UNIVERSITAT DE BARCELONA

ADVANCED MATHEMATICS

MASTER'S FINAL PROJECT

Variacions de diagrames de persistència en classificació

Author: Roura i Cubí

Supervisor: Carles Casacuberta Vergés,

Aina Ferrà Marcús

Facultat de Matemàtiques i Informàtica

September 2, 2022

Índice general

Abstract

L'anàlisi de dades topològica ha estat àmpliament usada per resoldre problemes de classificació de núvols de punts; no obstant, com usar-la en la classificació de punts a $\mathbb{R}^d$ ha estat un problema poc estudiat. En aquets teball usarem eines recentment desenvolupades en l'estudi de diagrames de persistència aleatoris per abordar el problema de classificar punts a $\mathbb{R}^d$.

Topological Data analysis has been extensively used to solve point cloud classification problems; however, how to use TDA in classification of points in $\mathbb{R}^2$ has not been studied yet. In this work we use recently developed tools on the study of random persistence diagrams in order to deal with the classification of points in $\mathbb{R}^d$.

1.0 Introducció

L'anàlisi de dades topològic compren un seguit de tècniques que s'usen per estudiar la topologia d'un conjunt de punts donat.

Una d'aquestes eines és l'homologia persistent que estudia la topologia d'un conjunt de punts través del còmput de l'homologia del conjunt en qüestió a diferents escales. Donat per exemple un núvol de punts $\{x_1, ..., x_n\} \subseteq \mathbb{R}^d$ i un paràmetre temporal, es pot construir un complex simplicial $K(X)_r$ a través d'alguna construcció com per exemple el complexe de Vietoris-Rips. Si fem créixer el paràmetre $r > 0$, aconseguim una successió creixent de complexos $\{K(X)\}_r$ que s'anomenarà filtració. Quan en aquesta filtració apareix un simplex en temps r pot donar-se que la topologia de $K(X)_r$ canvii respecte la de paràmetres inferiors: es poden generar o destruir components connexes, forats... L'homologia persistent captura aquests canvis en les característiques topològiques de la filtració i les codifica en un diagrama de persistència $\mathcal{D} = \{(b_i, d_i, k)\}$ i.e. un conjunt de ternes, cada una d'elles representatn un forat $k-$dimensional que ha aparegut a la filtració en temps b_i i ha desaparegut en temps d_j.Els diagrames de persistència doncs es poden veure com un descriptor multi-escala de les característiques topològiques d'un conjunt.

L'estudi dels diagrames de persistència ha estat un subjecte de recerca que ha generat molta atenció recentment, i nombrosos resultats han estat donats sobre aquests ([10], [?], [?], [?]). Això combinat amb l'anterior aparició d'algoritmes eficients per computar-los, ha permès que aquests hagin estat usats de manera molt àmplia en problemes de classificació. Les aproximacions que s'han seguit, han estat o bé usant directament estadística sobre diagrames de persistència ([?], [?], [?]), o bé construir un aplicació del conjunt de diagrames de persistència a un espai hilbertià ([1], [2]). Aquest últim mètode permet usar la informació continguda en el diagrama de persistència en algoritmes de aprenentatge automàtic tradicionals com ara PCA, RF, SVM... Totes les aproximacions anteriors comparteixen el fet que s'usen per a classificar núvols de punts, i.e. predir la classe de $\{x_1, ..., x_n\} \subseteq \mathbb{R}^d$. No obstant, quan es tracta de

la classificació d'un punt, i.e. predir la classe de $x_1 \in \mathbb{R}^d$ l'estudi és molt més reduit. Alguns dels pocs treballs que aborden el problema són [8] i [5].

En aquest context [5] desenvolupa un classificador que es basa en la variació de l'homologia persistent d'un conjunt al afegir'hi un punt. Més en concret classifica un punt a la classe sobre la qual la seva addició fa variar menys el diagrama de persistència. Així assumeix que en general l'homologia d'una classe variarà menys en afegir punts de la mateixa classe que no pas en afegir punts aliens.

En el treball veurem que l'anterior hipòtesi no sempre es compleix (en particular sol fallar en considerar homologies de dimensió >1), i que tot i fallar, la variació homològica respecte l'addició de punts pot seguir sent característica s'una classe. L'objectiu doncs serà construir un nou classificador que superi aquestes limitacions, i que aprofiti la informació que aporten les variacions homològiques d'una classe, i que no captura la hipòtesi de [?]. Proposarem un mètode basat en construir una mostra de diagrames de persistència variacionals, i donat un punt construir el diagrama variacional corresponent i a partir d'aquest i de les mostres predir a quina classe pertany.

Per tal de provar el classificador l'usarem sobre una sèrie de problemes de classificació reals i artificials, veient com funciona en general, i respecte el classificador proposat en [?].

L'estructura del teball serà la següent. En la secció 1 presentarem la noció d'homologia persistent, i definirem els objectes necessaris per treballar en la resta de treball. En la secció 2 definirem la estimació via nuclis de la funció de densitat donada en [9], i de la qual ens valdrem per construir el classificador en la secció següent. En la secció 4 presentarem el classificador donat en REF, estudiant les seves possibles limitacions. Partint d'aquest esquema de classificació i de les limitacions trobades definirem un nou classificador. Finalment en la secció 4 presentarem el resultats provinents de la prova del classificador. En la secció 5 donarem les conclusions del teball.

2.0 Homologia Persistent

En aquesta secció introduirem els conceptes necessaris per arribar a la definició d'homologia persistent i a la de diagrama de persistència, que seran els objectes fonamentals sobre els quals versarà la resta del treball. Començarem per donar una breu definició de l'homologia simplicial; seguirem per definir els possibles complexos simplicials que podem associar a un conjunt de punts de $\mathbb{R}^d$ per acabar donant la definició d'homologia persistent i de diagrama de persistència. Acabarem amb la definició dels paisatges i siluetes de persistència, presentats en [1] i [2] respectivament, i que usarem més endavant per definir el classificador presentat en [5].

2.1.0 Homologia simplicial

Definició 2.1. *Direm que un conjunt de punts $\{x_0, ..., x_n\}$ en un espai euclidià $\mathbb{R}^d$ és geomètricament independent si per a qualsevol conjunt $t_i \in \mathcal{R}$ tal que $\sum_{i=0}^{n} t_i = 0$ es compleix que si*

$$\sum t_i x_i = 0$$

llavors $t_i = 0$ per a tot $i \in \{0, ..., n\}$.

Definició 2.2. *Un k-símplex és una col·lecció de $k+1$ punts geomètricament independents juntament amb el seu con convex:*

$$[x_0, ..., x_k] = \{\sum_{i=0}^{k} a_i x_i \text{ tal que } \sum_{i=0}^{k} a_i = 1\}.$$

En aquest cas direm que els vèrtexs $\{x_0, ..., x_k\}$ generen el símplex $k-$dimensional $[x_0, ..., x_k]$. Definirem les cares d'un k-simplex $[x_0, ..., x_k]$ com els $(k-1)-$símplexs generats per $\{x_0, ..., x_k\}$.

Definició 2.3. *Un complex simplicial S és una col·lecció de k-símplexs tal que :*

i) Si $K \in S$ llavors totes les cares de K també pertanyen a S.

ii) La intersecció de dos símplexs en S és buida o pertany a S.

Definició 2.4. *Donat un complex simplicial K, el grup de cadenes k-dimensional es defineix com el grup abelià lliure generat per les cares k-dimensionals i es denota $C_k(K)$:*

$$C_k(K) = \{ \sum n_\sigma \sigma \quad tals\ que \quad \sigma \in \mathcal{Z} \}$$

Definició 2.5. *L'aplicació vora k-èsima es defineix com l'homomorfisme de grups $\delta_k : C_k(K) \to C_{k-1}(K)$ que envia cada símplex a la suma alternada de les seves cares d'una dimensió menor:*

$$\delta_k([v_0, ..., v_k]) = \sum_{n=0}^{k} (-1)^n (v_0, ..., v_{n-1}, v_{n+1}, ..., v_k).$$

Es pot comprovar que la composició de dos morfismes vora dona lloc al morfisme trivial, és a dir, $\delta_k \circ \delta_{k-1} = 0$ i $\text{im}(\delta_k) \subseteq \ker(\delta_{k-1})$. Així podem construir un complex de cadenes

$$... \xrightarrow{\delta_{n+1}} C_n(K) \xrightarrow{\delta_n} ... \xrightarrow{\delta_3} C_2(K) \xrightarrow{\delta_2} C_1(K) \xrightarrow{0} C_0(K) \to \{\delta_1\},$$

en el qual $\text{im}(\delta_{k+1}) \subseteq \ker(\delta_k)$. Sobre un complex així és natural definir l'homologia.

Definició 2.6. *El k-èsim grup d'homologia simplicial del complex K es defineix com $H_k(K) = ker(\delta_k)/im(\delta_{k+1})$.*

Els generadors del k-èsim grup d'homologia es corresponen amb característiques topològiques k-dimensionals del complex K. Per exemple, els generadors de $H_0(K)$ correspondran a components connexes, els de $H_1(K)$ a forats, etc.

Un cop definida l'homologia simplicial, serà del nostre interès estendre la seva definició a núvols de punts $\{x_1, ..., x_n\} \subseteq \mathbb{R}^d$. La idea és construir un cert complex simplicial sobre el núvol donat i computar l'homologia d'aquest.

2.2.0 De núvols de punts a complexos

Sigui $X = \{x_1, ..., x_n\} \subseteq \mathbb{R}^d$ un núvol de punts en un cert espai euclidià. La manera més natural de construir un complex sobre X serà considerar els seus punts com els vèrtexs d'un símplex (cares 0-dimensionals) i connectar cada parell de punts x_i, x_j

complint una certa condició de proximitat a través d'una aresta (cara 1-dimensional), per exemple, $d(x_i, x_j) < \epsilon$ per a un cert ϵ donat. Aquesta estructura formada per vèrtexs i arestes servirà per definir les cares de dimensió >1 del complex. Les diferents regles que es poden donar per formar aquestes cares donaran lloc a diferents definicions de complexos associats a núvols de punts. Les més habituals són les de complex de Cech i complex de Vietoris-Rips.

Definició 2.7. *Donada un col·lecció de punts en un espai euclidià $X = \{x_1, ..., x_n\} \subseteq \mathbb{R}^d$, el complex simplicial de Cech $\boldsymbol{C}_\epsilon$ és el complex simplicial els k-símplexs del qual estan formats per col·leccions no ordenades de punts $\{x_i\}_{i=0}^m$ amb $\cap_{i=0}^m \boldsymbol{B}(x_i, \epsilon/2) \neq \emptyset$.*

Definició 2.8. *Donada un col·lecció de punts en un espai euclidià $X = \{x_1, ..., x_n\} \subseteq \mathbb{R}^d$, el complex simplicial de Vietoris-Rips $\boldsymbol{R}_\epsilon$ és el complex simplicial els k-símplexs del qual estan formats per col·leccions no ordenades de punts $\{x_i\}_{i=0}^m$ tals que $\boldsymbol{B}(x_i, \epsilon/2) \cap \boldsymbol{B}(x_j, \epsilon/2) \neq \emptyset$ per a tot $j \neq i$.*

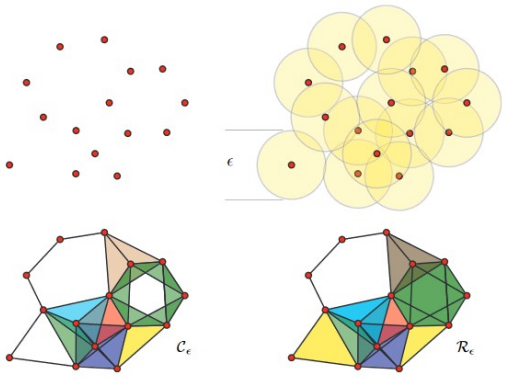


Figura 2.1: Construcció dels complexos de Cech (imatge inferior esquerra) i Vietoris-Rips (imatge inferior dreta) per un cert paràmetre ϵ sobre un núvol de punts donat (imatge superior esquerra). Figura extreta de [6].

De la definició se segueix que $\boldsymbol{C}_\epsilon \subseteq \boldsymbol{R}_\epsilon \subseteq \boldsymbol{C}_{2\epsilon}$ (vegi's [3] per a una demostració). Podem observar la relació en la figura 2.2. En general el complex de Cech tindrà major dimensió, i de fet, el teorema de Cech demostra que $\boldsymbol{C}_\epsilon$ té el mateix tipus d'homotopia que la unió de les boles tancades centrades en x_i i de radi $\epsilon/2$. Computacionalment, el complex de Cech serà més eficient que el de Rips. Mentre que $\boldsymbol{R}_\epsilon$ està completament determinat pel seu esquelet 1-dimensional (1-símplexs) i per tant pot ser guardat com un graf, per determinar $\boldsymbol{C}_\epsilon$ necessitarem la relació sencera de la vora. Degut a aquest motiu, en general, pels còmputs d'homologia persistent usem el complex de Rips.

2.3.0 Homologia persistent

Així doncs, donat un núvol de punts, podem construir el complex de Rips o de Cech amb paràmetre ϵ associat, i calcular l'homologia simplicial d'aquest. No obstant una qüestió natural apareix en aquest context: quin és el paràmetre ϵ adequat per capturar les característiques topològiques de la forma del nostre núvol? És clar que si ϵ és prou petit el complex serà un conjunt discret de punts, mente que si és prou gran serà un sol $(n-1)$-símplex; en els dos casos l'homologia del complex serà trivial. Vegi's la figura 2.3 que il·lustra la qüestió. Escollir un paràmetre ϵ òptim no és la solució, i aquesta és proposà en [4] via la construció de l'homologia persistent, que capturarà la persistència o fragilitat de les característiques homològiques que apareixen en els grups d'homologia per a successius paràmetres ϵ.

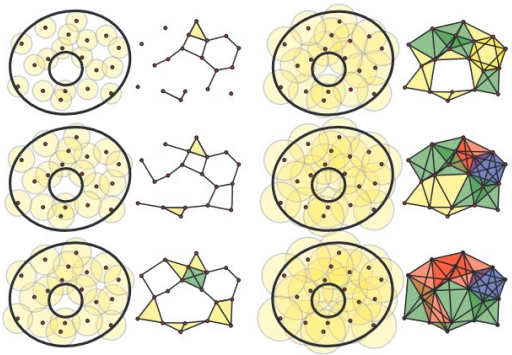


Figura 2.2: Filtració de Rips per a diferents valors ϵ sobre un núvol de punts que representa una corona circular. Extret de [6].

