Statistical Invariance of Betti Numbers

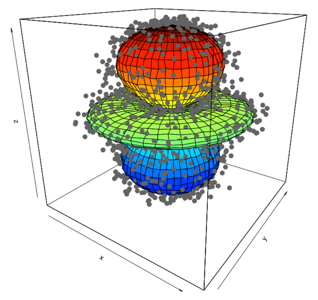
Siddharth Vishwanath [1] Kenji Fukumizu [2] Satoshi Kuriki [2] Bharath Sriperumbudur [1]

[1]Department of Statistics, The Pennsylvania State University [2]Depatrment of Statistical Inference, Institute for Statistical Mathematics

Topology and Data

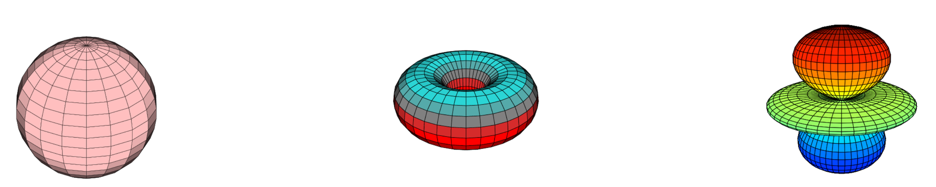
Topological Data Analysis (TDA) is a promising new paradigm comprising of *mathematical*, *statistical* and *algorithmic* tools to study the shape of data.

- Consider a "black-box" probability space $(\Omega, \mathcal{F}, \mathbb{P})$
- $\boldsymbol{X}_n = \{X_1 \dots X_n\}$ is a random collection of points in $\mathbb{R}^d$
- Topological summaries $\mathcal{T}op(\boldsymbol{X}_n)$ are random variables pushed-forward to a summary space $(\mathcal{S}, \mathscr{B}(\mathcal{S}), \mathbb{Q})$

$(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{\boldsymbol{X}_n}$  $\xrightarrow{\mathcal{T}op} \beta_k(\boldsymbol{X}_n), \chi(\boldsymbol{X}_n), \dots$

Betti Numbers and Topological Invariants

- The homology $\{H_k(\mathcal{X})\}_{k \in \mathbb{N}}$ is a topological invariant of $\mathcal{X}$
- The Betti numbers are given by $\beta_k(\mathcal{X}) = \dim(H_k(\mathcal{X}))$
- Informally, $\beta_k(\mathcal{X})$ counts the # of k-dimensional voids in $\mathcal{X}$



$\beta_k = 1, 0, 1, 0 \dots$ $\beta_k = 1, 2, 1, 0 \dots$ $\beta_k = 3, 2, 3, 0 \dots$

The TDA Pipeline

Persistent homology examines topological features across a wide spectrum of resolutions.

1. At resolution $r > 0$ form a *thickening* $\{B_r(X_k)\}_{k=1}^n$
2. Next, construct a simplicial complex K_r (ex. *Čech*, *Rips*, etc.)
3. Examine the simplicial homology for the filtration $\{K_r\}_{r>0}$

