Applications of Persistent Homology in Statistics

Giseon Heo

Department of Dentistry

Department of Mathematical and Statistical Sciences

University of Alberta

Two Studies

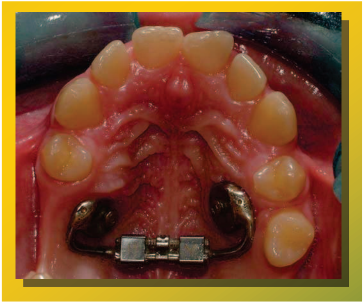
- Part I: **Persistence Diagram: orthodontic study**

  Joint work with Jennifer Gamble and Peter Kim.

- Part II: **Persistence Landscape: protein and 16S rRNA data**

  Joint work with Violeta Kovacev-Nikolic and Peter Bubenik.

Two types of maxillary expanders

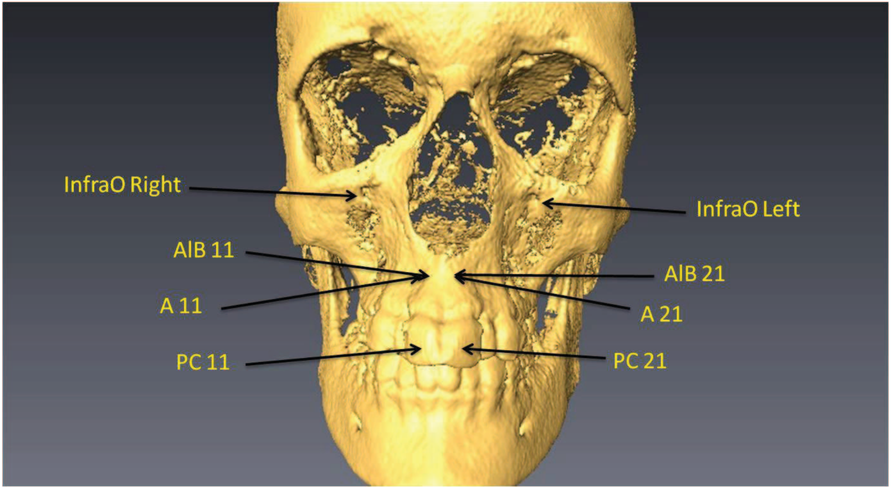


Part I: Clinical trial

- Groups: control (C), bone-anchored (B), and tooth-anchored (T).
- $k = 68$ landmarks in $\mathbb{R}^3$; 19, 21 and 20 subjects in each group.
- Each subject measured at four time points:
  baseline (T_1), mid-treatment (T_2), end of treatment (T_3),
  follow-up (T_4).

- **Research Aim**: Compare the patterns of variability between groups over time.

CBCT scan



3D orthodontic landmark data set

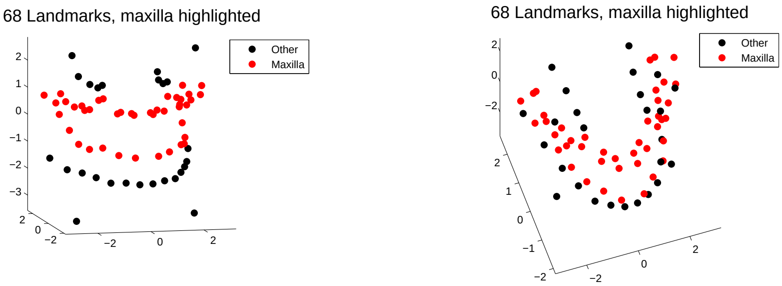


Figure: Three-dimensional plot of the 68 landmarks, with the maxillary landmarks highlighted (left). Overhead 3D view of the 68 landmarks (right).

Three approaches to 3d landmark data

- Shape Analysis (Dryden and Mardia)
- Euclidean Distance Matrix Analysis (Lele)
- **Computational Algebraic Topology**

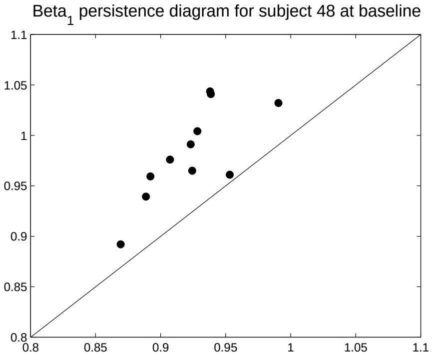
Data analysis applying persistent homology: step 1

