Extraction des nombres de Betti avec un modèle génératif

Maxime Maillot *, Michaël Aupetit*
Gérard Govaert**

* CEA, LIST, F-91191 Gif-sur-Yvette
maxime.maillot@cea.fr, michael.aupetit@cea.fr,
**UTC, Heudiasyc, UMR CNRS 6599, BP 20529, F-60205 Compiègne Cedex
gerard.govaert@utc.fr

Résumé. L'analyse exploratoire de données multidimensionnelles est un problème complexe. Nous proposons d'extraire certains invariants topologiques appelés nombre de Betti, pour synthétiser la topologie de la structure sous-jacente aux données. Nous définissons un modèle génératif basé sur le complexe simplicial de Delaunay dont nous estimons les paramètres par l'optimisation du critère d'information Bayésien (BIC). Ce Complexe Simplicial Génératif nous permet d'extraire les nombres de Betti de données jouets et d'images d'objets en rotation. Comparé à la technique géométrique des Witness Complex, le CSG apparait plus robuste aux données bruitées.

1 Introduction

Une approche récente de l'analyse exploratoire de données multidimensionnelles consiste en l'extraction d'invariants topologiques (Carlsson, 2009). Ces invariants permettent de caractériser la topologie de la population si l'on suppose qu'elle est définie par une collection de sous-variétés de l'espace $\mathbb{R}^D$ dont sont issues les données. Les nombres de Betti sont de tels invariants : le premier encode le nombre de composantes connexes, le second le nombre de cycles indépendants, le troisième le nombre de cavités indépendantes, etc... Ces nombres expriment numériquement des caractèristiques topologiques que nous extrayons visuellement lors de l'analyse de nuages de points présentés dans un repère cartésien à deux dimensions. Ces invariants topologiques le sont par homotopie une classe très large de transformations non linéaires incluant les homéomorphismes, les similitudes et les isométries, on s'attend donc à ce que l'informaiton topologique soit plus robuste à la chaîne de mesure que l'information de nature géométrique.

Il existe des techniques d'extraction des nombres de Betti à partir d'un complexe simplicial (de Silva). Un complexe simplicial est un ensemble de simplexes tels que l'intersection de deux simplexes du complexe est soit vide soit fait aussi partie du complexe. Le plongement dans $\mathbb{R}^D$ d'un k-simplexe est l'enveloppe convexe de $(k+1)$-points de cet espace. Un complexe simplicial plongé est donc une collection de variétés qui peut servir de modèle topologique et géométrique à la population génératrice des données. Nous devons lui associer un modèle de densité de probabilité afin de modéliser complètement le processus de génération des données.

Les cartes de Kohonen (Kohonen, 1989) ou leur pendant génératif que sont les Generative Topographic Map (GTM) (Bishop et al., 1998) ainsi que les Generative Principal Manifolds (Tibshirani, 1992) sont des modèles topologique dont la dimension intrinsèque doit être fixée *a priori*. Les modèles de mélange de gaussiennes classiques (McLachlan et Peel, 2000) permettent de modéliser la densité du nuage de points, mais n'encodent aucune information topologique. Nos précédents travaux sur le Graphe Génératif Gaussien (Aupetit, 2005), que l'on peut voir comme une version générative des Topology Representing Network (TRN) de (Martinetz et Schulten, 1994), sont limités à la modélsiation de la population sous forme d'un graphe, qui ne peut encoder que la connexité et non les invariants d'ordre supérieur (cycles, cavités...). Enfin l'approche des Witness Complex (de Silva et Carlsson, 2004) étend les TRN aux complexes simpliciaux. Cette technique est essentiellement géometrique, très sensible au bruit et non paramétrique. Nous proposons un modèle génératif obtenu par la convolution d'une loi normale multivariée au complexe simplicial plongé.

