

Le Graphe Génératif Gaussien

Pierre Gaillard¹, Michaël Aupetit², Gérard Govaert³

¹ CEA, DAM, DIF, F-91297 Arpaçon, France

² CEA, LIST, F-91191 Gif-sur-Yvette cedex, France

³ UTC, U.M.R. C.N.R.S. 6599 Heudiasyc, F-60205 Compiègne Cedex, France

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 un algorithme pour la maximisation de la vraisemblance des paramètres, et nous utilisons le critère BIC pour sélectionner la complexité du modèle. 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.

Keywords : connexité, graphe de Delaunay, modèle génératif, algorithme EM, critère BIC

1 Introduction

Dans les problèmes d'apprentissage statistique, on suppose que les données sont générées par une densité de probabilité $P : \mathbb{R}^D \rightarrow \mathbb{R}^+$. Cependant, le processus sous-jacent de génération des données défini par la fonction P , possède dans de nombreux cas d'intérêt moins de degrés de liberté que l'espace d'observation de dimension D . La formalisation de cette intuition est de supposer que les données sont sûres ou proches d'un ensemble de variétés, appelées *variétés principales* [1], chacune ayant une dimension intrinsèque inférieure à la dimension de l'espace d'observation.

Etant donné un ensemble x de M points observés, dans un espace 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 (modèles de mélange [2], méthodes à noyau [3]). Bien que la fonction densité de probabilité contienne la totalité de l'information extractible de la population dont le nuage de points est un échantillon, celle-ci ne rend pas explicite l'information géométrique et topologique relatives aux variétés principales. Pourtant, si l'on suppose qu'une structure existe dans les données, l'extraire et la caractériser à partir de la densité sont aussi importants que d'estimer la densité de probabilité elle-même. Par exemple, dans le contexte d'un problème de classification, la connexité de cette structure semble être le moyen naturel pour définir des groupes homogènes. L'intérêt d'utiliser cette structure sous-jacente qui gouverne la distribution des données est majeur, puisque celle-ci peut être aussi utilisée pour analyser [4], visualiser [5], discriminer les données [6].

De manière générale, on pourrait extraire des caractéristiques *géométriques* de cette structure telles que la position relative de ses différentes parties, mais aussi des caractéristiques dites *topologiques* telles que la dimension intrinsèque ou la connexité.

L'Apprentissage de la Topologie est un domaine récent en Apprentissage Automatique [7], dont l'objectif est de développer des méthodes basées sur les statistiques pour retrouver les invariants topologiques de ces variétés à partir du nuage de points. La connexité ou la dimension intrinsèque sont de tels invariants topologiques et dans ces travaux nous nous focalisons sur l'extraction de la connexité des variétés principales d'un nuage de points.

Dans la Section 2, nous proposons un bref état de l'art du domaine de l'Apprentissage de la Topologie. Dans la Section 3 et 4, nous présentons respectivement le modèle du Graphe Génératif Gaussien (GGG) et un algorithme pour extraire la connexité de variétés principales. Dans la Section 5 et 6, nous utilisons ce modèle pour analyser un ensemble de données ainsi que pour le débruitage de données.

2 Etat de l'art

Les approches d'Apprentissage de la Topologie sont généralement basées sur la construction d'un espace dont la topologie n'est pas contrainte *a priori* mais au contraire apprise des données, cela au prix de la visualisabilité (possibilité de structures non connexes et de dimensions intrinsèques non homogènes non préservables par projection). Par exemple, Martinetz [8] ou Aupetit [9] 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.

Martinetz et Schulten [8] ont proposé un algorithme de construction d'un graphe appelé Triangulation de Delaunay Induite (TDI). Cet algorithme appelé Competitive Hebbian Learning (CHL), consiste à localiser N_0 sommets $\underline{w} = \{w_n \in \mathbb{R}^D\}_{n=1}^{N_0}$ sur la distribution des données puis à connecter deux sommets w_i et w_j s'il existe une donnée $x \in \underline{x}$ dont ils sont les premier et deuxième plus proches voisins. Une telle donnée est appelé *témoin* de l'arc $\{i, j\}$, et cet arc fait partie du graphe de Delaunay $DG(\underline{w})$ des sommets \underline{w} .

D'un point de vue de l'apprentissage statistique, nous observons que le CHL [8] et la TDI résultante ont certaines limites.

1. **Sensibilité au bruit.** Une donnée est suffisante pour que le CHL crée un lien de la TDI, le rendant ainsi peu robuste au bruit. Un processus de vieillissement des âges a été proposé pour filtrer le bruit [10, 11]. Ce processus est équivalent à supprimer les liens créés par un nombre de données témoins inférieur à un seuil fixé. Ceci peut être vu comme un filtre basé sur la densité de probabilité des données dans la région d'influence des liens. Cependant aucun critère n'a été proposé pour régler le seuil.
2. **Non-consistance du modèle.** Même une infinité de données échantillonnées aussi finement que voulues sur la réalisation géométrique de la TDI ne garantissent pas d'être témoins de tous les liens de la TDI. Ainsi la réalisation géométrique du modèle n'est pas forcément représentable par le modèle lui-même.
3. **Aucune mesure de qualité.** Il n'existe pas de mesures de qualité du graphe obtenu et donc pas de critères permettant de comparer et sélectionner un graphe parmi une collection de graphes. Ceci est problématique lorsque la dimension de l'espace d'observation est supérieure à trois puisque la visualisation est impossible.

3 Le Graphe Génératif Gaussien

Afin de dépasser les limites du Competitive Hebbian Learning (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. A cette fin, nous proposons un modèle de graphe génératif qui combine des approches statistiques et géométriques en définissant un modèle de mélange gaussien construit à partir de la réalisation géométrique d'un graphe.

Le graphe est un moyen efficace de définir *un modèle flexible* permettant de caractériser la connexité de variétés même très compliquées. De plus, l'on sait facilement *extraire la connexité* de la structure discrète d'un graphe. Le modèle de mélange permet quant à lui d'inscrire rigoureusement le modèle dans le cadre de l'apprentissage statistique. En particulier, l'introduction d'une densité de probabilité définit *un modèle de bruit* pour les données et permet l'utilisation de critères statistiques pour *mesurer la qualité* du modèle.

3.1 Hypotheses du modèle

Le modèle que nous proposons repose sur les hypothèses suivantes. Elles sont représentées par la figure 1.

- **Les variétés principales sont inconnues** : nous supposons qu'il existe un graphe G engendré par des points \underline{w} de \mathbb{R}^D qui a la même connexité que les variétés principales.
- **La densité de probabilité le long des variétés est inconnue** : on suppose que la densité est uniforme sur chaque lien de la réalisation géométrique du graphe G .
- **La nature du bruit est inconnue** : nous supposons que le bruit κ est défini par une densité gaussienne de moyenne 0 et de variance σ^2 . Ceci a pour conséquence d'inscrire rigoureusement le modèle dans le cadre de l'apprentissage statistique.

