An Empirical Analysis of Topological Persistence as a Supplementary Measure of Dataset Drift

Christopher Shultz

The Hartford Financial Services Group

Hartford, CT, USA

c.shultz@live.com

Author Contributions: Single-author, 100% attribution.

Abstract

This paper proposes the utilization of Persistence Entropy (PE), a concept from Topological Data Analysis (TDA) as a supplementary method for detecting dataset drift. We present a brief review of dataset drift detection, and describe the potential benefits of utilizing PE to capture changes in the "shape" of the dataset over time, which are unobservable via traditional methods. Through a numerical experiment, we demonstrate that the proposed method both captures drift effectively and scales with the magnitude of the drift injected.

1: Introduction

Data drift refers to systematic changes in a dataset over time, which can adversely affect model performance. This phenomenon is typically categorized into two types: covariate shift, where the data distribution evolves over time while the relationship between inputs (X) and outputs (Y) remains constant; and concept drift, where the underlying relationship itself changes (Mallick et al., 2022). Our study focuses on covariate shift. If unaddressed, covariate shift can lead to significant declines in model accuracy.

Various methods exist for detecting covariate shift, each with distinct advantages and limitations. Statistical tests, such as the Kolmogorov-Smirnov and Chi-Squared tests, compare distributions of current and historical data to detect changes. Although these tests are straightforward and interpretable, they often fail to capture complex multivariate relationships. Alternatively, measures like the Population Stability Index (PSI) offer a more detailed view but may require domain-specific tuning and interpretation.

In this paper, we propose using Topological Data Analysis (TDA) as a supplementary approach for dataset drift detection. Specifically, we explore persistent homology to assess the "stability" of the data's shape over time. TDA examines the geometric and topological features of a dataset, providing insights not typically captured by traditional methods. Persistent homology, for instance, investigates the multi-scale topological characteristics of a dataset, such as connected components, holes, and voids.

Persistent Entropy (PE) is a metric derived from persistent homology that quantifies the complexity and variability of a dataset's topological features. It encapsulates information about the birth and death of these features in a persistence diagram, offering a concise summary of the data's topological landscape. By monitoring changes in PE over time, we can detect subtle structural alterations in the data indicative of drift. This method is particularly valuable as a supplementary measure to traditional metrics, as it identifies higher-order interactions and dependencies that might otherwise be overlooked.

To validate this approach, we conduct an empirical study using a simulated dataset. We introduce drift by applying various "shocks" to the data distribution and then plot the persistent entropy across the dataset's timeline. Our findings demonstrate that (a) persistent entropy consistently captures the introduced drift through notable changes, and (b) the magnitude of these changes aligns with the severity of the applied shocks. The results underscore the sensitivity and robustness of persistent entropy in detecting dataset drift.

This paper is structured as follows: Section 2 reviews related work on dataset drift detection and TDA. Section 3 outlines the theoretical foundation of persistent entropy and its application in drift detection, along with details of an empirical case study. Section 4 presents the results and discusses the findings. Finally, Section 5 concludes with insights into the implications of this work and potential future research directions.

By introducing persistent entropy as a supplementary drift detection mechanism, this paper aims to provide a novel approach that leverages topological information to enhance the detection of dataset drift, offering a more comprehensive toolkit for maintaining the reliability and performance of predictive models in dynamic data environments.

2: Background

The most basic form of dataset drift is called *covariate shift*, which occurs when data is generated via some model $P(y|x)P(x)$ and the distribution $P(x)$ changes over time. This is referred to as covariate shift because only the covariate distribution changes (Quinonero-Candela et al., 2009). Described another

way, covariate shift occurs when the mapping from inputs to output is shared by the source and target data, but the distribution of the inputs varies (Chen et al., 2016). The precise definition of covariate shift has been subject to some debate, and in a review paper aiming to unify terminology, Moreno-Torres et al. (2011) settle on a definition that aligns with the aforementioned explanation.

The other forms of drift are [a] *prior probability shift,* in which changes occur within the distribution of the variable y; and [b] *concept shift*, in which the relationship between the input and output variables changes from its prior form. This paper focuses exclusively on the simpler covariate shift problem. While simple in theory, the consequences of covariate shift are significant and prevalent in real-world applications. Theoretical justification for most predictive models relies on an assumed equality of the distributions from which "old" and "new" data arise (Tripuraneni et al., 2021). Without that equality, a core assumption is violated, and it is paramount to understand the potential impact.

