Limit Theorems for Betti Numbers of Extreme Sample Clouds

Takashi Owada

Technion-Israel Institute of Technology

May 2016

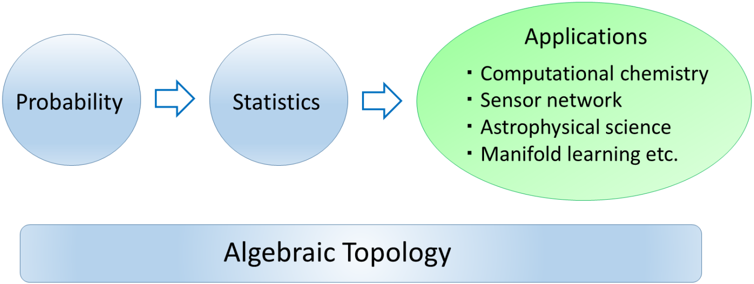
Topological data analysis

- Topological data analysis (TDA) is an approach to the analysis of datasets using techniques from topology and other mathematics.
- Typically, topologists classify objects into classes of "similar shapes" by the number of holes.

Topological data analysis

- Topological data analysis (TDA) is an approach to the analysis of datasets using techniques from topology and other mathematics.

- Typically, topologists classify objects into classes of “similar shapes” by the number of holes.



- As highlighted in a recent series of columns in the IMS Bulletin, the collaboration of three different disciplines, topology, probability, and statistics, is indispensable for the development of TDA.
    - The author of the column has invented a word, TOPOS (=topology, probability, and statistics).
- However, there are still only limited number of probabilistic and statistical works in TDA.

Betti numbers

- Basic quantifier in algebraic topology.
- Given a topological space X, the 0-th Betti number $\beta_0(X)$ is defined as

$$\beta_0(X) = \text{the number of connected components in } X.$$

- For $k \geq 1$, the k-th Betti number $\beta_k(X)$ is defined as

$$\beta_k(X) = \text{the number of } k\text{-dim holes in } X.$$

Betti numbers

- Basic quantifier in algebraic topology.
- Given a topological space X, the 0-th Betti number $\beta_0(X)$ is defined as

$$\beta_0(X) = \text{the number of connected components in } X.$$

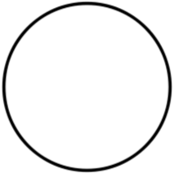
- For $k \geq 1$, the k-th Betti number $\beta_k(X)$ is defined as

$$\beta_k(X) = \text{the number of } k\text{-dim holes in } X.$$

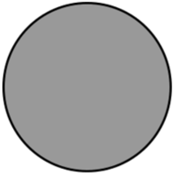
  - More intuitively,

$$\beta_1(X) = \text{the number of ``closed loops'' in } X.$$

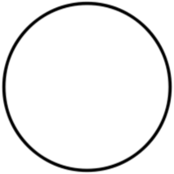
$$\beta_2(X) = \text{the number of ``hollows'' in } X.$$



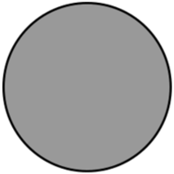
$\beta_0 = 1,\ \beta_1 = 1,$
$\beta_k = 0,\ k \geq 2.$



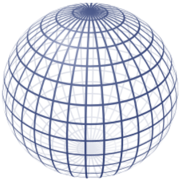
$\beta_0 = 1,\ \beta_k = 0,\ k \geq 1.$



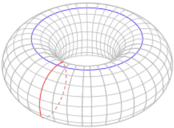
$\beta_0 = 1,\ \beta_1 = 1,$
$\beta_k = 0,\ k \geq 2.$



$\beta_0 = 1,\ \beta_k = 0,\ k \geq 1.$



$\beta_0 = 1,\ \beta_1 = 0,\ \beta_2 = 1,$
$\beta_k = 0,\ k \geq 3.$



$\beta_0 = 1,\ \beta_1 = 2,\ \beta_2 = 1,$
$\beta_k = 0,\ k \geq 3.$

Scheme

1. Generate random sample from a heavy tail distribution.

- (X_i): iid $\mathbb{R}^d$-valued random variables, $d \geq 2$, with spherically symmetric density f.
- f has a regularly varying tail: for some $\alpha > d$,

$$\lim_{r\to\infty} \frac{f(rte_1)}{f(re_1)} = t^{-\alpha} \quad \text{for every } t > 0\,,$$

