Topology-Driven Two-Sample Tests with Persistent Landscapes

Ke Xu

ACMS

December 8, 2025

Motivation: Goodness-of-Fit and Two-Sample Testing

- **Goodness-of-Fit (GoF)** tests assess whether a sample comes from a specified distribution or differs from it.
- In the **two-sample GoF problem**, we observe two independent samples:

$$X = \{x_1, \dots, x_n\} \sim F, \quad Y = \{y_1, \dots, y_m\} \sim G,$$

  and test $H_0 : F = G$ vs $H_1 : F \neq G$.
- The **Kolmogorov–Smirnov (KS)** test is a classical GoF test in 1D:

$$D_{n,m} = \sup_t |F_n(t) - G_m(t)|,$$

  where F_n, G_m are empirical CDFs of X and Y.
- **Limitation:** KS and other EDF-based tests rely on 1D ordering. Extensions to $\mathbb{R}^d$ are challenging due to the lack of a canonical cumulative distribution function.

Motivation: Why Not Euler Characteristic Curves?

- **Topological GoF tests** typically use summary functions such as Euler Characteristic Curves (ECC) (Dłotko et al., 2024).
- **The Euler-Equivalence Problem:** ECC is an alternating sum of Betti numbers ($\chi = \beta_0 - \beta_1 + \beta_2 \dots$).
- Distinct topological features can cancel each other out.
  - Example: A space with 2 components and 2 loops ($\chi = \beta_0 - \beta_1 = 0$) is indistinguishable from a space with 1 component and 1 loop ($\chi = \beta_0 - \beta_1 = 0$) using ECC alone.
- **Persistence Landscapes (PLs)** avoid this cancellation by preserving the full birth-death information in a functional form.

Topological Background: Persistence Landscapes

- **Construction:**
  1. Transform Persistence Diagram (PD) points into tent functions: $f_i(t) = \max(0, \min(t - b_i, d_i - t))$.
  2. Define the k-th landscape function $\lambda_k(t)$ as the k-th largest value of $\{f_i(t)\}_i$.
- **Key Property:** The sequence $\Lambda = (\lambda_1, \lambda_2, \dots)$ lies in a separable Banach space (L^p).
- This embedding allows us to define **means** and **variances**, unlike raw diagrams.

Formal Hypothesis Framework

- We map the problem from the space of distributions on $\mathbb{R}^d$ to the space of landscape functions.
- Let Λ_X and Λ_Y be random variables representing the persistence landscapes of point clouds sampled from F and G.
- We test the equality of expected landscapes:

$$H_0 : \mathbb{E}[\Lambda_X] = \mathbb{E}[\Lambda_Y] \quad \text{in } L^2(\mathbb{N} \times \mathbb{R})$$

$$H_1 : \mathbb{E}[\Lambda_X] \neq \mathbb{E}[\Lambda_Y]$$

- This creates a robust test: if the underlying topologies differ, the expected landscapes will differ.

Method: Two-Sample Test with PLs

- Given $X \sim F$, $Y \sim G$
- Compute persistence landscapes Λ_X, Λ_Y
- **Test Statistic (L^2 distance):**

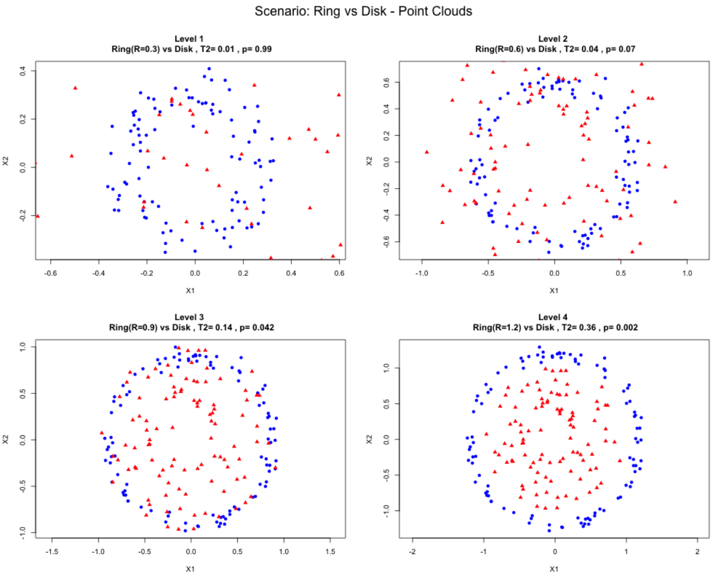
$$T_{obs} = \|\bar{\Lambda}_X - \bar{\Lambda}_Y\|_{L^2}^2 \approx \sum_{k=1}^{K} \sum_{j=1}^{J} (\bar{\lambda}_{X,k}(t_j) - \bar{\lambda}_{Y,k}(t_j))^2 \Delta t$$

- **Permutation Procedure:**
  1. Combine samples into pooled set $Z = X \cup Y$.
  2. Randomly permute indices to split into X^* and Y^*.
  3. Compute $T^* = \|\bar{\Lambda}_{X^*} - \bar{\Lambda}_{Y^*}\|_{L^2}^2$.
  4. Repeat B times to estimate the null distribution.
- **Decision:** Reject H_0 if p-value $= \frac{1}{B} \sum \mathbb{I}(T^* \geq T_{obs}) < \alpha$.

