Geometric Methods in Statistics and Machine Learning M2 Statistique, Apprentissage et Algorithmes (MS2A)

Jordan Serres[1]

February 18, 2026

[1]LPSM, Sorbonne Université, 4 place Jussieu 75005 Paris, France (serres@lpsm.paris)

Contents

Chapter 1

Introduction

1.1 Statistique en grande dimension

1.1.1 Exemple introductif

Au cours des dernières années, l'essor des données de grande dimension a confronté les statistiques et le machine learning à un défi majeur : le fléau de la dimension.

Ce terme désigne le phénomène selon lequel la plupart des algorithmes classiques voient leur vitesse de convergence devenir impraticable lorsque la dimension des données augmente fortement, comme c'est le cas, par exemple, pour les images, représentées par des vecteurs comportant plusieurs millions de composantes.

Prenons l'exemple jouet de classification suivant[1] : on prend la base de données d'images Fashion-MNIST, composée de 70 000 images de $28 \times 28 = 784$ pixels avec des niveaux de gris allant de 0 à 255 (soit 8 bits). Il s'agit donc de vecteurs éléments de l'ensemble

$$[[0, 255]]^{784} \subset \mathbb{R}^{784},$$

qui peut être considéré comme un espace de grande dimension, bien que dans la pratique les images de plus haute résolution se représentent plutôt comme des vecteurs de dimension de l'ordre de plusieurs millions.

Figure 1.1: Fashion-MNIST

Parmi ces images, il y a 10 classes différentes de vêtements (t-shirt, pantalon, etc.), et le but est de classifier ces images, c'est-à-dire d'assigner à chacune d'entre elles son label (le numéro de l'item auquel elle appartient). Pour cela, parmi les 70 000 images, on dispose de 60 000 qui sont déjà étiquetées (par des humains), et le but est d'apprendre à partir de cet ensemble d'entraînement afin de *généraliser* correctement, c'est-à-dire de prédire le label des images qui n'avaient pas été étiquetées.

[1] Merci à David Tewodrose à qui j'emprunte cet exemple !

Un algorithme réalisant cette tâche est le suivant. On fixe $N, k \in \mathbb{N}$, puis on divise l'ensemble d'entraînement en N sous-ensembles (les batches) $B_1, \dots, B_N$. Étant donnée une image $x \in \mathbb{R}^{784}$, on calcule pour chaque batch B_i, $i = 1, \dots, N$, ses k plus proches voisins dans le batch B_i, c'est-à-dire les k points dans B_i qui sont les plus proches de x pour la distance euclidienne. On regarde alors, pour chacun de ces k plus proches voisins, quel est son label, c'est-à-dire sa classe de vêtement. Pour ce batch B_i donné, on attribue ensuite à x le label majoritaire parmi les k plus proches voisins dans B_i. On obtient donc N labels, un pour chaque batch, et on attribue finalement à x le label majoritaire parmi ceux-ci.

Le problème de la grande dimension surgit justement du calcul des distances euclidiennes dans $\mathbb{R}^D$ pour D grand (ici $D = 784$). En effet, lorsque la dimension est grande, *"tous les points se retrouvent éloignés"*. Dit de façon plus rigoureuse, si on considère l'hypercube $[0,1]^D$ et qu'on le divise en petits hypercubes de côté 10^{-m}, alors on se retrouve avec 10^{mD} hypercubes, soit un nombre gigantesque dès lors que D est grand. Par exemple, ici pour $m = 1$ et $D = 784$, on obtient 10^{784} hypercubes, soit beaucoup, beaucoup plus que le nombre de données 7×10^4. Cela signifie que deux données différentes se trouveront quasi systématiquement dans des hypercubes différents et seront donc toujours éloignées. Ce phénomène très important est connu sous le nom de *malédiction de la dimension*.

1.1.2 Le phénomène de concentration de la mesure

En grande dimension, on ne voit que des événements typiques : c'est ce que l'on nomme *la concentration de la mesure*.

La loi des grands nombres est une instanciation de ce phénomène. Prenons un moment pour l'illustrer. Si l'on a une suite infinie de variables aléatoires $X_1, X_2, \dots, X_n, \dots$ qui sont iid et intégrables ($\mathbb{E}[|X_1|] < \infty$), alors la loi des grands nombres affirme que la moyenne empirique converge presque sûrement vers l'espérance, c'est-à-dire la moyenne théorique :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\longrightarrow} \mathbb{E}[X_1] \quad \text{presque sûrement.}$$

Cela signifie que lorsque le nombre de données augmente, c'est-à-dire quand n augmente, le nombre de possibilités augmente[2], mais le nombre de possibilités que l'on peut effectivement *observer* diminue, jusqu'à ne plus pouvoir observer qu'un seul événement à la limite où l'on a accès à la suite infinie des données[3]. Dans le cas où ce que l'on observe est la moyenne empirique, l'unique valeur effectivement observable à la limite est la moyenne théorique $\mathbb{E}[X_1]$, comme l'affirme la loi des grands nombres.

Dit autrement, lorsque la taille de l'échantillon augmente, les valeurs des *observables effectives* se *concentrent* autour d'un point. C'est exactement ce que l'on nomme le phénomène de concentration de la mesure, et cela dépasse largement la loi des grands nombres.

En effet, de façon générale, une observable est une fonction K_n-Lipschitz des données : $f_n(X_1, \dots, X_n)$. Il s'agit donc d'une fonction[4]

$$f_n : E^n \to \mathbb{R}$$

vérifiant

$$|f_n(X_1, \dots, X_n) - f_n(X_1', \dots, X_n')| \leq K_n \sum_{i=1}^{n} \|X_i - X_i'\|,$$

où E dénote l'ensemble où prennent leurs valeurs les variables aléatoires, et $||\cdot||$ une norme sur cet ensemble[5]. Notez bien que la constante K_n est autorisée à dépendre de n.

Exercice : La fonction moyenne empirique

$$(x_1, \dots, x_n) \mapsto \frac{1}{n}\sum_{i=1}^{n} x_i$$

[2] et augmente exponentiellement, par exemple pour des Bernoulli, le nombre de possibilités double à chaque nouvelle donnée
[3] ce qui est évidemment impossible mais peut être conçu comme une vue de l'esprit
[4] en réalité une suite de fonctions, mais pour alléger la notation, on n'écrit pas toujours l'indice n
[5] qui sera donc souvent $\mathbb{R}^d$ muni de la norme euclidienne

est-elle Lipschitz ? Si oui, quelle est sa constante ?

La concentration de la mesure est alors la propriété que l'observable $f(X_1, \dots, X_n)$ se concentre autour de sa moyenne avec une certaine vitesse $\alpha : (0, \infty) \to (0, \infty)$, au sens où pour tout $\varepsilon > 0$,

$$\mathbb{P}\Big(\big|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]\big| \geq \varepsilon\Big) \leq \alpha\Big(\frac{\varepsilon}{K}\Big).$$

Exercice : Dans le cas où les X_i sont iid à valeurs dans $[a, b] \subset \mathbb{R}$ et que l'observable est la moyenne empirique, déterminer la fonction de concentration $\alpha : (0, \infty) \to (0, \infty)$ correspondante.

Que signifie l'hypothèse que l'observable soit Lipschitz ? Cela signifie que si l'on ne modifie que peu de données, alors la valeur de la fonction ne change que peu. Autrement dit, il faut modifier beaucoup de données pour que la fonction soit modifiée significativement. C'est une propriété de robustesse.

La morale à retenir peut être formulée ainsi : en grande dimension (c'est-à-dire n grand), les observables robustes ne voient que les événements *typiques*. Inversement, les événements plus fins, non typiques[6], ne sont pas robustes. Seules les structures robustes sont statistiques.

1.2 La géométrie en analyse des données

L'apparition de la géométrie en science des données provient à la fois de la pratique et de la théorie.

Du côté pratique, cela vient du fait que, malgré que la malédiction de la dimension semble prohiber toute forme de statistique en grande dimension, l'utilisation de certains outils statistiques fonctionne malgré tout, alors que l'on s'attendrait à ce qu'ils ne fonctionnent pas à cause de la grande dimension.

Du côté théorique, l'existence d'une géométrie intrinsèque aux données constitue une bonne hypothèse de travail pour démontrer, de façon rigoureuse, des résultats non triviaux et utilisables en pratique dans le cadre de la grande dimension. Cette hypothèse justifie également *a posteriori* l'observation selon laquelle, en pratique, de nombreux algorithmes fonctionnent malgré la grande dimension.

De façon très vague, cela peut être résumé par ce que l'on appelle *l'hypothèse de la variété*.

Hypothèse. *(hypothèse de la variété) Les données de grande dimension $X_1, \dots, X_n \in \mathbb{R}^D$, $D \gg 1$, possèdent une géométrie intrinsèque, de dimension intrinsèque $d \ll D$.*

Un des objectifs de ce cours sera de comprendre plus précisément ce que signifie cette hypothèse, c'est-à-dire ce que l'on entend par "géométrie intrinsèque", mais également quelles sont ses implications, tant du point de vue pratique que théorique, en mettant davantage l'accent sur les idées conceptuelles plutôt que sur les aspects techniques.

1.2.1 Géométrie des données

Premier cas, Figure 1.2 : on mesure la configuration d'un bras robotique avec 3 articulations. Chaque articulation est caractérisée par son orientation, c'est-à-dire son angle, et un angle dans l'espace est naturellement vu comme un vecteur sur la sphère $\mathbb{S}^2$. Par conséquent, ici les données sont à valeurs dans la variété produit

$$\mathbb{S}^2 \times \mathbb{S}^2 \times \mathbb{S}^2 \subset \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 = \mathbb{R}^9.$$

[6] par exemple l'événement où tous les X_i sont égaux, correspondant à l'observable $\mathbf{1}_{X_1 = \dots = X_n}$

Figure 1.2: Bras robotique
Données dans $\mathbb{S}^2 \times \mathbb{S}^2 \times \mathbb{S}^2$
(image de M. Belkin)

Un second cas, Figure 1.3, est constitué d'un ensemble de photos d'un même visage 3D d'une statue, prises sous différents angles. Alors que l'image est naturellement représentée comme un vecteur de dimension égale au nombre de pixels (donc très grande dimension), le fait que les seuls degrés de liberté soient l'angle de prise de vue induit de facto une structure de dimension intrinsèque beaucoup plus petite.

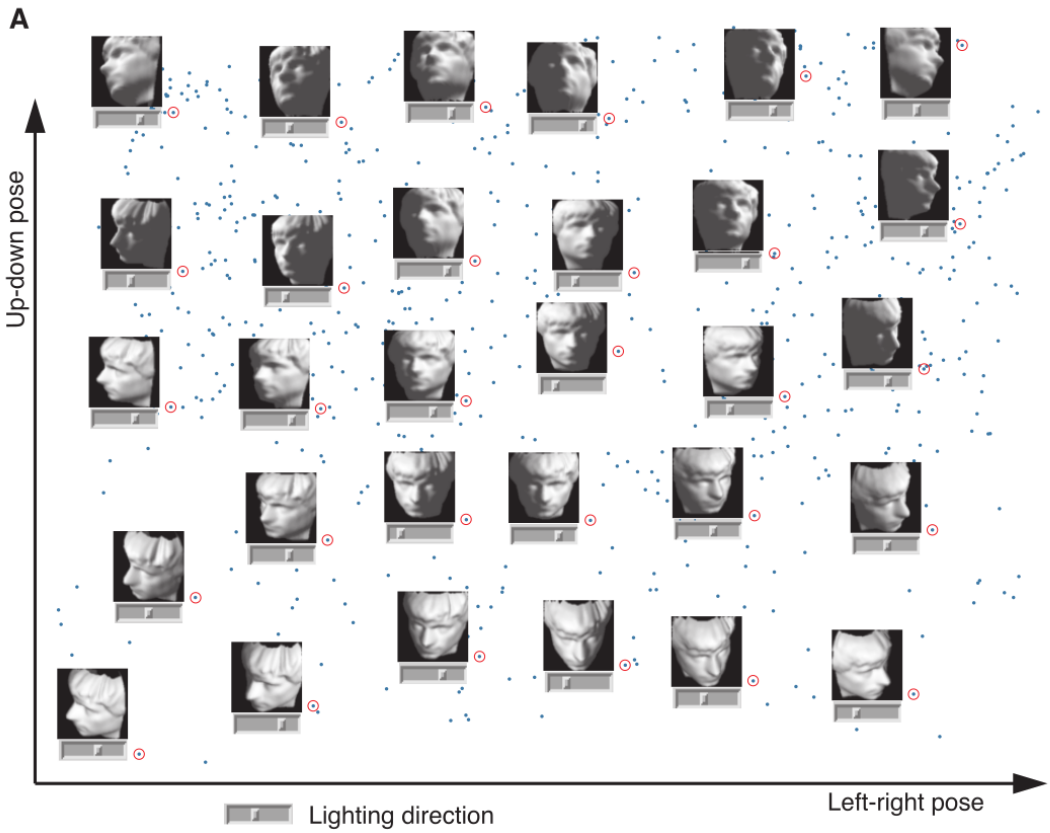

Figure 1.3: Visage vu sous différents angles
Données de dimension intrinsèque bien inférieure à la dimension ambiante
(image de [19])

Un autre exemple célèbre, voir [6, Section 2], montre que l'espace des patchs 3×3 d'une certaine collection d'images, après normalisation, possède la topologie d'une bouteille de Klein, voir Figure 1.4.

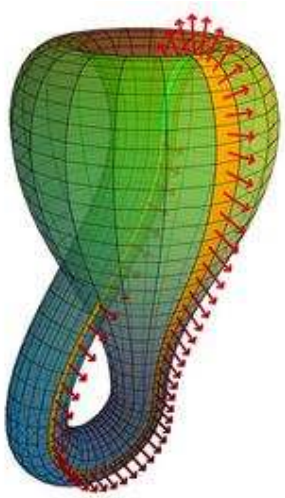

Figure 1.4: Bouteille de Klein dans $\mathbb{R}^3$
(image de K. Polthier, in [6])

Dans les deux premiers cas, on voit que la structure géométrique interne aux données résulte de contraintes physiques (l'articulation du bras, l'angle de la prise de photo). Il est donc attendu que ce phénomène soit récurrent. Notez bien que l'hypothèse de la variété dit deux choses : premièrement, les données ont une géométrie interne (c'est-à-dire qu'il est possible de donner du sens au fait que deux données X_1, X_2 soient proches, indépendamment de la façon dont on *représente* ces données), et deuxièmement, elles ont une dimension intrinsèque qui ne dépend pas de la dimension utilisée pour les collecter ou représenter, et qui est souvent beaucoup plus petite.

D'un point de vue probabiliste, l'hypothèse de la variété peut être vue de façon heuristique[7] comme une conséquence[8] du fait que les données possèdent des corrélations *internes*. Cela signifie la chose suivante. Soient

$$X_1, \dots, X_n \in \mathbb{R}^D$$

des données iid à valeurs dans $\mathbb{R}^D$ avec D grand. Il s'ensuit que

$$X_1 = (X_1^1, X_1^2, \dots, X_1^D).$$

Le nombre de coordonnées étant très grand, on observe souvent des corrélations entre elles, donc les $X_1^1, \dots, X_1^D$ seront des variables aléatoires réelles *corrélées*. On pourra alors écrire ces corrélations sous la forme

$$\text{a.s.} \quad \varphi(X_1^1, \dots, X_1^D) = 0$$

pour une certaine fonction $\varphi : \mathbb{R}^D \to \mathbb{R}^l$ représentant les contraintes suivies par les coordonnées. Notez que l'on parle de corrélations *internes* parce que ce sont les coordonnées des données dans la représentation que l'on a qui sont corrélées, et non pas les données elles-mêmes, que l'on suppose iid.

Un exercice d'analyse multivariée de niveau licence permet d'affirmer que si φ est suffisamment régulière, alors la pré-image $\varphi^{-1}(0)$ est une variété de dimension $D - l$, et donc les données sont supportées sur cette variété de dimension intrinsèque $d = D - l \ll D$ dès lors que l est suffisamment grand, ce qui correspond exactement à l'hypothèse de la variété.

Exercice : Quelles hypothèses doit-on rigoureusement supposer pour montrer que la pré-image $\varphi^{-1}(0)$ est effectivement une variété ? (**indice :** se rappeler du théorème des fonctions implicites).

Parmi les premières personnes à avoir étudié l'hypothèse de la variété, Belkin et Niyogi (2003) l'ont introduite comme un modèle permettant de pallier le fléau de la dimension. En effet, comme vu à la

[7] nous laissons les lecteurs comprendre pourquoi cela n'est pas totalement rigoureux
[8] et donc pas une hypothèse !

section 1.1.2, le fait que les données soient en grande dimension entraîne quasi systématiquement qu'elles sont éloignées les unes des autres, et par conséquent la taille de l'échantillon devrait dépendre de façon exponentielle de la dimension pour obtenir des résultats statistiquement efficaces. Autrement dit, si l'on est en dimension D, étant donné un seuil d'erreur $\varepsilon > 0$, on devrait avoir des échantillons de l'ordre de

$$n = \frac{1}{\varepsilon^D}$$

afin de pouvoir faire correctement des statistiques. Or, pour $D \approx 10^6$ et $\varepsilon \approx 10^{-1}$, cela donne des tailles d'échantillons inconcevables.

L'hypothèse de la variété, telle que présentée par Belkin et Niyogi, consiste donc, lorsque les données sont de grande dimension, à les modéliser comme étant à valeurs dans une variété de plus basse dimension, à laquelle on peut ajouter un bruit (voir Figures 1.6 et 1.5). Formulée de façon statistique, cela revient à dire que, étant donné $X_1, \ldots, X_n \in \mathbb{R}^D$, on considère des modèles statistiques $(P_\theta)_\theta$ où chaque loi de probabilité P_θ est supportée sur une sous-variété de dimension plus petite que D.

Hypothèse. *(hypothèse de la variété : modèle statistique en grande dimension)*
Les données proviennent d'un tirage sur une sous-variété, plus un bruit.

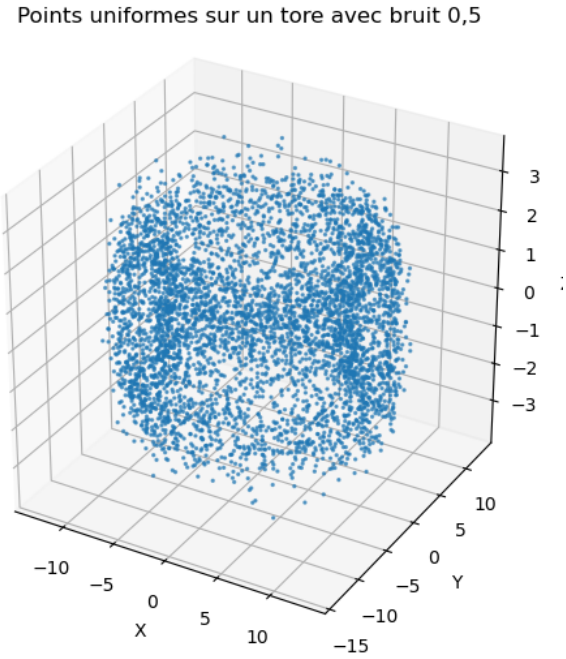

Figure 1.5: Tore + bruit

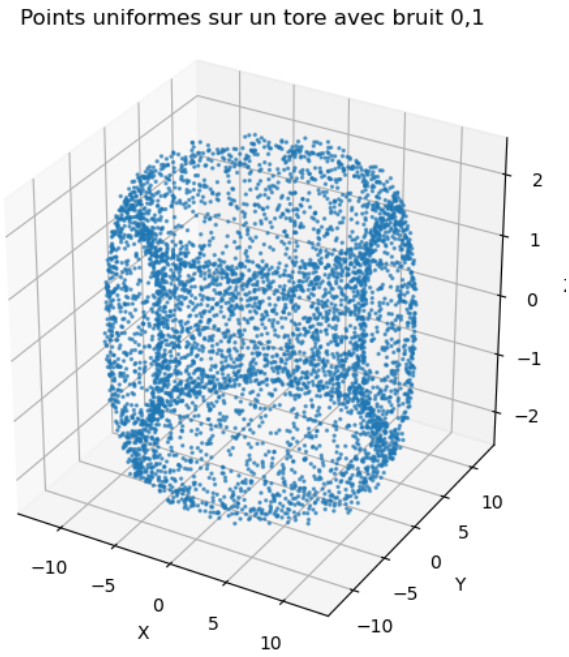

Figure 1.6: Tore + bruit

L'hypothèse de la variété a de nombreuses conséquences conceptuelles et interprétations. Pour illustrer ces conséquences, prenons l'exemple de la classification d'images de chiens et de chats. L'hypothèse de la variété affirme que l'ensemble des photos de chats possède une structure géométrique intrinsèque, et de même pour l'ensemble des photos de chiens. Le problème de classification revient alors à être capable de "séparer" ces deux sous-variétés de l'espace des images, qui sont entrelacées. Voir Figure 1.7.

