Topological Data Analysis for Learning Feature Extraction PhD proposal

Julien Tierny[1], Mélanie Plainchault[2]
[1] *CNRS, Sorbonne-Université, Paris, France*
[2] *Total SA, Pau, France*
julien.tierny@sorbonne-universite.fr, melanie.plainchault@total.com

Keywords: Topological Data Analysis, Feature Extraction, Machine Learning, Morse-Smale Complex.

1 CONTEXT

Total SA [1] is a French energy company producing and selling low carbon fuels, natural gas and electricity.

Its expertise covers a wide range of domains from battery design and lighting computation for solar electricity production to underground modeling. In all these domains, one of the main difficulty comes from the study of highly complex data sets whose sizes can be in the magnitude of 100 Go or 1To. This complexity requires, in order to properly understand and model the studied phenomena, to efficiently extract features of interest in the data set to handle both the data set size and complexity.

For battery life time maximization, a whole field of study comes from the understanding of the porous network of the electrodes whose complexity is critical, see Figure 1 for an example. Understanding the porous network effect on ions diffusion is particularly difficult, and has been studied for instance by Lagadec in [3], where topological invariants are used to characterize the tortuosity of the medium. Other studies have been realized at the molecular scale to understand that diffusion, in which the Morse-Smale Complex, an abstraction of Topological Data Analysis, has been used [4].

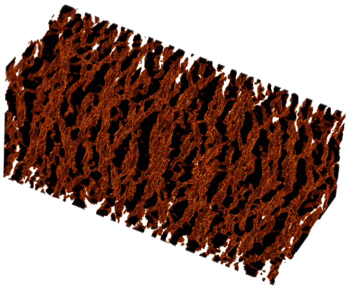

Figure 1: Volumic CT scan of an electrode sample showing the complexity and density of its pore network. [2]

.

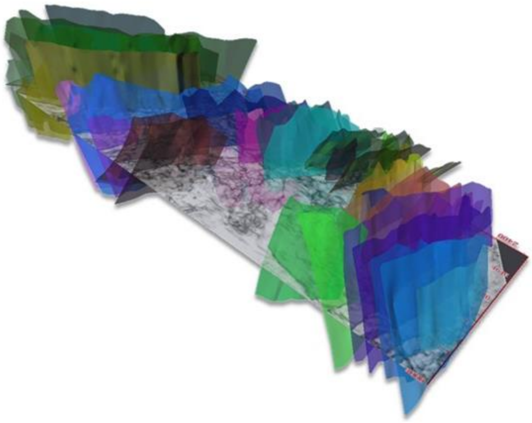

Figure 2: Fault model on a time-slice of the coherency seismic attribute, each fault surface is represented by a colored trimesh. *Courtesy of Sismage-CIG Team*

For underground modeling, a whole field study methodology is used in order to maximize the quantity of stored CO_2 in natural reservoirs. This methodology requires the design of a structural model of the underground, see for instance Figure 2. It is composed of horizons, corresponding to iso-time sediment deposition, and faults representing breaking events in the strati-graphic layers.

These horizons and faults are extracted by geophysicists manually picking through seismic images

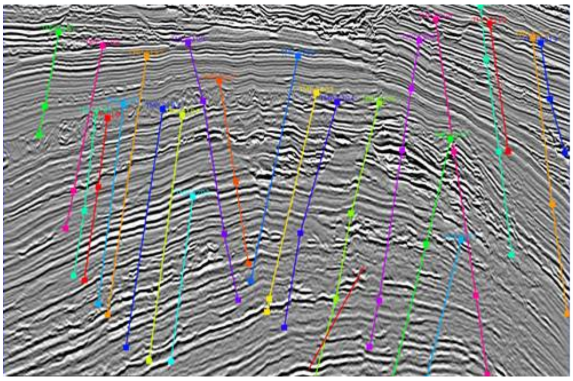

Figure 3: Fault sticks which have been picked by hand on a seismic image. Sediment horizons correspond to the black and white lines in the back ground, while faults appear as discontinuities in these horizons. Each stick corresponds to the presence of a fault on this seismic image. *Courtesy of Sismage-CIG Team*

presented on Figure 3.

This work is particularly tedious, and any automatic extraction method can be highly valuable to ease and accelerate this phase. For instance, David Hale in [5] proposes an automatic fault extraction 3D method from a fault likelihood metric which is based on crease surfaces [6]. Some other extraction method focus on using one or several seismic attributes in order to precisely define this fault position in the data set [7].

