Une extension de la décomposition tensorielle au phénotypage temporel

Hana Sebia*, Thomas Guyet*, Etienne Audureau**

* Inria, AIStroSight, Centre de Lyon, France
{hana.sebia, thomas.guyet}@inria.fr,
** AP-HP, Hôpital Henri Mondor, Université Paris Est Créteil, France

Résumé. La décomposition tensorielle a récemment fait l'objet d'une attention croissante dans la communauté de l'apprentissage automatique en raison de sa polyvalence dans le traitement des données à grande échelle. Cependant, cette tâche devient plus difficile lorsqu'il s'agit de prendre en comte la dimension temporelle. Dans cet article, nous étendons la décomposition tensorielle à l'extraction de phénotypes temporels, décrits comme un combinaison de caractéristiques sur une fenêtre de temps. Nous proposons un nouveau modèle de décomposition intégrant plusieurs régularisations pour améliorer l'interprétabilité des phénotypes extraits. Nous validons ce dernier sur des données synthétiques et réelles provenant de l'Assistance Publique – Hôpitaux de Paris (AP-HP). Les résultats montrent qu'il est plus performant que les modèles les plus récents de décomposition et qu'il découvre des phénotypes intéressants pour les cliniciens.

1 Introduction

Un tenseur est une représentation naturelle des données multidimensionnelles. La décomposition tensorielle est un outil statistique historique pour l'analyse de ces données complexes. La popularisation de techniques d'apprentissage automatique efficaces et évolutives l'a rendue attrayante pour les données du monde réel (Perros et al., 2017). Elle a donc été intensivement étudiée dans de nombreux domaines, tels que le traitement du signal, les neurosciences, la communication, la psychométrie, etc (Fanaee-T et Gama, 2016). Techniquement, la décomposition tensorielle est une méthode non-supervisée qui simplifie un tenseur multidimensionnel en tenseurs d'ordre inférieur (Anandkumar et al., 2014). Ceci revient à identifier des variables latentes. Ces variables latentes sont des caractéristiques non observées qui capturent les comportements cachés d'un système. Elles sont difficiles à extraire de données multidimensionnelles complexes en raison 1) des multiples interactions entre les dimensions et 2) de l'entrelacement des occurrences de comportements cachés.

Récemment, plusieurs approches basées sur la décomposition tensorielle ont montré leur efficacité et leur intérêt pour le phénotypage computationnel à partir des dossiers médicaux électroniques (DME) (Afshar et al., 2020, 2021; Chambard et al., 2021; Yin et al., 2019). Les motifs récurrents cachés qui sont découverts dans ces données sont appelés *phénotypes*. Ces phénotypes sont particulièrement intéressants pour 1) décrire les pratiques réelles des unités

médicales et 2) aider les administrateurs d'hôpitaux à améliorer leur gestion des soins. Par exemple, une meilleure description des parcours de soins des patients COVID-19 au début de la pandémie peut aider les cliniciens à améliorer la gestion des soins lors des futures vagues épidémiques. La principale limite des techniques existantes est la définition d'un phénotype comme une combinaison de traitements sans tenir compte de la dimension temporelle. On parle alors de phénotypes journaliers. Dans ce cas, un parcours de soins est considéré comme une succession de soins quotidiens indépendants. Néanmoins, il semble plus réaliste d'interpréter un parcours de soins comme des combinaisons de *séquences de traitements*. Par exemple, les patients COVID-19 hospitalisés pour un syndrome de détresse respiratoire aiguë sont traités pour plusieurs problèmes au cours de la même visite : infection virale, syndromes respiratoires et problèmes hémodynamiques. D'une part, le traitement de l'infection virale implique l'administration de médicaments sur plusieurs jours. D'autre part, le syndrome respiratoire aigu nécessite également une surveillance continue pendant plusieurs jours. Le parcours de soins d'un patient peut alors être abstrait comme une combinaison de ces traitements. Le phénotypage computationnel vise à découvrir un phénotype pour chaque traitement.

Dans cet article, nous présentons un modèle décomposition tensorielle basé sur l'apprentissage automatique pour extraire les phénotypes temporels. Contrairement à un phénotype journalier classique, un phénotype temporel décrit la combinaison de médicaments/procédures sur une fenêtre temporelle de quelques jours. Cela améliore nettement l'expressivité de la méthode. Suivant le principe de la décomposition tensorielle, le modèle découvre des phénotypes qui reconstruisent avec précision un tenseur d'entrée avec une dimension temporelle. Il permet le chevauchement d'occurrences distinctes de phénotypes pour représenter le début asynchrone de traitements. À notre connaissance, notre proposition est la première extension de la décomposition tensorielle au phénotypage temporel. Nous évaluons notre modèle en utilisant à la fois des données synthétiques et des données réelles de patients. Les résultats montrent qu'il reconstruit plus précisément les parcours que les modèles de l'état de l'art. De plus, l'analyse qualitative montre que les phénotypes découverts sont cliniquement significatifs.