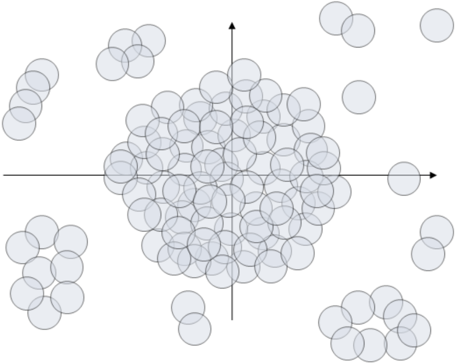
  where $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$.

- To make our story simpler, we will work on a special example in the following.

$$f(x) = C/\big(1 + ||x||^\alpha\big)\,, \quad x \in \mathbb{R}^d,\ \alpha > d\,.$$

2. Draw random balls of radius t about X_i's.

3. Establish the limit theorems for Betti numbers of the union of balls.



$\beta_0(\text{union of balls}) = 11$

$\beta_1(\text{union of balls}) = 3$

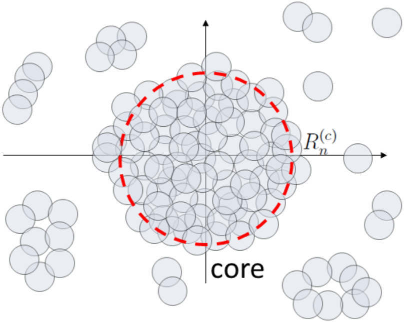
Proposition [Adler et. al, 2014].

There exists a sequence $R_n^{(c)} = C'(n/\log n)^{1/\alpha}$ for some $C' > 0$, such that

$$\mathbb{P}\left\{ B(0; R_n^{(c)}) \subset \bigcup_{X \in \mathcal{X}_n \cap B(0; R_n^{(c)})} B(X; 1) \right\} \to 1\,, \quad n \to \infty,$$

where $\mathcal{X}_n = \{X_1, \dots, X_n\}$.

- $B(0; R_n^{(c)})$ is called a core.



There are no holes inside a core

- The related notion, a weak core, plays a more decisive role in the characterization of the limit theorems.

Definition

Let f be a spherically symmetric density on $\mathbb{R}^d$ and $R_n^{(w)} \to \infty$ be a sequence determined by

$$nf\big(R_n^{(w)}e_1\big) \to 1\,, \quad n \to \infty\,.$$

Then $B(0; R_n^{(w)})$ is called a weak core.

- The related notion, a weak core, plays a more decisive role in the characterization of the limit theorems.

Definition

Let f be a spherically symmetric density on $\mathbb{R}^d$ and $R_n^{(w)} \to \infty$ be a sequence determined by

$$nf\big(R_n^{(w)} e_1\big) \to 1\,, \quad n \to \infty\,.$$

Then $B(0; R_n^{(w)})$ is called a weak core.

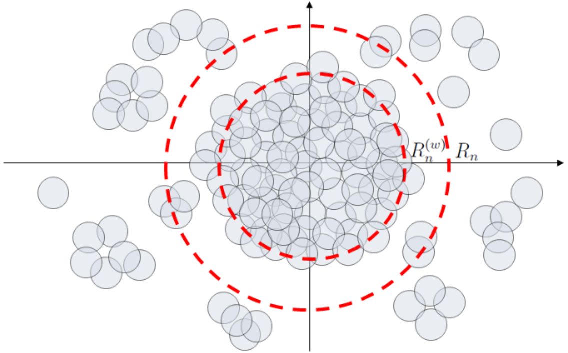
- If $f(x) = C/\big(1 + ||x||^\alpha\big)$, $x \in \mathbb{R}^d$, then $R_n^{(w)} = (Cn)^{1/\alpha}$.
- $R_n^{(c)}/R_n^{(w)} \to 0$, but they have the same regular variation exponent, $1/\alpha$.

