Détection des courts-circuits pour la réduction de dimension basée sur les graphes de voisinages

Yves MICHELS, Étienne BAUDRIER

ICube - MIV
300 Bd Sébastien Brandt - CS 10413 - F - 67412 Illkirch Cedex
yves.michels@unistra.fr, baudrier@unistra.fr

Résumé – Le traitement de données en grande dimension requiert généralement une étape de réduction de dimension afin de travailler dans la dimension intrinsèque des données. Lorsque les données sont bruitées, les méthodes de réduction de dimension non linéaires peuvent être induites en erreur par l'apparition de courts-circuits dans le graphe de voisinage. La méthode proposée a pour but de supprimer ces courts-circuits à l'aide d'un graphe parcimonieux qui approxime la structure des données, dont la construction est basée sur la densité estimée des données.

Abstract – Processing high dimensional datasets often makes use of a dimension reduction step. Indeed, high dimension data generally rely on a low dimension underlying structure. When the data are noisy, dimension reduction may fail because of shortcuts appearing on the graph catching the underlying structure. Our paper presents a method to suppress shortcuts in the underlying structure graph, based on a sparse graph that approximates the data structure and that is built using a data probability density estimation.

1 Introduction

Les quantités d'information disponibles pour l'apprentissage statistique, l'estimation de paramètres ou la visualisation de données amènent souvent à travailler dans des espaces de grande dimension. Néanmoins, les données appartiennent généralement à des variétés de faible dimension. Une étape préliminaire de réduction de dimension permet de synthétiser les données pour améliorer l'efficacité des traitements. Lorsque les données appartiennent à un sous-espace affine, il est possible d'utiliser des méthodes de réduction de dimension linéaire comme l'analyse en composantes principales qui projette les données sur le sous espace qui maximise leur variance. Nos travaux s'appliquent au cas plus général où la structure des données n'est pas forcément affine, où il est nécessaire d'utiliser des méthodes de réduction de dimension non linéaires, aussi connues sous le nom d'apprentissage de variété.
L'objectif de l'apprentissage de variété est d'exprimer les données dans un espace de plus faible dimension que l'espace d'origine en conservant leur géométrie. Il existe deux approches générales. Celles qui reposent sur une cartographie des données [7, 2], où une carte, dont la topologie est connue *a priori*, est placée dans l'espace de grande dimension pour rendre compte de la structure intrinsèque des données. La deuxième approche est basée sur la conservation d'information locale, et elle ne demande aucun *a priori* sur les données. L'information locale peut être une similarité, comme pour les cartes de diffusion [1], la structure locale d'espace vectoriel comme c'est le cas pour la méthode Locally Linear Embedding [5] ou Local Tangent Space Alignment [10], ou encore les distances comme pour Isomap [9]. L'information locale est représentée sous la forme d'un graphe de voisinage qui connecte les points qui sont proches dans l'espace de grande dimension. Lorsque les données sont trop peu nombreuses ou bruitées, le graphe de voisinage connecte des points qui sont distants sur la variété. Ces liaisons courts-circuits introduisent un biais important lors de la réduction de dimension.
L'objectif de nos travaux est de construire un graphe de voisinage sans courts-circuits sur des données bruitées issues d'une variété riemannienne de dimension intrinsèque plus faible que celle de l'espace des données.
Plusieurs approches existent dans la littérature, dont les plus simples reposent sur des statistiques locales comme les indices de Jaccard [8]. Pour une arrête donnée, l'indice de Jaccard mesure la similarité des voisinages des sommets. Cukierski et Foran [3] proposent de détecter les courts-circuits à l'aide de la mesure de centralité des arêtes. La centralité d'une arête est le nombre de plus courts chemins qui la contiennent. Par définition, un court-circuit connecte des points éloignés pour la distance géodésique –sur la variété–, la centralité des courts-circuit a donc tendance à être supérieure à celle des arêtes légitimes. Néanmoins, cette approche a un fort taux de fausses détections et est coûteuse en temps de calcul. Glasher et Martinez [6] proposent de supprimer l'ensemble minimal d'arêtes permettant de supprimer tous les cycles atomiques supérieurs à un seuil donné. En effet, les graphes contenant des courts-circuits possèdent des cycles larges. Cependant, cette approche n'est pas adaptée aux variétés fermées qui contiennent des cycles na-