Resolution r_1 Resolution r_2 Resolution r_3

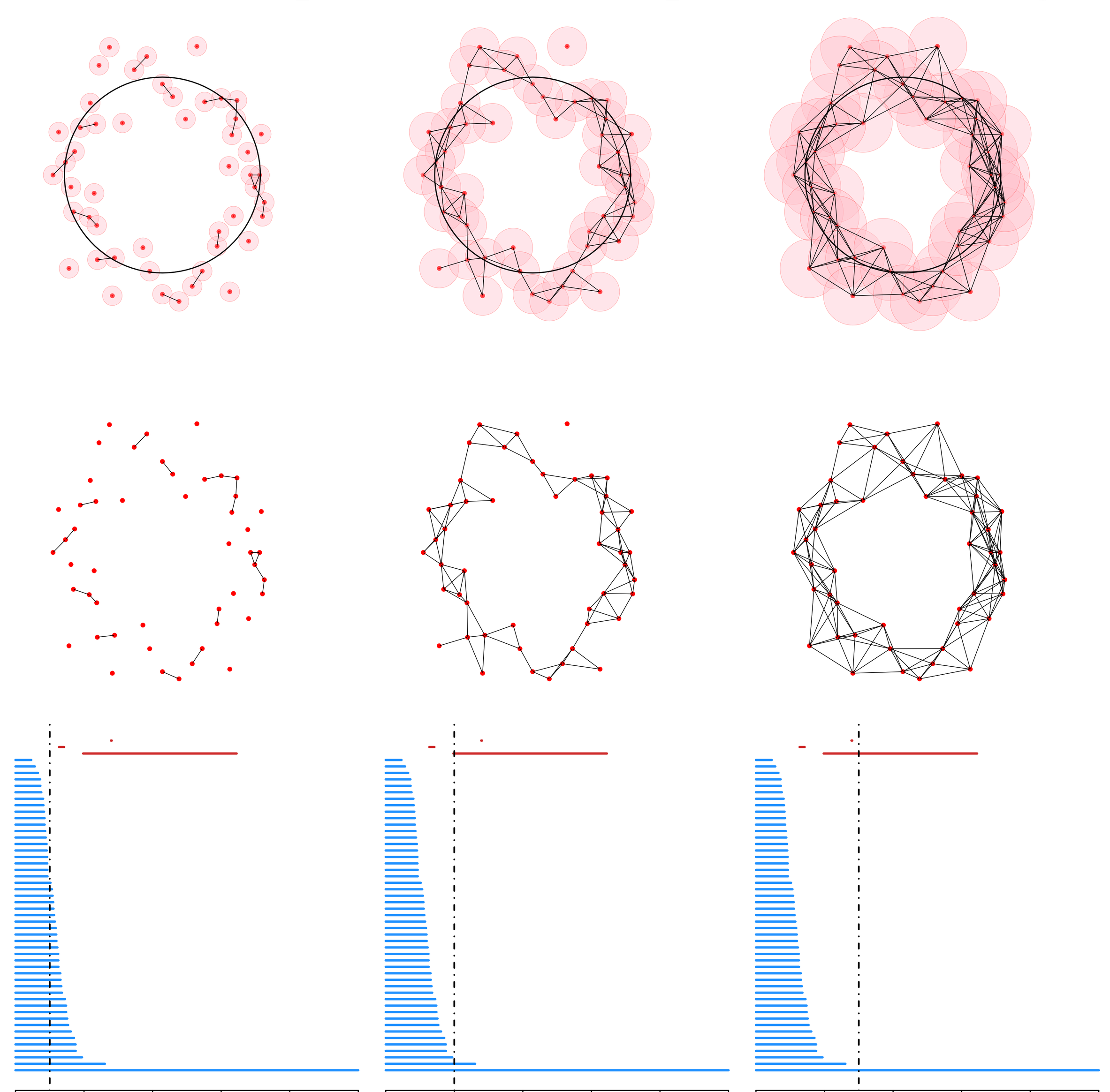


Figure 1. Data is uniformly sampled from a circle.The *persistence barcode* depicts the number of connected components (blue) and the number of loops (red) as the resolution r increases

Motivation

- $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ is a parametric family of distributions
- Let $\boldsymbol{X}_n \sim \mathbb{P}_{\theta_1}$ and $\boldsymbol{Y}_n \sim \mathbb{P}_{\theta_2}$ be two collections of points which are from fundamentally different distributions
- *Our aim is to examine the conditions under which they have identical asymptotic behaviour of the Betti numbers i.e.*