Betti number in the tail

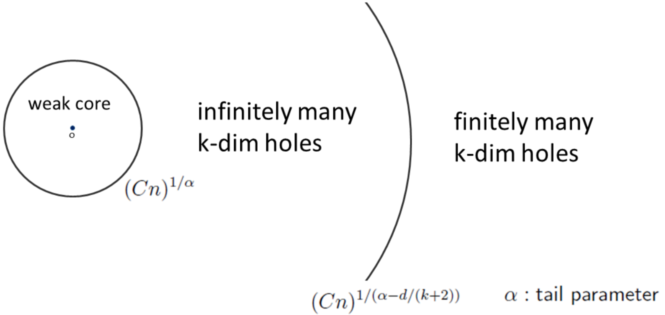
- $\mathcal{X}_n = \{X_1, \dots, X_n\}$: iid $\mathbb{R}^d$-valued random variables drawn from a power-law density with tail parameter α.
- For $k \geq 1$, define
$$\beta_{k,n}(t) := \beta_k\Bigg(\bigcup_{X \in \mathcal{X}_n \setminus B(0;R_n)} B(X;t) \Bigg), \quad t \geq 0\,,$$
where R_n is a non-random sequence with $R_n \geq R_n^{(w)}$ (= radius of a weak core).

Betti number in the tail

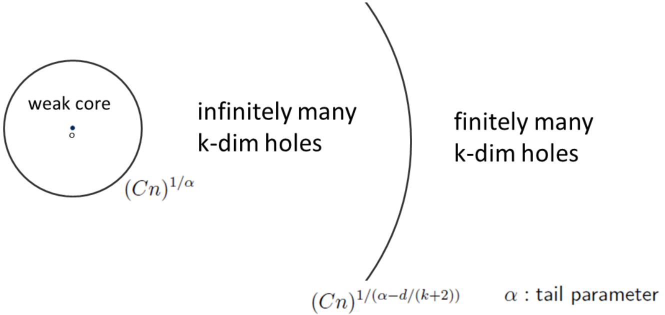
- $\mathcal{X}_n = \{X_1, \dots, X_n\}$: iid $\mathbb{R}^d$-valued random variables drawn from a power-law density with tail parameter α.
- For $k \geq 1$, define
$$\beta_{k,n}(t) := \beta_k \Bigg(\bigcup_{X \in \mathcal{X}_n \setminus B(0;R_n)} B(X;t) \Bigg), \quad t \geq 0,$$
where R_n is a non-random sequence with $R_n \geq R_n^{(w)}$ (= radius of a weak core).



- Roughly speaking, as $n \to \infty$, k-dim holes are distributed as follows.



- Roughly speaking, as $n \to \infty$, k-dim holes are distributed as follows.

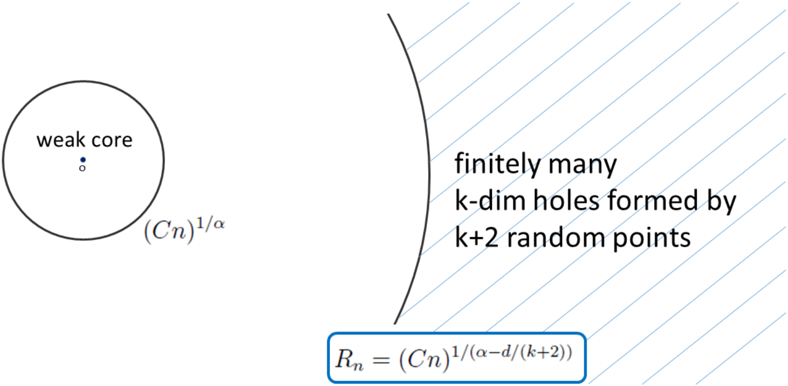


- Three different regimes must be considered.
  We set, respectively,
  - [1]: $R_n = (Cn)^{1/(\alpha-d/(k+2))}$,
  - [2]: $(Cn)^{1/\alpha} \ll R_n \ll (Cn)^{1/(\alpha-d/(k+2))}$,
  - [3]: $R_n = (Cn)^{1/\alpha}$,