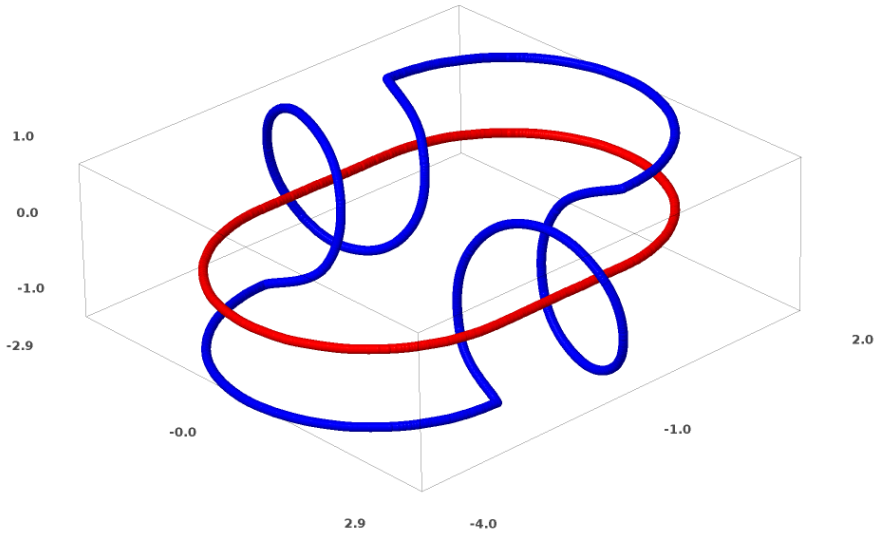

Figure 1.7: Sous-variétés entrelacées
Problème de classification
(image de C. Olah)

Comme mentionné précédemment, l'hypothèse de la variété naît d'un aller-retour entre pratique et théorie.

D'une part, la théorie statistique "classique", à cause du fléau de la dimension, prédit des convergences extrêmement lentes (voire inobservables), alors qu'en pratique, de nombreux algorithmes qui traitent des données de grande dimension fonctionnent malgré tout en temps raisonnable. L'hypothèse de la variété peut alors être vue comme une explication de ce phénomène, car les vitesses de convergence effectives dépendraient de la dimension intrinsèque, qui reste de taille raisonnable.

D'autre part, supposer que les données sont supportées sur une variété, donc utiliser le modèle "manifold + noise", constitue un cadre théorique dans lequel on peut obtenir des résultats statistiques concordant avec les résultats observés en pratique. Cela fournit un cadre théorique cohérent pour l'étude.

Ces deux aspects peuvent se résumer ainsi :

1. Apprendre la géométrie et la dimension intrinsèques des données afin d'en tirer des informations statistiques utiles ; ce sera le chapitre 3.

2. Supposer que l'on connaît la géométrie des données, et établir sous quelles conditions géométriques une théorie de l'estimation statistique ou de l'apprentissage est disponible ; ce sera le chapitre 4.

Concluons en mentionnant qu'apprendre la dimension intrinsèque des données correspond aux algorithmes de *réduction de dimension*. Dans notre cadre géométrique, il s'agira de méthodes de réduction de dimension *non linéaires*.

En particulier, la réduction de dimension peut être utilisée afin de rendre possibles des calculs qui seraient autrement trop coûteux, mais elle sert également comme un outil de *statistique descriptive*, c'est-à-dire qu'elle permet de projeter les données en dimension 2 ou 3, en respectant d'une certaine façon leur géométrie, et de les rendre ainsi *visualisables*, de façon analogue aux statistiques descriptives usuelles telles que les quantiles ou les diagrammes en boîte.

1.2.2 Géométrie de l'information

Un sujet très riche et proche, bien que différent de la géométrie des données traitée dans ce cours, est la *géométrie de l'information*. Bien qu'il ne sera pas question de géométrie de l'information dans ce cours, nous esquissons ici, de façon très grossière, sa "définition" afin d'en souligner les différences avec la géométrie des données.

Étant donné $X_1, \dots, X_n \in \mathbb{R}^D$, on considère un modèle statistique $(P_\theta)_{\theta \in \Theta}$ avec $\Theta \subset \mathbb{R}^k$ un ouvert. La géométrie de l'information consiste à voir le paramètre θ comme un système de coordonnées, et donc à voir le modèle $(P_\theta)_{\theta \in \Theta}$ comme une variété équipée d'une structure métrique[9]. Par exemple, la divergence de Kullback-Leibler induit une telle structure, et possède notamment de nombreux liens avec l'estimation par maximum de vraisemblance.

On retiendra donc qu'en géométrie des données, on considère une géométrie sur les données elles-mêmes $X_1, \dots, X_n$, tandis qu'en géométrie de l'information, on considère une géométrie sur le modèle statistique que l'on a choisi.

[9] définie à partir de l'information de Fisher, d'où le nom "géométrie de l'information"

Chapter 2

Qu'est-ce que la géométrie ?

Dans cette section, nous présentons quelques éléments de géométrie qui seront nécessaires pour la suite du cours. Nous discutons les concepts et présentons les principaux outils techniques.

Le mot *géométrie* vient de "geo", en référence à Gaïa, la déesse grecque de la Terre, et de "métrie", signifiant "mesure de". Ainsi, au sens premier, la géométrie est la mesure de la Terre. Rien d'abstrait, rien d'axiomatique, aucune quête de démonstrations parfaitement rigoureuses, simplement la capacité de *mesurer* le monde dans lequel nous vivons.

À première vue, le sens de la géométrie peut sembler avoir perdu ses racines, puisque les géomètres modernes étudient des sujets aussi abstraits que la topologie algébrique ou la géométrisation des variétés de dimension 3 de Thurston. Cependant, la sphère $\mathbb{S}^2 \subset \mathbb{R}^3$, définie par l'équation algébrique $x^2 + y^2 + z^2 = 1$, est elle-même une variété algébrique, modèle de notre terre ronde. Et puisque notre espace semble être tridimensionnel, la géométrisation de Thurston n'est rien de moins que la classification de tous les espaces possibles dans lesquels nous pourrions vivre; car, selon la relativité générale d'Einstein, l'espace (ou plutôt l'espace-temps) dépend de la matière qu'il contient.

Ainsi, si nous convenons de ne pas interpréter "geo", la Terre, trop littéralement comme la planète que nous habitons, mais plus largement comme le monde dans lequel nous vivons, alors "Géométrie" demeure le terme parfait pour décrire notre objectif : mesurer le monde.

2.1 Géométrie Riemannienne

Dans sa thèse d'habilitation de 1854 [16], intitulée *Über die Hypothesen, welche der Geometrie zu Grunde liegen* (Sur les hypothèses qui servent de fondement à la géométrie), Bernhard Riemann, qui venait d'achever sa thèse de doctorat sous la direction de Gauss, pose la question de ce qui définit la notion d'espace.

Pendant très longtemps, la géométrie s'est attachée à l'étude des figures dans l'espace, et elle a procédé a priori, comme dirait Kant, c'est-à-dire qu'elle partait d'un ensemble de postulats, principalement les axiomes d'Euclide, et en déduisait, de manière purement logique, des propriétés concernant les rapports de longueurs, les angles ou le concours de droites.

Les philosophes, constatant que la vérité des résultats obtenus en géométrie, c'est-à-dire leur adéquation avec le monde réel, dont la géométrie sert ultimement la compréhension par la mesure, dépendait de manière cruciale de la vérité des axiomes, ont commencé à examiner ces axiomes de plus près. C'est précisément ici que nous rencontrons Riemann, réfléchissant aux hypothèses qui constituent les fondements de la géométrie, dans une dissertation *mathématique* que beaucoup aujourd'hui jugeraient plutôt philosophique, notamment parce qu'elle contient très peu de formules mathématiques ; il s'agit essentiellement d'une quinzaine de pages de texte ! Et l'une des premières choses qui frappe le lecteur, et qui frappa Riemann lui-même lorsqu'il examina les axiomes, est qu'ils traitent des propriétés des figures contenues dans l'espace (droites, cercles, etc.), mais non de l'espace lui-même, qui les contient. Ne devrait-on pas pourtant attendre des fondements de la géométrie qu'ils portent sur *l'espace lui-même*, et qu'à partir de là on puisse déduire le comportement des objets qu'il contient ?

Pour Riemann, la notion d'espace coïncide avec le concept de grandeur. Ce qui importe, c'est la manière dont cette grandeur est mesurée : son *mode de détermination*. La difficulté tient alors au fait que, même si l'on sait mesurer une longueur dans une direction donnée, cela ne détermine pas entièrement le concept lorsque la grandeur possède plusieurs dimensions : il faut comprendre comment les mesures effectuées dans différentes directions interagissent entre elles. Il résume cette idée de la manière suivante :

> *Il en résultera qu'une grandeur de dimensions multiples est susceptible de diverses relations métriques, et que l'espace n'est donc qu'un cas particulier d'une grandeur à trois dimensions. Il s'ensuit nécessairement que les propositions de la géométrie ne peuvent être déduites des concepts généraux de grandeur, mais que les propriétés par lesquelles l'espace se distingue de toute autre grandeur tridimensionnelle concevable ne peuvent être tirées que de l'expérience. De là naît le problème de trouver les faits les plus simples par lesquels les relations métriques de l'espace peuvent être déterminées.* [...]
> — Bernhard Riemann, 1854

Il s'agit donc de déterminer les relations métriques de l'espace, c'est-à-dire la manière dont les longueurs se comportent les unes par rapport aux autres lorsqu'on les considère dans plusieurs directions différentes. Considérons le cas où le mode de mesure est *continu*, ce qui signifie que les résultats possibles sont des nombres réels, et supposons que nous ayons affaire à une grandeur de dimension n, c'est-à-dire possédant n directions indépendantes le long desquelles des mesures peuvent être effectuées *indépendamment*. Enfin, supposons que nous nous déplacions d'un point A vers un point B, et qu'à l'issue de ce déplacement nous ayons mesuré, le long de chacune de ces directions indépendantes, une "longueur" dx_i, $i = 1, \cdots, n$. Nous demandons alors : quelle longueur ds doit-on attribuer au segment reliant le point A au point B ?

Puisque nous avons supposé la grandeur de dimension n, et puisque nous avons effectué des mesures dans n directions indépendantes, nous devons nécessairement pouvoir exprimer ds comme une fonction des dx_i ; sinon, nous ne serions pas en dimension n mais au moins en dimension $n+1$. La formule qui exprime ds en fonction des dx_i est ce que Riemann appelle les *relations métriques* de l'espace.

Si nous voulons que ces relations correspondent à la géométrie euclidienne — en particulier en dimension 2, nous sommes rapidement conduits au théorème de Pythagore, et donc à la relation métrique *quadratique* suivante :

$$ds^2 = dx_1^2 + \cdots + dx_n^2 \tag{2.1.1}$$

Rappelons maintenant que l'objectif de Riemann est d'examiner les fondements ; il ne peut donc pas s'arrêter au théorème de Pythagore, déjà bien compris et découlant directement des axiomes d'Euclide. La question devient alors : quelles sont les relations métriques les plus générales que l'on puisse concevoir ? Bien sûr, on pourrait simplement affirmer qu'il existe une certaine fonction exprimant ds en fonction des dx_i, mais un tel niveau de généralité ne permet pas de saisir la notion d'espace que nous cherchons à décrire.

Il s'agit là d'un fait récurrent en mathématiques : il existe toujours un compromis entre la *généralité* d'un énoncé, c'est-à-dire le nombre d'objets qu'il englobe, et son *caractère informatif*, c'est-à-dire la quantité d'information qu'il fournit. Les deux cas extrêmes consistent soit à être très imprécis tout en parlant de beaucoup de choses (dire peu sur beaucoup), soit à être très précis tout en parlant de peu de choses (dire beaucoup sur peu).

Dans le cas présent, nous voulons une forme très générale pour les relations métriques de l'espace, mais en même temps nous souhaitons que cette forme demeure compatible avec une notion proche de celle que nous nous faisons de l'espace. Ce que Riemann observe, c'est que la relation métrique précédente, issue du théorème de Pythagore, possède la propriété cruciale d'être quadratique en les dx_i. Par conséquent, exprimer une relation quadratique générale de la forme

$$ds^2 = \sum_{i,j=1}^{n} g_{ij}\, dx_i\, dx_j$$

peut être vu comme l'extension la plus naturelle du théorème de Pythagore, et constitue le fondement de ce que l'on appelle aujourd'hui la géométrie riemannienne.[1]

[1] "Des relations encore plus compliquées peuvent apparaître lorsqu'on ne suppose plus que l'élément linéaire puisse être représenté par la racine carrée d'une expression différentielle du second degré." - Riemann, 1854. C'est précisément ce qui donnera plus tard naissance à ce que l'on appelle aujourd'hui la géométrie finslérienne.

Pour préciser les choses, nous concevrons ces éléments de longueur dx_i comme étant très petits, de sorte que la formule exprimant ds en fonction des dx_i ne soit valable qu'au sens *infinitésimal*. Afin de commencer à formuler ces idées dans le langage des mathématiques modernes, nous ne considérerons plus ces éléments de longueur comme des quantités infinitésimales, mais plutôt comme des directions que l'on peut emprunter en s'éloignant d'un point x. Nous représentons ces directions par des vecteurs attachés au point x, et l'ensemble de tous ces vecteurs forme ce que l'on appelle l'*espace tangent* en x.

Puisque nous sommes en dimension n, l'ensemble de tous les vecteurs possibles que l'on peut concevoir est, en tant qu'espace vectoriel, isomorphe à $\mathbb{R}^n$.[2]

Si les dx_i sont des vecteurs de dimension n attachés au point x, il s'ensuit que les quantités $g_{ij} = g_{ij}(x)$ sont également attachées au point x et forment une matrice $n \times n$,

$$g_x = (g_{ij}(x))_{ij},$$

appelée le *tenseur métrique*. La forme des relations métriques de l'espace peut alors se réécrire

$$ds^2 = g_x(dx, dx),$$

où le vecteur tangent est donné par $dx = (dx_1, \cdots, dx_n)$. Le tenseur métrique en un point x est donc une forme quadratique définie sur l'espace tangent en x. De plus, cette forme quadratique doit satisfaire certaines propriétés naturelles : elle doit être positive, $g_x(u, u) \geq 0$, puisqu'elle représente une longueur ; elle doit également être définie, $g_x(u, u) = 0 \Rightarrow u = 0$, de sorte que le seul vecteur de déplacement qui ne nous déplace pas soit le vecteur nul ; enfin, elle doit être symétrique, $g_x(u, v) = g_x(v, u)$. Tout cela peut se résumer en disant que le tenseur métrique est un produit scalaire sur l'espace tangent.[3]

La connaissance locale des relations métriques de l'espace suffit à en déduire la structure globale, en ce sens que si l'on souhaite connaître la distance parcourue le long d'un chemin d'un point A à un point B, il suffit d'intégrer la relation de l'élément de longueur infinitésimal ds le long de ce chemin, du point A à un point infiniment proche $A + dx$, et ainsi de suite jusqu'à atteindre B en suivant toujours la même trajectoire.

Plus rigoureusement, un chemin de A à B est représenté par une application différentiable à valeurs dans notre espace $\mathcal{M}$, donnée par

$$\gamma : [0, 1] \to \mathcal{M},$$

telle que $\gamma(0) = A \in \mathcal{M}$ et $\gamma(1) = B \in \mathcal{M}$. À chaque instant du chemin, $t \in (0, 1)$, nous pouvons calculer la dérivée $\dot{\gamma}(t)$ en cet instant, c'est-à-dire le vecteur vitesse au point $\gamma(t)$. Il s'agit d'un vecteur dans l'espace tangent attaché au point $\gamma(t) \in \mathcal{M}$.

Le tenseur métrique nous permet alors de calculer la longueur de ce vecteur vitesse, c'est-à-dire la vitesse (scalaire) effective, et il ne reste plus qu'à intégrer cette quantité pour obtenir la longueur totale parcourue :

$$\text{length}(\gamma) = \int_0^1 \sqrt{g_x(\dot{\gamma}(t), \dot{\gamma}(t))}\, dt.$$

Notons la présence de la racine carrée, qui garantit que l'expression possède bien la dimension d'une longueur, puisque le tenseur métrique g définit une forme quadratique.

Nous renvoyons à [12] pour une monographie complète.

2.1.1 Variétés différentielles

On part d'un ensemble de points $\mathcal{M}$, que l'on conçoit comme "l'espace" qui nous intéresse, et notre objectif est d'être capable de se repérer sur cet espace. On définit donc la notion de coordonnées (on parle de cartes par référence à la cartographie, la science s'occupant de représenter graphiquement des informations géographiques).

Définition 2.1.1. Soient $p_0 \in V \subset \mathcal{M}$, avec V un ouvert. On dit qu'une application bijective

$$\psi : V \to \mathbb{R}^d,$$

continue et dont la bijection réciproque est également continue, est une *carte locale* en p_0.

[2] C'est un théorème élémentaire d'algèbre linéaire que tout espace vectoriel de dimension n est isomorphe à $\mathbb{R}^n$.

[3] Cela munit ainsi chaque espace tangent T_x d'une structure d'espace de Hilbert $(\mathbb{R}^d, g_x)$.

Autrement dit, une carte est un homéomorphisme local.
La condition de bijectivité est importante : elle assure que tout point $p \in V$ a un unique ensemble de coordonnées

$$\psi(p) = (\psi_1(p), \dots, \psi_d(p)) \in \mathbb{R}^d,$$

et que réciproquement, à tout ensemble de coordonnées $(x_1, \dots, x_d) \in \mathbb{R}^d$ correspond un unique point

$$p = \psi^{-1}(x_1, \dots, x_d) \in \mathcal{M}.$$

Définition 2.1.2. On dit qu'une famille de cartes $(\psi_i : V_i \to \mathbb{R}^d)_{i\in I}$ est un *atlas* lorsque les cartes recouvrent tout l'espace :

$$\bigcup_{i\in I} V_i = \mathcal{M}.$$

Une fois que l'on a une façon de se repérer dans l'espace, on souhaite pouvoir effectuer des calculs à partir de ces coordonnées. Pour cela, on a besoin que les cartes soient des applications différentiables. Étant donné que l'on sait dériver une fonction définie sur $\mathbb{R}^d$, on utilisera les cartes $\psi : V \to \mathbb{R}^d$ pour effectuer les calculs dans $\mathbb{R}^d$, puis on reviendra dans $\mathcal{M}$ en utilisant la réciproque ψ^{-1}.

Rappelons qu'un *difféomorphisme* de classe $\mathcal{C}^k$ entre deux ouverts $U, V \subset \mathbb{R}^d$ est une application bijective $\psi : U \to V$, de classe $\mathcal{C}^k$, dont la bijection réciproque $\psi^{-1} : V \to U$ est également de classe $\mathcal{C}^k$. On dit que deux ouverts sont difféomorphes s'il existe un difféomorphisme de classe $\mathcal{C}^k$ entre eux. Notez que pour que deux ouverts $U \subset \mathbb{R}^{d_1}$ et $V \subset \mathbb{R}^{d_2}$ soient difféomorphes, il faut nécessairement que leurs dimensions coïncident : $d_1 = d_2$.

Définition 2.1.3. Une *variété différentielle* de classe $\mathcal{C}^k$ et de dimension d est un espace $\mathcal{M}$ qui est localement difféomorphe à $\mathbb{R}^d$: il existe un atlas $(\psi_i : V_i \to \mathbb{R}^d)_{i\in I}$, avec $V_i \subset \mathcal{M}$, tel que les *changements de coordonnées*

$$\psi_i \circ \psi_j^{-1} : \mathbb{R}^d \to \mathbb{R}^d$$

soient des difféomorphismes.

Un changement de coordonnées est défini dès lors que deux systèmes de coordonnées sont définis sur un même voisinage V d'un point p : $\psi_i : V \to \mathbb{R}^d$ et $\psi_j : V \to \mathbb{R}^d$. Dans ce cas, la composition $\psi_i \circ \psi_j^{-1} : \mathbb{R}^d \to \mathbb{R}^d$ est bien définie, voir Figure 2.1.

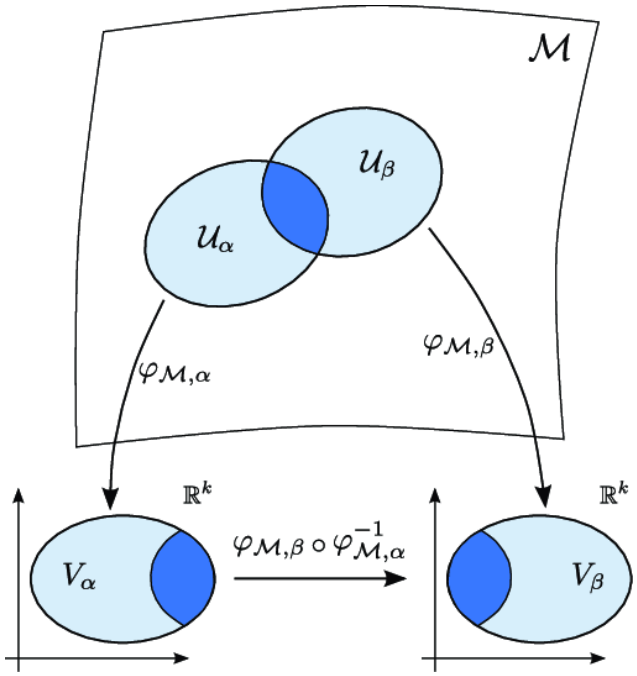

Figure 2.1: Changement de coordonnées, image issue de [14]

Cette condition de compatibilité, demandant que les changements de coordonnées soient des difféomorphismes, est essentielle : elle garantit que le fait qu'une application $f : \mathcal{M} \to \mathbb{R}$ soit différentiable en un point $p \in \mathcal{M}$ ne dépend pas de la carte utilisée pour le vérifier.

