Designing statistical tests for topological significance

Isaac Ren

January 9, 2025 — JMM AMS Special Session:
MRC Climate Science between TDA and Dynamical Systems

YOUNG TOPOLOGISTS MEETING 2025

Stockholm, Sweden
June 23-27

Short talks
Poster session
Invited speakers

Frédéric Chazal
Manuel Krannich
Maria Yakerson

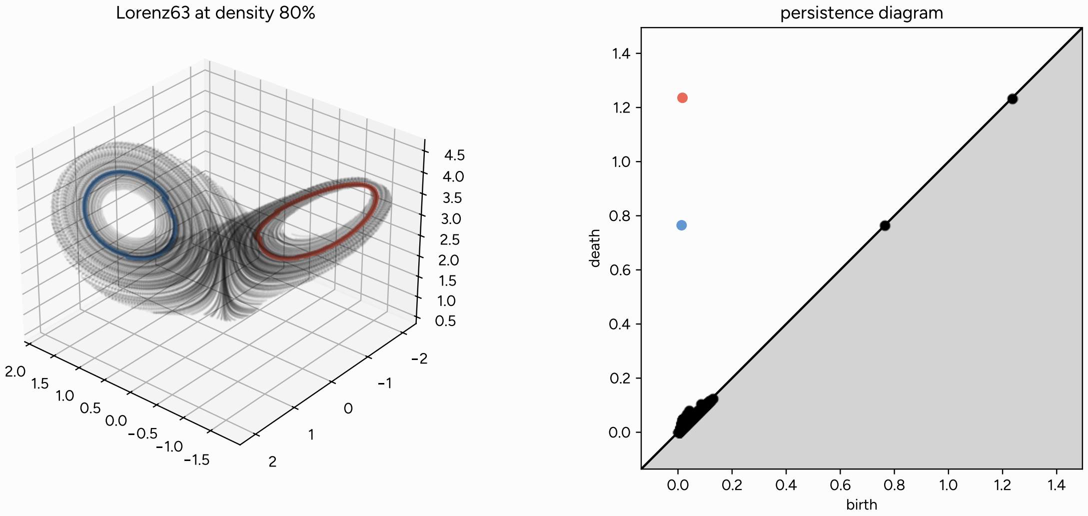
KTH VETENSKAP OCH KONST

Stockholm University

Topological significance and statistical tests

Topological significance

- Given a point cloud, what is a **topologically significant feature**?
- We say that it is a homological cycle whose corresponding persistence point is abnormal: e.g. an unusually long bar or a distant persistence point:



Statistical tests

Method

- Let X be a filtered point cloud, D its persistence diagram, and (b, d) a point of D.
- Our hypothesis test is

$$H_0: (b, d) \text{ does not correspond to a significant topological feature,}$$
$$H_1: (b, d) \text{ does correspond to a significant topological feature.}$$

Formalization

Definition

- Let $f: \mathbf{R}^2 \to \mathbf{R}$ be a measurable function.
- Let X be a random point process, D_q its persistence diagram for H_q.
- Consider a persistent homological cycle of X and $(b_0, d_0) \in D_q$ the corresponding persistence point.
- The cycle is **significant at level** α if the p-value of $f(b_0, d_0)$ is less than α:

$$P(f(b, d) \geq f(b_0, d_0) \mid (b, d) \in D_q) \leq \alpha.$$

- In practice we correct for multiple tests: we use Bonferroni (divide α by the number of tested points).

Properties of topological significance

Desired properties

- Ideally, f should be **translation** and **scale invariant**.
- Persistence is already translation invariant, **death-birth ratios** are scale invariant.
- We also want to know the distribution of $f(b, d)$.

Distribution of persistence points

Universal distributions of persistence points

Theorem [Bobrowski-Skraba 2024]

- Let X_n be a set of n i.i.d. points in $\mathbf{R}^m$ with a **“good”** probability density φ and consider its Vietoris-Rips or Čech complex.
  - Good densities include: those with closed support, bounded away from 0, and normal distributions.
- For $q \geq 1$, let $D_{q,n} = ((b_i, d_i))_i$ be the H_q persistence diagram.
- Define $\Pi_{q,n} = \{d_i/b_i\}_i$. Then

$$\Pi_{q,n} \xrightarrow[n\to\infty]{\text{weak}} \Pi_q^*,$$

where Π_q^* does not depend on the probability density φ.

Universal distributions of persistence points

Conjecture [Bobrowski-Skraba 2023a]

- Up to recentering, $\{A \log\log(\pi_i) \mid \pi_i \in \Pi_{q,n}\}$, with $A = 1$ for Vietoris-Rips and $A = \frac{1}{2}$ for Čech, weakly converges to the left-skewed Gumbel distribution with PDF e^{x-e^x} and CDF $1 - e^{-e^x}$.

