Représentation de courbes en dimension finie

Benjamin Auder, thésard CEA-UPMC

Quelques mots sur l'auteur :

En thèse depuis février 2008 au LCFR (laboratoire de conduite et fiabilité des réacteurs), au CEA Cadarache. Mon directeur de thèse est Gérard Biau (UPMC-ENS) et mon superviseur de thèse Bertrand Iooss (EDF).

Contexte de la thèse :

Compte-tenu de la complexité des systèmes industriels actuels et des progrès en calcul scientifique, les codes utilisés pour modéliser des phénomènes physiques en ingénierie nucléaire sont souvent coûteux en temps. Il est cependant nécessaire de réaliser des analyses statistiques sur certains événements. Ces analyses demandent de multiples applications du code pour être précises. C'est pourquoi nous cherchons à réduire le temps de simulation, en modélisant le code de calcul par une fonction de coût CPU négligeable. Cette modélisation s'effectue sur la base d'un échantillon de résultats du code **f** initial : $(\mathbf{x_i}, \mathbf{y_i})$ avec $\mathbf{f(x_i) = y_i}$, pour i=1…n.

f est un code de calcul à réponse fonctionnelle : dans notre application $\mathbf{f(x_i) \in R^{[a,b]}}$ représente l'évolution (continue) de paramètres physiques dans le temps pour l'état initial $\mathbf{x_i \in R^p}$. La construction d'un métamodèle pour **f** est divisée en plusieurs étapes :

1. classification non supervisée ou clustering des courbes $\mathbf{y_i}$ pour identifier différents comportement du système et faciliter la régression, en parallèle avec la classification supervisée des entrées $\mathbf{x_i}$ dans les clusters ;
2. réduction de la dimension des sorties $\mathbf{y_i}$ dans chaque groupe (pas forcément la même réduction d'un groupe à l'autre), pour appliquer les algorithmes usuels dans le cadre vectoriel ;
3. régression $\mathbf{x_i}$ → coordonnées réduites (représentation vectorielle obtenue à l'étape précédente) , puis restitution coordonnées réduites → courbes, permettant la prédiction de nouvelles réponses.

Communication :

L'objet de cette présentation est le développement des points 2 et 3 précédents. On suppose que les courbes $\mathbf{y_i}$ sont sur une variété fonctionnelle lisse. Cette variété n'est pas forcément un (sous-ensemble d'un) sous-espace vectoriel des fonctions continues de **[a,b]** dans **R**, cas dans lequel une base de fonctions orthonormées est indiquée. Il y a – au moins – trois principales façons de déterminer les coordonnées réduites d'une variété :

1. rechercher un système de courbes principales paramétrées décrivant les données ; ce système généralise la notion de base (orthonormée ou non). « La » définition d'une courbe

principale remonte à [1] et n'est pas unique (voir [2] et [3] par exemple) ; de plus elle nécessite des adaptations (surtout théoriques) dans le cas fonctionnel ;

2. résoudre un problème d'optimisation global (isomap [4]) ;

3. conserver des propriétés locales (LLE [5], laplacian eigenmaps [6]).

Certains algorithmes fonctionnent à la fois au niveau local et global, comme LTSA [7] qui approxime les coordonnées locales par projection sur le plan tangent avant d' « aligner » les coordonnées globales. Nous proposons d'utiliser le principe décrit dans l'algorithme RML (Riemaniann Manifold Learning [8]) pour déterminer des cartes locales « étendues » (en nombre restreint par rapport à la méthode LTSA), avant d'effectuer des transformations affines sur ces coordonnées locales pour en déduire les coordonnées globales dans l'esprit de l'algorithme LTSA.

Avant toutes choses il faut estimer la dimension de la variété échantillonnée par les courbes $\mathbf{y_i}$. Nous commençons par construire un graphe de similarité représentant les données : le graphe des **k** plus proches voisins avec **k** de l'ordre de **n/20 ~ n/10**. Les essais avec un voisinage adaptatif n'ont pas amélioré les résultats dans les applications. La méthode retenue pour estimer la dimension **d** est celle de l'article [9] (Manifold-adaptive dimension estimation).

Pour l'étape de reconstruction nous nous inspirons de l'article [10] : supposant disposer de nouvelles coordonnées prédites $\mathbf{x_{new}}$, on effectue la somme pondérée (inversement aux distances aux représentations des « centres ») des reconstructions pour les cartes locales proches. Une reconstruction consiste à conserver au mieux les angles et distances géodésiques localement.

Les performances de cette méthode seront illustrées sur quelques cas tests (artificiels et industriels), en comparaison avec LTSA, RML et la décomposition sur la base ACP.

Bibliographie :

[1] Principal curves and surfaces by T. Hastie, *PhD Thesis* at Stanford University (1984).
[2] Principal curves: learning, design, and applications by B. Kégl, *PhD Thesis* at Concordia University, Canada, 1999.
[3] Another look at principal curves and surfaces by P. Delicado in *Journal of Multivariate Analysis*, vol. 77-1 (2001), pp. 84-116.
[4] A global geometric framework for nonlinear dimensionality reduction by J. B. Tenenbaum, V. De Silva and J. C. Langford in *Journal of Science*, vol. 290 (2000), pp. 2319-2323.
[5] Nonlinear dimensionality reduction by locally linear embedding by S. T. Roweis and L. K. Saul in *Journal of Science*, vol. 290 (2000), pp. 2323-2326.
[6] Laplacian Eigenmaps for Dimensionality Reduction and Data Representation by M. Belkin and P. Niyogi in *Journal of Neural Computation*, vol. 15 (2002), pp. 1373-1396.
[7] Nonlinear dimension reduction via local tangent space alignment Z. Zhang and H. Zha at *International conference on intelligent data engineering and automated learning 4*, Hong Kong, China, vol. 2690 (2003), pp. 477-481.
[8] Riemannian Manifold Learning by Tong Lin and Hongbin Zha in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30-5 (2008), pp. 796-809.
[9] Manifold-adaptive dimension estimation by A. M. Farahmand, C. Szepesvári and J-Y. Audibert in *Proceedings of the 24th international conference on Machine learning* (2007), pp. 265-272.
[10] Charting a manifold by Matthew Brand in Advances in *Neural Information Processing Systems 15* (2003), pp. 961-968.