Sigui $\{R_i\}_{i=1}^N$ una filtració de complexos de Rips. Per a cada i tenim el complex simplicial de cadenes $C_*(R_i)$ de R_i (per simplificar notació l'anomenarem C_*^i), i per tant una seqüencia de complexos de cadenes $C = \{C_*^i\}_{i=1}^N$. Donat que $R_i \subseteq R_{i+1}$ tindrem que $C_*^i \subseteq C_*^{i+1}$ i naturalment tindrem una inclusió de complexos $x_{i,j} : C_*^i \to C_*^j$,. Les aplicacions induïdes en homologia per $x_{i,j}$,

$$f_{i,j} : H_*(C_*^i) \to H_*(C_*^j)$$

són les que finalment ens donaran informació sobre la persistència de les característiques topològiques. Direm que una característica p-homològica α neix en R_i, si $\alpha \in H_p(C_*^i)$ i $\alpha \notin \text{im}(f_{i-1,i})$. Si α neix en R_i, direm que desapareix en R_j si $f_{i,j-1}(\alpha) \notin$

$\text{im}(f_{i-1,j-1})$, però $f_{i,j}(\alpha) \subseteq \text{im}(f_{i-1,j})$. Així les imatges de les aplicacions $f_{i,j}$ consistiran en les característiques topològiques que han nascut en o abans de R_i i persisteixen en R_j.

Definició 2.9. *Per a $i < j$, la (i,j)-homologia persistent de C es defineix com la imatge del morfisme induït en homologia per $x_{i,j}$, $f_{i,j}$) : $H_*(C_*^i) \to H_*(C_*^j)$.*

Un cop definida l'homologia, és convenient representar-la. Una possible manera és codificar la persistència d'una característica topològica α que neix en b i mor en d com un punt en el pla amb coordenades (b,d). Si considerem la representació anterior en el pla de totes les característiques homològiques amb persistència major que 0, obtindrem el que definim com a diagrama de persistència.

Finalment definim algunes representacions alternatives de l'homologia persistent de les quals ens servirem en seccions posteriors. En concret, volem definir la silueta de persistència, donada en [2].

Definició 2.10. *Donat un diagrama de persistència $\mathcal{D}$, considerarem el diagrama girat 45º, és a dir, el nou diagrama on $p = (b,d) \in \mathcal{D}$ té coordenades $(\frac{b+d}{2}, \frac{d-b}{2})$. Sobre cada punt p d'aquest nou diagrama considerem la funció tenda*

$$\Lambda_p(t) = \begin{cases} t-b & si \quad t \in [b, \frac{b+d}{2}] \\ d-t & si \quad t \in (\frac{b+d}{2}, d] \\ 0 & altrament\,. \end{cases} \tag{2.1}$$

El paisatge de persistència es defineix com la col·lecció de funcions

$$\lambda_k(t) = \text{kmax}_{p \in \mathcal{D}}(\Lambda_p(t))$$

on kmax denota el k-èsim valor màxim del conjunt.

Definició 2.11. *Donat un diagrama de persistència $\mathcal{D}$ amb m punts, la silueta de persistència es defineix com la mitjana ponderada de les funcions tenda $\Lambda_p(t)$ donades en la definició anterior:*

$$\phi(t) = \frac{\sum_{j=1}^m w_j \lambda_j(t)}{\sum_{j=1}^m} W_j.$$

3.0 Diagrames de persistència com a elements aleatoris

3.1.0 Diagrames de persistència aleatoris

Donada la complexitat derivada de la no linealitat implícita en el procés de creació d'un diagrama de persistència sobre un núvol de punts, és convenient veure els diagrames de persistència com a elements aleatoris que depenen de la estructura global del conjunt subjacent. Notem però, que un diagrama de persistència té dues característiques que impedeixen definir-lo com un vector aleatori. Primerament, petits canvis en l'estructura del núvol subjacent poden conduir a canvis de cardinalitat en el diagrama de persistència i per tant, el cardinal d'un diagrama de persistència aleatori ha de ser varibable. D'altra banda, les característiques que defineixen un diagrama (parelles (b, d)) no tenen un ordre implícit, i per tant un diagrama aleatori no pot tenir en compte l'ordre. Resumint, tenim que un diagrama de persistència és un conjunt i definir-ne l'element aleatori associat ho ha de tenir en compte. Així caldrà definir un diagrama aleatori com un conjunt aleatori (de fet, un multiconjunt aleatori, donat que permetem la multiplicitat de característiques topològiques).

Definició 3.1. *Per a un núvol de punts en $\mathbb{R}^d$ definim l'espai de les seves característiques topològiques com*

$$\mathcal{W}_{0:d} = W \times \{0, ..., d-1\},$$

on $W = \{(b, d) \in \mathbb{R}^2 : d > b \geq 0)\}$.

Noti's que veiem $\mathcal{W}_{0:d}$ com d còpies disconnexes de W, on W té la topologia i la mètrica euclidiana. Els elements de $\mathcal{W}_{0:d}$ seran ternes (b, d, k) on (b, d) indicarà la persistència de la característica (naixement i desaparició), i k la seva dimensió homològica.

Així doncs considerarem un diagrama de persistència aleatori com un multicon-

junt aleatori format per característiques topològiques, $D = \{\xi_i\} \subset \mathcal{W}_{0:d}$. Notem que sempre que el diagrama estigui construït sobre un núvol de punts a $\mathbb{R}^d$ amb cardinal acotat, el seu diagrama de persistència tindrà un nombre acotat de característiques i de dimensions homològiques, llavors el seu cardinal també estarà acotat per un cert nombre $M \in \mathbb{N}$.

Definició 3.2. *Definim l'espai de diagrames de persistència amb cardinal acotat per $M \in \mathbb{N}$ com*

$$\mathcal{C}_M(\mathcal{W}_{0:d}) = \{D \text{ multiconjunt en } \mathcal{W}_{0:d} \,:\, |D| \leq M\}.$$

A aquest espai li associem de manera natural funcions $h_N : \mathcal{C}_M(\mathcal{W}_{0:d}) \to \mathcal{C}_N(\mathcal{W}_{0:d})$ que envien cada diagrama al subespai de diagrames del cardinal corresponent.

Definició 3.3. *Definim l'espai de diagrames de persistència amb cardinal $|D| = N$ com*

$$\mathcal{C}_N(\mathcal{W}_{0:d}) = \{D \text{ multiconjunt en } \mathcal{W}_{0:d} \,:\, |D| = N\}.$$

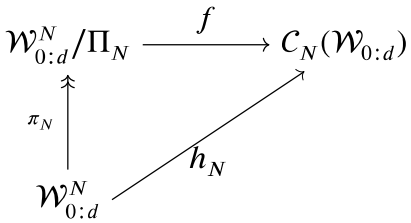
Considerem l'aplicació $h_N : \mathcal{W}_{0:d}^N \to \mathcal{C}_N(\mathcal{W}_{0:d})$ definida per

$$h_N(\xi_1, ..., \xi_N) = \{\xi_1, ..., \xi_N\}.$$

Notem que les aplicacions h_N defineixen classes d'equivalència en $\mathcal{W}_{0:d}^N$ segons l'acció de permutacions de Π_N. Més concretament, per a $Z = (\xi_1, ..., \xi_N) \in \mathcal{W}_{0:d}^N$,

$$[Z] = [(\xi_1, ..., \xi_N)]_{h_N} = \{(\xi_{\pi(1)}, ..., \xi_{\pi(N)}) \,:\, \pi \in \Pi_N\}.$$

Aquestes classes defineixen un espai quocient $\mathcal{W}_{0:d}^N/\Pi_N = \{[Z]_{h_N} \,:\, Z \in \mathcal{W}_{0:d}^N\}$, al qual podem associar la topologia quocient. Definirem la topologia en $\mathcal{C}_N(\mathcal{W}_{0:d})$ com aquella que fa que f sigui un homeomorfisme en el següent diagrama commutatiu:



Definirem la topologia en $\mathcal{C}_{\leq M}(\mathcal{W}_{0:d})$ com la topologia induïda per les aplicacions $h_N : \mathcal{C}_{\leq M}(\mathcal{W}_{0:d}) \to \mathcal{C}_N(\mathcal{W}_{0:d})$.

Un cop definida una topologia en l'espai de diagrames de persistència, podem considerar la σ-àlgebra definida pels seus borelians, i sobre aquesta definir una mesura de probabilitat.

Definició 3.4. *Un diagrama de persistència aleatori serà un multiconjunt aleatori distribuït segons una probabilitat $\mathbb{P}$ en $\mathcal{C}_{\leq M}(\mathcal{W}_{0:d})$.*

A continuació definirem els elements necessaris per arribar a la definició de la funció de densitat d'un diagrama de persistència aleatori donat. Com que $\mathcal{C}_{\leq N}(\mathcal{W}_{0:d}) \cong \mathcal{W}_{0:d}/\Pi_N$ ens restingirem a definir la densitat en els espais euclidians $\mathcal{W}_{0:d}^N$.

Definició 3.5. *Sigui D un diagrama de persistència aleatori i A un borelià de $\mathcal{W}_{0:d}$. Definim la funció de versemblança β_D com*

$$\begin{aligned} \beta_D : \quad & \mathcal{B}(\mathcal{W}_{0:d}) \to \mathbb{R} \\ & A \mapsto \mathbb{P}(D \subset A). \end{aligned}$$

Observació 3.1. Notem que $O_A = \{D \in \mathcal{C}_{\leq M}(\mathcal{W}_{0:d}) : D \subseteq A\} \in \mathcal{B}(\mathcal{C}_{\leq M}(\mathcal{W}_{0:d}))$. Efectivament O_A serà el quocient de $\cup_{N=0}^{M} A^N \subseteq \cup_{N=0}^{M} \mathcal{W}_{0:d}$ per h_N. Clarament A^N i la unió finita $\cup A^N$ seran borelians en $\mathcal{W}_{0:d}$), i donat que h_N indueix un homeomorfisme (és a dir, f és un homeomorfisme), tindrem que O_A és un borelià de $\mathcal{C}_{\leq M}(\mathcal{W}_{0:d})$. Així doncs la funció de versemblança està ben definida.

La funció de versemblança en el cas de diagrames aleatoris serà similar a la funció de distribució per a vectors aleatoris. Derivant aquesta a través del que definirem com l'equivalent per a conjunts de les derivades de Radon-Nykodým obtindrem l'equivalent a la funció de densitat.

Definició 3.6. *Sigui $\phi : \mathcal{B}(\mathcal{C}_{\leq M}(\mathcal{W}_{0:d})) \to \mathbb{R}$ una aplicació dels borelians de l'espai dels diagrames de persistència acotats per M a ls reals. Sigui $\xi \in \mathcal{W}_{0:d}$. Definim la derivada de ϕ respecte a ξ (avaluada a $\emptyset$) com*

$$\frac{\partial \phi}{\partial \xi}(\emptyset) = lim_{n\to\infty} \frac{\phi(B(\xi, 1/n))}{\lambda(B(\xi, 1/n))}.$$

Donat un multiconjunt Z tal que $Z \subseteq \mathcal{W}_{0:d}$, $Z = \{\xi_1, ..., \xi_N\}$, definirem la derivada de ϕ respecte al multiconjunt Z, (avaluada a $\emptyset$) com:

$$\frac{\partial \phi}{\partial Z}(\emptyset) = \frac{\partial^N \phi}{\partial \xi_1 ... \partial \xi_N} = [\frac{\partial}{\partial \xi_1} ... \frac{\partial}{\partial \xi_N} \phi](\emptyset).$$

En les anteriors definicions, $B(\xi, 1/n)$ són boles euclidianes centrades en ξ i de radi $1/n$, i λ és la mesura de Lebesgue en $\mathcal{W}_{0:d-1}$.

Observació 3.2. L'anterior definició és el cas particular de la derivada respecte a un conjunt d'una funció conjuntística, evaluada en el buit. Aquesta definició és suficient per definir la funció de densitat en un conjunt aleatori finit, i per tant per al nostre cas. Per veure la definició general, vegi's [7] (Secció 4.2.4, Definició 15).

Tal i com passa en les derivades habituals, tindrem un operació complementària d'integració.

Definició 3.7. *Sigui $f : C_{\leq M}(\mathcal{W}_{0:d-1}) \to \mathbb{R}$ una aplicació. Considerem un borelià de $\mathcal{W}_{0:d-1}$ A i un borelià de $C_{\leq M}(\mathcal{W}_{0:d-1})$ O. Les integral conjuntístiques de f sobre A i sobre O es defineixen respectivament com:*

$$\int_A f(Z)\partial Z = \sum_{N=0}^{M} \frac{1}{N!} \int_{A^N} f(h_N(\xi_1, ..., \xi_N)) \quad d\xi_1 ... d\xi_N$$

$$\int_O f(Z)\partial Z = \sum_{N=0}^{M} \frac{1}{N!} \int_{h_N^{-1}(O)} f(h_N(\xi_1, ..., \xi_N)) \quad d\xi_1 ... d\xi_N$$

on $Z = \{\xi_1, ..., \xi_N\} \subseteq \mathcal{W}_{0:d-1}$ és un diagrama de persistència.

Observació 3.3. Les derivades conjuntístiques evaluades al buit equivalen a les derivades de Radon-Nikodým amb l'ordre lligat al cardinal. Així doncs, el corresponent procés d'integració serà l'habitual, però sumant per a cada possible cardinal. La divisió per $N!$ que apareix en cada sumand contrarresta el fet que a l' integrar sobre $\mathcal{W}_{0:d}$, on els element són vectors, i no en $\mathcal{W}_{0:d}/\Pi_N$, obtenim $N!$ factors repetits.

Igual que en el cas habitual, es pot comprovar que integració i derivació són operacions inverses ([7, 4.3, Proposició 18]), i en particular tenim

$$\beta_D(A) = \int_A \frac{\partial^N \beta_D}{\partial\xi_1 ... \partial\xi_N}(\emptyset).$$

Definició 3.8. *Sigui D un diagrama de persistència aleatori. La seva funció de densitat global és la funció $f_D : \cup_{N\in\mathbb{N}} \mathcal{W}_{0:d-1}^N \to \mathbb{R}$ definida com*

$$\sum_{\pi\in\Pi_N} f_D(\xi_{\pi(1)}, ..., \xi_{\pi(N)}) = \frac{\partial^N \beta_D}{\partial\xi_1 ... \partial\xi_N}(\emptyset) \tag{3.1}$$

i descrita per les seves restriccions $f_N = f_D|_{\mathcal{W}_{0:d-1}^N} : \mathcal{W}_{0:d-1}^N \to \mathbb{R}$.

