Un classificateur non-supervisé utilisant les complexes simpliciaux (avec une application à la stylométrie).

Louis Hauseux, Bartlomiej Blaszczyszyn

Un classificateur non-supervisé utilisant les *complexes simpliciaux* (avec une application à la *stylométrie*).

Louis Hauseux
Encadré par Bartlomiej Blaszczyszyn

28 septembre 2017

Avant-propos

Nous nous proposons au cours des quelques pages de ce rapport de présenter au lecteur ce que sont les *complexes simpliciaux* ainsi qu'une de leurs possibles (et nombreuses !) applications : en classification non-supervisée.

Les complexes simpliciaux peuvent s'appréhender comme une généralisation des *graphes* ; un graphe étant la donnée d'un ensemble de sommets ainsi que d'une relation de voisinage entre des paires de ces sommets (deux points sont voisins si une arête les relie). Les complexes simpliciaux permettent de rendre compte de relations de voisinage plus élaborées (et faisant notamment intervenir un nombre arbitraire de points ; pas seulement deux).

La classification non supervisée est une branche du vaste domaine de l'*apprentissage automatique*. Étant donné un échantillon de *données* (le plus souvent des points de l'espace euclidien $\mathbb{R}^d$), elle consiste à regrouper ces données en différentes *classes* de sorte que les données d'une même classe présentent des *similarités* entre elles tandis que deux données appartenant à deux classes distinctes soient *dissemblables*.

Le présent rapport s'articulera donc en deux parties :

- la première introduira au lecteur non forcément familier cette notion de complexe simplicial d'un point de vue théorique. On l'illustrera ensuite avec la présentation des *complexes de Čech* et certaines propriétés mathématiques qui en font un outil puissant et pratique (la *théorie de Morse* permet, par exemple, de manier ces complexes de différentes façons). On verra encore quelques résultats des complexes simpliciaux *aléatoires* (c'est-à-dire que les sommets sont des points générés aléatoirement) dans le cas des régimes dits *surcritiques* justifiant certains algorithmes d'apprentissage de variétés (une des multiples applications promises des complexes simpliciaux). Enfin, nous présenterons très succinctement l'*homologie persistante*...

- ... homologie persistante qui est à l'origine (mais en l'inversant) de l'idée du classificateur non-supervisé présenté dans cette seconde partie. Ce classificateur a pour but d'isoler les « pics de densité » (c'est-à-dire les endroits où les données sont les plus *denses*). Après avoir essayé de justifier auprès du lecteur cet objectif en expliquant pourquoi il permettrait une bonne classification, nous tenterons de justifier l'algorithme aussi d'un point de vue mathématique en nous penchant sur certaines de ses propriétés. Finalement, nous montrerons quelques résultats concrets de ce classificateur quand on l'applique à identifier les différents auteurs d'un corpus de texte (ce domaine de la classification s'appelle la *stylométrie*)

Sommaire

1 Introduction

Voici une description plus précise de la manière dont va se dérouler ce rapport.

Nous avons déjà dit que la première partie allait porter sur les complexes simpliciaux. Pour aborder cette notion, nous commençons par définir les complexes simpliciaux *géométriques* qui sont des collections de *simplexes* (ensemble convexe engendré par une famille libre de points de l'espace) de sorte que toute *face* de l'un de ces simplexes (ensemble convexe engendré par un sous-ensemble de la famille de points) soit encore un simplexe de la collection. De même, l'intersection de deux simplexes est encore un simplexe.
On peut aussi considérer les complexes d'un point de vue combinatoire où les simplexes ne sont plus repérés que par leurs sommets (pour une certaine *orientation*) ; ce sont les complexes simpliciaux *abstraits*.
Nous verrons alors deux exemples de complexes géométriques : les complexes de Čech et ceux de Vietoris-Rips. Les premiers possèdent un grand nombre de propriétés mathématiques sur lesquelles nous nous pencherons plus ensuite. Les seconds sont parfois appréciés pour leur facilité de mise en œuvre dans des programmes informatiques.
Suit un paragraphe sur des rappels d'algèbre permettant d'introduire les notions de *groupes d'homologie* et notamment du cardinal de leur *groupe libre* : les *nombres de Betti*.
Enfin, nous approfondirons quelque peu sur les complexes de Čech. On peut les construire de façon purement abstraite en les voyant comme le *nerf du recouvrement* d'un voisinage de ses sommets ; on bénéficie alors du *théorème du nerf d'un recouvrement* affirmant que ce voisinage et le complexe sont *homotopiquement équivalents* (et ont donc les mêmes groupes d'homologie). La *théorie de Morse* est un autre outil pratique permettant de considérer les groupes d'homologie d'un complexe de Čech grâce au recours d'une certaine *fonction de Morse*, de ses *points critiques* et de leur *niveau*. Au total, nous disposons donc de trois objets distincts qu'on peut manipuler au choix selon le contexte ; cela facilite grandement un certain nombre de preuves.
À ce moment, nous verrons la notion de complexe simplicial *aléatoire* : les sommets sont tirés identiquement et indépendamment sur une variété. Quand on possède un échantillon en nombre suffisant, on peut alors recouvrer l'homologie de la variété. Cette propriété justifie certains algorithmes de reconnaissance de variétés (par exemple en imagerie médicale).
Pour finir cette partie, nous aborderons l'idée à l'origine de l'*homologie persistante* : considérer le « temps de vie » de chaque *cycle*.

La deuxième partie sera consacrée à la présentation d'un algorithme de classification non-supervisée à l'aide des complexes simpliciaux aléatoires. Ici nous supposerons que nos données sont des points de $\mathbb{R}^d$, la ressemblance entre les points étant fournie par la distance euclidienne $d(x, y) = ||x - y||$ (deux données sont d'autant plus semblables qu'elles sont proches spatialement). Nous ferons de plus l'hypothèse que notre échantillon est composé de n points tirés *indépendamment* selon une même loi de probabilité admettant une fonction de densité (pour la mesure de Lebesgue) f à support compact.
Nous postulons alors qu'une bonne classification est une classification qui isole les « *pics de densité* » (nous tenterons tout de même – autant que faire se peut – de justifier ce postulat auprès du lecteur incrédule).
Tout l'enjeu est alors de bien distinguer ces « pics de densité ». Pour ce faire, nous requerrons à quelques notions de *topologie* ; la topologie étant justement ce domaine des mathématiques qui fournit des informations qualitatives sur la structure géométrique de nos données.
La suite est dédiée à montrer quelques-unes des propriétés *asymptotiques* (quand $n \to \infty$) de l'algorithme et notamment le bon recouvrement des « pics de densité ».
Pour terminer ce rapport, nous présenterons un cas pratique : une application au domaine de la *stylométrie*.

2 Les complexes simpliciaux

2.1 Introduction aux complexes simpliciaux

L'approche la plus intuitive pour appréhender les *complexes simpliciaux* est celle géométrique ; ils sont alors définis sur la classe des *polyèdres*. Les polyèdres de base étant les *points*, les *segments*, les *triangles*, les *tétraèdres*, etc... et leurs équivalents de plus haute dimension. On peut construire un nouveau polyèdre à partir de deux autres polyèdres partageant une même *facette* en "recollant" ces derniers sur celle-ci.

Formalisons un petit peu tout cela !

2.1.1 Simplexes

Soit $\{s_0, \ ... \ , \ s_k\}$ une famille de $k+1$ points de $\mathbb{R}^N$ *géométriquement indépendants*, c'est-à-dire que la famille

$$s_1 - s_0, \ ... \ , \ s_k - s_0$$

est *(linéairement) libre*.

Définition . *Le simplexe σ de dimension k engendré par la famille $\{s_0, \ ... \ , \ s_k\}$ est l'ensemble :*

$$\sigma \ = \ Convexe\,(s_0, \ ... \ , \ s_k) \ = \ \left\{ x \in \mathbb{R}^N \ \middle| \ x = \sum_{i=0}^{k} t_i s_i \ , \ t_i \in [0;1] \ , \ \sum_{i=0}^{k} t_i = 1 \right\}.$$

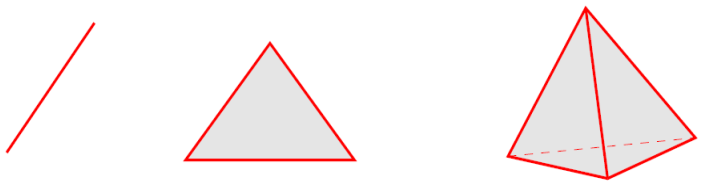


FIGURE 1 – Un segment, un triangle et un tétraèdre sont trois *simplexes* de dimensions respectives 1, 2 et 3. (source : [5])

Désignons par $S(\sigma)$ l'ensemble des *sommets* de σ. En reprenant les notations précédentes, on aurait donc : $S(\sigma) = \{s_0, \ \dots \ , \ s_k\}$.

Tout sous-ensemble $S' \subseteq S(\sigma)$ non vide de l'ensemble des sommets permet de définir un nouveau simplexe $\sigma' = Convexe(S') \subseteq \sigma$ appelé *face* de σ.

Ainsi, tout simplexe de dimension k possède exactement C_{k+1}^{i+1} faces de dimension i (pour $i \in \{0, \ \dots \ , k\}$) soit $2^{k+1} - 1$ faces en tout.

Remarque . *Il nous arrivera d'utiliser le terme de facette pour désigner une face de dimension exactement $k-1$.*

2.1.2 Complexes simpliciaux *géométriques*

Définition . *Une collection K de simplexes de $\mathbb{R}^N$ est un complexe simplicial si elle vérifie :*

- *toute face $\tau \subseteq \sigma$ d'un simplexe $\sigma \in K$ est elle-même dans K ($\tau \in K$) et*
- *l'intersection $\sigma \cap \tau$ de deux simplexes de K est soit vide soit une face de chacun des deux simplexes $\sigma, \tau \in K$.*

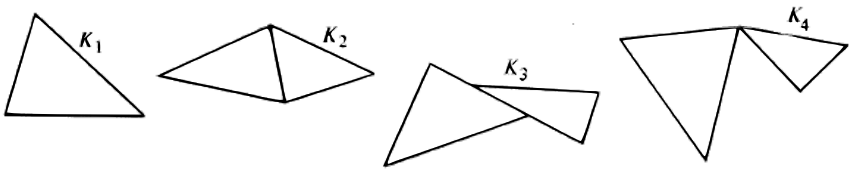


FIGURE 2 – K_1, K_2 et K_4 sont bien des *complexes simpliciaux*. K_3 en revanche ne vérifie pas la seconde propriété. (source : [14])

La *dimension* du complexe simplicial K sera la dimension *maximum* des dimensions de ses simplexes.