2 État de l'art

La découverte de phénotypes à partir des données longitudinales du DME est une tâche fondamentale qui a fait l'objet de nombreux travaux. Ces données peuvent être structurées en un tenseur tridimensionnel, c'est-à-dire un cube de données dont les dimensions sont : les identifiants des patients, les événements de soins (procédures, tests de laboratoire, médicaments administrés) et le temps. La décomposition tensorielle a été largement utilisée et a fait ses preuves pour extraire des modèles concis et interprétables à partir de telles données. La factorisation CP (Kolda et Bader, 2009) est la technique de décomposition générique qui décompose un tenseur $\mathcal{X}$ en une collection de tenseurs d'ordre inférieur $\mathcal{Y}_1, \dots \mathcal{Y}_m$ tels que $\mathcal{X} \approx \mathcal{Y}_1, \otimes \cdots \otimes \mathcal{Y}_m$ où $\otimes$ représente le produit vectoriel externe. PARAFAC2 (Kiers et al., 1999) est une variante de la factorisation CP qui traite les tenseurs irréguliers. Ceci est particulièrement intéressant pour les patients dont la durée de séjour à l'hôpital est variable. Divers travaux ont proposé des améliorations du modèle de base de PARAFAC2. Certains ont exploité des architectures parallèles pour passer à l'échelle (Perros et al., 2017; Afshar et al., 2018, 2021). D'autres ont introduit diverses contraintes (Afshar et al., 2018; Yin et al., 2020) sur les matrices résultantes pour améliorer leur interprétabilité, ou même introduit des tâches de pré-

diction spécifiques (Wang et al., 2015; Yang et al., 2017; Henderson et al., 2018) pour guider la décomposition. Ces avancées ont amélioré la décomposition des parcours de soins. Cependant, il existe un nombre limité de contributions améliorant l'exploitation de la dimension temporelle. COPA (Afshar et al., 2018) utilise une régularisation pour considérer l'irrégularité des durées entre deux visites. LogPar (Yin et al., 2020) l'a étendu aux tenseurs irréguliers binaires et incomplets. Cependant, ces deux modèles supposent que chaque pas de temps est indépendant des autres, ignorant le fait que l'état de santé des patients est fortement lié à leur historique médical. Pour résoudre ce problème, CNTF (Yin et al., 2019) exploite un RNN pour prendre en compte l'ordre des événements cliniques dans la construction des parcours des patients, tandis que TedPar (Yin et al., 2021) introduit la notion de transition temporelle d'un phénotype (journalier) à un autre pour capturer la dépendance temporelle lors de la modélisation de l'évolution des maladies chroniques. Dans ces travaux, l'aspect temporel n'est considéré que pour la construction des parcours des patients et non pour les phénotypes. En revanche, (Chambard et al., 2021) construit des phénotypes temporels comme des séquences typiques de phénotypes journaliers en utilisant un clustering a posteriori d'une décomposition tensorielle. Il se base néanmoins sur une décomposition en phénotypes journaliers.

Notre objectif est d'étendre la notion de phénotype journalier à celle de phénotype temporel, c'est-à-dire la description d'un comportement latent sur plusieurs pas de temps. Cet objectif est similaire à celui du *topic modeling* pour des documents temporels. Pour cela, TAMM (Emonet et al., 2014) a proposé un modèle graphique probabiliste conçu pour la découverte non-supervisée de modèles temporels récurrents dans les séries temporelles. Il extrait à la fois les motifs typiques, décrits sur une fenêtre temporelle, et leurs occurrences dans les documents. La principale limite de cette approche est son manque d'efficacité computationnelle. L'utilisation de technique d'échantillonnage de Gibbs la rend beaucoup moins performante que des techniques d'optimisation utilisées pour la décomposition tensorielle (Kolda et Bader, 2009).

3 Phénotypage temporel

Dans cette section, nous définissons les notations et le problème du phénotypage temporel.

Soit $\mathcal{X}$ un tenseur d'ordre 3, également considéré comme une collection de K matrices de dimensions $n \times T_k$, où K est le nombre d'individus (patients), n est le nombre de caractéristiques (événements de soins), et T_k est la durée des observations du k-ième individu.
$\boldsymbol{X}^{(k)}$ désigne la matrice du k-ième individu. Les matrices de $\mathcal{X}$ n'ont pas nécessairement la même durée, mais elles partagent le même ensemble de caractéristiques. Étant donné $R \in \mathbb{N}^*$, un nombre de phénotypes, et $\omega \in \mathbb{N}^*$, la durée des phénotypes, le phénotypage temporel vise à construire :

— $\mathcal{P} \in \mathbb{R}_+^{R \times n \times \omega}$: un tenseur d'ordre 3 représentant les R phénotypes temporels qui sont communs à tous les individus. Chaque phénotype temporel est une matrice de taille $n \times \omega$. Un phénotype représente la présence d'un événement à un moment relatif τ, $0 \leq \tau < \omega$ et ω est la même pour tous les phénotypes.

— $\mathcal{W} = \left\{ \boldsymbol{W}^{(k)} \in \mathbb{R}_+^{R \times T'_k} \right\}$: une collection de K matrices de dimension $R \times T'_k$ où $T'_k = T_k - \omega + 1$ est la durée du parcours du k-ième individu. Une valeur non-nulle à la position (r, t) dans $\boldsymbol{W}^{(k)}$ décrit le début du phénotype r au temps t pour le k-ième individu.