La següent proposició caracteritza la funció de veremblança d'un diagrama aleatori, i serà fonamental per al còmput dels diferents constituents de la seva funció de densitat.

Proposició 3.1. Sigui D un diagrama de persistència aleatori amb cardinal acotat per M i sigui β_D la funció de versemblança de D. Llavors β_D s'expressa com

$$\beta_D(S) = a_0 + \sum_{m=1}^{M} a_m q_m(S),$$

on $a_m = \mathbb{P}(|D| = m)$ i $q_m(S) = \mathbb{P}(D \subseteq S \,||D| = m)$.

Observació 3.4. Si el diagrama D està acotat per M i $n > M$, de l'anterior fórmula es desprèn que $f_N = f_D|\mathcal{W}_{0:d} = 0$.

Observació 3.5. En (3.1) la funció de densitat global no està definida en un sol espai euclidià, sino que ve definida per la col·lecció de funcions de densitat locals

$$f_N(Z) = f_D|_{\mathcal{W}_{0:d}^N}(Z).$$

Per l'anterior proposició, tindrem que $f_N(Z) = \mathbb{P}(|Z| = N) = f_D(Z||Z| = N)$, i per tant $\int_{\mathcal{W}_{0:d}^N} f_N(Z) \quad dZ = \mathbb{P}(|Z| = N)$.

3.2.0 Estimació no-paramètrica

Un cop difinida la funció de densitat per a un diagrama aleatori, serà del nostre interès obtenir una estimació d'aquesta a través de l'observació d'una mostra. En aquesta secció presentem l'estimació no paramètrica de la densitat d'un diagrama aleatori donat descrita en [9].

Sigui D un diagrama de persistència aleatori amb densitat f_D, i $\{\mathcal{D}_1, ..., \mathcal{D}_n\}$ una mostra d'aquest. Una estimació no-paramètrica via nuclis de f_D vindrà donada per

$$\hat{f}(\mathcal{D}) = \sum_{i=0}^{n} K_\sigma(\mathcal{D}, \mathcal{D}_i)$$

on K és un nucli (una funció no negativa i centrada en $\mathcal{D}_i$) i σ és l'ample de banda de la funció (un paràmetre que suavitza K).

Així doncs el primer que cal és definir un nucli centrat en un diagrama $\mathcal{D} = \{\xi_1, ..., \xi_M\}$ amb ample de banda σ. La construcció es basarà en el lema següent.

Definició 3.9. *Un diagrama de persistència D és un diagrama de persistència unitari si $|D| \leq 1$.*

NOTACIÓ: Distingirem els diagrames de persistència unitaris de la resta indexant-los amb un superíndex.

Lema 3.1. Sigui $\{D^j\}_{j=1}^M$ un multiconjunt de diagrames de persistència unitaris independents. Suposem que per a cada D^j tenim definits $q^{(j)} = \mathbb{P}(D^j \neq \emptyset)$ i $p^{(j)}(\xi)$ és la funció de densitat associada a la probabilitat condicionada $\mathbb{P}(D^j = \xi || D^j| = 1)$. La funció de densitat global associada a $D = \cup_{j=1}^M D^j$ ve donada per

$$f_d(\xi_1, ..., \xi_n) = \sum_{\gamma \in I(N,M)} Q^*(\gamma) \prod_{k=1}^N p^{(\gamma(k))}(\xi_k) \text{ per a } N \in \{0, ..., M\}, \tag{3.2}$$

on

$$Q^*(\gamma) = \frac{\prod_{j=1}^M (1 - q^{(j)})}{\prod_{k=1}^N (1 - q^{(\gamma(k))}} \prod_{k=1}^N q^{(\gamma(k))},$$

i $I(N, M)$ consisteix en el conjunt de les funcions creixents injectives entre $\{0, ..., N\}$ i $\{0, ..., M\}$.

Demostració. Per hipòtesi els diagrames unitaris D^j que constitueixen D són independents. Per tant, la funció de versemblança de D factoritza com

$$\beta_D(S) = \mathbb{P}(D \subset S) = \prod \mathbb{P}(D^j \subseteq S) = \prod_{j=1}^M \beta_{D^j}(S).$$

A partir d'aquesta, usant la regla de la cadena i tenint en compte que les derivades d'ordre major que 2 s'anul·len donat que D^j són diagrames unitaris (Proposició 3.1), tenim que

$$\frac{\partial^N \beta_D}{\partial \xi_1 ... \partial \xi_N}(\emptyset) = \sum_{1 \leq j_1 \neq ... \neq j_N \leq M} \frac{\beta_{D^1}(\emptyset) ... \beta_{D^M}(\emptyset)}{\beta_{D^{j_1}}(\emptyset) ... \beta_{D^{j_N}}(\emptyset)} \frac{\partial \beta_{D^{j_1}}}{\partial \xi_1}(\emptyset) ... \frac{\partial \beta_{D^{j_N}}}{\partial \xi_N}(\emptyset).$$

Com que $\beta_{D^j}(\emptyset) = \mathbb{P}(D \subset \emptyset) = 1 - q^{(j)}$ i per la proposició 3.1, $\frac{\partial \beta_{D^{j_i}}}{\partial \xi_i} = q^{(j_i)} p^{(j_i)}(\xi_{j_i})$, podem rescriure l'equació anterior com

$$\frac{\partial^N \beta_D}{\partial \xi_1 ... \partial \xi_N}(\emptyset) = \sum_{1 \leq j_1 \neq ... \neq j_N \leq M} \frac{\prod_{j=1}^M (1 - q^{(j)})}{\prod_{k=1}^N (1 - q^{(j_k)})} \prod_{k=1}^N q^{(j_k)} \prod_{k=1}^N p^{(j_k)}(\xi_{j_k}).$$

Caracteritzem el conjunt d'índexs sobre els quals sumem de la següent manera:

$$\{\{j_1, ..., j_N\} : j_i \neq j_k \text{ i } 1 \leq j_i \leq M\} = \{\gamma\pi(\{1, ..., N\}) : \pi \in \Pi_N, \gamma \in I(N, M)\},$$

on $I(N, M)$ és el conjunt de les injeccions creixents entre $\{1, ..., N\}$ i $\{1, ..., M\}$, i Π_N és el conjunt de les permutacions de $\{1, ..., N\}$. Amb això, i tenint en compte que el productori on apareixen els termes $q^{(j)}$ és independent de l'ordre, podem escriure

$$Q(\gamma) = \frac{\prod_{j=1}^{M}(1 - q^{(j)})}{\prod_{k=1}^{N}(1 - q^{(\gamma(k))})} \prod_{k=1}^{N} q^{(\gamma(k))}$$

i

$$\frac{\partial^N \beta_D}{\partial\xi_1 ... \partial\xi_N}(\emptyset) = \sum_{\pi\in\Pi_N} \sum_{\gamma\in I(N,M)} Q(\gamma) \prod_{k=1}^{N} p^{(\gamma(k))}(\xi\pi(k)). \tag{3.3}$$

Tenint en compte la caracterització de la funció de densitat donada en l'equació (3.1), l'enunciat del lema es dedueix d'aquesta última igualtat.

Observació 3.6. La suma sobre $I(N, M)$ en la funció de densitat donada en (3.2) tindrà en compte totes les possibles combinacions de característiques $\xi_1, ..., \xi_n$, mentre que $Q(\gamma)$ determinarà la probabilitat de cada elecció de característiques (noti's que és el producte de les probabilitats de les característiques presents q^j i les no presents $(1 - q^j)$), i $\prod p^j$ determinarà la probabilitat de la distribució de les característiques presents. Finalment en (3.3), la suma sobre les permutacions de Π_N dona una densitat simètrica, i té en compte totes les possibles assignacions entre característiques i variables d'entrada.

Amb la caracterització donada pel lema, i definint estimacions de q^j i p^j obtindrem $K_sigma(\cdot, D)$.

Definició 3.10. *Sigui D un diagrama de persistència aleatori i $\mathcal{D}$ un diagrama de persistència donat de dimensió homològica k fixada, $\mathcal{D} \subseteq \mathcal{W}_k = W \times \{k\}$. Sigui $\xi_j \in \mathcal{D}$. Considerem el diagrama unitari aleatori $D^j = \{\xi_j\}$ centrat en $\xi_j = (b_j, d_j)$. Aquest està definit per la probabilitat $q^{(j)}$ de no ser buit, i la funció de densitat $p^{(j)}$ segons la qual està distribuida la seva posició potencial. La funció de densitat estimada via nuclis de D centrada en $\mathcal{D}$ de f_D ve determinada per (3.2), on $q^{(j)}$, $p^{(j)}$ es defineixen de la següent manera:*

$$p^{(j)}(b, d) = \frac{\phi_j(b, d)}{\int_W \phi_j(u, v) \quad dudv} 1_W(b, d)$$

$$q^{(j)} = \mathbb{P}(D^j \neq \emptyset) = \int_{\{u>v\}} \phi_j(u,v) \quad dudv,$$

on ϕ_j la funció de densitat de $N((b_j, d_j), \sigma I)$.

En el següent exemple veurem un cas particular de l'estimació donada per l'anterior definició de la densitat d'un diagrama aleatori centrat en un diagrama de persistència.

Exemple 3.1. Considerem el diagrama de persistència $\mathcal{D} = \{(0.67, 0.68), (0.24, 1)\}$. En aquest exemple volem calcular la funció de densitat estimada pel diagrama de persistència aleatori D centrat en $\mathcal{D}$ segons la definició 3.10. Primer de tot cal que calculem les probabilitats q^j i les distribucions $p^j(b,d)$ dels diagrames unitaris D^j centrats en $\{(0.67, 0.68)\}$ i $\{(0.24, 1)\}$ respectivament:

$$q^{(0)} = \int_{u>v} \frac{1}{2\pi\sigma^2} e^{-((u-0.67)^2+(v-0.68)^2)/2\sigma^2} \approx 0.51$$

$$q^{(1)} = \int_{u>v} \frac{1}{2\pi\sigma^2} e^{-((u-0.24)^2+(v-1)^2)/2\sigma^2} \approx 0.86$$

Si tenim en compte que $\int_W \frac{1}{2\pi\sigma^2} e^{-((u-0.67)^2+(v-0.68)^2)/2\sigma^2} \approx 0.42$ i $\int_W \frac{1}{2\pi\sigma^2} e^{-((u-0.24)^2+(v-1)^2)/2\sigma^2} \approx 0.55$, i prenent $\sigma = 0.5$, tindrem que

$$p^{(0)}(b,d) = \frac{2}{0.42\pi} e^{-2((b-0.67)^2+(d-0.68)^2)}$$

$$p^{(1)}(b,d) = \frac{2}{0.55\pi} e^{-2((b-0.24)^2+(d-1)^2)}.$$

Segons (3.1), la funció de densitat de D vindrà descrita per $\{f_0, f_1(b,d), f_2((b_0,d_0),(b_1,d_1))\}$:

$$f_0 = \mathbb{P}(|D| = 0) = (1 - q^{(0)})(1 - q^{(1)}) \approx 0.07$$

$$\begin{aligned} f_1(b,d) &= (1-q^{(1)})q^{(0)}p^{(0)}(b,d) + (1-q^{(0)})q^{(1)}p^{(1)}(b,d) \\ &\approx 0.11e^{-2((b-0.67)^2+(d-0.68)^2)} + 0.49e^{-2((b-0.24)^2+(d-1)^2)} \end{aligned}$$

$$\begin{aligned} f_2((b_0,d_0),(b_1,d_1)) &= \frac{q^{(0)}q^{(1)}}{2}[p^{(0)}((b_0,d_0))p^{(1)}((b_1,d_1)) + p^{(0)}((b_1,d_1))p^{(1)}((b_0,d_0))] \\ &\approx 0.38[e^{-2((b_0-0.67)^2+(d_0-0.68)^2+(b_1-0.24)^2+(d_1-1)^2)} \\ &\quad + e^{-2((b_1-0.67)^2+(d_1-0.68)^2+(b_0-0.24)^2+(d_0-1)^2)}] \end{aligned}$$

Noti's que integrant les densitats locals (és a dir, sumant la probabilitat de cada cardinal) obtenim:

$$\mathbb{P}(|D| = 0) + \mathbb{P}(|D| = 1) + \mathbb{P}(|D| = 2) = f_0 + \int_W f_1(\xi) \quad d\xi + \int_{W^2} f_2(\xi_0, \xi_1) \quad d\xi_0 d\xi_1$$
$$= 0.07 + (0.11 \times 0.66 + 0.49 \times 0.86) + 0.76 \times (0.28 + 0.28) \approx 1$$

Les densitats locals per al cas $|D| = 1$ i $|D| = 2$ és mostren en la figura 5.1.2. Noti's que en el cas que D tingui només un punt, domina la funció gaussiana de la característica més persistent, tal i com és desitjable, mentre que en el cas que D tingui dues característiques, si una d'aquestes té una persistència mitja-alta (p.e. (0.4,0.1)), veurem que la funció prioritza lleugerament la característica de menor persistència.

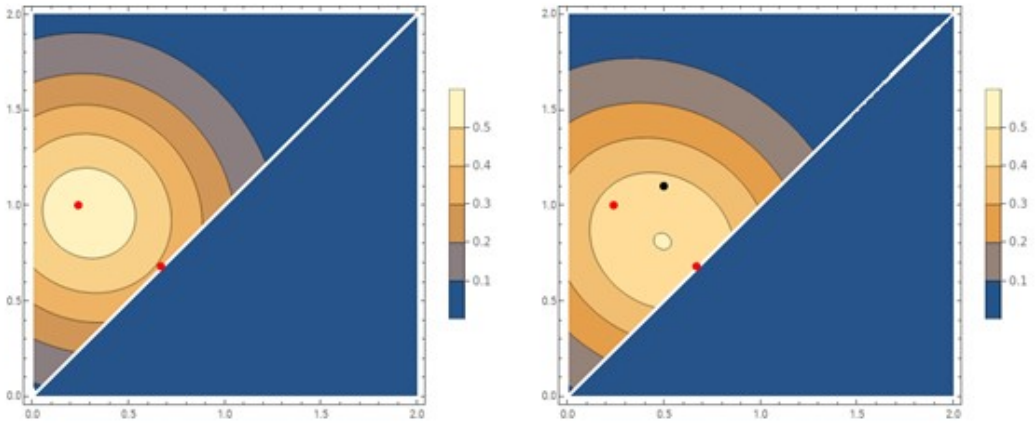


Figura 3.1: Esquerra: Contorns de la gràfica de f_0. Dreta: Contorns de la gràfica de $f_2((b, d), (0.4, 1))$. Els punts vermells indiquen el diagrama sobre el qual la densitat està centrada. El punt negre indica el punt auxiliar aobre el qual s'avalua f_2 per obtenir-ne una representació en W.

