Topološka analiza podatkov

# Analiza persistenčnih diagramov

Sara Bizjak in Žan Hafner Petrovski

12. januar 2021

## Povzetek

Vztrajnostni diagram Ripsovih kompleksov je prikaz razlik v zaporedju homoloških grup, ki jih porodi izbrana filtracija. Zanima nas stabilnost vztrajnostnih diagramov na konkretnih primerih, empirično torej raziščemo vpliv perturbacije vhodnih točk. Vpliv merimo z razdaljo bottleneck, katere porazdelitev primerjamo z normalno porazdelitvijo. Ugotovimo, da razdalje niso normalno porazdeljene.

## 1 Uvod

V članku so predstavljena orodja in rezultati, ki smo jih dobili po testiranju implementacije filtracij Ripsovega kompleska in nadaljnji analizi stabilnosti vztrajnostnih diagramov na izbranih množicah točk. Natančneje, topološke značilnosti Ripsovih kompleksov v izbrani filtraciji zakodiramo v vztrajnostni diagram in preučujemo stabilnost tega kodiranja. Zanimalo nas bo torej, če se z majhno perturbacijo izbranih točk tudi vztrajnostni diagram ustrezno malo spremeni. Konstruirane vztrajnostne diagrame med sabo primerjamo tako, da izračunamo razdalje med originalnim diagramom in tistimi, ki jih dobimo s perturbiranjem začetnih točk. Na teh meritvah razdalj izvedemo osnovno statistiko in dobljeno porazdelitev primerjamo z normalno.

Prvi del članka opiše teoretično ozadje problema. Najprej se seznanimo s prvotnim problemom. Opišemo pojme, ki jih je potrebno razumeti pred samo implementacijo in si ogledamo orodja in postopke, s katerimi smo rešili dani problem. V drugem delu predstavimo rezultate izvedene analize.

## 2 Metode

### 2.1 Teoretično ozadje

To podpoglavje je namenjeno uvedbi pojmov in postopkov, ki so potrebni za izvedbo implementacije in pripadajoče analize.

**Definicija.** *Ripsov kompleks* za množico točk $S$ in radij $r$, ki ga označimo z $\mathrm{Rips}(S, r)$, je abstraktni simplicialni kompleks, za katerega velja:

- množica oglišč je enaka množici $S$,
- podmnožica $\sigma \subseteq S$ je simpleks natanko tedaj, ko je premer $\sigma$ največ $r$.

Zgornja definicija nam pove, da je $\sigma \subseteq S$ simpleks v Ripsovem kompleksu $\mathrm{Rips}(S, r)$ natanko takrat, kadar je presek dveh zaprtih krogel s premerom $r$ v katerih koli dveh točkah iz $\sigma$ neprazen.

**Definicija.** *Filtracija* za kompleks $K$ je definirana kot naraščajoče zaporedje kompleksov

$$\emptyset \neq K_0 \leq K_1 \leq \ldots \leq K_n = K.$$

Filtraciji ustreza tudi zaporedje $p$-te homološke grupe, kar lahko zapišemo

$$\emptyset \neq H_p(K_0) \leq H_p(K_1) \leq \ldots \leq H_p(K_n) = H_p(K).$$

Vemo, da vsako homološko grupo generirajo t. i. generatorji grup. Rečemo, da se generator $\gamma$ v $H_p(K_i)$ za $i = 0, \ldots, r$

- *rodi* ob času (radiju) $i$, če $\gamma \notin H_p(K_{i-1})$ in $\gamma \in H_p(K_i)$,
- *umre* ob času (radiju) $j$, če $\gamma \in H_p(K_{j-1})$ in ni neodvisen generator v $H_p(K_j)$.

**Definicija.** *Vztrajnostni diagram* je prikaz vseh točk $(i, j)$ v ravnini $\mathbb{R}^2$, ki jih dodelimo vsakemu generatorju $\gamma \in H_p(K_l)$ za nek $l$. Pri tem $i$ označuje rojstvo, $j$ pa smrt generatorjev.