Définition 2.1.4. On dit qu'une fonction $f : \mathcal{M} \to \mathbb{R}$ est *différentiable* en $p \in \mathcal{M}$ lorsqu'il existe un voisinage $p \in V_p \subset \mathcal{M}$ et une carte $\psi : V_p \to \mathbb{R}^d$ tels que l'application composée $f \circ \psi^{-1} : \mathbb{R}^d \to \mathbb{R}$ soit dérivable en $\psi(p) \in \mathbb{R}^d$. La *différentielle* de f en p est définie par

$$\begin{aligned} df_p : \mathbb{R}^d &\to \mathbb{R} \\ u &\mapsto d(f \circ \psi^{-1})_{\psi(p)}(u). \end{aligned}$$

Cette formule ne dépend pas du choix de la carte ψ ni du voisinage V_p.

La différentielle de f en $p \in \mathcal{M}$, dans la direction $u \in \mathbb{R}^d$, représente la variation de f au voisinage de p dans la direction u. Ici, le vecteur $u \in \mathbb{R}^d$ représente donc une direction dans laquelle on peut se déplacer en partant de p. Il s'agit d'un vecteur *attaché* à p, appelé *vecteur tangent* à p.

L'ensemble des vecteurs tangents à un point p forme un espace vectoriel de dimension d, la dimension de la variété $\mathcal{M}$, et est donc isomorphe à $\mathbb{R}^d$. L'espace $\mathbb{R}^d$ sur lequel est défini la différentielle dans la définition précédente doit être compris comme l'*espace tangent* $T_p\mathcal{M}$ à $\mathcal{M}$ en p, et non comme un espace euclidien canonique.

En particulier, si l'on dérive la fonction f en un autre point $p' \in \mathcal{M}$, la différentielle en p' sera définie sur *un autre* espace tangent $T_{p'}\mathcal{M}$, toujours isomorphe à $\mathbb{R}^d$. Une façon correcte d'y penser est de considérer $T_p\mathcal{M}$ comme l'espace formé de toutes les directions dans lesquelles peut se déplacer un point situé en p. Chaque carte $\psi : V_p \to \mathbb{R}^d$ fixe un isomorphisme entre $T_p\mathcal{M}$ et $\mathbb{R}^d$ via la différentielle de la carte : chaque vecteur $(u_1, \dots, u_d) \in \mathbb{R}^d$ est identifié à un vecteur tangent

$$(u_1\partial_{x_1}\psi, \dots, u_d\partial_{x_d}\psi) \in T_p\mathcal{M}.$$

L'espace tangent attaché à $p \in \mathcal{M}$ peut aussi être défini comme l'ensemble des vitesses en p des courbes dérivables passant par p. Une courbe à valeurs dans $\mathcal{M}$ est une application $\gamma : (-1, 1) \to \mathcal{M}$, et on dit qu'elle passe par p si $\gamma(0) = p$. Sa dérivée en 0 est donnée par

$$\gamma'(0) = (\psi^{-1} \circ \gamma)'(0) \in \mathbb{R}^d,$$

où $\psi : V_p \to \mathbb{R}^d$ est une carte, et cette dérivée ne dépend pas du choix de la carte grâce à la condition de compatibilité de la définition 2.1.3.

Définition 2.1.5. Soit $\mathcal{M}$ une variété différentielle et $p \in \mathcal{M}$. L'*espace tangent* à $\mathcal{M}$ en p est défini par

$$T_p\mathcal{M} := \left\{ \gamma'(0) \,\middle|\, \gamma \in \mathcal{C}^1((-1, 1); \mathcal{M}),\, \gamma(0) = p \right\}.$$

2.1.2 Variétés Riemanniennes

Avec les définitions précédentes, on a réussi à donner du sens à la notion de dérivée sur des espaces non euclidiens (plus généraux que $\mathbb{R}^d$). À ce stade, cependant, nous n'avons aucun moyen de *mesurer* la taille d'un vecteur tangent. En effet, l'espace tangent est certes isomorphe à $\mathbb{R}^d$, mais l'isomorphisme est donné par le choix d'une carte, c'est-à-dire d'un système de coordonnées.

Par conséquent, si l'on utilise cet isomorphisme pour déterminer la taille d'un vecteur, cette taille dépendra nécessairement du choix du système de coordonnées, ce qui est absurde : la hauteur d'un objet ne dépend pas du système de mesure utilisé !

Pour remédier à ce problème, on doit supposer que sur chaque espace tangent $T_p\mathcal{M}$, il existe une façon intrinsèque de mesurer, c'est-à-dire un produit scalaire. On peut alors définir la notion de *variété riemannienne*.

Définition 2.1.6. On dit que $\mathcal{M}$ est une *variété riemannienne* de dimension d si c'est une variété différentielle de classe $\mathcal{C}^\infty$ et que, pour tout point $p \in \mathcal{M}$, l'espace tangent $T_p\mathcal{M}$ est muni d'un produit scalaire

$$g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R},$$

tel que l'application $p \mapsto g_p$ soit lisse.

À retenir : g_p est une matrice symétrique définie positive de taille $d \times d$ et constitue un objet *intrinsèque*, c'est-à-dire que si l'on exprime g_p dans un système de coordonnées, ses coefficients $(g_{ij})_{1\leq i,j\leq d}$ dépendent du choix de la carte, mais le résultat du produit scalaire $g_p(u,v) \in \mathbb{R}$ ne dépend pas du système de coordonnées.

En pratique, étant donnée une carte $\psi : \mathbb{R}^d \to V_p$, un vecteur $u \in T_p\mathcal{M}$ s'écrit $u = (u_i)_{i\leq d}$ et sa longueur est

$$g_p(u,u) = \sum_{i,j} g_{ij} u_i u_j.$$

Ici, les coefficients g_{ij} et les composantes u_i dépendent de la carte, mais le produit scalaire $g_p(u,u)$, lui, reste indépendant de la carte.

La philosophie derrière ce formalisme est la suivante :

> **Une propriété est de nature *géométrique* lorsqu'elle ne dépend pas du système de coordonnées choisi.**

Ainsi, en pratique, on utilise des coordonnées pour se repérer et effectuer des calculs, mais on cherche des propriétés *invariantes par changement de coordonnées*, car ce sont ces propriétés qui fournissent une véritable information sur la nature de notre problème.

Dans notre cadre, on dispose de données $X_1, \ldots, X_n$ qui vivent sur une variété $\mathcal{M}$, mais ces données sont fournies sous forme de vecteurs dans $\mathbb{R}^D$, avec D très grand (de l'ordre de plusieurs millions). La variété $\mathcal{M}$, de dimension d beaucoup plus petite ($d \ll D$), est *plongée* dans $\mathbb{R}^D$. En cherchant de l'information sur la structure intrinsèque des données, et non dépendante de la représentation dans $\mathbb{R}^D$, on recherche donc l'information de nature géométrique sur $\mathcal{M}$.

La propriété la plus évidente qui ne dépend pas du choix d'un atlas est le nombre de composantes connexes. Cette propriété est de nature *topologique* : elle ne nécessite pas de notion de métrique pour être définie, la donnée de la topologie (c'est-à-dire de l'ensemble des ouverts) suffit.

2.1.3 Distance, géodésiques, coordonnées normales

La notion de distance ne dépend pas non plus du choix des coordonnées : c'est bel et bien une notion géométrique, et heureusement ! Comme mentionné précédemment, une courbe lisse reliant un point p à un point q est une application de classe $\mathcal{C}^\infty$:

$$\gamma : [0,1] \to \mathcal{M}$$

telle que $\gamma(0) = p$ et $\gamma(1) = q$. La distance parcourue en suivant cette courbe est définie par

$$\text{length}(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\, dt.$$

L'idée est simple : la vitesse instantanée est la longueur du vecteur vitesse, on multiplie par le temps dt et on intègre.

On souhaite ensuite définir la distance entre deux points p et q. La distance doit être la longueur minimale à parcourir pour aller de l'un à l'autre, peu importe le chemin choisi. Cela se traduit par l'infimum sur tous les chemins possibles. La *distance riemannienne* d_g induite par le tenseur métrique g est donc

$$d_g(p,q) := \inf_{\gamma : p \to q} \text{length}(\gamma),$$

où l'infimum est pris sur toutes les courbes γ reliant p à q.

On peut facilement vérifier que d_g est symétrique : $d_g(p,q) = d_g(q,p)$, car $\text{length}(\gamma(t)) = \text{length}(\gamma(-t))$. Une courbe qui atteint l'infimum, c'est-à-dire une courbe qui réalise le plus court chemin entre deux points, est appelée une *géodésique minimisante*.[4]

Exercice : Vérifier que dans le cas euclidien $(\mathbb{R}^d, I_d)$, on retrouve la distance classique et que les géodésiques sont les droites.

[4] La notion de géodésique en géométrie riemannienne est plus générale que celle de courbe minimisant la longueur.

Étant donnés deux points p et q sur une variété riemannienne, il est toujours possible de trouver une géodésique minimisante. Cependant, elle n'est pas forcément unique : pensez aux pôles Nord et Sud de la Terre, reliés par tous les méridiens. Si p et q sont suffisamment proches, il existe toujours une unique géodésique minimisante. Pour l'exemple de la Terre, seuls les points antipodaux sont reliés par plusieurs géodésiques minimisantes.

Une construction naturelle est la suivante : étant donné un point p et un vecteur $v \in T_p\mathcal{M}$ tangent à p, on veut définir le point q atteint si l'on part de p dans la direction de v et que l'on parcourt une distance égale à $\|v\|_g = \sqrt{g_p(v,v)}$. Si cette distance est suffisamment petite, la géodésique minimisante est unique. Cela permet de définir l'*application exponentielle au point* p[5].

Définition 2.1.7. On définit l'application exponentielle au point p :

$$\exp_p : T_p\mathcal{M} \to \mathcal{M}$$

en posant, pour tout $u \in T_p\mathcal{M}$,

$$\exp_p(u) = \text{la valeur au temps 1 de la géodésique minimisante } \gamma \text{ telle que } \gamma(0) = p,\ \dot\gamma(0) = u.$$

Si $\|u\|_g$ est assez petit, cette application est bien définie, mais elle n'est en général pas définie sur tout $T_p\mathcal{M}$.

La plus grande boule $B(0,r) \subset T_p\mathcal{M} \simeq \mathbb{R}^d$ sur laquelle l'application exponentielle est bien définie est appelée le *domaine d'injectivité*, et le plus grand rayon $r_{\max}$ est le *rayon d'injectivité*. Alors

$$\exp_p : B(0, r_{\max}) \to \mathcal{M}$$

est un difféomorphisme. Comme la boule ouverte $B(0, r_{\max}) \subset \mathbb{R}^d$ est difféomorphe à $\mathbb{R}^d$[6], il s'agit d'une carte. L'application exponentielle fournit ainsi des coordonnées locales très particulières autour de p, appelées *coordonnées normales*.

Exercice : Montrer que dans l'espace euclidien $(\mathbb{R}^d, I_d)$, les coordonnées normales en 0 coïncident avec les coordonnées cartésiennes.

La morale est que les coordonnées normales constituent un choix canonique de coordonnées autour d'un point, mais ce n'est pas parce qu'une formule est vérifiée en coordonnées normales qu'elle est indépendante du choix de la carte. Seules les propriétés invariantes par changement de coordonnées sont véritablement de nature géométrique.

2.1.4 Volume Riemannien

De même que dans l'espace euclidien $\mathbb{R}^d$, la notion de distance permet de définir la notion de *volume* : une fois que l'on a défini les mètres, on peut définir les mètres carrés, les mètres cubes, etc. Dans $\mathbb{R}^d$, le volume naturel, c'est-à-dire celui associé à la distance euclidienne (la norme $\|\cdot\|_2$), est donné par la mesure de Lebesgue. Celle-ci associe un nombre réel positif ou nul à chaque partie *mesurable* $A \subset \mathbb{R}^d$, de façon cohérente avec la distance euclidienne. En particulier, si $R : \mathbb{R}^d \to \mathbb{R}^d$ est une isométrie, la mesure de Lebesgue est préservée :

$$\lambda(R(A)) = \lambda(A).$$

Dans le cas général où la métrique est Riemannienne, il existe également une notion naturelle de volume pour les sous-ensembles de la variété : le *volume Riemannien*.

Étant donnée une variété Riemannienne $(\mathcal{M}, g)$ et une carte $\psi : V \to \mathbb{R}^d$ pour un ouvert $V \subset \mathcal{M}$, le volume Riemannien d'un ensemble mesurable $A \subset V$ est défini par

$$\mathrm{vol}_g(A) = \int_{\psi(A)} \sqrt{\det\left(g_{\psi^{-1}(x_1,\dots,x_d)}\right)}\, dx_1 \cdots dx_d.$$

[5] Ce terme est utilisé par analogie avec la théorie des groupes de Lie.

[6] À vérifier.

Cette formule se comprend ainsi : en chaque point $p \in \mathcal{M}$, le tenseur métrique g_p est une matrice $d \times d$ symétrique définie positive. Son déterminant est strictement positif. On se ramène ensuite sur $\mathbb{R}^d$ via la carte ψ et on intègre par rapport à la mesure de Lebesgue $dx_1 \cdots dx_d$.

Pour que cette définition soit bien géométrique, il faut vérifier deux points :

i) La quantité $\mathrm{vol}_g(A)$ ne dépend pas du choix de la carte ψ. L'unité de mesure du volume peut varier selon la carte, mais pas le volume lui-même.

ii) Si A n'est pas entièrement contenu dans un unique ouvert V muni d'une carte, on découpe A selon un atlas $(V_i)_i$ et on définit

$$\mathrm{vol}_g(A) = \sum_i \mathrm{vol}_g(A \cap V_i),$$

il faut alors vérifier que ce découpage ne dépend pas du choix de l'atlas.

On retiendra que, de façon infinitésimale et exprimée en coordonnées, la densité de la mesure volume Riemannienne est

$$d\mathrm{vol}_g(x) = \sqrt{\det(g_x)}\, dx,$$

où dx est la mesure de Lebesgue usuelle. Cette écriture dépend du choix de la carte, mais la mesure volume Riemannienne sous-jacente reste intrinsèque à la variété.

2.1.5 Courbures

La propriété la plus cruciale en géométrie Riemannienne, qui est purement géométrique et ne dépend absolument pas des systèmes de coordonnées ni du plongement choisi, est la notion de *courbure*.

Intuitivement, la courbure mesure à quel point un espace n'est pas plat. L'espace plat étant l'espace Euclidien, dans lequel le théorème de Pythagore (2.1.1) est valable et où la courbure est nulle. Ceci équivaut à ce que le tenseur métrique soit égal à la matrice identité en tout point et dans toutes les cartes : $g = I_d$.

C'est une notion cruciale en géométrie Riemannienne car la connaissance du tenseur de courbure, appelé *tenseur de Riemann*, permet de reconstruire la métrique g. Différentes notions de courbure existent, et chacune peut se définir à partir du tenseur de courbure. La définition rigoureuse de ce tenseur dépasse le cadre de ce cours, mais nous allons toutefois définir la notion de *courbure sectionnelle*, qui sera utilisée à la section 4.2.

Définition 2.1.8. Soit $(\mathcal{M}, g)$ une variété Riemannienne et $p \in \mathcal{M}$. Pour tout couple de vecteurs orthonormés $u, v \in T_p\mathcal{M}$, c'est-à-dire tels que

$$g_p(u,u) = g_p(v,v) = 1 \quad \text{et} \quad g_p(u,v) = 0,$$

il existe un réel $K(u,v) \in \mathbb{R}$ tel que

$$d_g(\exp_p(tu), \exp_p(tv))^2 = t^2\|u-v\|_2^2 - \frac{1}{3}K(u,v)\, t^4 + o(t^4),$$

où $\|\cdot\|_2$ désigne la norme euclidienne sur $T_p\mathcal{M}$. Le nombre $K(u,v)$ est appelé la *courbure sectionnelle*.

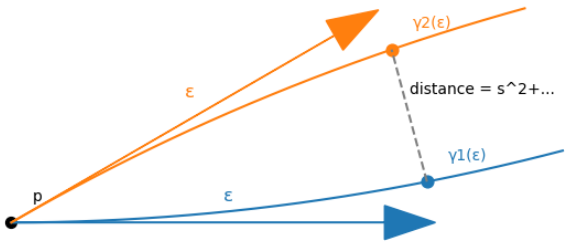

Figure 2.2: Courbure sectionnelle positive

La courbure sectionnelle mesure *localement* la déviation de deux géodésiques partant dans deux directions différentes.

- Si elle est nulle, l'écart entre les géodésiques partant du même point est linéaire (cas euclidien).
- Si elle est négative, cet écart est sur-linéaire, les géodésique tendent à s'éloigner (cas *hyperbolique*).
- Si elle est positive, cet écart est sous-linéaire, les géodésiques tendent à se rapprocher (cas *sphérique*).

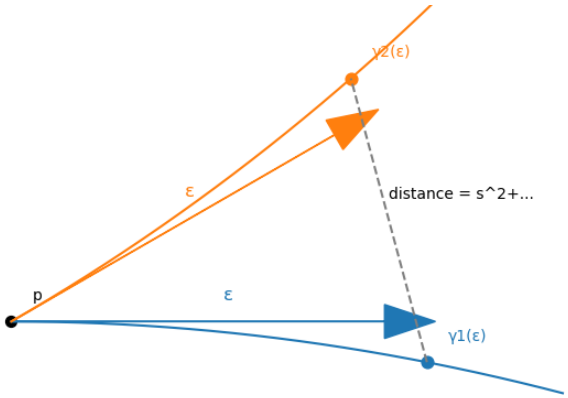

Figure 2.3: Courbure sectionnelle négative

2.1.6 Opérateur de Laplace-Beltrami

Étant donnés un point $p \in \mathcal{M}$ et un nombre $r > 0$, on peut considérer la *boule géodésique* $B(p, r)$ de centre p et de rayon r, c'est-à-dire l'ensemble des points situés à distance au plus r du point p pour la distance Riemannienne d_g :

$$B(p, r) := \{x \in \mathcal{M} \,|\, d_g(x, p) \leq r\}\,.$$

On peut également considérer les boules ouvertes, selon les besoins. Puisque l'on dispose d'une notion de volume intrinsèque, on définit le volume riemannien des boules géodésiques par $\mathrm{vol}_g(B(p, r))$.

Soit maintenant $f : \mathcal{M} \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. Pour tout $p \in \mathcal{M}$, on peut calculer le développement limité en $r = 0$ de la fonction

$$r \longmapsto \frac{1}{\mathrm{vol}_g(B(p, r))} \int_{B(p,r)} f(x)\, d\mathrm{vol}_g(x),$$

qui représente la valeur moyenne de f sur la boule $B(p,r)$. On obtient, pour un certain coefficient dépendant de f et de p, que l'on note $\Delta f(p)$,

$$\frac{1}{\mathrm{vol}_g(B(p,r))}\int_{B(p,r)} f(x)\,d\mathrm{vol}_g(x) = f(p) + \frac{r^2}{2(d+2)}\,\Delta f(p) + o(r^2), \tag{2.1.2}$$

où d est la dimension de la variété $\mathcal{M}$.

Définition 2.1.9. L'opérateur qui à une fonction $f \in \mathcal{C}^2(\mathcal{M},\mathbb{R})$ associe la fonction

$$p \longmapsto \Delta f(p)$$

défini par le développement de Taylor (2.1.2) est appelé *opérateur de Laplace-Beltrami* et est noté Δ.

Exercices :

1. Montrer que si $(\mathcal{M},g) = (\mathbb{R},1)$, alors l'opérateur de Laplace-Beltrami correspond à la dérivée seconde $f \mapsto f''$.
2. Montrer que si $(\mathcal{M},g) = (\mathbb{R}^d, I_d)$, alors l'opérateur de Laplace-Beltrami correspond au Laplacien

$$f \mapsto \sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}.$$

La morale à retenir est que le Laplacien d'une fonction en un point p mesure à quel point cette fonction dévie de sa moyenne sur une petite boule centrée en p.

En particulier, une fonction sur $\mathbb{R}^d$ dont le Laplacien est nul partout (une *fonction harmonique*, solution de $\Delta f = 0$) est égale à sa moyenne sur toutes les boules. Sur une variété générale, cela n'est plus vrai, car les termes d'ordre supérieur dans (2.1.2) font intervenir la courbure de la variété.

Théorème 2.1.10. *Soit $(\mathcal{M},g)$ une variété Riemannienne compacte de dimension d. Alors l'ensemble des fonctions solutions de l'équation*

$$\Delta f = 0$$

forme un espace vectoriel de dimension égale au nombre de composantes connexes de $\mathcal{M}$.

En particulier, si l'on note

$$\mathcal{M} = \bigcup_{i=1}^{k} C_i$$

la décomposition en composantes connexes, alors les k fonctions

$$f_i(p) = \begin{cases} 1 & \text{si } p \in C_i, \\ 0 & \text{sinon} \end{cases}, \quad i = 1,\dots,k$$

forment une base de l'espace des solutions de $\Delta f = 0$.

On retient que l'opérateur de Laplace-Beltrami encode, entre autres, les composantes connexes de la variété.

Un autre résultat fondamental, qui sera utilisé à la section 3.2 pour le *clustering spectral*, est le suivant, dû à Hermann Weyl.

Rappelons que le spectre d'un opérateur est défini comme pour une matrice : les valeurs λ pour lesquelles il existe une fonction non nulle f telle que

$$\Delta f = \lambda f$$

sont appelées *valeurs propres*, et les fonctions f associées sont les *fonctions propres*.

Théorème 2.1.11. *L'opposé $-\Delta$ de l'opérateur de Laplace-Beltrami sur une variété Riemannienne compacte $(\mathcal{M},g)$ de dimension d admet un spectre discret :*

$$0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots$$

qui tend vers l'infini ($\lambda_k \to \infty$ lorsque $k \to \infty$), et chaque valeur propre a une multiplicité finie.

