Une extension de la décomposition tensorielle au phénotypage temporel

H. Sebia[1], T. Guyet[1], E. Audureau[2]

[1] Inria, AIStroSight, Centre de Lyon, France
[2] AP-HP, Hôpital Henri Mondor, Université Paris Est Créteil, France

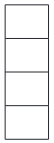
19/01/2023

AIstroSight Inria

Introduction

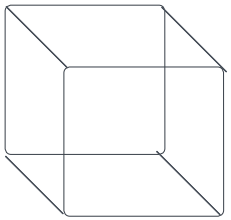
Représentation de Tenseur

- Un tenseur est la représentation naturelle de **données multidimensionnelles**
- Par exemple, les **mesures spatio-temporelles** peuvent être structurées en un cube de données appelé un tenseur d'ordre trois



First-order tensor

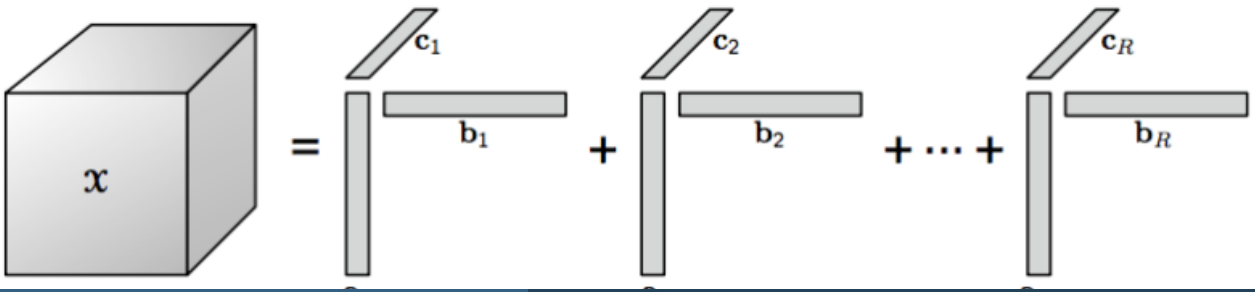
Second-order tensor



Third-order tensor

Décomposition Tensorielle

- Une généralisation de l'ACP et la SVD aux tenseurs d'ordre supérieur
- Une **technique non supervisée** pour l'analyse des variables cachées du modèle



Champs d'Application

Domain Ref.	Typical tensor	Application
Network security	OriginIP × DestIP × Time	Abnormal traffic discovery [STF06]
Social networks	Person × Person × Time	Event detection [KPF12]
Neuroscience	Frequency × Channel × Time	Seizure recognition [ABB+07]
Sensors	Measures × Location × Time	Anomaly detection [SPP06]
Transportation	Origin × Destination × Time	Detection of urban traffic problems [WGC+14]
Medical diagnosis	Patient × Medication × Time	Description of patient pathways [HGS+14]
Epidemiology	Space × Time × Indicators	Disease outbreak prediction [FTG15]
Seismology	Location × Time × Frequency	Predicting earthquake ground motion [BTCC13]

Dimension temporelle

- Souvent présente dans les tenseurs typiques

⇒ Le besoin de développer de modèles qui exploitent au mieux ce type de données

Champs d'Application

Domain Ref.	Typical tensor	Application
Network security	OriginIP × DestIP × Time	Abnormal traffic discovery [STF06]
Social networks	Person × Person × Time	Event detection [KPF12]
Neuroscience	Frequency × Channel × Time	Seizure recognition [ABB+07]
Sensors	Measures × Location × Time	Anomaly detection [SPP06]
Transportation	Origin × Destination × Time	Detection of urban traffic problems [WGC+14]
Medical diagnosis	Patient × Medication × Time	Description of patient pathways [HGS+14]
Epidemiology	Space × Time × Indicators	Disease outbreak prediction [FTG15]
Seismology	Location × Time × Frequency	Predicting earthquake ground motion [BTCC13]

Dimension temporelle

- Souvent présente dans les tenseurs typiques

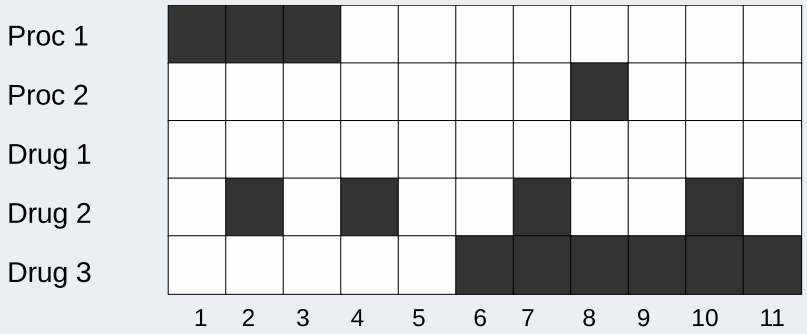
⇒ Le besoin de développer de modèles qui exploitent au mieux ce type de données