Further conjecture

Conjecture

- Suppose that the support of the point process is locally an r-dimensional space (i.e. a topological r-manifold).
- Then $\{A \log(\pi_i - 1) + \log(r+2) \mid \pi_i \in \Pi_{q,n}\}$ weakly converges to the left-skewed Gumbel distribution.

Comments

- $\log(x-1)$ is similar to $\log \log x$ at $x \approx 1$ but has a more spread out tail distribution, useful for identifying outliers.
- We no longer need to recenter the π_i's, which means we can use methods that only compute the most persistent features.

The case of H_0

For H_0, we cannot use the previous results, since $b_i = 0$.

Definition: cluster persistence

- **[Bobrowski-Skraba 2023b]** propose **k-cluster persistence**, where connected components are born only when they contain at least k points.
- The resulting persistence diagram can be computed using the dendrogram associated to the point cloud.
- See also **mergegrams** from **[Elkin-Kurlin 2020]**.
- **Upshot:** we get positive birth times, allowing for the definition of $\Pi_{0,n,k}$.

Conjecture for H_0

Conjecture

- Let $r \in \{2, 3\}$ be the dimension of the support of φ.
- Let $k = 3$ if $r = 2$ and $k = 2$ if $r = 3$.
- The set $\{\log(\pi_i - 1) \mid \pi_i \in \Pi_{0,n,k}\}$ weakly converges to the left-skewed Gumbel distribution.

Experiments and results

Quantifying the significance of weather regimes

Motivation

- In [Strommen-Chantry-Dorrington-Otter 2022] the goal is to topologically describe weather regimes; our goal is to quantify this description using statistical significance.
- We look at the point clouds of that paper, filtered at various density levels.
- We assume that the point clouds are samples from compact manifolds plus Gaussian noise.
- Restricting to the densest points then gives a distribution with compact support, bounded away from 0.

