Siddharth Vishwanath

Two Sigma Ph.D. Fellowship Proposal

✉ suv87@psu.edu

☎ +1 (814)-876-2068

The advent of information and computational technology has provided a wealth of modern statistical techniques, enabling us to examine data from unconventional sources, such as text and images, with statistical information often concealed in low-dimensional features embedded in higher dimensional space. My broad research goal is to use tools from geometry and topology in statistical learning, to uncover meaningful statistical information using efficient, robust, and theoretically rigorous methodology.

To this end, Topological Data Analysis (TDA) has emerged as a framework to infer geometric and topological features of complex data. TDA is particularly advantageous, as it is agnostic to the representation of data. In a nutshell, given a sample $\mathbb{X}_n = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$, the shape of $\mathbb{X}_n$ at resolution $\epsilon \in (0, \infty)$ is encoded in a *simplicial complex*, $\mathcal{K}_\epsilon$, obtained by looking at balls of radius ϵ centered at the points of $\mathbb{X}_n$. As ϵ varies, topological features of various dimensions (i.e., connected components, loops, holes, etc.) are born, or existing ones die. This **multiscale** information is summarized in a *persistence diagram*, $\mathsf{Dgm}(\mathbb{X}_n)$.

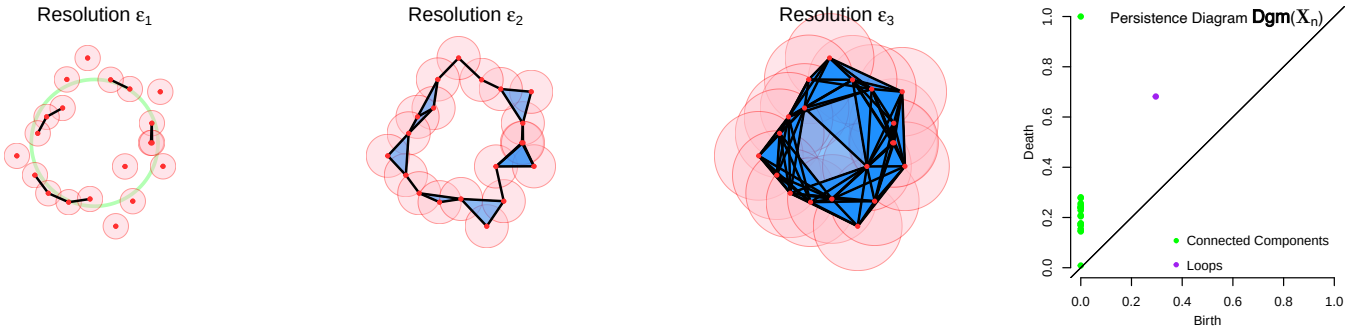

FIGURE 1: Points are sampled from a circle. As ϵ increases, the topological information at each resolution is encoded in the simplicial complex $\mathcal{K}_\epsilon$. $\mathsf{Dgm}(\mathbb{X}_n)$ depicts the evolution of topological features, indicating the presence of a significant circular feature in the data.

The adoption of TDA in mainstream statistical methodology is still limited. In [1, 2], we shed light on the behaviour of random topological quantities through the lens of classical statistics (i.e., asymptotic sufficiency, identifiability and inference). We study conditions under which the statistical behaviour of the topological summaries can be used to distinguish between random processes.