3.2 Description des composants du modèle

Etant donné un ensemble de N_0 sommets $\underline{w} = \{w_n \in \mathbb{R}^D\}_{n=1}^{N_0}$ et un graphe $G(\underline{w}, E)$ les connectant, le modèle utilisé pour extraire la connexité des variétés principales est basé sur deux types de composants que l'on appelle le *point-gaussien* et le *segment-gaussien*.

La valeur de la densité pour une donnée $x_i \in \underline{x}$ générée par un *point-gaussien* centré sur un sommet $w_n \in \underline{w}$ et de variance σ^2 est :

$$g^0(x_i|w_n; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} \exp\left(-\frac{(x_i - w_n)^2}{2\sigma^2}\right) \quad (1)$$

Un *segment-gaussien* est défini comme une somme infinie de points-gaussiens uniformément distribués le long d'un segment : c'est l'intégrale d'un point-gaussien le long d'un segment. La valeur de la densité d'une donnée $x_i \in \underline{x}$ générée par un segment-gaussien

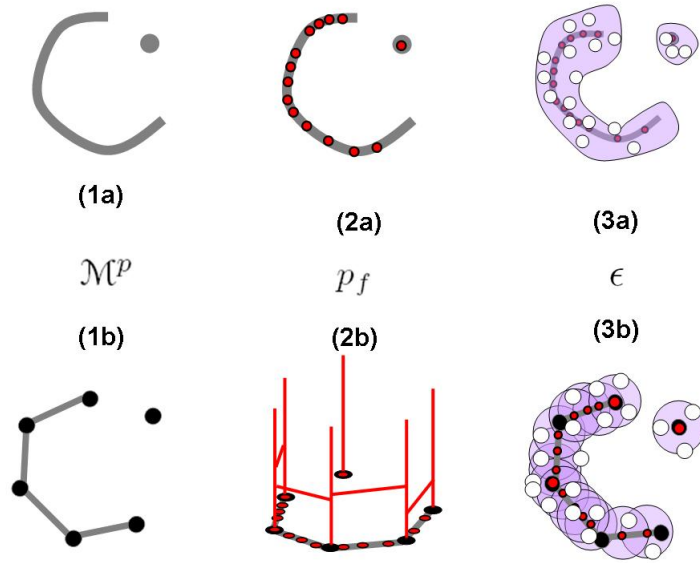


FIG. 1 – Illustration des hypothèses menant au Graphe Génératif Gaussien. Hypothèse 1 : la connexité des variétés principales \mathcal{M}^p est inconnue (1a); on suppose qu’il existe un graphe de même connexité (1b). Hypothèse 2 : la densité p_f sur la variété principale est inconnue; on suppose que la densité p_f est uniforme sur la réalisation géométrique des liens du graphe et qu’elle est définie par un Dirac sur ses sommets. Hypothèse 3 : le bruit ϵ est inconnu (3a); il est modélisé par une densité gaussienne isovariée (3b).

$[w_{a_n} w_{b_n}]$ de longueur non-nulle L_n et de variance σ^2 est :

$$\begin{aligned}
 g^1(x_i | \{w_{a_n}, w_{b_n}\}; \sigma^2) &= \frac{1}{L_n} \int_{w_{a_n}}^{w_{b_n}} g^0(x_i | t; \sigma^2) dt \\
 &= \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}} L_n} \int_{w_{a_n}}^{w_{b_n}} \exp\left(-\frac{(x_i - t)^2}{2\sigma^2}\right) dt
 \end{aligned} \tag{2}$$

où $L_n = \|w_{b_n} - w_{a_n}\|$.

Les fonctions g^0 et g^1 sont positives et on peut prouver que leur intégrale sur \mathbb{R}^D est égale à un, de telle sorte qu’elles définissent toutes deux des densités de probabilité. Des exemples de densités associées à un point-gaussien et un segment-gaussien sont illustrées par la figure 2.

Le calcul de la densité générée par un *segment-gaussien* peut être décomposée en deux parties : une partie qui correspond au bruit gaussien orthogonal au segment passant par le lien et l’autre correspond au bruit gaussien intégré le long du lien. Pour cela, on définit $q_{in} \in \mathbb{R}^D$, la projection orthogonale de la donnée x_i sur la droite passant par les sommets w_{a_n} et w_{b_n} :

$$q_{in} = w_{a_n} + (w_{b_n} - w_{a_n}) \frac{Q_{in}}{L_n} \tag{3}$$

avec $Q_{in} = \frac{(x_i - w_{a_n})^T (w_{b_n} - w_{a_n})}{L_n}$. Le scalaire Q_{in} est la coordonnée du point q_{in} sur un axe dont l’origine est w_{a_n} et de vecteur unitaire : $\frac{w_{b_n} - w_{a_n}}{\|w_{b_n} - w_{a_n}\|}$.

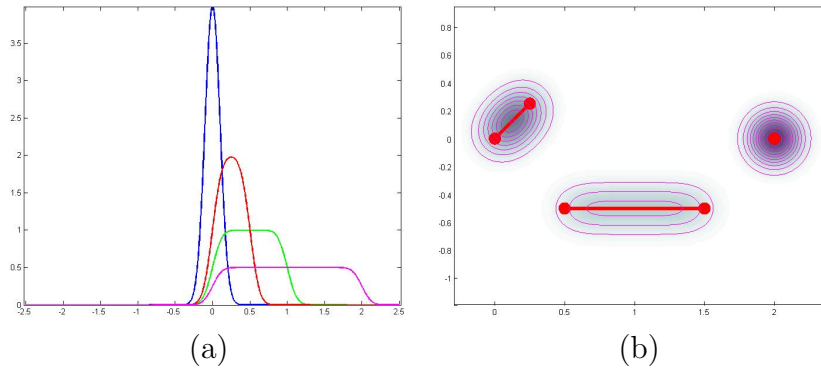


FIG. 2 – **Du point gaussien au segment gaussien** : (a) Densité générée par un segment gaussien unidimensionnel de variance $\sigma^2 = 0.01$ défini sur $[0; \ell]$, avec $\ell = 0$ (bleu), $\ell = 0.5$ (rouge), $\ell = 1$ (vert), $\ell = 2$ (violet). Lorsque la longueur du segment est nulle ($\ell = 0$) la densité d'un segment-gaussien équivaut à celle d'un point-gaussien. (b) Représentation de la densité générée par deux segments-gaussiens et un point-gaussien de variance $\sigma^2 = 0.01$ dans \mathbb{R}^2 .