2.1.7 Théorème de plongement de Nash et reach d'une variété

Dans cette section, nous présentons le théorème de Nash ainsi que la notion de *reach* d'une variété.

Définition 2.1.12 (Plongement)**.** On dit qu'une variété $\mathcal{M}$ est *plongée dans* $\mathbb{R}^D$ s'il existe une application

$$f : \mathcal{M} \to \mathbb{R}^D,$$

appelée *plongement*, telle que la restriction $f|_{\mathcal{M}} : \mathcal{M} \to f(\mathcal{M})$ soit une isométrie.

Par abus de notation, on identifiera souvent $\mathcal{M}$ avec son image $f(\mathcal{M})$ et on écrira $\mathcal{M} \subset \mathbb{R}^D$.

Le théorème fondamental de Nash affirme que toute variété riemannienne peut être plongée dans un espace euclidien :

Théorème 2.1.13 (Nash)**.** *Si $\mathcal{M}$ est une variété riemannienne de dimension d, alors il existe un entier $D > d$ et un plongement $f : \mathcal{M} \to \mathbb{R}^D$. On peut montrer que $D = d(d+1)(3d+11)/2$ suffit.*

Conséquences :

- **Théorique :** Il suffit de comprendre la théorie des sous-variétés de $\mathbb{R}^D$ pour comprendre la théorie générale des variétés riemanniennes.
- **Pratique :** Puisque les données sont toujours collectées comme vecteurs dans $\mathbb{R}^D$, toute configuration riemannienne peut, en principe, être représentée sous cette forme.

On distingue souvent le point de vue *extrinsèque* et *intrinsèque* :

- Extrinsèque : étudier la forme de la surface depuis un espace environnant (ex. la Terre vue depuis l'espace).
- Intrinsèque : étudier la surface sans quitter celle-ci, à partir de mesures locales (ex. distances, angles).

Pour une variété plongée, une quantité extrinsèque très utile est le *reach*.

Définition 2.1.14 (Reach d'une variété plongée)**.** Soit $\mathcal{M} \subset \mathbb{R}^D$ une variété plongée. Le *reach* de $\mathcal{M}$ est le plus grand réel $\varepsilon > 0$ tel que tous les points situés dans le ε-voisinage de $\mathcal{M}$ admettent une unique projection orthogonale sur $\mathcal{M}$.

Remarque : Le reach dépend du plongement choisi et non uniquement de la variété intrinsèque.

- Exemple : une feuille de papier plate plongée dans $\mathbb{R}^3$ a un reach infini (en ignorant les angles).
- Si on froisse légèrement la feuille, son reach devient fini, même si la géométrie intrinsèque reste inchangée (tant qu'on ne déchire pas la feuille).

Exercice : Montrer que le reach d'une sphère plongée naturellement dans $\mathbb{R}^3$ est égal à son rayon.

2.2 Géométrie métrique

Un bon cadre théorique pour étudier la géométrie est fourni par la géométrie Riemannienne, comme présenté à la section 2.1. Ce cadre est extrêmement riche et puissant, mais il présente aussi certaines limites qui découlent précisément de ses forces : la richesse de la géométrie riemannienne provient de la *rigidité* de sa structure (cartes lisses, tenseur métrique lisse, etc.), et ces conditions de régularité sont si fortes qu'elles excluent de nombreux objets qui ne sont pourtant pas pathologiques.

Par exemple, un carré n'est pas une variété riemannienne à cause de ses angles pointus, même s'il s'agit d'un objet simple et très régulier dans un sens intuitif.

Dans cette section, nous introduisons un cadre plus général, permettant de pallier ces limitations tout en conservant une notion de géométrie significative.

2.2.1 Définitions générales

Comme on l'a vu à la section précédente, le cadre riemannien a été introduit afin de déterminer les rapports métriques d'un espace. Pour cela, on a procédé de façon infinitésimale : on a d'abord défini le tenseur métrique comme produit scalaire sur l'espace tangent, puis on en a déduit une distance, la distance géodésique.

On peut se demander s'il n'est pas possible de sauter directement cette étape infinitésimale et de considérer un *espace métrique* (X, d), c'est-à-dire un ensemble X muni d'une distance

$$d : X \times X \to \mathbb{R}_+$$

satisfaisant les axiomes classiques :

(i) $d(x, y) = 0 \iff x = y$ (séparation),

(ii) $d(x, y) \leq d(x, z) + d(z, y)$ (inégalité triangulaire),

(iii) $d(x, y) = d(y, x)$ (symétrie).

Cette approche évite de supposer une structure lisse, trop rigide pour inclure des ensembles simples comme un carré. En revanche, le contre-coup est évident : maintenant, beaucoup d'exemples sont trop irréguliers pour être intéressants, comme les ensembles de Cantor ou d'autres objets fractals.

Pour rendre la théorie plus pertinente, on ajoute des axiomes supplémentaires.

Définition 2.2.1 (Géodésique)**.** Une courbe continue $\gamma : [0, 1] \to X$ est une *géodésique* si elle réalise la distance entre ses extrémités, c'est-à-dire si

$$d(\gamma(t), \gamma(s)) = |t - s|\, d(\gamma(0), \gamma(1)), \quad \forall s, t \in [0, 1].$$

Autrement dit, γ est une isométrie du segment $[0, d(\gamma(0), \gamma(1))]$ muni de la distance euclidienne.

Définition 2.2.2 (Espace géodésique)**.** Un espace métrique (X, d) est dit *géodésique* si, pour tout couple de points $x, y \in X$, il existe au moins une géodésique les reliant $(\gamma(0) = x, \gamma(1) = y)$. Notez qu'il peut en exister plusieurs : par exemple, sur la sphère, entre deux points antipodaux, il existe une infinité de géodésiques.

Définition 2.2.3 (Espace de longueur)**.** Un espace métrique (X, d) est dit *espace de longueur* si la distance est réalisée comme l'infimum des longueurs de toutes les courbes reliant deux points :

$$d(x, y) = \inf_{\gamma : x \to y} \ell(\gamma),$$

où la longueur d'une courbe $\gamma : [0, 1] \to X$ est définie par

$$\ell(\gamma) = \sup \left\{ \sum_{i=1}^{n} d(\gamma(t_{i-1}), \gamma(t_i)) \;\middle|\; 0 = t_0 < \cdots < t_n = 1 \right\}.$$

Exercice : Comparer les deux notions d'espace géodésique et d'espace de longueur. Sont-elles équivalentes ? L'une implique-t-elle l'autre ?

2.2.2 Géométrie d'Alexandrov

En machine learning, il arrive souvent que les fonctions de coût que l'on cherche à optimiser soient en fait des distances, ou bien des carrés de distance (comme le coût quadratique, omniprésent en statistiques mathématiques).

Or, un des cas les plus simples à étudier pour la minimisation d'une fonction est celui où cette fonction est *convexe*.

C'est le cas du coût quadratique : il s'agit du carré de la distance euclidienne. Sa Hessienne est égale à deux fois la matrice identité, qui est définie positive, et donc la perte quadratique est évidemment convexe, voire *fortement convexe*.

Dans cette section, nous abordons brièvement le cas où le carré d'une distance est fortement convexe, car cette hypothèse a de très fortes implications géométriques : il s'agit de la théorie des espaces métriques à courbure d'Alexandrov bornée supérieurement.[7] Ces notions seront réutilisées à la section 4.2.

[7] En anglais CBA pour curvature bounded above.

Définition 2.2.4 (Convexité géodésique)**.** Soit (X, d) un espace géodésique. Une fonction $f : X \to \mathbb{R}$ est dite *géodésiquement convexe* si, pour toute géodésique $\gamma : [0, 1] \to X$, la fonction $f \circ \gamma : [0, 1] \to \mathbb{R}$ est convexe au sens usuel. Autrement dit, pour tous $x, y \in X$ et toute géodésique γ reliant x à y,

$$\forall t \in [0, 1], \quad f(\gamma(t)) \leq (1 - t)f(\gamma(0)) + tf(\gamma(1)).$$

Définition 2.2.5 (Convexité géodésique forte)**.** Soit (X, d) un espace géodésique. Une fonction $f : X \to \mathbb{R}$ est dite *K-fortement géodésiquement convexe* si, pour toute géodésique $\gamma : [0, 1] \to X$, la fonction $f \circ \gamma$ est K-fortement convexe au sens usuel, c'est-à-dire :

$$\forall t \in [0, 1], \quad f(\gamma(t)) \leq (1 - t)f(\gamma(0)) + tf(\gamma(1)) - \frac{K}{2}d(x, y)^2.$$

Exercice : Montrer que les seules fonctions $f : \mathbb{S}^n \to \mathbb{R}$ géodésiquement convexes sont les fonctions constantes. *Indication :* la définition requiert la convexité le long de toutes les géodésiques, y compris lorsqu'il n'y a pas unicité.

Nous pouvons maintenant introduire la notion centrale pour les espaces à courbure non positive.

Définition 2.2.6 (Espace CAT(0))**.** Un espace métrique (X, d) est dit *CAT*(0) si :

(i) il est géodésique et complet,

(ii) pour tout point $p \in X$, la fonction
$$x \mapsto d(p, x)^2$$
est 1-fortement géodésiquement convexe.

L'acronyme CAT provient des noms de **Cartan, Alexandrov, Toponogov**, trois mathématiciens ayant contribué au développement de cette théorie.

Exercice : L'espace Euclidien $(\mathbb{R}^d, ||\cdot||_2)$ est-il CAT(0) ? Qu'en concluez-vous ?

Quelques propriétés importantes des espaces CAT(0) :

(i) Une variété Riemannienne est CAT(0) si, et seulement si, sa courbure sectionnelle est partout non positive.

(ii) Les arbres métriques sont CAT(0).

(iii) Dans un espace CAT(0), pour tous points $x, y \in X$, il existe une *unique* géodésique reliant x à y.

(iv) Pour toutes géodésiques $\gamma_1, \gamma_2 : [0, 1] \to X$, la fonction
$$t \mapsto d(\gamma_1(t), \gamma_2(t))$$
est géodésiquement convexe.

(v) **Comparaison des triangles** : un espace est CAT(0) si et seulement si tous ses triangles sont "plus fins" que les triangles Euclidiens. Formellement, pour tout triangle (x, y, z) dans X, il existe un triangle modèle (x', y', z') dans $\mathbb{R}^2$ avec les mêmes longueurs de côtés, et pour toutes géodésiques γ_{uv} reliant les sommets correspondants,
$$d(\gamma_{uv}(t), \gamma_{uw}(t)) \leq ||(1 - t)u' + tv' - ((1 - t)u' + tw')||, \quad \forall t \in [0, 1],$$
pour toutes paires de sommets (u, v, w) du triangle.

Exercice : En utilisant la comparaison des triangles et la définition de la courbure sectionnelle de la section 2.1.5, montrer la propriété (i).

2.3 Topologie

La topologie est une discipline des mathématiques née des travaux de Henri Poincaré entre la toute fin du 19ème siècle et le tout début du 20ème, alors baptisée par l'expression latine *Analysis Situs*, assez difficilement traduisible, éventuellement signifiant quelque chose comme "analyse de la position" ou "analyse des situations". (Noter qu'étymologiquement, la *topologie* est la science des lieux).

En langage moderne, l'idée de la topologie est de pouvoir définir la notion de *voisinages* entre points d'un espace, et la notion de *continuité* d'une transformation de cet espace, sans avoir à recourir à la notion de distance. Dit autrement, la topologie permet de parler de voisinages, c'est-à-dire de dire si deux points sont voisins ou non, sans *quantifier* cela, et donc seulement de façon qualitative. Il s'agit donc d'une théorie de géométrie encore plus générale que la notion d'espaces métriques. Il s'agit de la topologie générale, usuellement aujourd'hui enseignée en L3 de mathématiques.

Une idée plus poussée, décrite comme de la *topologie algébrique*, consiste à vouloir associer des invariants algébriques à des espaces, afin d'être en mesure de comparer ces espaces en comparant leurs invariants. Cette branche de la topologie est dite algébrique car les invariants sont de nature algébrique : groupes, espaces vectoriels, etc. Un invariant est une fonctionnelle[8] qui, à un espace X, associe un objet $A(X)$ (souvent de nature algébrique) vérifiant la propriété suivante : si X et X' sont isomorphes[9], alors ils ont le même invariant[10] $A(X) = A(X')$. Les invariants permettent donc de déterminer si deux espaces sont vraiment distincts, au sens où si l'on arrive à montrer que $A(X) \neq A(X')$, alors $X \neq X'$. Noter que la plupart du temps, un invariant ne caractérise pas un espace, c'est-à-dire qu'on peut très bien avoir $A(X) = A(X')$ mais $X \neq X'$.

2.3.1 Le groupe fondamental

Un invariant topologique que vous avez très probablement déjà rencontré est le groupe fondamental. Étant donné un espace topologique X et un point $x_0 \in X$, on considère l'ensemble de ses lacets, c'est-à-dire l'ensemble des applications continues $\gamma : [0,1] \to X$ telles que $\gamma(0) = \gamma(1) = x_0$. On munit cet ensemble de la loi de concaténation : si γ_1 et γ_2 sont des lacets, alors le lacet produit $\gamma := \gamma_1\gamma_2$ est défini par

$$\gamma(t) = \begin{cases} \gamma_1(2t) & \text{si } t \in [0,1/2], \\ \gamma_2(2t-1) & \text{si } t \in [1/2,1]. \end{cases}$$

Ceci munit l'ensemble des lacets basés en x_0 d'une structure de groupe. On ne veut alors décompter comme différents que les lacets qui ne peuvent pas être déformés continûment l'un en l'autre. De façon formelle, on passe au quotient l'ensemble des lacets par la relation "être homotopes", et cet ensemble quotient obtenu est alors appelé le *groupe fondamental* et dénoté $\pi_1(X, x_0)$. Il s'agit d'un invariant topologique, au sens où si $\varphi : X \to X'$ est un homéomorphisme, alors les groupes $\pi_1(X, x_0)$ et $\pi_1(X', \varphi(x_0))$ sont isomorphes.

Exercice : Deux lacets $\gamma_1 : [0,1] \to X$ et $\gamma_2 : [0,1] \to X$ basés en $x_0 \in X$ sont dits homotopes s'il existe une fonction continue $H : [0,1] \times [0,1] \to X$ vérifiant les trois conditions

$$\begin{cases} \forall t \in [0,1], H(t,0) = \gamma_1(t), \\ \forall t \in [0,1], H(t,1) = \gamma_2(t), \\ \forall s \in [0,1], H(0,s) = H(1,s) = x_0. \end{cases}$$

Le groupe fondamental est défini comme l'espace des lacets identifiés entre eux par homotopie. Que faut-il vérifier pour s'assurer que la loi de composition définie précédemment définit bien une structure de groupe ? Vérifiez-le.

[8] on dira souvent un foncteur par référence à la théorie des catégories
[9] par exemple homéomorphes, isométriques, etc.
[10] l'invariant est invariant, d'où son nom...

Exercice : Montrer que si X est connexe par arcs, alors pour tout $x, y \in X$, les groupes $\pi_1(X, x)$ et $\pi_1(X, y)$ sont isomorphes, permettant alors de parler du *groupe fondamental* de X.

D'un point de vue intuitif, le groupe fondamental est relié au nombre de "trous" dans l'espace. Par exemple, la sphère a un groupe fondamental trivial (c'est le groupe composé d'un seul élément, le lacet constant égal au point de base) ; on dit qu'elle est *simplement connexe*. Le tore a un trou[11], son groupe fondamental est $\mathbb{Z}^2$.

Exercice : Montrer que le groupe fondamental du tore est $\mathbb{Z}^2$.

Plus généralement, deux fonctions $f, g : X \to Y$ sont dites homotopes s'il existe une fonction continue $H : [0, 1] \times X \to Y$ telle que $H(0, \cdot) = f$ et $H(1, \cdot) = g$.

Deux espaces X et Y sont dits *homotopiquement équivalents* s'il existe deux applications $f : X \to Y$ et $g : Y \to X$ telles que $f \circ g$ est homotope à Id_Y et $g \circ f$ est homotope à Id_X.

Un espace X est dit contractile s'il est homotopiquement équivalent à un point (c'est-à-dire à l'espace singleton $\{0\}$).

Exercice :

1. Montrez que $\mathbb{R}^d$ est contractile.
2. Montrez que le groupe fondamental d'un espace contractile est réduit à un point.
3. Un tore est-il un espace contractile ?
4. Un cercle est-il contractile ?

2.3.2 Homologie simpliciale

Un k-simplexe est défini comme l'enveloppe convexe de $k + 1$ points affinement indépendants dans $\mathbb{R}^k$. Par exemple, un 1-simplexe est un segment, un 2-simplexe est un triangle, un 3-simplexe une pyramide, etc.

Un complexe simplicial K est une famille de simplexes telle que toutes les faces d'un simplexe de K sont elles-mêmes des simplexes de K, et l'intersection de deux simplexes de K est soit vide, soit une face commune aux deux.[12]

Un complexe simplicial *abstrait* est la codification en théorie des ensembles d'un complexe simplicial, sans avoir à demander que les éléments du complexe abstrait soient vraiment des simplexes. Étant donné un ensemble quelconque V, on dit que A est un complexe simplicial abstrait avec sommets dans V si $V \subset A$ et si pour tout $\sigma \in A$, tout sous-ensemble $s \subset \sigma$ est un élément de A : $s \in A$.

L'homologie simpliciale est alors définie de la façon suivante pour K un complexe simplicial.

Soit $k \in \mathbb{N}$. L'ensemble des k-chaînes $C_k(K)$ est défini comme étant l'ensemble des sommes finies formelles écrites à partir des k-simplexes $\sigma_i \in K$:

$$\sum_{i=1}^{p} n_i \, \sigma_i, \quad \text{avec} \quad n_i \in \mathbb{Z}/2\mathbb{Z}.$$

Dit de façon plus rigoureuse, il s'agit de l'espace vectoriel sur[13] le corps $\mathbb{Z}/2\mathbb{Z}$ des entiers modulo 2 engendré par les k-simplexes de K.

Le *bord* d'un k-simplexe $\sigma = [v_0, \cdots, v_k]$ est défini comme le $(k-1)$-simplexe donné par

$$\partial_k(\sigma) := \sum_{i=0}^{k} (-1)^i [v_0, \cdots, \hat{v}_i, \cdots, v_k],$$

[11] c'est un donut

[12] Une triangulation est un complexe simplicial ; il s'agit d'une généralisation de ce concept.

[13] On peut prendre d'autres coefficients, par exemple $\mathbb{Z}$

où $[v_0, \cdots, \hat{v}_i, \cdots, v_k]$ dénote le $(k-1)$-simplexe généré par les points $v_0, \cdots, v_k$ auxquels on a retiré v_i.

L'opérateur de bord peut alors être prolongé en une application linéaire

$$\partial_k : C_k(K) \to C_{k-1}(K),$$

son noyau $ker(\partial_k) \subset C_k(K)$ est appelé l'espace des k-cycles, et son image $Im(\partial_k) \subset C_{k-1}(K)$ est appelée l'espace des k-bords. Les opérateurs de bord vérifient la propriété fondamentale

$$\partial_k \circ \partial_{k+1} = 0.$$

Reformulée avec des mots, cette relation fondamentale dit que *les bords n'ont pas de bord*, et elle implique que les $(k+1)$-bords sont des k-cycles :

$$Im(\partial_{k+1}) \subset ker(\partial_k).$$

On peut alors considérer l'espace vectoriel quotient

$$H_k(K) = ker(\partial_k)/Im(\partial_{k+1}),$$

qui est appelé le k-ème groupe d'homologie simpliciale, et sa dimension $b_k(K)$ est appelée le k-ème nombre de Betti.

Il s'agit d'invariants topologiques : si K et K' sont homéomorphes, alors ils ont les mêmes groupes d'homologie et les mêmes nombres de Betti.

Énonçons les faits suivants[14] :

- Tous les $H_k(K)$ sont abéliens par construction.
- $H_0(K)$ est égal au nombre de composantes connexes de K.
- $H_1(K)$ est égal à l'abélianisé du groupe fondamental $\pi_1(K)$.

Exercice : Déterminer l'homologie simpliciale de la sphère $\mathbb{S}^n$.

[14] À défaut de preuve rigoureuse, le lecteur est au moins invité à réfléchir à *pourquoi* ces faits sont vrais

Chapter 3

Metric learning

3.1 Du discret au continu

La géométrie traite d'objets continus, d'espaces ayant une infinité de points, alors qu'en pratique les données seront toujours en nombre fini et induiront donc des structures discrètes, en particulier des graphes. Il est donc légitime de se demander si l'introduction d'objets aussi abstraits que les variétés riemanniennes pour rendre compte de la structure des données[1] est une idée pertinente.

Dans cette section, on présente un résultat dû à Gromov, connu sous le nom de *théorème de reconstruction*, permettant d'argumenter en faveur du fait que la théorie des espaces métriques mesurés reste cohérente avec la structure discrète des données, en ce sens que les données déterminent l'espace à la limite. Plus précisément :

Théorème 3.1.1. *(Théorème de reconstruction de Gromov)*
Soient $(\mathcal{M}, d_\mathcal{M}, \mu)$ et $(\mathcal{N}, d_\mathcal{N}, \nu)$ deux espaces métriques mesurés, et soient deux échantillons (infinis) $(X_i)_{i\in\mathbb{N}}$ avec les X_i i.i.d. à valeur dans $\mathcal{M}$ et de loi μ, et $(Y_i)_{i\in\mathbb{N}}$ avec les Y_i i.i.d. à valeur dans $\mathcal{N}$ et de loi ν. On considère les suites de matrices $M_n = (d_\mathcal{M}(X_i, X_j))_{1\leq i,j\leq n}$ et $N_n = (d_\mathcal{N}(Y_i, Y_j))_{1\leq i,j\leq n}$. Si pour tout $n \in \mathbb{N}$, M_n et N_n ont la même loi, alors les espaces sont isomorphes, au sens où il existe une application $f : \mathcal{M} \to \mathcal{N}$ qui soit une isométrie envoyant μ sur ν.