Parcours de patients à partir des données du DSE

Caractériser les patients à partir des dossiers de santé électroniques (DSE)

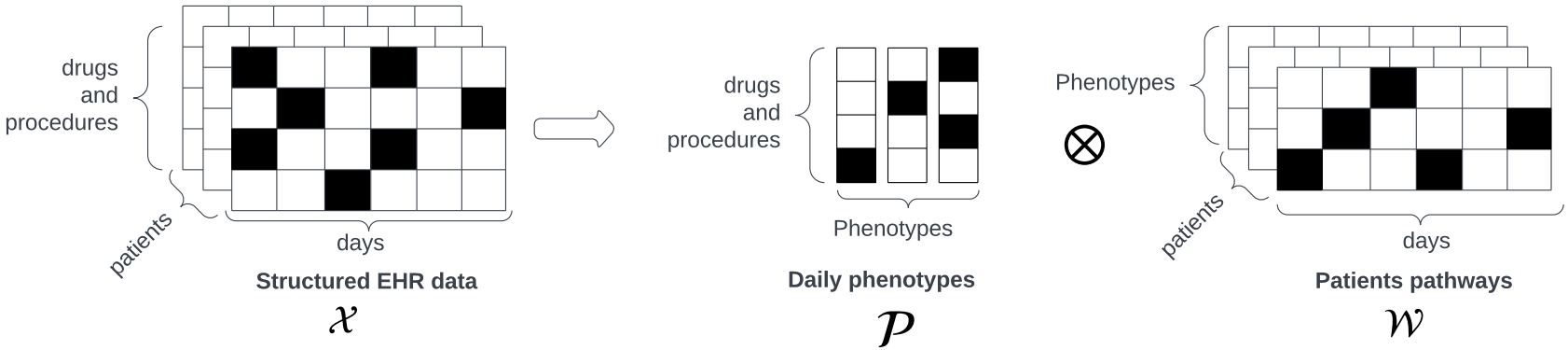
- Le DSE contient des données structurées sur les patients
  - **Procédures horodatées**
  - **Médicaments horodatés**
  - Caractéristiques des patients (âge, sexe, IMC, traitements, etc.)
  - Résultats des tests de laboratoire

Une représentation matricielle du séjour hospitalier d'un patient



→ $x_{i,j} = 1$ (carré coloré) si le i-ième événement (ligne) se produit au jour j (colonne)

Parcours de patients à partir des données du DSE

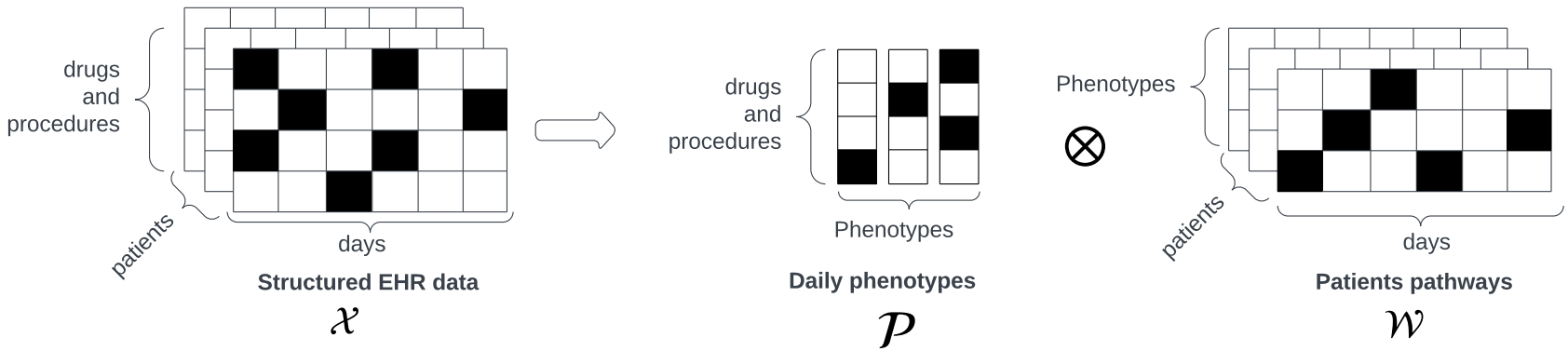


Décomposition de Tenseur

$\Rightarrow$ Problème de minimisation de $||\mathcal{X} - \mathcal{P} \otimes \mathcal{W}||_F$