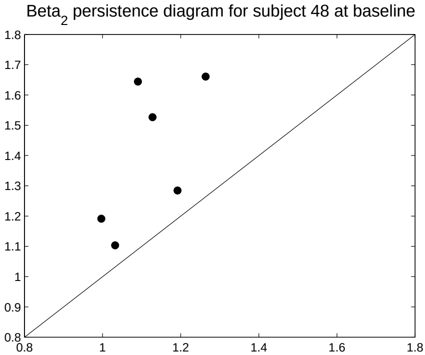
- Each subject ($n = 59 \times 4 = 236$) is represented as a landmark configuration 68×3 matrix.
- Obtain Delaunay triangulation on each configuration.
- The radius of the ball for each configuration is calculated as

$$\varepsilon_{ij} = d_{ij}/\bar{d}_{ij}$$

  where d_{ij} = interlandmark distance of each configuration and $\bar{d}_{ij}$ = average of interlandmark distance among 236 configurations.
- Obtain persistence diagrams using Rips filtration with d_{ij}.

β_1 and β_2 persistence diagrams for a subject 48





Step 2: matching two persistence diagrams

- Bipartite matching algorithm
- Each point in one persistence diagram matched to either a point in the other, or a point along the diagonal
- Let a and b be two landmark configurations with corresponding persistence diagrams of degree ℓ, $\mathrm{Dgm}_\ell(a)$ and $\mathrm{Dgm}_\ell(b)$,

Wasserstein distance between persistence diagrams

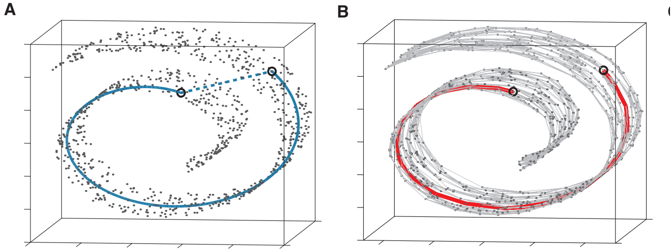
Cohen-Steiner, Edelsbrunner, and Harer 07

- $\gamma_\ell(x) : \mathrm{Dgm}_\ell(a) \longrightarrow \mathrm{Dgm}_\ell(b)$ is a matching between the β_ℓ persistence diagrams
- $$W_p(a,b) = \left[\sum_\ell \inf_{\gamma_\ell} \sum_{x \in \mathrm{Dgm}_\ell(a)} \|x - \gamma_\ell(x)\|_\infty^p \right]^{1/p}$$

Step 3: dimension reduction using Isomap

Tenenbaum, de Silva, and Langford 00

- Using Wasserstein distance between persistence diagrams, we obtain the isometric feature mapping (Isomap) coordinates.
- Geodesic distance between all pairs of points on manifold estimated by the shortest path distances in the Graph.



Step 4: Correlate the Isomap coordinates with interlandmark distances.

- The first β_1 coordinate is associated with both skeletal and dental maxillary expansion (Primary Research Objective).
- The first β_0 coordinate associated primarily with 'vertical elongation.' The vertical elongation is likely due to both growth and appliances.
- The first β_2 coordinate appears to correspond to degree of second molar eruption.

Scatter plot first two β_2 Isomap coordinates

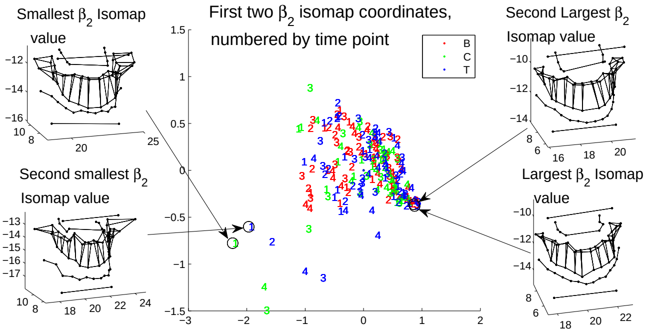


Figure: Triangle formed by root apex, pulp chamber, and alveolar bone landmarks is higher and wider in configurations (unerupted) with small coordinate values, and narrow and lower in configurations (progressive eruption) with larger coordinate values.

