Une nouvelle méthode de dépliement des données

Catherine Aaron
Université Blaise Pascal
Clermont-Ferrand. France
catherine.aaron@math.univ-bpclermont.fr

Abstract - *Le but de cet article est de passer en revue certaine méthodes de dépliement des données (ISOMAP et LLE) en spécifiant bien les hypothèses sur lesquelles elles reposent, leurs avantages et leurs inconvénients. On présentera alors une nouvelle méthode de dépliement "Curvimap" pouvant déplier des données plus "complexes" qu'isomap en nécessitant de moins gros échantillons que LLE.*

Key words - dimension intrinsèque, dépliement, Isomap, LLE

1 Introduction

1.1 Introduction

Le dépliement des données présente permet de résoudre de nombreux problèmes pratiques. Cela peut permettre représenter les données en dimension plus faible tout en prenant en compte des aspects non linéaires (on a là un pendant à l'ACP dans le cas de données tirées sur des ensembles "pliés"). Tout comme pour l'ACP l'application d'un dépliement permettra de dégager des axes indépendants (mais qui cette fois si ne seront plus des fonctions linéaires des variables) qui pourront servir de base à des regressions. Enfin il existe un lien très étroit entre dépliement et estimation de dimension intrinsèque. La connaissance de cette dernière est très utile dans les applications, par exemple dans la recherche du nombre de retards à prendre en compte lors d'une modélisation temporelle [4]. Dans ce papier on présentera deux méthodes de dépliement : ISOMAP [3] et LLE [2] (Locally Linear Embedding) et nous montrerons pour quels type d'ensemble, ou sous quelles hypothèses ces méthodes donnent de bons résultats. On présentera aussi une nouvelle méthode (qu'on appelera curvimap) qui permet de traiter des cas de manière légèrement moins restrictive que les méthodes sus-citées.

1.2 Introduction

Soient $X^i = (X_1, ..., X_p)^i$ les vecteurs des observations (p variables sont observées pour chaque individu i). On supposera qu'il existe une structure sous-jacente (que les variables sont liées) telle que : $X^i = f(\Theta_1^i, ..., \Theta_d^i) + \varepsilon_i$ Avec Θ_1,...,Θ_d d variables indépendantes (bien sur l'interêt repose dans le fait que d , f une fonction continue et ε un bruit de dimension $p - d$. Le but du dépliement sera alors de rechercher des estimations de θ, f et ε.

2 Les méthodes existantes et leurs hypothèses attachées

2.1 Si f est une fonction linéaire

Il est évident que si f est une fonction linéaire le problème est résolue par une ACP (et le bruit disparait)
Bien sur, des que f est une fonction non linéaire une ACP ne suffira pas. Une des raison de cet echec est que l'utilisation implicite de la distance euclidienne ne reflette pas la complexité éventuelle de l'ensemble traité. Ceci peut être illustré par la figure suivante.

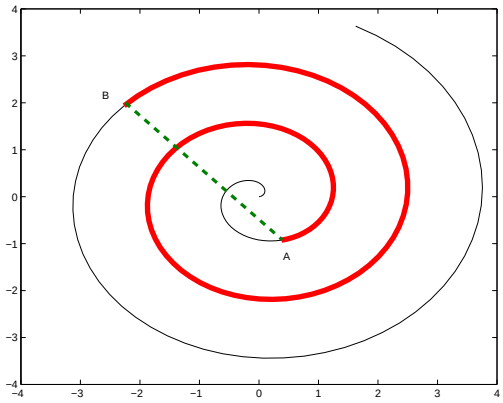

Figure 1: distance géodesique et distance euclidienne pour deux points sur une spirale

2.2 Si le support des données est développable

Si les données sont telles qu'il existe f et Θ_1,...,Θ_d tel que : $X^i = f(\Theta_1^i, ..., \Theta_d^i)$ et que, de plus, $f(\Theta)$ est un ensemble développable alors c'est la méthode ISOMAP développée par J. B. Tenenbaum, V. de Silva and J. C.Langford (see [3]) qui résoudra notre problème. En effet la méthode isomap effectue un Multidimensional Scaling (MDS) sur la base des distances géodesiques entre les points. Or des qu'un ensemble est développable, par définition, la distance géodesique entre les points à les propriétés d'une distance euclidienne sur le parametrage adequat.
Dès que les données ne sont pas tirées sur un ensemble développable ISOMAP dépliera de manière incorecte les données comme le montre la figure suivante :
L'echec d'ISOMAP sur le dépliement d'ensemble non développable viens du fait que si l'ensemble n'est pas développable la distance géodesique n'a plus de propriété euclidiennes et ne reflète pas forcément bien non plus la géométrie de l'ensemble.
Bien sur, la réussite d'autres méthodes de dépliement qui reposent sur l'utilisation de la distance géodésique (par exemple [1]) sera conditionnées à la même hypothèse.