- **Phénotypes** $\mathcal{P}$ **:** profils typiques de traitements
- **Parcours des patients** $\mathcal{W}$ **:** une série temporelle de phénotypes

Parcours de patients à partir des données du DSE



Décomposition de Tenseur

$\Rightarrow$ Problème de minimisation de $||\mathcal{X} - \mathcal{P} \otimes \mathcal{W}||_F$

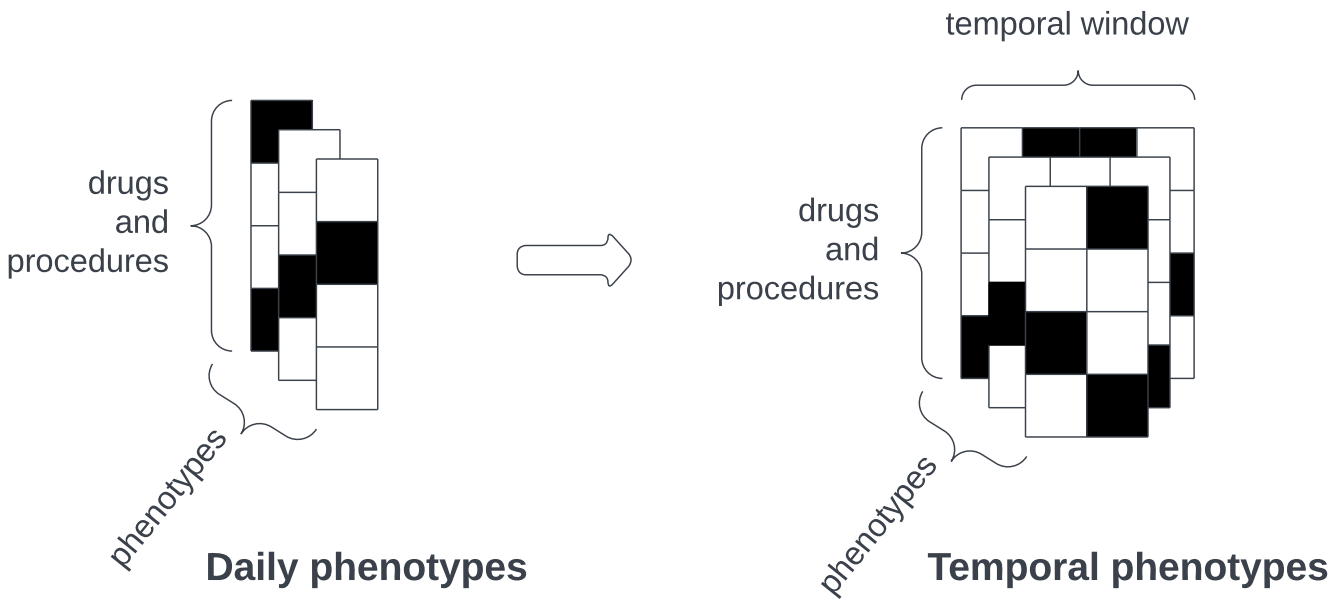
- **Phénotypes $\mathcal{P}$:** profils typiques de traitements
- **Parcours des patients $\mathcal{W}$:** une série temporelle de phénotypes

Limite

$\Rightarrow$ un phénotype décrit les procédures typiques et les médicaments administrés au cours d'**un jour** donnée

$\Rightarrow$ Les traitements sont mieux décrits comme des combinaisons de **séquences de traitements**