Profile plots: first Isomap $\beta_0, \beta_1, \beta_2$ Coordinate

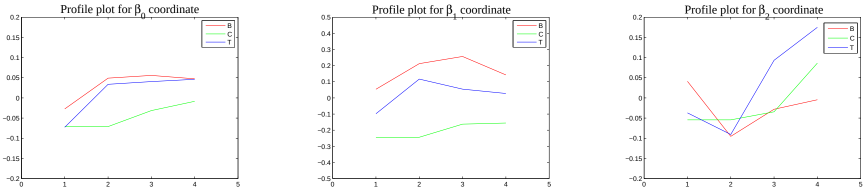


Figure: First β_0 Isomap coordinate is associated with vertical elongation; β_1 both skeletal and dental maxillary expansion; β_2 second molar eruption –smaller coordinate values for unerupted patients and larger coordinate values for progressive eruption.

Step 5: Repeated Measures ANOVA and ANOVA at fixed Time

Factor	form			shape			pca
	β_0	β_1	β_2	β_0	β_1	β_2	pc1
Time	< 0.001	< 0.001	.015	.006	.132	.004	$< .001$
Treatment	.130	.058	.943	.473	.288	.662	.644
Interaction	.021	.078	.317	.931	.329	.425	001
Time 1	.401	.212	.894	490	.957	.680	.326
Time 2	**.023**^{a}	**.012**^{b}	.978	.450	.216	.601	127
Time 3	.122	**.043**^{c}	.728	.462	.301	.708	.916
Time 4	.339	.147	.590	.913	.100	.472	.642

Table: At each fixed time: The Bone group is significantly different ($p < 0.017$) from Control at b, and suggestively different ($p < 0.05$) at a and c. Tooth and Control groups are weakly different ($p < 0.1$) for $a - b$.

Findings from shape analysis

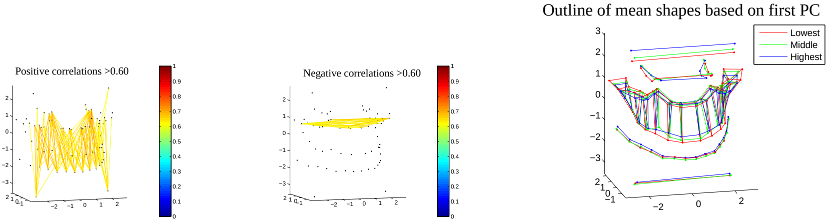


Figure: The landmarks with (left) positive correlation and (middle) negative correlation of $|r| \geq 0.6$ between inter-landmark distances and the first PC.

Small PC scores the configurations are wide in maxillary area but short vertically.The larger the first PC score, the configurations become narrower in maxillar and longer vertically.

Profile plot of first PC

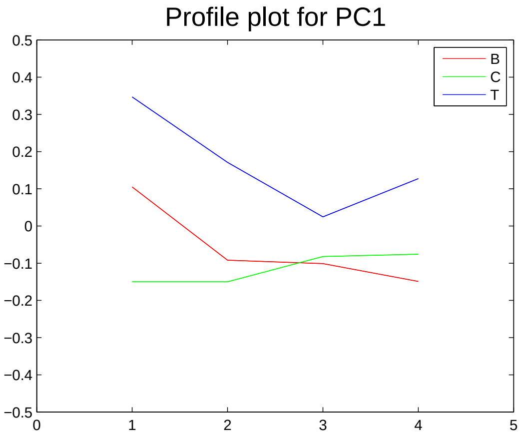
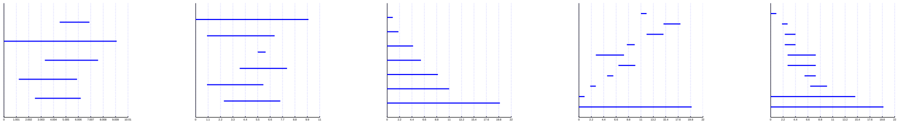


Figure: PC scores of treatments groups decrease over time, this corresponds to maxillary expansion and a relatively shorter vertical configuration.

Comparison to shape analysis

- Both PCA on the tangent space coordinates and persistent homology anova are able to identify the maxillary expansion occurring over time. Anova comparing three treatment groups at each time point, the first PC scores and Hotelling and James tests show no significance.
- The first Isomap coordinate based on β_1 distances explains the maxillary width. This coordinate shows significant effect over time and also treatment effect at time 2 and 3.
- The β_2 coordinate shows interesting feature such as molar eruption. Molar eruption stage is an important information on choice of orthodontic appliances.