Z besedami, na vztrajnostnem diagramu označimo stanje, ko se dva generatorja homoloških grup združita. Prikaz pomeni, da prejšnji generator ob tem času (oz. ob tem radiju) umre, novi združeni generator pa se rodi.

Da bi analizirali stabilnost vztrajnostnih diagramov, poskušamo odgovoriti na vprašanje, kako majhne spremembe začetnih podatkov vplivajo na diagram. Zanimalo nas bo torej, če se z majhno perturbacijo začetnih točk tudi vztrajnostni diagram ustrezno malo spremeni. Stabilnost lahko seveda poskusimo ugotoviti z opazovanjem vztrajnostnih diagramov, vendar si želimo formalnejšega in bolj objektivnega pristopa, zato vpeljemo še pojem razdalje bottleneck, ki nam poda razdaljo med dvema diagramoma.

**Definicija.** Razdalja bottleneck med vztrajnostnima diagramoma $X$ in $Y$ je definirana kot

$$W_\infty(X, Y) = \inf_{\varphi: X \to Y} \left( \sup_{x \in X} ||x - \varphi(x)||_\infty \right).$$

Da bi empirično pokazali stabilnost vztrajnostnih diagramov, moramo zadostiti naslednji trditvi.

**Trditev.** Naj bo $S = \{v_1, v_2, \ldots, v_n\}$ množica točk, ki ji priredimo $\epsilon$-perturbacijo $S' = \{v_1', v_2', \ldots, v_n'\}$, tako da velja

$$d(v_i, v_i') \leq \epsilon, \quad \forall i \in \{1, 2, \ldots, n\}.$$

Tedaj je razdalja bottleneck med vztrajnostnima diagramoma pripadajočih Ripsovih kompleksov največ $2\epsilon$:

$$d_b(D(S), D(S')) \leq 2\epsilon.$$

### 2.2 Opis problema in implementacija

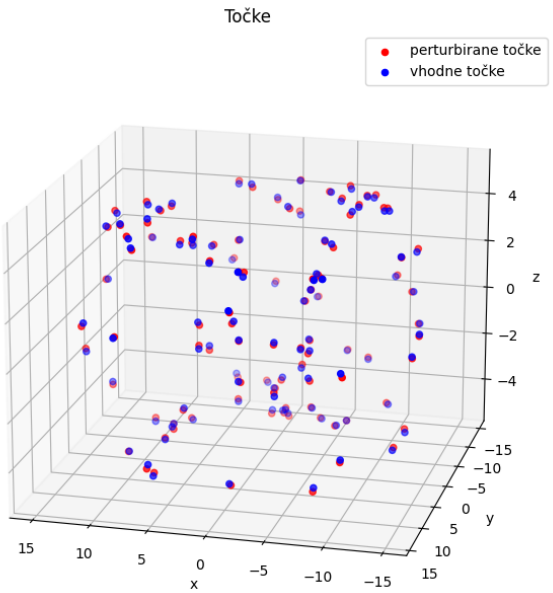
V problemu, ki ga rešujemo, je filtracija porojena z naraščanjem parametra $r$ za Ripsov kompeks, ki ji ustreza tudi zaporedje homoloških grup. Če si predstavljamo gradnjo Ripsa tako, da okoli točk konstruiramo zaprte krogle in opazujemo njihove preseke, potem lahko rečemo, da skozi to filtracijo večamo polmer oz. premer teh krogel. Zanima nas, kako se oblika Ripsovega kompleksa razvija skozi filtracijo, tj. kako se spreminja z večanjem parametra $r$. Vse te razlike v zaporedju homoloških grup, ki jih porodi filtracija, so zakodirane in predstavljene z vztrajnostnim diagramom. Cilj je predstaviti stabilnost vztrajnostnih diagramov Vietoris-Ripsovih kompleksov na izbranih množicah točk. Najprej je potrebno narediti particijo

$$0 = r_0 < r_1 < \ldots < r_{10} = R, \tag{1}$$

ki interval $[0, R]$, kjer je $R$ premer množice, razdeli na 10 enakih delov. Za vse izračunane $r_i$ zgradimo Ripsove komplekse, kar nam da filtracijo