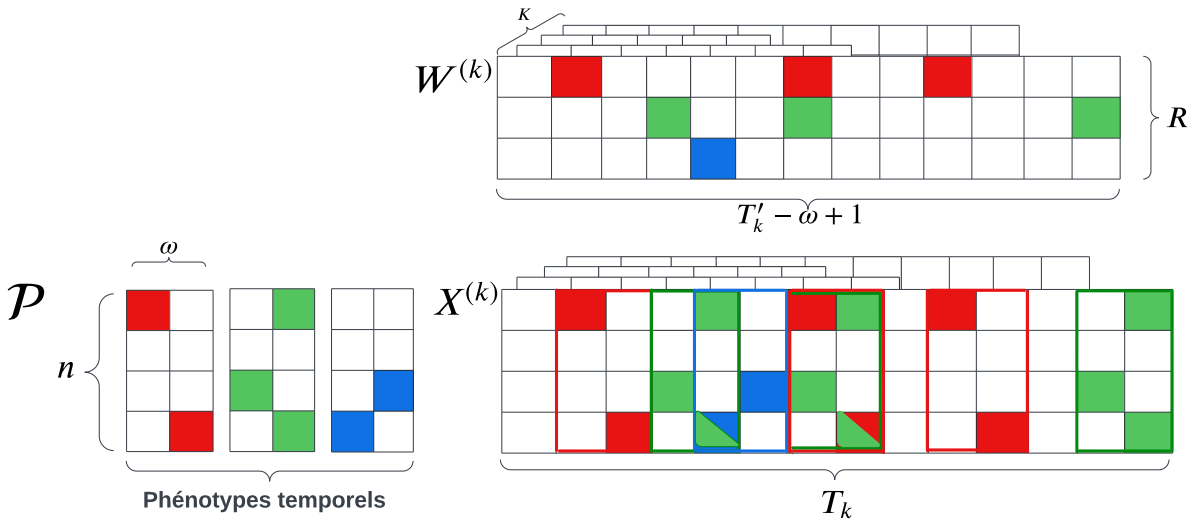
Extension au Phénotypage Temporel



- **Phénotype Temporel** décrit une disposition temporelle des événements médicaux
  - → plus expressif
  - → description plus précise d'un traitement

→ Nécessite de redéfinir la reconstruction

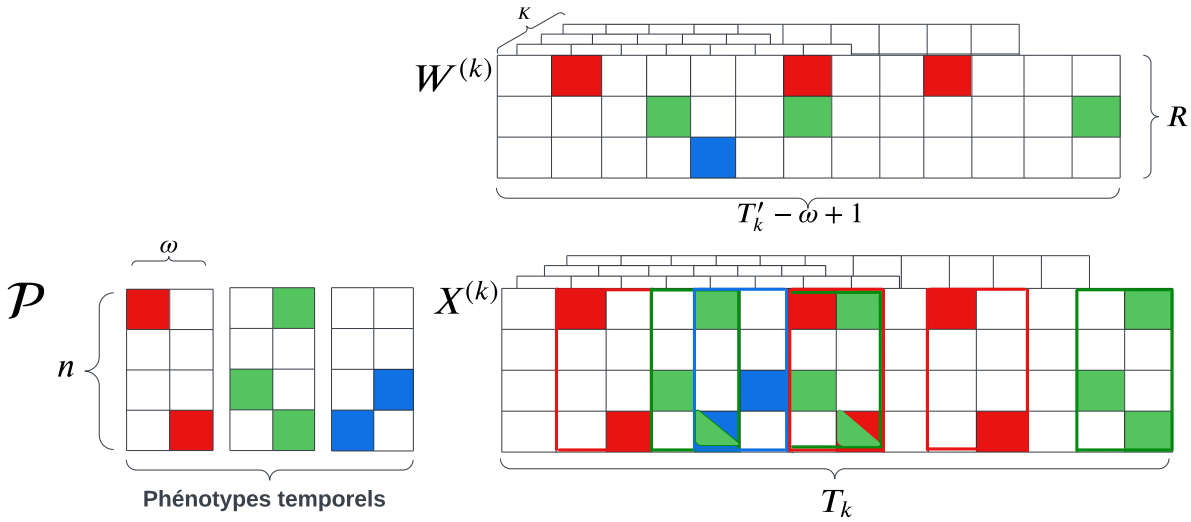
SWoTTeD: Reconstruction



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

- K : nombre des patients
- n : nombre des évenements médicaux
- T_k : la longueur du séjour du patient k
- ω : la taille de la fenêtre temporelle
- R : le nombre de phénotypes