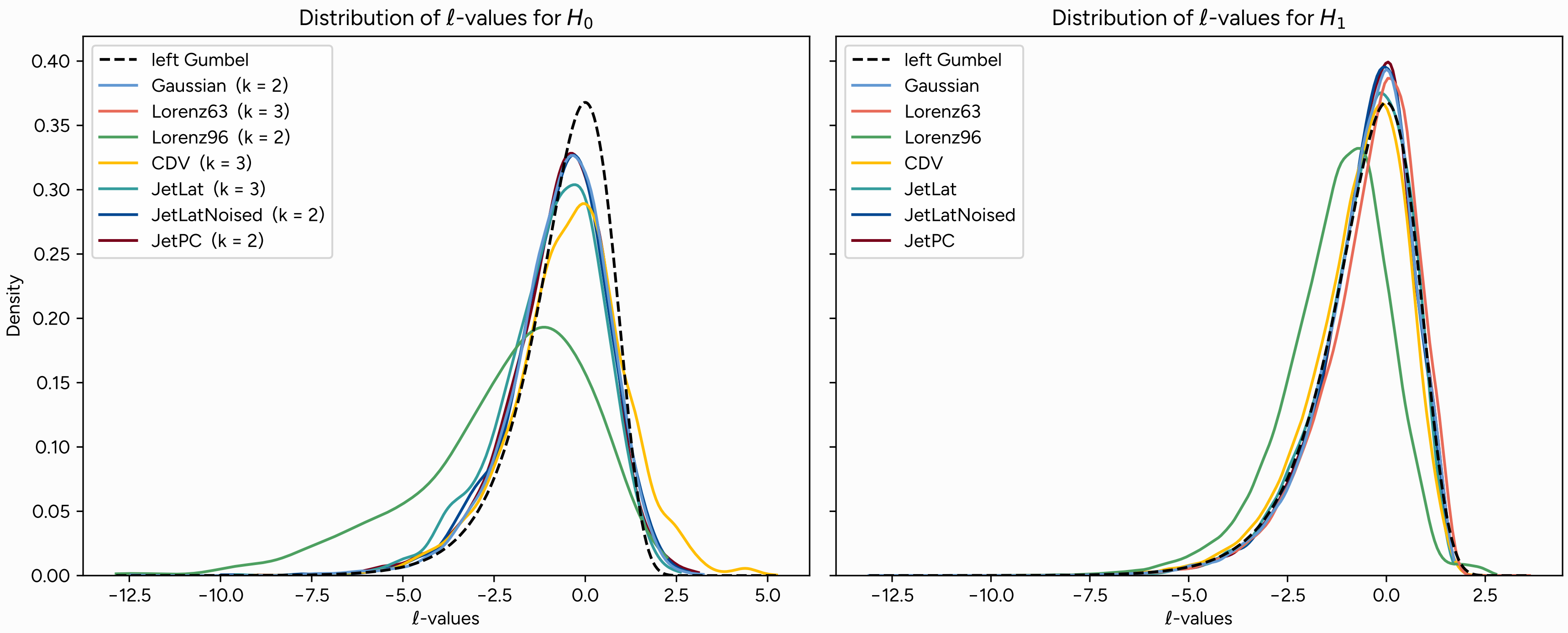
Setup

Assumptions

- Let X be a set of i.i.d. points with a good distribution.
- For $q \geq 1$ with the Čech filtration, we assume that the right tail of $\{\frac{1}{2}\log(\pi_i - 1) + \log(r+2) \mid \pi_i \in \Pi_{q,n}\}$ is upper bounded by left-skewed Gumbel.
- For $k \geq 2$, we assume that the right tail of $\{\log(\pi_i - 1) \mid \pi_i \in \Pi_{0,n,k}\}$ is upper bounded by left-skewed Gumbel.

Experiments and results

Checking Gumbelness

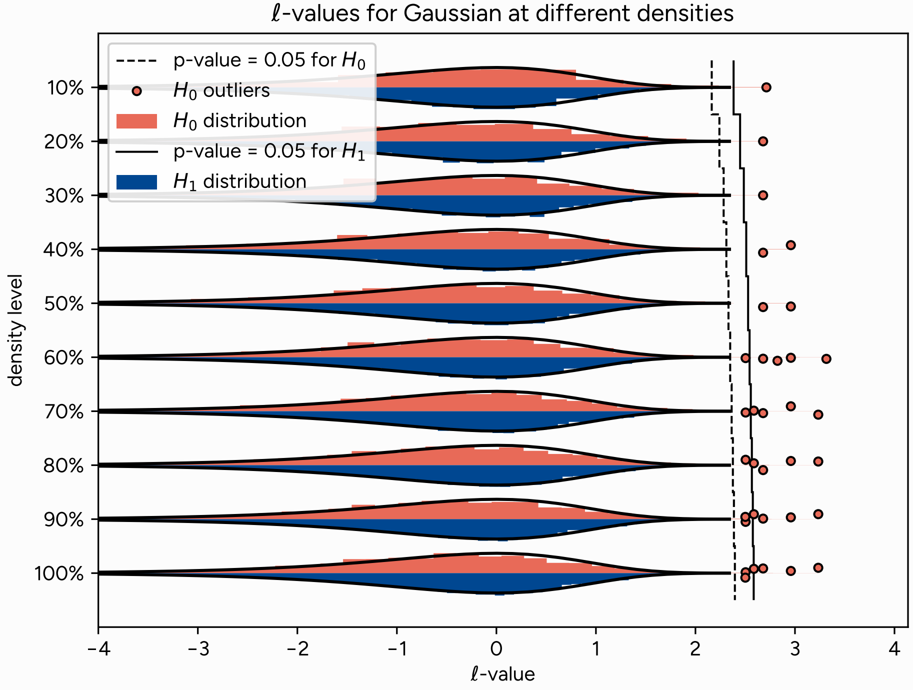


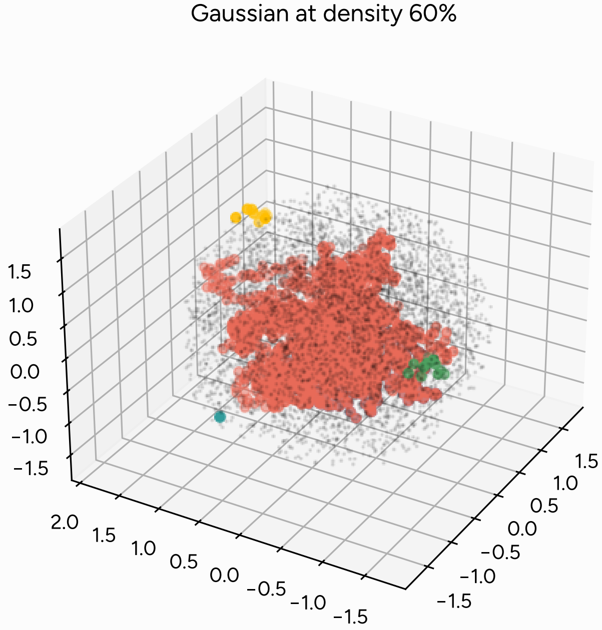
Reference set

Point cloud: Gaussian

- $10{,}000$ standard normally distributed points in $\mathbf{R}^3$.
- Ignoring the infinite bar in H_0, we expect no significant topological features.

