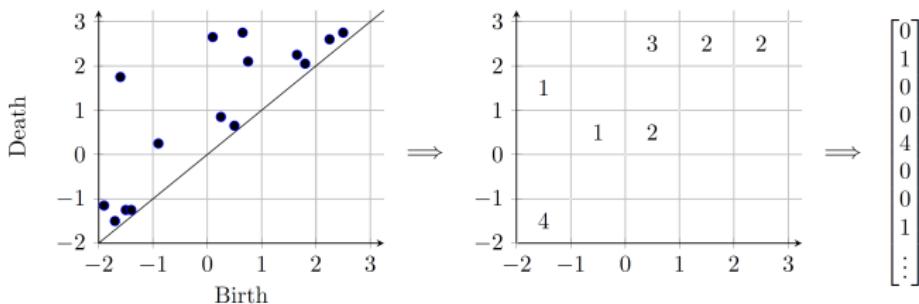
Figure: Fraction of each quadrant in one-dimensional persistence diagrams of Bentheimer Sandstone, as strain increases

Determining Sampling Size for Rock Images

- ▶ The three-dimensional rock images are expensive data to obtain
- ▶ A statistical representative elementary volume (sREV) is used to find the consistent size of materials according to the characteristics of interest
 - ▶ computationally expensive
- ▶ Can we find the sampling size for rock images?
- ▶ Underlying assumption: if the structural properties of sampled subvolumes are similar to each other, then persistence diagrams would be similar as well

Vectorized Persistence Diagrams

- ▶ Persistence diagrams are not easy to compare
- ▶ Consider persistence diagrams as an image and convert to an image vector



Similarity between Vectorized Persistence Diagrams

- ▶ Measure similarities between vectorized persistence diagrams using the structural similarity index (SSIM) of Wang et al. (2004)
- ▶ SSIM index 1 indicates the two images x and y are identical

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}.$$

- ▶ The Mean SSIM (MSSIM) is the average of the SSIM values of multiple local blocks
- ▶ Local blocks that have a non-zero element are used

$$\text{MSSIM}_{PH}(x, \mu) = \frac{1}{\#\{k | \mu_k \neq 0\}} \sum_{i \in \{k | \mu_k \neq 0\}} \text{SSIM}(x_i, \mu_i).$$

Real Applications: Datasets

- ▶ Selma Chalk data (Yoon and Dewers, 2013)
- ▶ Three sandstones data (Moon and Andrew, 2019b)

Rock Type	Modality	Resolution (m)	Sample	Subsamples
Selma Chalk	FIB-SEM	15.6×10^{-9}	$930 \times 520 \times 962$	$150^3, 300^3, 400^3, 500^3,$ $600 \times 520 \times 600,$ $765 \times 520 \times 962$
Bentheimer	XRM	8.9×10^{-6}	1024^3	$700 \times 600 \times 700$
Doddington	XRM	5.4×10^{-6}	1024^3	700^3
Rotleigend	XRM	2.0×10^{-6}	1024^3	$700 \times 700 \times 980$

Application to Rock Images

- ▶ Sampling sizes of rocks are determined considering their structure and connectivity
- ▶ Compute SSIM of the mean persistence diagram and persistence diagrams for different sized subvolumes and report their mean SSIM.
- ▶ Thresholds to used: 0.95 (strict) and 0.9 (weak)

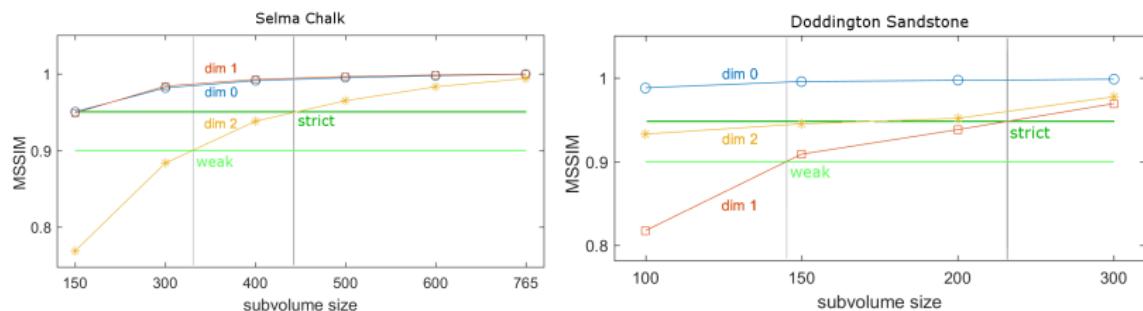
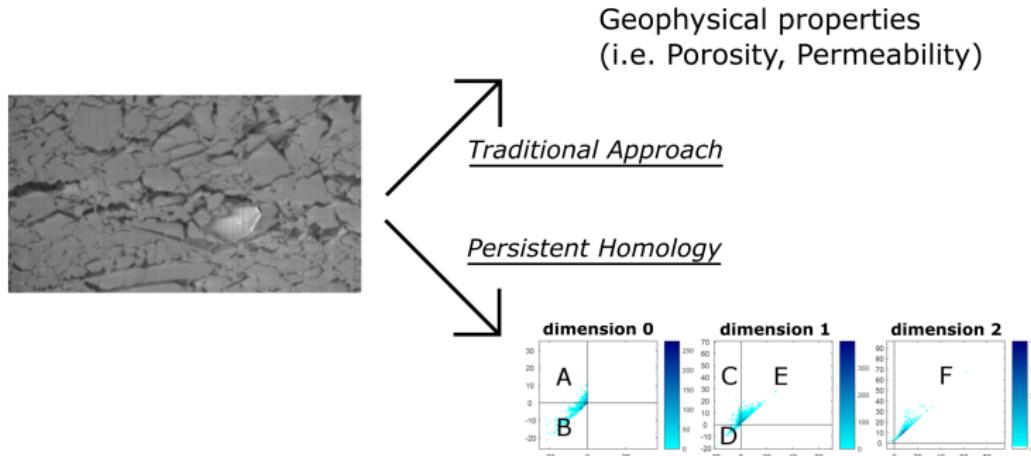


Figure: Mean SSIM of Selma chalk (left) and Doddington sandstone (right) subvolumes

Predicting Geophysical Properties



- ▶ Fit a statistical model to link persistent homology computation result with geophysical characteristics

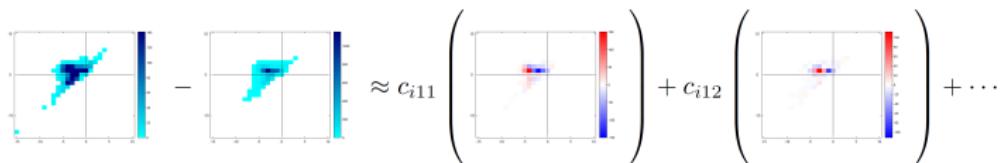
Feature Extraction and Dimension Reduction: PCA

- ▶ Principal component analysis (PCA)

i^{th} dimension k persistence diagram

– mean persistence diagram

$$= c_{ik1} * PC_{k1} + c_{ik2} * PC_{k2} + \dots + c_{ikn} * PC_{kn}$$



- ▶ Input variables of models: set of weights
- ▶ Vector $v_i = \{c_{i01}, c_{i02}, \dots, c_{i2n}\}$ summarizes the i^{th} rock