Rappelons que $f : \mathcal{M} \to \mathcal{N}$ est dite être une isométrie si c'est une application surjective vérifiant

$$\forall x, y \in \mathcal{M}, \quad d_\mathcal{N}(f(x), f(y)) = d_\mathcal{M}(x, y).$$

Exercice : Si $f : \mathcal{M} \to \mathcal{N}$ est une isométrie, montrer qu'elle est aussi injective, et que sa bijection réciproque $f^{-1} : \mathcal{N} \to \mathcal{M}$ vérifie

$$\forall x', y' \in \mathcal{N}, \quad d_\mathcal{M}(f^{-1}(x'), f^{-1}(y')) = d_\mathcal{N}(x', y').$$

Rappelons également qu'on dit que l'application $f : \mathcal{M} \to \mathcal{N}$ envoie μ sur ν si le push-forward de μ par f est égal à ν, autrement dit si $X \sim \mu$ alors $f(X) \sim \nu$, ou encore pour toute fonction continue bornée $\phi : \mathcal{N} \to \mathbb{R}$,

$$\int_\mathcal{M} \phi(f(x))\, d\mu(x) = \int_\mathcal{N} \phi(y)\, d\nu(y).$$

3.2 Spectral Clustering

La référence principale de cette section est le papier de Von Luxburg [21].

Le clustering consiste à ranger les données en différents sous-groupes, appelés *clusters*, qui partagent des caractéristiques communes.

Pour le graphe d'un réseau social par exemple, il s'agit de détecter les communautés, c'est-à-dire de retrouver quels sont les sous-groupes d'utilisateurs qui interagissent le plus entre eux. Par définition, on

[1] via l'hypothèse de la variété

veut trouver des sous-groupes qui partitionnent le graphe de telle sorte que chaque personne interagisse significativement plus avec les personnes de son propre cluster qu'avec celles des autres clusters.

Une des forces des méthodes que l'on va examiner, et donc aussi une des difficultés, est que l'on souhaite détecter les communautés de façon *non supervisée*, c'est-à-dire sans accès à des données étiquetées. Par exemple, dans le cas du graphe d'un réseau social, aucun utilisateur n'est étiqueté avec le groupe auquel il appartient : on ne sait pas quels sont les groupes, et on doit les construire. On parle d'*apprentissage non supervisé*.

D'un point de vue géométrique, les différents *clusters* que l'on cherche à apprendre correspondent aux *composantes connexes* du graphe des données.

Si les données sont de grande dimension $X_i \in \mathbb{R}^D$, on peut partir du principe[2] qu'elles sont supportées sur une sous-variété $\mathcal{M} \subset \mathbb{R}^D$, et par conséquent le graphe construit à partir des données sera une *discrétisation* de cette variété $\mathcal{M}$. Ainsi, apprendre quelles sont les composantes connexes du graphe revient à apprendre celles de la variété $\mathcal{M}$. Or, tandis qu'un graphe est un objet mathématique relativement peu structuré[3], on peut tirer avantage de la structure riche de la variété $\mathcal{M}$ pour estimer ses composantes connexes.

Nous avons vu à la section 2.1.6 qu'il existe un opérateur différentiel linéaire d'ordre 2 sur une variété compacte $\mathcal{M}$, appelé l'opérateur de Laplace-Beltrami, dont le noyau[4] caractérise entièrement les composantes connexes de $\mathcal{M}$, au sens où l'ensemble des fonctions constantes sur chacune des composantes connexes forme une base. Par conséquent, la question géométrique de la détermination des composantes connexes de $\mathcal{M}$ se réduit à la question analytique de trouver l'ensemble des solutions de l'équation de Poisson

$$\Delta f = 0.$$

Résoudre cette équation revient à résoudre une EDP elliptique linéaire d'ordre 2, faisable par des schémas numériques. Cependant, dans notre cas, nous ne connaissons pas réellement la variété $\mathcal{M}$ ni son tenseur métrique g, et nous n'avons donc pas accès à Δ. Nous n'avons accès qu'à un échantillon de points $X_1, \cdots, X_n \in \mathcal{M} \subset \mathbb{R}^D$, et il va falloir trouver une façon de discrétiser l'opérateur différentiel Δ à partir de cet échantillon, ce qui nous ramène à déterminer le noyau d'une matrice, computationnellement tractable.

L'opérateur Δ discrétisé est une matrice de taille $n \times n$, appelée le *graph Laplacian*. Cela s'explique ainsi : l'opérateur de Laplace-Beltrami agit sur des fonctions $f : \mathcal{M} \to \mathbb{R}$, vues comme des vecteurs de dimension le cardinal de $\mathcal{M}$: $f = (f(x))_{x \in \mathcal{M}} \in \mathbb{R}^{\mathcal{M}}$. Ici, nous n'avons pas accès à $\mathcal{M}$ tout entier, mais seulement à l'échantillon $\mathbb{X} := \{X_1, \cdots, X_n\} \subset \mathcal{M}$. Ainsi, les fonctions sur $\mathcal{M}$ sont remplacées par des vecteurs sur $\mathbb{X}$, c'est-à-dire des éléments de $\mathbb{R}^n$. Par conséquent, l'opérateur de Laplace-Beltrami, linéaire, devient une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^n$, soit une matrice $n \times n$.

Il existe plusieurs façons de construire un graph Laplacian. Dans ce qui suit, nous présentons le *normalized graph Laplacian* ; pour un aperçu des autres constructions, voir [21].

On commence par construire une matrice de similarité entre les points de $\mathbb{X}$:

$$K := (k_{i,j})_{1 \leq i,j \leq n} = (k(X_i, X_j))_{1 \leq i,j \leq n},$$

définie à partir d'un noyau $k : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+^*$ supposé positif, continu et symétrique. Le noyau est souvent défini sur $\mathbb{R}^D \times \mathbb{R}^D$ contenant $\mathcal{M} \times \mathcal{M}$. Le noyau le plus utilisé est le noyau gaussien :

$$k(x, y) = \exp\left(-\frac{|x - y|^2}{2\varepsilon_n}\right),$$

où ε_n est un paramètre de *bandwidth* à choisir de façon appropriée pour garantir la convergence lorsque $n \to \infty$.

[2] d'après l'hypothèse de la variété
[3] c'est un objet discret!
[4] c'est-à-dire l'espace propre associé à la valeur propre zéro

On définit ensuite la matrice des degrés, diagonale, dont les entrées sont $d_i = \sum_{j=1}^n k_{ij}$:

$$D = \mathrm{Diag}\,(d_1, \cdots, d_n)\,.$$

Le *normalized graph Laplacian* est défini par

$$L := I - D^{-1/2} K D^{-1/2}, \tag{3.2.1}$$

avec I la matrice identité $n \times n$.

Exercice :

1. Montrer que L définit une forme quadratique linéaire sur $\mathbb{R}^n$, donnée par
$$X^T L X = \frac{1}{2} \sum_{i,j=1}^n k_{ij} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2 .$$
2. En déduire que L est diagonalisable.
3. Montrer que 0 est valeur propre avec vecteur propre $D^{1/2}\mathbf{1}$, où $\mathbf{1}$ est le vecteur dont toutes les coordonnées valent 1.
4. Montrer que toutes les valeurs propres de L appartiennent à l'intervalle $[0, 1]$.

Remarque : la matrice L ne converge vers le véritable Laplacien qu'une fois correctement normalisée. Il faut choisir $\varepsilon_n \to 0$ à la bonne vitesse ; alors $\varepsilon_n^{-1} L \to -\Delta$. La matrice L converge vers un *noyau de transition markovien*, expliquant pourquoi ses valeurs propres sont dans $[0, 1]$ tandis que celles de $-\Delta$ sont dans $\mathbb{R}_+$. Les valeurs propres de L convergent vers 1, en accord avec le Théorème 2.1.11 vu à la section 2.1.6.

Pour n grand et ε_n bien choisi :

$$\varepsilon_n^{-1} L \begin{pmatrix} f(X_1) \\ \vdots \\ f(X_n) \end{pmatrix} \approx - \begin{pmatrix} \Delta f(X_1) \\ \vdots \\ \Delta f(X_n) \end{pmatrix} .$$

Par le théorème 2.1.10, le noyau du Laplacien encode les composantes connexes de $\mathcal{M}$, i.e. les clusters. On étudie donc le noyau de $\varepsilon_n^{-1} L$ (ou L) pour retrouver les clusters. La convergence de l'opérateur discretisé garantit que les k premières valeurs propres permettent de récupérer les k clusters.

Le *spectral clustering* consiste à diagonaliser L, retenir les k premières valeurs propres $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k$ et leurs vecteurs propres $u_1, \cdots, u_k \in \mathbb{R}^n$. On forme la matrice

$$U = (u_i^j)_{1 \leq i \leq n, 1 \leq j \leq k} \in \mathbb{R}^{n \times k}.$$

On définit les coordonnées spectrales des points :

$$y_i = (u_i^1, \cdots, u_i^k) \in \mathbb{R}^k.$$

On applique ensuite un algorithme k-means sur $y_1, \cdots, y_n$ pour obtenir des clusters $A_1, \cdots, A_k \subset \mathbb{R}^k$, puis on en déduit les clusters $C_1, \cdots, C_k$ des données originales :

$$X_i \in C_j \Leftrightarrow y_i \in A_j.$$

Ainsi, le spectral clustering revient à effectuer un k-means en *coordonnées spectrales*.

Récapitulatif des principes :

1. Les composantes connexes d'une variété correspondent au noyau de l'opérateur de Laplace-Beltrami (Théorème 2.1.10).
2. Convergence du graph Laplacian vers le Laplace-Beltrami :

$$\varepsilon_n L \underset{n\to\infty}{\longrightarrow} -\Delta.$$

Le premier point explique le fonctionnement intuitif de l'algorithme, et le second garantit la *consistance* statistique du spectral clustering, comme traité dans la littérature, voir par exemple [22].

3.3 Réduction de dimension non linéaire

Si l'hypothèse de la variété est vraie, alors les données vivent sur une variété de dimension beaucoup plus faible que l'espace ambiant. Le *metric learning*, ou apprentissage de la variété, permet donc de *réduire* la dimension : c'est ce que l'on appelle les méthodes de réduction de dimension non linéaires, en opposition aux méthodes linéaires, dont la plus célèbre et utilisée est l'Analyse en Composantes Principales (PCA).

Dans cette section, nous présentons l'algorithme *Isomap* ainsi que les *diffusion maps*, mais il existe évidemment beaucoup d'autres techniques. Pour une revue plus détaillée, voir par exemple [20].

3.3.1 Isomap

Isomap (Isometric Mapping) est une technique non linéaire de réduction de dimension introduite dans [19], qui vise à préserver la structure géométrique intrinsèque des données de haute dimension en approximant les distances géodésiques entre les points de données. Contrairement aux méthodes traditionnelles comme l'ACP (Analyse en Composantes Principales), qui reposent sur des hypothèses linéaires, Isomap capture la structure du "manifold" sous-jacent en construisant un graphe de voisinage où chaque point est connecté à ses voisins les plus proches. Les arêtes de ce graphe sont pondérées selon les distances euclidiennes entre les points connectés. La distance géodésique, représentant le plus court chemin le long du manifold, est ensuite estimée en calculant d_n, le plus court chemin entre des paires de points dans ce graphe pondéré. Ces distances géodésiques estimées sont ensuite utilisées comme entrée pour le "MultiDimensional Scaling" (MDS), qui projette les données dans un espace de dimension inférieure tout en préservant les relations géométriques globales. Cela rend Isomap particulièrement efficace pour les ensembles de données dont la géométrie intrinsèque est non linéaire ou fortement courbée.

Nous nous concentrons sur le ε-graphe, dans lequel deux points x_i et x_j sont adjacents si et seulement si $|x_i - x_j| \leq \varepsilon$. Dans ce graphe, nous attribuons le poids $w_{ij} = |x_i - x_j|$ à l'arête $\{x_i, x_j\}$ lorsque $x_i \sim x_j$.

Définition 3.3.1. Étant donné $\varepsilon > 0$, la *distance Isomap* est définie comme la distance du ε-graphe entre x_i et x_j, c'est-à-dire :

$$d_{n,\varepsilon}(x, y) = \inf_{\gamma} \sum_{i=0}^{r} |x_{i+1} - x_i|,$$

où l'infimum est pris sur tous les chemins $\gamma = (x_0, \ldots, x_{r+1})$ avec $x_0 = x$, $x_{r+1} = y$, $x_i \in \mathbb{X}_n$ pour tout $1 \leq i \leq r$, et $|x_{i+1} - x_i| \leq \varepsilon$ pour tout $0 \leq i \leq r$.

Une version alternative de la distance Isomap considère le graphe des K plus proches voisins au lieu du ε-graphe. Dans ce cas, deux points x_i et x_j sont connectés par une arête si l'un d'eux appartient à l'ensemble des K plus proches voisins de l'autre selon la distance euclidienne. Dans ce qui suit, nous nous concentrons sur la formulation du ε-graphe, bien que des résultats similaires s'appliquent au graphe des K plus proches voisins. Le théorème suivant a été établi dans [4].

Théorème 3.3.2. *Soit $\mathcal{M}$ une sous-variété lisse, compacte et connexe de dimension d de $\mathbb{R}^D$, et soit μ la mesure de volume normalisée sur $\mathcal{M}$. Soit $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ un échantillon i.i.d. tiré de μ, et soit d_{n,ε_n} la distance du plus court chemin sur le graphe des ε_n-voisinages construit sur $\mathbb{X}_n$, où deux points sont connectés dès que leur distance euclidienne est inférieure à ε_n. Si $\varepsilon_n \to 0$ et $n\varepsilon_n^d \to \infty$, alors*

$$\sup_{x,y\in\mathbb{X}_n} \left| d_{n,\varepsilon_n}(x, y) - d_{\mathcal{M}}(x, y) \right| \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

c'est-à-dire que la distance Isomap converge en probabilité vers la véritable distance géodésique d_M sur M. De plus, si le taux plus fort $n\varepsilon_n^d / \log n \to \infty$ est vérifié, la convergence ci-dessus est presque sûre.

Notez que dans la condition sur le taux pour ε, c'est la dimension intrinsèque d qui apparaît, et non la dimension ambiante (qui peut être très grande) D.

On pourrait s'arrêter là et faire des statistiques avec cette distance apprise par Isomap. Cependant, la plupart du temps, dans la pratique, on l'utilise pour projeter les données dans un espace de dimension inférieure à celle de départ. La méthode la plus couramment utilisée pour cela est le MultiDimensional Scaling (MDS). Il s'agit de partir du graphe complet des données, avec les arêtes pondérées par les distances apprises entre les points. On cherche alors des points $y_1, \dots, y_n \in \mathbb{R}^d$ avec $d \ll D$, choisis de manière à minimiser le "stress" entre les points :

$$\left(\sum_{i \neq j} (d_{n,\varepsilon}(X_i, X_j) - \|y_i - y_j\|_{\mathbb{R}^d})^2 \right)^{1/2}.$$

3.3.2 Méthodes spectrales: Laplacian eigenmaps/ Diffusion maps

Dans la section 3.2, nous avons vu l'utilisation de l'opérateur de Laplace-Beltrami pour retrouver les composantes connexes d'une variété, et en particulier l'utilisation de sa version discrétisée, le graph-Laplacien, ainsi que son spectre.

L'idée était que si l'on cherche k clusters, cela signifie que le noyau du Laplace-Beltrami est de dimension k et que les clusters correspondent aux k fonctions propres générant ce noyau. Dans la version discrétisée, on prend alors les k plus petites valeurs propres ainsi que les vecteurs propres associés.

On appliquait ensuite l'algorithme des k-means dans cet espace de coordonnées spectrales. Cela correspond à considérer les lignes de la matrice des vecteurs propres, et non plus les colonnes ; on se retrouve ainsi en dimension k au lieu de n.

Dans ce cas, l'entier k était choisi comme le nombre de clusters recherchés. Mais l'idée du plongement spectral est de prendre k bien plus petit que n, simplement pour *réduire la dimension* du problème. C'est précisément ce dont bénéficie l'algorithme de spectral clustering : on applique un k-means à n vecteurs de dimension k (beaucoup de vecteurs en "basse" dimension) plutôt qu'à k vecteurs propres de dimension n (peu de vecteurs mais de grande dimension[5]).

La réduction de dimension spectrale consiste donc à réduire la dimension du problème, mais une fois la dimension réduite, on peut analyser les données comme on le souhaite, et pas seulement appliquer le k-means.

En particulier, la réduction de dimension peut servir à rendre possibles des calculs trop coûteux autrement, mais elle est également utilisée comme outil de *statistique descriptive*, c'est-à-dire pour projeter les données en dimension 2 ou 3 tout en respectant leur géométrie, les rendant ainsi *visualisables*.

Réduire la dimension des données dans $\mathbb{R}^D$ signifie apprendre une fonction

$$\varphi : \mathbb{R}^D \to \mathbb{R}^d, \quad d \ll D.$$

Dans le cadre de la réduction de dimension spectrale, on choisit φ de la manière suivante. Soient $X_1, \dots, X_n \in \mathbb{R}^D$ les données, et posons

$$\mathbb{X} = \{X_1, \dots, X_n\} \subset \mathbb{R}^D.$$

On apprend la fonction φ uniquement sur l'ensemble des données, c'est-à-dire

$$\varphi : \mathbb{X} \to \mathbb{R}^d.$$

Sous l'hypothèse de la variété, où les données sont supportées sur une variété $\mathcal{M} \subset \mathbb{R}^D$, on s'attend à ce que, pour un nombre de données n grand, la fonction apprise φ converge vers une fonction définie sur la variété :

$$\varphi : \mathcal{M} \to \mathbb{R}^d.$$

[5] le scénario redouté en statistiques !

On construit alors un graph-Laplacien L, qui est une matrice $n \times n$. On choisit d, la dimension réduite. Typiquement, on prend $d = 2$ ou $d = 3$ pour de la visualisation, ou on choisit d de manière empirique selon le problème. Dans ce cas, D peut être de l'ordre du million, et d d'une centaine, par exemple.

On calcule alors les d premiers vecteurs propres :

$$\begin{aligned} u_1 &= (u_1^1, \dots, u_1^n) \in \mathbb{R}^n, \\ &\vdots \\ u_d &= (u_d^1, \dots, u_d^n) \in \mathbb{R}^n. \end{aligned}$$

La fonction φ est alors définie comme les coordonnées spectrales :

$$\begin{aligned} \varphi : \mathbb{X}_n &\to \mathbb{R}^d, \\ X_i &\mapsto (u_1^i, \dots, u_d^i). \end{aligned}$$

Remarquez que si l'on met tous les vecteurs u_i en colonnes dans une matrice $n \times d$, la fonction φ revient à associer à X_i la i-ème ligne de cette matrice.

Il existe un degré de liberté dans ce type de méthode : le choix du graph-Laplacien considéré. Dans la section 3.2, nous avons présenté le Laplacien discret normalisé défini par l'équation 3.2.1. Dans ce cas, l'apprentissage de la fonction φ est appelé la méthode *Laplacian Eigenmaps*, introduite par Belkin et Niyogi [1, 2].

Un autre choix standard consiste à considérer le noyau de transition markovien :

$$P := D^{-1}K,$$

qui définit une marche aléatoire sur le graphe des données. L'embedding obtenu à partir des vecteurs propres de P est appelé *Diffusion Map*, introduit par Coifman et Lafon [8].

Le Laplacien associé à cette dynamique est le *random walk graph Laplacian* :

$$L^{\mathrm{rw}} := I - D^{-1}K.$$

D'un point de vue calculatoire, on a la relation

$$L^{\mathrm{rw}} = D^{-1/2} L_{\mathrm{sym}} D^{1/2},$$

ce qui montre que ces deux opérateurs ont des propriétés spectrales étroitement liées.

Exercice : Montrer que λ est une valeur propre de L^{rw} avec vecteur propre u si et seulement si elle est une valeur propre de L avec vecteur propre $D^{1/2}u$. Conclure.

Du point de vue théorique, la justification de ces méthodes repose sur deux faits principaux :

1. les résultats de plongement des variétés dans des espaces de Hilbert via l'utilisation des noyaux de la chaleur [3],
2. la convergence des graph-Laplaciens vers des opérateurs définis sur la variété des données.

Remarque 3.3.3. Le noyau de la chaleur est défini par $P_t = e^{-t\Delta}$; il s'agit d'une famille d'opérateurs interpolant entre l'opérateur identité pour $t = 0$ et un opérateur P_∞ qui associe à une fonction sa moyenne par rapport à la mesure de volume sur la variété :

$$f \mapsto \frac{1}{\mathrm{vol}(\mathcal{M})} \int_{\mathcal{M}} f \, d\mathrm{vol}.$$

Pour un $t > 0$ fixé, on considère le spectre de l'opérateur de Laplace-Beltrami

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots,$$

ainsi que les fonctions propres associées $u_0, u_1, \ldots$. Les auteurs montrent que l'application

$$\begin{aligned} &\mathcal{M} \to \ell^2, \\ &x \mapsto \left(\sqrt{2}(4\pi)^{d/4} t^{\frac{d+2}{4}} e^{-\lambda_j t/2} u_j(x)\right)_{j\geq 0} \end{aligned}$$

est un plongement de la variété $\mathcal{M}$ de dimension d dans l'espace de Hilbert ℓ^2 des suites à carré sommable.