En utilisant cette projection et le théorème de Pythagore on obtient :

$$\begin{aligned}
 g^1(x_i | \{w_{a_n}, w_{b_n}\}; \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}} L_n} \int_{w_{a_n}}^{w_{b_n}} \exp\left(-\frac{(x_i - q_{in})^2 + (q_{in} - t)^2}{2\sigma^2}\right) dt \\
 &= \frac{g^0(x_i | q_{in}; \sigma^2)}{L_n} \int_0^{L_n} \exp\left(-\frac{(Q_{in} - t)^2}{2\sigma^2}\right) dt \\
 &= g^0(x_i | q_{in}; \sigma^2) \sqrt{\frac{\pi}{2}} \frac{\sigma}{L_n} \cdot \left[\operatorname{erf}\left(\frac{Q_{in}}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{Q_{in} - L_n}{\sigma\sqrt{2}}\right) \right]
 \end{aligned} \tag{4}$$

Ainsi, la densité $g^1(x)$ s'exprime analytiquement à l'aide de la fonction erf , puisque cette dernière admet une expansion en série de Taylor. La décomposition (4) met en évidence une façon de générer des données suivant la densité d'un segment-gaussien $[w_{a_n}, w_{b_n}]$ (2).

- On tire un point $q \in \mathbb{R}^D$ sur le segment $[w_{a_n}, w_{b_n}]$ suivant une loi uniforme : $p(q) = \frac{1}{L_n}$;
- On tire une donnée x suivant une distribution gaussienne centrée sur q et de variance σ^2 .

3.3 Description du modèle de mélange

Etant donné un graphe $G(\underline{w}, E)$, chaque sommet du graphe et chaque lien du graphe sont alors la base d'un modèle génératif. Un point-gaussien est associé à chaque sommet et un segment gaussien est associé à chaque lien du graphe $G(\underline{w}, E)$. Le modèle est donc basé sur les composants élémentaires du graphes, ses sommets et ses liens, qui ont chacun une dimension intrinsèque d différente, valant 0 pour les sommets et 1 pour les liens. Par la suite, on note N_0 le nombre de sommets et N_1 le nombre de liens du graphe.

Soit $z = \{z_n^d \in \{0, 1\} | n = 1, \dots, N_d; d = 0, 1\}$ la donnée manquante qui indique quel composant du modèle a généré la donnée observée x : z_n^d vaut 1 si le n^e composant de dimension d a généré la donnée x et vaut 0 sinon. On définit la densité de probabilité

associée à ces données manquantes comme étant :

$$p(z) = \prod_{d=0}^1 \prod_{n=1}^{N_d} (\pi_n^d)^{z_n^d} \quad (5)$$

π_n^0 (resp. π_n^1) est la probabilité qu'une donnée observée x soit issue du point-gaussien associé au sommet w_n (resp. un segment-gaussien associé au n^e lien du graphe).

Enfin si la donnée manquante z est connue, on tire la donnée observée x suivant la loi de ce composant :

$$p(x|z) = \begin{cases} g_n^0(x; \sigma^2) & \text{si } z_n^0 = 1 \\ g_n^1(x; \sigma^2) & \text{si } z_n^1 = 1 \end{cases} \quad (6)$$

Si l'on cherche à ajuster ce modèle, comme on ne dispose que des données observées, les valeurs de la variable z étant manquantes, l'estimation des paramètres du modèle devra se faire à partir de la densité $p(x)$. En utilisant les équations (5) et (6), on peut définir la densité jointe du modèle $p(x, z) = p(z)p(x|z)$ puis marginaliser par rapport à toutes les valeurs des données manquantes afin d'exprimer la densité $p(x)$.

$$\begin{aligned} p(x; \theta | G(\underline{w}, E)) &= \sum_{d=0}^1 \sum_{n=1}^{N_d} p(x, z_n^d; \theta) \\ &= \sum_{d=0}^1 \sum_{n=1}^{N_d} p(z_n^d) p(x|z_n^d; \theta) \\ &= \sum_{d=0}^1 \sum_{n=1}^{N_d} \pi_n^d g_n^d(x; \sigma^2) \end{aligned} \quad (7)$$

où θ dénote l'ensemble des paramètres du modèle.

La densité des données $p(x)$ s'exprime donc comme un modèle de mélange qui est définie comme la somme pondérée de N_0 points-gaussiens et N_1 segments-gaussiens. Ainsi, les proportions $\underline{\pi}$ vérifient naturellement les deux contraintes suivantes :

$$\sum_{d=0}^1 \sum_{n=1}^{N_d} \pi_n^d = 1 \quad \text{et} \quad 0 \leq \pi_n^d \leq 1 \quad \forall n, d \quad (8)$$

On appelle ce modèle le *Grappe Génératif Gaussien* (GGG).

Cette interprétation du modèle de mélange consiste à considérer que connaissant la position des sommets \underline{w} , le graphe $G(\underline{w}, E)$, les proportions $\underline{\pi}$ et la variance du bruit σ^2 , les données observées sont générées suivant un mécanisme à deux étapes.

1. On tire un composant du mélange (un sommet ou un lien) suivant la distribution donnée par l'équation (5).
2. On tire une donnée observée x suivant la loi du composant (6). Comme on l'a vu, des données issues d'un segment gaussien peuvent être générée à l'aide d'une donnée manquante q uniformément distribuée le long du segment.

4 Caractériser la connexité

Ayant introduit un modèle génératif, la question centrale demeure : comment déterminer le graphe G final modélisant la connexité des variétés principales ? Pour cela, nous proposons l'algorithme 1 dont l'idée principale est double :

- définir un modèle GGG sur $G(\underline{w}, E^+)$ tel que $G(\underline{w}, E) \subseteq G(\underline{w}, E^+)$;
- élaguer les liens de $G(\underline{w}, E^+)$ qui n'explique pas la connexité des variétés principales pour en déduire $G(\underline{w}, E)$.

Algorithme 1 (Principe)

Entrées : \underline{x}, N_0

Initialiser la position des sommets : \underline{w}

Initialiser le graphe génératif : construire un *sur-graphe* $G(\underline{w}, E^+)$ et fixer les paramètres $\underline{\pi}$ et σ^2 .

Apprendre les paramètres du graphe génératif : $\underline{\pi}, \sigma^2$ et \underline{w}

Élaguer le graphe génératif : supprimer les composants associés à une pondération négligeable, $\pi_n^d \leq \gamma$, où $\gamma \in \mathbb{R}^+$ est le seuil d'élagage.

Sortie : $G(\underline{w}, E)$.