SWoTTeD: Evaluation de la construction



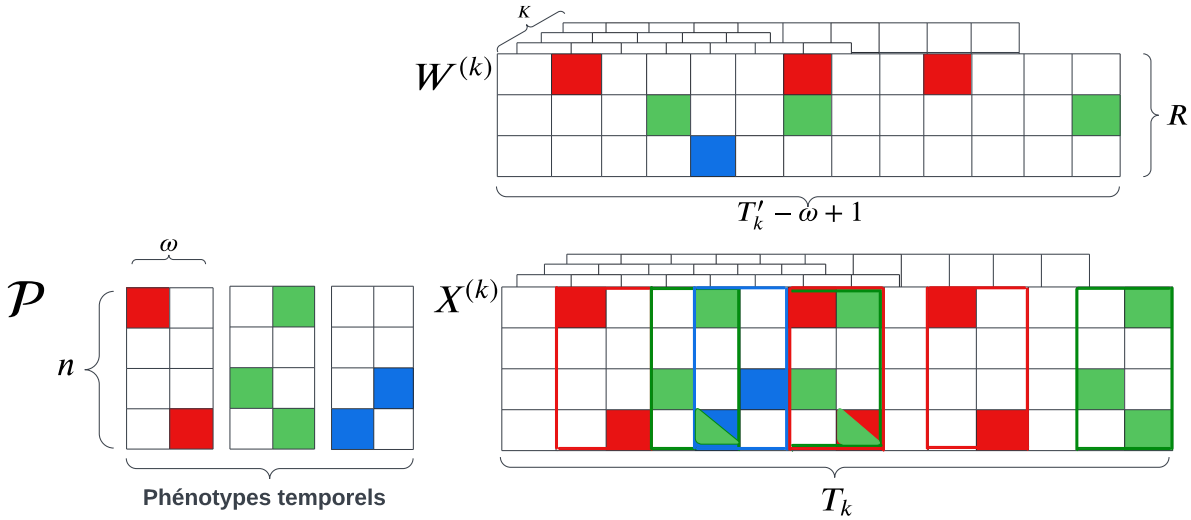
Erreur de reconstruction avec hypothèse de Bernoulli [HKD20]

$$\mathcal{L}^{SW} = \arg\min_{\mathcal{W},\mathcal{P}} \sum_{k=1}^{K} \sum_{t=1}^{T_k} \sum_{i=1}^{n} log(\hat{x}_{i,t}^{(k)} + 1) - x_{i,t}^{(k)} log(\hat{x}_{i,t}^{(k)})$$

$$\text{avec} \qquad \mathcal{W} \geqslant 0, \quad \mathcal{P} \geqslant 0.$$

- K : nombre des patients
- n : nombre des évenements médicaux
- T_k : la longueur du séjour du patient k
- ω : la taille de la fenêtre temporelle
- R : le nombre de phénotypes

SWoTTeD: Fonction de perte



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

où $\boldsymbol{W}^{(p)}$ et $\boldsymbol{P}^{(r)}$, pour tout p and tout r, doivent satisfaire :

- la contrainte de non-négativité
- la contrainte de normalisation

→ régularisation de de la parcimonie sur $\mathcal{P}$

→ régularisation de non-succession de phénotypes sur $\mathcal{W}$

Experimentations et Resultats

Plan d'expérimentation

- **Données synthétiques/réelles**
  - → Valider le modèle
  - → Caractériser des séjours des patients COVID-19 qui ont été admis en unités de soins intensifs
- **Compétiteurs**
  - → CNTF [YQC+19], LogPar [YAH+20], SWIFT [AYY+21]

Mesure de la qualité de reconstruction

- Nous utilisons la mesure $FIT \in (-\inf, 1]$. Plus la valeur est élevée, meilleure est la reconstruction:

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

- FIT_P (resp. FIT_W) désigne la qualité de reconstruction de $\mathcal{P}$ (resp.$\mathcal{W}$)

Expérimentations sur les données synthétiques

Génération de données (processus inverse de la décomposition)

1. Génération de $\mathcal{P}$ en tirant aléatoirement un sous-ensemble d'événements médicaux pour chaque instant de la fenêtre temporelle.
2. Génération de $\mathcal{W}$ en tirant aléatoirement les jours d'occurrence de chaque phénotype tout au long du séjour du patient
3. Génération de $\mathcal{X}$ en utilisant la formule de reconstruction proposée

Caractéristiques par défaut

- Nombre des patients K: 100
- Nombre d'événements médicaux n: 20
- Longueur de séjour des patients T_k: 6
- Nombre de phénotypes R: 4
- Longueur de la fenêtre temporelle ω:3

Qualité de la fonction de perte

Objectif: étudier l'importance des termes de la fonction de perte

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

	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	$\mathbf{0.71 \pm 0.07}$	$\mathbf{0.61 \pm 0.29}$	$\mathbf{0.56 \pm 0.18}$

Table: 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.

- La version Sp+Nr+PS donne les valeurs les plus élevées de FIT.