Penalized Regression: LASSO

- ▶ For n rock samples, we obtain $3n$ weights; “small n large p ” case

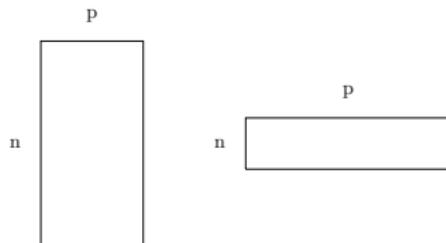


Figure: “large n small p ” vs. “small n large p ” data

- ▶ LASSO is the penalized regression models that give a L_1 penalty

$$\min_{\beta} \{ ||y - X\beta||_2^2 + \gamma ||\beta||_1 \}.$$

- ▶ The proposed model is

$$y = \text{geophysical property} \sim f_{LASSO}(\text{PCA weights})$$

Predictions using Selma Chalk Data (1)

- ▶ Use three sizes of subvolumes (150^3 , 300^3 and 400^3) of Selma Chalk Data from Yoon and Dewers (2013)
- ▶ The ratio of training, validation, and test sets is 60%, 20% and 20%
- ▶ Compared the LASSO model with three models
 - ▶ (Intercept) average of y
 - ▶ (LR) linear regression model using two x variables, porosity and specific surface area
 - ▶ (PCA) linear regression model using nine weights that correspond to the principal components that have the largest eigenvalues

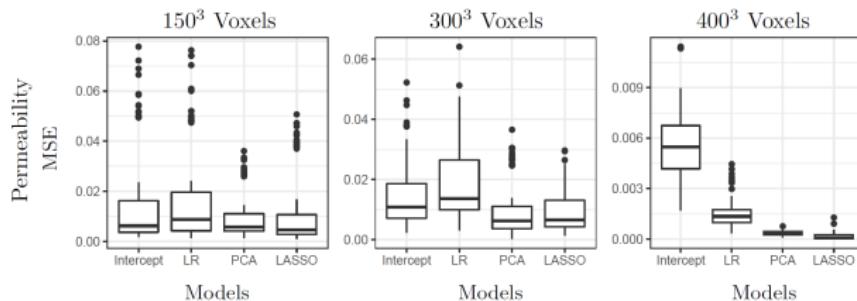


Figure: Permeability prediction results of three subvolumes



Predictions using Selma Chalk Data (2)

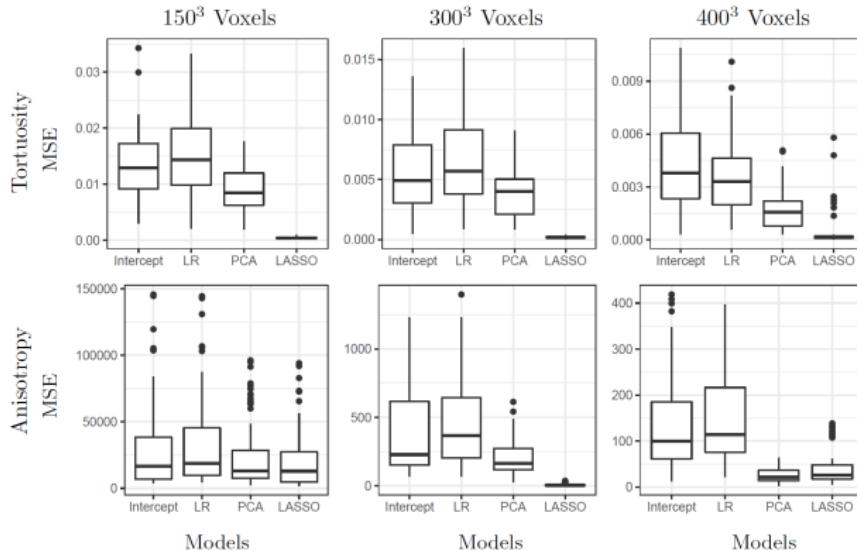


Figure: Tortuosity and anisotropy prediction results of three subvolumes

- ▶ Predict geophysical properties using a small number of rock samples

Summary

- ▶ Topological data analysis provides a tool to analyze shape/structure of data
- ▶ We propose a rock image analysis pipeline using topological data analysis
 - ▶ Hypothesis test to compare samples
 - ▶ Suggest a method to select an appropriate sampling size for porous materials
 - ▶ Predict geophysical properties using a small number of rock samples

Outline

Point Cloud Data: Application to Fingerprint Classification

3D Material Image: Application to Porous Materials

2D Medical Image: Application to Tumor Images

Application to Lung Cancer Image

Application to Brain Tumor Images

Time Series Data

Time Series as Morse Function

Point Cloud Embedding

Pathology Images and Preprocessing

- ▶ Lung cancer is one of the most deadly cancers and adenocarcinoma that accounts for about 40% of all lung cancers
- ▶ We use 247 pathology images of 143 lung adenocarcinoma patients in National Lung Screening Trial (NLST) data
- ▶ Pixels are segmented into tumor, normal, and empty regions using a deep convolutional neural network by Wang et al. (2018)

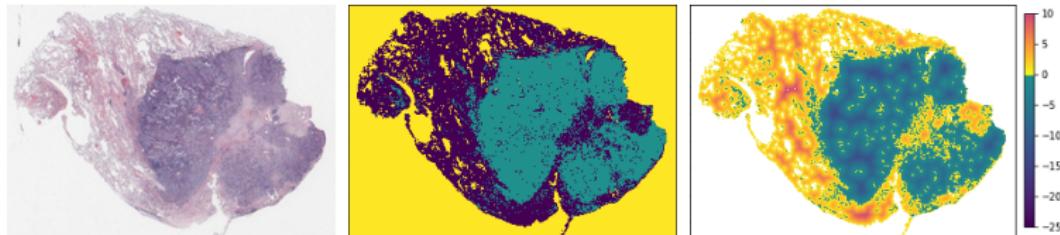


Figure: Denoised three-class pathology image and its SEDT-3

Homology of Ternary Image Data?

- ▶ Ternary image → Multiscale image
- ▶ We propose the Signed Euclidean Distance Transform for ternary image (SEDT-3)
 - ▶ Tumor – Negative vs. Normal – Positive vs. Empty – Infinite
 - ▶ Euclidean distance to the nearest the other phase pixel

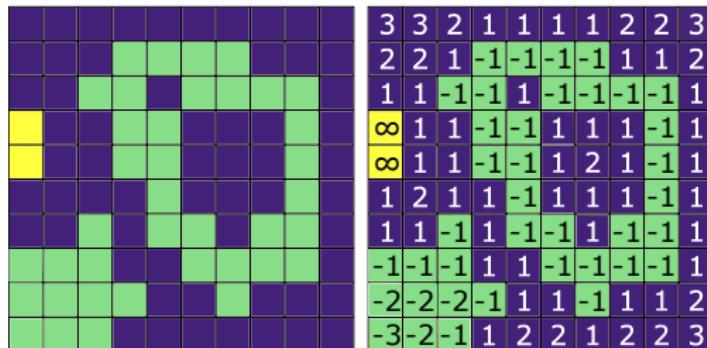


Figure: Ternary image and its signed taxicab distance transform

Constructing Complex

- ▶ Multiscale image → Sequence of complexes
- ▶ Construct complex using the SEDT-3 value: pixels whose SEDT-3 values are smaller than or equal to ϵ are used
- ▶ Start in the tumor pixels (negative), moving up to the normal pixels (positive)
- ▶ The empty pixels are not used to construct complexes