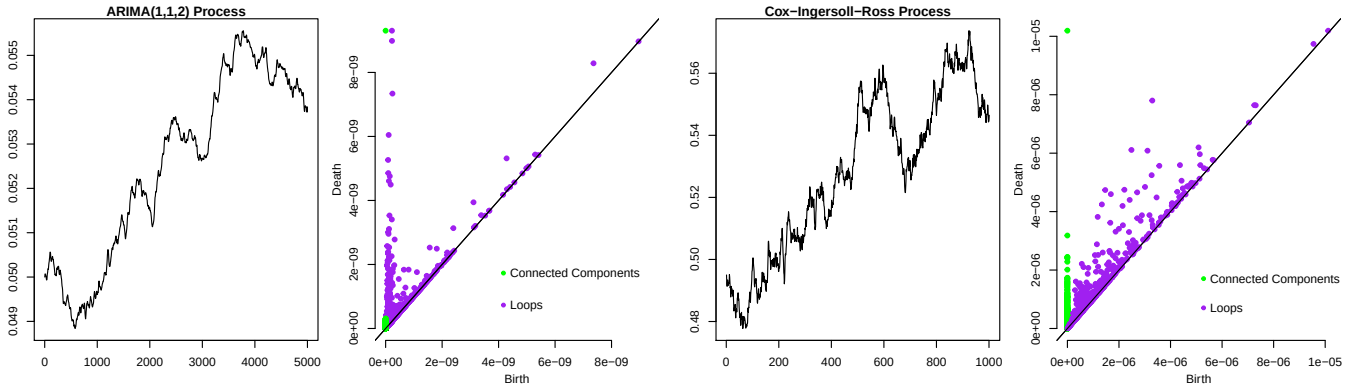

FIGURE 2: (Left) The ARIMA(1,1,2) persistence diagram has loops concentrated in a "V"-shaped region – characteristic of processes without stochastic volatility. (Right) The CIR persistence diagram has some well-defined loops away from the diagonal – characteristic of stochastic processes with mean-reversion.

For example, in Figure 2, the persistence diagrams obtained from the time-delay embeddings of an ARIMA(1,1,2) and CIR process are easily distinguishable. In the future, I would like to extend these ideas to develop general nonparametric tests for distinguishing between different random processes; potentially even for more nuanced processes, e.g., the CIR from the Hull-White process. This would be useful for understanding the patterns which underlie many real world-processes, and choosing appropriate models to generate insights from.

A significant limitation of TDA is that the persistence diagram from sample points, $\mathsf{Dgm}(\mathbb{X}_n)$, is highly sensitive to outliers. In [3] (to appear in NeurIPS 2020), we develop methodology to approximate $\mathsf{Dgm}(\mathbb{X}_n)$ using an outlier robust persistence diagram, $\mathsf{Dgm}(f_n)$, constructed using a sample function f_n. Here, given a function $f : \mathbb{R}^d \to \mathbb{R}$, the shape of f at scale ϵ is encoded in the *superlevel sets*, $f^{-1}([\epsilon, \infty))$. By varying ϵ, the shape of the function f can be summarized in the persistence diagram $\mathsf{Dgm}(f)$.

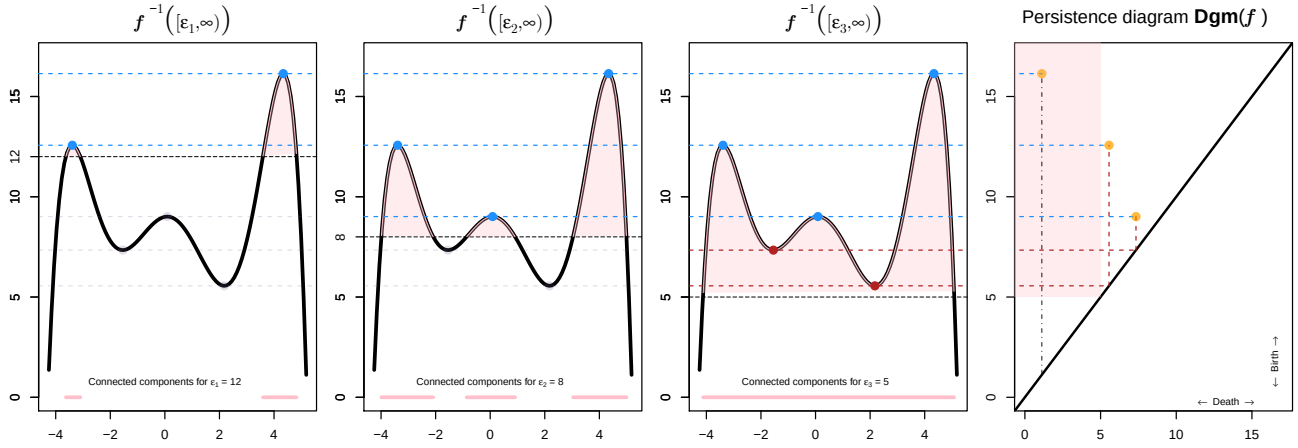

FIGURE 3: (Left) $f^{-1}([\epsilon, \infty))$ corresponds to the pink regions. In $\mathbb{R}^1$, the topological features are just connected components, but generalizes to higher-dimensional features in $\mathbb{R}^d$. (Right) For $\mathsf{Dgm}(f)$, new features are born at local maxima, and die at local minima.

