Un modèle génératif pour l'Apprentissage de la Topologie

Michaël Aupetit*, Pierre Gaillard**, Gérard Govaert***

* Commissariat à l'Energie Atomique
LIST, Laboratoire Intelligence Multi-capteurs et Apprentissage
F-91191 Gif-sur-Yvette
michael.aupetit@cea.fr
** Commissariat à l'Energie Atomique
Centre DAM - Ile de France
Bruyères-le-Châtel - 91297 Arpajon cedex
pierre.gaillard@cea.fr
*** UTC - Heudiasyc
Compiègne - France
gerard.govaert@hds.utc.fr

Résumé. Un nuage de points est plus qu'un ensemble de points isolés. La distribution des points peut être gouvernée par une structure topologique cachée, et du point de vue de la fouille de données, modéliser et extraire cette structure est au moins aussi important que d'estimer la seule densité de probabilité du nuage. Dans cet article, nous proposons un modèle génératif basé sur le graphe de Delaunay d'un ensemble de prototypes représentant le nuage de points, et supposant un bruit gaussien. Nous dérivons les équations de l'algorithme Expectation-Maximisation de maximisation de la vraisemblance, et nous utilisons le critère d'information bayésien (BIC) pour sélectionner le modèle de complexité optimale. Ce modèle ne nécessite aucun réglage manuel arbitraire de paramètres. Les expériences que nous menons sur des données jouets et des bases d'images montrent que la connexité du graphe reproduit correctement celle du nuage de points. Nous montrons aussi que ce modèle peut être utilisé en tant qu'outil de prétraitement en classification supervisée de caractères manuscrits. Ce travail a pour objectif de poser les premières pierres d'un cadre théorique basé sur les modèles génératifs statistiques, permettant la construction automatique de modèles topologiques d'un nuage de points.

1 Introduction

En apprentissage statistique, on suppose que les données sont générées par une fonction densité de probabilité (pdf) $p(.)$ ayant éventuellement beaucoup moins de degrés de liberté que l'espace ambiant (Belkin et Niyogi, 2004). Considérant des données de type vecteurs de réels, l'ensemble de données forme un nuage de points dans $\mathbb{R}^D$ que l'on suppose situé au voisinage d'un ensemble de variétés, appelées "variétés principales" (Tibshirani, 1992), plongées dans l'espace ambiant, et images de certaines variétés latentes au travers d'un processus

d'observation d'un phénomène physique. L'Apprentissage de la Topologie est un domaine récent en Apprentissage Automatique (Aupetit et al., 2007), dont l'objectif est de développer des méthodes basées sur les statistiques pour retrouver les invariants topologiques de ces variétés latentes à partir du nuage de points (Figure 2). La connexité ou la dimension intrinsèque sont de tels invariants topologiques.

Etant donné un ensemble $\underline{x}$ de M points observés, dans un espace ambiant euclidien à D dimensions, les méthodes statistiques permettent de résoudre des problèmes très généraux de discrimination, classification ou régression, en estimant la densité de probabilité de cet ensemble (Bishop, 2006). Cependant, la fonction densité de probabilité p (fonction de $\mathbb{R}^D$ dans $\mathbb{R}^+$) bien qu'elle contienne la totalité de l'information extractible de la population dont le nuage de points est un échantillon, est habituellement estimée par des méthodes (fenêtres de Parzen par exemple) qui ne rendent pas explicite l'information géométrique et topologique relatives à son support.

En effet, si l'on suppose que la population est une sous-variété de l'espace ambiant, il reste des informations à extraire de ces variétés et donc des variétés latentes correspondantes, que nous n'extrayons pas avec les estimateur de densité actuels. Pourtant nous pourrions apprendre de leur géométrie (position relative des variétés, courbure) et même plus encore de leur topologie (connexité, dimension intrinsèque, invariants topologiques comme les nombres de Betti (Munkres, 1993)). De plus, la topologie des variétés latentes est plus susceptible d'être préservée par le processus d'observation que leur géométrie, parce que la géométrie, à un facteur d'échelle près, est préservée seulement par similarité, tandis que la topologie l'est par homéomorphisme, une classe de transformations bien plus grande contenant les similarités.

Cette connaissance vaut-elle la peine d'être extraite ? Nous le pensons. Par exemple, en Reconnaissance de Formes, la forme des nuages de points est considérée comme l'élément pertinent pour les tâches de discrimination (Carlsson et al., 2004). En classification non-supervisée ou semi-supervisée, il est proposé d'attribuer à la même classe les points qui appartiennent à la même composante connexe (Belkin et al., 2006). Dans le contexte de l'Analyse Exploratoire de Données (Aupetit et Catz, 2005; Aupetit, 2007; Gaillard et al., 2008), les caractéristiques topologiques sont d'un intérêt primordial particulièrement en dimension supérieure à trois, afin de détecter des formes non visualisables directement. Extraire les invariants topologiques en modélisant les variétés principales fournit aussi un moyen de mesurer les distances géodésiques le long de ces variétés, ce qui a des applications dans la planification optimale de trajectoire et le calcul de cinématique inverse en robotique (Zeller et al., 1996), ou encore lors de projections non-linéaires pour faciliter le dépliage des variétés (Lee et al., 2002; de Silva et Tenenbaum, 2003).

Enfin, si nous sommes en mesure d'extraire des invariants topologiques d'un nuage de points, ces invariants sont susceptibles d'être plus robustes au bruit et à différentes conditions d'observation que la géométrie de ce nuage. Comme nous le montrons en perspective de ce travail, cela permet de générer un ensemble de données structurées (un ensemble de graphes) complémentaire des nuages de points de départ.

Dans ce travail, nous nous focalisons sur l'extraction de la connexité des variétés principales d'un nuage de points.

2 Etat de l'art

Les techniques de Quantification Vectorielle (Ahalt et al., 1990) comme les «K-moyennes» (Queen, 1967) ou le Neural-Gas (Martinetz et al., 1993), et leur version générative (Celeux et Govaert, 1992), tels les Modèles de Mélanges Gaussiens (McLachlan et Peel, 2000) tendent à représenter les variétés principales par un ensemble fini de N points $\underline{w} = \{w_i \in \mathbb{R}^D\}_{i=1}^N$, que nous appelons prototypes [1].

Le modèle génératif est un modèle du processus de génération des données en deux temps : tirage aléatoire d'un composant parmi les N, suivant une loi multinomiale, puis tirage des données suivant la loi gaussienne assignée à ce composant. Ainsi dans ce modèle, les variétés principales sont supposées être un ensemble de points $\underline{w}$ (les prototypes ou moyennes des composants gaussiens), ayant été corrompues par un bruit additif Gaussien qui a mené aux données finalement observées.

Ces deux techniques n'impliquent aucune structure topologique pour modéliser les variétés principales (aucune connexité, seulement un ensemble de sources ponctuelles isolées - les prototypes). Cependant, elles sont à l'origine d'une large famille de techniques que nous passons brièvement en revue maintenant.

Il y a deux familles d'approches en Apprentissage Automatique, qui impliquent des notions de topologie. D'une part, les approches «Apprentissage de Variétés» («Manifold Learning» en anglais) sont basées sur la projection non-linéaire des données dans un espace de plus faible dimension dont la topologie est dans une large mesure fixée *a priori*. Leur objectif principal est de visualiser les données comme pour les Carte Auto-Organisées (Kohonen, 2001; Bishop et al., 1998), leurs versions à topologie adaptative (Fritzke, 1992; Alahakoon et al., 1998) ou les versions continues appelées courbes et surfaces principales (Hastie et Stuetzle, 1989; Chang et Ghosh, 2001) et leur version générative (Tibshirani, 1992), ou de réduire leur dimension comme prétraitement de méthodes de classification (Tipping et Bishop, 1999; de Silva et Tenenbaum, 2003; Lee et al., 2002).

D'autre part, ce travail traite des approches d'«Apprentissage de la Topologie» («Topology Learning» en anglais), basées sur la construction d'un espace dont la topologie n'est pas contrainte *a priori* mais au contraire est apprise des données, cela au prix de la visualisabilité (possibilité de structures non connexes et de dimensions intrinsèques non homogènes non préservables par projection). Ainsi, suivant quelques hypothèses générales sur les variétés génératrices, on souhaite retrouver la topologie et la géométrie de celles-ci à partir des données.

Certains travaux se basent sur un graphe dont les sommets sont les données, par exemple le graphe des K-plus-proches-voisins (KPPV) utilisé pour estimer les distances géodésiques pour réduire la dimension par projection (de Silva et Tenenbaum, 2003). Cependant il est difficile de régler K, ces méthodes sont sensibles au bruit (Carreira-Perpiñán et Zemel, 2005) et elles ne permettent pas de traiter les données incomplètes. Des approches issues de la Géométrie Algorithmique ont aussi été proposées (Niyogi et al., 2006; Bubenik et Kim, 2007; Chazal et al., 2007), mais souffrent de ces mêmes limites. D'autres travaux (Martinetz et Schulten, 1994; Aupetit, 2003; de Silva et Carlsson, 2004) se sont basés sur la construction d'un graphe ayant pour sommets des prototypes, et dont la connexité tendait à reproduire celle de la structure sous-jacente aux données.

[1] Dans le cas des modèles de mélange, les prototypes sont aussi appelés «composants».

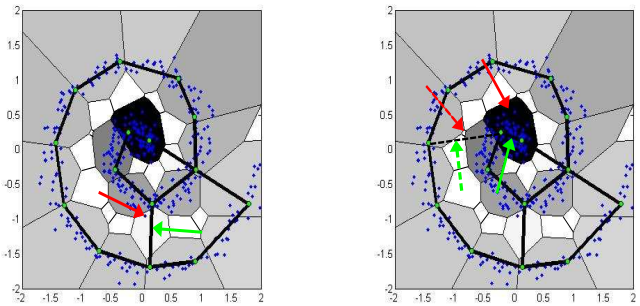


FIG. 1 – **Limite du Competitive Hebbian Learning :** *Un nuage de points (points bleus), un ensemble de prototypes (ronds verts) et la triangulation induite de Delaunay (traits gras). Les ROI des arcs (traits fins) sont coloriées avec un niveau de gris proportionnel au nombre de témoins qu'elles contiennent (blanc nul, noir maximum). (a)* **Sensibilité au bruit** *: un seul témoin (flèche rouge) suffit à créer un arc (flèche verte). (b)* **Forme non pertinente des ROI** *: les ROI (flèches rouges) ne sont pas simplement géométriquement reliées à leurs arcs (flèches vertes).* **Absence de self-consistance** *: certains arcs (flèche pointillée) ne coupent pas leur propre ROI, et la taille de ces ROI peut être si petite (flèche rouge pointillée) que même une trés forte densité de points au voisinage de l'arc ne suffirait pas à générer un témoin de cette arc.*

Martinetz et Schulten (Martinetz et Schulten, 1994) ont proposé un algorithme de construction d'un graphe appelé Triangulation de Delaunay Induite, qui approche le Graphe de Delaunay Restreint défini dans (Edelsbrunner et Shah, 1997). Cet algorithme appelé Competitive Hebbian Learning (CHL), consiste à connecter deux prototypes w_i et w_j s'il existe un point $x \in \underline{x}$ du nuage dont ils sont les premier et deuxième plus proches voisins. Un tel point est appelé «témoin» de l'arc $\{i, j\}$ (de Silva et Carlsson, 2004), et cet arc fait partie du graphe de Delaunay[2] $DG(\underline{w})$ des prototypes. La région de $\mathbb{R}^D$ qui contient tous les témoins d'un arc $\{i, j\}$, est appelée "Région d'Influence" (ROI) de cette arc.