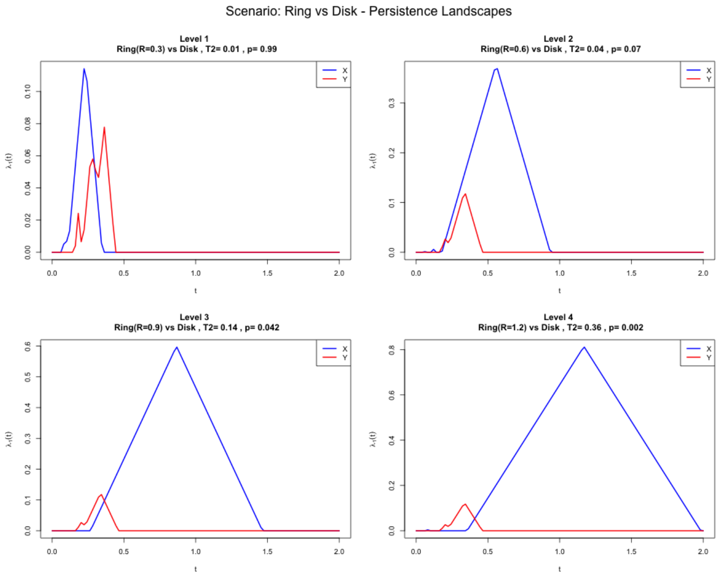
Scenarios in Simulation

1. **Ring vs Disk (Topological Difference)**
   - X: Points sampled on a ring with added Gaussian radial noise.
   - Y: Uniformly sampled in a disk of same radius.
   - **Levels:** Gradually increase the radius of X.
2. **Ring vs Ring (Noise Difference)**
   - X: Ring with small fixed radial noise.
   - Y: Ring with increasing radial noise.
   - **Levels:** Gradually increase noise in Y.