En l'exemple anterior veiem que aquesta estimació de la densitat és flexible, i captura el fet fonamental que el cardinal d'un diagrama de persistència donat és variable; no obstant, no contempla que el cardinal pugui excedir el del diagrama central sobre el qual construïm la densitat. Per exemple, si considerem un diagrama de persistència 0-dimensional sobre un núvol de punts D i un altre sobre aquest mateix núvol amb un punt afegit D^+, tindrem que $|D^+| = |D| + 1$. Si considerem que els dos diagrames pro-

venen de la mateixa distribució f_D hauria de considerar la possibilitat de diagrames com $\mathcal{D}^+$.

La proposta presentada en [9] per abordar aquest problema és la següent: separar el diagrama de persistència en dos subdiagrames independents, un incloent les característiques topològiques de llarga peristència i un altre les de curta persistència. Els punts de llarga persistència representen característiques topològiques prominents i estables respecte a variacions del núvol de punts base, i per tant seran tractades com a característiques independents, estimant la densitat com en la definició 3.10 (així la densitat no tindrà en compte o penalitzarà l'aparició de característiques topològiques de persistència llarga). Els punts de curta persistència segueixen essent importants per capturar la topologia del conjunt i per tant cal tenir-los en compte, però la seva estabilitat és menor i seran més susceptibles a aparèixer i desaparèixer, i per tant seran tractats de manera col·lectiva, com una característica única (aquesta part de la densitat serà la que considerarà l'aparició de nous punts). Per formalitzar l'anterior idea ens valem de la definició següent.

Definició 3.11. *Donat un ample de banda σ i un diagrama de persistència $\mathcal{D}$, definim el diagrama de persistència superior com $\mathcal{D}^u\{(b_i, d_i, k) \in \mathcal{D} : d_i - b_i \geq \sigma\}$ i el diagrama de persistència inferior com $\mathcal{D}^l\{(b_i, d_i, k) \in \mathcal{D} : d_i - b_i < \sigma\}$.*

Observació 3.7. L'ample de banda associat a la definició dels nuclis gaussians σ_1 i el que s'usa en l'anterior definició σ_2 no han de ser forçosament el mateix; no obstant i tal com s'indica en [9], és una bona pràctica donat que $\sigma_1 << \sigma_2$ projectar les característiques de persistència baixa a la diagonal pot conduir a un error significatiu, i $\sigma_1 >> \sigma_2$ elimina el benefici de partir el diagrama en dos.

Aplicarem la definició 3.10 per estimar la densitat de $\mathcal{D}^u$, mentre que construirem una nova estimació per a la de $\mathcal{D}^l$.

Definició 3.12. *Sigui $\xi_j \in \mathcal{D}^u$. Considerem el diagrama unitari aleatori $\mathcal{D}^j = \{\xi_j\}$ centrat en $\xi_j = (b_j, d_j)$. Aquest està definit per la probabilitat $q^{(j)}$ de no ser buit i la funció de densitat $p^{(j)}$ segons la qual està distribuïda la seva posició potencial. La funció de densitat estimada via nuclis gaussians de f_D ve determinada per l'equació (3.2), on $q^{(j)}$ i $p^{(j)}$ es defineixen de la següent manera:*

$$p^{(j)}(b, d) = \frac{\phi_j(b, d)}{\int_W \phi_j(u, v) \quad du dv} 1_W(b, d)$$

$$q^{(j)} = \mathbb{P}(D^j \neq \emptyset) = \int_{\{u>v\}} \phi_j(u,v) \quad dudv,$$

on ϕ_j és la funció de densitat de $N((b_j,d_j),\sigma I)$.

Definició 3.13. *El diagrama de persistència inferior aleatori D^l vindrà definit per l'elecció d'un cardinal N distribuït segons una funció de probabilitat ν i N punts i.i.d. segons una densitat p^l. Donat un diagrama de persistència $\mathcal{D}$, considerem $N_l = |\mathcal{D}^l|$, i una funció de densitat $\nu(\cdot)$ amb mitjana N_l i tal que $\nu(n) = 0$ per a $n > mN_l$ per cert a $m > 0$, independent de N_l. La densitat $p^l(b,d)$ es defineix com*

$$p^l(b,d) = \frac{1}{N_l} \sum_{(b_i,d_i)\in\mathcal{D}^l} \frac{1}{\pi\sigma^2} e^{-((b-\frac{b_i+d_i}{2})^2+(d-\frac{b_i+d_i}{2})^2)/2\sigma^2}.$$

Observació 3.8. La densitat $p^l(b,d)$ associada a un diagrama de persistència inferior està definida de manera que és l'estimació via nuclis gaussians de la densitat dels punts de $\mathcal{D}^l$ projectats a la diagonal.

De l'equació 3.1i la proposició 3.1, i de manera similar que en el lema 3.1 es pot provar el resultat següent.

Proposició 3.2. Donat un diagrama de persistència aleatori D^l, caracteritzat per ν i p^l, la seva funció de densitat ve donada per

$$f_{D^l}(\xi_1,...,\xi_N) = \nu(N)\prod_{j=1}^{N} p^l(\xi_j). \tag{3.4}$$

Demostració. Per la Proposició 3.1, tenm que

$$\beta_{D^l}(S) = \nu(0) + \sum_i = 0^{MN_l}\nu(i)\mathbb{P}(S||D| = i).$$

Teorema 3.1. *Sigui $\mathcal{D}$ un diagrama de persistència. Sigui $\sigma > 0$ un ample de banda fixat. Separem $\mathcal{D}$ en $\mathcal{D}^u$ i $\mathcal{D}^l$. Definim les funcions de densitat dels diagrames de persistència aleatoris D^l i D^u centrats en $\mathcal{D}^l$ i $\mathcal{D}^u$ i amb ample de banda σ segons les definicions 3.2, 3.12 respectivament. Tractant aquests dos diagrames com a independents, tenim que la funció de densitat del diagrama aleatori D, centrada en $\mathcal{D}$ i ample de banda σ, ve donada per*

$$K_\sigma(Z,\mathcal{D}) = \sum_{j=0}^{N_u} \nu(N-j) \sum_{\gamma\in I(j,N_u)} \mathcal{Q}(\gamma) \prod_{k=1}^{j} p^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p^l(\xi_k),$$

on $Z = (\xi_1, ..., \xi_N)$ és el diagrama que avaluem, $\xi_i = (b_i, d_i)$ són les característiques topològiques que el constitueixen i N_u és el cardinal de $\mathcal{D}^u$.

Demostració. Com que per hipòtesi D^u i D^l són independents, la funció de versemblança descompon com

$$\beta_D(S) = \beta_{D^l}(S)\beta_{D^u}(S).$$

Tenint en compte que les derivades de β_{D^u} d'ordre superior a N_u s'anul·len i aplicant la regla del producte, obtenim

$$\frac{\partial^N \beta_D}{\partial\xi_1...\partial\xi_N}(\emptyset) = \sum_{j=0}^{N_u} \sum_{1\le i_1\neq...\neq i_j\le N} \frac{\partial^j \beta_{D^u}}{\partial\xi_{i_1}...\partial\xi_{i_j}}(\emptyset) \frac{\partial^{N-j}\beta_{D^l}}{\partial\xi_1...\partial\hat{\xi}_{i_1}...\partial\hat{\xi}_{i_j}...\partial\xi_N}(\emptyset).$$

Si tenim en compte que l'ordre de la derivada és independent de l'elecció dels índexs i_j (vegi's [7, Corol·lari 10, Secció 4.2]), podem expressar-los en funció de $\pi \in \Pi_N$. D'aquesta manera obtindrem $j!$ termes redundants en el primer multiplicand i $(N-j)!$ en el següent, de manera que caldrà corregir cada summand amb un terme $1/(N-j)!j!$. Amb això obtenim la següent igualtat:

$$\frac{\partial^N \beta_D}{\partial\xi_1...\partial\xi_N}(\emptyset) = \sum_{\pi\in\Pi_N} \sum_{j=0}^{N_u} \frac{1}{(N-j)!j!} \frac{\partial^j \beta_{D^u}}{\partial\xi_{\pi(1)}...\partial\xi_{\pi(j)}}(\emptyset) \frac{\partial^{N-j}\beta_{D^l}}{\partial\xi_{\pi(j+1)}...\partial\xi_{\pi(N)}}(\emptyset). \tag{3.5}$$

Tenint en compte (3.1) i (3.4) , obtenim

$$\begin{aligned}\frac{\partial\beta_{D^l}^{N-j}}{\partial\xi_{\pi(j+1)}...\partial\xi_{\pi(N)}}(\emptyset) &= \sum_{\pi'\in\pi_{N-j}} f_{D^l}(\xi_{\pi'(j+1)}, ..., \xi_{\pi'(N)}) \\ &= \sum_{\pi'\in\pi_{N-j}} \nu(N-j)\prod_{j=1}^{N-j} p^l(\xi_{\pi'(j)}) = (N-j)!\nu(N-j)\prod_{j=1}^{N-j} p^l(\xi_j)\end{aligned}$$

Per altra banda, de la definició (3.2) es dedueix que

$$\begin{aligned}\frac{\partial^j \beta_{D^u}}{\partial\xi_{\pi(1)}...\partial\xi_{\pi(j)}}(\emptyset) &= \sum_{\pi'\in\Pi_j} f_{D^u}(\xi_{\pi'(1)}, ..., \xi_{\pi'(j)}) \\ &= \sum_{\pi'\in\Pi_j} \sum_{\gamma\in I(j,N_u)} \mathcal{Q}(\gamma)\prod_{k=1}^{j} p^{(\gamma(k))}(\xi_{\pi'(k)}).\end{aligned}$$

Substituint aquestes dues últimes igualtats en l'equació (3.5), obtindrem el resultat del teorema. La primer substitució és immediata, mentre que la segona cal tenir en compte

que afegeix $j!$ permutacions redundants, i ja recollides en π_N; així, multiplicant per un factor $j!$ cada sumand, obtenim l'expressió

$$\frac{\partial^N \beta_D}{\partial\xi_1 \ldots \partial\xi_N}(\emptyset) = \sum_{\pi\in\Pi_N} K_\sigma(Z,(D)).$$

De (3.2), deduim que $K_\sigma(Z,\mathcal{D})$ satisfà la definició de funció de densitat global per a $\boldsymbol{D}$.

El teorema anterior es pot estendre per a diagrames de persistència amb característiques homològiques de diferents dimensions.

Corol·lari 3.1. Sigui $\mathcal{D}$ un diagrama de persistència amb característiques de múltiples dimensions. Considerem $\mathcal{D} = \cup_{k=0}^{d-1}\mathcal{D}_k\times\{k\}$ on $\mathcal{D}_k$ és el diagrama restringit a dimensió k. Sigui D_k el diagrama de persistència aleatori centrat en $\mathcal{D}_k$. Suposant que D_k són independents, la funció de densitat global del diagrama $D = \cup D_k$ centrat en $\mathcal{D}$ amb ample de banda σ ve donada per

$$K_\sigma(Z,\mathcal{D}) = \Lambda(N)\prod_{k=0}^{d-1} K_\sigma(Z_k,\mathcal{D}_k),$$

on $Z = \cup_{k=0}^{d-1} Z_k\times\{k\} \subseteq \mathcal{W}_{0:d-1}$, $Z_k \subseteq W$ amb $|Z_k| = N_k$ i $\Lambda(N) = \frac{\prod N_k!}{(\sum N_k)!}$.

Demostració. Al considerar els diagrames en diferents dimensions independents, la funció de versemblança factoritza com $\beta_D(S) = \prod_{k=0}^{d-1}\beta_{D_k}(S)$. Tenint en compte que les derivades $\frac{\partial\beta_{D_k}}{\partial Z}(\emptyset) = 0$ per a qualsevol $Z \notin \mathcal{W}_k$ i aplicant la regla del producte obtindrem $\frac{\partial\beta_D}{\partial Z}(\emptyset) = \prod_{k=0}^{d-1}\frac{\partial\beta_{D_k}}{\partial Z_k}(\emptyset)$. Ara, tenint en compte que $\frac{\partial\beta_{D_k}}{\partial Z_k}(\emptyset) = \sum_{\pi\in\Pi_{N_k}} K_\sigma(\pi(Z_k),\mathcal{D}_k) = N_k!K_\sigma(Z_k,\mathcal{D}_k)$ i que $\frac{\partial\beta_D}{\partial Z}(\emptyset) = \sum_{\pi\in\Pi_{|N|}} K_\sigma(\pi(Z),\mathcal{D}) = |N|!K_\sigma(Z,\mathcal{D})$, obtenim

$$K_\sigma(Z,\mathcal{D}) = \frac{\prod_k N_k!}{(|N|)!}\prod_{k=0}^{d-1} K_\sigma(Z_k,\mathcal{D}_k).$$

Per acabar la secció donarem un exemple de còmput de la densitat d'un diagrama aleatori D centrat en $\mathcal{D}$, segons el teorema 3.1.

Exemple 3.2. Sigui D centrat en $\mathcal{D} = \{(0.24, 0.93), (0.82, 0.83), (0.87, 0.91)\}$. Fixem un ample de banda de $\sigma = 0.25$ i tenim

$$\mathcal{D}^u = \{(0.24, 0.93)\}$$

$$\mathcal{D}^u = \{(0.82, 0.83), (0.87, 0.91)\}.$$

Escollim una desviació estàndard $\sigma = 0.25$, i com a funció de probabilitat per al cardinal del diagrama inferior escollim

$$\nu(N) = \max\{\frac{N_l + 1 - |N - N_l|}{(N_l + 1)^2}\}.$$

Notem que $\mathbb{P}(|D| = N) = \sum_{N_u=0,1} P(|D^u| = N_u)\mathbb{P}(|D^l| = N - N_u) = (1 - q^{(0)})\nu(N) + q^{(0)}\nu(N-1)$. Tenint en compte que $q^{(0)} \approx 0.84$, tindrem que la distribució dels cardinals del diagrama aleatori centrat en $\mathcal{D}$ és la següent:

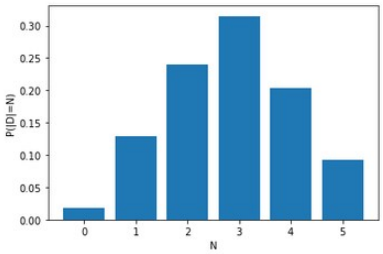


Figura 3.2: Probabilitats que $|D| = N$. Notem que la probabilitat es centra en $|\mathcal{D}| = 3$, i $\mathbb{P}(|D| = N) > 0$ per a $N \leq 5$. Per a $N > 6 = 2|\mathcal{D}|$, $\mathbb{P}(|D| = N) = 0$.