2 Hypothèses générales et description du modèle

Les données sont un nuage de points dans $\mathbb{R}^D$, supposées issues d'une collection de variétés, perturbées par un bruit centré gaussien isovarié, dont la variance σ^2 est inconnue. Nous supposons que cette collection de variétés peut être approchée par une complexe simplicial de Delaunay dont les sommets $\underline{w} = (w_1, ..., w_N)$ appartiennent à $\mathbb{R}^D$. Nous définissons le Complexe Simplicial Génératif (CSG) comme la convolution d'un bruit gaussien isovarié à un complexe simplicial de Delaunay de sommets $\underline{w}$. Les composants de ce modèle de mélange sont les différents simplexes du complexe : les sommets, les arêtes, les triangles, les tétraèdres... Les paramètres de ce modèle sont la position des sommets du complexe, la proportion des composants du mélange, et la variance du bruit gaussien. Nous utilisons l'algorithme EM pour trouver le jeu de paramètres qui maximise la vraisemblance, et nous utilisons le critère BIC (Schwarz, 1978) pour sélectionner le modèle de complexité optimale. La complexité du modèle est déterminée par son nombre de sommets, d'arêtes, de triangles... de simplexes que l'on conserve. Optimiser BIC revient donc à retirer certains simplexes du modèle, et donc à sculpter le complexe de Delaunay pour obtenir un sous-complexe de Delaunay BIC-optimal, dont on suppose que la topologie modélise ainsi au mieux celle de la population. Nous détaillons le modèle CSG et l'algorithme d'apprentissage.

2.1 Le Simplexe Génératif (SG)

Un simplexe génératif (SG) est une fonciton densité de probabilité basée sur un simplexe plongé. Soit S_n^d un simplexe de dimension d, $|S_n^d|$ son volume, g^0 une distribution gaussienne isovariée de variance σ^2 et de dimension D où D est la dimension de l'espace ambiant, et g_n^d la fonction densité de probabilité induite par le simplexe génératif associé à S_n^d :

$$g_n^d(x) = \frac{1}{|S_n^d|} \int_{S_n^d} g^0(x|t, \sigma^2) dt$$

Ceci peut être vu comme la convolution d'une densité gaussienne multidimensionnelle et d'un simplexe. Pour former le modèle de mélange nous remplaçons les composants gaussiens

classiques, sources ponctuelles de données perturbées par un bruit gaussien, par des sources simpliciales. Le modèle de mélange associé est le Complexe Simplicial Génératif.

2.2 Le Complexe Simplicial Génératif (CSG)

Un Complexe Simplicial Génératif est un mélange de simplexes génératifs $\underline{S} = \{S_i^d\}$ - de la même façon qu'un complexe simplicial est défini par un ensemble de simplexes - une variance σ^2, identique pour chaque simplexe, et un ensemble de proportions $\{\pi_i^d\}$ qui correspondent aux composants $\{S_i^d\}$. Les $\{\pi_i^d\}$ ont pour contrainte $\sum_{d=0}^{D}\sum_{i=1}^{N_d}\pi_i^d = 1$ où N_d est le nombre de simplexes de dimension d. Un CSG a pour densité de probabilité :

$$p(x|\sigma,\underline{S},\underline{w}) = \sum_{d=0}^{D}\sum_{i=1}^{N_d}\pi_i^d g_i^d(x|\sigma,\underline{w})$$

3 Le processus d'apprentissage

1. **Initialisation :** Il faut premièrement positionner les 0-simplexes $\underline{w}$. Un modèle de mélange gaussien classique (McLachlan et Peel, 2000) optimisant la vraisemblance est utilisé à cet étape. Le nombre de sommets N_0 est choisi grâce au critère BIC(Schwarz, 1978). Puis un graphe de Delaunay est construit à partir de ces sommets. Cet ensemble de sommets et arêtes est appelé S^1. Nous avons alors un CSG constitué de sommets et d'arêtes génératifs qui sont les composantes d'un modèle de mélange. Les proportions π_i^d sont tous égaux à l'initialisation. La variance initiale est celle apprise par le modèle de mélange gaussien classique ;
2. **Optimisation des proportions et de la variance :** L'algorithme EM est utilisé pour optimiser les proportions $\underline{\pi}$ ainsi que la variance σ^2. La mise à jour de la variance est coûteuse en temps - elle requiert une méthode de quasi-Monte Carlo (Morokoff et Caflisch, 1995) pour estimer le terme $g_i^d(x|\sigma,\underline{w})$ - elle n'est donc réalisée que toutes les 200 itérations. Une fois les nouvelles proportions $\underline{\pi}^*$ obtenues, nous cherchons à retirer du modèle les sommets ou arêtes qui ne sont pas source de données, donc ceux dont la proportion π_i^{d*} est faible ;
3. **Sélection des composants :** Les composants qui doivent être retirés sont choisis grâce au critère BIC. Les proportions $\underline{\pi}$ sont rangées par ordre croissant, et BIC est calculé pour chaque n ($n \leq N$) en conservant les n composants de plus forte proportion. L'ensemble C^1 de sommets et arêtes maximisant BIC est conservé pour l'étape suivante. Certains sommets sont retirés s'ils font partie d'une arête conservée dans C^1, puisque leur retrait ne modifie pas la topologie du complexe, et permet de réduire le nombre de paramètres du modèle de mélange ;
4. **Ajout de la dimension suivante :** Nous considérons maintenant les triangles du complexe de Delaunay. Un triangle est ajouté au modèle de mélange et à l'ensemble S^2, si toutes ses arêtes sont dans C^1. On répète ensuite les étapes 2, 3 et 4, en incrémentant d, puis en retirant de l'ensemble C^d les simplexes inutiles grâce au critère BIC. L'algorithme s'arête quand $d = D$ ou quand S^{d+1} est vide ;

