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

Christopher Shultz

c.shultz@live.com

Author Contributions: Single-author, 100% attribution.

Keywords: dataset drift, covariate shift, dataset shift, topological data analysis, TDA

MSC: 55N31, 62-08

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, particularly for its ability to capture higher-dimensional relationships that are unobservable to traditional drift detection methods. We present a brief review of drift, detection methods, and describe the potential benefits of utilizing PE to capture the dynamic "shape" of the dataset over time. 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, impacting both the performance and fairness of models (Deho et al., 2024). This phenomenon is typically categorized into two types: covariate shift, where the data distribution (X) evolves over time, but the relationship X~Y remains; and concept drift, where the relationship X~Y itself changes (Mallick et al., 2022). Mixed drift also occurs, in which some combination of covariate shift and concept drift take place simultaneously.

If unaddressed, drift can lead to significant declines in model accuracy and fairness. However, various methods exist for detecting drift, each with distinct advantages and limitations. Statistical tests such as the Kolmogorov-Smirnov (KS) test and the Hotelling T^2 test compare distributions of current and historical data to detect changes. Although these tests are straightforward and interpretable, they cannot capture more complex multivariable relationships. Alternatively, measures such as the Population Stability Index (PSI) are widely used and more complex, but are univariate in nature, and still fail to capture relationships between variables and their higher-dimensional characteristics.

In this paper, we propose the use of Topological Data Analysis (TDA) as a supplementary approach for dataset drift detection. Specifically, we explore quantified Persistence Entropy (PE) to assess the "stability" of the data's shape over time. TDA allows for the examination of the geometric and topological features of dataset, providing insights not captured by traditional methods. For example, by viewing a dataset as a noisy sample from some multidimensional "shape", we can examine this shape's representation and topological characteristics such as connected components, holes, and voids.

PE is a method derived from persistent homology that quantifies the complexity and variability of a dataset's topological features by first encapsulating information about the birth and death of topological features across different *resolutions* in a persistence diagram, and then examining the "stability" of these persistence diagrams through time. In theory, datasets that remain topologically stable through time will have a stable vector of PE values. Datasets that experience topological changes should have a PE vector that reflects those changes in topological structure. 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 simulated data. We generate a synthetic dataset of variables following a standard normal distribution and then inject various types (and sizes) of drift to the end of the series. Our findings demonstrate that [a] persistence entropy consistently captures the introduced drift through notable changes in computed average PE values; and [b] the magnitude of those changes aligns with the magnitude of the applied shocks. The results underscore the sensitivity and robustness of PE for detecting dataset drift, and capturing changes to high-dimensional relationships that are otherwise unobservable.

The remainder of this paper is structured as follows. Section 2 provides an overview of dataset drift detection methods and a background on TDA. Section 3 outlines the theoretical foundations of persistence entropy and its application in drift detection, along with the setup of an empirical analysis. Section 4 presents results and a discussion of findings. Finally, Section 5 concludes with a discussion of the implications of this work and potential future research directions.

By introducing persistence entropy as a supplementary drift detection mechanism, this work provides 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

2.1: Overview of the Concept of Drift

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 in another way, covariate shift occurs when the mapping from inputs-to-output stays consistent, but the distribution of the inputs changes (Chen et al., 2016). The precise definition of covariate shift has been subject to some debate, and aiming to unify terminology, Moreno-Torres et al. (2011) settle on a definition that aligns with the aforementioned explanation.

The other commonly discussed forms of drift are [a] *prior probability shift*, in which changes occur within the distribution of Y; and [b] *concept shift*, in which the relationship mapping $X \rightarrow Y$ changes from its prior form. The consequences of drift are significant and prevalent in real-world applications. Theoretical justification for most models relies on the assumed absence of drift in the forms stated (Tripuraneni et al., 2021). Without the equality of distributions from "old" and "new" data, for example, core assumptions are violated and it is vital to understand the potential impacts on model validity and performance.

The impacts of drift have been relevant in the literature for decades, both in discussing the problem itself, as well as methods for its resolution. The most common problem we aim to resolve is a generalized degradation of model performance metrics like accuracy, precision, recall, etc. Widmer and Kubat (1996) discuss drift using the language of "changing context" and a family of algorithms that can flexibly react to a drifting environment through "on-line" learning. The performance of classifiers, for example, have shown drastic improvement when covariate shift is detected and corrected for (Dharani et al., 2019). Failing to correct for degraded performance resulting from drift can lead to real world consequences such as poor decision making, financial losses, and loss of trust in models by stakeholders. Gama et al. (2014) provide a comprehensive overview of the impacts of concept drift and covariate shift on model performance and methods for their detection.

Rather than another review of drift types, detection methods, and handling, we provide a brief overview resting on foundations set forth in other papers. We discuss the pros and cons of "standard" drift detection mechanisms and then present TDA as a supplementary measure capable of capturing otherwise unobservable information in the dynamics of the evolving dataset.

2.2: Traditional Methods for Drift Detection

Commonly used measures include the Kolmogorov-Smirnov (KS) test, the Hotelling T^2 test, and the Population Stability Index (PSI). Each has its own set of pros and cons, and all fail to capture the higher-dimensional relationships that TDA can provide, though we view the use of TDA as a supplement to these core methods, rather than a replacement.