La densitat que descriurà la distribució de les característiques topològiques de persistència baixa vindrà donada per l'estimació via nuclis de la densitat associada a les projeccions de ξ_1, ξ_2 a la diagonal, $(0.825, 0.825)$ i $(0.89, 0.89)$ respectivament, i restringida a W;

$$p^l(b, d) = \frac{8}{\pi}[e^{-2((b-0.825)^2+(d-0.825)^2} + e^{(b-0.89)^2+(d-0.89)^2)}].$$

La funció de densitat vindrà donada per les funcions locals donades pels diferents cardinals $N \in \{0, ..., 5\}$. Per a $N = 0, 1$ tindrem

$$K_\sigma(\emptyset, \mathcal{D}) = \nu(0)(1 - q^{(0)}),$$

$$K_\sigma(\xi_0, \mathcal{D}) = \nu(0)q^{(0)}p^{(0)}(b, d) + \nu(1)(1 - q^{(0)})p^l(b_0, d_0),$$

mentre que per a $N > 1$ la densitat vindrà donada per

$$K_\sigma(((b_0,d_0),...,(b_N,d_N)),\mathcal{D}) = \nu(N-1)q^{(0)}p^{(0)}(b_0,d_0)\prod_{i=1}^{N}p^l(b_i,d_i) + \nu(N)(1-q^{(0)})\prod_{i=0}^{N}p^l(b_i,d_i).$$

Tenint en compte l'anterior descripció de p^l, que

$$p^{(0)}(b,d) = \frac{8}{0.52\pi}e^{-8((b-0.24)^2+(d-0.93)^2},$$

$q^{(0)} = 0.84$ i que $\nu(\{0,1,2,3\}) = \{1/9, 2/9, 3/9, 2/9\}$, obtindrem les següents aproximacions numèriques per a les respectives densitats locals:

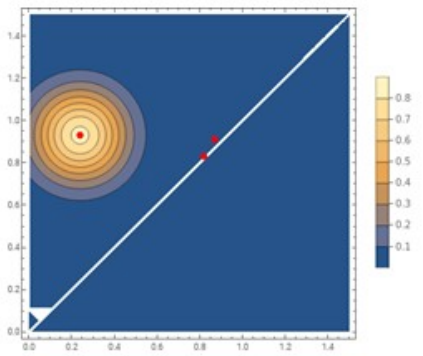


Figura 3.3: Gràfic de contorn de la densitat de $K_\sigma(b,d)$.

$$K_\sigma(\emptyset,\mathcal{D}) \approx 0.018$$

$$K_\sigma(\xi_0,\mathcal{D}) = 0.114e^{-2((b_0-0.24)^2+(d_0-0.93)^2} + 0.022[e^{-2((b_0-0.825)^2+(d_0-0.825)^2} + e^{(b_0-0.89)^2+(d_0-0.89)^2)}]$$

$$K_\sigma((\xi_0,\xi_1),\mathcal{D}) = 0.228e^{-2((b_0-0.24)^2+(d_0-0.93)^2}[e^{-2((b_1-0.825)^2+(d_1-0.825)^2} + e^{(b_1-0.89)^2+(d_1-0.89)^2)}] + 0.033[e^{-2((b_0-0.825)^2+(d_0-0.825)^2} + e^{(b_0-0.89)^2+(d_0-0.89)^2)}] [e^{-2((b_1-0.825)^2+(d_1-0.825)^2} + e^{(b_1-0.89)^2+(d_1-0.89)^2)}]$$

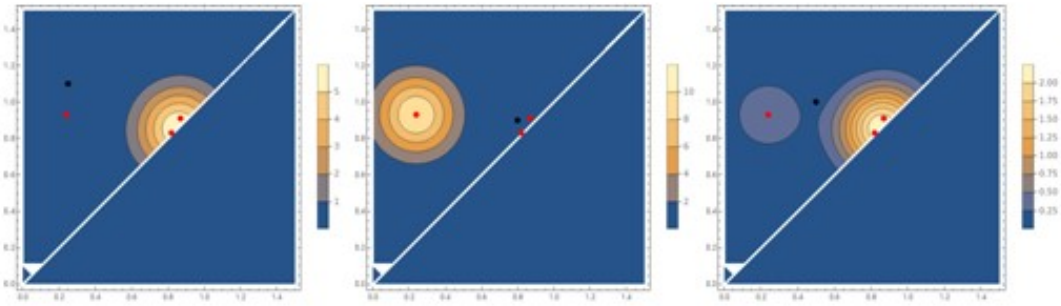


Figura 3.4: Gràfics de contorn de la densitat de $K_\sigma(\xi_1, \xi_2)$ per a $\xi_2 = (0.25, 1.1), (0.8, 0.9), (0.6, 1)$ respectivament.

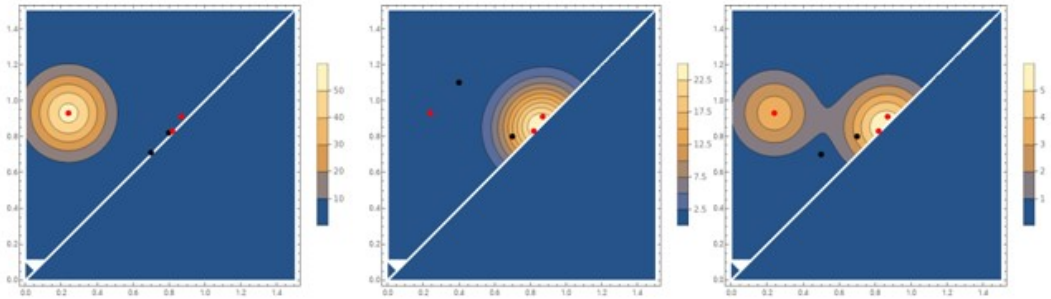


Figura 3.5: Gràfics de contorn de la densitat de $K_\sigma(\xi_1, \xi_2, \xi_3)$ per a $(\xi_1, \xi_2) = ((0.7, 0.71), (0.8, 0.82)), ((0.7, 0.8), (0.4, 1.1))$.

3.3.0 Convergència

Suposem donat un conjunt de diagrames de persistència $\{D_i\}_{i=1}^n$ independents i idènticament distribuïts, provinents d'una distribució definida per una funció de densitat global f. L'estimació via nuclis de la funció de densitat convergirà a la funció f al fer tendir n a infinit, tal i com mostra el resultat següent [9, Teorema 2, Secció 4].

Teorema 3.2. *Sigui D un diagrama de persistència aleatori amb funció de densitat global f que satisfà:*

i) $f(Z) = 0$ per a tot $|Z| > M$.

ii) La funció de densitat local $f_N : \mathcal{W}_k^N \to \mathbb{R}$ *està acotada a* $N \in \{1, ..., M\}$.

iii) Existeix $C_N > 0$ *tal que* $f(\xi_1, ..., \xi_N) \leq C_N \|(\xi_1, ..., \xi_N)\|^{-2N}$ *per a qualsevol* $N \in \{1, ..., M\}$.

Definim $K_\sigma(Z, \mathcal{D})$ *com en el teorema 3.1 i considerem* $\hat{f}(Z) = \sum_{i=0}^{n} \frac{1}{n} K_\sigma(Z, \mathcal{D}_i)$, *on* $\mathcal{D}_i$ *són i.i.d. amb la distribució definida per* f *i* $\sigma = \mathcal{O}(N^{-\alpha})$ *on* $0 < \alpha < \alpha_{2M}$. *Sota aquestes condicions* $\hat{f} \xrightarrow{n\to\infty} f$ *uniformement en compactes de* W.

Observació 3.9. El valor α_M és un valor inherent a la selecció de l'ample de banda per a estimacions via nuclis de densitats $2M$-dimensionals [9, Remark 4.2].

Observació 3.10. La validesa del resultat anterior està condicionada a la validesa de les hipòtesis i), ii), iii), que donen condicions sobre la funció de densitat del diagrama de persistència aleatori que es vol estimar. Cal notar que aquestes condicions es compleixen per a diagrames aleatoris associats a núvols de punts amb els quals es treballa de manera usual. Per exemple, és trivial observar que i) es compleix per a qualsevol diagrama associat a un núvol de punts finit en $\mathbb{R}^d$. Les condicions ii) i iii) per exemple es compliran per diagrames de persistència construïts sobre núvols de punts obtinguts de distorsionar amb soroll gaussià compactes de $\mathbb{R}^d$ (vegi's [9, Consideracions prèvies al Teorema 4.2]).

4.0 Classificadors

En aquesta secció definim i motivaem la construcció d'un classificador homològic nou. Definirem primerament el context on es situa un problema de classificació. Sota aquest context definirem el classificador proposat en [5] per seguidament discutir les limitacions que presenta. Amb aquestes com a motivació i finalitzant la secció construirem i definirem el nou classificador.

4.1.0 El problema de classificació

El problema de classificació (en el context de l'aprenentatge automàtic) consisteix en predir la classe d'una observació donada (i no classificada) en base al coneixament d'un conjunt previ d'observacions classificades, és a dir, pertanyents a un conjunt de classes finit. Podem definir formalment un classificador de la següent manera.

Definició 4.1. *Donat (X, Y) un parell aleatori que pren valors en $\mathbb{R}^d \times \{1, ..., k\}$, un classificador és una funció mesurable*

$$g : \mathbb{R}^d \to \{1, ..., k\}.$$

Definició 4.2. *El problema de classificació consisteix en, donat (X, Y) un parell aleatori que pren valors en $\mathbb{R}^d \times \{1, ..., k\}$, trobar una funció $g : \mathbb{R}^d \to \{1, ..., k\}$, a la que anomenarem classificador, que minimitzi l'error en la predicció*

$$L(g) = \mathbb{P}(g(X) \neq Y).$$

Idealment aspirem a trobar $g^* = \operatorname{argmin}_{g:\mathbb{R}^d \to \{1,\dots,M\}} \mathbb{P}(g(X) \neq Y)$. En el cas que el parell (X, Y) sigui conegut, pot arribar a ser possible computar g^*, però en el context habitual només disposarem d'una mostra finita del parell, i de cap altre coneixement sobre aquest. Així tindrem d'un conjunt d'observacions etiquetades $T =$

$\{(x_1, y_1), ..., (x_n, y_n)\} \subseteq \mathbb{R}^d \times \{1, ..., k\}$, i en aquest context el problema de classificació consisteix en escollir una funció g com l'anterior a partir de T. Anomenarem T el conjunt d'entrenament i g serà un classificador supervisat.

Definició 4.3. *Un classificador supervisat és una seqüència de funcions mesurables*

$$g_n(x, T) : \mathbb{R}^d \times \{\mathbb{R}^d \times \{1, ..., k\}\}^n \to \{1, ..., k\}.$$

Donat un conjunt d'entrenament T amb n observacions, per avaluar el rendiment del classificador cal computar $\mathbb{P}(g_n(X, T) \neq Y)$. En el cas habitual, al només poder accedir a una mostra finita del parell (X, Y), ens valdrem d' un subconjunt d'aquesta mostra $S = \{(x_{i_1}, y_{i_1}), ..., (x_{i_k}, y_{i_k})\}$ per estimar l'error de classificació. Computarem $\{g(x_{i_j}, T)\}_{j=1...k}$ i el compararem a través de diferents mètriques amb $\{y_{i_j}\}_{j=1...k}$. Normalment, donada una mostra finita P del parell (X, Y), la dividirem en dos subconjunts disjunts $P = T \cup S$ i usarem T com a conjunt d'entrenament i S com a conjunt d'avaluació.

4.2.0 Homologia persistent i classificació

Sigui E el conjunt dels núvols de punts finits de $\mathbb{R}^d$, $E = \{N \subset \mathbb{R}^d : |N| \in \mathbb{N}\}$. L'homologia persistent ha provat la seva eficàcia en la construcció de classificadors per a parells (X, Y) on X pren valors en E, i equivalentment, en la classificació de conjunts de núvols de punts. Bon exemple d'aquest fet és l'extensa gamma d'exemples que podem trobar en la literatura; vegi's per exemple REFERENCIES. Efectivament l'homologia persistent pot arribar a ser una bona caracterització d'un conjunt de núvols de punts que comparteixen certes característiques geomètriques o topològiques, i per tant pot arribar a ser una bona eina per classificar-los.

No obstant, quan es tracta de classificar punts a $\mathbb{R}^d$ (és a dir, quan X pren valors en $\mathbb{R}^d$, com en la definició 4.2) el problema no ha estat tant estudiat. En el que resta de secció i treball abordarem el problema de definir un classificador d'aquest tipus basat en homologia persistent. Noti's que construir un classificador d'aquestes característiques és equivalent a construir un classificador per a núvols de punts (com en la literatura citada anteriorment) en el cas particular que els núvols tinguin cardinal 1, és a dir, $E = E_1 = \{N \subset \mathbb{R}^d : |N| = 1\}$. Els classificadors existents per a classificació de núvols de punts via homologia persistent majoritàriament es basaran en codificar les característiques topològiques d'un núvol donat amb alguna de les eines aportades per la TDA, i posteriorment usar aquesta caracterització per predir la classe

del núvol. En el cas que aquest núvol tingui un sol element, aquesta informació topològica serà irrellevant. Amb això volem fer notar que els classificadors proposats fins ara no es poden aplicar directament al problema que ens concerneix. En aquest sentit les úniques propostes de classificadors han estat donades per [8] i [5].

En el nostre cas, estudiarem la proposta donada en REF, les seves limitacions i veurem com a partir d'aquestes construir un nou classificador supervisat per al problema de classificació de punts a $\mathbb{R}^d$.

4.3.0 El classificador FP

Anomenarem classificador FP al classificador proposat en [5] que en la notació de la secció anterior es defineix de la següent manera.