Un sous-ensemble $K' \subseteq K$ de K qui est lui-même un complexe simplicial est appelé un *sous-complexe* de K.

Exemples . *L'exemple le plus simple de complexe simplicial est la collection constituée de toutes les faces d'un simplexe.*

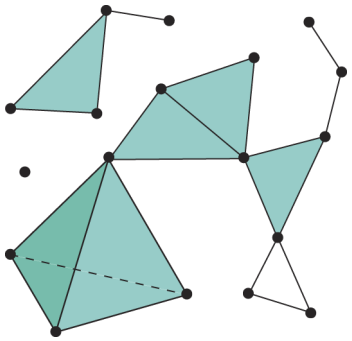


FIGURE 3 – Un exemple d'un *complexe simplicial*. (source : [17])

2.1.3 Complexes simpliciaux *abstraits*

Après avoir vu une approche géométrique des complexes simpliciaux, voyons à présent une approche combinatoire.

Tout simplexe est entièrement déterminé par ses sommets. Ainsi, on peut représenter un simplexe simplement par l'ensemble de ses sommets... ce qui conduit à la définition suivante.

Définition . *Un complexe simplicial abstrait $\mathcal{K}$ est une collection d'ensembles finis non vides vérifiant :*

$$\forall\ \sigma \in \mathcal{K},\ \forall\ \tau \subseteq \sigma,\ \tau \neq \emptyset \implies \tau \in \mathcal{K}.$$

Les éléments de $\mathcal{K}$ sont appelés ses *simplexes* (ou ses *faces*).

Étant donné un simplexe σ, sa *dimension* est égale à : $dim(\sigma) = \#\sigma - 1$ (où $\#E$ désigne le cardinal de E).

Toutes les définitions précédemment introduites s'étendent naturellement aux complexes abstraits.

Remarque . *Tout complexe géométrique admet une représentation abstraite ; soit K un complexe géométrique, il suffit de considérer la collection $\mathcal{K}$ constituée des ensembles $\{s_0,\ \dots\ , s_k\}$ pour tout simplexe $\sigma \in K$ de sommets $S(\sigma) = \{s_0,\ \dots\ , s_k\}$.*

La réciproque n'est pas vraie, du moins pas dans $\mathbb{R}^N$ car la dimension d'un complexe abstrait n'est pas forcément bornée par N. Cependant, on peut alors travailler sur un espace plus grand : soit J un ensemble quelconque (au moins aussi grand que l'espace des sommets), on peut considérer $\mathbb{E}^J$ le sous-ensemble de $\mathbb{R}^J$ des éléments x n'ayant que des coordonnées nulles sauf sur un sous-ensemble fini de J. Cet espace étant hilbertien pour le produit scalaire :

$$< x; y >:= \sum_{j \in J} x_j y_j.$$

2.1.4 Complexes de *Čech* et complexes de *Vietoris-Rips*

Nous allons voir deux construction possibles de complexes simpliciaux abstraits à partir de points de l'espace $\mathbb{R}^d$; les *complexes de Čech* et ceux de *Vietoris-Rips*.

Remarque . *Le complexe de Čech $\check{C}(\mathcal{P}, r)$ d'un ensemble de sommets $\mathcal{P}$ et de rayon r est celui qui présente le plus d'intérêt du point de vue mathématique ; en effet, il est défini comme le nerf du recouvrement du voisinage des points de $\mathcal{P}$ par des boules de rayon r centrées en ces derniers. Cette propriété permet notamment de dire qu'il est homotopiquement équivalent au voisinage des points (l'union de boules centrées en ces points). Aussi, nous travaillerons dans la suite exclusivement avec ces complexes de Čech. Nous présentons néanmoins les complexes de Vietoris-Rips qui se rencontrent assez fréquemment pour des raisons de facilité pour les calculs numériques.*

Définitions . *Soit $\mathcal{P} = \{x_1 \ , \ ... \ , \ x_n\}$ un nuage de points.*

- *Le complexe de Čech $\check{C}(\mathcal{P}, r)$ de sommets $\mathcal{P}$ et de rayon r est le complexe simplicial abstrait dont les sommets sont les éléments de $\mathcal{P}$ et dont le simplexe de dimension k $[x_{i_0} \ , \ ... \ , \ x_{i_k}]$ (avec $i_0 < ... < i_k$) est dans $\check{C}(\mathcal{P}, r)$ si :*

$$\bigcap_{j=0}^{k} B_r\left(x_{i_j}\right) \neq \emptyset.$$

- *Le complexe de Vietoris-Rips $\check{R}(\mathcal{P}, r)$ de sommets $\mathcal{P}$ et de rayon r est le complexe simplicial abstrait dont les sommets sont les éléments de $\mathcal{P}$ et dont le simplexe de dimension k $[x_{i_0} \ , \ ... \ , \ x_{i_k}]$ (avec $i_0 < ... < i_k$) est dans $\check{C}(\mathcal{P}, r)$ si :*

$$B_r\left(x_{i_\alpha}\right) \cap B_r\left(x_{i_\beta}\right) \neq \emptyset \text{ pour tout } \alpha, \beta \in \{0, ..., k\}.$$

2.2 Groupes d'homologies

2.2.1 Rappels d'algèbre

G désignera dans toute la suite de cette section un groupe *abélien* (*commutatif*).

G est dit *libre* s'il possède une *base* $B = \{g_\alpha\}_\alpha \subseteq G$, c'est-à-dire un sous-ensemble de G tel que tout élément $g \in G$ ait une *unique* décomposition dans cette base :

$$g = \sum_\alpha n_\alpha g_\alpha.$$

L'unicité entraîne que tout élément $g_\alpha \in B$ est d'ordre *infini*.

Si une famille B permet de décomposer – non forcément de façon unique – tout élément $g \in G$, on dira qu'elle *engendre* G (ou que B est une famille *génératrice* de G).

L'ensemble T de tous les éléments $g \in G$ d'ordre fini (il existe $n > 0$ tel que : $ng = 0$) est un groupe appelé *groupe de torsion*. Si $T = \{0\}$, G est dit *sans torsion*.

Remarque . *Un groupe abélien libre est sans torsion. La réciproque est fausse !*

Théorème . *(dit théorème fondamental des groupes abéliens de type fini)*
Soit G un groupe abélien de type fini (engendré par une famille finie) et soit T son groupe de torsion.

- *G admet un sous-groupe libre $H \subseteq G$ d'ordre fini β tel que : $G = H \oplus T$ (H et T sont en somme directe)*
- *Il existe des sous-groupes $T_1, \ ... \ , \ T_k$ cycliques de T d'ordre respectif $t_i > 1$ tels que $t_1 \mid t_2 \mid ... \mid t_k$ et*
$$T = \bigoplus_{i=1}^{k} T_i$$
- *les nombres $\beta, \ t_1, \ ... \ , \ t_k$ sont déterminés de manière unique*

Définition . *β est le nombre de Betti de G ; $t_1, \ ... \ , \ t_k$ sont ses coefficients de torsion.*

Remarque . *H est isomorphe au quotient G/T.*

2.2.2 Groupes d'homologie et *nombres de Betti*

Après tous ces préliminaires, nous allons enfin pouvoir définir les *groupes d'homologie*. Pour cela, nous aurons d'abord besoin d'*orienter* les simplexes σ.

Dotons-nous d'un ordre total sur les sommets. On dispose d'une relation d'équivalence sur les ordres : deux ordres sont équivalents si l'on passe de l'un à l'autre par une permutation de *signature* positive (à l'aide d'un nombre *pair* de transpositions). Il en résulte deux classes d'équivalence (ou deux *orientations*). Soit alors σ un simplexe de sommets $S(\sigma) = \{s_0, \ ... \ , \ s_k\}$ avec $s_0 < ... < s_k$.

Définition . *On désignera ainsi le simplexe orienté (c'est-à-dire le simplexe muni de son orientation) :*
$$[s_0, \ ... \ , s_k].$$

Remarque . *la notation n'est pas unique puisqu'elle dépend du représentant de l'orientation. Toutefois, toutes les définitions et propriétés qui vont suivre ne dépendent, elles, que de l'orientation (et non pas de l'ordre choisi comme représentant de cette orientation).*

Soit K un complexe simplicial, une *p-chaîne* est une fonction des simplexes orientés de K dans $\mathbb{Z}$ vérifiant :

- $c(\sigma) = -c(\sigma')$ si σ et σ' représentent le même simplexe mais munis d'une orientation opposée.

- $c(\sigma) = 0$ sauf éventuellement pour un nombre fini de simplexes orientés de dimension p.

L'ensemble des p-chaînes peut naturellement être muni d'une structure additive (on additionne des fonctions) faisant de ce dernier un groupe abélien libre.