Dans (Fritzke, 1995), une variante du CHL est présentée. Elle est basée sur un processus d'ajout dynamique de prototypes et de liens appelé Growing Neural Gas (GNG). Du point de vue de l'Apprentissage Automatique, le Competitive Hebbian Learning et le graphe résultant (appelé IDT pour «Induced Delaunay Triangulation») ont certaines limites :

1. **Sensibilité au bruit** (Figure 1 (a)) Un seul point témoin suffit à créer un arc de l'IDT.
2. **Forme des ROI non pertinente** (Figure 1 (b)) Un seuil sur le nombre de témoins minimal de chaque arc a été proposé pour filtrer le bruit (Martinetz et al., 1993; Fritzke, 1995) mais le choix de ce seuil ne repose sur aucun critère objectif. Les ROI sont des polytopes de $\mathbb{R}^D$ rendant difficile le calcul de leur volume pour baser ce seuil sur des probabilités. Elles peuvent aussi être de taille très réduite, pouvant ne contenir aucun

[2]Le graphe de Delaunay des prototypes connecte deux prototypes si leurs cellules de Voronoï sont adjacentes. La cellule de Voronoï d'un point s d'un ensemble S dans un espace vectoriel E suivant une mesure de distance d, est le lieu des points de E dont s est le plus proche suivant d parmi les points S.

témoin alors que la distribution locale de l'échantillon légitimerait pourtant l'existence de l'arc correspondant.

3. **Absence de parcimonie et mauvaise topologie des sources ponctuelles** Comme tout point du nuage a un premier et un second plus proches prototypes, tout prototype ayant des témoins est nécessairement connecté à un autre prototype. Donc une source ponctuelle est représentée par deux prototypes interconnectés au lieu d'un, et une dimension intrinsèque de 1 au lieu de 0 est faussement induite de l'arc les reliant.
4. **Absence de self-consistance**[3] (Figure 1 (b)) La réalisation géométrique $\mathbb{R}_{IDT}$ de l'IDT dans l'espace ambiant $\mathbb{R}^D$, n'est pas une variété self-consistante : même un échantillonnage aussi dense que voulu de $\mathbb{R}_{IDT}$ ne garantit pas de générer des points témoins pour toutes les arcs de cette IDT, car il peut exister des arcs $\{i, j\}$ de l'IDT dont la ROI ne coupe pas leur réalisation géométrique $[w_i, w_j]$.
5. **Absence de mesure de qualité** Il n'a pas été proposé de critère objectif pour mesurer la qualité du graphe obtenu, particulièrement en dimension supérieure à 3 où l'inspection visuelle n'est plus possible, ce qui fait obstacle à l'automatisation du processus sur une base statistique objective permettant la sélection d'un modèle optimal.

En résumé, le CHL est avant tout un moyen pratique et peu complexe en termes d'algorithmie et d'implémentation car basé sur une recherche de plus proches voisins, de générer des arcs du graphe de Delaunay d'un ensemble de prototypes. Il se trouve que le CHL a pour effet de bord de générer des variétés dont la connexité semble visuellement proche de ce que seraient les variétés principales du nuage de points. Au mieux sait-on (Martinetz et Schulten, 1994) que le CHL reproduit cette connexité sous certaines conditions d'échantillonnage des variétés principales. Mais aucune mesure ne permet de quantifier le respect de ces conditions sans connaître a priori les variétés et leur échantillonnage. Il apparaît donc que cet effet de bord n'est pas suffisament contrôlable pour être exploitable dans le contexte de l'apprentissage automatique.

2.1 Notre contribution

Afin de dépasser les limites du CHL, nous avons changé de point de vue. Si nous considérons la densité de probabilité de la population dont le nuage de points est un échantillon, nous souhaitons détecter les régions de faible densité qui séparent les régions de forte densité, et surtout rendre explicite le résultat de cette séparation en termes de connexité. Il nous faut donc un modèle de densité particulier en ce qu'il rend extractible (calculable) l'information sur la connexité. Pour cela, nous nous plaçons dans le cadre des modèles génératifs.

Les modèles de mélanges classiques peuvent être vus comme le pendant génératif des techniques de quantification vectorielle à l'origine du CHL (le «Neural-Gas» (Martinetz et al., 1993)), aussi nous proposons un modèle génératif pour remplacer le CHL[4]. Ce modèle génératif nous fournit naturellement un critère de qualité : la vraisemblance, et permet d'exprimer le

[3] La «self-consistance» a été définie pour les courbes principales par (Hastie et Stuetzle, 1989). Tout point d'une courbe principale est au centre de gravité des points qui se projettent sur lui. Les points de la courbe se projettent sur eux-mêmes. Une courbe principale est donc toujours courbe principale d'elle-même, d'où le terme de "self-consistance"

[4] Une nuance toutefois, notre modèle n'est théoriquement équivalent au CHL pour aucune valeur de ses paramètres, alors que c'est le cas entre les K-moyennes et les modèles de mélange gaussiens à variance nulle et proportions égales

problème d'apprentissage de la topologie dans le cadre théorique clairement défini des modèles de mélange. Dans ce modèle, nous supposons que la réalisation géométrique d'un graphe[5] est la source génératrice du nuage de points. Dans un modèle de mélange classique, seuls les sommets sont pondérés, ici les arcs de ce graphes sont pondérés en plus des sommets. La clef de l'apprentissage de la connexité réside dans l'utilisation d'un graphe comme variété source car on sait calculer, donc extraire, la connexité d'un graphe, ainsi que dans l'attibution de proportions à chacuns de ses arcs, *i.e.* leur propension à générer des données. L'idée est que si un arc à une proportion nulle, il peut être supprimé du graphe et donc du modèle génératif sans en modifier la vraisemblance. Ainsi à vraisemblance égale, le graphe le plus parcimonieux est celui dont les arcs inutiles ont été élagués, et sa connexité est alors proposée comme estimateur de la connexité des variétés principales du nuage de points.

Nous utilisons le critère d'information Bayésien (BIC) pour régler le compromis entre vraisemblance et parcimonie, et nous supposons que les observations issues des variétés génératrices sont perturbées suivant une loi de densité gaussienne isovariée. Nous nommons ce modèle génératif le "Graphe Génératif Gaussien" (GGG).

Nous avons déjà introduit ce modèle dans (Aupetit, 2006) et une version supervisée dans (Gaillard et al., 2008). Ici nous proposons pour la première fois d'utiliser le critère BIC pour le réglage des paramètres et méta-paramètres. Nous proposons un cadre génératif général pour l'apprentissage de la topologie puis nous nous focalisons sur le problème de l'apprentissage de la connexité, et comparons notre approcheà celles de l'état de l'art sur des données réelles multi-dimensionnelles.

3 Le Graphe Génératif Gaussien

3.1 Vers un modèle d'apprentissage automatique de la topologie

Nous décrivons les fondements de notre approche du problème de l'Apprentissage de la Topologie.

De notre étude de l'état de l'art, nous déduisons les propriétés qu'un modèle devrait posséder : (1) le modèle devrait être dégagé au maximum de toute contrainte topologique a priori, i.e. le modèle devrait être le plus flexible possible pour pouvoir modéliser la topologie de n'importe quel variété aussi compliquée soit-elle ; (2) la topologie de ce modèle devrait être extractible, i.e. calculable ; ce modèle devrait être (3) robuste au bruit ; (4) parcimonieux ; (5) self-consistant ; (6) fournir un moyen de mesure de sa qualité et (7) ne pas nécessiter de réglages arbitraires de ces méta-paramètres ; enfin, (8) sa complexité en temps de calcul devrait être raisonnable.

Dans le cadre de l'approximation de fonction, on estime une fonction complexe en combinant des fonctions élémentaires de base issues d'une famille de fonctions suffisament riche (approximation universelle). Pour modéliser des variétés inconnues, nous suivons la même approche en considérant un modèle obtenu par l'assemblage de variétés issues d'une famille suffisament riche pour que l'on puisse obtenir la complexité nécessaire. Afin de rendre calculable l'extraction des caractéristiques topologiques de cette variété modèle, nous devons utiliser des

[5] Dans la suite, nous ne faisons plus la distinction entre un graphe et sa réalisation géométrique, c'est-à-dire son plongement dans un espace vectoriel, ici l'espace des données, que l'on obtient en donnant à chaque sommet une position dans cet espace.

variétés élémentaires «discrète» et en nombre fini. Ainsi la famille des d-boules (point $(d = 0)$, segment $(d = 1)$, disque $(d = 2)$, boule $(d = 3)$...) ne convient pas car elle permet certes d'extraire la dimension intrinsèque locale, mais la connexité est difficile à obtenir (intersection de d-boules). La famille des d-pavés (point $(d = 0)$, segment $(d = 1)$, carré plein $(d = 2)$, cube plein $(d = 3)$...) permet de retrouver la dimension intrinsèque locale (la dimension d du d-pavé qui localement explique le nuage de point), la connexité (on peut assembler les d-pavés par leurs sommets, arcs, faces... et représenter cet assemblage par une structure de graphe qui permet de calculer rapidement la connexité), mais c'est la parcimonie qui n'est pas optimale car il faut 2^d sommets pour représenter un objet de dimension d. C'est pourquoi il est plus intéressant d'utiliser un complexe simplicial (Munkres, 1993), *i.e.* un assemblage de k-simplexes acollés les uns aux autres par leurs facettes. La réalisation géométrique d'un k-simplexe dans $\mathbb{R}^D$ est l'enveloppe convexe d'un ensemble de $k + 1$ points de $\mathbb{R}^D$ (ses sommets) : point $(d = 0)$, segment $(d = 1)$, triangle plein $(d = 2)$, tétrahèdre plein $(d = 3)$... Un exemple de complexe simplicial est le complexe de Delaunay défini comme le dual du complexe formé par les cellules de Voronoï [6]. Du fait de sa nature discrète, de nombreuses caractéristiques topologiques d'un complexe simplicial sont calculables, donc extractibles (de Silva et Carlsson, 2004; de Silva, 2003). En particulier, la dimension intrinsèque locale est donnée par la dimension des simplexes principaux (ceux qui ne sont les facettes d'aucun autre simplexe dans le complexe), et l'arc-connexité est donnée par celle du 1-squelette du complexe, i.e. le graphe sous-jacent au complexe, défini par ses sommets (0-simplexes) et arcs (1-simplexe). Enfin le complexe simplicial est le plus parcimonieux car $d + 1$ sommets (un d-simplexe) suffisent à représenter un objet de dimension d.

