INSTITUT DE FRANCE
Académie des sciences

Inria

VILLE DE NICE

UNIVERSITÉ CÔTE D'AZUR

Académie en région
à Nice et Sophia Antipolis

Statistiques géométriques : des fondations mathématiques aux applications en anatomique computationnelle

Xavier Pennec
Inria

L'anatomie computationnelle est une discipline émergente à l'interface de la géométrie, des statistiques, de l'analyse d'images et de la médecine dont l'objectif est de modéliser la variabilité biologique des organes. On s'intéresse par exemple à la forme moyenne et à ses variations dans une population de manière à décrire et à quantifier la diversité des formes et de leurs évolutions normales ou pathologiques. Ces formes sont décrites par les classes d'équivalence sous l'action de transformations sur des ensembles de points, des courbes, des surfaces, des images, voire directement par des déformations. La difficulté est que ces objets géométriques appartiennent en général à des espaces non-linéaires alors que les statistiques ont été essentiellement développées dans un cadre euclidien. Par exemple, additionner ou soustraire deux courbes n'a pas vraiment de sens. On ne peut donc pas facilement parler de leur moyenne. Il convient donc de redéfinir le cadre mathématique dans lequel nous devons développer nos outils d'analyse statistique. L'objectif de cet exposé est de détailler les bases géométriques et quelques avancées récentes dans le développement d'un cadre de travail rigoureux pour effectuer des statistiques sur des objets géométriques vivant dans des variétés.

Les espaces de formes sont la plupart du temps localement euclidiens, et une mesure de distance infinitésimale (une métrique) permet de les munir d'une structure de variété Riemannienne. Celle-ci permet de mesurer des directions, des angles, des distances intrinsèques et les plus courts chemins géodésiques, généralisant ainsi les outils géométriques fondamentaux à des espaces courbes. Sur cette base, on peut redéfinir des notions statistiques consistantes. Par exemple, la moyenne de Fréchet est l'ensemble des points minimisant la somme du carré des distances aux observations. On peut ensuite développer la variété linéairement autour du point moyen : la somme de vecteurs initiaux des géodésiques pointant vers les données est alors nulle. On peut ainsi revenir aux notions statistiques classiques sur ces vecteurs tangents pour définir les moments d'ordre supérieurs. La reformulation de la notion de moyenne permet également d'étendre de nombreux algorithmes de traitement

d'image à des images à valeur dans une variété. C'est le cas par exemple de l'imagerie du tenseur de diffusion (DTI) dont chaque voxel mesure la matrice de covariance anisotrope de la diffusion de l'eau dans les tissus. On peut ainsi établir des algorithmes bien posés d'interpolation, de filtrage, de diffusion et de restauration de données manquantes par l'utilisation de moyennes pondérées [1]. Ces algorithmes sont maintenant très utilisés dans les logiciels de visualisation et d'analyse des images du tenseur de diffusion en IRM.

Du point de vue de l'inférence statistique, l'incertitude de l'estimation de la moyenne empirique est une question clef. Dans des conditions de concentration suffisante des échantillons, un théorème limite central dans les variétés a été obtenu par Bhattacharya & Patrangenaru en 2005. Nous avons récemment établi un développement asymptotique qui met plus clairement en évidence la modulation par la courbure de la vitesse de convergence de la moyenne empirique. Nous avons également établi un développement non-asymptotique en forte concentration qui montre quant à lui un biais statistique de l'estimation dans la direction moyenne du gradient de la courbure [2]. Ces effets de la courbure deviennent importants en cas de forte courbure et peuvent changer drastiquement l'estimation. Ils pourraient ainsi expliquer dans certain cas les moyennes dites collantes (« *sticky means* ») qui ont récemment été mises en évidence et étudiées dans les espaces stratifiés, notamment pour les variétés à coins en courbure négative.

