Persistent homology fingerprinting of microstructural controls on larger-scale fluid flow in porous media

Chul Moon[1], **Scott A Mitchell**[2], **Nickolas Callor**[3], **Thomas A Dewers**[2], **Jason E Heath**[2], **Hongkyu Yoon**[2] and **Gregory R Conner**[3]

[1]Department of Statistics
University of Georgia
email: chulmoon@uga.edu
[2]Sandia National Laboratories
[3]Brigham Young University

Introduction

Algebraic topology offers powerful mathematical tools for describing structures or shapes. Persistent homology analyzes the dynamics of topological features and summarizes it by numeric values. The dynamics of topological features are recorded as an interval between each feature's birth and death.

We present a persistence homology analysis framework of 3D image sets. We demonstrate it over a focused ion beam scanning electron microscopy dataset of the Selma Chalk. We compute and extract structural characteristics of sampling volumes via persistent homology and principal component analysis (PCA). We fit a statistical model using the summarized values to estimate porosity, permeability, anisotropy, and tortuosity. The suggested framework efficiently predicts fluid flow and transport properties based on geometry and connectivity.

Analysis pipeline

1. **Persistent homology computation and vectorization**
   - Original Images → Binary images → Transformed grayscale images → Cubical complexes → Persistence diagrams → Vectorized persistence diagrams
2. **Sampling**
   - Vectorized persistence diagrams → Similarity metric → Determine sREV
3. **Feature extraction**
   - Vectorized persistence diagrams → Principal component analysis
4. **Modeling**
   - Prediction: Loadings → Fit a statistical model → Prediction
   - Classification: Loadings → Classification

Data

We analyze a FIB-SEM dataset of the Selma Chalk, based on previously binarized images used in the study by Yoon and Dewers [4]. We use previously calculated parameters, which were calculated as a function of an increasingly-larger subvolume. These include porosity, permeability, tortuosity, and anisotropy. There are in total six different sizes of subvolumes: 150^3, 300^3, 400^3, 500^3, $600 \times 520 \times 600$, and $765 \times 520 \times 765$.

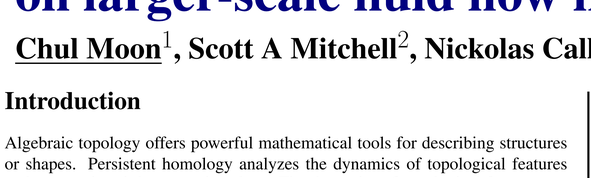


Figure 1: Slice of selma group chalk data (left) and binarized image (right)

Persistent homology computation and vectorization

Persistent homology computation We use the persistent homology computation framework used by Robins et al. [2, 1]. In Robins et al. [2], the geometric characteristics of the binary image are defined by the Signed Euclidean Distance Transform (SEDT). The SEDT assigns a numeric value to each pixel: negative for pore and positive for grain. Its magnitude represents the Euclidean distance between the pixel and the closest opposite status pixel; a large negative value indicates a large pore size, and a large positive value indicates a large grain size. Then, a cubical cell complex is defined based on the discrete Morse function. The SEDT value of a cell is the maximum value of all of its vertices.

As we change the filtration value, k-cell components are added to the cubical cell complex. That is, we add cells one by one in order of increasing SEDT value, thus a cell is added when the current filtration value reaches the maximum value of all of its vertices. By tracking the homology of the sequence of cubical complexes, we can compute the persistent homology.

The barcodes for each dimension are reported separately: the zero, one, and two dimensional homology groups. For each dimension's barcode, the different quadrants of its image reveal different aspects of the rock. Figure 2 shows examples of persistence diagrams, and Table 1 gives interpretations of corresponding regions of persistence diagrams.

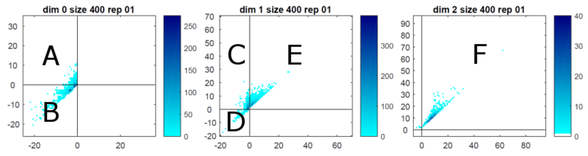


Figure 2: Examples of persistence diagrams dimension 0, 1, and 2

Dim.	Region	Value of X (Birth)	Value of Y (Death)
0	A	Size of pore (max radius)	Narrowest grain contact to other pores
	B	Size of pore (max radius)	Pore throat radius
1	C	Pore tube radius	Grain contact radius
	D	Pore tube radius	Non-convex pore throat radius
	E	Grain tube radius	Non-convex grain throat radius
2	F	Grain-contact radius	Size of grain (max radius)

Table 1: Interpretation of persistence diagrams

Vectorization of persistence diagrams We convert persistence diagrams to image vectors. The vectorization process has two advantages: 1) we may apply image analysis technique 2) we may compute the mean persistence diagram. We bin the elements of the persistence diagram into 100×100 bins; each bin is an output pixel. We count the number of dots that correspond to barcodes in each bin and assign it as the output pixel intensity.