Definició 4.4. *Sigui $X \subseteq \mathbb{R}^d$ un núvol de punts i $x \in \mathbb{R}^d$. Anomenarem diagrama de persistència de X ampliat respecte a x al diagrama de persistència $D(X \cup \{x\})$, és a dir, al diagrama construït sobre el núvol de punts X en afegir-hi x.*

Definició 4.5. *Donat un problema de classificació determinat per un parell (X, Y), el classificador de FP sobre un conjunt d'entrenament $T = \{(x_1, y_1), ..., (x_n, y_n)\}$ ve determinat per la funció*

$$C_n^{FP}(x, T) = argmin_{\{1,...,k\}} d(l(D(T^i)), l(D(T^i \cup \{x\})))$$

on $T^i = \{x_j : (x_j, y_j) \in T, y_j = i\}$ és el conjunt de punts d'entrenament amb etiqueta $i \in \{1, ..., k\}$, $D : E \to \cup_{N=0}^{M} W^N$ retorna el diagrama de persistència d'un núvol de punts donat, $l : \cup_{N=0}^{M} W^N \to \mathbb{R}^m$ és una representació lineal del diagrama de persistència, i $d : \mathbb{R}^m \to \mathbb{R}_0$ és la distància euclidiana.

Observació 4.1. El classificador en [5] es defineix com en la definició anterior en els casos particulars que l sigui el paisatge persistent, la silueta, l'entropia persistent, o bé no prenent l i considerant d com la distància *bottleneck*. Vegi's [5, Secció 3].

Com podem observar, el classificador basa la seva regla de decisió en la validesa (en major o menor mesura) de la següent hipòtesi.

Hipòtesi FP: L'homologia persistent d'una classe variarà menys en afegir punts d'aquesta mateixa classe que en afegir-ne d'altres, o equivalentment que

$$r = argmin_{\{1,...,k\}} d(l(D(X_{train_i})), l(D(X_{train_i}) \cup \{x_i\}))$$

per qualsevol $r \in \{1, ..., k\}$ i $x_i \in X_r$.

4.4.0 Limitacions

Com es pot veure en [5, Apèndix A] i 5.3, en 0-homologia el comportament del classificador és correcte. No obstant, el comportament canvia totalment quan s'usen els diagrames de persistència 1-dimensionals per a la classificació.

Com veurem, tot i que la hipòtesi de FP pot semblar raonable en dimensió 0, la casuística d'aquesta es pot tornar més complexa en casos de dimensió superior.

En el cas de la 0-homologia l'addició d'un punt a un núvol genera una nova component connexa que tindrà una certa persistència i per tant el diagrama ampliat serà diferent de l'original: $d(l(D(X_i)), l(D(X_i)) \cup \{x\})) \neq 0$ per a qualsevol X_i i $x \notin X_i$. En els casos d'homologies de dimensió superior, l'addició d'un punt no sempre garantirà que el diagrama ampliat sigui diferent de l'original. En aquests casos el comportament del diagrama ampliat tendirà a ser el contrari que el suposat per la hipòtesi FP, i punts x aliens a un cert núvol de punts X no generaran ni modificaran cap característica 1-homològica, donant diagrames ampliats sense modificacions: $d(l(D(X)), l(D(X)) \cup \{x\})) = 0$, mentre que punts pertinents a la classe del núvol si que en modificaran la 1-homologia i $d(l(D(X)), l(D(X)) \cup \{x\})) \neq 0$. La casuística de la variació 1-homològica en afegir un punt es torna més complexa, i aquí la hipòtesi de FP falla àmpliament. Veurem ara aquest fet en un exemple concret.

Exemple 4.1. Suposem que volem classificar el conjunt de dades *Cercles*, consistent en dos núvols de punts distribuïts sobre dos cercles concèntrics i distorsionats amb soroll gaussià X^+, X^- (vegi's Figura 4.1, columna 1). Sobre aquest conjunt intentarem classificar dos punts: un pertanyent al cercle interior (Figura 4.1, punt vermell, gràfics 2n i 4t de la 1a columna) i un altre al cercle superior (Figura 4.1, punt vermell, gràfics 1r i 3r de la 1a columna). Considerem els diagrames de persistència $\mathcal{D}^+, \mathcal{D}^-$ de les classes exterior i interior respectivament (Figura 4.1, 3a columna).

Ens centrarem en analitzar la classificació 1-homològica. Suposem primer que volem classificar el punt del cercle exterior x^+. Prenem la classe exterior i la interior, i hi afegim el punt (Figura 4.1, columna 2, diagrames 1r i 3r), en computem els respectius diagrames ampliats en dimensió 1 (Figura 4.1, columna 4, diagrames 1r i 3r) i els comparem amb els diagrames de les classes originals (Figura 4.1, columa 3). Per comparar-los ho fem a través de les respectives siluetes (columna 4). Podem observar que mentre que afegir x^+ a la classe exterior varia lleugerament la seva silueta, en afegir-lo a la classe interior la silueta es manté igual (efectivament al ser un punt llunyà la seva addició a X^- no varia la seva 1-homologia persistent). Suposem ara

que volem classificar un punt del cercle interior. En aquest cas com podem veure en la comparació de siluetes (4.11, columna 4, diagrames 2n i 4t) afegir el punt a la classe interior varia lleugerament la seva 1-homologia mentre que afegir-lo a la classe exterior varia significativament la 1-homologia.

Aquests dos exemples ens mosten una tendència que en aquest núvol es complirà en general:

1.1. A l'afegir un punt exterior a la classe exterior, variarà poc la 1-homologia. Pot fer variar les característiques menys persistents, i en tot cas afectarà poc a la principal característica 1-homològica de la classe (forat central).

1.2. A l'afegir un punt exterior a la classe interior no variarà la 1-homologia.

2.1. A l'afegir un punt interior a la classe exterior no variarà significativament la 1-homologia. Els punts del cercle interior escurcen la vida de la principal característica 1-homològica de la classe exterior.

2.2. A l'afegir un punt interior a la classe interior la 1-homologia no variarà, o bé ho farà poc significativament.

Així doncs el classificador tendirà a classificar bé els punts interiors, i a estar indecís o classificar malament els punts exteriors. La hipòtesi de FP falla per a punts de la classe exterior.

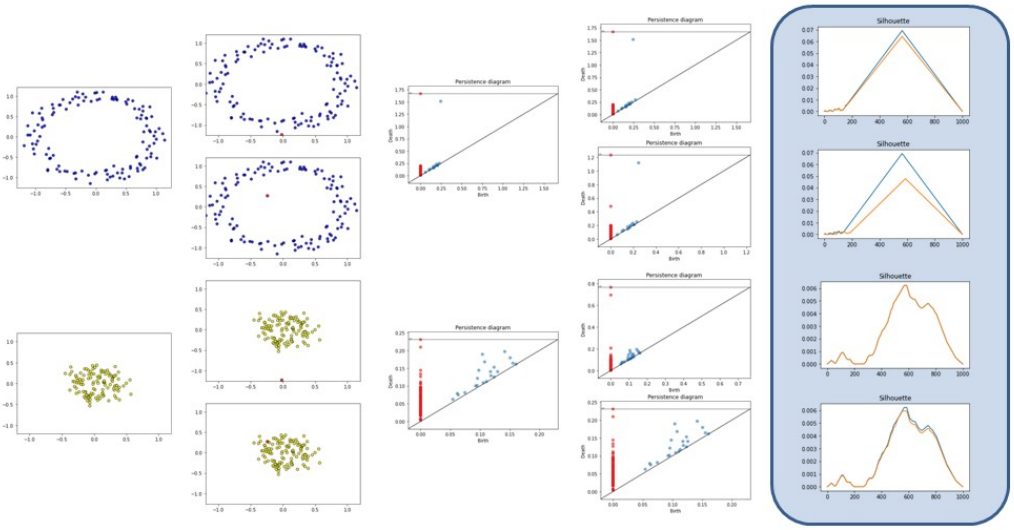


Figura 4.1: Exemple de classificació de FP en dos punts donats.

De l'anterior exemple es desprèn que tot i que la hipòtesi de FP no sigui capaç de caracteritzar els punts de les dues classes, es pot observar que en general les variacions homològiques respecte l'addicció de punts (i per tant els diagrames ampliats) sí que ho permeten fer: els punts de la classe exterior no varien l'homologia de la classe interior, i varien poc la de la seva classe; els punts de la classe interior varien significativament l'homologia de la classe exterior i no varien o varien poc l'homologia de la classe interior. Amb això volem fer notar que la variació homològica pot ser una característica distintiva de la classe, però a vegades no sota la hipòtesi de FP.

4.5.0 Classificador basat en la densitat de les variacions homològiques

En aquest punt ens proposem la construcció d'un classificador basat en la variació de l'homologia persistent respecte a l'addició puntal al conjunt subjacent, que eviti l'ús de HFP i que sigui capaç de capturar la variació homològica característica d'una classe respecte a l'addició de punts, que tal i com hem apuntat en el final de la secció anterior pot ser rellevant en classificació.

Suposem que ens trobem en el context d'un problema de classificació binari. És a dir, suposem que tenim (X, Y) un parell on Y pren valors a $\{0, 1\}$. Considerem el vector aleatori associat a cada classe $X^i = (X|Y = i)$ i una mostra d'aquest $T^i = \{x^i_1, ..., x^i_{n_i}\}$. El procés de classificació que proposem consistirà en dues parts. En un primer pas, obtindrem un conjunt de mostres $\{\mathcal{D}^{ij}\}_{i,j=0,1}$ dels diagrames de persistència de T^i ampliats sobre punt de les diverses classes X^j. Noti's que podem considerar aquestes com a mostres dels següents diagrames de persistència aleatoris.

Definició 4.6. *En el context anterior, anomenarem diagrama aleatori de T^i ampliat respecte a X^j a*

$$D^{ij} = D(T^i \cup X^j)$$

(els elements d'aquest diagrama aleatori seràn diagrames construïts sobre la mostra T^i en afegir un punt x distribuït segons X^j, és a dir, seran diagrames de persistència de T^i ampliats respecte punts distribuïts segons X^j).

Per obtenir la mostra $\mathcal{D}^{ij}$ de D^{ij} considerarem

$$T^i_a = \{x^i_1, ..., x^i_{l_i}\},$$

$$T^i_b = \{x^i_{l_i+1}, ..., x^i_{n_i}\},$$

on $l_i \in \{1, ..., i_{n_i} - 1\}$. Els punts T_b^i s'usaran com a núvol de punts base per computar l'homologia persistent de la classe i, i els T_a^j com a punts a afegir al núvol base de la classe i per tal de construir la mostra dels diagrames persistents de la classe i ampliats respecte punts de la classe j. Amb això, aconseguim mostres $\mathcal{D}^{ij}$ de D^{ij} tal i com buscàvem:

$$\mathcal{D}^{ij} = \{\mathcal{D}_k^{ij} = D(T_b^i \cup \{x_k^j\})\}_{k=1,...,l_j}$$

Ara, donat un punt a classificar x, distribuït segons X (segons X^i per a cert i desconegut), podem calcular els seus diagrames ampliats respecte a les classes 0, 1

$$\mathcal{D}_x^0 = D(T_b^i \cup \{x\})$$

$$\mathcal{D}_x^1 = D(T_b^i \cup \{x\}).$$

Un cop tinguem $\mathcal{D}^{00}$, $\mathcal{D}^{01}$, $\mathcal{D}^{10}$, $\mathcal{D}^{11}$ i $\mathcal{D}_x^0$, $\mathcal{D}_x^1$, per tal de classificar x inferirem de quina de les dues mostres $\mathcal{D}^{i0}$, $\mathcal{D}^{i1}$, és més probable que provingui $\mathcal{D}^{\geq}$. Aquesta serà la segona part del nostre mètode. Notem que inferirem l'anterior per $i = 0, 1$ obtenint dos classificadors diferents, un per a cada núvol de punts base.

Per a fer l'inferència en qüestió ens basarem en l'avaluació de l'estimació de la densitat dels diagrames aleatoris D^{ij} (estimada sobre la mostra $\mathcal{D}^{ij}$) sobre el diagrama $\mathcal{D}_x^i$. Si el diagrama de persistència en questió té un alt nombre de punts de llarga persistència aquest comput no serà factible (vegi's Observació 4.2). Per tant, cal que ens restringim a diagrames que continguin la mateixa informació (respecte la variació homològica d'una classe al afegir un punt) amb un menor cardinal. Aquesta qüestió es resol amb la següent definició.

Definició 4.7. *Sigui $X \subseteq \mathbb{R}^d$ un núvol de punts i $x \in \mathbb{R}^d$. Anomenarem diagrama de persistència variacional de X respecte a x, al diagrama de persistència $D(X \cup \{x\}) \backslash D(X)$; és a dir, al diagrama construït sobre el núvol de punts X en afegir-hi x, només tenint en compte les característiques topològiques que han variat.*

Noti's que podem desenvolupar la primera part del mètode de manera completament anàloga substituint el còmput de les mostres de diagrames ampliats $\mathcal{D}^{ij}$ pel còmput de les mosters equivalents però de diagrames variacionals. Obtindrem mostres

$$\mathcal{D}^{ij\prime} = \{\mathcal{D}^{ij\prime}_k = D(T_b^i \cup \{x_k^j\}) \backslash D(T_b^i)\}_{k=1,...,l_j}, \tag{4.1}$$

que es poden veure com a mostres dels diagrames variacionals aleatoris D^{ij} definits de la manera següent.

Definició 4.8. *En el context en que ens trobem, anomenarem diagrama aleatori variacional de T_i respecte a X^j, al diagrama*

$$D^{ij\prime} = D(T^i \cup X^j) \backslash D(T^i)$$

(els elements d'aquest diagrama aleatori seràn diagrames constituits pels punts que es modifiquen al diagrama $D(T^i)$ al afegir a T^i un punt x distribuït segons X^j; és a dir, diagrames variacionals de T^i respecte X^j).

Per al punt a classificar x, també computem els corresponents diagrames variacionals de T^i respecte a x, $\mathcal{D}^{i\prime}_x$. Com abans, el problema es tractarà d'inferir de quina de les dues mostres $\mathcal{D}'^{i0}$, $\mathcal{D}'^{i1}$ és més probable que provingui $\mathcal{D}^{i\prime}_x$.

Donat el parell $(D^{i\prime}, Y)$ (on $D^{i\prime}$ representa el diagrama variacional aleatori de la classe i, i.e. el diagrama aleatori tal que $D^{ij\prime} = (D^{i\prime}|Y = j)$) del qual coneixem les mostres $\mathcal{D}^{i0\prime}$, $\mathcal{D}^{i1\prime}$, volem classificar $\mathcal{D}^{i\prime}_x$.

