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

Chul Moon1, Scott A Mitchell2, Nickolas Callor3, Thomas A Dewers4, Jason E Heath5, Hongkyoon Yoon1, and Gregory R Conner4

1Department of Statistics University of Georgia email: chulmoon@uga.edu
2Sandia National Laboratories
3Brigham 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 ascending birth and a descending death time. 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, tortuosity, and tortuosity. The suggested framework efficiently predicts fluid flow and transport properties based on geometry and connectivity.

Analysis pipeline
1. Persistent homology computation and visualization
   - Original Images → Binary images → Transformed grayscale images → Cubic 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. Modelling
   - Prediction: Loadings → Fit a statistical model → Prediction
   - Classification: Loadings → Classification

Data
We analyze a FI-BSEM 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 we evaluated as a function of an increasingly-larger subvolume. These include porosity, permeability, tortuosity, and tortuosity. There are in total six different sizes of subvolumes: 150 × 300, 400 × 600, 500 × 800, 625 × 1000, 750 × 1250, and 3000 × 5000.

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 minimal. 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 REV by comparing persistence diagrams. We use the statistical similarity measure for image which is sensitive to the shift and scale differences while robust to perturbations. The SSIM index of two images x,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 x,y images and identical. We use the default setup so that

\[
\text{SSIM}(x, y) = \left(\frac{2\mu_x \mu_y + c_1 \sigma_x + c_2 \sigma_y + c_3}{\mu_x^2 + \mu_y^2 + \sigma_x^2 + \sigma_y^2 + c_1} \right) \left(\frac{2\sigma_{xy} + c_2}{\sigma_x \sigma_y + c_2} \right)
\]

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 vector persistence diagrams (image) and every persistence diagram (image) 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 show, give the biggest difference in Selma group chalk. There is no need for deciding REV using SSIM. We set the threshold to be 0.9 because Yoon and Dewers [4] finds that the subvolume size 40×40 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 diagrams of subvolumes. The principal components form a basis to represent the vectorized persistence diagrams:

\[\psi(k, l) = \sum_{i=1}^{n} c_{i,k} \phi_i(k) \]

After computing principal components, we select a subset of principal components that explain at least 5% 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 \(\mathbf{v} = (c_{1,k}, c_{2,k}, \ldots, c_{n,k})\) can be a proxy to the coefficients. LASSO is a penalized regression model using a \(\ell_1\) penalty. The result of LASSO can be obtained by solving:

\[
\min \left\{ \| \mathbf{Y} - \mathbf{X} \mathbf{b} \|_2^2 + \lambda \| \mathbf{b} \|_1 \right\}
\]

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 × 300, 400 × 600, and 500 × 800. We use cross-validation for the other sizes. The number of loadings obtained for all dimensions is 6, which is larger than the number of samples. 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 small n large p regression but give a penalty to the coefficients. LASSO is a penalized regression model using a \(\ell_1\) penalty. The result of LASSO can be obtained by solving:

\[
\min \left\{ \| \mathbf{Y} - \mathbf{X} \mathbf{b} \|_2^2 + \lambda \| \mathbf{b} \|_1 \right\}
\]

References