→ Tous les termes de la fonction de perte sont importants pour avoir les meilleurs résultats

Précision des phénotypes découverts

Objectif: évaluer la capacité de SWoTTeD à extraire les motifs cachés par rapport aux modèles récents de l'état de l'art

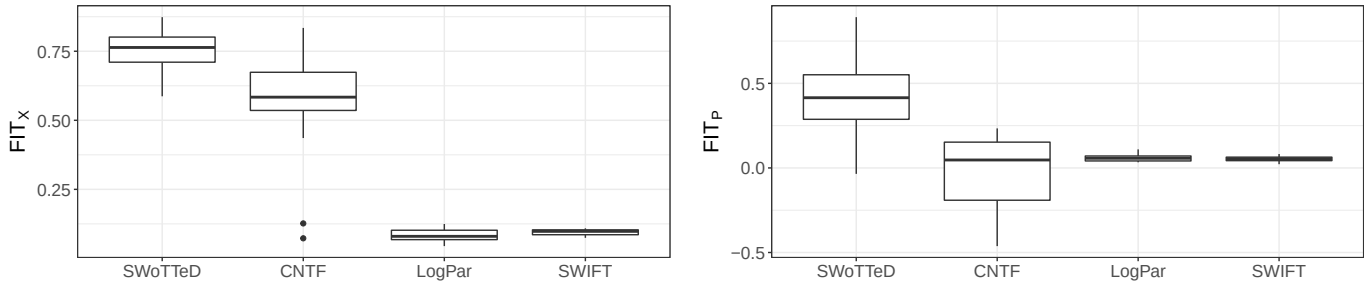


Figure: Valeurs *FIT* du modèle et ses concurrents sur des données synthétiques avec $\omega = 1$.

- SWoTTeD obtient les meilleures performances en termes de mesures FIT_X et FIT_P

→ la différence est significative selon le test de Wilcoxon.

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

- **Objectif:** décrire les phénotypes des patients lors des premières vagues de COVID-19
- **Constitution de la cohorte:**
  - Données provenant de l'Assistance Publique – Hôpitaux de Paris
  - Un jeu de données par vague épidémique de COVID-19
  - Les patients adultes avec en moins un test PCR Positif
  - Sélection de 58 types de médicaments et 27 types de procédures
  - Considération des 10 premiers jours en unité de soins intensifs

→ Les résultats pour la quatrième vague (du 05/07/2021 au 06/09/2021) : extraction de 8 phénotypes de longueur $\omega = 3$.