Dans cet algorithme, on peut différencier 3 problèmes qui seront traitées dans les paragraphes suivants : (1) *l'initialisation* (Comment positionner les sommets ? Quel sur-graphe choisir ?), (2) *l'apprentissage des paramètres du modèle GGG* et (3) *la sélection de modèle* (Combien de sommet ? Comment choisir le seuil d'élagage ?).

4.1 Initialisation

Nous proposons la méthode suivante pour débiter l'algorithme avec un *bon* graphe.

- Nous utilisons un modèle de mélange gaussien sphérique dont la variance est commune à chaque composant pour positionner les sommets \underline{w} .
- Nous initialisons la variance σ^2 du bruit gaussien à la valeur obtenue par le modèle de mélange.
- Nous initialisons les proportions de façon équiprobable : $\pi_n^d = \frac{1}{N_0 + N_1} \forall n, d$.

Après localiser les sommets \underline{w} , il nous faut enfin choisir le graphe $G(\underline{w}, E^+)$. On peut évidemment considérer le cas du graphe complet des sommets \underline{w} , car il est simple à construire et il est le plus à même de contenir la connexité des variétés principales. Cependant, l'état de l'art nous incite à envisager une autre alternative. Le graphe de Delaunay, bien que plus long à construire $O(N_0^3)$ [12], semble être un choix pertinent puisqu'il est composé de moins de liens que le graphe complet sans pour autant supprimer des liens caractérisant la connexité [8]. Ainsi, on peut considérer que le graphe G recherché pour extraire la connexité vérifie : $G(\underline{w}, E) \subseteq G(\underline{w}, E^+) \equiv GD(\underline{w})$.

4.2 Maximisation de la vraisemblance

Etant donné un Graphe Génératif Gaussien (GGG) construit sur $G(\underline{w}, E)$, la fonction $p(x_i; \underline{\pi}, \underline{w}, \sigma)$ est la densité de probabilité au point x_i sachant les paramètres du modèle. Afin de maximiser la vraisemblance par rapport aux paramètres $\theta = (\underline{\pi}, \sigma^2, \underline{w})$, nous utilisons le cadre de l'algorithme EM [13]. Si l'on peut démontrer que l'étape de maximisation effectuée lors de l'algorithme EM est analytique pour les proportions $\underline{\pi}$ et la variance du bruit σ^2 , celle impliquant les sommets \underline{w} n'est pas directe. Nous proposons donc une étape M *approchée* pour modifier leur position, et on observe empiriquement que la règle de mise à jour *approchée* augmente la plupart du temps la vraisemblance. Si

ce n'est pas le cas, la mise à jour n'est pas effectuée et la position du sommet n'est pas modifiée. Les équations de mise à jour des paramètres sont les suivantes :

$$\begin{aligned}
 \pi_j^{d[\text{new}]} &= \frac{1}{M} \sum_{i=1}^M \tilde{z}_{ij}^d && \text{[Etape M exacte]} \\
 \sigma^{2[\text{new}]} &= \frac{1}{DM} \sum_{i=1}^M \left[\sum_{j=1}^{N_0} \tilde{z}_{ij}^0 (x_j - w_i)^2 \right. \\
 &\quad \left. + \sum_{j=1}^{N_1} \tilde{z}_{ij}^1 \frac{g^0(x_i | q_j^i; \sigma) (I_1 [(x_i - q_j^i)^2 + \sigma^2] + I_2)}{L_j \cdot g_j^1(x_i, \sigma)} \right] && \text{[Etape M exacte]} \quad (9) \\
 w_n^{[\text{new}]} &= \frac{\sum_{i=1}^M [\tilde{z}_{in}^0 x_i + \sum_{j \in E_n} \tilde{z}_{ij}^1 \frac{g^0(x_i | q_j^i; \sigma)}{L_j \cdot g_j^1(x_i, \sigma)} (-E_2 w_{b_j} + E_3 x_i)]}{\sum_{i=1}^M [\tilde{z}_{in}^0 + \sum_{j \in E_n} \tilde{z}_{ij}^1 E_1]} && \text{[Etape M approchée]}
 \end{aligned}$$

où $\tilde{z}_{ij}^d = p(d, j | x_i) = \frac{\pi_j^d g_j^d(x_i; \sigma^2)}{\sum_{d=0}^1 \sum_{j=1}^{N_d} \pi_j^d g_j^d(x_i; \sigma^2)}$, où E_n représente l'ensemble des arcs $[w_{a_j}, w_{b_j}]$ ayant $w_n = w_{a_j}$ comme extrémité, et où

$$\begin{aligned}
 I_1 &= \sigma \sqrt{\frac{\pi}{2}} \left(\text{erf}\left(\frac{Q_j^i}{\sigma \sqrt{2}}\right) - \text{erf}\left(\frac{Q_j^i - L_j}{\sigma \sqrt{2}}\right) \right) \\
 I_2 &= \sigma^2 \left((Q_j^i - L_j) \exp\left(-\frac{(Q_j^i - L_j)^2}{2\sigma^2}\right) - Q_j^i \exp\left(-\frac{(Q_j^i)^2}{2\sigma^2}\right) \right) \\
 E_1 &= \frac{\sigma^2}{L_j^2} \left[e^{-\frac{(Q_j)^2}{2\sigma^2}} (Q_j - 2L_j) - e^{-\frac{(Q_j - L_j)^2}{2\sigma^2}} (Q_j - L_j) \right] + \frac{1}{L_j^2} ((L_j - Q_j)^2 + \sigma) I_1 \quad (10) \\
 E_2 &= \frac{\sigma^2}{L_j^2} \left[e^{-\frac{(Q_j - L_j)^2}{2\sigma^2}} Q_j - e^{-\frac{(Q_j)^2}{2\sigma^2}} (Q_j - L_j) \right] - \frac{1}{L_j^2} (Q_j^2 - L_j Q_j + \sigma^2) I_1 \\
 E_3 &= \frac{1}{L_j} \left[e^{-\frac{(Q_j - L_j)^2}{\sigma^2}} - e^{-\frac{Q_j^2}{\sigma^2}} + (Q_j - L_j) I_1 \right]
 \end{aligned}$$

4.3 Sélection de modèle

En apprentissage statistique, sélectionner un modèle parcimonieux parmi une collection de modèle est un thème récurrent. En particulier, il est connu qu'un modèle génératif ne doit pas être uniquement évalué en fonction de sa vraisemblance mais aussi en terme de complexité. Dans notre cas, il est clair que les paramètres N_0 and γ sont liés à la complexité du graphe génératif, de telle sorte que nous avons à faire face à un problème de sélection de modèle. Dans ce contexte, de nombreux critères et approches ont été proposés, comme par exemple le critère BIC [14]. Ce critère réalise un compromis entre la vraisemblance et la complexité d'un modèle et retient le modèle \mathbf{M} qui maximise :

$$BIC(\mathbf{M}) = \mathcal{L}(\underline{x} | \hat{\theta}) - \frac{\nu}{2} \log(M) \quad (11)$$

où M est le nombre total de données observées, ν est le nombre de paramètres libres du modèles, \mathcal{L} la log-vraisemblance des paramètres θ qui sont à leur maximum de vraisemblance $\hat{\theta}$.