Gaussian


ℓ-values for Gaussian at different densities
p-value = 0.05 for H_0
H_0 outliers
H_0 distribution
p-value = 0.05 for H_1
H_1 distribution
density level
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
ℓ-value
−4
−3
−2
−1
0
1
2
3
4

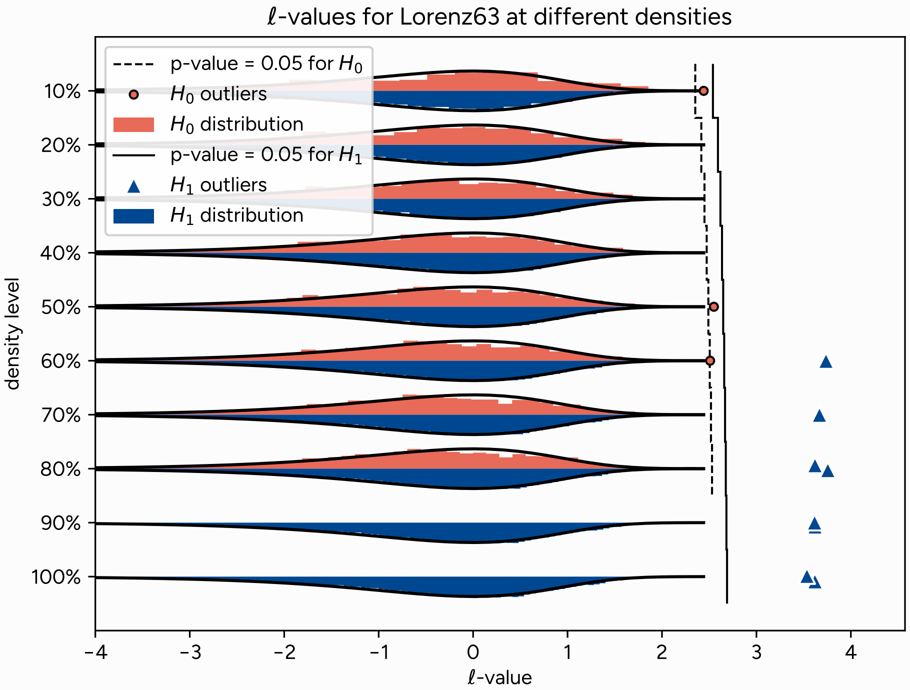

Gaussian at density 60%

Toy models

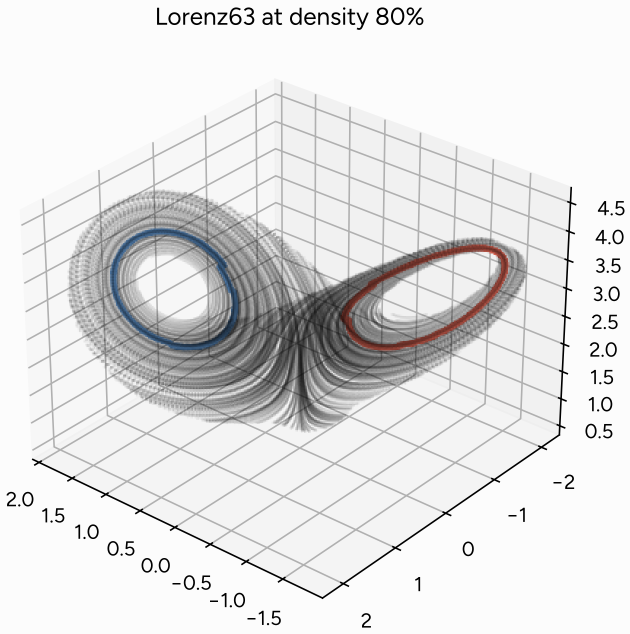
Point clouds: Lorenz '63, Lorenz '96, Charney-de Vore

- These point clouds model atmospheric dynamics, showcasing their chaotic structure.
- **Lorenz '63** is the classic butterfly wing model. $100{,}000$ points in $\mathbf{R}^3$.
- **Lorenz '96** is a more complex model, in $\mathbf{R}^{40}$. We consider $20{,}000$ points projected onto the first 4 principal components (empirical orthogonal functions).
- **Charney-de Vore** models large-scale midlatitude blocking dynamics in $\mathbf{R}^6$. We consider $40{,}000$ points projected onto the first 3 principal components.
- Representative 1-cycles are computed with Persloop.