Per fer-ho considerarem la probabilitat $\mathbb{P}_{D,Y}$ del parell $(D^{i\prime}, Y)$, i les probabilitats condicionades i marginals corresponents. La idea serà classificar segons

$$BF^i(x) = \frac{\mathbb{P}_{Y|D}(Y = 0|D^{i\prime} = \mathcal{D}^{i\prime}_x)}{\mathbb{P}_{Y|D}(Y = 1|D^{i\prime} = \mathcal{D}^{i\prime}_x)}.$$

Si $BF(x) > 1$ considerarem que $\mathcal{D}^{i\prime}_x$, té etiqueta $Y = 0$, i per tant clasificarem x com a 0 (segons el classificador donat per la classe i); si $BF(x) \leq 1$ farem l'invers. Com que no tenim eines per calcular el factor de Bayes directament, caldrà que n'usem una aproximació:

$$\begin{aligned} BF^i(x) =& \frac{\mathbb{P}_{Y|D}(Y = 0|D^{i\prime} = \mathcal{D}^{i\prime}_x)}{\mathbb{P}_{Y|D}(Y = 1|D^{i\prime} = \mathcal{D}^{i\prime}_x)} = \frac{\mathbb{P}_{D|Y}(D^{i\prime} = \mathcal{D}^{i\prime}_x|Y = 0)\mathbb{P}_Y(Y = 0)}{\mathbb{P}_{D|Y}(D^i = \mathcal{D}^{i\prime}_x|Y = 1)\mathbb{P}_Y(Y = 1)} \\ \approx& \frac{f_{D^{i0\prime}}(\mathcal{D}^{i\prime}_x)}{f_{D^{i1\prime}}(\mathcal{D}^i_x)} \approx \frac{\mathrm{KDE}_\sigma(\mathcal{D}^{i\prime}_x, \mathcal{D}^{i0\prime})}{\mathrm{KDE}_\sigma(\mathcal{D}^{i\prime}_x, \mathcal{D}^{i1\prime})}, \end{aligned}$$

on la tercera equivalència prové de considerar les dues classes equiprobables, i de considerar $\mathbb{P}_{D|Y}(\cdot|Y = j) \propto f_{D^{ij\prime}}$. Finalment l'ultima equivalència prove d'estimar les corresponents densitats via nuclis sobre les respectives mostres $\mathcal{D}^{ij\prime}$. Al prendre una aproximació considerarem un llindar c (enlloc de 1) que serà un hiperparàmetre del model. La classificació doncs resultarà 1 si $BF(x) \leq c$ i 0 si $BF(x) > c$.

Observació 4.2. En aquest punt es justifica l'elecció de diagrames variacionals. Per evaluar la KDE sobre un diagrama de N punts caldran $|\mathcal{D}^{ij\prime}|N!$ avaluacions de funcions de $\mathbb{R}^N \to \mathbb{R}$, cosa que eleva molt el cost computacional de l'operació. Noti's

que generalment el cardinal d'un diagrama variacional serà molt menor que el d'un ampliat, cosa que reduirà significativament aquest cost.

Amb tot, obtindrem dos classificadors, depenent de la classe base escollida, i depenents d'un conjunt d'hiper-paràmetres (n, c, σ) que es definiran de la següent manera.

Definició 4.9. *Sigui (X, Y) un problema de classificació binari com en la definició 4.2, i T una mostra d'aquest. Si separem la mostra en quatre submostres $T_a^0, T_a^1, T_b^0, T_b^1$ i amb aquestes construim les respectives mostres de diagrames variacionals $D^{00\prime}, D^{01\prime}, D^{10\prime}, D^{11\prime}$ com en (4.1), definirem el classificador DF (classificador basat en el factor de les densitats) com*

$$C^{DF}_{n,i,\sigma,c}(x) = \begin{cases} 0 & si \quad \frac{KDE_\sigma(D^i_x, D^{i0\prime})}{KDE_\sigma(D^i_x, D^{i1\prime})} > c \\ 1 & si \quad \frac{KDE_\sigma(D^i_x, D^{i0\prime})}{KDE_\sigma(D^i_x, D^{i1\prime})} \leq c \end{cases} .$$

Observació 4.3. En l'anterior definició n, i, σ i c seran considerats hiperparàmetres del model: n indica la dimensió homològica dels diagrames de persistència que prenem, i indica la classe del núvol de punts base i σ indica l'ample de banda i el paràmetre de partició de l'aproximació via nuclis de les densitats.

4.6.0 ALGORITME

Algorithm 1 Classificador DF

Input: una mostra finita donada $\{(x_1, y_1), ..., (x_n, y_n)\}$ on $x_i \in \mathbb{R}^d$ i $y_i \in \{1, ..., k\}$ i $x^* \in \mathbb{R}^d$. Hiper-paràmetres n, i, c, σ

Output: l'etiqueta del punt x^*, c^*.

Algorisme:

1.Separem la mostra segons les seves etiquetes $T^i = \{(x_j, y_j) : y_j = i\}$.

2. Separem

$$T^i = T^i_b \cup T^i_a$$

on $T^i_b \cap T^i_a = \emptyset$, $i = 1, ..., n$.

2. Calculem $\mathcal{D}^{ij\prime} = D(T^i_b \cup \{x\}) \backslash D(T^i_b)$ per $x \in T^j_a$.

Calculem $\mathcal{D}^{i\prime}_{x^*} = D(T^i_b \cup \{x^*\}) \backslash D(T^i_b)$

3. Estimem les densitats f_{ij} corresponents a les mostres $\mathcal{D}^{ij\prime}$ a través de KDE$(\cdot, \mathcal{D}^{ij\prime})$.

4. $BF \leftarrow \frac{\text{KDE}_\sigma(\mathcal{D}^i_x, \mathcal{D}^{i0\prime})}{\text{KDE}_\sigma(\mathcal{D}^i_x, \mathcal{D}^{i1\prime})}$

5. **if** BF>c: **return** 0

else **return** 1 =0

5.0 Resultats

En aquesta secció provarem el classificador basat en densitat sobre diferents problemes de classificació analitzant el seu comportament davant el classificador FP, i un classificador base poc complex que en aquest cas serà un 1-NN.

5.1.0 Datasets

Per avaluar el funcionament del classificador davant diferents problemes hem escollit una sèrie de datasets reals i artificials que descriurem tot seguit.

5.1.1. Datasets artificials

Escollim tres datasets artificials diferents.

El conjunt de dades **Cercle** consisteix en dos conjunts de punts distribuïts sobre dos cercles concènctrics, i distorssionats amb soroll provinent d'una distribució gaussiana. Cada un dels cercles tindrà una etiqueta $\{0, 1\}$. Treballarem sobre aquest conjut amb sorolls $0.1, 0.3, 0.6$.

El conjunt de dades **Lluna** consisteix en dos conjunts de punts distribuïts sobre dos semicercles desplaçats i distorsionats amb soroll provinent d'una distribució gaussiana. Cada un dels semicercles tindrà una etiqueta $\{0, 1\}$. Treballarem sobre aquest conjut amb sorolls $0.1, 0.3, 0.6$.

El conjunt de dades **Quantils Gaussians** consisteix en un conjunt de punts provinents d'una distribució gaussiana a $\mathbb{R}^3$ i on definim 2 classes amb el mateix nombre de punts separades per una esfera. Cada un dels conjunts tindrà una etiqueta $\{0, 1\}$.

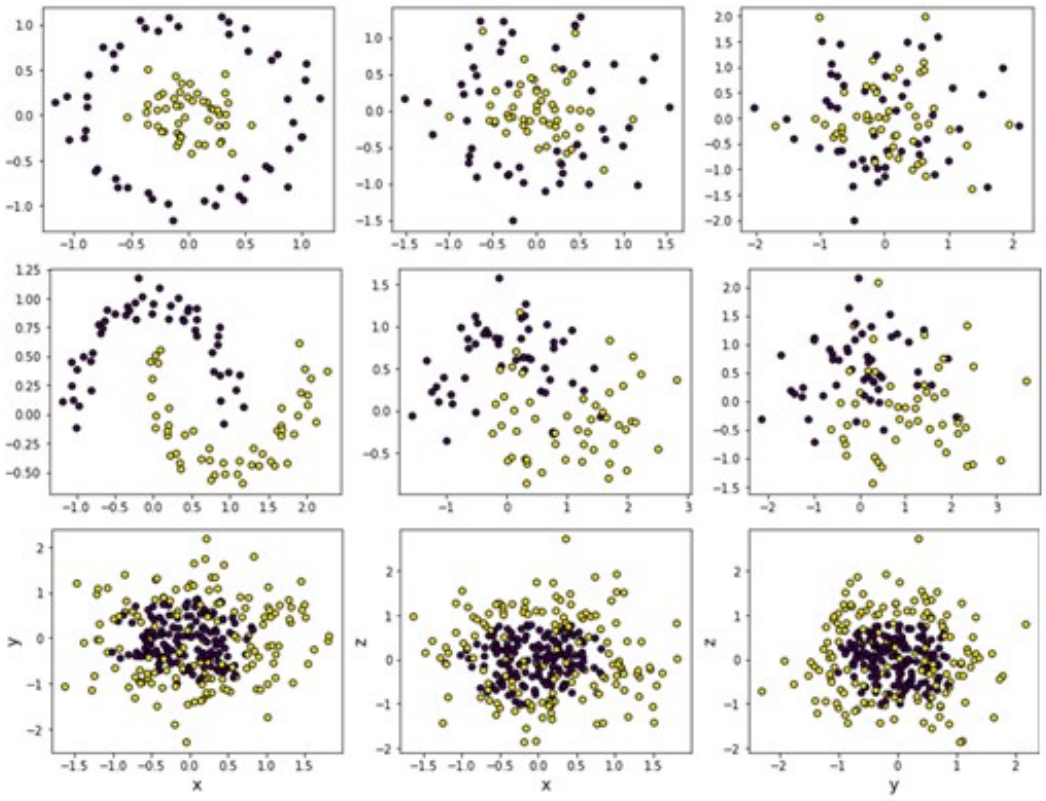


Figura 5.1: Gràfics dels conjunts de dades artificials Cercles (per a valors de distorsió 0.1, 0.3, 0.6), Lluna (per valors de distorsió 0.1, 0.3, 0.6), i QG

5.1.2. Datasets reals

Per a avaluar el classificador sobre dades reals, escollim tres conjunts diferents.

El conjunt de dades **Iris** consisteix en un conjunt de 50 dades de 3 classes diferents a $\mathbb{R}^4$. Cada classe indica un tipus de flor d'Iris, i les components de $\mathbb{R}^4$ indiquen la llargada i amplada del sèpal i del pètal respectivament. Cada una de les classes tindrà una etiqueta $\{0, 1, 2\}$.

El conjunt de dades **Vi** prové d'un anàlisi químic d'un conjunt de 178 vins d'una certa regió d'Itàlia. Consta de 178 dades 13−dimensionals i etiquetades dintre 3 classes desbalancejades. Cada una de les dimensions descriurà una característica del vi en

qüestió, i cada etiqueta representarà un tipus de vi. Cada una de les classes tindrà una etiqueta $\{0, 1, 2\}$.

El conjunt de dades **Càncer** consta de 569 dades 30−dimensionals i etiquetades dintre 3 classes desbalancejades. Cada un dels 569 punts representarà un pacient, i les diferent dimensions representaran característiques del nucli celul·lar provinents d'una imatge FNA realitzades a aquest pacient. L'etiqueta representarà si la imatge presenta evidència de tumor maligne, o no.

Per tal de adaptar els datasets Iris i Vi un problema de classificació binari, en Iris considerarem l'etiqueta 1 com a 0, i.e. classificarem "Setosa" contra la resta, mentre que en el cas de Vi considerarem l'etiqueta 2 com a 0, altra vegada classificant 1 contra la resta.

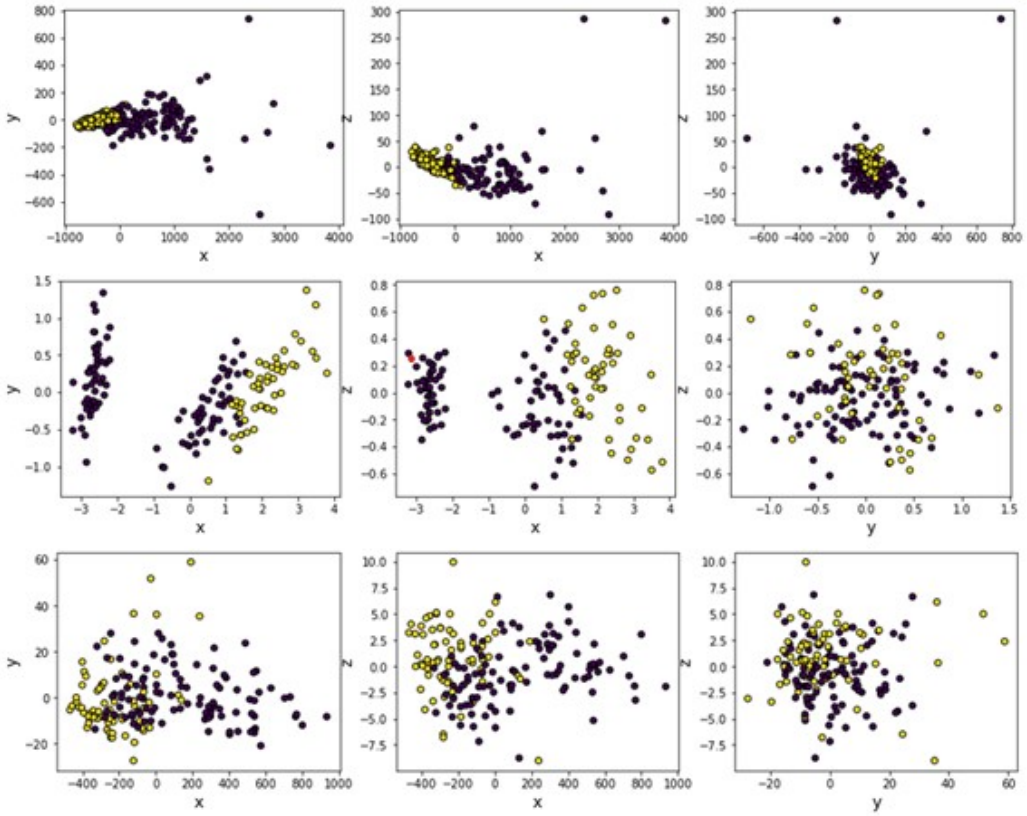


Figura 5.2: Gràfics dels conjunts de dades reals Càncer, Iris i Vi. La visualització es realitza a través d'una reducció de dimensionalitat via PCA.