2.2.1: Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) test is a nonparametric statistical test for comparing the cumulative distribution functions (CDFs) of two univariate datasets. For each value therein, it calculates the difference between the CDFs, with the test statistic D as the maximum absolute difference observed:

$$D = \sup_x |F_1(x) - F_2(x)|$$

In the equation for D, $F_1(x)$ and $F_2(x)$ represent the empirical CDFs of the two datasets. The null hypothesis H_0 states that the two datasets came from the same distribution. If D is larger than the test critical value (determined by sample size and significance level), H_0 is rejected, indicating a statistically significant difference between the two distributions.

Because the KS test is univariate, it is unable to be applied simultaneously to all variables in a multidimensional dataset, though it can be used to compare the "old" and "new" series for each variable in isolation. For a multidimensional representation of drift, we can consider the *Hotelling* T^2 test, which moves us closer to a dynamical data-level view and is discussed in the following section.

2.2.2: Hotelling's T^2 Test

Hotelling's T^2 test is a multivariate statistical test used to compare the means of multiple variables simultaneously, creating a more detailed window in the evolution over time of the joint distribution underlying a dataset. It is a generalization of the Student's t-test to a multivariate setting, and its test statistic is defined as:

$$T^2 = n(\bar{X} - \mu)^T S^{-1} (\bar{X} - \mu)$$

where n is the sample size, $\bar{X}$ is the sample mean vector, μ is the population mean vector, and S is the sample covariance matrix. The null hypothesis states that the sample mean vector is equivalent to the population mean vector, and the test statistic following a distribution related to the F:

$$\frac{\big((n-1)p\big)}{n-p} T^2 \sim F(p, n-p)$$

where p is the number of variables. If the calculated T^2 value is greater than the critical value from the F distribution, the null is rejected, indicating a significant difference between mean vectors. This provides a multidimensional comparison of means, but still captures nothing about the relationship between variables and their higher-dimensional properties and co-relationships.

2.2.3: The Population Stability Index (PSI)

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:

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

With $\big(f_{0,i}, f_{1,i}\big)$ representing the fraction of entities in bin i in the original and new populations, respectively. As a result, this method necessitates the binning of each continuous variable into discrete groups before application. The PSI is univariate in nature and results in the comparison of individual variables in isolation, though their results can be examined in aggregate.

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.

The PSI does provide a more granular view of the stability of the underlying distributions than the simple mean method using in Hotelling's T^2 test, but still fails to capture higher-dimensional interactions and "shape" features such as joint distribution changes and topological features such as holes or voids. Yurdakul and Naranjo (2020) provide a thorough overview of the statistical properties of the PSI and a summary of its pros and cons.

2.2.4: Summary of Standard Drift Detection Methods

As we have seen, most commonly used drift detection methods are powerful, but have a limited scope. Traditional univariate methods such as the KS test or PSI fail to examine the joint distribution and evolving "shape" of the combined dataset, focusing rather on the changing distributions of individual variables alone. Multivariate tools like Hotelling's T^2 provide that "combined" view of the representation of data, but are limited by a focus only on comparison of means vectors, rather than distributions at large. Rabanser et al. (2019) provide a thorough review of statistical methods, outlining their empirical structure, benefits, and weaknesses.

2.3: Introducing Topological Data Analysis (TDA)

The stated limitations of traditional drift detection methods are related to their scope, not their validity. The proposal within this paper is that methods from TDA can *supplement* traditional methods by providing a window into the dynamic shape and topological characteristics of data, adding new information that is otherwise unobservable. Using standard drift detection tools has been widely adopted, but fails to paint the entire picture.

Viewing the dataset as a noisy sample from some multidimensional "object" with shape and form, methods from TDA provide a framework for capturing previously unobserved information over time to inform our concept of "drift." In brief, the question we aim to consider is this: *does the shape and topological structure of a given multidimensional dataset change over time, and can that be quantified?*

The general framework employed in many applications of TDA is to represent a set of data as a *point cloud,* this point cloud as a *simplicial complex*, and a collection of these simplicial complexes across varying levels of "resolution" as a *filtration*. This TDA pipeline is summarized by Lum et al. (2013).

The original dataset forms a set of discrete points in some space, which we refer to as a point cloud. That point cloud is viewed as a noisy representation of some underlying structure which can be estimated (Carlsson, 2009). To accomplish this, we construct a *simplicial complex,* which is a "smoothed" representation of the point cloud, viewed as the vertices of a combinatorial graph whose edges are determined by some proximity measure defining the "resolution" of the complex via parameter ϵ (Lum et al., 2013). The common choice in deriving the simplicial complex is the Vietoris-Rips algorithm, summarized by Shultz (2023).

The development of multiple 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, etc.), 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 within the filtration across our point cloud (Zomorodian and Carlsson, 2005).