Définition . *On notera $C_p(K)$ le groupe des p-chaînes de K.*

Remarque . *$C_p(K)$ admet pour base l'ensemble des p-chaînes élémentaires $\{c_\sigma\}_\sigma$ qui valent 1 en σ, -1 en σ' (le même simplexe d'orientation opposée), 0 ailleurs et σ parcourt l'ensemble des simplexes de dimension p pour une certaine orientation choisie.*

Définition . *On peut maintenant définir l'opérateur du bord ∂_p qui est un homomorphisme de $C_p(K)$ dans $C_{p-1}(K)$:*

$$\partial_d[s_0, \ \dots \ , s_p] = \sum_{i=0}^{p} (-1)^i [s_0, \ \dots \ , s_{i-1}, s_{i+1}, \ \dots \ , s_p]$$

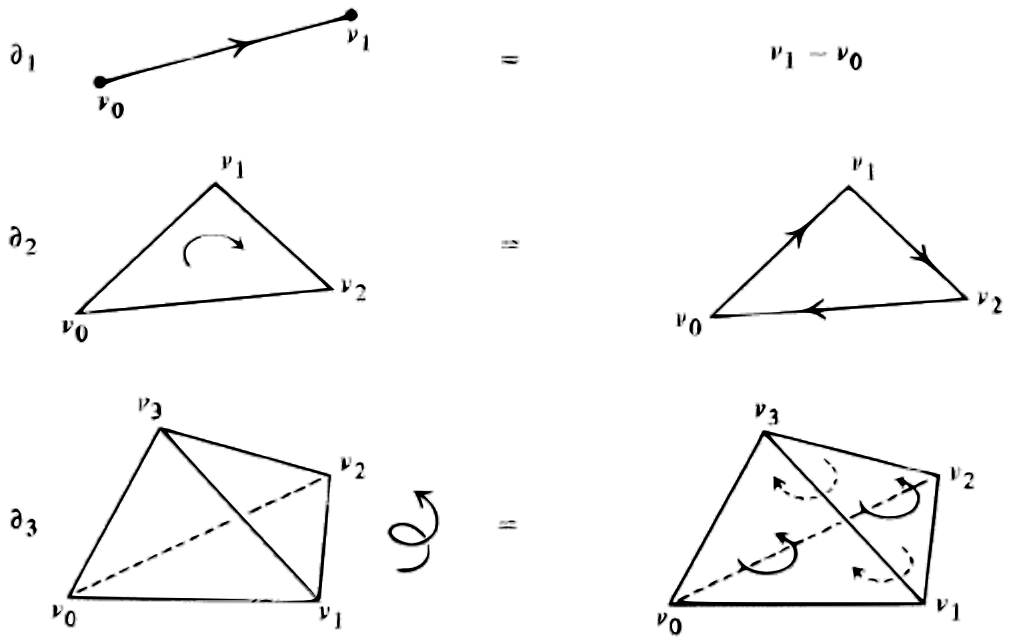


FIGURE 4 – Calcul du bord de trois simplexes orientés : un segment, un triangle et un tétraèdre. (source : [14])

Proposition .

$$\partial_p \circ \partial_{p-1} = 0$$

Définitions . *Le noyau de $\partial_p : C_p(K) \to C_{p-1}(K)$, noté $Z_p(K)$, est appelé groupe des p-cycles de K.*
L'image de $\partial_{p+1} : C_{p+1}(K) \to C_p(K)$, noté $B_p(K)$, est appelé groupe des p-bords (ou p-frontières) de K.
Grâce à la proposition précédente, on sait que $B_p(K) \subseteq Z_p(K)$ est un sous-groupe de $Z_p(K)$, on peut donc définir le p-ième groupe d'homologie de K :

$$H_p(K) := Z_p(K)/B_p(K).$$

On notera β_p son nombre de Betti.

2.3 Sur les complexes géométriques de Čech

2.3.1 Rappel sur les complexes de Čech

Définition . *Soit $\mathcal{P} = \{x_1, x_2, \ ...\}$ une collection de points de $\mathbb{R}^d$ et $\epsilon > 0$, le complexe de Čech $\check{C}(\mathcal{P}, \epsilon)$ formé par les points de $\mathcal{P}$ et de rayon ϵ est le complexe simplicial abstrait dont :*

- *les sommets sont les points de $\mathcal{P}$,*
- *le simplexe $\{x_{i_0}, \ ... \ , x_{i_n}\}$ est dans $\check{C}(\mathcal{P}, r)$ si*

$$\bigcap_{k=0}^{n} B_\epsilon\left(x_{i_k}\right) \neq \emptyset.$$

2.3.2 Théorème du nerf d'un recouvrement

Comme dans la section précédente, on considère $\mathcal{P}$ un nuage de points de $\mathbb{R}^d$ et $\epsilon > 0$.

Définition . *Le voisinage de $\mathcal{P}$ de rayon ϵ est l'ensemble :*

$$\mathcal{U}\left(\mathcal{P}, \epsilon\right) = \bigcup_{p \in \mathcal{P}} B_\epsilon\left(p\right).$$

Définition . *Étant donné un recouvrement ouvert $\{U_j\}_{j \in J}$ d'un espace topologique E ($E = \bigcup_{j \in J} U_j$), son nerf $\mathcal{N}(U_j)$ est le complexe simplicial abstrait formé de l'ensemble des parties finies non vides $I \subseteq J$ telles que :*

$$\bigcap_{i \in I} U_i \neq \emptyset.$$

Remarque . *Le complexe de Čech $\check{C}(\mathcal{P}, \epsilon)$ – précédemment introduit – correspond au nerf du recouvrement $\{B_\epsilon(p)\}_{p \in \mathcal{P}}$ de $\mathcal{U}\left(\mathcal{P}, \epsilon\right)$*

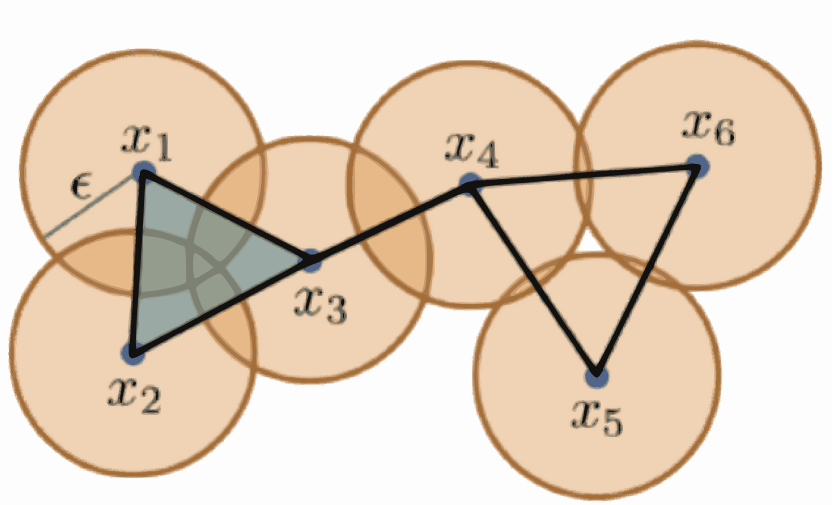


FIGURE 5 – Un exemple de complexe de Čech dans $\mathbb{R}^2$ comportant 6 sommets, 7 arêtes et 1 triangle. (source : [3])

De plus, on dispose du théorème suivant :

Théorème . *(Nerf d'un recouvrement).*
Soit un recouvrement ouvert $\{U_j\}_{j\in J}$ d'un espace topologique E. Supposons que ce recouvrement est localement fini, et que les intersections sont toutes contractiles (de même type d'homotopie qu'un point), alors E et le nerf de son recouvrement $\mathcal{N}(U_j)$ ont le même type d'homotopie (ou sont homotopiquement équivalents).

Corollaire . *$\check{C}(\mathcal{P},\epsilon)$ et $\mathcal{U}(\mathcal{P},\epsilon)$ ont les mêmes groupes d'homologie.*

2.3.3 Théorie de Morse

Nous venons de voir que si l'on se concentre sur les groupes d'homologie, on peut travailler indifféremment avec le complexe abstrait $\check{C}(\mathcal{P},\epsilon)$ ou l'objet géométrique $\mathcal{U}(\mathcal{P},\epsilon)$. Nous allons à présent voir qu'il y a une troisième façon possible pour aborder le calcul des groupes d'homologie ; c'est la *théorie de Morse* qui le permet.

Soit $\mathcal{M} \subset \mathbb{R}^d$ un sous-ensemble compact de $\mathbb{R}^d$. $\mathcal{M}$ est une *variété de dimension $m < d$ sans bord* si tout point $p \in \mathcal{M}$ admet comme voisinage dans $\mathcal{M}$ une boule ouverte (topologique) de dimension m.

Supposons que l'on dispose d'une *fonction de Morse*, c'est-à-dire une fonction $f : \mathcal{M} \to \mathbb{R}$ vérifiant certaines propriétés et possédant notamment des *points critiques* $c \in \mathcal{M}$ associés à des *niveaux critiques* $\rho \in \mathbb{R}$ (que le lecteur ne soit pas effrayé ; nous définirons plus précisément sous peu). Pour tout $r \in \mathbb{R}$, on

peut considérer le sous-ensemble de $\mathcal{M}$:

$$\mathcal{M}_r := f^{-1}(]-\infty; r]) = \{x \in \mathcal{M} \mid f(x) \leq r\} \subseteq \mathcal{M}.$$

La théorie de Morse permet de dire que s'il n'existe pas de niveaux critiques $\rho \in]a; b]$, alors $\mathcal{M}_a$ et $\mathcal{M}_b$ ont le même type d'homotopie et, en particulier, ont les mêmes groupes d'homologie. Par contre, si $\rho = f(c)$ est un niveau critique et c un point critique de Morse d'indice k, les groupes d'homologie de $\mathcal{M}_r$ changent en ρ :

- soit un nouveau *cycle* de dimension k apparaît, faisant croître β_k (le nombre de Betti de H_k) de 1
- soit B_{k-1} s'enrichit d'un nouveau *bord*, diminuant le nombre de Betti β_{k-1} de $H_{k-1} = Z_{k-1}/B_{k-1}$ de 1

Fonction de Morse $f : \mathbb{R}^d \to \mathbb{R}$

Nous disions naguère que pour être dite « de Morse », une fonction devait vérifier certaines conditions ; les voici.

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^2. Un point $c \in \mathbb{R}^d$ est appelé *point critique* de f si $\nabla f(c) = 0$ et $f(c)$ est sa *valeur critique*. Un point critique est qualifié de *non dégénéré* si la Hessienne $H_f(c)$ est *inversible*. En ce cas, l'*indice de Morse* de f en c est le nombre de valeurs propres négatives comptées avec multiplicité de $H_f(c)$.

Définition . *Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction de classe C^2, f est une fonction de Morse si tous ses points critiques sont non dégénérés et tous de niveau distinct.*

Relation entre $\mathcal{M}_r$ et $\mathcal{U}(\mathcal{P}, r)$

Le lien entre les fonction de Morse et l'homologie du complexe de Čech $\check{C}(\mathcal{P}, r)$ n'est pas encore apparent ; il s'agit de choisir judicieusement f.