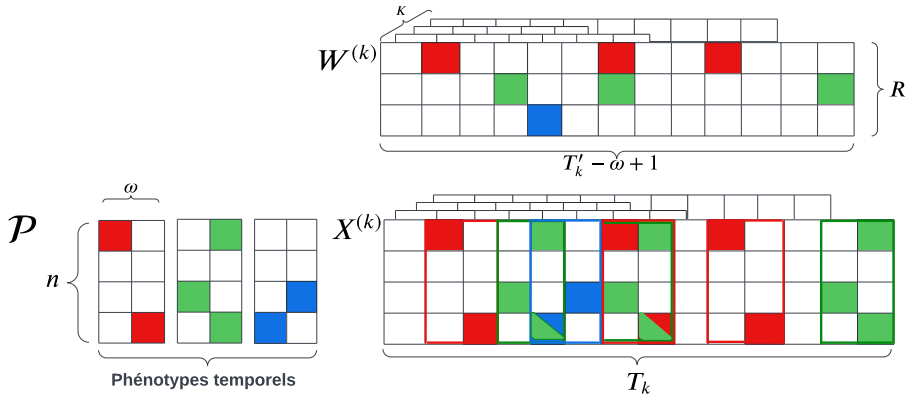

FIG. 1 – *Illustration d'une reconstruction matricielle ($\boldsymbol{X}^{(k)}$) à partir de 3 phénotypes (à gauche) et d'un parcours de soins ($\boldsymbol{W}^{(k)}$) (en haut). Chaque cellule colorée dans $\boldsymbol{W}^{(k)}$ désigne le début d'une occurrence de phénotype dans la reconstruction (entourée d'un rectangle coloré dans $\boldsymbol{X}^{(k)}$). Une cellule avec deux couleurs décrit les contributions de deux occurrences de phénotypes différents. Les mêmes phénotypes sont utilisés pour les K patients.*

Les phénotypes $\mathcal{P}$ et les parcours $\mathcal{W}$ sont définis de sorte à reconstruire avec précision le tenseur d'entrée, à savoir $\mathcal{X}$. Pour la décomposition tensorielle classique, la reconstruction est définie par un produit de matrices. Dans le cas de phénotype temporel, cette opération de reconstruction doit être redéfinie. Nous introduisons pour cela un nouvel opérateur, noté $\circledast$, qui prenne en compte la dimension temporelle de $\mathcal{P}$. On a alors $\boldsymbol{X}^{(k)} \approx \widehat{\boldsymbol{X}}^{(k)} = \mathcal{P} \circledast \boldsymbol{W}^{(k)}$ pour tout $k \in [K]$. Formellement, cet opérateur reconstruit chaque vecteur de la matrice $\widehat{\boldsymbol{X}^{(k)}}$ au temps t, noté $\widehat{\boldsymbol{x}}^{(k)}_{.,t}$, comme suit :

$$\widehat{\boldsymbol{x}}^{(k)}_{.,t} = \sum_{r=1}^{R} \sum_{\tau=1}^{\min(\omega,t-1)} \boldsymbol{w}^{(k)}_{r,t-\tau} \boldsymbol{p}^{(r)}_{\tau}. \tag{1}$$

La figure 1 illustre la reconstruction pour une matrice du tenseur d'entrée. Cette matrice est de longueur $T_k = 14$ avec $n = 4$ caractéristiques. Sa décomposition est constituée de $R = 3$ phénotypes de taille 4×2 chacun ($\omega = 2$ et $n = 4$) et d'un parcours de longueur $T'_k = 14 - 2 + 1 = 13$. Les cellules colorées contiennent des valeurs non-nulles (1 par exemple) tandis les autres cellules contiennent des 0. La figure montre que pour chaque valeur non-nulle de $\boldsymbol{W}^{(k)}$ au temps t, le phénotype correspondant est positionné au temps t dans la reconstruction. Les occurrences des phénotypes peuvent se chevaucher ou commencer à la même date. L'équation 1 formalise la reconstruction à un instant (une colonne) comme la somme du τ-ème jour des R phénotypes pondérés par la matrice $\boldsymbol{W}^{(k)}$. En considérant que la longueur des phénotypes est de ω, $\widehat{\boldsymbol{x}}^{(k)}_{.,t}$ est une combinaison de phénotypes qui se sont manifestés au plus ω unités de temps auparavant, sauf au début. Il est important de noter que la décomposition se fait pour tous les individus en même temps.

4 Découverte des phénotypes temporels

Dans cette section, nous présentons la méthode de découverte des phénotypes temporels par une approche par optimisation.

Phénotypage temporel vu comme un problème de minimisation Comme dans le cas de la décomposition tensorielle classique, le phénotypage temporel est un problème de minimisation de l'erreur entre le tenseur d'entrée et sa reconstruction. L'équation 1 détaille la reconstruction de la matrice d'un patient. Nous devons maintenant définir une mesure de cette erreur.