Part II: Statistics with descriptors



- How do we calculate the mean and variance?
- Can we apply it to hypothesis testing?

New descriptor (Bubenik 2012)
Statistical topology using persistence landscapes

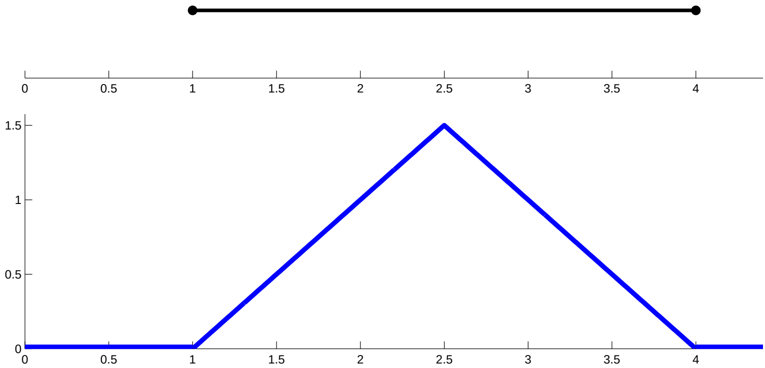


Figure: For (a, b), define $f_{(a,b)} : \mathbb{R} \to \mathbb{R}$ by $f_{(a,b)}(t) = \min(t - a, b - t)_+$

Persistence Landscape (Bubenik)

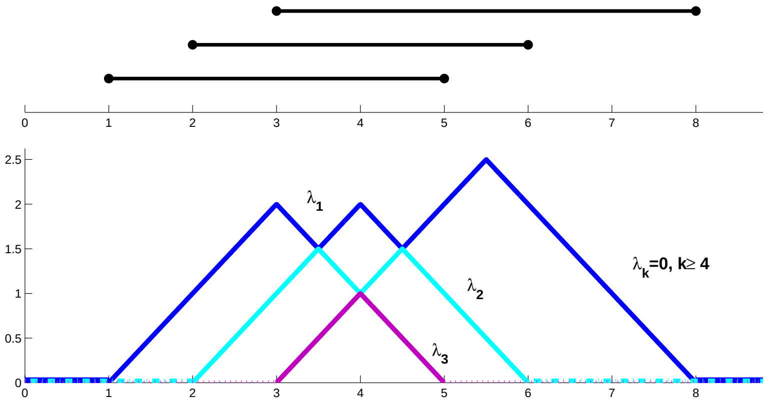
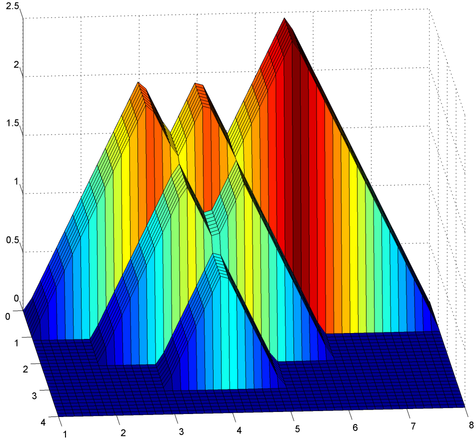
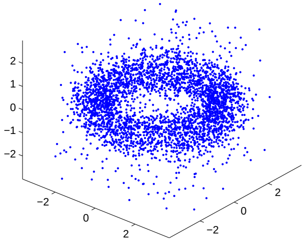


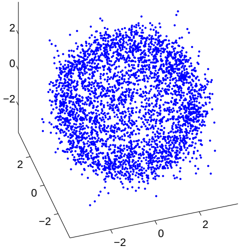
Figure: For $\{(a_i, b_i)\}_{i=1}^{m}$, $\left\{\lambda(k,t) = k^{th} \textit{ largest value of } \{f_{(a_i,b_i)}(t)\}_{i=1}^{m}\right\}$

Extend $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ to $\lambda : \mathbb{R}^2 \to \mathbb{R}$