Prenons pour f, la fonction $d_{\mathcal{P}}$ de distance à notre nuage de points $\mathcal{P}$:

$$d_{\mathcal{P}}(x) := \min_{p \in \mathcal{P}} ||x - p||.$$

Observons qu'on a alors :

$$\mathcal{M}_r = d_{\mathcal{P}}^{-1}(]-\infty; r]) = \{x \in \mathbb{R}^d \mid d_{\mathcal{P}}(x) \leq r\} = \mathcal{U}(\mathcal{P}, r)$$

En additionnant à cela que, en vertu du théorème du nerf d'un recouvrement, le complexe de Čech $\check{C}(\mathcal{P}, r)$ et $\mathcal{U}(\mathcal{P}, r)$ ont même type d'homologie, il suffit d'étudier le type d'homologie des niveaux $\mathcal{M}_r$ de $d_{\mathcal{P}}$ pour connaître celui de $\check{C}(\mathcal{P}, r)$. Pour étudier les groupes d'homologie, on peut donc travailler indifféremment avec $\mathcal{M}_r$, $\mathcal{U}(\mathcal{P}, r)$ ou $\check{C}(\mathcal{P}, r)$; par exemple, pour certaines preuves combinatoires, il est plus facile de regarder $\check{C}(\mathcal{P}, r)$.

Points critiques de $d_{\mathcal{P}}$

Toutefois, la théorie de Morse telle que nous l'avons présentée ne s'applique qu'à des fonctions de classe C^2... ce qui n'est pas le cas de $d_{\mathcal{P}}$. Intuitivement, on sent bien qu'il suffirait de lisser très légèrement la fonction $d_{\mathcal{P}}$ pour lui appliquer la théorie de Morse. Nous allons procéder autrement et proposer une définition alternative des points critiques pour le cas de la fonction $d_{\mathcal{P}}$.

Définition . *Les points critiques de $d_{\mathcal{P}}$ d'indice 0 sont les points $p \in \mathcal{P}$.*
Un point $c \in \mathbb{R}^d$ est un point critique d'indice $k \in \{1, \ ... \ , d\}$ s'il existe un ensemble $\mathcal{Y} \subseteq \mathcal{P}$ de $k+1$ points de $\mathcal{P}$ vérifiant :

- $\forall\ y \in \mathcal{Y},\ d_{\mathcal{P}}(c) = ||c - y||$ *et* $\forall\ p \in \mathcal{P} \backslash \mathcal{Y},\ d_{\mathcal{P}}(c) < ||c - p||$
- *les points de $\mathcal{Y}$ sont géométriquement libres*
- $c \in \overset{\circ}{Convexe}(\mathcal{Y})$ *appartient à l'intérieur de l'enveloppe convexe de $\mathcal{Y}$*

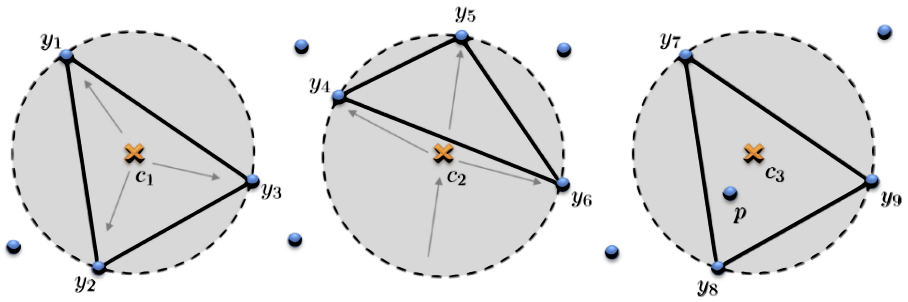


FIGURE 6 – Exemples des conditions pour être un point critique. Seul le point c_1 vérifie les trois conditions (il est d'indice 2). (source : [3])

Travaillons maintenant avec un rayon et un nuage de points qui soient fonction de n : $r = r_n \underset{n\to\infty}{\to} 0$ et $\mathcal{P} = \mathcal{P}_n$ avec $\#\mathcal{P}_n = n \to \infty$.

Soit $\beta_{k,n}$ le nombre de Betti de $\mathcal{U}(\mathcal{P}_n, r_n)$:

$$\beta_{k,n} := \beta_k\left(\mathcal{U}(\mathcal{P}_n, r_n)\right) = \beta_k\left(\check{C}(\mathcal{P}_n, r_n)\right) = \beta_k\left(\mathcal{M}_{r_n}\right)$$

Soit $\mathcal{C}_{k,n}$ l'ensemble des points critiques de $\mathcal{U}(\mathcal{P}_n, r_n)$ d'indice k et $\mathcal{C}^L_{k,n}$ les points de $\mathcal{C}_{k,n}$ à proximité de $\mathcal{P}_n$

$$\mathcal{C}^L_{k,n} := \{c \in \mathcal{C}_{k,n} \ \mid d_{\mathcal{P}_n}(c) < r_n\} = \mathcal{C}_{k,n} \cap \mathcal{U}(\mathcal{P}_n, r_n)$$

et $N_{k,n}$ son cardinal

$$N_{k,n} := \#\mathcal{C}^L_{k,n}.$$

Les deux ensembles $\{\beta_{k,n}\}_{k=0}^d$ et $\{N_{k,n}\}_{k=0}^d$ ont des comportements très semblables.

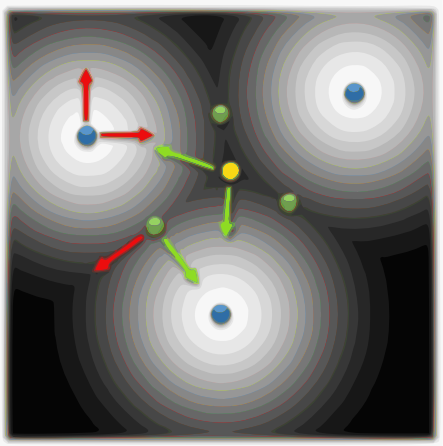


FIGURE 7 – Exemples de points critiques. Les points bleus sont d'indice 0, les verts d'indice 1 et le jaune d'indice 2. (source : [1])

2.3.4 Propriétés sur les *nombres de Betti* dans le cas du régime *sur-critique*

Dans toute cette section, nous supposerons que les points de $\mathcal{P}_n$ sont tirés indépendamment selon une mesure de probabilité sur une variété compacte $\mathcal{M} \subset \mathbb{R}^d$ de dimension m, sans bord et admettant une densité f sur $\mathcal{M}$ par rapport à la mesure de Lebesgue.

La quantité à regarder est : nr_n^m. Selon que cette quantité tende vers 0, $\lambda \in]0;\infty[$ ou que $nr_n^m \to \infty$, les groupes d'homologies connaissent des comportements tout à fait distincts qualifiés respectivement de régimes *sous-critique*, *critique* (ou *thermodynamique*) et *sur-critique*.

Dans la suite, nous porterons notre attention exclusivement sur le régime *sur-critique* ; c'est le régime qui nous a paru le plus intéressant (sans vouloir dénigrer les autres cas de figure !) puisqu'il possède certaines propriétés légitimant (d'un point de vue mathématique) des programmes de reconnaissances de variétés (nous en reparlerons d'ailleurs brièvement).

Dans notre cas (celui du régime sur-critique), nous supposerons aussi que

$$f_{min} := \inf_{x \in \mathcal{M}} f(x) > 0.$$

Nous disions donc que le régime sur-critique est le plus intéressant ; nous allons voir dans la section suivante qu'avec une hypothèse un peu plus puissante, le complexe de Čech $\check{C}(\mathcal{P}_n, r_n)$ a asymptotiquement le même type d'homologie que la variété $\mathcal{M}$ elle-même !

Ne nous contentons donc pas simplement de l'hypothèse $nr_n^m \to \infty$... mais demandons un peu plus : si $nr_n^m \geq C \log(n)$, alors le voisinage de $\mathcal{P}_n$ va recouvrir toute la variété $\mathcal{M}$.

Proposition . *Si* $C > \frac{2}{\omega_m f_{min}}$*, alors il existe une variable aléatoire* $M > 0$ *p.s. finie telle que :*

$$\forall\ n > M,\ \mathcal{M} \subseteq \mathcal{U}(\mathcal{P}_n, r_n).$$

Où ω_m *est la valeur du volume de la boule unité dans* $\mathbb{R}^m$*.*

(cf. [3] pour la démonstration de cette proposition, du théorème suivant ainsi que de beaucoup d'autres résultats)

2.3.5 Apprentissage de la topologie d'une variété $\mathcal{M}$

Étant donné un échantillonnage $\mathcal{P}$ d'une variété, peut-on retrouver sa structure topologique ?
Un début de réponse est donné par le résultat suivant ; sous les hypothèses précédentes, on peut retrouver ses nombres de Betti.

Théorème . *Avec les mêmes hypothèses que la proposition précédente, si* $C > \frac{2}{\omega_m f_{min}}$*, alors il existe une variable aléatoire* $M > 0$ *p.s. finie telle que :*

$$\forall\ n > M,\ \forall\ k \in \mathbb{N},\ \beta_{k,n} = \beta_k(\mathcal{M}).$$

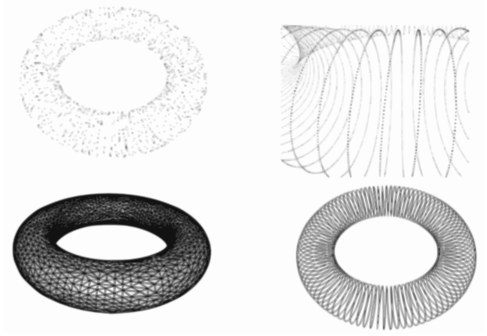


FIGURE 8 – Deux exemples d'apprentissage de variétés : un tore et un solénoïde. (source : [4])

2.4 Homologie persistante

Dans cette section, nous supposerons – situation réelle – que nous disposons d'un nuage de points $\mathcal{P}$ de taille fixée. Toute la question est de trouver le *juste* rayon r pour construire notre complexe de Čech $\check{C}(\mathcal{P}, r)$. Question primordiale ; prenez-en un trop petit et $\check{C}(\mathcal{P}, r)$ sera trop déconnecté pour bien recouvrir la variété $\mathcal{M}$ dont on essaye de retrouver les caractéristiques d'homologie, prenez-en un trop grand et il risque d'apparaître des cycles intempestifs qui fausseront eux-aussi l'homologie.

