Bertrand MICHEL
Professeur à l'Ecole Centrale Nantes - Nantes Université
Laboratoire de Mathématiques Jean Leray
Email : bertrand.michel@ec-nantes.fr

CENTRALE
NANTES

Nantes, 24/11/2024

Report on the Thesis « Probability Meets Topological Data Analysis »
Written by Niklas Hellmer

This Ph.D. thesis lies at the intersection of Probability, Statistics, and Applied Topology. The primary objective of the thesis is to develop Topological Data Analysis (TDA) methods that account for the stochastic nature of data. This overarching framework encompasses several challenging problems. First, when data is considered as being generated from a probability distribution, TDA signatures themselves become random objects, requiring tailored statistical treatment. Furthermore, this stochastic perspective on TDA involves constructing statistical tests based on topological signatures. It also includes the goal of defining approximating structures (e.g., simplicial complexes) and metrics that are sensitive to the underlying data distribution.

In this context, this thesis addresses these challenges from several perspectives:

1 -The thesis introduces and studies the Prokhorov distance for TDA to compare topological information.

2 -From a distribution comparison perspective, the thesis introduces a statistical hypothesis test to compare excess masses, from a statistical standpoint. Industrial applications of this approach are also described.

3 -The thesis proposes a new (bif-)filtration that incorporates both measures and distances to define simplicial complexes.

Structure of the Thesis

The thesis is split into seven chapters.

After the Introduction, **Chapter 2** introduces the main concepts necessary for the thesis: Prokhorov distances and other optimal transport metrics, the construction of filtrations using category theory formalism, as well as the fundamentals of persistent homology. It also remind recent results on the asymptotic behavior of the expected Euler characteristic curve (ECC) in the thermodynamic regime. This chapter is very dense, but it successfully presents the state of the art in TDA across several aspects essential to this thesis.

Chapter 3 introduces and studies a family of discrete Prokhorov distances for persistence diagrams. On the theoretical side, the chapter examines conditions under which such Prokhorov distances are extended pseudometrics. It also explores algorithms for computing the bottleneck profile and provides a computational complexity analysis of these methods. Numerical experiments suggest that these metrics are promising for data science applications, although selecting an appropriate admissible function for the Prokhorov metric can be challenging in practice.

Chapter 4 examines the ECC in the thermodynamic regime and demonstrates that it fully determines the excess mass, a key quantity in the field. This result complements a recent and significant finding by Vishwanath et al., which established the reverse implication. The proof is simple yet elegant.

Chapter 5 proposes a goodness-of-fit test - TopoTest - based on the Functional Central Limit Theorem of Krebs et al. on the asymptotic normality of Betti persistent numbers. The aim here is to test equality of a mass excesses. The chapter gives a consistency result for the test. The algorithmic implementation of the test, on the other hand, follows a non-asymptotic approach and is in fact disconnected from the results obtained in the chapter: it is a permutation test, which thus non-parametrically estimates the distribution of the statistic under H0. Numerical experiments suggest that the test is indeed powerful in many situations, notably in the univariate setting. It would have been relevant here to include in the comparisons the same permutation approach for the KS test (but also for an MMD-type test). Indeed, it is debatable whether the power comes from the topological nature of TopoStat, or from the fact that the permutation approach enables non-asymptotic calibration. A bootstrap approach could also be proposed for TopoTest. This being said, I think it is a very interesting chapter which also opens up avenues for future developments.

Chapter 6 proposes an application of TDA methods to the problem of bearing damage detection. The method relies on both Takens' reconstruction theorem and persistent homology representations, more specifically Betti curves, with distance measures used as metrics. The approach is highly successful and convincing. In particular, I find it interesting that TDA enhances the performance of state-of-the-art methods, such as those based on conditional variance, by combining the two approaches.

Chapter 7 introduces a new Dowker bifiltration that accounts for the mass of the relation in the construction. Dowker complexes are defined by considering the points in the intersections between two sets. This new bifiltration incorporates the size of these intersections. A stability result, sensitive to measures, is provided, taking the form of a bound on the interleaving distance by the Hausdorff and Prokhorov distances. As a consequence, a consistency result is obtained, and rates of convergence could likely be derived here. The chapter also elaborate on the computational aspects of this new bifiltration. Several numerical experiments are presented, demonstrating the robustness of this approach, particularly when compared to the degree-Rips bifiltration. This part of the chapter is brief but sufficient to highlight the potential of the method.

Recommendation

The thesis is well-structured, well-balanced, and enjoyable to read. Its positioning at the intersection of statistics, probability, and applied topology is timely and results in relevant contributions to TDA. The numerous contributions of the thesis address both theoretical and computational aspects. I believe the thesis reflects the candidate's deep understanding of the methods and challenges in the field. Given the relevance and quality of the contributions presented, I ***deem the thesis as sufficient to grant a PhD and provide a positive recommendation for proceeding with the defense of the thesis.***

Bertrand Michel