3	3	2	1	1	1	1	2	2	3
2	2	1	-1	-1	-1	-1	1	1	2
1	1	-1	-1	1	-1	-1	-1	-1	1
∞	1	1	-1	-1	1	1	1	-1	1
∞	1	1	-1	-1	1	2	1	-1	1
1	2	1	1	-1	1	1	1	-1	1
1	1	-1	1	-1	-1	1	-1	-1	1
-1	-1	-1	1	1	-1	-1	-1	-1	1
-2	-2	-2	-1	1	1	-1	1	1	2
-3	-2	-2	-1	1	2	2	1	2	3

Constructing Complex

- ▶ Multiscale image → Sequence of complexes
- ▶ Construct complex using the SEDT-3 value: pixels whose SEDT-3 values are smaller than or equal to ϵ are used
- ▶ Start in the tumor pixels (negative), moving up to the normal pixels (positive)
- ▶ The empty pixels are not used to construct complexes

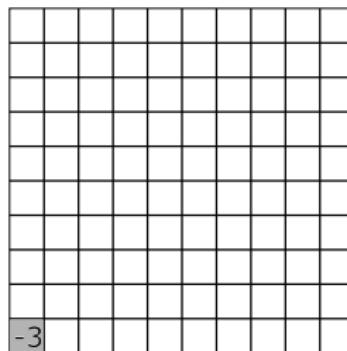


Figure: $\epsilon = -3$

Constructing Complex

- ▶ Multiscale image → Sequence of complexes
- ▶ Construct complex using the SEDT-3 value: pixels whose SEDT-3 values are smaller than or equal to ϵ are used
- ▶ Start in the tumor pixels (negative), moving up to the normal pixels (positive)
- ▶ The empty pixels are not used to construct complexes

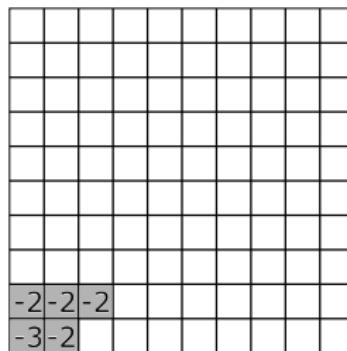


Figure: $\epsilon = -2$

Constructing Complex

- ▶ Multiscale image → Sequence of complexes
- ▶ Construct complex using the SEDT-3 value: pixels whose SEDT-3 values are smaller than or equal to ϵ are used
- ▶ Start in the tumor pixels (negative), moving up to the normal pixels (positive)
- ▶ The empty pixels are not used to construct complexes

		-1	-1	-1	-1				
	-1	-1		-1	-1	-1	-1		
		-1	-1				-1		
		-1	-1				-1		
			-1				-1		
		-1		-1	-1		-1	-1	
-1	-1	-1			-1	-1	-1	-1	
-2	-2	-2	-1			-1			
-3	-2	-1							

Figure: $\epsilon = -1$

Constructing Complex

- ▶ Multiscale image → Sequence of complexes
- ▶ Construct complex using the SEDT-3 value: pixels whose SEDT-3 values are smaller than or equal to ϵ are used
- ▶ Start in the tumor pixels (negative), moving up to the normal pixels (positive)
- ▶ The empty pixels are not used to construct complexes

			1	1	1	1			
		1	-1	-1	-1	-1	1	1	
1	1	-1	-1	1	-1	-1	-1	-1	1
	1	1	-1	-1	1	1	1	-1	1
	1	1	-1	-1	1		1	-1	1
1		1	1	-1	1	1	1	-1	1
1	1	-1	1	-1	-1	1	-1	-1	1
-1	-1	-1	1	1	-1	-1	-1	-1	1
-2	-2	-2	-1	1	1	-1	1	1	
-3	-2	-1	1			1			

Figure: $\epsilon = 1$

Constructing Complex

- ▶ Multiscale image → Sequence of complexes
- ▶ Construct complex using the SEDT-3 value: pixels whose SEDT-3 values are smaller than or equal to ϵ are used
- ▶ Start in the tumor pixels (negative), moving up to the normal pixels (positive)
- ▶ The empty pixels are not used to construct complexes

		2	1	1	1	1	2	2	
2	2	1	-1	-1	-1	-1	1	1	2
1	1	-1	-1	1	-1	-1	-1	-1	1
	1	1	-1	-1	1	1	1	-1	1
	1	1	-1	-1	1	2	1	-1	1
1	2	1	1	-1	1	1	1	-1	1
1	1	-1	1	-1	-1	1	-1	-1	1
-1	-1	-1	1	1	-1	-1	-1	-1	1
-2	-2	-2	-1	1	1	-1	1	1	2
-3	-2	-1	1	2	2	1	2	2	

Figure: $\epsilon = 2$

Constructing Complex

- ▶ Multiscale image → Sequence of complexes
- ▶ Construct complex using the SEDT-3 value: pixels whose SEDT-3 values are smaller than or equal to ϵ are used
- ▶ Start in the tumor pixels (negative), moving up to the normal pixels (positive)
- ▶ The empty pixels are not used to construct complexes

3	3	2	1	1	1	1	2	2	3
2	2	1	-1	-1	-1	-1	1	1	2
1	1	-1	-1	1	-1	-1	-1	-1	1
	1	1	-1	-1	1	1	1	-1	1
	1	1	-1	-1	1	2	1	-1	1
1	2	1	1	-1	1	1	1	-1	1
1	1	-1	1	-1	-1	1	-1	-1	1
-1	-1	-1	1	1	-1	-1	-1	-1	1
-2	-2	-2	-1	1	1	-1	1	1	2
-3	-2	-1	1	2	2	1	2	2	3

Figure: $\epsilon = 3$

Topological Data Analysis Result of Binary Images

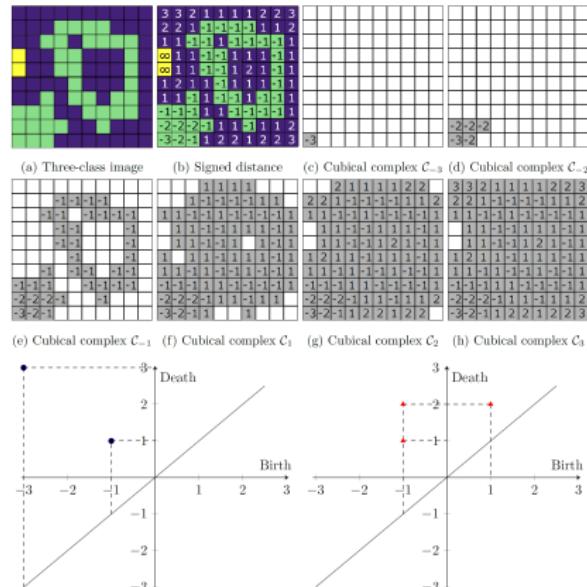
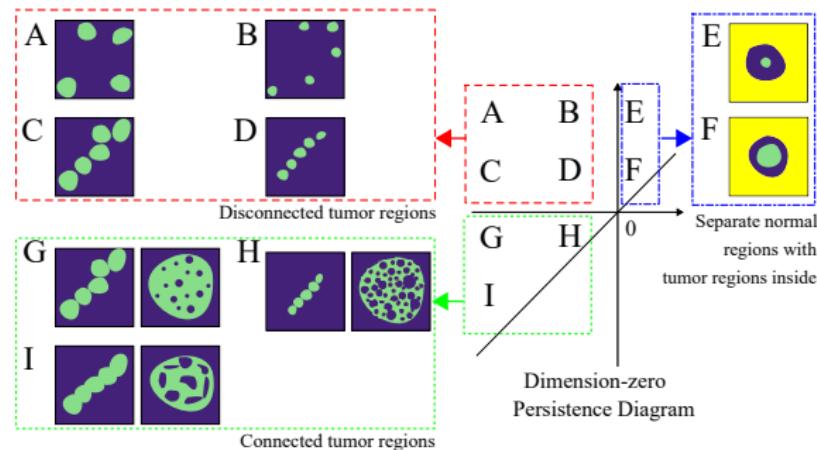