Nous proposons de diviser le problème de sélection de modèle en deux sous-problèmes. Le premier, est la détermination du seuil délagage γ lorsque N_0 est connu et le second est la sélection du nombre de sommets N_0 .

Soit un graphe génératif gaussien à N_0 sommets construit sur $G^+(\underline{w}, E^+)$ et soit $G_\gamma(\underline{w}, E^+ | \pi_j^d \geq \gamma)$ le graphe génératif gaussien qui ne contient que les composants génératifs dont la pondération est supérieure à $\gamma : \pi_j^d \geq \gamma$. En faisant varier le paramètre γ de 1

à 0, on obtient une séquence emboîtée de graphes⁴ allant de l'ensemble vide au graphe $G(\underline{w}, E^+)$:

$$G_1 = \emptyset \subseteq \dots \subseteq G_\gamma \subseteq \dots \subseteq G_0 = G(\underline{w}, E^+) \quad (12)$$

Pour comparer les modèles de la séquence à l'aide du critère BIC, les paramètres θ_γ doivent être à leur maximum de vraisemblance. Ceci n'est évidemment pas le cas, puisque ayant élagué le graphe initial $G(\underline{w}, E^+)$, tous les modèles ne vérifient pas : $\sum_{d=0}^1 \sum_{n=1}^N \pi_n^d = 1$. Il faut donc pour chaque G_γ ré-estimer les paramètres θ_γ . Pour des raisons de complexité, nous proposons que seules les proportions $\underline{\pi}$ soient optimisées via l'équation . Dans ce cas, la fonction de vraisemblance est convexe et l'algorithme converge rapidement. Ceci est aussi motivé par le fait que la variance du modèle sélectionné ne devrait pas être très différente de l'estimation obtenue par le modèle GGG construit sur $G(\underline{w}, E^+)$.

Parmi la séquence emboîtée, nous supposons donc que la connexité des variétés principales est représentée par le meilleur modèle au sens du critère BIC.

$$BIC(G_\gamma) = \log \mathcal{L}(\hat{\theta}_\gamma; G_\gamma, \underline{x}) - \frac{\nu_\gamma}{2} \log(M) \quad (13)$$

où ν_γ est le nombre de paramètres libres du modèle génératif associé à G_γ . Le nombre de paramètres libres pour un graphe génératif gaussien ayant N_0 sommets et N_1 liens est :

$$\begin{aligned} \nu &= [N_0 + N_1 - 1] + [N_0 \times D] + 1 \\ &= (N_0 + N_1) + N_0 \times D \end{aligned} \quad (14)$$

Le premier terme correspond aux proportions⁵ $\underline{\pi}$, le second correspond aux coordonnées des N_0 sommets $w_n \in \mathbb{R}^D$ et la dernière constante correspond à la variance $\sigma^2 \in \mathbb{R}^+$.

Pour sélectionner le nombre de sommets N_0 , l'on peut répéter cette procédure pour différentes valeurs de $N_0 \in \{N_0^{[1]}, \dots, N_0^{[k]}, \dots, N_0^{[K]}\}$:

$$G_{(\gamma^*, N_0^{[k]})} \equiv \max_{\gamma} BIC(G_{(\gamma, N_0^{[k]})}) \quad \forall N_0^{[k]} \quad (15)$$

On obtient K modèles, chacun associé à une valeur de critère $BIC(G_{(\gamma^*, N_0^{[k]})})$. Finalement, on choisit celui dont le nombre de sommets $N_0^{[*]}$ mène à la plus grande valeur :

$$G_{(\gamma^*, N_0^{[*]})} \equiv \max_k BIC(G_{(\gamma^*, N_0^{[k]})}) = \max_k \max_{\gamma} BIC(G_{(\gamma, N_0^{[k]})}) \quad (16)$$

4.4 Illustration

Les figures 3 et 4 illustrent l'algorithme proposé avec un ensemble de données simulées : 50 points sont générées par un point de coordonnées $[0.5, 1.5]$ et 200 points sont générées par une spirale d'équation $[t \cos(2\pi t), t \sin(2\pi t)]$ ($t \in [0; 1.1]$), tous les deux corrompus par un bruit gaussien de variance $\sigma^2 = 0.01$.

Pour un nombre fixe de sommets $N_0 = 7$ (figure 3) : (a) Les sommets sont localisés à l'aide d'un modèle de mélange. (b) Le graphe de Delaunay est construit et un graphe

⁴Il peut s'agir d'un objet géométrique pouvant contenir un lien sans ses sommets. Par abus, nous dirons que G_γ est un graphe.

⁵La somme des pondérations étant contrainte à valoir 1, seules $N_0 + N_1 - 1$ pondérations sont indépendantes.

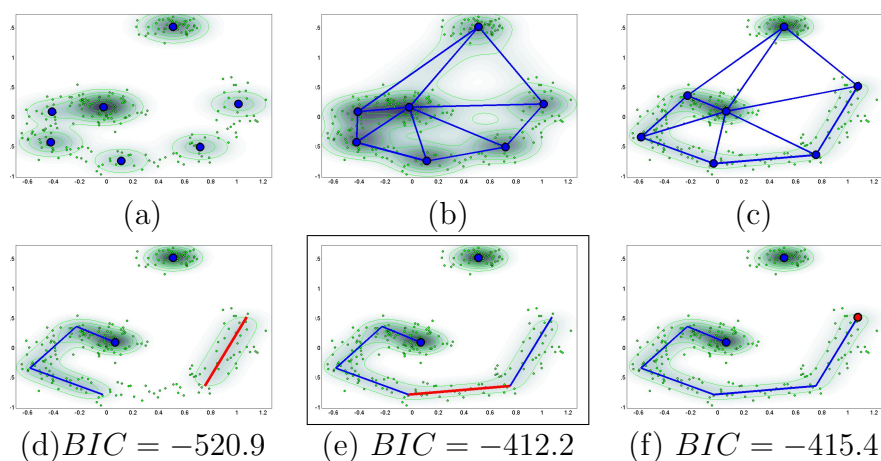


FIG. 3 – Illustration de l’algorithme pour un nombre de sommets fixe $N_0 = 7$.