Travaux financés par l'Agence Nationale de la Recherche sur le contrat ANR-14-CE27-0012-01.

turels, et ne peut pas détecter les courts-circuits lorsqu'ils sont trop dispersés pour engendrer des cycles larges.
L'approche proposée repose sur la construction d'un graphe parcimonieux approximant la variété, qui rend compte de la structure globale des données. Le graphe approximant est construit à l'aide d'une estimation de la densité de probabilité des données dans l'espace.
La Section 2 présente la méthode et son cadre d'application, la Section 3 propose une évaluation de la méthode et la Section 4 conclut l'article.

2 Méthode proposée

Soit $l < m$ des entiers positifs. On note $\mathbb{R}^l$ l'espace des paramètres et $\mathbb{R}^m$ l'espace des données. Les topologies sur $\mathbb{R}^l$ et $\mathbb{R}^m$ sont les topologies usuelles munies des distances euclidiennes d_l et d_m. Soit $I \in \mathbb{R}^l$ un compact connexe, et f une fonction injective et continue de I dans $\mathbb{R}^m$. On note $\mathcal{M} = f(I)$ la variété image de I par f.

On note P le sous-ensemble de $\mathbb{R}^m$, image de $\Theta \subset I$, un ensemble fini de cardinal n_p, par la fonction f :

$$P = f(\Theta)\ ,\ \ \Theta = \{\theta_1, ..., \theta_{n_p}\} \subset I\ ,\ \ n_p \in \mathbb{N}^*$$

L'ensemble $\Pi \in (\mathbb{R}^m)^{n_p}$ est une réalisation bruitée de P par un bruit blanc additif Gaussien d'écart-type $\sigma \geq 0$:

$$\Pi = \{\pi_i = P_i + \eta_i\}_{1 \leq i \leq n_p}$$

avec $\eta_i \sim \mathcal{N}(0, \sigma^2)$.
Connaissant Π, on cherche à construire un graphe connectant deux éléments de Π lorsque leurs antécédents sont voisins dans I.
On note $G(\alpha) = (\Pi, A_\alpha, W)$ ce graphe de voisinage, avec α un paramètre de voisinage pouvant être le nombre de plus proches voisins pour les k-plus proches voisins ou la taille du voisinage pour les ϵ-voisinages.

On définit la mesure géodésique entre deux points, x et y de $\mathbb{R}^m$, comme extension de la distance géodésique par :

$$\tilde{d}_{\mathcal{M}}(x,y) = \min\left(d_{\mathcal{M}}(x^*_{\mathcal{M}},\ y^*_{\mathcal{M}}) + d_m(x, x^*_{\mathcal{M}}) + d_m(y, y^*_{\mathcal{M}})\right)\ , \tag{1}$$

où $x^*_{\mathcal{M}} \in \mathrm{argmin}_{x_{\mathcal{M}} \in \mathcal{M}}\left(d_m(x, x_{\mathcal{M}})\right)$,
$y^*_{\mathcal{M}} \in \mathrm{argmin}_{y_{\mathcal{M}} \in \mathcal{M}}\left(d_m(y, y_{\mathcal{M}})\right)$ et
$d_{\mathcal{M}} : \mathcal{M}^2 \to \mathbb{R}^+$ est la distance géodésique sur $\mathcal{M}$. On appelle courts-circuits les arêtes connectant des points éloignés pour la distance géodésique. Nous définissons maintenant le Graphe Parcimonieux Approximant la variété (GPA). Le GPA sera noté $G_{\mathcal{M}} = (S_{\mathcal{M}}, A_{\mathcal{M}}, W_{\mathcal{M}})$. Lorsque le graphe de voisinage repose sur les distances, comme pour les k-plus proches voisins, la présence de bruit sur les données génère des courts-circuits. Les données bruitées peuvent être modélisées comme une réalisation d'une loi de probabilité dont la densité est donnée par la convolution entre la densité de probabilité des points sur la variété non bruitée et d'une gaussienne de moyenne nulle. La détection des courts-circuits est effectuée en trois étapes qui sont le calcul des sommets du GPA, $S_{\mathcal{M}}$, la recherche des arêtes du GPA, $A_{\mathcal{M}}$, et la détection des courts-circuits dans le graphe de voisinages sur les données à l'aide du GPA. Ces trois étapes sont représentées dans la figure 1.