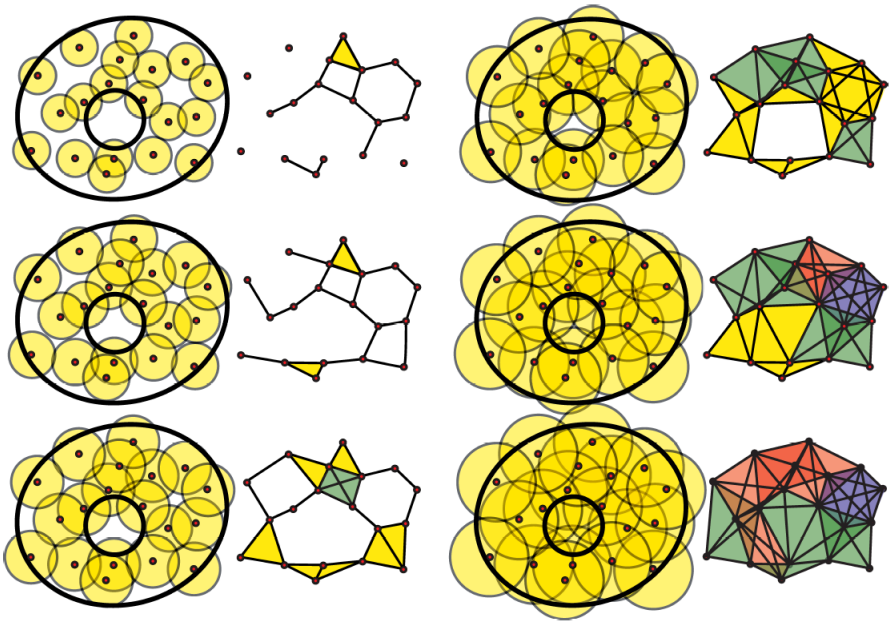


FIGURE 9 – Nuage de points $\mathcal{P}$ pris sur un anneau et 6 différents complexes construits à partir de ce même nuage en faisant varier le rayon r. (source : [10])

Le principe de l'homologie persistante consiste à contourner le problème en considérant toutes les valeurs de r possibles (au lieu de s'échiner à tenter vainement de déterminer un rayon idéal).

Sans rentrer dans trop de formalisme, remarquons que, pour $r \leq r'$, on a une injection naturelle :

$$\check{\mathcal{C}}(\mathcal{P}, r) \hookrightarrow \check{\mathcal{C}}_(\mathcal{P}, r').$$

On a $\check{\mathcal{C}}(\mathcal{P}, 0) \cong \{1, \ \ldots \ , n\}$ (avec $n = \#\mathcal{P}$) et $\check{\mathcal{C}}(\mathcal{P}, \infty) \cong \Delta_{n-1}$ (où Δ_{n-1} est la famille contenant un simplexe sur n sommets ainsi que toutes ses faces).

Tout cycle Z apparaît donc à un certain moment r. Il y alors deux cas de figure possibles :

- il existe un cycle plus grand Z' qui apparaît au temps $r' \geq r$ et qui fait disparaître Z de son groupe d'homologie :

  $$\partial Z' = Z,$$

  auquel cas Z a l'*intervalle de vie* $[r; r'[$.

- il n'existe pas de tel Z' auquel cas Z a l'*intervalle de vie* $[r; \infty[$.

On désignera alors sous le vocable de *durée de vie* l'élément $\delta \in \mathbb{R}_+ \cup \{\infty\}$:

$$\delta = r' - r \text{ (ou } \delta = \infty \text{ si } r' = \infty).$$

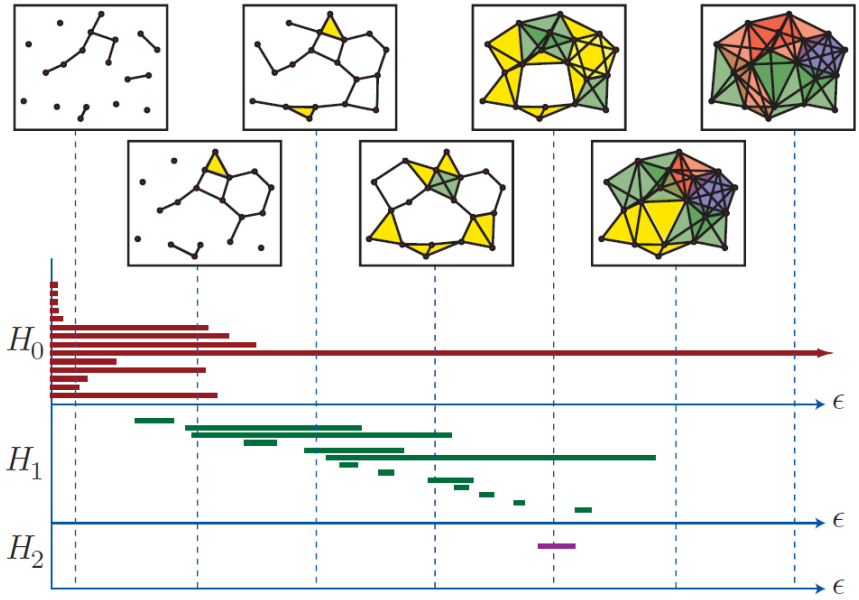


FIGURE 10 – Durée de vie de vie de quelques-uns des 0-cycles, 1-cycles et 2-cycle.
Il y a un seul 0-cycle *significatif* ; c'est-à-dire une seule composante connexe. On constate la présence d'un voire deux 1-cycles *significatifs*. (source : [10])

L'idée est la suivante : les cycles ayant la plus longue durée de vie disent quelque chose de l'homologie, ceux qui n'ont qu'une vie éphémère ne sont pas à prendre en compte.

Nous allons maintenant passer à la seconde partie proposant un algorithme de classification non-supervisée consistant à repérer des « pics de densité ». Pour ce faire, à rebours de l'homologie persistante, il faut rechercher là où les cycles ont la durée de vie la plus brève (car c'est là où il y aura la plus grande densité de points qu'il apparaîtra et disparaîtra le plus de cycles). Toutefois, l'algorithme reprend cette idée de faire varier un rayon pour comprendre quelque chose de la distribution. Voici comment.

3 L'algorithme de classification non-supervisée

3.1 Présentation du contexte

Toutes nos données seront maintenant des points de l'espace $\mathbb{R}^d$ munis de la distance euclidienne :

$$d(x,y) = ||x-y|| = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}.$$

L'échantillon $\mathcal{P}$ (certains parlent de *nuages de points*) à analyser est constitué de n points qui ont tous été tirés de façon indépendante selon une même loi de densité f par rapport à la mesure de Lebesgue. C'est-à-dire que pour tous points $p_1, \ldots, p_k \in \mathcal{P}$ et tous boréliens $B_1, \ldots, B_k \subseteq R^d$, la probabilité que tous les p_i soient dans les B_i vaut :

$$\mathbb{P}(\forall\, i \in \{1, \ldots, k\},\ p_i \in B_i) = \prod_{i=1}^{k} \int_{B_i} f(x)\,\mathrm{d}x.$$

Nous supposerons de plus que cette fonction f est continue et à support compact.

Pour tout réel r compris entre 0 et $f_{max} := \max_{\mathbb{R}^d} f$, on appellera *probabilité de niveau r* le nombre :

$$\mathcal{A}(r) := \int_{f>r} f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \mathbb{1}_{f(x)>r} f(x)\,\mathrm{d}x.$$

La fonction $\mathcal{A} : [0; f_{max}] \to [0; 1]$ est continue-à-doite/limite-à-gauche (*càdlàg*), strictement décroissante avec $\mathcal{A}(0) = 1$ et $\mathcal{A}(f_{max}) = 0$.

Définition . *Soit $r \in [0; f_{max}]$, on appellera pics de densité de niveau r l'ensemble des composantes connexes de $f^{-1}(]r; f_{max}])$.*
On dira de plus que ce niveau n'est pas dégénéré si les pics de densité sont bornés, en nombre fini, à distance strictement positive les uns des autres et que : $\partial f^{-1}(]r; f_{max}]) = f^{-1}(\{r\})$.

3.2 Justification de la recherche des « pics de densité »

Nous parlions naguère de *stylométrie* ; c'est-à-dire de classification de textes en fonction de leurs caractéristiques linguistiques (le vocabulaire ou la ponctuation employés, les constructions grammaticales, etc..). L'idée sous-jacente est bien qu'on peut reconnaître l'auteur d'un texte en fonction de son *style*.

Il en va de même pour l'*apprentissage statistique* en général : on suppose que nos données sont tirées parmi k classes différentes, ces k classes ayant des distributions $f_1, \ldots, f_k$ distinguables notamment par des « pics de concentration » sur des compacts $C_1, \ldots, C_k$ deux à deux disjoints.

Les seules informations qui nous sont données sont les points constituant le *nuage* $\mathcal{P}$, dont on peut supposer (à défaut d'informations supplémentaires) qu'ils sont tirés indépendamment selon la densité :

$$f = \sum_{i=1}^{k} \lambda_i f_i$$

où λ_i est la proportion dans l'échantillon $\mathcal{P}$ de points tirés à partir de la classe i.

Il suffit alors de ne conserver qu'une partie des données (on notera P cette proportion des points gardés) ; celles appartenant à ces « pics de concentration » (qu'on peut reconnaître grâce à la complexité des structures géométriques

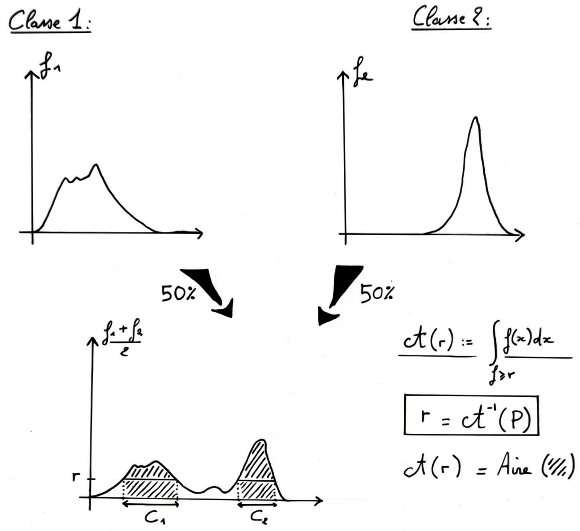


FIGURE 11 – Nouvelle fonction de densité $f = \frac{f_1+f_2}{2}$ pour un échantillon dont les éléments sont tirés aléatoirement selon une densité f_1 ou f_2

formées par les points en ces compacts) pour retrouver l'entier k et, par suite, les k compacts C_1, ... , C_k.