Dans le présent travail, nous nous focalisons sur l'extraction de la connexité de $\mathcal{M}^{prin}$, nous proposons de modéliser cette variété avec un sous-graphe du graphe de Delaunay de quelques prototypes localisés au voisinage de celle-ci. Grâce à ce modèle, nous remplissons les propriétés désirables (1) et (2) décrites ci-dessus. La parcimonie est en partie réalisée en terme du nombre de paramètres requis pour modéliser un objet de dimension donnée, mais pas pour le moment en terme du nombre total de variétés élémentaires requises dans l'assemblage.

3.2 L'apprentissage de la topologie dans le cadre génératif

Nous posons le problème de l'Apprentissage de la Topologie comme un problème génératif (Figure 2abc) : soit $\mathcal{M}$ un espace topologique latent et $\mathbb{R}^D$ l'espace Euclidien (espace des observations ou espace ambient). Les variétés principales $\mathcal{M}^{prin}$ sont définies comme l'image de $\mathcal{M}$ par une fonction f : $f(\mathcal{M}) = \mathcal{M}^{prin}$. Cependant, en pratique, nous ne connaissons ni $\mathcal{M}$ ni f, et nous devons nous contenter d'un ensemble fini de points $\underline{x}$ représentant les M données observées, au lieu d'un ensemble de variétés $\mathcal{M}^{prin}$. Les points $x \in \underline{x}$ sont les images par l'application f de points "cachés" $\underline{z} \subset \mathcal{M}$, issus de $\mathcal{M}^{prin}$ suivant une certaine fonction densité de probabilité p^{prin}, et potentiellement corrompus par un bruit de nature inconnue ϵ : $\underline{x} = f(\underline{z}) + \epsilon$. Nous souhaitons extraire la connexité de $\mathcal{M}$ à partir de l'observation du nuage de points $\underline{x}$. L'application f peut modifier la géométrie de $\mathcal{M}$ et éventuellement sa topologie. Cependant, nous supposons que f est un homeomorphisme, donc que la topologie de $\mathcal{M}$ est la même que celle de $\mathcal{M}^{prin}$, et donc que nous n'avons pas besoin d'estimer f. Il suffit de nous

[6] Dans le plan, à chaque sommet des cellules de Voronoï correspond un triangle dans le complexe de Delaunay (aussi appelé "triangulation"), à chaque arc de Voronoï , un arc de Delaunay, et à chaque cellule de Voronoï un sommet de Delaunay. De manière générale, à chaque k-cellule de Voronoï correspond un $(D - k)$-simplexe de Delaunay.

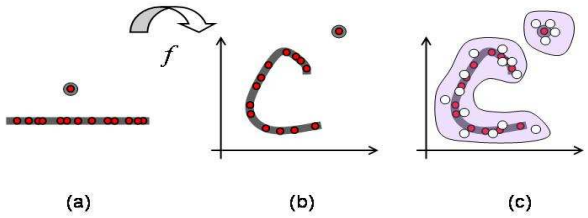


FIG. 2 – **L'Apprentissage de la Topologie vu sous l'angle génératif** *: (a) Un échantillon $\underline{z}$ (ronds rouges) de l'espace topologique $\mathcal{M}$ (un segment de droite et un point). (b) $\mathcal{M}$ et $\underline{z}$ sont plongés dans l'espace des observations suivant f. L'image par ce plongement définit les variétés principales $\mathcal{M}^{prin}$ (courbe gris foncé et point isolé) et une distribution de points $f(\underline{z})$ (ronds rouges). (c) Ces points sont perturbés par un bruit ϵ, menant au nuage de point $\underline{x}$ finalement observé (ronds blancs). L'objectif de l'Apprentissage de la Topologie est de retrouver la topologie de $\mathcal{M}$ (ici un point et un segment de droite) à partir de la seule observation du nuage de points $\underline{x}$ ($\mathcal{M}$, f et les distributions de $\underline{z}$ et de ϵ sont inconnus).*

focaliser sur l'extraction de la topologie de $\mathcal{M}^{prin}$. Pour cela, nous proposons de modéliser $\mathcal{M}^{prin}$ avec un modèle statistique basé sur un graphe permettant l'extraction de cette topologie (en particulier la connexité) avec des méthodes de l'état de l'art (de Silva et Carlsson, 2004)).

Nous introduisons le modèle génératif en simplifiant les hypothèses générales ci-dessus (Figure 2). Au lieu de considérer tout type de variété pour $\mathcal{M}^{prin}$, et du fait des propriétés désirables des complexes simpliciaux, nous supposons que $\mathcal{M}^{prin}$ est une sous-graphe G du graphe de Delaunay de quelques prototypes positionnés dans l'espace ambient. Au lieu de supposer toute densité de probabilité p^{prin} sur $\mathcal{M}^{prin}$, nous supposons que p^{prin} est uniforme le long de chaque arc du graphe G. Et au lieu de supposer n'importe quel type de bruit ϵ, nous supposons un bruit gaussien additif isovarié de moyenne nulle et de variance σ. Ce dernier point satisfait la propriété (3) de prise en compte du bruit (Un modèle de bruit plus sophistiqué peut être envisagé). Le modèle génératif obtenu est donc une somme pondérée de fonctions de densité basées sur les composantes élémentaires du graphe (ses arcs et ses sommets), convoluées avec un bruit gaussien. En tant que modèle génératif, et en considérant un réglage des paramètres et de la complexité du modèle par le critère BIC (parcimonie et mesure de la qualité), ce modèle est par construction le meilleur modèle de densité des points qu'il génère (self-consistance)[7]. Cela satisfait les propriétés (4), (5) et (6). De plus, le cadre statistique fournit des critères objectifs comme le critère BIC pour séléctionner la complexité du modèle, si bien qu'aucun méta-paramètre ne nécessite de réglage manuel arbitraire.

Cependant, les algorithmes (propriété (8)) pour l'apprentissage des paramètres de ce modèle sont relativement complexes (voir section 3.6).

[7]Cela ne signifie pas que le modèle est identifiable, il pourrait exister des modèles ayant même densité et même complexité donc même score BIC mais ayant une topologie différente. Nous discuterons de cela en fin d'article

3.3 Définition formelle du modèle

Etant donné un ensemble de prototypes $\underline{w}$ positionnés au voisinage d'un nuage de points (les données) avec un modèle de mélange gaussien isovarié, le graphe de Delaunay (DG) des prototypes est construit[8]. Chaque arc et chaque sommet du graphe est la base d'un modèle génératif, de sorte que le graphe génère un mélange de densités gaussiennes. Le modèle de mélange résultant représente les données à partir de ces éléments génératifs que nous appelons "points Gaussiens" et "segments Gaussiens", constituant le "Graphe Génératif Gaussien" (GGG).

La valeur de la densité d'un point Gaussien centré sur un prototype $w_j \in \underline{w}$ et de variance σ^2, calculée en un point $x_i \in \underline{x}$ est définie par :

$$g_j^0(x_i;\sigma) = g^0(x_i|w_j;\sigma) = (2\pi\sigma^2)^{-D/2}\exp(-\frac{(x_i-w_j)^2}{2\sigma^2}) \tag{1}$$

Un segment gaussien normalisé est défini comme la somme d'un nombre infini de points gaussiens régulièrement répartis le long d'un segment de droite. Il s'agit donc de l'intégrale d'un point gaussien le long d'un segment de droite (Figure 3). La valeur en un point x_i du segment gaussien $[w_{a_j} w_{b_j}]$ associé au j^e arc $\{a_j, b_j\}$ de longueur L_j du graphe de Delaunay, de variance σ^2 est donnée par :

$$\begin{aligned} g_j^1(x_i;\sigma) \quad &= g^1(x_i|\{w_{a_j}, w_{b_j}\};\sigma) \quad = \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}} L_j}\int_{w_{a_j}}^{w_{b_j}} \exp\left(-\frac{(x_i-t)^2}{2\sigma^2}\right) dt \\ &= \frac{\exp\left(-\frac{(x_i-q_j^i)^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{\frac{D-1}{2}}} \cdot \frac{\mathrm{erf}\left(\frac{Q_j^i}{\sigma\sqrt{2}}\right) - \mathrm{erf}\left(\frac{Q_j^i - L_j}{\sigma\sqrt{2}}\right)}{2L_j} \end{aligned} \tag{2}$$

où $L_j = \|w_{b_j} - w_{a_j}\|$, $Q_j^i = \frac{(x_i - w_{a_j})'(w_{b_j} - w_{a_j})}{L_j}$ et $q_j^i = w_{a_j} + (w_{b_j} - w_{a_j})\frac{Q_j^i}{L_j}$ est la projection orthogonale de x_i sur la droite passant par w_{a_j} et w_{b_j}. Dans le cas où $w_{a_j} = w_{b_j}$, nous posons $g_j^1(x_i;\sigma) = g^0(x_i|w_{a_j};\sigma)$.

Dans l'équation (2), la partie gauche du produit représente le bruit gaussien orthogonal au segment, et la partie droite le bruit gaussien intégré le long du segment (convolution). Les fonctions g^0 et g^1 sont positives et l'on démontre que leur intégrale sur $\mathbb{R}^D$ vaut 1, donc que ce sont des fonctions densités de probabilité.

Un point gaussien est associé à chaque prototype de $\underline{w}$ et un segment gaussien à chaque arc du graphe de Delaunay (DG). Le mélange de gaussiennes est obtenu par une somme pondérée des N_0 points gaussiens et N_1 segments gaussiens, de telle sorte que la somme des poids $\underline{\pi}$ vaut 1 et qu'ils soient positifs ou nuls :

$$p(x_i;\Theta) = \sum_{d=0}^{1}\sum_{i=1}^{N_d} \pi_j^d g_j^d(x_i;\sigma) \tag{3}$$

avec $\sum_{d=0}^{1}\sum_{j=1}^{N_d}\pi_j^d = 1$ et $\pi_j^d \geq 0\ \forall\, j,\, d$ et où $\Theta = \{\underline{\pi}, \underline{w}, \sigma, DG\}$ représente l'ensemble des paramètres du modèle. Le poids π_j^0 (resp.π_j^1) est la probabilité a priori qu'une donnée x soit tirée du point gaussien associé à w_j (resp. du segment gaussien associé au j^e arc de DG).

