Confidence band for persistent homology of density filtration on Rips complex

Jaehyeok Shin
Department of Statistics
Carnegie Mellon University
jaehyeos@andrew.cmu.edu

September 22, 2016

Bobrowski et al. [2014] provided a new way to estimate persistent homology of upper level set density filtration. Instead of directly calculate persistent homology of KDE filtration on $\mathbb{R}^d$ which require grid-approximation, they suggest calculate persistent homology of KDE filtration on Rip complex whose parameter is equal to the bandwidth of KDE. They showed that under some regularity condition, their estimate is consistent up to a small error coming from discrete nature of homology. In this report, I provide a brief idea about how to calculate confidence band for their estimate. In section 1, I provide a short summary of Bobrowski et al. [2014] for persistent homology of density filtration part only. In section 2, I suggest a simple way to calculate confidence band for their estimate.

1 Summary of Bobrowski et al. [2014]

In Bobrowski et al. [2014], the authors focused on estimating homology of upper level set and persistent homology of upper level set filtration with respect to density and regression functions. In this section, I will only focus estimating persistent homology of upper level set density filtration case.

1.1 Notation and Assumptions

Let p be the density of the observed data $\mathscr{D}_n := \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$. The (upper) level sets of p is defined by

$$D_L := \left\{ x \in \mathbb{R}^d : p(x) \geq L \right\} \tag{1}$$

As a basic block, we will use the kernel density estimator defined by

$$\widehat{p}_r(x) := \frac{1}{nr^d C_K} \sum_{i=1}^{n} K_r(x - X_i) \quad \text{where} \quad K_r(x) := K(x/r)$$

Define an estimator of D_L

$$\widehat{D}_L(n,r) := U(\widehat{\mathscr{D}}_n^L, r) := \bigcup_{X \in \widehat{\mathscr{D}}_n^L} B_r(X) \tag{2}$$

where $\widehat{\mathscr{D}}_n^L := \{X_i : \widehat{p}_r(X_i) \geq L; 1 \leq i \leq n\}$, and $B_r(X)$ is the ball centered at X with radius r

We need some assumptions on kernel function K, density p, and level L.

Definition 1. *Let $p : \mathbb{R}^d \to \mathbb{R}$ be the density function and D_L be the upper level set of p.*

1. *We say that L is a homological regular value if there exists $\varepsilon > 0$ such that for every $v_2 \leq v_1$ in $(L-\varepsilon, L+\varepsilon)$ the map $H_k(D_{v_1}) \to H_k(D_{v_2})$ induced by inclusion is an isomorphism for every $k \geq 0$. Otherwise, we say that L is a homological critical value.*
2. *A function f is called* tame *if it has a finite number of homological critical values, and rank$(H_k(D_L))$ is finite for all L and k.*

Definition 2. *Given a level $L > 0$ and $\varepsilon \in (0, L/2)$, we say that L is ε-regular if*

$$\partial D_{L+2\varepsilon} \cap \partial D_{L+\frac{3}{2}\varepsilon} = \partial D_{L+\frac{1}{2}\varepsilon} \cap \partial D_L = \partial D_L \cap \partial D_{L-\frac{1}{2}\varepsilon} = \partial D_{L-\frac{3}{2}\varepsilon} \cap \partial D_{L-2\varepsilon} = \emptyset$$

where ∂ is the set boundary.

Assumption 1. *The marginal density p of X satisfies the following conditions,*

1. *supp(p) is bounded.*
2. *p is tame.*
3. $p_{max} := \sup_{x \in \mathbb{R}^d} p(x) < \infty$
4. *Every L are ε-regular, and the set $D_L \subset \mathbb{R}^d$ are bounded.*

Assumption 2. *The kernel function $K : \mathbb{R}^d \to \mathbb{R}$ satisfies the following conditions,*

1. *supp$(K) \subset B_1(0)$*
2. $K(0) = 1$ *and* $K(x) \leq 1, \ \forall x$
3. $\int K := C_K \in (0,1)$