Figure: Computation results summarized as persistence diagrams

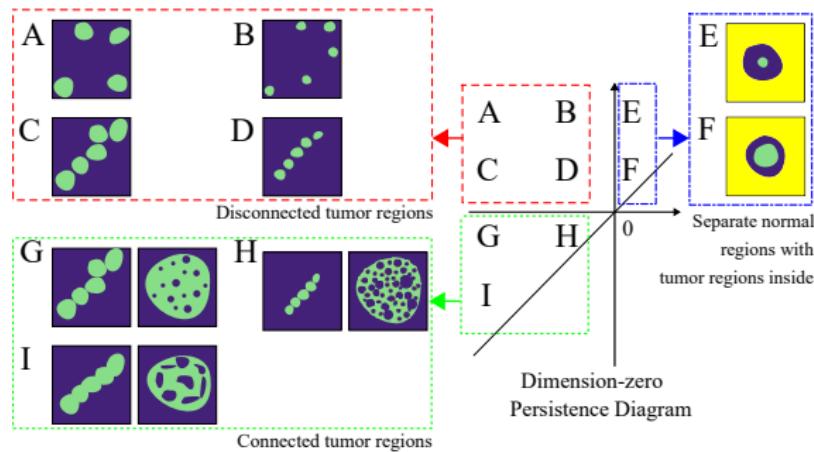
Which information can we obtain here?

Dimension Zero Features: Size and Distributions of Tumors



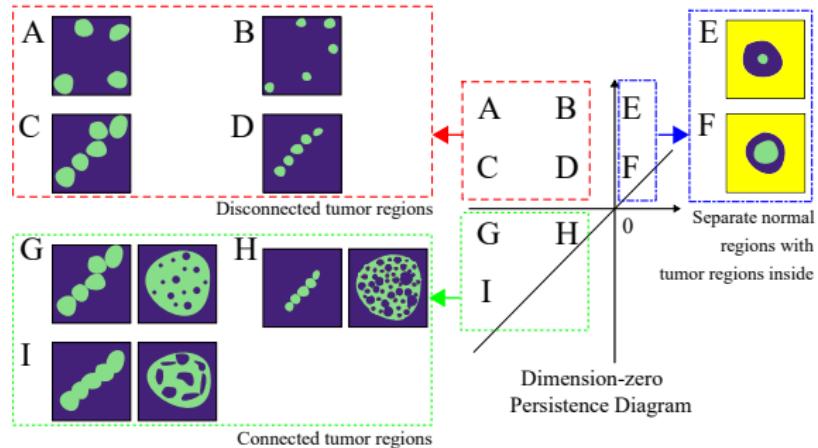
- ▶ Quadrant II (e.g., A, B, C, and D)
 - ▶ Scattered tumors
 - ▶ Birth (X): size of tumor (larger the tumor, larger the absolute birth values)
 - ▶ Death (Y): distances between tumor regions (further the tumor regions, larger the death values)

Dimension Zero Features: Size and Distributions of Tumors



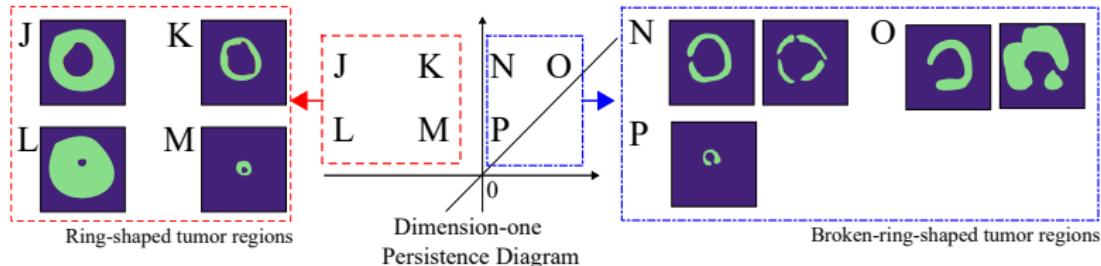
- ▶ Quadrant I (e.g., E and F)
 - ▶ Separate normal regions that include tumor regions
 - ▶ Larger the death values, thicker the normal regions

Dimension Zero Features: Size and Distributions



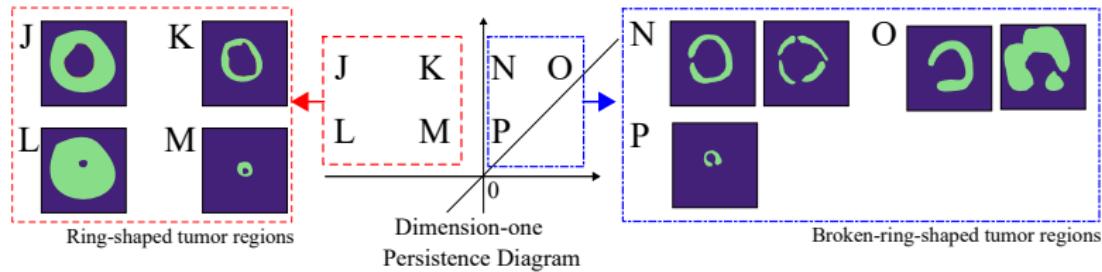
- ▶ Quadrant III (e.g., G, H, and I)
 - ▶ Connected tumor regions
 - ▶ Birth (X): size of tumor (larger the tumor, larger the absolute birth values)
 - ▶ Death (Y): contact area of two tumor regions (larger the contact areas, larger the absolute death values)

Dimension One Features: Shape



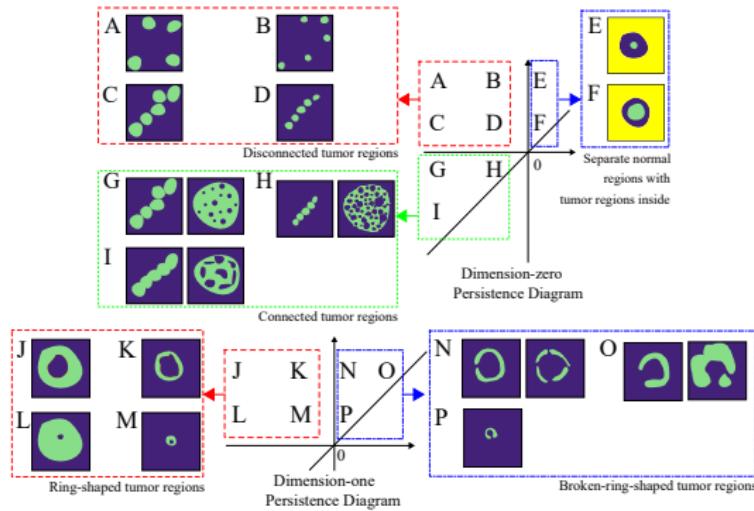
- ▶ Quadrant II (e.g., J, K, L, and M)
 - ▶ Tumor region that surrounds normal region
 - ▶ Birth (X): thickness of tumor region (thicker the tumor region, larger the absolute birth values)
 - ▶ Death (Y): size of trapped normal region (larger the normal region, larger the death values)

Dimension One Features: Shape



- ▶ Quadrant I (e.g., N, P, and O)
 - ▶ Broken-ring shaped tumor regions
 - ▶ Birth (X): size of pore
 - ▶ Death (Y): distance to the nearest pore

Additional Notes on Topological Features



- ▶ The size of features is measured by the radius of the largest circle that can be placed inside of it.
- ▶ Dimension zero and one features are not exclusive; the pixels used to construct dimension zero features can be used to build dimension one feature and vice versa.