[8]Les algorithmes pour construire le graphe de Delaunay sont fournis dans (Barber et al., 1996; Agrell, 1993)

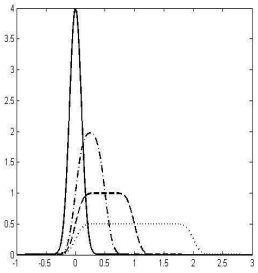


FIG. 3 – **Du point gaussien au segment gaussien :** *segments gaussiens définis sur* $[0; k]$*, avec* $k = 0$ *(trait plein),* $k = 0.5$ *(trait point-tiret),* $k = 1$ *(trait tireté),* $k = 2$ *(trait pointillé).*

Ainsi l'espace latent est un ensemble de points et de segments, qui sont plongés dans l'espace ambient par la réalisation géométrique du graphe de Delaunay. De plus, en toute généralité , les segments latents sont définis de longueur 1, ce qui permet de définir la distribution a priori d'une variable cachée t pour chaque point et chaque segment. Dans notre cas, la distribution a priori sur t pour un point est la distribution de Dirac, et pour un segment, la distribution uniforme. Si nous introduisons une variable latente discrète z indiquant quel élément génératif a généré les données, le processus de génération est le suivant :

- (i) tirage du j^e élément génératif de dimension d avec une probabilité π_j^d, *i.e.* la variable latente z suit une distribution multinomiale de paramètre $\underline{\pi}$
- (ii) tirage de la variable latente t suivant une distribution de Dirac ou uniforme fonction de la nature point ou segment du j^e élément ;
- (iii) tirage de la donnée x_i suivant une distribution gaussienne de variance σ^2 et de moyenne w avec $w = w_j$ si le j^e élément est un point, et $w = w_{a_j} + \frac{(w_{b_j} - w_{a_j})}{L_j} t$ sinon (où a_j et b_j sont les extrémités du segment j^e élément).

Le GGG en tant que modèle génératif est présenté sur la figure 4.

3.4 La vraisemblance comme mesure de qualité et sa maximisation avec l'algorihthme EM

La fonction $p(x_i; \underline{\pi}, \underline{w}, \sigma, DG)$ est la densité de probabilité au point x_i sachant les paramètres du modèle. Nous mesurons la vraisemblance P des données $\underline{x}$ par rapport aux paramètres $\Theta = \{\underline{\pi}, \underline{w}, \sigma, DG\}$ du modèle GGG :

$$P(\Theta; \underline{x}) = \prod_{i=1}^{M} p(x_i; \underline{\pi}, \underline{w}, \sigma, DG) \tag{4}$$

Afin de maximiser la vraisemblance P par rapport à $\underline{\pi}$ et σ, nous utilisons le cadre *EM* (Dempster et al., 1977). L'idée clef de l'algorithme *EM* est de considérer que les données observées x_i ne sont qu'une partie des données dites *complètes*, et que la maximisation de la

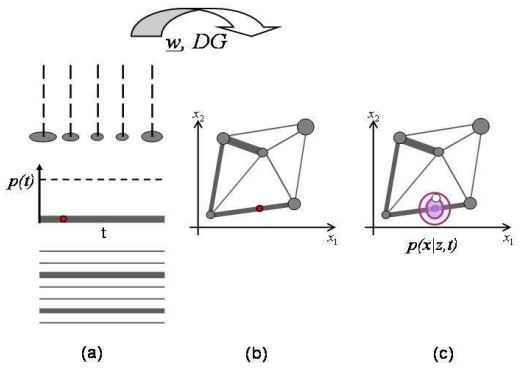


FIG. 4 – **Génération des données avec le GGG :** *Pour résoudre le problème d'apprentissage de la topologie (figure 2), nous définissons un espace latent qui correspond à un ensemble de points et de segments (a), plongé dans l'espace d'observation par la réalisation géométrique du graphe de Delaunay (b).(a) Une donnée observée x est générée en sélectionnant un composant z suivant la distribution a priori $p(z)$ puis sachant ce composant, en tirant une valeur t (disque rouge) suivant la distribution a priori $p(t)$. (c) Enfin, en tirant le vecteur x (disque blanc) d'une distribution gaussienne isotrope (cercles magenta).*

vraisemblance associée à ces données complètes est facile. On définit les données complètes (x_i, z_i, t_i), où z_i et t_i sont des variables cachées.

L'algorithme *EM* effectue itérativement une étape de calcul de l'espérance (étape *E*) puis une étape de maximisation (étape *M*), qui garantie la convergence vers un maximum local de la vraisemblance (Boyles, 1983). Durant l'étape E, les données manquantes sont estimées à partir des données observées et de l'estimation courante des paramètres du modèle. Durant l'étape M, la vraisemblance est maximisée sous l'hypothèse que les données manquantes sont connues. L'estimation des données manquantes par l'étape E est utilisée à la place des véritables valeurs que l'on ne connaît pas. Les règles d'adaptation prennent en compte les contraintes de positivité et de somme unitaire des paramètres :

$$\begin{array}{rcl} \pi_j^{d[\text{new}]} & = & \frac{1}{M}\sum_{i=1}^{M} \tilde{z}_{ij}^{d} \\ \sigma^{2[\text{new}]} & = & \frac{1}{DM}\sum_{i=1}^{M}\Big[\sum_{j=1}^{N_0}\tilde{z}_{ij}^{0}(x_j - w_i)^2 \\ & & + \sum_{j=1}^{N_1}\tilde{z}_{ij}^{1}\frac{g^0(x_i|q_j^i;\sigma)(I_1[(x_i-q_j^i)^2+\sigma^2]+I_2)}{L_j \cdot g_j^1(x_i,\sigma)}\Big] \end{array} \tag{5}$$

où $\tilde{z}_{ij}^{d} = p(d, j|x_i; \Theta) = \frac{\pi_j^d g_j^d(x_i;\sigma)}{\sum_{d=0}^{1}\sum_{j=1}^{N_d}\pi_j^d g_j^d(x_i;\sigma)}$ est la probabilité a posteriori que la donnée x_i soit générée par le j^{me} élément génératif de dimension d (écrit (d, j)), et avec :

$$\begin{array}{rcl} I_1 & = & \sigma\sqrt{\frac{\pi}{2}}(\text{erf}(\frac{Q_j^i}{\sigma\sqrt{2}}) - \text{erf}(\frac{Q_j^i - L_j}{\sigma\sqrt{2}})) \\ I_2 & = & \sigma^2\left((Q_j^i - L_j)\exp(-\frac{(Q_j^i - L_j)^2}{2\sigma^2}) - Q_j^i\exp(-\frac{(Q_j^i)^2}{2\sigma^2})\right) \end{array} \tag{6}$$

Jusqu'à présent, la position des prototypes qui sont les sommets du graphe de Delaunay, est figée une fois que le graphe est construit. Afin d'accroître la vraisemblance du modèle par rapport aux données, il serait intéressant de mettre à jour la position des prototypes. Cependant, l'étape M impliquant les prototypes n'est pas triviale, donc nous proposons une *approximation* de l'étape M. Nous observons empiriquement que la règle suivante accroît la plupart du temps la vraisemblance. Si ce n'est pas le cas pour un prototype, la règle n'est pas appliquée et la position de ce prototype n'est pas modifiée (on passe au suivant). L'étape M *approchée* prend en compte la probabilité que les données soient générées par un prototype (w_k , $k \in [1, 2, ..., N_0]$) et par les arcs ayant ce prototype comme extrémité :

$$w_k^{[\text{new}]} = \frac{\sum_{i=1}^{M} \left[\tilde{z}_{ik}^0 x_i + \sum_{j \in W_k} \tilde{z}_{ij}^1 \frac{g^0(x_i|q_j^i;\sigma)}{L_j \cdot g_j^1(x_i;\sigma)}(-E_2 w_{b_j} + E_3 x_i)\right]}{\sum_{i=1}^{M} \left[\tilde{z}_{ik}^0 + \sum_{j \in W_k} \tilde{z}_{ij}^1 E_1\right]} \tag{7}$$

où W_k représente l'ensemble des arcs $[w_{a_j}, w_{b_j}]$ ayant $w_k = w_{a_j}$ comme extrémité, et où

$$\begin{aligned}
E_1 &= \frac{\sigma^2}{L_j^2}[e^{\frac{-(Q_j)^2}{2\sigma^2}}(Q_j - 2L_j) - e^{\frac{-(Q_j - L_j)^2}{2\sigma^2}}(Q_j - L_j)] + \frac{1}{L_j^2}((L_j - Q_j)^2 + \sigma)I_1 \\
E_2 &= \frac{\sigma^2}{L_j^2}[e^{-\frac{(Q_j - L_j)^2}{2\sigma^2}} Q_j - e^{-\frac{(Q_j)^2}{2\sigma^2}}(Q_j - L_j)] - \frac{1}{L_j^2}(Q_j^2 - L_j Q_j + \sigma^2)I_1 \\
E_3 &= \frac{1}{L_j}[e^{\frac{-(Q_j - L_j)^2}{\sigma^2}} - e^{\frac{-Q_j^2}{\sigma^2}} + (Q_j - L_j)I_1]
\end{aligned} \tag{8}$$

La règle de mise à jour des positions des prototypes peut être intégrée dans un schéma *EM*, appelé algorithme *EM généralisé* (GEM) (Dempster et al., 1977) (voir (Gaillard, 2008) pour le développement des équations). Cet algorithme correspond dans notre cas à d'abord effectuer l'étape E puis à mettre à jour durant l'étape M, les paramètres $\underline{\pi}$ plus l'un des deux paramètres σ ou $w_k \in \underline{w}$. Le développement in extenso des règles de mise à jour sont fournis dans (Gaillard, 2008).

Le principe du GGG est présenté sur la figure 5.

3.5 Emergence de la topologie et sélection de modèle par le critère BIC

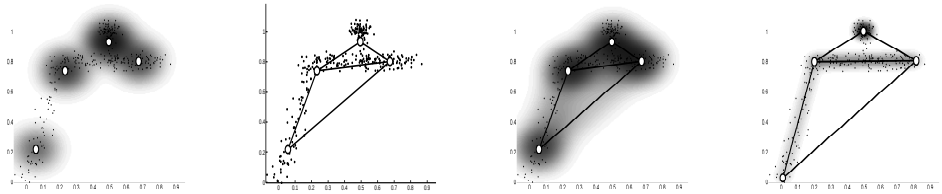
Pour obtenir le graphe représentant la topologie (TRG) à partir du modèle génératif, l'idée clef est de supprimer du graphe DG des prototypes, les éléments gaussiens qui n'ont aucune chance d'avoir généré des données, *i.e* les éléments associés à un poids faible : $\pi_j^d < \gamma_{N_0}$.

Soit $G = \{\sigma, \underline{w}, E, \underline{\pi}\}$ le graphe génératif défini par sa variance du bruit σ^2, ses N_0 sommets $\underline{w}$, son ensemble d'arcs E et ses proportions $\underline{\pi}$. Et soit $G_{N_0,\gamma_{N_0}} = \{\sigma, \underline{w}, E, \underline{\pi} \,|\pi_j^d \geq \gamma_{N_0}\}$, le graphe génératif qui contient seulement les éléments génératifs ayant une proportion plus élevée que le seuil γ_{N_0} : $\pi_j^d \geq \gamma_{N_0},\ \forall d \in \{0, 1\}$.