1.1.1 Key results

Lemma 1. *For every $L > 0, \varepsilon \in (0,L)$, if $r \to 0$ and $nr^d \to \infty$, then for large enough n we have*

$$\mathbb{P}\left(D^{\downarrow}_{L+\varepsilon}(2r) \subset \widehat{D}_L(n,r) \subset D^{\uparrow}_{L-\varepsilon}(2r)\right) \geq 1 - 3ne^{-C^*_\varepsilon nr^d} \tag{3}$$

where $\partial D^r_L := \bigcup_{x \in \partial D_L} B_r(x),\ \ r > 0$, *$D^{\uparrow}_L(r) := D_L \cup \partial D^r_L$, and $D^{\downarrow}_L(r) := D_L \setminus \partial D^r_L$*

Using the above lemma, we can estimate persistent homology of level sets filtration. Let $\mathrm{PH}_*(p)$ be the persistent homology of p constructed by the continuous filtration $\{D_L\}_{L \in \mathbb{R}}$. Define

$$N_\varepsilon := \sup_{x \in \mathbb{R}^d} \lceil p(x)/2\varepsilon \rceil,\ \ L_{\max} = 2\varepsilon N_\varepsilon,\ \text{ and }\ L_i = L_{\max} - 2i\varepsilon$$

and consider the following discrete filtration

$$\left\{\widehat{D}_{L_i}(n,r)\right\}_{i \in \mathbb{Z}} \tag{4}$$

Denoting the persistent homology of the above filtration by $\widehat{\mathrm{PH}}^{\varepsilon}_*(p)$, the authors proved the following theorem.

Theorem 2. *If the assumption 1 and 2 hold, and $r \to 0$, $nr^d \to \infty$, then*

$$\mathbb{P}\left(d_B\left(\widehat{\mathrm{PH}}^{\varepsilon}_*(p), \mathrm{PH}_*(p)\right) \leq 5\varepsilon\right) \geq 1 - 3N_\varepsilon ne^{-C^\star_{\varepsilon/2} nr^d}, \tag{5}$$

where

$$C^\star_\varepsilon = \frac{\varepsilon^2 C_K}{3p_{\max} + \varepsilon}, \quad \textit{and } d_B \textit{ is the bottleneck distance.}$$

In particular, if $nr^d \geq D \log n$ with $D > (C^\star_{\varepsilon/2})^{-1}$, we have

$$\lim_{n \to \infty} \mathbb{P}\left(d_B\left(\widehat{\mathrm{PH}}^{\varepsilon}_*(p), \mathrm{PH}_*(p)\right) \leq 5\varepsilon\right) = 1$$

Proof. Define a discrete version of the filtration (4) given by $\{D_{L_i+\varepsilon}\}_{i \in \mathbb{Z}}$, and denote the persistent homology of it by $\mathrm{PH}^{\varepsilon}_*(p)$. Since $\{D_{L_i+\varepsilon}\}_{i \in \mathbb{Z}}$ is a discrete approximation of the continuous filtration $\{D_L\}_{L \in \mathbb{R}}$, with step size 2ε, the maximum difference between $\mathrm{PH}_*(p)$ and $\mathrm{PH}^{\varepsilon}_*(p)$ would be the step size, and thus we have

$$d_B\left(\mathrm{PH}^{\varepsilon}_*(p), \mathrm{PH}_*(p)\right) \leq 2\varepsilon$$

Thus, it is enough to show that with a high probability, we have

$$d_B\left(\widehat{\mathrm{PH}}^{\varepsilon}_*(p), \mathrm{PH}^{\varepsilon}_*(p)\right) \leq 3\varepsilon$$

Let E be the event that we have the following sequence of inclusions

$$\begin{tikzcd}
D_{L_0+\varepsilon} \arrow[dr, hook] & & D_{L_1+\varepsilon} \arrow[dr, hook] & & D_{L_2+\varepsilon} \arrow[dr, hook] & & \cdots \\
& \widehat{D}_{L_0}(n,r) \arrow[ur, hook] & & \widehat{D}_{L_1}(n,r) \arrow[ur, hook] & & \widehat{D}_{L_2}(n,r) \arrow[ur, hook] &
\end{tikzcd}, \tag{6}$$