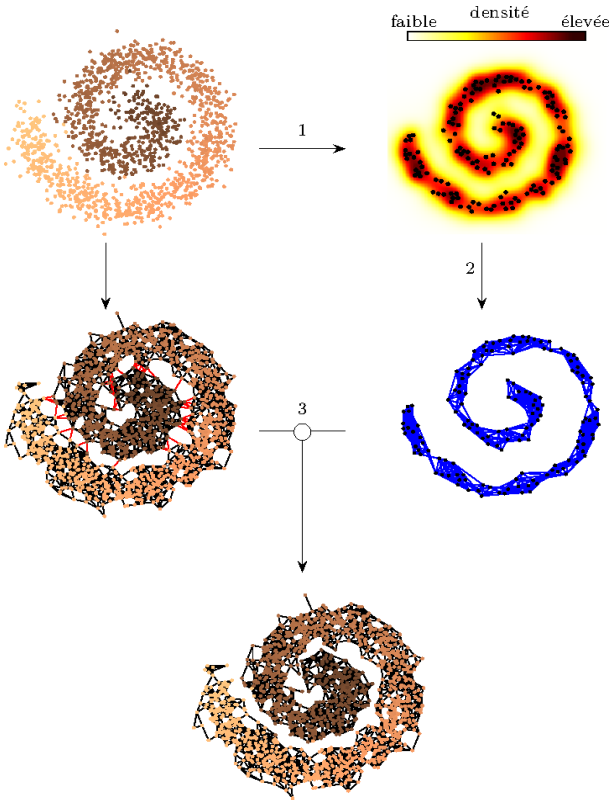


FIGURE 1 – étapes de la détection des courts circuits illustrée sur un ensemble de 1500 points bruités issus d'une spirale 2D.

2.1 Recherche des sommets du GPA

Les sommets du GPA ont pour objectif de représenter la variété avec un nombre $n_m < n_p$ points de $\mathbb{R}^m$. Le nombre de points, n_m contrôle le compromis biais-variance. L'ensemble $S_{\mathcal{M}}$ doit être distribué sur l'ensemble de la variété et chaque point de $S_{\mathcal{M}}$ doit être proche de la variété. L'estimation est effectuée en deux temps. Un premier ensemble, M, de n_m points est obtenu à l'aide de l'algorithme des k-moyennes, puis les sommets du GPA sont calculés en déplaçant les points de M dans des zones de forte densité. Il est assuré, par les k-moyennes, que les sommets sont répartis sur l'ensemble de la variété, cependant, il n'est pas assuré que tous les sommets sont dans des zones de forte densité. Les points dans les zones de faible densité ont une forte probabilité d'être connectés à des points éloignés pour la mesure géodésique. L'obtention des $s_i \in S_{\mathcal{M}}$ se fait par migration des points M_i dans les zones de forte densité par la minimisation de la fonction de coût E_i :

$$\begin{aligned} E_i(x) \ &= \ (1-\lambda_1-\lambda_2)\tfrac{d_m(M_i,x)^2}{\delta^2} \\ &\quad +\lambda_1.\left(-\tfrac{D_\Pi(x)}{D_\Pi(M_i)}\right) + \lambda_2 \textstyle\sum_{j\neq i} \tfrac{\delta^2}{d_m(s_j,x)^2} \end{aligned}\ , \tag{2}$$

où D_Π est la densité du nuage de point estimée par une méthode à noyaux, λ_1, λ_2 sont des pondérations et δ est la distance moyenne entre les projections voisines.
La fonction de coût est composée de trois termes qui sont :