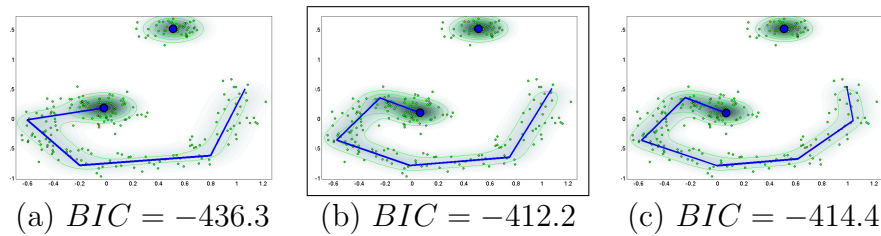


FIG. 4 – Illustration de l’algorithme pour $N_0 = 6, 7, 8$.

génératif est associé à sa réalisation géométrique. (c) La vraisemblance des paramètres est maximisée via un algorithme EM *approché*. L’objectif de l’étape d’élagage est de supprimer automatiquement les composants du modèle génératif inutiles et en particulier, les arcs du graphe traversant un trou de densité. A cette fin, on construit une séquence de modèles génératifs emboîtés en fonction d’un seuil γ . Les figures (d-f) montrent trois modèles consécutifs avec en rouge le composant ajouté par rapport au modèle précédent. Notons que le dernier modèle de cette séquence (G^+) est celui de la figure (c). La vraisemblance de chaque modèle de la séquence est à nouveau optimisée par rapport aux proportions et nous indiquons en dessous de chaque figure la valeur du critère BIC correspondant : le meilleur modèle au sens du critère BIC est celui encadré.

Pour différents nombres de sommets $N_0 = 6, 7, 8$, nous répétons l’algorithme ci-dessus et les modèles correspondants sont représentés dans la figure 4(a-c). Le meilleur modèle au sens du critère BIC est celui encadré. Notons tout de même la connéxité des deux variétés principales est aussi correctement modélisée pour $N_0 = 6, 8$.

5 Connéxité des données *Teapot*

5.1 Données

L’ensemble de données *Teapot* est constitué de 400 images, chacune de taille 101×76 . Les images, qui correspondent aux données observées, représentent une théière photographiée sous différents angles d’un plan [15]. Dix images de cet ensemble sont représentées

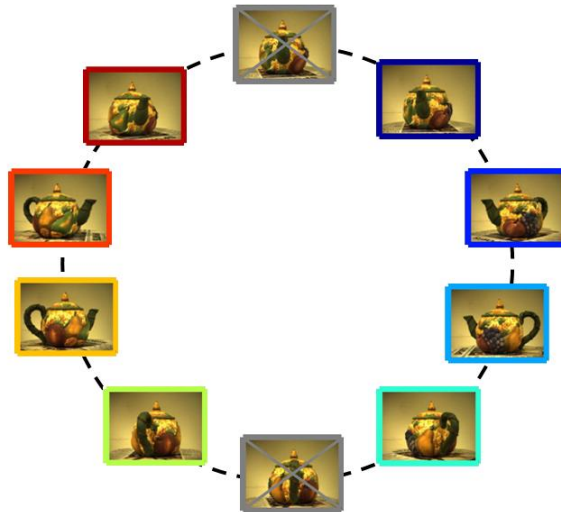


FIG. 5 – *Teapot*. Dix images de l'ensemble Teapot original [15]. Dans l'ensemble utilisé [16], quelques images ont été supprimées. Elles correspondent à celles où l'anse est face à la caméra (images barrées). La couleur des autres cadres code l'angle de rotation de la théière.

par la figure 5.

Malgré la grande dimensionnalité des images ($D = 7676$), le protocole mis en oeuvre pour les générer, laisse penser que dans l'espace d'observation, les données sont proches d'une variété principale ayant la topologie d'un cercle. En effet, les images ne sont paramétrisées que par un seul degré de liberté, l'angle de rotation de la théière.

Dans cette expérience, nous utilisons l'ensemble de données décrit dans [16]. Les images ont été converties en niveau de gris et leur taille a été réduite. De plus, ce nouvel ensemble étant utilisé pour un problème de discrimination où l'objectif est d'identifier si l'anse de la théière est à gauche ou à droite de l'image, les auteurs ont retiré les quelques images où celle-ci est à peu près au centre. A la fin, ils disposent de 365 images ($M = 365$) de taille 16×12 ($D = 192$). De la sorte, les auteurs créent artificiellement deux variétés principales déconnectées, chacune ayant la topologie d'un demi-cercle : l'une correspondant aux positions où la théière a l'anse à droite de l'image et l'autre aux positions où l'anse est à gauche.

5.2 Visualisation

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 ambiant [17]. 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 (ACP), le Generative Topogra-

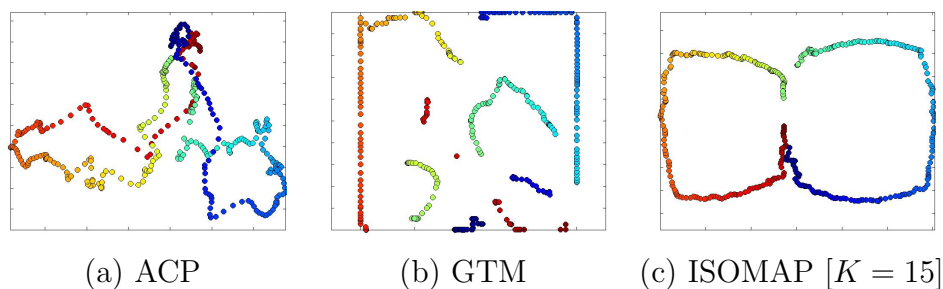


FIG. 6 – **Projections de l'ensemble de données Teapot.** La couleur des données projetées correspond aux couleurs de la figure 5 : elle code l'angle de rotation de la théière. (a) Projection des données en utilisant les deux premiers axes principaux. (b) Projection en utilisant le GTM (c) Projection en utilisant l'algorithme ISOMAP à l'aide d'un graphe des 15. Aucune de ces projections ne permet de montrer l'existence de deux variétés principales déconnectées.

phic Mapping [18] (GTM) et ISOMAP [19] pour visualiser les données Teapot (figure 6) : aucune de ces projections n'est en mesure de montrer la structure déconnectée originelle.

5.3 Extraction de la connexité

Nous optimisons le GGG avec l'algorithme décrit en section 3 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 graphe résultant nous informe sur l'existence de 2 composantes connexes en dépit de ce que montrent les techniques de projection classiques (figure 6). L'analyse des degrés des sommets⁶ du graphe (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 $1.5720e - 010$.

Nous construisons le CHL et le CHL filtré à partir des mêmes sommets. La figure 8 (a) montre que le CHL ne permet pas de retrouver la connexité des variétés principales. En effet, le graphe construit n'a qu'une seule composante connexe. De plus, notons qu'il existe des cycles : le graphe n'est donc pas homéomorphe aux variétés principales ayant généré les données observées. Pour le CHL filtré, on constate avec la figure 8 (b) qu'aucun seuil T n'est convenable pour obtenir un graphe ayant la même connexité que les variétés principales.