La convergence des graph-Laplaciens assure ensuite la *consistance* des procédures de plongement spectral.

3.3.3 UMAP, SNE, t-SNE

Les algorithmes UMAP, SNE et t-SNE sont également parmi les techniques de réduction de dimension non linéaire les plus utilisées. Voici une liste de sujets de présentation, avec pour référence l'article [15].

1. expliquer l'algo SNE
2. expliquer l'idée géométrique de SNE
3. présenter une simulation en Python de SNE
4. expliquer l'algo t-SNE
5. expliquer l'idée géométrique de t-SNE
6. présenter une simulation en Python de t-SNE
7. expliquer l'algo UMAP
8. expliquer l'idée géométrique de UMAP
9. présenter une simulation en Python de UMAP
10. à quoi servent ces algos, comment sont-ils utilisés ?
11. quels mauvais usages sont pointés dans le papier [13] ?

3.4 Analyse Topologique des données

Cette section suit de très près le déroulé de l'exposé de Frédéric Chazal et Bertrand Michel [7], dont elle reprend également les illustrations. Les lecteurs intéressés sont vivement invités à consulter ce très bel exposé.

Dans les méthodes de clustering, l'objectif était d'apprendre les composantes connexes d'une variété riemannienne, supposée être le support de la loi des données observées.

On a vu à la section 2.3 que, d'un point de vue topologique, les composantes connexes sont encodées par le zéro-ième groupe d'homologie H_0.

En analyse topologique des données, on cherche à apprendre sur les données des propriétés topologiques plus fines que la simple connaissance des composantes connexes. En particulier, on verra comment récupérer de l'information sur les groupes d'homologie d'ordre supérieur. Puis, on définira la notion d'homologie persistante et on montrera comment cette notion permet d'éliminer le bruit inhérent aux données grâce aux résultats de stabilité.

3.4.1 Le théorème du nerf

L'homologie que l'on cherche à apprendre concerne avant tout les complexes simpliciaux. On commence donc par voir comment construire des complexes simpliciaux à partir des données, c'est-à-dire à partir d'un nuage de points et non plus d'un espace topologique continu, comme c'était le cas en théorie (cf. Section 2.3). Les deux constructions les plus utilisées sont les suivantes :

Définition 3.4.1 (Complexe de Vietoris-Rips)**.** Étant donnés $\alpha > 0$, $k \in \mathbb{N}$, et un nuage de points $\mathbb{X}_n = \{x_1, \ldots, x_n\}$ dans un espace métrique $(\mathbb{X}, d)$, le complexe de Vietoris-Rips, noté $Rips_\alpha(\mathbb{X}_n)$, est défini comme l'ensemble des simplexes $[x_0, \ldots, x_k]$ tels que $d(x_i, x_j) \leq \alpha$ pour tout i, j.

Par définition, le complexe de Vietoris-Rips est un complexe simplicial abstrait (cf. Section 2.3). Même si les données sont à valeurs dans $\mathbb{R}^d$, ce complexe n'admet pas forcément de réalisation dans $\mathbb{R}^d$ et peut être de dimension plus grande que d.

Exercice : Montrer que pour $k = 1$, le complexe de Vietoris-Rips est égal au graphe des α-voisinages.

Définition 3.4.2 (Complexe de Čech)**.** Étant donnés $\alpha > 0$, $k \in \mathbb{N}$, et un nuage de points $\mathbb{X}_n = \{x_1, \ldots, x_n\}$ dans un espace métrique $(\mathbb{X}, d)$, le complexe de Čech, noté $\check{C}ech_\alpha(\mathbb{X}_n)$, est défini comme l'ensemble des simplexes $[x_0, \ldots, x_k]$ tels que les $(k+1)$ boules fermées $B(x_i, \alpha)$ aient une intersection non vide.

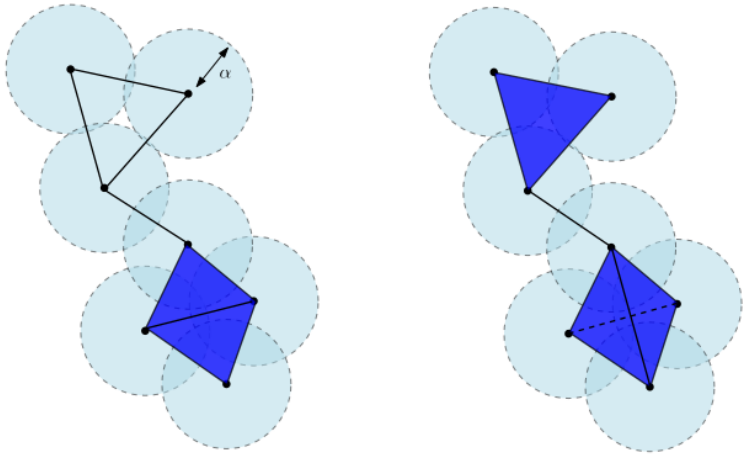

Figure 2: The Čech complex $\text{Cech}_\alpha(\mathbb{X})$ (left) and the and Vietoris-Rips complex $\text{Rips}_{2\alpha}(\mathbb{X})$ (right) of a finite point cloud in the plane $\mathbb{R}^2$. The bottom part of $\text{Cech}_\alpha(\mathbb{X})$ is the union of two adjacent triangles, while the bottom part of $\text{Rips}_{2\alpha}(\mathbb{X})$ is the tetrahedron spanned by the four vertices and all its faces. The dimension of the Čech complex is 2. The dimension of the Vietoris-Rips complex is 3. Notice that this later is thus not embedded in $\mathbb{R}^2$.

Figure 3.1: Illustration provenant de [7]

Exercice : Montrer les inclusions suivantes :

$$Rips_\alpha(\mathbb{X}_n) \subset \check{C}ech_\alpha(\mathbb{X}_n) \subset Rips_{2\alpha}(\mathbb{X}_n).$$

Définition 3.4.3 (Nerf d'un recouvrement)**.** Un recouvrement d'un espace topologique $\mathbb{X}$ est une famille de sous-ensembles $U = (U_i)_{i \in I}$ telle que $\mathbb{X} = \bigcup_{i \in I} U_i$. Le nerf d'un recouvrement U est défini comme le complexe simplicial abstrait $C(U)$ dont les sommets sont les sous-ensembles U_i, et tel que

$$\sigma = [U_{i_0}, \ldots, U_{i_k}] \in C(U) \quad \Leftrightarrow \quad \bigcap_{j=0}^{k} U_{i_j} \neq \emptyset.$$

On peut maintenant énoncer le théorème du nerf, qui affirme que si l'on choisit le recouvrement de façon suffisamment régulière, alors l'espace est *topologiquement équivalent* au nerf de ce recouvrement.

Théorème 3.4.4 (Théorème du nerf)**.** *Soit $U = (U_i)_{i \in I}$ un recouvrement de $\mathbb{X}$, tel que pour tout $J \subset I$, $\bigcap_{j \in J} U_j$ soit vide ou bien* contractible. *Alors $\mathbb{X}$ est homotopiquement équivalent au nerf $C(U)$ du recouvrement.*

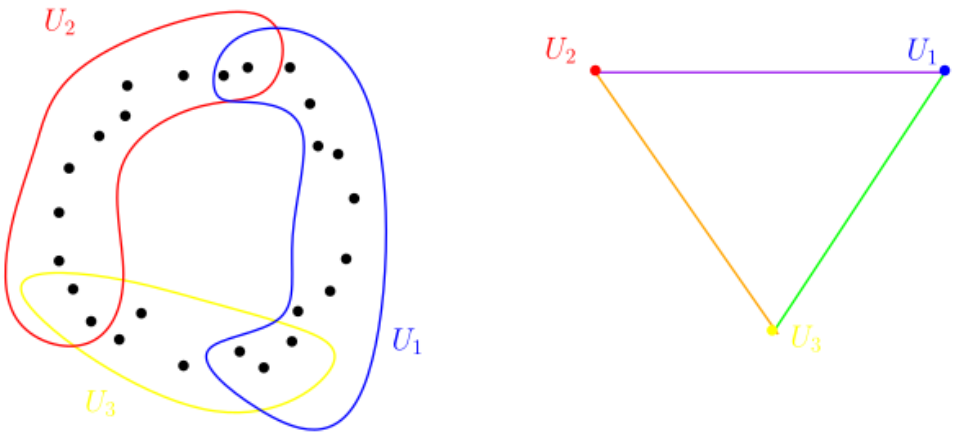

Figure 3: The nerve of a cover of a set of sampled points in the plane.

Figure 3.2: Illustration provenant de [7]

Le théorème du nerf garantit que si l'on parvient à inférer de l'information topologique à partir du nerf d'un recouvrement régulier, alors cette information est également valide pour l'espace sous-jacent (le support des données). Ceci est particulièrement utile, car nous n'avons pas accès au support des données[6], mais on peut manipuler les nerfs de recouvrement, assurant ainsi qu'aucune information n'est perdue par ce procédé.

Exercice :

1. Montrer que le complexe de Čech est le nerf d'un recouvrement, à déterminer.
2. Montrer que les ensembles convexes de $\mathbb{R}^d$ sont contractibles.
3. En déduire que si $\mathbb{X}_n \subset \mathbb{R}^d$, alors le complexe de Čech est homotopiquement équivalent à l'union des boules $\bigcup_{x \in \mathbb{X}_n} B(x, \alpha)$.

Le nerf d'un recouvrement des données est très utilisé pour la visualisation et l'exploration des données. Cette idée est mise en pratique par le célèbre algorithme Mapper, voir Figure 3.3.

[6] On n'a *quasiment jamais* accès à cette information en pratique

Algorithm 1 The Mapper algorithm

Input: A data set $\mathbb{X}$ with a metric or a dissimilarity measure between data points, a function $f : \mathbb{X} \to \mathbb{R}$ (or $\mathbb{R}^d$), and a cover $\mathscr{U}$ of $f(\mathbb{X})$.
for each $U \in \mathscr{U}$, decompose $f^{-1}(U)$ into clusters $C_{U,1}, \cdots, C_{U,k_U}$.
Compute the nerve of the cover of X defined by the $C_{U,1}, \cdots, C_{U,k_U}$, $U \in \mathscr{U}$
Output: a simplicial complex, the nerve (often a graph for well-chosen covers $\to$ easy to visualize):
- a vertex $v_{U,i}$ for each cluster $C_{U,i}$,
- an edge between $v_{U,i}$ and $v_{U',j}$ iff $C_{U,i} \cap C_{U',j} \neq \emptyset$

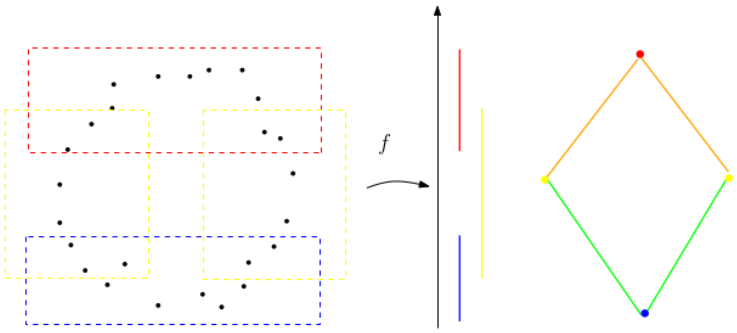

Figure 3.3: Algorithme Mapper, illustration provenant de [7]

3.4.2 Apprendre l'homologie

Dans cette section, nous présentons un résultat garantissant, sous des hypothèses de régularité, que les nombres de Betti d'un espace peuvent se calculer à partir du complexe de Čech.

Théorème 3.4.5. *Soit $M \subset \mathbb{R}^d$ une sous-variété lisse[7] de $\mathbb{R}^d$, de dimension $m < d$. On suppose que, pour certains $\alpha \in (0,1)$ et $R > 0$, la variété M possède un α-reach d'au moins R : $reach_\alpha(M) \geq R$.*

Soit un échantillon de points $\mathbb{X}_n = \{x_1, \ldots, x_n\} \subset M$ qui est ε-proche de M en distance de Hausdorff :

$$\varepsilon := d_H(M, \mathbb{X}_n) \leq \frac{R}{5 + 4/\alpha^2}.$$

Alors, pour tout $r \in [4\varepsilon/\alpha^2, R - 3\varepsilon]$ et tout $k = 0, \ldots, m$, les nombres de Betti de M coïncident avec les nombres de Betti du complexe de Čech $Cech_r(\mathbb{X}_n)$ construit à partir des données :

$$\forall k = 0, \ldots, m, \quad b_k(Cech_r(\mathbb{X}_n)) = b_k(M).$$

Ce résultat garantit que ce que l'on calcule en pratique à partir du complexe de Čech approche les véritables nombres de Betti du support des données.

En pratique, certaines difficultés apparaissent, parmi lesquelles :

1. L'hypothèse sur le reach de la variété peut être assez restrictive.
2. Construire le complexe de Čech est compliqué, car la condition d'intersection des boules (vide ou contractile) est difficile à vérifier.
3. Le choix du paramètre r peut être délicat.
4. L'instabilité aux outliers (voir Figure 3.4).

Bien que des méthodes aient été développées pour traiter ce type de problématiques, nous verrons dans la section suivante une autre approche, basée sur un autre invariant topologique, l'homologie persistante, qui permet également de pallier ce type de difficultés.

[7] c'est-à-dire de classe $\mathcal{C}^\infty$

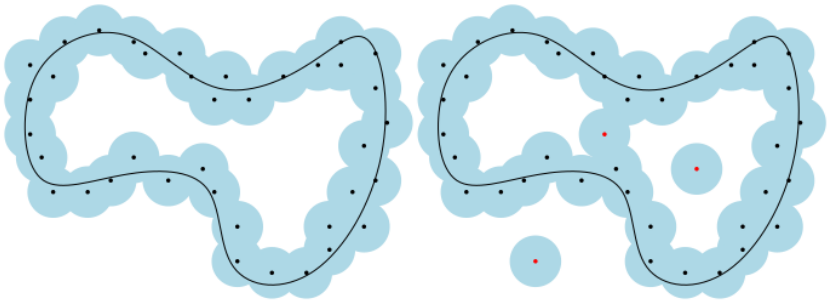

Figure 10: The effect of outliers on the sublevel sets of distance functions. Adding just a few outliers to a point cloud may dramatically change its distance function and the topology of its offsets.

Figure 3.4: Illustration provenant de [7]

3.4.3 Homologie persistante

Définitions

On commence par définir la notion de filtration avant de donner celle de module de persistance.

Définition 3.4.6 (Filtration)**.**

- Une filtration d'un complexe simplicial K est une famille de sous-complexes simpliciaux $(K_r)_{r\in T}$ avec $T \subset \mathbb{R}$, tels que si $r \leq r'$ alors $K_r \subset K_{r'}$ et $K = \bigcup_{r\in T} K_r$.
- Une filtration d'un espace topologique X est une famille de sous-espaces topologiques $(X_r)_{r\in T}$ avec $T \subset \mathbb{R}$, tels que si $r \leq r'$ alors $X_r \subset X_{r'}$ et $X = \bigcup_{r\in T} X_r$.

Un exemple très important de filtration est la filtration par les sous-niveaux d'une fonction. Si $f : X \to \mathbb{R}$ est une fonction, alors la famille

$$\left(f^{-1}((-\infty, r])\right)_{r\in\mathbb{R}}$$

est automatiquement une filtration.

Exercice : Le prouver.

Dans la pratique, on veut des filtrations sur des données $\mathbb{X}_n = \{x_1, \ldots, x_n\}$. Il se trouve que la famille des complexes de Čech $(Cech_r(\mathbb{X}_n))_{r\geq 0}$ est une filtration, de même que la famille des complexes de Rips $(Rips_r(\mathbb{X}_n))_{r\geq 0}$.

Dans ces deux cas, le paramètre r s'interprète très naturellement comme un paramètre d'échelle : pour $r \sim 0$, on observe la structure locale, "microscopique", tandis que pour $r \gg 1$, on observe la structure globale.

Exercice : Prouver que les familles des complexes de Čech et des complexes de Rips sont des filtrations.
On définit maintenant la notion de module de persistance.

Définition 3.4.7 (Module de persistance)**.** Un module de persistance est la donnée d'une famille d'espaces vectoriels[8] $(V_r)_{r\in T}$ indexée par $T \subset \mathbb{R}$, ainsi que d'une famille d'applications linéaires

$$(l_s^r : V_r \to V_s)_{r\leq s}, \quad r, s \in T,$$

vérifiant la loi de composition

$$l_t^s \circ l_s^r = l_t^r, \quad r \leq s \leq t,$$

[8]Espaces vectoriels sur le corps $\mathbb{Z}/2\mathbb{Z}$

avec $l_s^s = \mathrm{Id}$.

Notez que l'ensemble des paramètres $T \subset \mathbb{R}$ correspond toujours, en pratique, à celui d'une filtration $(K_r)_{r \in T}$.

Sous de bonnes hypothèses[9], tout module de persistance se décompose en somme directe de modules élémentaires appelés modules d'intervalle.

Un module d'intervalle est défini de la manière suivante : l'ensemble des paramètres T est un intervalle $T = [b, d) \subset \mathbb{R}$, la famille des espaces vectoriels est constituée uniquement du corps de base

$$\forall r \in [b, d), \quad V_r = \mathbb{Z}/2\mathbb{Z},$$

et les applications linéaires sont toutes l'identité. On peut aussi prolonger le module en l'indexant sur $\mathbb{R}$ entier, et poser $V_r = \{0\}$ si $r \notin [b, d)$. On peut alors représenter le module ainsi :

$$\cdots \to 0 \to \cdots \to \mathbb{Z}/2\mathbb{Z} \to \cdots \to \mathbb{Z}/2\mathbb{Z} \to 0 \to \cdots$$

Les flèches entre les zéros représentent l'application nulle, et les flèches entre les $\mathbb{Z}/2\mathbb{Z}$ représentent l'application identité. Le début de l'intervalle, noté b pour "birth", représente l'apparition d'une caractéristique topologique dans la filtration à cette échelle, et la fin de l'intervalle, notée d, représente sa disparition.

La décomposition d'un module de persistance en une somme directe de modules d'intervalles joue un rôle analogue à la réduction d'un endomorphisme en algèbre linéaire, où l'on met une matrice sous forme diagonale par blocs.

Dès qu'un module est décomposable en modules d'intervalles, on peut définir son *code-barres* de persistance comme l'ensemble des intervalles intervenant dans cette décomposition.

Puisque chaque intervalle peut se représenter comme un couple de points (b, d), on définit le *diagramme de persistance* comme l'union de tous ces couples avec la diagonale

$$\{(x, y) \in \mathbb{R}^2 \mid x = y\}.$$

Il s'agit donc d'un sous-ensemble du demi-quadrant supérieur du plan, voir Figure 3.5.

On pourrait se demander ce qui se passe si un module de persistance peut être décomposé de plusieurs façons différentes en modules d'intervalles. La réponse est un résultat qui assure que le diagramme de persistance est indépendant de la décomposition choisie ; nous renvoyons le lecteur à [7] pour plus de détails.

Interprétation et exemples

Dans cette section, nous esquissons la construction des diagrammes de persistance en pratique sur deux exemples.

[9] que l'on passera sous silence ici

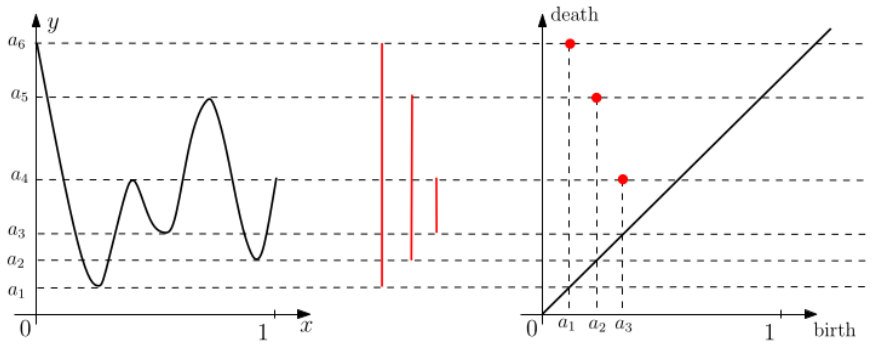

Figure 11: The persistence barcode and the persistence diagram of a function $f : [0, 1] \to \mathbb{R}$.

Figure 3.5: Illustration provenant de [7]

En pratique, le calcul du diagramme de persistance se fait de la manière suivante. On part d'une filtration sur les données, très souvent la filtration de Rips (voir Figure 3.6) ou une filtration par sous-niveaux d'une fonction (voir Figure 3.5).

