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

Christopher Shultz

**Abstract**

This paper proposes the utilization of Persistence Entropy, a concept from Topological Data Analysis (TDA) as a supplementary method for detecting dataset drift. We present a review of TDA and summarize the proposed method, which allows for the capture of topological changes in the underlying “shape” of a dataset, which may be missed by traditional methods. We conduct a demonstration on simulated data to show that the proposed method [a] captures drift injected into the dataset consistently; and [b] scales with the size of the drift injected.

# 1: Introduction

Dataset drift is the phenomenon that manifests as systematic changes to a dataset occur over time, which can significantly degrade model performance. It is often described as a mismatch between the training and test data which can be classified into two groups: *covariate shift*, in which the data distribution changes over time, but the underlying mapping (concept) from remains; and *concept drift*, where the underlying concept changes over time (Mallick et al., 2022). This paper focuses on the question of covariate shift.

A number of methods exist for detecting dataset drift, each with its own pros/cons. Statistical tests such as the Kolmogorov-Smirnov and Chi-Squared Test compare the distributions of current and historical data to identify changes. Such tests are simple and interpretable, but are traditionally limited to univariate analysis and fail to capture complex multivariate relationships. Direct measurements of the data distribution over time like the Population Stability Index (PSI) and Kullback-Leibler Divergence can show a more granular view, but may require domain knowledge to set appropriate thresholds and are sensitive to hyperparameter decisions.

We propose Topological Data Analysis (TDA) as a supplemental method for dataset drift detection, particularly methods that explore the concept of persistent homology. TDA focuses on the intrinsic geometric and topological features of a dataset, and can thus provide unique insights that aren’t otherwise captured by traditional drift mechanisms. Persistent homology studies the multi-scale topological features of a dataset such as its connected components, holes, and voids.

Persistent Entropy (PE) is a measure derived from persistent homology that quantifies the complexity and variability of the topological features of a dataset. It encapsulates information about the birth and death of these features in a persistence diagram, providing a compact summary of the data’s topological landscape. By tracking changes in persistence entropy over time, we can detect subtle structural changes in the data that may indicate drift. This approach is particularly useful as a supplemental measure, as it captures higher-order interactions and dependencies that are often missed by conventional methods.

To test this theory, we conduct an empirical exercise in which we develop a simulated dataset, apply a range of “shocks” to the data distribution to introduce drift, and plot the persistent entropy across the horizon of the dataset, demonstrating that [a] persistent entropy captures the introduced drift by either dropping or spiking; and [b] the magnitude of the drop or spike is consistent with the magnitude of the shock applied, centered around 1. Results are presented to highlight the sensitivity and robustness of persistence entropy in identifying dataset drift.

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

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 evolving data environments.

# 2: Background

The most basic form of dataset drift is called *covariate shift*, which occurs when the data is generated via some model and where the distribution changes over time. This is called 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 themselves may vary (Chen et al., 2016). There has been some debate about the exact definition of covariate shift, and in a review paper aiming to unify terminology, Moreno-Torres et al. (2011) settle on a definition that aligns with our aforementioned explanation. The other commonly discussed types of drift are [a] *prior probability shift*, in which changes occur within the distribution of the variable ; and [b] *concept shift* is when the relationship between the input and output variables changes. In this paper, we focus on the covariate shift problem exclusively.

The consequences of covariate shift are significant, and it is prevalent in real-world applications, though a theoretical understanding of exactly *why*. At the core of the issue, theoretical justification for most predictive models relies on the assumed equality of the distributions from which training and test data are pulled (Tripuraneni et al., 2021). The most commonly cited impact of covariate shift on a trained model is degradation of model performance metrics such as accuracy, precision, recall, etc. In empirical analyses, performance of classification models drastically improves after covariate shift is corrected for (Dharani et al., 2019). Aside from the core modeling problem of degraded performance, depending on the application, real world consequences can arise such as poor decision making, financial losses, and reduced trust in predictive models by stakeholders.