In [3], the sample function f_n used is a robust nonparametric estimator obtained as the solution to an optimization problem in an infinite dimensional space. We show that this methodology provably improves robustness without compromising on statistical efficiency. Persistence diagrams are gaining popularity in finance, especially for change-point detection (e.g., [4–6]). Many such practical applications would arguably benefit from robust topological machinery, where I hope to make some contributions. Additionally, this has piqued my interest in robust nonparametrics & machine learning. I am keen to investigate how these ideas (i.e., robust losses, and/or topologically-inspired penalties) can be incorporated in existing deep learning and kernel methods to enhance their performance. This would be a promising methodology for "generating alpha" by identifying patterns in highly contaminated and complex data.

I am also currently working on incorporating ideas from differential geometry in posterior sampling and MCMC. In many applications, the energy barrier between modes is too large for existing methods such as Hamiltonian and Langevin Monte-Carlo to overcome for reasonable sample sizes. By incorporating additional geometric information of the target distribution, we can propose more efficient moves. Decidedly, this would be useful for many applications, e.g., multivariate time-series with cointegration. I hope to be able to explore such applications in future work.

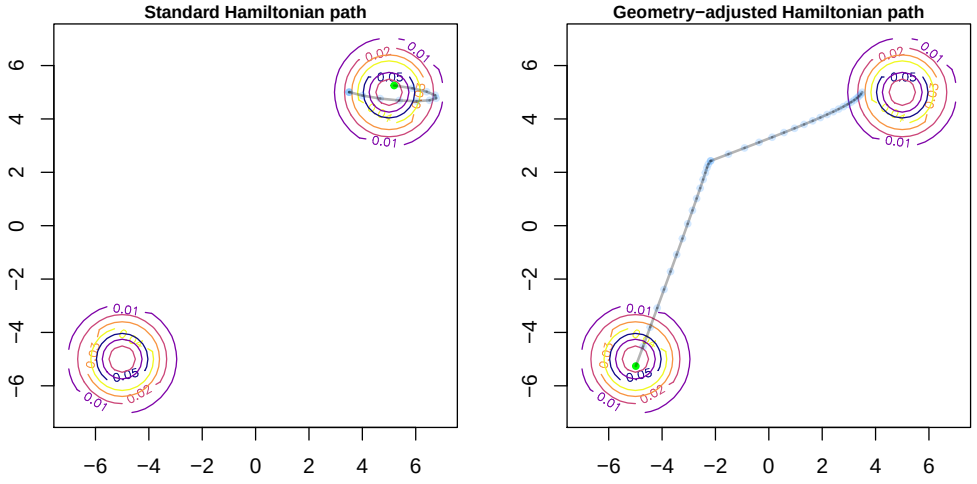

FIGURE 4: Incorporating the geometry improves proposals for target distributions with multiple modes.

REFERENCES

[1] Siddharth Vishwanath, Kenji Fukumizu, Satoshi Kuriki, and Bharath Sriperumbudur. Statistical invariance of Betti numbers in the thermodynamic regime. *arXiv preprint arXiv:2001.00220*, 2020.

[2] Siddharth Vishwanath, Kenji Fukumizu, Satoshi Kuriki, and Bharath Sriperumbudur. Statistical invariance of Betti numbers. *Algebraic Topology: Methods, Computation, and Science Conference*, 2020. URL https://www.youtube.com/watch?v=9K2ynjC5R0c.

[3] Siddharth Vishwanath, Kenji Fukumizu, Satoshi Kuriki, and Bharath Sriperumbudur. Robust persistence diagrams using reproducing kernels. *arXiv preprint arXiv:2006.10012*, 2020 (To appear in NeurIPS 2020).

[4] Marian Gidea and Yuri Katz. Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications*, 491:820–834, 2018.

[5] Kwangho Kim, Jisu Kim, and Alessandro Rinaldo. Time series featurization via topological data analysis. *arXiv preprint arXiv:1812.02987*, 2018.

[6] Yuhei Umeda. Time series classification via topological data analysis. *Information and Media Technologies*, 12: 228–239, 2017.