Lorenz '63



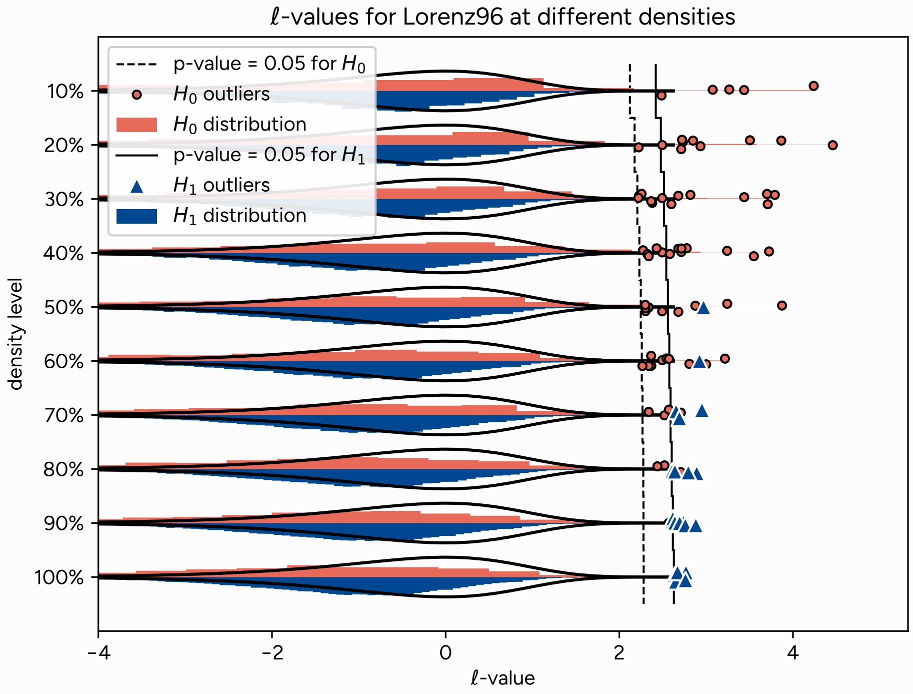
ℓ-values for Lorenz63 at different densities
p-value = 0.05 for H_0
H_0 outliers
H_0 distribution
p-value = 0.05 for H_1
H_1 outliers
H_1 distribution
density level
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
ℓ-value
−4
-3
-2
-1
0
1
2
3
4