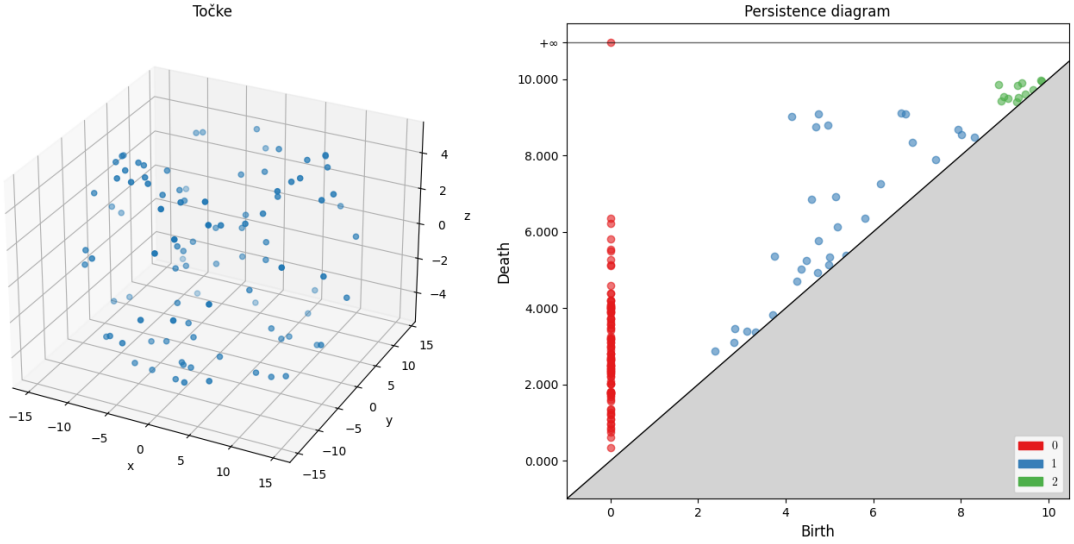
$$Rips(S, r_0) \leq Rips(S, r_1) \leq \ldots \leq Rips(S, r_{10}). \tag{2}$$

V podani filtraciji je kompleks $Rips(S, r_0)$ enak množici točk $S$, v Rips$(S, r_{10})$ pa je simpleks kar vsaka podmnožica množice $S$, saj je $r_{10}$ definiran kot premer $S$.
Opisan postopek še 100-krat ponovimo na točkah, ki jim dodamo šum velikosti $\epsilon < \frac{R}{100}$. Dodajanje šuma oz. perturbacija točk je prikazana na sliki 1.