Functional Survival Model

- ▶ Predict survival risks of patients using clinical variables and persistent homology shape features (i.e., persistence functions X^0 and X^1)
- ▶ Functional Cox proportional-hazards (FCoxPH) model

$$h(t) = h_0(t) \exp \left(Z^T \gamma + \int X^0(u) \alpha(u) du + \int X^1(v) \beta(v) dv \right).$$

where h_0 is the baseline hazard function, $t \in [0, \tau]$ for $0 < \tau < \infty$, $Z = (z_1, \dots, z_p)^T$ is a p -dimensional scalar predictor, and X^0 and X^1 are functional topological shape data

- ▶ Use both scalar variable Z and two functional predictors X^0 and X^1

FCoxPH Model with Selected FPCs

- Dimension reduction of functional data is required due to their infinite dimensions
 - Functional data is approximated by the functional principal component (FPC) analysis with q and r selected number of FPCs

$$X_i^0 \approx \mu^0(u) + \sum_{j=1}^q \xi_{ij} \phi_j(u) \text{ and } X_i^1 \approx \mu^1(v) + \sum_{k=1}^r \zeta_{ik} \pi_k(v)$$

- The FCoxPH model can be approximated as

$$h_i(t) \approx h_0^*(t) \exp \left(Z_i^T \gamma + \sum_{j=1}^q \xi_{ij} \alpha_j + \sum_{k=1}^r \zeta_{ik} \beta_k \right)$$

- Use Akaike Information Criterion (AIC) to select the number of FPCs
 - $AIC(q, r) = 2(q+r) - 2 \log\{L(\hat{\gamma}_1, \dots, \hat{\gamma}_p, \hat{\alpha}_1, \dots, \hat{\alpha}_q, \hat{\beta}_1, \dots, \hat{\beta}_r | q, r)\}$

Two Survival models

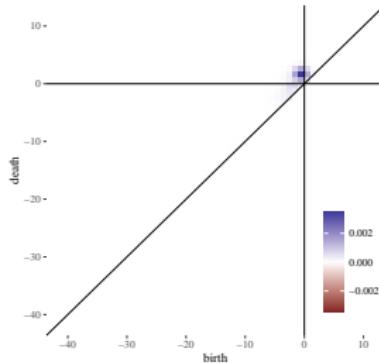
1. Cox regression (CoxPH) model
 - ▶ Scalar (clinical) predictors: age, sex, smoking status, stage of cancer (I to IV), and tumor size
2. Functional Cox regression (FCoxPH) model
 - ▶ Scalar (clinical) predictors: age, sex, smoking status, stage of cancer (I to IV), and tumor size
 - ▶ Functional (topological) predictors: three FPCs are selected
 $q = 1$ and $r = 2$ by AIC.

Estimated Survival Models

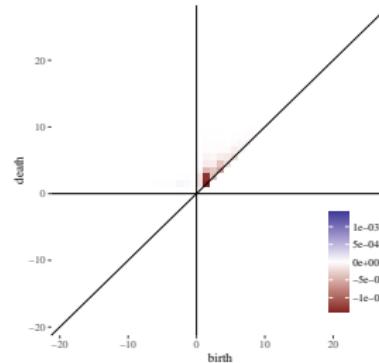
- ▶ The overall p-values of the Wald tests for global statistical significance are 2×10^{-4} for the CoxPH model and 3×10^{-7} for the FCoxPH model
- ▶ The chi-square test indicates that the topological features are a strong signal (p-value = 1×10^{-13})

	CoxPH model				FCoxPH model			
	coef.	exp(coef.)	SE	p-value	coef.	exp(coef.)	SE	p-value
Age	0.084	1.088	0.023	0.007	0.083	1.087	0.024	0.013
Smoker vs. non-smoker	-0.067	0.935	0.262	0.831	-0.253	0.776	0.234	0.426
Female vs. male	0.081	1.085	0.235	0.803	0.088	1.092	0.242	0.791
Tumor size	0.000	1.000	0.000	0.084	<0.001	1.000	<0.001	0.572
Stage II vs. stage I	0.545	1.725	0.382	0.366	0.477	1.611	0.390	0.432
Stage III vs. stage I	1.195	3.304	0.270	0.002	1.170	3.222	0.277	0.002
Stage IV vs. stage I	1.442	4.227	0.342	0.002	1.424	4.152	0.346	0.003
Dimension 0, 1 st FPC	-	-	-	-	0.005	1.005	0.001	<0.001
Dimension 1, 1 st FPC	-	-	-	-	<0.001	1.001	<0.001	0.002
Dimension 1, 2 nd FPC	-	-	-	-	-0.002	0.998	<0.001	<0.001

Estimated Coefficient Functions and Clinical Implications



(a) Estimated dimension zero
functional coefficient $\hat{\alpha}(u)$

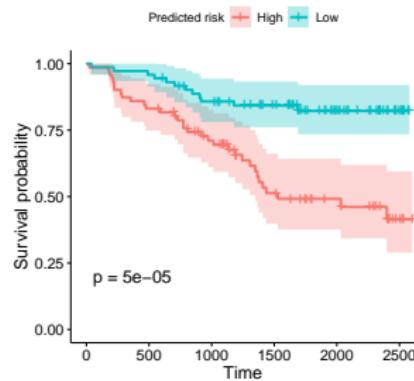


(b) Estimated dimension one
functional coefficient $\hat{\beta}(v)$

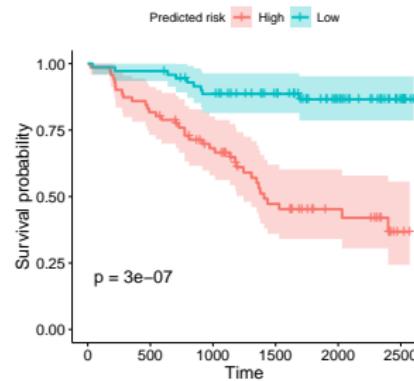
- ▶ Estimated coefficient functions show which topological features are related with the survival risks
 - ▶ Blue-colored areas: aggressive tumor patterns
 - ▶ Red-colored areas: less-aggressive tumor patterns
- ▶ These results coincide with the existing findings on how tumor shapes associate with tumor progression and patient survival

Prediction using Cross-Validation

- ▶ Predict the risk scores using the leave-one-out cross-validation (LOOCV) for two models and assign the patients into 71 high-risk and 72 low-risk patients groups
- ▶ The p-value of the log-rank test of the FCoxPH model is 3×10^{-7} , which is smaller than that of the CoxPH model 5×10^{-5} .