#### Traditional Methods for Dataset Drift Detection

Rather than providing a complete review of dataset drift detection mechanisms, we focus on the two core categories: statistical tests like the Kolmogorov-Smirnov Test and the Chi-Squared test, which compare distributions; and data distribution monitoring via tools like the Population Stability Index (PSI) or Kullback-Leibler divergence.

The statistical methods utilized for drift detection are fundamentally about testing the equality of continuous univariate probability distributions. In particular, the two-sample KS test, checks whether two samples come from the same distribution.

KEEP GOING HERE. NEED TO SUMMARIZE KS test next.

Summarize statistical methods: KS Test, Chi-Squared Test, Others.

Summarize distribution monitoring: PSI, KL Divergence, Others.

#### Topological Data Analysis

Describe TDA, concepts/motivation, understanding the shape / structure of data.

Define and explain point clouds to simplicial complexes (VR, Cech), and the concepts of homology (features like connected components, holes, and voids) and persistent homology (across smultiple scales). Construction and interpretation of persistence diagrams.

Persistent entropy – define and discuss. Role of PE in quantifying the complexity and variability of topological features. Viewing it as the “stability” of the topological object. How persistence entropy captures information about the data's topological landscape.

#### New Information Captured

Higher Order Interactions: How persistence entropy reveals complex relationships and dependencies in data that traditional methods miss. Examples of higher-order interactions: Multivariate dependencies, topological anomalies.

Sensitivity to Structural Changes: Persistence entropy's sensitivity to subtle structural changes in data. Comparison with traditional methods: Enhanced detection of gradual and abrupt drifts.

Quantitative Measure of Topological Complexity: How persistence entropy provides a quantitative measure of the data's topological complexity. Applications in different domains: Detecting shifts in biological data, changes in financial markets, etc.

#### Existing Studies on TDA and Drift Detection

Has anyone studied TDA / Drift Detection together? If so, discuss/explain.

Identification of Gaps in the Current Research: No one has done this, why? Opportunities for future exploration.

This detailed background section will provide a comprehensive foundation for understanding the motivation, methodology, and potential impact of using persistence entropy as a supplementary drift detection mechanism. By covering the theoretical underpinnings, practical applications, and existing research, the section will set the stage for presenting your experimental validation and results.

# 3: Methodology

To explore the question of whether persistent entropy can capture drift, we simulate a dataset of normal variates, and artificially introduce drift via a number of “shocks.” We consider a dataset , where each is a normally distributed random variable with mean , standard deviation and length .

We then apply a series of shocks by looping through this vector and generating a supplementary dataset with the same structure as , but with “shocked” by the selected parameter and . This new dataset is appended to the end of to form a drifted dataset , in which drift is introduced by factor as of the 1001st observation. Also note that the shock parameter is applied to the and parameters multiplicatively, meaning that when , no drift is injected into the system and this serves as a “control” instance.

For each computed dataset, we loop through the time index via a rolling window approach, utilizing an arbitrarily selected window size of 30 observations. Each window constitutes a “point cloud” of observations across on which simplicial complexes are formed via the Vietoris-Rips algorithm. Persistence diagrams are then computed for each window, to track the birth and death resolutions of topological features therein. The persistence entropy calculation quantifies the “entropy” or “stability” of the persistence diagrams over the sliding time horizon. Our hypothesis is that when plotted against the time horizon, this quantified persistence entropy will spike or drop when drift is introduced.

# 4: Results

Dataset Summary:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***shock*** | ***avg*** | ***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 |
| BASELINE | 0.01933 | -3.2413 | 3.8527 | 0.957900 |

We observe that when plotted across the time horizon, the average persistence entropy value demonstrates two important phenomena: First, quantified PE drops when drift is added to the dataset, regardless of whether that drift is in or ; 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 are likely to go unnoticed, while large changes are visually obvious.

A graph with different colored lines

Description automatically generated

A graph of a graph showing different colored lines

Description automatically generated with medium confidence

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 value, such as the <5th or >95th percentiles.

# 5: Conclusion

# References

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.*

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.

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.

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

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