THE UNIVERSITY OF CHICAGO
PHYSICAL SCIENCES

DEPARTMENT OF MATHEMATICS
5734 SOUTH UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637
TELEPHONE: 773-702-7100
FAX: 773-702-9787
WEB: math.uchicago.edu
E-MAIL: ucmath@math.uchicago.edu

Report on: Inférence topologique à partir de mesures et de fibrés vectoriels by Raphaël TINARRAGE

This thesis consists of three distinct and interrelated contributions to topological data analysis, which contribute to the basic inference problem of deciding a topological structure on the basis of samples from a space (especially a manifold) and to understanding objects associated to such a space. The main contributions are:

(1) Inferring topological invariants of a manifold, especially the homology. Here Tinarrage gives a novel approach for dealing with noise or outliers using a novel kind of distance to measure (DTM) approach which is computationally feasible. This area itself is one which has now been studied for decades with a number of different methods popular. The approach given here is both elegant and has interesting properties. Tinarrage describes an interesting application of this in an industrial research project and also uses his DTM approach in his work on the immersed reconstruction.

(2) Topological inference from an immersed manifold. This is a beautiful and natural idea: We usually picture a Klein bottle via its singular projection into 3-space, although that is not topologically accurate. The problem to be considered is that the measurements are not adequate to determine the data. In some sense this is mathematically not completely well defined: the image f(M) of an immersion of a manifold does not determine the manifold, although with genericity conditions this becomes a reasonable question. To do this, Tinarrage uses a mix of delicate geometric reasoning about the estimation of the tangent space on most of the point cloud, and then uses this to give a "platonic" metric measure space in a product of Euclidean space and a Grassmanian (or space of matrices). The construction is very nice, and the arguments are technically involved.

(3) Computing the basic mod 2 cohomological invariants of a vector bundle, i.e. the Stiefel-Whitney classes. (The method could surely be extended to other cohomological invariants of real and complex bundles, such as Euler classes, Chern classes and Pontrjagin classes, although incorporating these into explicit algorithms would require significant additional work. Indeed, the treatment of higher Stiefel Whitney classes here is not as explicit as the treatment of the first such class. Pontrjagin classes would be of interest because they are strong enough to distinguish examples of non-homeomorphism among homotopy equivalent simply connected manifolds). The method requires data on the vector bundle requiring an image not to lie in the medial axis of a Grassmanian embedded in a matrix space. [This is shown to be a computable condition; it seems to be topologically obstructed and is a measurement of whether the sampling of the space is fine enough to capture the nontriviality of the bundle.] The use of this auxiliary space (classifying the vector bundle) is much in the same spirit as the construction in part (2) of the reconstruction of the original manifold that underlies a given immersed image in Euclidean space.

The techniques used are a mixture of differential geometry (both of manifolds and of bundles), the algebra of persistence modules, measure theory (in terms of the original DTM techniques and also in the understanding of the desingularization of an immersed image), as well as a useful discussion of linear algebra's connection to the geometry of Grassmanians. This is an excellent thesis, in terms of the interrelated problems that were studied, the naturality of the ideas, the technical skill demonstrated implementing these, and its potential for extension and application. The only mistakes that I detected were in English spelling. I strongly support its acceptance.

Sincerely,

Shmuel Weinberger
Andrew MacLeish Professor of Mathematics
And Chair, Department of Mathematics.

September 15, 2020

Simon Masnou
Professeur
Université Claude Bernard Lyon 1
Institut Camille Jordan (UMR 5208)
43 boulevard du 11 novembre 1918
69622 Villeurbanne cedex
masnou@math.univ-lyon1.fr
http://math.univ-lyon1.fr/~masnou

Lyon, le 6 octobre 2020,

Rapport sur le manuscrit, titré *Inférence topologique à partir de mesures et de fibrés vectoriels*, présenté par M. Raphaël Tinarrage en vue de l'obtention du grade de Docteur en Mathématiques appliquées de l'Université Paris-Saclay