Le modèle que nous proposons considère la décomposition de tenseurs binaires, *i.e.* $\mathcal{X} \in \{0, 1\}$. Ceci correspond à des données qui décrivent la présence/absence d'événements. Dans ce cas, nous supposons que le tenseur d'entrée $\mathcal{X}$ suit une distribution de Bernoulli et nous utilisons la fonction de perte proposée par (Hong et al., 2020) pour les données binaires. Le problème d'optimisation résultant est alors défini comme suit :

$$\begin{aligned} \mathcal{L}^{SW} = \quad & \arg\min_{\mathcal{W},\mathcal{P}} \sum_{k=1}^{K} \sum_{t=1}^{T_k} \sum_{i=1}^{n} log(\hat{x}_{i,t}^{(k)} + 1) - x_{i,t}^{(k)} log(\hat{x}_{i,t}^{(k)}) \\ \text{subject to} \quad & \mathcal{W} \geq 0, \quad \mathcal{P} \geq 0. \end{aligned} \tag{2}$$

Contrainte de Normalisation Le problème de minimisation présenté dans l'équation 2 impose la non-négativité de $\mathcal{W}$ et $\mathcal{P}$ pour assurer une interprétation cohérente. Néanmoins, la restriction des valeurs à l'intervalle $[0, 1]$ rend les résultats plus interprétables. L'idée est d'interpréter les valeurs de $\mathcal{P}$ (resp. $\mathcal{W}$) comme la probabilité d'avoir un événement (resp. un phénotype) à un moment donné. Nous proposons donc d'ajouter une contrainte de normalisation qui impose que les valeurs de $\mathcal{P}$ et $\mathcal{W}$ se situent dans l'intervalle $[0, 1]$.

Termes de régularisation Le modèle comprend également deux termes de régularisation : la parcimonie et la non-succession de phénotypes.

L'introduction d'une régularisation sur la parcimonie des phénotypes améliore leur interprétation. Nous avons choisi la technique de régularisation L_1 pour réduire le nombre de valeurs non-nulles. Ce choix de norme s'est montré plus efficace dans nos cas d'application que celui d'autres normes.

Nous proposons également une régularisation limitant l'utilisation successive du même phénotype. La figure 2 illustre une décomposition indésirable que nous cherchons à éviter. Considérons deux phénotypes identiques le premier jour, et une matrice de patient décrivant l'apparition du second phénotype. Nous aimerions que le modèle soit capable de trouver la seconde représentation de la matrice du parcours du patient, et non de proposer une succession du premier phénotype sur trois jours consécutifs, comme l'illustre la première matrice de parcours. Le terme de régularisation proposé pénalise un modèle de reconstruction qui utilise un même phénotype sur plusieurs jours successifs sans considérer l'étendue de la fenêtre temporelle. Cette régularisation est appliquée sur les parcours des patients $\boldsymbol{W^{(k)}}$ et est définie comme suit :

$$\mathcal{S}(\boldsymbol{W}^{(k)}) = \sum_{r=1}^{R} \sum_{t=1}^{T_k} w_{r,t} \log \left(\sum_{\tau=t-\omega}^{t+\omega} w_{r,\tau} \right). \tag{3}$$

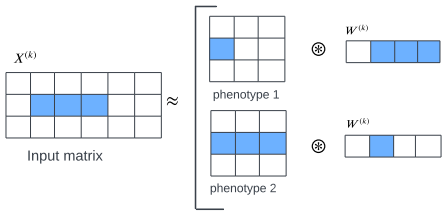

FIG. 2 – *Exemple d'une reconstruction incorrecte avec événements similaires consécutifs. Le phénotype 1 ne capture pas la séquence des événements contrairement au phénotype 2.*

La fonction logarithmique est utilisée pour empêcher ce terme de l'emporter sur les autres. Notons que cette régularisation n'a de sens qu'avec un parcours normalisé, *i.e.* $\boldsymbol{W}^{(k)} \in [0, 1]$. Sinon, la régularisation pénalise chaque présence d'un phénotype avec des poids élevés.

Finalement, la fonction de perte à optimiser est donnée par la somme pondérée de l'erreur de reconstruction, la régularisation de parcimonie et de la non-succession de phénotypes :

$$\ell = \mathcal{L}^{SW} + \alpha||\mathcal{P}||_1 + \gamma \sum_{k=1}^{K} \mathcal{S}(\boldsymbol{W}^{(k)}) \tag{4}$$

où $\mathcal{W}$ et $\mathcal{P}$ doivent satisfaire les contraintes de non-négativité et de normalisation. α et γ sont deux hyperparamètres réels positifs.

Framework d'optimisation Pour optimiser la fonction de perte ℓ, nous avons utilisé une stratégie de minimisation alternée avec une descente de gradient projeté pour gérer les contraintes de non-négativité et de normalisation. L'apprentissage étant non supervisé, l'ensemble du jeu de données est utilisé pour l'extraction des phénotypes.

5 Expérimentations et résultats

Nous avons expérimenté notre modèle sur des jeux de données synthétiques et réels afin d'évaluer sa précision de reconstruction par rapport à ses concurrents. [1] Les jeux de données synthétiques sont utilisés pour proposer des expériences reproductibles et contrôlables. De plus, comme les motifs cachés sont connus dans ce cas, cela permet d'évaluer la qualité des motifs découverts. Les résultats obtenus sont ensuite confirmés sur un jeu de données réel afin de démontrer leur fiabilité.

