Optimal transport for temporal graph decomposition – application to care trajectories analysis

The internship will be supervised by:

- Thomas Guyet, Inria, Lyon, thomas.guyet@inria.fr

The internship will be hosted at the Inria Lyon (hosted at the Universty Hospital, 56 Bld Pinel, Lyon) in the AIStroSight project that aims at developing numerical tools to derisk drugs repurposing. The AIStroSight project gathers competencies in artificial intelligence and computational biology to tackle this challenge.

Interactions with, Titouan Vayer, specialists in graph decomposition with optimal transport is planned during the internship.

Contexte applicatif

Un entrepôt de données de santé (EDS) contient les informations médicales des patients admis dans un hôpital. La base de données contient des informations sur les visites des patients, y compris les soins et les médicaments délivrés lors de chacune de leurs visites (avec leur date et heures de délivrance). Par exemple, l'APHP a identifié une cohorte de plus de 20 000 patients hospitalisés pendant la crise du Covid-19. Un jeu de données a été créé à partir des informations sur leur état de santé et les soins qu'ils ont reçus. L'ensemble de ces informations constitue leur "trajectoire de soins".

La notion de "trajectoire de soins" désigne la séquence temporelle des soins dispensés à un patient. Ces parcours de soins sont des objets complexes mais ils contiennent de nombreuses informations qu'il est utile d'exploiter pour comprendre les interactions entre les soins prodigués et l'état de santé du patient.

Pour exploiter des objects complexes, disponibles en masses mais contenant une grande variabilité, il est nécessaire de mettre en place des outils informatiques qui vont permettre d'analyser ces parcours tels que disponibles dans un EDS.

Le développement de l'analyse des parcours de soins peut avoir deux objectifs : 1) caractériser les soins de santé des patients souffrant d'une maladie, et 2) caractériser de petits groupes de patients qui ont des réponses similaires à une stratégie de soins (potentiels similaires de guérison ou risques similaires d'induire des événements indésirables). Le premier objectif vise à proposer une meilleure prise en charge des soins. Par exemple, dans le cas de la crise du Covid-19, il était intéressant d'identifier les stratégies de soins qui auraient évité aux patients de nécessiter des soins intensifs [2, 8]. Le second objectif est utile pour la médecine personnalisée et le repositionnement de médicaments. Dans ce dernier cas, le fait de connaître à l'avance certains effets indésirables possiblement induits par l'introduction de certains médicaments dans le parcours de soins des patients peut éviter des investigations peu prometteuses.

Dans les deux cas, nous avons besoin d'extraire des parcours de soins et des patients types. Dans ce stage nous nous intéressons à des méthodes d'analyse des données d'un entrepôt de données de santé pour identifier des parcours de soins typiques, vus comme des séquences d'événements médicaux, et de groupes de patients également typiques.

Contexte scientifique

Une technique de l'état de l'art pour répondre à cette problématique est celle de la factorisation tensorielle (ou décomposition tensorielle) [3]. Cette technique générique consiste à décomposer un tenseur $\mathcal{X}$ de dimension n en un ensemble de tenseurs de dimension inférieure $\mathcal{Y}_1, \ldots, \mathcal{Y}_k$ tels que $X \approx \mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_k$ où $\otimes$ est un produit matriciel.

Pour des données longitudinales, $\mathcal{X}$ est considéré comme un tenseur tridimensionnel dont les dimensions sont l'identifiant de l'individu (patient), le temps et les événements. La décomposition des tenseurs bidimensionnels permet d'identifier des profils types.

Ces dernières années, des techniques de factorisation parcimonieuses ont permis d'adapter ces problématiques pour leur permettre de passer à l'échelle et de gagner en stabilité [12]. D'autre part, en apprentissage automatique, plusieurs architectures récentes de réseaux de neurones ont été proposées [7, 1, 11, 8]. Elles ont prouvé la

faisabilité de l'approche pour décomposer efficacement des tenseurs larges et complexes. Un intérêt pratique de ces méthodes de décomposition est de fournir des résultats qui soient facilement analysables par des cliniciens ou épidémiologistes. Ce qui les rend attrayantes dans le contexte médical.

Néanmoins, ces approches de décomposition offrent une expressivité limitées du fait d'une représentation matricielle des données, et plus spécifiquement :

- les types d'événements médicaux sont strictement distincts. Par exemple, une prescription de Prednisone, ou de Prednisolone peut s'avérer équivalente. Sans cette information, une méthode automatique pourrait manquer des régularités pertinentes dans les données.
- l'information temporelle est contrainte par une succession rigide. Par exemple, arriver à l'hopital puis deux jours après être intubé est différents de si l'intubation à lieu 1 jour après l'hospitalisation. Pour ces délais similaires on aimerait pouvoir que nos méthodes identifie une certaine régularité.