Remarque . *On ne conserve – certes ! – pour la recherche des différentes classes qu'une certaine proportion P des points... cependant tout n'est pas perdu ; une fois les k classes identifiées, on peut recourir à d'autres méthodes d'apprentissage (supervisé, cette fois-ci) pour classifier les points qui tout d'abord ne l'avaient pas été.*

3.3 Présentation de l'algorithme

Entrons à présent dans le vif du sujet, savoir : le fonctionnement de l'algorithme.

Après une brève description générale présentant notamment les paramètres que doit fournir l'utilisateur, nous verrons comment le programme construit un complexe de Čech qu'il fait grossir jusqu'à remplir une condition demandée par l'utilisateur.

Enfin, une fois la construction du complexe finie, nous verrons le fonctionnement du classificateur lui-même qui utilise ce complexe

3.3.1 Représentation des données et paramètres fournis

La seule chose dont nous disposons est une matrice M de taille $n \times d$ représentant l'ensemble des données $\mathcal{P}$ constitué de n points de l'espace $\mathbb{R}^d$. L'utilisateur a tout de même deux paramètres supplémentaire à fournir au programme :

- un réel $P \in]0;1]$; c'est la proportion des n points qui seront classifiés lorsque le programme aura fini
- un entier $k \geq 1$; c'est la dimension des polyèdres qui seront construits pour la classification... nous y reviendrons !

Remarques . *Pour montrer les propriétés asymptotiques (quand $n \to \infty$) de notre algorithme (cf. la section suivante), k doit vérifier deux hypothèses :*

$$k = o(n) \qquad et \qquad k \gg \log(n).$$

Pour ce qui est du choix de P, l'utilisateur doit se fier à l'idée qu'il se fait de la bonne séparation des données entre les différentes classes ; s'il fournit un P proche de 1, c'est qu'il estime avoir des données clairement distinctes d'une classe à l'autre.

3.3.2 Construction du *complexe de Čech*

Le programme récupère donc trois ingrédients :

- une matrice M de taille $n \times d$ représentant les n points de $\mathbb{R}^d$ constituant notre nuage de points $\mathcal{P}$
- un entier $k \geq 1$
- un réel $P \in]0;1]$

... charge à lui maintenant de classifier !

Pour commencer, il va construire un complexe de Čech $\check{C}(\mathcal{P}, r)$ sur $\mathcal{P}$ avec un rayon r assez important pour qu'il y ait au moins une proportion P des points de $\mathcal{P}$ qui soient dans des simplexes de $\check{C}(\mathcal{P}, r)$ de dimension k. Concrètement, cela est très simple : on initialise r à 0, le complexe de Čech $\check{C}(\mathcal{P}, 0) \simeq \{1, \ldots, n\}$ est alors simplement constitué de simplexes de dimension 0 ; les sommets de $\mathcal{P}$. On fait alors croître r. Grâce à l'injection naturelle :

$$\check{C}(\mathcal{P}, r) \hookrightarrow \check{C}(\mathcal{P}, r') \quad \text{pour } r \leq r',$$

le nombre d'éléments de $\mathcal{P}$ apparaissant dans au moins un simplexe de dimension k va augmenter avec r... on s'arrête dès que l'on en a $P \times n$.

3.3.3 Classification grâce aux différents *polyèdres* du complexe de Čech

Maintenant que le choix du *juste rayon* r a été fait, que le complexe $\check{C}(\mathcal{P}, r)$ a été construit, nous en arrivons à la classification elle-même qui se fera à l'aide de *polyèdres* de dimension k.

Définition . *Définissons inductivement ce que nous entendons par un polyèdre de dimension k :*

- *un simplexe de dimension k est un polyèdre de dimension k*
- *si deux polyèdres de dimension k partagent une même facette de dimension $k-1$, alors leur union est encore un polyèdre de dimension k*

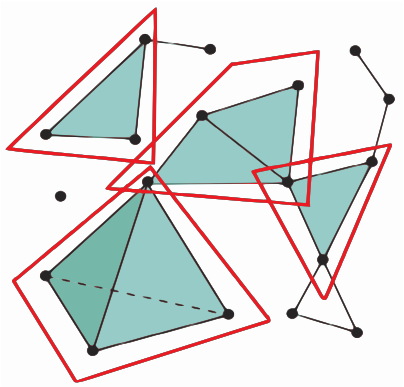


FIGURE 12 – Les polyèdres que retourne le classificateur pour $k = 2$.

À partir du complexe de Čech $\check{C}(\mathcal{P}, r)$ que l'on a calculé, on peut construire des polyèdres de dimension k en les faisant les plus grands possibles. Tout élément présent dans un simplexe de dimension k se retrouvera donc dans au moins un tel polyèdre ; voilà notre classificateur !

3.4 Propriétés de l'algorithme

Passons maintenant aux propriétés de notre algorithme.
Nous supposerons que la fonction f de densité selon laquelle sont tirées les données est une fonction étagée. Il n'est bien sûr pas indispensable de la supposer telle. Toutefois, et de manière générale, on prendra des hypothèses très fortes par souci d'avoir des démonstrations sans trop de longueurs. Le lecteur attentif pourra à sa guise affiner les hypothèses.

Dans toute la suite de cette section, n désignera le nombre de données de l'espace $\mathbb{R}^d$ (toutes tirées indépendamment selon une même loi de densité f), P et k seront les deux paramètres fournis par l'utilisateur ; respectivement la proportion des points à trier et la dimension des polyèdres pour la classification.

3.4.1 Cas où la fonction de densité f est étagée

Supposons qu'il existe un entier N, des réels $a_1 > ... > a_N > 0$ et des ensembles ouverts convexes $A_1, ..., A_N \subset \mathbb{R}^d$ à distance strictement positive les uns des autres tels que f puisse s'écrire sous cette forme :

$$\forall\, x \in \mathbb{R}^d,\ f(x) = \sum_{i=1}^{N} a_i \mathbb{1}_{\{x \in A_i\}}.$$

P a été donné par l'utilisateur de telle sorte qu'il existe $\alpha \in \{0, \dots, N-1\}$ tel que (la notation $|V|$ désigne la mesure de Lebesgue de l'ensemble V) :

$$\sum_{i=1}^{\alpha} a_i |A_i| < P < \sum_{i=1}^{\alpha+1} a_i |A_i|.$$

Alors asymptotiquement, quand $n \to \infty$, $k \to \infty$, $k = o(n)$ et $k \gg \log(n)$, le classificateur sera composé de :

- α classes qui vont asymptotiquement recouvrir parfaitement chacun des A_i ($i \in \{1, \dots, \alpha\}$)
- aucune classe sur les A_i pour $i > \alpha + 1$
- un certain nombre de classes sur $A_{\alpha+1}$ (le comportement sur $A_{\alpha+1}$ n'est pas très clair à déterminer... néanmoins, f étant en pratique continue, on peut considérer que l'ensemble sur lequel le classificateur a un comportement chaotique est petit et diminue avec n)

La preuve suit le cours suivant : on commence par essayer d'encadrer le rayon r qui sera choisi pour la construction du complexe $\check{C}(\mathcal{P}, r)$ entre un majorant $r_{\max}$ et un minorant $r_{\min}$.

Asymptotiquement, $\check{C}(\mathcal{P}, r_{\max})$ comportera une proportion de points classifiés égale à $\sum_{i=1}^{\alpha+1} a_i |A_i|$. Pour $r_{\min}$, cette proportion est égale à $\sum_{i=1}^{\alpha} a_i |A_i|$.

Comme P est strictement compris entre ces deux valeurs, $\mathbb{P}(r \in [r_{\min}; r_{\max}]) \to 1$ quand $n \to \infty$.

De plus, pour tout $i \in \{1, \dots, \alpha\}$,

$$\mathbb{P}\left(\{x \in Int\,(A_i, r_{\min})\} \text{ appartiennent à un même polyèdre de } \check{C}(\mathcal{P}, r_{\min})\right) \to 1$$

où $Int(A_i, r_{\min}) = \{x \in A_i \mid B_{r_{\min}}(x) \subseteq A_i\}$.
Comme asymptotiquement $r \geq r_{\min}$ et qu'alors $\check{C}(\mathcal{P}, r_{\min}) \hookrightarrow \check{C}(\mathcal{P}, r)$ et $Int(A_i, r) \subseteq Int(A_i, r_{\min})$, il s'ensuit que cette propriété est aussi vérifiée en substituant r à $r_{\min}$.

De même, pour tout $i > \alpha + 1$,

$$\mathbb{P}\left(\{\text{aucun des points de } A_i \text{ n'est classifié pour } \check{C}(\mathcal{P}, r_{\max}\}\right) \to 1.$$

Puisque $\mathbb{P}(r \leq r_{\max}) \to 1$, cette propriété est aussi vérifiée pour la classification en polyèdres de $\check{C}(\mathcal{P}, r)$.

Bon !.. allons-y !

Soit $\epsilon > 0$ assez petit tel que si l'on définit $r_{\min}$ et $r_{\max}$ comme ci-dessous (ω_d est le volume de la boule unité de $\mathbb{R}^d$) :

$$r_{\min} := \left(\frac{(1+\epsilon)k}{n\,\omega_d\,a_\alpha}\right)^{\frac{1}{d}} \text{ et}$$

$$r_{\max} := \left(\frac{(1+\epsilon)k}{n\,\omega_d\,a_{\alpha+1}}\right)^{\frac{1}{d}},$$

alors, si ϵ est suffisamment proche de 0, on peut aussi avoir :

$$r_{\min} \leq \left(\frac{(1-\epsilon)k}{n\,\omega_d\,a_{\alpha+1}} \right)^{\frac{1}{d}} \text{ et}$$

$$r_{\max} \leq \left(\frac{(1-\epsilon)k}{n\,\omega_d\,a_{\alpha+2}} \right)^{\frac{1}{d}}.$$

Ces propriétés sont réalisable car $\frac{a_\alpha}{a_{\alpha+1}}$ et $\frac{a_{\alpha+1}}{a_{\alpha+2}}$ sont *strictement* plus grands que 1.
Elles signifient qu'une boule de rayon $r_{\min}$ dans A_α va contenir *en moyenne* $(1+\epsilon)k$ points et moins de $(1-\epsilon)k$ si elle se trouve dans $A_{\alpha+1}$. De même, une boule de rayon $r_{\max}$ va contenir en moyenne $(1+\epsilon)k$ éléments si elle est dans $A_{\alpha+1}$ et moins de $(1-\epsilon)k$ sur $A_{\alpha+2}$.