The most commonly cited impact of covariate shift is a general degradation of model performance metrics such as accuracy, precision, recall, etc. In empirical analyses, the performance of classifiers drastically improves after covariate shift is detected and corrected for (Dharani et al., 2019). Aside from the core problem of performance degradation, depending on the specific use case, real world consequences can arise such as poor decision making, financial losses, and reduced trust in models by stakeholders.

Rather than provide a complete review of dataset drift detection mechanisms, we provide a brief overview resting on the foundations set forth by other papers. In general, statistical approaches like the Kolmogorov-Smirnov test and the Chi-Squared test compare distributions directly, whereas more generalizable tools like the Population Stability Index (PSI) monitor the data distributions directly over time. There are subtle differences within these two approaches.

The statistical methods utilized for drift detection are fundamentally about testing the equality of the distributions of data during and after training. Rabanser et al. (2019) provide a thorough review of statistical methods, outlining their empirical structure, benefits, and weaknesses.

Perhaps most commonly used, the Population Stability Index (PSI) measures the amount of change in a population based on a single variable, quantifying the change of the fraction of entities therein at several possible values/ranges (Haas and Sibbald, 2024). The PSI is usually stated as

$$PSI = \sum_{bins,i} \left(f_{1,i} - f_{0,i}\right) \cdot \ln\left(\frac{f_{1,i}}{f_{0,i}}\right)$$

With $\left(f_{0,i}, f_{1,i}\right)$ representing the fraction of the entities in bin i in the original and new populations, respectively. Kurian and Allali (2024) frame PSI as a variant of KL divergence, wherein given two probability distributions the KL divergence measures the "excess surprise" in using the actual distribution vs the expected distribution. They point out the core limitation that KL divergence is not symmetric (i.e. given two datasets Q, P, the $D_{KL}(Q||P) \neq D_{KL}(P||Q)$). PSI resolves this problem by modifying KL divergence into a symmetric measure.

While such methods have been widely adopted, they do not paint the entire picture, particularly failing to capture higher-dimensional information / relationships related to the shape and structure of the dataset. Viewing the data as a multidimensional "object" with a given shape and form, methods from TDA may provide a framework for capturing previously unobservable information in monitoring for drift. In brief, the question we want to ask is: *does the shape/structure of the multidimensional dataset change over time, and can that be quantified?*

The general framework employed in many applications of Topological Data Analysis (TDA) is to represent a set of data as a "point cloud", a point cloud as a simplicial complex, and then the extraction of topological information from a filtration of simplicial complexes across varying levels of resolution. This TDA pipeline is summarized in full by Shultz (2023).

At a high level, data is viewed as a set of discrete points in some space (a point cloud). That point cloud is a noisy representation of some underlying structure that can be estimated (Carlsson, 2009). To do so, we consider the construction of a simplicial complex, which is a "smoothed" representation of our point cloud, viewed as the vertices of a combinatorial graph whose edges are determined by some proximity measure which defines the "resolution" of the complex via the parameter ϵ (Lum et al., 2013).

The development of simplicial complexes at various levels of scaling (ϵ) creates a *filtration* of complexes $\emptyset \subseteq \mathbb{X}_0 \subseteq \cdots \subseteq \mathbb{X}_n \subseteq \mathbb{X}$. In this context, we consider the "birth" and "death" values of the topological features that emerge (e.g. connected components, holes, voids), and the values of resolution over which they persist (Gidea and Katz, 2018). Persistence diagrams encode this information visually to display the birth and death pairs for each observed topological feature.

For drift detection, we propose the use of the *persistence entropy* measure (Munch et al., 2019), which generates a vector of quantified entropy representing the amount of order/disorder of the underlying topological object over time. E.g. with a series of data X on some time index, we can compute, based on an arbitrary rolling window size (e.g. $w = 10$), a data frame showing the values of each column / row for the subset selected. We can then apply this rolling window the dataset and create an array of point clouds to be first established, then converted to simplicial complexes, then converted to persistence diagrams, and finally to lead into a persistent entropy calculation. More formally, given a persistence diagram of birth-death dimension triples (b,d,q), persistence entropies are calculated as the base 2 Shannon entropies of the collections of differences (d-b) "lifetimes", normalized by the sum of all such differences (Munch et al., 2019).

Methods from TDA allow us to capture higher order interactions and dependencies; provide an enhanced understanding of complex structural changes in data; and the PE method provides a quantitative "measurement" of topological complexity, revealing new information that is otherwise unobservable in traditional metrics used for covariate shift assessment.

To the authors' knowledge, no existing research examines the utilization of persistence entropy as a potential supplementary method for the detection of dataset drift.