2 RESEARCH PROBLEM

In this project, we want to derive new algorithms for feature extraction, based on Topological Data Analysis and Machine Learning. This project is motivated by multiple applications at Total, including ions diffusion in porous material and fault extraction (described above). In this latter context, a fault probability presence volume data set which is represented on Figure 4 where high probabilities are yellow spots. Such probability volumes are produced by deep-learning algorithms in the context of a collaborative project involving Total and Google. While these algorithms manage to produce highly relevant fault presence probability estimates, their interpretation for geometrical analysis remains challenging. In particular, geophysicists would like to extract an explicit representation of these faults (in the form of a triangular surface, to perform various measurements on them: size, curvature, etc.) as well as a higher level understanding of their global structure (how faults intersect and connect together). However, no off-the-shelf algorithm exists for such a post-processing of these deep-learning results.

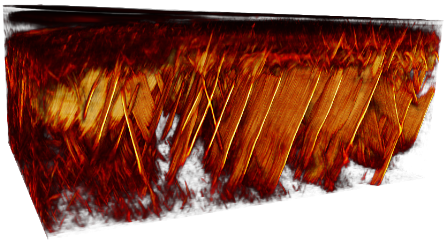

Figure 4: Volumic representation of fault probability (yellow stops represent the highest probability locations)

In this work, we want to explore how Topological Data Analysis [8] can be used for the exploitation of feature presence probability fields generated by deep-learning algorithms, in the context of fault extraction for geosciences. In particular, we would like to focus on the Morse-Smale complex [9], which is a topological object that is, in principle, well suited for the extraction of surfaces locally maximizing a scalar function [10] (here the presence probability field). While our preliminary experiments confirm the relevance of this research directions, many research questions remain open.

In this research work, we will focus on:

1. How to exploit the Morse-Smale complex to extract the network of fault surfaces or porous material:
   - Designing algorithms using machine learning for feature extraction, which would exploit topological objects (in the fault use-case, the 2-separatrices of the Morse-Smale complex, in the porous media example, the 1-separatrices) as a core data representation. In other words, the designed algorithms will use machine learing to learn which parts of the Morse-Smale complex separatrices exactly coincide with features of interest;
   - Designing algorithms to extract the global structure of the set of fault surfaces or porous network;
   - Developing, in collaboration with the scientists, new representations of networks of fault surfaces (possibly including physical properties) or pore network, as well as new methods for their interpretation.
2. How to make this approach scale for real-life datasets used at Total:
   - Designing algorithms for the simplification of the Morse-Smale complex, to account for the presence of noise in the probability fields;

- Designing algorithms capable of handling large-scale fields (hundred of gigabytes in size), possibly out-of-core or in a distributed manner;
- Designing time-efficient algorithms (possibly shared-memory parallel);
- Applying all the designed algorithms on real-life use cases.

3 ORGANIZATION

The Ph.D. thesis would last 3 years (maximum duration in France) and it could be organized as follows:

- Preliminary study:
  1. Review of the literature in Topological Data Analysis [8], especially regarding Persistent Homology and Morse-Smale complexes.
  2. Design of an algorithm (using machine learning) for the extraction of locally maximizing surfaces in a probability field (based on the Morse-Smale complex);
  3. Design of an algorithm for the extraction of the global structure of the set of maximizing surfaces;
  4. Experiments on manufactured synthetic examples created from a ground truth.
- Systematic study:
  1. In-depth review of the literature in Topological Data Analysis [8] (with a focus on Persistent Homology and Morse-Smale complexes);
  2. Preliminary use-case study on a selected real-life example of small size;
  3. Exploration of the following research questions:
     - How to make this approach more accurate? More time efficient?
     - How to make this approach scale to real-life data sets of large size?
  4. Full size case study on real-life data sets in collaboration with geophysicists;
  5. Exploration of the following perspective questions:
     - How to help pre-process the training data for the deep learning approach generating the probability field?
     - How to generalize this approach to other problems involving surface presence probability estimates (for example: density estimations of LIDAR point clouds)?

4 ENVIRONMENT

This PhD will be co-supervised by Julien Tierny [11] and Mélanie Plainchault, who already co-supervised a Ph.D. thesis previously, on topological methods for material sciences [12, 13, 14, 15]. It will be a CIFRE doctoral program promoting research collaboration between universities and companies, see [16] for more information.