Traitons, le cas de $r_{\min}$!
Commençons par une petite propriété :

Lemme . $\mathbb{P}\left(\exists\, c \in Int(A_\alpha, r_{\min}),\ B_{r_{\min}}(c) \textit{ contient moins de } k \textit{ points}\right) \to 0$

Remarque . *Si cette propriété est vraie pour A_α, elle l'est donc aussi pour tous les A_i ($i \in \{1, \ldots, \alpha\}$) car la densité des points y est plus élevée.*

Pour montrer ce résultat, il nous faudra d'abord un majorant de la probabilité qu'une boule $B \subseteq A_\alpha$ *fixe* de rayon $r_{\min}$ ait moins de k points.
Les n points sont tirés indépendamment, la probabilité p_n pour chacun de tomber dans B est égale à :

$$p_n = \frac{(1+\epsilon)k}{n}.$$

Le nombre de points dans B suit donc une loi binomiale $\mathcal{B}(n,\, p_n)$ et

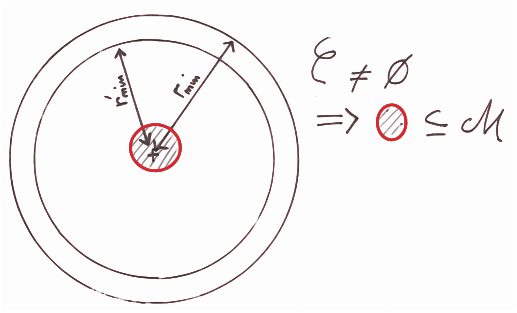
$$\begin{aligned}
\mathbb{P}(\#B \leq k) &= \sum_{i=0}^{k} \mathrm{C}_n^i (p_n)^i (1-p_n)^{n-i} \\
&= \sum_{i=0}^{k} \mathrm{C}_n^i \left(\frac{(1+\epsilon)k}{n}\right)^i \left(1 - \frac{(1+\epsilon)k}{n}\right)^{n-i} \\
&\leq (k+1)\frac{n!}{k!(n-k)!}(p_n)^k(1-p_n)^{n-k} \qquad \text{(on majore chaque terme par celui où } i=k\text{)} \\
&\leq (k+1)\frac{n!}{k!(n-k)!}\frac{k^k(1+\epsilon)^k}{n^k}\frac{(n-(1+\epsilon)k)^{n-k}}{n^{n-k}} \\
&\leq (k+1)\sqrt{n}\left(\frac{n}{e}\right)^n\left(\frac{k}{e}\right)^{-k}\left(\frac{n-k}{e}\right)^{-(n-k)}\frac{k^k(1+\epsilon)^k}{n^k}\frac{(n-(1+\epsilon)k)^{n-k}}{n^{n-k}} \text{ (formule de Stirling)} \\
&\leq (k+1)\sqrt{n}\,(n-k)^{-(n-k)}\,(1+\epsilon)^k\,(n-(1+\epsilon)k)^{n-k} \\
&\leq (k+1)\sqrt{n}\exp\left[(n-k)\left(-\log(n-k)+\log\left(n-(1+\epsilon)k\right)\right)+k\log(1+\epsilon)\right] \\
&\leq (k+1)\sqrt{n}\exp\left[(n-k)\left(-\log\left(1-\frac{k}{n}\right)+\log\left(1-\frac{(1+\epsilon)k}{n}\right)\right)+k\log(1+\epsilon)\right] \\
&\leq (k+1)\sqrt{n}\exp\left[(n-k)\left(-\epsilon\frac{k}{n}+o\left(\frac{k}{n}\right)\right)+k\log(1+\epsilon)\right] \qquad (k \text{ est négligeable devant } n\,!) \\
&\leq (k+1)\sqrt{n}\exp\left[k(-\epsilon+\log(1+\epsilon))+o(k)\right]
\end{aligned}$$

Or $-\epsilon+\log(1+\epsilon)<0$, de plus $k \gg \log(n)$, il existe donc une constante $C(\epsilon)$ ne dépendant que de ϵ telle qu'on puisse avoir le majorant suivant pour k et n assez grands :

$$\mathbb{P}(\#B \leq k) \leq \exp\left[-kC(\epsilon)\right]. \qquad \text{(Victoire !)}$$

Revenons à nos moutons (et à leur lemme).
Soit $\mathcal{C} = \{c \in Int(A_\alpha, r_{\min}) \mid \#B_{r_{\min}}(c) \leq k\}$. On veut montrer que $\mathbb{P}(\{\mathcal{C} \neq \emptyset\}) \to 0$. Pour cela considérons :

$$r'_{\min} := r_{\min}\left(\frac{1+\frac{\epsilon}{2}}{1+\epsilon}\right)^{\frac{1}{d}} < r_{\min}.$$



Avec ce nouveau rayon, une boule de A_α a en moyenne $\left(1+\frac{\epsilon}{2}\right)k$ points.
Soit : $\mathcal{M} = \{c \in Int(A_\alpha, r'_{\min}) \;\middle|\; \#B_{r'_{\min}}(c) \leq k\}$. On a :

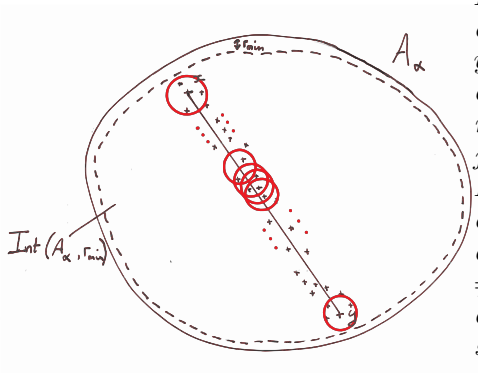
$$c \in \mathcal{C} \Rightarrow B_{r_{\min}-r'_{\min}}(c) \subseteq \mathcal{M}.$$

Ainsi :

$$
\begin{aligned}
\mathbb{P}(\mathcal{C} \neq \emptyset) &\leq \mathbb{P}\left(|\mathcal{M}| \geq \omega_d \left(r_{\min} - r'_{\min}\right)^d\right) \\
&\leq \frac{\mathbb{E}(|\mathcal{M}|)}{\omega_d \left(r_{\min} - r'_{\min}\right)^d} \qquad \text{(par l'inégalité de Markov)} \\
&\leq \frac{|A_\alpha| \mathbb{P}(\#B \leq k)}{\omega_d \, r_{\min}^d \left(1 - \left(\frac{1+\frac{\epsilon}{2}}{1+\epsilon}\right)^{\frac{1}{d}}\right)^d} \qquad \text{où } B \text{ est une boule } \textit{fixée} \text{ de rayon } r'_{\min} \\
&\leq \frac{|A_\alpha|}{\frac{k(1+\epsilon)}{n}\left(1 - \left(\frac{1+\frac{\epsilon}{2}}{1+\epsilon}\right)^{\frac{1}{d}}\right)^d} \exp\left[-kC\left(\frac{\epsilon}{2}\right)\right] \\
&\to 0 \quad \text{quand } n \to \infty.
\end{aligned}
$$

En couplant ce résultat avec la proposition suivante, on obtient le résultat escompté (avoir α classes qui recouvrent parfaitement asymptotiquement les ensembles A_1, ... A_α).

Proposition . *Si $\mathcal{C}$ est vide, alors tous les points de $Int(A_\alpha, r_{\min})$ sont dans un même polyèdre de dimension k pour le complexe $\check{C}(\mathcal{P}, r_{\min})$.*



Preuve . *Soient x et y dans $Int(A_\alpha, r_{\min})$. Regardons la boule centrée en x et faisons-la glisser vers y. Initialement en x, $\#B(x) \geq k+1$ donc x est dans un simplexe de dimension k. On fait glisser la boule jusqu'à ce qu'un seul des points de $B(x)$ n'y soit plus (avec probabilité 1, on peut supposer qu'un seul point sort à la fois). Soit c le nouveau centre. $\#(B(x) \cap B(c)) \geq k$ et $\#B(c) \geq k+1$ donc il existe un simplexe de dimension $k-1$ reliant les points de $B(x)$ à ceux de $B(c)$. On procède ainsi jusqu'à relier x et y.*

On peut de la même façon montrer que la probabilité qu'il existe dans $\check{C}(\mathcal{P}, r_{\min})$ un simplexe de dimension k constitué de points de $A_{\alpha+1}$ (ou de A_i pour tout $i \geq \alpha+1$) tend vers 0 quand $n \to \infty$.

Ainsi, la proportion des points classés pour $\check{C}(\mathcal{P}, r_{\min})$ tend vers $\sum_{i=1}^{\alpha} a_i |A_i| < P$ (tous les points des A_1, ... , A_α sont asymptotiquement parfaitement classés et aucun point des autres A_i, $i > \alpha$ n'est classifié).

Le cas de $r_{\max}$ se traite exactement de la même façon ; cette fois, tous les points des classes 1 à $\alpha+1$ seront asymptotiquement parfaitement classés et aucun des autres classes... ce qui fait qu'on en classifie une proportion qui tend

en probabilité vers $\sum_{i=1}^{\alpha+1} a_i|A_i| > P$. De ces deux encadrements, on obtient que :

Proposition . $\mathbb{P}(r \in [r_{\min}; r_{\max}]) \to 1$ *quand* $n \to \infty$. *Par suite,* $r \geq r_{\min}$ *entraîne que les* α *premières classes sont parfaitement reconnues et* $r \leq r_{\max}$ *implique qu'aucun point des classes* $\alpha + 2$, $\alpha + 3$, ... , N *n'est classifié...*

Bref ! on a montré les beaux résultats promis !!!

3.5 Un exemple : la *stylométrie*

3.5.1 Présentation de la stylométrie et des données disponibles

Nous allons présenter au lecteur un cas concret d'utilisation de l'apprentissage statistique qui tient au cœur de l'auteur : celui de la *stylométrie*.

Comme cela a déjà été dit, la *stylométrie* est cet *art* (pris dans son sens grec de τέχνη) à la croisée de la statistique et de la linguistique dont le but est de produire à partir de textes une *information* (qui puisse être traitée statistiquement) rendant compte du *style* de ces derniers ; c'est-à-dire qui caractériserait – idéalement, cela s'entend... – à la fois son auteur, mais aussi son genre, son époque, etc..