Au-delà de la moyenne, l'analyse en composantes principales est un outil ubiquitaire pour l'analyse statistique et la réduction de dimension. La généralisation la plus simple aux variétés consiste à maximiser la variance expliquée dans l'espace tangent au point moyen (ACP tangente, dénotée tPCA). Il est toutefois souvent plus justifié de minimiser aux moindres carrés les résidus non-expliqués par la projection dans un sous-espace. C'est l'objet de l'analyse en composantes géodésiques principales (PGA), qui considère pour cela des sous-espaces totalement géodésiques au point moyen. Pour réduire l'importance de la moyenne sur la définition des modes, nous avons tout d'abord introduit des sous-espaces plus généraux : les sous-espaces barycentriques sont le lieu des moyennes pondérées d'un nombre fixé de points de référence, généralisant ainsi aux variétés la notion d'espace affine engendré par ces points. Nous avons aussi revisité l'hypothèse usuelle que les données en grande dimension vivent en fait sur une variété de faible dimension (« *the manifold hypothesis* »). En effet, la dimension optimale dépend en pratique de l'échelle à laquelle on approxime les données. Il semble donc plus justifié de construire une séquence de sous-espaces emboités de dimension croissante qui approche de mieux en mieux les données et de choisir a posteriori la dimension, s'il y a lieu. La notion géométrique naturelle qui encode cette structure est celle des variétés de drapeaux pour des sous-espaces linéaires. Dans une variété, les sous-espaces barycentriques peuvent aussi être naturellement imbriqués en ajoutant ou en enlevant des points de référence pour constituer une hiérarchie de sous-espaces proprement imbriqués. Cette vision conduit à reformuler l'ACP comme une optimisation dans un espace de drapeaux généralisés, ce que nous avons nommé analyse en sous-espaces barycentriques. Une application en imagerie cardiaque illustre la puissance de cette approche : le calcul des coordonnées barycentriques par recalage vers trois images de référence sélectionnées dans une séquence ciné-IRM permet par exemple de synthétiser l'ensemble des 30 images du cycle cardiaque avec une erreur de reconstruction 40% plus faible et une compression 2.5 fois plus forte que l'analyse en composantes principales avec 2 modes de déformation.

L'anatomie algorithmique utilise aussi des statistiques sur des groupes de transformations pour caractériser les déformations longitudinales ou inter-sujets. L'usage de métriques Riemanniennes invariantes à droite sur les groupes de difféomorphismes a notamment donné naissance au cadre LDDMM (« *Large deformation Diffeomorphic Metric Mapping* »). La

consistance avec les opérations du groupe n'est toutefois que partielle, car la métrique ne peut pas être en général à la fois invariante à droite et à gauche, ce qui induit un défaut de symétrie : la moyenne de l'inverse d'un ensemble de déformations n'est pas l'inverse de la moyenne de ces déformations. En changeant la structure Riemannienne pour une structure plus faible d'espace symétrique à connexion affine, on peut encore définir des géodésiques, même en l'absence d'une distance. Ces géodésiques sont maintenant des lignes droites et non plus des plus courts chemins. Sur cette base, on peut définir une moyenne locale grâce aux barycentres exponentiels. Dans un groupe de transformations, la structure symétrique canonique est donnée par la connexion de Cartan-Schouten et ses géodésiques sont les sous-groupes à un paramètre réalisés par le flow de champs de vecteurs stationnaires (SVFs), et leurs translations à gauche ou à droite. Ce formalisme justifie ainsi l'usage des SVFs que nous avons introduits il y a quelques années en recalage d'images médicales et qui s'avèrent être très efficaces en pratique pour paramétrer des difféomorphismes.