3: Methodology

To directly explore the question of whether persistence entropy (PE) can capture dataset drift, we simulate a test dataset for experimentation. The simulated data $X_{1000\times3}$ consists of three standard normal variates, then modified with a varying set of "shocks" to change the distribution of the underlying dataset over time. These shocks are arbitrarily selected real numbers $s \in [0.001, 0.01, 0.1, 1, 5, 10, 20, 30]$.

By looping through this vector and generating 8 supplementary datasets with the same data generating process as X, but with x_1 shocked by the selected parameter and $n = 400$, we create a new dataset X_s, which is appended to the end of X to form a drifted set of data X_d where the first 1000 observations follow the original data generating process and the remaining 400 are drifted in x_1 per the selected parameter. Note that the shock parameter s is applied to the μ and σ parameters multiplicatively, and since $X \sim N(0,1)$, we are primarily shifting the spread of the data around 0 for x_1 alone. When $s = 1$, this is the same as a "no drift" scenario, as we will still result in a x_1 that is distributed $N(0,1)$.

For each of the 8 generated datasets, we loop through the time index via a rolling window of arbitrarily selected size $n = 30$. Each window constitutes a "point cloud" of observations across x_1, x_2, x_3 on which simplicial complexes are formed via the Vietoris-Rips algorithm. Persistence diagrams are computed for each window to track the birth-death resolutions of the topological features therein. The persistence entropy calculation quantifies the entropy of the persistence diagrams over the sliding time horizon. Intuitively, if the distribution of the underlying dataset changes, we should expect some change to the quantified entropy around the same time.

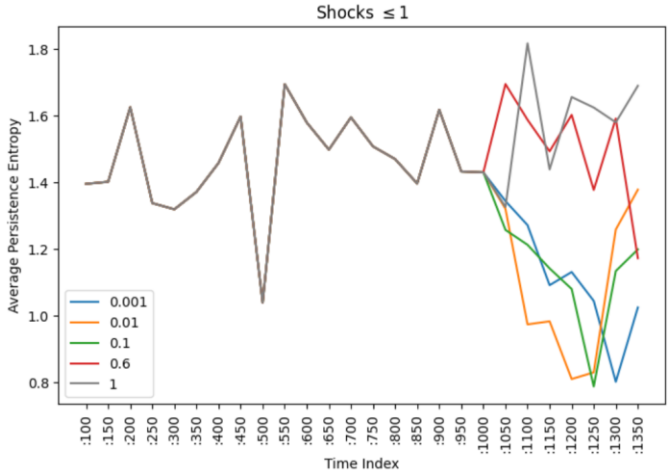
4: Results

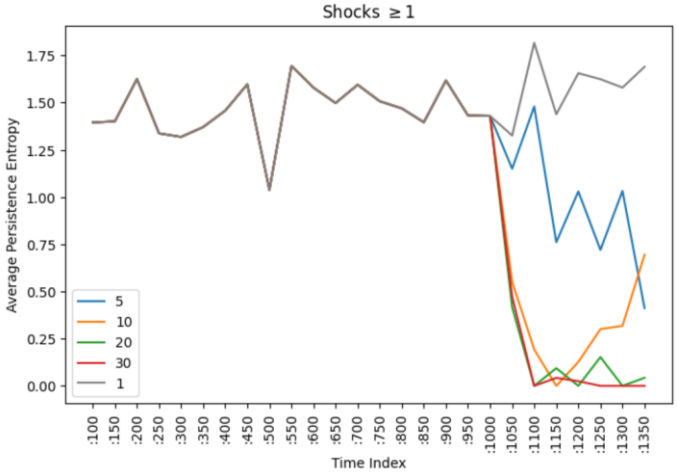
Table 1 provides a summary of the datasets utilized in our analysis, each with 1400 observations, the first 1000 of which are non-drifted / identical.

Table 1: Dataset Summary Statistics

shock	***mean***	***min***	***max***	***var***
0.001	0.00002	-0.0029	0.0032	0.000001
0.01	0.00006	-0.0318	0.0311	0.000102
0.1	-0.00487	-0.2713	0.2780	0.009771
0.6	-0.00203	-2.2130	1.7191	0.386608
1	-0.00069	-2.5209	2.2409	0.880971
5	-0.03321	-14.0531	14.0069	23.105002
10	-0.22916	-30.3399	26.0248	111.017928
20	2.07924	-49.4133	68.5782	429.601690
30	0.02698	-99.8851	72.5507	970.806298

Examining the resulting PE vectors (below) and their averages over 50-length blocks (for easier visualization), we observe that the average PE value demonstrates two interesting phenomena. First, the quantified PE drops when drift is added to the dataset, regardless of whether that drift is greater than or less than 1. Second, the magnitude of the observed change in the PE metric scales with the distance of the applied shock from 1. This means that small changes in topological structure are likely to go unnoticed, whereas large changes should be visibly obvious. It is also possible to apply quantitative thresholds to the quantified drift measure, for example by computing the historical percentiles and raising a flag when the PE breaches some predefined thresholds, such as the 5^{th} and 95^{th} percentiles.