En réglant le paramètre γ_{N_0} de 1 à 0, on obtient une séquence de graphes génératifs emboîtés allant de l'ensemble vide au graphe DG complet :

$$G_1 = \emptyset \subseteq \ldots \subseteq G_{\gamma_{N_0}} \subseteq \ldots \subseteq G_0 = DG \tag{9}$$

Supposons que l'un des graphes de la séquence ait la «bonne» topologie, i.e. celle inconnue des variété principales, nous utilisons le Critère d'Information Bayésien (BIC) pour sélectionner ce graphe. Le critère BIC (Schwartz, 1978) est adapté et satisfaisant en pratique pour la



(a) (b) (c) (d)

FIG. 5 – **Principe du Graphe Génératif Gaussien** *: (a) 300 données (points noirs) sont tirés d'un ensemble de variétés du plan $\mathcal{M}^{prin}$: un segment oblique, un segment horizontal et un point isolé, de proportions respectives $\{0.25; 0.5; 0.25\}$, perturbées par un bruit gaussien isotrope de moyenne nulle et de variance $\sigma^2 = 0.001$. Les prototypes (disques blancs) sont positionnés à l'aide d'un modèle de mélange gaussien. (b) Les prototypes sont connectés avec les arcs du graphe de Delaunay. (c) La densité de probabilitégénérée par le graphe génératif gaussien initial. (d) La densité générée par le GGG optimal obtenu par maximisation de la vraisemblance en fonction de σ, $\underline{\pi}$ et $\underline{w}$. Noter comment les arcs se déplacent pour mieux expliquer les données générées par les segments, et comment les proportions diminues lorsqu'il n'y a pas de données sous-jacente à expliquer.*

sélection du nombre de composants d'un modèle de mélange classique (Roeder et Wasserman, 1997; Fraley et Raftery, 2002). Ici nous l'utilisons pour régler le compromis entre la vraisemblance $P(\underline{x}; G_{\gamma_{N_0}})$ et la complexité du graphe génératif $G_{\gamma_{N_0}}$. Donc nous supposons que la topologie du modèle génératif d'un ensemble de points est un estimateur de la topologie des variétés principales de cet ensemble, dont la qualité est mesurée par le critère BIC de vraisemblance pénalisée :

$$BIC(N_0, G_{\gamma_{N_0}}) = -log(P(\underline{x}; G_{\gamma_{N_0}})) + \frac{v_{\gamma_{N_0}}}{2} log(M) \tag{10}$$

où $v_{\gamma_{N_0}}$ est le nombre de paramètres libres du modèle $G_{\gamma_{N_0}}$ et M le nombre de données.

La vraisemblance de chaque graphe génératif $G_{\gamma_{N_0}}$ de la séquence est optimisée en fonction des proportions[9] $\underline{\pi}$ de telle sorte que tous les graphes génératifs sont à leur maximum de vraisemblance. A la fin, le graphe représentant la topologie pour un nombre de prototypes N_0 donné, est celui défini par le graphe génératif $G_{\gamma^*_{N_0}}$ minimisant le critère BIC à N_0 fixé (eq. (10)). On recommence pour chaque valeur N_0 et l'on retient finalement le TRG optimal associé au couple de paramètres $(N_0^*, \gamma^*_{N_0^*})$ qui minimise BIC. L'algorithme complet est donné sur la figure 6 et est illustré sur la figure 7.

3.6 Complexité des temps de calcul

La complexité en temps de calcul du *GGG* est $O(D(N_0 + N_1)Mt_{max})$ plus le temps $O(DN_0^3)$ requis pour construire le graphe de Delaunay (Agrell, 1993) qui domine le temps total au pire cas. D'autres graphes non-paramétriques moins gourmands en temps de calcul

[9]Pour des raisons de temps de calcul, seuls les proportions $\underline{\pi}$ sont optimisées durant cette étape. En commençant par les proportions normalisés obtenus avec le GGG complet, le problème est convexe et l'algorithme converge rapidement. La variance du bruit est aussi supposée peu variable.

Algorithme : Graphe Génératif Gaussien	
Entrée	Choisir $N_0^{\max}$, le nombre de prototypes maximal
POUR chaque	$N_0 \in \{1, ..., N_0^{\max}\}$
Initialisation	Positionner les prototypes $\underline{w}$ avec un GM isotropique
	Construire le DG (ou l'IDT) des prototypes
	Initialiser $\underline{\pi}$ à $1/(N_0 + N_1)$
	Initialiser σ à la valeur trouvée par le GM
EM	Utiliser la règle de mise à jour (5) pour trouver σ^*,
	$\underline{\pi}^*$, $\underline{w}^*$ maximisant la vraisemblance P.
Elagage	Construire une séquence emboîtée de GGG par rapport à $\underline{\pi}^*$
BIC	Calculer le critère $BIC(G_{N_0,\gamma^*_{N_0}}) = min_{\gamma_{N_0}}(BIC(N_0, G_{\gamma_{N_0}}))$ (10)
FIN POUR	
Sortie	Retourner le TRG minimisant BIC pour tout $N_0 \in \{1, ..., N_0^{\max}\}$
	et ce critère BIC minimal

FIG. 6 – **Algorithme du Graphe Génératif Gaussien** *GM : modèle de mélange gaussien ; DG : graphe de Delaunay ; IDT : Triangulation Induite de Delaunay ; TRG : graphe représentant la topologie ; BIC : critère d'information bayésien.*

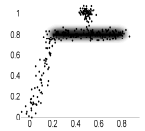
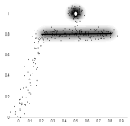
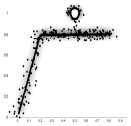
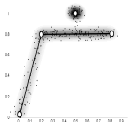
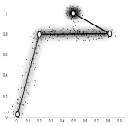
$\gamma_{\mathbf{N_0^*}} = \mathbf{0.48}$	$\gamma_{\mathbf{N_0^*}} = \mathbf{0.25}$	$\gamma_{\mathbf{N_0^*}} = \mathbf{0.23}$	$\gamma_{\mathbf{N_0^*}} = \mathbf{0.0022}$	$\gamma_{\mathbf{N_0^*}} = \mathbf{1^{-100}}$
$log(P) = -9871$	$log(P) = -8310$	$log(P) = 527$	$log(P) = 529$	$log(P) = 529$
$BIC = 19783$	$BIC = 16662$	$BIC = -1009$	$BIC = -1013$	$BIC = -1007$
				
(a)	(b)	(c)	(d)	(e)

FIG. 7 – **Emergence de la topologie avec le critère d'information bayésien (BIC)** *: (a-e) 5 graphes génératifs $G_{\gamma_{N_0^*}}$ issus de la séquence de graphes emboîtés basées sur $\gamma_{N_0^*}$ (ici $N_0^* = 4$). Pour chaque graphe génératif présenté, la valeur du seuil $\gamma_{N_0^*}$, la log-vraisemblance et la valeur de BIC associée sont fournies au-dessus de chaque tracé. (d) Le graphe génératif de la séquence emboîtée qui minimise BIC. Ce graphe est le graphe représentant la topologie (TRG) dont la connexité estime celle des variétés principales.*

peuvent être envisagés à la place du DG, comme le graphe des plus-proche-voisins (NNG) $O(DN_0^2)$, l'arbre recouvrant minimal (MST) $O(DN_0^2)$, le graphe des voisins relatifs (RNG) $O(DN_0^3)$, les graphe de Gabriel (GG) $O(DN_0^3)$ ou la triangulation induite de Delaunay (IDT) $O(DN_0M)$. Cependant, ces graphes portent moins d'information que le graphe de Delaunay dont ils sont des sous-graphes[10].

Quand la dimension des données augmente, nous proposons de considérer l'IDT (obtenu avec l'algorithme CHL) au lieu du graphe de Delaunay comme graphe initial du GGG. En effet, l'IDT contient en générale plus d'arcs que nécessaire pour modéliser la connexité. Nous utilisons l'IDT lorsque la dimension D est supérieure à 4.

4 Expériences

4.1 Problème jouet

Dans ces expériences, nous souhaitons vérifier la pertinence du GGG pour apprendre la connexité d'un ensemble de données jouets. L'ensemble (Figure 8) consiste en 300 points du plan, tirés d'une spirale et 200 d'un point isolé. Les points sont perturbés par un bruit gaussien additif de moyenne 0 et de variance 0.0025.

Sur la figure 8, nous comparons le TRG obtenu avec le GGG, à ceux obtenus avec le CHL, le CHL filtré pour lequel les arcs qui ont un nombre de témoins inférieur à un seuil T sont supprimés, et le GNG. Le GGG est optimisé avec l'algorithme décrit en section 3.4 afin de retrouver les variétés principales. Nous utilisons le critère BIC pour sélectionner la complexité du modèle avec différentes valeurs pour le nombre de prototypes N_0. A la fin, le graphe génératif optimal comprend 13 prototypes. Pour le CHL et le CHL filtré, nous utilisons les prototypes trouvés pour le GGG optimal. Pour le CHL filtré, nous fixons le seuil T de telle sorte que le graphe obtenu corresponde visuellement au mieux à la solution attendue. Cependant, rappelons que régler le seuil T requiert un contrôle visuel et ne correspond à l'optimum d'aucune fonction d'énergie, ce qui rend totalement arbitraire l'utilisation de ce modèle en dimension supérieure à 3. Pour le GNG, nous définissons le nombre maximum de prototypes comme le nombre de prototypes trouvés pour le GGG optimal. Tous les autres paramètres du GNG sont réglés aux valeurs indiquées dans l'article original (Fritzke, 1995). En particulier, le paramètre d'âge[11] est fixé arbitrairement à 50. Dans toutes les expériences suivantes, la méthodologie est la même.

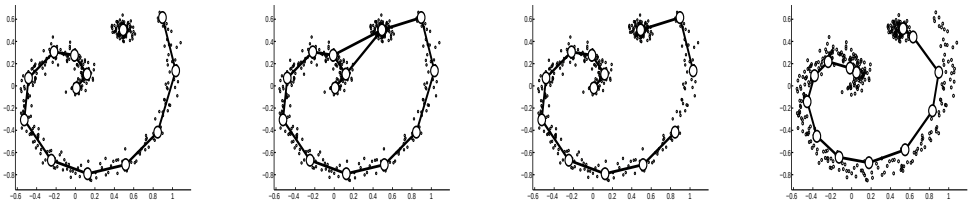
Le CHL filtré ou non, et le GNG ne sont pas en mesure de retrouver la vraie connexité de l'ensemble de points. En particulier, un groupe de points isolé (Figure 8 (b)) si $T = 0$. En remarquant que $T_1 < T_2 \Rightarrow IDT(T_2) \subseteq IDT(T_1)$, on peut voir sur la figure 8c qu'aucun seuil T ne permet de retrouver la vraie connexité de l'ensemble de points.

4.2 Connexité des données Teapot

Les données Teapot originales ont été créées à partir de la vue sous différents angles, d'une théière en rotation autour d'un axe vertical (Weinberger et Saul, 2006) (Figure 9). Dans cette

[10] En particulier $NNG \subset MST \subset RNG \subset GG \subset DG$ (Veltkamp, 1991)

[11] Notons qu'aucune méthode objective n'est fournie dans (Fritzke, 1995) pour régler ces paramètres.