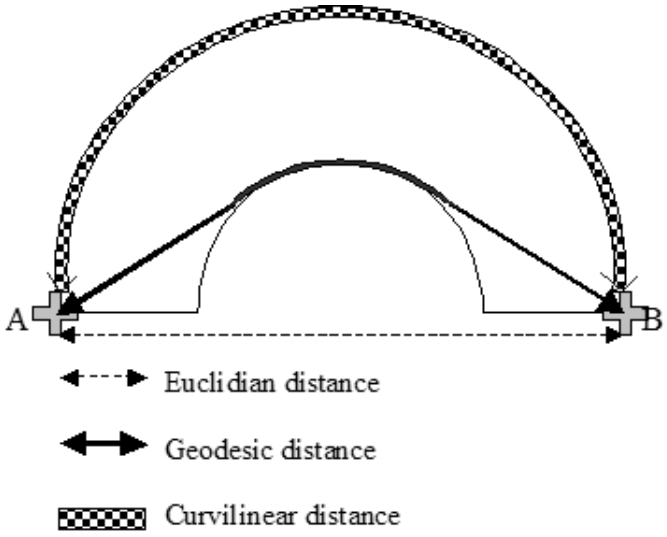

Figure 2: exemple de résultats d'Isomap sur un ensemble non développable

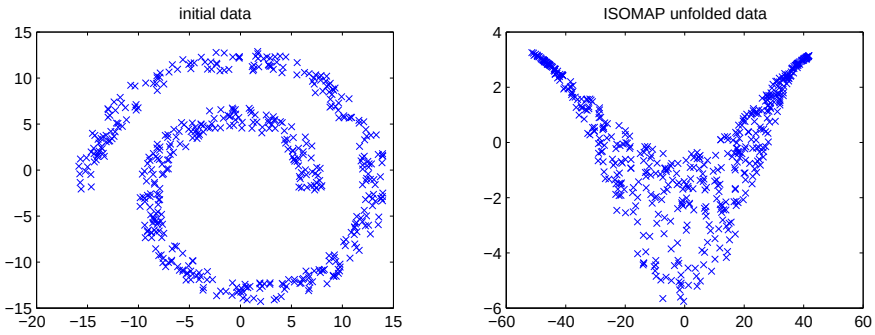

Figure 3: For non torsale surfaces the geodesic distance is not optimal

2.3 Dans le cas général

Dans le cas le plus général qu'il soit (avec uniquement des conditions de régularité de f et de connexité de l'ensemble, il existe une méthode de dépliement : LLE (Locally Linear Embedding) crée par L.K. Saul and S.T. Roweis in [2]. Le principe est très simple, sous les conditions de régularité on utilise le fait que localement l'ensemble peut être considéré comme linéaire et on recherche une matrice W (associée à un structure de voisinage) reflettant la géométrie des données construite de la manière suivante :

- $W_{i,j} = 0$ si X_i n'est pas un voisin X_j
- $\sum_j W_{i,j} = 1$
- W doit minimiser $C(W) = sum_i||X_i - WX_i||$

On cherche alors Y de plus petite dimension que X qui minimise :

$$\Phi(Y) = sum_i||Y_i - WY_i||$$

Le principal inconvénient d'une telle méthode est plus d'aspect pratique que théorique et est lié au très grand nombre de points nécessaire (en fonction de d) à un dépliement correct pour éviter que les problème de bords ne soient prépondérants.

3 Curvimap

On va se placer dans un cas un peu plus général qu'ISOMAP mais suffisament restrictif pour obtenir une méthode nécessitant moins de données que LLE. On supposera, comme pour ISOMAP que $f(\Theta)$ est une surface développable mais on autorisera des bruits ε inon nuls.
La spirale présentée en section 2.2 (figure 2), par exemple, satisfait ces conditions. En effet une telle surface est constituée d'une spirale de dimension une (partie développable) auquel s'additionne un bruit (qui rend l'ensemble non développable).