Since we assume every L is ε-regular, if r is small enough, by applying Lemma 1 N_ε times, we can show that if n is large enough

$$\mathbb{P}(E) \geq 1 - 3nN_\varepsilon e^{-C^\star_{\varepsilon/2} nr^d}$$

Using the notation in Chazal et al. [2009], (6) implies that $\{\widehat{D}_{L_i}(n,r)\}_{i\in\mathbb{Z}}$ and $\{D_{L_i+\varepsilon}\}_{i\in\mathbb{Z}}$ are *weakly ε-interleaving*. Using Theorem 4.3 in [chazal2009proximity] yields

$$d_B\left(\widehat{\mathrm{PH}}^\varepsilon_*(p), \mathrm{PH}^\varepsilon_*(p)\right) \leq 3\varepsilon \tag{7}$$

□

2 How to calculate confidence band

In this section, we tried to provide a simple way how to calculate confidence band of the persistent homology of filtration

$$\left\{\widetilde{D}_L(n,r)\right\}_{L\in\mathbb{R}} \tag{8}$$

where $\widetilde{D}_L(n,r)$ is the same as Eq.(2) except $\widetilde{D}_L(n,r) := \mathbb{R}^d$ for $\forall L \leq 0$.

Let $\widetilde{\mathrm{PH}}_*(p)$ be the corresponding persistent homology, and let $\mathrm{PH}_*(p_r)$ be the persistent homology of upper level set filtration of $p_r := \mathbb{E}(\widehat{p}_r)$. Our intermediate objective is find $\widehat{C}_\alpha$ such that

$$\mathbb{P}\left(d_B\left(\widetilde{\mathrm{PH}}_*(p), \mathrm{PH}_*(p_r)\right) > \widehat{C}_\alpha/\sqrt{n}\right) \to \alpha$$

We need further assumption on K.

Assumption 3. *The support of kernel function K is exactly equal to unit ball.*

Theorem 3. *If the assumption 1, 2, and 3 hold,*

$$d_B\left(\widetilde{\mathrm{PH}}_*(p), \mathrm{PH}_*(p_r)\right) \leq \|\widehat{p}_r - p_r\|_\infty + \widehat{c}_r \tag{9}$$

where $\widehat{c}_r := \max_i \sup_{\|x-X_i\|\leq r} |\widehat{p}_r(x) - \widehat{p}_r(X_i)|$

Proof. By the strong stability theorem in Chazal et al. [2009], it is enough to show that

$$D_L \subset \widetilde{D}_{L-\|\widehat{p}_r-p_r\|_\infty-\widehat{c}_r}, \text{ and } \widetilde{D}_L \subset D_{L-\|\widehat{p}_r-p_r\|_\infty-\widehat{c}_r}, \text{ for } \forall L\in\mathbb{R}$$

For the first part, if $L < \|\widehat{p}_r - p_r\|_\infty$, then $D_L \subset \widetilde{D}_{L-\|\widehat{p}_r-p_r\|_\infty-\widehat{c}_r} = \mathbb{R}^d$. If not,

$$\begin{aligned}
x \in D_L &\Leftrightarrow p_r(x) \geq L \\
&\Rightarrow \widehat{p}_r(x) \geq L - \|\widehat{p}_r - p_r\|_\infty (>0) \\
&\Rightarrow \exists X_i \text{ such that } \|x - X_i\| \leq r \quad \text{(because } \widehat{p}_r(x) \text{ is positive and } supp(K) \text{ is equal to unit ball.)} \\
&\Rightarrow \exists X_i \text{ such that } \|x - X_i\| \leq r \;\&\; \widehat{p}_r(X_i) \geq L - \|\widehat{p}_r - p_r\|_\infty - \widehat{c}_r \\
&\Leftrightarrow x \in \widetilde{D}_{L-\|\widehat{p}_r-p_r\|_\infty-\widehat{c}_r}
\end{aligned}$$