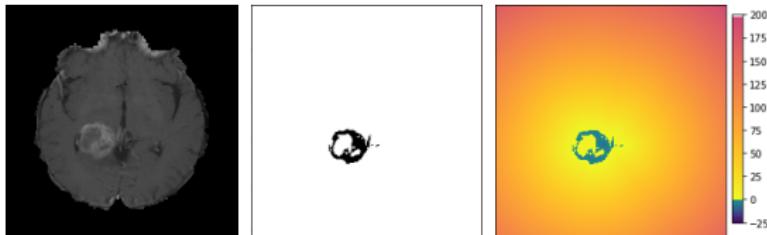
(a) CoxPH model



(b) FCoxPH model

Figure: The Kaplan–Meier plots for the high- and low-risk groups

Glioblastoma MRI Data



- ▶ Glioblastoma multiforme (GBM) is the most common malignant grade IV brain tumor
- ▶ The rapid outward growth of GBM develops necrosis, and most GBM cases show a ring-shaped enhancement
- ▶ We use 77 GBM patients' MRIs from The Cancer Imaging Archive and clinical data from The Cancer Genome Atlas
- ▶ MRI Images are segmented using the Medical Imaging Interaction Toolkit with augmented tools for segmentation
- ▶ Scalar predictors: age, gender, Karnofsky performance score (KPS)

Estimated Survival models

- ▶ Four FPCs ($q = 1$ and $r = 3$) are selected by AIC
- ▶ The overall p-values of the Wald tests are 0.02 for the CoxPH model and 6×10^{-4} for the FCoxPH model
- ▶ The p-value of the chi-square test of degrees of freedom four is 2×10^{-4} , suggesting that the topological features are a strong signal.

	CoxPH model				FCoxPH model			
	coef.	exp(coef.)	SE	p-value	coef.	exp(coef.)	SE	p-value
Age	0.030	1.030	0.013	0.024	0.035	1.036	0.013	0.006
Female vs. male	-0.030	0.744	0.311	0.340	-0.345	0.708	0.312	0.269
KPS	-0.021	0.980	0.010	0.048	-0.009	0.991	0.011	0.454
Dimension 0, 1st FPC	-	-	-	-	0.002	1.002	0.117	0.990
Dimension 1, 1st FPC	-	-	-	-	0.161	1.175	0.115	0.163
Dimension 1, 2nd FPC	-	-	-	-	0.045	1.175	0.084	0.597
Dimension 1, 3rd FPC	-	-	-	-	-0.305	0.737	0.128	0.018

Estimated Coefficient Functions and Clinical Implications

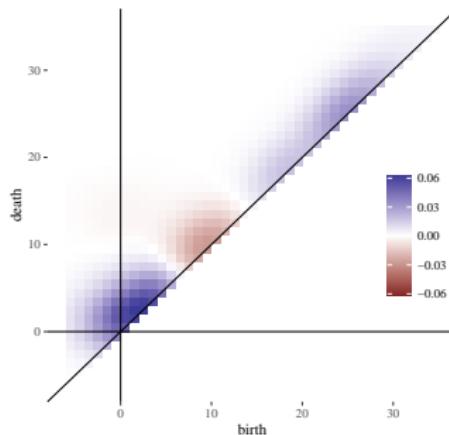


Figure: Estimated dimension one functional coefficient $\hat{\beta}(v)$

- ▶ Blue-colored areas: tumor patterns positively related with risks
- ▶ Red-colored areas: tumor patterns negatively related with risks

Prediction using Cross-Validation

- ▶ Compare three models: Gaussian Process (GP) of Crawford et al. (2020), CoxPH, and FCoxPH models
- ▶ Predict the risk scores using the LOOCV and assign the patients into 38 high-risk and 39 low-risk patients groups
- ▶ The p-values of the log-rank tests are 0.0069 for the FCoxPH model and 0.17 for the CoxPH model

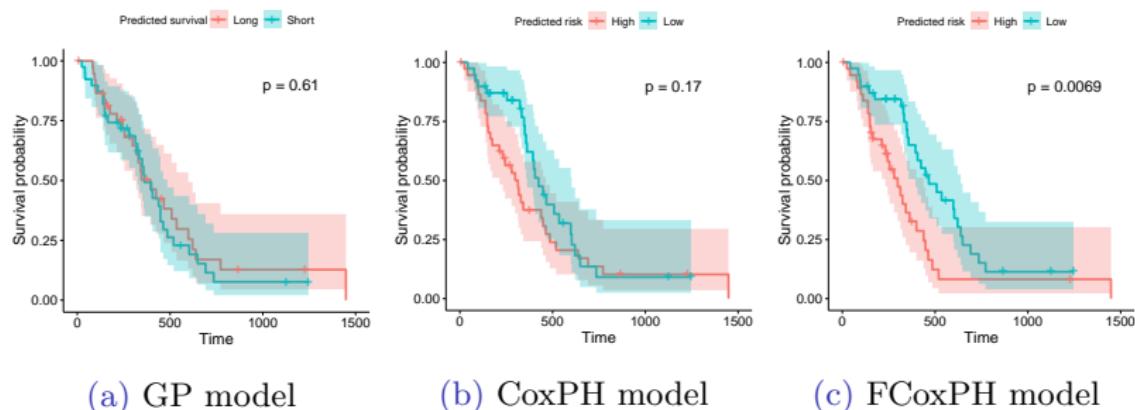


Figure: The Kaplan–Meier plots for the high- and low-risk groups

Conclusions

- ▶ Propose a new tumor shape summary statistic of medical images using topological features
- ▶ The topological features are shown to have a significant non-zero effect on the hazard function; This suggests that the topological features well summarize tumor aggressiveness
- ▶ The proposed model provides interpretable topological shape features while considering right-censored observations

Outline

Point Cloud Data: Application to Fingerprint Classification

3D Material Image: Application to Porous Materials

2D Medical Image: Application to Tumor Images

Application to Lung Cancer Image

Application to Brain Tumor Images

Time Series Data

Time Series as Morse Function

Point Cloud Embedding

For more detailed review, see Ravishanker and Chen (2019).

Morse Function

- ▶ Time series data → Morse function

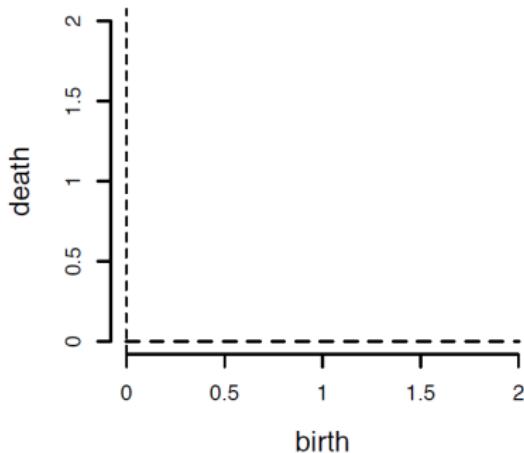
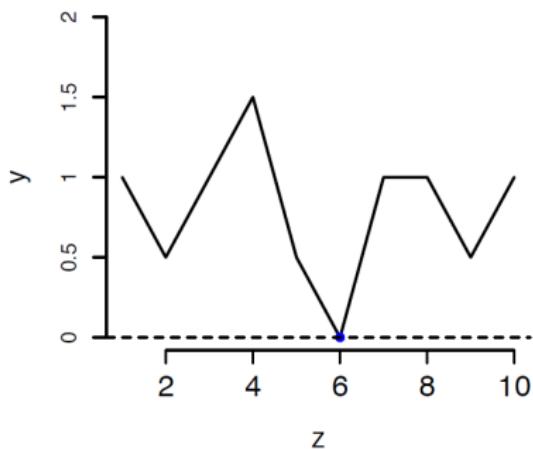