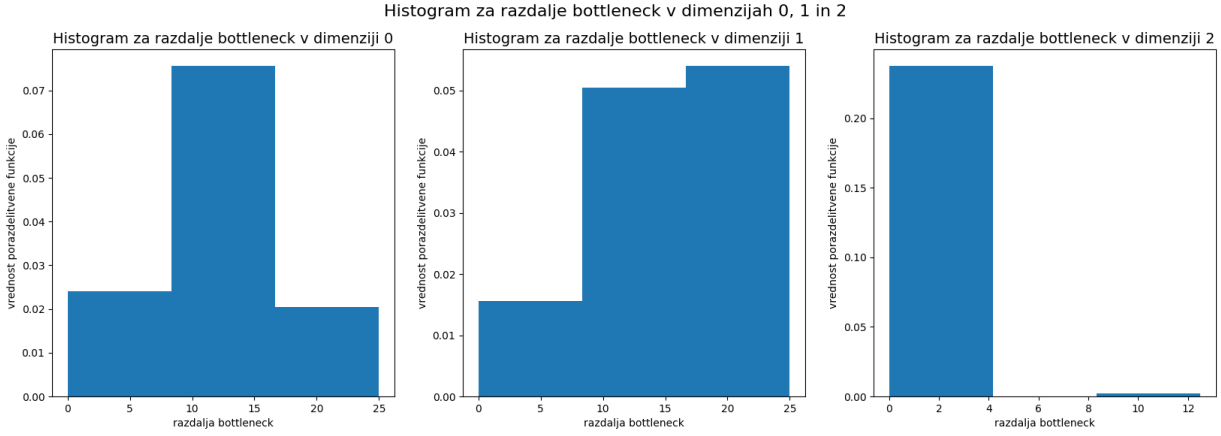

Slika 1: Prikaz ene izmed podanih množic točk in ene izmed njenih perturbacij.

Za vse filtracije Ripsovih kompleksov (zgrajenih na prvotnih in perturbiranih točkah) zgradimo vztrajnostne diagrame v dimenzijah 0, 1 in 2 ter jih primerjamo. Gradnja vztrajnostnega diagrama za izvorno množico točk je prikazana na sliki 2.




Slika 2: Prikaz izvornih točk in vztrajnostnega diagrama pripadajoče Ripsove filtracije.

Za vsak vztrajnostni diagram, zgrajen na perturbiranih točkah, izračunamo razdaljo bottleneck od prvotnega vztrajnostnega diagrama za vsako dimenzijo posebej. Na sliki 3 so prikazani histogrami vseh razdalj bottleneck za dimenzije 0, 1 in 2, ki jih dobimo z zgoraj opisanim postopkom.




Slika 3: Razdalje bottleneck vztrajnostnih diagramov v dimenziji 0, 1 in 2 za množico s 100 točkami pri filtraciji, kot je zapisano v 2.

Ker je na rezultatih, pridobljenih na podlagi particije na 10 delov, težko narediti zaključke, se odločimo raje za finejšo particijo. Naredimo najfinejšo možno filtracijo za Ripsove komplekse, kot jo ponuja knjižnica `gudhi`. Za še boljše in bolj relevantne rezultate – namesto 100 – zgradimo 1000 množic s perturbiranimi točkami. Na dobljenih rezultatih izvedemo nekaj osnovne

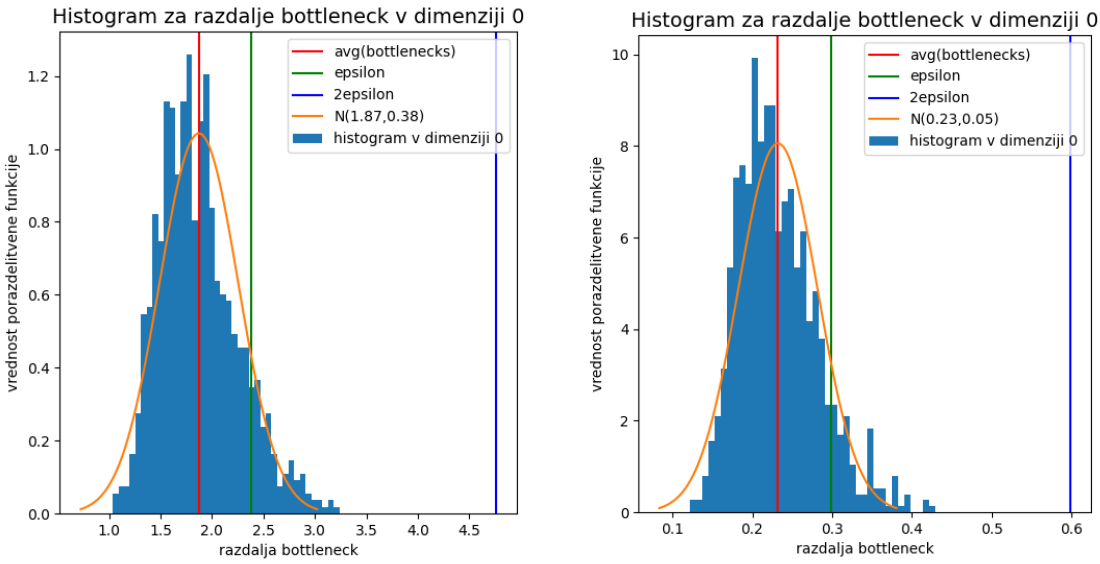
statistike in jih primerjamo z normalno porazdelitvijo.