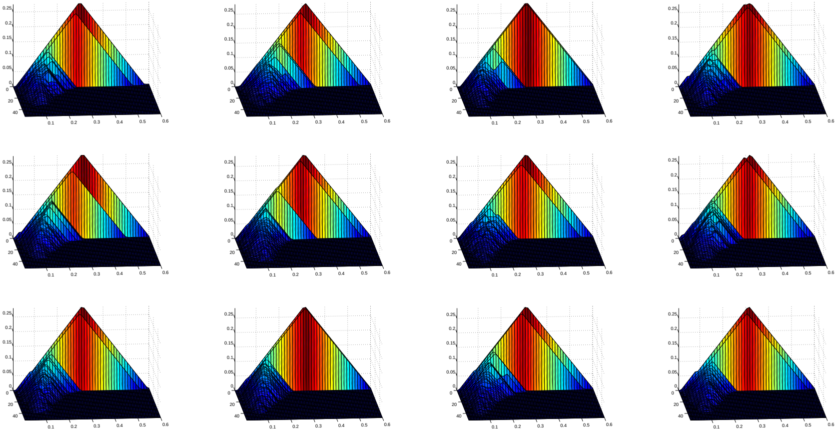


Noisy torus vs noisy sphere: Point clouds





Several persistence landscapes of torus dim 1



Average and Variance

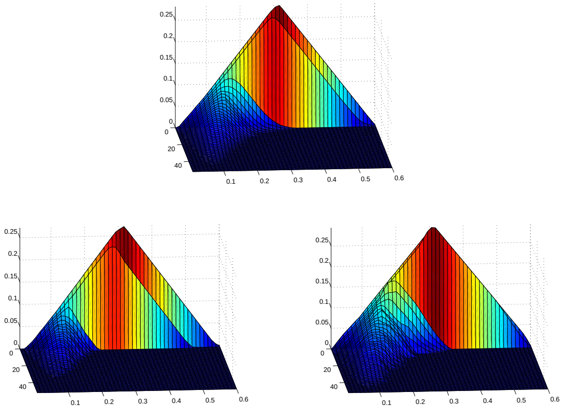


Figure: Pointwise average and ±2 standard deviations

Diagrams of vector spaces (Bubenik 2012)

- Let M be a diagram of finite dimensional vector spaces indexed by $(\mathbb{R}, \leq)$.
- That is, for each $a \in \mathbb{R}$, $M(a)$ is a finite dimensional vector space over some field, $\mathbb{F}$, and for all $a \leq b$, we have a linear map $M(a) \to M(b)$, denoted by $M(a \leq b)$, satisfies

  (i) for all a, $M(a \leq a) = \mathrm{Id}_{M(a)}$,
  (ii) for all $a \leq b \leq c$, $M(a \leq c) = M(b \leq c) \circ M(a \leq b)$.

  Denote $M \in \mathbf{Vec}^{(\mathbb{R}, \leq)}$.

- Example of diagrams:
  (i) Homology group of filtered simplicial complexes $H_*(K_i, \mathbb{F}) \in \mathbf{Vec}^{[\mathbf{n}]}$.
  (ii) Homology group of sublevel sets $H_*(f^{-1}(-\infty, a], \mathbb{F}) \in \mathbf{Vec}^{(\mathbb{R}, \leq)}$.

Betti number

- Define *p-persistence* of $M(a)$ to the image of the map $M(a \leq a + p)$.
- For $a \leq b$ define the corresponding Betti number of M for $a \leq b$ as

$$\beta^{a,b}(M) = \dim(\mathrm{im}(M(a \leq b))).$$

Diagram of interval (Bubenik 2012)

- For an interval $\mathrm{I} \subseteq \mathbb{R}$, define $\chi_{\mathrm{I}} \in \mathbf{Vec}^{(\mathbb{R},\leq)}$ by

$$\chi_{\mathrm{I}}(c) = \begin{cases} \mathbb{F} & \text{if } c \in \mathrm{I} \\ 0 & \text{if } c \notin \mathrm{I} \end{cases}$$

  and

$$\chi_{\mathrm{I}}(c \leq d) = \begin{cases} \mathrm{Id}_{\mathbb{F}} & \text{if } c, d \in \mathrm{I} \\ 0 & \text{if otherwise.} \end{cases}$$

- For $\mathcal{B} = \{\mathrm{I}_i\}_{i=1}^{k}$, define $\chi_{\mathcal{B}} \in \mathbf{Vec}^{(\mathbb{R},\leq)}$ by