Nous répétons cette expérience 10 fois afin d'évaluer la robustesse de l'algorithme. Le tableau 1 présente les résultats pour les différents algorithmes. On remarque que le modèle *GGG* permet, à l'exception d'une fois, de retrouver la topologie attendue. Ceci peut s'expliquer par trois éléments : (1) en utilisant les mêmes données, le résultat de

⁶Dans un graphe, le degré d'un sommet est le nombre d'arcs qui ont ce sommet comme extrémité

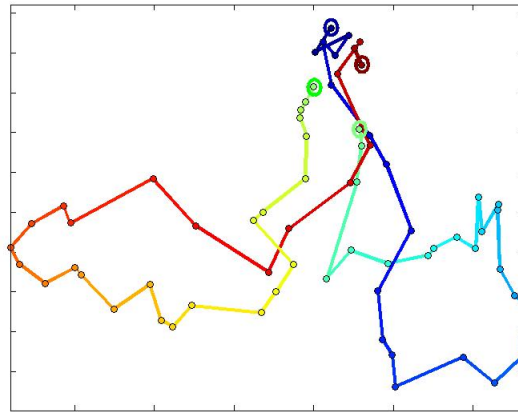
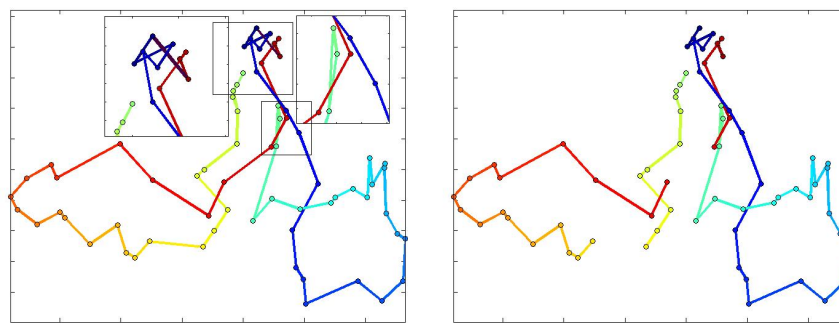


FIG. 7 – **Caractérisation de la connexité de l'ensemble de données *Teapot* avec le GGG.** Projection par ACP du graphe GGG optimal. La couleur des sommets correspond à la couleur de la donnée la plus proche suivant le code défini par la figure 5. Les extrémités des deux composantes connexes définies par le modèle GGG optimal sont encerclées.



(a) CHL [$T = 0$]

(b) CHL [$T = 3$]

FIG. 8 – **Caractérisation de la connexité de l'ensemble de données *Teapot* avec le *CHL*.** Projection par ACP du graphe obtenu par l'algorithme *CHL* (a) et par sa version filtrée (b). La couleur des sommets correspond à la couleur de la donnée la plus proche suivant le code défini par la figure 5. (b) Le seuil $T = 3$ est la première valeur déconnectant les deux variétés principales, celle élaguant le lien violet représenté dans l'agrandissement de gauche de la figure (b). En utilisant un tel seuil, l'une des deux variétés se trouve être morcelée : la connexité est donc perdue.

	GGG	CHL [$T = 0$]	CHL [T^*]
Connexité	100	70	70
Topologie	90	20	20

TAB. 1 – **Modélisation de la connexité des variétés des données *Teapot*.** Le tableau donne en pourcentage, le nombre de fois où les modèles *GGG*, *CHL* et *CHL filtré* permettent de retrouver la connexité (2 composantes connexes) et la topologie (deux chaînes de sommets ayant au plus deux voisins) des données *Teapot*.

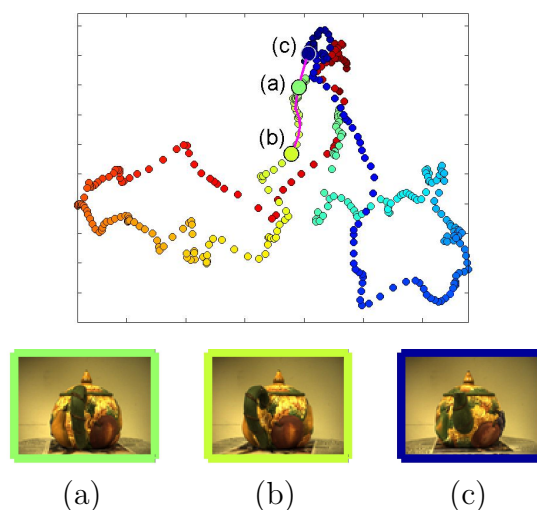


FIG. 9 – **Données *Teapot* et distances.** Trois images de l'ensemble *Teapot* original. Si la distance entre deux images est mesurée par la somme des différences au carré de l'intensité lumineuse entre les pixels de deux images, alors l'image (a) est plus proche de l'image (c) que de l'image (b). Pourtant, l'angle de rotation séparant les positions décrites par les images (a) et (b) est plus faible qu'entre le couple d'images (a) et (c).

l'algorithme dépend uniquement des conditions initiales. Or, l'influence des conditions initiales est minimisée puisque nous utilisons la stratégie *short EM* proposée dans [20] pour déterminer la position initiale des sommets; (2) les données respectent assez bien les hypothèses génératives utilisées par le modèle *GGG*; (3) l'échantillonnage est dense ce qui favorise une estimation fiable du critère BIC.

Le *CHL* permet majoritairement de retrouver la connexité, cependant le graphe obtenu présente généralement des cycles qui faussent la topologie. La version filtrée ne permet pas d'éviter ces cycles.

5.4 Naviguer sur la variété

Le graphe génératif permet aussi de naviguer aisément au travers des données. Supposons que l'on dispose de trois images représentées par la figure 9. On peut par exemple se demander laquelle des images (b) ou (c) vient naturellement après l'image (a). On peut aussi souhaiter savoir s'il existe une série continue d'images passant par ces trois images.