For the second part, if $L < \|\widehat{p}_r - p_r\|_\infty$, then $\widetilde{D}_L \subset D_{L-\|\widehat{p}_r-p_r\|_\infty - \widehat{c}_r} = \mathbb{R}^d$. If not,

$$\begin{aligned} x \in \widetilde{D}_L &\Leftrightarrow \exists X_i \text{ such that } \|x - X_i\| \leq r \ \& \ \widehat{p}_r(X_i) \geq L \\ &\Rightarrow \exists X_i \text{ such that } \|x - X_i\| \leq r \ \& \ \widehat{p}_r(x) \geq L - \widehat{c}_r \\ &\Rightarrow p_r(x) \geq L - \|\widehat{p}_r - p_r\|_\infty - \widehat{c}_r \\ &\Leftrightarrow x \in D_{L-\|\widehat{p}_r-p_r\|_\infty - \widehat{c}_r} \end{aligned}$$

□

Since we can use bootstrap to estimate $\widehat{C}_\alpha$ which satisfies

$$\mathbb{P}\left(\sqrt{n}\|p_r - \widehat{p}_r\|_\infty) > \widehat{C}_\alpha\right) \to \alpha \text{ as } n \text{ goes to infinity and } r \text{ is fixed}$$

we can construct an asymptotic confidence band for the persistent homology via the following theorem.

Theorem 4.

$$\mathbb{P}\left(d_B\left(\widetilde{\mathrm{PH}}_*(p), \mathrm{PH}_*(p_r)\right) > \widehat{C}_\alpha/\sqrt{n} + \widehat{c}_r\right) \to \alpha \tag{10}$$

Proof. By the theorem 3,

$$\begin{aligned} &\mathbb{P}\left(d_B\left(\widetilde{\mathrm{PH}}_*(p), \mathrm{PH}_*(p_r)\right) > \widehat{C}_\alpha/\sqrt{n} + \widehat{c}_r\right) \\ &\leq \mathbb{P}\left(\|p_r - \widehat{p}_r\|_\infty + \widehat{c}_r > \widehat{C}_\alpha/\sqrt{n} + \widehat{c}_r\right) \\ &= \mathbb{P}\left(\sqrt{n}\|p_r - \widehat{p}_r\|_\infty > \widehat{C}_\alpha\right) \to \alpha \end{aligned}$$

□

Corollary 5. *If we replace $\widetilde{D}_L(n,r)$ with $\widetilde{R}_L(n,r)$, the Rips complex with the radius r constructed on $\widehat{\mathscr{D}}_n^L := \{X_i : \widehat{p}_r(X_i) \geq L; 1 \leq i \leq n\}$, then*

$$\mathbb{P}\left(d_B\left(\widetilde{\mathrm{PH}}_*^R(p), \mathrm{PH}_*(p_r)\right) > \widehat{C}_\alpha/\sqrt{n} + \widehat{c}_{\sqrt{2}r}\right) \to \alpha \tag{11}$$

where $\widetilde{\mathrm{PH}}_*^R(p)$ is the persistence diagram of $\left\{\widetilde{R}_L(n,r)\right\}_{L\in\mathbb{R}}$

Proof. By the nerve theorem, we know that the Čech complex $\widetilde{C}_L(n,r) := C(\widehat{\mathscr{D}}_n^L, r)$ is homotopy equivalent to the homology of the union of balls $\widetilde{D}_L(n,r)$. From the relationship $\widetilde{C}_L(n,r) \subset \widetilde{R}_L(n,r) \subset \widetilde{C}_L(n,\sqrt{2}r)$, the corollary follows. □

Remark 6. *We can use different value h for bandwidth of kernel density estimator and r for radius of ball centered at each data point if $h \leq r$.*

2.1 Toy examples

By the corollary 5 , we can calculate persistent homology for the estimated density function filtration on the Rips complex with radius r. To choose appropriate r, we used the diagram of the usual Rips filtration. Since our filtration is based on Rips complex with radius r, our diagram can only capture the persistent homology classes whose birth time is smaller than r and death time is greater than r in the Rips diagram. Once the Rips diagram reveals some persistent homology classes whose lifetimes are longer than the others, we can choose appropriate r which allow us to check the significance of these classes.