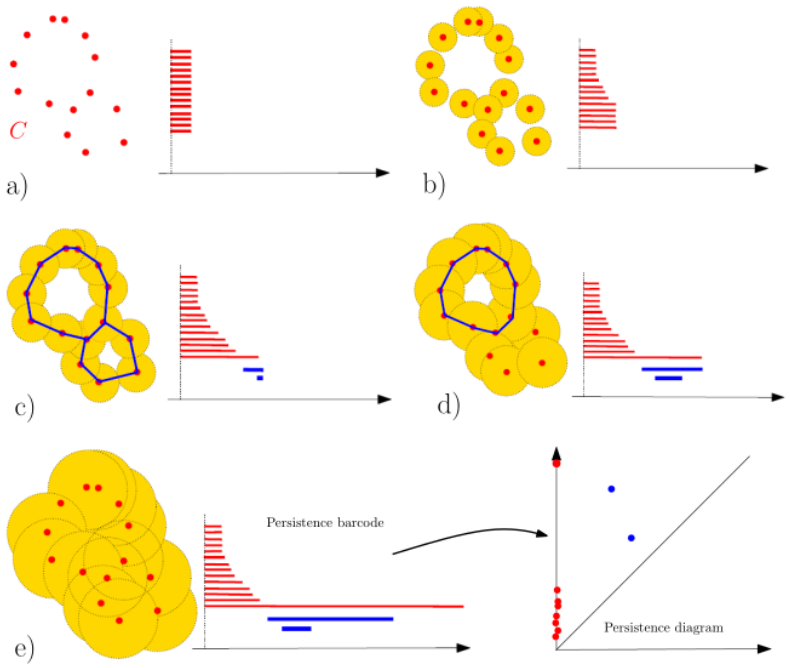

Figure 13: The sublevel set filtration of the distance function to a point cloud and the "construction" of its persistence barcode as the radius of balls increases.

Figure 3.6: Illustration provenant de [7]

Prenons le cas de la filtration de Rips. Pour chaque $r > 0$, on fait grossir des boules de rayon r centrées en chaque point. Au début, pour des r très petits, les boules ne s'intersectent pas : cela donne autant de boules que de points, et donc autant de segments initiaux que de points.

On laisse ensuite ces boules grandir avec r, ainsi que les segments. Dès que deux boules commencent à s'intersecter, on arrête l'un des segments correspondants (par convention, celui le plus en haut dans la Figure

3.6). On poursuit ce processus jusqu'à ce que r devienne si grand que toutes les boules s'intersectent : il ne restera alors plus qu'un seul segment qui ne s'arrêtera jamais.

Si l'on regarde ce qui se passe à des échelles intermédiaires, on enregistre également l'apparition et la disparition des 1-cycles, correspondant à l'image c) de la Figure 3.6. Lorsque des boules $(B_i)_i$ s'intersectent de manière cyclique[10], on en garde trace. Du point de vue de l'homologie, il s'agit des 1-cycles.

Dans cet exemple, puisque l'on est dans $\mathbb{R}^2$, il n'est pas nécessaire de considérer des k-cycles pour $k > 1$. En général, on peut aller plus loin : par exemple, pour des points sur une sphère $\mathbb{S}^2 \subset \mathbb{R}^3$, lorsque les boules grossissent, on observe d'abord les composantes connexes (H_0), puis les cycles (H_1), puis le trou de la sphère (la cavité à l'intérieur), correspondant au 2-cycle.

Dans notre exemple, il y a l'apparition de deux 1-cycles à l'étape c), que l'on représente par les barres bleues. Lorsque l'un des cycles disparaît, on arrête l'un des deux segments, suivant le même principe que pour les 0-cycles.

Stabilité et bruit

Rappelons l'idée générale. On dispose de données que l'on suppose supportées sur une sous-variété $\mathcal{M} \subset \mathbb{R}^D$. L'analyse topologique des données permet *d'apprendre* certaines propriétés topologiques de cette variété. Ceci constitue le cadre théorique "parfait".

Dans la pratique, les données sont bruitées : elles ne vivent donc pas *exactement* sur une sous-variété, mais plutôt dans le voisinage d'une sous-variété[11]. Pour que les méthodes soient utiles en pratique, il faut qu'elles soient *robustes*[12], c'est-à-dire que même avec des données bruitées, l'information topologique soit récupérable.

Nous allons voir dans cette section que c'est le cas des diagrammes de persistance en analyse topologique des données.

Afin de quantifier la stabilité, il est nécessaire de disposer de deux notions de distance : l'une entre le nuage de points et la variété sous-jacente[13], et l'autre entre les diagrammes de persistance. L'objectif est de formuler des résultats généraux de la forme : *si X et Y sont proches, alors les diagrammes de persistance $dgm(X)$ et $dgm(Y)$ sont proches.*

Définition 3.4.8 (Bottleneck distance)**.** La bottleneck distance entre deux diagrammes de persistance dgm_1 et dgm_2 est définie par

$$d_b(dgm_1, dgm_2) := \inf_m \max_{(p,q)\in m} \|p - q\|_\infty,$$

où l'infimum est pris sur tous les *matchings* $m \subset dgm_1 \times dgm_2$ entre les diagrammes dgm_1 et dgm_2.

Par définition, un matching entre deux diagrammes dgm_1 et dgm_2 est un sous-ensemble $m \subset dgm_1 \times dgm_2$ tel que tous les points de $dgm_1 \setminus \{(x,x) \,|\, x \in \mathbb{R}\}$ et tous les points de $dgm_2 \setminus \{(x,x) \,|\, x \in \mathbb{R}\}$ apparaissent exactement une fois dans m, voir Figure 3.7.

[10] c'est-à-dire $B_i \cap B_{i+1} \neq \emptyset$ pour $i = 1, \dots, n-1$ et $B_n \cap B_1 \neq \emptyset$
[11] C'est le cadre de l'hypothèse de la variété : variété + bruit.
[12] on parle aussi de stabilité
[13] dont les données représentent une version bruitée et discrétisée

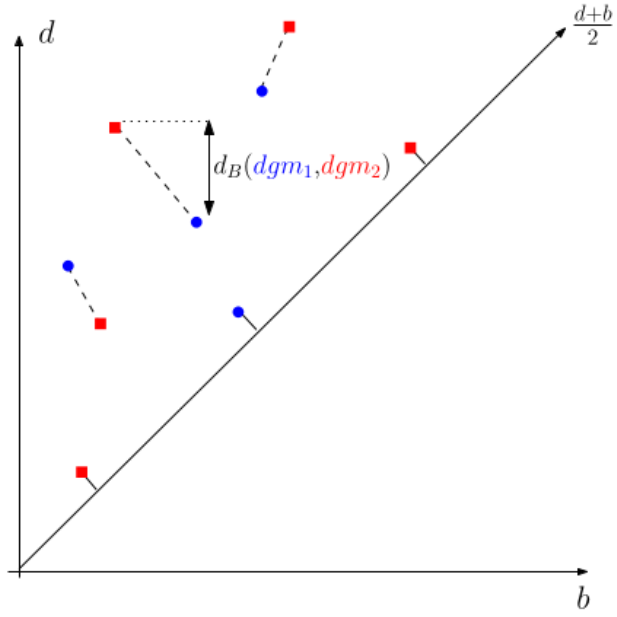

Figure 3.7: Exemple de matching entre deux diagrammes, illustration provenant de [7]

La bottleneck distance est une distance de type L^∞ qui mesure l'écart dans le pire des cas.

Exercice : Montrer que la bottleneck distance est bien une distance.

On introduit maintenant la distance de Hausdorff pour mesurer l'écart entre des formes.

Définition 3.4.9 (Distance de Hausdorff)**.** La distance de Hausdorff entre deux ensembles $A, B \subset \mathbb{R}^d$ est le plus petit ε tel que chacun des ensembles soit contenu dans le ε-voisinage de l'autre :

$$d_H(A, B) := \inf\{\varepsilon > 0 \,|\, A \subset B^\varepsilon \text{ et } B \subset A^\varepsilon\},$$

avec

$$A^\varepsilon := \{x \in \mathbb{R}^d \,|\, \mathrm{dist}(x, A) \leq \varepsilon\}, \quad \mathrm{dist}(x, A) := \inf_{a \in A} \|x - a\|_2.$$

Exercice : Montrer que la distance de Hausdorff est bien une distance.

On peut alors énoncer deux résultats fondamentaux de stabilité.

Théorème 3.4.10. *Soient $f, g : M \to \mathbb{R}$ deux fonctions régulières*[14] *sur un espace topologique M, et soient $dgm_k(f)$ et $dgm_k(g)$ les diagrammes de persistance associés pour un certain $k \in \mathbb{N}$. Alors*

$$d_b\big(dgm_k(f), dgm_k(g)\big) \leq \|f - g\|_\infty.$$

Théorème 3.4.11. *Soient $\mathbb{X}, \mathbb{Y} \subset \mathbb{R}^D$ deux sous-ensembles compacts, et soient $Filt(\mathbb{X})$ et $Filt(\mathbb{Y})$ les filtrations de Rips ou de Čech associées. Alors*

$$d_b\big(dgm(Filt(\mathbb{X})), dgm(Filt(\mathbb{Y}))\big) \leq 2\, d_H(\mathbb{X}, \mathbb{Y}),$$

où $dgm(Filt(\mathbb{X}))$ désigne le diagramme de persistance construit à partir de la filtration considérée.

[14] Certaines notions de régularité sont omises ici

Ces théorèmes garantissent que si les données $\mathbb{X}$ sont une version bruitée d'une certaine variété sous-jacente, le diagramme de persistance calculé à partir des données ne peut pas être trop éloigné (en bottleneck distance) du diagramme de la vraie variété.

En pratique, cela permet d'ignorer les points proches de la diagonale en dessous d'un certain seuil dans le diagramme, voir Figure 3.8.

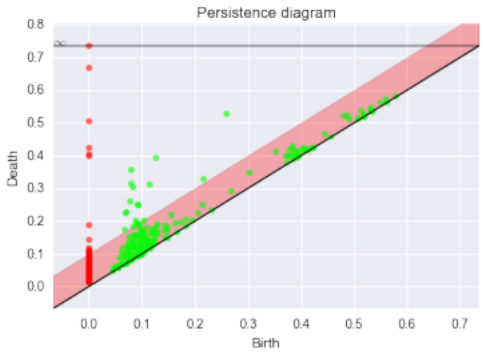

Figure 18: Persistence diagram and confidence region for the persistence diagram of a MBP.

Figure 3.8: Illustration provenant de [7]

Chapter 4

Estimation statistique en contexte géométrique

4.1 Tester l'hypothèse de la variété

Dans cette section, on considère l'hypothèse de la variété au sens statistique, et on va donc présenter un exemple de construction d'un test statistique permettant de trancher si les données observées sont ou non supportées sur une variété.

Dans la littérature, les premiers à construire un tel test sont Fefferman, Mitter et Narayanan dans [10]. Dans cette section, on présente un test un peu plus simple, mais utilisé récemment, voir [17].

Rappelons que l'on part de données $X_1, \cdots, X_n \in \mathbb{R}^D$. Le terme "hypothèse de la variété" regroupe en fait deux hypothèses de nature très différente.

La première est que l'on suppose que les données sont concentrées sur une sous-variété $\mathcal{M} \subset \mathbb{R}^D$ de dimension plus petite $d \ll D$.

La seconde est que cette variété est une vraie variété au sens de la géométrie riemannienne, et donc qu'elle est lisse.

Le premier point est le plus important en pratique, et c'est celui pour lequel on va proposer un test.

Considérons donc un jeu de données $X_1, \cdots, X_n \in \mathbb{R}^D$. On se fixe un $k \in \mathbb{N}$, et pour tout X_i, on considère l'ensemble de ses k plus proches voisins

$$N_k(X_i) = \{X_{i_1}, \cdots, X_{i_k}\} \subset \mathbb{R}^D.$$

Afin de réduire en partie le bruit, on recentre chaque $X_j \in N_k(X_i)$ par rapport aux autres points de ce voisinage :

$$\tilde{X}_j := X_j - \frac{1}{k} \sum_{l=1}^{k} X_{i_l}.$$

On fait alors une analyse en composantes principales[1] (PCA) pour les $\tilde{X}_j$, et on récupère donc D valeurs propres

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$$

mesurant chacune la variance dans une direction propre.

On se donne un seuil $\tau \in (0,1)$[2] et on définit alors la *dimension locale* comme étant le plus petit d_i tel que

$$\sum_{j=1}^{d_i} \lambda_j \geq \tau \sum_{j=1}^{D} \lambda_j.$$

[1] Dans cette configuration, on parle de *local PCA*.

[2] Typiquement 0,9.

Ce nombre d_i dépend de X_i, c'est pourquoi il est dit *local*. Il caractérise une notion de dimension, car il encode le nombre de directions indépendantes dans lesquelles les données varient de façon significative, au voisinage de X_i.

On répète cette procédure au voisinage de chaque X_i dans le jeu de données, et on obtient donc n dimensions locales $d_1, \cdots, d_n \in [1, D]$.

Si l'hypothèse de la variété est vraie, alors tous les d_i doivent simultanément :

- être très proches de la même valeur (dimension constante),
- et cette valeur doit être significativement plus petite que D.

En se fixant a priori un seuil de variabilité des dimensions locales, ainsi qu'un seuil d'écart à D, cela permet de définir un test qui rejette, ou non, l'hypothèse de la variété.

4.2 Loi des grands nombre pour les espaces à courbure négative ou nulle

Dans cette section, on se place dans le cadre théorique où l'on suppose connue la géométrie, et l'on cherche à établir des garanties de convergence a priori. Que la géométrie soit supposée connue signifie que l'on va imaginer disposer d'un oracle nous donnant directement accès, notamment, aux géodésiques[3].

On va supposer que les données $X_1, \cdots, X_n$ prennent leurs valeurs dans un espace métrique $(\mathcal{M}, d)$ à courbure sectionnelle non positive, donc vérifiant la propriété $CAT(0)$ telle qu'introduite à la section 2.2.2.

L'objectif est d'être capable de démontrer, dans ce cadre, le fondement même de la théorie des statistiques (et des probabilités), à savoir la loi des grands nombres.

4.2.1 Moyenne de Fréchet et moyenne inductive

Rappelons que la loi des grands nombres affirme que, si les données $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées, et si de plus X_1 admet un premier moment fini, $\mathbb{E}[X_1] < \infty$, alors la moyenne empirique converge vers la moyenne théorique :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\longrightarrow} \mathbb{E}[X_1],$$

où évidemment le mode de convergence[4] reste à préciser.

Ici, dans le cadre où $X_1, \cdots, X_n \in \mathcal{M}$, la première difficulté est que la moyenne empirique *n'est pas* définie : en effet, on n'est plus dans le cadre d'un espace linéaire, et il n'a donc plus de sens de prendre une somme !

Il va donc falloir redéfinir la notion de moyenne, de façon à ce que la moyenne de points dans $\mathcal{M}$ soit un point de $\mathcal{M}$.

Étant donné que les X_i sont souvent obtenus comme des vecteurs de $\mathbb{R}^D \supset \mathcal{M}$, on pourrait vouloir simplement calculer la moyenne dans $\mathbb{R}^D$: il s'agit du point de vue *extrinsèque*. Cependant, cela a pour inconvénient que la moyenne de points de $\mathcal{M}$ ne sera pas forcément un point de $\mathcal{M}$, ce qui est gênant, car le fait que les points soient dans $\mathcal{M}$ est une information très importante que l'on ne veut pas perdre. En particulier, la notion de moyenne encode l'idée qu'elle doit être *représentative* des données. Si on a une moyenne qui perd l'information d'être dans $\mathcal{M}$, cela contredit complètement cette idée. Pour le dire de façon plus concrète : si les données sont des images de chats, et que l'on considère $\mathcal{M}$ comme la variété des images de chats, alors il n'aurait pas de sens de dire que la moyenne de ces images est une image qui n'est pas un chat (pas dans $\mathcal{M}$).

Afin de donner du sens à la notion de moyenne de n points $x_1, \cdots, x_n$ dans un espace métrique $(\mathcal{M}, d)$, on cherche une formulation de la moyenne dans $\mathbb{R}^d$ qui ne fasse pas intervenir la notion de linéarité. Ainsi,

[3] Qui, évidemment, dans la pratique, ne seront que approximées.
[4] En probabilité, presque sûrement, ...

la définition sera bien une extension de la moyenne usuelle. Pour cela, on se souvient que la moyenne est le minimiseur du problème des moindres carrés :

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \operatorname*{argmin}_{p\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} d(p,x_i)^2.$$

Exercice : Redémontrer cette identité.

Le problème des moindres carrés ne fait intervenir que la notion de distance ; il a donc du sens dans n'importe quel espace métrique général. C'est ce que l'on appelle une *moyenne de Fréchet*.

Définition 4.2.1 (Moyenne de Fréchet)**.** Soit $(\mathcal{M}, d)$ un espace métrique. On définit :

- L'ensemble des moyennes de Fréchet de n points $x_1, \cdots, x_n \in \mathcal{M}$ comme étant l'ensemble

$$\operatorname*{argmin}_{p\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n} d(p,x_i)^2.$$

- L'ensemble des moyennes de Fréchet d'une mesure de probabilité μ sur $\mathcal{M}$ comme étant l'ensemble

$$\operatorname*{argmin}_{p\in\mathcal{M}} \mathbb{E}\left[d(p,X)^2\right], \quad X \sim \mu.$$

Il s'agit bien d'un ensemble : en effet, il peut y avoir plusieurs moyennes de Fréchet. Par exemple, sur la sphère $\mathbb{S}^2$, tous les points de l'équateur sont des moyennes de Fréchet des pôles sud et nord.

Si l'on est sur un espace non compact, il peut évidemment y avoir des mesures de probabilité qui n'admettent pas de moyenne : c'est déjà le cas dans $\mathbb{R}$ avec la loi de Cauchy. Pour un espace métrique général, cela peut également être le cas pour la moyenne entre deux points si l'espace possède des "trous". Par exemple, si l'on considère $\mathbb{R}\setminus\{0\}$ muni de la distance usuelle, alors les points -1 et $+1$ n'ont pas de moyenne, car la fonctionnelle de Fréchet admet un infimum non atteint (atteint en $0 \notin \mathbb{R}\setminus\{0\}$).

Dans le cas d'un espace géodésique et complet, le problème admet toujours au moins un minimiseur.

Un autre moyen de définir la moyenne de n points dans un espace géodésique consiste à remarquer l'identité suivante dans $\mathbb{R}^d$:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \left(1-\frac{1}{n}\right)\frac{1}{n-1}\sum_{i=1}^{n-1} x_i + \frac{1}{n}x_n.$$

Si l'on pose $S_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, l'identité se réécrit

$$S_n = \left(1-\frac{1}{n}\right) S_{n-1} + \frac{1}{n}x_n.$$

Or, la fonction $t \in [0,1] \mapsto (1-t)x + ty$ est exactement le segment de droite reliant x à y. Par conséquent, l'identité précédente signifie que S_n est *le* point situé à distance $1/n$ de S_{n-1} sur le segment reliant S_{n-1} à x_n. Cette formulation ne fait intervenir que la notion de géodésique (segment de droite dans le cas euclidien) et peut donc être généralisée aux espaces métriques généraux.

Définition 4.2.2 (Moyenne inductive)**.** Soit $(\mathcal{M}, d)$ un espace géodésique tel que, entre tout point x et y, il existe une unique géodésique $\gamma_{x,y} : [0,1] \to \mathcal{M}$ avec $\gamma_{x,y}(0) = x$ et $\gamma_{x,y}(1) = y$. Soient $x_1, \cdots, x_n \in \mathcal{M}$. On définit leur moyenne inductive S_n de façon récurrente :

- $S_1 = x_1$,
- $S_n = \gamma_{S_{n-1},x_n}\left(\frac{1}{n}\right)$.

Exercice : Montrer que la moyenne inductive dépend de l'ordre des données x_i. En déduire qu'elle ne coïncide pas nécessairement avec la moyenne de Fréchet.

4.2.2 Loi des grands nombres pour les espaces $CAT(0)$

Maintenant que l'on dispose de notions de moyenne, on peut formuler une loi des grands nombres, et l'objectif sera de la démontrer dans le cas d'un espace vérifiant la propriété $CAT(0)$. On va ici la démontrer pour la moyenne inductive, mais le résultat est également vrai pour la moyenne de Fréchet.

Théorème 4.2.3. *Soit $(\mathcal{M}, d)$ un espace $CAT(0)$ et $X_1, \cdots, X_n$ des variables i.i.d. de loi μ admettant une unique moyenne de Fréchet m. Soit S_n la moyenne inductive. Alors on a*

$$\mathbb{E}\left[d(m, S_n)^2\right] \leq \frac{1}{n}\,\mathbb{E}\left[d(m, X_1)^2\right].$$

En particulier, la moyenne inductive S_n converge dans L^2 vers la moyenne théorique de Fréchet m.

Avant de passer à la preuve, on a besoin de démontrer l'unicité de la moyenne de Fréchet dans les espaces $CAT(0)$, ainsi que *l'inégalité de la variance*.

Lemme 4.2.4. *[Unicité de la moyenne de Fréchet en $CAT(0)$] Soit $(\mathcal{M}, d)$ un espace $CAT(0)$ et X une variable aléatoire de loi μ. On suppose que pour un certain $x_0 \in \mathcal{M}$, on ait $\mathbb{E}[d(x_0, X)] < \infty$. Alors X admet une unique moyenne de Fréchet.*

Proof. On suit la preuve de Sturm [18, Prop. 4.3]. La première remarque est que la définition de la moyenne de Fréchet fait intervenir le carré de la distance (norme L^2), alors que l'hypothèse porte sur la distance (norme L^1). On commence donc par modifier la fonctionnelle objectif en observant que les minimiseurs de

$$z \mapsto \mathbb{E}[d(z, X)^2]$$

sont les mêmes (quand ils existent) que ceux de

$$F_y(z) := \mathbb{E}[d(z, X)^2 - d(y, X)^2],$$

où $y \in \mathcal{M}$ est fixé. En effet, $F_y(z) - F_{y'}(z) = \mathbb{E}[d(y', X)^2 - d(y, X)^2]$ ne dépend pas de z.