Pour répondre à la première question, il faut définir une distance adéquate entre les

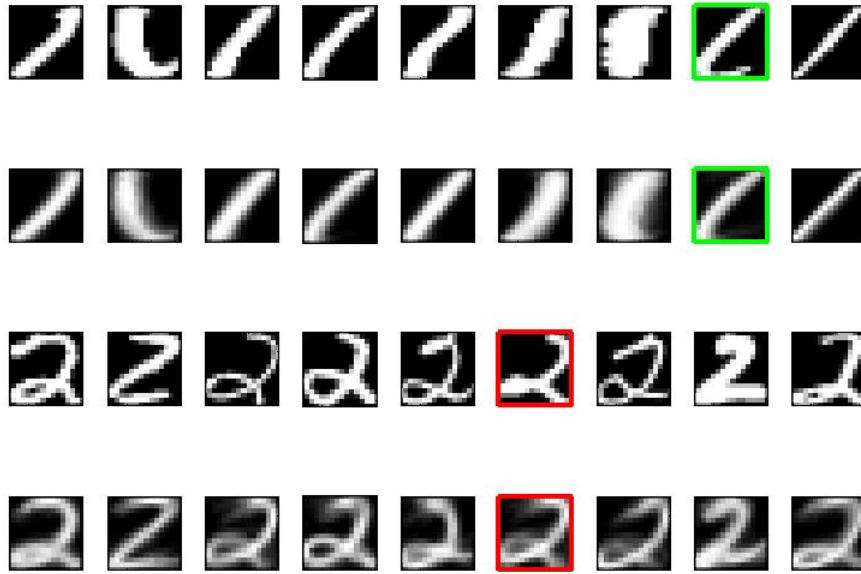


FIG. 10 – Débruitage des chiffres "1" et "2". Première et troisième ligne : les images MNIST originales. Deuxième et quatrième ligne, les images débruitées correspondantes.

images. Si la distance entre deux images est mesurée par la somme des différences au carré de l'intensité lumineuse entre les pixels des deux images, alors l'image (a) est plus proche de l'image (c) que de l'image (b). Mais ceci est contradictoire avec notre perception. En effet, l'angle de rotation séparant les positions décrites par les images (a) et (b) est plus faible qu'entre le couple d'images (a) et (c). Il s'agit donc de mesurer les distances géodésiques le long des variétés, ce qui peut être fait à l'aide du graphe après avoir projeté les données sur le graphe. Pour répondre à la deuxième question, il suffit de savoir si les données se projettent sur une même composante du graphe. Dans l'exemple illustré par la figure 9, la réponse est négative.

6 Débruitage

Dans cette section, nous montrons que l'extraction des variétés peut aussi permettre le débruitage de données. Pour l'illustrer, nous considérons un ensemble d'images représentant les chiffres "1" et "2". On dispose d'un ensemble d'apprentissage de 1100 exemples (550 de chaque classe) et nous souhaitons classifier un ensemble de 1100 chiffres qui forment l'ensemble de test. Nous utilisons l'algorithme de la section 2 pour apprendre les variétés principales de l'ensemble d'apprentissage de manière non-supervisée (sans tenir compte des classes). Le graphe est défini par 50 sommets. Les données de l'ensemble d'apprentissage et de test sont ensuite projetées sur le graphe, et ces projections définissent les données "débruitées". La figure 10 montre quelques images originales et leur version débruitée. Par exemple le "1" original encadré en vert a une "queue" qui disparaît avec le débruitage. De manière similaire, le débruitage a tendance à reformer la boucle du "2" encadré en rouge.

Ensuite, on classe les deux ensembles de test (données originales et débruitées) à l'aide du classifieur des K plus proches voisins (classement par vote majoritaire) par

	Données MNIST	Données débruitées	Prototypes
$K = 3$	98.8	99.0	98.6
$K = 5$	98.6	99.0	98.4
$K = 10$	98.0	99.1	93.4
$K = 15$	97.9	99.1	91.9
$K = 30$	96.8	99.0	83.7

TAB. 2 – **Classification de données MNIST "1" et "2"**. Le tableau donne le pourcentage de taux de bonne classification en utilisant le classifieur des k plus proches voisins pour les différents ensembles.

rapport aux ensembles d'apprentissage respectifs. Le tableau 2 donne le résultat pour différentes valeurs de K . On compare aussi le résultat par rapport à l'algorithme *supervisé* des k -means construit sur les données originales : on représente chaque classe (de l'ensemble d'apprentissage des données originales) par 25 prototypes et on classe les données originales de l'ensemble de test en utilisant ces prototypes (qui sont affectés à une seule classe). Le tableau 2 montre qu'en utilisant les données débruitées, le classifieur est moins sensible au paramètre de voisinage.

7 Conclusion

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, permettant d'apprendre la connexité des variétés principales d'un nuage de points. 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.

Nous avons montré que le modèle était utile pour l'analyse exploratoire de données en fournissant une vue des données complémentaire des méthodes de visualisation par projection. Nous avons aussi montré que le graphe génératif, en modélisant les variétés principales permet le débruitage des données.

Nous étudions désormais 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 [6]. Nous montrons⁷ 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).

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.

⁷Soumission en cours

Références

- [1] R. Tibshirani. Principal curves revisited. *Statistics and Computing*, 2 :183–190, 1992.
- [2] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.
- [3] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3) :1065–1076, 1962.
- [4] P. Gaillard, M. Aupetit, and G. Govaert. Learning topology of a labeled data set with the supervised generative gaussian graph. *Neurocomputing (in press)*, 2008.
- [5] J. Lee, A. Lendasse, N. Donckers, and M. Verleysen. A robust nonlinear projection method. *Eighth European Symposium on Artificial Neural Networks*, 2000.
- [6] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
- [7] M. Aupetit, F. Chazal, G. Gasso, D. Cohen-Steiner, and P. Gaillard. Topology learning : New challenges at the crossing of machine learning, computational geometry and topology, 2007.
- [8] T. Martinetz and K. Schulten. Topology representing networks. *Neural Networks, Elsevier London*, 7 :507–522, 1994.
- [9] M. Aupetit. Robust topology representing networks. In *Proceedings of the European Symposium on Artificial Neural Networks*, pages 45–50, Bruges (Belgium), 2003. d-side.
- [10] T. Martinetz, S. Berkovitch, and K. Schulten. Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4) :558–569, 1993.
- [11] B. Fritzke. A growing neural gas network learns topologies. In G. Tesauero, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, Cambridge, MA, 1995. MIT Press.
- [12] E. Agrell. A method for examining vector quantizer structures. *Proceedings of IEEE International Symposium on Information Theory*, pages 394–394, 1993.
- [13] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1) :1–38, 1977.
- [14] G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6 :461–464, 1978.
- [15] K. Weinberger and L. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1) :77–90, 2006.
- [16] X. Zhu and J. Lafferty. Harmonic mixtures : combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 1052–1059, New York, USA, 2005. ACM.
- [17] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing, Elsevier*, 70 :1304–1330, 2007.

- [18] C. Bishop, M. Svensén, and C. Williams. GTM : the generative topographic mapping. *Neural Computation*, MIT Press, 10(1) :215–234, 1998.
- [19] V. de Silva and J. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 705–712. MIT Press, Cambridge, MA, 2003.
- [20] C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41 :561–575, 2003.