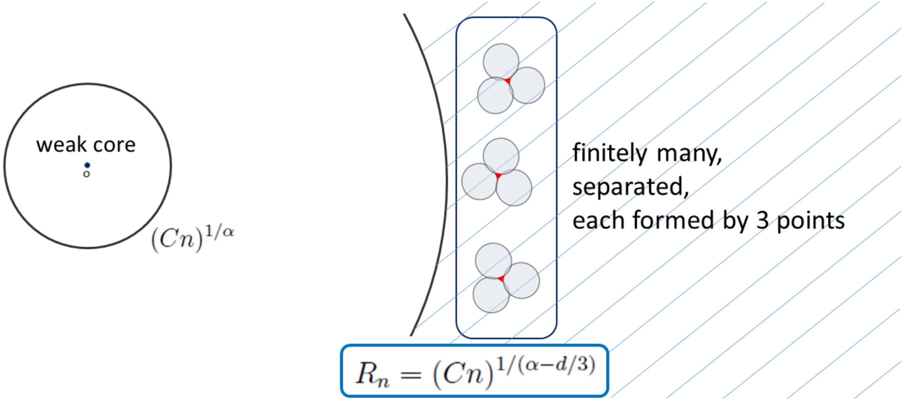
  and compute $\beta_{k,n}(t)$ by counting k-dim holes outside $B(0; R_n)$.

In the regime $[1]$,

- There exist finitely many k-dim holes formed by $k+2$ random points outside $B(0; R_n)$ as $n \to \infty$.



Example $(k = 1)$



- The appearance of holes is a rare event.

Limiting process for $\beta_{k,n}(t)$:

$$N_k(t) := \int_{(\mathbb{R}^d)^{k+1}} h_t(0, y_1, \dots, y_{k+1})\, M_k(d\mathbf{y}).$$

- M_k is a Poisson random measure with Lebesgue intensity measure on $(\mathbb{R}^d)^{k+1}$.
- $h_t(0, y_1, \dots, y_{k+1}) = \mathbf{1}\Big\{ \beta_k\Big(B(0;t) \cup \bigcup_{i=1}^{k+1} B(y_i;t)\Big) = 1 \Big\}$

  with $0, y_1, \dots, y_{k+1} \in \mathbb{R}^d$.

- $h_t(0, \mathbf{y})$ can be expressed as

$$h_t(0, \mathbf{y}) = h_t^+(0, \mathbf{y}) - h_t^-(0, \mathbf{y}),$$

where h_t^+ and h_t^- are some other indicator functions, increasing in t.

- $h_t(0, \mathbf{y})$ can be expressed as

$$h_t(0, \mathbf{y}) = h_t^+(0, \mathbf{y}) - h_t^-(0, \mathbf{y}),$$

  where h_t^+ and h_t^- are some other indicator functions, increasing in t.

- Accordingly,

$$\begin{aligned} N_k(t) &= \int_{(\mathbb{R}^d)^{k+1}} h_t^+(0, \mathbf{y})\, M_k(d\mathbf{y}) - \int_{(\mathbb{R}^d)^{k+1}} h_t^-(0, \mathbf{y})\, M_k(d\mathbf{y}) \\ &:= N_k^+(t) - N_k^-(t). \end{aligned}$$

- We can prove that $N_k^+(t)$ and $N_k^-(t)$ are represented as a (time-changed) Poisson process.
  - However, $N_k(t)$ is not a (time-changed) Poisson process.

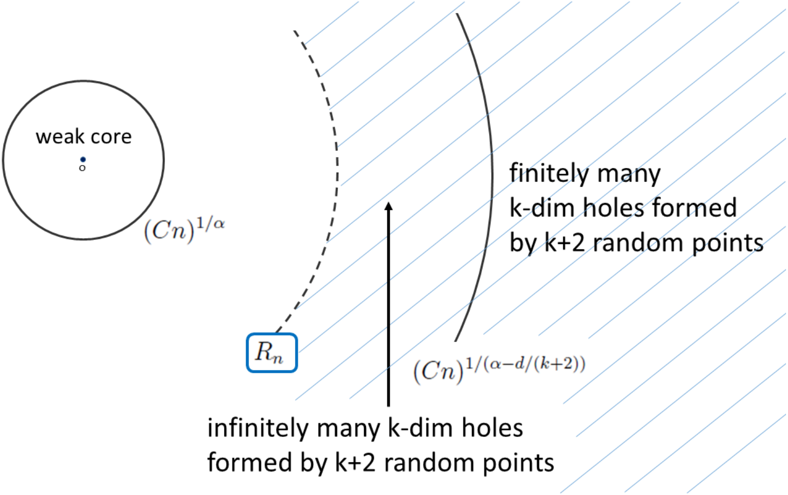
Theorem 1. [O., 2016]