5.2.0 Evaluació

Per tal d'evaluar i comparar el classificador amb el de FP, usarem els dos classificadors i el clasificador base (1-NN) sobre cada dataset a través d'un procès de $5-$*fold cross-validation*) (5-CV). El procés consistirà en separar cada conjunt de dades en 5 subconjunts disjunts i amb el mateix cardinal (cinc 5-folds). Per a cada $n = 1, ..., 5$ pendrem l' n-èsim 5-*fold* com a conjunt de test i la resta com a conjunt d'entrenament. Els classificadors seran entrenats en el conjunt d'entrenament, i prediran un conjunt d'etiquetes $\hat{Y}_i$ per cada fulla. Per a cada predicció $\hat{Y}_i$ computarem la matriu de confusió $CM(\hat{Y}_i, Y_i)$ i els valors d'exactitud, precisió, F1 i *recall* $e_i(\hat{Y}_i, Y_i)$, $pr_i(\hat{Y}_i, Y_i)$, $F1_i(\hat{Y}_i, Y_i)$, $rec_i(\hat{Y}_i, Y_i)$. Finalment, el valor mitjà sobre $i = 1, ..., 5$ serà computat per a cada conjunt de dades i cada classificadors obtenint valors que mesuraran el funcionament dels diferents classificador lliures de biaix degut a l'elecció del conjunt d'entrenament i/o test.

Noti's que en el cas del classificador FD, cal seleccionar una sèrie d'hiper-paràmetres. Seleccionarem els hiper-paràmetres òptims σ, i c a través d'un procés de *cross-validation* independent i previ a l'anterior. No farem cap selecció prèvia dels hiperparàmetres i (classe base sobre la que construim els diagrames ampliats) i k (dimensió homològica) per tal d'obtenir resultats per a tots els possibles valors d'aquest paràmetes, que seran del nostre interès en la posterior anàlisi dels resultats.

Així doncs, avaluarem un total de 7 classificadors sobre cada conjunt de dades (1-NN, FP 0 i 1 dimensional, i DF 0 i 1 dimensional per a classes base 0 i 1) 5 vegades, i d'aquestes avaluacions obtindrem les mètriques mostrades en les taules (5.1, 5.2, 5.3, 5.4).

5.3.0 Comparació

5.4.0 Discussió

En la Taula 5.5, podem veure una mitjana de les mètriques sobre les que hem avaluat els diferents clasificadors respecte als diferents conjunts de dades. A trets generals podem observar que els quatre classificadors homològics no assoleixen les mètriques del classificador base. Mentre que en dimensió 0 les mètriques obtingudes són comparables a les d'un 1-NN, per a dimensió 1 aquestes són significativament inferiors tal

	C0.1	C0.3	C0.6	M0.1	M0.3	M0.6	GQ	Càncer	Iris	Vi	Mitjana	STD
1NN	1	0.73	0.57	1	0.83	0.71	0.9	0.93	0.93	0.7	0.75	0.14
FP0	0.96	0.71	0.55	0.99	0.74	0.65	0.77	0.89	0.88	0.7	0.71	0.13
FP1	0.64	0.59	0.54	0.38	0.48	0.43	0.4	0.28	0.29	0.49	0.41	0.11
FD00	0.99	0.61	0.5	0.99	0.73	0.54	0.83	0.72	0.68	0.6	0.65	0.16
FD01	0.98	0.76	0.6	0.97	0.6	0.61	0.54	0.86	0.91	0.6	0.68	0.16
FD10	0.85	0.78	0.51	0.56	0.48	0.52	0.84	0.7	0.67	0.61	0.59	0.13
FD11	0.66	0.68	0.53	0.56	0.63	0.54	0.64	0.67	0.67	0.6	0.56	0.05

Cuadro 5.1: Exactitud mitjana de cada classificador després del 5-cv sobre els diferents conjunts de dades. FPi indica el classificador FP en dimensió i, mentre que FDij indica el classificació en dimensió i classe base j.

	C0.1	C0.3	C0.6	M0.1	M0.3	M0.6	GQ	Càncer	Iris	Vi	Mitjana	Desviació
1NN	1	0.70	0.56	1	0.79	0.71	0.93	0.92	0.95	0.72	0.829	0.145
FP0	0.96	0.77	0.55	1	0.75	0.66	0.74	0.93	0.95	0.79	0.810	0.138
FP1	0.58	0.55	0.52	0.36	0.46	0.42	0.32	0.40	0.18	0.36	0.418	0.113
DF00	1	0.54	0.48	1	0.71	0.64	0.76	0.71	0.64	0.49	0.698	0.175
DF01	0.97	0.73	0.58	0.98	0.66	0.60	0.57	0.81	0.66	0.51	0.707	0.157
DF10	1	0.78	0.69	0.60	0.72	0.51	0.82	0.70	0.39	0.34	0.655	0.190
DF11	1	0.94	0.39	0.525	0.78	0.6	0.73	0.70	0.60	0.39	0.665	0.196

Cuadro 5.2: Precisió mitjana de cada classificador després del 5-cv sobre els diferents conjunts de dades. FPi indica el classificador FP en dimensió i, mentre que FDij indica el classificació en dimensió i classe base j.

	C0.1	C0.3	C0.6	M0.1	M0.3	M0.6	GQ	Càncer	Iris	Vi	Mitjana	Desviació
1NN	1	0.72	0.58	1	0.82	0.70	0.89	0.94	0.94	0.71	0.83	0.14
FP0	0.96	0.65	0.53	0.99	0.74	0.65	0.78	0.90	0.92	0.66	0.78	0.15
FP1	0.72	0.66	0.58	0.33	0.42	0.40	0.21	0.32	0.24	0.41	0.43	0.16
DF00	0.99	0.45	0.59	0.99	0.75	0.35	0.85	0.81	0.68	0.50	0.70	0.21
DF01	0.99	0.67	0.54	0.97	0.46	0.67	0.59	0.90	0.70	0.67	0.72	0.17
DF10	0.82	0.76	0.26	0.17	0.28	0.66	0.84	0.78	0.52	0.25	0.53	0.26
DF11	0.51	0.52	0.43	0.68	0.45	0.12	0.57	0.78	0.37	0.55	0.50	0.17

Cuadro 5.3: F1 mitjana de cada classificador després del 5-cv sobre els diferents conjunts de dades. FPi indica el classificador FP en dimensió i, mentre que FDij indica el classificació en dimensió i classe base j.

i com ja apuntaven els resultats de [5]. Si comparem només els classificadors homològics veurem que mente que el nostre classificador obté resultats similars o lleugerament inferiors a FP en homologia 0, les mètriques resulten significativament superiors en homologia 1. Aquest fet reforça la idea plantejada en la secció 4.4, que mentre que

	C0.1	C0.3	C0.6	M0.1	M0.3	M0.6	GQ	Càncer	Iris	Vi	Mitjana	Desviació
1NN	1	0.76	0.64	1	0.86	0.73	0.85	0.96	0.94	0.71	0.845	0.123
FP0	0.96	0.62	0.55	0.99	0.75	0.67	0.85	0.88	0.91	0.58	0.776	0.155
FP1	1	0.87	0.68	0.31	0.41	0.40	0.16	0.27	0.35	0.48	0.494	0.257
DF00	0.98	0.42	0.77	0.99	0.84	0.28	0.97	0.94	0.85	0.57	0.760	0.240
DF01	1	0.66	0.54	0.97	0.43	0.81	0.69	1	0.85	1	0.795	0.197
DF10	0.74	0.81	0.18	0.10	0.25	0.99	0.86	0.90	0.81	0.29	0.593	0.326
DF11	0.38	0.40	0.50	1	0.38	0.07	0.49	0.93	0.27	0.93	0.533	0.297

Cuadro 5.4: *Recall* mitjana de cada classificador després del 5-cv sobre els diferents conjunts de dades. FPi indica el classificador FP en dimensió i, mentre que FDij indica el classificació en dimensió i classe base j.

	1-NN	FP0	FP1	DF0	DF1
ACC	0.839±0.135	0.798±0.139	0.448±0.114	0.774±0.137	0.619±0.112
PR	0.829 ± 0.145	0.81±0.138	0.418±0.113	0.736±0.157	0.648± 0.19
F1	0.831±0.138	0.778±0.149	0.429±0.164	0.772±0.144	0.605±0.199
REC	0.845±0.123	0.776±0.155	0.494±0.257	0.862±0.151	0.674±0.273

Cuadro 5.5: Resum de les mitjanes i desviacions estàndard de les diferents mètriques de cada classificador sobre tots els conjunts de dades. FPi indica el classificador FP en dimensió i, mentre que FDi indica el classificació en dimensió i (en aquest cas el paràmeter j indicant la classe de l'homologia base ha estat optimitzat).

	C0.1	C0.3	C0.6	M0.1	M0.3	M0.6	GQ	Càncer	Iris	Vi	Mitjana	Desviació
1NN	1	0.73	0.57	1	0.83	0.71	0.90	0.93	0.96	0.77	0.839	0.135
FP	0.97	0.75	0.56	1	0.72	0.58	0.71	0.77	0.94	0.75	0.777	0.144
%	0.68	0.58	0.63	0.37	0.50	0.46	0.41	0.29	0.27	0.43	0.463	0.129
VAR	0.008	0.042	0.013	0.010	-0.018	-0.067	-0.057	-0.121	-0.009	-0.012	-0.021	0.045
CD	1	0.86	0.63	1	0.79	0.64	0.93	0.86	0.89	0.84	0.844	0.124
%	0.83	0.76	0.51	0.55	0.56	0.31	0.78	0.77	0.59	0.36	0.602	0.171
VAR	0.020	0.078	0.027	0.010	0.059	0.026	0.091	0.004	0.133	0.228	0.068	0.066

Cuadro 5.6: Exactitud dels classificadors per a factors de certesa 1. La fila 1-NN mostra l'exactitud del classificador base. Les files FP i DF mostren les exactituds dels classificadors sobre punts amb factor de certesa 1. Les files en les files % es mostren els percentatges de punts sobre els quals el respectiu classificador té certesa 1. En les files VAR s'indica la variació de l'exactitud del classificador sobre els punts amb factor de certesa 1 respecte l'exactitud del classificador sobre tots els punts.

la hipòtesi de FP pot valdre en homologia 0, sota la casuística de l'homologia en major dimensió aquesta falla. També avala que el plantejament que hi ha darrere del

classificador DF és capaç de capturar millor aquesta casuística que FP.

Podem observar també que la taula 5.5 DF mostra valors de desviació estàndard majors que la resta, això indica el fet que DF pot no ser tan estable com el classificador base, i dependre del conjunt sobre el qual classifica. Efectivament, si mirem les mètriques sobre els conjunts concrets (taules 5.1-5.4) veiem per exemple que DF és el classificador amb millors mètriques per a Cercles (soroll 0.3) i millora significativament el resultat de FP en Quantils Gaussians. Efectivament DF capturarà especialment bé la caracterització de les diferents classes en aquells conjunts de dades on la variació homològica sigui una característica distintiva d'aquestes (en Cercles ho és, com hem vist en l'exemple 4.1, i similarment ho és en Quantils Gaussians).

Afirmem doncs que DF serà un classificador amb un comportament comparable al de 1-NN i FP en homologia 0, i significativament millor que FP en homologia 1. Obtenir un bon classificador basat en 1-homologia pot ser interessant per aplicar-lo en conjunts de dades on aquest encaixi bé (per exemple QG), però també ens pot servir com a mesura de certesa en classificació. Donat un punt a classificar, la certesa serà una mesura de com de fiable és l'etiqueta que dona un classificador. En classificadors homològics com DF o FP, una mesura de certesa pot ser la següent:

$$c(x) = \begin{cases} 0.5 & \text{si} \quad C_0(x) \neq C_1(x) \\ 1 & \text{si} \quad C_0(x) = C_1(x), \end{cases}$$

on x serà el punt a classificar, i C_0, C_1 el classificador en homologia 0 i 1 respectivament. Així tenir un classificador operatiu en diferents dimensions pot ser útil per assignar un grau de certesa major a la classificació d'un punt quan els classificadors coincideixen en les diferents dimensions. Com podem veure en la taula 5.6 si estudiem les mètriques de classificació de DF en els punts on la certesa és 1, aquestes augmenten considerablement, superant fins i tot el classificador base. Per a FP aquesta mesura de certesa no funciona a causa del limitat poder predictiu en dimensió 1.

6.0 Conclusions

En aquest treball s'usen les eines de TDA per a definir un classificador basat purament en homologia. El classificador s'avalua en 9 conjunts de dades diferents per evaluar el seu comportament respecte al classificador topològic de característiques similars presentat en [5].

La proposta basada en estimar la distribució de les variacions dels diagrames de persistència d'una classe respecte a l'addició de punts, i donat un punt no etiquetat inferir de quina distribució prové obté resultats similars al classificador base i a [5] en 0-homologia, i millorem substancialment el resultats dels classificadors 1-homològics existents. El classificador resultarà d'especial interès quan una classe tingui diagrames variacionals molt diferenciats d'alters classes (per exemple Cercles o GQ).

Finalment més enllà de l'interès d'obtenir un classificador 1-homològic amb bon comportament, veiem que la combinació de CD_0 i CD_1 ens dona una mesura de certesa en la classificació prou rellevant.

References

Bibliografía

[1] Bubenik, P. Statistical topological data analysis using persistence landscapes.

[2] Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., and Wasserman, L. Stochastic convergence of persistence landscapes and silhouettes. SOCG'14, Association for Computing Machinery, p. 474–483.

[3] Edelsbrunner, H., and Harer, J. *Computational Topology - an Introduction.* American Mathematical Society, 2010.

[4] Edelsbrunner, H., Letscher, D., and Zomorodian, A. Topological persistence and simplification. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science* (USA, 2000), FOCS '00, IEEE Computer Society, p. 454.

[5] Fisas, F. P. N. The brain network of motivation: A topological approach. Master's thesis, Facultat de Matemàtiques i Informàtica, 2021.

[6] Ghrist, R. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society 45* (2007), 61–75.

[7] I.R.Goodman, Ronald. P.S. Mahler, H. T. N. *Mathematics of Data fusion*, vol. 37.

[8] Kindelan, R., Frías, J., Cerda, M., and Hitschfeld, N. Classification based on topological data analysis, 2021.

[9] Maroulas, V., Mike, J. L., and Oballe, C. Nonparametric estimation of probability density functions of random persistence diagrams. *Journal of Machine Learning Research 20*, 151 (2019), 1–49.

[10] Maroulas, V., Nasrin, F., and Oballe, C. A bayesian framework for persistent homology. *SIAM Journal on Mathematics of Data Science 2*, 1 (2020), 48–74.