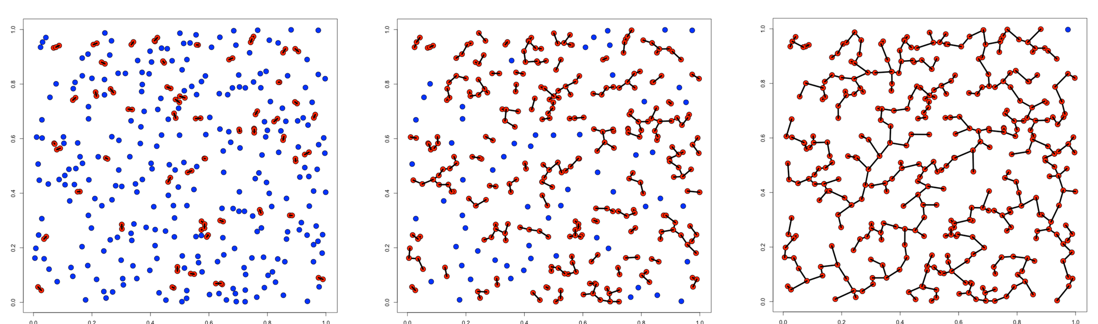
$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left(\beta_k\left(\check{\mathfrak{C}}(\boldsymbol{X}_n, r_n)\right)\right) \overset{?}{=} \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left(\beta_k\left(\check{\mathfrak{C}}(\boldsymbol{Y}_n, r_n)\right)\right)$$

- *This outlines conditions when* ***topological inference*** *is possible*

Asymptotic Regimes

The asymptotic behaviour (as $n \to \infty$) is qualitatively different as the behaviour of the resolution $r_n \to 0$ varies. These are:

- Sparse regime : $r_n = o\left(n^{-1/d}\right)$
- Thermodynamic regime : $r_n = \Theta\left(n^{-1/d}\right)$
- Dense regime : $r_n = \omega\left(n^{-1/d}\right)$



Sparse Thermodynamic Dense

Table 1. Asymptotic regimes of β_0 for $2D$-Poisson Process with $\lambda = 700$

Minimal Spanning Trees

Minimal spanning trees are intrinsically related to $\beta_0(\boldsymbol{X}_n, r)$

(Lemma): Let $\boldsymbol{X}_n = \{X_1, X_2 \dots X_n\}$ be a random collection of points in $\mathbb{R}^d$. Then the following hold true:

1. The Euclidean MST – $\mathcal{M}(\boldsymbol{X}_n)$ is unique a.s. $\mathbb{P}$
2. The smallest edge e^* is an element of $\mathcal{M}(\boldsymbol{X}_n)$
3. For each $\boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{X}_n$ s.t. $\boldsymbol{Y} \cap \boldsymbol{Z} = \varnothing$ and $\boldsymbol{Y} \cup \boldsymbol{Z} = \boldsymbol{X}_n$, the edge e defined by $\|e\| = \min_{y \in \boldsymbol{Y}, z \in \boldsymbol{Z}} \|y - z\|$ is in $\mathcal{M}(\boldsymbol{X}_n)$.

This reveals the relationship between the Euclidean MST and the 0^{th} persistence barcode.

(Theorem): Under the conditions of the previous Lemma:

1. The edges of $\mathcal{M}(\boldsymbol{X}_n)$ generate the 0^{th} persistence barcode.
2. The 0^{th} persistent Betti number at resolution r is given by

$$\beta_0\left(\check{\mathfrak{C}}(\boldsymbol{X}_n)\right) = n - \sum_{e \in \mathcal{M}(\boldsymbol{X}_n)} \mathbb{1}_{[0,r]}(\|e\|)$$