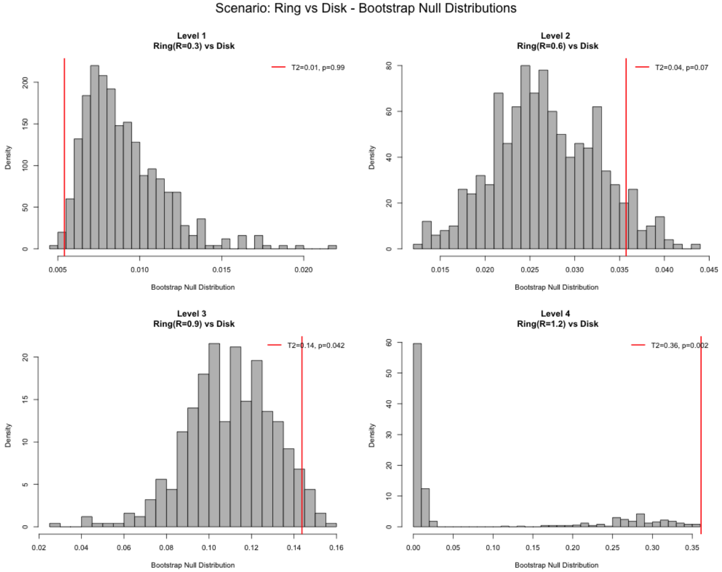
Ring vs Disk: Point Clouds



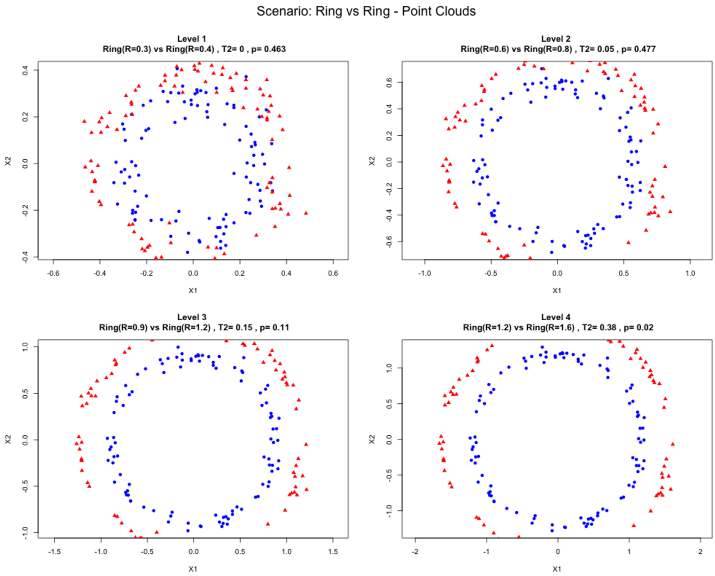
Ring vs Disk: Landscapes



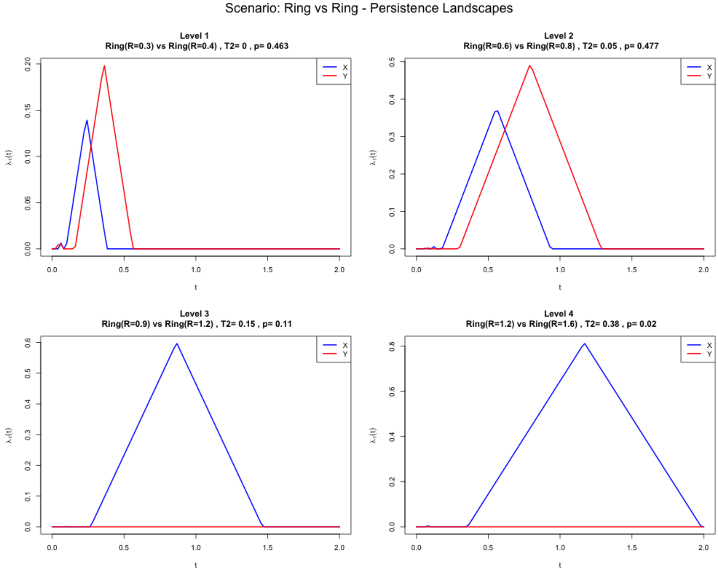
Ring vs Disk: Bootstrap Null



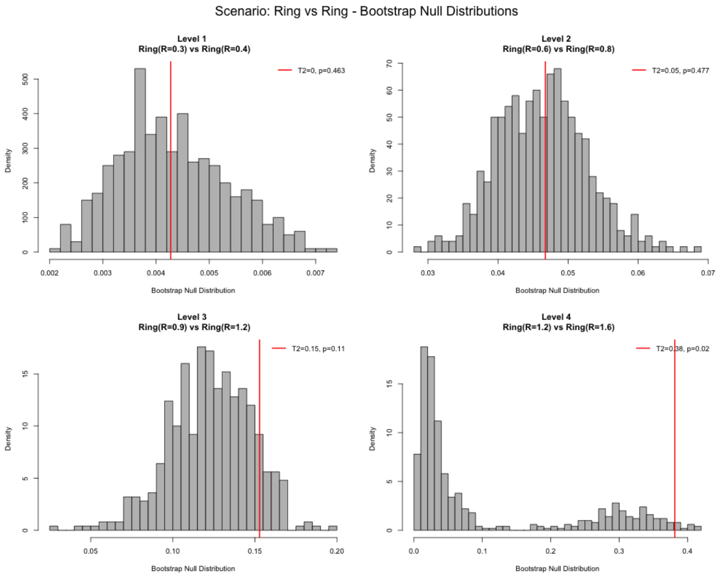
Ring vs Ring: Point Clouds



Ring vs Ring: Landscapes



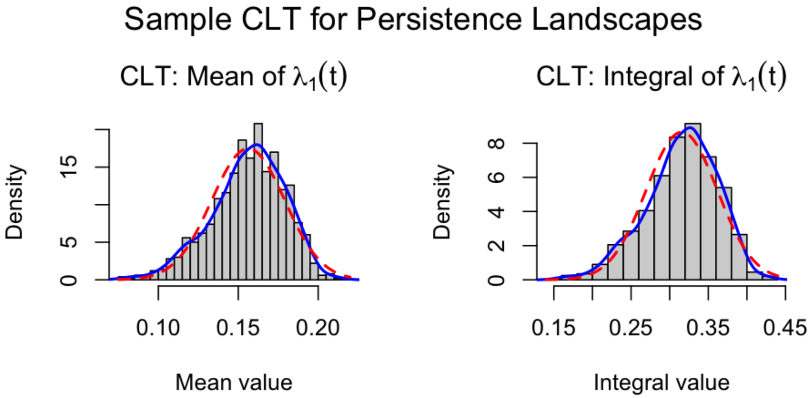
Ring vs Ring: Bootstrap Null



Statistical Properties

- **Stability:** The mapping Diagram $\to$ Landscape is 1-Lipschitz (Bubenik, 2015). Small noise in data $\to$ Small change in T_{obs}.
- **Central Limit Theorem:** For $p \geq 2$, $\sqrt{n}(\bar{\Lambda}_n - \mathbb{E}[\Lambda])$ converges to a Gaussian process in L^p (Bubenik, 2015). However, the covariance operator is unknown.
- **Consistency:** Under H_1, as $n, m \to \infty$, the test power converges to 1 provided $\|\mathbb{E}[\Lambda_X] - \mathbb{E}[\Lambda_Y]\| > 0$.

CLT Validation (Normal Sample)



- Left: distribution of mean landscape values
- Right: distribution of integrated landscape values
- Both show convergence to Gaussian, validating CLT

Conclusion and Outlook

- **Summary:** Persistence landscapes provide a functional framework for multivariate two-sample testing.
- **Advantages:** Stability and compatibility with standard Hilbert space statistics (We can use standard tools from functional data analysis).
- **Future Work:**
  - Kernel-based extensions (Maximum Mean Discrepancy with topological kernels).
  - Other Statistics: can we estimate the covariance operator more efficiently?
  - Application to time-varying topology (dynamic systems).