Données synthétiques La génération de données synthétiques est basée sur le processus inverse de la décomposition. 1) Un tenseur de phénotypes du troisième ordre $\mathcal{P}$ est généré en tirant aléatoirement un sous-ensemble d'événements médicaux pour chaque instant de la fenêtre temporelle de chaque phénotype. 2) Les parcours des patients $\mathcal{W}$ sont générés en tirant aléatoirement les jours d'occurrence de chaque phénotype tout au long du séjour du patient

1. Code disponible ici : `https://gitlab.inria.fr/hsebia/swotted`

avec comme contrainte : le même phénotype ne peut pas se produire plusieurs jours consécutifs. Pour 1) et 2), nous utilisons des distributions de Bernoulli avec $p = 0.5$. 3) Les matrices de patients de $\mathcal{X}$ sont ensuite calculées en utilisant la formule de reconstruction proposée dans l'Eq. 1. Cette reconstruction peut conduire à des valeurs supérieures à 1 lorsqu'on accumule plusieurs occurrences de phénotypes qui partagent le même événement. Nous binarisons donc le tenseur résultant en projetant les valeurs non nulles à 1. Les caractéristiques par défaut des jeux de données synthétiques générés par la suite dans les différentes expérimentations sont les suivantes : $K = 100$ patients, $n = 20$ événements de soins, $R = 4$ phénotypes de longueur $\omega = 3$ et des séjours de $T_k = 6$ jours.

Concurrents Nous comparons les performances de notre méthode à celles de trois modèles récents : **LogPar** (Yin et al., 2020), une version logistique de PARAFAC2 pour la décomposition de tenseurs binaires supposés suivre la distribution de Bernoulli ; **CNTF** (Yin et al., 2019), un modèle de décomposition de tenseurs ayant une dimension temporelle variable, qui suppose que le tenseur d'entrée suit une distribution de Poisson mais qui a également montré son efficacité sur des données binaires ; et **SWIFT** (Afshar et al., 2021), un modèle de décomposition minimisant la distance de Wasserstein entre le tenseur d'entrée et sa reconstruction. Pour chaque expérience, nous configurons manuellement leurs hyperparamètres afin d'assurer une comparaison équitable. L'implémentation de TedPar n'étant pas disponible, il n'a pu être testé.

Mesure de précision Nous utilisons la $FIT \in (-\infty, 1]$ (Bro et al., 1999) pour mesurer la qualité de la reconstruction de notre modèle. Plus la valeur de FIT est élevée, meilleure est la reconstruction.

$$FIT_X = 1 - \frac{\sum_{k=1}^{K} ||\boldsymbol{X}^{(k)} - \widehat{\boldsymbol{X}}^{(k)}||_F}{\sum_{k=1}^{K} ||\boldsymbol{X}^{(k)}||_F} \tag{5}$$

où le tenseur original $\mathcal{X}$ sert de vérité terrain, le tenseur résultant est noté $\widehat{\mathcal{X}}$ et $|| \cdot ||_F$ est la norme de Frobenius. La mesure FIT est également utilisée pour comparer les phénotypes et les parcours des patients lorsque les motifs cachés sont connus a priori. Ainsi, FIT_P (resp. FIT_W) désigne la qualité de reconstruction de $\mathcal{P}$ (resp. $\mathcal{W}$).

Implémentation Le modèle est implémenté à l'aide de *PyTorch*. Nous avons entraîné le modèle avec un optimiseur *Adam* pour la mise à jour de $\mathcal{P}$ et $\mathcal{W}$. Le taux d'apprentissage est fixé à 10^{-3}. Nous avons ajusté les hyperparamètres α et γ en testant différentes valeurs et en sélectionnant celles qui donnent la meilleure mesure de reconstruction. Leurs valeurs par défaut sont $\alpha = 0.5$ et $\gamma = 0.5$.

5.1 Étude des termes de la fonction de perte

Cette expérimentation compare différentes versions régularisées du modèle. L'objectif est de montrer que tous les termes de la fonction de perte sont importants pour obtenir de bons résultats. Nous évaluons trois versions de notre modèle avec différents termes : Sp pour la régularisation de parcimonie seule, Sp+Nr pour la parcimonie avec la contrainte de normalisation et Sp+Nr+PS pour la version incluant également la régularisation de non-succession de phénotypes. La parcimonie est toujours prise en compte pour garantir une interprétation

	FIT_X	FIT_P	FIT_W
Sp	0.66 ± 0.08	0.47 ± 0.29	0.48 ± 0.14
Sp+Nr	0.69 ± 0.07	0.59 ± 0.18	0.53 ± 0.11
Sp+Nr+PS	$\mathbf{0.71 \pm 0.07}$	$\mathbf{0.61 \pm 0.29}$	$\mathbf{0.56 \pm 0.18}$

TAB. 1 – *Valeurs moyennes et écarts types de FIT_X, FIT_P et FIT_W pour différentes versions régularisées du modèle appliquées à des jeux de données synthétiques.*

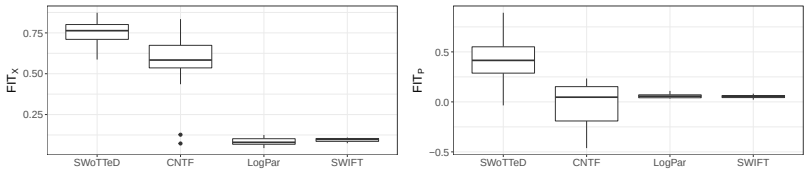

FIG. 3 – *Valeurs FIT du modèle et ses concurrents sur des données synthétiques avec $\omega = 1$. SWoTTeD désigne le modèle présenté dans cet article.*

des phénotypes. Comme la régularisation de non-succession n'a de sens que pour les $\mathcal{W}$ normalisés, la combinaison Sp+PS n'est pas évaluée. Chaque version est exécutée sur 20 jeux synthétiques et les valeurs FIT sont collectées.