(a) GGG (b) CHL : $T = 0$ (c) CHL : $T^* = 19$ (d) GNG

FIG. 8 – **Une spirale et un point isolé** *: le TRG défini par (a) le GGG, (b) le CHL, (c) le CHL filtré, (d) le GNG. Le GGG est le seul modèle capable de retrouver la vraie connexité.*

FIG. 9 – **Huit images de la base de données «teapot» originale** *: la couleur des boîtesencode le degré de rotation de la théière. Quelques images ont été retirées entre les boîtes rouges et bleues foncées, et entre les boîtes bleue-vertes et vertes clair, pour créer artificiellement un ensemble composé de deux composantes connexes.*

expérience nous utilisons les données fournies par Zhu et Lafferty (2005) [12] où les images sont converties en échelle de gris et réduites à une taille de 12×16 pixels. Ce nouvel ensemble de données a été conçu dans le cadre de la reconnaissance de formes, pour tester des algorithmes devant déterminer automatiquement si l'anse se trouve à droite ou à gauche. Donc les images dans lesquelles l'anse se trouve dans l'axe de la prise de vue sont supprimées. Finalement, 365 images sont disponibles. Bien qu'en dimension 192, les données se trouvent au voisinage d'une variété de dimension 1 paramétrée par l'angle de rotation de la prise de vue, et cette variété est séparée en deux composantes connexes, l'une contenant les images avec l'anse à droite, l'autre celles avec l'anse à gauche. Nous voulons retrouver ces caractéristiques topologique (2 composantes connexes de dimension intrinsèque 1).

Afin d'analyser la structure sous-jacente aux données, les techniques de réduction de dimension sont largement utilisées. Cependant, du fait de la perte d'information qu'elles engendrent, la plupart des distances visualisées sont soit comprimées soit étirées, et il est donc difficile de savoir si les formes observées existent ou non dans l'espace ambient (Aupetit, 2007). Dans les expériences suivantes, nous montrons que le GGG est une méthode complémentaire aux techniques de projection "classiques" pour analyser un ensemble de données.

Nous utilisons l'Analyse en Composantes Principales (Bishop, 1995) (ACP), le Generative Topographic Mapping (Bishop et al., 1998) (GTM)[13] et ISOMAP (de Silva et Tenenbaum, 2003)[14] pour visualiser les données Teapot (figure 10).

[12] Données fournies sur le site internet http ://pages.cs.wisc.edu/∼jerryzhu

[13] Une implémentation Matlab du GTM est fournie à l'adresse internet http ://www.ncrg.aston.ac.uk/GTM/

[14] Une implémentation Matlab d'ISOMAP est fournie à l'adresse internet http ://web.mit.edu/cocosci/isomap/code/

PCA (a) | GTM (b) | ISOMAP 10-NN (c) | ISOMAP 15-NN (d) | ISOMAP 20-NN (e)

FIG. 10 – **Cinq projections des données «teapot»** *: la couleur des points correspond à la couleur utilisée sur la figure 9. Projection des données «teapot» suivant les deux premiers axes principaux (ACP) (a), par le GTM défini par une grille* 20×20 *et une transformation non linéaire obtenue par une grille* 10×10 *de fonctions gaussiennes (b), et par ISOMAP utilisant un graphe des* K*-plus-proches voisins avec* $K = \{10, 15, 20\}$ *(c-e). Aucune de ces projections n'est en mesure de montrer la structure déconnectée originelle.*

Nous optimisons le GGG avec l'algorithme décrit en section 3.4 avec un N_0 candidat compris entre 40 et 80, afin de retrouver la connexité des variétés principales de cet ensemble. Le graphe génératif optimal est finalement défini par 67 prototypes.

Le TRG résultant nous informe sur l'existence de 2 composantes connexes en dépit de ce que montrent les techniques de projection classiques (figure 10). L'analyse des degrés des sommets[15] du TRG (les degrés valent 2, sauf pour les 4 sommets extrémités des deux composantes connexes, dont le degré vaut 1) montre que chaque composante est une chaîne de sommets, donc une variété homéomorphe à un segment, montrant que la dimension intrinsèque de ces deux variétés vaut 1.

De plus, le modèle étant génératif nous savons aussi que les deux variétés ont à peu près la même probabilité a priori : 0.507 et 0.493, et que le long des variétés, les données sont à peu près uniformément distribuées, puisque la moyenne et la variance de la quantité $\frac{\pi_j^1}{L_j}$ sont respectivement : $4.5966e-005$ et $1.5720e-010$.

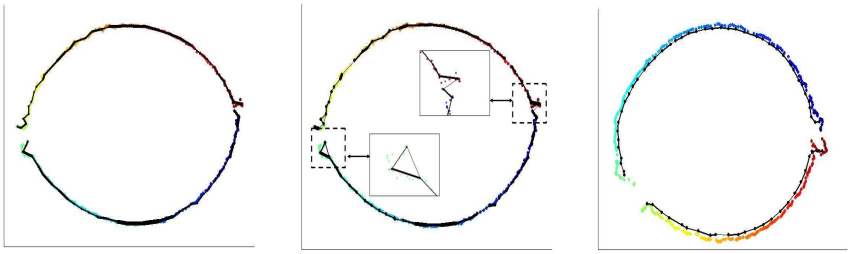
Enfin, nous comparons le TRG obtenu par le GGG avec celui obtenu par le CHL et le GNG. Pour le CHL, nous utilisons les prototypes obtenus par le GGG optimal. Pour le GNG, le processus de croissance est maintenu jusqu'à obtenir 67 prototypes.

La figure 11 montre que le CHL ne permet pas de retrouver la connexité des données. En particulier, le TRG obtenu par le CHL ne possède qu'une seule composante connexe qui contient des sommets de degré 3 ou plus. Le GNG permet de retrouver la bonne connexité mais ses méta-paramètres doivent être réglés manuellement sans critère objectif.

4.3 Classification non supervisée d'images avec le TRG

Dans cette expérience, nous avons sélectionné les images de 5 objets (figure 12) de la base d'objets fournie par l'Amsterdam Library of Object Images (ALOI) Geusebroek et al. (2005). Pour chaque objet, un camera a enregistré 72 images en faisant tourner l'objet autour d'un axe vertical avec un pas angulaire de 5 degrés. A nouveau, la taille des images est réduite à 12×16

[15] Dans un graphe, le degré d'un sommet est le nombre d'arcs qui ont ce sommet comme extrémité



(a) (b) (c)

FIG. 11 – **Projections de différent Graphes Représentant la Topologie (TRG)** *: le code de couleurs est celui défini sur la figure 9. (a-c) Les données et les prototypes sont projetés par ISOMAP à partir du graphe des* 15*-PPV. Les projections diffèrent de celles des figures 10 (d-f) car les prototypes sont inclus dans l'ensemble projeté. Projection du TRG obtenu avec le GGG (a), le CHL (b) et le GNG (c). l'épaisseur des arcs est proportionnelle aux proportions* $\underline{\pi}^1$ *du GGG (a), au nombre de points témoins des arcs du CHL (b) et à l'âge final des arcs du GNG (c). Le GGG et le GNG permettent de retrouver la vraie connexité, mais les méta-paramètres du GNG ont être réglés à la main donc arbitrairement.*

FIG. 12 – **Cinq objets de la base ALOI.** *De gauche à droite : une carte à jouer, un kiwi, une balle, une alarme, un tube.*

pixels. Nous nous attendons à trouver dans l'espace des pixels à 192 dimensions, 5 groupes de 72 points chacuns correspondant aux images d'un objet particulier.

Nous utilisons l'ACP, le GTM et ISOMAP pour visualiser l'ensemble de points. Les figures 13(a-c) montrent que les techniques classiques de projection ne permettent pas de détecter visuellement 5 groupes de points.

Nous optimisons le GGG avec l'algorithme de la section 3.4 avec un N_0 candidat compris entre 60 and 90, pour retrouver la connexité des variétés principales de cet ensemble de points. Enfin, nous assimilons chaque composante connexe du TRG obtenu, à un groupe de points.

Comme le TRG correspond à un modèle génératif, chaque donnée peut être assignée à un groupe par la règle du maximum a posteriori. Nous comparonscette classification non supervisée avec celle obtenue avec le CHL et le GNG. Pour ces deux modèles, les données appartiennent à la composante connexe contenant leur prototype le plus proche. Pour comparer la qualité des différentes classifications, nous utilisons la fonction d'erreur suivante :

$E = M(M-1)/2 \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \delta_{ij}$ où δ_{ij} vaut 1 si les données i et j appartiennent par erreur au même groupe ou à un groupe différent, et 0 sinon.

Nous répétons 10 fois la procédure (l'intialisation aléatoire des paramètre initiaux et la

convergence vers un optimum local de la fonction optimisée mènent à des paramètres otpimaux et donc des classifications différentes.). Nous reportons sur la figure 14 la moyenne de l'erreur de classification et la moyenne du nombre de composantes connexes trouvées par chaque algorithme. Les résultats expérimentaux montrent que le GGG est meilleur que les autres méthodes suivant ces deux indices de qualité : il fournit des estimations plus précises et plus stables de la vraie connexité.

Sur la figure 13(d), nous traçons la meilleure classification obtenue par chaque algorithme après 10 essais. Le graphe génératif optimal est défini par 80 prototypes. Le TRG résultant nous montre qu'il existe 5 variétés séparées, et la classification ne fait aucune erreur. Remarquons que les classifications basées sur des prototypes (K-Means ou Mélanges de gaussiennes) nous auraient fourni autant de groupes que de prototypes. En comparaison, le CHL fournit seulement 3 composantes connexes, et le GNG en trouve 7 (il sépare un groupe légitime en 4 groupes, et connecte deux groupes normalement séparés).

4.4 Des données vectorielles aux données de type graphe

Dans cette expérience, nous montrons comment un TRG peut être utilisé pour coder l'information topologique contenue dans des données de type nuages de points. Nous utilisons la base de données MNIST qui regroupe des images de chiffres manuscrits. Les 100 premières images de chaque chiffre (soit 1000 images de 28×28 pixels) sont transformées en images binaires (noir et blanc) [16], et pour chaque image, on positionne une données à la place de chaque pixel blanc dans le plan image. L'ensemble constitue une base de 1000 nuages de points, chacun de ces nuages représentant un chiffre et pouvant contenir jusqu'à 784 points. Nous construisons le TRG à partir du GGG sur chacun de ces 1000 nuages de points (Quelques exemple de TRG obtenus sont présentés sur la figure 15).

Nous classifions manuellement chaque TRG en fonction de sa topologie, et nous obtenons 7 classes principales d'homotopie representées sur la première ligne du tableau 16. Chaque classe d'homotopie peut être représentée par une structure de graphe, et donc chaque image d'un chiffre initialement vue comme un nuage de points, est transformée en un donnée de type graphe, contenant l'essentiel de l'information topologique.

En quoi ces classes peuvent-elles être utiles ? La plupart des chiffres appartiennent en majorité à une ou deux classes différentes. Donc pour une image d'un chiffre inconnu, le calcul de sa classe d'homotopie avec un TRG fournit un a priori sur la probabilité que ce caractère manuscrit soit tel ou tel chiffre. Aussi il nous semble prometteur de combiner un classifieur basé sur les pixels avec un classifieur basé sur la classe d'homotopie extraite du TRG obtenu par un GGG. Pour l'instant, ce travail reste à effectuer pour évaluer l'apport de cette approche. Notons aussi que les classes d'homotopie ont été déterminées manuellement, il faudrait concevoir une méthode de calcul automatique de ces classes.