We calculated Rips diagram, density diagram, and our diagram over 6 toy data sets. The number of data is equal to 300. 330, 500, or 530 depending on the number of circles and the existence of outliers. The alpha level is set to 0.2. The below figures illustrate performances of our methods for several situations.

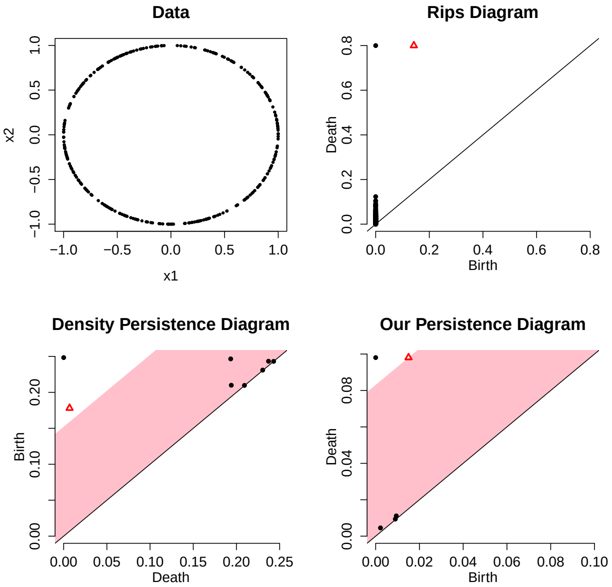

Figure 1: The data points are uniformly selected from a circle.

References

Omer Bobrowski, Sayan Mukherjee, and Jonathan E Taylor. Topological consistency via kernel estimation. *arXiv preprint arXiv:1407.5272*, 2014.

Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J Guibas, and Steve Y Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the twenty-fifth annual symposium on Computational geometry*, pages 237–246. ACM, 2009.

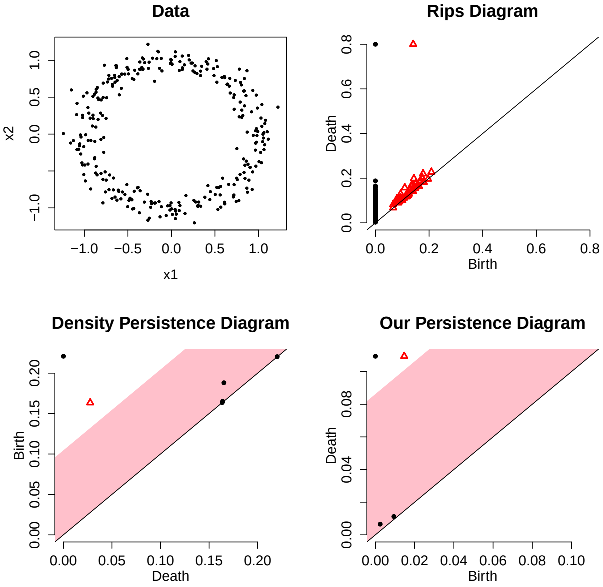

Figure 2: The data points are uniformly distributed over a circle with small gaussian noise.

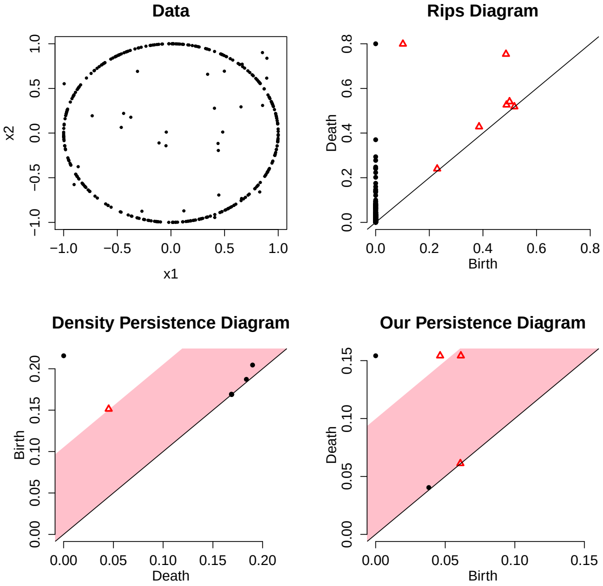

Figure 3: The data points are uniformly selected from a circle. Few outliers are added to the data set.