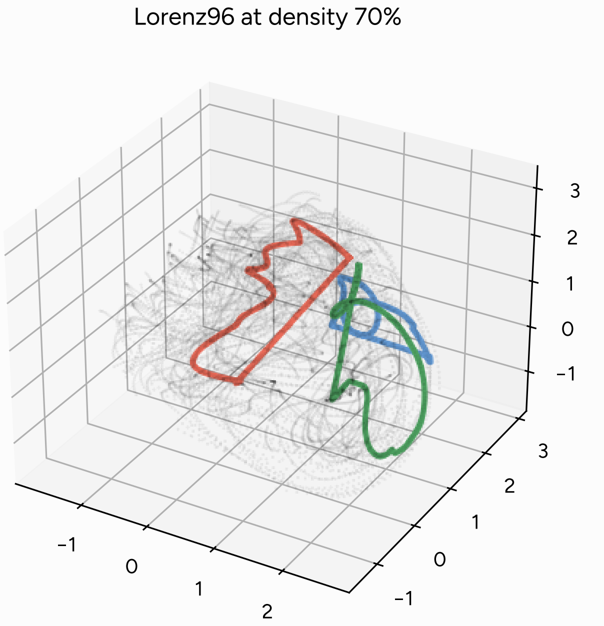


Lorenz63 at density 80%
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1
2

Experiments and results

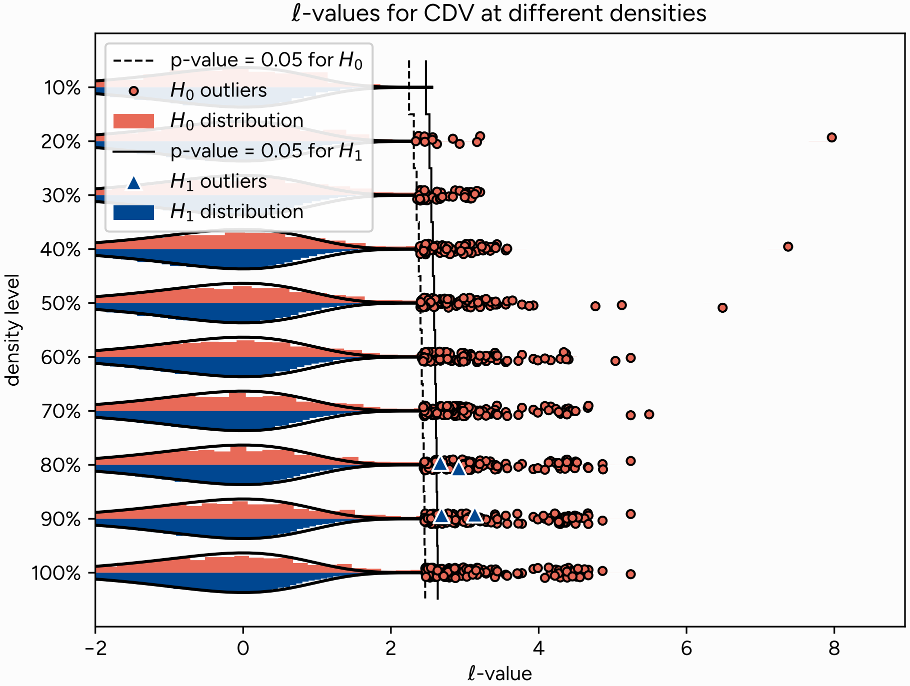
Lorenz '96

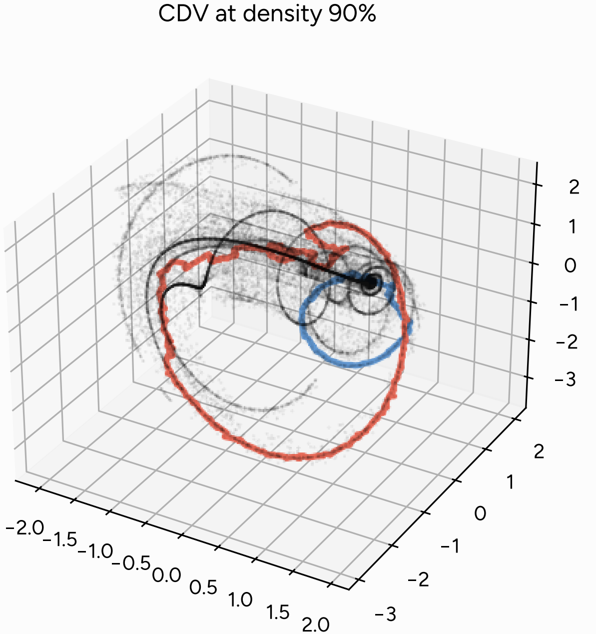




Experiments and results

Charney-de Vore



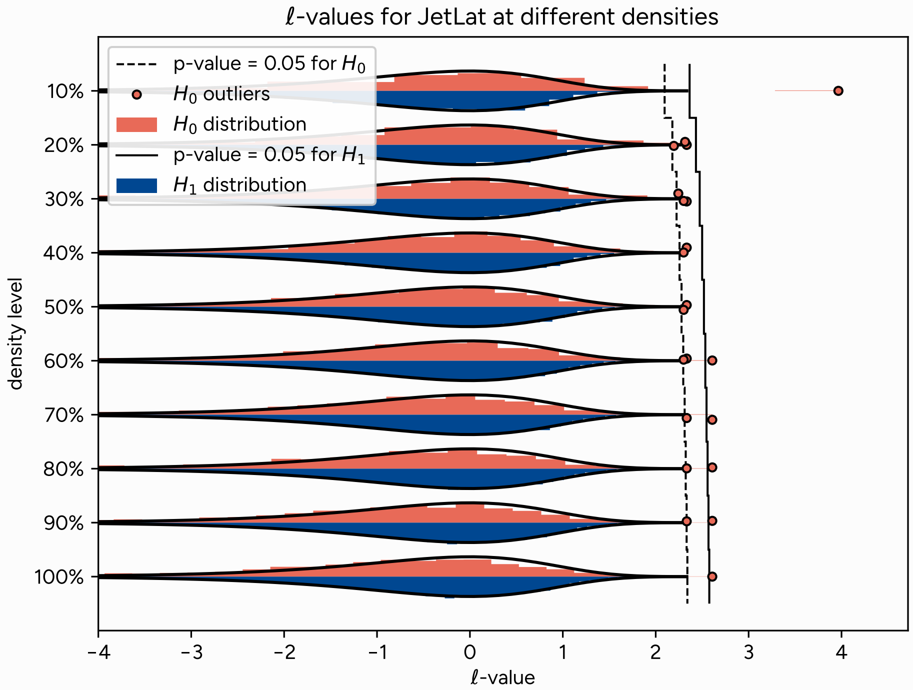


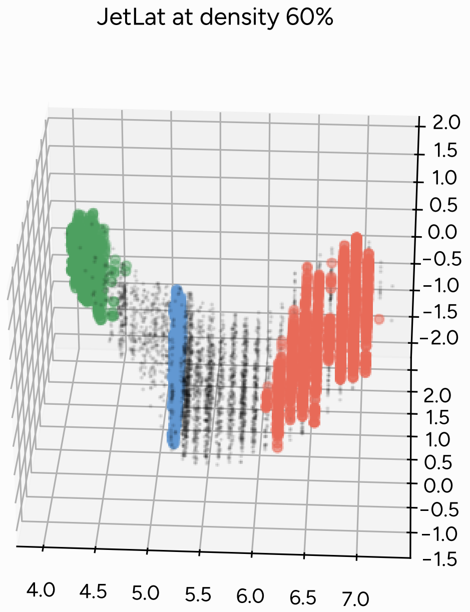
Observational data

Point clouds: North Atlantic jet