Figure: Figures from Ravishanker and Chen (2019)

Morse Function

- ▶ Time series data → Morse function

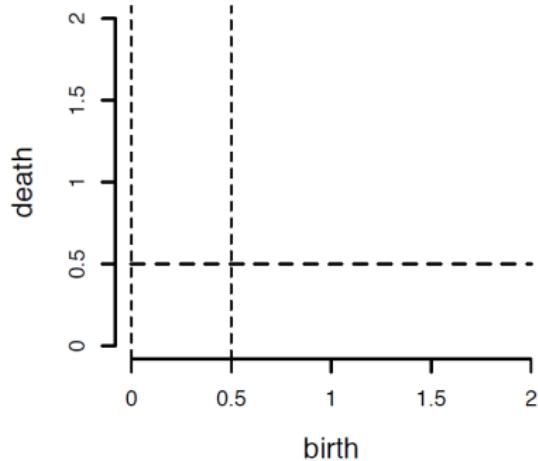
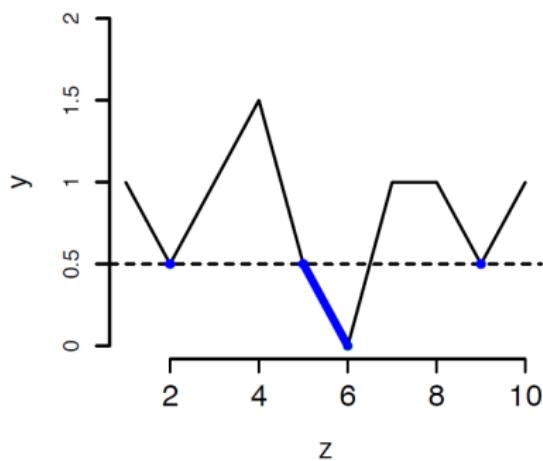


Figure: Figures from Ravishanker and Chen (2019)

Morse Function

- ▶ Time series data → Morse function

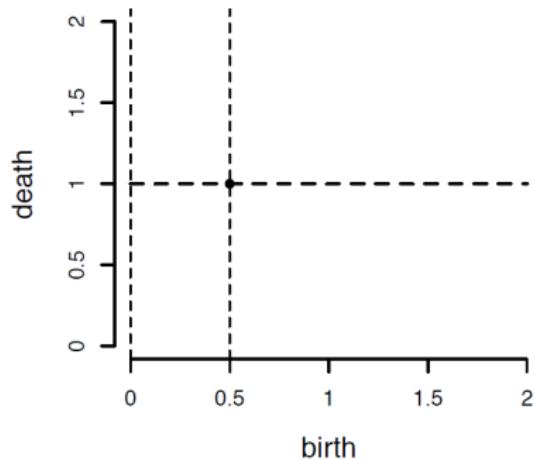
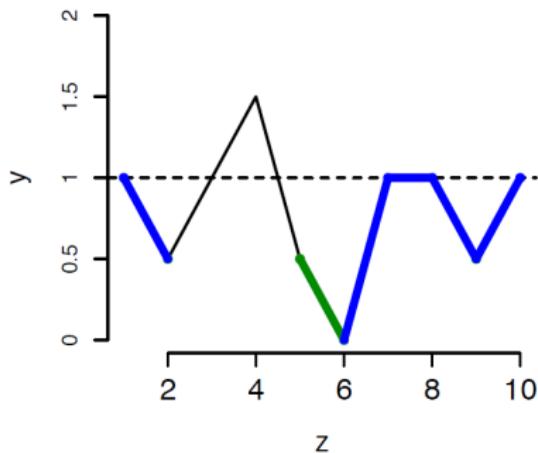


Figure: Figures from Ravishanker and Chen (2019)

Morse Function

- ▶ Time series data → Morse function

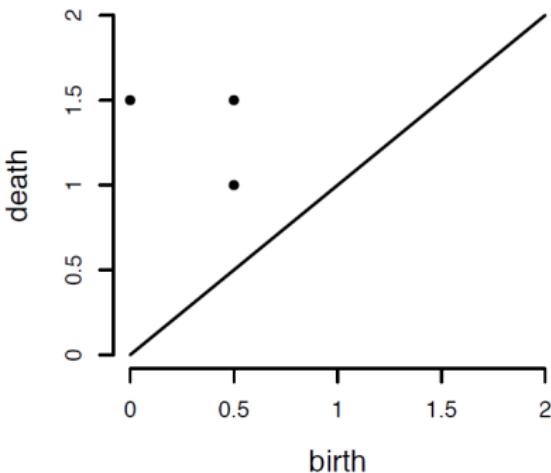
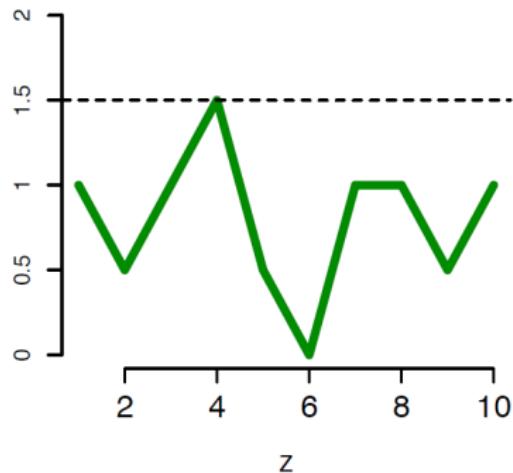
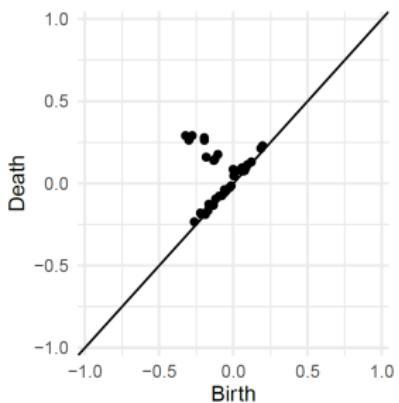
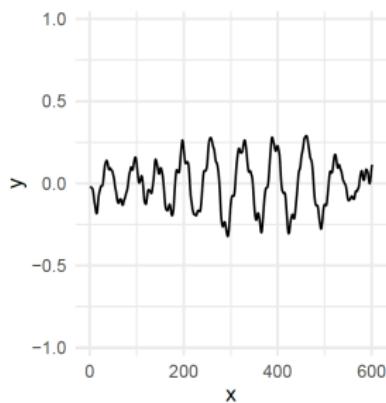
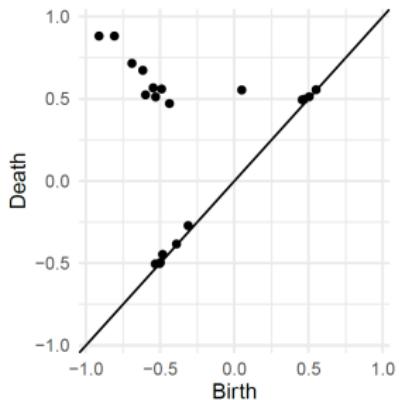
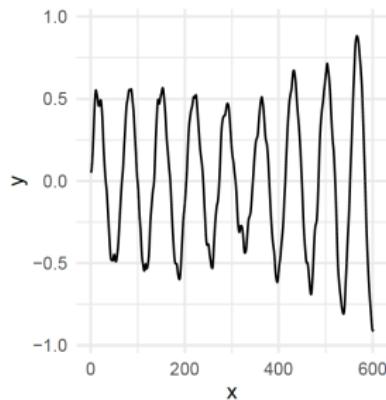