[16] La base MNIST est un sous-ensemble d'une base plus grande du NIST, dans laquelle les images sont en noir et blanc.

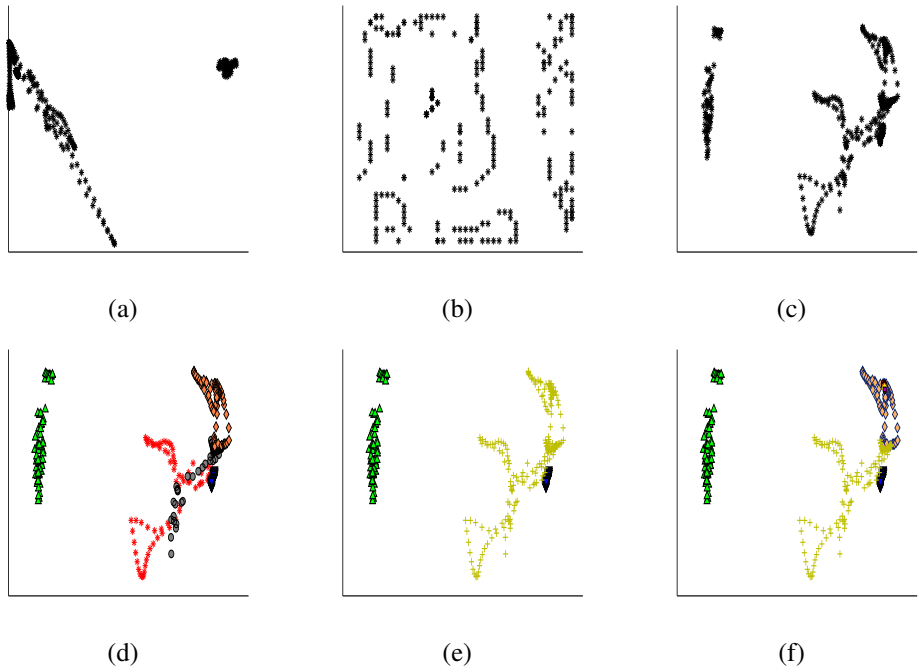


FIG. 13 – **Analyse des données ALOI** *: (a-c) Projections des données ALOI avec (a) l'ACP, (b) le GTM (grille* 40×40 *et transformation non linéaire sur une grille de* 10×10 *fonctions gaussiennes, (c) ISOMAP (*70*-PPV). (d-f) Projection des données par ISOMAP des graphes obtenus par les différents algorithmes. La couleur indique les classes en termes de composantes connexes des graphes GGG (d), CHL (e), et GNG (f). Sur le tracé (d), où les couleurs sont aussi en accord avec la classe réelle, les cartes à jouer sont représentées par un* $*$*, les kiwis par un* ∇*, les balles par un* Δ*, l'alarme par un* O *et le tube par un* $\diamond$*. Le GGG est le seul capable de retrouver les vraies classes. Par ailleurs, on peut noter que sans connaître les classes à l'avance, les différentes projections (a-c) sont trompeuses quant au nombre original de classes à trouver. L'ACP (a) semble montrer qu'il existe deux classes distinctes, le GTM (b) fournit un trés grand nombre de classes, et ISOMAP montre* 3 *classes distinctes, alors que les points de la classe de gauche appartiennent en fait à une unique classe comme on le voit en vert sur le tracé (d).*

	GGG	CHL	GNG
E (%)	0.1 ± 0.2	13.0 ± 8.3	4.4 ± 2.1
# connected components	5.4 ± 0.9	3.0 ± 0.7	7.1 ± 0.8

FIG. 14 – **Robustesse de la classification par composantes connexes des images ALOI**

FIG. 15 – **GGG sur les images MNIST** *: Quelques exemples de TRG obtenus avec le GGG travaillant dans le plan image sur les images MNIST.*

	a	b	c	d	e	f	g	Other
0	95		3					2
1			98		1			1
2		7	52	19	11			11
3			42	18	31			9
4	2	2	16		52	10		18
5			64	26	3			7
6			25	63				12
7	6		84		2	4		2
8	4	8	18	9			45	16
9	2	2		85	4			7

FIG. 16 – **Topologie des images MNIST** *: la topologie du TRG résultant de l'application du GGG sur chacun des 1000 chiffres, a été classée manuellement en 7 classes principales d'homotopie représentées sur la première ligne. Par exemple, on peut lire que* 19% *des TRG modélisant le chiffre* 2 *ont la même connexité qu'un cercle connecté à un segment (colonne d).*

4.5 Discussion

4.5.1 Avantages et limites du GGG

Le Graphe Génératif Gaussien (GGG) permet de contourner les limites de l'algorithme Competitive Hebbian Learning (CHL) pour modéliser la connexité. En particulier, il permet de prendre en compte le bruit, et de mesurer la qualité du modèle, même lorsqu'aucune visualisation n'est possible. Cependant, la complexité en temps de calcul est plus élevée : $O(D(N_0 + N_1)Mt_{max})$ plus le temps requis pour construire le graphe initial pour le GGG, tandis qu'il faut un temps $O(DN_0M)$ pour le CHL.

4.5.2 Modèle de mélange généralisé

Le GGG peut être vu comme une généralisation des modèles de mélange classiques, à des points et des segments : un mélange de gaussiennes est un GGG sans arcs. Le GGG fournit une estimation de la densité de probabilité des données plus précise que celle d'un mélange de gaussiennes basé sur le même ensemble de prototypes et la même hypothèse de bruit gaussien isovarié (parce que le GGG ajoute les segments gaussiens à l'ensemble des points gaussiens). Le GGG et surtout son extension aux complexes simpliciaux (non traité ici mais dont le principe reste identique basés sur des k-simplexes génératifs) permettent naturellement de modéliser des densités uniformes par morceaux plus précisément qu'un modèle de mélange de gaussiennes classique. Le GGG fournit aussi intrinsèquement par sa structure génératrice un modèle explicite de la connexité des variétés principales. Par contraste, les autres modèles génératifs ne fournissent aucun indice sur cette connexité, excepté le Generative Topographic Mapping (Bishop et al., 1998) et les courbe principales probabilistes (Tibshirani, 1992). Cependant, dans ces deux cas, la connexité du modèle est contrainte *a priori* et non apprise des données.

4.5.3 Répartition des degrés de liberté

On pourrait envisager un modèle de bruit plus flexible, en individualisant les variances pour chaque prototype, et en considérant la matrice de covariance complète au lieu d'une matrice identité comme c'est le cas pour le GGG. Cependant, nous considérons que les modèles de mélange gaussiens classiques donnent trop de flexibilité au modèle au mauvais endroit (sauf dans le cas où l'on a une connaissance précise du processus de génération des données qui justifie cette répartition des degrés de liberté). En effet, ils tentent d'expliquer la topologie non triviale des variétés principales par une structure trop simpliste (un ensemble de points génératifs) et un modèle de bruit trop complexe (des gaussiennes ellipsoïdales indépendantes les unes des autres). Au contraire, on reporte dans le GGG les degrés de liberté sur le modèle structurel plutôt que sur le modèle de bruit : une structure complexe (un graphe voire un complexe simplicial) et un bruit simple (isovarié). Donc si l'on complexifiait le modèle GGG en libérant les variances de chaque prototypes (et en interpolant linéairement ces variances le long des arcs par exemple), ou même en supposant des densités non uniformes (linéaires par exemple) le long des arc, on aurait un modèle plus flexible et probablement identifiable, mais il faudrait un nombre de données beaucoup plus important pour estimer les paramètres et surtout éviter que deux modèles radicalement différents en termes de topologie (structure complexe et

bruit simple, ou structure simple et bruit complexe) soient tout aussi vraisemblables (au sens de BIC) l'un que l'autre.

4.5.4 Topologie, identifiabilité et critère BIC

Dans ce travail nous faisons explicitement l'hypothèse que la connexité d'un graphe génératif optimal au sens du critère BIC par rapport à des données issues de certaines variétés principales, est proche de la connexité de ces variétés. Nous avons montré que c'était le cas sur quelques exemples, mais il reste à le démontrer théoriquement. Mesurer la "proximité" topologique de deux variétés n'est pas trivial : un sphère percée d'un trou a la topologie d'un disque aussi petit que soit ce trou tant qu'il existe. Donc si l'on considère deux sphères superposées, l'une ayant un trou et l'autre pas, on peut avoir une distance (de Hausdorff par exemple) aussi petite que l'on veut entre ces deux objets en réduisant la taille du trou sans pour autant rendre identique la topologie de ces deux variétés (elles ne sont pas homéomorphes). Ainsi deux objets géométriquement très proches peuvent avoir une topologie radicalement différente. Cependant, si l'on considère un échantillon fini de telles variétés, on ne connaitra leur topologie qu'au travers de cet échantillon et la topologie trouvée dépendra des hypothèses définissant le modèle. Vu sous l'angle génératif, le problème se pose alors ainsi : étant données deux variétés génératrices qui expliquent tout aussi bien le nuage de points en termes de vraisemblance (proximité géométrique), mais avec une topologie différente, lequel choisir ?

Le principe du rasoir d'Occam nous incite à choisir le plus parcimonieux, celui ayant le moins de paramètres. Le critère BIC est un critère qui pénalise la complexité et donc qui joue le rôle du rasoir d'Occam. Cependant nous devons démontrer qu'il ne peut y avoir deux modèles GGG ayant la même vraisemblance et la même complexité mais avec une topologie différente, ou au moins que cela est peu probable. Ce problème est lié à l'identifiabilité du GGG et nous pensons que la clé se situe dans le pouvoir explicatif d'un segment gaussien : un segment gaussien à le pouvoir explicatif d'un mélange d'une infinité de gaussiennes uniformément réparties le long du segment, donc à vraisemblance égale, la complexité d'un mélange de K gaussiennes réparties uniformément sur un segment $(KD + 1)$ est toujours plus grande que celle d'un segment gaussien $(2D + 1)$. Il est donc toujours plus avantageux du point de vue de la complexité et donc du critère BIC d'expliquer localement des données issues d'une variété linéique uniforme perturbée par un bruit gaussien, avec un segment gaussien (même topologie) qu'avec un mélange de K points gaussiens (K variétés de dimension 0). De même, expliquer ces données avec un d-simplexe gaussien ($d > 1$) dont les sommets seraient alignés (donc une variété élémentaire dont la dimension est trop grande par rapport à celle de la variété principale) augmenterait la complexité du modèle ($(d + 1)D + 1$) sans améliorer la vraisemblance, donc dégraderait aussi le score BIC. BIC serait donc un bon critère d'adéquation topologique du modèle GGG aux données. Cette piste reste encore à défricher et formaliser.