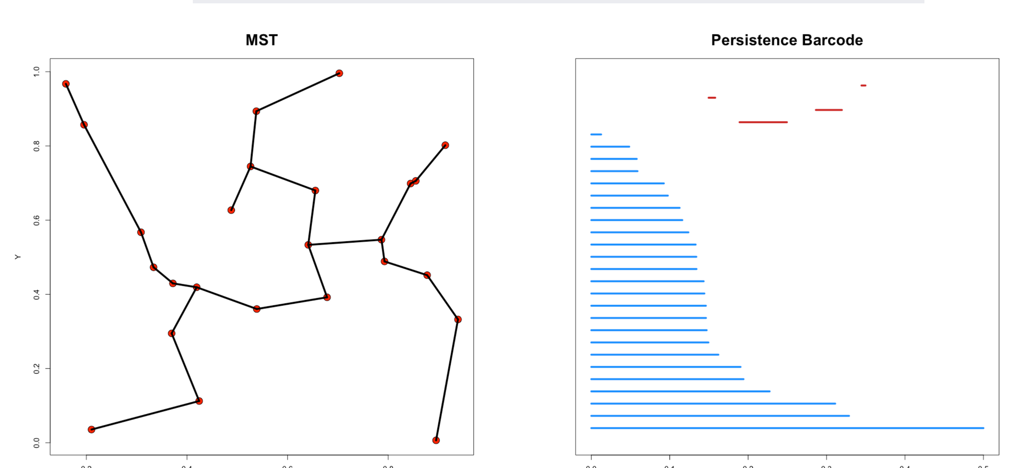


Table 2. Correspondence between $\mathcal{M}(\boldsymbol{X}_n)$ and 0^{th} Barcode

Thermodynamic Behaviour

The thermodynamic regime exhibits interesting behaviour

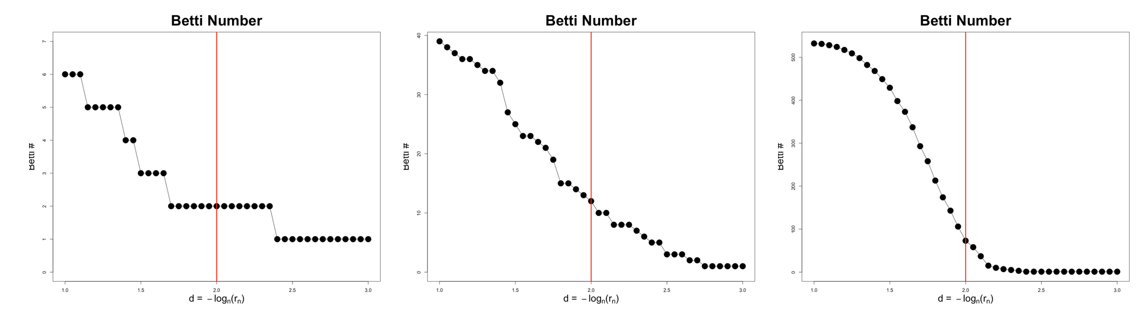


Table 3. β_0 in the theormodynamic regime for PP with $\lambda = 10, 50, 500$

(Theorem): Let $\boldsymbol{X}_n \overset{iid}{\sim} f$, where $f(\boldsymbol{x})$ is bounded, Riemann integrable with compact support. When $n^{1/d} r_n \to t \in (0, \infty)$:

$$\frac{1}{n}\mathbb{E}\left(\beta_0\left(\check{\mathfrak{C}}(\boldsymbol{X}_n, r_n)\right)\right) \to \int_{\mathbb{R}^d} \mathbb{E}\left(\sum_{e \in \mathcal{M}(\mathcal{P}_{1,0})} \mathbb{1}_{[0,t]}\left(f(\boldsymbol{x})^{\frac{-1}{d}}\|e\|\right)\right) f(\boldsymbol{x}) d\boldsymbol{x}$$

where $M(\mathcal{P}_{1,0})$ is the MST for the unit intensity Poisson process with a point at the origin. Now, when we look at any β_k

(Theorem 3.3, [3]) Under the conditions of the Theorem above there exist functions $\hat{\beta}_k$ such that:

$$\frac{1}{n}\mathbb{E}\left(\beta_k\left(\check{\mathfrak{C}}(\boldsymbol{X}_n, r_n)\right)\right) \xrightarrow{n\to\infty} \int_{\mathbb{R}^d} \hat{\beta}_k\left(f(\boldsymbol{x})^{1/d} t\right) f(\boldsymbol{x}) d\boldsymbol{x} := \Psi_k(f, t)$$

Statistical Invariance : Characterization

Define $\mathcal{F}_k$ as a Ψ_k-invariant family of densities such that $\Psi_k(f, t) = \Psi_k(g, t)\ \forall f, g \in \mathcal{F}_k$. These densities admit identical behaviour for β_k in the thermodynamic regime.