  $\chi_{\mathcal{B}}(c) = \oplus^k \chi_{\mathrm{I}_i}(c)$ and $\chi_{\mathcal{B}}(c \leq d) = \oplus^k \chi_{\mathrm{I}_i}(c \leq d)$

Persistence landscape from diagram (Bubenik 2012)

- Let $M \in Vec^{(\mathbb{R},\leq)}$. The persistence of landscape of M is the sequence of functions $\{\lambda_k(M) : \mathbb{R} \to \mathbb{R} \cup \{-\infty, \infty\}\}_{k \in \mathbb{N}}$ given by

$$\lambda_k(M)(t) = \sup(s \geq 0 | \beta^{t-s,t+s}(M) \geq k).$$

  Define $\lambda(M) : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ by $\lambda(M)(k,t) = \lambda_k(t)$.

- For $\mathcal{B} = \{(a_i, b_i)\}_{i=1}^{m}$, where $-\infty \leq a_i \leq b_i \leq \infty$, then the diagram $\chi_{\mathcal{B}} \in \mathbf{Vec}^{(\mathbb{R},\leq)}$ and

$$\lambda(\chi_{\mathcal{B}})(k,t) = k^{th} \text{ largest element of } \min(t - a_i, b_i - t)_+.$$

A space for persistence landscapes (Bubenik 2012)

- Let PL be the convex hull of the set $\{\lambda(M) | M \in \textbf{Vec}^{(\mathbb{R},\leq)}\}$.
- Let $\mathrm{PL}^{\mathrm{p}} = \mathrm{PL} \cap \mathrm{L}^{\mathrm{p}}$ together with the metric from L^{p}.
- PL^{p} is complete and separable, that is, it is a Polish space.
- In PL^{p}, the Fréchet mean is given by the pointwise mean, and the Fréchet variance is the integral of the pointwise variances.

Central Limit Theorem (Bubenik 2012)

- Let $\lambda_1, \ldots, \lambda_n$ be a sample of persistence landscapes drawn from a probability measure with Fréchet mean μ.
- Let $\bar{\lambda}_n$ be pointwise mean of the sample.

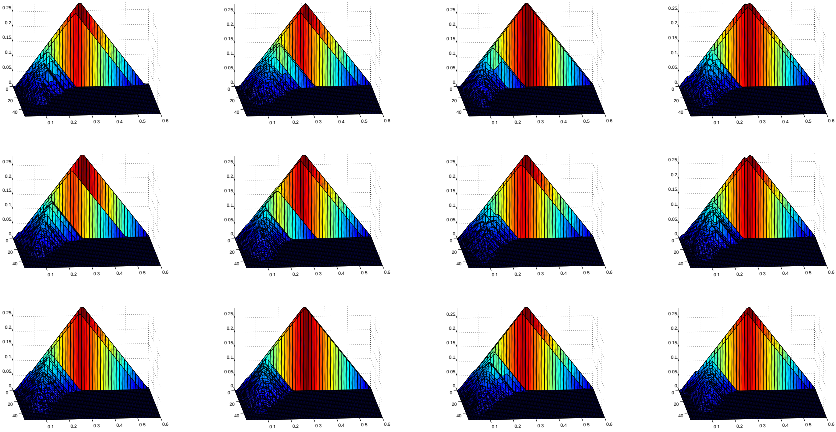
  For all x and t,
- Strong Law of Large Numbers