Za implementacijo funkcij in analizo je bil uporabljen programski jezik `Python`. Za gradnjo Ripsovih kompleksov v filtraciji in vztrajnostnih diagramov je bila v največji meri uporabljena že vgrajena knjižnica `gudhi`, za test normalne porazdelitve pa smo uporabili *Spahiro-Wilkov test*, ki je vgrajen v knjižnici `scipy.stats`. Analiza in testi so dostopni v priloženi datoteki `analysis.py`, funkcije, ki smo jih uporabljali, pa v datoteki `helper_functions.py`.

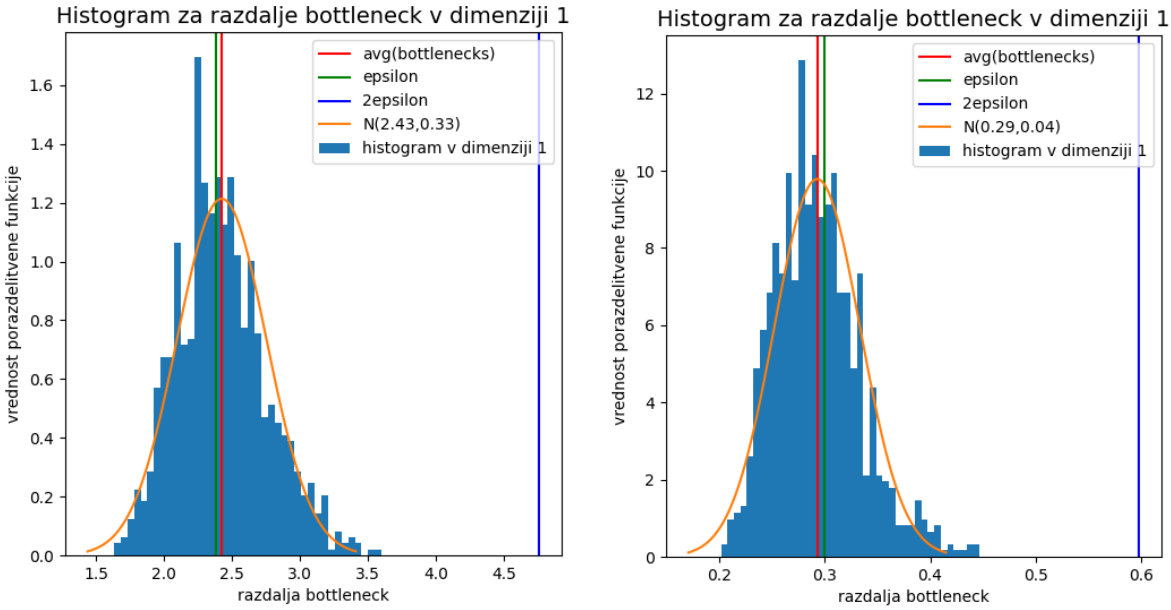
# 3 Rezultati

V tem poglavju so prikazani rezultati, dobljeni z najfinejšo možno filtracijo, in statistika, narejena na 1000 ponovitvah s perturbiranimi točkami iz datotek `persistence01_100.out` in `persistence02_100.out`.