Figure: Figures from Ravishanker and Chen (2019)

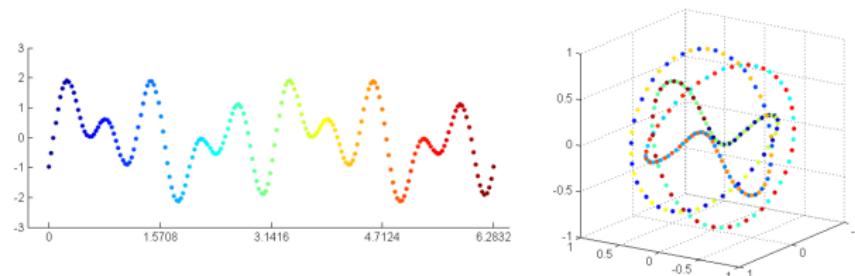
Morse Function Example



Takens's Delay Embedding

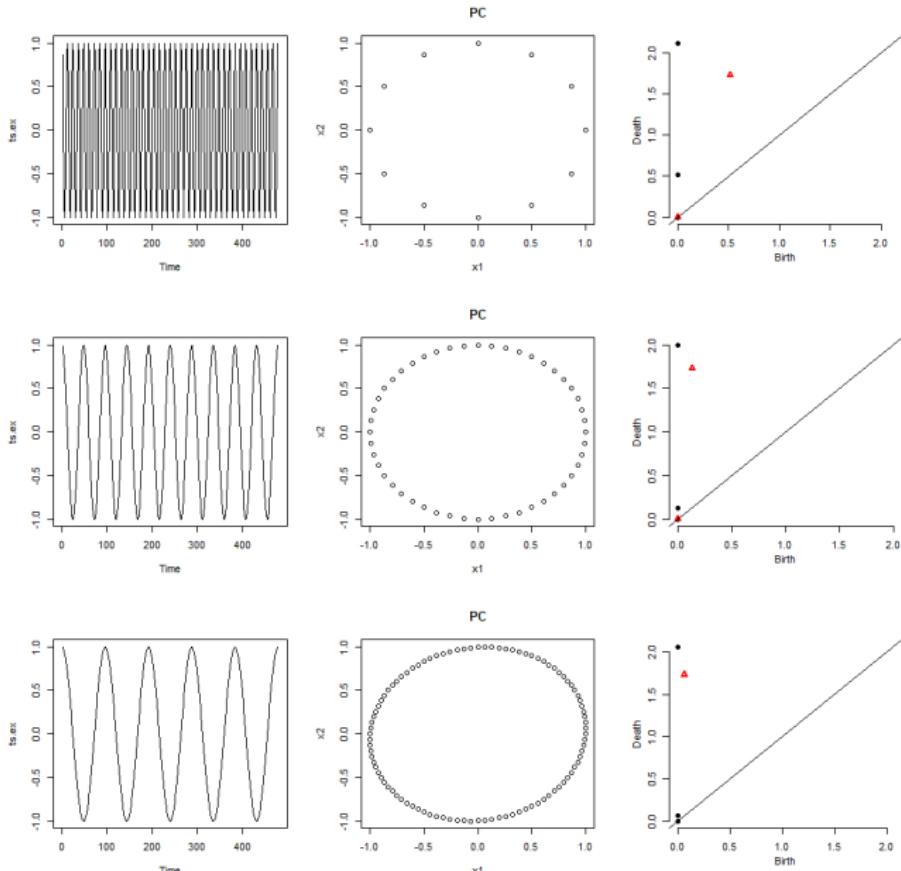
Time series data can be transformed to point clouds by Taken's embedding theorem (Takens, 1981)

- ▶ Input: time series
- ▶ Parameters: dimension of points d and delay parameter τ
- ▶ Time series $\{x_t, t = 1, 2, \dots, T\}$ is converted to points in d -dimensional space $v_i = (x_i, x_{i+\tau}, \dots, x_{i+(d-1)\tau})$



- ▶ Topology of point cloud data may reveal the characteristics of time series data such as periodicity
- ▶ Compute persistent homology using the transformed point clouds

Takens's Delay Embedding Example



Sliding Windows and 1-Persistence Scoring

- ▶ The point cloud embedding has developed to SW1PerS (Sliding Windows and 1-Persistence Scoring) proposed by Perea et al. (2015)
- ▶ SW1PerS detects periodicity of noisy time series data and topological features are robust to differences in time series data such as amplitude, phase, and frequency.

References I

- Crawford, L., Monod, A., Chen, A. X., Mukherjee, S., and Rabadán, R. (2020), “Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis,” *Journal of the American Statistical Association*, 115, 1139–1150.
- Davami, P. (2015), “3D Tomography,” *Arya Electron Optic Ltd.*,
<http://www.arya-eo.com/3D%20Tomography.htm>.
- Galton, F. (1892), *Finger Prints*, McMillan, London.
- Moon, C. and Andrew, M. (2019a), “Bentheimer networks,”
<http://www.digitalrocksportal.org/projects/223>.
- (2019b), “Intergranular Pore Structures in Sandstones,”
<http://www.digitalrocksportal.org/projects/222>.
- Perea, J. A., Deckard, A., Haase, S. B., and Harer, J. (2015), “SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data,” *BMC bioinformatics*, 16, 1–12.
- Ravishanker, N. and Chen, R. (2019), “Topological data analysis (TDA) for time series,” *arXiv preprint arXiv:1909.10604*.

References II

- Robins, V., Saadatfar, M., Delgado-Friedrichs, O., and Sheppard, A. P. (2016), “Percolating Length Scales from Topological Persistence Analysis of Micro-CT Images of Porous Materials,” *Water Resources Research*, 52, 315–329.
- Stelma, D. (2004), “Appraisal Assessment of Geology at Black Rock Damsite,” *U.S. Department of the Interior, Bureau of Reclamation, Pacific Northwest Region*.
- Takens, F. (1981), “Detecting strange attractors in turbulence,” in *Dynamical systems and turbulence, Warwick 1980*, Springer, pp. 366–381.
- Talabi, O., AlSayari, S., Iglauer, S., and Blunt, M. J. (2009), “Pore-scale simulation of NMR response,” *Journal of Petroleum Science and Engineering*, 67, 168 – 178.
- Wang, S., Chen, A., Yang, L., Cai, L., Xie, Y., Fujimoto, J., Gazdar, A., and Xiao, G. (2018), “Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome,” *Scientific Reports*, 8, 10393.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004), “Image Quality Assessment: From Error Visibility to Structural Similarity,” *IEEE transactions on image processing*, 13, 600–612.

References III

- Weller, R. (2004), “Photos of Rocks,” *Cochise College*.
- Yoon, H. and Dewers, T. A. (2013), “Nanopore Structures, Statistically Representative Elementary Volumes, and Transport Properties of Chalk,” *Geophysical Research Letters*, 40, 4294–4298.