Le point de départ de ce travail est l'utilisation d'une représentation de trajectoires longitudinales sous la forme d'un graphe (temporel) pour répondre aux questions initiales d'identification de parcours typiques et de groupes de patients typiques. Un graphe temporel représente des informations comme un ensemble d'objets reliés par des relations temporelles et éventuellement sémantiques. Il existe differentes manière de coder l'information temporelle sous la forme de graphe [4].

Dans ce stage, on se propose d'explorer des méthodes de comparaison de graphes basés sur le transport optimal et de développer des approches originales d'apprentissage automatique pour identifier des parcours types de patients. Les techniques liées au transport optimal ont déjà été développées dans le cadre de graphes labellisés aux noeuds [9, 10]. Elles permettent d'apprendre des régularités au sein d'une collection de graphes et, à la manière de la décomposition tensorielles, de découvrir un dictionnaire de graphes médians (des phénotypes) qui composent une collection de graphes.

Plusieurs nouvelles questions se posent pour adapter ce type de modèle aux trajectoires longitudinales sous forme de graphes temporels. Tout d'abord, les graphes temporels sont également labellisés sur les arcs. Une adaptation des méthodes existantes de décomposition de graphe est donc nécessaire. Nous envisageons deux alternative de représentation de l'information temporelle : sous la forme de graphes dynamiques [6] ou sous la forme d'information dans le graphe. Pour chacune de ces représentations, de nouveaux modèles de comparaison de graphes basés sur le transport optimal seront proposés, notamment pour considérer des alignements temporels flexibles. Plus spécifiquement, nous envisageons d'explorer des modélisations par processus de points, tel quel des modèles de Hawkes, pour lesquels des processus de décompositions simiilaires aux techniques de graphes ont été proposées [5]. La fusion de ces méthodes doit permettre d'identifier des groupes typiques de sous-graphes représentatifs des trajectoires dans les données..

Chacun des modèles de comparaison de graphes sera évalué sur des données synthétiques pour en évaluer les propriétés et sur des données réelles.

Candidate profile

- You are student in a Master 2 in computer science, data science or statistics, or student in a engineering school.
- You are enthusiastic about research, you love to understand in depth the problems and to find them elegant solutions.
- You have an strong background in math and computer science (Python for machine learning environment).
- You are interested in artificial intelligence and, more precisely, in machine learning, optimization techniques, data analysis, ...
- You have interest in the field of health and to contribute to the development of solutions that may help clinicians or epidemiologists.
- You speak and write English and/or French.

References

[1] Ardavan Afshar, Ioakeim Perros, Evangelos E. Papalexakis, Elizabeth Searles, Joyce Ho, and Jimeng Sun. COPA: Constrained PARAFAC2 for sparse & large datasets. page 793–802, 2018.

[2] Mathieu Chambard, Thomas Guyet, Yên-Lan Nguyen, and Etienne Audureau. Temporal phenotyping for characterisation of hospital care pathways of covid19 patients. In *AALTD 2021-The 6th International Workshop on Advanced Analytics and Learning on Temporal Data*, 2021.

[3] Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

[4] Jong Ho Jhee, Alberto Megina, Pacôme Constant Dit Beaufils, Matilde Karakachoff, Richard Redon, Alban Gaignard, and Adrien Coulet. Predicting clinical outcomes from patient care pathways represented with temporal knowledge graphs. In *European Semantic Web Conference*, pages 282–300. Springer, 2025.

[5] Dixin Luo, Hongteng Xu, and Lawrence Carin. Fused gromov-wasserstein alignment for hawkes processes. *arXiv preprint arXiv:1910.02096*, 2019.

[6] Franco Manessi, Alessandro Rozza, and Mario Manzo. Dynamic graph convolutional networks. *Pattern Recognition*, 97:107000, 2020.

[7] Ioakeim Perros, Evangelos E Papalexakis, Fei Wang, Richard Vuduc, Elizabeth Searles, Michael Thompson, and Jimeng Sun. Spartan: Scalable parafac2 for large & sparse data. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining -ACM SIGKDD)*, pages 375–384, 2017.

[8] Hana Sebia, Thomas Guyet, and Etienne Audureau. Swotted: an extension of tensor decomposition to temporal phenotyping. *Machine Learning*, 113(9):5939–5980, 2024.

[9] Titouan Vayer, Nicolas Courty, Romain Tavenard, Laetitia Chapel, and Rémi Flamary. Optimal Transport for structured data with application on graphs. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 6275–6284, Long Beach, USA, 09–15 Jun 2019.

[10] Cédric Vincent-Cuaz, Titouan Vayer, Rémi Flamary, Marco Corneli, and Nicolas Courty. Online Graph Dictionary Learning. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 10564–10574, Online, 18–24 Jul 2021. PMLR.

[11] Kejing Yin, Dong Qian, William K. Cheung, Benjamin C. M. Fung, and Jonathan Poon. Learning phenotypes and dynamic patient representations via rnn regularized collective non-negative tensor factorization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1246–1253, 2019.

[12] Léon Zheng, Elisa Riccietti, and Rémi Gribonval. Efficient Identification of Butterfly Sparse Matrix Factorizations. working paper or preprint, April 2022.