While this is an interesting and promising result, its core limitation is the numerical experiment utilized. The intuition behind such an approach flows naturally, but the literature may benefit from a more formal mathematical analysis of the underlying mechanics to understand the bounds and limitations of this method's ability to capture various types of drift and data conditions.

5: Conclusion

In this paper, we introduced Persistence Entropy (PE), a measure from Topological Data Analysis (TDA), as a supplementary method for detecting dataset drift. Through our empirical analysis, we demonstrated that PE effectively captures drift introduced into a dataset, showing consistent changes corresponding to the magnitude of the applied drift. Our findings indicate that PE provides a robust and sensitive measure of structural changes in data, complementing traditional statistical methods and distribution monitoring tools.

Our results support the hypothesis that PE can detect subtle and complex topological changes in the dataset's structure that might be overlooked by conventional methods. By tracking PE over time, practitioners can gain deeper insights into the evolving data landscape, enhancing their ability to maintain model performance and reliability in dynamic environments.

The implications of this work suggest that integrating TDA into existing drift detection frameworks can provide a more comprehensive toolkit for addressing dataset drift. Future research could explore the application of PE in various real-world scenarios, assess its scalability to larger datasets, and refine the method for specific types of drift.

Overall, this study contributes to the growing body of literature on dataset drift detection, offering a novel approach that leverages topological information to enhance the detection and understanding of dataset drift.

References

Carlsson, G. 2009. Topology and Data. *Bulletin of the American Mathematical Society.* 46(2): 255-308.

Chen, X., Monfort, M., Liu, A., and B. Ziebart. 2016. Robust Covariate Shift Regression. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS) 2016.* Cadiz, Spain.

Dharani, G., Nair, N., Satpathy, P., Christopher, J. 2019. Covariate Shift: A Review and Analysis on Classifiers. *2019 Global Conference for Advancement in Technology. Bangalore, India.*

Gidea, M. and Y. Katz. 2018. Topological Data Analysis of Financial Time Series: Landscapes of Crashes. *Physica A: Statistical Mechanics and Its Applications.* 491.

Haas, M. and L. Sibbald. 2024. Measuring Data Drift with the Unstable Population Indicator. *Data Science.* 7: 1-12.

Huang, Y., Yuan, Z., Leung, C., Wu, Q., Ma, S., Wang, S., Wang, D., and Z. Huang. 2023. Towards Balanced Representation Learning for Credit Policy Evaluation. *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023.* Valencia, Spain.

Kurian, J. and M. Allali. 2024. Detecting Drifts in Data Streams Using Kullback-Leibler (KL) Divergence Measure for Data Engineering Applications. *Journal of Data, Information, and Management.*

Lum, P., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J. and G. Carlsson. 2013. Extracting Insights from the Shape of Complex Data Using Topology. *Scientific Reports.* 3(1): 1-8.

Mallick, A., Hsieh, K., Behnaz, A., and G. Joshi. 2022. Matchmaker: Data Drift Mitigation in Machine Learning for Large-Scale Systems. *Proceedings of the 5th MLSys Conference,* Santa Clara, CA, USA.

Moreno-Torres, J., Raeder, T., Alaiz-Rodriguez, R., Chawla, N., and F. Herrera. 2011. A Unifying View on Dataset Shift in Classification. *Pattern Recognition.* 45: 521-530.

Munch, E., Myers, A., and F. Khasawneh. 2019. Persistent Homology of Complex Networks for Dynamic State Detection. *Physical Review E.* 100.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and N.D. Lawrence. 2009. *Dataset Shift in Machine Learning.* The MIT Press.

Rabanser, S., Gunnemann, S., and Z. Lipton. 2019. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. *33rd Conference on Neural Information Processing Systems.* Vancouver, CAN.

Shultz, C. 2023. Applications of Topological Data Analysis in Economics. *SSRN.* https://ssrn.com/abstract=4378151.

Tripuraneni, N., Adlam, B., and J. Pennington. 2021. Overparameterization Improves Robustness to Covariate Shift in High Dimensions. *35th Conference on Neural Information Processing Systems.*