5. **Extraction des nombres de Betti :** Le complexe simplicial C^d obtenu quand l'algorithme se termine est ensuite utilisé comme entrée dans le logiciel Plex (de Silva) qui fournit les nombres de Betti recherchés.

4 Expériences

4.1 Variétés topologiques connues : le tore et la sphère

Nous testons la technique sur des variétés plongée dans $\mathbb{R}^3$ dont la topologie est connue : une sphère de rayon 1 et un tore de petit rayon 3 et de grand rayon 10. Les nombres de Betti attendus pour la sphère sont $(1, 0, 1, 0...)$, et pour le tore $(1, 2, 1, 0...)$. Le CSG est comparé au Witness Complex (WitC) tel qu'implémenté dans le code Javaplex (Adams et Tausz). Le nombre de prototypes (30 pour la sphère, 40 pour le tore) et leur position sont identiques pour le CSG et le WitC, ils sont obtenus par un modèle de mélange gaussien classique maximisant le critère BIC. La méthode *infiniteBarcodes* de Javaplex retourne les nombres de Betti recherchés à partir d'une filtration du complexe simplicial.

Pour la sphère, 1000 points sont tirés aléatoirement avec un bruit gaussien centré d'écart-type $\sigma \in \{0.05, 0.1, 0.2\}$. Pour le tore, 2000 points sont tirés aléatoirement avec un bruit gaussien centré d'écart-type $\sigma \in \{0.01, 0.05, 0.1\}$. Dans chaque cas, CSG et WitC sont relancés 100 fois à partir d'une initialisation aléatoire de leurs paramètres. Une réponse est considérée comme correcte uniquement si les nombres de Betti attendus sont trouvés. Le résultat calculé est le pourcenatge de réponses correctes sur les 100 essais.

4.2 Base d'images : COIL-100

COIL-100 (Nene et al., 1996) est un corpus d'images de 100 objets différents photographiés en rotation. Chaque image est prise après que l'objet a été tourné de 5 degrés autour d'un axe vertical, il y a donc 72 images par objet. Les images couleurs sont transformées en niveaux de gris et leur taille est réduite. Un objet est représenté par un nuage de 72 points dans l'espace de dimension le nombre de pixels des images. Afin de pouvoir calculer le complexe de Delaunay, nous projetons les données sur les 5 premières composantes principales du nuage de points. Du fait de la rotation complète de l'objet, ce nuage de points à la topologie d'un anneau dont les nombres de Betti sont $(1, 1, 0...)$. Nous testons le CSG sur 5 objets (*oignon*, *flacon*, *tomate*, *boîte*, *chat décoratif*), et nous le lançons 10 fois pour chacun.

5 Resultats

5.1 La sphère et le tore

On peut trouver les résultats obtenus sur la sphère et le tore dans les tableaux 1 et 2. Pour les variances les plus faibles, WitC domine le CSG, mais ils sont tous les deux fiables. Pour $\sigma = 0.2$, les performances de CSG diminuent alors que celles de WitC sont stables. Ici, la filtration avantage WitC : la cavité à l'intérieur de la sphère persiste. Alors que pour le CSG, si la variance grandit trop, une corde à l'intérieur de la sphère peut apparaître malgré les différentes étapes d'élagage et ajouter des cycles non désirés dans le modèle.