5 Conclusion

5.1 Résumé

Nous avons proposé un cadre dans lequel le problème de l'apprentissage de la topologie d'un nuage de points peut être posé comme un problème d'apprentissage statistique. Nous

avons défini un modèle génératif basé sur le graphe de Delaunay de prototypes, permettant d'apprendre la connexité des variétés principales d'un nuage de points. Ce modèle est flexible, parcimonieux, self-consistent, robuste au bruit, ne nécessite pas de réglage manuel arbitraire des méta-paramètres, fournit une mesure objective de qualité par la vraisemblance pénalisée au sens de BIC, et dont la connexité est calculable. Nous avons montré sur des exemples jouets et des données réelles de grande dimension que ce modèle fournit effectivement une bonne estimation de la connexité des variétés principales, avec de meilleurs score que les méthodes de l'état de l'art. Nous l'avons utilisé en classification non supervisé, et montré qu'il était utile pour l'analyse exploratoire de données en fournissant une vue des données plus juste et complémentaire des méthodes de visualisation par projection. Nous avons aussi proposé de l'utiliser pour extraire d'images de caractères manuscrits des caractéristiques topologiques utilisables comme entrées supplémentaires de méthodes de reconnaissance de formes.

5.2 Perspectives

La triangulation induite de Delaunay a été étendue (de Silva et Carlsson, 2004) pour générer des simplexes de dimension supérieure à 1, qui forment un complexe simplicial appelé "witness complex". Une piste à suivre consiste à étendre le modèle de graphe génératif décrit ici, au cas d'un complexe simplicial, ce qui permettrait d'accéder à des caractéristiques topologiques plus riches (nombre de Betti, dimension intrinsèque supérieure à 1, au lieu de la seule arc-connexité).

Nous étudions aussi l'utilisation de ce modèle comme support d'un apprentissage semi-supervisé où la structure des données non étiquetées joue un rôle dans la construction d'un classifieur (Belkin et Niyogi, 2004; Belkin et al., 2006). Nous montrons[17] que la propagation des étiquettes le long des arcs d'un GGG en tenant compte de la densité de ces arcs (propagation d'autant plus forte que la densité est forte) est aussi efficace que les autres approches de l'état de l'art généralement basées sur le graphe des K plus proches voisins, mais ne nécessite aucun réglage arbitraire de méta-paramètres (K par exemple).

Nous envisageons d'étudier l'impact des caractéristiques topologiques additionnelles sur un classifieur en reconnaissance de formes. Nous poursuivons l'étude du lien entre critère BIC et topologie du modèle. Ces résultats seront aussi à comparer à ceux obtenus en Géométrie Algorithmique avec la mesure de "persistance topologique" (Edelsbrunner et al., 2000) et l'inscription de cette mesure dans un cadre statistique (Bubenik et Kim, 2007).

Concernant la famille de graphes dans laquelle la solution est recherchée, nous avons considéré les graphes de Delaunay. Cependant la famille la plus riche est représentée par le graphe complet. Peut-on éviter le choix a priori arbitraire de Delaunay, en partant du graphe complet pour obtenir un graphe représentant la topologie suivant les principes décrits ici ? Ou bien peut-on justifier théoriquement le choix de Delaunay comme bon candidat pour le graphe initial ? Comment pourrait se décliner ce modèle dans un cadre parcimonieux basé sur les données, comme celui des Machines à Vecteurs Supports (SVM), l'émergence des arcs définissant le TRG dans le modèle présenté ici faisant penser à l'émergence des vecteurs supports qui seuls suffisent à définir la frontière de décision dans les SVM.

[17] Soumission en cours

D'un point de vue plus général, ce travail se veut une contribution au rapprochement des domaines de l'Apprentissage Statistique et de la Topologie Algorithmique, à la frontière desquels nous pensons qu'il ouvre de nombreuses perspectives.

Références

Agrell, E. (1993). A method for examining vector quantizer structures. *Proceedings of IEEE International Symposium on Information Theory*, 394–394.

Ahalt, A., A. Krishnamurthy, et D. M. P. Chen (1990). Competitive learning algorithms for vector quantization. *Neural Networks 3*.

Alahakoon, D., S. Halgamuge, et B. Srinivasan (1998). A structure adapting feature map for optimal cluster representation. In S. Usui et T. Omori (Eds.), *ICONIP*, pp. 809–812. IOA Press.

Aupetit, M. (2003). Robust topology representing networks. In *Proceedings of the European Symposium on Artificial Neural Networks*, Bruges (Belgium), pp. 45–50. d-side.

Aupetit, M. (2006). Learning topology with the generative gaussian graph and the em algorithm. In Y. Weiss, B. Schölkopf, et J. Platt (Eds.), *Advances in Neural Information Processing Systems 18*, pp. 83–90. Cambridge, MA : MIT Press.

Aupetit, M. (2007). Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing, Elsevier 70*, 1304–1330.

Aupetit, M. et T. Catz (2005). High-dimensional labeled data analysis with topology representing graphs. *Neurocomputing, Elsevier 63*, 139–169.

Aupetit, M., F. Chazal, G. Gasso, D. Cohen-Steiner, et P. Gaillard (2007). Topology learning : New challenges at the crossing of machine learning, computational geometry and topology.

Barber, C., D. Dobkin, et H. Huhdanpaa (1996). The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software 22*, 469–483.

Belkin, M. et P. Niyogi (2004). Semi-supervised learning on riemannian manifolds. *Journal of Machine Learning Special Issue on Clustering 56*, 209–239.

Belkin, M., P. Niyogi, et V. Sindhwani (2006). Manifold regularization : A geometric framework for learning from examples. *Journal of Machine Learning Research 7*, 2399–2434.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. New York : Oxford Univ. Press.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bishop, C., M. Svensén, et C. Williams (1998). Gtm : the generative topographic mapping. *Neural Computation, MIT Press 10*(1), 215–234.

Boyles, R. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society, Series B 45*, 47–50.

Bubenik, P. et P. Kim (2007). A statistical approach to persistent homology. *Homology, homotopy and Applications 9*(2), 337–362.

Carlsson, G., A. Zomorodian, A. Collins, et L. Guibas (2004). Persistence barcodes for shapes. In *SGP '04 : Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, New York, NY, USA, pp. 124–135. ACM Press.

Carreira-Perpiñán, M. A. et R. S. Zemel (2005). Proximity graphs for clustering and manifold learning. In L. K. Saul, Y. Weiss, et L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, pp. 225–232. Cambridge, MA : MIT Press.

Celeux, G. et G. Govaert (1992). A classification em algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis 14*(3), 315–332.

Chang, K. et J. Ghosh (2001). A unified model for probabilistic principal surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 22 – 41.

Chazal, F., D. Cohen-Steiner, et A. Lieutier (2007). A sampling theory for compact sets in euclidean spaces. *Discrete and Computational Geometry*.

de Silva, V. (2003). Plex : Simplicial complexes in matlab.

de Silva, V. et G. Carlsson (2004). Topological estimation using witness complexes. In M. Alexa et S. Rusinkiewicz (Eds.), *Eurographics Symposium on Point-Based Graphics*, pp. 157–166.

de Silva, V. et J. B. Tenenbaum (2003). Global versus local methods in nonlinear dimensionality reduction. In S. T. S. Becker et K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*, pp. 705–712. Cambridge, MA : MIT Press.

Dempster, A., N. Laird, et D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B 39*(1), 1–38.

Edelsbrunner, H., D. Letscher, et A. Zomorodian (2000). Topological persistence and simplification. *IEEE Symp. on Found. of Comp. Sci.*, 454–463.

Edelsbrunner, H. et N. Shah (1997). Triangulating topological spaces. *International Journal on Computational Geometry and Applications 7*, 365–378.

Fraley, C. et A. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*, 611–631.

Fritzke, B. (1992). Growing cell structures-a self-organizing network in k dimensions. In I. Aleksander et J. Taylor (Eds.), *Artificial Neural Networks*, Volume 2, Amsterdam, Netherlands, pp. 1051–1056. North-Holland.

Fritzke, B. (1995). A growing neural gas network learns topologies. In G. Tesauro, D. Touretzky, et T. Leen (Eds.), *Advances in Neural Information Processing Systems 7*, Cambridge, MA. MIT Press.

Gaillard, P. (2008). Apprentissage de la connexité dŠun nuage de points par modèle génératif. applications à lŠanalyse exploratoire de données et à la classification semi-supervisée. *Université de Technologie de Compiègne - Commissariat à l'Energie Atomique*.

Gaillard, P., M. Aupetit, et G. Govaert (2008). Learning topology of a labeled data set with the supervised generative gaussian graph. *Neurocomputing 71*(7-9), 1283–1299.

Geusebroek, J., G. Burghouts, et A. Smeulders (2005). The Amsterdam library of object images. *International Journal of Computer Vision 61*(1), 103–112.

Hastie, T. et W. Stuetzle (1989). Principal curves. *Journal of the American Statistical Association 84*, 502–516.

Kohonen, T. (2001). *Self-Organizing Maps*. Berlin, Heidelberg, New York : Springer Series in Information Sciences.

Lee, J., A. Lendasse, et M. Verleysen (2002). Curvilinear distance analysis versus isomap. In *Proceedings of the European Symposium on Artificial Neural Networks*, Bruges (Belgium), pp. 185–192. d-side.

Martinetz, T., S. Berkovitch, et K. Schulten (1993). Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks 4*(4), 558–569.

Martinetz, T. et K. Schulten (1994). Topology representing networks. *Neural Networks, Elsevier London 7*, 507–522.

McLachlan, G. et D. Peel (2000). *Finite Mixture Models*. New York : John Wiley & Sons.

Munkres, J. (1993). *Elements of Algebraic Topology*. Westview Press.

Niyogi, P., S. Smale, et S. Weinberger (2006). Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*.

Queen, J. M. (1967). Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathemtical Statistics and Probability*, pp. 281–297.

Roeder, K. et L. Wasserman (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association 92*(439), 894–902.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Tibshirani, R. (1992). Principal curves revisited. *Statistics and Computing 2*, 183–190.

Tipping, M. et C. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B 21*(3), 611–622.

Veltkamp, R. (1991). The gamma-neighborhood graph. *Computational Geometry 1*, 227–246.

Weinberger, K. et L. Saul (2006). Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision 70*(1), 77–90.

Zeller, M., R. Sharma, et K. Schulten (1996). Topology representing network for sensor-based robot motion planning. *World Congress on Neural Networks, INNS Press*, 100–103.

Zhu, X. et J. Lafferty (2005). Harmonic mixtures : combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *ICML '05 : Proceedings of the 22nd International Conference on Machine learning*, New York, USA, pp. 1052–1059. ACM.

Summary

A point set is more than a set of points. Some hidden topological structure may govern the point distribution, and from a data mining perspective, catching this structure is at least as important as estimating the sole spatial point density. In this work, we propose a generative model based on the Delaunay graph of a set of prototypes representative of the point set, assuming a Gaussian noise. We determine the Expectation-Maximization equations and we use the Bayesian Information Criterion to select the best model. Moreover, it does not need any hand-tuning of the meta-parameters. Empirical experiments on toys and real image data

show that the connectedness of the proposed graph accounts correctly for that of the point set. This model provides a principled way to cluster a point set regarding its connectedness. We also show how it could be used as a pre-processing step in classification of point sets, to complete the point sets with their topological invariants coded as graph structures. At last, this work is an attempt to lay foundation stones towards the construction of a topological model of a point set, grounded on statistical generative models.