For drift detection, we propose the use of the *persistence entropy (PE)* measure (Munch et al., 2019), which generates a vector of quantified entropy representing the amount of order/disorder in the underlying topological object over time. For example, with a series of data X over some time index t, we can utilize a rolling window approach with user-selected length (e.g. $w = 10$) to create sequential sub-frames of length w. We may then apply this rolling window to the TDA pipeline, each represented as a point cloud, converted to simplicial complexes, then converted to persistence diagrams. These persistence diagrams can be utilized to compute persistent entropy.

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. Formally, the Persistence Entropy is represented as $H(X)$ below:

$$H(X) = -\sum_{i} \left(\frac{l_i}{L}\right) \log\left(\frac{l_i}{L}\right)$$

where l_i/L represents the normalized persistence of each bar in the diagram, with l_i being the length of the i_{th} bar and L as the sum of the lengths of all bars in the diagram. This creates a quantification of the "complexity" of the persistence diagram by using the concept of entropy. This creates a window into a higher level of dimensionality than traditional drift detection methods allow; and tracking the PE measure can provide a view into the stability of the topological complexity of the underlying object, revealing previously invisible information. For example, it is possible that means remain stable but topological structure drifts, and this would be missed in traditional approaches. For a basic introduction to TDA, please refer to Chazal and Michel (2017).

To our knowledge, no existing research examines the utilizing of TDA and/or Persistent Entropy as a potential supplementary method for the detection of dataset drift. The utilization of TDA-based method for comparing distributions is an emerging area of research, with Dlotko et al. (2024) examining a comparison of distributions topologically using the Euler Characteristic Curve (ECC). Their analysis shows that TDA-based comparisons perform similarly to state-of-the-art methods. This lends further support for the exploration of topology-based comparisons in drift detection.

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.

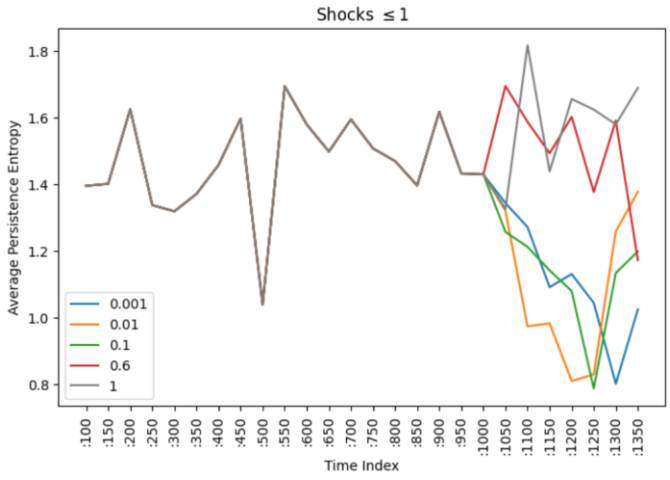
4: Results

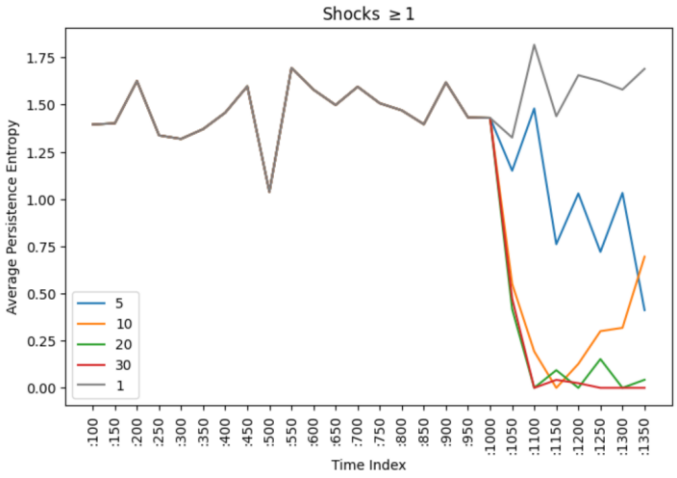
Table 1 provides a summary of the datasets utilized in our analysis, each with 1400 observations, the first 1000 of which are non-drifted / identical.

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 5th and 95th 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.

Chazal, F. and B. Michel. 2017. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *arXiv: 1710.04019.*

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.

Deho, O., Liu, L., Li, J., Liu, J., Zhan, C., and S. Joksmiovic. 2024. When the Past != The Future: Assessing the Impact of Dataset Drift on the Fairness of Learning Analytics Models. Accepted in *IEEE Transactions on Learning Technologies.*

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.*

Dlotko, P., Hellmer, N., Stettner, L. and R. Topolnicki. 2024. Topology-Driven Goodness-of-Fit Tests in Arbitrary Dimensions. *Statistics and Computing.* 34:34.

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., and A. Bouchachia. 2014. A Survey on Concept Drift Adaptation. *ACM Computing Surveys (CSUR).* 46(4): 1-37.

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.*

Widmer, G. and M. Kubat. 1996. Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning.* 23(1): 69-101.

Yurdakul, B. and J. Naranjo. 2020. Statistical Properties of the Population Stability Index. *Journal of Risk Model Validation.* 14(4): 89-100.

Zomorodian, A. and G. Carlsson. 2005. Computing Persistent Homology. *Discrete & Computational Geometry.* 33(2): 249-274.