3.0.1 Principe

Rappelons qu'on veut obtenir :

$$X = f(\theta) + \varepsilon$$

Les hypothèses sur la surface permettent de rechercher des fonction, paramétrages et bruits avec les contraintes suivantes :

- $f(\Theta)$ un ensemble développable
- ε sera localement dans l'orthogonal de $\overrightarrow{grad}(f)$ et orientés

Pour se représenter un peu mieux les choses si l'on suppose que $f(\Theta)$ est de dimension 1 on calcule les bruits dans le repère de Frenet (repère local) associé à un paramétrage de $f(\Theta)$.
Le dépliement correspondra alors aux données de θ et ε
L'écriture précédente rend alors relativement aisée l'estimation de θ, f, $\overrightarrow{grad}(f)$ et ε :

- θ est estimée en utilisant ISOMAP et en concervant les d axes de plus grande variance
- Une fois θ estimé on peut estimer f et $\overrightarrow{grad}(f)$ en utilisant une méthode de regression non linéaire de X sur θ
- Enfin, on obtient les résidus par $X = f(\theta) + \nu$ qu'il suffit de ré-ecrire dans un repère local type Frenet

Un resultat de l'approximation de la base de frenet est présenté dans la figure 4. On peut observer que dans cet exemple les calculs du gradient de f et de son orthogonal (orienté) donne de très bons résultats.

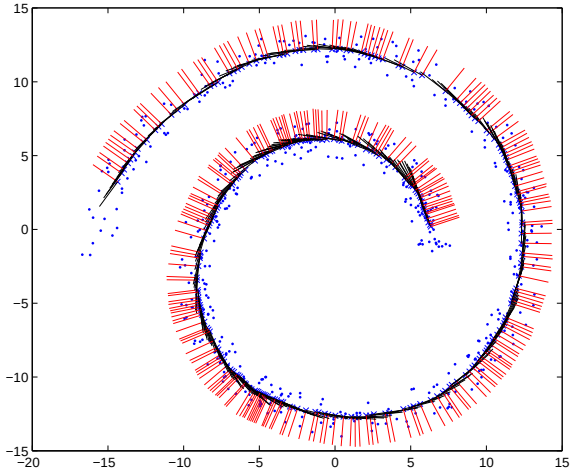

Figure 4: Exemple de repère de frenet approché

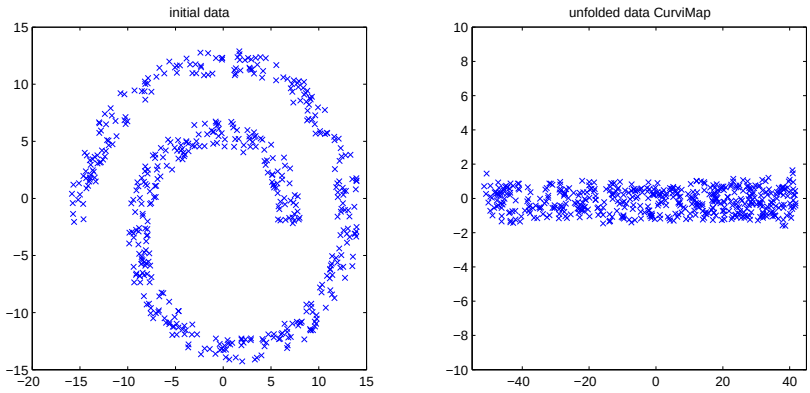

Figure 5: Dépliement associé

3.0.2 Détails de l'algorithm

On présente ici de manière plus détaillée l'algorithme ayant permis de déplier les données.
La première étape consiste en le calcul des distances géodesiques à partir desquelles on va appliquer ISOMAP.
Dans un deuxième temps il faut choisir la dimension de la surface développable sous-jacente (dans le cas de notre exemple "jouet" la dimension est 1). Cette étape est à la fois primordiale et compliquée étant donnée qu'on ne dispose pas d'indicateur fiable pour ce choix. Plusieurs tests peuvent être effectués pour ne conserver au final que le meilleur dépliement.