We *define* $\mathcal{F}^*$ as the family of densities such that for each $t \geq 0$
$\mathbb{E}(\mathbb{1}(f(\boldsymbol{X}) \geq t)) = \mathbb{E}(\mathbb{1}(g(\boldsymbol{Y}) \geq t))\ \ \forall \boldsymbol{X} \sim f, \boldsymbol{Y} \sim g.$

For families indexed by Θ we denote them $\mathcal{F}^*(\Theta)$ and $\mathcal{F}_k(\Theta)$

(Lemma): $\mathcal{F}^* \subset \bigcap_{k=0}^{\infty} \mathcal{F}_k$

Thus, $\mathcal{F}^*$ admits identical behaviour for each Betti number β_k

Statistical Invariance : I

We employ groups to characterize strong invariance properties.

(Theorem): Suppose $\mathcal{G}$ is a group of Borel-measurable isometries acting on $\mathcal{X} \subseteq \mathbb{R}^d$, and $T : \mathcal{X} \to \mathcal{T}$ is $\mathcal{G}$-maximal invariant. If $\boldsymbol{X}_\theta \sim f_\theta(\boldsymbol{x})$ where

$$f_\theta(\boldsymbol{x}) = \phi\left(g_\theta \circ \Psi(\boldsymbol{x})\right)$$

where $\Psi \in \mathcal{C}^1(\mathcal{X})$; and, $\phi : \mathcal{X} \to \mathbb{R}_{\geq 0}$ is some function which ensures that f_θ is a valid density.

Then, $\{f_\theta : \theta \in \Theta\}$ admits $\mathcal{F}^*(\Theta)$-invariance if and only if $det\left(\boldsymbol{J}_{\Psi^{-1}}(x)\right) = \zeta\left(T(\boldsymbol{x})\right)$ for some function $\zeta : \mathcal{T} \to \mathbb{R}$

This gives us *necessary and sufficient conditions* for **identifying** f_θ upto isometry from the asymptotic behaviour of Betti numbers.

(Theorem): Let $\mathcal{P}$ be a family of distributions such that, for each $f_\theta \in \mathcal{P}$, f_θ statisfies stochastic regularity conditions. Then, $\mathcal{P}$ admits $\mathcal{F}^*$-invariance if and only if

$$\langle f_\theta^k, S_\theta \rangle_{L^2} = 0 \quad \forall k \in \mathbb{N}$$

where, S_θ is the score-function given by $S_\theta(\boldsymbol{x}) = \nabla_\theta \log(f_\theta(\boldsymbol{x}))$

(Example 1): Consider $\mathcal{X} = \mathbb{R}^2$ and $R_\theta = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}$ is the 2D-rotation matrix with $v_\theta := (\cos(\theta), \sin(\theta))^\top$.
Denote $\Phi(\boldsymbol{x})$ as the CDF of a $\mathcal{N}(\boldsymbol{0}, I_2)$ distribution. Then,

$$f_\theta(x, y) = \left(v_\theta^\top \Phi^{-1}(x, y)\right)^2 = \left(\cos(\theta)\Phi^{-1}(x) + \sin(\theta)\Phi^{-1}(y)\right)^2$$

admits invariance for each $2D$-rotation R_θ.

In the general case, $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{G} = \mathcal{SO}(d)$ with $v_\theta \in S^{d-1}$

Statistical Invariance : Non-isometric Cases

We illustrate conditions where fundamentally different distributions admit asymptotic invariance.