Le tableau 1 présente les résultats de cette expérimentation. Nous observons que la moyenne de FIT_X varie entre 0.66 et 0.71 pour toutes les versions. De plus, les valeurs moyennes de FIT_P et FIT_W sont comprises entre 0.47 et 0.61. Ces valeurs, proches de 1, signifient que les reconstructions des trois tenseurs sont précises sur les données synthétiques. Nous concluons à partir des valeurs de FIT_P que les trois versions ont la capacité de découvrir précisément les phénotypes cachés. Nous observons également que la valeur moyenne de FIT augmente lorsqu'on ajoute la normalisation et la régularisation de la non-succession des phénotypes. La version Sp+Nr+PS est meilleure que les autres. Cependant, la différence n'est pas statistiquement significative selon le test pairé de Wilcoxon.

Nous concluons que tous les termes sont nécessaires pour obtenir les meilleurs résultats. De plus, l'ajout de la régularisation de non-succession désambiguïse la situation illustrée dans la figure 2 et aide le modèle à reconstruire correctement les variables latentes.

5.2 Précision des phénotypes découverts

L'expérimentation compare la précision du modèle à celle de ses concurrents sur des données synthétiques générées avec des phénotypes journaliers ($\omega = 1$). L'objectif est d'évaluer sa capacité à extraire les motifs cachés par rapport aux modèles récents de l'état de l'art.

Les résultats, résumés dans la figure 3 montrent que notre modèle obtient les meilleures performances en termes de mesures FIT_X et FIT_P. Selon le test de Wilcoxon, cette différence est significative. Le modèle parvient même à atteindre une reconstruction parfaite sur 3 jeux de données. Ces bonnes performances s'expliquent premièrement par la flexibilité de la reconstruction permettant le chevauchement de différents phénotypes et leur arrivée avec un décalage, et deuxièmement par l'utilisation d'une fonction de perte qui suppose une distribution de Bernoulli adaptée aux données binaires.

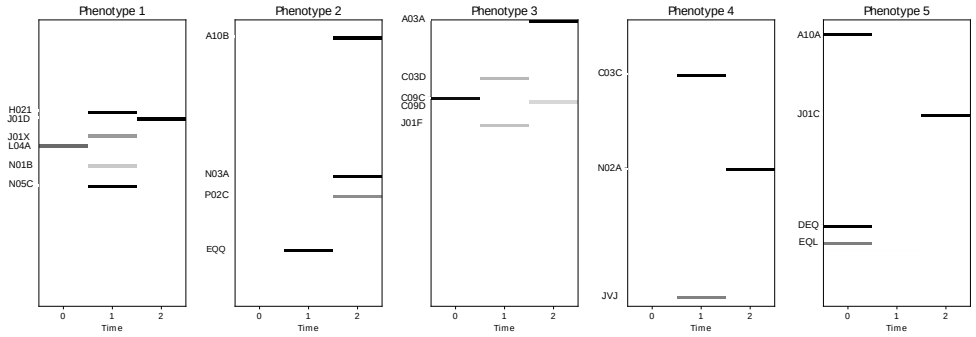

FIG. 4 – *Cinq phénotypes découverts pour la 4ème vague épidémique. Chaque cellule grise représente la présence d'un médicament à un instant relatif. Plus la cellule est foncée, plus la valeur est élevée. Les valeurs des cellules sont comprises dans l'intervalle* $[0, 1]$.

6 Application à l'analyse de parcours de patients COVID-19

L'objectif de cette étude de cas est de décrire les parcours typiques des patients qui ont été admis dans une unité de soins intensifs lors des premières vagues de COVID-19 en région parisienne. Ces parcours typiques sont représentatifs des protocoles de traitement qui ont été effectivement mis en œuvre. Leur description peut aider les hôpitaux à mieux appréhender leur gestion des traitements en période de crise. Dans le cadre de COVID-19, nous savons que les cas les plus critiques sont les patients présentant des comorbidités (diabète, hypertension, etc.). Cela complique l'analyse des parcours de soins de ces patients car ils cumulent plusieurs traitements indépendants. Dans une telle situation, les outils développés pour l'analyse des parcours sont utiles pour démêler les différents traitements qui ont été délivrés. Les parcours de soins des patients COVID-19 ont été obtenus à partir de l'entrepôt de données de l'Assistance Publique – Hôpitaux de Paris. Nous créons un jeu de données par chacune des 4 premières vagues épidémiques de COVID-19. Les périodes de ces vagues sont celles définies officiellement par le gouvernement. Les patients sélectionnés pour cette étude sont des adultes (plus de 18 ans) ayant un test PCR positif. Pour chaque patient, nous créons une matrice binaire qui représente les événements des soins du patient (délivrance de médicaments et procédures) au cours des 10 premiers jours de son séjour dans l'unité de soins intensifs. Les épidémiologistes ont sélectionné 85 types d'événements de soins (58 types de médicaments et 27 types de procédures) en fonction de leur fréquence et de leur pertinence pour la COVID-19. Les médicaments sont codés en utilisant le troisième niveau des codes ATC et les procédures en utilisant le troisième niveau des codes CCAM. [2]