In the regime $[1]$, we have, as $n \to \infty$,

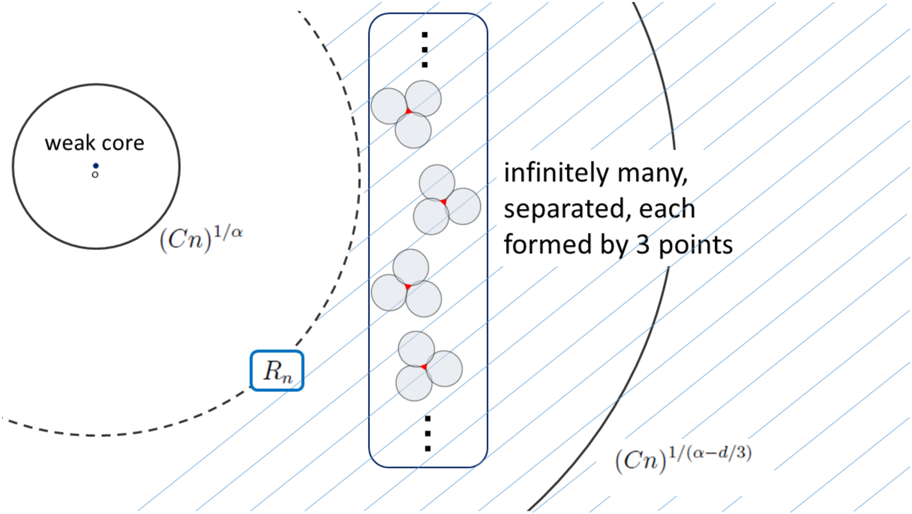
$$\beta_{k,n}(t) \Rightarrow N_k^+(t) - N_k^-(t).$$

In the regime [2],

- There exist infinitely many k-dim holes formed by $k+2$ points outside $B(0; R_n)$ as $n \to \infty$.



Example $(k = 1)$



- The appearance of holes is no longer a rare event.

Limiting process for $\beta_{k,n}(t)$: Define a Gaussian process

$$Y_k(t) := \int_{(\mathbb{R}^d)^{k+1}} h_t(0, y_1, \dots, y_{k+1})\, G_k(d\mathbf{y}).$$

- G_k is a Gaussian random measure with Lebesgue control measure on $(\mathbb{R}^d)^{k+1}$.
- $h_t(0, y_1, \dots, y_{k+1}) = \mathbf{1}\left\{ \beta_k\Big(B(0;t) \cup \bigcup_{i=1}^{k+1} B(y_i;t)\Big) = 1 \right\}$

  with $0, y_1, \dots, y_{k+1} \in \mathbb{R}^d$.

- Using the decomposition $h_t = h_t^+ - h_t^-$, we can write

$$Y_k(t) = \int_{(\mathbb{R}^d)^{k+1}} h_t^+(0, \mathbf{y})\, G_k(d\mathbf{y}) - \int_{(\mathbb{R}^d)^{k+1}} h_t^-(0, \mathbf{y})\, G_k(d\mathbf{y})$$
$$:= Y_k^+(t) - Y_k^-(t).$$

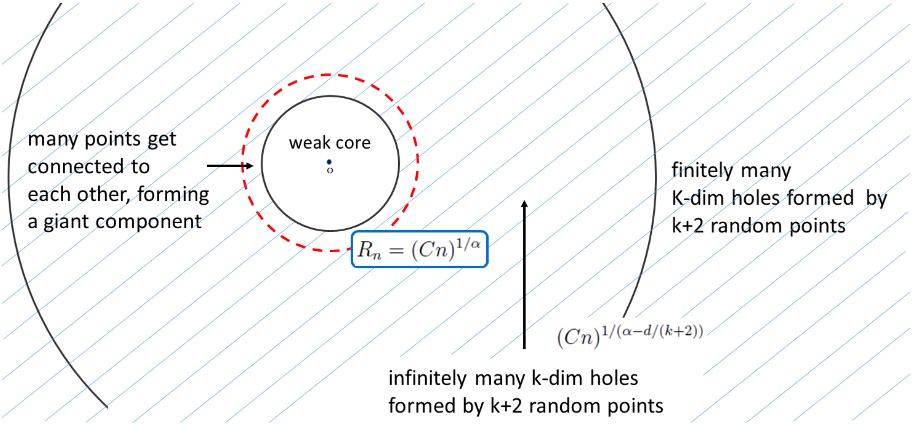
- Then, $Y_k^+(t)$ and $Y_k^-(t)$ are represented as a (time-changed) Brownian motion.
  - $Y_k(t)$ is a Gaussian process, but it is not a (time-changed) Brownian motion.