Research time will be shared between the computer science department (LIP6) of Sorbonne University (downtown Paris – Jussieu subway station – France) and Total (Pau, France) in order to take benefit from both the academic environment and the feedback from end users, i.e. geoscientists. The balance between academic and company time is adjustable and will be decided in collaboration with the student.

This work will lead to publications and participations to international conferences (such as IEEE VIS [17]). Most of the developed code will be released open-source in the TTK library [18].

5 APPLICATION

We are looking for a highly motivated student, with strong C++ programming skills, a clear interest for Topological Data Analysis, Machine Learning and their applications, as well as a good English (spoken/written) level. Some background in geosciences would be a plus.

To apply, candidates are invited to send us their CV and a short cover letter by email to julien.tierny@sorbonne-universite.fr and melanie.plainchault@total.com.

REFERENCES

[1] https://www.total.com/.

[2] L. M. Francine, "Microstructure of celgard® pp1615 lithium-ion battery separator," https://www.research-collection.ethz.ch/handle/20.500.11850/265085.

[3] M. F. Lagadec, R. Zahn, S. Müller, and V. Wood, "Topological and network analysis of lithium ion battery components: the importance of pore space connectivity for cell operation," *Energy & Environmental Science*, vol. 11, no. 11, pp. 3194–3200, 2018.

[4] A. Gyulassy, A. Knoll, K. C. Lau, B. Wang, P. T. Bremer, M. E. Papka, L. A. Curtiss, and

V. Pascucci, “Morse-smale analysis of ion diffusion for dft battery materials simulations,” in *Topology-Based Methods in Visualization (TopoInVis)*, 2015.

[5] D. Hale, “Methods to compute fault images, extract fault surfaces, and estimate fault throws from 3d seismic images,” *Geophysics*, 2013.

[6] T. Schultz, H. Theisel, and H. P. Seidel, “Crease surfaces: From theory to extraction and application to diffusion tensor MRI,” *IEEE Transactions on Visualization and Computer Graphics*, 2009.

[7] M. Bahorich and S. Farmer, “3-d seismic discontinuity for faults and stratigraphic features: The coherence cube,” *The Leading Edge*, 1995.

[8] H. Edelsbrunner and J. Harer, *Computational Topology: an Introduction.*

[9] A. Gyulassy, P. T. Bremer, B. Hamann, and V. Pascucci, “A practical approach to Morse-Smale complex computation: Scalability and Generality,” *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2008.

[10] A. Gyulassy, N. Kotava, M. Kim, C. D. Hansen, H. Hagen, and V. Pascucci, “Direct Feature Visualization Using Morse-Smale Complexes,” *IEEE Transactions on Visualization and Computer Graphics*, 2012.

[11] J. Tierny, “Homepage,” https://julien-tierny.github.io/.

[12] M. Soler, “Large Data Reduction and Structure Comparison with Topological Data Analysis,” Ph.D. dissertation, Sorbonne University, 2019, https://hal.archives-ouvertes.fr/tel-02171190.

[13] M. Soler, M. Plainchault, B. Conche, and J. Tierny, “Topologically controlled lossy compression,” in *Proc. of IEEE PacificVis*, 2018, https://julien-tierny.github.io/stuff/papers/soler_pv18.pdf.

[14] ——, “Lifted Wasserstein Matcher for Fast and Robust Topology Tracking,” in *IEEE Symposium on Large Data Analysis and Visualization*, 2018, Best Paper Honorable Mention Award, https://arxiv.org/pdf/1808.05870.

[15] M. Soler, M. Petitfrere, G. Darche, M. Plainchault, B. Conche, and J. Tierny, “Ranking Viscous Finger Simulations to an Acquired Ground Truth with Topology-Aware Matchings,” in *IEEE Symposium on Large Data Analysis and Visualization*, 2019, Best Paper Award, https://julien-tierny.github.io/stuff/papers/soler_ldav19.pdf.

[16] ANRT, “CIFRE Program,” http://www.anrt.asso.fr/fr/cifre-7843.

[17] IEEE, “VIS conference,” http://ieeevis.org/.

[18] J. Tierny, G. Favelier, J. A. Levine, C. Gueunet, and M. Michaux, “The Topology ToolKit,” *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2017, Best Paper Honorable Mention Award, https://topology-tool-kit.github.io/.