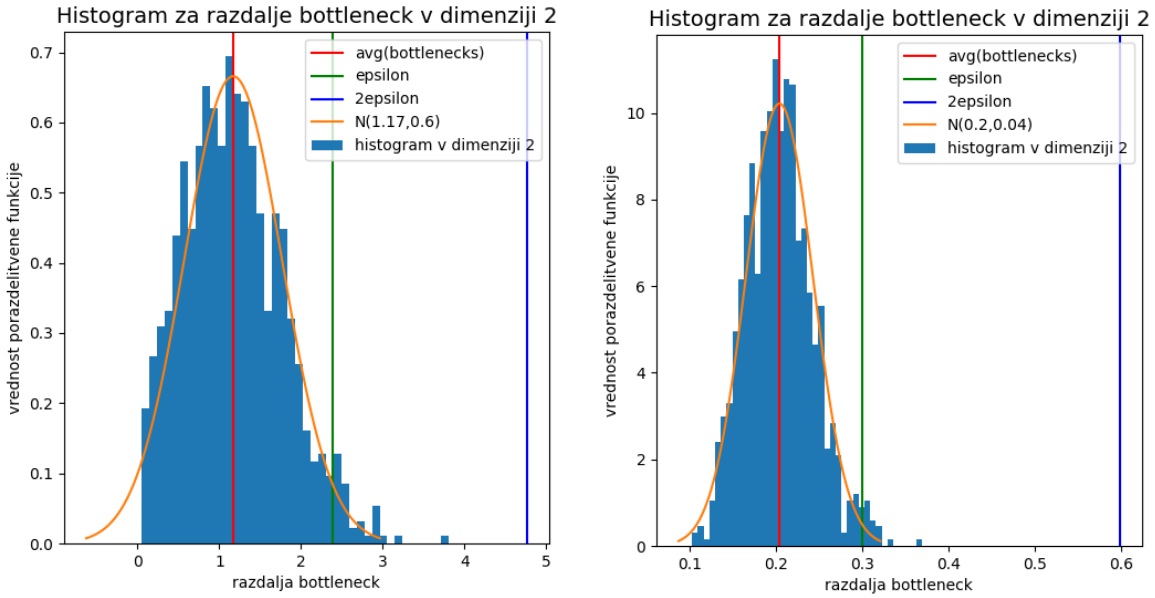
Na spodnjih slikah so prikazani histrogrami vseh dobljenih razdalj bottleneck med izvornim in vsemi vztrajnostnimi diagrami, dobljenimi na podlagi perturbiranih točk, za dimenzije 0, 1 in 2. Na vsakem histogramu je z rdečo barvo označeno povprečje razdalj bottleneck in z zeleno barvo $\epsilon$, ki predstavlja šum, s katerim smo zmotili izvorne točke. Z oranžno barvo je označena normalna porazdelitev za ustrezne parametre, torej za izračunano povprečje razdalj in varianco za posamezni primer.




Slika 4: Primerjava razdalje bottleneck vztrajnostnih diagramov v dimenziji 0 z normalno porazdelitvijo za dve različni izvorni množici s 100 točkami.




Slika 5: Primerjava razdalje bottleneck vztrajnostnih diagramov v dimenziji 1 z normalno porazdelitvijo za dve različni izvorni množici s 100 točkami.




Slika 6: Primerjava razdalje bottleneck vztrajnostnih diagramov v dimenziji 2 z normalno porazdelitvijo za dve različni izvorni množici s 100 točkami.

Če je pri 100 ponovitvah na oko še izgledalo, da bi bila porazdelitev morda lahko normalna, pa se na teh histogramih za 1000 ponovitev perturbiranja točk hitro vidi, da je histogram preveč nagnjen v levo, da bi bil lahko normalno porazdeljen. Da bi opazovanja tudi računsko potrdili, smo izvedli Shapiro-Wilkov test normalnosti, ki potrdi, da porazdelitev v nobenem primeru dimenzije ni normalna. V vseh dimenzijah je bila $p-$vrednost manjša od števila $10^{-5}$, kar