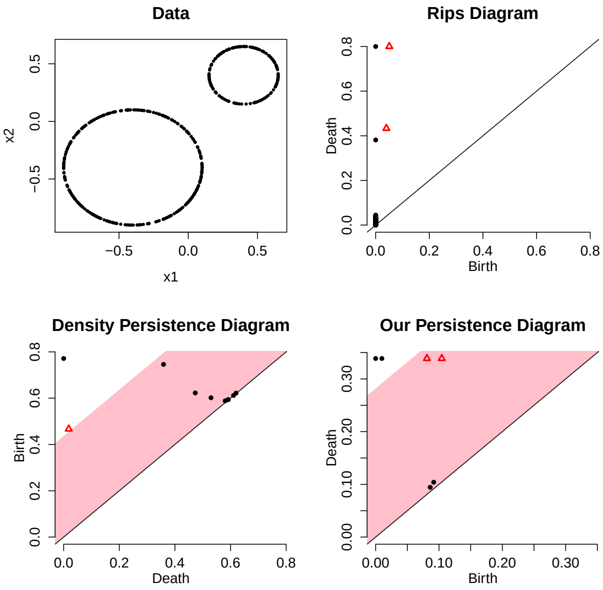

Figure 4: The data points are uniformly selected from two circles.

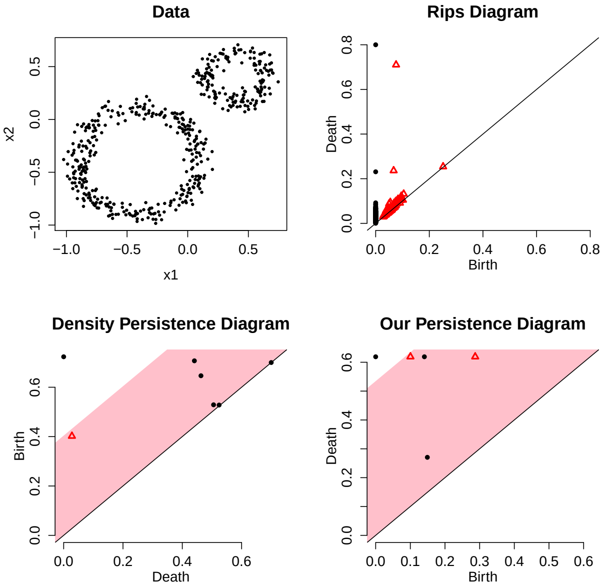

Figure 5: The data points are uniformly distributed over two circles with small gaussian noise.

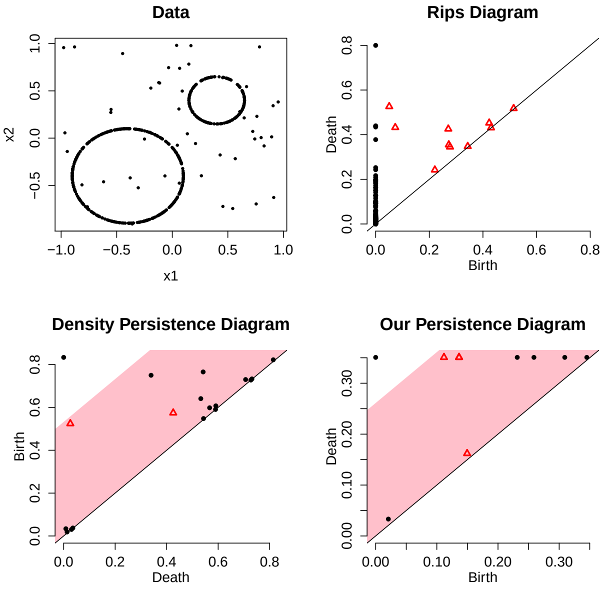

Figure 6: The data points are uniformly selected over two circles. Few outliers are added to the data set.