— $\frac{d_m(M_i,x)^2}{\delta^2}$ un terme d'attache à l'initialisation
— $-\frac{D_\Pi(x)}{D_\Pi(M_i)}$ une récompense pour les zones de forte densité
— $\sum_{j\neq i} \frac{\delta^2}{d_m(s_j,x)^2}$ une pénalisation pour les regroupements

La minimisation des fonctions de coût est effectuée par une descente de gradient. Le calcul des arêtes est détaillé dans le paragraphe suivant

2.2 Construction du graphe approximant

Les arêtes $A_\mathcal{M}$ connectent seulement les sommets voisins pour la mesure géodésique. Une arête qui traverse une zone de faible densité a une probabilité plus forte de court-circuiter la variété qu'une arête qui traverse une zone de forte densité. Les arêtes sont donc pondérées par :

$$w_{i,j} = \frac{2D_\Pi\big((s_i+s_j)/2\big)}{D_\Pi(s_i)+D_\Pi(s_j)} \quad . \tag{3}$$

où $w_{i,j}$ est une densité relative.
La construction du GPA repose sur trois contraintes qui sont :

— $G_\mathcal{M}$ est un graphe connexe.
— Deux sommets sont connectés seulement s'ils sont proches pour la distance euclidienne.
— Deux sommets sont connectés seulement si le poids de leur arête commune est élevé.

Pour des niveaux de bruits élevés, le graphe des k-plus fortes liaisons –pour les poids $w_{i,j}$– contient des courts-circuits. Plusieurs étapes sont donc nécessaires à la construction des arêtes. La recherche des arêtes est présentée dans Algo 1 que nous détaillons dans le paragraphe suivant.
Ligne 4 : La recherche des k-plus proches voisins pour la distance euclidienne assure la proximité des sommets connectés.
Ligne 10 : L'arbre couvrant de poids maximum assure que $G_\mathcal{M}$ est connexe.
Ligne 11 : L'utilisation d'un nombre de plus fortes liaisons, k, suffisamment faible assure une probabilité de présence de court-circuit faible, mais limite le nombre d'arêtes. Or le nombre de sommets non connectés qui sont proches pour la mesure géodésique augmente lorsque le nombre d'arêtes diminue.
Ligne 13-17 : Des arêtes sont ajoutées à $G_\mathcal{M}$ à l'aide d'une recherche des k plus fortes liaisons adaptatives dépendant des distances sur le graphe obtenu Ligne 11. La grandeur $r(i,j)$ correspond au rang du point s_j dans le voisinage du point s_i classé par poids croissant.

2.3 Détection des courts-circuits

Le GPA forme une partition de l'espace donnée par le diagramme de Voronoi de ses sommets $S_\mathcal{M}$. Cette partition est munie d'une distance entre cellules. Soit deux cellules représentées par deux sommets de $S_\mathcal{M}$, leur distance est égale au nombre d'arêtes de $A_\mathcal{M}$ dans le plus court chemin entre leurs sommets respectifs.
étant donné un graphe de voisinage, $G(\alpha) = (\Pi, A_\alpha, W)$, sur les données, une arête de A_α est détectée comme étant un court-circuit si les cellules contenant ses extrémités sont séparées d'une distance supérieur à un seuil fixé. Le choix du seuil dépend du nombre de sommets de $S_\mathcal{M}$. Pour $n_m = 400$, le seuil optimal empirique est de 4 arêtes.