Le cadre statistique obtenu peut être utilisé pour modéliser l'atrophie du cerveau au cours du temps dans la maladie d'Alzheimer. La trajectoire longitudinale de chaque sujet est modélisée par une géodésique dans l'espace des difféomorphismes. C'est l'équivalent d'un modèle linéaire dans un espace non-linéaire. Les vecteurs initiaux de ces géodésiques doivent ensuite être transportés dans la géométrie du cerveau de référence, le long de la déformation géodésique du sujet vers le gabarit. Ceci requiert des algorithmes numériques de transport parallèle appropriés à la structure choisie pour l'espace de déformations. Avec M. Lorenzi, nous avons adapté pour cela l'échelle de Schild et proposé une modification originale plus symétrique (l'échelle de perroquet ou « *pole ladder* ») en simplifiant l'utilisation des parallélogrammes géodésiques. Nous avons récemment découvert que cette nouvelle méthode particulièrement efficace est en fait un schéma d'ordre 3. Elle est donc bien plus stable que les schémas de transport parallèle habituels qui sont d'ordre 1. De plus, cet algorithme est exact en une seule étape dans les espaces symétriques car il réalise une transvection. Il est donc particulièrement avantageux dans le cadre des SVFs. Après transport de toutes les trajectoires géodésiques longitudinales individuelles dans la géométrie de référence, il est possible d'établir des trajectoires longitudinales moyennes et d'y chercher les caractéristiques qui différencient différentes conditions cliniques. On peut ainsi modéliser le vieillissement normal d'un groupe de contrôle et détecter les déformations supplémentaires occasionnées par la maladie d'Alzheimer ou par d'autres maladies neurodégénératives. Grâce à ces analyses multivariées sur les paramètres des déformations, Il est même possible de mettre en évidence des différences statistiquement significatives entre des sous-groupes de contrôle présentant des facteurs de risque différents, alors que les méthodes classiques reposant sur les changements de volumes ne détectent aucune différence.

Ces résultats applicatifs montrent l'importance de la géométrie dans la modélisation de données en sciences de la vie. Des structures géométriques comme les quotients par des groupes de difféomorphismes, les arbres phylogéniques ou les espaces de formes apparaissent en effet naturellement dans les applications. Or, nous avons montré avec les thèses récentes de Nina Miolane et de Loic Devillier que la moyenne de Fréchet dans un espace quotient stratifié présente en général un aspect répulsif au niveau des singularités qui rejette l'estimation empirique dans une strate de dimension supérieure. Ce comportement est à l'opposé des moyennes collantes, où la singularité attire la moyenne empirique, et peut conduire à amplifier le bruit au lieu de le réduire. Il convient donc de mieux comprendre l'interaction de la géométrie avec l'estimation statistique et de consolider les fondements des statistiques géométriques en étendant leur domaine de validité à des espaces non-euclidiens plus complexes que les variétés Riemanniennes. C'est l'objectifs de mon projet ERC *G-Statistics*. Par exemple, des résultats importants ont été établis pour des espaces à courbure

négative en utilisant la convexité. Il serait intéressant d'étendre ces méthodes à d'autres structures géométriques non convexes, de courbure potentiellement positive, et qui peuvent présenter des singularités et des changements de dimension (stratifications). A l'instar de la physique, on espère ainsi pouvoir découvrir des invariances approximatives, c'est-à-dire des lois empiriques, caractérisant mieux les données extrêmement variables et très bruitées des sciences de la vie, et assoir encore un peu plus la déraisonnable efficacité des mathématiques rendue célèbre par Eugène Wigner.

Références clés

1. Xavier Pennec, Stephan Sommer, and Tom Fletcher. *Riemannian Geometric Statistics in Medical Image Analysis*. 636 p. Academic Press, September 2019.
2. Xavier Pennec. *Curvature effects on the empirical mean in Riemannian and affine Manifolds: a non-asymptotic high concentration expansion in the small-sample regime*. ARXIV preprint 1906.07418, June 2019.
3. Xavier Pennec. *Barycentric Subspace Analysis on Manifolds*. Annals of Statistics, 46(6A):2711-2746, July 2018.

Illustration

Statistiques dans un groupe de difféomorphismes muni de la connexion affine symétrique canonique pour modéliser la trajectoire du vieillissement normal et la composante de déformation additionnelle imputable à la maladie d'Alzheimer. Les déformations géodésiques longitudinales régressées dans les séquences d'images de chacun des sujets sont transportées parallèlement vers le gabarit de référence où un modèle linéaire dans l'espace tangent permet de calculer la trajectoire moyenne pour les différentes conditions cliniques. D'après des images et illustrations originales de Marco Lorenzi et Raphaël Sivera.