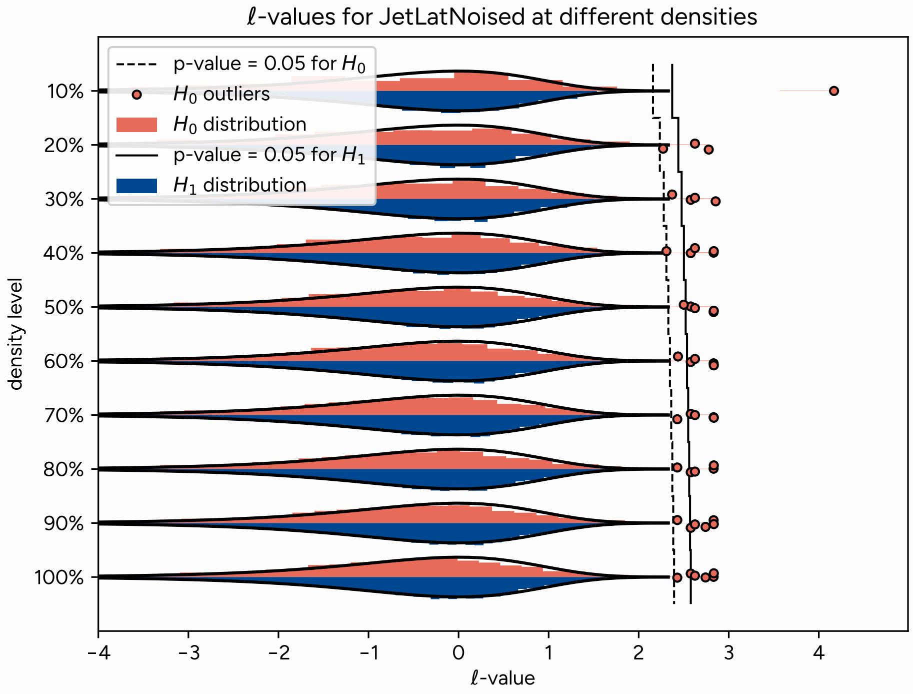
- Data based on observed atmospheric data.
- **JetLat** consists of the observation's latitude and the first 2 principal components.
- The latitude is discretized, so **JetLatNoised** adds uniform noise in $[-\frac{1}{2}, \frac{1}{2}]$ to the latitude.
- **JetPC** consists of the first 3 principal components.
- We are expecting to identify two or three weather regimes.