Nous présentons maintenant les résultats obtenus pour la quatrième vague (du 05/07/2021 au 06/09/2021) qui contient $2\,593$ patients et $21\,325$ événements de soins. Nous exécutons le modèle pour extraire $R = 8$ phénotypes de longueur $\omega = 3$ avec comme paramètres $1\,000$ epochs et un taux d'apprentissage de 10^{-3}. La figure 4 illustre cinq des huit phénotypes extraits de la quatrième vague. Une première observation est que ces phénotypes contiennent peu

2. L'ATC (Anatomical, Therapeutical and Chemical) est une classification standard des médicaments. La CCAM est la classification française des procédures médicales.

Code	Description	Jours		
H02A	*Prednisone, antibiotique*	0.00	1.00	0.00
J01D	*Cefotaxime, antibiotique*	0.00	0.00	1.00
J01X	*Metronidazole*	0.00	0.40	0.00
L04A	*Tocilizumab*	0.59	0.00	0.00
N01B	*Lidocaine*	0.00	0.21	0.00
N05C	*Midazolam, sédation*	0.00	1.00	0.00

TAB. 2 – *Phénotype 1 : ventilation mécanique après le traitement à la COVID-19.*

Code	Description	Jours		
A10A	*Insuline*	1.00	0.00	0.00
J01C	*Amoxicilline*	0.00	0.00	1.00
DEQ	*Électrocardiogramme*	1.00	0.00	0.00
EQL	*Dopamine*	0.50	0.00	0.00

TAB. 3 – *Phénotype 5 : choc septique sévère.*

d'éléments non-nuls ce qui les rend presque faciles à interpréter. Deuxièmement, chaque phénotype décrit la présence d'événements de soins a moins deux instants différents, ce qui souligne l'importance de leur dimension temporelle. Ces phénotypes ont été montrés à un clinicien pour interprétation. Il a été confirmé que ces derniers révèlent des combinaisons pertinentes de soins. Deux types différents de combinaisons ont été identifiés : certaines combinaisons de soins esquissent le contexte pathologique des patients (hypertension, insuffisance hépatique, etc.) tandis que d'autres sont représentatives des protocoles de traitement. Nous détaillons un phénotype de chaque type dans les tableaux 2 et 3. Chaque ligne du tableau correspond à un événement de soins qui a au moins une valeur non nulle dans le phénotype. La première colonne donne le code (ATC ou CCAM) de l'événement, et la deuxième colonne sa description. Les autres colonnes détaillent sa présence au cours du temps (jours). Le tableau 3 illustre un phénotype temporel qui a été interprété comme un protocole typique de COVID-19. En effet, le *Tocilizumab* est devenu un médicament standard pour aider les patients souffrant de problèmes respiratoires aigus à éviter le recours à la ventilation mécanique. Dans ce phénotype, les cliniciens détectent une transition de l'administration prophylactique de *Tocilizumab* (le premier jour) à une ventilation mécanique identifiée par l'utilisation de médicaments sédatifs typiques (*Lidocaïne*, *Métronidazole* et *Midazolam*). Ce changement, incluant l'arrêt du traitement par *Tocilizumab*, est un protocole typique. Néanmoins, des investigations complémentaires sont nécessaires pour expliquer la présence d'antibiotiques. Le tableau 2 illustre le phénotype temporel d'un choc septique sévère : un patient dans cette situation sera monitoré, on lui administrera de la *dopamine* pour induire une activité cardiaque et on lui injectera de l'*insuline* pour gérer sa glycémie. Ce protocole est couramment rencontré dans les unités de soins intensifs et a été appliqué pour les patients COVID-19 en état critique.

En conclusion, les phénotypes détaillés précédemment illustrent le fait que notre méthode démêle les protocoles génériques des soins intensifs et des traitements spécifiques de la COVID-19. D'autres phénotypes ont également été facilement identifiés par les cliniciens comme correspondant au traitement de patients ayant des antécédents médicaux spécifiques. Leur conclusion générale est que notre modèle extrait des phénotypes pertinents qui décrivent de véritables pratiques.