$$\bar{\lambda}_n(x,t) \xrightarrow{a.s.} \mu(x,t)$$

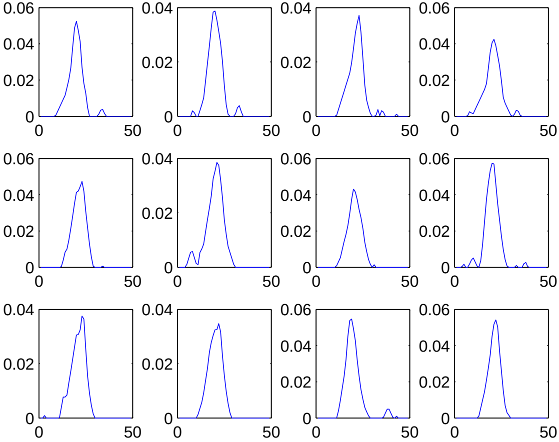
- Central Limit theorem

$$\sqrt{n}(\bar{\lambda}_n(x,t) - \mu(x,t))/\sigma \xrightarrow{d} N(0,1)$$

Several persistence landscapes of torus dim 1



Average of $\lambda(k, t)$ over k



Hypothesis testing

- Let $\mu_1(t)$ and $\mu_2(t)$ be population mean of persistence landscapes for sphere and torus, where $t \in (a, b) := \tau$.
- $H_0 : \mu_1(t) = \mu_2(t)$ for all $t \in \tau$ $\quad H_a : \mu_1(t) \neq \mu_2(t)$ for some $t \in \tau$
- The null hypothesis above is the intersection of hypothesis, $H_0 = \bigcap_{t\in\tau} H_0(t)$, where $H_0(t) : \mu_1(t) = \mu_2(t)$ for each fixed t.
- The type I error for the multiple comparisons is then given by

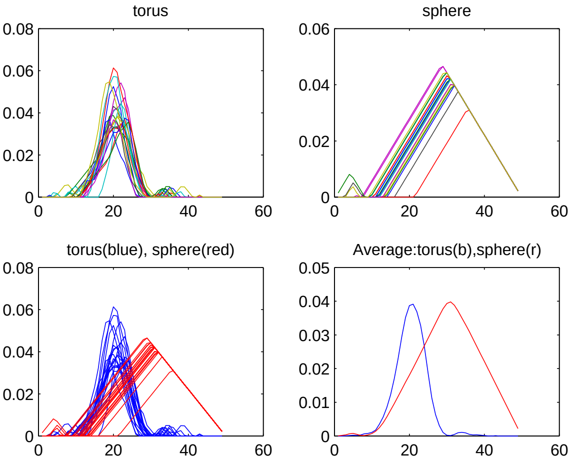
$$\alpha = P\left\{\bigcup_{t\in\tau}(T(t) > h)\right\} = P\left\{\sup_{t\in\tau} T(t) > h\right\}$$

.

Permutation test

- Define test statistic $T(t) = \frac{|\overline{\lambda}_1(t) - \overline{\lambda}_2(t)|}{\mathrm{sd}(\overline{\lambda}_1(\mathrm{t}) - \overline{\lambda}_2(\mathrm{t}))}$.
- Shuffle the labels of $\lambda(t)$ k times and each permutation calculate $T(t)$ at d number of points, store it in $d \times k$ matrix, $T_{nullvalues}(t)$.
- Take maximum of $T_{nullvalues}(t)$ over d number of time points, which forms the null distribution and call it k vector, T_{null}.
- Let $T_{obs}(t)$ be the test statistic of the original data (without shuffling) and set $T_{maxobs}(t)$ as the maximum value of $T_{obs}(t)$.
- Calculate p-value by averaging the number of times that $T_{null}(t) \geq T_{maxobs}(t)$.

β_2 PL of noisy torus and sphere (p-value < 0.001)



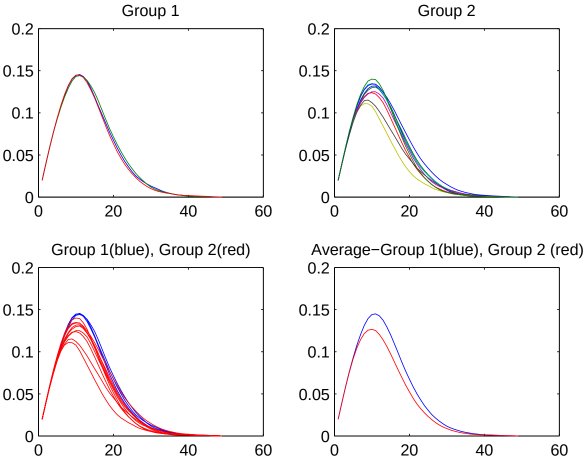
HIV-1 Protease (Yi Mao 2011)
Dynamical basis for drug resistance of HIV-1 protease

- Twelve proteins (correlation of 198 amino acid sequences)

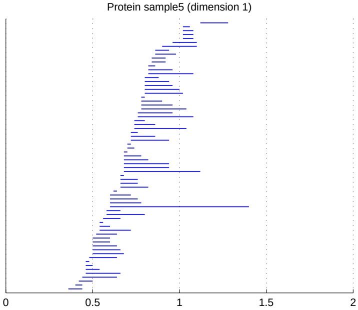
1hpv	1hxb	1hxw	1mui
3jvy	1hvr	2b7z	2fns
2o4k	2o4p	2pyn	1hhp

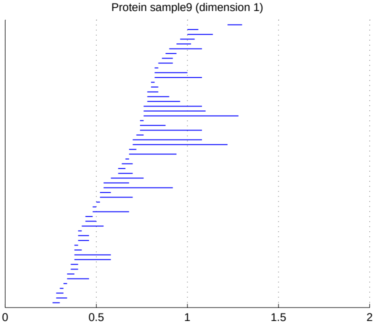
- 3jvy, 2b7z, 2pyn–contain drug-resistant mutations while others do not.