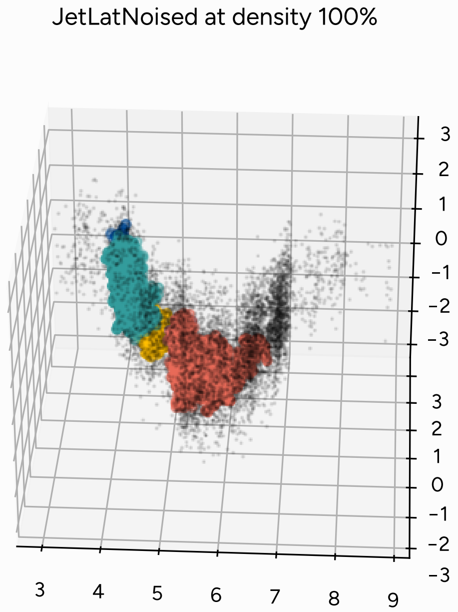
North Atlantic jet



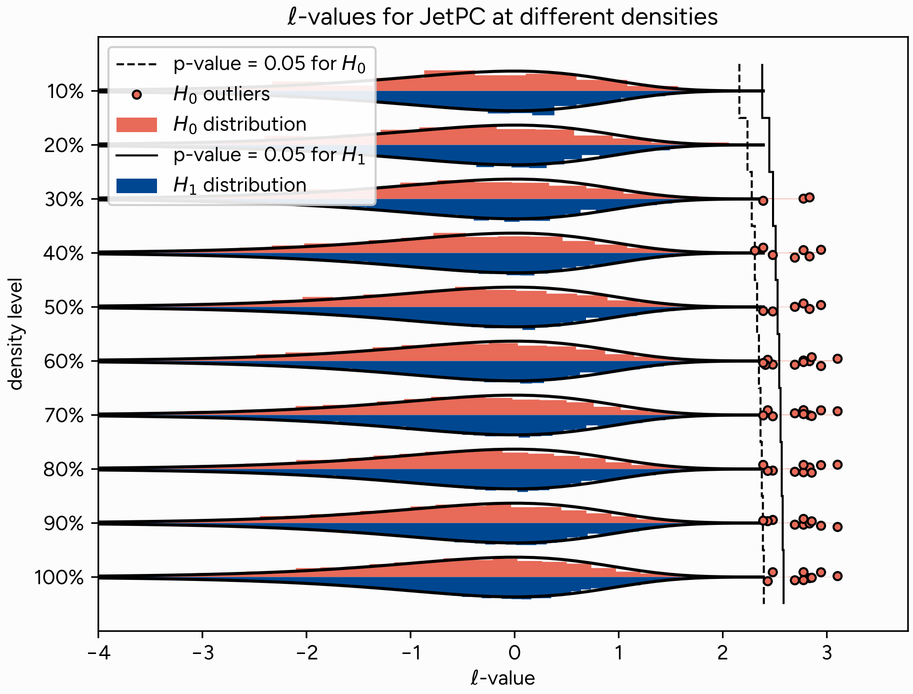


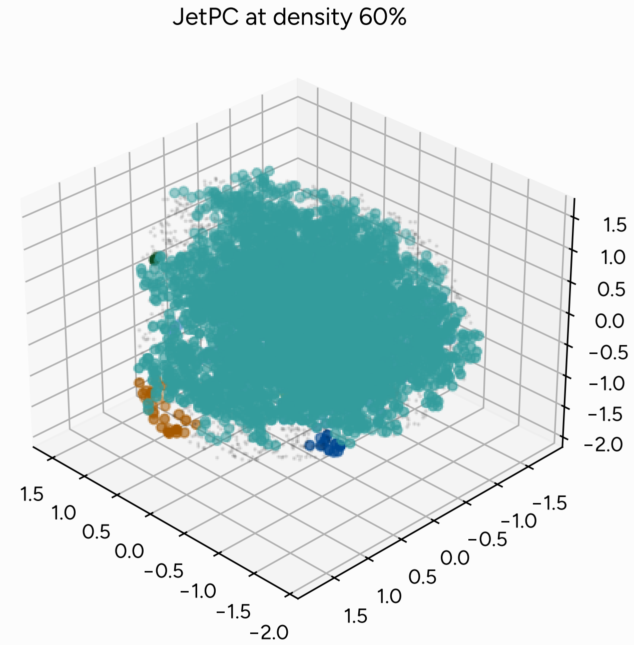
North Atlantic jet (noised)





North Atlantic jet (PCs)





Conclusions

Observations

- The assumptions for H_0 do not really hold, and also need to be extended to higher dimensional spaces.
- The method works better for H_1, although it also works best for lower dimensions.
- Still hard to conclude for the observational data.

Summary

- Following Bobrowski and Skraba, we run **hypothesis tests for topological significance** in all degrees, using known and conjectured results about **scale-invariant** functionals.
- We test this on various toy models and observational data, filtering the point clouds by density.

Outlook

- This is exploratory work: future work includes looking at larger, **higher-dimensional datasets**, and developing the theory behind this analysis.
- We will also study the 2-**parameter** nature of the data: faster computation of persistence, statistics on the decomposition or presentation of 2-parameter persistence modules, etc.
    - Ongoing project with Kristian Strommen, Tung Lam, and Fabian Lenzen.

References

- K. Strommen, M. Chantry, J. Dorrington, N. Otter. *A topological perspective on weather regimes*, 2022.
- O. Bobrowski and P. Skraba. *A universal null-distribution for topological data analysis*, 2023.
- —. *Cluster-persistence for weighted graphs*, arXiv 2023.
- —. *Universality in random persistent homology and scale-invariant functionals*, arXiv 2024.
- Y. Elkin, V. Kurlin. *The mergegram of a dendrogram and its stability*, arXiv 2020.