Algorithm 1 Construction des arêtes du GPA

1: Entrées : Paramètres :
2: $\Pi,\ S_\mathcal{M}$ $n_v,\ k$;
3:
4: A_0 = unweighted n_v nearest neighbor graph($S_\mathcal{M}$) ;
5: $W_0 = null$;
6: **for** $a_{i,j} \in A_0$ **do**
7: $w_{i,j} = \frac{2D_\Pi\left((s_i+s_j)/2\right)}{D_\Pi(s_i)+D_\Pi(s_j)}$;
8: $W_0 \leftarrow w_{i,j}$;
9: **end for**
10: $A_\mathcal{M}$ = maximum spanning tree(W_0) ;
11: $A_\mathcal{M} = A_\mathcal{M} \cup k$ highest weight graph(W_0) ;
12: D_G = graph distance matrix($A_\mathcal{M}$) ;
13: **for** $a_{i,j} \in A_0$ **do**
14: **if** $\Big(D_{G\ i,j} < \big(n_v - r(i,j)\big)/2\Big) \vee \Big(D_{G\ j,i} < \big(n_v - r(j,i)\big)/2\Big)$ **then**
15: $A_\mathcal{M} = A_\mathcal{M} \cup \{si, sj\}$;
16: **end if**
17: **end for**
return $A_\mathcal{M}$

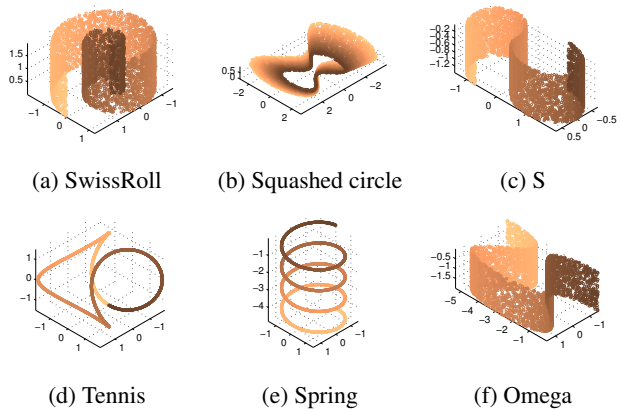
3 Évaluation

La détection des courts-circuits est évaluée sur la qualité de l'estimation des distances géodésiques à partir du graphe $G(\alpha)$ débruité.
Les données utilisées sont des ensembles de 5000 points issus de 6 variétés non-linéaires synthétiques de dimension 3. L'évaluation est répétée 100 fois pour 4 niveaux de bruits différents avec pour chaque expérience un échantillonnage et une réalisation de bruit aléatoire. Des exemples non bruités sont donnés dans la Figure 2.

Les distances géodésiques sont estimées par l'algorithme de Dijkstra [4] et sont évaluées par la mesure d'erreur ci-dessous.

$$err = \frac{1}{n_p}\sqrt{\sum_{i=1}^{n_p}\sum_{j=1}^{n_p}\left(\frac{D_G(i,j)-E(D_G)}{\sqrt{V(D_G)}} - \frac{D_{G0}(i,j)-E(D_{G0})}{\sqrt{V(D_{G0})}}\right)^2} \quad , \tag{4}$$

avec D_G, la matrice des distances géodésiques du graphe évalué, D_{G0} la matrice des distances géodésiques obtenues par le graphe des plus proches voisins sur les données non bruitées, $E(.)$, la moyenne et $V(.)$ la variance.



(a) SwissRoll (b) Squashed circle (c) S

(d) Tennis (e) Spring (f) Omega

FIGURE 2 – 6 variétés synthétiques.

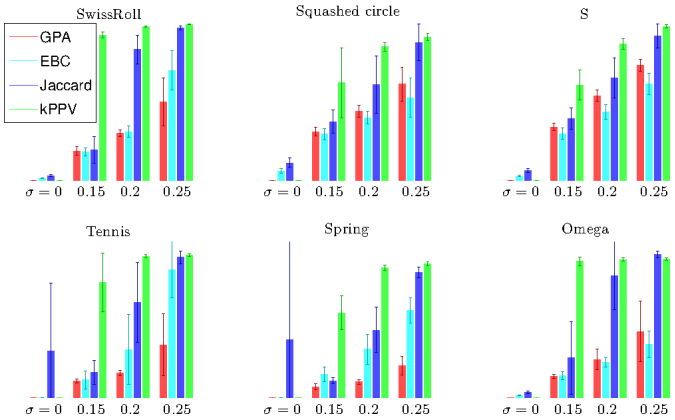


FIGURE 3 – Erreurs d'estimation des distances géodésiques pour des graphes débruités avec différentes methodes.