On peut alors estimer f (fonction qui liera les différents axes issus d'ISOMAP aux coordonnées des observations). Cette estimation va recquierir des méthodes de regression non linéaire. Il faudra alors faire attention à choisir une méthode donnant une fonction suffisament régulière pour que l'estimation de ses dérivées partielles soient robustes. Nous avons choisit dans ce papier une methode par les $k-$plus proches voisins.
$f^{*,(k)}(\theta_i) = (\frac{1}{k+1}\sum G(i,j)(X_{.,j})$ avec $G(i,j) = 1$ si $j = i$ ou si θ_j est un des k plus proches voisins de θ_i.
avec, pour notre exemple $k = 30$

On a aussi choisit d'estimer les dérivées partielles de f par regression locale sur les k'plus proches voisins (dans notre exemple on a choisit $k' = 10$. Ceci permet d'obtenir un premier jeu de vecteurs orthogonaux à notre surface (qu'on notera Z_1).
Enfin il nous faut orienter ces vecteurs. Pour cela on a choisit de calculer le Minimal Spanning Tree (MST) sur les θ. On choisit alors au hasard un point de départ j_0 et on initialise l'ensemble des points "régularisé" à $J_0 = \{j_0\}$ et on définit $Z_2(j_0) = Z_1(j_0)$ puis on itère : pour tout j lié via le MST à un point de J_0 (k_0) :

- On recherche le changement de base P qui minimise $||(Z_1(j)P)'Z_2(k_0) - Id||^2$.
- On calcul alors $Z_2(j) = Z_1(j) * P$.
- Enfin on ajoute à J_0 l'élément j

4 Quelle méthode de dépliement choisir ?

De manière évidente la meilleure méthode est celle qui correspond le mieux à la géométrie malheureusement inconnue de l'ensemble sur lequel les points sont observés. Pour résumer on dispose des méthodes suivante

- $ISOMAP$ ensemble développable
- $CURVIMAP$ ensemble "presque" développable
- LLE ensemble régulier mais avec beaucoup de points en fonction de la dimension intrinsèque (inconnue)

On peut présenter l'heuristique suivante pour choisir la méthode à utiliser

- si le tableau des distance géodésique est très proche de celui des distances euclidienne l'ensemble est déjà déplié et une ACP pourra achever de décomposer les variables en axes indépendants
- Si la distance géodesique a des propriétés de distances euclidiennes (i.e. si à l'issue du MDS on n'obtient que des valeurs propres positives ou nulles) ISOMAP dépliera parfaitement les données
- Si la distance géodesique a des propriétés "presque" euclidienne, les valeurs propres négatives semblent de l'ordre du bruit. On a conscience de l'aspect bien floue de ce propos, on pourra tenter d'applique la méthode "curvimap"
- Enfin si aucune des hypothèses ci-dessus n'est vérifiée et si l'on dispose de suffisament de points au regard de la dimension intrinsèque des données qu'on pourra estimer par ailleurs on appliquera de préférance LLE

5 Conclusion and Further developments

Les résultats en terme de dépliement des données sont encourageant tant d'un point de vue de la représentation des données que d'un point de vue de l'estimation de la dimension intrinsèque (point qui n'est pas développé ici). Néanmoins plusieurs problèmes se posent pour la méthode "curvimap" notament comment savoir si il existe un ensemble développable sous-jacent et quelle est sa dimension. Plusieurs problèmes théoriques apparaissent aussi et ne semblent guère aisé comme la construction d'un test d'hypothèse : la matrice des distance géodésique est la matrice des distances euclidienne, ou, plus difficile encore : la matrice des distance géodesique est une matrice euclidienne.

References

[1] J. A. Lee, A. Lendasse and M. verleysen, (2004), Nonlinear projection with curvilinear distances : Isomap versus curvilinear distance analysis, *Neurocomputing*, **vol. 57** p. 49-76.

[2] Sam T. Roweis and Lawrence K. Saul, (2000), Nonlinear Dimensionality Reduction by Locally Linear Embedding, *Science*, **vol. 290** p. 2323-2326.

[3] Sam T. Roweis and Lawrence K. Saul, (2000), A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science*, **vol. 290** p. 2319-2323.

[4] M. Verleysen, E. de Bodt, A. Lendasse, (1999), Forecasting financial time series through intrinsic dimension estimation and non-linear data projection, *Proceedings of IWANN'99 International Work-conference on Artificial and Natural Neural Networks*, **vol. 290** p. 596-605.