Résultats de la 4ème vague de COVID-19

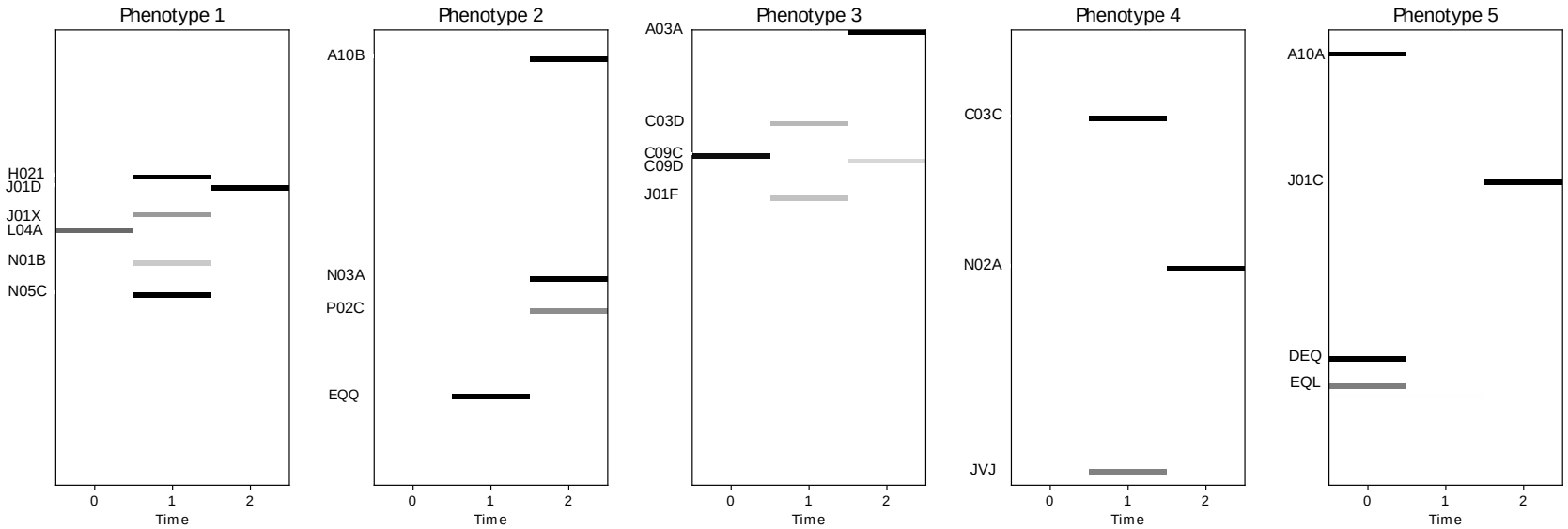


Figure: Cinq phénotypes découverts pour la 4ème vague épidémique

- Les phénotypes sont **épars** et sont décrits sur au moins **deux instants différents**
- Deux types de combinaisons identifiés : certaines esquissent le **contexte pathologique des patients** (hypertension, insuffisance hépatique, etc.) tandis que d'autres sont représentatives des **protocoles de traitement** selon les **cliniciens**

Résumé

Contribution

- Extension de la décomposition tensorielle au **phénotypage temporel**
- La proposition de SWoTTeD
- L'intégration de **contraintes** et de **régularisation** pour améliorer l'intérprétation des phénotypes

Validation

- Reconstruction **correcte** des données synthétiques
- **Meilleure qualité de reconstruction** comparé à des modèles récents de l'état de l'art

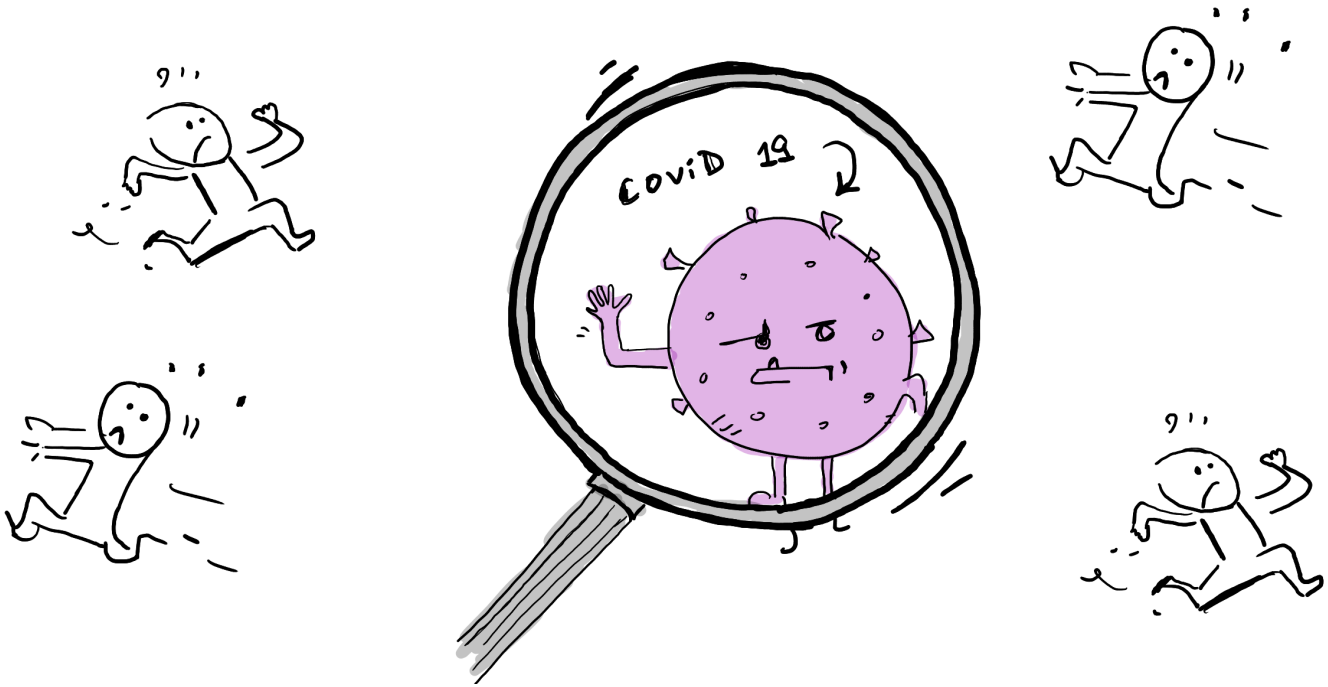
Application

- Étude sur des données des patients COVID-19

→ Les phénotypes extraits décrivent de **véritables pratiques** selon les **cliniciens**

→ SWoTTeD démêle les **protocoles génériques** des soins intensifs et des **traitements spécifiques** de la COVID-19.

Merci pour votre attention !



References I

Evrim Acar, Canan Aykut Bingol, Haluk Bingol, Rasmus Bro, and Bulent Yener, *Seizure recognition on epilepsy feature tensor*, 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2007, pp. 4273–4276.