Les travaux de Raphaël Tinarrage relèvent de l'analyse topologique des données, un domaine qui a connu au cours des 20 dernières années un essor fantastique tant les besoins de traiter, d'analyser, de comprimer, de comparer ou de classifier des données de plus en plus complexes ont augmenté. Dans la gamme très vaste des outils mathématiques et numériques utilisés pour l'analyse des données, R. Tinarrage s'intéresse en particulier à la théorie de la persistance topologique dont l'objectif est de déterminer des caractéristiques topologiques d'une façon qui soit la plus invariante possible pour une large classe de perturbations (perturbations géométriques, sous-échantillonnage, ajout de bruit dans les données, etc.). Il s'agit d'une théorie riche et éclectique : les questions auxquelles R. Tinarrage s'est intéressé et les réponses qu'il a apportées font appel à des outils de topologie algébrique, de géométrie riemannienne, de théorie géométrique de la mesure, de transport optimal mais aussi à de nombreux outils numériques non triviaux de manipulation de complexes simpliciaux.

Le manuscrit de R. Tinarrage comporte 5 chapitres. Le chapitre 1 est une longue introduction qui donne un excellent aperçu des applications de l'analyse topologique des données et propose une synthèse très claire des résultats de la thèse. Le chapitre 2 contient des rappels sur tous les outils théoriques utilisés et permet de prendre la juste mesure de leur diversité et de leur technicité. Les résultats nouveaux obtenus durant la thèse sont longuement décrits aux chapitres 3, 4 et 5. Un premier constat s'impose : ce manuscrit est remarquablement écrit, avec un souci constant de clarté, d'illustration et de mise en perspective des résultats et des preuves qui est enthousiasmant. Par ailleurs, chacun des chapitres 3, 4 et 5 est complété par un *notebook* Python en ligne extrêmement utile : ces notebooks offrent un magnifique éclairage numérique sur les résultats de la thèse et donnent la possibilité au lecteur de rentrer (en partie) dans le détail des codes numériques.

Décrivons maintenant les résultats obtenus par R. Tinarrage. Le chapitre 3 est né des questions soulevées par une application industrielle de détection d'anomalies dans des données de capteurs inertiels installés sur des ponts ou dans des bâtiments. Ce chapitre apporte plus précisément une réponse à la question suivante : est-il possible de définir une filtration d'un nuage de points dont l'homologie persistante ait de bonnes propriétés d'approximation, de stabilité, de robustesse au bruit et aux points aberrants, et qui soit relativement simple à calculer? Les filtrations classiques de Čech et de Vietoris-Rips ne sont pas suffisamment robustes au bruit, plusieurs variantes pondérées ont donc été proposées dans lesquelles le recouvrement utilisé fait intervenir un rayon inhomogène adapté aux données. Une façon intéressante de s'adapter aux données consiste à utiliser la notion de *distance-to-measure* (DTM) qui module la distance en fonction de la densité locale de la mesure naturelle associée à l'ensemble étudié. La DTM est spatialement lipschitzienne et, vue comme application sur l'espace des mesures, continue en norme L^∞ par rapport à la 2-distance de Wasserstein. Utiliser comme filtration les ensembles de niveau de la DTM est cependant trop coûteux numériquement pour le calcul de l'homologie persistante et plusieurs alternatives ont donc été proposées. La variante proposée par R. Tinarrage consiste à utiliser la DTM comme fonction de pondération dans les filtrations pondérées de Čech et Vietoris-Rips. Cela permet de garantir à la fois un coût de calcul raisonnable et, en pratique, la robustesse au bruit associée à la DTM comme l'illustrent plusieurs exemples. R. Tinarrage établit des résultats théoriques de stabilité relative et d'approximation qui offrent un contrôle de la pseudo-distance d'entrelacement entre les filtrations DTM associées à deux ensembles par la 2-distance de Wasserstein relativement à deux ensembles de comparaison dont on pénalise la non-concentration. Ces résultats de stabilité sont précisés lorsque l'un des ensembles est suffisamment dense relativement à une mesure de Hausdorff (ce que R. Tinarrage appelle un ensemble (a, d)-standard).