Sampling: determining sREV

The Representative Elementary Volume (REV) is the smallest volume for which a measurement is representative of the whole. The statistical REV, sREV, is a scale where the means of properties are constant, and their variations are small. The sREV is closely related to defining the sampling unit from the rock images, and the "right scale" for rock analysis.

sREV determination If the persistence diagrams of sampled subvolumes are similar to each other, then their structural properties would be consistent. We can determine the resemblance by comparing persistence diagrams. We suggest using the structural similarity (SSIM) [3], the similarity measure for image which is sensitive to the shift and scale differences while robust to perturbations. The SSIM index of two images x and y is defined as a multiplication of three components: the luminance $l(x, y)$, contrast $c(x, y)$, and structure $s(x, y)$. The SSIM varies from zero to one, where one indicates two images x and y are identical. We use the default setup so that

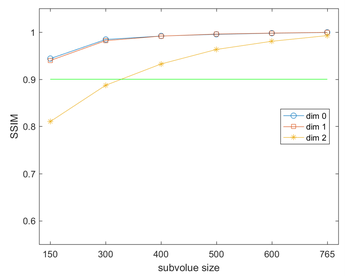


Figure 3: Average SSIM: Selma group chalk

$$\begin{aligned}\text{SSIM}(x, y) &= l(x, y)^{\alpha} * c(x, y)^{\beta} * s(x, y)^{\gamma} \\ &= \frac{(2\mu_x\mu_y + 0.01)(2\sigma_{xy} + 0.03)}{(\mu_x^2 + \mu_y^2 + 0.01)(\sigma_x^2 + \sigma_y^2 + 0.03)}.\end{aligned}$$

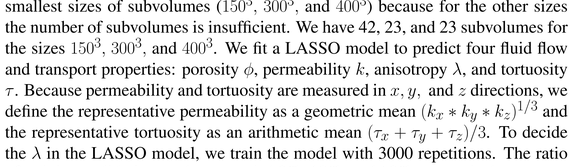
Instead of computing SSIM for the whole image, Wang et al. [3] suggests using a local block. Similarly, we compute SSIM of the local blocks that have a non-zero element.

sREV for Selma group chalk We compute persistent homology for six differently-sized subvolumes. For each size, SSIM is computed between the mean vectorized persistence diagram (image x) and every persistence diagram (image y) and average SSIM is reported. We use 5-by-5 sized local blocks in computation. Figure 3 shows the average SSIM values of three dimensions for six subvolumes. The dimension 2 persistence diagrams, related to size of grain, show the biggest difference in Selma group chalk. There is no standard for deciding REV using SSIM. We set the threshold to be 0.9 because Yoon and Deweres [4] finds that the subvolume size 400^3 is the sREV.

Feature extraction: principal component analysis

We extract features from persistence diagrams using principal component analysis. We subtract the mean (vectorized) persistence diagram from all the persistence diagram vectors and compute the principal components. The principal components form a basis to represent the vectorized persistence diagrams:

$$\begin{aligned}&i^{th} \text{ persistence diagram of dim } k - \text{mean persistence diagram of dim } k \\ &= c_{ik1} * PC_{k1} + c_{ik2} * PC_{k2} + \cdots + c_{ikn} * PC_{kn}\end{aligned}$$



After computing principal components, we select a subset of principal components that explain at least 85% of total variabilities and discard the rest to achieve a reduction in dimension.

We call the coefficients of the principal components the *loadings*. We can use the set of loadings to summarize persistence diagrams. The Euclidean vector $v = \{c_{i01}, c_{i02}, \cdots c_{ikn}\}$ summarizes the i^{th} porous material.

3D rock Image ⟶ Persistence Diagrams ⟶ Euclidean Vector

Once we convert data into the Euclidean vector, we can then apply classical statistical approaches to make an inference. For example, the numeric values can be used as an explanatory variable for a classification or regression.

Prediction of fluid flow and transport characteristics: penalized regression

We would like to fit a model that explains the geometric properties (y variables) using loadings obtained from the principal component analysis (x variables). However, the number of loadings obtained for all dimensions is $3n$, which is larger than the number of samples n. Even after the dimension reduction via PCA, the number of variables could be larger than the number of samples. We call this the "*small n large p*" case, which can lead to over-fitting a model.