On montre alors que F_y est fortement convexe et continue, ce qui, d'après l'analyse convexe classique, implique existence et unicité du minimiseur.

Premièrement, F_y est continue car

$$|F_y(z) - F_y(z')| \leq \mathbb{E}|d(z, X)^2 - d(z', X)^2|.$$

Deuxièmement, grâce à la forte convexité géodésique de $z \mapsto d(z, x)^2$ (voir Définition 2.2.6), pour $z_0, z_1 \in \mathcal{M}$ et $\gamma : [0, 1] \to \mathcal{M}$ la géodésique les reliant, on a

$$\begin{aligned} F_y(\gamma(t)) &= \mathbb{E}[d(\gamma(t), X)^2 - d(y, X)^2] \\ &\leq (1-t)\mathbb{E}[d(\gamma(0), X)^2 - d(y, X)^2] + t\mathbb{E}[d(\gamma(1), X)^2 - d(y, X)^2] - t(1-t)d(z_0, z_1)^2 \\ &= (1-t)F_y(\gamma(0)) + tF_y(\gamma(1)) - t(1-t)d(z_0, z_1)^2, \end{aligned}$$

ce qui établit la forte convexité et termine la preuve. □

Exercice : Montrer que si pour un certain $x_0 \in \mathcal{M}$, $\mathbb{E}[d(x_0, X)] < \infty$, alors pour tout $x \in \mathcal{M}$, $\mathbb{E}[d(x, X)] < \infty$.

Lemme 4.2.5. *[Inégalité de la variance] Soit $(\mathcal{M}, d)$ un espace $CAT(0)$ et X une variable aléatoire de loi μ. On note $m \in \mathcal{M}$ sa moyenne de Fréchet. Alors, pour tout $z \in \mathcal{M}$,*

$$d(z, m)^2 \leq \mathbb{E}[d(z, X)^2 - d(m, X)^2].$$

Proof. On suit la preuve de Sturm [18, Prop. 4.4]. Pour $y = m$, la fonctionnelle

$$F_m(z) := \mathbb{E}[d(z, X)^2 - d(m, X)^2]$$

est fortement convexe le long de la géodésique $\gamma : [0,1] \to \mathcal{M}$ reliant m à z, donnant

$$F_m(\gamma(t)) \leq (1-t)F_m(\gamma(0)) + tF_m(\gamma(1)) - t(1-t)d(m,z)^2.$$

Or, $F_m(\gamma(0)) = 0$ et $F_m(\gamma(1)) = \mathbb{E}[d(z,X)^2 - d(m,X)^2]$. Ainsi,

$$0 \leq t\,\mathbb{E}[d(z,X)^2 - d(m,X)^2] - t(1-t)d(m,z)^2.$$

En divisant par t et laissant $t \to 0$, on obtient l'inégalité de la variance. □

Exercice : Montrer que dans $\mathbb{R}^d$, l'inégalité de la variance est vraie et devient une égalité. Cela est-il cohérent avec la notion d'espace $CAT(0)$?

Preuve du théorème 4.2.3. On suit la preuve de Sturm [18, Théorème 4.7]. On note

$$\sigma^2 := \mathbb{E}[d(m,X_1)^2].$$

La preuve se fait par récurrence sur $n \geq 1$. Pour $n = 1$, l'inégalité est une égalité. Supposons qu'elle soit vraie pour n et montrons-la pour $n+1$:

$$\begin{aligned}
\mathbb{E}[d(m,S_{n+1})^2] &= \mathbb{E}\Big[d(m,\gamma_{S_n,X_{n+1}}(\tfrac{1}{n+1}))^2\Big] \\
&\leq \frac{n}{n+1}\mathbb{E}[d(m,S_n)^2] + \frac{1}{n+1}\mathbb{E}[d(m,X_{n+1})^2] - \frac{n}{(n+1)^2}\mathbb{E}[d(X_{n+1},S_n)^2] \\
&\leq \frac{n}{n+1}\mathbb{E}[d(m,S_n)^2] + \frac{1}{n+1}\mathbb{E}[d(m,X_{n+1})^2] - \frac{n}{(n+1)^2}\Big(\mathbb{E}[d(m,S_n)^2] + \mathbb{E}[d(m,X_{n+1})^2]\Big) \\
&= \left(\frac{n}{n+1} - \frac{n}{(n+1)^2}\right)\mathbb{E}[d(m,S_n)^2] + \left(\frac{1}{n+1} - \frac{n}{(n+1)^2}\right)\mathbb{E}[d(m,X_1)^2] \\
&= \left(\frac{n}{n+1}\right)^2\mathbb{E}[d(m,S_n)^2] + \frac{1}{(n+1)^2}\mathbb{E}[d(m,X_1)^2] \\
&\leq \frac{1}{n+1}\mathbb{E}[d(m,X_1)^2],
\end{aligned}$$

où la première inégalité provient de la forte convexité en $CAT(0)$, la seconde de l'inégalité de la variance (Lemme 4.2.5) et la dernière de l'hypothèse de récurrence. □

4.2.3 Normalité asymptotique de la moyenne de Fréchet empirique, ou BP-TCL

Le terme de BP-TCL vient des auteurs Bhattacharya et Patrangenaru, qui ont largement contribué à développer ces résultats [5].

À la section précédente, on a vu la notion de loi des grands nombres pour les espaces métriques, et en particulier on a prouvé une loi des grands nombres L^2 pour la moyenne inductive dans les espaces $CAT(0)$.

L'étape suivante dans la théorie des probabilités après la loi des grands nombres est évidemment le théorème central limite (TCL). En statistique, la loi des grands nombres correspond à établir la *consistance* de l'estimateur de la moyenne, tandis que le TCL correspond à la *normalité asymptotique* de cet estimateur.

La question est donc : peut-on établir un TCL, c'est-à-dire une normalité asymptotique, dans le cadre qui nous occupe ? Dans cette section, on va esquisser des arguments en faveur d'une réponse positive pour la moyenne empirique de Fréchet dans le cadre d'une variété riemannienne.

Soit $(\mathcal{M}, g)$ une variété riemannienne, et $X_1, \cdots, X_n$ des variables aléatoires i.i.d. à valeurs dans $\mathcal{M}$, de loi μ. On suppose que μ admet une unique moyenne de Fréchet $m \in \mathcal{M}$.

Soit S_n la[5] moyenne de Fréchet empirique des données $X_1, \cdots, X_n$. La loi des grands nombres pour les moyennes de Fréchet dit que S_n converge[6] vers la moyenne théorique m. Le but du TCL est d'étudier l'ordre

[5] sous de bonnes hypothèses, elle est unique

[6] p.s., en probabilité ou en L^2

des fluctuations ainsi que leur forme[7] entre S_n et m. Le problème est que la fluctuation $S_n - m$ n'a pas de sens directement dans $\mathcal{M}$; on doit donc passer en coordonnées pour écrire ces fluctuations.

On se place donc dans un voisinage $V \subset \mathcal{M}$ de m, et on considère les coordonnées normales centrées en m (voir Section 2.1.3). On a ainsi l'application exponentielle définie sur un ouvert U contenant zéro de l'espace tangent $T_m\mathcal{M} \approx \mathbb{R}^d$ en m :

$$\exp_m : U \subset T_m\mathcal{M} \to V \subset \mathcal{M},$$

et sa réciproque, appelée logarithme au point m :

$$\mathrm{Log}_m : V \subset \mathcal{M} \to U \subset T_m\mathcal{M},$$

qui associe à un point de la variété un vecteur de l'espace tangent. C'est donc le logarithme qui nous permet de *remonter* les points sur l'espace tangent, et l'on étudie ainsi les fluctuations

$$\mathrm{Log}_m(S_n) - \mathrm{Log}_m(m) = \mathrm{Log}_m(S_n) \in T_m \approx \mathbb{R}^d.$$

Une fois cela écrit en coordonnées, la moyenne de Fréchet devient un m-estimateur classique, et sa normalité asymptotique se déduit de l'analyse usuelle des M-estimateurs.

Soit

$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n} d(p, X_i)^2$$

la fonctionnelle empirique, dont S_n est le minimiseur, et

$$F(z) = \mathbb{E}[d(p, X_1)^2]$$

la fonctionnelle de Fréchet, dont m est le minimiseur.

Dans la carte exponentielle, on écrit le développement de Taylor de ∇F_n au voisinage de m, en notant $\tilde{S}_n := \mathrm{Log}_m(S_n)$ et $\tilde{m} := \mathrm{Log}_m(m)$[8] :

$$\nabla F_n(S_n) = \nabla F_n(m) + (\tilde{S}_n - \tilde{m})\nabla^2 F_n(m) + \text{Reste}.$$

Comme S_n est minimiseur, on obtient :

$$0 = \frac{1}{n}\sum_{i=1}^{n} \nabla d(m, X_i)^2 + (\tilde{S}_n - \tilde{m})\frac{1}{n}\sum_{i=1}^{n} \nabla^2 d(m, X_i)^2 + \text{Reste}.$$

En multipliant par $\sqrt{n}$:

$$0 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla d(m, X_i)^2 + \Big(\sqrt{n}(\tilde{S}_n - \tilde{m})\Big)\frac{1}{n}\sum_{i=1}^{n} \nabla^2 d(m, X_i)^2 + \text{Reste}.$$

L'idée est que la loi des grands nombres usuelle[9] assure la convergence presque sûre de

$$\frac{1}{n}\sum_{i=1}^{n} \nabla^2 d(m, X_i)^2 \to \mathbb{E}[\nabla^2 d(m, X_1)^2],$$

et que les hypothèses géométriques (bornitude des courbures sectionnelles) permettent de contrôler le reste, qui converge vers zéro en probabilité.

On en déduit donc le BP-TCL :

$$\sqrt{n}(\tilde{S}_n - \tilde{m}) \rightharpoonup \mathcal{N}\Big(0, \Lambda^{-1}C(\Lambda^t)^{-1}\Big),$$

avec C la matrice de covariance de $\nabla d(m, X_1)^2$ et $\Lambda = \mathbb{E}[\nabla^2 d(m, X_1)^2]$.

Notons que seule la matrice C encode de l'information sur la loi μ, tandis que Λ encode uniquement la géométrie de la variété. Dans le cas euclidien, la Hessienne de la distance au carré est constante, donc seul le terme de covariance apparaît, conformément au TCL classique.

[7] Gaussienne

[8] On a bien sûr $\tilde{m} = 0$, mais l'écrire facilite le lien avec les formules de Taylor en cadre euclidien

[9] Tout est écrit en coordonnées, donc on est dans $\mathbb{R}^d$

4.3 Illustration : barycentres de Wasserstein

Dans cette section, l'objectif est d'être capable de faire des moyennes d'images.

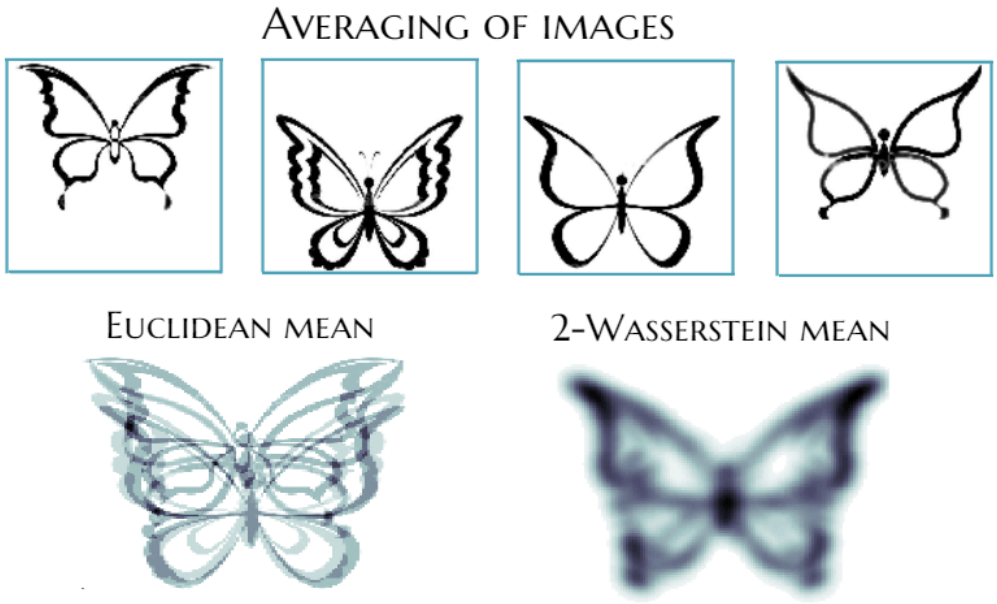

Figure 4.1: Illustration provenant de [9]

Une image est représentée comme un tableau de nombres, les pixels, codant les niveaux de gris. C'est donc un vecteur de $\mathbb{R}^D$ avec D le nombre de pixels.

Une première approche très naïve consisterait à faire la moyenne linéaire dans $\mathbb{R}^D$, ce qui revient à superposer les images. Cette méthode ne fonctionne pas correctement (voir figure ci-dessus).

Une approche plus fructueuse est la suivante : une image peut être représentée comme une mesure de probabilité.

Exercice : Une image est une fonction

$$I : \{1, \cdots, n\} \times \{1, \cdots, m\} \to \mathbb{R}_+,$$

où $I(i, j)$ représente l'intensité de niveau de gris[10] du pixel en position (i, j). Expliquez comment on peut représenter cela de façon équivalente sous forme d'une mesure de probabilité.

On peut alors définir la moyenne d'images comme une moyenne de Fréchet pour une certaine distance bien choisie sur l'espace des mesures de probabilité. En pratique, la distance appropriée est la distance de Wasserstein.

Étant données μ et ν, deux lois de probabilité sur un espace métrique $(\mathcal{M}, d)$, leur distance de Wasserstein est définie comme suit. On prend $X \sim \mu$ et $Y \sim \nu$ deux variables aléatoires, *non nécessairement indépendantes*. Le couple (X, Y) est appelé *couplage* de μ et ν[11]. La distance de Wasserstein est alors l'infimum sur tous les couplages :

$$W_2(\mu, \nu) := \sqrt{\inf_{X \sim \mu,\, Y \sim \nu} \mathbb{E}[d(X, Y)^2]}.$$

(Il s'agit ici de la distance W_2, mais on peut définir de façon similaire W_p pour tout $p \geq 1$.)

[10] Dans le cas du système de couleur RGB, on a trois fonctions pour chaque niveau d'intensité de Rouge, Vert et Bleu.

[11] Il existe toujours au moins un couplage, par exemple le couplage trivial où X et Y sont indépendants.

Exercice : Montrer que W_2 est bien une distance.

On définit alors la moyenne entre des images comme la moyenne de Fréchet pour la distance W_2 entre leur représentation sous forme de probabilité.

Si $\mu_1, \cdots, \mu_n$ sont des lois de probabilité et $\lambda_1, \cdots, \lambda_n$ des poids tels que $\lambda_1 + \cdots + \lambda_n = 1$, leur moyenne de Wasserstein est définie comme la moyenne de Fréchet :

$$\operatorname*{argmin}_{p \in \mathcal{M}} \sum_{i=1}^{n} \lambda_i W_2(p, \mu_i)^2.$$

On termine cette section par quelques illustrations comparant la moyenne naïve et la moyenne de Wasserstein.

En figure 4.1, la moyenne avec des poids uniformes de quatre dessins de papillons est représentée. On voit que la moyenne euclidienne n'est qu'une superposition, qui ne ressemble donc plus à un papillon, tandis que la moyenne de Wasserstein, bien que légèrement floue, capture réellement l'idée des images dont elle est la moyenne : autrement dit, cela reste un papillon.

En figure 4.2, on considère des images des lettres A, B, C et D, et on fait des moyennes de Wasserstein en faisant varier les poids par pas de 1/4. Dans chaque coin, on met une masse de 1 pour une image et on retrouve donc la lettre correspondante, tandis que pour l'image centrale, on met des poids uniformes égaux à 1/4. Le point central représente donc bien la moyenne des quatre lettres. On remarquera qu'en faisant varier les poids, on obtient des déformations continues des images.

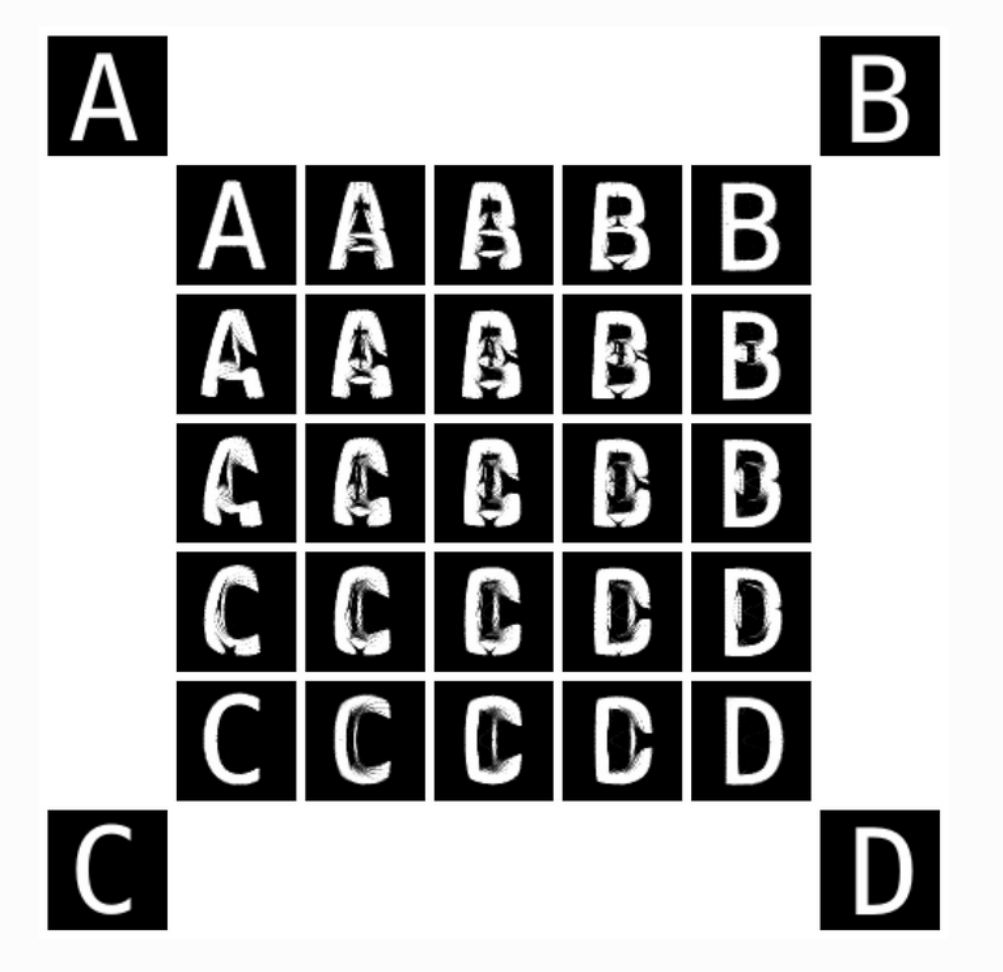

Figure 4.2: Source GeomLoss documentation (kernel-operations.io) [11]

Bibliography

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14, 2001.

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] Pierre Bérard, Gérard Besson, and Sylvain Gallot. Embedding Riemannian manifolds by their heat kernel. *Geometric & Functional Analysis GAFA*, 4(4):373–398, 1994.

[4] Mira Bernstein, Vin De Silva, John C. Langford, and Joshua B. Tenenbaum. Graph approximations to geodesics on embedded manifolds, 2000.

[5] Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds—II. *The Annals of Statistics*, 33(3):1225 – 1259, 2005.

[6] Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International journal of computer vision*, 76(1):1–12, 2008.

[7] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:667963, 2021.

[8] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[9] Johannes Ebert, Vladimir Spokoiny, and Alexandra Suvorikova. Construction of non-asymptotic confidence sets in 2-wasserstein space, 2017.

[10] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[11] Jean Feydy. Wasserstein barycenters in 2d – geomloss example. https://www.kernel-operations.io/geomloss/_auto_examples/optimal_transport/plot_wasserstein_barycenters_2D.html, 2026. Accessed: 12 February 2026.

[12] Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian geometry*, volume 2. Springer, 1990.

[13] Hyeon Jeon, Jeongin Park, Sungbok Shin, and Jinwook Seo. Stop misusing t-sne and umap for visual analytics. *arXiv preprint arXiv:2506.08725*, 2025.

[14] Jasper Kreeft, Artur Palha, and Marc Gerritsma. Mimetic framework on curvilinear quadrilaterals of arbitrary order. 11 2011.

[15] Marina Meilă and Hanyu Zhang. Manifold learning: What, how, and why. *Annual Review of Statistics and Its Application*, 11(1):393–417, 2024.

[16] Bernhard Riemann. Über die Hypothesen, welche der Geometrie zugrunde liegen. In *Über die Hypothesen, welche der Geometrie zu Grunde liegen*, pages 1–47. Springer, 1854.

[17] Michael Robinson, Sourya Dey, and Tony Chiang. Token embeddings violate the manifold hypothesis. *arXiv preprint arXiv:2504.01002*, 2025.

[18] Karl-Theodor Sturm. *Probability measures on metric spaces of nonpositive curvature*. SFB 611, 2003.

[19] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[20] Laurens Van Der Maaten, Eric O Postma, H Jaap Van Den Herik, et al. Dimensionality reduction: A comparative review. *Journal of machine learning research*, 10(66-71):13, 2009.

[21] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[22] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.