7 Conclusion

Les méthodes de décomposition tensorielle les plus récentes se limitent à l'extraction de phénotypes qui ne décrivent qu'une combinaison de caractéristiques survenant un même jour. Dans cet article, nous avons proposé une méthode de décomposition tensorielle dédiée à l'extraction de phénotypes temporels. Elle a été testée sur des jeux de données synthétiques et réels. Les résultats montrent qu'elle est plus performante que les techniques de décomposition de l'état de l'art :les phénotypes extraient sont plus expressifs et permettent une reconstruction plus précise des tenseurs d'entrée. Une étude de cas sur les patients COVID-19 illustre l'efficacité du modèle pour extraire des phénotypes temporels significatifs et la pertinence de la dimension temporelle pour décrire des protocoles de soins typiques. Ces résultats prometteurs ouvrent de nouvelles voies de recherche en apprentissage automatique, en phénotypage temporel et en analyse de parcours de soins. Pour les travaux futurs, nous prévoyons d'étendre le modèle pour extraire des phénotypes temporels décrits sur des fenêtres de taille variable.

Remerciements Une partie des recherches présentées dans cet article est subventionnée par la Fondation de l'AP-HP, dans le cadre de la Chaire AI-RACLES et a reçu l'accord du Comité scientifique et éthique du CDW de l'AP-HP (CSE-20-11-COVIPREDS).

Références

Afshar, A., I. Perros, E. E. Papalexakis, E. Searles, J. Ho, et J. Sun (2018). COPA : Constrained PARAFAC2 for sparse and large datasets. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pp. 793–802.

Afshar, A., I. Perros, H. Park, C. R. deFilippi, X. Yan, W. F. Stewart, J. Ho, et J. Sun (2020). TASTE : temporal and static tensor factorization for phenotyping electronic health records. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pp. 193–203.

Afshar, A., K. Yin, S. Yan, C. Qian, J. C. Ho, H. Park, et J. Sun (2021). SWIFT : Scalable wasserstein factorization for sparse nonnegative tensors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6548–6556.

Anandkumar, A., R. Ge, D. Hsu, S. M. Kakade, et M. Telgarsky (2014). Tensor decompositions for learning latent variable models. *Journal of machine learning research 15*, 2773–2832.

Bro, R., C. Andersson, et H. Kiers (1999). PARAFAC2 – Part II. modeling chromatographic data with retention time shifts. *Journal of Chemometrics 13*(3-4), 295–309.

Chambard, M., T. Guyet, Y.-L. NGuyen, et E. Audureau (2021). Temporal phenotyping for characterisation of hospital care pathways of COVID-19 patients. In *Proceedings of the Workshop on Advanced Analytics and Learning on Temporal Data (AALTD)*, pp. 55–70.

Emonet, R., J. Varadarajan, et J.-M. Odobez (2014). Temporal analysis of motif mixtures using dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*(1), 140–156.

Fanaee-T, H. et J. Gama (2016). Tensor-based anomaly detection : An interdisciplinary survey. *Knowledge-Based Systems 98*, 130–147.

Henderson, J., H. He, B. A. Malin, J. C. Denny, A. N. Kho, J. Ghosh, et J. C. Ho (2018). Phenotyping through semi-supervised tensor factorization (PSST). In *Proceedings of the Annual Symposium of AMIA*, pp. 564–573.

Hong, D., T. G. Kolda, et J. A. Duersch (2020). Generalized canonical polyadic tensor decomposition. *SIAM Review 62*(1), 133–163.

Kiers, H. A., J. M. Ten Berge, et R. Bro (1999). PARAFAC2–part I. A direct fitting algorithm for the PARAFAC2 model. *Journal of Chemometrics : A Journal of the Chemometrics Society 13*(3-4), 275–294.

Kolda, T. G. et B. W. Bader (2009). Tensor decompositions and applications. *SIAM review 51*(3), 455–500.

Perros, I., E. E. Papalexakis, F. Wang, R. Vuduc, E. Searles, M. Thompson, et J. Sun (2017). SPARTan : Scalable PARAFAC2 for large and sparse data. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 375–384.

Wang, Y., R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, et J. Sun (2015). Rubik : Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 1265–1274.

Yang, K., X. Li, H. Liu, J. Mei, G. Xie, J. Zhao, B. Xie, et F. Wang (2017). TaGiTeD : Predictive task guided tensor decomposition for representation learning from electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2824–2830.

Yin, K., A. Afshar, J. C. Ho, W. K. Cheung, C. Zhang, et J. Sun (2020). LogPar : Logistic PARAFAC2 factorization for temporal binary data with missing values. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 1625–1635.

Yin, K., W. K. Cheung, B. C. Fung, et J. Poon (2021). TedPar : Temporally dependent PARAFAC2 factorization for phenotype-based disease progression modeling. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pp. 594–602.

Yin, K., D. Qian, W. K. Cheung, B. C. M. Fung, et J. Poon (2019). Learning phenotypes and dynamic patient representations via RNN regularized collective non-negative tensor factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1246–1253.

Summary

Tensor decomposition has recently been gaining attention in the machine learning community due to its versatility in processing large-scale data. In particular, it has become popular for the analysis of Electronic Health Records (EHR). However, this task becomes significantly more difficult when the data follows complex temporal patterns. This paper introduces the notion of a temporal phenotype as an arrangement of features over time. We propose a novel model integrating several constraints and regularizations to discover interpretable hidden temporal patterns. We validate our proposal using both synthetic and real patient data from the Greater Paris University Hospital. The results show that this technique outperforms the recent state-of-the-art tensor decomposition models.