One of the solutions to the overfitting problem is to use a penalized regression. We fit the same regression but give a penalty to the coefficients. LASSO is a penalized regression model using a L_1 penalty. The result of LASSO can be obtained by solving

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

It fits a regression and does variable selection at the same time. The advantage of the LASSO model is that we can see which principal components play a role in predicting fluid flow and transport properties.

Prediction of fluid flow of Selma group chalk We fit a model only for three smallest sizes of subvolumes (150^3, 300^3, and 400^3) because for the other sizes the number of subvolumes is insufficient. We have 42, 23, and 23 subvolumes for the sizes 150^3, 300^3, and 400^3. We fit a LASSO model to predict four fluid flow and transport properties: porosity ϕ, permeability k, anisotropy λ, and tortuosity τ. Because permeability and tortuosity are measured in $x, y,$ and z directions, we define the representative permeability as a geometric mean $(k_x * k_y * k_z)^{1/3}$ and the representative tortuosity as an arithmetic mean $(\tau_x + \tau_y + \tau_z)/3$. To decide the λ in the LASSO model, we train the model with 3000 repetitions. The ratio of training, validation, and test sets is 72%, 18%, and 10%, respectively.

Figure 4 summarizes prediction results. Porosity prediction results are fairly accurate at all scales, from 150–400. Permeability prediction results show the sudden increase in predictive accuracy at size 400. Anisotropy prediction results exhibit the gradual increase in predictive accuracy between sizes 300 and 400. The accuracy for anisotropy is not as good as for permeability. Tortuosity prediction result displays a sudden increase in predictive accuracy at size 400 as with permeability. At the sREV, the subvolume size of 400, we found that the models explained all four fluid flow and transport variables. Also, as we increase the subvolume size, the predictions tend to be more accurate.

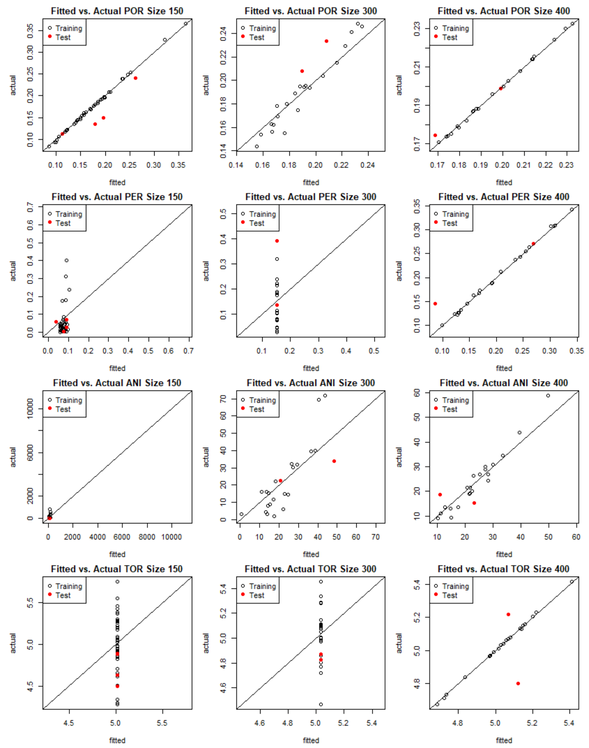


Figure 4: Prediction vs. actual plots of four geometric properties.

Acknowledgement

This material is based upon work supported by the U.S. Department of Energy, Office of Science: both the Basic Energy Sciences Program under contract DE-SC0006883, and the Office of Advanced Scientific Computing Research (ASCR), Applied Mathematics Program.

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

This research was supported in part by an appointment with the NSF Mathematical Sciences Summer Internship Program sponsored by the National Science Foundation, Division of Mathematical Sciences (DMS). This program is administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and NSF. ORISE is managed by ORAU under DOE contract number DE-SC0014664.

References

[1] Vanessa Robins, Mohammad Saadatfar, Olaf Delgado-Friedrichs, and Adrian P. Sheppard. Percolating length scales from topological persistence analysis of micro-CT images of porous materials. *Water Resources Research*, 52:315–329, 2016.

[2] Vanessa Robins, Peter John Wood, and Adrian P Sheppard. Theory and algorithms for constructing discrete Morse complexes from grayscale digital images. *IEEE Transactions on pattern analysis and machine intelligence*, 33:1646–1658, 2011.

[3] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, 13:600–612, 2004.

[4] Hongkyu Yoon and Thomas A. Dewers. Nanopore structures, statistically representative elementary volumes, and transport properties of chalk. *Geophysical Research Letters*, 40:4294–4298, 2013.