(Example 2): Let g be a density on $\mathbb{R}_+$ with $\phi_a(x) = ax$ and $\phi_b(x) = -bx$ with the condition that $\frac{1}{a} + \frac{1}{b} = 1$. Then,

$f(x) = \{g(ax)\mathbb{1}(x \geq 0) + g(-bx)\mathbb{1}(x \leq 0)\}$ admits invariance

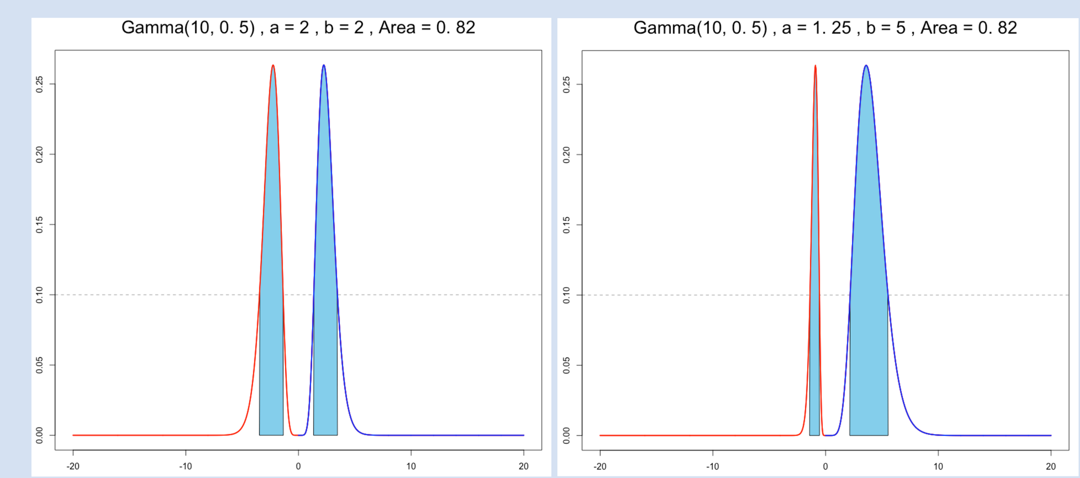


Table 4. Illustration of invariance for $Gamma(10, 0.5)$ distribution

(Abridged Theorem): Under some technical conditions on ν, μ and Θ; Suppose g is a density with respect to ν which satisfies $\nu(d\boldsymbol{y}) = \Psi\left(\left|\det\left(J_\phi\right)\right|\right)\nu(d\boldsymbol{x})$, where $\phi_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^d$ is a full-rank linear bijection for each $\boldsymbol{\theta} \in (\Theta, \Xi, \mu)$. Define,

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = g(\phi_{\boldsymbol{\theta}}(\boldsymbol{x}))$$

Then f admits $\mathcal{F}^*$-invariance for each $(\phi_{\boldsymbol{\theta}}, \mu)$ if

$$\int_\Theta \Psi\left(\left|\det\left(J_{\phi_{\boldsymbol{\theta}}^{-1}}\right)\right|\right)\mu(d\boldsymbol{\theta}) = 1$$

(Example 3): Suppose $g(r)$ be a density with $supp(g) \subseteq \mathbb{R}_+$ w.r.t. the measure ν such that $\nu(dr) = d(r^d) = r^{d-1}dr$.

Let $\Theta = S^{d-1}$ and $a : S^{d-1} \to \mathbb{R}_+$ be a non-negative function.

For each $\boldsymbol{\theta} \in S^{d-1}$ define the mapping $\phi_{\boldsymbol{\theta}}(r) = ra(\boldsymbol{\theta})$ such that

$$\int_{S^{d-1}} \frac{\mu(d\boldsymbol{\theta})}{a(\boldsymbol{\theta})^d} = 1$$

Then for each such $(a(\boldsymbol{\theta}), \mu)$ we have that

$$f(\boldsymbol{x}) = f(r, \boldsymbol{\theta}) = g(ra(\boldsymbol{\theta})) \text{ admits invariance}$$

References

[1] Morris L Eaton. Group invariance applications in statistics. In *Regional conference series in Probability and Statistics*, pages i–133. JSTOR, 1989.

[2] Mathew D Penrose, Joseph E Yukich, et al. Weak laws of large numbers in geometric probability. *The Annals of Applied Probability*, 13(1):277–303, 2003.

[3] Khanh Duy Trinh. A remark on the convergence of betti numbers in the thermodynamic regime. *Pacific Journal of Mathematics for Industry*, 9(1):4, 2017.

[4] Robert A Wijsman. Invariant measures on groups and their use in statistics. IMS, 1990.