β_0 PL analysis (p-value=0.041)

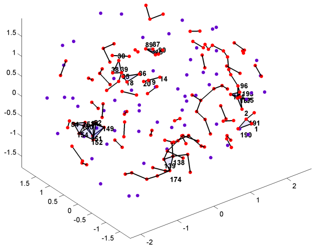


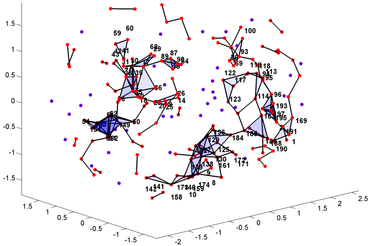
Comparing samples 3jvy and 2o4k (β_1 barcode)

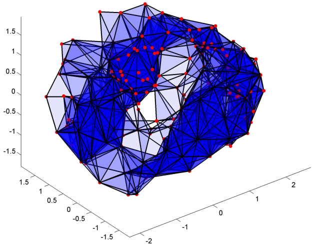




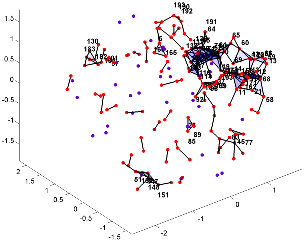
Snap shot of clustering–3jvy

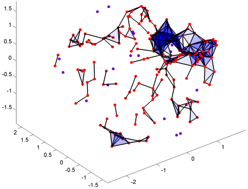


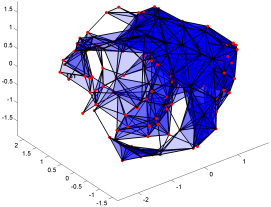




Snap shot of clustering–2o4k



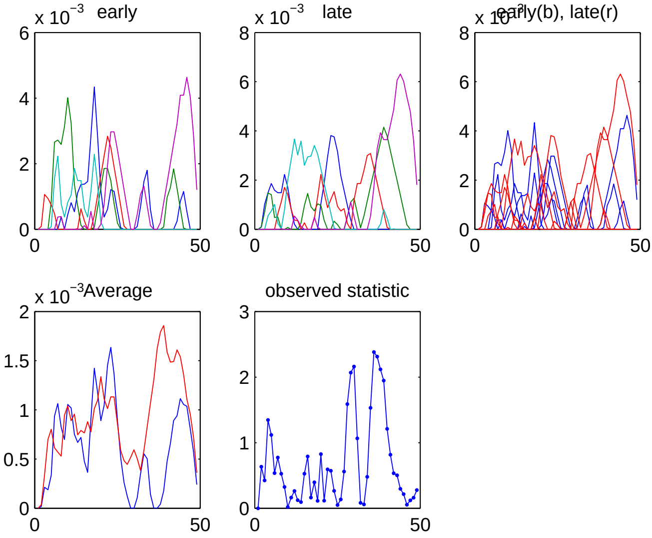




16S rRNA gene sequences: MOTHUR (Schloss lab)

- Collected fresh feces from mice on a daily basis for 365 days post weaning (dpw).
- During the first 150 days post weaning, mice were allowed eat, get fat, and be merry.
- The rapid change in weight observed during the first 10 dpw affected the stability of microbiome compared to the microbiome observed between 140 and 150 days.
- Early (0, 2, 4, 6, and 8 days) and late (142, 144, 146, 148 and 150) were compared.
- Approx. 700,000 genomic DNA sequences.

Permutation test with β_1 PL



Acknowledgement

- The ATMCS conference organizers.
- Orthodontists, Paul Major and Manuel Lagravere.
- Jennifer Gamble, Violeta Kovacev-Nikolic, Peter Bubenik, Peter Kim, Yi Mao, and Yin Li.
- McIntyre Memorial Fund and NSERC.