Plus pratiquement, ces statistiques portent généralement sur le vocabulaire utilisé (qui peut être affiné en repérant le *sens* en lequel l'auteur emploie tel mot), les catégories grammaticales (noms/pronoms, adjectifs, adverbes, verbes... à quel temps/mode ?), la ponctuation, etc... on peut même simplement regarder dans un premier temps (et ce qui ne donne pas de mauvais résultats) la fréquence des lettres employées.

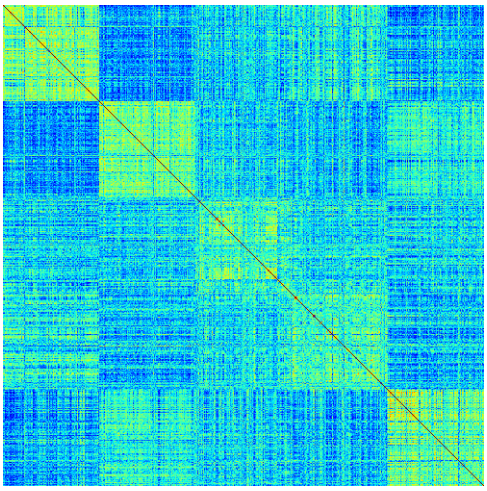


FIGURE 13 – Matrice de Gram pour 500 textes (5 auteurs × 100 textes/auteurs). Les blocs diagonaux aux couleurs plus chaudes montrent que deux textes d'un même auteur sont, en général, reconnus comme plutôt ressemblants.

Dans notre cas, nous avons travaillé sur des articles de journaux de cinq auteurs différents en prenant cent textes de chaque auteur. Nous donnons en image la matrice 500×500 des produits scalaires (appelée *matrice de Gram*) : l'indice (i, j) représente la *ressemblance* entre les textes i et j. Comme les textes sont triés en fonction de leur auteur, on voit se détacher cinq bloc diagonaux ; un pour chaque auteur. Le but d'un classificateur non supervisé est de retrouver une telle matrice à partir de la même matrice dont les coordonnées auraient subi une permutation aléatoire (évidemment, on ne donne pas les textes triés en fonction de leur auteur).

3.5.2 Résultats

Pour commencer, l'auteur prie le lecteur de bien vouloir aller de nouveau jeter un œil sur la matrice de Gram ; il s'apercevra que les données sur lesquelles le programme a tourné ne sont pas parfaites (c'est une litote !) : il y a beaucoup de carrés chauds (dans les tons verts) qui apparaissent en dehors de la diagonale là où théoriquement il devrait apparaître des carrés bleus... et vice-versa... l'auteur demande donc l'indulgence du jury pour les résultats qui seront présentés.

Ce préambule étant fait, passons au vif du sujet !

Et tout d'abord, choisissons la facilité ; sur la matrice de Gram il apparaît que les auteurs les plus distinguables sont ceux numéros 1 et 2 (la matrice extraite en haut à gauche se divise peu ou prou en 4 carrés : 2 verts sur la diagonales et 2 bleus à l'extérieur). On peut donc demander une précision importante : $P = 99\%$ sur l'ensemble des 50 premiers textes de chacun des deux auteurs (et $k = 10$). Le programme identifie qu'il y a effectivement 2 classes (les textes 0 à 49 étant de l'auteur 1 et ceux indexés de 50 à 99 de l'auteur 2) :

i) 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49

ii) 22, 50, 51, 52, 53, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99

... il se trompe donc uniquement sur le texte 22 (ce qui est excusable lorsqu'on regarde la matrice de Gram précisément sur cette ligne) et ne classe pas le texte 54. Le lecteur pourra trouver une projection à l'aide d'une Analyse en Composante Principale des deux premières coordonnées avec leur classification par le programme.

Passons à quelque chose de beaucoup plus ardu !
Cette fois, mettons les cinq auteurs (40 textes chacun) !
On ne peut guère plus que demander une précision $P = 70\%$ (et $k = 5$). Le programme réussit quand même à déterminer la présence de 5 classes... les voici :

i) 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23,

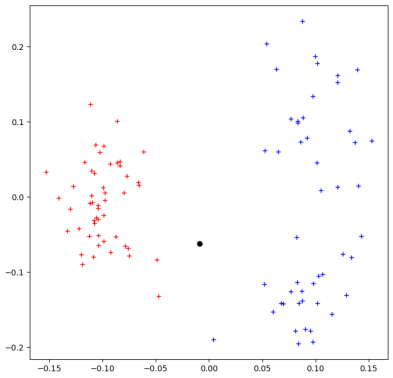


FIGURE 14 – Projection à l'aide d'une Analyse en Composantes Principales (ACP) de 50 textes des auteurs 1 et 2. Les différentes couleurs représentent les classes renvoyées par notre classificateur. Le point noir est celui non classé (texte 54)

24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 121, 122, 130, 144, 146, 153, 156, 157

ii) 40, 41, 42, 43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 61, 63, 64, 65, 66, 68, 70, 72, 74, 76, 77, 78, 79, 125, 182, 188, 192, 196

iii) 85, 87, 94, 96, 99, 101, 102, 103, 104, 105, 108, 109, 110, 111, 112, 114, 116, 117, 123, 124

iv) 85, 87, 99, 104, 113, 114, 117

v) 55, 80, 106, 115, 119, 143, 160, 161, 162, 163, 164, 165, 166, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 194, 195, 196, 197, 199

Il saute aux yeux que la classe 4 (des textes 120 à 159) a été complètement passée à la trappe (en fait le programme n'identifie qu'à peu près correctement 4 des 5 classes).

Sur les éléments qu'il a classés, il a commis les erreurs suivantes :

- 8 erreurs sur les 46 éléments mis dans la classe 1
- 5 erreurs sur 36 pour la classe 2
- 2 erreurs sur 20 pour la classe 3
- 6 erreurs sur 41 pour la classe 5

Remarque . *En retirant de la classification les points qui tombent dans différentes classes, cela permettrait de diminuer fortement le nombre d'erreurs de classement (cela diminuerait – certes ! – aussi la proportion de points classés...)*

Peut-être le lecteur sera-t'il moins dur lorsqu'il verra les projections (grâce toujours à une Analyse en Composantes Principales) des données qu'il fallait classifier...

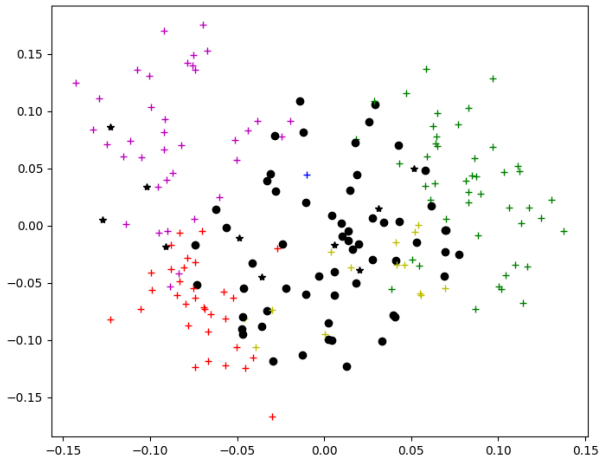


FIGURE 15 – Projection grâce à une ACP de 40 textes de chacun des 5 auteurs. En couleur apparaissent les différentes classes que le programme renvoie pour $k = 5$ et $P = 70\%$. Les ronds noirs sont les points non classés, les étoiles noires sont les points tombant dans plusieurs classes.

4 Conclusion

Et voilà que ce rapport se termine. Cette conclusion sera pour moi l'occasion de remercier très vivement BARTLOMIEJ BLASZCZYSZYN, mon professeur de graphes aléatoires, qui, après m'avoir dispensé son cours fort intéressant, a très gentiment accepté de m'encadrer (tout en me laissant toujours beaucoup de liberté) pour ce mémoire.

Pour finir, nous nous excusons auprès du lecteur s'il a par trop souvent été rebuté au fil de sa lecture par l'aridité de ce rapport ; nous avons (autant que faire se peut) essayé de l'agrémenter d'exemples et d'illustrations pour pallier ce problème... au moins La Fontaine nous accorde-t'il quelque mérite :

Et si de t'agréer je n'emporte le prix,
J'aurai du moins l'honneur de l'avoir entrepris.

5 Bibliographie

Références

[1] Omer BOBROWSKI and Matthew KAHLE. Topology of randam geometric complexes: a survey. Sept. 2014.

[2] Omer BOBROWSKI, Matthew KAHLE, and Primoz SKRABA. Maximally persistent cycles in random geometric complexes. Mai 2016.

[3] Omer BOBROWSKI and Sayan MUKHERJEE. The topology of probability distributions on manifolds. Mars 2014.

[4] Jean-Daniel BOISSONAT. Manifold reconstruction. *Cours dispensé au Master Parisien de Recherche Informatique (MPRI)*, 2017.

[5] Jean-Daniel BOISSONAT. Simplicial complexes. *Cours dispensé au MPRI*, 2017.

[6] Peter BUBENIK and Peter T. KIM. A statistical approach to persistent homology. *Homology, Homotopy and Applications*, 9(2), 2007.

[7] Gunnar CARLSSON. Topology and data. *BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY*, 46(2), avril 2009.

[8] Herbert EDELSBRUNNER and John HARER. *Computational topology : an introduction.*

[9] Aurélie FISCHER. Classification non supervisée et données fonctionnelles. Oct. 2008.

[10] Robert GHRIST. Barcodes : the persistent topology of data. *BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY*, 45(1), Janv. 2008.

[11] Matthew KAHLE. Topology of random clique complexes. Nov. 2008.

[12] Matthew KAHLE and Elizabeth MECKES. Limit theorems for betti numbers of random simplicial complexes. Janv. 2011.

[13] Dimitry MOROZOV. A practical guide to persistent homology.

[14] James R. MUNKRES. *Elements of algebraic topology.*

[15] P. NIYOGI, S. SMALE, and S. WEINBERGER. A topological view of unsupervised learning from noisy data. Mai 2015.

[16] Matthew PENROSE. *Random geometric graphs.*

[17] WIKIPÉDIA. Complexe simplicial, 2017.