Theorem 2. [O., 2016]

In the regime $[2]$, we have, as $n \to \infty$,

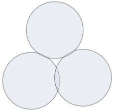
$$\frac{\beta_{k,n}(t) - \mathbb{E}\big\{\beta_{k,n}(t)\big\}}{\big(n^{k+2}R_n^{d-\alpha(k+2)}\big)^{1/2}} \Rightarrow Y_k^+(t) - Y_k^-(t).$$

In the regime [3],

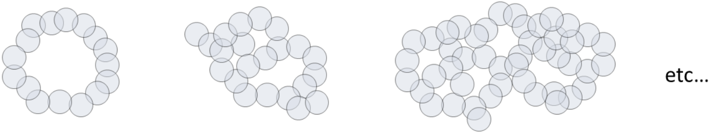


Example $(k = 1)$

- In the regimes $[1]$ and $[2]$, all the one-dim holes contributing to $\beta_{1,n}(t)$ in the limit are always of the form



- In the regime $[3]$, many different kinds of one-dim holes (which exist close enough to a weak core) contribute to $\beta_{1,n}(t)$ in the limit.



The limiting Gaussian process is given by

$$Z_k(t) := \sum_{i=k+2}^{\infty} \sum_{j>0} Z_k^{(i,j)}(t) \, .$$

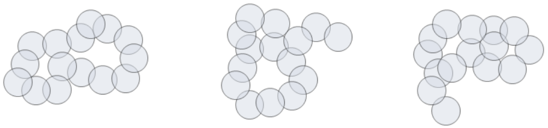
- $Z_k^{(i,j)}(t)$ is a Gaussian process representing the connected components that are formed by i points and contain j holes.

The limiting Gaussian process is given by

$$Z_k(t) := \sum_{i=k+2}^{\infty} \sum_{j>0} Z_k^{(i,j)}(t)\,.$$

- $Z_k^{(i,j)}(t)$ is a Gaussian process representing the connected components that are formed by i points and contain j holes.

Example: $Z_1^{(15,2)}(t)$ (i.e., $k=1$, $i=15$, $j=2$).



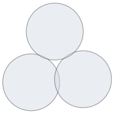
etc...

Rewrite $Z_k(t)$ as

$$Z_k(t) = Z_k^{(k+2,1)}(t) + \sum_{i=k+3}^{\infty} \sum_{j>0} Z_k^{(i,j)}(t) \, .$$

- $Z_k^{(k+2,1)}(t)$ represents the connected components that are formed by $k+2$ points and contain a single k-dimensional hole.

Example: $Z_1^{(3,1)}(t)$ (i.e., $k = 1$, $i = 3$, $j = 1$).

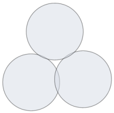


Rewrite $Z_k(t)$ as

$$Z_k(t) = Z_k^{(k+2,1)}(t) + \sum_{i=k+3}^{\infty} \sum_{j>0} Z_k^{(i,j)}(t) \,.$$

- $Z_k^{(k+2,1)}(t)$ represents the connected components that are formed by $k+2$ points and contain a single k-dimensional hole.

Example: $Z_1^{(3,1)}(t)$ (i.e., $k=1$, $i=3$, $j=1$).



- $Z_k^{(k+2,1)}(t)$ is "similar" to the $Y_k(t)$ in the regime $[2]$.

Theorem 3. [O., 2016]

In the regime $[3]$, we have, as $n \to \infty$,

$$\frac{\beta_{k,n}(t) - \mathbb{E}\big\{\beta_{k,n}(t)\big\}}{n^{d/(2\alpha)}} \Rightarrow Z_k(t).$$