Les erreurs d'estimation des distances géodésiques moyennes obtenues pour les graphes de 25 plus proches voisins débruités avec notre méthode (GPA), à l'aide de la centralité des arêtes (EBC), avec les indices de Jaccard et sans débruitage (kPPV) sont comparées dans la Figure 3. Pour les variétés SwissRoll, Tennis et Spring, les erreurs les plus faibles sont obtenues pour les graphes débruités avec la méthode GPA. Les erreurs obtenues sur les trois autres variétés sont plus faibles pour les débruitages par GPA et EBC, qui détectent tous les deux les courts-circuits les plus importants – qui permettent de rendre compte de la topologie de la variété après réduction de dimension –. Le seuil utilisé pour notre méthode a été choisis pour avoir un taux de détection nulle sur les variétés non bruitées, il permet de détecter seulement les courts-circuits qui connectent des points éloignés pour la distance géodésique, ce qui induit une légère sous-estimations des distances géodésiques pour les variétés qui possèdent des faibles rayons de courbures comme Squashed circle, S et Omega. Néanmoins, le taux de fausses détection est inférieur pour notre méthode, avec en moyenne, 20 à 100 fois moins d'arêtes supprimées que pour EBC.

Les opérations sur le graphe parcimonieux sont moins coûteuses que sur le graphe de voisinage des données et ne dépendent ni du nombre de courts-circuits, ni de la géométrie de la variété. En moyenne, un rapport 10 a été observé entre les temps de calculs obtenus pour le débruitage par EBC et le notre.

4 Conclusion

La méthode proposée permet de mettre en évidence la structure intrinsèque des données afin de détecter et supprimer les courts-circuits dans les graphes de voisinage. La géométrie de la variété est représentée par un Graphe Parcimonieux Approximant (GPA) construit dans les zones de forte densité. L'utilisation de ce graphe permet de supprimer un nombre minimal d'arêtes qui ne respectent pas les distances géodésiques sur la variété. Nos expériences montrent que le débruitage de graphe par GPA rend les algorithmes de réduction de dimension basés sur les graphes plus robustes au bruit pour des temps de calcul raisonnables par rapport à ceux des méthodes existantes. Les travaux à venir concerneront l'utilisation d'a priori topologique pour améliorer la construction du GPA dans le but de l'utiliser dans le cadre de la cryo-microscopie électronique où les niveaux de bruit sont trop élevés pour que la méthode proposée soit directement applicable.

Références

[1] M. Belkin and P. Niyrogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. 2002.

[2] C. M. Bishop, M. Svensen, and C. K. I. Williams. GTM : the generative topographic mapping. *Neural Comput.*, 10(1) :215–234, 1998.

[3] W. J. Cukierski and D. J. Foran. Using betweenness centrality to identify manifold shortcuts. In *Proc. - IEEE Int. Conf. Data Min. Work. ICDM Work. 2008*, pages 949–958, 2008.

[4] W. Dijkstra. A note on two problems in connexion with graphs. *Numer. Math.*, 1 :269–271, 1959.

[5] D. L. Donoho and C. Grimes. Hessian eigenmaps : locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. U. S. A.*, 100(10) :5591–5596, 2003.

[6] M. Gashler and T. Martinez. Robust Manifold Learning With CycleCut. *Conn. Sci.*, 24(01) :57–69, 2012.

[7] T. Kohonen. The self-organizing map. *Proc. IEEE*, 78(9) :1464–1480, 1990.

[8] A. Singer and H. Wu. Two-Dimensional Tomography from Noisy Projections Taken at Unknown Random Directions. *SIAM J. Imaging Sci.*, 6(1) :136–175, 2013.

[9] J. B. Tenenbaum, V. D. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science (80-.).*, 290(5500) :2319–2323, 2000.

[10] Z. Zhang and H. Zha. Nonlinear Dimension Reduction via Local Tangent Space Alignment. *4th Int. Conf. IDEAL 2003*, 2690 :477–481, 2003.