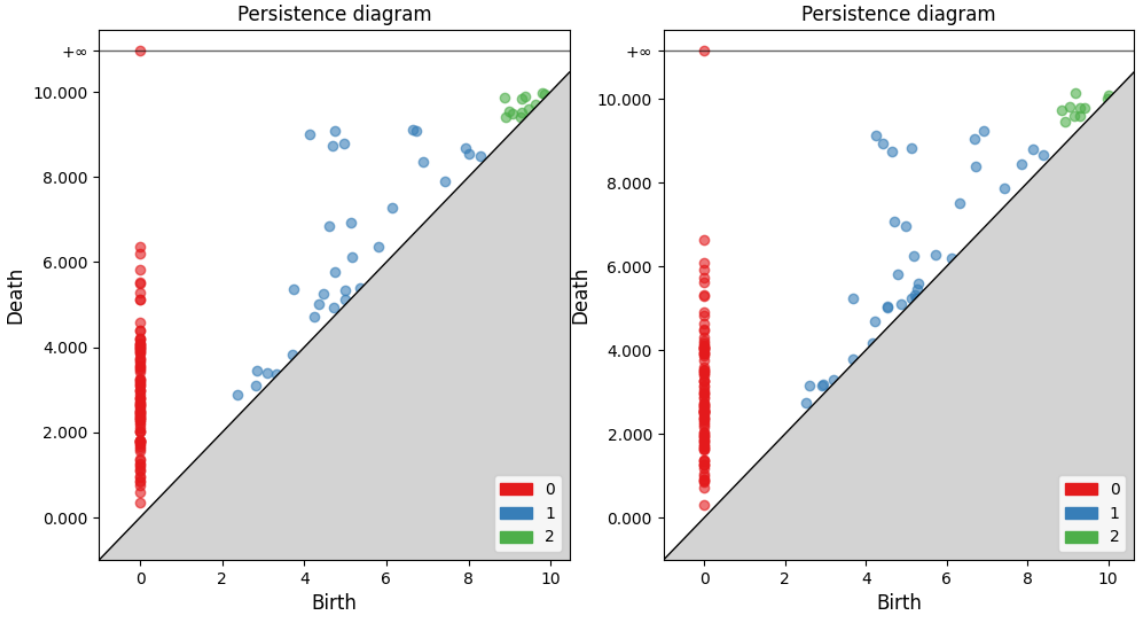
vidimo tudi v tabeli 1. Ker so $p-$vrednosti manjše od 0.05, zavrnemo hipotezo o normalni porazdelitvi.

| dimenzija \ datoteka | `persistence01_100` | `persistence02_100` |
|---|---|---|
| 0 | 3.645e-13 | 9.648e-15 |
| 1 | 9.993e-08 | 2.873e-12 |
| 2 | 7.845e-10 | 3.802e-06 |

Tabela 1: Izračunane $p-$vrednosti s Shapiro-Wilkovin testom.

Zanimivo je še opaziti, da se povprečna bottleneck razdalja giblje pod ali pa okoli vrednosti $\epsilon$.

Nadalje opazimo tudi, da so vse razdalje bottleneck manjše od $2\epsilon$, kar empirično potrdi že zapisana trditev o stabilnosti vztrajnostnih diagramov za Ripsove komplekse. Tudi subjektivna ocena na podlagi vztrajnostnega diagrama na izvornih in perturbiranih točkah sledi tej opazki, saj se vztrajnostna diagrama tudi na oko ustrezno malo razlikujeta.




Slika 7: Levo je prikazan vztrajnostni diagram za izvorne točke, desno pa za perturbirane.

## 4 Zaključek

Odločitev o finejši filtraciji in izračunu 1000 vrednosti razdalje bottleneck namesto 100 se je izkazala za dobro, saj smo histograme lahko tako bolj kredibilno primerjali z normalno poradelitvijo. Videli smo, da porazdelitve razdalj bottleneck niso normalne, empirično pa smo potrdili tudi trditev o stabilnosti vztrajnostnih diagramov Ripsovih kompleksov.

## Literatura

[1] Zapiski po predavanjih profesorja Ž. Virka in profesorice N. Mramor Kosta pri predmetu *Topološka analiza podatkov*, šolsko leto 2020/2021.