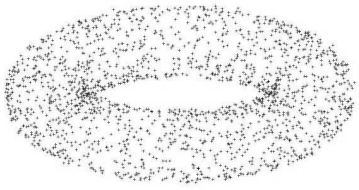

FIG. 1 – *Exemple de 2000 points tirés uniformément à la surface du tore.*

FIG. 2 – *Cinq objets sélectionnés dans la base de données COIL-100.*

Le tore est plus complexe à modéliser que la sphère : si la variance du bruit est trop grande, l'intérieur du tore se remplit. Des cycles et des cavités parasites peuvent apparaître entrainant une baisse de performance que l'on peut observer ici. Alors que les résultats sont toujours acceptables pour le CSG, ceux du WitC ne sont plus suffisants. Les erreurs du WitC se font en général sur le nombre de cycles, alors que le nombre de composantes connexes et de cavités sont corrects.

	WitC	GSC
$\sigma = 0.05$	100%	95%
$\sigma = 0.1$	99%	90%
$\sigma = 0.2$	98%	55%

TAB. 1 – *Sphère.*

	WitC	GSC
$\sigma = 0.01$	5%	63%
$\sigma = 0.05$	8%	60%
$\sigma = 0.1$	9%	57%

TAB. 2 – *Tore.*

5.2 COIL-100

La rotation de la *boîte*, du *flacon* et du *chat décoratif* génère pour chacun de ces objets des images nettement différentes les unes des autres, et crée donc pour chaque objet un nuage de points de topologie annulaire. Le CSG retrouve systématiquement les nombres de Betti attendus dans ce cas $(1, 1, 0...)$. En analysant la structure du complexe simplicial, on observe qu'il est constitué de 11 sommets et 11 arêtes formant un cycle. La dimension maximale des simplexes vaut 1 montrant que la projection sur les 5 premières composantes principales n'a pas fait perdre d'information. Pour les deux autres objets, *oignon* et *tomate*, le CSG optimal contient seulement 5 sommets et tous les simplexes jusqu'à la dimension 5, avec la signature topologique d'une boule $(1, 0, 0...)$ sans cycle ni cavité. Ce résultat peut s'expliquer par le fait que ces deux objets sont plutôt invariants par rotation autour de l'axe vertical, les 72 images associées à chacun sont donc très similaires et forment un nuage de points compact dans l'espace des pixels, sans structure topologique complexe.

6 Conclusion

Contrairement à celle du Witness Complex (WitC), l'efficacité du Complexe Simplicial Génératif (CSG) est bonne à la fois pour la sphère et le tore montrant l'intérêt de l'approche générative pour l'extraction des nombres de Betti d'un nuage de points. Le temps de calcul du WitC est plus rapide que celui du CSG et permet de traiter des données en très grande dimension. Nous travaillons à l'adaptation d'une technique de l'état de l'art pour le calcul du complexe de Delaunay, afin de rendre accessible au GSC des données de grande dimension.

Références

Adams, H. et A. Tausz. Javaplex tutorial. *"http ://javaplex.googlecode.com/"*.

Aupetit, M. (2005). Learning topology with the generative gaussian graph and the EM algorithm. *NIPS*.

Bishop, C. M., M. Svensén, et C. K. I. Williams (1998). GTM : The generative topographic mapping. *Neural Computation 10*(1), 215-234.

Carlsson, G. (2009). Topology and data. *Bull. Amer. Math. Soc.* 46, 255-308.

S. A. Nene, S. K. Nayar and H. Murase, Columbia Object Image Library (COIL-100). Technical Report CUCS-006-96.

de Silva, V. Plex : Simplicial complexes in matlab. *"http ://comptop.stanford.edu/u/programs/plex/"*.

de Silva, V. et G. Carlsson (2004). Topological estimation using witness complexes. *Proc. Eurographics Symp. Point-Based Graphics*, 157-166.

Kohonen, T. (1989). *Self-organization and associative memory*. Springer-Verlag New York.

Martinetz, T. et K. Schulten (1994). Topology representing networks. *Neural Networks 7*(3), 507-522.

McLachlan, G. et D. Peel (2000). *Finite Mixture Models*. Applied probability and statistics. John Wiley & Sons.

Morokoff, W. J. et R. E. Caflisch (1995). Quasi-monte carlo integration. *Journal of Computational Physics*,122, 218-230.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. of Stat.*, 6(2), 461-464.

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

Summary

Exploring multidimensional data is a complex analytic task. We propose a generative model called Generative Simplicial Complex, to extract topological invariants called Betti numbers from the data. The GSC is used to analyze toys data and image data. The GSC appears to be more robust to noise than the Witness Complex, a state of the art geometrical technique to extract Betti numbers of point set data.