Ardavan Afshar, Kejing Yin, Sherry Yan, Cheng Qian, Joyce C Ho, Haesun Park, and Jimeng Sun, *SWIFT: Scalable wasserstein factorization for sparse nonnegative tensors*, Proceedings of the AAAI conference, 2021.

Yang Bai, Jale Tezcan, Qiang Cheng, and Jie Cheng, *A multiway model for predicting earthquake ground motion*, 2013 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE, 2013, pp. 219–224.

Hadi Fanaee-T and Joao Gama, *Eigenevent: an algorithm for event detection from complex data streams in syndromic surveillance*, Intelligent Data Analysis 19 (2015), no. 3, 597–616.

Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun, *Limestone: High-throughput candidate phenotype generation via tensor factorization*, Journal of biomedical informatics 52 (2014), 199–211.

David Hong, Tamara G. Kolda, and Jed A. Duersch, *Generalized canonical polyadic tensor decomposition*, SIAM Review 62 (2020), no. 1, 133–163.

Danai Koutra, Evangelos E Papalexakis, and Christos Faloutsos, *Tensorsplat: Spotting latent anomalies in time*, 2012 16th Panhellenic Conference on Informatics, IEEE, 2012, pp. 144–149.

Jimeng Sun, Spiros Papadimitriou, and S Yu Philip, *Window-based tensor analysis on high-dimensional and multi-aspect streams*, Sixth International Conference on Data Mining (ICDM'06), IEEE, 2006, pp. 1076–1080.

References II

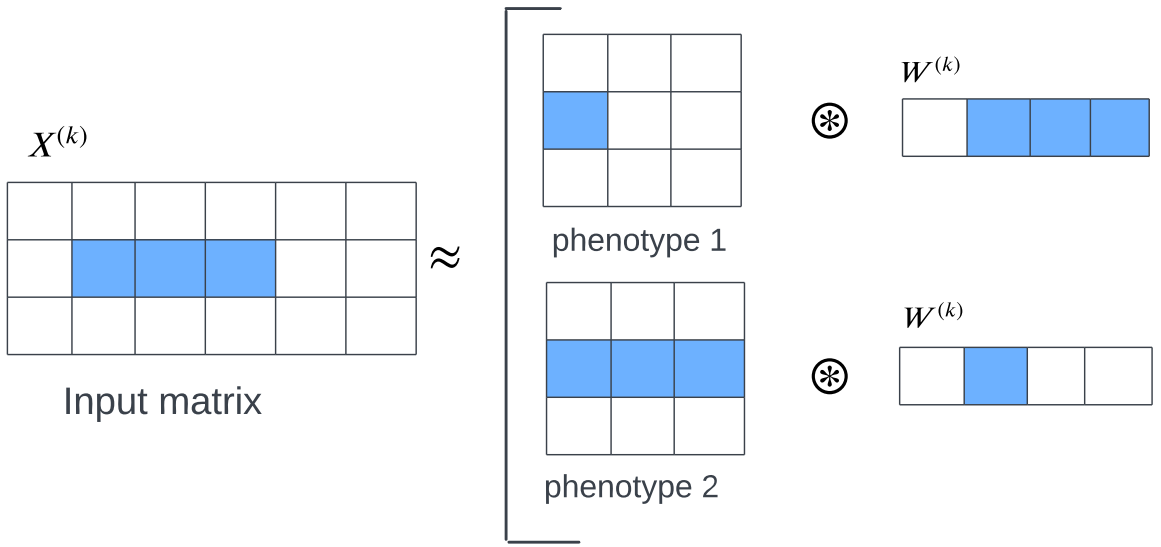
Jimeng Sun, Dacheng Tao, and Christos Faloutsos, *Beyond streams and graphs: dynamic tensor analysis*, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 374–383.

Jingyuan Wang, Fei Gao, Peng Cui, Chao Li, and Zhang Xiong, *Discovering urban spatio-temporal structure from time-evolving traffic networks*, Asia-pacific web conference, Springer, 2014, pp. 93–104.

Kejing Yin, Ardavan Afshar, Joyce C Ho, William K Cheung, Chao Zhang, and Jimeng Sun, *Logpar: Logistic parafac2 factorization for temporal binary data with missing values*, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1625–1635.

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 (2019), no. 01, 1246–1253.

Régularisation de Non-Succession de Phénotypes



→ Ajoute un terme de pénalité à la loss lorsqu'il y a plusieurs phénotypes identiques sur la même fenêtre

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