La seconde contribution de la thèse (chapitre 4) porte sur la détermination de l'homologie singulière d'un ensemble à partir de points échantillonnés sur une immersion de cet ensemble dans un espace euclidien (immersion qui n'a bien sûr aucune raison d'avoir la même homologie que l'ensemble de départ). R. Tinarrage propose une solution intéressante sous des hypothèses de transversalité de l'immersion, ainsi que de régularité et de densité de l'ensemble de départ : il propose d'estimer à partir des points échantillonnés une mesure associée au relèvement $x \mapsto (x, p_{T_xM})$ où p_{T_xM} désigne la projection orthogonale sur l'espace tangent. L'intérêt évident de ce relèvement sous l'hypothèse de transversalité est qu'il sépare les points multiples et doit donc permettre de retrouver l'homologie de l'ensemble de départ. Dans ce contexte, R. Tinarrage introduit une notion de *reach normal* dont il montre par des résultats topologiques et géométriques qu'elle est mieux adaptée aux immersions que la notion classique de *reach* (qui vaut 0 en cas de point multiple). Il prouve en outre en dimension 1 une propriété d'homogénéité. Il s'intéresse ensuite à l'estimation des espaces tangents à l'aide de matrices de covariance locale et il prouve pour la mesure relevée associée des résultats très intéressants de cohérence, de stabilité et d'approximation relativement à la p-distance de Wasserstein. La dernière partie du chapitre 4 porte sur l'homologie persistante de la filtration-DTM de la mesure relevée associée à l'immersion (sous les mêmes hypothèses que précédemment) et son approximation par la mesure relevée estimée à partir d'un échantillonnage. Les résultats démontrés et les illustrations numériques sont vraiment très convaincants et montrent toute la pertinence et la cohérence générale des travaux présentés aux chapitres 3 et 4. Par ailleurs, R. Tinarrage mentionne le lien évident entre la mesure relevée qu'il étudie et la notion de varifold. Une question me vient naturellement : ne pourrait-on pas s'affranchir de la contrainte de transversalité en adaptant la notion de varifold (localement) orienté introduite par Hutchinson?

Poursuivant dans la veine du relèvement, le dernier chapitre du manuscrit (chapitre 5) porte sur l'étude de la persistance cohomologique (pour les classes de Stiefel-Whitney) de filtrations de Čech de fibrés vectoriels $\mathbb{R}^n \times \mathcal{M}(\mathbb{R}^m)$. R. Tinarrage propose une construction élégante de ces filtrations de Čech puis étudie la persistance de leurs classes de Stiefel-Whitney en démontrant des propriétés convaincantes de stabilité et de cohérence. Il propose ensuite un schéma d'algorithme pour le calcul des classes persistantes, schéma dont la mise en œuvre pratique est trop difficile dans le cas général mais réalisable lorsque les espaces tangents sont de dimension 1. Quelques exemples numériques, ainsi que ceux disponibles dans le notebook Python, illustrent la pertinence de l'approche.

Parmi les enjeux actuels de l'inférence topologique figure le traitement de données très complexes, bruitées et en très grand volume. Il est donc essentiel de disposer de méthodes théoriquement robustes et numériquement efficaces. Bien que la question du volume des données traitées n'ait pas été abordée spécifiquement par R. Tinarrage, les outils qu'il introduit et étudie en détail dans son manuscrit ont bien pour ambition le traitement de données plus générales par des méthodes numériques relativement accessibles et offrant des garanties théoriques profondes. L'ensemble des résultats et des méthodes présentés dans ce manuscrit, à la frontière de plusieurs champs mathématiques, sont extrêmement séduisants. Chacun des chapitres 3, 4 et 5 est riche en contributions de tout premier plan qui vont sans aucun doute ouvrir la voie à d'autres développements. Il s'agit d'un travail de thèse remarquable et je recommande avec enthousiasme que Raphaël Tinarrage soit autorisé à soutenir.

Simon Masnou